Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

multiGMF: A multi-similarity geometric matrix factorization for identifying drug-associated indications

Mengyun Yang

mengyun_yang@126.com

Hunan First Normal University

Bin Yang

Guangdong University of Technology

Xiwei Tang

Hunan First Normal University

Guihua Duan

Central South University

Article

Keywords:

Posted Date: April 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9409242/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on August 24th, 2026.
See the published version at https://doi.org/10.1038/s41598-026-66963-7.

multiGMF: A multi-similarity geometric matrix factorization for identifying drug-associated indications

Mengyun Yang[1,+,*], **Bin Yang**[2,+], **Xiwei Tang**[1], **and Guihua Duan**[3]

[1]Hunan Provincial University Key Laboratory for Big Data Analysis and Application, Hunan First Normal University, Changsha, 410083, China
[2]School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou, 510006, China
[3]School of Computer Science and Engineering, Central South University, Changsha, 410000, China
[*]mengyun_yang@126.com
[+]these authors contributed equally to this work

ABSTRACT

Drug repositioning serves as a promising strategy in drug development, exploring the potential uses of existing drugs through numerical computation. Compared to experimental screening, drug repositioning proves to be more efficient and cost-effective, playing a vital role in the field of pharmaceutical development. Designing an effective approach to integrate multi-source prior information about drugs and diseases holds significance in drug repositioning, given the low coupling of latent features in existing methods for handling the associated information and multi-similarity information of drugs and diseases. In this article, we propose a novel method based on multi-similarity geometric matrix factorization (multiGMF) for identifying the potential indications of existing and new drugs. Through weighted k-nearest neighbors (WKNN) algorithm and soft regularization technique, it couples the multi-similarity features of drugs and diseases with associated features. Moreover, it explores their latent feature information in high-dimensional space using graph regularization technique aimed at inferring potential drug-disease associations. To evaluate the performance of multiGMF, we contrast it with five most advanced drug repositioning approaches in both 10-fold cross-validation and cold-start tests. The numerical outcomes demonstrate that multiGMF exhibits outstanding predictive performance. Furthermore, case studies further support the viability of our method in practical applications. The multiGMF code is freely available at https://github.com/YangPhD84/multiGMF.

Introduction

The process of developing new drugs is intricately complex, requiring both time and resources while facing significant risks[1]. According to statistical data, approximately 90% of candidate drugs encounter failure during early clinical trials, and the process of developing a drug and bringing it to market requires a significant investment of billions of dollars and spans a period of 10 to 15 years[2]. To facilitate drug development, the concept of "drug repositioning" has been introduced, aiming to explore new purposes for already approved drugs[3]. Utilizing approved drugs already on the market can offer valuable safety data for experimental purposes, mitigating risks and enhancing overall efficiency[4].

Based on the hypothesis of using similar molecular pathways shared by similar drugs to treat related diseases, finding new indications for commercial drugs through computational methods holds an indispensable role in drug repositioning[5]. The continuous advancement of high-throughput technologies has led to the emergence of a plethora of multi-omic data in the field of bioinformatics. The availability of such data has significantly catalyzed the introduction and advancement of various computational methods, providing robust support for drug repositioning studies.

Network-based approach involves the transformation of drugs and diseases into nodes within a biological network, where interactions are represented by edges. Extracting topological details from this constructed biological network allows the prediction of unknown drug-disease associations. Its foundational principle is rooted in the "guilt-by-association" concept, indicating that drugs with similar mechanisms of action in the biological network are likely to share therapeutic effects on the same diseases[6]. Wang et al.[7] developed a diverse graph incorporating drugs and targets, employing existing data on drugs and targets for heterogeneous graph-based inference, namely HGBI, to identify novel drug-target interactions. Zhou et al.[8] introduced a drug repositioning method using the multi-network concept, namely DRMN. This model utilized diverse biological data to create a multiplex network and the PageRank algorithm calculated drug-disease association scores.

Machine learning-based methods are commonly employed in predicting drug-disease associations, utilizing supervised

learning algorithms such as logistic regression[9], random forests[10], and naive bayes[11]. These algorithms employ known associations as training samples to explore the patterns of associations between drugs and indications. They view the prediction of unknown associations as a binary classification challenge. Jiang and Huang[12] introduced a graph-based learning model anchored in random forests, utilizing composite features from multi-level biological molecular networks to uncover undisclosed drug-disease connections. Thafar et al.[13] presented a method named Affinity2Vec, which framed the prediction task as a graph-based regression problem. Through the integration of various data sources and the combination of techniques from feature representation, graph mining, and machine learning, potential associations between drugs and targets were predicted.

Deep learning methods are widely applied in drug repositioning studies, commonly utilizing neural networks to extract high-dimensional features of drugs and diseases. These methods employ neural network models to learn intricate nonlinear relationships between drugs and indications, facilitating the prediction of potential associations. Ghasemi and Lakizadeh[14] proposed an end-to-end heterogeneous graph convolutional encoder (LHGCE) model, which extracts features for drugs and diseases through the encoder and uses an aggregation function to achieve association prediction through feature embedding. To address the issue of sparse associations, Wang et al.[15] introduced a model based on the multi-similarity graph convolutional autoencoder (MSGCA). This model integrates multi-kernel learning and weighted k-nearest neighbor algorithms to consolidate the multi-similarities and association data of drugs and indications. It combines a graph convolutional autoencoder with an attention mechanism to infer new drug-disease associations.

However, the strategies mentioned above heavily rely on feature extraction and the selection of negative samples, leading to the emergence of more intricate and efficient techniques[16]. Among them, matrix factorization and matrix completion methods stand as relatively novel and popular approaches. These methods use low-rank matrices to capture latent feature representations of drug-disease associations, thereby predicting unknown drug-disease associations by contrasting prior information about drugs and diseases. In response to these challenges, We had proposed the bounded nuclear norm regularization approach (BNNR)[17], aimed at minimizing the rank of the drug-indication relevance matrix through the incorporation of a regularization term. This effectively balances prediction error and rank properties. In addition, we have developed the multi-similarity bilinear matrix factorization (MSBMF) approach[18]. It decomposed the association matrix into two low-rank matrices and incorporated multiple similarity features of drugs and indications dynamically to predict their potential associations. However, It simply splices the similarity matrices of drugs and indications and do not take into account the intrinsic geometric structure of their feature spaces.

In this study, we present a novel approach known as multi-similarity geometric matrix factorization (multiGMF). This method effectively integrates spatial geometric structural information shared in the varied similarities of drugs and diseases, providing comprehensive features for both. Consequently, it captures the inherent characteristics of drugs and indications, facilitating the prediction of associations between them. Acknowledging the scarcity of established drug-indication associations, we employ WKNN preprocessing to enrich the association matrix with relevant predictive values by leveraging multi-similarity information. Subsequently, the updated association matrix is subjected to matrix factorization to obtain the association features of drugs and diseases. Not only utilize graph regularization techniques to explore the latent feature information of drugs and indications but also employ soft regularization techniques to achieve linear fusion of the association feature matrix and the multi-similarity matrix. This process generates low-rank composite feature matrices for drugs and diseases. Finally, by multiplying these two matrices, the association matrix can be obtained. Figure 1 illustrates the overall process of our approach. The primary contributions of the multiGMF model include:

- multiGMF provides an effective solution by efficiently integrating multi-source prior information of drugs and diseases.
- multiGMF effectively extracts latent geometric features of drugs and diseases in high-dimensional space through the utilization of the WKNN algorithm and graph regularization.
- In the multiGMF model, features of drugs and diseases come from both the association matrix and the multi-similarity matrix. We used soft regularization to better couple these features.
- An efficient iterative algorithm is designed to achieve the convergence of the multiGMF model.

Materials

In this study, we validated our proposed method using the established gold standard dataset (Fdataset)[19] for drug repositioning. This dataset comprises 1,933 extensively confirmed associations between drugs and diseases. These associations involve 593 drugs and 313 diseases. The data for these associations are sourced from well-established databases, namely DrugBank[20] and the Online Mendelian Inheritance in Man (OMIM)[21]. To represent the drug-disease associations, we constructed a binary matrix denoted as A. Within this matrix, established drug-indication relationships are expressed by 1, while uncertain associations are depicted by 0.

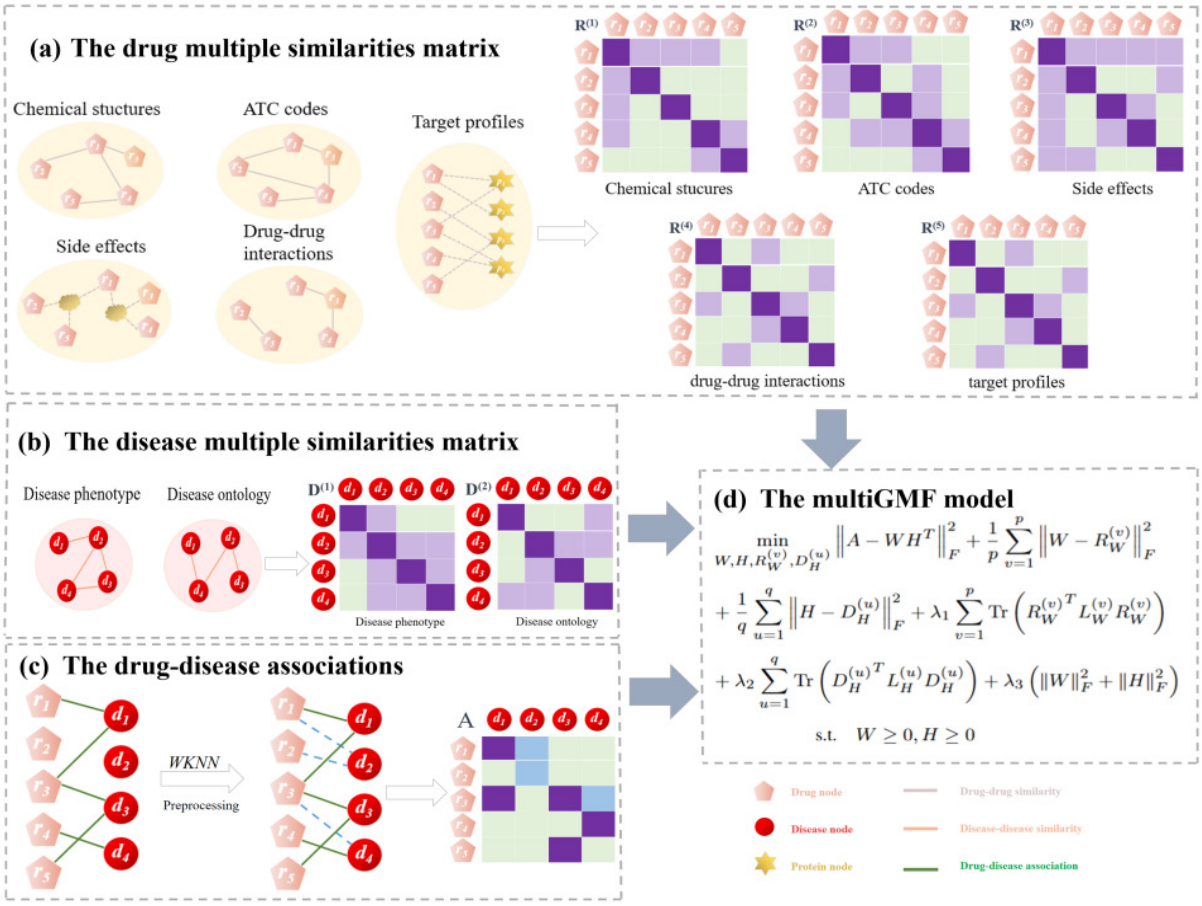


Figure 1. The overall workflow of multiGNMF. (a) The process of constructing the drug multi-similarity matrix. (b) The process of constructing the disease multi-similarity matrix. (c) Drug–disease associations and WKNN preprocessing. (d) The multiGMF model.

Here, we adopted a comprehensive evaluation by employing five distinct similarity measures to assess the connections between drugs:

- Chemical structure similarity $R^{(1)}$: The Chemical Development Toolkit (CDK)[22] was utilized to gauge chemical structure similarities. This involved computing hash fingerprints using the Canonical SMILES[23] of all drugs.
- ATC code similarity $R^{(2)}$: The DrugBank database was utilized to retrieve ATC codes, and a semantic similarity algorithm[24] was applied to determine the resemblances among ATC codes.
- Side effect similarity $R^{(3)}$: Side effect data was drawn from the SIDER database[25], and the likeness of side effects was assessed by computing Jaccard similarity coefficients[26].
- Drug-drug interactions similarity $R^{(4)}$: Drug interaction details were acquired from DrugBank, enabling us to evaluate drug interaction similarities using the Jaccard score within the drug–drug interaction network.
- Target profiles similarity $R^{(5)}$: Similarly, by employing the Jaccard similarity coefficient, we determined the similarities in drug-target information sourced from DrugBank, resulting in target profile similarities.

For diseases, we obtained two similarity measures based on disease phenotypes and ontologies:

- Disease phenotype similarity $D^{(1)}$: The similarity scores for indication phenotypes were obtained from the MimMiner database[27]. Text mining was employed on MeSH terms[28] within OMIM to gather disease phenotype vectors from this database, and their similarity was determined by the cosines of the angles.
- DO similarity $D^{(2)}$: Considering the composition of terms in the disease ontology, we applied the gene ontology-based algorithm[29] to calculate similarities within the disease ontology.

Methods

This section describes our novel drug repositioning method, structured as follows. First, we present non-negative matrix factorization (NMF) and graph regularised non-negative matrix factorization (GrNMF). Then, we propose a drug-disease association prediction model based on multi-similarity geometric matrix factorization (multiGMF). Finally, the alternating least squares (ALS) is demonstrated in detail to solve the multiGMF model.

Non-negative matrix factorization (NMF)

As we all know, the NMF is a highly effective and efficient approach. It has been widely employed in research such as image classification[30], signal processing[31], and recommender systems[32]. Its goal is to conduct two non-negative matrices that optimally approximate the original matrix through multiplication. The mathematical formula of NMF is generally articulated as:

$$\min_{W,H} \left\|A - WH^T\right\|_F^2 \quad \text{s.t.} \quad W \geq 0, H \geq 0 \tag{1}$$

where $\|\cdot\|_F$ represents the Frobenius norm. The original matrix $A \in \mathbb{R}^{m\times n}$ can be decomposed by NMF into two low-dimensional non-negative,$W \in \mathbb{R}^{m\times l}$ and $H \in \mathbb{R}^{n\times l}$ $(l << \min(m,n))$, such that $A \approx WH^T$.

Graph regularized non-negative matrix factorization (GrNMF)

To address the limitation of NMF in neglecting the manifold structure of the data, Graph Regularized Nonnegative Matrix Factorization (GrNMF) was proposed by Cai et al.[33] This is widely used in areas such as facial recognition[34], feature extraction[35], and anomaly detection[36].

The expression of the graph regularization is

$$\begin{aligned} &\min_H \frac{1}{2}\sum_{i,j=1}^{n} \left\|h_i - h_j\right\|^2 S_{ij} \\ &= \min_H \operatorname{Tr}\left(HUH^\top\right) - \operatorname{Tr}\left(HVH^\top\right) \\ &= \min_H \operatorname{Tr}\left(HLH^\top\right), \end{aligned} \tag{2}$$

where h_i denotes the feature vector of the i-th node, and S is a similarity matrix describing the degree of similarity between data points, e.g. S_{ij} represents the similarity of nodes i and j. $L = U - V$ denotes the graph Laplacian matrix, $V = \left(S + S^T\right)/2$ and U is a diagonal matrix with $u_{jj} = \sum_{i=1}^{n} v_{ij}$. As the similarity degree S_{ij} between node i and node j increases, they become closer in high-dimensional space, i.e., the smaller $\left\|h_i - h_j\right\|^2$. Eq. (2) ensures that the latent feature vectors of nodes with higher similarity in high-dimensional space are closer, enabling the capture of the latent geometric features of nodes i and j in high-dimensional space. The GrNMF model is formulated as follows:

$$\begin{aligned} &\min_{W,H} \left\|A - WH^T\right\|_F^2 + \lambda \operatorname{Tr}\left(HLH^\top\right) \\ &\quad \text{s.t.} \quad W \geq 0, H \geq 0, \end{aligned} \tag{3}$$

where the regularisation parameter $\lambda \geq 0$ controls the smoothness. Through graph regularization, it's possible to effectively retain the intrinsic high-dimensional structure of low-dimensional sampled data[37]. The model can better learn structural features and maintain the local continuity of the data.

multiGMF for identifying drug-associated indications

Weighted k-nearest neighbor preprocessing. We denote the set of drug nodes as $\{r_1, r_2, \ldots, r_m\}$ and disease nodes as $\{d_1, d_2, \ldots, d_n\}$. The original drug–disease association matrix is represented by $A \in \mathbb{R}^{m\times n}$, where $A_{ij} = 1$ denotes confirmed connection between drug r_i and disease d_j, and $A_{ij} = 0$ otherwise. The association properties of drug r_i are defined by the i-th row vector of matrix A, labeled as $A(r_i) = (A_{i1}, A_{i2}, \ldots, A_{im})$. Likewise, the properties of disease d_j is represented by the j-th column vector of matrix A, as $A(d_j) = \left(A_{1j}, A_{2j}, \ldots, A_{nj}\right)$. In practical applications, the drug-disease association matrix is extremely sparse, and it suffers from a limited number of training samples. Additionally, with the introduction of new drug or disease nodes, the associated row or column values are set to zero in the adjacency matrix. These situations can potentially impact prediction performance, leading to less satisfactory results[38]. To address this issue, we introduce the Weighted k-Nearest

Neighbors (WKNN) algorithm to supplement additional association values to the original matrix from the perspectives of both drugs and diseases.

For each drug, represented as r_m, we revise the profile of drug associations by incorporating the similarity it shares with the k-nearest verified drugs (where the presence of no less than one proven association):

$$A_r(r_m) = \frac{1}{Q_r}\sum_{i=1}^{k} w_i A(r_i), \tag{4}$$

Where r_1 to r_k represent drugs arranged in descending order according to their similarity. The weight coefficient $w_i = \alpha^{i-1} * \bar{R}(r_i, r_m)$, indicating that if r_i shows more resemblance to r_m then a higher weight is given. The decay terms $\alpha \in [0,1]$, and $\bar{R}$ represent the average matrix derived from all drug similarity. The $Q_r = \sum_{1 \le i \le k} \bar{R}(r_i, r_m)$ is the normalization term. In the same way, the updated association profile for each disease d_n is obtained as follows, using the described procedure:

$$A_d(d_n) = \frac{1}{Q_d}\sum_{j=1}^{k} w_j A(d_j), \tag{5}$$

where d_1 to d_k represent diseases arranged in descending order according to their similarities. $\bar{D}$ represents the average matrix derived from all disease similarity. The normalization term $Q_d = \sum_{1 \le j \le k} \bar{D}(d_j, d_n)$ and the weight coefficient $w_j = \alpha^{j-1} * \bar{D}(d_j, d_n)$ are shown. Following this, we generate the previously mentioned matrices A_r and A_d for the drug and disease, respectively. Through the computation of the likelihood scores and substitution of instances where $A_{ij} = 0$, the adjacency matrix A is subsequently modified in the manner:

$$A = \max(A, A_{rd}), \tag{6}$$

where $A_{rd} = (A_r + A_d)/2$. The update matrix A has a higher density after WKNN preprocessing.

The multiGMF model. To effectively manage data relationships and minimize redundancy, facilitating the coupling of association information and similarity information latent features between drugs and diseases, we propose a drug repositioning model based on multi-similarity geometric matrix factorization (multiGMF). By incorporating both soft regularization[39] and graph Laplacian regularization terms, we achieve a reliable integration of similarity data. Additionally, we use the Tikhonov regularization to prevent overfitting. The multiGMF model is expressed as follows:

$$\begin{aligned} &\min_{W,H,R_W^{(v)},D_H^{(u)}} \left\|A - WH^T\right\|_F^2 + \frac{1}{p}\sum_{v=1}^{p}\left\|W - R_W^{(v)}\right\|_F^2 \\ &+ \frac{1}{q}\sum_{u=1}^{q}\left\|H - D_H^{(u)}\right\|_F^2 + \lambda_1 \sum_{v=1}^{p} \mathrm{Tr}\left(R_W^{(v)^T} L_W^{(v)} R_W^{(v)}\right) \\ &+ \lambda_2 \sum_{u=1}^{q} \mathrm{Tr}\left(D_H^{(u)^T} L_H^{(u)} D_H^{(u)}\right) + \lambda_3\left(\|W\|_F^2 + \|H\|_F^2\right) \\ &\qquad \text{s.t.} \quad W \ge 0, H \ge 0 \end{aligned} \tag{7}$$

where $A \in \mathbb{R}^{m \times n}$ signifies the updated drug–disease associations matrix after WKNN. Tr $(\cdot)$ denotes the trace of a matrix. $R^{(v)} \in \mathbb{R}^{m \times m}$ represents the corresponding drug similarity, whlie $D^{(u)} \in \mathbb{R}^{n \times n}$ is the matching disease similarity. The Laplacian matrix of $R^{(v)}$ is denoted by $L_W^{(v)} = U^{(v)} - R^{(v)}$. The $U^{(v)}$ is a diagonal matrix with the i-th diagonal entry defined as $\sum_{j=1}^{n}\left(R_{ji}^{(v)} + R_{ij}^{(v)}\right)/2$. Likewise, the Laplacian matrix of $D^{(u)}$ is denoted by $L_H^{(u)}$. In Eq. (7), by introducing graph regularization terms, we can capture higher-order geometric features of each node from the high-dimensional space of drugs. Building upon this, we tightly couple the geometric features with the associated features of drugs using soft regularization term, thereby constructing a more comprehensive low-rank feature matrix of drugs (i.e., W). Similarly, a low-rank composite similarity matrix H can be constructed for diseases. Furthermore, we are introducing λ_1, λ_2, and λ_3 as smoothing parameters to prevent overfitting.

Solutions for multiGMF

In this section, an iterative algorithm is devised to achieve efficient convergence of the multiGMF model. The Lagrange multipliers $\Psi = [\psi_{jk}]$ and $\Phi = [\varphi_{ik}]$ correspond to the constraints $h_{jk} \ge 0$ and $w_{ik} \ge 0$, respectively. The Lagrange function L_f

of Eq. (7) is described as:

$$
\begin{aligned}
L_f &= \mathrm{Tr}\left(AA^T\right) - 2\,\mathrm{Tr}\left(A^T W H^T\right) + \mathrm{Tr}\left(HW^T W H^T\right) \\
&+ \frac{1}{p}\sum_{v=1}^{p}\left(\mathrm{Tr}\left(WW^T\right) - 2\,\mathrm{Tr}\left(W^T R_W^{(v)}\right) + \mathrm{Tr}\left({R_W^{(v)}}^T R_W^{(v)}\right)\right) \\
&+ \frac{1}{q}\sum_{u=1}^{q}\left(\mathrm{Tr}\left(HH^T\right) - 2\,\mathrm{Tr}\left(H^T D_H^{(u)}\right) + \mathrm{Tr}\left({D_H^{(u)}}^T D_H^{(u)}\right)\right) \\
&+ \lambda_1 \sum_{v=1}^{p}\mathrm{Tr}\left({R_W^{(v)}}^T L_W^{(v)} R_W^{(v)}\right) + \lambda_2 \sum_{u=1}^{q}\mathrm{Tr}\left({D_H^{(u)}}^T L_H^{(u)} D_H^{(u)}\right) \\
&+ \lambda_3\left(\mathrm{Tr}\left(W^T W\right) + \mathrm{Tr}\left(H^T H\right)\right) + \mathrm{Tr}\left(\Phi W^T\right) + \mathrm{Tr}\left(\Psi H^T\right)
\end{aligned}
\tag{8}
$$

The function's partial derivatives with respect to $R_W^{(v)}$, $D_H^{(u)}$, W and H are:

$$\frac{\partial L_f}{\partial R_W^{(v)}} = \frac{1}{p}\left(-2W + 2R_W^{(v)}\right) + \lambda_1\left(L_W^{(v)} + {L_W^{(v)}}^T\right)R_W^{(v)}, \tag{9a}$$

$$\frac{\partial L_f}{\partial D_H^{(u)}} = \frac{1}{q}\left(-2H + 2D_H^{(u)}\right) + \lambda_2\left(L_H^{(u)} + {L_H^{(u)}}^T\right)D_H^{(u)}, \tag{9b}$$

$$\frac{\partial L_f}{\partial W} = -2AH + 2WH^T H + \frac{1}{p}\sum_{v=1}^{p}\left(2W - 2R_W^{(v)}\right) + 2\lambda_3 W + \Phi, \tag{9c}$$

$$\frac{\partial L_f}{\partial H} = -2A^T W + 2HW^T W + \frac{1}{q}\sum_{u=1}^{q}\left(2H - 2D_H^{(u)}\right) + 2\lambda_3 H + \Psi. \tag{9d}$$

By equating the partial derivatives obtained from Eq. (9a) and Eq. (9b) to 0, we can arrive at the following equations for $R_W^{(v)}$ and $D_H^{(u)}$:

$$R_W^{(v)} = \left(\frac{1}{p}E + \lambda_1 L_W^{(v)}\right)^{-1}\left(\frac{1}{p}W\right), \tag{10a}$$

$$D_H^{(u)} = \left(\frac{1}{q}E + \lambda_2 L_H^{(u)}\right)^{-1}\left(\frac{1}{q}H\right), \tag{10b}$$

The constraints $\varphi_{ik} w_{ik} = 0$ and $\psi_{jk} h_{jk} = 0$ of Karush-Kuhn-Tucker (KKT)[40] are conducted. Through the Eq. (9c) and Eq. (9d), the equations are obtained for w_{ik} and h_{jk} :

$$-\left(AH + \frac{1}{p}\sum_{v=1}^{p}R_W^{(v)}\right)_{ik} w_{ik} + \left(WH^T H + W + \lambda_3 W\right)_{ik} w_{ik} = 0, \tag{11a}$$

$$-\left(A^T W + \frac{1}{q}\sum_{u=1}^{q}D_H^{(u)}\right)_{jk} h_{jk} + \left(HW^T W + H + \lambda_3 H\right)_{jk} h_{jk} = 0. \tag{11b}$$

Consequently, the following updating rules are determined:

$$w_{ik} \leftarrow w_{ik}\frac{\left(AH + \frac{1}{p}\sum_{v=1}^{p}R_W^{(v)}\right)_{ik}}{\left(WH^T H + W + \lambda_3 W\right)_{ik}}, \tag{12a}$$

$$h_{jk} \leftarrow h_{jk}\frac{\left(A^T W + \frac{1}{q}\sum_{u=1}^{q}D_H^{(u)}\right)_{jk}}{\left(HW^T W + H + \lambda_3 H\right)_{jk}}. \tag{12b}$$

Updating the W and H iteratively up to convergence using Eq. (12a) and (12b). Finally, we obtain the projected drug-disease association matrix as $A^* = WH^T$. Based on the association scores in matrix A^*, we can infer unknown drug-disease associations.

The algorithm 1 gives the overall iterative scheme of the multiGMF model.

1 Results

Evaluation metrics

To comprehensively evaluate the performance of multiGMF, we conducted two computational experiments, including 10-fold cross-validation and cold-start tests. The existing drug-disease associations are randomly divided into 10 subsets for cross-validation, with one serving as the test set and the rest as training sets. This process is repeated ten times, and averages are

Algorithm 1 Solution for multiGMF model.

Input: The preprocessed association matrix $A \in \mathbb{R}^{m\times n}$, the five drug similarity matrices: $R^{(v)} \in \mathbb{R}^{m\times m}$ (i.e., $R^{(1)}, R^{(2)}, R^{(3)}, R^{(4)}, R^{(5)}$), the two disease similarity matrices: $D^{(u)} \in \mathbb{R}^{n\times n}$ (i.e., $D^{(1)}, D^{(2)}$), regularization coefficients λ_1, λ_2, and λ_3.

Output: Final association matrix A^*.

1. randomly initialize two matrices with non-negative elements W $\in \mathbb{R}^{m\times l}$ and H $\in \mathbb{R}^{n\times l}$;
2. construct matrices $L_W^{(v)}$, $L_H^{(u)}$ from $R^{(v)}$, $D^{(v)}$;
3. **repeat**

Update $R_W^{(v)}$ by Eq. (10a).

Update $D_H^{(u)}$ by Eq. (10b).

Update W by Eq. (12a).

Update H by Eq. (12b).

4. **until convergence**
5. $A^* = WH^T$;
6. **return** A^*.

Table 1. The summation of AUC and AUPR values with varying λ_1 and τ parameters in the 10-fold cross-validation.

$\lambda_1 \setminus \tau$	0.1	0.3	0.5	0.7	0.9
0.1	1.268	1.356	1.366	1.366	1.372
0.01	1.271	1.363	1.371	1.378	1.381
0.001	1.269	1.358	1.377	1.387	1.389
0.0001	1.269	1.357	1.381	**1.391**	1.391
0.00001	1.268	1.268	1.357	1.379	1.392

computed to determine the final results. For the cold-start test, drugs with only one known disease association are considered target drugs. Each known drug-disease association linked to a target drug is systematically removed to assess the algorithm's performance in cold-start scenarios. $R^{(3)}$ and $R^{(4)}$ are not included in the cold-start test due to the potential lack of experimental data for novel drugs. Predicted scores for unknown drug-disease pairs are used to rank potential indications for the test drugs in descending order. To assess the superiority of our model, three evaluation metrics, namely, AUC, AUPR, and precision, were set up for consideration.

Parameters setting

Within the framework of the multiGMF model, certain preset parameters coexist alongside three pivotal parameters that require adjustment. The selection of these default parameters is grounded in empirical exploration. In this study, we set $p = 5$ and $q = 2$, and keep the decay term α at 0.9 in the WKNN step. We designate the dimension of the low-rank matrices W and H in Eq. (7) as $l = [\tau \min(m,n)]$, where $\tau \in [0,1]$ and $[\cdot]$ denotes the rounding function. To mitigate overfitting resulting from an excessive number of parameters, set λ_1 and λ_2 to the same value and keep $\lambda_3 = 1$. Ultimately, three parameters require determination: neighborhood sizes k, the smooth parameter λ_1, and the low-rank parameter τ.

To optimize these three crucial parameters, we employ a "fix one parameter, determine others" strategy. We initially put k to 10 and subsequently proceeded with the selection λ_1 from {0.1, 0.01, 0.001, 0.0001, 0.00001} and τ from {0.1, 0.3, 0.5, 0.7, 0.9} through cross-validation. The computed values for coefficients λ_1 and τ are presented in Table 1. After conducting the grid search, we found that when $\lambda_1 \leq 0.0001$, the combined values of AUC and AUPR reach a stable state. Moreover, consistently achieving favorable performance when $\tau \geq 0.7$ implies that multiGMF can accurately represent the association matrix through the low-rank matrix. Following this, we fixed $\tau = 0.7$ and $\lambda_1 = 0.0001$, and selected values of k from {1, 5, 10, 15, 20} to evaluate the performance of multiGMF under different k values. As depicted in Table 2, the summation of AUC and AUPR values attains its maximum at $k = 10$.

In the meantime, the multiGMF algorithm concludes either upon meeting the prescribed stopping criteria or reaching the maximum iteration steps:

$$f_k \leq tol1, \quad \frac{|f_{k+1} - f_k|}{max\{1, |f_k|\}} \leq tol2, \tag{13}$$

Table 2. Calculating the total AUC and AUPR values at various k values while keep $\tau = 0.7$ and $\lambda_1 = 0.0001$.

k	1	5	10	15	20
AUC+AUPR	1.231	1.367	**1.391**	1.385	1.381

where $f_k = \frac{\|A^*_{k+1} - A^*_k\|_F}{\|A^*_k\|_F}$, $tol1$ and $tol2$ are the given tolerances. Here, $tol1$ is configured as 2×10^{-3}, while $tol2$ is set to 10^{-4}.

Compared to multiple methods of drug repositioning

We contrasted multiGMF with five most advanced drug repositioning approaches: DDA-SKF[41], MSBMF[18], HGIMC[42], BNNR[17], and MBiRW[43]. To evaluate the effectiveness of these methods, the 10-fold cross-validation and cold-start tests are executed on the Fdataset. As shown in Table 3, multiGMF outperforms other methods across various metrics (i.e., AUC, AUPR, and precision) in the cross-validation, and it also attains the highest AUC value in the cold-start test. The optimum outcomes are highlighted in bold, while the next best are underscored. multiGMF demonstrates superior performance in the cross-validation compared to other methods, with AUC, AUPR, and precision advantages over the second-best MSBMF being 0.956%, 3.325%, and 3.736%, respectively. These results emphasize that by extracting the intrinsic geometric structure of drug and disease feature spaces, more accurate predictions of drug-association indications can be achieved. In the cold-start testing, multiGMF achieved its optimal AUC value of 0.909, surpassing the second-ranked HGIMC by 3.649%. Additionally, it obtained the second-highest AUPR and precision values, trailing only behind HGIMC. However, during the cross-validation, multiGMF's AUC, AUPR, and precision exceeded those of HGIMC by 0.956%, 10.406%, and 7.763%, respectively. These findings show that multiGMF continuously performs exceptionally well across a variety of tests. Figure 2(a) and 2(b) respectively display ROC curves for the 10-fold cross-validation and cold-start tests.

Table 3. The metrics for each method during 10-fold cross-validation and cold-start tests on the Fdataset.

Tests	Metrics	**multiGMF**	DDA-SKF	MSBMF	HGIMC	BNNR	MBiRW
10-fold CV	AUC	**0.950**	0.929	<u>0.941</u>	<u>0.941</u>	0.932	0.917
	AUPR	**0.435**	0.354	<u>0.421</u>	0.394	0.402	0.264
	Precision	**0.472**	0.397	<u>0.455</u>	0.438	0.440	0.304
cold-start	AUC	**0.909**	0.832	0.875	<u>0.877</u>	0.830	0.818
	AUPR	<u>0.340</u>	0.256	0.301	**0.356**	0.199	0.189
	Precision	<u>0.398</u>	0.310	0.368	**0.433**	0.251	0.234

After all, multiGMF proves to be a promising prediction method, efficiently integrating diverse similarity information between drugs and diseases, thus enhances the efficiency of drug-disease association prediction.

Ablation experiment

To demonstrate the effectiveness of multiGMF in integrating highly coupled multiple similarity features and relational features, we designed four ablation experiments: "multiGMF (w/o WKNN)", "multiGMF (w/o soft)", "singleGMF" and "singleGMF (w/o WKNN)". Specifically, "multiGMF (w/o WKNN)" indicates that multiGMF does not perform WKNN preprocessing on the association matrix when using multiple similarities. "multiGMF (w/o soft)" denotes multiGMF without incorporating soft regularization, relying solely on the WKNN algorithm to inject additional multi-similarity information into the association matrix to facilitate their coupling, as shown in Eq. (14).

$$\begin{aligned} &\min_{W,H} \|A - WH^T\|_F^2 + \lambda_1 \sum_{v=1}^{p} \mathrm{Tr}\left(W^T L_W^{(v)} W\right) \\ &+ \lambda_2 \sum_{u=1}^{q} \mathrm{Tr}\left(H^T L_H^{(u)} H\right) + \lambda_3 \left(\|W\|_F^2 + \|H\|_F^2\right) \\ &\qquad \text{s.t.} \quad W \geq 0, H \geq 0 \end{aligned} \tag{14}$$

Here, $L_W^{(v)} = U^{(v)} - R^{(v)}$ and $L_H^{(u)} = O^{(v)} - D^{(v)}$ are the Laplacian matrices corresponding to $R^{(v)}$ and $D^{(u)}$, respectively.

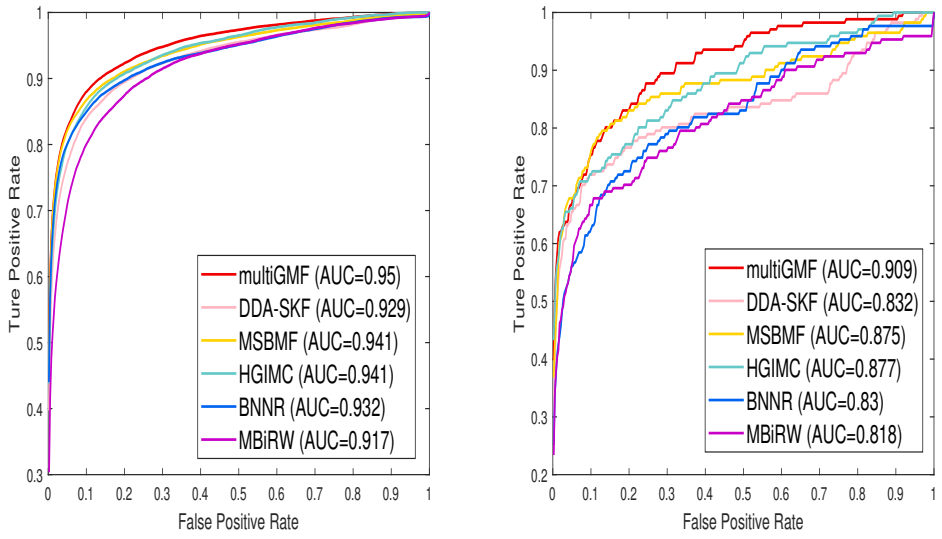


Figure 2. The results of each method are evaluated using the 10-fold cross-validation and cold-start tests on the Fdataset. (a) The 10-fold cross-validation ROC curves. (b) The cold-start test ROC curves.

According to the iterative solution described earlier, their iterative update process is given by Eq. (15a) and (15b):

$$w_{ik} \leftarrow w_{ik} \frac{\left(AH + \lambda_1 \sum_{v=1}^{p} R^{(v)} W\right)_{ik}}{\left(WH^T H + \lambda_1 \sum_{v=1}^{p} U^{(v)} W + \lambda_3 W\right)_{ik}}, \tag{15a}$$

$$h_{jk} \leftarrow h_{jk} \frac{\left(A^T W + \lambda_2 \sum_{u=1}^{q} D^{(u)} H\right)_{jk}}{\left(HW^T W + \lambda_2 \sum_{u=1}^{q} O^{(u)} H + \lambda_3 H\right)_{jk}}. \tag{15b}$$

The "singleGMF" involves predicting associations solely using $R^{(1)}$ and $D^{(1)}$. "singleGMF (w/o WKNN)" denotes using only $R^{(1)}$ and $D^{(1)}$ for prediction without utilizing WKNN optimization. Given that "singleGMF" and "singleGMF (w/o WKNN)" utilize single similarities for prediction, the parameters need suitable adjustments. Using the same way as before, we update the parameters to achieve optimal performance, where $\tau = 0.7$, $\lambda_1 = \lambda_2 = 0.0001$ and $\lambda_3 = 1$.

As shown in Table 4, multiGMF demonstrated the most superior performance across all ablation experiments, achieving the best results in terms of AUC, AUPR, and precision. This demonstrates that the model effectively integrates various heterogeneous information between drugs and diseases, resulting in a reliable composite similarity, thereby enhancing predictive outcomes. As depicted in Table 4, the comparison of AUC, AUPR, and precision values between the "multiGMF" and "singleGMF"

Table 4. The comparison of the results for multiGMF, multiGMF (w/o WKNN), singleGMF and singleGMF (w/o WKNN) is conducted through 10-fold cross-validation and cold-start tests on the Fdataset.

Test	10-fold CV			cold-start		
Mertricics	AUC	AUPR	Precision	AUC	AUPR	Precision
multiGMF	**0.950**	**0.435**	**0.472**	**0.909**	**0.340**	**0.398**
singleGMF	0.930	0.387	0.431	0.814	0.184	0.240
multiGMF (w/o soft)	0.950	0.432	0.469	0.890	0.300	0.357
multiGMF (w/o WKNN)	0.916	0.366	0.411	0.823	0.245	0.292
singleGMF (w/o WKNN)	0.908	0.349	0.399	0.760	0.100	0.140

experiments, as well as between the "multiGMF (w/o WKNN)" and "singleGMF (w/o WKNN)" experiments, shows that the use of multiple similarities effectively improves the prediction performance. The comparison between multiGMF and "multiGMF (w/o WKNN)" demonstrates the significant role of WKNN prefilling in the multiGMF model. According to the ablation experiment results, the WKNN algorithm not only pre-fills the original association matrix but also provides it with high-dimensional latent features. This facilitates the coupling between association features and similarity features, effectively enhancing prediction accuracy. Similarly, by comparing the results of multiGMF with those of "multiGMF (w/o soft)", it can be observed that incorporating soft regularization term in the model facilitates the high coupling of various heterogeneous information.

Case studies

In order to present the practical utility of multiGMF, we carried out a sequence of case studies to uncover fresh applications for established medications. We harnessed the entirety of accessible drug-disease associations and multiple similarity metrics present in the standardized dataset. This information was then applied through the multiGMF algorithm, enabling the anticipation of previously uncharted relationships between drugs and indications. By extrapolating from the results yielded by multiGMF, we compiled potential indications for each drug and ranked them according to scores in a descending manner. For an extended period, cancer and leukemia have remained pivotal diseases of societal concern, with a widespread focus on the progression of drug development for treating these ailments. Within this study, we selected five common anticancer drugs (i.e., fluorouracil, methotrexate, doxorubicin, cisplatin, and vincristine) to conduct a practical analysis of our approach. We conducted an exhaustive search within the CTD database[44] to amass substantiating evidence pertaining to the suggested indications for these selected drugs.

In Table 5, the multiGMF predictions reveal the leading 10 potential indications for the five drugs, with proven applications emphasized in bold. Notably, each drug demonstrates 4 to 5 substantiated indications among the top 10 forecasts. As an instance, considering methotrexate (DB00563) as an antimetabolite anticancer agent, it demonstrates effectiveness across diverse cancer

Table 5. The top 10 potential applications for fluorouracil, methotrexate, and doxorubicin.

Drugs (DrugBank IDs)	Top 10 potential diseases (OMIM IDs)
Fluorouracil (DB00544)	**Prostate Cancer (176807)**; **Renal Cell Carcinoma, Nonpapillary (144700)**; Osteogenic Sarcoma (259500); **Bladder Cancer (109800)**; Leukemia, Acute Myelocytic, with Polyposis Coli and Colon Cancer (246470); Melanoma, Cutaneous Malignant, Susceptibility to, 1 (155600); **Lung Cancer (211980)**; Esophageal Cancer (133239); **Wilms Tumor 1 (194070)**; Small Cell Cancer of The Lung (182280).
Methotrexate (DB00563)	**Lung Cancer (211980)**; Neuroblastoma, Susceptibility to, 1 (256700); **Prostate Cancer (176807)**; **Wilms Tumor 1 (194070)**; Leukemia, Chronic Lymphocytic (151400); Myeloma Multiple (254500); **Thyroid Cancer, Nonmedullary, 2 (188470)**; Moved to 619182 (175505); Renal Cell Carcinoma, Nonpapillary (144700); Pronation-Supination of The Forearm, Impairment of (176800).
Doxorubicin (DB00997)	**Esophageal Cancer (133239)**; Colorectal Cancer (114500); Kaposi Sarcoma, Susceptibility to (223350); Leukemia, Chronic Lymphocytic, Susceptibility to, 2 (109543); **Reticulum Cell Sarcoma (267730)**; **Kaposi Sarcoma, Susceptibility to (267730)**; Small Cell Cancer of The Lung (182280); Testicular Germ Cell Tumor (273300); **Prostate Cancer (176807)**; **Renal Cell Carcinoma, Nonpapillary (144700)**.
Cisplatin (DB00515)	Rhabdomyosarcoma 2 (268220); **Diffuse Gastric and Lobular Breast Cancer Syndrome (137215)**; **Mycosis Fungoides (254400)**; Breast Cancer (114480); **Hydatidiform Mole, Recurrent, 1 (231090)**; Lymphoblastic Leukemia, Acute, With Lymphomatous Features (247640); **Reticulum Cell Sarcoma (267730)**; **Leukemia, Chronic Lymphocytic, Susceptibility to, 2 (109543)**; Lung Cancer (211980); Colorectal Cancer (114500).
Vincristine (DB00541)	Leukemia, Chronic Lymphocytic (151400); Mycosis Fungoides (254400); Myelofibrosis (254450); **Kaposi Sarcoma, Susceptibility To (148000)**; **Thyroid Cancer, Nonmedullary, 2 (188470); Bladder Cancer(109800)**; **Lung Cancer (211980)**; Osteogenic Sarcoma (259500); **Diffuse Gastric and Lobular Breast Cancer Syndrome (137215)**; Breast Cancer (114480);

The results highlighted in **bold** align with the CTD database findings.

types, namely Lung Cancer, Prostate Cancer, Wilms Tumor 1, and Nonmedullary Thyroid Cancer. Additionally, candidate indications encompass Neuroblastoma, Susceptibility To, 1, ranking second, with only this entry among the leading four awaiting verification. These candidate indications, which still need to be verified, hold great promise for targeted investigation in future research efforts.

Experiments on the other dataset

To further validate the feasibility of multiGMF on additional datasets and assess the model's generalization capability, we conducted a series of comparative experiments on two other datasets, including Cdataset[43] and our latest collected dataset, namely CTDdataset2023. The Cdataset comprises 409 diseases gathered from the OMIM database, 663 drugs sourced from DrugBank, and a set of 2352 known drug-disease associations. This CTDdataset2023 is constructed by integrating the latest information from the CTD database, encompassing 1327 drugs, 278 diseases, and 3740 associations. We employed a combination of 10-fold cross-validation and cold-start tests to assess the effectiveness of each method on the Cdataset and CTDdataset2023.

As indicated in Table 6, we evaluated various comparative methodologies on the Cdataset, encompassing measurements such as AUC, AUPR, and precision. In the 10-fold cross-validation, multiGMF achieves excellent results on all indices, demonstrating its superior predictive performance. In cold-start test, multiGMF achieved AUC, AUPR, and precision values of 0.899, 0.273, and 0.339, respectively. It is worth noting that the method ranked second (i.e., MSBMF), displayed AUC, AUPR, and precision that are lower than those of our approach by 0.523%, 1.794%, and 2.114%, respectively, in the context of the cross-validation.

Table 6. The metrics for each method during 10-fold cross-validation and cold-start tests on the Cdataset.

Tests	Metrics	**multiGMF**	DDA-SKF	MSBMF	HGIMC	BNNR	MBiRW
10-fold CV	AUC	**0.961**	0.938	0.956	0.953	0.948	0.933
	AUPR	**0.454**	0.392	0.446	0.428	0.441	0.310
	Precision	**0.483**	0.429	0.473	0.463	0.471	0.351
cold-start	AUC	**0.899**	0.832	0.883	0.879	0.812	0.804
	AUPR	0.273	0.200	**0.298**	0.294	0.193	0.174
	Precision	0.339	0.246	**0.367**	0.356	0.254	0.232

As indicated in Table 7, we conducted the same tests in CTDdataset2023. In the 10-fold cross-validation, multiGMF shows comparable AUC values to MSBMF, while achieving higher AUPR and precision by 15.789% and 18%, respectively. In the cold-start test, the multiGMF method demonstrates similar AUC values to MSBMF, while surpassing in AUPR and precision by 20.502% and 21.786%, respectively. In CTDdataset2023, DDA-SKF employs the same similarity as the other methods. As evident from Table 7, the AUPR of multiGMF outperforms the other methods in both tests. However, apart from the similarity in precision between multiGMF and MSBMF in the 10-fold cross-validation, MSBMF consistently lags behind multiGMF in other metric values, particularly in terms of AUPR and precision. Overall, in comparison to alternative methods, our model demonstrates superior performance on additional datasets.

Table 7. The metrics for each method during 10-fold cross-validation and cold-start tests on the CTDdataset2023.

Tests	Metrics	**multiGMF**	DDA-SKF	MSBMF	HGIMC	BNNR	MBiRW
10-fold CV	AUC	0.891	0.874	**0.892**	0.886	0.849	0.874
	AUPR	**0.171**	0.168	0.144	0.134	0.107	0.110
	Precision	0.200	**0.203**	0.164	0.155	0.121	0.128
cold-start	AUC	0.900	0.879	**0.904**	0.890	0.816	0.883
	AUPR	**0.239**	0.233	0.190	0.222	0.111	0.118
	Precision	**0.280**	0.277	0.219	0.267	0.137	0.139

Conclusions

In this work, we provide a computational model named multiGMF that uses multiple similarity measures to infer potential drug-association indications. The multiGMF method not only efficiently handles the sparse drug-disease association matrix, but also extracts node features from multiple similarities to generate a comprehensive low-rank feature matrix. Pre-filling the association matrix using WKNN enhances its density, facilitating the coupling of association features and similarity features. In cross-validation and cold-start tests, multiGMF demonstrates outstanding predictive performance. This highlights its significant advantages and emphasizes its superior efficacy in extracting geometric features from high-dimensional drug and disease spaces. Furthermore, case studies further validate the reliability of multiGMF in identifying new uses for known drugs. In summary, we believe that multiGMF holds promise as a potent and efficient instrument for computational drug repositioning.

References

1. Chong, C. & Sullivan, D. New uses for old drugs. *Nature*, 448, 645–646. http://dx.doi.org/10.1007/bf03297424 (2007).
2. Li, J. *et al.* A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics*, 17(1), 2–12. http://dx.doi.org/10.1093/bib/bbv020 (2016).
3. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8), 673–683. http://dx.doi.org/10.1038/nrd1468 (2004).
4. Luo, H. *et al.* Biomedical data and computational models for drug repositioning: a comprehensive review. *Briefings in Bioinformatics*, 22(2), 1604–1619. http://dx.doi.org/10.1093/bib/bbz176 (2021).
5. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery*, 1(9), 727–730. http://dx.doi.org/10.1038/nrd892 (2002).
6. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1), 56–68. http://dx.doi.org/10.1038/nrg2918 (2010).
7. Wang, W., Yang, S. & Li, J. Drug target predictions based on heterogeneous graph inference. *Pacific Symposium on Biocomputing*, 18, 53–64. http://dx.doi.org/10.1142/9789814447973_0006 (2013).
8. Zhou, G. *et al.* Drug repositioning based on a multiplex network by integrating disease, gene, and drug information. *Current Bioinformatics*, 18(3), 267–275. http://dx.doi.org/10.2174/1574893618666230223114427 (2023).
9. Yang, J., He, S., Zhang, Z. & Bo, X. Negstacking: Drug–target interaction prediction based on ensemble learning and logistic regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(6), 2624–2634. http://dx.doi.org/10.1109/TCBB.2020.2968025 (2021).
10. Zhao, B. W., Hu, L., You, Z. H., Wang, L. & Su, X. R. Hingrl: predicting drug-disease associations with graph representation learning on heterogeneous information networks. *Briefings in Bioinformatics*, 23(1), bbab515. http://dx.doi.org/10.1093/bib/bbab515 (2022).
11. Issa, N. T., Stathias, V., Schürer, S. & Dakshanamurthy, S. Machine and deep learning approaches for cancer drug repositioning. *Seminars in Cancer Biology*, 68, 132–142. http://dx.doi.org/10.1016/j.semcancer.2019.12.011 (2021).
12. Jiang, H. & Huang, Y. An effective drug-disease associations prediction model based on graphic representation learning over multi-biomolecular network. *BMC Bioinformatics*, 23(1), 9. http://dx.doi.org/10.1186/s12859-021-04553-2 (2022).
13. Thafar, M. A. *et al.* Affinity2vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Scientific Reports*, 12(1), 4751. http://dx.doi.org/10.1038/s41598-022-08787-9 (2022).
14. Ghasemi, S. & Lakizadeh, A. Drug–disease association prediction based on end–to–end multi–layerr heterogeneous graph convolutional encoders. *Informatics in Medicine Unlocked*, 37, 101177. http://dx.doi.org/10.1016/j.imu.2023.101177 (2023).
15. Wang, Y., Gao, Y. L., Wang, J., Li, F. & Liu, J. X. Msgca: Drug-disease associations prediction based on multi-similarities graph convolutional autoencoder. *IEEE Journal of Biomedical and Health Informatics*, 27(7), 3686–3694. http://dx.doi.org/10.1109/JBHI.2023.3272154 (2023).
16. Yan, Y. *et al.* Drug repositioning based on multi-view learning with matrix completion. *Briefings in Bioinformatics*, 23(3), bbac054. http://dx.doi.org/10.1093/bib/bbac054 (2022).
17. Yang, M., Luo, H., Li, Y. & Wang, J. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics*, 35(14), i455–i463. http://dx.doi.org/10.1093/bioinformatics/btz331 (2019).

18. Yang, M., Wu, G., Zhao, Q., Li, Y. & Wang, J. Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Briefings in Bioinformatics*, 22(4), bbaa267. http://dx.doi.org/10.1093/bib/bbaa267 (2021).

19. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1), 496. http://dx.doi.org/10.1038/msb.2011.26 (2011).

20. Wishart, D. S. *et al.* Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34, D668–D672. http://dx.doi.org/10.1093/nar/gkj067 (2006).

21. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(1), 514–517. http://dx.doi.org/10.1093/nar/gki033 (2005).

22. Steinbeck, C. *et al.* The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics. *Chemical Information and Computer Sciences*, 43(2), 493–500. http://dx.doi.org/10.1021/ci025584y (2003).

23. Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Chemical Information and Computer Sciences*, 28(1), 31–36. http://dx.doi.org/10.1021/ci00057a005 (1988).

24. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. *In IJCAI '95: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada*. http://dx.doi.org/10.48550/arXiv.cmp-lg/9511007 (1995).

25. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sider database of drugs and side effects. *Nucleic Acids Research*, 44, D1075–D1079. http://dx.doi.org/10.1093/nar/gkv1075 (2016).

26. Jaccard, P. Nouvelles recheres sur la distribution florale. *Bulletins Société Vaudoise des Sciences Naturelles*, 44, 223–270. https://cir.nii.ac.jp/crid/1573668925301098880 (1908).

27. Van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. A. A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5), 535–542. http://dx.doi.org/10.1038/sj.ejhg.5201585 (2006).

28. Lipscomb, C. E. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3), 265. https://pmc.ncbi.nlm.nih.gov/articles/PMC35238/ (2000).

29. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C. F. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10), 1274–1281. http://dx.doi.org/10.1093/bioinformatics/btm087 (2007).

30. Chaudhary, P., Varshney, Y. V., Srivastava, G. & Bhatia, S. Motor imagery classification using sparse nonnegative matrix factorization and convolutional neural networks. *Neural Computing and Applications*, 1–11. http://dx.doi.org/10.1007/s00521-022-07861-7 (2022).

31. Wang, H., Wang, M., Li, J., Song, L. & Hao, Y. A novel signal separation method based on improved sparse non-negative matrix factorization. *Entropy*, 21(5), 445. http://dx.doi.org/10.3390/e21050445 (2019).

32. Parvina, H., Moradi, P., Esmaeilib, S. & Jalilic, M. An efficient recommender system by integrating non-negative matrix factorization with trust and distrust relationships. *IEEE Data Science Workshop*, 135–139. http://dx.doi.org/10.1109/DSW.2018.8439905 (2018).

33. Cai, D., He, X., Han, J. & Huang, T. S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8), 1548–1560. http://dx.doi.org/10.1109/TPAMI.2010.231 (2011).

34. Wan, M., Lai, Z., Ming, Z. & Yang, G. An improve face representation and recognition method based on graph regularized non-negative matrix factorization. *Multimedia Tools and Applications*, 78, 22109–22126. http://dx.doi.org/10.1007/s11042-019-7454-2 (2019).

35. Wan, M., Lai, Z., Ming, Z. & Yang, G. Feature extraction via multi-view non-negative matrix factorization with local graph regularization. *IEEE International Conference on Image Processing*, 3500–3504. http://dx.doi.org/10.1109/ICIP.2015.7351455 (2015).

36. Liu, Z., Luo, X. & Zhou, M. Symmetry and graph bi-regularized non-negative matrix factorization for precise community detection. *IEEE Transactions on Automation Science and Engineering*. http://dx.doi.org/10.1109/TASE.2023.3240335 (2023).

37. Li, C., Che, H., Leung, M. F., Liu, C. & Yan, Z. Robust multi-view non-negative matrix factorization with adaptive graph and diversity constraints. *Information Sciences*, 634, 587–607. http://dx.doi.org/10.1016/j.ins.2023.03.119 (2023).

38. Xiao, Q., Luo, J., Liang, C., Cai, J. & Ding, P. A graph regularized non-negative matrix factorization method for identifying microrna-disease associations. *Bioinformatics*, 34(2), 239–248. http://dx.doi.org/10.1093/bioinformatics/btx545 (2018).

39. Gao, L. G., Yang, M. Y. & Wang, J. X. Collaborative matrix factorization with soft regularization for drug-target interaction prediction. *Journal of Computer Science and Technology*, 36, 310–322. http://dx.doi.org/10.1007/s11390-021-0844-8 (2021).

40. Facchinei, F., Kanzow, C. & Sagratella, S. Solving quasi-variational inequalities via their kkt conditions. *Mathematical Programming*, 144, 369–412. http://dx.doi.org/10.1007/s10107-013-0637-0 (2014).

41. Gao, C. Q., Zhou, Y. K., Xin, X. H., Min, H. & Du, P. F. Dda-skf: Predicting drug–disease associations using similarity kernel fusion. *Frontiers in Pharmacology*, 12, 784171. http://dx.doi.org/10.3389/fphar.2021.784171 (2022).

42. Yang, M., Huang, L., Xu, Y., Lu, C. & Wang, J. Heterogeneous graph inference with matrix completion for computational drug repositioning. *Bioinformatics*, 36(22–23), 5456–5464. http://dx.doi.org/10.1093/bioinformatics/btaa1024 (2021).

43. Luo, H. *et al.* Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17), 2664–2671. http://dx.doi.org/10.1093/bioinformatics/btw228 (2016).

44. Davis, A. P. *et al.* The comparative toxicogenomics database: update 2019. *Nucleic Acids Research*, 47(D1), D948–D954. http://dx.doi.org/10.1093/nar/gky868 (2019).

Funding Declaration

This work was supported by the National Natural Science Foundation of China (Grant no. 62272309) and the Natural Science Foundation of Hunan (Grant no. 2026JJ50233).

Author contributions statement

Conceptualization, M.Y. and B.Y.; methodology, M.Y. and B.Y.; software, B.Y.; validation, M.Y. ; formal analysis, G.D.; investigation, M.Y. and X.T.; resources, M.Y.; data curation, M.Y. and B.Y.; writing–original draft preparation, M.Y. and B.Y.; writing–review and editing, X.T. and G.D.; visualization, M.Y. and B.Y.; supervision, M.Y.; project administration, M.Y. ; funding acquisition, M.Y.; All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.