OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–14

https://doi.org/10.1093/bib/bbac298

Problem Solving Protocol

Prediction of biomarker–disease associations based on graph attention network and text representation

Minghao Yang†, Zhi-An Huang†, Wenhao Gu, Kun Han, Wenying Pan, Xiao Yang and Zexuan Zhu

Corresponding author: Xiao Yang, GeneGenieDx Corp, 160 E Tasman Dr, San Jose, CA 95134. E-mail: xyang@genegeniedx.com; Zexuan Zhu, National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, 518060, China. E-mail: zhuzx@szu.edu.cn

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Motivation: The associations between biomarkers and human diseases play a key role in understanding complex pathology and developing targeted therapies. Wet lab experiments for biomarker discovery are costly, laborious and time-consuming. Computational prediction methods can be used to greatly expedite the identification of candidate biomarkers. **Results:** Here, we present a novel computational model named GTGenie for predicting the biomarker–disease associations based on graph and text features. In GTGenie, a graph attention network is utilized to characterize diverse similarities of biomarkers and diseases from heterogeneous information resources. Meanwhile, a pretrained BERT-based model is applied to learn the text-based representation of biomarker–disease relation from biomedical literature. The captured graph and text features are then integrated in a bimodal fusion network to model the hybrid entity representation. Finally, inductive matrix completion is adopted to infer the missing entries for reconstructing relation matrix, with which the unknown biomarker–disease associations are predicted. Experimental results on HMDD, HMDAD and LncRNADisease data sets showed that GTGenie can obtain competitive prediction performance with other state-of-the-art methods. **Availability:** The source code of GTGenie and the test data are available at: https://github.com/Wolverinerine/GTGenie.

Keywords: miRNA–disease associations, microbe–disease associations, lncRNA–disease associations, graph attention network, text-based relation representation, bimodal fusion network

Introduction

Biomarkers play a vital role in disease detection and the follow-up care [1]. Many diseases are signified by the dysregulation of complex functional mechanisms involving multiple biomarkers at genetic, genomic and microbial levels. Identifying biomarker–disease associations is of great values for understanding the pathogenesis of diseases. Since the conventional laboratory validation approach is labor-intensive, expensive and time-consuming, various computational methods have been developed to greatly expedite the identification of candidate biomarkers. For simplicity, we use the term biomarker as a reference to either the ones that have been verified or the candidates that we are trying to infer.

Generally, the existing computational methods for the prediction of biomarker–disease associations can be categorized into network propagation and machine-learning-based methods. Network-propagation-based methods construct networks with diseases and biomarkers being nodes and their similarities or interactions being edges, such that the network topology can be utilized to propagate an arbitrary node to its neighbors through the updated edge scores in the network. The final edge scores reflect the credibility of the corresponding associations. For example, Mugunga et al. [2] constructed the disease similarity and microRNA (miRNA) similarity as the edges in heterogeneous networks and employed random walk to predict the

Minghao Yang is a master student in the College of Computer Science and Software Engineering, Shenzhen University. His research interests include biomarker discovery and deep learning.

Zhi-An Huang received the Ph.D. degree in the Department of Computer Science at City University of Hong Kong in 2021. He is currently a research fellow at City University of Hong Kong Dongguan Research Institute. His research interests include medical data analysis, bioinformatics, and machine learning.

Wenhao Gu received the M.S. degree in computer science from Shenzhen University. He is a software engineer in Guangdong Jiyin Biotech Co. Ltd, specializes in NLP.

Kun Han received the Ph.D. degree from the Ohio State University in computer science in 2014 . He has been with Facebook and DiDi Research after graduation. He is currently a chief scientist in a start-up company in robotics. His research interests include machine learning and its applications on language and vision.

Wenying Pan received the Ph.D. degree from the Stanford University in Bioengineering in 2016. She worked in Grail Inc for early cancer detection and is currently the CSO of GeneGenieDx Corp, San Jose, USA. Her research interests include biomarker discovery and machine learning.

Xiao Yang is the cofounder and head of computational biology in GeneGenieDx, Corp, San Jose, USA. He is focusing on providing effcient and scalable solutions in biomarker and therapeutic target discovery and validation.

Zexuan Zhu is a Professor at the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, PR China. His current research interest includes computational intelligence and bioinformatics.

Received: April 3, 2022. **Revised:** June 28, 2022. **Accepted:** June 30, 2022

latent associations between miRNAs and diseases. Sumathipala et al. [3] integrated the known long noncoding (lncRNA)–protein, protein–protein and protein–disease associations into a multilevel heterogeneous network, and inferred the lncRNA–disease associations using random walk. Random walk with restarting was also used to select the reliable negative microbe–disease associations by positive-unlabeled learning in [4]. Rashmi and Rangarajan [5] considered all disease–disease and miRNA–miRNA similarities together with local edges among nodes to predict the miRNA–disease associations via random walk with restarting. Liu et al. [6] fused multiple data sources including gene–gene, miRNA–gene, miRNA–lncRNA and lncRNA–lncRNA associations to calculate the miRNA similarity matrix. Afterward, they used the heterogeneous networks composed of the miRNA and disease similarity matrices to infer the miRNA–disease associations via random walk. Zhang et al. [7] adopted the KATZ index [8] to identify the miRNA–disease associations based on the meta-path method. Network propagation methods have achieved great accomplishments, yet they are incompetent to capture complex interaction patterns due to the use of simple linear network architecture. Machine-learning-based methods can alleviate this problem by adopting nonlinear classifier and estimating the probability score for each candidate association. For instance, Lan et al. [9] combined multiple biological similarities to predict lncRNA–disease associations via the bagging support vector machines. Le et al. [10] predicted miRNA–disease associations based on random forest and ensemble learning technique. Guo et al. [11] adopted the rotation forest algorithm to perform nonlinear feature transformation for the prediction of lncRNA–disease associations. Wang et al. [12] used Word2vec based embedding [13] to encode miRNA sequences and adopted logistic model tree classifier with biological similarities to predict the latent miRNA–disease associations. Uthayopas et al. [14] employed an extreme gradient boosting classifier to improve the miRNA–disease associations prediction by introducing target and symptom information.

More recently, deep learning techniques have emerged as the new solution for the prediction of biomarker–disease associations and achieved promising results especially on biological big-data. Deep learning based methods use deep neural networks to extract hierarchical feature representation from large-scale biological data. For example, Zeng et al. [15] constructed heterogeneous networks by aggregating the neighbor information in neural networks for miRNA–disease association prediction. Dong and Khosla [16] proposed a multitask graph convolutional learning framework for predicting miRNA–disease associations based on heterogeneous networks. Deepthi and Jereesh [17] employed an auto-encoder combined with deep neural networks to identify circular RNA (circRNA)–disease associations. Besides being classifier, auto-encoders can also be stacked as feature extractor to learn the latent miRNA and disease feature representations [18]. Madhavan and Gopakumar [19] put forward a computational model based on a deep belief network to identify lncRNA–disease associations. Fan et al. [20] performed graph convolutional matrix completion for the prediction of lncRNA–disease associations with conditional random field and attention mechanism. Mudiyanselage et al. [21] presented a graph convolution network with message passing to uncover circRNA–disease associations via learning the global graph structure. Li et al. [22] integrated the miRNA–lncRNA, lncRNA–disease and miRNA–disease association matrices to construct multilevel heterogeneous graphs. Then the hierarchical graph attention networks (GATs) including node-layer attention network and semantic-layer attention network were used to infer the miRNA–disease associations.

The existing computational methods for the prediction of biomarker–disease associations are mainly focused on experimental data, whereas the massive biomedical literature can provide a complementary knowledge resource to improve the prediction accuracy. With this in mind, we developed a new framework called GTGenie with hybrid graph and text features for the prediction of biomarker–disease associations. In GTGenie, we used a GAT to capture the graph features from the integrated biomarker/disease similarities and proposed a text-based relation representation method (TRR) to extract the text features from the relevant literature. The text features reflect the biomarker–disease correlations reported in published literature. Based on both graph and text features, we further adopted a bimodal fusion network (BFN) and inductive matrix completion [23] to identify the unknown biomarker–disease associations. The graph features and text features can profile the biomarker–disease associations from different perspectives. Their integration thus can lead to more accurate prediction. In this study, we investigated three types of biomarker–disease associations, namely, miRNA–disease, microbe–disease and lncRNA–disease associations. The experimental results on the corresponding HMDD [24], HMDAD [25] and LncRNADisease [26] datasets suggested the competitiveness of GTGenie with the other state-of-the-art methods. We also verified the predicted associations via cross checking from other independent databases and published literature.

Materials and methods

Data

We used three representative datasets collected from databases HMDD v2.0 [24], HMDAD [25] and LncRNADisease v2017 [26], to investigate the effectiveness of GTGenie. In particular, the HMDD v2.0 dataset contains 5430 known miRNA–disease associations among 383 diseases and 495 miRNAs. The HMDAD dataset includes 450 known microbe–disease associations between 39 diseases and 292 microbes. The LncRNADisease v2017 dataset contains 1765 known lncRNA–disease associations involving 328 diseases and 881 lncRNAs (the duplicated associations were removed). The HMDD and LncRNADisease datasets are accompanied with supportive descriptions of the associations in the form of both free text and structured text. There is no counterpart text-based information available for HMDAD dataset, so we also manually collected the related sentences from the public literature. Specifically, we searched in PubMed for the eligible sentences with the names of diseases and biomarkers and then confirmed the implied relation based on the description.

The above three datasets are referred to as explicit information sources as the biomarker–disease associations are explicitly indicated in the data. On top of the explicit information sources, we also considered implicit information sources, i.e. the MeSH, UniProt and Expression Atlas datasets, with which the biomarker–disease associations cannot be directly identified but can be estimated via various similarity metrics. In particular, the MeSH dataset (https://www.ncbi.nlm.nih.gov/mesh) provides hierarchical structure information to calculate disease semantic similarity. The UniProt dataset (https://www.uniprot.org/) offers the hierarchical tree structure of microbes for microbe semantic similarity (MSS) calculation. The Expression Atlas dataset (https://www.ebi.ac.uk/gxa/home) provides information on gene expression patterns to measure miRNA expression profile similarity and lncRNA expression profile similarity.

Proposed GTGenie

The proposed GTGenie combines both graph and text features for the representation of biomarker–disease associations. We integrated both explicit and implicit information sources to extract graph features of the biomarkers and diseases via a GAT. To capture the text features of the biomarker–disease associations, we introduced a TRR based on the text records in curated datasets (e.g. HMDD and LncRNADisease) and the published research articles from PubMed.

The overall framework of GTGenie is shown in Figure 1. GTGenie consists of three main components including GAT (Figure 2), TRR (Figure 3) and BFN (Figure 4). The GAT extracts graph features from both disease and biomarker similarity matrices whose elements represent the disease–disease similarity (DDS), biomarker–biomarker similarity (BBS) and Gaussian interaction profile kernel similarity (GIPKS). The TRR extracts text features of biomarkers and diseases based on the relevant sentences in the literature. Finally, the BFN combines the graph features and text features captured by the GAT and the TRR to reconstruct the final biomarker–disease associations matrix by filling the missing values.

Similarity measurements

In this section, we introduce the different measurements of the similarities between biomarkers and diseases.

Disease–disease similarity (DDS): We measured the DDS with disease semantic similarity [27]. For each disease, a unique directed acyclic graph (DAG) was built on the MeSH dataset to capture the relation of a disease with others. Based on the disease DAG, the contribution of a disease d_i to the semantic value of another disease d_j can be defined as $D(d_i, d_j)$ and the semantic value of disease d_i itself as $D(d_i)$. $D(d_i, d_j)$ and $D(d_i)$ are calculated as follows:

$$D(d_i, d_j) = \begin{cases} 1, & \text{if } d_i = d_j \\ \max\left\{\Delta * D(d', d_j) | d_i \in T(d')\right\}, & \text{else} \end{cases} \quad (1)$$

$$D(d_i) = 1 + \sum\nolimits_{t \in T(d_i)} D(t, d_i), \quad (2)$$

where Δ represents a decay factor and is commonly set to 0.5. $T(d_i)$ indicates the ancestors of disease d_i in the DAG. Two diseases sharing more common parts in their DAGs are supposed to have higher similarity. Accordingly, the semantic DDS between two diseases d_i and d_j is given below:

$$\mathbf{DDS}\,(d_i, d_j) = \frac{\sum_{t \in \left(T(d_i) \cap T(d_j)\right)} \left(D(t, d_i) + D(t, d_j)\right)}{D\,(d_i) + D\,(d_j)} \quad (3)$$

Biomarker–biomarker similarity (BBS): To evaluate the BBS of miRNAs and lncRNAs, we obtained the gene expression profiles from Expression Atlas that involves 44 801 genes among 53 human tissues or cell types. Here the genes were restricted to miRNA and lncRNA. Each gene is represented by a vector of size 53 to capture its expression values in the form of Fragments Per Kilobase of exon per Million fragments mapped (FPKM) in all human tissues and cell types. Following [28], we used the Spearman correlation coefficient to compute the BBS ($\in [0, 1]$) as follows:

$$\mathbf{BBS}(b_i, b_j) = \left| 1 - \frac{6 \sum_{k=1}^{N} (b_{ik} - b_{jk})^2}{N(N^2 - 1)} \right|, \quad (4)$$

where b_i and b_j represent two given miRNA/lncRNA vectors and N indicates the size of the gene expression vectors.

To calculate the BBS of microbes, each microbial organism is assigned to a taxonomy rank within {Domain, Phylum, Class, Order, Family, Genus, Species}. We downloaded the taxonomic ranks for each investigated microbe from UniProt. The hierarchy of ranks was represented as a tree structure and also converted into a unique DAG. The same procedure used in calculating DDS was also applied to the calculation of BBS, which can also be referred to as MSS.

Gaussian interaction profile kernel similarity (GIPKS): The calculations of DDS and BBS rely on the MeSH, Expression Atlas and UniProt datasets; however, these datasets

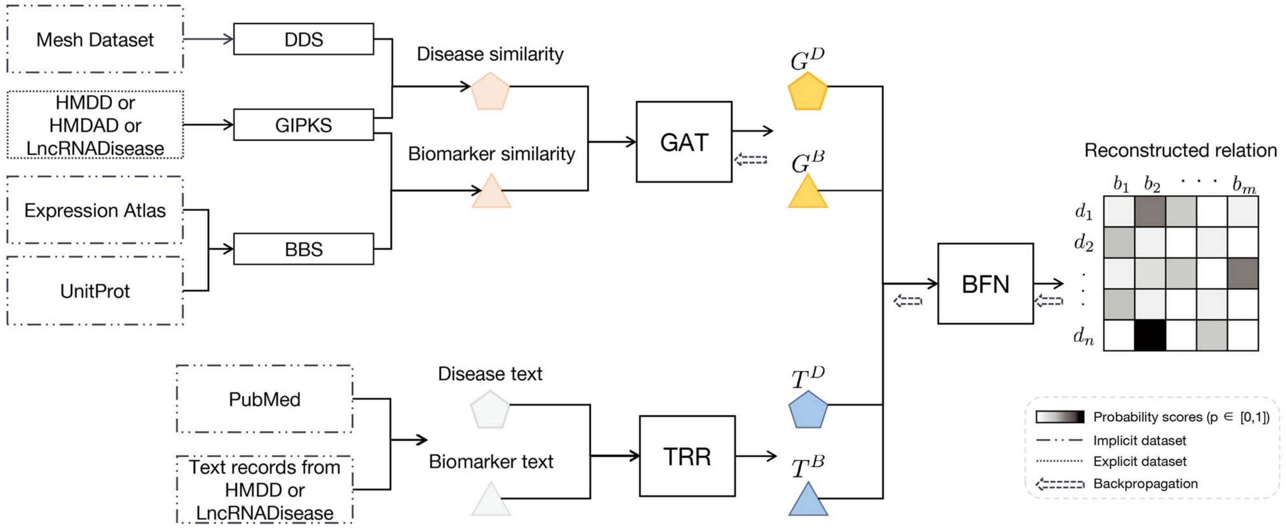


Figure 1. GTGenie Framework.

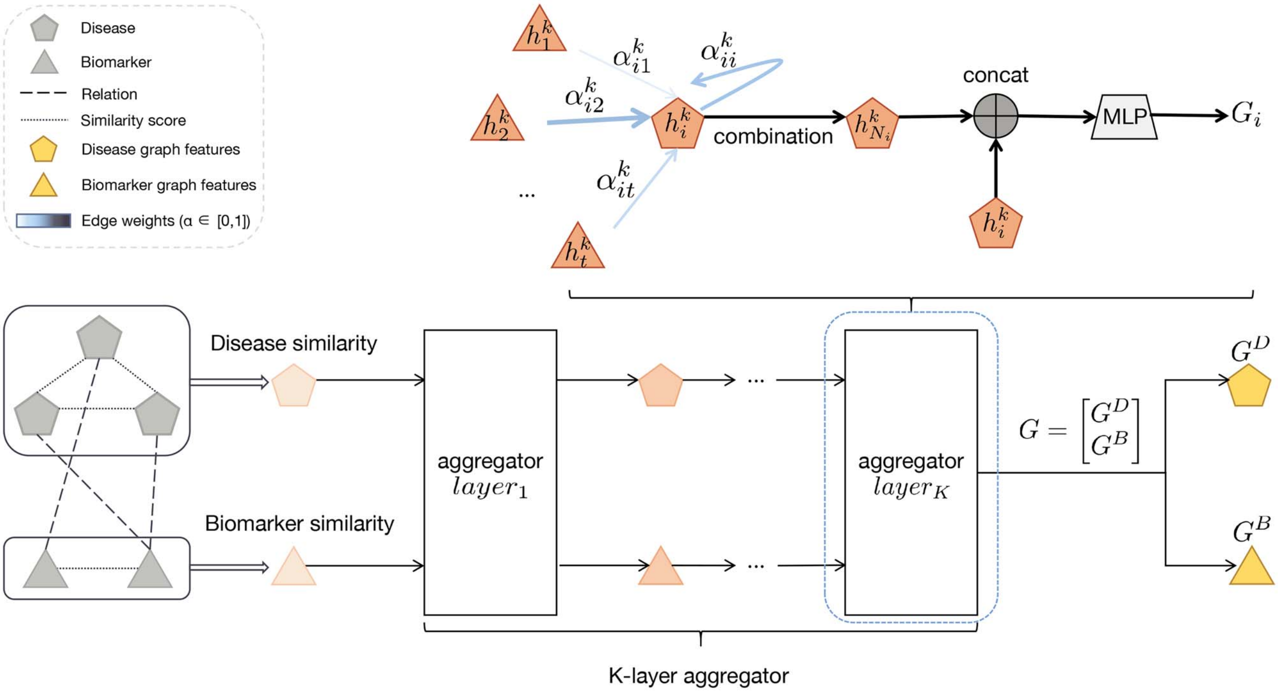


Figure 2. Graph Attention Network architecture.

contain only a certain portion of the biomarkers and diseases. Therefore, the contributed similarity matrices tend to be sparse, which might deteriorate the performance of the GAT. To address this issue, GIPKS [29] is applied as an alternative similarity measurement. GIPKS is based on the hypothesis that two diseases/biomarkers with more similar association patterns can be considered more pathologically similar. Given two diseases d_i and d_j, their similarity based on GIPKS is calculated as follows:

$$\text{GIPKS}\left(d_i, d_j\right) = \exp\left(-\gamma_d \|Y_{i\cdot} - Y_{j\cdot}\|^2\right) \tag{5}$$

$$\gamma_d = 1/\left(\sum_{k=1}^{n} \|Y_{k\cdot}\|^2\right), \tag{6}$$

where $Y_{i\cdot}$ represents the i-th row vector of the biomarker–disease adjacency matrix Y. In Y, the element Y_{ij} equals to 1 if the i-th disease is experimentally verified to have an association with the j-th biomarker, otherwise Y_{ij} is 0. The parameter γ_d controls the kernel bandwidth in GIPKS and n denotes the number of diseases. Similarly, the similarity between two biomarkers say b_i and b_j based on GIPKS is defined as

$$\text{GIPKS}\left(b_i, b_j\right) = \exp\left(-\gamma_d \|Y_{\cdot i} - Y_{\cdot j}\|^2\right) \tag{7}$$

$$\gamma_d = 1/\left(\sum_{k=1}^{m} \|Y_{\cdot k}\|^2\right), \tag{8}$$

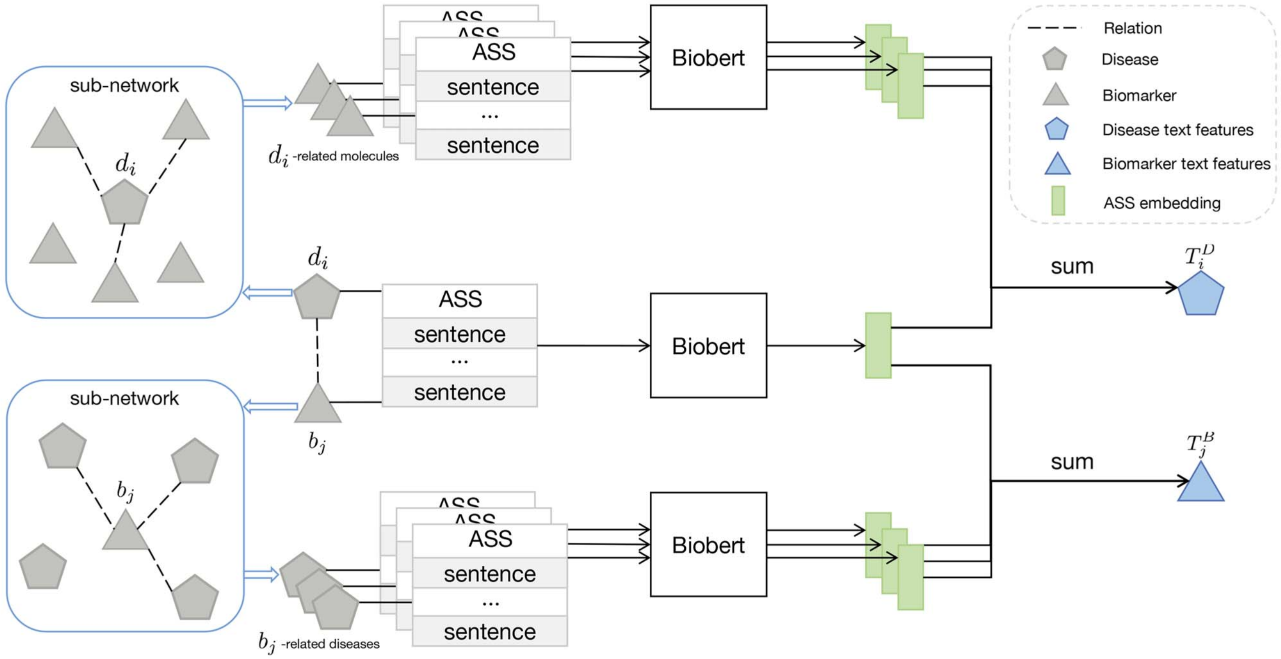


Figure 3. Text-based relation representation method (TRR) architecture. T_i^D and T_j^B represent the text features of the i-th disease and j-th biomarker, respectively.

where $Y_{\cdot i}$ indicates the i-th column vector of Y and m is the number of biomarkers. Since GIPKS is calculated based on the known biomarker–disease associations recorded in Y, only the training data are used to construct Y.

Integrated similarity of biomarkers and diseases: As shown in Figure 1, GIPKS, DDS and BBS values are input into the GAT to extract the graph features of biomarker–disease associations. Let $\mathbf{D} \in \mathcal{R}^{n\times n}$ and $\mathbf{B} \in \mathcal{R}^{m\times m}$ denote the DDS matrix and BBS matrix, respectively. The symbols n and m indicate the numbers of diseases and biomarkers, respectively. A feature matrix $\mathbf{X} \in \mathcal{R}^{(n+m)\times(n+m)}$ is formed as the input to the GAT:

$$\mathbf{X} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}, \tag{9}$$

where the element of **D** and the element of **B** are defined as follows:

$$D_{ij} = \begin{cases} 0.5\text{DDS}(d_i, d_j) + 0.5\text{GIPKS}(d_i, d_j), & \text{if DDS}(d_i, d_j) \neq 0 \\ \text{GIPKS}(d_i, d_j), & \text{else} \end{cases} \tag{10}$$

$$B_{ij} = \begin{cases} 0.5\text{BBS}(b_i, b_j) + 0.5\text{GIPKS}(b_i, b_j), & \text{if BBS}(b_i, b_j) \neq 0 \\ \text{GIPKS}(b_i, b_j), & \text{else} \end{cases} \tag{11}$$

Graph attention network

We used a GAT [30] to learn the graph features based on biomarker similarity and disease similarity as shown in Figure 2. Given a node being a biomarker or a disease, the GAT firstly learns the attention values of its neighbors based on their importance and then combines the features of the neighbors according to their attention values. Afterward, the GAT outputs the entity graph features through K-layers aggregation based on the features of the node and its neighbors.

The multilayer aggregator is described as follows. The importance of the features of the j-th node v_j to the i-th node v_i in the k-th layer aggregator is calculated as follows:

$$e_{ij}^k = f(W_h h_i^k, W_h h_j^k), \tag{12}$$

where $f(\cdot,\cdot)$ is a single-layer feedforward neural network, W_h is a parameterized matrix that transforms the input features into hierarchical feature representation for the biomarkers and diseases and $h_i^k \in \mathcal{R}^{(n+m)}$ denotes the representation of the i-th node in the k-th layer aggregator. Note that the input of the first layer aggregator $h_*^1 \in \mathcal{R}^{(n+m)\times(n+m)}$ is the initial feature matrix **X** defined in Eq. (9). To regulate the effects across different nodes, the edge weights α_{ij}^k in the k-th layer aggregator is normalized across all the sampled associations of node v_j using the *softmax* function:

$$\alpha_{ij}^k = \frac{\exp(\text{LeakyReLU}(e_{ij}^k))}{\sum_{t\in\Omega(v_i)} \exp(\text{LeakyReLU}(e_{it}^k))}, \tag{13}$$

where $\Omega(v_i)$ denotes the neighborhood of node v_i and LeakyReLU$(\cdot)$ is a leaky rectified linear activation function (with the negative input slope set to 0.2). The normalized attention coefficients are used to compute a linear combination of the neighbors of v_i, serving as the

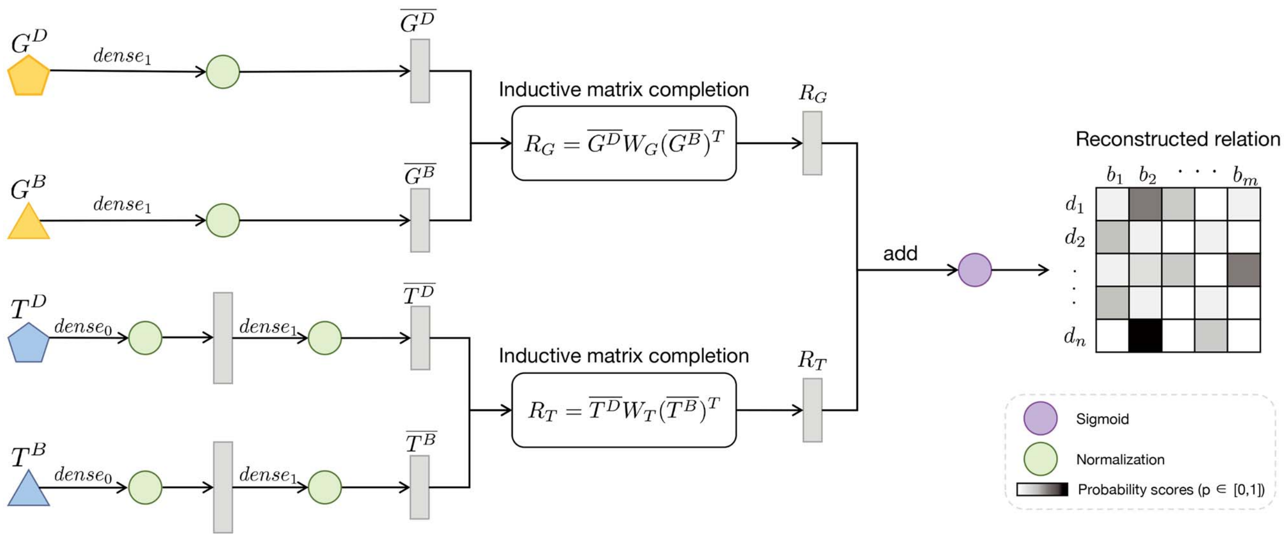


Figure 4. Bimodal fusion network architecture.

neighborhood representation.

$$h^k_{\Omega(v_i)} = \sum_{t \in \Omega(v_i)} \alpha^k_{it} h^k_t \tag{14}$$

Let $h^k_* \in \mathcal{R}^{(n+m)\times r}$ and $h^k_{\Omega(*)} \in \mathcal{R}^{(n+m)\times r}$ denote the representation of the k-th layer aggregator and its corresponding neighborhood representation, respectively, where r is a predefined feature dimension of the nodes (r is fixed to $m+n$ for the first layer, i.e. when $k=1$). Based on Eq. (14), we can concatenate h^k_* with $h^k_{\Omega(*)}$ for updating the $(k+1)$-th layer embedding as follows:

$$h^{k+1}_* = \mathrm{Elu}((h^k_* \oplus h^k_{\Omega(*)})\omega^k + \beta), \tag{15}$$

where $\oplus$ represents the concatenation operation and $\mathrm{Elu}(\cdot)$ denotes an exponential linear activation function. $\omega^k \in \mathcal{R}^{2r\times r} (\omega^1 \in \mathcal{R}^{2(m+n)\times r})$ and $\beta \in \mathcal{R}^{(n+m)\times r}$ are the parameterized weight and bias matrices, respectively. After K-layers aggregation, we can get the final graph representation matrix $G \in \mathcal{R}^{(n+m)\times r}$ for diseases and biomarkers as follows:

$$G = \begin{bmatrix} G^D \\ G^B \end{bmatrix}, \tag{16}$$

where $G^D \in \mathcal{R}^{n\times r}$ and $G^B \in \mathcal{R}^{m\times r}$ represent the final graph features of all involved diseases and biomarkers, respectively.

Text-based relation representation

We developed TRR method to automatically measure the correlations between biomarkers and diseases based on the relevant sentences retrieved from the literature. The architecture of TRR is shown in Figure 3. Each confirmed pairwise association records the related description with at least one sentence identified in the literature. Here, the association sentence set (ASS) is defined to represent the set of relevant sentences verifying a given biomarker–disease association. To limit the effect of the high-frequency associations, at most three sentences were included in each ASS.

To obtain the TRR of a biomarker–disease association, a sub-network is firstly constructed for each disease/biomarker. As shown in Figure 3, the sub-network of a biomarker encodes the interactions of this biomarker with all related diseases. Similarly, the sub-network of a disease records the interactions of this diseases with all related biomarkers. Secondly, the ASS for each association between a disease d_i and a biomarker b_j is identified. Based on the sub-networks of d_i and b_j, we also sampled ASSs for all associations involving d_i and b_j. Thirdly, the pretrained language model BioBERT [31] is adopted to estimate the correlation coefficient for each sentence in ASS to extract the sentence embedding. For the sake of efficiency, BioBERT was directly used in the training process without additional fine-tuning. In this way, the embedding scores of the associations involving d_i and b_j are aggregated to reach the final text features T^D_i and T^B_j of d_i and b_j, respectively. Compared with the manually or computationally curated databases, which can provide only ASS as the qualitative pieces of evidence of the biomarker–disease associations, the text features generated by TRR can provide quantitative measure of the biomarker–disease associations, enabling more accurate comparison or evaluation of the associations.

Bimodal fusion network

The core idea of GTGenie is to combine both graph features and text features for representing a given biomarker–disease association. To this end, we used a BFN to reconstruct the relation matrix. The architecture of the BFN is presented in Figure 4. Firstly, the graph features $G = [G^D, G^B]$ and the text features $T = [T^D, T^B]$ of diseases and biomarkers are fed into the feedforward neural networks to obtain the refined feature representations $\overline{G} = [\overline{G^D}, \overline{G^B}]$ and $\overline{T} = [\overline{T^D}, \overline{T^B}]$, respectively.

The sub-components $\overline{G^D} \in \mathcal{R}^{n\times u_1}$, $\overline{G^B} \in \mathcal{R}^{m\times u_1}$, $\overline{T^D} \in \mathcal{R}^{n\times u_1}$ and $\overline{T^B} \in \mathcal{R}^{m\times u_1}$ indicate the disease graph features, biomarker graph features, disease text features and biomarker text features output by the corresponding feedforward neural networks, respectively. The value u_1 denotes the size of the first fully connected layer $dense_1$. Secondly, the inductive matrix completion technique incorporates the side features associated with the rows (diseases) and the columns (biomarkers) to recover a low-rank parameter matrix (i.e. $W_G \in \mathcal{R}^{u_1\times u_1}$ or $W_T \in \mathcal{R}^{u_1\times u_1}$) [23, 32]. Through matrix completion, nonlinear embedding representation of graph feature and text features can be reflected and captured by the low-dimensional restriction of the embedding space. The inductive matrix completion technique is used to reconstruct the incomplete association matrix by inferring the missing values from the known information, and obtain the graph relation $R_G \in \mathcal{R}^{n\times m}$ and the text relation $R_T \in \mathcal{R}^{n\times m}$ as follows:

$$R_G = \overline{G^D} W_G (\overline{G^B})^T \tag{17}$$

$$R_T = \overline{T^D} W_T (\overline{T^B})^T \tag{18}$$

Finally, the sum of R_G and R_T is input to the sigmoid activation function to reconstruct the relation matrix. The loss function of the BFN is given below:

$$\mathcal{L}(G, T) = \|Y - \sigma(R_G + R_T)\|_F^2, \tag{19}$$

where σ represents the sigmoid activation function, Y is the biomarker–disease adjacency matrix constructed based on the training data and $\| \cdot \|_F$ calculates the Frobenius norm. BFN is iterativly optimized until the maximum number of epochs is reached.

Results and discussion

Experiment setup

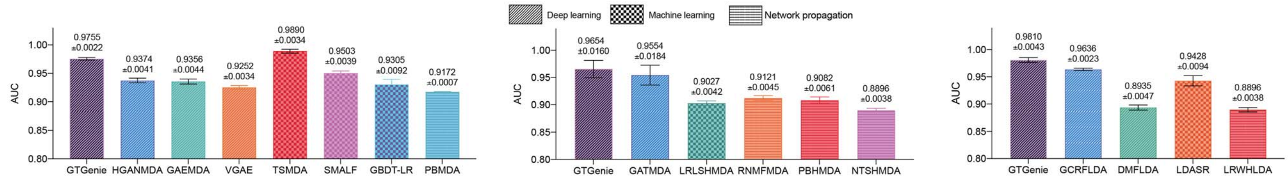
To evaluate the performance of GTGenie on the prediction of biomarker–disease associations, we compared it with other state-of-the-art methods on HMDD, HMDAD and LncRNADisease datasets with 5-fold cross-validation. Since the three data sources were curated to depict different types of biomarker–disease associations, i.e. HMDD for miRNA–disease associations, HMDAD for microbe–disease associations and LncRNADisease for lncRNA–disease associations, there are no overlap between these three data sources. An independent classification task was therefore proceeded on each dataset and the corresponding state-of-the-art methods specified on each dataset were involved in the comparison with GTGenie. In each round of cross validation, there is no overlap between the testing set and training set, i.e. the 1-fold test samples were unseen during the training of the model. The performance was measured in terms of the area under the receiver-operating characteristic (AUC) and the area under the precision-recall curve (AUPR). Since there are no available resources purposed for negative samples (i.e. known irrelevant biomarker–disease associations), we followed the most commonly used sampling strategy to generate the negative samples randomly from unknown biomarker–disease associations [33, 34]. The negative samples were randomly sampled in the beginning of the study and fixedly used throughout the study. We evaluated the average performance of GTGenie over 10 random runs of cross-validation. To verify the prediction results of GTGenie, we also cross checked the predicted biomarker–disease associations with other independent databases and published literature. All experiments were carried out on an Ubuntu 16.04 platform with Tesla K80 GPU.

GTGenie was implemented in Python and Tensorflow v1.15. The learning rates, training epochs and dropout rates were empirically set to 1e-3/500/0.1, 3e-3/300/0.0 and 3e-3/300/0.1 for the HMDD, HMDAD and LncRNADisease datasets, respectively. The effects of using different settings for these three parameters were investigated in the Supplementary Materials. The Adam optimizer [35] was used for the optimization of BFN. The number of graph aggregator layers K in GAT was set to 4. The number of neurons r in the GAT was set to 256, 64 and 64 for the HMDD, HMDAD and LncRNADisease datasets, respectively. The sizes u_0 and u_1 of the fully connected layers $dense_0$ and $dense_1$ in BFN, respectively, were set according to the value of r, i.e. $u_0 = 2u_1 = r$. The effects of the key components of GTGenie as well as the settings of the parameters including K, r, u_0 and u_1 to the performance of GTGenie were investigated in the last part of the experimental study. With the above parameter settings, the average time costs of GTGenie for performing 5-fold cross validation on HMDD, HMDAD and LncRNADisease datasets were 318, 65 and 199 s, respectively.

Comparisons with other state-of-the-art methods

GTGenie was compared with other state-of-the-art methods that have also been applied to the three datasets. Note that, since each of these methods was originally designed for a single specific prediction task, GTGenie was compared with different methods on different datasets. On HMDD dataset, GTGenie was compared with HGANMDA [22], GAEMDA [36], VGAE [37], TSMDA [14], SMALF [18], GBDT-LR [38] and PBMDA [39]. On HMDAD dataset, GTGenie was pitted against five representative methods GATMDA [33], LRLSHMDA [40], RNMFMDA [4], PBHMDA [41] and NTSHMDA [42]. On LncRNADisease dataset, GTGenie competed with four representative methods including GCRFLDA [20], DMFLDA [43], LDASR [11] and LRWHLDA [44]. The parameters of all the compared methods were set following the corresponding original references.

The results on HMDD dataset are shown in Fig. 5a, where GTGenie obtained the second best performance.



(a) HMDD (b) HMDAD (c) LncRNADisease

Figure 5. The AUC results of different methods on HMDD, HMDAD and LncRNADisease datasets.

TSMDA won on this dataset mainly due to the use of positive-unlabeled learning [45, 46] to choose reliable negative samples. Among the deep learning methods, both HGANMDA and GTGenie use graph features extracted from curated databases for predicting biomarker–disease associations, whereas, on top of graph features, GTGenie also introduces text features extracted from published literature to depict the biomarker–disease associations from different angles. Taking the advantage of both graph and text features, GTGenie achieved a higher AUC (by 3.8%) than HGANMDA on the HMDD dataset. Nevertheless, HGANMDA can reconstruct the knowledge in one database based on the information of another through a multi-view learning framework, which could be borrowed to improve the generalization ability of GTGenie. GTGenie achieved the best performance on HMDAD (as shown in Fig. 5b) and LncRNADisease (as shown in Fig. 5c). It is worth noting that GATMDA also uses GAT but not text features. It was the runner-up on HMDAD dataset. Overall, GTGenie showed better generalization ability and achieved a robust prediction performance on the three benchmark datasets. This result may attribute to the use of both explicit and implicit information sources, the concurrent capturing of both graph and text features and the orthogonal information learned in the bimodal representations.

Cross checking

In this section, the top-rank predicted associations of GTGenie that do not occur in the training set were cross checked by other independent databases, including dbDEMC v3.0 ([47]), MNDR v3.1 ([48]) and Lnc2Cancer v3.0 ([49]). Moreover, supportive pieces of evidence of the identified associations were also sought from the literature in PubMed. Given an identified association, we reported the PMID of a published paper where the association was confirmed.

We firstly investigated whether the 20 top predicted associations of GTGenie in terms of probability score could be confirmed by the independent databases or the literature. The validation results trained on HMDD v2.0, HMDAD and LncRNADisease v2017 are shown in Table 1. All the top 20 predicted associations were confirmed by the independent databases or the literature. For example, the association of *has-mir-146a* and Adenocarcinoma was verified in [50] (PMID:32382320) where the overexpression of *has-mir-146a* was reported to reduce the proliferation of human Lung adenocarcinoma cell line *A549*. The association of lncRNA *H19* and Alzheimer's disease was indicated in [51] (PMID:30107531). The silenced *H19* could accelerate the viability and repress apoptosis of *PC12* cells by stimulating *A*β25-35 in Alzheimer's disease. Since there are no proper databases available for cross checking of the microbe–disease associations, the PubMed literature were used to validate the predicted microbe-disease associations. For example, multiple bacteria taxa in the phylum *Actinobacteria* were found to be associated with the risk of Type 2 diabetes [52], which confirms the 3rd predicted microbe–disease associations on the HMDAD dataset.

We also investigated the 50 top predicted associations, where 100%, 90% and 96% associations identified on HMDD v2.0, HMDAD and LncRNAdisease v2017 were confirmed, respectively. The details are provided in the Supplementary Materials.

Effects of key model components and parameters

In this section, we investigated the effects of the implicit information sources, the graph and text features, the proposed MSS, the pretrained model BioBERT and the parameter settings (the number of graph aggregator layers *K* and number of neurons in neural networks *r*) to the performance of GTGenie.

Effects of using implicit information sources

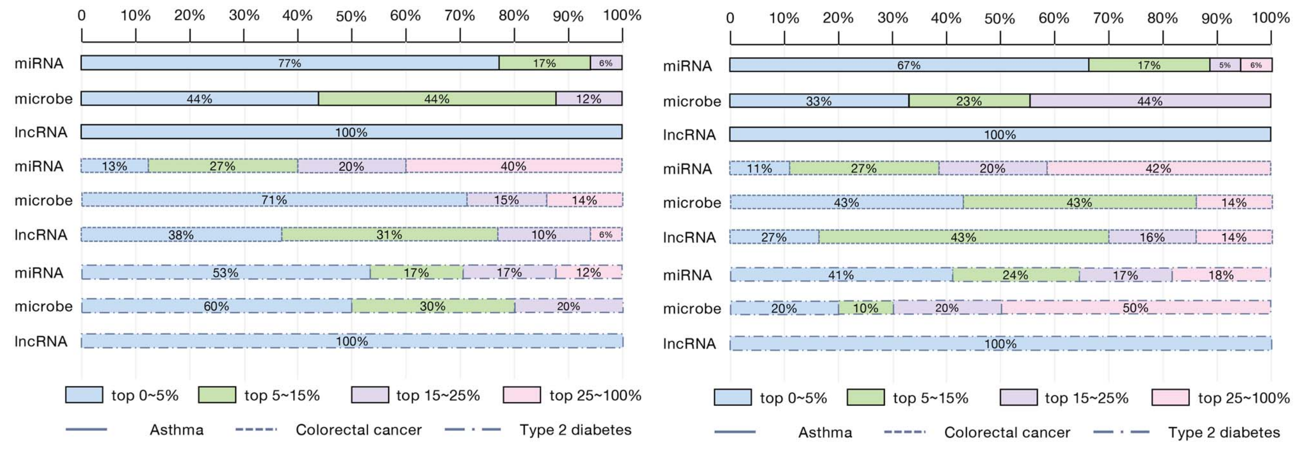
GTGenie uses implicit information sources including the MeSH, UniProt and Expression Atlas datasets for heterogeneous similarities estimation. We conducted ablation experiments to evaluate how GTGenie performs with and without using the MeSH, UniProt and Expression Atlas datasets. The model trained merely with the known associations and text features, i.e. without BBS and DDS, was compared with the conventional GTGenie. The results shown in Figure S1 of the Supplementary Materials suggest that GTGenie attained 1% improvement of AUC by using the implicit information sources.

Moreover, the implicit information sources enable GTGenie to handle the unseen associations in the training data especially for minor diseases where there are few associations existing in the training data and the corresponding GIPKS matrix constructed is sparse. In this subsection, three human diseases, i.e. Asthma, Colorectal cancer and Type 2 diabetes coexisting in HMDD, HMDAD and LncRNADisease, with relatively few associations were selected for investigating the effects of using implicit information sources. For each

Table 1. The global top-20 predicted biomarker–disease associations

	Top 1-10			**Top 11-20**		
	Disease	**Biomarker**	**Evidence**	**Disease**	**Biomarker**	**Evidence**
HMDD	Adenocarcinoma	*hsa-mir-1*	PMID:33305905	Breast neoplasms	*hsa-mir-130a*	dbDEMC
	Adenocarcinoma	*hsa-mir-146a*	PMID:32382320	Breast neoplasms	*hsa-mir-15b*	dbDEMC
	Adenocarcinoma	*hsa-mir-18a*	PMID:33942935	Breast neoplasms	*hsa-mir-144*	dbDEMC
	Adenocarcinoma	*hsa-mir-34a*	PMID:30700696	Breast neoplasms	*hsa-mir-138*	dbDEMC
	Adenoviridae infections	*hsa-mir-29a*	PMID:30405317	Breast neoplasms	*hsa-mir-142*	dbDEMC
	Adrenocortical carcinoma	*hsa-mir-21*	dbDEMC	Breast neoplasms	*hsa-mir-212*	dbDEMC
	Breast neoplasms	*hsa-mir-150*	dbDEMC	Breast neoplasms	*hsa-mir-130b*	dbDEMC
	Breast neoplasms	*hsa-mir-106a*	dbDEMC	Breast neoplasms	*hsa-mir-32*	dbDEMC
	Breast neoplasms	*hsa-mir-192*	dbDEMC	Carcinoma	*hsa-mir-9*	PMID:30795814
	Breast neoplasms	*hsa-mir-99a*	dbDEMC	Carcinoma, hepatocellular	*hsa-mir-143*	dbDEMC
HMDAD	Bacterial vaginosis	*Proteobacteria*^{P}	PMID:32296412	Asthma	*Actinobacteria*^{P}	PMID:29318023
	Type 2 diabetes	*Prevotella*^{G}	PMID:34040023	Colorectal carcinoma	*Actinobacteria*^{P}	PMID:35049922
	Type 2 diabetes	*Actinobacteria*^{P}	PMID:28177125	Bacterial vaginosis	*Bacteroidetes*^{P}	PMID:20819230
	Colorectal carcinoma	*Proteobacteria*^{P}	PMID:34650531	Clostridium difficile infection	*Prevotella*^{G}	PMID:33854066
	Crohn's disease	*Actinobacteria*^{P}	PMID:31911822	Clostridium difficile infection	*Bacteroides ovatus*^{S}	PMID:29076071
	Crohn's disease	*Proteobacteria*^{P}	PMID:31530835	Crohn's disease	*Lactobacillus*^{G}	PMID:15113451
	Crohn's disease	*Prevotella*^{G}	PMID:28542929	Psoriasis	*Prevotella*^{G}	PMID:32545459
	Crohn's disease	*Bacteroides*^{P}	PMID:33669168	Clostridium difficile infection	*Bacteroides*^{G}	PMID:32660525
	Liver cirrhosis	*Actinobacteria*^{P}	PMID:31726747	Clostridium difficile infection	*Bacteroides vulgatus*^{S}	PMID:32660525
	Asthma	*Firmicutes*^{P}	PMID:32072252	Psoriasis	*Bacteroidetes*^{P}	PMID:33384669
LncRNADisease	Alzheimer's disease	*H19*	PMID:30107531	Lung adenocarcinoma	*H19*	Lnc2Cancer;MNDR
	Alzheimer's disease	*MALAT1*	MNDR	Lung adenocarcinoma	*UCA1*	Lnc2Cancer;MNDR
	Cancer	*HOTTIP*	MNDR	Glioma	*PVT1*	Lnc2Cancer;MNDR
	Cancer	*NEAT1*	MNDR	Glioma	*UCA1*	Lnc2Cancer;MNDR
	Cancer	*TUG1*	MNDR	Nasopharyngeal carcinoma	*GAS5*	Lnc2Cancer;MNDR
	Gastric cancer	*AFAP1-AS1*	Lnc2Cancer;MNDR	Nasopharyngeal carcinoma	*PVT1*	Lnc2Cancer;MNDR
	Gastric cancer	*BCYRN1*	Lnc2Cancer;MNDR	Nasopharyngeal carcinoma	*UCA1*	Lnc2Cancer;MNDR
	Gastric cancer	*PCAT1*	Lnc2Cancer;MNDR	Hepatocellular carcinoma	*BCYRN1*	PMID:31339046
	Gastric cancer	*SOX2-OT*	MNDR	Hepatocellular carcinoma	*CRNDE*	PMID:30230527
	Lung adenocarcinoma	*CDKN2B-AS1*	MNDR	Papillary thyroid carcinoma	*H19*	PMID:31403942

Note: The superscripts *D, P, C, O, F, G* and *S* represent domain, phylum, class, order, family, genus and species, respectively.



(a) GTGenie with BBS and DDS

(b) GTGenie without BBS and DDS

Figure 6. Percentage of masked association data among the top 5*th*, 15*th* and 25*th* percentile of GTGenie predicted associations. The scale on the top shows the percentage of the inferred associations.

case, the target disease along with its associations were completely masked, i.e. excluded from the training set. Figure 6 shows the prediction results of GTGenie with and without the implicit information sources, i.e. BBS and DDS, on the test data. Given the predicted associations ranked by the probability scores, we measured the percentage of the masked associations among the top 5*th*, 15*th* and 25*th* percentile of the predictions, respectively. As shown in Fig. 6a, with BBS and DDS, the majority of the masked associations were identified at the top 25*th* percentile of the prediction results of GTGenie, where all Asthma-related associations were among the top 25*th* percentile. On average, 48%, 58% and 79% masked associations occurred at the top 5*th* percentile of the inferred miRNA-related, microbe-related and lncRNA-related associations, respectively. Without BBS and DDS, the correctly predicted masked associations of GTGenie at the top 5th percentile of the inferred miRNA-related, microbe-related and lncRNA-related associations decreased 8%, 26% and 3%, respectively (Fig. 6b). The observation suggests that using merely GIPKS cannot provide sufficient information for predicting associations of the minor diseases. The implicit information sources make a good complement to GIPKS on minor diseases. We also compared GTGenie with the recently developed method GATMDA [33], which also can make prediction on unseen associations, with the same experimental setting. GATMDA predicted 43%, 46% and 77% of the masked associations at the top 5*th* percentile of the miRNA-related, microbe-related and lncRNA-related associations, respectively. Compared with GATMDA, the proposed GTGenie with the help of implicit information sources managed to identify more masked associations on the three cases.

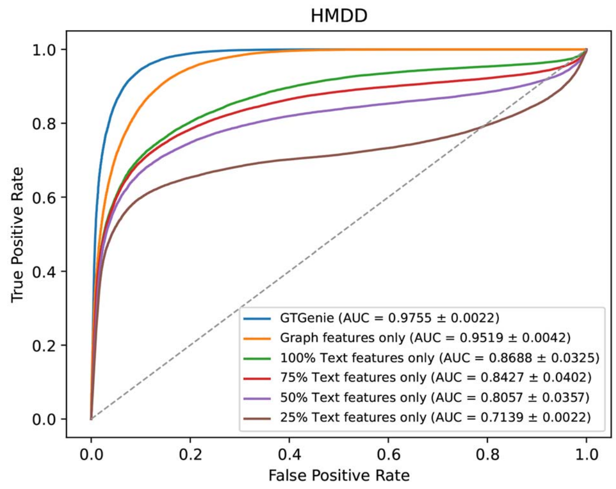


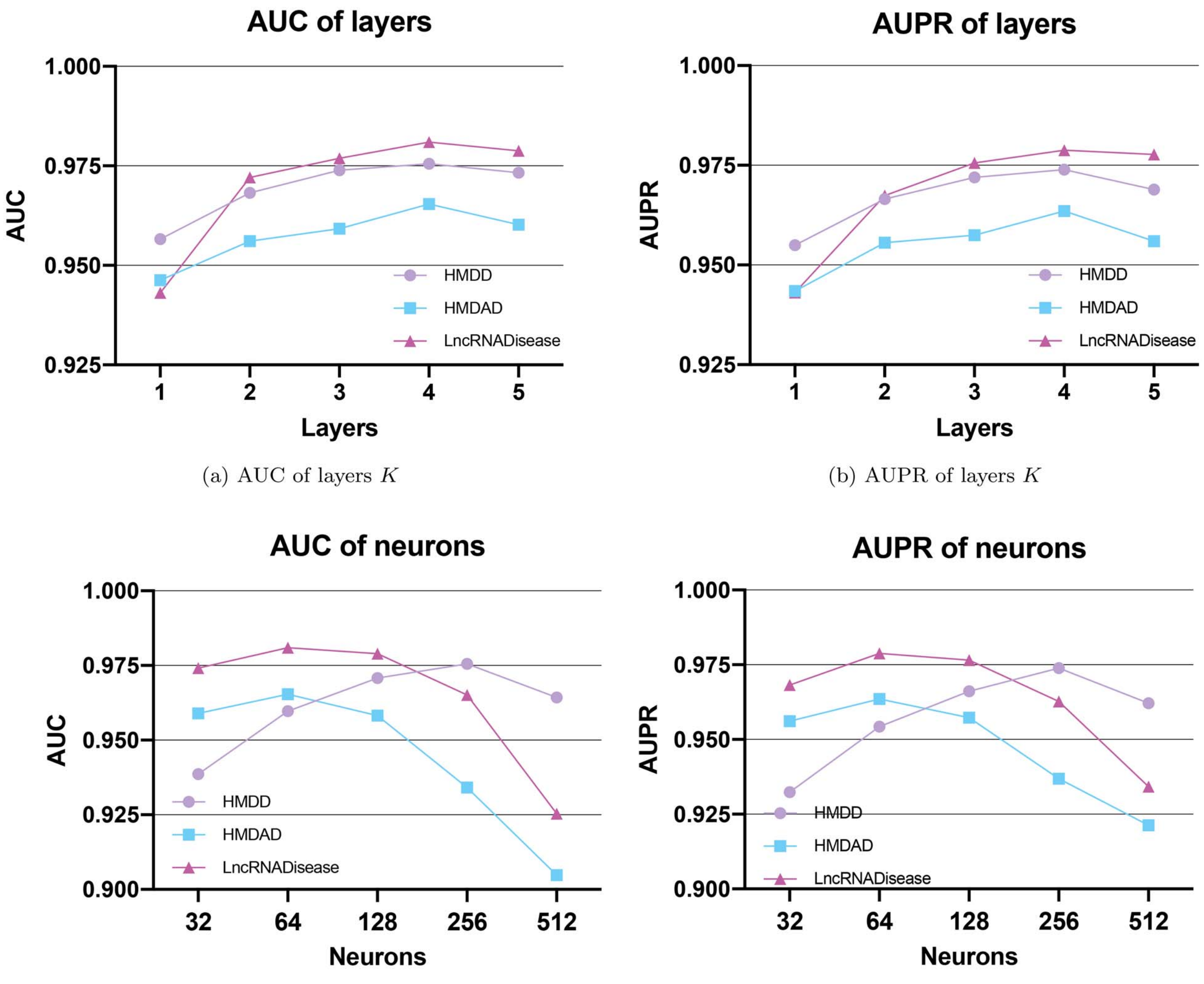
Figure 7. ROC curve analysis on the HMDD dataset with graph features and different combinations of text features.

Contributions of graph and text features

The contributions of graph and text features extracted by GAT and TRR, respectively, were investigated through ablation experiments. Excluding the text features, GTGenie suffered from degradation of AUC by 2.3%, 0.8% and 1.1% on HMDD, HMDAD and LncRNADisease, respectively. Without graph features, the prediction performance of GTGenie was restricted to the availability of text information and AUCs of 0.8688, 0.6057 and 0.6388 were obtained on HMDD, HMDAD and LncRNADisease, respectively. With respect to the existing methods, one main novelty of GTGenie is the involvement of text features. Therefore, we further explored the effects of reducing the amount of available text features on HMDD in Figure 7. As expected, the performance of GTGenie was highly sensitive to the amount of text information when only text features were used.

Effectiveness of MSS

The effectiveness of the MSS, which is a new semantic similarity metric of microbe defined in this study, was evaluated in the comparison with the other four representative microbe similarities, i.e. functional similarity [53], cosine similarity [54], GIPKS [29] and microbe taxonomic similarity [55]. As illustrated in Table 2, GTGenie integrated with MSS achieved the highest AUC mean



(a) AUC of layers K

(b) AUPR of layers K

(c) AUC of neurons r

(d) AUPR of neurons r

Figure 8. The AUC and AUPR curve with different neurons r and layers K.

Table 2. Comparison of different measurements of microbe similarities

Microbe Similarity	AUC	AUPR
Functional	0.9558 ± 0.0173	0.9529 ± 0.0219
GIPKS	0.9563 ± 0.0155	0.9545 ± 0.0211
Cosine	0.9601 ± 0.0217	0.9570 ± 0.0228
Taxonomic	0.9620 ± 0.0151	0.9589 ± 0.0175
MSS (Ours)	**0.9654 ± 0.0160**	**0.9635 ± 0.0174**

value of 0.9654. Microbe taxonomic similarity, also utilizing taxonomic information, achieves the second-best AUC of 0.9620. However, microbe taxonomic similarity only focuses on the lowest common ancestor rather than the hierarchy of taxonomic information considered in MSS. The result suggests that the hierarchy of taxonomic information is more effective in the estimation of microbe–microbe similarity.

Effects of BioBERT

In this part, the use of the other two pretrained BERT-based language models, i.e. PubMedBERT [56] and SciBERT [57], in GTGenie was tested to see the effects of BioBERT. PubMedBERT is pretrained model using the abstract and full text from PubMed papers. With an in-domain scientific vocabulary, SciBERT is pretrained on a large multi-domain corpus of scientific publications. For domain-specific language representation, BioBERT is pretrained on large-scale biomedical corpora. Table 3 shows that, using BioBERT and SciBERT, GTGenie achieved comparable performance, that is, slightly better than that of using PubMedBERT. This is because domain-specific corpora could refrain from a semantic distribution shift for frequent words. The results also suggest that GTGenie has strong robustness on the effects of BERT-based text embedding. BioBERT is suggested in GTGenie for the sake of generalization.

Influence of the key parameters

GTGenie involves two key parameters, namely, the number of graph aggregator layers K and the number of neurons r in the GAT (also used in the BFN). Their effects to the performance of GTGenie were investigated as follows. The performance of GTGenie with K = {1,

Table 3. The AUC of different pretrained BERT-based language models on HMDD, HMDAD, and LncRNADisease

	HMDD	HMDAD	LncRNADisease
PubMedBERT	0.9694 ± 0.0031	0.9604 ± 0.0157	0.9740 ± 0.0094
SciBERT	0.9746 ± 0.0025	0.9612 ± 0.0167	0.9805 ± 0.0045
BioBERT	**0.9755** ± **0.0022**	**0.9654** ± **0.0160**	**0.9810** ± **0.0043**

2, 3, 4, 5} and r = {32, 64, 128, 256, 512} is plotted in Figure 8. The best performance was achieved when $K = 4$ and r = 256/64/64 on HMDD, HMDAD and LncRNADisease, respectively. GTGenie showed good tolerance for parameter fluctuation. Larger values of K and r do not necessarily lead to better performance.

On top of the aforementioned parameters and model components, we also investigated the impact of using different ratios of positive and negative samples, and the effects of using different fusion modules and strategies in the Supplementary Materials. The experimental results justify the current configurations of GTGenie.

Conclusions

In this paper, we presented a new computational model named GTGenie for the prediction of biomarker–disease associations by considering both graph and text features. The methodological novelties of GTGenie lie in the extraction of TRR from the relevant sentences of reported literatures as well as the integration of both graph and text features into a BFN. Particularly, to effectively extract the text features, we proposed a TRR method to measure the correlations between biomarkers and diseases automatically from the relevant sentences in literature. Based on both graph and text features, a BFN was developed to reconstruct the relation matrix by filling the missing entries, i.e. identifying the unknown biomarker–disease associations. Furthermore, to improve the representation of BBS for microbe, we also proposed a novel MSS by introducing the hierarchy of taxonomic information. Extensive simulation experiments were designed to demonstrate the effectiveness of GTGenie.

Despite the promising performance achieved by GTGenie, there is room for improvement in the following potential directions. Firstly, since the feature dimensionality is fixed by the size of input similarity matrix, a trained GTGenie on one dataset should be retrained from scratch to fit another dataset. GTGenie is limited to infer new associations within a single biomarker–disease association type. To better repurpose on a second related prediction task, transfer learning techniques (e.g. multi-task learning [58], knowledge distillation [59] and multi-view learning [22]) or additional linkages (e.g. miRNA–microbe, miRNA–lncRNA and lncRNA–microbe interactions) can be introduced to speed up the training and improve the generalization performance of GTGenie. Secondly, GTGenie can also be extended to predict other types of associations/relationships (e.g. drug–disease and protein–protein) via hybrid modality fusion. The graph features and text features can be obtained from explicit and implicit datasets; however, GTGenie is likely restricted to the availability of text information. Some previous studies regarding biomedical text mining (e.g. miRiaD [60] and miRCancer [61]) could provide insights into developing an efficient text mining tool for automatically detecting more ASS in literature. Moreover, the text-based entity representation method (TER) is effective for characterizing the attributes and interaction patterns of single entity itself. Integrating TER and TRR into GTGenie can jointly decipher the linkage between biomarker and disease from perspectives of entity property and pairwise correlation, so as to extract more refined text representation. Thirdly, most GIPKS-based methods tend to have a bias toward those well-annotated entities [62, 63]. The main reasons are largely attributed to the inherent class imbalance issue in the datasets and the adopted GIPKS method that encourages higher certainty to those entities with more known associations. This issue could be further addressed using more concrete and denser similarity matrix instead of GIPKS. Fourthly, using various similarity extraction methods can help further improve the performance of GTGenie by addressing the sparsity of BBS. For example, the linguistic properties of biomarkers can be captured by the existing biological sequence analysis tools (e.g. BioSeq-Analysis2.0 [64] and BioSeq-BLM [65]), providing new insights into calculating informative BBS based on natural language processing. Finally, GTGenie is merely a computational model for assisting the identification of biomarker–disease associations. The predicted associations should be further verified with wet lab experiments or clinical trials.

Key Points

- Heterogeneous data resources and various similarity matrices were integrated to effectively measure the characteristics of biomarkers and diseases from different perspectives.
- GTGenie combined both graph and text features for the prediction of biomarker–disease associations. These two venues provided complementary information that can help improve the representation learning capability.
- We demonstrated the effectiveness and flexibility of GTGenie in the comparison with other state-of-the-art methods and the investigation of the effects of the key components.

Supplementary materials

Supplementary data are available at online https://academic.oup.com/bib and https://github.com/Wolverinerine/GTGenie.

Acknowledgments

We would like to express our gratitude toward the anonymous reviewers and the editors whose valuable comments greatly contributed to this manuscript.

Funding

This work was partially supported by the National Key Research and Development Project [2019YFE0109600], the National Natural Science Foundation of China [61871272 and 61911530218], the Shenzhen Fundamental Research Program [JCYJ20190808173617147], the open project of BGIShenzhen [BGIRSZ20200002], and the Guangdong Provincial Key Laboratory [2020B121201001].

References

1. Nimse SB, Sonawane MD, Song K-S, *et al.* Biomarker detection technologies and future directions. *Analyst* 2016;**141**(3): 740–55.
2. Mugunga I, Ying J, Liu X, *et al.* Computational prediction of human disease-related microRNAs by path-based random walk. *Oncotarget* 2017;**8**(35):58526–35.
3. Sumathipala M, Maiorino E, Weiss ST, *et al.* Network diffusion approach to predict lncRNA disease associations using multi-type biological networks: LION. *Front Physiol* 2019;**10**: 888.
4. Peng L, Shen L, Liao L, *et al.* RNMFMDA: a microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Front Microbiol* 2020;**11**:592430.
5. Rashmi JR, Rangarajan L. Global random walk for the prediction of MiRNA disease association using heterogeneous networks. In: Kaiser MS, Xie J, Rathore VS (eds). *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*. Singapore: Springer, 2021, 379–92.
6. Liu Y, Zeng X, He Z, *et al.* Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**(4):905–15.
7. Zhang X, Zou Q, Rodriguez-Paton A, *et al.* Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**(1):283–91.
8. Katz L. A new status index derived from sociometric analysis. *Psychometrika* 1953;**18**(1):39–43.
9. Lan W, Li M, Zhao K, *et al.* LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics* 2017;**33**(3): 458–60.
10. Le D-H, Pham V-H, Nguyen TT. An ensemble learning-based method for prediction of novel disease-microRNA associations. In: Thanh TN, Anh PL, Satoshi T *et al.* (eds). *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*. Vietnam. IEEE: Hue, 2017, 7–12.
11. Guo Z-H, You Z-H, Wang Y-B, *et al.* A learning-based method for lncRNA-disease association identification combing similarity information and rotation forest. *IScience* 2019;**19**:786–95.
12. Wang L, You Z-H, Chen X, *et al.* LMTRDA: using logistic model tree to predict miRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput Biol* 2019;**15**(3):e1006865.
13. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. In: Yoshua B, Yann L (eds). *1st International Conference on Learning Representations, ICLR*. United States: Scottsdale, Arizona, 2013.
14. Uthayopas K, de Sá AGC, Alavi A, *et al.* TSMDA: target and symptom-based computational model for miRNA-disease-association prediction. *Molecular Therapy-Nucleic Acids* 2021;**26**:536–46.
15. Zeng X, Wang W, Deng G, *et al.* Prediction of potential disease-associated microRNAs by using neural networks. *Molecular Therapy-Nucleic Acids* 2019;**16**:566–75.
16. Dong TN, Khosla M. MuCoMiD: a multitask convolutional learning framework for miRNA-disease association prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2021;1–1.
17. Deepthi K, Jereesh AS. An ensemble approach for circRNA-disease association prediction based on autoencoder and deep neural network. *Gene* 2020;**762**:145040.
18. Liu D, Huang Y, Nie W, *et al.* SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. *BMC Bioinformatics* 2021;**22**(1):1–18.
19. Madhavan M, Gopakumar G. DBNLDA: deep belief network based representation learning for lncRNA-disease association prediction. *Applied Intelligence* 2022;**52**:5342–52.
20. Fan Y, Chen M, Pan X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Brief Bioinform* 2022;**23**(1):bbab361.
21. Mudiyanselage TB, Lei X, Senanayake N. Graph convolution networks using message passing and multi-source similarity features for predicting circRNA-disease association. In: Park T, Cho YR, Hu X *et al.* (eds). *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Seoul, Korea: IEEE, 2020, 343–8.
22. Li Z, Zhong T, Huang D, *et al.* Hierarchical graph attention network for miRNA-disease association prediction. *Mol Ther* 2022;**30**(4):1775–86.
23. Jain P, Dhillon IS. Provable inductive matrix completion. In Hanna M.W, Hugo L, Alina B, Florence D'A, Edward A.F, Roman G, editors. *Advances in Neural Information Processing Systems 32 (NIPS)*. Vancouver, BC, Canada. 2019.
24. Yang L, Qiu C, Jian T, *et al.* HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**(D1):D1070–4.
25. Wei Ma L, Zhang PZ, Huang C, *et al.* An analysis of human microbe–disease associations. *Brief Bioinform* 2017;**18**(1): 85–97.
26. Chen G, Wang Z, Wang D, *et al.* LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 2012;**41**(D1):D983–6.
27. Wang D, Wang J, Ming L, *et al.* Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**(13):1644–50.
28. Chen X, Yan G-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 2013;**29**(20):2617–24.
29. Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 2011;**27**(21):3036–43.
30. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks. In: Yoshua B, Yann L (eds). *6th International Conference on Learning Representations, ICLR*. BC, Canada: Vancouver, 2018, 2018.
31. Lee J, Yoon W, Kim S, *et al.* Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**(4):1234–40.

32. Natarajan N, Dhillon IS. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* 2014;**30**(12):i60–8.
33. Long Y, Jiawei Luo Y, Zhang, and Yan Xia. Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Brief Bioinform* 2021;**22**(3):bbaa146.
34. Yang K, Zheng Y, Lu K, *et al.* PDGNet: predicting disease genes using a deep neural network with multi-view features. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**19**(1):575–84.
35. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Yoshua B, Yann L (eds). *3th International Conference on Learning Representations, ICLR 2015*. CA: San Diego, 2015.
36. Li Z, Li J, Nie R, *et al.* A graph auto-encoder model for miRNA-disease associations prediction. *Brief Bioinform* 2021;**22**(4):1–13.
37. Ding Y, Tian L-P, Lei X, *et al.* Variational graph auto-encoders for miRNA-disease association prediction. *Methods* 2021;**192**:25–34.
38. Zhou S, Wang S, Qi W, *et al.* Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput Biol Chem* 2020;**85**:107200.
39. You Z-H, Huang Z-A, Zhu Z, *et al.* PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol* 2017;**13**(3):e1005455.
40. Wang F, Huang Z-A, Chen X, *et al.* LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction. *Sci Rep* 2017;**7**(1):1–11.
41. Huang Z-A, Chen X, Zhu Z, *et al.* PBHMDA: path-based human microbe-disease association prediction. *Front Microbiol* 2017;**8**:233.
42. Luo J, Long Y. NTSHMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**17**(4):1341–51.
43. Zeng M, Chengqian L, Fei Z, *et al.* DMFLDA: a deep learning framework for predicting lncRNA–disease associations. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**18**(6):2353–63.
44. Li J, Zhao H, Xuan Z, *et al.* A novel approach for potential human lncRNA-disease association prediction based on local random walk. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**18**(3):1049–59.
45. Fusilier DH, Gómez M M-Y, Rosso P, *et al.* Detecting positive and negative deceptive opinions using pu-learning. *Inf Process Manag* 2015;**51**(4):433–43.
46. Liu B, Lee WS, Yu PS, *et al.* Partially supervised classification of text documents. In Claude S, Achim G.H, (ed). *ICML*, Vol. **2**. Sydney, NSW. Morgan Kaufmann Publishers Inc. 2002, 387–94.
47. Yang Z, Wu L, Wang A, *et al.* dbDEMC 2.0: updated database of differentially expressed mirnas in human cancers. *Nucleic Acids Res* 2017;**45**(D1):D812–8.
48. Lin N, Cui T, Zheng B, *et al.* MNDR v3. 0: mammal lncRNA–disease repository with increased coverage and annotation. *Nucleic Acids Res* 2021;**49**(D1):D160–4.
49. Gao Y, Shang S, Guo S, *et al.* Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res* 2021;**49**(D1):D1251–8.
50. Yuan F, Zhang S, Xie W, *et al.* Effect and mechanism of miR-146a on malignant biological behaviors of lung adenocarcinoma cell line. *Oncol Lett* 2020;**19**(6):3643–52.
51. Zhang Y-Y, Bao H-L, Dong L-X, *et al.* Silenced lncRNA H19 and up-regulated microRNA-129 accelerates viability and restrains apoptosis of PC12 cells induced by a β25-35 in a cellular model of Alzheimer's disease. *Cell Cycle* 2021;**20**(1):112–25.
52. Long J, Cai Q, Steinwandel M, *et al.* Wei Zheng, and Xiao Ou Shu. Association of oral microbiome with type 2 diabetes risk. *J Periodontal Res* 2017;**52**(3):636–43.
53. Zhang W, Yang W, Xiaoting L, *et al.* The bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access* 2018;**6**:38052–61.
54. Chuanyan W, Gao R, Zhang D, *et al.* PRWHMDA: human microbe-disease association prediction by random walk on the heterogeneous network with PSO. *Int J Biol Sci* 2018;**14**(8):849.
55. Ma Y, Jiang H. NinimHMDA: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe–disease association. *Bioinformatics* 2020;**36**(24):5665–71.
56. Yu G, Tinn R, Cheng H, *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 2021;**3**(1):1–23.
57. Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text. In: Inui K, Jiang J, Ng V *et al.* (eds). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: China, 2019, 3615–20.
58. Huang F, Yang Q, Li Q, *et al.* Predicting drug-disease associations via multi-task learning based on collective matrix factorization. *Front Bioeng Biotechnol* 2020;**8**:218.
59. Yang C, Liu J, Shi C. Extract the knowledge of graph neural networks and go beyond it: An effective knowledge distillation framework. In: Leskovec J, Grobelnik M, Najork M *et al.* (eds). *Proceedings of the Web Conference 2021*. New York, NY: United States. Association for Computing Machinery, 2021, 1227–37.
60. Gupta S, Ross KE, Tudor CO, *et al.* miRiaD: a text mining tool for detecting associations of microRNAs with diseases. *Journal of Biomedical Semantics* 2016;**7**(1):1–15.
61. Xie B, Ding Q, Han H, *et al.* miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics* 2013;**29**(5):638–44.
62. Huang Z, Liu L, Gao Y, *et al.* Benchmark of computational methods for predicting microRNA-disease associations. *Genome Biol* 2019;**20**(1):1–13.
63. Zhou S, Xuan Z, Wang L, *et al.* A novel model for predicting associations between diseases and lncRNA-miRNA pairs based on a newly constructed bipartite network. *Comput Math Methods Med* 2018;**2018**:1–11.
64. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res* 2019;**47**(20):e127–7.
65. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform* 2022;**23**(2):bbab569.