Truncated Arctangent Rank Minimization and Double-strategy Neighborhood Constraint Graph Inference for Drug-Disease Association prediction

Tiyao Liu[a], Shudong Wang[a,*], Shanchen Pang[a] and Xiaodong Tan[a]

[a]College of Computer Science and Technology, Qingdao Institute of Software, China University of Petroleum, Qingdao 266580, China

ARTICLE INFO

Keywords:
drug-disease associations
truncated arctangent rank minimization
neighborhood constraint
double-strategy graph inference
association prediction

ABSTRACT

Accurately identifying new therapeutic uses for drugs is essential to advancing pharmaceutical research and development. Graph inference techniques have shown great promise in predicting drug-disease associations, offering both high convergence accuracy and efficiency. However, most existing methods fail to sufficiently address the issue of numerous missing information in drug-disease association networks. Moreover, existing methods are often constrained by local or single-directional reasoning. To overcome these limitations, we propose a novel approach, truncated arctangent rank minimization and double-strategy neighborhood constraint graph inference (TARMDNGI), for drug-disease association prediction. First, we calculate Gaussian kernel and Laplace kernel similarities for both drugs and diseases, which are then integrated using nonlinear fusion techniques. We introduce a new matrix completion technique, referred to as TARM. TARM takes the adjacency matrix of drug-disease heterogeneous networks as the target matrix, and enhances the robustness and formability of the edges of DDA networks by truncated arctangent rank minimization. Additionally, we propose a double-strategy neighborhood constrained graph inference method to predict drug-disease associations. This technique focuses on the neighboring nodes of drugs and diseases, filtering out potential noise from more distant nodes. Furthermore, the DNGI method employs both top-down and bottom-up strategies to infer associations using the entire drug-disease heterogeneous network. The synergy of the dual strategies can enhance the comprehensive processing of complex structures and cross-domain associations in heterogeneous graphs, ensuring that the rich information in the network is fully utilized. Experimental results consistently demonstrate that TARMDNGI outperforms state-of-the-art models across two drug-disease datasets, one lncRNA-disease dataset, and one microbe-disease dataset.

1. Introduction

In the traditional drug development process, substantial resources often wasted on drugs that ultimately prove ineffective [1]. As a result, there is an urgent need for new strategies to reduce the time and cost of drug development. One such approach is drug repositioning, which involves finding new therapeutic uses for existing drugs [2]. This strategy has gained significant attention because it focuses on repurposing drugs that have already been developed or approved, offering the potential to treat new diseases [3]. Drug repositioning leverages existing safety and efficacy data from approved drugs, enabling faster clinical translation and significantly lower development costs compared to developing a drug from scratch [4, 5]. Identifying drug-disease associations (DDAs) is crucial for both drug discovery and repositioning. By predicting these associations, medical researchers can prioritize the most promising therapeutic candidates, streamlining resource allocation and increasing the likelihood of success [6].

Advances in bioassay and screening technologies are expanding our understanding of drugs and diseases, allowing for more possibilities in treatment discovery [7]. Identifying DDAs can harness this valuable data to explore new therapeutic options. However, traditional DDA predictions based on biological experiments are often labor-intensive and time-consuming, significantly limiting the efficiency of the prediction process. Therefore, the development of robust computational approaches is essential to accelerating progress in this field.

Graph inference techniques are widely employed in association prediction due to their high efficiency and interpretability. However, existing methods often fail to adequately address the substantial number of missing edges in drug-disease networks. Matrix completion can effectively fill these gaps. For instance, SLHGISMMA [8] employs a kernel-paradigm-based sparse learning approach to complement missing values in the correlation matrix, subsequently deriving predicted scores through heterogeneous graph inference. Similarly, TSPN [9] constructs a heterogeneous network using enriched biological information and achieves predictions based on the adjacency matrix of this network through truncated schatten p-paradigm minimization. However, current matrix completion techniques struggle to efficiently approximate the rank function, often leading to suboptimal solutions. Furthermore, existing graph inference techniques typically rely on local network inference or one-way reasoning, which presents several limitations. For example, NetPro [10] first preprocesses the similarity and association matrices before employing label propagation to

*Corresponding author
z22070050@s.upc.edu.cn (T. Liu); shudongwang2013@sohu.com (S. Wang); pangsc@upc.edu.cn (S. Pang); reyes.tan@foxmail.com (X. Tan)
ORCID(s):

predict DDAs. In contrast, DR-IBRW [11] utilizes double randomized wandering for association predictions. Both methods reason locally, roaming the similarity network to the association network, and do not fully leverage the information available from the entire heterogeneous network. Although HGIMC [12] generates predicted scores based on the comprehensive heterogeneous network using heterogeneous graph inference techniques, it considers only single-direction reasoning paths, lacking a holistic perspective. Additionally, these existing methods are vulnerable to the influence of remote nodes during the inference process.

To overcome these limitations, Therefore, we propose a novel approach named truncated arctangent rank minimization and double-strategy neighborhood constraint graph inference (TARMDNGI) to predict DDAs. First, we calculate Gaussian kernel and Laplace kernel similarities for both drugs and diseases, which are then integrated using nonlinear fusion techniques. We introduce a new matrix complementation technique, referred to as TARM. TARM takes the adjacency matrix of drug-disease heterogeneous networks as the target matrix, and enhances the robustness and formability of the edges of DDA networks by truncated arctan rank minimization. Additionally, we propose a double-strategy neighborhood-constrained graph inference method to predict DDAs. This technique focuses on the neighboring nodes of drugs and diseases, filtering out potential noise from more distant nodes. Furthermore, the DNGI method employs both top-down and bottom-up strategies to infer associations across the entire drug-disease heterogeneous network. The synergy of the dual strategies can enhance the comprehensive processing of complex structures and cross-domain associations in heterogeneous graphs, ensuring that the rich information in the network is fully utilized.

In summary, the key contributions of this paper are as follows:

(1) We introduce a new matrix completion technique, called TRAM, to enhance the robustness and formative nature of drug-disease association network edges. Unlike previous methods that make it difficult to accurately approximate the rank function, TRAM enhances this approximation by using well-defined mathematical expressions to rationalize the relationship between singular values and rank. Structurally, TRAM truncates and retains the most significant singular values while minimizing the less important ones and preserving certain prior values. This approach significantly reduces noise without distorting the underlying structure of the matrix.

(2) We developed a novel graph inference technique called DNGI, which focuses on leveraging the neighboring nodes of drugs and diseases within the network, while disregarding potential noise from distant nodes. DNGI integrates both top-down and bottom-up strategies to reason across the entire drug-disease heterogeneous network. The synergy of the dual strategies can enhance the comprehensive processing of complex structures and cross-domain associations in heterogeneous graphs, ensuring that the rich information in the network is fully utilized.

(3) Experimental results demonstrate that TARMDNGI consistently outperforms state-of-the-art models across two drug-disease datasets, one lncRNA-disease dataset, and one microbe-disease dataset. The outstanding performance observed in two case studies further highlights the effectiveness of TARMDNGI in practical applications.

The remainder of this paper is organized as follows: Section 2 reviews related work. The proposed method is detailed in Section 3, followed by the presentation of experimental results in Section 4. Finally, Section 5 offers the conclusions of the paper.

2. Related Work

2.1. Matrix completion

Matrix completion techniques are widely applied in association prediction tasks due to their efficiency in predicting missing values within a matrix. For instance, SLHGISMMA [8] leverages a kernel-paradigm-based sparse learning approach to fill in missing values in the correlation matrix, followed by deriving predicted scores through heterogeneous graph inference. Similarly, AMCSMMA [13] integrates three types of biological data to construct heterogeneous networks and recovers missing values by minimizing the truncated kernel paradigm in the network's adjacency matrix. TSPN [9] also fuses three types of biological information to build a heterogeneous network, using truncated schatten p-norm minimization on the adjacency matrix to derive predicted scores. However, these methods have limited capacity to approximate the rank function, leading to suboptimal solutions.

We introduce a new matrix completion technique, called TRAM. Unlike previous methods that make it difficult to accurately approximate the rank function, TRAM enhances this approximation by using well-defined mathematical expressions to rationalize the relationship between singular values and rank. Structurally, TRAM truncates and retains the most significant singular values while minimizing the less important ones and preserving certain prior values. This approach significantly reduces noise without distorting the underlying structure of the matrix. Therefore, TRAM approximates the rank function better and is able to effectively fill in missing values in the matrix.

2.2. Graph inference

Graph inference is a powerful method for predicting unknown molecular associations within complex bioinformatics networks. For example, NetPro [10] first preprocesses the similarity and association matrices and then uses label propagation to predict DDAs. Similarly, DR-IBRW [11] utilizes double randomized walks to make such predictions. MDHGI [14] enhances the association matrix through matrix decomposition, followed by heterogeneous graph inference based on bioinformatics networks to generate predicted scores. However, these approaches have certain limitations. They are confined to either local or unidirectional reasoning within heterogeneous networks, preventing them from leveraging global network information and capturing the intricate

bidirectional interactions inherent in such networks. Furthermore, they are susceptible to the influence of distant nodes during the inference process, reducing their robustness.

We developed a novel graph inference technique called DNGI, which focuses on leveraging the neighboring nodes of drugs and diseases within the network, while disregarding potential noise from distant nodes. DNGI integrates both top-down and bottom-up strategies to reason across the entire drug-disease heterogeneous network. The synergy of the dual strategies can enhance the comprehensive processing of complex structures and cross-domain associations in heterogeneous graphs, ensuring that the rich information in the network is fully utilized.

3. Methods

3.1. TARMDNGI

Figure 1 illustrates the overall framework of the TARMDNGI model. First, Gaussian and Laplace kernel similarities are calculated for both drugs and diseases, which are then integrated using nonlinear fusion techniques. Next, we introduce a novel matrix completion technique, TRAM, to enhance the robustness and structural integrity of the DDA network edges. Finally, we propose a double-strategy neighborhood constrained graph inference method to predict DDAs.

3.2. Datasets

To evaluate the performance of TARMDNGI, we conducted a series of experiments on four datasets, encompassing diverse association prediction tasks across diverse domains. Specifically, we tested the method on three different types of datasets: drug-disease, lncRNA-disease, and microbe-disease. For the drug-disease datasets, We deliberately chose two datasets at different scales to validate the scalability of the model, containing 269 drugs and 598 diseases, respectively. The larger dataset contains 18,416 known DDAs from the Comparative Toxicogenomics Database (CTD) [15]. In addition, we used a smaller therapeutic dataset containing 6244 annotated therapeutic DDAs from the CTD for performance comparison. For the lncRNA-disease dataset, we gathered 605 associations involving 82 lncRNAs and 157 diseases from lncRNADisease [16]. For the microbe-disease dataset, we collected 450 microbe-disease associations involving 292 microbes and 39 diseases from HMDAD [17]. In this context, the CDA matrix is denoted as $A_{CD} \in R^{m \times n}$, where m corresponds to the number of drugs and n denotes the number of diseases.

3.3. Similarity measures

3.3.1. Gaussian kernel similarity

The Gaussian kernel function is widely used for calculating biosimilarity due to its strong performance and generalizability across various types of biological data. In this study, we applied the Gaussian kernel to compute drug and disease similarities based on known DDAs. $S(r_i)$ denotes the vector storing row i of the association matrix and $S(d_j)$ denotes the vector storing column j of the association matrix. The calculation equations are presented below:

$$A_R^1(r_i, r_j) = exp\left(-q_r \left\|S(r_i) - S(r_j)\right\|^2\right) \tag{1}$$

$$A_D^1(d_i, d_j) = exp\left(-q_d \left\|S(d_i) - S(d_j)\right\|^2\right) \tag{2}$$

where A_R^1 denotes the drug gaussian kernel similarity, and A_D^1 represents the disease gaussian kernel similarity. The adjustable parameters g_r and g_d are calculated using the following equations:

$$g_r = 1/\frac{1}{m}\sum_{i=1}^{m} \left\|S(r_i)\right\|^2 \tag{3}$$

$$g_d = 1/\frac{1}{n}\sum_{i=1}^{n} \left\|S(d_i)\right\|^2 \tag{4}$$

3.3.2. Laplace kernel similarity

Compared to the Gaussian kernel function, the Laplace kernel function exhibits greater robustness in handling noisy or anomalous data. In this section, we compute the Laplace similarity for both drugs and diseases, with the calculation equations provided below:

$$A_R^2(r_i, r_j) = exp(\frac{\left\|S(r_i) - S(r_j)\right\|}{-\sigma}) \tag{5}$$

$$A_D^2(d_i, d_j) = exp(\frac{\left\|S(d_i) - S(d_j)\right\|}{-\sigma}) \tag{6}$$

where A_R^2 denotes the Laplace kernel similarity for drugs, while A_D^2 represents the disease laplace kernel similarity. The parameter σ is used to control the bandwidth of the function, and throughout this research, we consistently set σ to 0.5.

3.4. Similarity kernel fusion

To mitigate the bias introduced by relying on a single similarity measure, we apply a nonlinear fusion method to separately integrate the two drug and disease similarities.

Here, the similarity kernel fusion method was employed to combine two drug similarity kernels (A_R^1 and A_R^2). The process begins by standardizing each miRNA similarity kernel, as described below:

$$P_h(i, j) = \frac{A_R^h(i, j)}{\sum_{k=1}^{m} A_R^h(k, j)} \tag{7}$$

Next, we construct neighbor similarity kernels for each of the two drugs to effectively capture their local similarities.

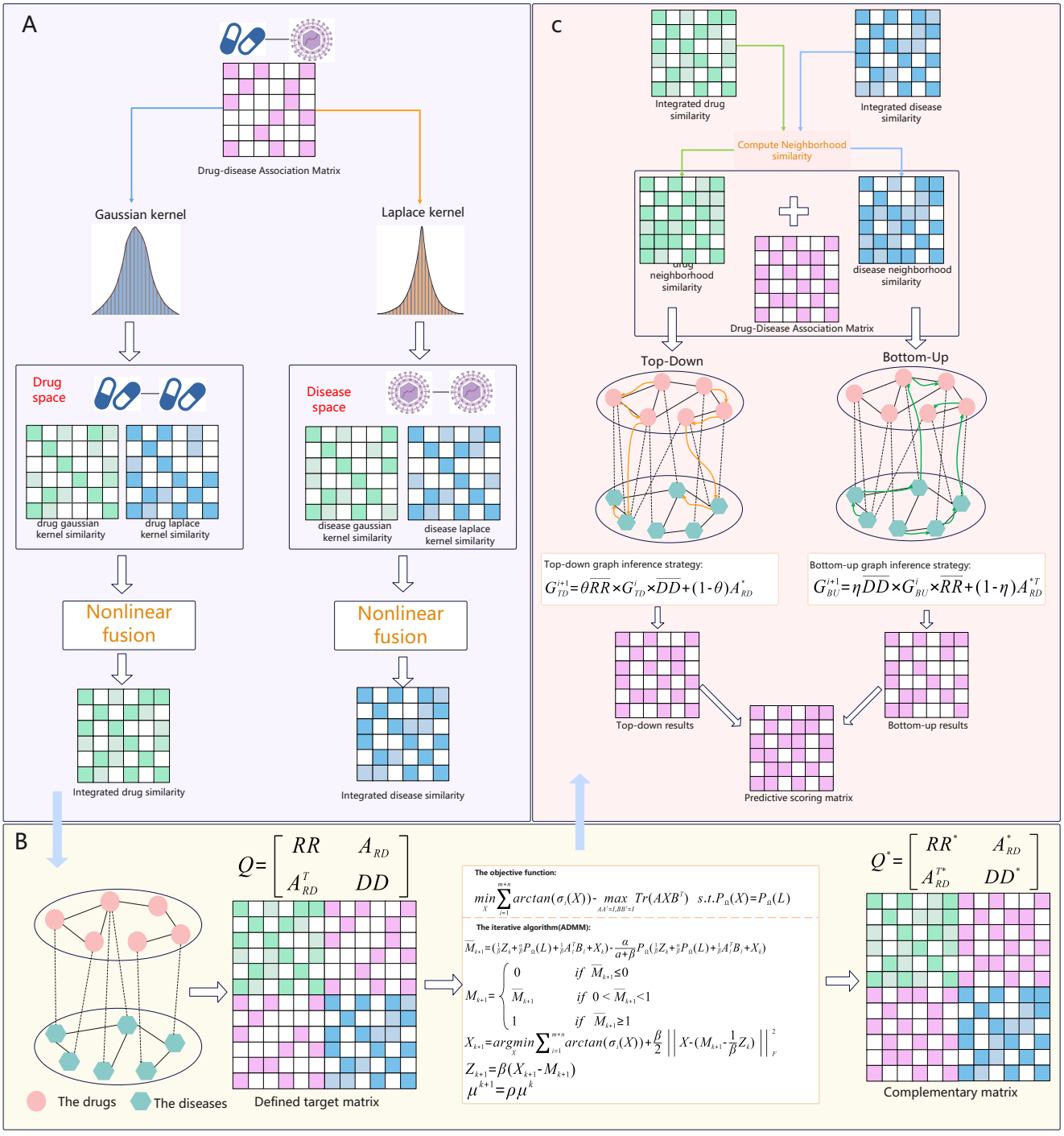

Figure 1: Framework of TARMDNGI. (A) Calculate the Gaussian and Laplace kernel similarities for both drugs and diseases, then combine them using a nonlinear fusion method. (B) Constructing drug-disease heterogeneous networks, where the formation and robustness of association network edges are enhanced through truncated tangent rank minimization. (C) The predicted scores are generated using a two-strategy neighborhood-constrained graph inference approach.

$$S_h(i,j)=\begin{cases}\dfrac{A_R^h(i,j)}{\sum_{k\in N_i}A_R^h(i,k)} & j\in N_i\\ 0 & otherwise\end{cases} \tag{8}$$

where S_h denotes the h-th neighbor similarity kernel. Here, the neighbor set N_i of drug r_i is characterized as the k most similar drugs to r_i.

We combine the normalized similarity kernel P_h and the neighbor similarity kernel S_h literally, as detailed below:

$$P_h^{t+1} = \alpha(S_h \times \frac{\sum_{r \neq h} P_r^t}{2} \times S_h^T) + (1-\alpha)\frac{\sum_{r \neq h} P_r^0}{2} \tag{9}$$

where P_h^{t+1} indicates the h-th kernel acquired after t-th iterations, P_r^0 denotes the initial value of P_r^t, and $\alpha \in (0,1)$ indicates the equilibrium coefficient. After t-th iterations, the drug similarity kernel is derived the following:

$$SR = \frac{1}{2}\sum_{h=1}^{2} P_h^{t+1} \tag{10}$$

Next, the weight matrix embedding the neighbor information is constructed. The formula is detailed below.

$$w(i,j) = \begin{cases} 1 & if\ i \in N_j \cap j \in N_i \\ 0 & if\ i \notin N_j \cap j \notin N_i \\ 0.5 & otherwise \end{cases} \tag{11}$$

Finally, we obtain the integrated drug similarity kernel $RR \in R^{m*m}$ as follows:

$$RR = w(i,j) * SR \tag{12}$$

Similarly, we derive the integrated disease similarity kernel $DD \in R^{n*n}$.

3.5. Truncated arctangent rank minimization

In this work, We introduce a new matrix completion technique, called TRAM, to enhance the robustness and formative nature of drug-disease association network edges. Unlike previous methods that faced challenges in accurately approximating the rank function, TRAM improves this approximation by effectively linking singular values to the rank through precise mathematical expressions. Structurally, TRAM truncates and retains the most significant singular values while minimizing the less important ones and preserving certain prior values. This approach significantly reduces noise without distorting the underlying structure of the matrix. We start by integrating the drug-disease association network with the similarity network to construct a heterogeneous drug-disease network. The adjacency matrix of this combined network is used as the target matrix for truncated arctangent rank minimization. The corresponding objective function is then efficiently solved using the Alternating Direction Method of Multipliers (ADMM).

3.5.1. Constructing the heterogeneous network

In this section, we integrate the similarity matrices and association matrix to construct a drug-disease heterogeneous network, enriched with biological information. This network consists of two types of nodes: drug nodes and disease nodes. Let $R = \{R_1, R_2, \ldots, R_m\}$ represent m drug nodes and $D = \{D_1, D_2, \ldots, D_n\}$ represent n disease nodes. The similarity scores between drugs and between diseases are used as edge weights within the drug and disease similarity networks, respectively. Additionally, when a drug-disease association exists in the dataset, an edge is established between the corresponding drug and disease nodes, with the association score serving as the weight of this edge. We define E_{rr}, E_{dd}, E_{rd} as the sets of edges representing drug-drug, disease-disease, and drug-disease relationships, respectively. The weights on these edges are denoted by W_{rr}, W_{dd}, and W_{rd}, corresponding to the drug-drug, disease-disease, and drug-disease networks. Thus, the drug-disease heterogeneous network can be represented as $G_{RD} = \{\{R, D\}, \{E_{rr}, E_{dd}, E_{rd}\}, \{W_{rr}, W_{dd}, W_{rd}\}\}$.

The adjacency matrix $L \in R^{(m+n)(m+n)}$ of this heterogeneous network is constructed in the form shown below.

$$L = \begin{bmatrix} RR & A_{RD} \\ A_{RD}^T & DD \end{bmatrix} \tag{13}$$

3.5.2. Building TARM

Matrix Completion (MC) is an efficient data recovery technique that reconstructs a complete matrix by filling in its missing values. In this study, we apply MC to tackle the issue of extreme sparsity in the drug-disease association matrix.

Algorithm 1 ADMM

Require: $Q = \begin{bmatrix} RR & A_{RD} \\ A_{RD}^T & DD \end{bmatrix}$

Ensure: A_{RD}^*

$X_1 = P_\Omega(L),\ M_1 = X_1,\ Z_1 = X_1,\ l = 0,\ k = 0$

outer-loop:

$l \leftarrow l + 1$

X_l: $[U_l,\ \triangle_l,\ V_l] = SVD(X_l)$, where $U_l = (\mu_1, \ldots, \mu_{m+n}), V_l = (\nu_1, \ldots, \nu_{m+n})$.

$A_l = (\mu_1, \ldots, \mu_r)^T, B_l = (\nu_1, \ldots, \nu_r)^T$.

inner-loop:

$k \leftarrow k + 1$

$\overline{M}_{k+1} = (\frac{1}{\beta}Z_k + \frac{\alpha}{\beta}P_\Omega(L) + \frac{1}{\beta}A_l^T B_l + X_k) - \frac{\alpha}{\alpha+\beta}P_\Omega(\frac{1}{\beta}Z_k + \frac{\alpha}{\beta}P_\Omega(L) + \frac{1}{\beta}A_l^T B_l + X_k)$

$$M_{k+1} = \begin{cases} 0 & if \quad \overline{M}_{k+1} \leq 0 \\ \overline{M}_{k+1} & if \quad 0 < \overline{M}_{k+1} < 1 \\ 1 & if \quad \overline{M}_{k+1} \geq 1 \end{cases}$$

$X_{k+1} = \underset{X}{\operatorname{argmin}} \sum_{i=1}^{m+n} arctan(\sigma_i(X)) + \frac{\beta}{2}\left\| X - \left(M_{k+1} - \frac{1}{\beta}Z_k\right)\right\|_F^2$

$Z_{k+1} = \beta\left(X_{k+1} - M_{k+1}\right)$

$\mu^{k+1} = \rho\mu^k$

Until: $\frac{\|X_{k+1} - M_{k+1}\|_F}{\|X_{k+1}\|_F} \leq \Phi$

Until: Iteration number l=1(Main dataset)

$Q^* = \begin{bmatrix} RR^* & A_{RD}^* \\ A_{RD}^{T*} & DD^* \end{bmatrix}$

return A_{RD}^*

The underlying model of MC is summarized as follows:

$$\begin{aligned} &\min_X \operatorname{rank}(X) \\ &\text{s.t. } P_\Omega(X) = P_\Omega(L) \end{aligned} \tag{14}$$

Here $rank(\cdot)$ denotes the rank function, Ω represents the set of coordinate indices for the elements in L, and the operator P_Ω refers to the orthogonal projection operator applied to Ω.

$$(P_\Omega(X))_{ij} = \begin{cases} X_{ij} & \text{if (i, j)} \in \Omega \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Equation (14) presents an NP-hard problem, for which no efficient optimization algorithm currently exists to solve it directly. A notable advancement was made by Fazel et al. [18], who addressed this challenge by approximating the rank function through the nuclear norm paradigm. The corresponding formula is provided below.

$$\begin{aligned} &\min_X \|X\|_* \\ &\text{s.t. } P_\Omega(X) = P_\Omega(L) \end{aligned} \tag{16}$$

where $\|X\|_* = \sum_{i=1}^{m} \sigma_i(X)$ denotes the nuclear norm of X. Numerous studies have shown that the nuclear norm can serve as an approximate substitute for the rank function. However, it is not an optimal replacement, and practical applications yield only suboptimal solutions. In recent years, several alternative rank relaxation paradigms to the kernel paradigm have been proposed, including the truncated nuclear norm [19], the schatten p-norm [20], the truncated schatten p-norm [21], and the γ-norm [22]. Nonetheless, all these rank relaxation paradigms have certain limitations and do not fully account for both the mathematical properties and the physical structure of matrix rank.

To recover the target matrix more accurately and efficiently, we use the arctangent function as a replacement for the rank function, as shown in equation (24).

$$\begin{aligned} &\min_X \sum_{i=1}^{m+n} arctan(\sigma_i(X)) \\ &\text{s.t. } P_\Omega(X) = P_\Omega(L) \end{aligned} \tag{17}$$

where $\sigma_i(X)$ denotes the i-th singular value of matrix X, and $arctan(|\cdot|)$ is differentiable, concave, monotonically increasing and you-invariant on [0,1]. Inspired by previous studies [13], we have developed an effective matrix completion technique. Building on Equation 17, we have further transformed it. The complete proof procedure can be found in Appendix 1.

$$\begin{aligned} &\min_X \sum_{i=1}^{m+n} arctan(\sigma_i(X)) - \max_{AA^T=I, BB^T=I} Tr(AXB^T) \\ &\text{s.t. } P_\Omega(X) = P_\Omega(L) \end{aligned} \tag{18}$$

where $A \in R^{r\times(m+n)}$, $B \in R^{r\times(m+n)}$, and $I \in R^{r\times r}$ denotes the identity matrix.

Inspired by previous studies, we introduce regularization terms to accommodate potential noise in the block matrix [23]. We restricted all values of the matrix to the interval [0, 1] to ensure their practical relevance [24]. In summary, we constructed the following model:

$$\begin{aligned} \min_X \sum_{i=1}^{m+n} arctan(\sigma_i(X)) &- \max_{AA^T=I, BB^T=I} Tr(AXB^T) \\ &+ \frac{\alpha}{2} \left\| P_\Omega(X) - P_\Omega(L) \right\|_F^2 \\ \text{s.t. } 0 \le X_{ij} &\le 1 (0 \le i, j \le m+n) \end{aligned} \tag{19}$$

where α is a balancing parameter and $0 \le X_{i,j} \le 1$ (where $0 \le i, j \le m+n$) indicates that all elements in X are in the [0, 1] range.

3.5.3. Solving TARM

We developed a framework that employs ADMM to solve the model. By introducing new variables to streamline the computations, we alternated the update direction in each iteration.

$$\begin{aligned} \min_X \sum_{i=1}^{m+n} arctan(\sigma_i(X)) &- \max_{AA^T=I, BB^T=I} Tr(AXB^T) \\ &+ \frac{\alpha}{2} \left\| P_\Omega(X) - P_\Omega(L) \right\|_F^2 \\ \text{s.t. X} = \text{M}, 0 \le X_{ij} &\le 1 (0 \le i, j \le m+n) \end{aligned} \tag{20}$$

The augmented Lagrangian form of equation (20) can be expressed as shown below:

$$\begin{aligned} F(M, X, Z, \alpha, \beta) = \min_X \sum_{i=1}^{m+n} arctan(\sigma_i(X)) &- \max_{AA^T=I, BB^T=I} Tr(AXB^T) \\ + \frac{\alpha}{2} \left\| P_\Omega(X) - P_\Omega(L) \right\|_F^2 &+ \text{Tr}\left(\text{Z}^{\text{T}}(X - M)\right) + \frac{\beta}{2} \|X - M\|_F^2 \end{aligned} \tag{21}$$

where Z indicates the Lagrange multiplier, β denotes the penalty parameter, and we initialized all matrices (M_1,X_1 and Z_1) to L. During the k-th iterations, M_{k+1}, X_{k+1}, and Z_{k+1} are computed sequentially. The solution process is shown in Algorithm 1. Thus, we obtain the complemented association matrix A^*_{RD}.

3.6. Double-strategy Neighborhood Constraint Graph Inference

We propose DNGI, a double-strategy neighborhood constrained graph inference method for predicting unknown DDAs. DNGI introduces several key improvements over previous graph inference techniques: (1) It emphasizes that similar drugs are more likely to be associated with similar diseases, focusing on the neighborhood nodes of both drugs and diseases to minimize noise from distant, less relevant nodes. (2) DNGI operates within a drug-disease heterogeneous network, utilizing both top-down and bottom-up strategies for a more comprehensive graph inference. The synergy between these dual strategies enhances the

processing of complex structures and cross-domain associations in heterogeneous graphs, ensuring full utilization of the network's rich information. The process begins by calculating the neighbor similarity for drugs and diseases based on their adjacent nodes. A comprehensive heterogeneous network is then constructed by integrating the well-established DDA network with drug and disease neighbor similarities. Finally, inference is performed on this network using both top-down and bottom-up approaches, effectively capturing key associations.

3.6.1. Compute neighborhood similarity of drug/disease

Building on the assumption that similar drugs are more likely to be associated with similar diseases, we further hypothesize that the likelihood of association between a disease $d(i)$ and drug $r(i)$ should closely correspond to the association score derived from the neighbors of $d(i)$ and $r(i)$. We compute the neighborhood similarity of a drug by analyzing its neighboring nodes in the drug similarity network. For a given drug $r(i)$, we select the K_r most similar drugs based on the integrated drug similarity, denoting this neighbor set as $N(r(i))$. The neighborhood similarity matrix, $\overline{RR}$, is then defined, with its elements calculated as shown below.

$$\overline{RR}(i,l) = \begin{cases} \frac{RR(i,j)}{\sum_{r(p)\in N(r(i))} RR(i,p)} & r(l) \in N(r(i)) \\ 0 & otherwise \end{cases} \tag{22}$$

Similarly, the K_d nearest neighbors of disease $d(i)$ form the set of neighbors denoted as $N(d(i))$. The disease neighborhood similarity matrix, $\overline{DD}$, is computed in a similar manner.

$$\overline{DD}(i,l) = \begin{cases} \frac{DD(i,j)}{\sum_{d(p)\in N(d(i))} DD(i,p)} & d(l) \in N(d(i)) \\ 0 & otherwise \end{cases} \tag{23}$$

3.6.2. Double-strategy Graph Inference

We constructed comprehensive drug-disease heterogeneous network by integrating the complemented DDA networks with drug and disease neighbor similarity networks. Predictions were then made using both top-down and bottom-up graph inference strategies, as depicted in Figure 2. The graph inference model, which incorporates these dual strategies, is illustrated below.

Top-down graph inference strategy:

$$G_{TD}^{i+1} = \theta\overline{RR} \times G_{TD}^{i} \times \overline{DD} + (1-\theta)A_{RD}^{*} \tag{24}$$

Bottom-up graph inference strategy:

$$G_{BU}^{i+1} = \eta\overline{DD} \times G_{BU}^{i} \times \overline{RR} + (1-\eta)A_{RD}^{*T} \tag{25}$$

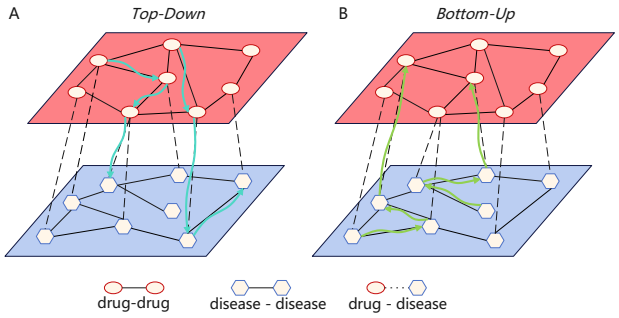

Figure 2: (A) Global inference is conducted along a top-down path, progressing from drug similarity networks to association networks and then to disease similarity networks. (B) Global inference is conducted along a bottom-up path, progressing from disease similarity networks to association networks and then to drug similarity networks.

Here, G_{TD} and G_{BU} represent the prediction results from the top-down and bottom-up graph inference strategies, respectively. The parameters θ and η correspond to the weights assigned to the top-down and bottom-up strategies. Before initiating the inference process, the neighborhood similarity matrices $\overline{RR}$ and $\overline{DD}$, must be normalized.

$$\overline{RR}(i,j) = \frac{\overline{RR}(i,j)}{\sqrt{\sum_{h=1}^{m}\overline{RR}(i,h)}\sqrt{\sum_{h=1}^{m}\overline{RR}(h,j)}} \tag{26}$$

$$\overline{DD}(i,j) = \frac{\overline{DD}(i,j)}{\sqrt{\sum_{h=1}^{n}\overline{DD}(i,h)}\sqrt{\sum_{h=1}^{n}\tilde{d}(h,j)}} \tag{27}$$

Equation (15) anb (16) will converge when the iteration stopping condition $\left\|G_{TD}^{i+1} - G_{TD}^{i}\right\|_1 \leq 10^{-6}$ and $\left\|G_{BU}^{i+1} - G_{BU}^{i}\right\|_1 \leq 10^{-6}$ are satisfied.

Next, we combine the association matrices computed by the top-down and bottom-up strategies to obtain the final predictive scoring matrix $A_{RD}^{\sim}$.

$$A_{RD}^{\sim} = \frac{G_{TD} + G_{BU}^{T}}{2} \tag{28}$$

4. Results and discussions

4.1. Experimental Settings

Implementation details. In this section, K-Fold cross-validation (CV) is employed to evaluate the performance of TARMDNGI. Specifically, K-Fold CV involves dividing all known microbial-disease associations into K equal parts. Each part is then sequentially used as a test set, while the remaining K-1 parts are used as the training set. In this paper, we utilize the widely adopted 5-fold and 10-fold CV methods to assess the accuracy of the model.

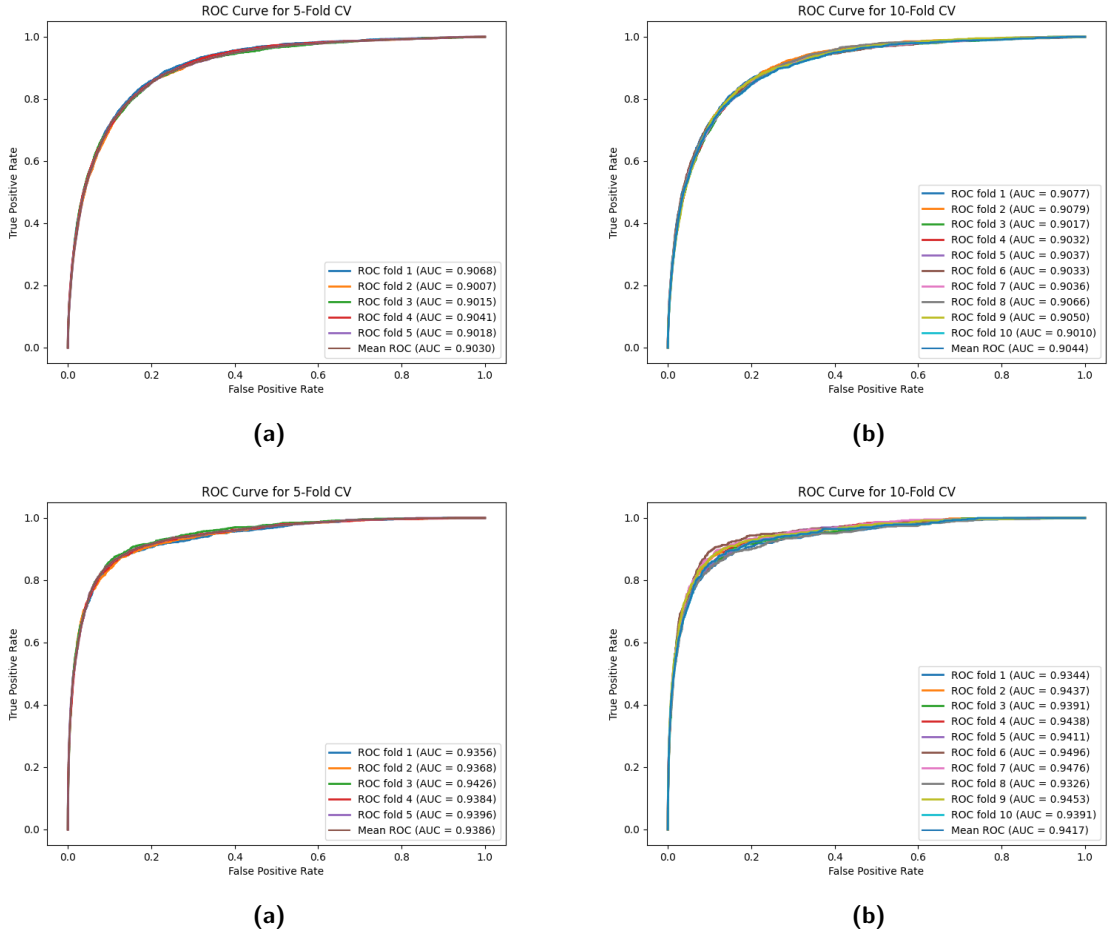

Figure 3: (a) TARMDNGI executes 5-fold CV on the Main dataset. (b) TARMDNGI executes 10-fold CV on the Main dataset. (c) TARMDNGI executes 5-fold CV on the Therapeutic dataset. (d) TARMDNGI executes 10-fold CV on the Therapeutic dataset.

Table 1
Multi-metric performance evaluation of the seven algorithms based on two drug-disease datasets

Dataset	Module	AUPR	AUC	Accuracy	F1-Score	Precision	Recall
Main dataset	TARMDNGI	**0.2890**	**0.9030**	**0.9621**	**0.3463**	**0.2957**	**0.4191**
	HDGAT	0.2665	0.8642	0.9614	0.2964	0.2762	0.3221
	NIMCGCN	0.2002	0.8533	0.9572	0.2661	0.2341	0.3083
	LAGCN	0.2287	0.8486	0.9588	0.2701	0.2445	0.3021
	DeepDR	0.1351	0.8211	0.9400	0.1991	0.1500	0.2959
	TL-HGBI	0.0665	0.7029	0.9114	0.1266	0.0843	0.2545
	DRRS	0.1321	0.8429	0.9324	0.2178	0.1631	0.3276
Therapeutic dataset	TARMDNGI	**0.2851**	**0.9386**	**0.9895**	**0.3436**	**0.3446**	0.3437
	HDGAT	0.2707	0.8877	0.9783	0.2204	0.1549	0.3829
	NIMCGCN	0.0899	0.8075	0.9798	0.1525	0.1160	0.2225
	LAGCN	0.2777	0.8810	0.9713	0.1910	0.1236	**0.4216**
	DeepDR	0.1011	0.8572	0.9806	0.1610	0.1231	0.2327
	TL-HGBI	0.0388	0.7401	0.9761	0.0720	0.0524	0.1151
	DRRS	0.1383	0.8754	0.9849	0.2249	0.1914	0.2726

Evaluation indicators. In our experiments, we use the AUC, AUPR, Accuracy, F1-Score, Precision, and Recall metrics to assess the model's performance.

Baselines. We compare our proposed TARMDNGI with several baseline methods, categorized as follows:

• HDGAT [25] predicts unknown DDAs using a coding-reconstruction-regression framework, effectively capturing the complex interactions between drugs and diseases.

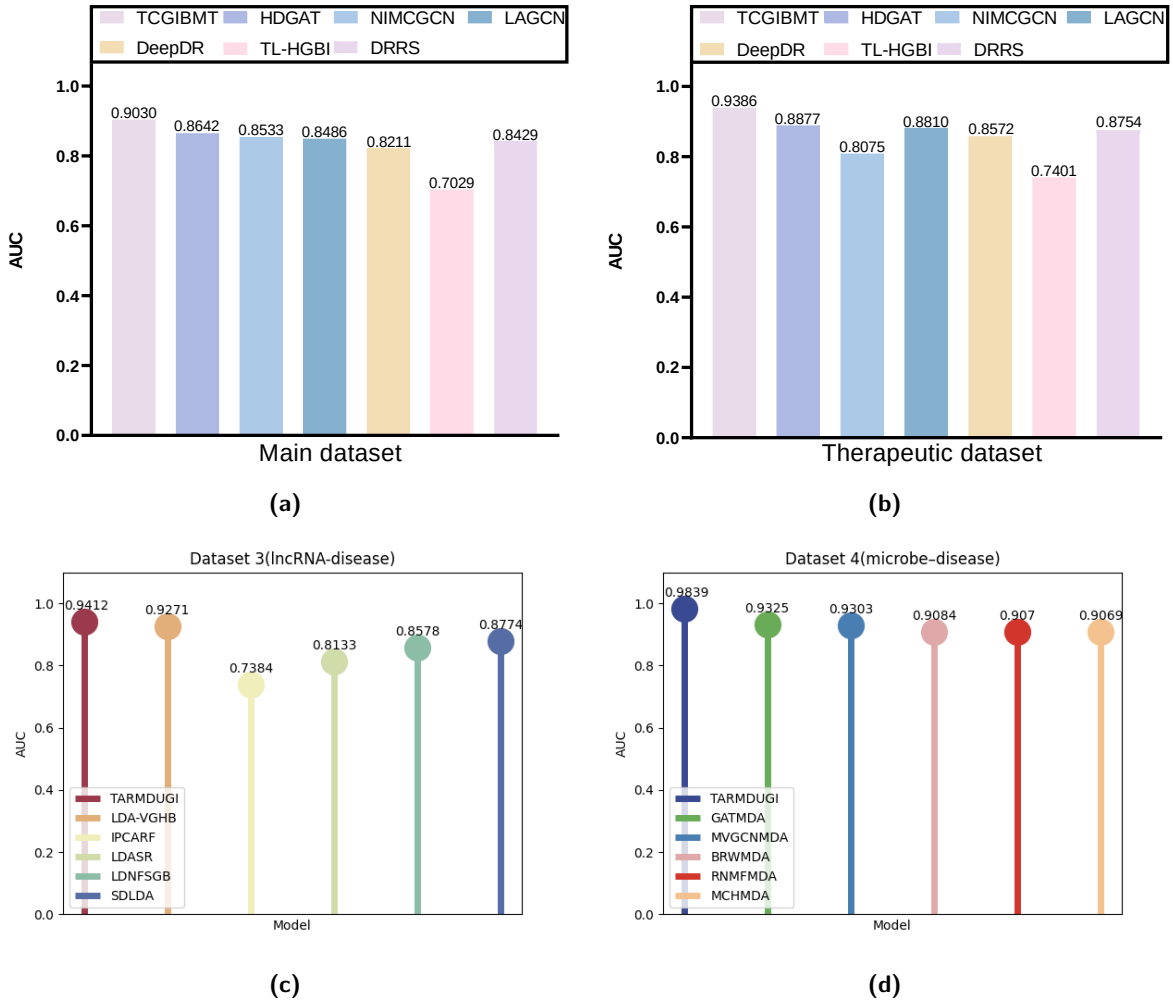

Figure 4: (a) Performance comparison of TARMDNGI with several methods in Main dataset. (b) Performance comparison of TARMDNGI with several methods in Therapeutic dataset. (c) Performance comparison of TARMDNGI with several methods in lncRNA-disease dataset. (d) Performance comparison of TARMDNGI with several methods in microbe-disease dataset.

• NIMCGCN [26] leverages graph convolutional network to learn biological features, and predictions are subsequently made using a novel neural induction matrix complementation approach.

• LAGCN [27] learns node embeddings within biologically heterogeneous networks and combines the embeddings from multiple graph convolutional layers using an attention mechanism.

• DeepDR [28] extracts biometric features using deep autoencoders and further processes them through a variant autoencoder for prediction.

• TL-HGBI [29] employs a label propagation algorithm to predict DDAs in heterogeneous information networks.

• DRRS [30] constructs heterogeneous networks based on biosimilarity and associations, with predictions made using a singular value thresholding algorithm.

4.2. Performance Evaluation

In this study, we conducted both 5-fold and 10-fold CV on two drug-disease datasets to assess TARMDNGI's predictive performance. As shown in Figure 3 (a) and (b), for the Main dataset, TARMDNGI achieved AUC values of 0.9030 and 0.9044 for 5-fold and 10-fold CV, respectively. For the Therapeutic dataset, the AUC values were 0.9386 and 0.9417 (Figure 3 (c) and (d)).

To further validate TARMDNGI's effectiveness, we compared it with six recently proposed models: HDGAT, NIMCGCN, LAGCN, DeepDR, TL-HGBI, and DRRS. As displayed in Figure 4 (a) and (b), TARMDNGI achieved AUC values of 0.9030 on the Main dataset and 0.9386 on the Therapeutic dataset, outperforming many state-of-the-art methods. Additionally, we conducted 5-fold CV experiments on both datasets using other evaluation metrics to further establish the model's superiority. The detailed results are shown in Table 1. On the Main dataset, TARMDNGI achieved the best performance in terms of AUPR, AUC, Accuracy, F1-Score, Precision, and Recall, with values of 0.2890, 0.9030, 0.9621, 0.3463, 0.2957, and 0.4191, respectively. Similarly, on the Therapeutic dataset, TARMDNGI attained an AUPR of 0.2851, an AUC of 0.9386,

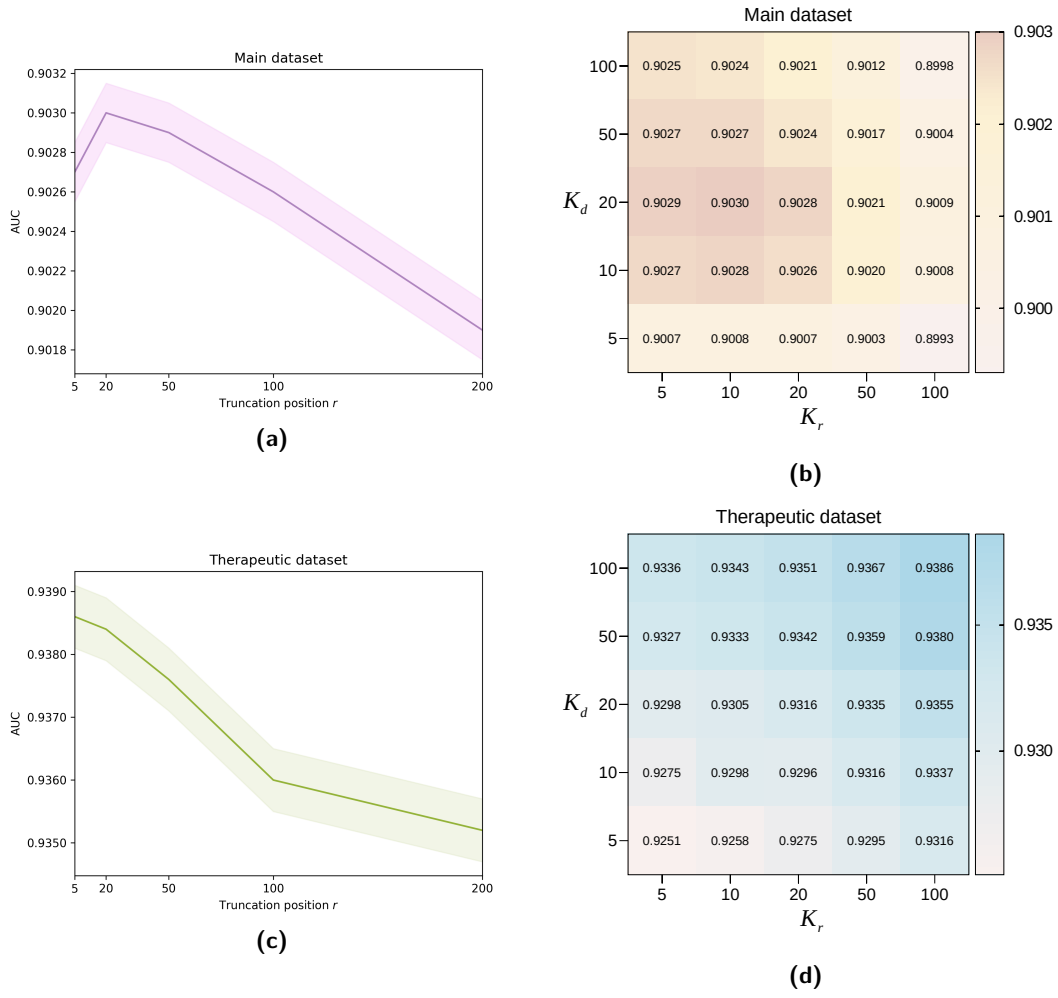

Figure 5: (a) Sensitivity analysis of parameter r on Main dataset. (b) Sensitivity analysis of parameter k_r and k_d on Main dataset. (c) Sensitivity analysis of parameter r on Therapeutic dataset. (d) Sensitivity analysis of parameter k_r and k_d on Therapeutic dataset.

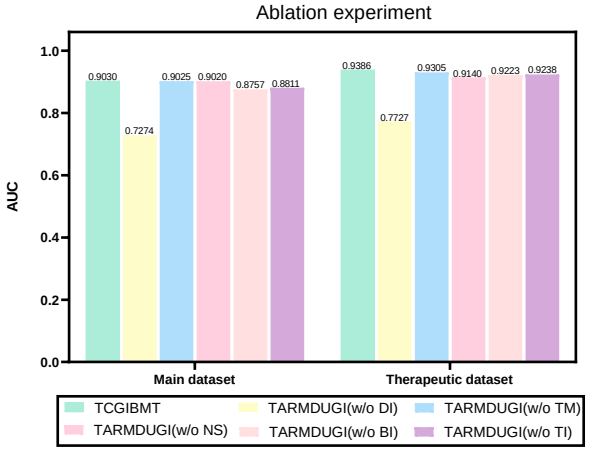

Figure 6: Results of ablation experiments.

an Accuracy of 0.9895, an F1-Score of 0.3436, a Precision of 0.3446, and a Recall of 0.4216, surpassing the comparison models. In summary, TARMDNGI demonstrated strong performance across different datasets, consistently outperforming other state-of-the-art methods. These results confirm the models accuracy and effectiveness.

4.3. Generalized Molecular Association Prediction Study

To demonstrate the strong generalization ability of TARMDNGI, we applied it to both lncRNA-disease and microbe-disease datasets and compared its performance with several state-of-the-art models. For the lncRNA-disease dataset, we benchmarked TARMDNGI against LDA-VGHB [31], IPCARF [32], LDASR [33], LDNFSGB [34], and SDLDA [35]. As shown in Figure 4 (c), TARMDNGI achieved an AUC value of 0.9412, outperforming all other models. Likewise, for the microbe-disease dataset, we compared TARMDNGI with GATMDA [36], MVGCNMDA [37], BRWMDA [38], RNMFMDA [39], and MCHMDA [40]. As illustrated in Figure 4 (d), TARMDNGI attained an AUC value of 0.9839, surpassing all competing models. These results underscore TARMDNGI's versatility and exceptional effectiveness in molecular interaction prediction.

4.4. Parameter sensitivity analysis

We investigated the effects of several key parameters on TARMDNGI: (1) Truncation position for truncated arctangent rank minimization, denoted as r, which controls the number of singular values and consequently determines the matrix dimension. (2) Number of neighbors for specific drugs and diseases in the local similarity network, denoted as θ and η, which are used in constructing the local similarity networks. We conducted parameter sensitivity analyses using two datasets with 5-fold CV.

The importance of the truncation position r. We evaluated the performance of TARMDNGI across various singular value truncation positions r from $\{5, 20, 50, 100, 200\}$. As illustrated in Figure 5 (a), TARMDNGI achieves optimal performance on Main dataset when r=20. For Therapeutic dataset, the best performance is observed with $r = 5$ (Figure 5 (c)). Based on these findings, we set the singular value truncation position to r=20 for Main dataset and $r = 5$ for Therapeutic dataset.

The importance of the number of neighbors k_r and k_d. We analyzed the impact of varying the number of neighbors k_r and k_d in equation (24-26) on the graph inference technique. Specifically, k_r and k_d were explored within the range $\{5, 10, 20, 50, 100\}$. As shown in Figure 5(b), the model performs optimally on the Main dataset with k_r=10 and k_d=20. For the Therapeutic dataset, the best performance is achieved with k_r=100 and k_d=100 (Figure 5(d)). Our analysis provides the following insights: (1) For the Main dataset, the optimal number of drug neighbors is higher than the number of disease neighbors, suggesting that drug similarity data points in this dataset exhibit less variation than disease similarity data points. (2) The number of neighbors is set to 20 for drugs and 5 for diseases in the Main dataset, while both are set to 100 in the Therapeutic dataset. This indicates that drug and disease similarity points in the Main dataset show greater variability, making it more effective to focus on the most relevant neighbors, whereas in the Therapeutic dataset, a broader range of neighbors is beneficial.

4.5. Ablation Study

To evaluate the significance of various components in our model, we analyzed the following variants of TARMDNGI:

- **TARMDNGI without double-strategy unbalanced graph inference (w/o DI)**: This variant omits the double-strategy unbalanced graph inference module.
- **TARMDNGI without truncated arctangent rank minimization (w/o TM)**: This variant excludes the truncated arctangent rank minimization module.
- **TARMDNGI without calculating neighbor similarity (w/o NS)**: This variant removes the imbalance property from the graph reasoning module.
- **TARMDNGI without bottom-up graph inference strategy (w/o BI)**: This variant uses only top-down graph inference, removing bottom-up graph inference.
- **TARMDNGI without top-down graph inference strategy (w/o TI)**: This variant uses only bottom-up graph inference, removing top-down graph inference.

The results of the ablation study under 5-fold cross-validation, as shown in Figure 6, indicate that all variants of TARMDNGI perform worse than the full model, highlighting the importance of each component in CDA prediction. Additionally, we made the following observations: (1) The results for TARMDNGI (w/o DI) and TARMDNGI (w/o TM) demonstrate that both the double-strategy unbalanced graph inference and truncated arctangent rank minimization modules are essential. Notably, TARMDNGI (w/o TM) outperforms TARMDNGI (w/o DI), suggesting that double-strategy unbalanced graph inference is particularly significant. (2) TARMDNGI (w/o NS) shows lower performance than the full TARMDNGI, emphasizing the model's focus on local network details while minimizing distant noise. (3) The results for TARMDNGI (w/o BI) and TARMDNGI (w/o TI) confirm that the combination of the two graph inference strategies is both effective and necessary.

4.6. Case studies

To assess TARMDNGI's ability to identify novel drugs in practical application, we conducted additional experiments using Main dataset. In these experiments, all confirmed DDAs were used as training data, while unconfirmed associations were selected for prediction. Our goal was to identify potential new drug candidates for Alzheimer's disease (AD) and type 2 diabetes (T2D). Drug candidates for each disease were ranked in descending order based on their predicted scores by TARMDNGI, and the top ten ranked drugs for each disease were subsequently validated through relevant open literature.

Alzheimer's disease is a neurodegenerative disorder with a gradual onset, marked by the progressive decline of cognitive function, impaired ability to perform daily activities, and changes in mental behavior [41]. Table 2 lists the top 10 AD drugs predicted by TARMDNGI, of which nine were successfully validated. Notably, Quercetin, typically used to treat cardiovascular diseases, was identified by our model as a potential treatment for AD. Research suggests that Quercetin, a flavonoid, holds significant therapeutic potential for improving cognitive function in Alzheimer's disease [42].

Type 2 diabetes, also known as adult-onset diabetes mellitus, has a gradual onset with no early symptoms and accounts for approximately 95% of all diabetes cases [43]. Table 2 lists the top 10 T2D drugs predicted by TARMDNGI, of which seven were successfully validated. For instance, Losartan, a drug typically used to treat hypertension, was identified by our model as a potential treatment for T2D. Supporting literature confirms that Losartan alleviates T2D-induced hypoactivity and normalizes corticosterone levels, nociception, and memory function [44]. In summary, our model has identified several promising treatments for both AD and T2D.

Table 2
TARMDNGI predicts top 10 drugs associated with alzheimer disease and type 2 diabetes

Disease	Drug	Evidence
Alzheimer diseasen	Cocaine	PMID: 31702488
	Aspirin	PMID: 36846997
	Methotrexate	PMID: 32423175
	Diclofenac	PMID: 36875711
	Quercetin	PMID: 31905923
	Doxorubicin	unconfirmed
	Dexamethasone	PMID: 1713799
	Acetaminophen	unconfirmed
	Caffeine	PMID: 35744865
	Diazepam	PMID: 37607528
Type 2 diabetes	Quercetin	PMID: 34824300
	Losartan	PMID: 34824300
	Niacin	PMID: 21241985
	Olanzapine	PMID: 24890070
	Benazepril	PMID: 30793466
	Risperidone	PMID: 29770221
	Ramipril	PMID: 16965776
	Cocaine	unconfirmed
	Prednisone	unconfirmed
	Aspirin	unconfirmed

5. Conclusion

Our models demonstrate strong effectiveness in predicting unknown DDAs, offering potential for personalized treatment plans. We developed a novel matrix completion technique, TRAM, designed to enhance the robustness and structural integrity of DDA network edges. Unlike previous methods, which struggle to accurately approximate the rank function, TRAM improves this approximation through well-defined mathematical expressions that clarify the relationship between singular values and rank. Structurally, TRAM truncates and retains the most significant singular values while minimizing the influence of less important ones, preserving key prior values. This approach significantly reduces noise without distorting the underlying structure of the matrix. TARMDNGI, on the other hand, focuses on leveraging neighboring nodes of drugs and diseases within the network, minimizing the impact of noise from distant nodes. It integrates both top-down and bottom-up strategies to reason across the entire drug-disease heterogeneous network. The synergy of these dual strategies enhances the comprehensive analysis of complex structures and cross-domain associations in heterogeneous graphs, ensuring that rich network information is fully utilized. Experimental results consistently show that TARMDNGI outperforms state-of-the-art models across two drug-disease datasets, one lncRNA-disease dataset, and one microbe-disease dataset. Moreover, the exceptional performance observed in two case studies further underscores TARMDNGIs effectiveness in practical applications.

While our approach has many strengths, there are still areas that could be improved. First, the parameters used may not be fully optimized, which could affect the model's performance. Second, we did not account for the specific mechanisms underlying changes in drug and disease expression. In the future, we aim to establish multilevel associations between drugs, diseases, and other biomolecules, while exploring more complex biological mechanisms of action.

6. Appendix 1

By Von Neumanns trace inequality, we obtain:

$$Tr(AXB^T) = Tr(XB^TA) \leq \sum_{i=1}^{m+n} \sigma_i(X)\sigma_i(B^TA) \tag{29}$$

where $X \in R^{(m+n)\times(m+n)}$, $A \in R^{r\times(m+n)}$, $B \in R^{r\times(m+n)}$, $AA^T = I$, $BB^T = I$. Here, $r(r \leq (m+n))$ is a non-negative integer, $I \in R^{r\times r}$ denotes the identity matrix, and σ_i represents the i-th singular value, satisfying the relationship $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_{(m+n)} \geq 0$. Since $rank(A) = rank(B) = r$, we have $rank(B^TA) = s \leq r$. Given that $B^TA(B^TA)^T = I$, it follows that $\sigma_i(B^TA) = 1$ for $i \leq s$ and 0 otherwise. Thus, we have:

$$\begin{aligned}\sum_{i=1}^{m+n} \sigma_i(X)\sigma_i(B^TA) &= \sum_{i=1}^{s} \sigma_i(X)\sigma_i(B^TA) + \sum_{i=s}^{m+n} \sigma_i(X)\sigma_i(B^TA) \\ &= \sum_{i=1}^{s} \sigma_i(X)\cdot 1 + \sum_{i=s}^{m+n} \sigma_i(X)\cdot 0 = \sum_{i=1}^{s} \sigma_i(X)\end{aligned} \tag{30}$$

Combining inequalities inequality (29) and equation (30), we have:

$$Tr\left(AXB^T\right) \leq \sum_{i=1}^{m+n} \sigma_i(X)\sigma_i(B^TA) = \sum_{i=1}^{s} \sigma_i(X) \leq \sum_{i=1}^{r} \sigma_i(X) \tag{31}$$

The singular values of a matrix decrease diagonally, and the sum of the first r singular values provides the best approximation to the rank function. Similarly, the inverse tangent rank function offers a closer approximation to the rank. Therefore, we adopt a hybrid approach to the singular values: we retain the first r larger singular values and minimize the remaining ones, while still preserving some prior information to ensure that the matrix structure remains intact. Then, we have $\sum_{i=1}^{m+n} arctan(\sigma_i(X)) - \max_{AA^T=I,BB^T=I} Tr(AXB^T)$

7. Data availability

The code for TARMDNGI and datasets are accessible on GitHub at https://github.com/Lty-Skyrocket/TARMDNGI.git.

8. Declaration of competing interest

The authors state that they do not have any known competing interests.

9. Funding

This research received backing from the National Key Research and Development Project of China (2021YFA1000102, 2021YFA1000103).

References

[1] Bram Van de Sande, Joon Sang Lee, Euphemia Mutasa-Gottgens, Bart Naughton, Wendi Bacon, Jonathan Manning, Yong Wang, Jack Pollard, Melissa Mendez, Jon Hill, et al. Applications of single-cell rna sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22(6):496–520, 2023.

[2] Divya Ajmeera and Rajanna Ajumeera. Drug repurposing: A novel strategy to target cancer stem cells and therapeutic resistance. *Genes & Diseases*, 11(1):148–175, 2024.

[3] Trisha Bhatia and Shweta Sharma. Drug repurposing: Insights into current advances and future applications. *Current Medicinal Chemistry*, 2024.

[4] Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, pages 1–13, 2024.

[5] Xin Feng, Zhansen Ma, Cuinan Yu, and Ruihao Xin. Mrndr: multi-head attention-based recommendation network for drug repurposing. *Journal of Chemical Information and Modeling*, 64(7):2654–2669, 2024.

[6] Arsalan Wafi and Reza Mirnezami. Translational–omics: future potential and current challenges in precision medicine. *Methods*, 151:3–11, 2018.

[7] Min Seong Kim, Hyesoo Kim, and Gabsang Lee. Precision medicine in parkinson's disease using induced pluripotent stem cells. *Advanced Healthcare Materials*, page 2303041, 2024.

[8] Jun Yin, Xing Chen, Chun-Chun Wang, Yan Zhao, and Ya-Zhou Sun. Prediction of small molecule–microrna associations by sparse learning and heterogeneous graph inference. *Molecular pharmaceutics*, 16(7):3157–3166, 2019.

[9] Shudong Wang, Tiyao Liu, Chuanru Ren, Wenhao Wu, Zhiyuan Zhao, Shanchen Pang, and Yuanyuan Zhang. Predicting potential small molecule–mirna associations utilizing truncated schatten p-norm. *Briefings in Bioinformatics*, 24(4):bbad234, 2023.

[10] Yiran Huang, Yongjin Bin, Pingfan Zeng, Wei Lan, and Cheng Zhong. Netpro: neighborhood interaction-based drug repositioning via label propagation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(3):2159–2169, 2023.

[11] Yuehui Wang, Maozu Guo, Yazhou Ren, Lianyin Jia, and Guoxian Yu. Drug repositioning based on individual bi-random walks on a heterogeneous network. *BMC bioinformatics*, 20:1–13, 2019.

[12] Mengyun Yang, Lan Huang, Yunpei Xu, Chengqian Lu, and Jianxin Wang. Heterogeneous graph inference with matrix completion for computational drug repositioning. *Bioinformatics*, 36(22-23):5456–5464, 2020.

[13] Shudong Wang, Chuanru Ren, Yulin Zhang, Shanchen Pang, Sibo Qiao, Wenhao Wu, and Boyang Lin. Amcsmma: Predicting small molecule–mirna potential associations based on accurate matrix completion. *Cells*, 12(8):1123, 2023.

[14] Xing Chen, Jun Yin, Jia Qu, and Li Huang. Mdhgi: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS computational biology*, 14(8):e1006418, 2018.

[15] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Benjamin L King, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45(D1):D972–D978, 2017.

[16] Geng Chen, Ziyun Wang, Dongqing Wang, Chengxiang Qiu, Mingxi Liu, Xing Chen, Qipeng Zhang, Guiying Yan, and Qinghua Cui. Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic acids research*, 41(D1):D983–D986, 2012.

[17] Wei Ma, Lu Zhang, Pan Zeng, Chuanbo Huang, Jianwei Li, Bin Geng, Jichun Yang, Wei Kong, Xuezhong Zhou, and Qinghua Cui. An analysis of human microbe–disease associations. *Briefings in bioinformatics*, 18(1):85–97, 2017.

[18] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

[19] Feilong Cao, Jiaying Chen, Hailiang Ye, Jianwei Zhao, and Zhenghua Zhou. Recovering low-rank and sparse matrix based on the truncated nuclear norm. *Neural Networks*, 85:10–20, 2017.

[20] Chen Xu, Zhouchen Lin, and Hongbin Zha. A unified convex surrogate for the schatten-p norm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[21] Beijia Chen, Huaijiang Sun, Guiyu Xia, Lei Feng, and Bin Li. Human motion recovery utilizing truncated schatten p-norm and kinematic constraints. *Information Sciences*, 450:89–108, 2018.

[22] Liangsheng He, Hao Wu, and Xiaotao Wen. Seismic random noise simultaneous attenuation in the time–frequency domain using lp-variation and γ norm constraint. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2022.

[23] Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. *Advances in neural information processing systems*, 28, 2015.

[24] Mengyun Yang, Huimin Luo, Yaohang Li, and Jianxin Wang. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics*, 35(14):i455–i463, 2019.

[25] Shuhan Huang, Minhui Wang, Xiao Zheng, Jiajia Chen, and Chang Tang. Hierarchical and dynamic graph attention network for drug-disease association prediction. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[26] Jin Li, Sai Zhang, Tao Liu, Chenxi Ning, Zhuoxuan Zhang, and Wei Zhou. Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction. *Bioinformatics*, 36(8):2538–2546, 2020.

[27] Zhouxin Yu, Feng Huang, Xiaohan Zhao, Wenjie Xiao, and Wen Zhang. Predicting drug–disease associations through layer attention graph convolutional network. *Briefings in bioinformatics*, 22(4):bbaa243, 2021.

[28] Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019.

[29] Wen Zhang, Xiang Yue, Yanlin Chen, Weiran Lin, Bolin Li, Feng Liu, and Xiaohong Li. Predicting drug-disease associations based on the known association bipartite network. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 503–509. IEEE, 2017.

[30] Huimin Luo, Min Li, Shaokai Wang, Quan Liu, Yaohang Li, and Jianxin Wang. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, 34(11):1904–1912, 2018.

[31] Lihong Peng, Liangliang Huang, Qiongli Su, Geng Tian, Min Chen, and Guosheng Han. Lda-vghb: identifying potential lncrna–disease associations with singular value decomposition, variational graph auto-encoder and heterogeneous newton boosting machine. *Briefings in Bioinformatics*, 25(1):bbad466, 2024.

[32] Rong Zhu, Yong Wang, Jin-Xing Liu, and Ling-Yun Dai. Ipcarf: improving lncrna-disease association prediction using incremental principal component analysis feature selection and a random forest classifier. *BMC bioinformatics*, 22:1–17, 2021.

[33] Zhen-Hao Guo, Zhu-Hong You, Yan-Bin Wang, Hai-Cheng Yi, and Zhan-Heng Chen. A learning-based method for lncrna-disease association identification combing similarity information and rotation forest. *IScience*, 19:786–795, 2019.

[34] Yuan Zhang, Fei Ye, Dapeng Xiong, and Xieping Gao. Ldnfsgb: prediction of long non-coding rna and disease association using network feature similarity and gradient boosting. *BMC bioinformatics*, 21:1–27, 2020.

[35] Min Zeng, Chengqian Lu, Fuhao Zhang, Yiming Li, Fang-Xiang Wu, Yaohang Li, and Min Li. Sdlda: lncrna-disease association prediction based on singular value decomposition and deep learning. *Methods*, 179:73–80, 2020.

[36] Yahui Long, Jiawei Luo, Yu Zhang, and Yan Xia. Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Briefings in bioinformatics*, 22(3):bbaa146, 2021.

[37] Meifang Hua, Shengpeng Yu, Tianyu Liu, Xue Yang, and Hong Wang. Mvgcnmda: multi-view graph augmentation convolutional network for uncovering disease-related microbes. *Interdisciplinary Sciences: Computational Life Sciences*, 14(3):669–682, 2022.

[38] Cheng Yan, Guihua Duan, Fang-Xiang Wu, Yi Pan, and Jianxin Wang. Brwmda: Predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(5):1595–1604, 2019.

[39] Lihong Peng, Ling Shen, Longjie Liao, Guangyi Liu, and Liqian Zhou. Rnmfmda: a microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Frontiers in microbiology*, 11:592430, 2020.

[40] Cheng Yan, Guihua Duan, Fang-Xiang Wu, Yi Pan, and Jianxin Wang. Mchmda: Predicting microbe-disease associations based on similarities and low-rank matrix completion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(2):611–620, 2019.

[41] Alireza Atri. The alzheimers disease clinical spectrum: diagnosis and management. *Medical Clinics*, 103(2):263–293, 2019.

[42] Haroon Khan, Hammad Ullah, Michael Aschner, Wai San Cheang, and Esra Küpeli Akkol. Neuroprotective effects of quercetin in alzheimers disease. *Biomolecules*, 10(1):59, 2019.

[43] Dianna J Magliano, Julian W Sacre, Jessica L Harding, Edward W Gregg, Paul Z Zimmet, and Jonathan E Shaw. Young-onset type 2 diabetes mellitusimplications for morbidity and mortality. *Nature Reviews Endocrinology*, 16(6):321–331, 2020.

[44] Daniela Pechlivanova, Ekaterina Krumova, Nedelina Kostadinova, Jeny Mitreva-Staleva, Petar Grozdanov, and Alexander Stoynev. Protective effects of losartan on some type 2 diabetes mellitus-induced complications in wistar and spontaneously hypertensive rats. *Metabolic Brain Disease*, 35(3):527–538, 2020.