communications biology

Article

https://doi.org/10.1038/s42003-024-06734-0

TriFusion enables accurate prediction of miRNA-disease association by a tri-channel fusion neural network

Check for updates

Sheng Long, Xiaoran Tang, Xinyi Si, Tongxin Kong, Yanhao Zhu, Chuanzhi Wang, Chenqing Qi, Zengchao Mu ✉ & Juntao Liu ✉

The identification of miRNA-disease associations is crucial for early disease prevention and treatment. However, it is still a computational challenge to accurately predict such associations due to improper information encoding. Previous methods characterize miRNA-disease associations only from single levels, causing the loss of multi-level association information. In this study, we propose TriFusion, a powerful and interpretable deep learning framework for miRNA–disease association prediction. It develops a tri-channel architecture to encode the association features of miRNAs and diseases from different levels and designs a feature fusion encoder to smoothly fuse these features. After training and testing, TriFusion outperforms other leading methods and offers strong interpretability through its learned representations. Furthermore, TriFusion is applied to three high-risk sexually associated cancers (ovarian, breast, and prostate cancers) and exhibits remarkable ability in the identification of miRNAs associated with the three diseases.

MicroRNAs (miRNAs) are a group of tiny non-coding RNAs that are typically made up of around 20 to 24 nucleotides. They are important for cell function, development, fighting infections, immune responses, and health issues, including diseases and cancers[1,2]. Thus, identifying the association between miRNAs and diseases is crucial to gain a more comprehensive insight into the intricate mechanisms of disease pathology. The accurate identification of miRNA-disease associations can be effectively performed using various biological techniques, including high-throughput RNA sequencing, quantitative real-time Polymerase Chain Reaction, and innovative multiplexed detection methods. However, these methods are quite time-intensive and costly. Fortunately, rapid improvements in computing power and the creation of databases related to miRNAs and diseases have led to the development of computational approaches[3]. These methods offer a more efficient way to investigate the connections between miRNAs and diseases and greatly reduce the reliance on labor-intensive laboratory work[4–6]. Recent computational prediction methods mainly fall into two categories: those based on machine learning and those on deep learning.

Machine learning-based methods typically extract reliable biometric and association features and apply existing models to predict miRNA-disease relationships. For example, the RWRMDA[7] approach first builds a network capturing the functional similarities between miRNAs and diseases and then employs the random walk with restart algorithm on this network to detect miRNAs that are likely associated with particular diseases. EGBMMDA[8] is an early computational approach for predicting miRNA-disease associations by utilizing decision tree learning. It calculates the probability of associations between miRNAs and diseases via the extreme gradient boosting method. Zeng[9] introduces a model that leverages structural consistency as a metric to infer associations between miRNAs and diseases. Zhong et al.[10] introduces a sparse penalization model based on non-negative matrix factorization to predict disease-associated miRNAs. DF-MDA develops a heterogeneous network incorporating miRNAs, diseases, and other small molecules to infer potential associations by utilizing a diffusion-based method[11]. Xu et al.[12] designs a method called MTDN, which only requires extracting features through the miRNA-target association network before inputting them into the prediction model. The use of stacked auto-encoders[13–15] has also gradually improved prediction accuracy. HDMP[16] starts from calculating disease semantic similarity and phenotype similarity, followed by choosing the k most similar neighbors to group miRNAs and identify the miRNAs that are associated with specific diseases. To recover missing associations between miRNAs and diseases, Chen[17] proposes a new method, NCMCMDA, which adds neighborhood constraints to the joint similarity of miRNAs and diseases. ABMDA[18] employs a random sampling technique to generate balanced positive and negative samples and combines multiple weak classifiers to improve the classification accuracy. Overall, the machine learning methods can effectively predict miRNA-disease associations based on the well-designed association features for small-scale data. However, they are difficult to learn unknown miRNA-disease association patterns hidden in large datasets due to their limited

School of Mathematics and Statistics, Shandong University, Weihai, China. ✉e-mail: muzengchao@sdu.edu.cn; juntaosdu@126.com

fitting abilities. Therefore, in recent years, deep learning-based computational methods have been increasingly utilized by researchers.

In the field of deep learning-based methods, Graph Convolutional Networks (GCNs) have recently gained wide attention due to their outstanding ability to learn graph representations. Lou et al.[19] propose the MINIMDA model, which improves existing GCNs by explicitly aggregating information from high-order neighborhoods. Tang et al.[20] introduce the MMGCN model to adaptively learn different feature representations by integrating multi-source similarity networks with a combination of a GCN encoder and CNN decoder. Applications like drug repositioning[21], drug-target interaction prediction[22,23], and cancer-related gene prediction[24] have benefited from the exceptional performance of GCNs in association prediction tasks[25–27]. Additionally, the use of transformer architectures to predict associations within their respective domains has been explored, which takes advantage of heterogeneous networks' multi-typed meta-path instance exploration for feature embedding[28–30] and overcomes the limitations of graph models in effectively exploring and learning global information[31–33]. MD-former[34] employs a transformer-based deep neural network with specialized encoders to effectively predict miRNA-disease associations by analyzing their complex features.

Although great efforts have been made in the design of computational methods and impressive improvements have been achieved in predicting miRNA-disease associations, most of the existing methods fail to capture the complex representations of miRNA-disease associations, leading to unsatisfactory predictions. In fact, the similarity network of miRNAs (diseases) and the known associations between miRNAs and diseases are the vital information that determines prediction results. The two kinds of information encompass diverse association features related to miRNA-disease association patterns, which should be captured from multiple manners. However, previous methods usually encode the association features from single levels, making it difficult to fully characterize the complete miRNA-disease associations.

To solve this challenging task, we propose a model, TriFusion, which implements a tri-channel framework for association features encoding from three levels. The first channel designs a graph convolution module to encode the similarity relationships between each miRNA (disease) and its neighbors of different orders based on the miRNA (disease) similarity network. The second channel develops a hypergraph convolution module for encoding the high-level similarity information between two miRNAs (diseases) hidden in their common neighbors again based on the miRNA (disease) similarity network. The third channel introduces an miRNA-disease interaction encoding module to capture the inherent association information between miRNAs and diseases based on the known miRNA-disease associations. And then, a feature fusion encoder is implemented for effectively fusing the tri-channel features (see Fig. 1 and the Methods section for details).

TriFusion is tested under HMDD v3.2[35] and compared with multiple leading prediction methods. The evaluation results show that TriFusion clearly outperforms all the other models and demonstrates stronger ability in discovering new associations. Meanwhile, we conduct case studies on three high-risk sexually associated cancers (ovarian, breast, and prostate cancers) based on the HMDD v3.2 database. Remarkably, 100% of the top 30 miRNAs in the predicted miRNA scores by TriFusion are confirmed by relevant databases, showcasing its outstanding reliability in practical applications. Through visualization, we find that the learned representations from the three channels, the fused representations, and the GCN enhanced representations are all characterizing the miRNA-disease association patterns in different manners, which explains the necessity of feature encoding from multiple levels.

Results

Overview of TriFusion

The main framework of TriFusion comprises the following four parts. (1) feature extraction for miRNAs and diseases; (2) encoding high-level representations for miRNAs and diseases via a tri-channel feature encoder; (3) fusion of features for miRNAs and diseases via a feature fusion encoder; and (4) prediction of miRNA-disease associations.

Since similar miRNAs (diseases) often have close associative properties, we first construct multiple types of similarity matrices for miRNAs (diseases). For diseases, both the semantic similarity and Gaussian similarity are used to measure the similarity of two diseases. The semantic similarity and Gaussian similarity of two diseases are respectively defined based on their hierarchical relations and their interactions with miRNAs. For miRNAs, the similarity of two miRNAs is described by three types: sequence similarity, functional similarity, and Gaussian similarity. Sequence similarity is defined based on the similarity of their sequences, functional similarity is defined based on the similarity of their functions, and Gaussian similarity for miRNAs is defined through their interactions with diseases. The extracted similarity matrices serve as the original feature matrices for miRNAs and diseases.

To comprehensively learn the association patterns between miRNAs and diseases, TriFusion develops a tri-channel feature encoder to encode the representations of miRNAs and diseases from different levels, including low-order graph encoding, high-order hypergraph encoding, and miRNA-disease interaction encoding. The direct relationships of an miRNA (disease) with its neighboring miRNAs (diseases) can effectively characterize the miRNA-disease association patterns. The low-order graph encoding channel of the tri-channel module is designed to calculate the representations of miRNAs (diseases) by message passing between miRNAs (diseases) and their multi-order neighbors. The high-level relationships between two miRNAs (diseases) hidden in their common neighbors can also effectively describe the association patterns. The high-order hypergraph encoding channel learns the representations by a hypergraph convolution on the constructed hypergraph of miRNAs (diseases). The relationships between target miRNAs and diseases contain inherent association information to measure their association patterns. The miRNA-disease interaction encoding channel can effectively capture the association representations by encoding the degrees and neighbor similarities of the nodes in the constructed miRNA-disease heterogeneous graph.

The three representations learned by the tri-channel encoder describe the miRNA-disease association patterns from different levels, which should be carefully fused together to generate a complete representation. To achieve this, we design a feature fusion encoder that encompasses a biased Transformer encoder with an embedded residual connection, followed by a multi-layer graph convolution. The final classification is conducted by fusing the representations of an miRNA and a disease through a Hadamard product and then deriving an miRNA-disease association probability via a multi-layer MLP.

Experimental settings

To validate the performance of a method, we conduct 5-fold cross-validation tests on the HMDD v3.2 database via different manners for various purposes as follows.

Random zero cross-validation. All known miRNA-disease associations are considered as positive samples, which are randomly divided into five non-overlapping subsets. During each iteration of cross-validation, a subset is chosen as the test set, complemented by an equal number of randomly selected negative samples. The remaining of all positive and negative samples serve as the training set. This process, known as random zero cross-validation, evaluates the capacity of a model to identify undetected miRNA-disease associations.

Random multi-column zero cross-validation. Given the miRNA-disease association matrix, the test set is generated by randomly selecting and zeroing out 1/5 of the columns in this matrix, with the training set based on the remaining 4/5 columns. In addition, an equivalent number of randomly selected negative samples is added for balance. This process aims to test the effectiveness of a model in discovering the associations between known miRNAs and new diseases.

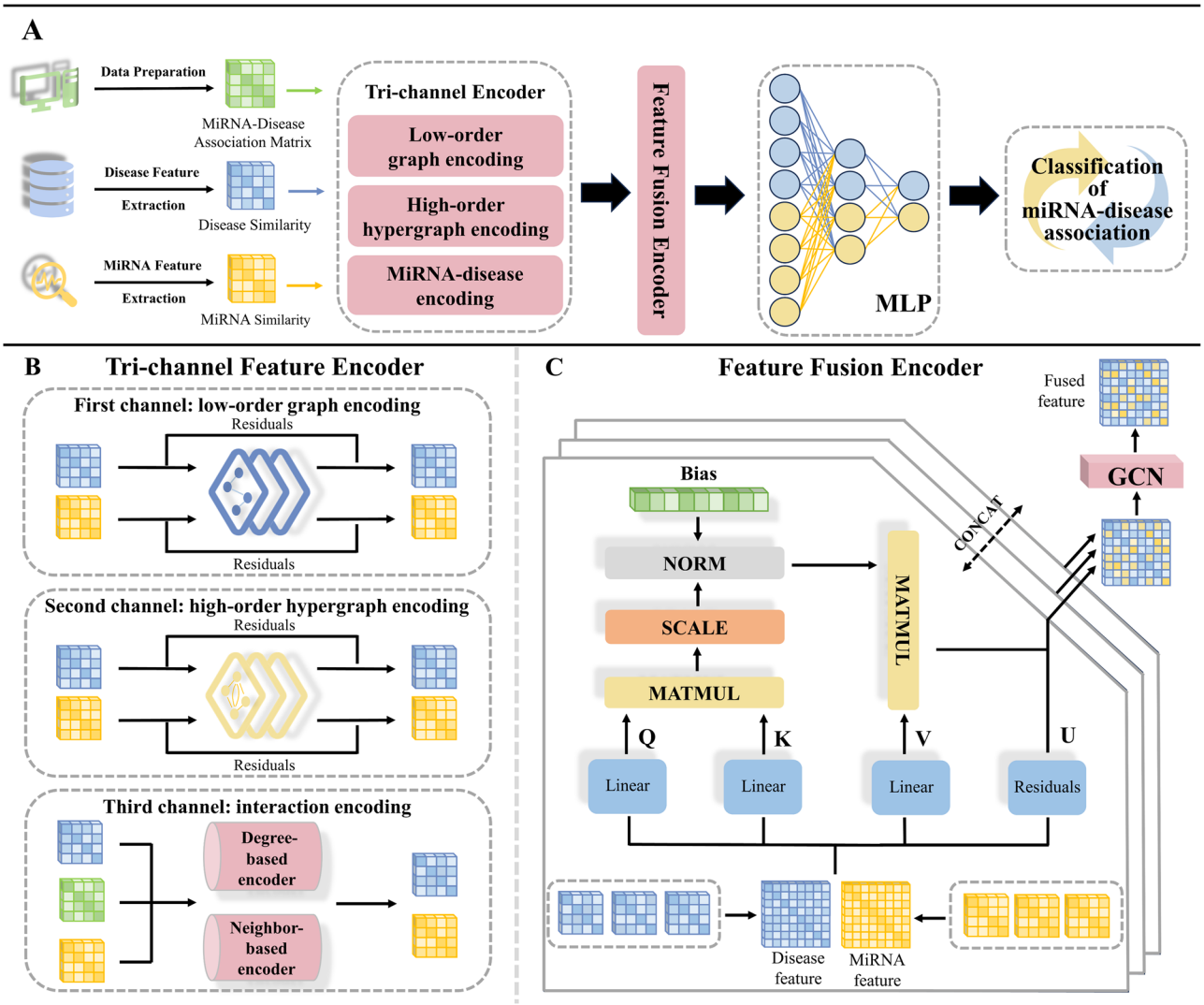

Fig. 1 | Flowchart of the TriFusion model. A The overall framework of the tri-channel architecture, divided into four sections including feature extraction, tri-channel feature encoder, feature fusion encoder, and classification. **B** The detailed structure of the tri-channel feature encoder, where the three channels respectively conduct multi-order graph convolutions, hypergraph convolutions, and miRNA-disease interactions. **C** The detailed structure of the feature fusion encoder, which incorporates a biased Transformer encoder in a U-dimensional space and a graph convolutional network (GCN) to effectively fuse the information from the three channels.

Random multi-row zero cross-validation. Similar to the above, the test set is generated by randomly selecting and zeroing out 1/5 of the rows in this matrix, with the training set based on the remaining 4/5 rows. This process aims to test the effectiveness of a model in discovering the associations between new miRNAs and known diseases.

State-of-the-art methods including MINIMDA[19], MD-former[34], DAEMDA[36], AGAEMD[37], AMHMDA[38], and ELMDA[39] are collected to compare with TriFusion. In this study, six common evaluation metrics are used to evaluate the performance of a model, namely area under the ROC Curve (AUC), area under the PR Curve (AUPR), Accuracy (ACC), F1 score, precision, and recall (see Supplementary Note 1 for detailed definitions of the metrics).

TriFusion shows the best performance

We compare the performance of TriFusion with the above six leading miRNA-disease association prediction methods on the same test set under the three types of cross-validations. According to the evaluation results, TriFusion achieves great improvements over all the methods across all the tests.

Random zero cross-validation. The comparison results of Random Zero Cross-Validation are shown in Fig. 2 (see Supplementary Table 1 for detailed results). Among the compared methods, ELMDA and AGAEMDA are machine learning-based models, while the others are based on deep learning. We find that deep learning methods illustrate better performance than machine learning models, with both AUC and AUPR exceeding 94% (see Supplementary Fig. 1). Specifically, MINIMDA, which applies improved graph convolution to encode node information, achieves a very high AUC value of 94.97%, only lower than that of TriFusion. MD-former, which extracts features from heterogeneous graphs through random walks, obtains the second-highest AUPR value of 94.75%. Among these models, only TriFusion achieves both AUC and AUPR exceeding 95% (with its AUC and AUPR being 95.41% and 95.25%, respectively). Compared to these models, the relative increase in AUC and AUPR of TriFusion range from 0.47% to 3.97% and from 0.53% to 4.30%, respectively. Moreover, the recall of TriFusion even exceeds 90%, with an improvement of 2.01% over the second best method. To further illustrate the significance of the improvement, we select MDformer, the model with the second-best overall performance,

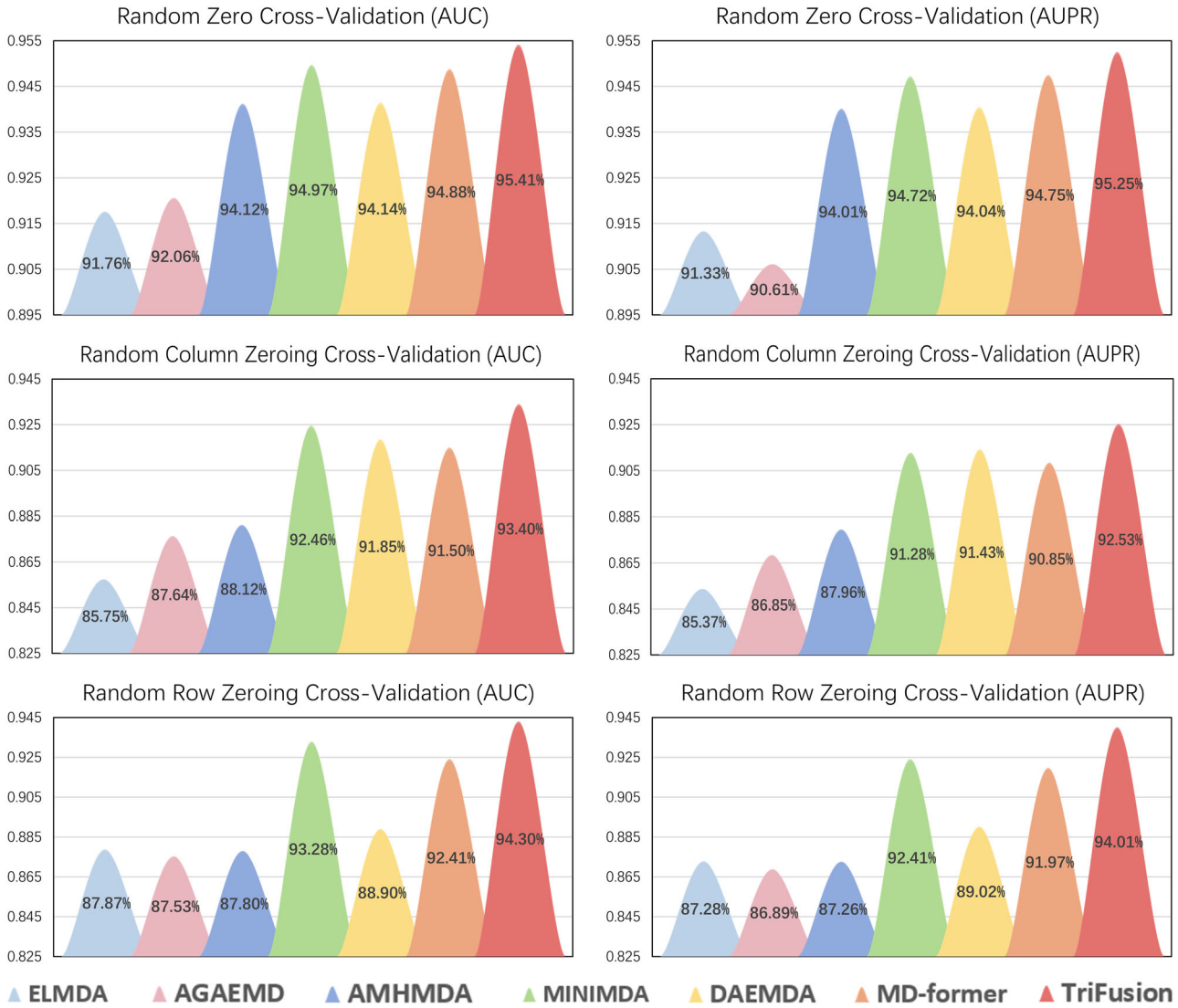

Fig. 2 | Comparison of TriFusion with other methods under three types of validations. This figure displays the values of AUC and AUPR of all the compared methods under three types of cross-validation conditions: Random Zero Cross-Validation, Random Multi-Column Zero Cross-Validation, and Random Multi-Row Zero Cross-Validation.

and Trifusion, each running 10 times, for an independent samples *t*-test. The *p*-values for the tests based on AUC and AUPR are all smaller than 1e-10, indicating the the significance of the improvement made by TriFusion (see Supplementary Table 2 for details).

Random multi-column zero cross-validation. The comparison results of Random Multi-Column Zero Cross-Validation are shown in Fig. 2 (see Supplementary Table 3 for detailed results). It is observed that most deep learning models again show much better performance than machine learning-based methods, with the AUC and AUPR values reaching over 90%. It is worth noting that, compared to other models, the AUC improvement of TriFusion ranges from 1.02% to 8.92%, and its AUPR improvement ranges from 1.20% to 8.39%, which demonstrates that TriFusion can better predict the associations between known miRNAs and unknown diseases.

Random multi-row zero cross-validation. Performance evaluation is also conducted by Random Multi-Row Zero Cross-Validation and the results are shown in Fig. 2 (see Supplementary Table 3 for detailed results). After comparison, we find that TriFusion consistently performs better than all the other compared methods, with both AUC and AUPR exceeding 94%. Specifically, its AUC reaches 94.30%, with its improvement over the other methods ranging from 1.10% to 7.73%, and its AUPR achieves 94.01%, with an improvement ranging from 1.73% to 7.74%. This indicates that TriFusion shows better ability in predicting associations between new miRNAs and known diseases.

Ablation study

To measure the impact of the tri-channel feature encoder, each channel of the tri-channel feature encoder, and the feature fusion encoder, we conduct ablation experiments by removing certain encoding modules from the TriFusion model. Here, ablation studies are carried out in the manner of removing or altering only one component each time.

Impact of the tri-channel feature encoder. To examine the influence of this encoder, we directly input the extracted similarity data between miRNA and disease through a fully connected layer into the feature fusion encoder, which results in a significant decrease in performance (Fig. 3). This indicates that the tri-channel approach is able to extract effective multi-level miRNA-disease association information, which contributes a lot in accurate association predictions.

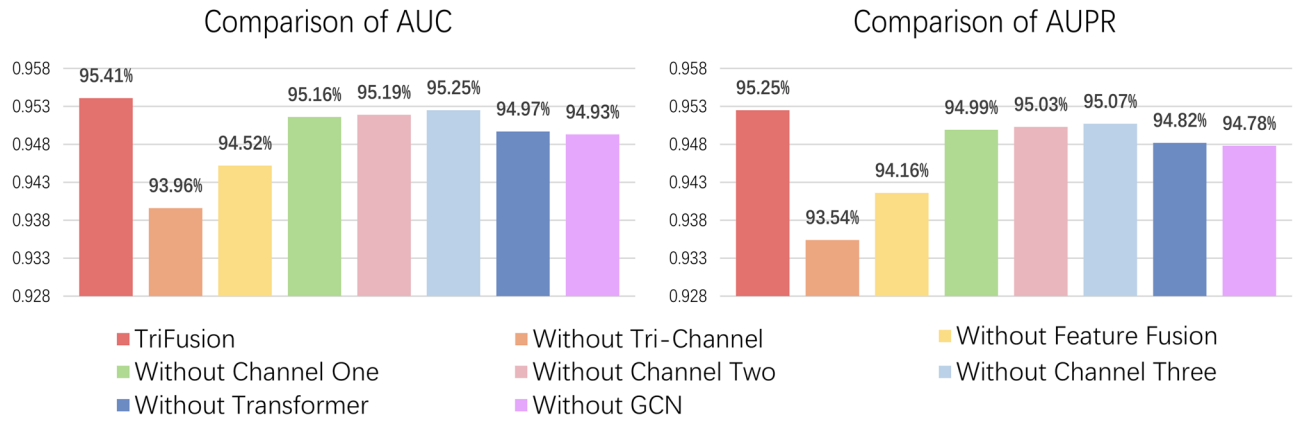

Fig. 3 | Results of the ablation experiments. This figure illustrates the results of several ablation experiments. This two figures show the performance of TriFusion with several modules removed.

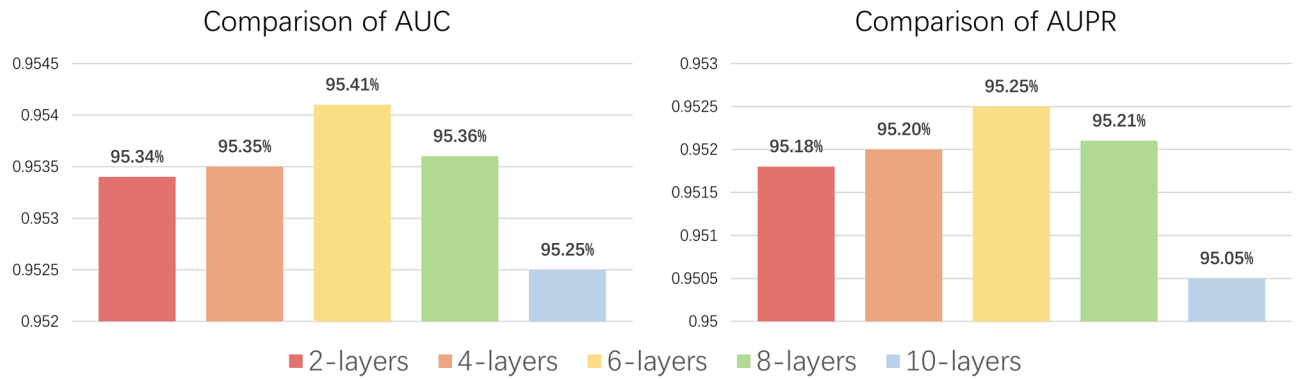

Fig. 4 | Results of the ablation experiments. This two figures show the AUC and AUPR values for different numbers of GCN layers within the feature fusion encoder.

Impact of each encoding channel. To further explore the impact of each channel, we conduct three experiments by respectively removing the graph convolution module, the hypergraph convolution module, and the miRNA-disease interaction encoding module. Results show that the performance of all three experiments clearly declines (see Fig. 3). It is worth noting that the impact of any channel is much lower than that of the whole tri-channel feature encoder (see Fig. 3), which indicates that any two channels among the three can capture most association features, and the application of all three channels achieves the best feature representations.

Impact of the feature fusion encoder. The feature fusion encoder contains two parts: the biased Transformer and GCN. First, we simply add the three different kinds of features obtained by the tri-channel encoder and input the features directly into the classification module, which results in a great decline in performance (see Fig. 3). Next, to individually test the role of the biased Transformer module, we input the representations obtained from the tri-channel feature encoder directly into the GCN part for prediction, again resulting in a great decrease (see Fig. 3). This indicates that the biased Transformer encoder plays a crucial role in learning the complete representations of miRNAs and diseases. To further test the contribution of the GCN module, we remove it by inputting the fused representations directly into the classification module, and results show that the performance of TriFusion also declines (Fig. 3).

Impact of the number of GCN layers. To assess the impact of the number of GCN layers within the feature fusion encoder on the overall predictive performance of the model, we carry out experiments with GCN layers of 2, 4, 6, 8, and 10, respectively. The experimental results, as shown in Fig. 4, indicate that the model performs best when the GCN has 6 layers.

Interpretation of the TriFusion model

To deeply understand the learning mechanism of TriFusion in capturing the miRNA-disease association patterns, we try to interpret it in different manners. Firstly, we extract all the learned representations from the test set at continuous training stages and visualize their 2-dimensional projections via the t-SNE tool (Fig. 5). From the visualization, it is evident that TriFusion is gradually learning the association patterns and the segmentation of associations and non-associations is becoming increasingly clear according to the 2D t-SNE projections of the learned representations. Secondly, to verify what and how each module of TriFusion is learning, we respectively visualize the 2-dimensional projections of the representations learned from the tri-channel feature encoder, each of the three channels, and the feature fusion encoder. The visualization results show that each module is learning the miRNA-disease association patterns in different manners. Notably, in the interaction encoding channel, it seems that the associations and non-associations are not well classified. However, over 80% of the samples are arranged near the center, which are well classified.

Case studies

In this section, we conduct case studies on three different types of cancer: ovarian cancer, breast cancer, and prostate cancer to demonstrate the prediction capability of TriFusion. We used all known positive associations in HMDDv3.2, a total of 12,446 positive associations, as the positive training set. From the remaining unknown samples, we randomly selected an equal number of samples as negative and added them into the training set. After training, we obtained an 853*591 association prediction matrix, where the score of (i, j) represents the predicted association value between sample i and

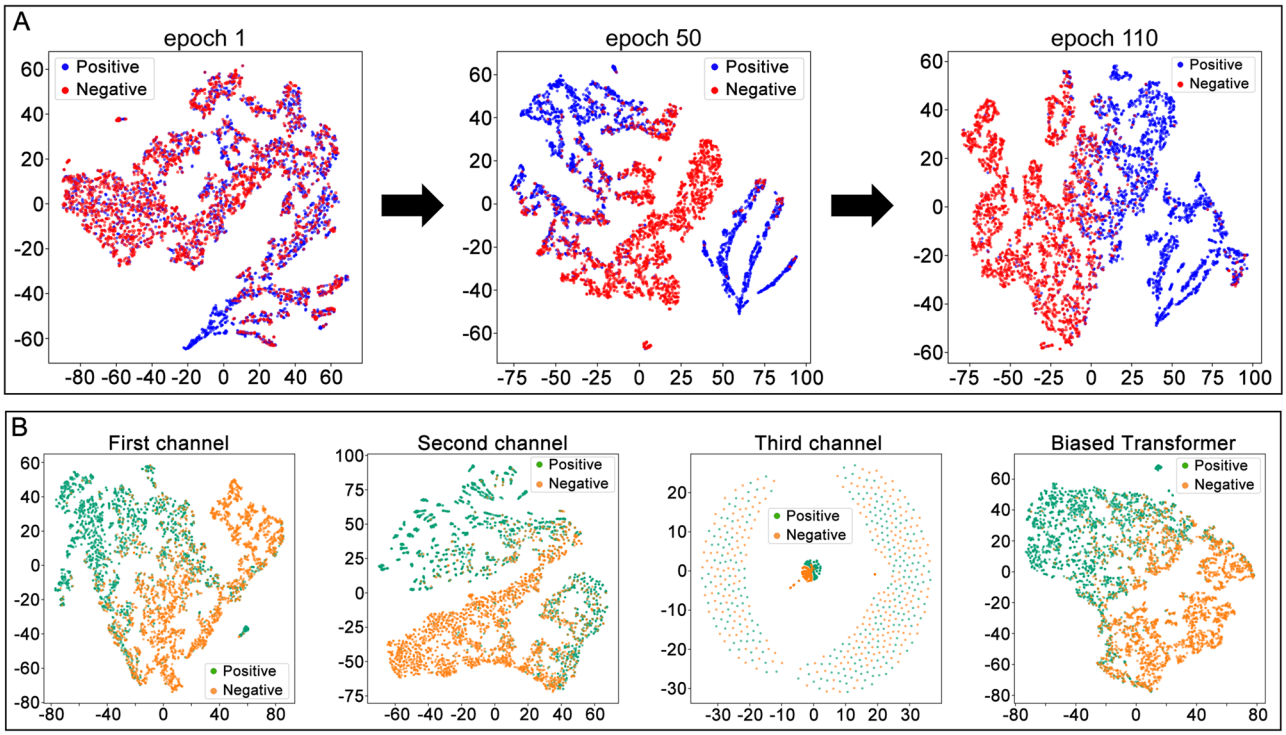

Fig. 5 | Interpretation experiments of TriFusion. A The three figures illustrate the TriFusion training process, with blue points indicating positive samples and red points indicating negative samples. **B** The four figures display the visualization results of the learned representations from each of the three channels in the tri-channel feature encoder as well as the Transformer module in the feature fusion encoder, where green points represent positive samples and orange points represent negative samples.

sample *j*. We then index the *k*-th column corresponding to the target disease, remove all known positive association points in the *k*-th column, and select the top 50 points with the highest scores from the remaining points. After that, we screen the top 50 predicted miRNAs and verify these prediction associations based on two other miRNA–disease association datasets, dbDEMC[40] and HMDDv4.0[41] (Fig. 6).

Ovarian cancer poses a serious risk to women's health. However, its early detection is quite difficult because there are currently no clear early symptoms and screening methods that are proved effective. Fortunately, in ovarian cancer patients, the presence of miR-148b is as high as 92.21%, which makes it a key indicator for detecting the disease early[42]. In this case, all the top 50 miRNAs associated with ovarian cancer predicted by TriFusion are confirmed in the dbDEMC database, with the detailed verification of the remaining miRNAs listed in Supplementary Table 4.

Breast cancer is among the most common cancers in women, accounting for approximately 25% of all cancer cases in females and presenting a significant threat to life. Recent studies indicate that in patients with breast cancer, the levels of certain miRNAs such as hsa-miR-126 and hsa-miR-10b are reduced in their tissues[43]. This provides a new method for the early detection of this type of cancer. In this case, except for hsa-miR-181a-1 and hsa-miR-153-1, which lack supporting data, the datasets have validated all of the top 50 miRNAs associated with breast cancer predicted by TriFusion. For specific verification details, refer to Supplementary Table 5.

Prostate cancer is a leading type of cancer and the second primary cause of cancer-related deaths in men. It is especially prevalent in those over seventy, ranking as the third most common urological tumor. Current studies highlight a clear link between the serum miRNA expression patterns in prostate cancer and the tumor's severity. Notably, changes include variations in 156 miRNAs, miR-16 and miR-141 levels are decreased in patients with prostatic hyperplasia and throughout various prostate cancer stages, whereas miR-34 levels are found to increase under the same conditions[44]. In this case, all the top 50 miRNAs predicted to be associated with prostate cancer by TriFusion, except for hsa-miR-181a-1, hsa-miR-138-1, and hsa-miR-337, which have no supporting data, are again confirmed in the datasets. Specific verification can be found in Supplementary Table 6.

In summary, it is clear that TriFusion demonstrates excellent performance in the above case studies. Specifically, the top 30 predicted miRNAs associated with the three diseases are all validated, and it achieves a prediction accuracy of 96.7% for the top 50 miRNAs. These findings highlight the effectiveness of TriFusion in predicting miRNA-disease associations and its great potential in identifying new biomarkers and therapeutic targets.

Discussion

The identification of miRNA-disease associations is critical for early disease prevention and treatment. However, in previous models predicting miRNA-disease associations, researchers only encode the association features from single levels that are not capable of fully extracting the miRNA-disease association information. In this study, we propose TriFusion, a model that extracts features from different levels through a tri-channel feature encoder and carefully fuses them by a feature fusion encoder. After training and testing, it performs much better than six leading methods in terms of AUC and AUPR. Moreover, we find that the learned representations of TriFusion from its different modules are all fitting the miRNA-disease association patterns in different manners, which again explains the necessity of feature encoding from multiple levels and demonstrates its strong interpretability. We also apply TriFusion to three high-risk sexually associated cancers including ovarian, breast, and prostate cancers. Remarkably, 100% of the top 30 miRNAs and most of the top 50 miRNAs predicted by TriFusion are confirmed by relevant studies, showcasing its outstanding reliability in practical applications.

The strong predictive capability of TriFusion can be attributed to its two main factors. (1) To fully describe the association patterns between miRNAs and diseases, TriFusion develops a tri-channel architecture to encode the representations of miRNAs and diseases from three different levels, including low-order graph features, high-order hypergraph features,

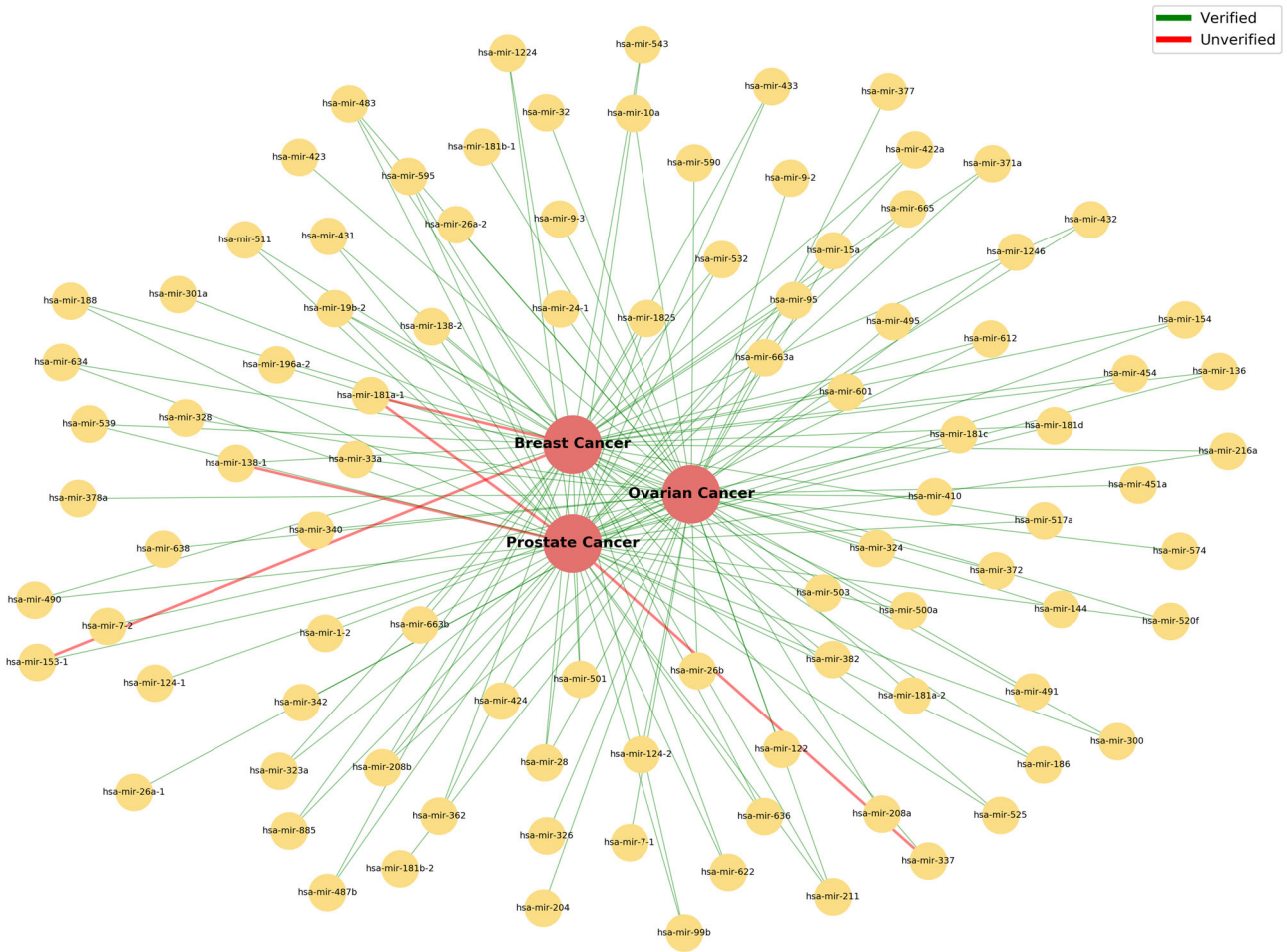

Fig. 6 | Validation results for the top 50 miRNAs associated with three types of cancers (Ovarian cancer, Breast cancer, and Prostate cancer) predicted by TriFusion. Green lines indicate that the corresponding associations have been validated, while red lines denote the associations have not yet been validated.

and miRNA-disease interaction features. Through t-SNE visualizations, we find that the three representations from the three channels are all characterizing the miRNA-disease association patterns with different manners. And the ablation experiments also confirm that removal of each channel causes clear decline of its performance. Therefore, it is necessary to perform feature encoding from different levels. (2) Ablation experiments show that simply adding these three types of features together results in a significant decline in performance. Therefore, feature fusion is another important task. To carefully fuse the learned representations from the three different channels, TriFusion designs a feature fusion encoder to generate a complete representation, which can accurately characterize miRNA-disease association patterns.

In fact, we still have a long way to go to completely solve this challenging problem. To date, only the association network between miRNAs and diseases is taken into consideration for miRNA-disease association prediction. In practice, other types of data, including targeted drugs and target genes, can also be applied to systematically solve this problem. In addition, existing methods fail to predict association types between miRNAs and diseases, such as whether these associations lead to an increase or decrease in miRNA levels. Moreover, due to the fact that these associations are inherently time-dependent, introducing a temporal dimension to the associations can lead to a deeper understanding of the issues. The above three points may be the future directions for miRNA-disease association predictions.

To assist researchers in exploring potential associations, we designed a Python program based on all the association edges from HMDDv3.2. By entering the name of a disease, the program directly outputs the most likely associated diseases. We released user-friendly software for TriFusion and hope that it can contribute to the understanding of miRNA-disease associations as well as early disease prevention and treatment.

Methods

Data preparation

The Human MicroRNA Disease Database-HMDD (http://www.cuilab.cn/hmdd) offers valuable insights into the associations between microRNAs and human diseases, with all relationships validated by experiments or supported by referenced sources. In this study, we downloaded HMDD v3.2 from this database to train and evaluate the performance of TriFusion and other methods. After preprocessing, it contains 12,446 associations between 853 miRNAs and 591 diseases. All the known associations are considered positive samples, with the remaining unknown/unassociated as negative. To achieve a more comprehensive training, the positive and negative samples are balanced by randomly selecting an equal number of negative samples to match the positive ones.

Feature extraction

In this study, the information from disease similarity and miRNA similarity is extracted as features for MD association learning and prediction. Both semantic similarity and Gaussian similarity are used to measure the similarity between two diseases, and the similarity between two miRNAs is described by three types of similarities including sequence similarity, functional similarity, and GIP similarity.

Disease similarity. According to MeSH (http://www.ncbi.nlm.nih.gov/mesh), there are two common methods to calculate the similarity of two diseases based on their hierarchical relations[34]. Therefore, two similarity matrices $DSS1 \in R^{591\times591}$ and $DSS2 \in R^{591\times591}$ are generated accordingly (see Supplementary Note 2 for detailed calculations). Then, the average matrix DSS of $DSS1$ and $DSS2$ is calculated to measure the semantic similarity between diseases. The Gaussian Interaction Profile (GIP) is another metric used to measure the association degree between miRNAs and diseases. According to Van Laarhoven et al.[45], two binary vectors $IP(d_i)$ and $IP(d_j)$ can first be defined to describe the interaction profile of two diseases d_i and d_j, based on which the Gaussian similarity matrix $DGS \in R^{591\times591}$ is calculated (see Supplementary Note 2 for details).

miRNA similarity. All miRNA sequences matching the dataset were first downloaded from the miRBase database (https://mirbase.org/), and then the sequence similarity matrix $MSS \in R^{853\times853}$ is generated by using the Needleman-Wunsch algorithm[34] (see Supplementary Note 3 for details). miRNA functional similarity is another reliable representation of miRNAs and is widely used in multiple fields. MiRNAs with similar functions are typically associated with similar diseases. Based on the information provided by Wang et al.[46] in the MISIM database (https://www.cuilab.cn/), the miRNA functional similarity matrix is calculated for this study, denoted as $MFS \in R^{853\times853}$ (see Supplementary Note 3 for details). Similar to the Gaussian similarity with diseases, we also define two binary vectors $IP(m_i)$ and $IP(m_j)$ to describe the interaction spectrum between miRNAs m_i and m_j, and then calculate the Gaussian similarity matrix for miRNAs, denoted as $MGS \in R^{853\times853}$ (see Supplementary Note 3 for details).

The TriFusion framework

Tri-channel feature encoding of miRNA and disease. A tri-channel feature encoder is developed to capture three types of representations of miRNAs and diseases that encompass low-order graph encoding, high-order hypergraph encoding, and miRNA-disease interaction encoding.

Low-order graph encoding via graph convolution. An miRNA (disease) association graph is first constructed with nodes representing miRNAs (diseases) and edges denoting close relationships between two nodes. In this study, the top K most similar miRNAs (diseases) for each miRNA (disease) are defined as the neighbors of the miRNA (disease) and are connected by K edges (K is set to 40 in this study). In this section, the similarities between two miRNAs and two diseases are respectively measured by $MS = (MSS + MGS)/2$ and $DS = (DSS + DGS)/2$, which also serve as the feature matrices for miRNAs and diseases. Then, the two corresponding adjacency matrices are respectively generated for miRNA and disease associations, denoted as G_m, and G_d, and the multi-order graph convolution is applied by the following formula.

$$H_i^{(l+1)} = RELU\left[\left(D^{-\frac{1}{2}}G_aD^{-\frac{1}{2}}\right)^i H^{(l)}W_i^{(l)}\right]$$
$$H^{(l+1)} = \sum_{i=1}^{N}\lambda_i H_i^{(l+1)}$$

where G_a represents the miRNA or disease adjacency matrix with $a = m$ or $a = d$, D is the degree matrix of G_a, $(D^{-1/2}G_aD^{-1/2})^i$ denotes matrix $D^{-1/2}G_aD^{-1/2}$ multiplied by itself i times, which is the normalized adjacency matrix at the i-th order, $H_i^{(l+1)}$ is the feature matrix for the $(l+1)$-th layer at the i-th order, $H^{(l)}$ is the combined feature matrix for the l-th layer, $W_i^{(l)}$ is the l-th trainable parameter matrix at the i-th order, N is a hyperparameter representing the largest neighbor order ($N = 3$ is set in this study), λ_i is another hyperparameter indicating the weight assigned to the feature matrix of the i-th order ($\lambda_1 = \lambda_2 = \ldots = \lambda_N = 1/N$ is set in this study).

Through the graph convolutional network, the two feature matrices for miRNA and disease are respectively obtained as $MF1$ and $DF1$. Meanwhile, the two original feature matrices MS and DS are embedded by an MLP to generate another two feature matrices $MF2$ and $DF2$. Finally, the encoded features for miRNAs and diseases are calculated as follows.

$$GF = \begin{bmatrix}MF\\DF\end{bmatrix} \in R^{(N_m+N_d)\times h}$$
$$MF = \frac{MF1+MF2}{2}, DF = \frac{DF1+DF2}{2}$$

where N_m and N_d respectively denote the number of miRNAs and diseases, and h refers to the dimension of the hidden features.

High-order hypergraph encoding via hypergraph convolution. As the Gaussian similarity contains important high-level relationships among miRNAs (diseases), we utilize it to obtain high-level representations of miRNAs (diseases) by applying the hypergraph convolution. First of all, a graph G_m (G_d) is constructed for miRNAs (diseases) with nodes representing miRNAs (diseases), and an edge is connected between any two nodes if their Gaussian similarity is larger than s_g (s_g is set to 0 in this study). Then, a hypergraph HG_m (HG_d) is built for miRNAs (diseases) with nodes denoting miRNAs (diseases) and each hyperedge consisting of the neighbor set of a node in G_m (G_d). Taking HG_m as an example, $HG_m = \{V_m, E_m\}$, where V_m represents all the nodes (miRNAs) in G_m, and E_m is the set of hyperedges, manually set to match the number of nodes, and the i-th hyperedge $e_i = \{v_j | \ v_j$ is a neighbor of v_i in $G_m\}$ represents the set of the neighbors of the i-th node in G_m. The corresponding incidence matrix Y_m (Y_d) is obtained with rows representing the nodes in V_m (V_d) and columns denoting the hyperedges in E_m (E_d). $Y_m(i, j)$=1 if the i-th node is included in the j-th hyperedge and $Y_m(i, j) = 0$ otherwise. The feature matrix MHF of miRNA is constructed by concatenating $(MSS + MFS)/2$ and MGS, while DHF is constructed by concatenating DSS and DGS for disease. Based on the hypergraphs of miRNA and disease, the hypergraph convolution is applied as follows.

$$H^{(l+1)} = \sigma\left[D^{-\frac{1}{2}}YWB^{-1}Y^TD^{-\frac{1}{2}}H^{(l)}P^{(l)}\right]$$

where D is the node degree matrix, B is the hyperedge degree matrix, Y is the incidence matrix of miRNA or disease, W is hyperedge weight matrix, $H^{(l)} \in R^{N\times2N}$ is the feature matrix for the l-th layer, $P^{(l)}$ is the l-th trainable parameter matrix and σ is the activation function.

By applying a 2-layer hypergraph convolution, feature matrices $MHF1$ and $DHF1$ of miRNA and disease are generated. At the same time, feature matrices MHF and DHF are embedded by an MLP to $MHF2$ and $DHF2$. Finally, the high-level encoded features for miRNAs and diseases are represented as follows.

$$HGF = \begin{bmatrix}MF\\DF\end{bmatrix} \in R^{(N_m+N_d)\times h}$$
$$MF = \frac{MHF1+MHF2}{2}, DF = \frac{DHF1+DHF2}{2}$$

miRNA-disease interaction encoding. Feature encoding of miRNAs (diseases) by utilizing miRNA-disease interaction information helps obtain inherent representations of miRNAs (diseases), contributing to the accurate identification of miRNA-disease associations. To effectively characterize the miRNA-disease interactions, a heterogeneous graph G_{md} is constructed with nodes representing all the miRNAs and diseases and edges denoting all the positive and negative associations between miRNA and disease in the training set. In this channel, the model is driven to capture association patterns according to the number of associations and the neighbor similarity of a node (an miRNA or a disease) in the heterogeneous graph. Therefore, the node degree encoding and the neighbor similarity encoding are applied in this channel.

It is considered that different attentions should be allocated to nodes with different numbers of associations in the heterogeneous graph. And therefore, a node degree-based encoding module is conducted by calculating the degrees of all nodes in G_{md} and generate a vector $v_d \in R^{(853+591)\times1}$, which is then embedded into a feature matrix $DeF \in R^{(853+591)\times h/2}$ via an MLP.

In terms of neighbor similarity encoding, we first extract all the disease (miRNA) neighbors of each miRNA (disease) in the heterogeneous graph

G_{md}, and then calculate the average similarity of the disease (miRNA) neighbors according to the similarity matrix DS (MS). Suppose that an miRNA m_i has three disease neighbors d_j, d_k, and d_l, then the neighbor similarity $S(m_i)$ of the node m_i is defined as the average similarity of the three diseases based on the similarity matrix DS as follows.

$$S(m_i) = \frac{DS(d_j, d_k) + DS(d_j, d_l) + DS(d_k, d_l)}{3}$$

Therefore, a vector $v_s \in R^{(853+591)\times 1}$ is obtained after completing the computation of all the miRNAs and diseases, which is also projected to another feature matrix $NeF \in R^{(853+591)\times h/2}$ via an MLP. Finally, the high-level encoded miRNA-disease interaction features are generated by concatenating DeF and NeF into $HeF \in R^{(853+591)\times h}$.

Fusion of the tri-channel features. A feature fusion encoder is developed to effectively fuse the three features GF, HGF, and HeF by employing a biased TransFormer encoder and an embedded residual connection as follows.

$$F_{fusion} = TransFormer(F)^{(m)} + U$$
$$F = GF + HGF + HeF$$
$$U = F \otimes sigmoid(F \cdot W_F)$$
$$TransFormer(F)^{(m)} = concat(head_1, \ldots, head_m)$$
$$head_i = softmax\left(\frac{Q_i K_i^T}{\sqrt{d}} + b_i\right) V_i$$
$$\begin{cases} Q_i = F \times W_q{}^i \\ K_i = F \times W_k{}^i \\ V_i = F \times W_v{}^i \end{cases}$$

where W_F, $W_q{}^i$, $W_k{}^i$, $W_v{}^i$, and b_i represent learnable parameter matrices, d is the dimension of Q_i, and $\otimes$ denotes the Hadamard product.

The fused features F_{fusion} of miRNAs and diseases serve as the node representations in the heterogeneous graph G_{md} and a 6-layer graph convolution is performed to complete the encoding of all miRNAs and diseases.

Classification of the miRNA-disease associations. In this study, the miRNA-disease association prediction task is formulated into an edge classification problem in the heterogeneous graph G_{md} with each edge (m_i, d_j) described by a vector e_{ij} = Hadamard product $[F_{fusion}(m_i), F_{fusion}(d_j)]$ and a multi-layer MLP is applied to complete the edge classification.

Statistics and reproducibility

All the experiments, including validation experiments, ablation experiments, interpretation experiments, and case study were conducted based on the HMDDv3.2 dataset, which includes 853 microRNAs and 591 diseases, with a total of 12,446 validated positive associations. We compared our results with another state-of-the-art model (MDformer) using a t-test with $n = 10$ five-fold random cross-validations, obtaining a p-value < 0.01, as detailed in Supplementary Table 2. The code for reproducibility is available at https://doi.org/10.5281/zenodo.13092401[47] and the source data for the figures can be found in the Supplementary Data 1 file.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All samples were obtained from HMDDv3.2 (http://www.cuilab.cn/hmdd). The similarity data for microRNA and diseases are available at https://doi.org/10.5281/zenodo.13092401[47]. In addition, the numerical data used to generate the main figures can be found in the Supplementary Data 1 file.

Code availability

Trifusion is implemented by Python using the PyTorch framework. All supporting source codes can be downloaded from https://doi.org/10.5281/zenodo.13092401[47].

Received: 3 April 2024; Accepted: 13 August 2024;
Published online: 30 August 2024

References

1. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
2. Calin, G. A. & Croce, C. M. MicroRNA signatures in human cancers. *Nat. Rev. Cancer* **6**, 857–866 (2006).
3. Chen, X., Xie, D., Zhao, Q. & You, Z. H. MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* **20**, 515–539 (2019).
4. Huang, L., Zhang, L. & Chen, X. Updated review of advances in microRNAs and complex diseases: towards systematic evaluation of computational models. *Brief. Bioinform.* **23**, bbac407 (2022).
5. Huang, L., Zhang, L. & Chen, X. Updated review of advances in microRNAs and complex diseases: experimental results, databases, webservers and data fusion. *Brief. Bioinform.* **23**, bbac397 (2022).
6. Huang, L., Zhang, L. & Chen, X. Updated review of advances in microRNAs and complex diseases: taxonomy, trends and challenges of computational models. *Brief. Bioinform.* **23**, bbac358 (2022).
7. Chen, X., Liu, M.-X. & Yan, G.-Y. RWRMDA: predicting novel human microRNA–disease associations. *Mol. Biosyst.* **8**, 2792–2798 (2012).
8. Chen, X., Huang, L., Xie, D. & Zhao, Q. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* **9**, 3 (2018).
9. Zeng, X., Liu, L., Lü, L. & Zou, Q. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **34**, 2425–2432 (2018).
10. Zhong, Y. et al. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. *Bioinformatics* **34**, 267–277 (2018).
11. Li, H. Y., You, Z. H., Wang, L., Yan, X. & Li, Z. W. DF-MDA: an effective diffusion-based computational model for predicting miRNA-disease association. *Mol. Ther.* **29**, 1501–1511 (2021).
12. Xu, J. et al. Prioritizing candidate disease mirnas by topological features in the miRNA target–dysregulated network: case study of prostate cancer. *Mol. Cancer Ther.* **10**, 1857–1866 (2011).
13. Chen, X., Gong, Y., Zhang, D. H., You, Z. H. & Li, Z. W. DRMDA: deep representations-based miRNA-disease association prediction. *J. Cell Mol. Med.* **22**, 472–485 (2018).
14. Fu, L. & Peng, Q. A deep ensemble model to predict miRNA-disease association. *Sci. Rep.* **7**, 14482 (2017).
15. Chen, X., Wang, L., Qu, J., Guan, N. N. & Li, J. Q. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* **34**, 4256–4265 (2018).
16. Xuan, P. et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE* **8**, e70204 (2013).
17. Chen, X., Sun, L.-G. & Zhao, Y. NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. *Brief. Bioinform.* **22**, 485–496 (2020).
18. Zhao, Y., Chen, X. & Yin, J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* **35**, 4730–4738 (2019).
19. Lou, Z. et al. Predicting miRNA-disease associations via learning multimodal networks and fusing mixed neighborhood information. *Brief. Bioinform.* **23**, bbac159 (2022).
20. Chen, X. et al. A novel computational model based on super-disease and miRNA for potential miRNA-disease association prediction. *Mol. Biosyst.* **13**, 1202–1212 (2017).

21. Sun, M. et al. Graph convolutional networks for computational drug development and discovery. *Brief. Bioinform.* **21**, 919–935 (2020).
22. Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T. & Peng, J. Identifying drug-target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* **22**, 2141–2150 (2021).
23. Li, Y., Qiao, G., Wang, K. & Wang, G. Drug-target interaction predication via multi-channel graph neural networks. *Brief. Bioinform.* **23**, bbab346 (2022).
24. Peng, W., Tang, Q., Dai, W. & Chen, T. Improving cancer driver gene identification using multi-task learning on graph convolutional network. *Brief. Bioinform.* **23**, bbab432 (2022).
25. Tang, X., Luo, J., Shen, C. & Lai, Z. Multi-view multichannel attention graph convolutional network for miRNA–disease association prediction. *Brief. Bioinform.* **22**, bbab174 (2021).
26. Chu, Y. et al. MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. *Brief. Bioinform.* **22**, bbab165 (2021).
27. Li, J. et al. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* **36**, 2538–2546 (2020).
28. Vaswani, A. et al. in *Proceedings of the 31st International Conference on Neural Information Processing Systems.* p. 6000–6010 (Curran Associates Inc., 2017).
29. Hu, J. et al. DTSyn: a dual-transformer-based neural network to predict synergistic drug combinations. *Brief. Bioinform.* **23**, bbac302 (2022).
30. Li, F., Zhang, Z., Guan, J. & Zhou, S. Effective drug-target interaction prediction with mutual interaction neural network. *Bioinformatics* **38**, 3582–3589 (2022).
31. Zhang, R., Wang, Z., Wang, X., Meng, Z. & Cui, W. MHTAN-DTI: metapath-based hierarchical transformer and attention network for drug-target interaction prediction. *Brief. Bioinform.* **24**, bbad079 (2023).
32. Li, Y., Guo, Z., Wang, K., Gao, X. & Wang, G. End-to-end interpretable disease-gene association prediction. *Brief. Bioinform.* **24**, bbad118 (2023).
33. Gu, P. et al. Multi-head self-attention model for classification of temporal lobe epilepsy subtypes. *Front. Physiol.* **11**, 604764 (2020).
34. Dong, B., Sun, W., Xu, D., Wang, G. & Zhang, T. MDformer: a transformer-based method for predicting miRNA-Disease associations using multi-source feature fusion and maximal meta-path instances encoding. *Comput. Biol. Med.* **167**, 107585 (2023).
35. Huang, Z. et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* **47**, D1013–d1017 (2019).
36. Dong, B., Sun, W., Xu, D., Wang, G. & Zhang, T. DAEMDA: a method with dual-channel attention encoding for miRNA-disease association prediction. *Biomolecules* **13**, 1514 (2023).
37. Zhang, H. et al. Predicting miRNA-disease associations via node-level attention graph auto-encoder. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 1308–1318 (2023).
38. Ning, Q. et al. AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA–disease associations identification. *Brief. Bioinform.* **24**, bbad094 (2023).
39. Wu, Y., Zhu, D., Wang, X. & Zhang, S. An ensemble learning framework for potential miRNA-disease association prediction with positive-unlabeled data. *Comput. Biol. Chem.* **95**, 107566 (2021).
40. Xu, F. et al. dbDEMC 3.0: functional exploration of differentially expressed mirnas in cancers of human and model organisms. *Genomics Proteom. Bioinform.* **20**, 446–454 (2022).
41. Cui, C., Zhong, B., Fan, R. & Cui, Q. HMDD v4.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* **52**, D1327–D1332 (2023).
42. Chang, H. et al. Increased expression of miR-148b in ovarian carcinoma and its clinical significance. *Mol. Med. Rep.* **5**, 1277–1280 (2012).
43. Shang, C., Chen, Q., Zu, F. & Ren, W. Integrated analysis identified prognostic microRNAs in breast cancer. *BMC Cancer* **22**, 1170 (2022).
44. Arrighetti, N. & Beretta, G. L. miRNAs as therapeutic tools and biomarkers for prostate cancer. *Pharmaceutics* **13**, 380 (2021).
45. van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043 (2011).
46. Wang, D., Wang, J., Lu, M., Song, F. & Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **26**, 1644–1650 (2010).
47. Long, S. T. *Zenodo* https://doi.org/10.5281/zenodo.13092401 (2024).

Acknowledgements

This work was supported by the National Key R&D Program of China with code 2020YFA0712400, and the National Natural Science Foundation of China with code 62272268.

Author contributions

J.L. and Z.M. conceived and designed the experiments. S.L. performed the experiments. S.L., X.T., X.S., T.K., Y.Z., C.W., C.Q., Z.M., and J.L. analyzed the data. S.L., X.T., X.S., and T.K. contributed reagents/materials/analysis tools. S.L., Y.Z., and J.L. wrote the paper. S.L. designed the software used in the analysis. J.L. oversaw the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-024-06734-0.

Correspondence and requests for materials should be addressed to Zengchao Mu or Juntao Liu.

Peer review information *Communications Biology* thanks Xing Chen and the other, anonymous, reviewer for their contribution to the peer review of this work. Primary Handling Editors: Laura Rodríguez Pérez.