TYPE Methods
PUBLISHED 30 May 2023
DOI 10.3389/fgene.2023.1201934

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
Wei Liu,
Xiangtan University, China
Yajie Meng,
Wuhan Textile University, China

*CORRESPONDENCE
Xu Zhou,
✉ zhouxu2008@sina.com

RECEIVED 07 April 2023
ACCEPTED 17 May 2023
PUBLISHED 30 May 2023

CITATION
Niu Z, Gao X, Xia Z, Zhao S, Sun H, Wang H, Liu M, Kong X, Ma C, Zhu H, Gao H, Liu Q, Yang F, Song X, Lu J and Zhou X (2023), Prediction of small molecule drug-miRNA associations based on GNNs and CNNs. *Front. Genet.* 14:1201934. doi: 10.3389/fgene.2023.1201934

COPYRIGHT
© 2023 Niu, Gao, Xia, Zhao, Sun, Wang, Liu, Kong, Ma, Zhu, Gao, Liu, Yang, Song, Lu and Zhou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Prediction of small molecule drug-miRNA associations based on GNNs and CNNs

Zheyu Niu, Xin Gao, Zhaozhi Xia, Shuchao Zhao, Hongrui Sun, Heng Wang, Meng Liu, Xiaohan Kong, Chaoqun Ma, Huaqiang Zhu, Hengjun Gao, Qinggong Liu, Faji Yang, Xie Song, Jun Lu and Xu Zhou*

Department of Hepatobiliary Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China

MicroRNAs (miRNAs) play a crucial role in various biological processes and human diseases, and are considered as therapeutic targets for small molecules (SMs). Due to the time-consuming and expensive biological experiments required to validate SM-miRNA associations, there is an urgent need to develop new computational models to predict novel SM-miRNA associations. The rapid development of end-to-end deep learning models and the introduction of ensemble learning ideas provide us with new solutions. Based on the idea of ensemble learning, we integrate graph neural networks (GNNs) and convolutional neural networks (CNNs) to propose a miRNA and small molecule association prediction model (GCNNMMA). Firstly, we use GNNs to effectively learn the molecular structure graph data of small molecule drugs, while using CNNs to learn the sequence data of miRNAs. Secondly, since the black-box effect of deep learning models makes them difficult to analyze and interpret, we introduce attention mechanisms to address this issue. Finally, the neural attention mechanism allows the CNNs model to learn the sequence data of miRNAs to determine the weight of sub-sequences in miRNAs, and then predict the association between miRNAs and small molecule drugs. To evaluate the effectiveness of GCNNMMA, we implement two different cross-validation (CV) methods based on two different datasets. Experimental results show that the cross-validation results of GCNNMMA on both datasets are better than those of other comparison models. In a case study, Fluorouracil was found to be associated with five different miRNAs in the top 10 predicted associations, and published experimental literature confirmed that Fluorouracil is a metabolic inhibitor used to treat liver cancer, breast cancer, and other tumors. Therefore, GCNNMMA is an effective tool for mining the relationship between small molecule drugs and miRNAs relevant to diseases.

KEYWORDS

small molecule drug, miRNAs, graph neural networks, convolutional neural networks, CNN, liver cancer

Introduction

With the development of sequencing technology, the biomedical field has accumulated a large amount of medical data, which provides more convenience for researchers to study the relationship between diseases and drugs using these data. The prediction of the relationship between small molecule (SM) drugs and microRNAs (miRNAs) has become an important and rapidly developing area in pharmacology and pharmacogenomics research (Bartel, 2004; Beermann et al., 2016; Kozomara et al., 2019; Liu et al., 2022). miRNAs are small non-coding RNA molecules that regulate gene expression and play a key role in various biological processes, including the development of diseases (Cai et al., 2021; Peng et al., 2023). On the other hand, small molecule drugs have been widely used to treat diseases, but their impact on miRNA expression is not clear. However, there are still blind issues in using traditional biological experiments to identify small molecule drug-related miRNAs, which require a lot of experimental time and cost. With the increasing availability of large datasets, it is possible to predict the relationship between small molecule drugs and miRNAs and use this information to improve the efficacy and safety of drugs (Wang et al., 2019; Chen et al., 2020). This field has tremendous potential in discovering new therapeutic targets and developing personalized drugs (Chen et al., 2021; Liu et al., 2023; Xu et al., 2023).

Computational methods have played a crucial role in predicting the association between small molecule drugs and miRNAs (Xu et al., 2020; Zhang et al., 2023). As the available data on drugs and miRNAs continues to increase, various computational methods have been proposed to identify and predict their interactions. Lv et al. (2015) constructed a complete network by combining small molecule similarity networks, miRNA similarity networks, and known small molecule-miRNA association networks. They calculated the similarity of small molecules and miRNAs using a weighted combination strategy, and then used the RWR (Random Walk With Restart) algorithm to predict the potential associations between small molecule drugs and miRNAs. BNNRSMMA first defined a new matrix to represent the small molecule-miRNA heterogenous network using miRNA-miRNA similarity, small molecule-small molecule similarity, and known small molecule-miRNA associations. They then completed this matrix by minimizing its kernel parameter count and used alternating direction multiplication to further minimize the kernel parameter count and obtain prediction scores. They introduced a regularization term to tolerate noise in the integrated similarity. Wang et al. (2022a) proposed a novel dual-network collaborative matrix factorization (DCMF) method for predicting potential SM-miRNA associations. They first preprocessed the missing values in the SM-miRNA association matrix using the WKNKN method, and then constructed a matrix factorization model for the dual network to obtain feature matrices containing potential features of small molecules and miRNAs, respectively. Finally, the predicted SM-miRNA association score matrix was obtained by calculating the inner product of the two feature matrices. Li et al. (2016) proposed a network-based inference model for small molecule-miRNA networks (SMiR-NBI), which relies solely on known SM-miRNA associations. For a given SM, the initial resources are evenly allocated to its associated miRNAs. Then, the resources of each miRNA are allocated to all its associated SMs, and the resources are then redistributed from SMs to their associated miRNAs. The final resources obtained by the miRNAs reflect the likelihood of associations between the given SM and miRNAs. Guan et al. (2018) developed a new graphlet interaction-based inference model for predicting small molecule-miRNA associations (GISMMA). The complex relationships among SMs or miRNAs are described by graphlet interactions, which consist of 28 isomers. The association score for an SM-miRNA pair is calculated by counting the number of graphlet interactions. However, if neither the SM nor the miRNA has a known association, the model cannot predict the SM-miRNA association. Wang et al. (2022b) proposed an ensemble method for predicting small molecule-miRNA associations based on kernel ridge regression (EKRRSMMA). This method combines feature dimension reduction and ensemble learning to reveal potential SM-miRNA associations. Firstly, the authors constructed different feature subsets for SMs and miRNAs. Then, homogeneous base learners were trained on different feature subsets, and the average scores obtained from these base learners were used as the association scores for SM-miRNA pairs. Peng et al. (2022) proposed a new computational method based on deep autoencoder and scalable tree boosting model (DAESTB) to predict the associations between small molecules and miRNAs. Firstly, a high-dimensional feature matrix was constructed by integrating small molecule-small molecule similarity, miRNA-miRNA similarity, and known small molecule-miRNA associations. Secondly, the feature dimension of the integrated matrix was reduced using a deep autoencoder to obtain potential feature representations for each small molecule-miRNA pair. Finally, a scalable tree boosting model was used to predict potential associations between small molecules and miRNAs. Although these models have achieved promising results and played important roles in the development of computational methods for small molecule-miRNA association identification, they have certain issues or limitations: the experimental validation of small molecule-miRNA associations is very limited, and there are many negative associations. When performed on this noisy and sparse small molecule-miRNA association network, the predictors often detect many false negative associations.

Therefore, we propose a miRNA-molecule association prediction model (GCNNMMA) by integrating graph convolutional networks (GCNs) (Scarselli et al., 2008) and convolutional neural networks (CNNs) (Chen, 2015) (Figure 1). Firstly, GCNs are used to effectively learn the molecular structural graph data of small molecule drugs, and CNNs are used to learn the sequence data of miRNAs. Due to the black-box nature of deep learning models, it is difficult to analyze and interpret them. Therefore, GCNNMMA introduces a neural attention mechanism (Bahdanau et al., 2014) to address this issue. The neural attention mechanism enables CNNs to learn the weights of sub-sequences in miRNAs, thus predicting the associations between miRNAs and small molecule drugs.

Materials and methods

Datasets

For dataset 1, we obtained a total of 664 known small molecule-miRNA associations from SM2miR database (version 1.0) (Liu et al., 2013). Then a total of 831 small molecules were extracted and

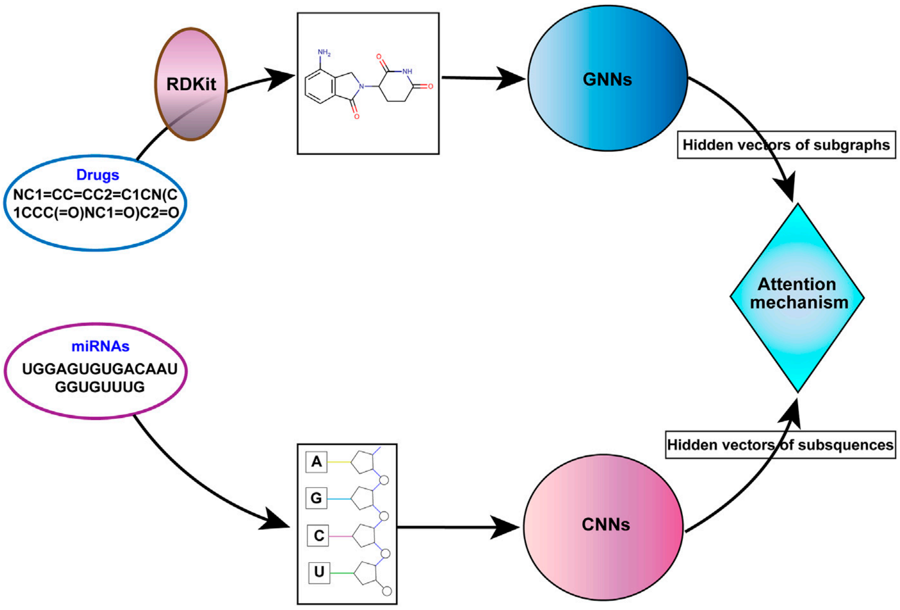

FIGURE 1
The overall workflow of GCNNMMA.

TABLE 1 Statistics of datasets used in this study.

Dataset	No. of miRNAs	No. of molecules	No. of associations
Dataset 1	541	831	664
Dataset 2	2,460	680	60,212

integrated from SM2miR, DrugBank (Wishart et al., 2018), and PubChem (Kim et al., 2019). 541 miRNAs were collected from SM2miR, HMDD, miR2Disease, and PhenomiR (Ruepp et al., 2010). To evaluate our model performance more comprehensively, we constructed dataset 2, which contains 680 small molecules, 2,460 miRNAs, and 60,212 known small molecule-miRNA associations. Additionally, we downloaded corresponding small molecule drug SMILES data from DrugBank. The SMILES format data was used to describe the spatial structural information of small molecule drugs. Furthermore, we obtained corresponding miRNA sequence data from the miRbase database (Table 1).

Prediction model based on the integration of CNNs and GNNs

GNNs process small molecule drug data

End-to-end learning model GNNs has been shown to achieve good performance in many scenarios. Therefore, we first use two functions [the transformation function $tran(x)$ and the output function $f(x)$] in GNNs to map the molecular structure graph $G(V,E)$ of small molecule drugs to a low-dimensional vector $y\epsilon\mathbb{R}^d$. The transformation function $tran(x)$ updates the feature information of each node in the molecular graph $G(V,E)$ using information from neighboring nodes (atoms in the molecular structure graph) and neighboring edges (chemical bonds in the molecular structure graph). The output function $f(x)$ converts the updated node information in the molecular graph after the transformation function into a low-dimensional vector. In GNNs, both the transformation function and the output function are implemented as differentiable neural networks, and the parameters in the functions are automatically learned through the backpropagation process (Figure 2). The specific steps are as follows:

Subgraph embedding with radius r: Here, we use $G(V,E)$ to represent a molecular graph, where V is a set of nodes and E is a set of edges. In the molecular structure graph, $v_i\epsilon V$ represents the i-th atom and $e_{ij}\epsilon E$ represents the chemical bond between atom i and atom j. Because there are only a few types of nodes (hydrogen and carbon) and edges (double and single bonds) in the molecular graph, representative learning models cannot obtain effective learning results. To solve this problem, GCNNMMA introduces the concept of r-radius subgraphs. An r-radius subgraph describes the set of atoms and chemical bonds within a radius of r with a certain atom as the center. Here, we use $\Gamma(i,r)$ to represent the set of indices of all adjacent nodes in the subgraph with node i as the center and a radius of r. $\Gamma(i,0)$

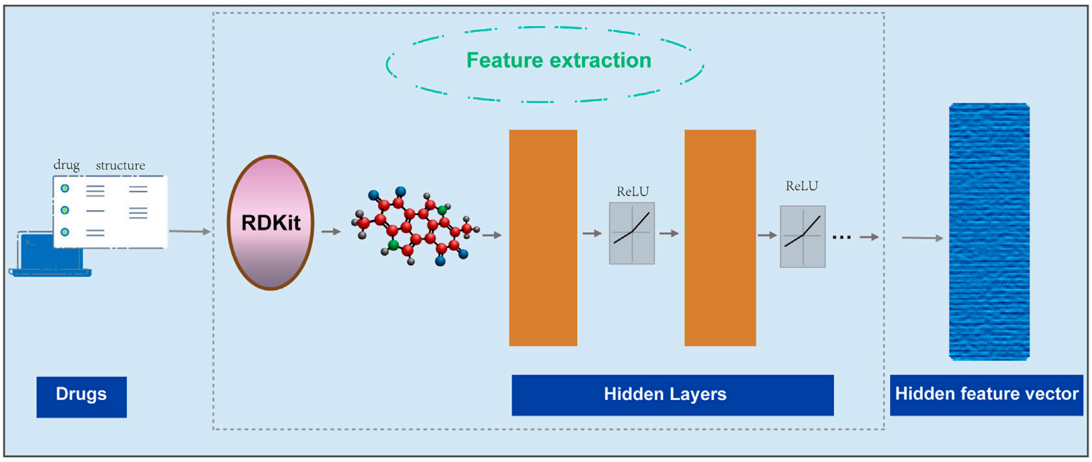

FIGURE 2
Using GNNs to extract features of small molecule drugs.

is the node i itself. We use the following definition to describe the subgraph with node v_i and a radius of r:

$$v_i^r = \left(V_i^r, E_i^r\right) \tag{1}$$

Where, $V_i^r = \left\{v_j | j\epsilon\Gamma(i,r)\right\}$, $E_i^r = \{e_{mn}\epsilon E | (m,n)\epsilon(\Gamma(i,r) \times \Gamma(i,r-1))\}$ Similarly, the subgraph with a radius of r can be defined for the edge e_{ij}: $e_{ij}^r = (V_i^{r-1} \cup V_j^{r-1}, E_i^r \cap E_j^r)$.

Vertex transformation function: In the molecular structure graph G, subgraph embedding can start from any vertex. $v_i^{(t)}\epsilon R^d$ is used to describe the vertex i at the t-th step of subgraph embedding information update. The update process is described as follows:

$$v_i^{(t)} = \sigma\left(v_i^{(t-1)} + \sum_{j\epsilon\Gamma(i)} h_{ij}^{(t)}\right) \tag{2}$$

Where $\sigma(x) = \frac{1}{(1+e^x)}$, $\Gamma(i)$ represents the set of neighbor node indices for vertex i. $h_{ij}^{(t)}$ is a hidden vector describing the information of neighbor node j and the edge e_{ij} between the two nodes for vertex i. It can be calculated using the following formula:

$$h_{ij}^{(t)} = max\left(0, W_{neighbor}{}^{*}\begin{bmatrix} v_j^{(t)} \\ e_{ij}^{(t)} \end{bmatrix} + b_{neighbor}\right) \tag{3}$$

Were, $W_{neighbor}\epsilon\mathbb{R}^{d\times 2d}$ is a weight matrix and $b_{neighbor}\epsilon\mathbb{R}^d$ is a bias matrix. $e_{ij}^{(t)}$ represents the t-th subgraph embedding information update between vertex i and vertex j. By summing the hidden vectors of adjacent nodes and iteratively updating, vertex embedding can gradually learn the global information of the molecular structure graph.

The edge transformation function: The process of updating edge embeddings are similar to the process of updating vertex embeddings. Here, $e_{ij}^{(t)}$ is used to represent the embedding of the edge between vertex i and vertex j. At the same time, the embeddings of adjacent vertices to the edge, $v_i^{(t)}$ and $v_j^{(t)}$, are used to update the edge embedding information. The update process is described as follows:

$$e_{ij}^{(t)} = \sigma\left(e_{ij}^{(t-1)} + g_{ij}^{(t-1)}\right) \tag{4}$$

The formula describes $g_{ij}^{(t-1)}$ as follows: $g_{ij}^{(t)} = max(0, W_{side}{}^{*}[v_i^{(t)} + v_j^{(t)}] + b_{side})$. $W_{side}\epsilon\mathbb{R}^{d\times 2d}$ is a weight matrix, and $b_{side}\epsilon\mathbb{R}^d$ is a bias vector.

Small molecule output function: To obtain the final output $y_{sm}\epsilon\mathbb{R}^d$, the model sums up the embeddings of each vertex in the molecular graph $V = \left\{v_1^{(t)}(t), v_2^{(t)}(t), \cdot\cdot\cdot, v_{|V|}^{(t)}(t)\right\}$. The process is described as follows:

$$y_{sm} = \frac{1}{|V|}\sum_{i=1}^{|V|} v_i^{(t)} \tag{5}$$

$|V|$ represents the number of vertices in the molecular graph.

Using CNNs to process miRNA sequence data

First, CNNs use filter functions to compute a hidden vector $y \in R^d$ based on the sub-sequences of the input sequence C and a weight matrix (learned parameters). The filter functions are implemented by neural networks. In CNNs, the overall function = $f(C)$ is differentiable and all parameters in $f(x)$ are learned through backpropagation (Figure 3). The specific steps are shown as follows:

Sequence input function

To apply CNNs to miRNA sequence data, First, miRNA sequences are defined as "words" consisting of n-length bases (Dong et al., 2006; Costa and De Grave, 2010), where n refers to the number of bases. Then, the miRNA sequence is divided into overlapping n-mers. In this study, to maintain a manageable and informative word vocabulary and to avoid using low-frequency sequence fragmentation in learning representations, a relatively small value of $n = 3$ was set for the number of bases. The miRNA sequence $S = x_1, x_2, \cdot\cdot\cdot, x_{|s|}$, where x_i is the i-th base pair

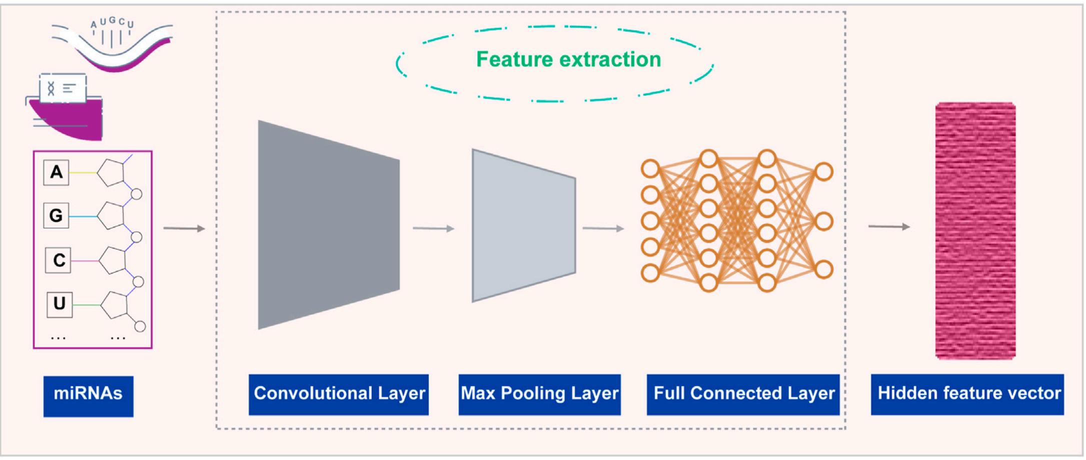

FIGURE 3
Using CNNs to extract features of miRNAs.

and $|s|$ is the length of the sequence, is then split into overlapping n-base pair segments. All words are then translated into randomly initialized embeddings, referred to as "word embeddings." The word embeddings are ordered as $X_1, X_2, \cdots, X_{|s|-1} X_{|s|}$, where $X_i \in \mathbb{R}^d$ is a d-dimensional embedding for the i-th word. Alternatively, we can consider a sequence whose elements consist of concatenated word embeddings. For example, a sequence composed of three consecutive embeddings would be $[X_1; X_2;\ X_3], [X_2;\ X_3; X_4] \cdots [X_{|S|-2}; X_{|S|-1}; X_{|s|}]$, where $[X_{i+1};\ X_{i+2};\ X_{i+3}] \epsilon \mathbb{R}^{3d}$ is the concatenation of X_{i+1}, X_{i+2}, and X_{i+3}. Here, $X_{i:\ i+w-1}$ refers to $[X_i;\ \cdots;\ X_{i+w-1}]$, where w is the window size. This processed sequence can be used as input for CNNs.

Filter function

Using $X_{i:\ i+w-1} = [X_i; X_{i+w-1}] = c_i^{(0)} \epsilon \mathbb{R}^{dw}$ as the input to the filter function $f(x)$, the output of the filter function is a hidden vector $c_i^{(1)} \epsilon \mathbb{R}^d$. The description of the hidden vector is as follows:

$$c_i^{(1)} = f\left(W_{conv} {}^* c_i^{(0)} + b_{conv}\right) \tag{6}$$

Where $f(x)$ is a non-linear activation function, $W_{conv} \epsilon \mathbb{R}^{d \times dw}$ is the weight matrix, and b_{conv} is the bias vector. By using the filter function repeatedly, multiple hidden vectors can be obtained:

$$c_i^{(t)} = f\left(W_{conv} {}^* c_i^{(t-1)} + b_{conv}\right) \tag{7}$$

Multiple hidden vectors form a hidden vector set $C = \left\{c_1^{(t)}, c_2^{(t)}, c_3^{(t)}, \ldots\ldots c_{|c|}^{(t)}\right\}$.

miRNA sequence output function. In order to obtain the final output $y_{miRNA} \epsilon \mathbb{R}^d$ from $C = \left\{c_1^{(t)}, c_2^{(t)}, c_3^{(t)}, \ldots\ldots c_{|c|}^{(t)}\right\}$, the average of C is taken. The process is described as follows:

$$y_{miRNA} = \frac{1}{|C|} \sum_{i=1}^{|C|} c_i^{(t)} \tag{8}$$

$|C|$ denotes the number of elements in set C.

Neural attention mechanism for predicting potential associations between miRNAs and small molecule drugs

GCNNMMA employs a neural attention mechanism to infer interactions between small molecules and subsequences in miRNA sequences. In the collection of hidden vector sequences $C = \left\{c_1^{(t)}, c_2^{(t)}, c_3^{(t)}, \ldots\ldots c_{|c|}^{(t)}\right\}$ for miRNA sub-sequences, each hidden vector sequence represents its corresponding miRNA sub-sequence. Different miRNA sub-sequences have different binding abilities and probabilities with small molecules. A neural attention mechanism is used to assign corresponding weights to each sub-sequence in the miRNA hidden vector sequence collection, which represents the importance of its association with small molecules. The weight calculation process is described as follows:

$$h_{sm} = f\left(W_{inter} {}^* y_{sm} + b_{inter}\right) \tag{9}$$

$$h_i = f\left(W_{inter} {}^* c_i + b_{inter}\right) \tag{10}$$

$$\alpha_i = \sigma\left(h_{sm}^T {}^* h_i\right) \tag{11}$$

Where W_{inter} is the weight matrix and b_{inter}_inter is the bias vector. α_i represents the strength of interaction between small molecules and miRNA sub-sequences. Based on the calculated attention weights, the final weighted sum can be obtained, as shown below:

$$y_{miRNA} = \sum_{i=1}^{|C|} \alpha_i {}_* h_i \tag{12}$$

Finally, the model obtains the final classification output vector $Z \epsilon R^2$ by jointly considering y_{miRNA} and y_{sm}:

$$Z = W_{output} {}^* \left[y_{miRNA}; y_{sm}\right] + b_{output} \tag{13}$$

Where $W_{output} \in R^{2 \times 2d}$ is the weight matrix and $b_{output} \in R^2$ is the bias vector. Finally, the output vector $Z = [y0, y1]]$ is passed through the softmax function to compute the associated probabilities:

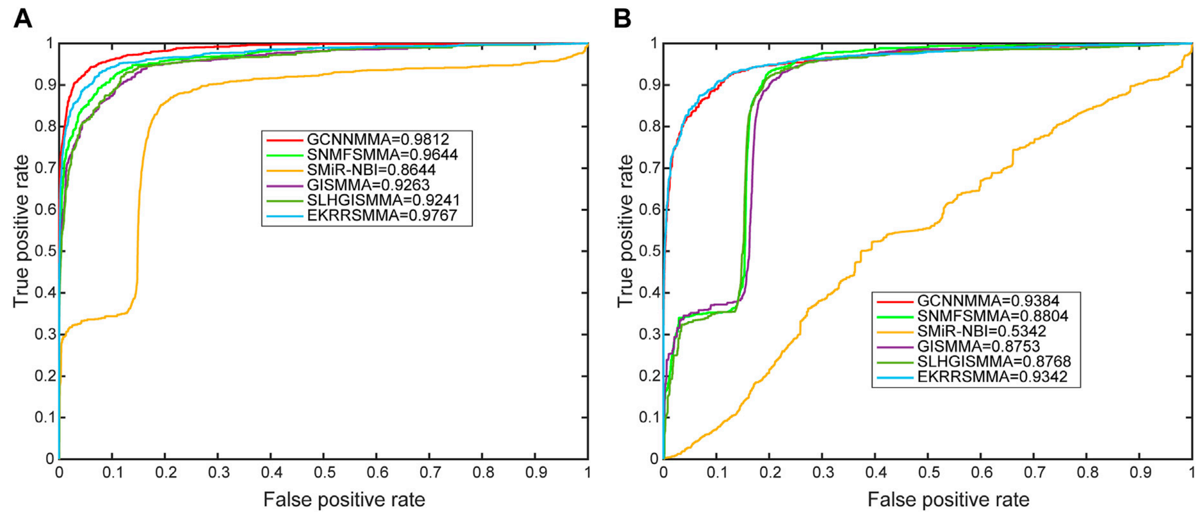

FIGURE 4
The ROC curves for GCNNMMA and benchmark algorithms for 5-fold CV on the **(A)** dataset 1 and **(B)** dataset 2.

$$P_t = \frac{exp(y_t)}{\sum_i y_i} \tag{14}$$

Results

Performance of GCNNMMA in the cross-validation

In this work, we compared the performance of the latest five models [SMiR-NBI (Li et al., 2016), GISMMA (Guan et al., 2018), SLHGISMMA (Yin et al., 2019), SNMFSMMA (Zhao et al., 2020), EKRRSMMA (Wang et al., 2022b)] with GCNNMMA, and conducted 5-fold cross-validation (CV) on both dataset 1 and dataset 2 to evaluate the predictive performance of GCNNMMA. All predicted small molecule miRNA pairs were ranked according to the obtained scores. Based on the rankings, we used receiver operating characteristic (ROC) curves to illustrate the performance of our models in the cross-validation runs. As shown in Figure 4, we found that GCNNMMA achieved the best predictive performance on both dataset 1 (AUC = 0.9812) and dataset 2 (AUC = 0.9384). This suggests that GCNNMMA performed the best in predicting the correlation between small molecule drugs and miRNAs.

GCNNMMA is superior to other popular methods in predicting miRNAs associated with new small molecule drugs

It is important to examine the performance of the above method in predicting new miRNAs related to small molecule drugs, in addition to testing the performance of global prediction of small molecule drug-miRNA relationships. A leave-one-out experiment is used to evaluate the ability of the algorithm to predict miRNAs related to new small molecule drugs. To compare the fairness of the test, we still use ROC as the indicator of predictive performance. The local LOOCV experiment was carried on the dataset 1 and dataset 2 (see Figure 5). GCNNMMA showed a higher performance over other approaches in terms of AUC on the dataset 2. Specifically, GCNNMMA obtained AUC value of 0.9367, outperforming that of SMiR-NBI (AUC = 0.6754), GISMMA (AUC = 0.8473), SLHGISMMA (AUC = 0.8532), SNMFSMMA (AUC = 0.9254), EKRRSMMA (AUC = 0.8751). In addition, we can find that the performance of GCNNMMA is also second only to SNMFSMMA on the dataset 1. This also sufficient GCNNMMA is also the best way to predict m miRNAs related to new small molecule drugs.

Case studies: identifying the relationship between small molecule drugs and miRNAs associated with liver cancer

To further verify the reliability capability of GCNNMMA, we take all known miRNAs-small molecule drug associations in the SM2miR dataset 1 as the training set, and regard the missing miRNAs-small molecule drug associations as candidate sets. After GCNNMMA predicted the interaction probabilities of all candidate miRNAs-small molecule drug associations, we then ranked them according to the predicted probabilities so that the top-ranked associations were most likely to interact. We also validated these top 30 associations by searching for corresponding PubMed literature, as shown in Table 2. Among the top 10, 20, and 30 predicted associations, we were able to validate 6, 12, and 20 associations, respectively through literature search. In the top 10 predicted associations, we found that 5 different miRNAs were associated with Fluorouracil (CID: 3385), a small molecule drug that belongs to the class of pyrimidine analogs and is an anti-metabolic drug used to treat tumors. It interferes with DNA synthesis by

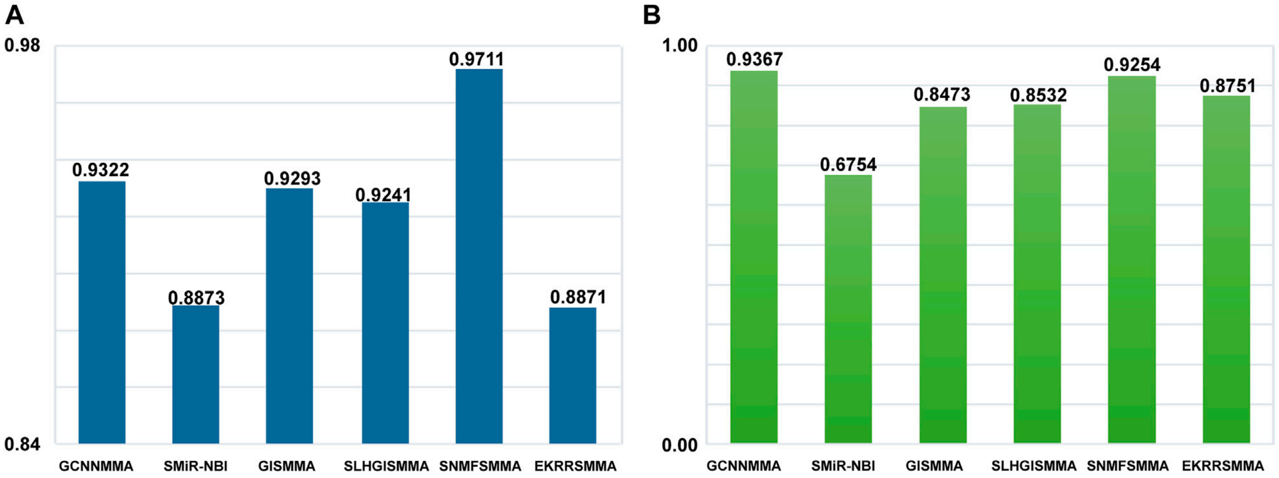

FIGURE 5
The ROC curves for GCNNMMA and benchmark algorithms for local LOOCV on the **(A)** dataset 1 and **(B)** dataset 2.

TABLE 2 Predicting the top 30 small molecule drugs associated with miRNAs.

Rank	CID	miRNA	Evidence (PubMed)	Rank	CID	miRNA	Evidence (PubMed)
1	3,229	hsa-mir-212	28,131,841	16	5,757	hsa-mir-542	17,765,232
2	3,385	hsa-mir-149	27,415,661	17	5,757	hsa-mir-663a	32,215,262
3	3,385	hsa-mir-1915	22,121,083	18	6,013	hsa-mir-135a-1	32,735,753
4	3,385	hsa-mir-203a	25,526,515	19	6,013	hsa-mir-29a	26,296,572
5	3,385	hsa-mir-320a	unconfirmed	20	10,635	hsa-mir-32	20,945,501
6	3,385	hsa-mir-483	unconfirmed	21	10,635	hsa-mir-630	20,945,501
7	3,385	hsa-mir-519c	26,386,386	22	31,401	hsa-mir-603	20,689,055
8	3,385	hsa-mir-617	21,743,970	23	36,462	hsa-mir-26b	31,985,026
9	5,311	hsa-mir-126	unconfirmed	24	36,462	hsa-mir-663a	31,639,426
10	5,311	hsa-mir-409	unconfirmed	25	60,750	hsa-mir-139	33,300,085
11	5,311	hsa-mir-574	unconfirmed	26	60,750	hsa-mir-211	25,789,319
12	5,311	hsa-mir-595	unconfirmed	27	60,750	hsa-mir-299	28,131,841
13	5,311	hsa-mir-744	unconfirmed	28	60,750	hsa-mir-326	unconfirmed
14	5,311	hsa-mir-760	unconfirmed	29	60,953	hsa-mir-137	22,740,910
15	5,757	hsa-mir-17	24,283,290	30	216,239	hsa-mir-664a	unconfirmed

blocking the conversion of deoxyuridine monophosphate to thymidine monophosphate (Ellison, 1961). Currently, Fluorouracil is used to treat diseases such as actinic keratosis, breast cancer, colon cancer, pancreatic cancer, gastric cancer, liver cancer, and superficial basal cell carcinoma (Lecluse and Spuls, 2015; Guo et al., 2020). Among the top 20 predicted associations, we discovered novel small molecule drugs associated with miRNAs and Estradiol (CID:5757), Testosterone (CID: 6013), and Dihydrotestosterone (CID: 10635). These three hormones have high bioavailability and can enhance cellular metabolism. These three hormones have high bioavailability and can enhance cellular metabolism (Pentikäinen et al., 2000). Among the top 30 predicted associations, we found that the small molecule drugs Etoposide (CID: 36462) (Wang et al., 2003) and Gemcitabine (CID: 60750) are used for cancer treatment. Etoposide is a semi-synthetic derivative with anti-tumor activity. It inhibits DNA synthesis by forming a complex with topoisomerase II and DNA, inducing double-stranded DNA breaks and preventing repair by blocking the binding of topoisomerase II. Accumulation of DNA breaks prevents cells from entering mitosis, leading to cell death (Uesaka et al., 2007).

Gemcitabine (CID: 60750) is a nucleoside analog used in chemotherapy that, like fluorouracil and other pyrimidine analogs, replaces a structural group of nucleic acids in DNA replication to form cytidine in this case. The formation of cytidine stops tumor growth as new nucleosides cannot attach to the "defective" nucleosides, leading to cell apoptosis (cell "suicide") (Hastak et al., 2010; Vogl et al., 2010). Currently, Gemcitabine is used to treat cancers such as non-small cell lung cancer, pancreatic cancer, bladder cancer, and breast cancer.

Discussion

The development of deep learning provides new approaches for predicting the association between small molecule drugs and miRNAs. We developed a prediction model called GCNNMMA based on graph neural networks (GNNs) and convolutional neural networks (CNNs), and validated its performance on two datasets. Experimental results show that GCNNMMA exhibited the best performance in the datasets. Compared with previous similarity-based models, our model extracts the characteristic information of small molecule drugs and miRNAs through GNN and CNN networks, avoiding the dependence on known association information. Furthermore, when predicting the top 30 associations in the dataset, GCNNMMA identified Gemcitabine (CID: 60750) related to hsa-mir-139 and Fluorouracil (CID: 3385) related to hsa-mir-149, both of which are used in cancer treatment by targeting the relevant miRNAs to inhibit cell division and induce cancer cell death. While GCNNMMA achieved good performance, there is still room for improvement, such as integrating multi-source data which remains a challenging problem. In the future, incorporating more data sources, such as miRNA spatial structure data and miRNA precursor data, could improve GCNNMMA. In addition, three-dimensional structural information can better reflect spatial information. One of the future research directions is to utilize the three-dimensional structural information of miRNAs and small molecule drugs to improve prediction accuracy.

Data availability statement

The program and data used in this study are publicly available at: https://github.com/niuzheyu123/GCNNMMA.git.

Author contributions

XZ conceived the study. ZN, XG, ZX, SZ, and HS performed experiments and data analysis. HW, ML, KX, and CM interpreted the data analysis. ZN, HZ HG, QL FY, XS, JL, and XZ drafted the manuscript and critically revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by Natural Science Foundation of Shandong Province, China (Grant Nos ZR2020MH259, ZR2016HB40), Natural Science Foundation of China Youth Project (Grant No. 81802379).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). "Neural machine translation by jointly learning to align and translate.", arXiv preprint arXiv:1409.0473.

Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 1162, 281–297. doi:10.1016/s0092-8674(04)00045-5

Beermann, J., Piccoli, M. T., Viereck, J., and Thum, T. (2016). Non-coding RNAs in development and disease: Background, mechanisms, and therapeutic approaches. *Physiol. Rev.* 96, 1297–1325. doi:10.1152/physrev.00041.2015

Cai, L., Lu, C., Xu, J., Meng, Y., Wang, P., Fu, X., et al. (2021). Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Briefings Bioinforma.* 22, bbab319. doi:10.1093/bib/bbab319

Chen, X., Zhou, C., Wang, C. C., and Zhao, Y. (2021). Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization. *Briefings Bioinforma.* 22, bbab328. doi:10.1093/bib/bbab328

Chen, X., Guan, N. N., Sun, Y. Z., Li, J. Q., and Qu, J. (2020). MicroRNA-small molecule association identification: From experimental results to computational models. *Briefings Bioinforma.* 21, 47–61. doi:10.1093/bib/bby098

Chen, Y. (2015). "Convolutional neural network for sentence classification,". MS thesis (University of Waterloo) Computer Science.

Costa, F., and De Grave, K. (2010). "Fast neighborhood subgraph pairwise distance kernel," in Proceedings of the 26th International Conference on Machine Learning (Madison, WI, United States: Omnipress), 255–262.

Dong, Q.-W., Wang, X., and Lin, L. (2006). Application of latent semantic analysis to protein remote homology detection. *Bioinformatics* 22.3, 285–290. doi:10.1093/bioinformatics/bti801

Ellison, R. R. (1961). Clinical applications of the fluorinated pyrimidines. *Med. Clin. N. Am.* 45.3, 677–688. doi:10.1016/s0025-7125(16)33880-9

Guan, N.-N., Sun, Y. Z., Ming, Z., Li, J. Q., and Chen, X. (2018). Prediction of potential small molecule-associated microRNAs using graphlet interaction. *Front. Pharmacol.* 9, 1152. doi:10.3389/fphar.2018.01152

Guo, P., Pi, C., Zhao, S., Fu, S., Yang, H., Zheng, X., et al. (2020). Oral co-delivery nanoemulsion of 5-fluorouracil and curcumin for synergistic effects against liver cancer. *Expert Opin. Drug Deliv.* 17.10, 1473–1484. doi:10.1080/17425247.2020.1796629

Hastak, K., Alli, E., and JamesFord, M. (2010). Synergistic chemosensitivity of triple-negative breast cancer cell lines to poly(ADP-Ribose) polymerase inhibition, gemcitabine, and cisplatin. *Cancer Res.* 70.20, 7970–7980. doi:10.1158/0008-5472.CAN-09-4521

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic acids Res.* 47, D1102–D1109. doi:10.1093/nar/gky1033

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). "miRBase: from microRNA sequences to function." *Nucleic acids Res.* 47. D155–D162. doi:10.1093/nar/gky1141

Lecluse, L. L. A., and Spuls, P. I. (2015). Photodynamic therapy versus topical imiquimod versus topical fluorouracil for treatment of superficial basal-cell carcinoma: A single blind, non-inferiority, randomised controlled trial: A critical appraisal. *Br. J. Dermatology* 172.1, 8–10. doi:10.1111/bjd.13460

Li, J., Lei, K., Wu, Z., Li, W., Liu, G., Liu, J., et al. (2016). Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* 7.29, 45584–45596. doi:10.18632/oncotarget.10052

Liu, W., Hui, L., and Li, H. (2022). Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Briefings Bioinforma.* 23, 3. doi:10.1093/bib/bbac104

Liu, W., Yang, Y., Lu, X., Fu, X., Sun, R., Yang, L., et al. (2023). Nsrgrn: A network structure refinement method for gene regulatory network inference. *Briefings Bioinforma.*, bbad129. doi:10.1093/bib/bbad129

Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., et al. (2013). SM2miR: A database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 29.3, 409–411. doi:10.1093/bioinformatics/bts698

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31.22, 3638–3644. doi:10.1093/bioinformatics/btv417

Peng, L., Cheng, Y., Yifan, C., and Wei, L. (2023). Predicting CircRNA-Disease associations via feature convolution learning with heterogeneous graph attention network. *IEEE J. Biomed. Health Inf.*, 1–11. doi:10.1109/JBHI.2023.3260863

Peng, L., Tu, Y., Huang, L., Li, Y., Fu, X., and Chen, X. (2022). Daestb: Inferring associations of small molecule–miRNA via a scalable tree boosting model based on deep autoencoder. *Briefings Bioinforma.* 23.6, bbac478. doi:10.1093/bib/bbac478

Pentikäinen, V., Erkkilä, K., Suomalainen, L., Parvinen, M., and Dunkel, L. (2000). Estradiol acts as a germ cell survival factor in the human testis *in vitro*. *J. Clin. Endocrinol. Metabolism* 85.5, 2057–2067. doi:10.1210/jcem.85.5.6600

Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., et al. (2010). PhenomiR: A knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.* 11, R6–R11. doi:10.1186/gb-2010-11-1-r6

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. neural Netw.* 20.1, 61–80. doi:10.1109/TNN.2008.2005605

Uesaka, T., Shono, T., Kuga, D., Suzuki, S. O., Niiro, H., Miyamoto, K., et al. (2007). Enhanced expression of DNA topoisomerase II genes in human medulloblastoma and its possible association with etoposide sensitivity. *J. neuro-oncology* 84, 119–129. doi:10.1007/s11060-007-9360-0

Vogl, T. J., Naguib, N. N. N., Nour-Eldin, N. E. A., Eichler, K., Zangos, S., and Gruber-Rouh, T. (2010). Transarterial chemoembolization (TACE) with mitomycin C and gemcitabine for liver metastases in breast cancer. *Eur. Radiol.* 20, 173–180. doi:10.1007/s00330-009-1525-0

Wang, C.-C., Chen, X., Qu, J., Sun, Y. Z., and Li, J. Q. (2019). Rfsmma: A new computational model to identify and prioritize potential small molecule–mirna associations. *J. Chem. Inf. Model.* 59, 1668–1679. doi:10.1021/acs.jcim.9b00129

Wang, C.-C., Zhu, C.-C., and Chen, X. (2022). Ensemble of kernel ridge regression-based small molecule–miRNA association prediction in human disease. *Briefings Bioinforma.* 23, bbab431. doi:10.1093/bib/bbab431

Wang, S.-H., Wang, C. C., Huang, L., Miao, L. Y., and Chen, X. (2022). Dual-network collaborative matrix factorization for predicting small molecule-miRNA associations. *Briefings Bioinforma.* 23, bbab500. bbab500. doi:10.1093/bib/bbab500

Wang, X., Furukawa, T., Nitanda, T., Okamoto, M., Sugimoto, Y., Akiyama, S. I., et al. (2003). Breast cancer resistance protein (BCRP/ABCG2) induces cellular resistance to HIV-1 nucleoside reverse transcriptase inhibitors. *Mol. Pharmacol.* 63.1, 65–72. doi:10.1124/mol.63.1.65

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic acids Res.* 46.D1, D1074–D1082. doi:10.1093/nar/gkx1037

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: An accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36.10, 3139–3147. doi:10.1093/bioinformatics/btaa109

Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., et al. (2023). Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cell Rep. Methods* 3, 100382. doi:10.1016/j.crmeth.2022.100382

Yin, J., Chen, X., Wang, C. C., Zhao, Y., and Sun, Y. Z. (2019). Prediction of small molecule–microRNA associations by sparse learning and heterogeneous graph inference. *Mol. Pharm.* 16.7, 3157–3166. doi:10.1021/acs.molpharmaceut.9b00384

Zhang, Z., Xu, J., Wu, Y., Liu, N., Wang, Y., and Liang, Y. (2023). CapsNet-LDA: Predicting lncRNA-disease associations using attention mechanism and capsule network based on multi-view data. *Briefings Bioinforma.* 24, bbac531. doi:10.1093/bib/bbac531

Zhao, Y., Chen, X., Yin, J., and Qu, J. (2020). Snmfsmma: Using symmetric nonnegative matrix factorization and kronecker regularized least squares to predict potential small molecule-microRNA association. *RNA Biol.* 17.2, 281–291. doi:10.1080/15476286.2019.1694732