Molecular Therapy
Nucleic Acids
Original Article

MHGTMDA: Molecular heterogeneous graph transformer based on biological entity graph for miRNA-disease associations prediction

Haitao Zou,[1,2] Boya Ji,[2] Meng Zhang,[3] Fen Liu,[4] Xiaolan Xie,[1] and Shaoliang Peng[2]

[1]Guilin University of Technology, College of Information Science and Engineering, Guilin 541006, China; [2]Hunan University, College of Computer Science and Electronic Engineering, Changsha 410082, China; [3]Xiangya Hospital, The Department of Thoracic Surgery, Changsha 410082, China; [4]Hunan Provincial People's Hospital, Institute of Cardiovascular Epidemiology, Changsha 410082, China

MicroRNAs (miRNAs) play a crucial role in the prevention, prognosis, diagnosis, and treatment of complex diseases. Existing computational methods primarily focus on biologically relevant molecules directly associated with miRNA or disease, overlooking the fact that the human body is a highly complex system where miRNA or disease may indirectly correlate with various types of biomolecules. To address this, we propose a novel prediction model named MHGTMDA (miRNA and disease association prediction using heterogeneous graph transformer based on molecular heterogeneous graph). MHGTMDA integrates biological entity relationships of eight biomolecules, constructing a relatively comprehensive heterogeneous biological entity graph. MHGTMDA serves as a powerful molecular heterogeneity map transformer, capturing structural elements and properties of miRNAs and diseases, revealing potential associations. In a 5-fold cross-validation study, MHGTMDA achieved an area under the receiver operating characteristic curve of 0.9569, surpassing state-of-the-art methods by at least 3%. Feature ablation experiments suggest that considering features among multiple biomolecules is more effective in uncovering miRNA-disease correlations. Furthermore, we conducted differential expression analyses on breast cancer and lung cancer, using MHGTMDA to further validate differentially expressed miRNAs. The results demonstrate MHGTMDA's capability to identify novel MDAs.

INTRODUCTION

MicroRNAs (miRNAs) are non-coding RNAs that are composed of about 22 nucleotides.[1] They participate in the process of biological cell growth, differentiation, and apoptosis by regulating target genes.[2] In recent years, many researchers have shown that aberrant expression of miRNAs is closely related to the occurrence of diseases.[3] In-depth research on the association between miRNAs and diseases can uncover the potential role of miRNAs in disease pathogenesis and aid in the exploration of potential biomarkers.[4] This, in turn, offers new perspectives and methods for early disease diagnosis and personalized treatment.[5] In addition, miRNAs can be also used as drug targets to open up new avenues for disease treatment.[6] Thus, investigating the associations between miRNAs and diseases holds significant theoretical and clinical implications.

Currently, computational methods for predicting possible miRNA-disease connections fall into two broad categories: similarity-based methods and machine learning-based methods.[7–10] For similarity-based methods, they consider the assumption that functionally similar miRNAs are generally associated with phenotypically similar diseases.[11,12] For example, Shi et al.[13] employed a random walk-based technique to map disease-causing genes and miRNA target genes onto a PPI similarity network to discover possible correlations between miRNAs and diseases. Xuan et al.[14] proposed a weighted k-nearest neighbor-based approach to predict miRNA-disease associations by allocating greater weights to clustering of the same miRNA family. Ha et al.[15] utilized a global similarity network based on environmental factors and known miRNA-disease associations to predict potential associations between them. In general, these models based on similarity metrics are overly dependent on similarity scores and have limitations in predicting miRNA-disease associations.

On the other hand, with artificial intelligence technology becoming more widely used, more and more machine learning approaches are being applied to the field of miRNA-disease association prediction.[16,17] For example, Zhou et al.[18] proposed a gradient-enhanced decision tree combined with a logistic regression model to predict miRNA-disease associations. Chen et al.[19] proposed a decision tree-based model for predicting miRNA-disease associations. This was achieved by reducing the dimensionality of the miRNA and disease feature subsets through principal-components analysis. Peng

Received 28 November 2023; accepted 31 January 2024;
https://doi.org/10.1016/j.omtn.2024.102139.

Correspondence: Boya Ji, Hunan University, College of Computer Science and Electronic Engineering, Changsha 410082, China.
E-mail: byj@hnu.edu.cn
Correspondence: Xiaolan Xie, Guilin University of Technology, College of Information Science and Engineering, Guilin 541006, China.
E-mail: xie_xiao_lan@foxmail.com
Correspondence: Shaoliang Peng, Hunan University, College of Computer Science and Electronic Engineering, Changsha 410082, China.
E-mail: slpeng@hnu.edu.cn

et al.[20] introduced a convolutional neural network approach to predict associations between miRNAs and diseases. This approach leveraged a self-encoder to capture pertinent features shared between miRNAs and diseases. Li et al.[21] introduced the Matrix Completion for MiRNA-Disease Association prediction (MCMDA) model using the singular value thresholding technique. To generate the ultimate miRNA-disease association matrix, they applied the matrix completion algorithm to modify the miRNA-disease adjacency matrix. In summary, machine learning-based prediction models for miRNA-disease associations are highly efficient, significantly reducing computational costs. However, the quality of feature extraction by the model has a significant impact on the predictive outcomes.

Despite the numerous methods proposed for predicting miRNA-disease associations (MDAs), most have focused on analyzing individual molecules, overlooking the intricate correlations between different molecules. Additionally, various molecules, such as proteins and genes, offer multi-level, multi-perspective biological information, aiding in a more in-depth exploration of the potential mechanisms underlying MDAs. This diversity contributes to providing richer input features for prediction models. Building upon this concept, we propose a prediction model named MHGTMDA, as shown in Figure 1, which utilizes a molecular heterogeneous graph transformer. MHGTMDA integrates biological entity relationships of eight major biomolecules, constructing a relatively comprehensive heterogeneous biological entity graph. Serving as a powerful molecular heterogeneous graph extractor, MHGTMDA can extract graph structural elements of miRNA and disease. By combining these elements with their prior attribute information, it detects potential correlations. MHGTMDA's contribution is categorized into the following points.

(1) We collected eight types of biological entities and their relationships, constructing a relatively comprehensive heterogeneous biological entity graph. Utilizing a molecular heterogeneous graph transformer, we extracted graph structural features of miRNA and disease. These features were then combined with prior attribute characteristics to achieve enhanced predictive performance.
(2) The MHGTMDA method, when compared with the current state-of-the-art approaches, demonstrates outstanding performance. Case analyses also thoroughly showcase the exceptional robustness of our model.
(3) The MHGTMDA method is an automated, accessible, and open-source tool for bioinformatic researchers, which is freely available at https://github.com/zht-code/HGTMDA.git.

RESULTS

In this section, we conducted a series of experiments to test the comprehensive performance of MHGTMDA on the benchmark dataset Human MicroRNA Disease Database (HMDD) v3.2.

Performance evaluation under 5-fold cross-validation

We used a 5-fold cross-validation strategy to evaluate the generalization ability of our model (MHGTMDA). In the results, we plot the receiver operating characteristic curves (ROCs) and precision-recall curves (PRCs) as shown in Figures 2 and3. Furthermore, the area under the ROCs (AUC) was also used to measure the ability of MHGTMDA. The area under the PRCs (AUPR) was used to represent the relationship between precision and recall criteria of MHGTMDA. We also computed the Matthews correlation coefficient (MCC) to evaluate our models. As shown in Table 1, the average AUC value of MHGTMDA under 5-fold cross-validation reaches 95.4%, which demonstrates the high level of accuracy and robustness of our proposed model.

Parameter analysis

To optimize the performance of our model, we conducted a parameter analysis, with a particular focus on two crucial parameters: embedding size and the number of layers in the multi-layer perceptron (MLP). This in-depth analysis aims to gain a comprehensive understanding of the performance of our proposed model under various parameter configurations, revealing its sensitivity and robustness to input data.

Impact of embedding size

The embedding size represents the dimensionality of our input data in the model. By systematically analyzing the impact of different embedding sizes on model performance, we gain a better understanding of the model's sensitivity to input features. We explored a series of embedding sizes, ranging from small to large, observing the model's ability to learn complex features and generalize across them. The experimental results, as shown in Table S2, indicate that increasing the embedding size appropriately can enhance the model's performance. However, when the embedding size becomes excessively large, the model tends to overfit the training data, resulting in a decrease in performance on the test set. Through detailed experiments and analysis, we set the embedding size to 901 to achieve the optimal balance between performance and computational efficiency.

Impact of MLP layers

The selection of the number of layers in the MLP, as a core component of deep learning models, directly relates to the depth and complexity of the model. By adjusting the number of MLP layers, we explored the model's performance across different layer configurations. As shown in Table 2, increasing the number of layers allowed us to observe a corresponding enhancement in the model's expressive capacity for data representation. This increase in layers facilitated better capturing of advanced features within the data. Notably, our model achieved peak performance when the number of layers was set to four. Consequently, we set the number of MLP layers in our model to four, based on these findings.

Ablation experiment

To further demonstrate the effectiveness of MHGTMDA, we conducted two sets of ablation experiments, removing attribute features and structural features, respectively, to compare the effects with MHGTMDA

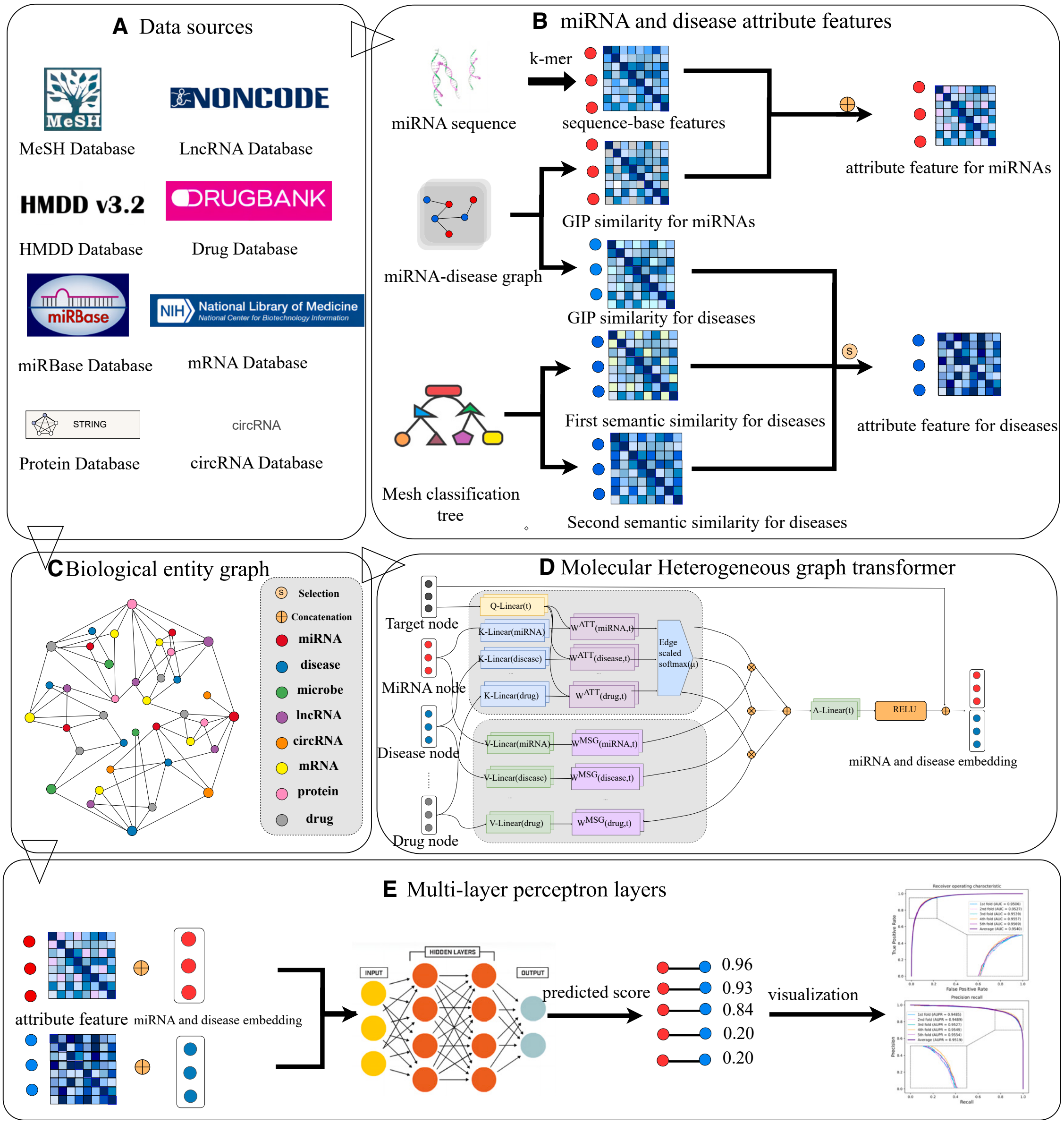

Figure 1. Overall architecture of MHGTMDA

(A) Data sources for MHGTMDA. (B) The integrated miRNA sequence and similarity network and the integrated disease similarity network were constructed respectively to extract the inherent attribute features of both. (C) The sub-module constructed biological entity graphs, including miRNA, disease, microbe, lncRNA, circRNA, MRNA, protein, and drug. (D) The sub-module mainly extracts the embedding feature of miRNA and disease in biological entity graphs. (E) The multimodal embedding representations of miRNAs and diseases were concatenated and fed into the MLP for training and prediction.

under 5-fold cross-validation experiment. First, we eliminated the attribute features and only used the biological entity graph for feature extraction by molecular heterogeneous graph transformer for potential MDA prediction. Second, we eliminated the biological entity graph and only used the attribute features containing fused Gaussian interaction profile (GIP) for MDA prediction. Finally, we fused the attribute features and biological entity graph for feature fusion and then performed miRNA-disease association prediction (MHGTMDA). The comparative results as shown in Figure 4 indicate that our proposed MHGTMDA fusing the attribute feature and biological entity graph structure feature is better in predicting potential MDAs.

Performance comparison with the state-of-the-art methods

In this section, we further compare MHGTMDA with a number of existing methods, including the latest and state-of-the-art methods. More specifically, we compare MHGTMDA with ERMDA,[22]

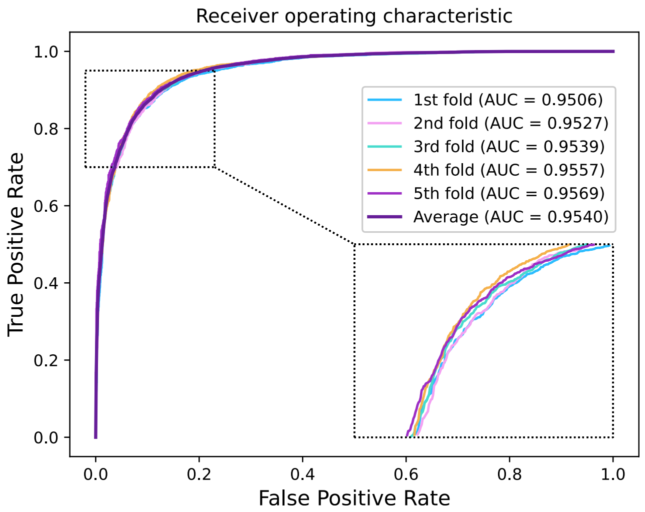

Figure 2. The 5-fold cross-validation ROC curves of MHGTMDA

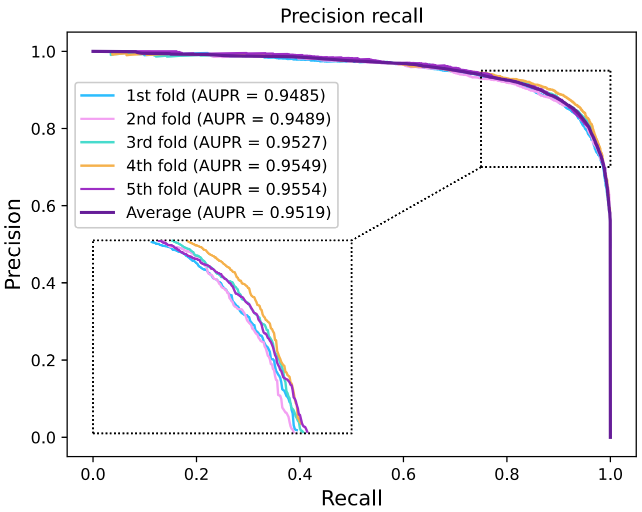

Figure 3. The 5-fold cross-validation PR curves of MHGTMDA

AGAEMD,[23] PATMDA,[24] NIMCGCN,[25] and MAGCN[26] on the basis of the same dataset. We have performed six evaluation indicators to compare from different perspectives as shown in Table S1, in which AUC is the area under the ROC curve and AUPRC is the area under the PR curve. The results demonstrate that MHGTMDA outperforms these competitors and improves the accuracy and AUC by at least 3.45% and 2.12%, respectively.

Case study

Breast cancer stands as one of the most common and deadliest cancers globally. Research indicates that the differential expression of certain miRNAs influences the occurrence, progression, and prognosis of breast cancer.[27] We conducted a case study focusing on breast cancer. Some researchers have found that miRNAs such as miR-31 can inhibit the renewal and development of breast cancer stem cells.[28] Additionally, miR-221 has been identified to be overexpressed in breast cancer, and its overexpression correlates with the malignancy of breast cancer.[29] The aforementioned evidence suggests that miRNAs can serve as biomarkers for breast cancer, and the discovery of potential miRNAs provides new targets for the treatment of breast cancer.

To further validate the effectiveness of MHGTMDA in practical applications, we randomly selected the top 50 miRNAs predicted by MHGTMDA that are associated with breast cancer. Subsequently, we verified these miRNAs in the dbDEMC database[30] and found that out of the 50 miRNAs, 47 received support from the latest literature (as shown in Table S3). Our case study demonstrates the practicality and effectiveness of the proposed MHGTMDA method.

To validate the correlation between the model's predictions for specific diseases and differentially expressed miRNAs, we conducted a differential expression analysis on breast cancer. Utilizing the limma package in R, we performed a differential expression analysis on the miRNA expression profile of breast cancer GSE118782, as illustrated in Figure 5. The analysis identified 41 differentially expressed miRNAs, as depicted in Figure 5A. Among these, nine miRNAs were not present in the dbDEMC database, making it impossible to establish effective judgments. Consequently, we proceeded to predict the remaining 32 miRNAs. Upon assessment, 68% (22 of 32) of the miRNAs were found to be associated with breast cancer, and these 22 miRNAs were subsequently validated through the dbDEMC database.

Lung cancer, being one of the deadliest cancers globally, has consistently drawn the attention of scientists due to its high mortality rate. Early diagnosis and precise treatment have been focal points of research. miRNA, as a crucial class of non-coding RNA, plays a key role in regulating gene expression, contributing significantly to cellular processes. Recent studies suggest that the aberrant expression of certain miRNAs may be associated with the occurrence, development, and metastasis of lung cancer. Therefore, we conducted a differential expression analysis on the miRNA expression profile of lung adenocarcinoma. As shown in Figure 6, the analysis identified 50 differentially expressed miRNAs. We utilized the MHGTMDA model to predict these 50 differentially expressed miRNAs. Upon evaluation, it was found that 68% (34 of 50) of the miRNAs were associated with lung adenocarcinoma. Subsequently, these 34 miRNAs were validated through the dbDEMC database (as shown in Table S4).

DISCUSSION

In this study, we propose a novel prediction model named MHGTMDA (miRNA and disease association prediction using heterogeneous graph transformer based on molecular heterogeneous graph). MHGTMDA integrates biological entity relationships of eight biomolecules, constructing a relatively comprehensive heterogeneous biological entity graph. Serving as a powerful molecular heterogeneous graph transformer, MHGTMDA extracts graph structural elements of miRNA

Table 1. The 5-fold cross-validation performance of MHGTMDA (s.t.d: standard deviation)

Fold	Accuracy	Sensitive	Specificity	Precision	MCC	AUC
1st	0.881	0.8757	0.8862	0.8842	0.762	0.9506
2ND	0.8833	0.8816	0.8849	0.8813	0.7665	0.9527
3rd	0.8857	0.8829	0.8886	0.8905	0.7715	0.9539
4th	0.8928	0.8979	0.8875	0.893	0.7855	0.9557
5th	0.8872	0.8809	0.8933	0.889	0.7744	0.9569
Average	**0.8927**	**0.8838**	**0.8881**	**0.8926**	**0.772**	**0.9551**
s.t.d	**0.0045**	**0.0083**	**0.0032**	**0.0048**	**0.0089**	**0.0025**

and disease, combining their attribute information to detect potential correlations. In a 5-fold cross-validation study, MHGTMDA achieved an AUC of 0.9569, surpassing state-of-the-art methods by at least 3%. Feature ablation experiments suggest that considering features among multiple biomolecules is more effective in uncovering miRNA-disease correlations. Furthermore, we conducted differential expression analyses on breast cancer and lung cancer, using MHGTMDA to further validate differentially expressed miRNAs. The results demonstrate MHGTMDA's capability to identify novel MDAs.

MATERIALS AND METHODS

Datasets

We utilized the MDA database from the HMDD v3.2[31] as our training data, consisting of 35,547 experimentally validated MDAs. We selected 901 miRNAs and 877 diseases from this dataset, thereby constructing the association matrix $\in R_{901\times 877}$ $A \in R_{901\times 877}$. Additionally, we obtained 3,348 protein nodes from the STRING database,[32] 2,633 lncRNA nodes from NONCODEV5,[33] 421 circRNA nodes from CircBase,[34] 1,319 drug nodes from DrugBank,[35] 3,024 mRNA nodes from the NCBI database,[36] and 100 microbe nodes from the NIH Medical Subject Headings (MeSH) database,[37] thus forming the biological entity graph. Finally, an equal number of non-MDAs were randomly selected as negative controls.

Biological entity graph construction

We constructed the biological entity graph consisting of eight types of biological entities together with 16 types of associations as shown in Figure 1C, including circRNA-disease associations, circRNA-miRNA associations, disease-mRNA associations, disease-microbe associations, drug-disease associations, drug-mRNA associations, drug-microbe associations, drug-protein associations, lncRNA-disease associations, lncRNA-mRNA associations, lncRNA-miRNA associations, lncRNA-protein associations, miRNA-drug associations, miRNA-mRNA associations, miRNA-protein associations, and mRNA-protein associations. Note that the graph does not contain known MDAs in the training set to avoid label leakage. Specifically, we constructed the graph utilizing the torch geometric tool. First, we enter the collected biological entities into HeteroData as nodes (HeteroData is a PyG built in data structure for representing heterogeneous graphs). Next, we constructed node mappings by different node types to construct edge indexes in HeteroData. Finally, we constructed node type labels to represent the type of each node in HeteroData.

Sequence feature of miRNAs

Based on the hypothesis that the higher the sequence similarity of two miRNAs, the higher the likelihood that they have the same function,[38] we downloaded the sequences of 901 miRNAs from miRBase and calculated the sequence feature of miRNAs by comparing the sequence of each pair of miRNAs. Based on previous studies,[39,40] we used sequence alignment to quantify the similarity between miRNA m_a and miRNA m_b. In addition, we used the min-max normalization function to normalize the sequence similarity between miRNAs. The formula is shown below:

$$M(m_a, m_b) = \frac{Score(m_a, m_b) - Score_{min}}{Score_{max} - Score_{min}} \qquad \text{(Equation 1)}$$

where Score $_{min}$ and Score $_{max}$ respectively represent the maximum and minimum similarity scores of all miRNA sequence pairs.

Semantic similarity of diseases

We obtained MeSH descriptors for numerous diseases from the National Library of Medicine and represented the complicated interactions between diseases using a directed acyclic graph (DAG). Each MeSH descriptor in a DAG is linked to another via an edge from the parent node to the child node. Each MeSH descriptor has one

Table 2. Performance analysis on MLP layers

MLP layers	Accuracy	Precision	Recall	F1-score	AUC	AUPRC
Layer = 1	0.8742	0.8743	0.8745	0.8742	0.9412	0.9351
Layer = 2	0.8851	0.8850	0.8851	0.8850	0.9464	0.9423
Layer = 3	0.8889	0.8888	0.8889	0.8888	0.9516	0.9496
Layer = 4	**0.8927**	**0.8926**	**0.8925**	**0.8926**	**0.9559**	**0.9551**

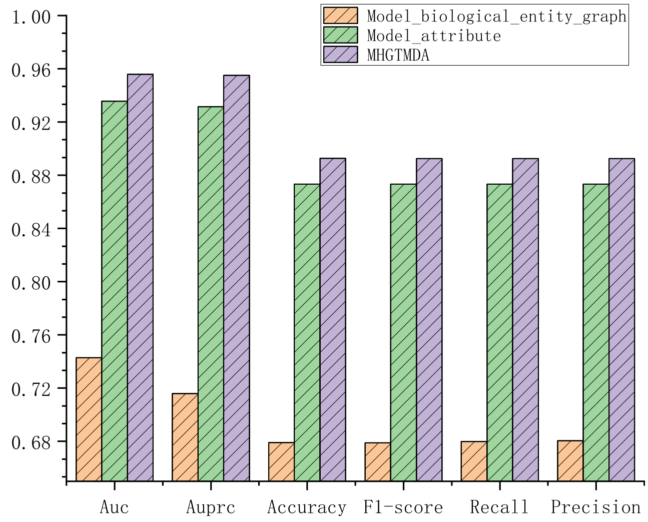

Figure 4. Ablation experiments with different features of MHGTMDA

Model_biological_entity_graph is a performance analysis that eliminates the attribute features only keeping the biological entity graph. Model_attribute is a performance analysis that eliminates the biological entity graph only keeping attribute features. MHGTMDA is a performance analysis that integrates attribute characterization and biological entity graphs.

or more tree numbers and is stored in the DAG as a number. The DAG of illness A, for example, is represented as $DAG(A) = (D(A), E(A))$, where $D(A)$ signifies the disease and its ancestors, and $E(A)$ defines the link between two nodes.

We compute the illness semantic similarity based on the aforementioned definitions. Specifically, we define the semantic contribution of a disease term t in $DAG(A)$ to disease A as follows:

$$\begin{cases} DAG_A = 1 \text{ if } t = A \\ DAG_A(t) = max\{\Delta^* DAG_A(t') | t' \in \text{children of } t\} \text{if } t \neq A \end{cases} \quad \text{(Equation 2)}$$

where *Δ* represents the semantic contribution decay factor, which decreases as the distance between a disease and its ancestor node increases. Therefore, the semantic value of disease at the first level is 1. In order to distinguish the semantic contribution of different levels of diseases to disease A, the semantic contribution value is obtained by multiplying the contribution of different levels of diseases by the semantic contribution attenuation factor.

In addition, the semantic value of a disease is defined as follows:

$$\mathrm{SC}_A = \sum_{t \in \mathrm{D}(A)} DAG_A(t) \quad \text{(Equation 3)}$$

Therefore, the semantic similarity between different diseases SC_A and SC_B is defined as follows:

$$SS(A,B) = \frac{\sum_{t \in \mathrm{D}(A) \cap \mathrm{D}(B)} (DAG_A(t) + DAG_B(t))}{SC_A + SC_B} \quad \text{(Equation 4)}$$

GIP kernel similarity for miRNAs and diseases

We calculated the GIP kernel similarity between miRNA and disease. First, the disease GIP similarity $GIP_d(\,d_a, d_b\,)$ between disease d_a and disease d_b was computed by:

$$GIP_d(d_a, d_b) = \exp(-\,\gamma_d * \| IP(d_a) - IP(d_b) \|^2) \quad \text{(Equation 5)}$$

where $IP(\,\cdot\,)$ represents the binary vector and γ_d represents the kernel bandwidth, which was calculated by normalizing the original parameter γ'_d as follows:

$$\gamma_d = \frac{\gamma'_d}{\left(\frac{1}{nd} \sum_{i=1}^{nd} \| IP(d_a) \|^2 \right)} \quad \text{(Equation 6)}$$

Similarly, the miRNA GIP similarity m (a, b) between miRNA *a* and miRNA *b* was computed by:

$$GIP_m(m_a, m_b) = \exp\left(-\,\gamma_m * \| IP(m_a) - IP(m_b) \|^2\right) \quad \text{(Equation 7)}$$

$$\gamma_m = \frac{\gamma'_m}{\left(\frac{1}{nm} \sum_{i=1}^{nm} \| IP(m_a) \|^2 \right)} \quad \text{(Equation 8)}$$

Integrated GIP features and attribute features of miRNAs and diseases

We integrated the original features of miRNA and diseases with their corresponding GIP features to obtain comprehensive attribute characteristics for both miRNA and diseases. For miRNAs, we constructed $SM(m_a, m_b)$ as the fused miRNA similarity matrix. It can be described as follows:

$$SM(m_a, m_b) = \begin{cases} M(m_a, m_b) \text{ if } (m_a, m_b) \text{ in M} \\ GIP_m(m_a, m_b) \text{ others} \end{cases} \quad \text{(Equation 9)}$$

Similarly, we constructed $SD(d_a, d_b)$ as the fused disease similarity matrix. it can be described as follows:

$$SD(d_a, d_b) = \begin{cases} SS(A,B) \text{ if } (d_a, d_b) \text{ in SS(A, B)} \\ GIP_d(d_a, d_b) \text{ others} \end{cases} \quad \text{(Equation 10)}$$

Molecular heterogeneous graph transformer

In this section, we present the detailed steps to build a molecular heterogeneous graph transformer (HGT).[41] First, the whole HGT is

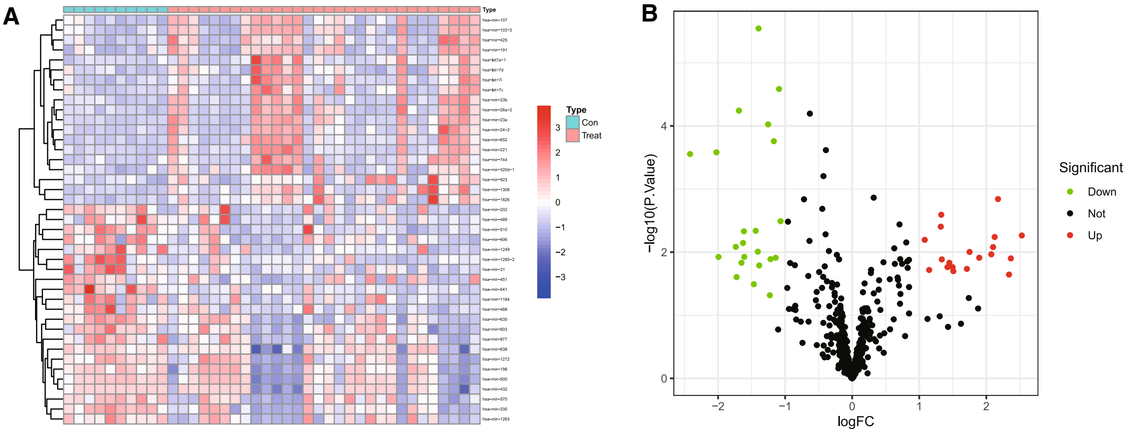

Figure 5. Differentially expressed RNA(DERs) in breast cancer samples

(A) Heatmap of the 42 DERs in breast cancer samples. Red blocks indicate high-expression RNA, and blue blocks indicate low-expression RNA. (B) Volcano plot of the 42 DERs. Red dots represent significantly up-regulated genes, and blue dots represent significantly down-regulated genes.

divided into three modules: heterogeneous mutual attention, heterogeneous message passing, and target-Specific aggregation.

Heterogeneous mutual attention

Unlike the traditional transformer model, the HGT model assigns a unique set of projection weights W to each meta-relationship, whereas in the traditional transformer, all words share the same set of weights. With a narrower focus, we apply a linear projection technique, denoted as Q-Linear$^i_{\tau(t)}$, to transform the target node t into the i-th Query vector.

$$Attention_{HGT}(s,e,t) = \underset{\forall s\in N(t)}{\mathrm{Softmax}}\left(\underset{i\in[1,h]}{\Vert} ATT\text{-head}^i(s,e,t)\right) \qquad \text{(Equation 11)}$$

$$\begin{aligned} ATT\text{-head}^i(s,e,t) &= \left(K^i(s)W^{\mathrm{ATT}}_{\varphi(e)}Q^i(t)^T\right)\cdot\frac{\mu_{\langle\tau(s),\varphi(e),\tau(t)\rangle}}{\sqrt{d}} \\ K^i(s) &= K\text{-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right) \\ Q^i(t) &= \text{Q-Linear}^i_{\tau(t)}\left(H^{(l-1)}[t]\right) \end{aligned} \qquad \text{(Equation 12)}$$

where ATT-head denotes the i-th attention head, K(s) represents the i-th Key vector that source nodes is projected into, Q(t) represents the i-th Query vector that targets node t is projected into, e represents the relationship between the source node s and the target node t, and μ(·) denotes the general significance of each relational ternary as a means of adaptive scaling of the attention. The operation of Attention(·) mainly consists of connecting the h ATT-heads connections to get the attention vector of each node pair (s,t). At its fundamental level, the process involves executing another softmax function, leading to the generation of a probability distribution specific to each target node t. This distribution is formed by utilizing the attention vectors accumulated from the adjacent nodes N(t).

Heterogeneous message passing

The meta-relationship of edges is introduced to the message-passing mechanism in this module to alleviate discrepancies in the distribution of different types of nodes and edges.

$$\begin{aligned} Message_{HGT}(s,e,t) &= \underset{i\in[1,h]}{\Vert} \text{MSG-head}^i(s,e,t) \\ \text{MSG-head}^i(s,e,t) &= \text{M-Linear}^i_{\tau(s)}\left(H^{(l-1)}[s]\right)W^{\mathrm{MSG}}_{\varphi(e)} \end{aligned} \qquad \text{(Equation 13)}$$

Target-Specific aggregation

In this module, first, the above heterogeneous mutual attention and heterogeneous message passing are aggregated from the source node to the target node, and the above steps can be shown as:

$$\tilde{H}^{(l)}[t] = \underset{\forall s\in N(t)}{\bigoplus}\left(\text{Attention}_{HGT}\ (s,e,t)\cdot\text{Message}_{HGT}\ (s,e,t)\right) \qquad \text{(Equation 14)}$$

Then, the results obtained above are input into the ELU activation layer, and then the output is linearly transformed and residuals are connected, and the above steps can be shown as:

$$H^{(l)}[t] = \sigma\left(\text{A-linear}_{\tau(t)}\tilde{H}^{(l)}[t]\right) + H^{(l-1)}[t] \qquad \text{(Equation 15)}$$

Finally, the extraction of embedding for the biological entity graph part is completed.

DATA AND CODE AVAILABILITY

The datasets and source codes used in this study are freely available at https://github.com/zht-code/HGTMDA.git.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2024.102139.

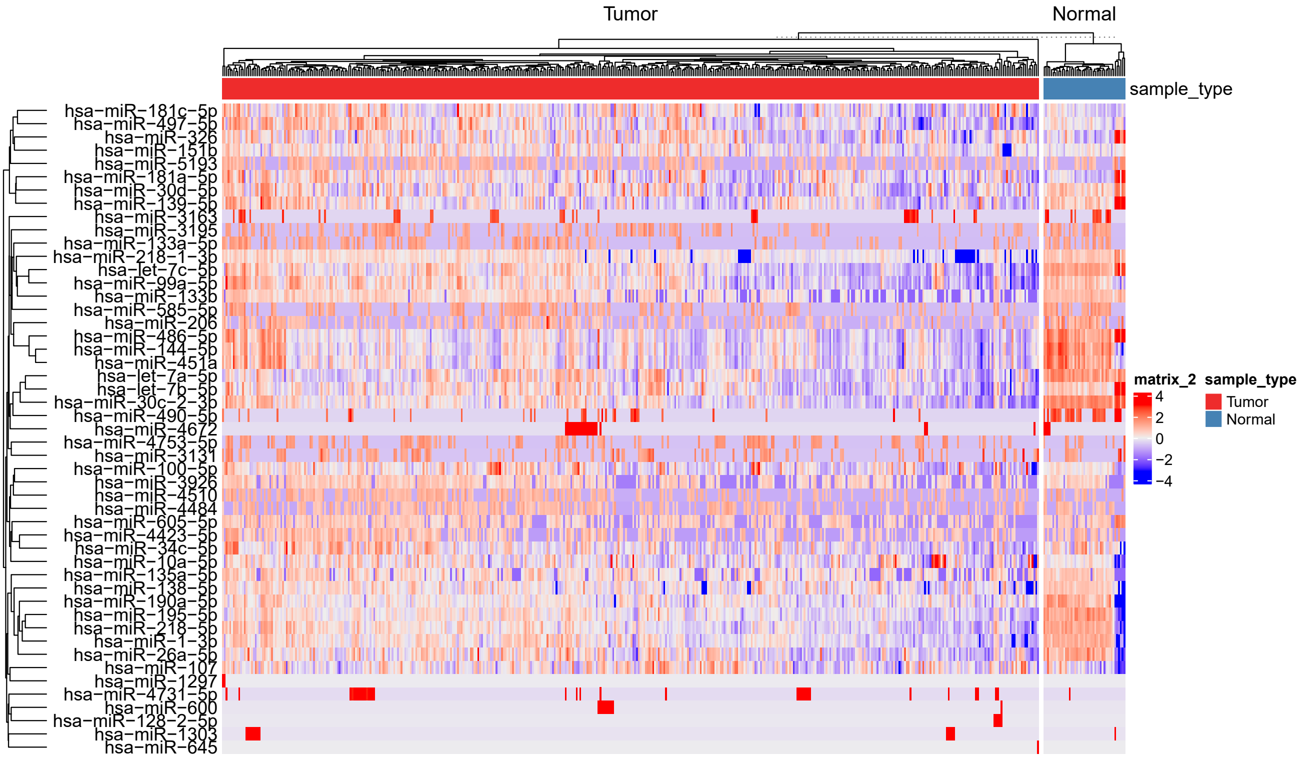

Figure 6. Differentially expressed RNA(DERs) in lung cancer samples

Heatmap of the 50 DERs in lung cancer samples. Red blocks indicate high-expression RNA, and blue blocks indicate low-expression RNA.

ACKNOWLEDGMENTS

This work was supported by NSFC-FDCT Grants 62361166662; National Key R&D Program of China 2023YFC3503400, 2022YFC3400400; Key R&D Program of Hunan Province 2023GK2004, 2023SK2059, 2023SK2060; Top 10 Technical Key Project in Hunan Province 2023GK1010; Key Technologies R&D Program of Guangdong Province (2023B1111030004 to F.F.H.); funds from State Key Laboratory of Chemo/Biosensing and Chemometrics, the National Supercomputing Center in Changsha (http://nscc.hnu.edu.cn/), Peng Cheng Lab, and the Guangxi Key Research and Development Program (No.2022AB43023).

AUTHOR CONTRIBUTIONS

These authors contributed equally to this work. H.T.Z. and B.Y.J. were responsible for the experimental design and wrote the paper. M.Z. and F.L. analyzed and interpreted the data. S.L.P. and X.L.X. provided critical advice on the experimental design and participated in proofreading the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Ambros, V. (2001). microRNAs: tiny regulators with great potential. Cell *107*, 823–826.
2. Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nat. Rev. Genet. *16*, 421–433.
3. Yoshino, H., Seki, N., Itesako, T., Chiyomaru, T., Nakagawa, M., and Enokida, H. (2013). Aberrant expression of microRNAs in bladder cancer. Nat. Rev. Urol. *10*, 396–404.
4. Zhou, S.S., Jin, J.P., Wang, J.Q., Zhang, Z.G., Freedman, J.H., Zheng, Y., and Cai, L. (2018). miRNAS in cardiovascular diseases: potential biomarkers. Acta Pharmacol. Sin. *39*, 1073–1084.
5. Mishra, P.J., and Bertino, J.R. (2009). MicroRNA polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine. Pharmacogenomics *10*, 399–416. https://doi.org/10.2217/14622416.10.3.399.
6. Quiat, D., and Olson, E.N. (2013). MicroRNAs in cardiovascular disease: from pathogenesis to prevention and treatment. J. Clin. Invest. *123*, 11–18.
7. Wong, L., Wang, L., You, Z.H., Yuan, C.A., Huang, Y.A., and Cao, M.Y. (2023). GKLOMLI: a link prediction model for inferring miRNA–lncRNA interactions by using Gaussian kernel-based method on network profile and linear optimization algorithm. BMC Bioinf. *24*, 188.
8. Zheng, K., Zhang, X.L., Wang, L., You, Z.H., Ji, B.Y., Liang, X., and Li, Z.W. (2023). SPRDA: a link prediction approach based on the structural perturbation to infer disease-associated Piwi-interacting RNAs. Briefings Bioinf. *24*, bbac498.
9. Chu, Y., Wang, X., Dai, Q., Wang, Y., Wang, Q., Peng, S., Wei, X., Qiu, J., Salahub, D.R., Xiong, Y., and Wei, D.Q. (2021). MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. Briefings Bioinf. *22*, bbab165.
10. Wang, W., Dai, Q., Li, F., Xiong, Y., and Wei, D.Q. (2021). MLCDForest: multi-label classification with deep forest in disease prediction for long non-coding RNAs. Briefings Bioinf. *22*, bbaa104.

11. Su, X., Hu, L., You, Z., Hu, P., Wang, L., and Zhao, B. (2022). A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. Briefings Bioinf. *23*, bbab526.

12. Wang, L., You, Z.H., Huang, D.S., and Li, J.Q. (2023). MGRCDA: metagraph recommendation method for predicting CircRNA-disease association. IEEE Trans. Cybern. *53*, 67–75.

13. Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., Zhao, Z., Jiang, W., Guo, Z., and Li, X. (2013). Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst. Biol. *7*, 101–112.

14. Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., Liu, Y., Dai, Q., Li, J., Teng, Z., and Huang, Y. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. PLoS One *8*, e70204.

15. Ha, J., Kim, H., Yoon, Y., and Park, S. (2015). A method of extracting disease-related microRNAs through the propagation algorithm using the environmental factor based global miRNA network. Bio Med. Mater. Eng. *26*, S1763–S1772.

16. Wang, L., Wong, L., You, Z.H., and Huang, D.S. (2023). AMDECDA: Attention mechanism combined with data ensemble strategy for predicting CircRNA-disease association. IEEE Transactions on Big Data (IEEE), pp. 1–11. https://doi.org/10.1109/TBDATA.2023.3334673.

17. Dai, Q., Chu, Y., Li, Z., Zhao, Y., Mao, X., Wang, Y., Xiong, Y., and Wei, D.Q. (2021). MDA-CF: predicting miRNA-disease associations based on a cascade forest model by fusing multi-source information. Comput. Biol. Med. *136*, 104706.

18. Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. Comput. Biol. Chem. *85*, 107200.

19. Chen, X., Zhu, C.C., and Yin, J. (2019). Ensemble of decision tree reveals potential miRNA-disease associations. PLoS Comput. Biol. *15*, e1007209.

20. Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., Shang, X., and Wei, Z. (2019). A learning-based framework for miRNA-disease association identification using neural networks. Bioinformatics *35*, 4364–4371.

21. Li, J.Q., Rong, Z.H., Chen, X., Yan, G.Y., and You, Z.H. (2017). MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget *8*, 21187–21199.

22. Dai, Q., Wang, Z., Liu, Z., Duan, X., Song, J., and Guo, M. (2022). Predicting miRNA-disease associations using an ensemble learning framework with resampling method. Briefings Bioinf. *23*, bbab543.

23. Zhang, H., Fang, J., Sun, Y., Xie, G., Lin, Z., and Gu, G. (2023). Predicting miRNA-disease associations via node-level attention graph auto-encoder. IEEE ACM Trans. Comput. Biol. Bioinf *20*, 1308–1318.

24. Xie, X., Wang, Y., He, K., and Sheng, N. (2023). Predicting miRNA-disease associations based on PPMI and attention network. BMC Bioinf. *24*, 113–119.

25. Li, J., Zhang, S., Liu, T., Ning, C., Zhang, Z., and Zhou, W. (2020). Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics *36*, 2538–2546.

26. Zhan, Q., Wu, G., and Gan, C. (2021). Magcn: A multi-adaptive graph convolutional network for traffic forecasting. In 2021 International Joint Conference on Neural Networks (IJCNN) (IEEE), pp. 1–8.

27. Sun, Y.S., Zhao, Z., Yang, Z.N., Xu, F., Lu, H.J., Zhu, Z.Y., Shi, W., Jiang, J., Yao, P.P., and Zhu, H.P. (2017). Risk factors and preventions of breast cancer. Int. J. Biol. Sci. *13*, 1387–1397.

28. Li, L., Yuan, L., Luo, J., Gao, J., Guo, J., and Xie, X. (2013). MiR-34a inhibits proliferation and migration of breast cancer through down-regulation of Bcl-2 and SIRT1. Clin. Exp. Med. *13*, 109–117.

29. Si, M.L., Zhu, S., Wu, H., Lu, Z., Wu, F., and Mo, Y.Y. (2007). miR-21-mediated tumor growth. Oncogene *26*, 2799–2803.

30. Xu, F., Wang, Y., Ling, Y., Zhou, C., Wang, H., Teschendorff, A.E., Zhao, Y., Zhao, H., He, Y., Zhang, G., and Yang, Z. (2022). dbDEMC 3.0: functional exploration of differentially expressed miRNAs in cancers of human and model organisms. Dev. Reprod. Biol. *20*, 446–454.

31. Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., Zhou, Y., and Cui, Q. (2019). HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Res. *47*, D1013–D1017.

32. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

33. Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. *46*, D308–D314.

34. Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. Rna *20*, 1666–1670.

35. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.

36. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

37. Lipscomb, C.E. (2000). Medical subject headings (MeSH). Bull. Med. Libr. Assoc. *88*, 265–266.

38. Qu, J., Chen, X., Sun, Y.Z., Li, J.Q., and Ming, Z. (2018). Inferring potential small molecule–miRNA association based on triple layer heterogeneous network. J. Cheminf. *10*, 1–14.

39. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

40. Ji, B.Y., You, Z.H., Cheng, L., Zhou, J.R., Alghazzawi, D., and Li, L.P. (2020). Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. Sci. Rep. *10*, 6658.

41. Hu, Z., Dong, Y., Wang, K., and Sun, Y. (2020). Heterogeneous graph transformer. In Proceedings of the Web Conference 2020, pp. 2704–2710.