scientific reports

Check for updates

OPEN

Improving drug repositioning accuracy using non-negative matrix tri-factorization

Qingmei Li[1], Yangyang Wang[2], Jihan Wang[3] & Congzhe Zhao[1]✉

Drug repositioning is a transformative approach in drug discovery, offering a pathway to repurpose existing drugs for new therapeutic uses. In this study, we introduce the IDDNMTF model designed to predict drug repositioning opportunities with greater precision. The IDDNMTF model integrates multiple datasets, allowing for a more comprehensive analysis of drug-disease associations. We evaluated the IDDNMTF model using various combinations of datasets and found that its performance, as measured by AUC, AUPR, and F1 scores, improved with the inclusion of more data. This trend underscores the importance of data diversity in strengthening predictive capabilities. Comparatively, the IDDNMTF model demonstrated superior performance against the NMF model, solidifying its potential in drug repositioning. In summary, the IDDNMTF model offers a promising tool for identifying new therapeutic uses for existing drugs. Its predictive accuracy and interpretability are poised to accelerate the transition from bench to bedside, contributing to personalized medicine and the development of targeted treatments.

Keywords Drug repositioning, Non-negative matrix Tri-factorization (NMTF), Computational modeling, Drug-protein interactions, Drug-disease associations, Personalized medicine

The pharmaceutical landscape is continuously reshaped by the innovative approach of drug repositioning, a strategy that involves identifying new therapeutic applications for existing drugs[1–3]. This paradigm shift is driven by the potential to mitigate the extensive costs and lengthy timelines associated with traditional drug development[4,5]. Drug repositioning offers a unique opportunity to tap into the vast repository of drugs already proven safe[6], thereby accelerating the discovery of treatments for a myriad of diseases, including those that are currently underserved by existing therapies[7,8].

The quest for novel uses of existing drugs has been propelled by a variety of sophisticated methods. Non-negative Matrix Factorization (NMF) has been a cornerstone in this endeavor[9–11], adept at uncovering latent factors within gene expression data that correlate with drug efficacy and disease mechanisms[12,13]. The Bayesian Network with Neighboring Relations (BNNR) has provided a probabilistic lens through which the complex interdependencies between drugs and diseases can be viewed[14]. DANN-DDI, a deep learning model, has excelled in predicting drug-drug interactions by considering the intricate patterns that emerge from multi-label classification tasks[15]. The Drug-Drug Interactions by Random Survival Forest (DRRS) has harnessed the power of machine learning to forecast potential interactions based on a comprehensive set of drug features[16]. Additionally, the Sparse Correlation Matrix Factorization for Drug-Disease association (SCMFDD) has emerged as a method that adeptly captures the sparse yet significant relationships within large-scale datasets, further enriching the arsenal of drug repositioning techniques[17]. Despite these advancements, the dynamic and multifaceted nature of biological systems calls for an approach that can dissect the intricate layers of drug-disease interactions with greater precision[8]. It is within this context that we introduce the Non-negative Matrix Tri-factorization (NMTF) method[18]. NMTF transcends the capabilities of its predecessors by introducing an additional dimension to the factorization process, allowing for a more granular and nuanced exploration of the relationships between drugs, diseases, genes and biological pathways[19,20]. This tri-factorization not only captures the interdependencies within the data but also enhances the interpretability of the results, providing a clearer understanding of the underlying mechanisms that may link a drug to a new therapeutic use[21]. The advantages of NMTF are multifaceted. It offers improved accuracy in predicting drug repositioning candidates by accounting for the multi-dimensionality of biological data. Its ability to handle high-dimensional datasets makes it particularly suited for the analysis of large-scale omics data, which is increasingly available in the era of big data in biomedicine[22]. The interpretability

[1]Honghui Hospital, Xi'an Jiaotong University, Xi'an 710054, China. [2]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China. [3]Shaanxi Provincial Key Laboratory of Infection and Immune Diseases, Shaanxi Provincial People's Hospital, Xi'an 710068, China. ✉email: 849855369@qq.com

No.	Matrix name	Matrix description	Matrix Size
1	R_{12}	Drug-Label association matrix	141 × 3261
2	R_{23}	Drug-protein association matrix	3261 × 3691
3	R_{24}	Drug-Pathway association matrix	3261 × 307
4	R_{25}	Drug-Disease association matrix	3261 × 841

Table 1. Associations between drugs and labels, proteins, pathways, and diseases.

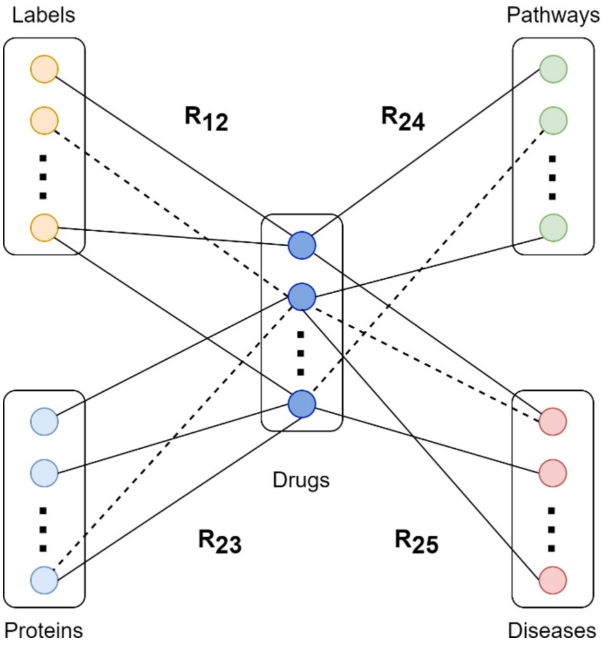


Fig. 1. The network of datasets base on R_{12}, R_{23}, R_{24}, R_{25}. Solid lines represent known correlations between different indicators, and dotted lines represent unknown or predictable correlations.

of NMTF also stands out, allowing researchers to translate computational predictions into biological insights, thus facilitating the translation of findings from bench to bedside[23].

In this study, we proposed the IDDNMTF (NMTF model integrating diverse datasets), which aims to leverage the ability of NMTF methods to accurately analyze complex biological data to identify new therapeutic indications for existing drugs. We anticipate that the application of IDDNMTF will reveal new pathways for drug utilization that are obscured by the limitations of traditional methods, ultimately contributing to the advancement of personalized medicine and the development of more effective and targeted treatments.

Materials and methods

In this section, we first give an overview of the datasets used in our study. Then, we introduce the NMF and NMTF models, and finally give the objective function, iteration rules, and evaluation criteria based on NMTF in this paper based on the datasets.

Datasets

DrugBank[24,25] is a comprehensive online database that provides detailed information on drugs, including their chemical structures, pharmacological profiles, and therapeutic uses. This paper integrates 3261 drugs and their classification labels, proteins, pathways, and disease associations from DrugBank[15] and Therapeutic Target Database (TTD)[26], as shown in Table 1; Fig. 1.

NMF and NMTF

Before introducing NMTF, it is necessary to understand the basic principles of NMF. NMF is a computational method used in various fields, including data mining, machine learning, and signal processing. It aims to approximate a non-negative matrix $R^{n\times m}$ by multiplying two lower-rank non-negative matrices G_1 and G_2 , $R\approx G_1G_2^T$, as shown in Fig. 2, and the objective function of NMF is

$$\min_{G_1\geqslant 0,G_2\geqslant 0}||R-G_1G_2^T||_F^2,\ s.t.\ \mathrm{G}_1\geqslant 0,G_2\geqslant 0 \tag{1}$$

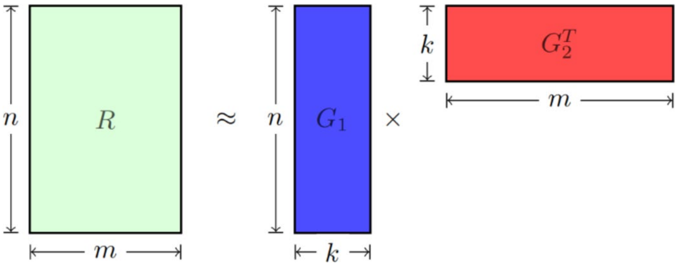


Fig. 2. The framework of NMF model. G_1 and G_2 are low-rank feature matrices.

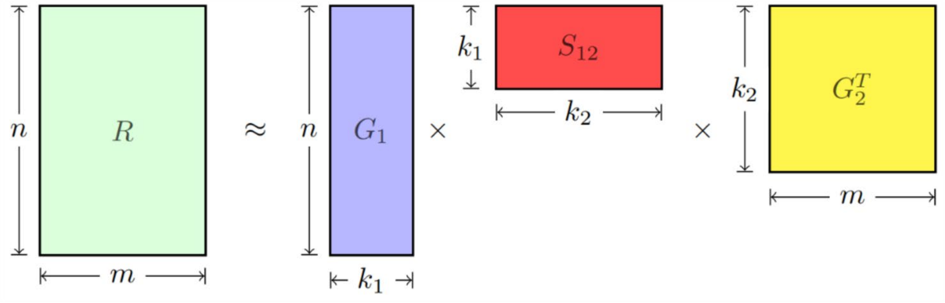


Fig. 3. The framework of NMTF model. G_1 and G_2 are low-rank feature matrices, and S_{12} is the correlation matrix.

where $G_1 \in \mathbb{R}_+^{n\times k}$represents the basis vectors and $G_2 \in \mathbb{R}_+^{m\times k}$ represents the coefficients, $k \ll \min\{m, n\}$. In bioinformatics, NMF is used to analyze gene expression data, clustering genes with similar expression patterns and identifying potential biomarkers.

In the research conducted by Ding et al.[27], it was demonstrated that the imposition of orthogonal constraints $G_1^T G_1 = I$ and $G_2^T G_2 = I$, mirrors the execution of a k-means clustering algorithm on both the row and column dimensions of R, as shown in Eq. (2):

$$\min_{G_1\geqslant 0, G_2\geqslant 0} ||R - G_1 G_2^T||_F^2, \ s.t. \ G_1^T G_1 = I, G_2^T G_2 = I \tag{2}$$

G_1 serves as the indicator matrix that signifies the clustering of rows, and G_2 as the indicator matrix that denotes the clustering of columns. To solve the problem that the biorthogonality constraint is too strict for low-rank approximation, Ding proposed the NMTF framework[18], which decomposes R into three components, $R \approx G_1 S_{12} G_2^T$, as shownFig. ig. 3, and the objective function is

$$\min_{G_1\geqslant 0, G_2\geqslant 0, S\geqslant 0} ||R - G_1 S_{12} G_2^T||_F^2, \ s.t. \ G_1^T G_1 = I, G_2^T G_2 = I \tag{3}$$

where $G_1 \in \mathbb{R}_+^{n\times k_1}$, $G_2 \in \mathbb{R}_+^{m\times k_2}$, $S_{12} \in \mathbb{R}_+^{k_1\times k_2}$ and $k \ll \min\{m, n\}$.

The solution for Eq. (3) can be achieved through the following iterative rules:

$$G_{1(i,j)} \leftarrow G_{1(i,j)} \sqrt{\frac{(RG_2 S_{12}^T)_{i,j}}{(G_1 G_1^T R G_2 S_{12}^T)_{i,j}}} \tag{4}$$

$$G_{2(i,j)} \leftarrow G_{2(i,j)} \sqrt{\frac{(RG_1 S_{12})_{i,j}}{(G_2 G_2^T R^T G_1 S_{12})_{i,j}}} \tag{5}$$

$$S_{12(i,j)} \leftarrow S_{12(i,j)} \sqrt{\frac{(G_1^T R G_2)_{i,j}}{(G_1^T G_1 S_{12} G_2^T G_2)_{i,j}}} \tag{6}$$

The extended NMTF model for our datasets

Equation 3 is specifically tailored for the resolution of a single association matrix, thus necessitating an extension to interrelate $R_{12}, R_{23}, R_{24}, R_{25}$ in Fig. 1, effectively. From Fig. 3, the extended objective function is:

$$\min_{G_i \geqslant 0, i=1,2,\ldots,5} ||R_{12} - G_1 S_{12} G_2^T||_F^2 + ||R_{23} - G_2 S_{23} G_3^T||_F^2$$
$$+||R_{24} - G_2 S_{24} G_4^T||_F^2 + ||R_{25} - G_2 S_{25} G_5^T||_F^2, \ s.t. \ G_i^T G_i = I, i = 1, 2, \ldots, 5 \tag{7}$$

The iterative algorithms for Eq. (7) are outlined below, enabling the computation of a suite of positive matrices:$G_1, G_2, G_3, G_4, G_5$and$S_{12}, S_{23}, S_{24}, S_{25}$. These matrices are mutually orthogonal and collectively serve to optimize the objective function. The update rules for Eq. (7) are shown in Eqs. (8)-(16).

$$G_{1_{(i,j)}} \leftarrow G_{1_{(i,j)}} \sqrt{\frac{(R_{12} G_2 S_{12}^T)_{i,j}}{(G_{11} R_{12} G_2 S_{12}^T)_{i,j}}} \tag{8}$$

$$G_{2(i,j)} \leftarrow G_{2(i,j)} \sqrt{\frac{(R_{12}^T G_1 S_{12} + R_{23} G_3 S_{23}^T + R_{24} G_4 S_{24}^T + R_{25} G_5 S_{25}^T)_{i,j}}{(G_{22} R_{12}^T G_1 S_{12} + G_{22} R_{23} G_3 S_{23}^T + G_{22} R_{24} G_4 S_{24}^T + G_{22} R_{25} G_5 S_{25}^T)_{i,j}}} \tag{9}$$

$$G_{3_{(i,j)}} \leftarrow G_{3_{(i,j)}} \sqrt{\frac{(R_{23} G_2 S_{23}^T)_{i,j}}{(G_{33} R_{23} G_2 S_{23}^T)_{i,j}}} \tag{10}$$

$$G_{4_{(i,j)}} \leftarrow G_{4_{(i,j)}} \sqrt{\frac{(R_{24} G_2 S_{24}^T)_{i,j}}{(G_{44} R_{24} G_2 S_{24}^T)_{i,j}}} \tag{11}$$

$$G_{5_{(i,j)}} \leftarrow G_{5_{(i,j)}} \sqrt{\frac{(R_{25} G_2 S_{25}^T)_{i,j}}{(G_{55} R_{25} G_2 S_{25}^T)_{i,j}}} \tag{12}$$

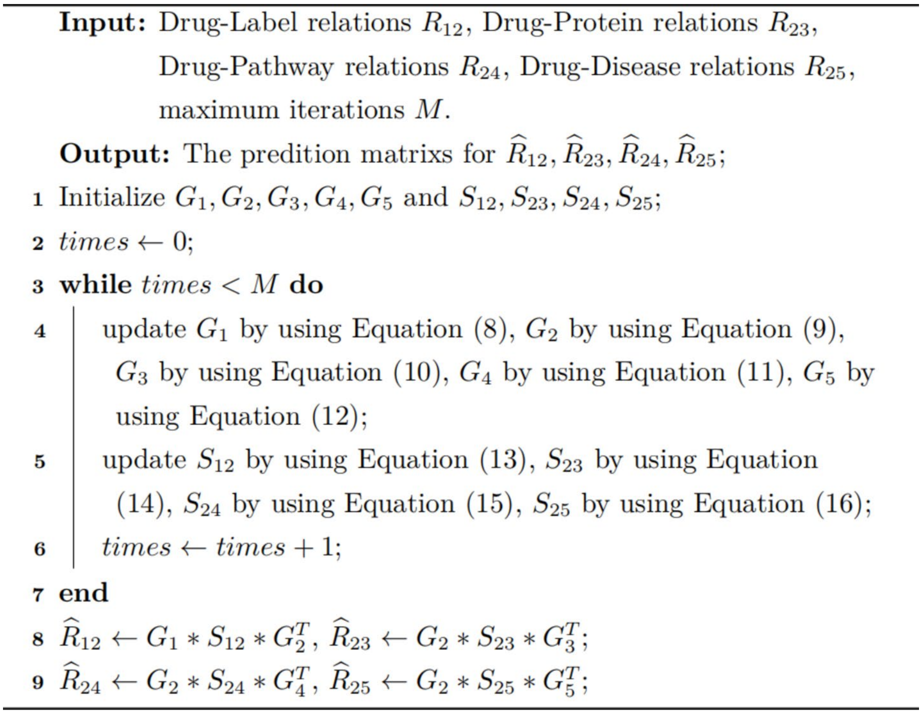
$$S_{12_{(i,j)}} \leftarrow S_{12_{(i,j)}} \sqrt{\frac{(G_1^T R_{12} G_2)_{i,j}}{(G_1^T G_1 S_{12} G_2^T G_2)_{i,j}}} \tag{13}$$

$$S_{23_{(i,j)}} \leftarrow S_{23_{(i,j)}} \sqrt{\frac{(G_2^T R_{23} G_3)_{i,j}}{(G_2^T G_2 S_{23} G_3^T G_3)_{i,j}}} \tag{14}$$

$$S_{24_{(i,j)}} \leftarrow S_{24_{(i,j)}} \sqrt{\frac{(G_2^T R_{24} G_4)_{i,j}}{(G_2^T G_2 S_{24} G_4^T G_4)_{i,j}}} \tag{15}$$

$$S_{25_{(i,j)}} \leftarrow S_{25_{(i,j)}} \sqrt{\frac{(G_2^T R_{25} G_5)_{i,j}}{(G_2^T G_2 S_{25} G_5^T G_5)_{i,j}}} \tag{16}$$

where $G_{ii} = G_i G_i^T, i = 1, 2, 3, 4, 5$. Algorithm 1 shows the procedure of IDDNMTF.



```
Input: Drug-Label relations $R_{12}$, Drug-Protein relations $R_{23}$,
       Drug-Pathway relations $R_{24}$, Drug-Disease relations $R_{25}$,
       maximum iterations $M$.
Output: The predition matrixs for $\widehat{R}_{12}, \widehat{R}_{23}, \widehat{R}_{24}, \widehat{R}_{25}$;
Initialize $G_1, G_2, G_3, G_4, G_5$ and $S_{12}, S_{23}, S_{24}, S_{25}$;
$times \leftarrow 0$;
while $times < M$ do
    update $G_1$ by using Equation (8), $G_2$ by using Equation (9),
      $G_3$ by using Equation (10), $G_4$ by using Equation (11), $G_5$ by
      using Equation (12);
    update $S_{12}$ by using Equation (13), $S_{23}$ by using Equation
      (14), $S_{24}$ by using Equation (15), $S_{25}$ by using Equation (16);
    $times \leftarrow times + 1$;
end
$\widehat{R}_{12} \leftarrow G_1 * S_{12} * G_2^T$, $\widehat{R}_{23} \leftarrow G_2 * S_{23} * G_3^T$;
$\widehat{R}_{24} \leftarrow G_2 * S_{24} * G_4^T$, $\widehat{R}_{25} \leftarrow G_2 * S_{25} * G_5^T$;
```

Algorithm 1. Optimization algorithm for IDDNMTF.

Evaluation metrics

The metrics of interest include Area Under the Curve (AUC), Area Under the Precision-Recall Curve (AUPR), Accuracy (Acc), Sensitivity or Recall (Sen), Specificity (Spe), Precision (Pre), and F1 Score (Fl), as shown in Eq. (17) to Eq. (23).

$$SN = \frac{TP}{TP + FN} \tag{17}$$

$$SP = \frac{TN}{TN + FP} \tag{18}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$precision = \frac{TP}{TP + FP} \tag{20}$$

$$recall = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{22}$$

These metrics provide a holistic view of the classification capabilities of each algorithm. All the experiments were conducted based on MATLAB 2023b and Python 3.7 on Windows 10 with Intel(R) Core (TM) i5-12400 F 2.50 GHz.

Parameter settings

Here, we need to explain the parameter settings for solving Eq. (7). To find new associations between indications and drugs, the parameter initialization of IDDNMTF is crucial[28]. Spherical k-means initialization is a method used to initialize the centroids in the k-means clustering algorithm in such a way that they are distributed uniformly on the surface of a hypersphere[29]. This approach is particularly useful for high-dimensional data where traditional initialization methods may not perform well due to the "curse of dimensionality. In this research, we adopt spherical k-means method to implement NMTF parameter initialization, and the hyperparameters of spherical k-means are fixed as $k_1, k_2, k_3, k_4, k_5 = 500, 141, 500, 500, 300$[30]. In addition, we define the maximum number of iterations as 100 to balance the computational accuracy and the algorithm running time.

Model	AUC	AUPR	Acc	Sen (Recall)	Spe	Pre	F1
R_{12}	0.9269	0.6572	0.9643	0.5693	0.9848	0.6603	0.6124
R_{12}, R_{23}	0.9305	0.6679	0.9670	**0.5990**	0.9861	0.6901	0.6426
R_{12}, R_{23}, R_{24}	0.9315	0.6780	0.9674	0.5875	0.9871	0.7031	0.6418
R_{12}, R_{23}, R_{24}, R_{25}	**0.9315**	**0.6871**	**0.9694**	0.5813	**0.9896**	**0.7425**	**0.6538**

Table 2. The results of different combinations of datasets. Significant values are in bold.

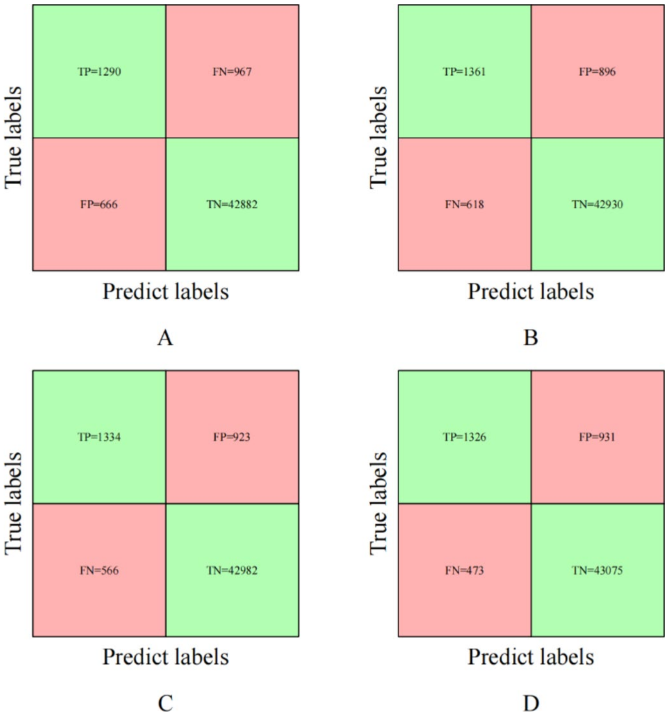


Fig. 4. The confuse matrix for different combinations of datasets. (**A**) The confuse matrix for R12; (**B**) The confuse matrix for R12-R23; (**C**) The confuse matrix for R12-R23-R24; (**D**) The confuse matrix for R12-R23-R24-R25.

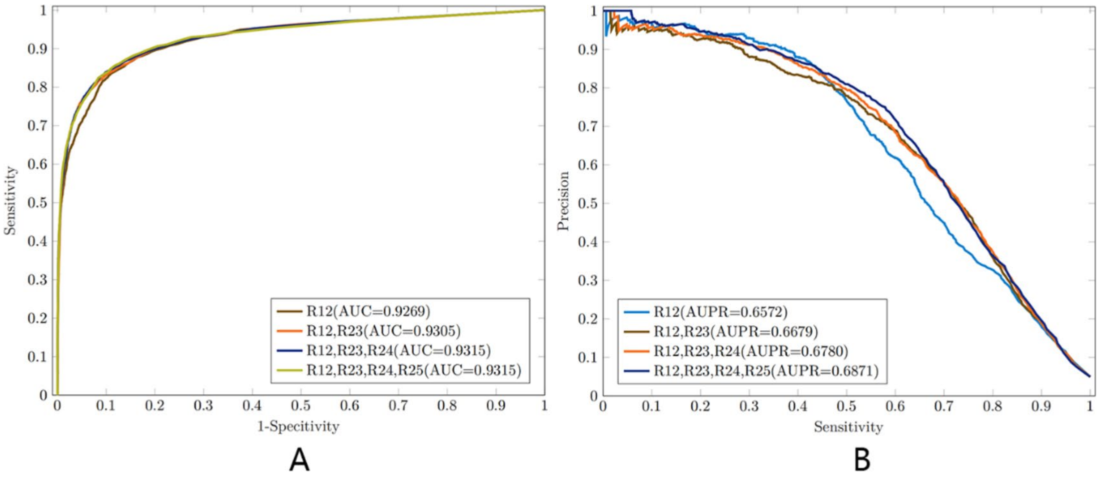


Fig. 5. The AUC and AUPR for different combinations of datasets (R_{12}, R_{12} ~ R_{23}, R_{12} ~ R_{24}, R_{12} ~ R_{25}). (**A**) AUC for different combinations of datasets; (**B**) AUPR for different combinations of datasets.

Model	AUC	AUPR	Acc	Sen (Recall)	Spe	Pre	F1
NMF (*k* = 10)	**0.8867**	0.5245	0.9541	0.4812	0.9786	0.5382	0.5089
NMF (*k* = 50)	0.8840	**0.6327**	**0.9664**	**0.5476**	0.9881	**0.7043**	**0.6165**
NMF (*k* = 100)	0.7545	0.4374	0.9594	0.3833	**0.9893**	0.6489	0.4826
NMF (*k* = 141)	0.7523	0.2615	0.9231	0.3301	0.9538	0.2702	0.2983

Table 3. The results of NMF algorithm on R_{12} dataset under different *k* values. Significant values are in bold.

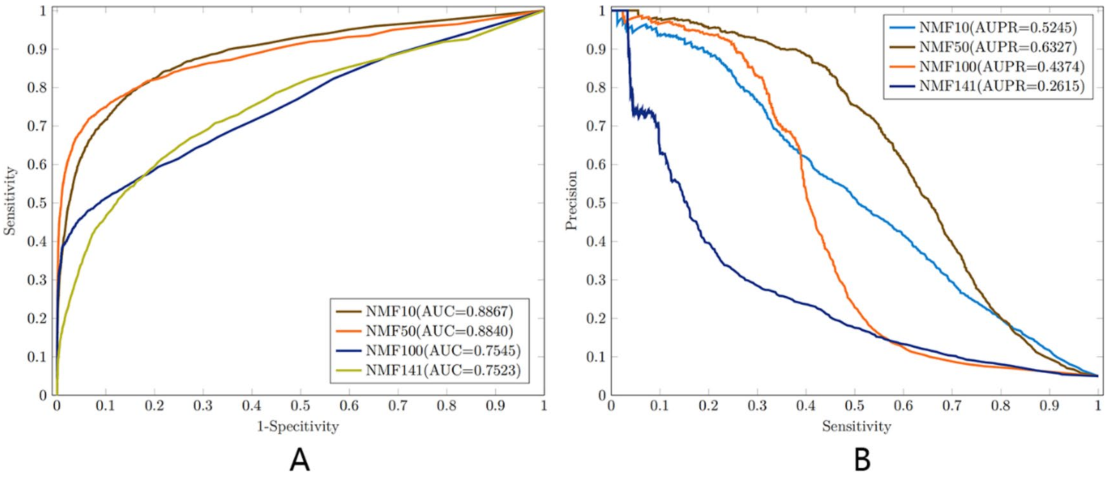


Fig. 6. The AUC and AUPR based on R_{12} for NMF model with Parameters (*k* = 10, *k* = 50, *k* = 100 and *k* = 141). (**A**) AUC based on R_{12} for NMF model; (**B**) AUPR based on R_{12} for NMF model.

Results

Benefit of adding association dataset

Table 2; Figs. 4 and 5 display the results of the IDDNMTF model using different combinations of datasets. To predict whether a specific relationship exists, A range of thresholds is defined, spanning from the minimum to the maximum value of the predicted list. Here, 1000 evenly spaced threshold values are generated to evaluate the model's performance under varying levels of sensitivity and specificity. For each threshold value, the continuous predictions are binarized. If a predicted value is greater than or equal to the threshold, it is classified as 1 (True); otherwise, it is classified as 0(False). Based on the confuse matrix in Fig. 4, we can calculate the corresponding results by using evaluation metrics. For IDDNMTF only utilizing the R_{12} dataset, the AUC score is 0.9269, which is a good indicator of the model's ability to distinguish between classes. The AUPR is 0.6572, suggesting a good balance between precision and recall, and the F1 score is 0.6124. These metrics show excellent performance when using only the R_{12} dataset. By adding the R_{23} dataset to the model, we observe improvements in some metrics. The accuracy increases to 0.9305, and the recall improves to 0.5990. The specificity remains high at 0.9861, with precision increasing to 0.6901 and the F1 score being 0.6426. This indicates that the model benefits from additional data, leading to better predictive performance. The inclusion of the R_{24} dataset further enhances the model's performance. The AUC remains high at 0.9315, the AUPR increases to 0.6780, and the F1 score is 0.6418. This demonstrates that the model's predictive power continues to improve with the addition of more datasets. Finally, adding the R_{25} dataset to the mix results in an increase in AUPR to 0.6871 and an F1 score of 0.6538. Although there is no significant improvement in AUC, the overall trend shows that incorporating more datasets generally leads to better model performance. The results from Table 2; Fig. 5 suggest that the performance of the NMTF model tends to improve with the inclusion of more datasets. This is likely due to the increased amount of information and diversity in the data, which allows the model to learn more complex patterns and make more accurate predictions. The use of multiple datasets can help in capturing a wider range of features and relationships, which is crucial for enhancing the model's predictive capabilities.

Comparison of NMF variants and IDDNMTF

Table 3; Fig. 6 illustrate the outcomes of applying the NMF algorithm to the R_{12} dataset with various values of *k*, where *k* represents the number of latent features in the model. Identifying the optimal *k* value is challenging because it significantly impacts the model's performance. For NMF (k = 10), the AUC is 0.8867, and the F1 score is 0.5089, indicating a moderate level of performance. Increasing *k* to 50 results in a slight decrease in AUC to 0.8840, but a noticeable improvement in the F1 score to 0.6165. This suggests that a higher *k* value allows the model to capture more complexity in the data, potentially enhancing certain aspects of performance.

However, further increasing *k* to 100 leads to a drop in AUC to 0.7545 and a decrease in the F1 score to 0.4826. This indicates that too many latent features might cause overfitting, where the model becomes too tailored to

Model	AUC	AUPR	Acc	Sen (Recall)	Spe	Pre	F1
FRO_MU_NMF (*k* = 10)	0.8840	0.5095	0.9486	0.5126	0.9712	0.4801	0.4966
FRO_MU_NMF (*k* = 50)	**0.8985**	**0.6039**	**0.9640**	**0.5255**	**0.9867**	**0.6723**	**0.5902**
FRO_MU_NMF (*k* = 100)	0.8814	0.5275	0.9530	0.4564	0.9788	0.5271	0.4899
FRO_MU_NMF (*k* = 141)	0.8851	0.5525	0.9546	0.5193	0.9772	0.5413	0.5306

Table 4. The results of FRO_MU_NMF algorithm on R_{12} dataset under different *k* values. Significant values are in bold.

Model	AUC	AUPR	Acc	Sen (Recall)	Spe	Pre	F1
HALS_NMF (*k* = 10)	**0.8907**	0.5379	0.9577	0.4661	0.9832	0.5894	0.5222
HALS_NMF (*k* = 50)	0.8808	**0.6065**	**0.9651**	**0.5228**	0.9880	**0.6929**	**0.5968**
HALS_NMF (*k* = 100)	0.6941	0.3438	0.9570	0.2676	**0.9927**	0.6565	0.3810
HALS_NMF (*k* = 141)	0.5556	0.1618	0.9180	0.1400	0.9583	0.1482	0.1471

Table 5. The results of HALS_NMF algorithm on R_{12} dataset under different *k* values. Significant values are in bold.

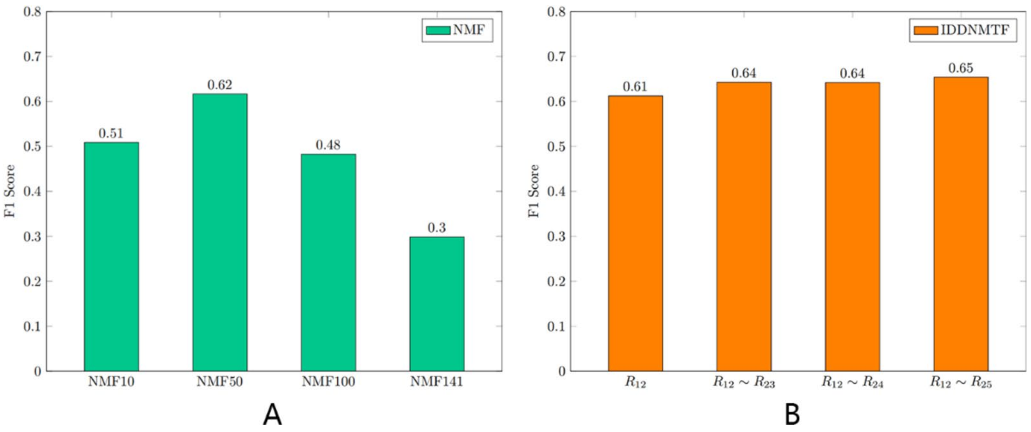


Fig. 7. The F1 scores for NMF and IDDNMTF models. (**A**) F1 scores for NMF; (**B**) F1 scores for IDDNMTF.

the training data, reducing its generalizability. The most significant drop in performance is observed with NMF (*k* = 141), where the AUC falls to 0.7523, and the F1 score drops to 0.2983. The sharp decline in performance highlights the difficulty in selecting a suitable *k* value, as it requires a balance between model complexity and generalization capability.

To demonstrate the superiority of the IDDNMTF model, we compared the results from Table 3 with those of the R_{12} model from Table 2. The results show that the IDDNMTF model outperforms the NMF model in most metrics. For instance, the R_{12} model from Table 2 has an AUC of 0.9269, which is higher than any of the AUC scores from the NMF models in Table 3. The F1 score of the R_{12} model is also higher at 0.6124 compared to the F1 scores of the NMF models, which range from 0.2983 to 0.6165. In addition, we also applied two variant methods of NMF (FRO_MU_NMF and HALS_NMF)[31] to the R12 dataset, and the results are shown in Tables 4 and 5. The results show that the F1 scores of the two methods did not exceed 0.6, and the AUC did not exceed 0.9, which reflects that IDDNMF still has significant advantages over these two variants. Figure 7 shows a more intuitive comparison of the F1 scores between the NMF and IDDNMTF models. This suggests that the IDDNMTF model is better at capturing the underlying structure of the data, leading to improved predictive performance. The superior performance of the IDDNMTF model can be attributed to its ability to factorize the data into three separate but complementary matrices, allowing for a more nuanced representation of the data. This tri-factorization approach likely provides a richer set of features that better capture the complexities of the dataset compared to the traditional two-factor NMF approach.

QuartataWeb[32] is a bioinformatics tool that utilizes machine learning to predict and analyze protein-drug and protein-chemical interactions. It helps researchers efficiently mine known interactions and uncover new ones, potentially aiding in drug repurposing and the development of novel treatments. All the prediction proteins in Table 6 are consistent with those on the QuartataWeb, and Table S1 shows more prediction results between drugs and proteins. Pancuronium is a steroidal neuromuscular blocker used to induce muscle paralysis in surgeries or mechanical ventilation by antagonizing nicotinic acetylcholine receptors (nAChRs) at the neuromuscular

DrugID	DrugName	Existing relations in R_{23}	Predicted proteins	Weight	Evidence
DB01337	Pancuronium	Q15822, P08172, P20309	P11229	1.131	Quartata
DB01338	Pipecuronium	Q15822, P08172, P20309	P11229	1.129	Quartata
DB06216	Asenapine	P14416, P28223, P08908	P35368	0.941	Quartata
DB01392	Yohimbine	P08913, P18089, P18825	P35348	0.930	Quartata
DB00546	Adinazolam	P14867, P47869, P34903	Q16445	0.914	Quartata
DB00482	Celecoxib	P35354, O15530, P00918	P23219	0.887	Quartata
DB01559	Clotiazepam	P14867, P47869, P34903	Q16445	0.879	Quartata
DB00392	Ethopropazine	P11229, Q8TCU5, P08172	P20309	0.856	Quartata
DB01628	Etoricoxib	P35354	P23219	0.847	Quartata
DB08439	Parecoxib	P35354, P02788	P23219	0.807	Quartata
DB00211	Midodrine	P35348, P35368, P25100	P07550	0.773	Quartata
DB01238	Aripiprazole	P28223, P14416, P08908	P25100	0.767	Quartata

Table 6. The predictions of drug-protein relations based on R_{23}.

DrugID	DrugName	Existing relations in R_{25}	Predicted diseases	Weight	Evidence
DB08922	Perospirone	Psychotic disorders	Schizophrenia	1.047	[42]
DB00796	Candesartan	Not Exist	Hypertension	0.881	DrugBank
DB01407	Clenbuterol	Chronic breathing disorders	Asthma	0.854	DrugBank
DB00852	Pseudoephedrine	Nasal congestion	Asthma	0.839	[43]
DB00607	Nafcillin	Arthritis	Bacterial infections	0.715	DrugBank
DB00623	Fluphenazine	Psychotic disorders	Schizophrenia	0.655	DrugBank
DB00948	Mezlocillin	Urinary tract infections	Bacterial infections	0.645	DrugBank
DB13025	Tiapride	Alcohol use disorders	Schizophrenia	0.633	[44]
DB01139	Cephapirin	Not Exist	Bacterial infections	0.585	DrugBank
DB01064	Isoproterenol	Not Exist	Asthma	0.560	[45]
DB01244	Bepridil	Chronic stable angina	Hypertension	0.550	DrugBank
DB00467	Enoxacin	Urinary tract infections	Bacterial infections	0.544	DrugBank
DB09242	Moxonidine	Alcohol use disorders	Hypertension	0.537	DrugBank
DB00231	Temazepam	Insomnia	Anxiety disorder	0.494	DrugBank
DB01595	Nitrazepam	Insomnia, Epilepsy	Anxiety disorder	0.488	DrugBank
DB01215	Estazolam	Insomnia	Anxiety disorder	0.487	[46]
DB01558	Bromazepam	Anxiety disorder, Panic attacks	Insomnia	0.479	DrugBank
DB01064	Isoproterenol	Not Exist	Cardiovascular disorder	0.454	[35]

Table 7. The predictions of drug-disease relations based on R_{25}.

junction[33]. Muscarinic acetylcholine receptors (mAChRs), members of G protein-coupled receptor (GPCR) family, mediate cellular responses like adenylate cyclase inhibition and phosphoinositide breakdown[34]. Although pancuronium mainly targets nAChRs, it can indirectly affect the broader cholinergic system, which includes mAChRs, by altering acetylcholine signaling. Understanding pancuronium's interactions with various proteins is crucial for predicting side effects and drug interactions. While DrugBank primarily associates pancuronium with nAChRs, its clinical effects may involve other cholinergic system components, emphasizing the complexity of the cholinergic system's role in drug action. Asenapine is an atypical antipsychotic that targets multiple receptors, including alpha adrenergic receptors (ADR), which are involved in physiological processes[35]. These receptors activate a phosphatidylinositol-calcium second messenger system through G proteins, with subtypes alpha-1 A adrenergic receptor (ADRA1A) and alpha-1B adrenergic receptor (ADRA1B) playing a role in cardiomyocytes signaling[36]. Asenapine's antagonism of alpha-1 ADR contributes to its therapeutic effects in conditions like schizophrenia and bipolar disorder by modulating neurotransmitter balance[37]. It also leads to side effects such as hypotension and dizziness due to its impact on blood pressure and central nervous system activity. DrugBank notes Asenapine's high affinity for these receptors, which is significant for its clinical use and side effect management.

The predictions in Table 7 highlight the potential for drug repurposing, which can accelerate the development of new therapies by leveraging existing drugs for new therapeutic applications, and Table S2 shows more prediction results between drugs and diseases. Perospirone, known for treating psychosis, has a predicted potential for schizophrenia, indicating a possible opportunity for repurposing. Given its role as a serotonin and dopamine receptor antagonist, it may offer a novel treatment approach for schizophrenia, complementing existing antipsychotic therapies[38]. Candesartan, currently not associated with any diseases in R_{25} (Drug-Disease association matrix), is predicted to be related to hypertension. As an angiotensin II receptor blocker (ARB), Candesartan is typically used for the treatment of hypertension and heart failure[39]. This prediction aligns with its known use, but further exploration might reveal additional benefits or patient populations where it is particularly effective. Clenbuterol, primarily used for chronic respiratory disorders, has a predicted disease of asthma, which is a logical extension. Clenbuterol's bronchodilatory effects may make it a candidate for repurposing in the treatment of asthma, especially for patients where current therapies are suboptimal. Temazepam, used for insomnia, has a predicted potential for anxiety disorders, which is a reasonable extension considering the drug's sedative properties[40]. Isoproterenol, a synthetic catecholamine, is also not associated with any diseases in R_{25} and used for the treatment of heart block and cardiogenic shock[41]. It is predicted to cause cardiovascular diseases, and its anticipated use in cardiovascular diseases could lead to new applications in managing heart conditions, particularly where current treatments have limitations. In addition, our model also provides predictions for drug-pathway and drug-indicator associations, as detailed in Tables S3 and S4, respectively.

Discussion

The traditional NMF model is limited by its two-factor approximation, which may not fully capture the complexity of multifaceted biological systems. NMTF has gained widespread recognition for its ability to identify latent factors within complex biological data, such as gene expression profiles, and has been extensively applied in the field of drug repositioning[30,47].

In this study, we introduce the IDDNMTF model, which integrates diverse datasets to enhance the prediction of drug repositioning candidates. The comparison between NMF and IDDNMTF reveals the latter's superiority in handling the complexity of biological data. The IDDNMTF model's tri-factorization process allows for a more granular and nuanced exploration of relationships between drugs, diseases, genes, and biological pathways. This additional dimension in factorization captures not only the interdependencies within the data but also enhances the interpretability of the results, providing a clearer understanding of the underlying mechanisms that may link a drug to a new therapeutic use. The IDDNMTF model's predictions on drug-protein interactions, as evidenced by the alignment with QuartataWeb's data, showcase its potential in identifying novel targets and pathways. This alignment is crucial as it not only validates our model's predictions but also highlights its ability to contribute to the understanding of polypharmacological treatments and drug repurposing[33,34]. Furthermore, the model's predictions on drug-disease associations provide valuable insights into potential new indications for existing drugs. For instance, the prediction of schizophrenia as a potential indication for Perospirone, a drug known for treating psychosis, underscores the model's capability to reveal opportunities for drug repositioning[38]. Similarly, the prediction of hypertension for Candesartan, a drug not previously associated with this disease in our dataset, aligns with its known function as an angiotensin II receptor blocker, indicating the model's accuracy in predicting known associations and its potential to uncover new ones[39].

Despite these advancements, the IDDNMTF model faces certain limitations. One of the primary challenges is the quality and quantity of the input data. The model is only as good as the data it is trained on, and biases or inaccuracies in the data can lead to suboptimal predictions. The risk of overfitting is another limitation, particularly when the model is trained on a limited number of examples or when the number of latent factors is too high. Overfitting can result in a model that performs well on training data but fails to generalize to new, unseen data. Addressing this issue requires careful model tuning and validation. Although the IDDNMTF model made promising predictions about potential drug repurposing candidates and novel drug-disease associations, the validity of these predictions needs to be further validated experimentally.

Conclusion

In conclusion, our model has demonstrated superior performance across a variety of evaluation metrics compared to the traditional NMF. The IDDNMTF model has showcased its ability to more accurately classify and predict drug-disease associations, with improved performance that is crucial for drug development. The capability of the IDDNMTF model to analyze complex biological data renders it a powerful tool in the realm of drug repositioning. Its multi-dimensional data analysis approach, coupled with enhanced predictive accuracy and interpretability, lays a solid foundation for advancing drug discovery and development. As we continue to refine and expand the application of the IDDNMTF model, we anticipate it will play a significant role in uncovering new therapeutic opportunities and accelerating the journey from bench to bedside.

Data availability

The datasets analysed during the current study are available in the FigShare repository, https://figshare.com/articles/dataset/_b_Enhancing_Drug_Repositioning_by_using_Non-negative_Matrix_Tri-factorization_b_/27018994.

Received: 8 September 2024; Accepted: 24 February 2025
Published online: 06 March 2025

References

1. Yu, J. L., Dai, Q. Q. & Li, G. B. Deep learning in target prediction and drug repositioning: recent advances and challenges. *Drug Discovery Today* **27**, 1796–1814. https://doi.org/10.1016/j.drudis.2021.10.010 (2022).
2. Cheng, F. et al. Prediction of drug-Target interactions and drug repositioning via Network-Based inference. *PLoS Comput. Biol.* **8**, e1002503. https://doi.org/10.1371/journal.pcbi.1002503 (2012).
3. Jarada, T. N., Rokne, J. G. & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J. Cheminform.* **12**, 46. https://doi.org/10.1186/s13321-020-00450-7 (2020).

4. Chen, H., Cheng, F., Li, J. & iDrug Integration of drug repositioning and drug-target prediction via cross-network embedding. *PLoS Comput. Biol.* **16**, e1008040. https://doi.org/10.1371/journal.pcbi.1008040 (2020).
5. Kim, Y., Jung, Y. S., Park, J. H., Kim, S. J. & Cho, Y. R. Drug-Disease association prediction using heterogeneous networks for computational drug repositioning. *Biomolecules* **12**, 1497 (2022).
6. Yang, J. et al. Computational drug repositioning based on the relationships between substructure–indication. *Brief. Bioinform.* **22**. https://doi.org/10.1093/bib/bbaa348 (2020).
7. Peng, Y. et al. Drug repositioning by prediction of drug's anatomical therapeutic chemical code via network-based inference approaches. *Brief. Bioinform.* **22**, 2058–2072. https://doi.org/10.1093/bib/bbaa027 (2020).
8. Ko, Y. Computational drug repositioning: current progress and challenges. *Appl. Sci.* **10**, 5076 (2020).
9. Sadeghi, S., Lu, J. & Ngom, A. A network-based drug repurposing method via non-negative matrix factorization. *Bioinformatics* **38**, 1369–1377. https://doi.org/10.1093/bioinformatics/btab826 (2021).
10. Santos, S. S. et al. Machine learning and network medicine approaches for drug repositioning for COVID-19. *Patterns* 3. https://doi.org/10.1016/j.patter.2021.100396 (2022).
11. Tang, X. et al. Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front. Immunol.* **11**, 603615 (2021).
12. Huang, J., Chen, J., Zhang, B., Zhu, L. & Cai, H. Evaluation of gene–drug common module identification methods using pharmacogenomics data. *Brief. Bioinform.* **22**. https://doi.org/10.1093/bib/bbaa087 (2020).
13. Xue, Y., Tong, C. S., Chen, Y. & Chen, W. S. Clustering-based initialization for non-negative matrix factorization. *Appl. Math. Comput.* **205**, 525–536. https://doi.org/10.1016/j.amc.2008.05.106 (2008).
14. Yang, M., Luo, H., Li, Y. & Wang, J. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* **35**, i455–i463 (2019).
15. Liu, S. et al. Enhancing Drug-Drug interaction prediction using deep attention neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinform* **20**, 976–985. https://doi.org/10.1109/tcbb.2022.3172421 (2023).
16. Luo, H. et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* **34**, 1904–1912. https://doi.org/10.1093/bioinformatics/bty013 (2018).
17. Zhang, W. et al. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinform.* **19**, 1–12 (2018).
18. Ding, C., Li, T., Peng, W. & Park, H. In: *Proc. 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* 126–135.
19. Gligorijević, V., Malod-Dognin, N. & Pržulj, N. In: *Biocomputing 2016: Proceedings of the Pacific Symposium.* 321–332 (World Scientific).
20. Žitnik, M. et al. Gene prioritization by compressive data fusion and Chaining. *PLoS Comput. Biol.* **11**, e1004552. https://doi.org/10.1371/journal.pcbi.1004552 (2015).
21. Dang, Q. et al. Improved computational Drug-Repositioning by Self-Paced Non-Negative matrix Tri-Factorization. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **20**, 1953–1962 (2022).
22. Ceddia, G., Pinoli, P., Ceri, S. & Masseroli, M. In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).* 1–7.
23. Ceddia, G., Pinoli, P., Ceri, S. & Masseroli, M. Matrix Factorization-based technique for drug repurposing predictions. *IEEE J. Biomed. Health Inf.* **24**, 3162–3172. https://doi.org/10.1109/jbhi.2020.2991763 (2020).
24. Wishart, D. S. et al. DrugBank: a comprehensive resource for in Silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–672. https://doi.org/10.1093/nar/gkj067 (2006).
25. Wishart, D. S. et al. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* **46**, D1074–d1082. https://doi.org/10.1093/nar/gkx1037 (2018).
26. Li, Y. H. et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* **46**, D1121–d1127. https://doi.org/10.1093/nar/gkx1076 (2018).
27. Ding, C., He, X. & Simon, H. D. In: *Proc. SIAM International Conference on Data Mining (SDM).* 606–610. (2005).
28. Dissez, G., Ceddia, G., Pinoli, P., Ceri, S. & Masseroli, M. In: *Proc. 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics.* 25–33.
29. Wild, S., Wild, W. S., Curry, J., Dougherty, A. & Betterton, M. *Seeding non-negative Matrix Factorizations With the Spherical k-means Clustering.* (University of Colorado, 2003).
30. Dissez, G., Ceddia, G., Pinoli, P., Ceri, S. & Masseroli, M. in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics.* 25–33 (Association for Computing Machinery, 2019).
31. Gillis, N. & Glineur, F. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Comput.* **24**, 1085–1105 (2012).
32. Li, H., Pei, F., Taylor, D. L. & Bahar, I. QuartataWeb: Integrated Chemical–Protein-Pathway Mapping for Polypharmacology and Chemogenomics. *Bioinformatics* 36, 3935–3937. https://doi.org/10.1093/bioinformatics/btaa210 (2020).
33. Sakata, S. & Ono, F. Allosteric Inhibition of muscle-type nicotinic acetylcholine receptors by a neuromuscular blocking agent Pancuronium. *PLoS One* **18**, e0292262. https://doi.org/10.1371/journal.pone.0292262 (2023).
34. Szczurowska, E., Szanti-Pinter, E., Randakova, A., Jakubik, J. & Kudova, E. Allosteric modulation of muscarinic receptors by cholesterol, neurosteroids and neuroactive steroids. *Int. J. Mol. Sci.* **23**. https://doi.org/10.3390/ijms232113075 (2022).
35. Balaraman, R. & Gandhi, H. Asenapine, a new Sublingual atypical antipsychotic. *J. Pharmacol. Pharmacother.* **1**, 60–61. https://doi.org/10.4103/0976-500X.64538 (2010).
36. Foldes, G. et al. Aberrant alpha-adrenergic hypertrophic response in cardiomyocytes from human induced pluripotent cells. *Stem Cell. Rep.* **3**, 905–914. https://doi.org/10.1016/j.stemcr.2014.09.002 (2014).
37. Bishara, D. & Taylor, D. Asenapine monotherapy in the acute treatment of both schizophrenia and bipolar I disorder. *Neuropsychiatr Dis. Treat.* **5**, 483–490. https://doi.org/10.2147/ndt.s5742 (2009).
38. Kishi, T. & Iwata, N. Efficacy and tolerability of Perospirone in schizophrenia: a systematic review and meta-analysis of randomized controlled trials. *CNS Drugs* **27**, 731–741. https://doi.org/10.1007/s40263-013-0085-7 (2013).
39. Grosso, A. M. et al. Comparative clinical- and cost-effectiveness of Candesartan and Losartan in the management of hypertension and heart failure: a systematic review, meta- and cost-utility analysis. *Int. J. Clin. Pract.* **65**, 253–263. https://doi.org/10.1111/j.1742-1241.2011.02633.x (2011).
40. Fluyau, D., Ponnarasu, S. & Patel, P. *StatPearls* (2024).
41. Szymanski, M. W. & Singh, D. P. *StatPearls* (2024).
42. Onrust, S. V., McClellan, K. & Perospirone *CNS Drugs* **15**, 329–337. https://doi.org/10.2165/00023210-200115040-00006 (2001).
43. Corren, J. et al. Efficacy and safety of Loratadine plus pseudoephedrine in patients with seasonal allergic rhinitis and mild asthma. *J. Allergy Clin. Immunol.* **100**, 781–788. https://doi.org/10.1016/s0091-6749(97)70274-4 (1997).
44. Karia, S., Shah, N., De Sousa, A. & Sonavane, S. Tiapride for the treatment of auditory hallucinations in schizophrenia. *Indian J. Psychol. Med.* **35**, 397–399. https://doi.org/10.4103/0253-7176.122238 (2013).
45. Katsunuma, T. et al. Efficacy and safety of isoproterenol continuous inhalation treatment for acute severe exacerbation of asthma in children; a randomized, double-blind controlled study. *Eur. Respir. J.* **46**, PA1280. https://doi.org/10.1183/13993003.congress-2015.PA1280 (2015).

46. Post, G. L. et al. Estazolam treatment of insomnia in generalized anxiety disorder: a placebo-controlled study. *J. Clin. Psychopharmacol.* **11**, 249–253 (1991).
47. Dang, Q. et al. Improved computational Drug-Repositioning by Self-Paced Non-Negative matrix Tri-Factorization. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **20**, 1953–1962. https://doi.org/10.1109/TCBB.2022.3225300 (2023).

Author contributions

Q.L. conceptualized the study and wrote the main manuscript text. Y.W. and J.W. performed methodology, data curation, and software. C.Z. supervised the project, reviewed and edited the manuscript. All authors reviewed the manuscript.

Funding

This research was funded by the Open Funds for Shaanxi Provincial Key Laboratory of Infection and Immune Diseases (No. 2023-KFMS-1).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91757-8.

Correspondence and requests for materials should be addressed to C.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.