

HybridGNN: a graph neural network approach for human miRNA–disease association prediction

Basharat Ahmad¹, Muhammad Hammad Musaddiq², Sebu Aboma Temesgen¹,
Grace-Mercure Bakanina Kissanga¹, Huma Fida¹, Hao Lin¹, Ye-Chen Qi^{1,*}

¹School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China

²School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

*Corresponding author. School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China.

E-mail: ycqi@uestc.edu.cn

Associate Editor: Jianlin Cheng

Abstract

Motivation: MicroRNAs (miRNAs) are small non-coding RNAs, typically 18–24 nucleotides in length, that play a pivotal role in RNA silencing and the post-transcriptional regulation of gene expression by targeting messenger RNAs (mRNAs). Dysregulation of these miRNAs has consistently been implicated in the onset and progression of a variety of complex human diseases.

Results: In this study, we propose a novel *HybridGNN* model that integrates a Graph Convolutional Network (GCN), a Graph Attention Network (GAT), and Matrix Decomposition with Matrix Factorization (MDMF) to predict potential miRNA–disease associations (MDAs). We incorporate five types of similarity in which three are derived from miRNAs and two are derived from diseases, to comprehensively explore and optimize multi-source feature information. The complementary interactions among these modules also help to mitigate the oversmoothing problem. The model utilizes neighboring nodes in a heterogeneous network to generate node embeddings via a message-passing mechanism. To improve computational efficiency, we employ a mini-batch gradient descent approach that partitions the graph into smaller sub-graphs, thereby enhancing the model's accuracy, speed, and scalability. As a result of these advanced techniques, HybridGNN achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.9715 using a dot-product classifier, outperforming several existing methods and underscoring its potential as a robust and accurate tool for predicting MDAs.

Availability: Code and data are freely available at <https://github.com/mbasharatahmad/HybridGNN-miRNA-disease/>

1 Introduction

The human genome comprises of vast genetic material but only 1.5% of genes are responsible for coding proteins and majority are transcribed into non-coding RNAs (ncRNAs). These ncRNAs play pivotal roles in fundamental biological process, including tissue differentiation and regulation of cell cycle (George *et al.* 2024). These RNAs are commonly classified based on the transcript length, those shorter than 200 nucleotides are referred to as small ncRNAs, whereas transcripts longer than 200 nucleotides are termed long non-coding RNAs (lncRNAs) (Liu *et al.* 2024b). Among small ncRNAs, microRNAs (miRNAs) represent a well-known characterized class of single-stranded, evolutionarily conserved molecules, typically comprising 18–24 nucleotides (Liu *et al.* 2024a, Saw and Song 2025). miRNAs regulate gene expression at the post-transcriptional level by binding to target

messenger RNAs (mRNAs), resulting in either mRNA degrading or translation inhibition (Wei *et al.* 2021, Bakhti and Latifi-Navid 2022, Lin *et al.* 2025, Luo *et al.* 2025). Numerous miRNAs are associated with various diseases including, miR-143/145 family is linked to hypertension, miRlet-7 and miR-103/107 regulate of glucose metabolism; miR-29 is involved in diabetes; and miR-17–92 cluster, contribute to cancers development (Chistiakov *et al.* 2016, Wojciechowska *et al.* 2017, Saliminejad *et al.* 2019).

Identification of disease-associated miRNAs can facilitate pathological classification of diseases and contribute to the development of individualized treatment strategies (Tang *et al.* 2018, Mamatkarimov *et al.* 2025, Popa *et al.* 2025, Qi *et al.* 2025, Wang *et al.* 2025b, Xie *et al.* 2025). Experimental approaches can accurately identify ncRNA–disease associations but are constrained by high cost and time-consumption (Lv *et al.* 2024, Qiao *et al.* 2024, Sheng *et al.* 2024). Recently significant advances

have been made in computational methods for identifying potential miRNA–disease association (MDAs) (Chen *et al.* 2025b) based on network biology, mainly including similarity-based and machine learning-based approaches (Perez-Iratxeta *et al.* 2005, Ai *et al.* 2024, Wang *et al.* 2023, Yizheng Wang *et al.* 2024). Computational models such as, DeepMDA which is proposed by Fu *et al.* (2017) that predict MDAs by constructing heterogeneous network and extracting miRNA and disease features using the Gaussian Interaction Profile Kernel (GIPK) similarity (Fu and Peng 2017). Similarly, Ha and Park (2021) introduced (MLMD) that predict MDAs and capture miRNA–miRNA and disease–disease association. In 2018, Chen *et al.* proposed IMCMDA which uses inductive matrix completion to predict MDAs by integrating Directed Acyclic Graph (DAG) based disease similarity and GIPK based miRNA similarity (Li *et al.* 2023). In the subsequent study, the authors employed bipartite recommendation and a rating integrated bipartite network projection method to predict MDAs (Chen *et al.* 2018b).

Furthermore, various approaches based on MDAs similarity have been explored (Zeng *et al.* 2016). A novel computational method was proposed that integrates microRNA-based similarity inference, phenotype-based similarity inference and network-consistency-based inference for predicting MDAs (Wang *et al.* 2010). The researchers utilized a DeepWalk-based graph embedding method (DWMMDA) to predict MDAs by extracting low dimensional features from miRNAs and disease network (Ha 2025). Additionally, a deep neural network was employed to identify MDAs, while MNEMDA, applied meta-path-based network embedding to extract features from heterogeneous

network (Wang *et al.* 2025a). Recently, Chen *et al.* (2025), proposed a hybrid framework combining principal neighborhood aggregation and graph attention networks, to capture complex node feature and improve MDAs prediction (Chen *et al.* 2025a).

To address the aforementioned limitations, we introduced a novel HybridGNN integrating MDMF and GCN–GAT, for MDAs prediction. The heterogeneous network was constructed based on miRNA, disease and their association using five similarity network and fused adjacency matrices. Three modules were used to extract sequence and topological feature of miRNA–disease, while a fully connected layer was employed to fuse the extracted feature and prevent overfitting and generate the final prediction. The complete HybridGNN model process was depicted in Fig. 1.

2 Methods

2.1 Dataset collection and processing

The data set was obtained from Human microRNA disease database (HMDD v4.0) it comprises of curated 53 531 miRNA–disease association (MDAs). After removal of duplicate entry, 728 miRNAs and 175 diseases, encompassing 18 738 known miRNA–disease associations were considered. Duplicate miRNA–disease pairs were systematically removed to construct a non-redundant dataset, in which each association is represented only once. The preprocessing step ensures data consistency and provides a reliable foundation for training and evaluation of the

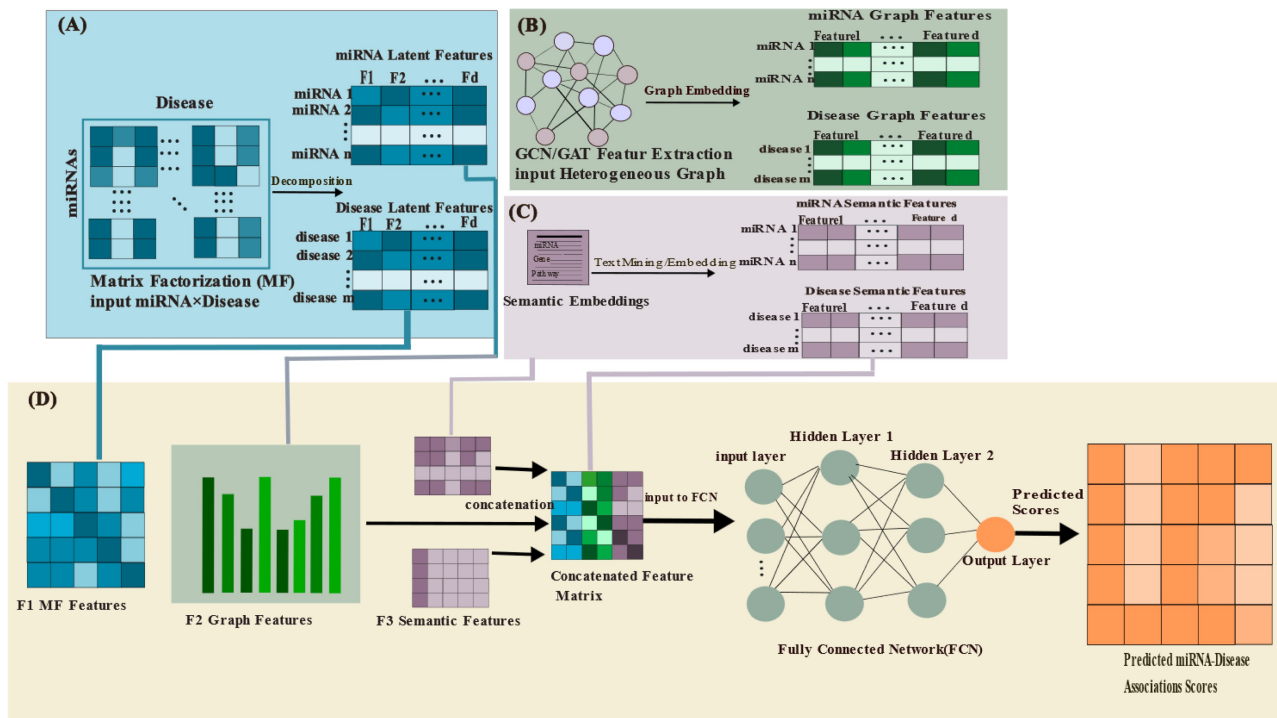


Figure 1 Architecture of the proposed HybridGNN model for predicting miRNA–disease associations. (A) Data preparation, including the construction of miRNA–disease association matrices and the integration into a heterogeneous network. (B) Feature extraction using MDMF, GCN, and GAT modules (C) feature representation and (D) multimodal fusion and prediction of miRNA–disease associations through a fully connected network.

HybridGNN model (Cui *et al.* 2024). Additionally, to further assess the generalization ability of the proposed model we collected the IMCMDA benchmark dataset consisting of 5430 MDA introduced by Chen *et al.* (2018a).

2.2 miRNA association

miRNAs association play crucial role genes regulation and cellular homeostasis and their disruption can lead to various diseases therefore, computational modeling of MDA is essential, in this study the miRNA similarity matrix was constructed by averaging sequence, GIPK and functional similarities. Accordingly, the sequence similarity matrix nS^{seq} was calculated using the Equation (1) (Sun *et al.* 2022).

$$nS^{\text{seq}} = \frac{\langle s_i s_j \rangle}{\|s_i\|^2 + \|s_j\|^2 - \langle s_i s_j \rangle} \quad (1)$$

where $\langle s_i s_j \rangle$ holds the dot product values of miRNAs 4-mers sequence, vector of miRNA s_i and s_j which reflects the Euclidean distance between the two miRNAs. Similarly, $\|s_i\| = \sqrt{\langle s_i s_i \rangle}$. It exhibits the Euclidean distance between the two miRNAs.

$$\langle s_i s_j \rangle = \sum_{k=1}^n s_{i,k} \times s_{j,k} \quad (2)$$

Where k represents the index of feature (from 1 to n), where n is the total number of features used to represent each miRNA. $s_{i,k}$ represents the k -th feature value of miRNA i , and $s_{j,k}$ represents the feature value of miRNA j .

2.3 miRNA–disease association

The adjacency matrix is a 2D array in which row indicate miRNAs and columns represents diseases. The matrix element $A(i, j)$ indicates whether an association exists between the i -th miRNA and j -th disease, a value of 1 denotes a confirmed association, while 0 indicates an unknown association.

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,n} \\ \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots a_{m,n} \end{bmatrix} \quad (3)$$

GIPK is used to capture similarity between miRNAs based on their disease association patterns. GIPK is calculated based on the following Equation (4) (Sun *et al.* 2024).

$$nS^{\text{GIPK}}(n_i, n_j) = \exp\left(-\lambda_n |A(n_i) - A(n_j)|^2\right) \quad (4)$$

where $A(n_i)$ and $A(n_j)$ represent the i -th and j -th row vector of the adjacency matrix A and λ_n is the Kernel coefficient, which is specified in Equation (5).

$$\lambda_n = \frac{1}{\frac{1}{N} \sum_{l=1}^N |A(n_l)|^2} \quad (5)$$

where N is the total miRNA number and $A(n_l)$ is the l -th row vector of the adjacency matrix A .

We used Wang *et al.* MISIM (microRNA Functional Similarity) method (Wang *et al.* 2007) to calculate miRNA functional similarity based on their shared disease association.

$$nS^{\text{fun}}(n_1, n_2) = \frac{\sum_{i=1}^p S(\text{de}_{i1}, \text{DE}_2) + \sum_{j=1}^q S(\text{de}_{2j}, \text{DE}_1)}{p + q} \quad (6)$$

In the equation, p and q denotes number of diseases linked to miRNAs n_1 and miRNA n_2 respectively, and $S(\text{de}, \text{DE})$ measure the semantic similarity between a given disease de and the set of disease DE .

$$S(\text{de}, \text{DE}) = \max_{1 \leq i \leq j} (S(\text{de}_i, \text{de}_j)) \quad (7)$$

$S(\text{de}_i, \text{de}_j)$ represent semantic similarity between disease i and j .

The integrated miRNA similarity is calculated as the average of sequence, GIPK and functional similarity as shown in Equation (8).

$$nS = \frac{nS^{\text{seq}} + nS^{\text{GIPK}} + nS^{\text{fun}}}{3} \quad (8)$$

2.4 Disease association

Diseases associations were obtained by averaging DAG-based semantic similarity and GIPK similarity derived from disease association matrices. In the DAG based semantic similarity (Nilsson *et al.* 2021), diseases sharing common ancestor are considered more closely related as shown in Equation (9).

$$DS^{\text{sem}}(d_i, d_j) = \frac{\sum_{t \in T_i \cap T_j} (S_{d_i}(t) + S_{d_j}(t))}{\sum_{t \in T_i} S_{d_i}(t) + \sum_{t \in T_j} S_{d_j}(t)} \quad (9)$$

T_i and T_j denote disease set for d_i and d_j , sharing ancestor disease in DAG structure. $S_{d_i}(t)$ indicates the semantic contribution of ancestor t to disease d_i , computed in following Equation (10).

$$\begin{cases} S_{d_k}(t) = 1, & \text{if } d_k = d_j \\ S_{d_k}(t) = 1, & \text{if } d_k = d_j \\ & \text{otherwise} \end{cases} \quad (10)$$

Where S denotes the semantic score of disease, d_k denotes specific disease node under the consideration, and t is the node in the DAG, d_j represents the target disease node for which the semantic score is calculated, refers to a child node of t , the max function selects the highest weighted child score and ∂ is a decay factor set to 0.5.

The GIPK similarity among various diseases is calculated based on the following Equation (11).

$$DS^{GIPK}(d_i, d_j) = \exp\left(-\lambda_d |A(:, d_i) - A(:, d_j)|^2\right) \quad (11)$$

where $A(:, d_i)$ and $A(:, d_j)$ hold the i -th and j -th column vectors of the adjacency matrix A and λ_d represents the Kernel coefficient which is specified in Equation (12).

$$\lambda_d = \frac{1}{\frac{1}{D} \sum_{l=1}^D |A(:, d_l)|^2} \quad (12)$$

where D represents the total number of diseases and $A(:, d_l)$ represents the l -th column vector of adjacency matrix A . Then, average association among diseases is calculated as shown in:

$$DS = \frac{DS^{sem} + DS^{GIPK}}{2} \quad (13)$$

2.5 Constructing of miRNA-disease heterogeneous network

A heterogeneous graph comprises of multiple nodes and edges, each representing unstructured data sources. Following the work of Chen et al. (2022), a heterogeneous graph is defined as:

$$G = (V, E, D, R) = \begin{bmatrix} SS & A^T \\ A & DS \end{bmatrix} \quad (14)$$

where V and E represent the sets of nodes and edges, the D and R signify node and edge types. The heterogeneous network integrates SS (miRNA similarity), A (miRNAs-disease association), A^T (representing reverse association), and DS (diseases similarity). following Chen et al. (2022), a network embedding function $f: V \rightarrow X \in R^{|V| \times b}$ (where $b \ll |V|$), maps node into a low-dimensional space effectively modeling both known and unknown miRNA-disease interactions, in Equation (14).

This hypergraph learning, capture high order relations, it may produce inaccuracies when applied to limited or noisy heterogeneous data (Kesimoglu and Bozdog 2024). To address this, we designed a hybrid multimodule framework combining Matrix Decomposition and Matrix Factorization (MDMF), Graph Convolutional Networks (GCN), and Graph Attention Networks (GAT) module.

2.6 Feature extraction

2.6.1 Matrix factorization-based feature extraction (MDMF)

We employed matrix decomposition and matrix factorization (MDMF) approach to extract latent representation from the MDAs matrix A . First, matrix factorization approximates the original matrix as:

$$A \approx WH \quad (15)$$

where, W represents latent feature matrix of miRNAs and H indicates latent feature matrix of diseases. This step

projects the sparse and high-dimensional association matrix into a shared low-dimensional latent space. Based on this low-rank representation, the matrix is further expressed as:

$$A = BC \quad (16)$$

where, B denotes basic components matrix, and C represents the reconstruction coefficients matrix. The matrices B and C are derived from and structured upon the latent space learned through W and H . Thus, the second formulation is a refined reconstruction of the first factorization, reorganizing the learned latent features into basis components and coefficients to enhance pattern extraction and dimensionality reduction. The resulting refined embeddings (Feature1, F1) are then used as input to the HybridGNN model for downstream prediction.

2.6.2 Association network representation module based on GCN

We constructed a three layer of GCN-based module, that takes miRN-disease heterogeneous association network as input and using integrated miRNA and disease similarity matrices as features, with message passing as shown in Equation (17).

$$H_{l+1} = \sigma\left(\tilde{D}^{-(1/2)} \tilde{A}_h \tilde{D}^{-(1/2)} H_l W_l\right) \quad (17)$$

here $\tilde{A}_h = A_h + I_N$, where A_h is the adjacency matrix and I_N is identity matrix. $\tilde{D}$ denotes the degree matrix with diagonal element. $\tilde{D}_{ii} = \sum_j \tilde{A}_{hj}$. H_l is feature matrix at the l -th layer, W_l the trainable weight matrix and $\sigma(\cdot)$ displays the Relu activation function. This module produces the feature matrix $F2$.

2.6.3 Topological structure representation module based on GAT

To dynamically weights nodes importance, a 3-layer GAT architecture with a single attention head per layer were used (Momanyi et al. 2024a, 2024b, Dai et al. 2025).

$$h_i = \sigma\left(\sum_{j \in N_i} \alpha_{i,j} W_l H_j\right) \quad (18)$$

Where h_i represents the feature vector of node i at the l -th layer, N_i denotes the set of neighbor's nodes connected to node i in heterogeneous network, W_l is the trainable weight matrix at l -th layer attention coefficients $\alpha_{i,j}$ is the normalized attention score as shown in Equation (19).

$$\alpha_{i,j} = \text{softmax}(e_{i,j}) = \frac{\exp(e_{i,j})}{\sum_{k \in N_i} \exp(e_{i,k})} \quad (19)$$

Where $e_{i,j}$ denotes self-attention score, this matrix captures higher order relational dependencies for improved link prediction, and outputs the feature matrix $F3$.

2.7 Predicting miRNA-disease association by fully connected network

MDAs are predicted using fully connected network, where extracted miRNA-diseases feature are integrated through a fusion mechanism as follow.

$$F = \frac{\alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3}{\alpha_1 + \alpha_2 + \alpha_3} \quad (20)$$

where (F_1, F_2, F_3) are feature vector from different module. The coefficient α_i determines the feature weight. All modules were assigned equal weight ($\alpha_i = \alpha$), reflecting their equal significance.

After feature extraction, outputs from modules were linearly transformed and fed into a fully connected network to predict MDAs (Yang et al. 2023).

$$y^* = \text{sigmoid}(W.F + b) \quad (21)$$

where W and b represent learnable weight matrix, the model was trained using cross-entropy loss function, and Adam optimizer used to minimize loss function.

$$L = -\frac{1}{N} \sum_{i=1}^N y_i \log y_i + (1 - y_i) \log(1 - y_i) \quad (22)$$

where N and y_i represent the number of ncRNA-disease association pairs and true labels.

Our proposed HybridGNN synergistically integrates MDMF, GCN and GAT to jointly capture structural graph topology and semantic association patterns in the miRNA-disease network.

$$A \approx UV^T \quad (23)$$

where $U \in R^{m \times l}$, $V \in R^{d \times l}$ Here, l is the latent dimension (e.g., $l=42$). The MDMF loss function combines reconstruction error and regularization. The loss function is explained below:

$$\mathcal{L}_{\text{MDMF}} = \|A - UV^T\|_2^2 + \lambda(\|U\|_2^2 + \|V\|_2^2 + \|UU^T - S_m\|_2^2 + \|VV^T - S_d\|_2^2), \quad (24)$$

where λ (e.g., 0.0121) balanced reconstruction and regularization strength.

2.8 Training protocol

The heterogeneous graph was constructed using PyG. miRNAs and diseases presented as distinct node types, and miRNA-disease associations presented as relational edges. To ensure rigorous training and evaluation, we employed the RandomLinkSplit function to split edges (Betkier et al. 2023). (RandomLinkSplit is a dataset splitting tool specifically designed for the graph link prediction task in the PyTorch Geometric (PyG) library. The core of link prediction is to predict “whether there is an edge between two nodes”, so the core of the process is splitting edges.) The edges were split into 80% training, 10% validation, and 10% test sets, to evaluate the model

ability to predict unseen association. Nodes and edges were partitioned differently because due to their distinct roles in the learning. A disjoint-train-ratio parameter of 0.3, assigned 70% of node for training and 30% for supervision. Node splitting controls message passing and embedding learning without evaluating prediction, whereas edge splitting provides supervision for link prediction; thus, they serve complementary but distinct roles.

Regarding the setting of the negative sampling ration, negative edges are dynamically sampled at a 20% (0.2 ratio) of the total negative association during training, to provide sufficient negative supervision and model generalization. During evaluation, the number of negative edges is fixed at 10% (0.1 ratio). The model was trained using Binary Cross-Entropy with Logits Loss, optimized by Adam optimizer (learning rate of 0.0076) and weight decay (1.75×10^{-5}) (Huang et al. 2022). The mini-batch subgraph loaders employed to improve computational efficiency and scalability. Hyperparameter were optimized using optuna tuning with 100 trials, selecting the best configuration (e.g. hidden size 117) based on the validation AUC (Pham et al. 2023, Harun-Or-Roshid et al. 2024, Pham et al. 2024, Dhanka et al. 2026). Early stopping was applied with a patience of 20 epochs to prevent overfitting. The model was trained using the following setup. Loss Function: Binary cross-entropy with logits as shown in Equation (25):

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (25)$$

where y_i is the true binary label, and $\hat{y}_i$ represents the predicted probability of sample i . This equation is particularly suitable for the binary classification task and effectively handle prediction express as logits.

Furthermore, the residual connection was added to prevent vanishing gradients and retain original node features by merging GCN outputs with the transformed outputs as shown in Equation (26).

$$H_{\text{out}} = H_{\text{GCN}} + W_r H_{\text{in}} \quad (26)$$

where H_{GCN} is the GCN output, H_{in} is the input feature matrix, W_r is a learnable linear projection matrix ensuring stable training and retains key feature. The details of hyperparameters adjustment for the HybridGNN model are also shown in the Table 1.

2.9 Accuracy metrics

To evaluate model performance, multiple metrics employed such as accuracy (ACC), Precision (Pre), Recall (Rec), F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Area Under the Precision-Recall curve (AUPRC). AUC-ROC quantifies the model’s overall discriminative ability by measuring the trade-off between true positives rate (TPR) and against false positive rate (FPR), whereas AUPRC emphasizes precision-recall behavior, and is particularly informative for imbalanced datasets. In Addition, ACC, Pre, Rec, and F1-score were computed to provide a balanced assessment of classification performance. All evaluation metrics, were computed in

Table 1 Hyperparameter configuration and optimal settings of the HybridGNN model.

Parameter	Search space	Optimal value
Hidden Dimension	[64, 256]	117
Learning Rate	$[10^{-5}, 10^{-2}]$	0.0076
Dropout	[0.1, 0.5]	0.4493
Number of Heads	[1, 16]	12
Number of Layers	[2, 4]	2
PCA Dimension	{16, 32, 64}	32
Latent Dimension	[16, 64]	42
Output Channels	[8, 64]	22
λ_{reg}	[0.01, 0.1]	0.0121
Weight Decay	$[10^{-5}, 10^{-3}]$	1.75×10^{-5}

Equation (27) (Liu *et al.* 2019, Liu *et al.* 2024a, Meng *et al.* 2024, Shao *et al.* 2025, Xing *et al.* 2025, Huang *et al.* 2025).

$$\left\{ \begin{array}{l} ACC = \frac{TP + TN}{TP + TN + FP + FN} \\ Rec = \frac{TP}{TP + FN} \\ Pre = \frac{TP}{TP + FP} \\ F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \end{array} \right. \quad (27)$$

where TP represents the number of correctly identified positive samples, TN denotes the number of correctly classified negative samples, FP refers to the number of falsely predicted positive samples, and FN indicates the number of misclassified negative samples.

3 Results

3.1 HybridGNN model results

The dataset was obtained from HMDD v4.0 database, The dataset captures pairwise similarities between miRNAs and diseases. Analyzing their distribution reveals functional relationships, offering valuable insights into biological interactions. We calculated the similarity values for miRNAs and disease, explored the distribution in Fig. 2. The verticals axis represents the number of samples within the specific similarity ranges, while the horizontal axis explores the range of similarity values of miRNAs and disease. The miRNAs-miRNAs and disease-disease similarity scores mostly range between 0 and 0.8. This similarity results were higher compared with the similarity distribution reported in previous study. For instance, the SAGESDA model reported low snoRNA-snoRNA and disease-disease similarity values with most score concentrated between 0.0 and 0.2 (Momanyi *et al.* 2024b). The higher similarity score in the current study can be attributed to the fact that we considered the average of three similarity scores measures for miRNA-miRNA pairs and average of two similarity measures for disease-

disease pairs, which more effectively capture the underlying biological relationships.

Building upon this foundation, we developed HybridGNN, model for predicting MDAs by integrating three heterogeneous networks and a dot-product classifier. The model, trained with mini-batch subgraph loaders, ensures scalability, efficiency, and reduces over-smoothing. HybridGNN employs multi-head attention in GAT for informative neighbor aggregation, with a penalty term (λ) to avoid overfitting. Although attention increases time complexity, it improves edge-specific weight learning. Shared weights across networks provide richer node embeddings. Cosine similarity in the dot product quantifies miRNA-disease alignment. HybridGNN can compute embeddings for new miRNAs or diseases without retraining, offering enhanced scalability, accuracy, and adaptability.

To visualize miRNA-disease associations, we selected the sub-network of heart-related diseases, because these diseases exhibited highest number of miRNA association (Fig. 3). The highest miRNA frequencies refer to the number of known and validated miRNA-disease associations recorded for each disease in the integrated heterogeneous network. The bipartite sub-graph included 820 pairs linking 324 miRNAs (green) and 8 diseases (red). Central clustering of myocardial infarction and ischemic stroke indicated strong connectivity, while peripheral miRNAs showed disease-specific patterns.

Based on constructed miRNA-disease heterogenous networks, the model performance was evaluated by comparing the predicted association scores with ground-truth data from the evaluation dataset, encompassing both positive and negative associations. To ensure robustness and stability five-fold cross-validation (FF-CV) was employed on proposed dataset. As shown in Fig. 4A and C, HybridGNN achieves AUC values above 0.97 and Average precision values around 0.97 across all folds on proposed dataset, demonstrating excellent discriminative ability, stable performance, and strong capability in identifying true miRNA-disease associations, even under class-imbalance conditions. The generalization ability of HybridGNN was evaluated on independent IMCMDA dataset. As shown in Fig. 4B and D, model maintains strong predictive performance, achieving AUC values around 0.96 and AP values above 0.95 across five folds. The ROC and PR curves remain smooth and stable, indicating consistent predictive behavior across different data splits. Compared with previously reported methods on the IMCMDA dataset, which achieved an AUC of 0.8845 and an AUPR of 0.8832, HybridGNN demonstrates a substantial improvement, achieving an AUC of 0.96, an accuracy of 0.89, and an F1-score of 0.89. These results underscore the robustness and effectiveness of the HybridGNN framework for miRNA-disease association prediction.

In this study, the cold disease setting refers to an evaluation scenario in which all associations related to certain diseases are completely excluded from the training set and used only for testing. This setting evaluates the model's ability to predict miRNA associations for previously unseen diseases. Similarly, the cold miRNA setting refers to a scenario where all associations corresponding to specific miRNAs are removed from the training data and reserved for testing. This strategy assesses the model's capability to generalize to novel miRNAs without prior association information as shown in Table 2.

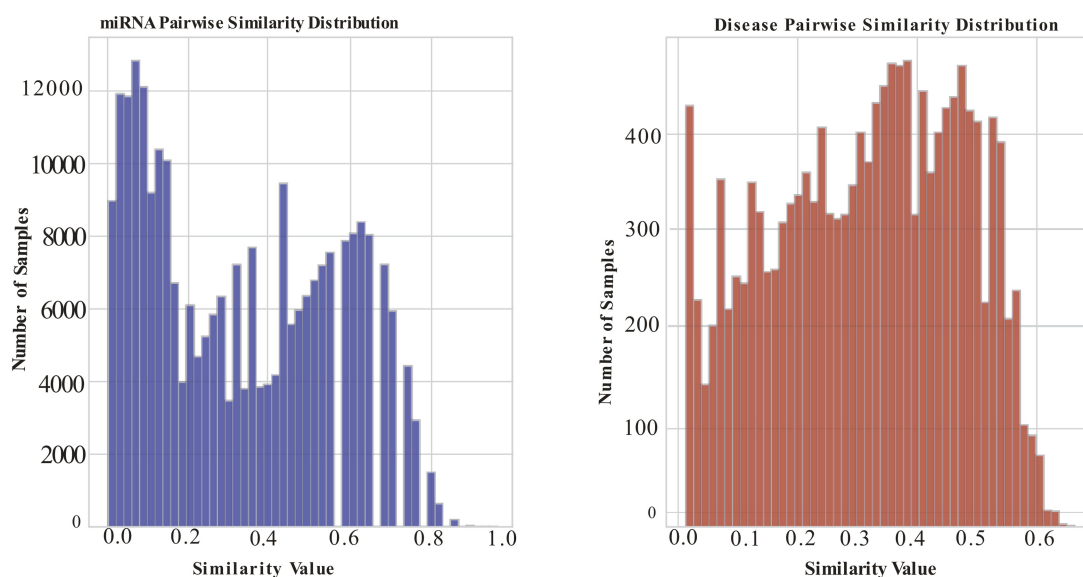


Figure 2 Distribution of pairwise similarity values between miRNAs and diseases matrices. A pair correlation that is weaker and stronger indicated by the values closer to 0 and 1.

3.2 Ablation study

To investigate the contribution of each architectural component and feature type, we conducted an ablation study of HybridGNN model. The performance was evaluated using AUC evaluation. As shown in Table 3, the full HybridGNN model, which integrates a hybrid GCN-GAT architecture, GIP and MDMF features, similarity-based edges, and residual connections, achieves the best overall performance with AUCs of 0.9765. Removing either GAT and GCN leads to noticeable performance degradation. The GAT-only variant shows a slight drop in AUC (0.9708) indicating that attention alone is insufficient to capture all topological patterns. Similarly, the GCN-only variant exhibits lower Random AUC (0.9386), demonstrating that combining GCN and GAT enables more expressive node representations. MDMF features are also important, as their removal leads to a noticeable drop in performance as evidenced by decreased in AUC 0.9285. This indicates that MDMF features significantly enhance association prediction. The baseline GCN and GAT models without similarity edges and MDMF features achieve the lowest AUC scores across all evaluation settings. In particular, Baseline-GAT yields an AUC of only 0.8951, confirming that the proposed enhancements feature fusion, similarity edges, and hybrid architecture are essential for robust performance. Overall, the ablation results demonstrate that each component of the proposed HybridGNN contributes positively to performance, with the hybrid GCN-GAT architecture and MDMF features playing particularly critical roles in improving generalization under different evaluation setting.

3.3 Comparative performance evaluation with previous methods

We compare HybridGNN model with several state-of-the-art methods, in addition to quantitative metrics, receiver operating characteristic (ROC) and precision-recall (PR) curves are plotted to provide a comprehensive and visual assessment of predictive

performance. The comparative IMCMA methods (Chen *et al.* 2018a) employs neural inductive matrix completion framework to infer potential miRNAs-disease association (MDAs). GCAEMDA (Li *et al.* 2021a) utilized graph convolutional autoencoder to model association patterns and MINIMDA (Lou *et al.* 2022) constructs a similarity network by integrating multi-source miRNA information and capture high-order domain features followed by a multilayer perceptron for associations prediction. MTLMDA (He *et al.* 2023) applies a multi-task learning strategy to predict potential MDAs. While VMTMDA (Huang *et al.* 2021) model employed the multi-view, multi-task approach, to predict MDAs. AMHMDA (Ning *et al.* 2023) incorporates graph convolutional networks with attention mechanisms to predict MDAs. furthermore, MDformer (Dong *et al.* 2023) model leverages a transformer-based architecture and combined with meta-path information to predict the MDAs. HGTMADA (Lu *et al.* 2024) integrates graph convolutional attention with a transformer encoder to predict MDAs scores. Finally, GONNMADA (Zeng *et al.* 2025) applies graph neural network to the heterogeneous biological molecular graph, for effective MDAs. In Fig. 5, the comparison highlights that HybridGNN model is superior from all models. These complementary metrics allowed for a more holistic evaluation of the model's predictive performance. Its advantage comes from dynamic embedding updates and advanced feature learning. However, HybridGNN integrates MDMF with PCA to fuse heterogeneous features and remove redundancy. GNN layers with GAT assign dynamic node-specific weights for refined feature processing. The findings indicate that the proposed HybridGNN approach consistently outperforms from existing methods, providing compelling evidence of its robustness and superiority across multiple evaluation criteria.

3.4 Impact of data split strategies and stratified performance analysis

To evaluate data leakage and strengthen confidence in model performance, we employed three distinct data-split strategies.

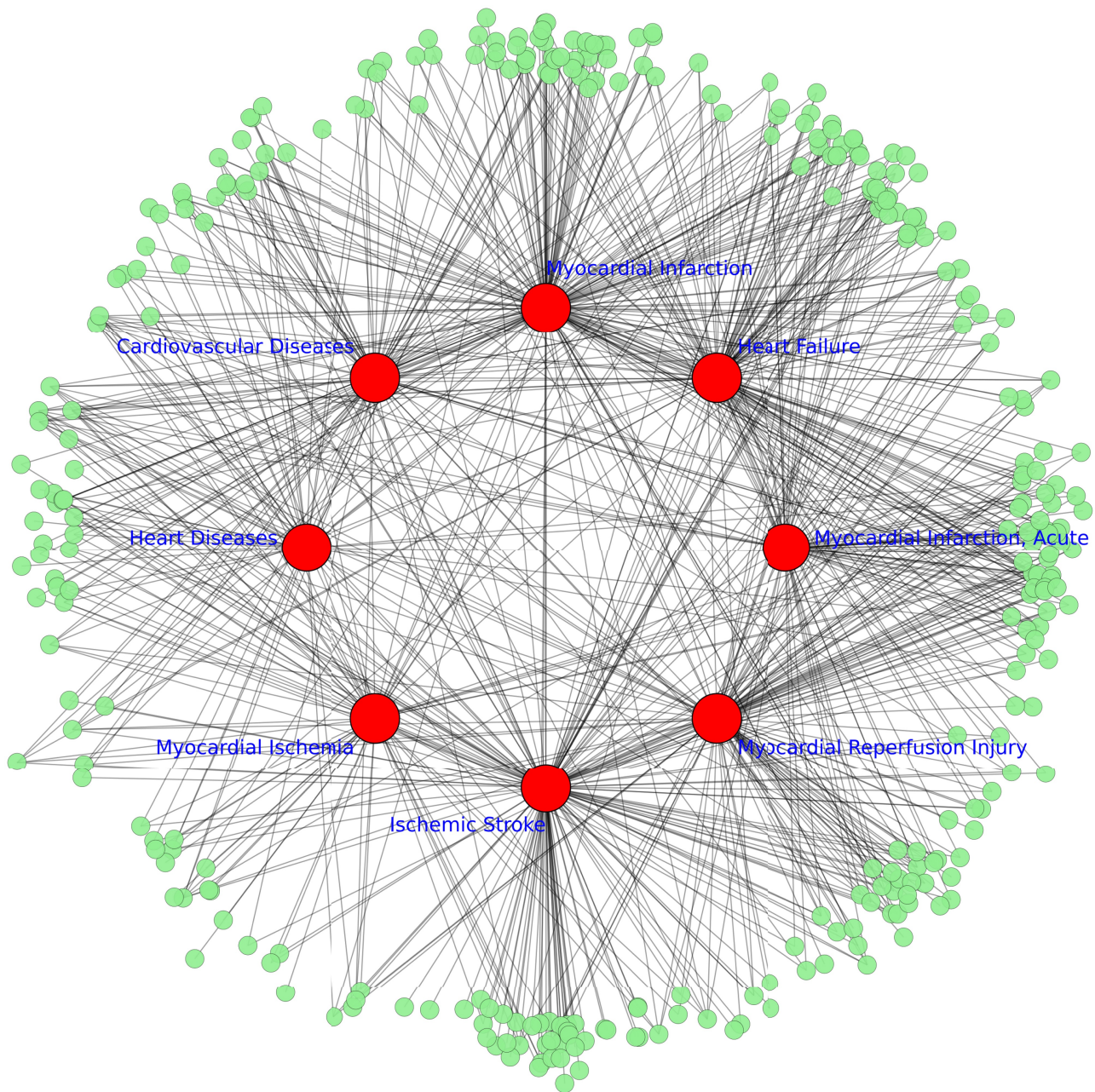


Figure 3 Illustration of the heterogeneous graph network showing two node types miRNA (green) and diseases (red) connected by undirected links that enable bidirectional message propagation and effective feature aggregation in HybridGNN model.

These includes a random split using 20% of miRNA-disease pairs, a 15% cold-disease split in which disease entities in the test set do not appear during training, and a 15% cold-miRNA split in which miRNA entities in test set are unseen during training. The results of these splitting strategies are presented in [Table 4](#) based on 5-fold cross-validation performance metrics.

Moreover, we conducted a stratification analysis based on seven disease categories to assess the generalization ability of HybridGNN and its performance analysis. [Table 5](#) shows category dependent performance variations across evaluation settings. While random split yield uniformly high performance, cold-disease split more severely affect genetic/syndrome, neurological disorder and other diseases. Fibrotic diseases

generalize well under cold-start conditions, whereas Inflammatory and Cardiovascular diseases remain relatively stable. Performance under cold-miRNA splits is consistently high, indicating greater robustness to unseen miRNAs than to unseen diseases.

3.5 Case study

To validate the predictive ability of HybridGNN model in the prediction of MDAs, we have conducted case study on five diseases: Myocardial Infarction (MI), Ischemic Stroke (IS), Lung Neoplasms (LN), Breast Neoplasm (BN) and Fibrosis (F). These diseases are widely reported in the literature to be strongly regulated by

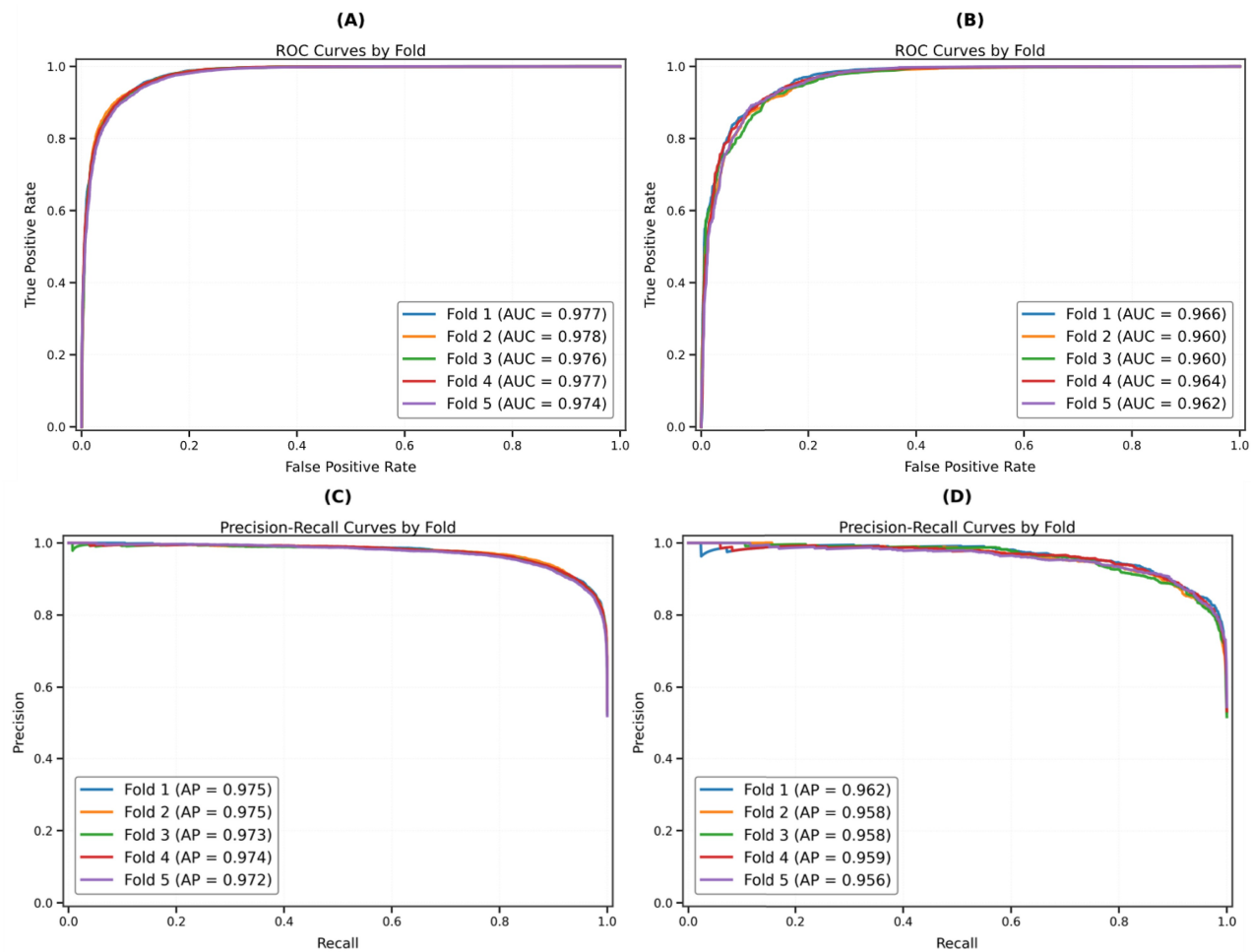


Figure 4 Evaluation results of the HybridGNN model on proposed dataset and IMCMDA dataset: (A) ROC curve on the proposed dataset, (B) ROC curve on the IMCMDA dataset (C) PR curve on proposed dataset and (D) represent PR curve on the IMCMDA dataset.

Table 2 HybridGNN model performance on an IMCMDA dataset.

Split mode	AUC (mean $\pm$ std)	AUPR (mean $\pm$ std)	ACC (mean $\pm$ std)	F1 (mean $\pm$ std)
random	0.9625 $\pm$ 0.0022	0.9586 $\pm$ 0.0018	0.8928 $\pm$ 0.0024	0.8953 $\pm$ 0.0021
cold_disease	0.9437 $\pm$ 0.0092	0.9352 $\pm$ 0.0072	0.8613 $\pm$ 0.0235	0.8684 $\pm$ 0.0180
cold_mirna	0.9583 $\pm$ 0.0042	0.9552 $\pm$ 0.0069	0.8866 $\pm$ 0.0077	0.8894 $\pm$ 0.0071

miRNAs, making them suitable benchmarks for miRNA-disease association prediction studies. Additionally, these diseases contain a relatively large number of experimentally validated miRNA associations in the Human microRNA Disease Database, ensuring sufficient data for meaningful case study validation and reducing bias caused by extremely sparse disease nodes and selected diseases cover cardiovascular disorders. This diversity allows comprehensive evaluation of HybridGNN across different biological mechanisms and disease categories rather than focusing on a single disease type (Ha 2011, Çakmak and Demir 2020). Table 6 presents the top 10 miRNAs with the highest predicted scores for each disease according our model prediction (Groenewegen et al. 2020). Previous studies have also confirmed the involvement of miRNAs in disease development. For instance, the first two MI-associated miRNA identified in the

Table 6, homosapiens (hsa-miR-29a) and hsa-miR-29b, were downregulated in response to myocardial infarction (van Rooij et al. 2008). While has-miR-195 was up-regulated during cardiac hypertrophy (Long et al. 2012). Similarly, hsa-miR-223 shown downregulated in acute myocardial infarction (Rangrez et al. 2013).

Among IS-related miRNAs, hsa-mir-21 is upregulated and plays regulatory role highlighting the importance of miRNAs in Ischemic stroke (Zhan et al. 2023). Hsa-mir-126 has been proposed as a potential biomarker for ischemic stroke (Zhang et al. 2023). Hsa-mir-34a was reported up-regulated in blood samples of IS patients (Liang et al. 2024). The hsa-miR-125 family has been implicated in the pathogenesis of ischemic stroke (Wang et al. 2021). Hsa-miR-206, has identified a potential diagnostic marker for ischemic stroke (Zhu et al. 2020).

Table 3 Ablation study component contribution analysis (AUC).

Variant	Model architecture	GIP Features	MDMF features	Similarity edges	Residual connections	AUC
FULL (HybridGNN)	GCN + GAT Hybrid	✓	✓	✓	✓	0.9765
GAT-Only	GAT Only	✓	✓	✓	✓	0.9708
Only-MDMF	GCN + GAT Hybrid	✗	✓	✓	✓	0.9705
GCN-Only	GCN Only	✓	✓	✓	✓	0.9386
No-MDMF	GCN + GAT Hybrid	✓	✗	✓	✓	0.9285
Baseline-GCN	GCN Only	✓	✗	✗	✓	0.9254
Baseline-GAT	GAT Only	✓	✗	✗	✓	0.8951

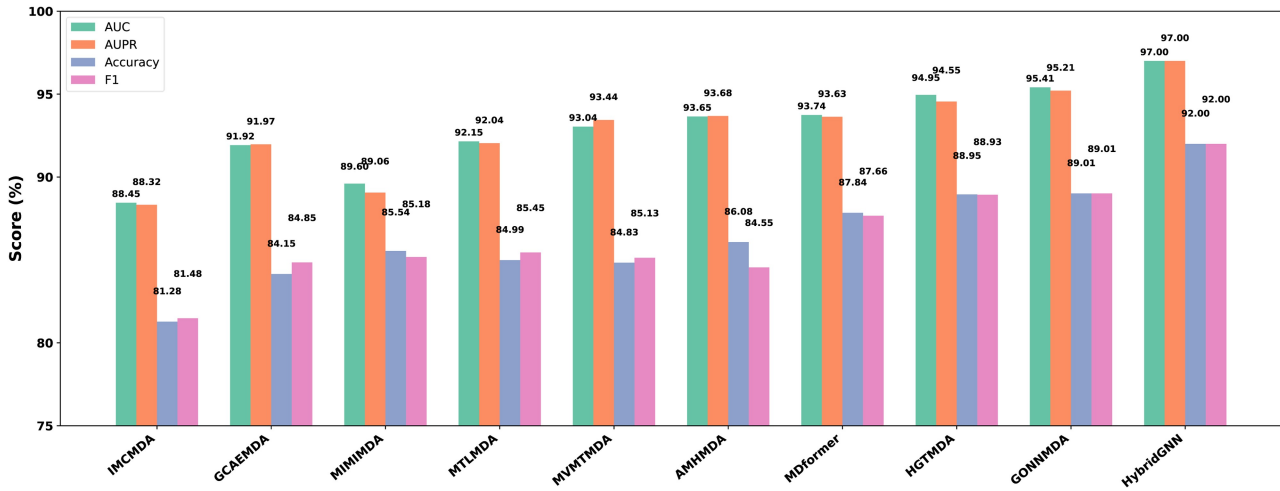


Figure 5 Comparative analysis of the different models with HybridGNN model based on the AUC, AUPR, Accuracy and F1.

Table 4 Evaluation metrics of HybridGNN across random and cold-start setting.

Metric	Random split	Cold-disease split	Cold-miRNA split	Performance pattern
Mean AUC-ROC (±SD)	0.9765 ± 0.0006	0.9365 ± 0.0054	0.9678 ± 0.0009	Random > Cold-miRNA > Cold-Disease
Mean PR-AUC (±SD)	0.9734 ± 0.0008	0.9397 ± 0.0160	0.9675 ± 0.0007	Consistent with AUC trends
Mean Accuracy (±SD)	0.9186 ± 0.0028	0.5756 ± 0.0279	0.9034 ± 0.0024	Cold-Disease shows classification difficulty
Mean Precision (±SD)	0.9013 ± 0.0057	0.5424 ± 0.0170	0.8966 ± 0.0097	Cold-Disease: low precision, high recall
Mean Recall (±SD)	0.9403 ± 0.0056	0.9809 ± 0.0079	0.9123 ± 0.0106	Cold-Disease: recall-oriented predictions
Mean F1-Score (±SD)	0.9204 ± 0.0026	0.6983 ± 0.0123	0.9042 ± 0.0022	Balanced performance in non-cold scenarios
Performance Stability	Highest (lowest std)	Moderate	High	Random split most consistent

For LN-miRNAs, hsa-miR-29b is downregulated in NSCLC tissues, and its reduced expression correlates with lymphatic metastasis (Pastuszak-Lewandoska et al. 2016). The overexpression of hsa-miR-23a/27a/24-2 cluster promotes non-small cell lung cancer

progression suggesting its role as an oncogenic regulatory unit in tumor growth (Fan et al. 2022). Our study revealed that hsa-miR-27b is associated with lung neoplasm, and participates in angiogenesis. Highlighting its potential therapeutic target in cancer

Table 5 HybridGNN performance across disease categories under different split.

Disease category	Random split	Cold-disease split	Cold-miRNA split	Performance pattern	Biological interpretation
Cancer	0.9780	0.9661	0.9770	Consistently high across all splits	Well-characterized associations with clear miRNA patterns; extensive training data enables robust generalization
Fibrotic	0.9643	0.9725	0.9798	Improves in cold scenarios	Distinct, focused miRNA involvement allows easier pattern recognition even with novel entities
Neurological	0.9791	0.9140	0.9778	Cold-disease vulnerable	Complex, multi-system involvement; associations less consistent across different neurological conditions
Genetic/Syndrome	0.9747	0.8904	0.9576	Most cold-disease sensitive	Rare conditions with limited training examples; heterogeneous molecular mechanisms
Inflammatory	0.9662	0.9246	0.9693	Stable across splits	Moderately characterized with consistent immune-related miRNA patterns
Cardiovascular	0.9762	0.9207	0.9730	Moderate cold-drop	Well-studied but with diverse pathological mechanisms affecting generalization
Other	0.9745	0.9117	0.9633	Significant cold-disease drop	Heterogeneous category; lack of clear unifying patterns hampers novel disease prediction

(Ding *et al.* 2017). The involvement of hsa-mir-22 in lung neoplasms is consistent with prior studies, suggesting its potential anti-angiogenic therapeutic target in non-small cell lung cancer (Gu *et al.* 2021). Notably, hsa-miR-375 is associated with lung neoplasm, and may act as an oncogene in certain types of cancer, as reported in previous studies (Nishikawa *et al.* 2011).

Among BN-miRNAs, hsa-miR-146a identified potential regulator linked to lung neoplasm consistent with its reported role in breast cancer proliferation (Chen *et al.* 2020). The hsa-miR-23a is associated with lung cancer, aligning with previous evidence that its down-regulation contributes to breast cancer (Chen *et al.* 2017). A strong association between hsa-miR-205 and breast cancer was predicted, while the previous results showing its overexpression in breast cancer (Xiao *et al.* 2019). Our result demonstrates that hsa-miR-23b act as an oncogene in breast cancer, and its predicted association is consistent with existing experimental evidence (Hannafon *et al.* 2019).

For F-miRNAs, hsa-miR-29b showed a strong association with fibrosis, existing literature confirmed that miR-29a and miR-29b are linked to renal fibrosis (Wang *et al.* 2012). The detection of the hsa-miR-150 in the current case study underscore its potential role in modulation of liver fibrosis (Tian *et al.* 2025). Consistently, miR-29 is significantly downregulated in human fibrotic disorders of multiple organs (Cushing *et al.* 2015). Our analysis reveals that miR-451 has a potential role in fibrosis aligning with previous studies that this miRNA identified as a promising biomarker for cardiac fibrosis (Jin 2021).

4 Discussion

Emerging evidence suggests that many miRNA genes are located within specific genomic regions associated with various

diseases, indicating their key regulatory role in disease progression (Zhong *et al.* 2024, Liu *et al.* 2025). Consequently, accurate prediction of miRNA-disease associations (MDAs) has become a key area of research. Although, numerous machine learning and deep learning models have achieved promising results, many still face challenges related to interpretability, scalability, and computational cost. Therefore, developing more efficient and interpretable computational frameworks remains essential for advancing disease diagnosis, prognosis, and therapeutic discovery (Sheng *et al.* 2024). In this study, we employed HybridGNN a heterogeneous network-based model for predicting potential MDAs. Three types of miRNA similarities and two types of disease similarities were integrated to construct unified similarities metrics which served as adjacency representations of miRNA-disease network. The model incorporated MDMF coupled with principal component analysis (PCA), GCN, and GAT to capture diverse information from network. A transition layer fused these multi-source features into a unified dimension and a fully connected neural network predicted the final association scores between miRNA-disease pairs.

Ablation analysis revealed that no single architectural configuration consistently dominates across all evaluation scenarios, emphasizing scenario-dependent optimality. While the GAT-only variant achieved performance comparable to the full model under random splits, GCN demonstrated stronger generalization in cold-disease settings, likely due to its simpler aggregation mechanism. In contrast, optimal performance in miRNA prediction required the full HybridGNN architecture, confirming the synergistic contribution of GCN, GAT, and feature fusion components. Furthermore, MDMF features were shown to capture distinct yet complementary similarity information, reinforcing the benefit of multi-source feature integration.

Table 6 HybridGNN model identified the top 10 most significant miRNA-disease associations.

MI-miRNAs	Evidence	IS-miRNAs	Evidence	LN-miRNAs	Evidence	BN-miRNAs	Evidence	F-miRNAs	Evidence
hsa-mir-29a	Confirmed (Shao et al. 2025)	hsa-mir-21	Confirmed (Xing et al. 2025)	hsa-mir-23a	Confirmed (Ning et al. 2023)	hsa-mir-146a	Confirmed (Groenewegen et al. 2020)	hsa-mir-29b	Confirmed (Zhang et al. 2023)
hsa-mir-29b	Confirmed (Shao et al. 2025)	hsa-mir-29b	Confirmed (Huang et al. 2021)	hsa-mir-29b	Confirmed (Huang et al. 2021)	hsa-mir-23a	Confirmed (van Rooij et al. 2008)	hsa-mir-150	Confirmed (Liang et al. 2024)
hsa-mir-195	Confirmed (Long et al. 2012)	hsa-mir-126	Confirmed (Ha 2025)	hsa-mir-27b	Confirmed (Dong et al. 2023)	hsa-mir-24	Confirmed	hsa-mir-375	Confirmed (Wang et al. 2021)
hsa-mir-143	Confirmed	hsa-mir-34a	Confirmed (Li et al. 2025)	hsa-mir-22	Confirmed (Lu et al. 2024)	hsa-mir-205	Confirmed (Rangrez et al. 2013)	hsa-mir-29	Confirmed (Zhu et al. 2020)
hsa-mir-223	Confirmed (Shao et al. 2025)	hsa-mir-125a	Confirmed (Lou et al. 2022)	hsa-mir-375	Confirmed (Zeng et al. 2025)	hsa-mir-23b	Confirmed (Zhan et al. 2023)	hsa-mir-451	Confirmed (Pastuszak-Lewandoska et al. 2016)
hsa-mir-144	Confirmed	hsa-mir-206	Confirmed (He et al. 2023)	hsa-mir-126	Confirmed	hsa-mir-192	Confirmed	hsa-mir-200a	Confirmed
hsa-mir-15b	Confirmed	hsa-mir-146b	Confirmed	Has-mir-101	Confirmed	hsa-mir-210	Confirmed	hsa-mir-122	Confirmed
hsa-mir-214	Confirmed	hsa-mir-204	Confirmed	hsa-mir-641	Confirmed	hsa-mir-138	Confirmed	hsa-mir-29b-1	Confirmed
hsa-mir-218	Confirmed	hsa-mir-9	Confirmed	hsa-mir-340	Confirmed	hsa-mir-146b	Confirmed	hsa-mir-374a	Confirmed
hsa-mir-7	Confirmed	hsa-mir-19a	Confirmed	hsa-mir-152	Confirmed	hsa-mir-556	Confirmed	hsa-mir-1202	Confirmed

**MI = Myocardial Infarction, IS = ischemic stroke, LN = Lung Neoplasms, BN = Breast Neoplasm, F = Fibrosis ** confirmed indicates that miRNA is associated with disease as predicted by model.

In recent years, multiple deep learning-based models have been proposed to enhance (MDA) prediction. Notably, IMCMDA (Chen *et al.* 2018a), employed an inductive matrix completion to integrates functional similarity, semantic similarity, achieving an AUC of 0.8034. GCAEMDA (Li *et al.* 2021) used graph convolutional autoencoder to learn embeddings from heterogeneous similarity network reporting an AUC of 0.9415. MINIMDA (Lou *et al.* 2022) incorporated a high-order neighborhood information to improve representation learning achieving an AUC of 0.9. MTLMDA (He *et al.* 2023) introduced a multi task GCN based framework, while MVMTMDA (Huang *et al.* 2021) and AHMDA (Ning *et al.* 2023) further explored multitask and attention-based learning strategies for MDA prediction.

To compared with these methods, HybridGNN consistently achieved superior AUC AUPR, demonstrating its effectiveness in modelling complex biological associations. Nonetheless, the model presents limitation including increasing computational complexity due to multi-head attention mechanism, reliance on local neighborhood information, and evaluation on a single dataset. Future work will focus on optimizing computational efficiency, incorporating gradual component analysis, validating across multiple datasets and integrating additional biological features such as expression or pathway data. Extending HybridGNN framework to related tasks including protein-disease and drug-target association prediction, may further enhance its generalizability and predictive power.

5 Conclusion

In this study, we proposed a HybridGNN that integrates features from three powerful methods Matrix Decomposition with Matrix Factorization (MDMF), Graph Convolutional Network (GCN), and Graph Attention Network (GAT). This integrated framework is designed to significantly enhance the prediction of miRNA-disease associations. The model first employs a multi-similarity network to extract features, followed by three specialized feature extraction modules. A fully connected network is then used to predict miRNA-disease association scores, uncovering previously unknown associations. Owing to this multi-step process, the model achieved a higher AUC score compared to existing approaches, indicating superior predictive performance. Furthermore, we conducted case studies on heart-related diseases such as Myocardial Infarction (MI) and Ischemic Stroke (IS), Lung Neoplasms, Breast Neoplasm, Fibrosis the model identified miRNAs association, confirming its potential for predicting meaningful associations that could aid in future clinical diagnostics. Overall, the development of HybridGNN model cover a critical gap in the prediction of miRNA-disease associations and paves the way for future research, ultimately contributing to advancements in precision medicine and improved health-care outcomes.

Author contributions

Basharat Ahmad (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [equal], Software [lead], Validation [equal], Visualization [equal], Writing—original draft [lead]), Muhammad Hammad Musaddiq

(Formal analysis [equal], Methodology [equal], Software [equal]), Sebu Aboma Temesgen (Formal analysis [equal], Investigation [equal], Methodology [equal]), Grace-Mercure Bakanina Kissanga (Formal analysis [equal], Investigation [equal]), Huma Fida (Validation [equal], Visualization [equal]), Hao Lin (Conceptualization [equal], Funding acquisition [lead], Supervision [equal], Writing—review & editing [equal]), and Ye-Chen Qi (Conceptualization [equal], Formal analysis [equal], Methodology [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [lead])

Conflict of interests

The authors declare no conflict of interest.

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (ZYGX2024Z011).

Data availability

The dataset of miRNA-disease associations in this article can be obtained from Human microRNA disease database (HMDD v4.0).

References

- Ai C, Yang H, Liu X *et al.* MTMol-GPT: de novo multi-target molecular generation with transformer-based generative adversarial imitation learning. *PLoS Comput Biol* 2024;**20**:e1012229.
- Bakhti SZ, Latifi-Navid S. Non-coding RNA-encoded peptides/proteins in human cancer: the future for cancer therapy. *Curr Med Chem* 2022;**29**:3819–35.
- Betkier I, Oszczypała M, Pobożniak J *et al.* PocketFinderGNN: a manufacturing feature recognition software based on graph neural networks (GNNs) using PyTorch geometric and NetworkX. *SoftwareX* 2023;**23**:101466.
- Çakmak HA, Demir M. MicroRNA and cardiovascular diseases. *Balkan Med J* 2020;**37**:60–71.
- Chen C, Ma M, Nie J *et al.* PNAGMDA: a principal neighborhood aggregation based graph neural network for miRNA-disease association prediction. *IEEE Trans Comput Biol Bioinform* 2025a;**22**:1189–200.
- Chen J, Jiang Q, Jiang X-Q *et al.* miR-146a promoted breast cancer proliferation and invasion by regulating NM23-H1. *J Biochem* 2020;**167**:41–8.
- Chen P, He Y-H, Huang X *et al.* MiR-23a modulates X-linked inhibitor of apoptosis-mediated autophagy in human luminal breast cancer cell lines. *Oncotarget* 2017;**8**:80709–21.
- Chen X, Hao T, Han L *et al.* Heterogeneous network embedding based on random walks of type and inner constraint. *Mathematics* 2022;**10**:2623.
- Chen X, Wang L, Qu J *et al.* Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 2018a;**34**:4256–65.

- Chen X, Xie D, Wang L *et al.* BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 2018b;**34**:3178–86.
- Chen Y, Wang Z, Wang J *et al.* Self-supervised learning in drug discovery. *Sci China Inf Sci* 2025b;**68**:170103.
- Chistiakov DA, Orekhov AN, Bobryshev YV. Cardiac-specific miRNA in cardiogenesis, heart function, and cardiac pathology (with focus on myocardial infarction). *J Mol Cell Cardiol* 2016;**94**:107–21.
- Cui C, Zhong B, Fan R *et al.* HMDD v4.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2024;**52**:D1327–D1332.
- Cushing L, Kuang P, Lu J. The role of miR-29 in pulmonary fibrosis. *Biochem Cell Biol* 2015;**93**:109–18.
- Dai B, Li H, Wang P *et al.* DualNetM: an adaptive dual network framework for inferring functional-oriented markers. *BMC Biol* 2025;**23**:254.
- Dhanka S, Sharma A, Kumar A *et al.* Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection methodologies: a comprehensive review. *Arch Comput Methods Eng* 2026;**33**:289–324.
- Ding L, Ni J, Yang F *et al.* Promising therapeutic role of miR-27b in tumor. *Tumour Biol* 2017;**39**:1010428317691657–18.
- Dong B, Sun W, Xu D *et al.* MDformer: a transformer-based method for predicting miRNA-disease associations using multi-source feature fusion and maximal meta-path instances encoding. *Comput Biol Med* 2023;**167**:107585.
- Fan X, Tao S, Li Q *et al.* The miR-23a/27a/24-2 cluster promotes postoperative progression of early-stage non-small cell lung cancer. *Mol Ther Oncolytics* 2022;**24**:205–17.
- Fu L, Peng Q. A deep ensemble model to predict miRNA-disease association. *Sci Rep* 2017;**7**:14482.
- George TP, Subramanian S, Supriya MH. A brief review of non-coding RNA. *Egypt J Med Hum Genet* 2024;**25**:98.
- Groenewegen A, Rutten FH, Mosterd A *et al.* Epidemiology of heart failure. *Eur J Heart Fail* 2020;**22**:1342–56.
- Gu Y, Pais G, Becker V *et al.* Suppression of endothelial miR-22 mediates non-small cell lung cancer cell-induced angiogenesis. *Mol Ther Nucleic Acids* 2021;**26**:849–64.
- Ha T-Y. MicroRNAs in human diseases: from cancer to cardiovascular disease. *Immune Netw* 2011;**11**:135–54.
- Ha J, Park C. MLMD: metric learning for predicting MiRNA-disease associations. *IEEE Access* 2021;**9**:78847–58.
- Ha J. DeepWalk-based graph embeddings for miRNA-disease association prediction using deep neural network. *Biomedicines* 2025;**13**:536.
- Hannafon BN, Cai A, Calloway CL *et al.* miR-23b and miR-27b are oncogenic microRNAs in breast cancer: evidence from a CRISPR/Cas9 deletion study. *BMC Cancer* 2019;**19**:642.
- Harun-Or-Roshid M, Maeda K, Phan LT *et al.* Stack-DHUpred: advancing the accuracy of dihydrouridine modification sites detection via stacking approach. *Comput Biol Med* 2024;**169**:107848.
- He Q, Qiao W, Fang H *et al.* Improving the identification of miRNA-disease associations with multi-task learning on gene-disease networks. *Brief Bioinform* 2023;**24**:bbad203.
- Huang YA, Chan KCC, You ZH *et al.* Predicting microRNA-disease associations from lncRNA-microRNA interactions via multi-view multitask learning. *Brief Bioinform* 2021;**22**:bbaa153.
- Huang D, An J, Zhang L *et al.* Computational method using heterogeneous graph convolutional network model combined with reinforcement layer for MiRNA-disease association prediction. *BMC Bioinformatics* 2022;**23**:299.
- Huang Z, Xiao Z, Ao C *et al.* Computational approaches for predicting drug-disease associations: a comprehensive review. *Front Comput Sci* 2025;**19**:195909.
- Jin ZQ. MicroRNA targets and biomarker validation for diabetes-associated cardiac fibrosis. *Pharmacol Res* 2021;**174**:105941.
- Kesimoglu ZN, Bozdog S. Fusing multiplex heterogeneous networks using graph attention-aware fusion networks. *Sci Rep* 2024;**14**:29119.
- Li R, Ning Q, Zhao Y *et al.* MHMDA: “Similarity-association-similarity” metapaths and heterogeneous-hyper network learning for mirna-disease association prediction. *IEEE Trans Comput Biol Bioinform* 2025;**22**:203–15.
- Li L, Wang Y-T, Ji C-M *et al.* GCAEMDA: predicting miRNA-disease associations via graph convolutional autoencoder. *PLoS Comput Biol* 2021;**17**:e1009655.
- Li Z, Zhang Y, Bai Y *et al.* IMC-MDA: prediction of miRNA-disease association based on induction matrix completion. *Math Biosci Eng* 2023;**20**:10659–74.
- Liang L, Yi X, Wang C *et al.* The impact of miR-34a on endothelial cell viability and apoptosis in ischemic stroke: unraveling the MTHFR-Homocysteine pathway. *Clin Invest Med* 2024;**47**:27–37.
- Lin X, Bao Y, Wang S *et al.* Investigating the unique transcriptional miRNA-mRNA regulatory network of ALK-positive lung adenocarcinoma using machine learning methods. *Cbio* 2025;**21**:314–329.
- Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res* 2019;**47**:e127.
- Liu T, Huang J, Luo D *et al.* Cm-siRPred: predicting chemically modified siRNA efficiency based on multi-view learning strategy. *Int J Biol Macromol* 2024a;**264**:130638.
- Liu T, Qiao H, Ren L *et al.* Deciphering cancer therapy resistance via patient-level single-cell transcriptomics with CellResDB. *Commun Biol* 2025;**8**:1049.
- Liu T, Qiao H, Wang Z *et al.* CodLncScape provides a self-enriching framework for the systematic collection and exploration of coding lncRNAs. *Adv Sci* 2024b;**11**:2400009.
- Long G, Wang F, Duan Q *et al.* Circulating miR-30a, miR-195 and let-7b associated with acute myocardial infarction. *PLoS One* 2012;**7**:e50926.
- Lou Z, Cheng Z, Li H *et al.* Predicting miRNA-disease associations via learning multimodal networks and fusing mixed neighborhood information. *Brief Bioinform* 2022;**23**:bbac159.
- Lu D, Li J, Zheng C *et al.* HGTMDA: a hypergraph learning approach with improved GCN-transformer for miRNA-Disease association prediction. *Bioengineering (Basel)* 2024;**11**:680.
- Luo H, Wu C, Yu S *et al.* PlateletBase: a comprehensive knowledgebase for platelet research and disease insights. *Genomics Proteomics Bioinformatics* 2025;**23**:qzaf031.
- Lv Y, Wang Y, Zhang Y *et al.* Predicting the risk of breast cancer recurrence and metastasis based on miRNA expression. *CBIO* 2024;**19**:482–9.

- Mamatkarimov D, Kang J, Han K. Predicting distant metastatic sites in cancer using miRNA and mRNA expression data. *Curr Bioinformatics* 2025;**20**:473–85.
- Meng X, Yan X, Zhang K *et al.* The application of large language models in medicine: a scoping review. *iScience* 2024;**27**:109713.
- Momanyi BM, Temesgen SA, Wang T-Y *et al.* iGATTLDA: integrative graph attention and transformer-based model for predicting lncRNA–disease associations. *IET Syst Biol* 2024a;**18**:172–82.
- Momanyi BM, Zhou Y-W, Grace-Mercure BK *et al.* SAGESDA: multi-GraphSAGE networks for predicting SnoRNA-disease associations. *Curr Res Struct Biol* 2024b;**7**:100122.
- Nilsson A, Bonander C, Strömberg U *et al.* A directed acyclic graph for interactions. *Int J Epidemiol* 2021;**50**:613–9.
- Ning Q, Zhao Y, Gao J *et al.* AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA–disease associations identification. *Brief Bioinform* 2023;**24**:bbad094.
- Nishikawa E, Osada H, Okazaki Y *et al.* miR-375 is activated by ASH1 and inhibits YAP1 in a lineage-dependent manner in lung cancer. *Cancer Res* 2011;**71**:6165–73.
- Pastuszek-Lewandoska D, Kordiak J, Czarnecka KH *et al.* Expression analysis of three miRNAs, miR-26a, miR-29b and miR-519d, in relation to MMP-2 expression level in non-small cell lung cancer patients: a pilot study. *Med Oncol* 2016;**33**:96.
- Perez-Iratxeta C, Wjst M, Bork P *et al.* G2D: a tool for mining genes associated with disease. *BMC Genet* 2005;**6**:45.
- Pham NT, Phan LT, Seo J *et al.* Advancing the accuracy of SARS-CoV-2 phosphorylation site detection via meta-learning approach. *Brief Bioinform* 2023;**25**:bbad43.
- Pham NT, Terrance AT, Jeon Y-J *et al.* ac4C-AFL: a high-precision identification of human mRNA N4-acetylcytidine sites based on adaptive feature representation learning. *Mol Ther Nucleic Acids* 2024;**35**:102192.
- Popa ML, Ichim C, Anderco P *et al.* MicroRNAs in the diagnosis of digestive diseases: a comprehensive review. *J Clin Med* 2025;**14**:2054.
- Qi R, Liu S, Hui X *et al.* AI in drug development: advances in response, combination therapy, repositioning and molecular design. *Sci China Inf Sci* 2025;**68**:170102.
- Qiao J, Jin J, Yu H *et al.* Towards retraining-free RNA modification prediction with incremental learning. *Inf Sci* 2024;**660**:120105.
- Rangrez AY, Kumari M, Frey N. An emerging role of microRNA miR-223 in cardiovascular pathophysiology. *microRNAs Cardiovasc Res* 2013;**1**:1–11.
- Saliminejad K, Khorram Khorshid HR, Soleymani Fard S *et al.* An overview of microRNAs: biology, functions, therapeutics, and analysis methods. *J Cell Physiol* 2019;**234**:5451–65.
- Saw PE, Song E. Distinguishing functional roles of Protein-Coding Vs. Non-Coding RNAs: the possibility of bifunctional RNAs. In *RNA Therapeutics in Human Diseases*. Singapore: Springer Nature Singapore, 2025, 307–330.
- Shao J, Chen J, Liu B. ProFun-SOM: protein function prediction for specific ontology based on multiple sequence alignment reconstruction. *IEEE Trans Neural Netw Learn Syst* 2025;**36**:8060–71.
- Sheng N, Xie X, Wang Y *et al.* A survey of deep learning for detecting miRNA–disease associations: databases, computational methods, challenges, and future directions. *IEEE/ACM Trans Comput Biol Bioinform* 2024;**21**:328–47.
- Sun W, Guo C, Wan J *et al.* piRNA-disease association prediction based on multi-channel graph variational autoencoder. *PeerJ Comput Sci* 2024;**10**:e2216.
- Sun Z, Huang Q, Yang Y *et al.* PSnoD: identifying potential snoRNA-disease associations based on bounded nuclear norm regularization. *Brief Bioinform* 2022;**23**:bbac240.
- Tang W, Wan S, Yang Z *et al.* Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2018;**34**:398–406.
- Tian M, Zhou Y, Guo Y *et al.* MicroRNAs in adipose tissue fibrosis: mechanisms and therapeutic potential. *Genes Dis* 2025;**12**:101287.
- van Rooij E, Sutherland LB, Thatcher JE *et al.* Dysregulation of microRNAs after myocardial infarction reveals a role of miR-29 in cardiac fibrosis. *Proc Natl Acad Sci USA* 2008;**105**:13027–32.
- Wang D, Wang J, Lu M *et al.* Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**:1644–50.
- Wang JZ, Du Z, Payattakool R *et al.* A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;**23**:1274–81.
- Wang L, Huang X, Liu X. Predicting miRNA–disease associations via meta-path embedding. *IEEE Trans Comput Biol Bioinform* 2025a;**22**:911–22.
- Wang L, Qian Y, Xie H *et al.* Structured sparse regularization-based deep fuzzy networks for RNA N6-methyladenosine sites prediction. *IEEE Trans Fuzzy Syst* 2025b;**33**:131–44.
- Wang R, Jiang Y, Jin J *et al.* DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. *Nucleic Acids Res* 2023;**51**:3017–29.
- Wang Y, Tan J, Wang L *et al.* MiR-125 family in cardiovascular and cerebrovascular diseases. *Front Cell Dev Biol* 2021;**9**:799049.
- Wang B, Komers R, Carew R *et al.* Suppression of microRNA-29 expression by TGF-beta1 promotes collagen expression and renal fibrosis. *J Am Soc Nephrol* 2012;**23**:252–65.
- Wang Y, Zhai Y, Ding Y, Zou Q. SBSM-Pro: support bio-sequence machine for proteins. *Sci China Inf Sci* 2024;**67**:212106.
- Wei L, He W, Malik A *et al.* Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief Bioinform* 2021;**22**:bbaa275.
- Wojciechowska A, Braniewska A, Kozar-Kaminska K. MicroRNA in cardiovascular biology and disease. *Adv Clin Exp Med* 2017;**26**:865–74.
- Xiao Y, Humphries B, Yang C *et al.* MiR-205 dysregulations in breast cancer: the complexity and opportunities. *Noncoding RNA* 2019;**5**:53.
- Xie H, Wang L, Qian Y *et al.* Methyl-GP: accurate generic DNA methylation prediction based on a language model and representation learning. *Nucleic Acids Res* 2025;**53**:gkaf223.
- Xing J, Yang S, Liang Y *et al.* Deciphering sequence determinants of zygotic genome activation genes: insights from machine

- learning and the ZGAExplorer platform. *Cell Prolif* 2025; **58**:e70039.
- Yang H, Luo Y-M, Ma C-Y *et al.* A gender specific risk assessment of coronary heart disease based on physical examination data. *NPJ Digit Med* 2023;**6**:136.
- Zeng S, Zhang S, Wang Z *et al.* GONNMDA: a ordered message passing GNN approach for miRNA–disease association prediction. *Genes (Basel)* 2025;**16**:425.
- Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief Bioinform* 2016; **17**:193–203.
- Zhan L, Mu Z, Jiang H *et al.* MiR-21-5p protects against ischemic stroke by targeting IL-6R. *Ann Transl Med* 2023;**11**:101–16.
- Zhang J, Shen Y, Kang K *et al.* Circulating MicroRNAs as potential biomarkers for ischemic stroke in patients with asymptomatic intracranial artery stenosis. *Cell Mol Neurobiol* 2023; **43**:1573–82.
- Zhong B, Dai Y, Chen L *et al.* ncRS: a resource of non-coding RNAs in sepsis. *Comput Biol Med* 2024;**172**:108256.
- Zhu X, Liu X, Liu Y *et al.* Uncovering the potential differentially expressed miRNAs and mRNAs in ischemic stroke based on integrated analysis in the gene expression omnibus database. *Eur Neurol* 2020;**83**:404–14.