frontiers | Frontiers in Pharmacology

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fphar.2022.872785

Drug Repositioning with GraphSAGE and Clustering Constraints Based on Drug and Disease Networks

Yuchen Zhang[1], *Xiujuan Lei*[1]*, *Yi Pan*[2] *and Fang-Xiang Wu*[3]

[1]School of Computer Science, Shaanxi Normal University, Xi'an, China, [2]Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, [3]Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada

The understanding of therapeutic properties is important in drug repositioning and drug discovery. However, chemical or clinical trials are expensive and inefficient to characterize the therapeutic properties of drugs. Recently, artificial intelligence (AI)-assisted algorithms have received extensive attention for discovering the potential therapeutic properties of drugs and speeding up drug development. In this study, we propose a new method based on GraphSAGE and clustering constraints (DRGCC) to investigate the potential therapeutic properties of drugs for drug repositioning. First, the drug structure features and disease symptom features are extracted. Second, the drug–drug interaction network and disease similarity network are constructed according to the drug–gene and disease–gene relationships. Matrix factorization is adopted to extract the clustering features of networks. Then, all the features are fed to the GraphSAGE to predict new associations between existing drugs and diseases. Benchmark comparisons on two different datasets show that our method has reliable predictive performance and outperforms other six competing. We have also conducted case studies on existing drugs and diseases and aimed to predict drugs that may be effective for the novel coronavirus disease 2019 (COVID-19). Among the predicted anti-COVID-19 drug candidates, some drugs are being clinically studied by pharmacologists, and their binding sites to COVID-19-related protein receptors have been found *via* the molecular docking technology.

Keywords: drug reposition, graph neural network, GraphSAGE, matrix factorization, clustering constraint, COVID-19

OPEN ACCESS

Edited by:
Sajjad Gharaghani,
University of Tehran, Iran

Reviewed by:
Marcus Scotti,
Federal University of Paraíba, Brazil
Fatemeh Zare-Mirakabad,
Amirkabir University of Technology, Iran

****Correspondence:***
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Experimental Pharmacology and Drug Discovery,
a section of the journal Frontiers in Pharmacology

Received: *10 February 2022*
Accepted: *11 April 2022*
Published: *10 May 2022*

Citation:
Zhang Y, Lei X, Pan Y and Wu F-X (2022) Drug Repositioning with GraphSAGE and Clustering Constraints Based on Drug and Disease Networks.
Front. Pharmacol. 13:872785.
doi: 10.3389/fphar.2022.872785

INTRODUCTION

Traditional drug discovery is often based on a specific disease. It generally has a number of stages, including target discovery, target validation, lead compound identification, lead optimization, preclinical drug development, advancing to clinical trials, and clinical trials. Typically, the development of an effective drug takes an average of 15 years and costs 800 million to 1.5 billion US dollars (Dudley et al., 2011) (Yu et al., 2015). However, the success rate is often not high due to the lack of systematic evaluation of other indications that drugs can treat, as well as the impact of our life, disease development, and market factors. These difficulties have caused pharmaceutical companies very worrisome when developing new drugs, and the development speed is slow (Booth and Zemmel, 2004).

From cheminformatics and life sciences (Bader et al., 2008), it is well acknowledged that one drug may work on multiple target proteins, and one target protein is related to multiple diseases, which is the basis of drug repositioning. Actually, drug repositioning brings significant benefits to drug research and related pharmaceutical companies. For example, minoxidil (Varothai and Bergfeld, 2014), a drug originally used to relieve hypertension and excessive tension was later found to effectively treat symptoms such as hair loss. Antifungal and antitumor drug itraconazole (ITZ) can act as a broad-spectrum enterovirus inhibitor (Strating et al., 2015). However, this kind of drug repositioning is mostly based on clinical accidental discoveries and the experience of pharmacists, and it is difficult for large-scale investigation.

With the development of cross-technology, more and more researchers tend to use computational technologies to predict new indications of existing drugs. These methods mainly include network propagation, low-rank matrix approximation, and graph neural network. Based on biological networks, similarity measures and bi-random walk were proposed for drug repositioning (Luo et al., 2016). Yu *et al.* combined miRNAs and group specificity to predict potential therapeutic drugs for breast cancer (Yu et al., 2018). A genome-wide positioning systems network algorithm was developed for drug repurposing (Cheng et al., 2019). Fiscon *et al.* presented a new network-based algorithm SAveRUNNER and applied it to COVID-19 (Fiscon et al., 2021). However, due to the complexity and noise of interactions between organisms, the prediction accuracy based on those existing methods cannot meet the requirements. Some methods were developed based on low-rank matrix approximation. Luo *et al.* proposed a drug repositioning recommendation system (DRRS) to predict novel drug indications based on low-rank matrix approximation and randomized algorithms (Luo et al., 2018). Wang *et al.* proposed a projection onto convex sets (Wang et al., 2019) to relocate the functions of drugs. Weight graph regularized matrix factorization was also used in drug response prediction (Guan et al., 2019). Wu *et al.* used meta paths and singular value decomposition to predict drug–disease associations (Wu et al., 2019). Yang *et al.* used a bounded nuclear norm regularization (BNNR) method to complete the drug–disease matrix (Yang et al., 2019). An improved drug repositioning approach using Bayesian inductive matrix completion also was proposed (Zhang W. et al., 2020). Meng *et al.* used the similarity-constrained probabilistic matrix factorization for drug repositioning and applied it to COVID-19 (Meng et al., 2021). However, these matrix-based methods did not take the biochemical properties of drugs and diseases into consideration.

With the widespread application of artificial intelligence technology, more and more machine learning and deep learning methods are also applied to drug development and other fields of bioinformatics. Regularized kernel classifier was proposed to predict new drug–disease associations (Lu and Yu, 2018). Madhukar *et al.* used a Bayesian machine learning approach to identify drug targets with diverse data types (Madhukar et al., 2019). Huang *et al.* proposed a network embedding-based method CMFMTL for predicting drug–disease associations. CMFMTL handled the problem as multi-task learning where each task is to predict one type of association, and two tasks complement and improve each other by capturing the relatedness between them (Huang et al., 2020). Zhu *et al.* constructed a drug knowledge graph for drug repurposingand transformed information in the drug knowledge graph into valuable inputs to allow machine learning models to predict drug repurposing candidates (Zhu et al., 2020). Zeng *et al.* developed a network-based deep learning approach, termed deepDR (Zeng et al., 2019), for *in silico* drug repurposing. Li *et al.* used molecular structures and clinical symptoms *via* a deep convolutional neural network to identify drug–disease associations (Li Z et al., 2019). A network embedding method called NEDD (Zhou et al., 2020) was proposed to predict novel associations between drugs and diseases using meta paths of different lengths.

Graph convolutional network (GCN) methods have also been further used in the field of medicine. A layer attention graph convolutional network (LAGCN) (Yu et al., 2020) was also used by fusing heterogeneous information to the GCN. They introduced a layer attention mechanism to combine embeddings from multiple graph convolution layers for further improving the prediction performance (Cai et al., 2021). Wang *et al.* also proposed a global graph feature learning method to predict associations (Wang et al., 2022). Meta path-based methods such as metapath2vec and meta-structure have also been developed (Zhang Y. et al., 2020; Lei et al., 2021). Algorithms based on graph neural networks (GNNs) or graph embeddings consider both biochemical characteristics and network interactions, but they often have high time complexity and do not consider the characteristic of drug clusters or combination drugs. At the same time, when extracting features of drug–disease associations, a large number of methods only directly connect drug features and disease features without considering the influence of different features. The feature representation of association needs to be improved.

While existing methods cannot accurately predict the potential drug–disease associations, and the network is often unchangeable after model training, we proposed a drug repositioning method DRGCC based on network clustering constraints and GraphSAGE. First, we extracted the molecular structure features of drugs and the symptom features of diseases as the biological attribute features. After that, we used the associations between drugs and genes, as well as the relationships between diseases and genes, to reconstruct a drug–drug interaction network and establish a disease similarity network. The third step was to use a clustering algorithm to divide the two networks into some clusters. The network clustering features of drugs and diseases were obtained by matrix factorization with the divided cluster set as a condition constraint, respectively. Finally, we built two GraphSAGE models based on drug and disease networks and fed the attributes and clustering features of drugs and diseases to the two models, respectively, to obtain the potential treatment probability of the existing drugs for the diseases. The method was applied to the prediction of anti-COVID-19 drugs, and some case studies

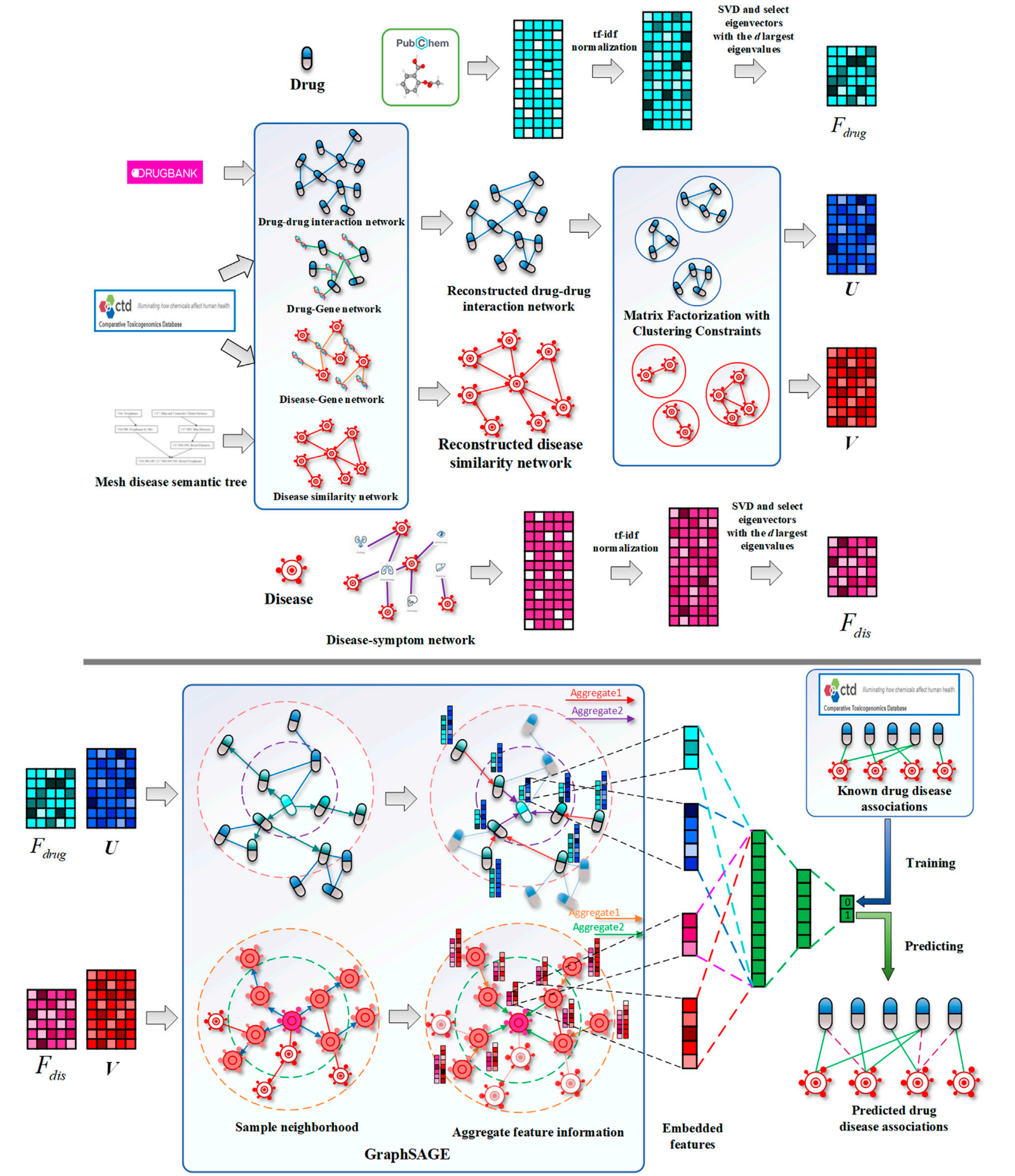

Drug
PubChem
tf-idf normalization
SVD and select eigenvectors with the d largest eigenvalues
F_{drug}
DRUGBANK
Drug-drug interaction network
Drug-Gene network
Disease-Gene network
Disease similarity network
ctd Illuminating how chemicals affect human health. Comparative Toxicogenomics Database
Mesh disease semantic tree
Reconstructed drug-drug interaction network
Reconstructed disease similarity network
Matrix Factorization with Clustering Constraints
U
V
Disease
Disease-symptom network
tf-idf normalization
SVD and select eigenvectors with the d largest eigenvalues
F_{dis}
F_{drug}
U
F_{dis}
V
Aggregate1
Aggregate2
Aggregate1
Aggregate2
Sample neighborhood
Aggregate feature information
GraphSAGE
Embedded features
Known drug disease associations
Training
0
1
Predicting
Predicted drug disease associations

FIGURE 1 | Schematic diagrams of data processing and DRGCC model. 1) Reconstruct the drug interaction network and establish the disease similarity network, 2) extract the attributes feature of drugs and diseases, 3) obtain network clustering features of drugs and diseases, and 4) construct the GraphSAGE prediction model to predict potential drug–disease relationships.

TABLE 1 | Statistics of pre-processed CTD and HDVD database.

Dataset	Drugs	Diseases/viruses	Known associations	Density
CTD	780	717	17594	0.0315
HDVD	219	34	455	0.0611

were conducted. The framework of the method DRGCC is shown in **Figure 1**. The main contributions of this work are summarized as the following two points: 1) DRGCC integrates the clustering features of networks, which can effectively improve the prediction accuracy of drug–disease associations. 2) DRGCC can embed new nodes in the existing network and predict their associations. In addition, DRGCC is complementary to existing experimental methods to enable rapid and accurate discovery of drug candidates for anti-COVID-19 and other emerging viral infectious diseases.

MATERIALS AND METHOD

In this section, we introduce the database used in the study and how they were processed. The known associations of drug and disease were obtained. The drug–drug interaction network was reconstructed. The disease similarity network was calculated. Their attribute features and network features were also extracted. The purpose of our study is to predict potential associations from known drug–disease associations, which can be formulated as a classification problem. Therefore, we developed a GNN model based on GraphSAGE, which takes the obtained drug and disease attribute features and clustering features as input, and outputs the possibility of potential relationships between them.

Known Associations of Drugs and Diseases

Known drug and disease relationship data can be obtained from the Comparative Toxicogenomics Database (CTD) (Davis et al., 2021). CTD is a publicly available database that aims to advance understanding of how environmental exposures affect human health. It provides manually curated information about drug compound–gene/protein interactions, drug compound–disease, and gene–disease relationships. We first screened 36,392 drug–disease associations marked with therapeutic relationships in CTD (version 2021.2.26). They corresponded to 6,699 drugs and 2,472 diseases. In order to make it more focused and easier to verify the method later, we extracted drugs with more than 10 disease treatment effects and diseases that are affected by more than 10 drugs. We made the corresponding PubChem Compound ID (CID) and PubChem Substance ID (SID) (Kim et al., 2021) for each drug compound. In the end, we extracted 780 drugs, 717 diseases, and 17,594 therapeutic associations. The known drug–disease association matrix is marked as Y, if drug i has a therapeutic effect on disease j, then $Y_{ij} = 1$; otherwise, it is 0. In addition, we also considered the relational database of viruses and drugs, HDVD (Meng et al., 2021), which includes 34 viruses, 219 drugs, and 455 human drug–virus interactions. In the HDVD database, SARS-CoV-2, which has recently attracted much attention, is included. The statistics of the two datasets are shown in **Table 1**.

Reconstruction of Drug–Drug Interaction Network

In daily life, we have known for a long time that there are interactions between drugs and drugs. Some combinations of drugs can promote the cure of diseases. The interactions between drugs can also provide the basis for feature extraction and fusion of drugs. DrugBank (Wishart et al., 2018) provides us with a large number of drug–drug interactions (DDIs). We found 2,669,764 interactions in the database. We denote the drug–drug interaction matrix by M_{DDI}. Due to the non-correspondence of IDs, only 489 of the 780 drugs were mapped to DrugBank. There are 56,439 interactions among 489 drugs. Therefore, we aimed to use other biological properties of drugs to infer possible associations between drugs. The clinical relevance of drug–drug interactions also depends on the patient's genetic profile. Drug–drug–gene and drug–gene–gene interactions affect the therapeutic properties of drugs (Hahn and Roll, 2021). A method for calculating drug similarity using drug–gene associations was proposed by Groza et al. (2021). Inspired by these studies, we aimed to use the drug–gene relationship to complement the existing drug interactions. The CTD also provides the relationships between drug compounds and genes. We obtained 383,525 drug–gene relationships from it. They covered 768 drugs and 34,184 genes. We denote the drug–gene association matrix by $M_{drug-gene}$; if drug i has an association with gene j, then $M_{drug-gene_{ij}} = 1$; otherwise, it is 0. The reconstructed drug–drug interaction (RDDI) matrix M_{RDDI} is calculated as follows:

$$M_{RDDI_{ij}} = \begin{cases} M_{DDI_{ij}} & if\ M_{DDI_{ij}} \neq 0, \\ \dfrac{\left|M_{drug-gene_{i,}} \cap M_{drug-gene_{j,}}\right|}{\left|M_{drug-gene_{i,}} \cup M_{drug-gene_{j,}}\right|} & if\ M_{DDI_{ij}} = 0. \end{cases} \tag{1}$$

These associated genes often encode target proteins, and thus, we considered the relationship between drugs and target proteins, making the drug interaction network more complete.

Construction of Disease Similarity Network

There are also similarities between diseases, and a large number of calculation methods for disease similarity have been developed in the literature. In studying the relationship between miRNAs and diseases, Cui *et al.* successively developed two versions of the method (Wang et al., 2010) (Li J. et al., 2019), both of which applied disease semantic similarity. All the denominations of diseases were in accordance with the MeSH (Yu, 2018) database (https://www.nlm.nih.gov/mesh/meshhome.html). Finally, we obtained the semantic similarity matrix M_{DS} of diseases according to the method of Wang et al. (2010). Different from the method in Disease Ontology (Schriml et al., 2019) that only builds an overall semantic tree, MESH divides diseases into 17 subcategories or sub-trees, so there are null values in the calculated disease similarity for some different subcategory diseases. Previous work has shown elucidating disease and gene associations (Li et al., 2021). Similar to reconstructing the M_{RDDI}, we use disease–gene relationship to reconstruct the disease similarity network. The CTD contains 13,775,363 disease–gene relationships, which cover 715 diseases and 50,827 genes. The disease–gene association matrix is denoted by $M_{dis-gene}$. If disease i is related to gene j, then $M_{dis-gene_{ij}} = 1$; otherwise, it is 0. The reconstructed disease similarity matrix M_{RDS} is calculated as follows:

$$M_{RDS_{ij}} = \begin{cases} M_{DS_{ij}} & if\ M_{DS_{ij}} \neq 0, \\ \sum_{k=1}^{N_{gene}} \dfrac{M_{dis-gene_{ik}} M^{T}_{dis-gene_{kj}}}{\sum_{p=1}^{N_{gene}} M_{dis-gene_{ip}} \sum_{q=1}^{N_{gene}} M^{T}_{dis-gene_{qj}}} & if\ M_{DS_{ij}} = 0, \end{cases} \tag{2}$$

where N_{gene} is the number of all genes.

Processing of Attribute Features

The attribute features of drugs can be described by their structures. The PubChem system generates a binary substructure fingerprint for chemical structures. These fingerprints are used by PubChem for similarity neighboring and similarity searching (Kim et al., 2021). The structure of a drug can be described by 881 substructures, and a substructure is a fragment of a chemical structure. The fingerprint is an ordered list of binary bits (0/1). A Boolean value for each bit determines or tests the presence of a chemical structure. Binary data are stored in one-byte increments. Therefore, the length of the fingerprint is 111 bytes (888 bits), which include padding 7 bits at the end to complete the last byte. The four-byte prefix including the fingerprint bit length (881 bits) increases the size of the stored PubChem fingerprint to 115 bytes (920 bits). To learn embeddings of drugs, we also used latent semantic analysis (Deerwester et al., 1990). Let N_{sub} denote the number of substructures generated from all drugs. We employ a matrix $M_{drug-sub} \in R^{N_{drug} \times N_{sub}}$, and $M_{drug-sub}$ is defined as follows:

$$M_{drug-sub_{ij}} = tf(i,j) \cdot idf\left(N_{drug}, j\right), \tag{3}$$

where $tf(i,j)$ stands for the strength of the i-th drug having j-th substructure. If substructure j appears in drug i, then $tf(i,j) = 1/N_{sub_i}$; otherwise, it is 0. N_{sub_i} is the number of substructures in drug i.

$$idf\left(N_{drug}, j\right) = log\frac{N_{drug}}{\left|\{i \in drug: tf(i,j) \neq 0\}\right|}. \tag{4}$$

$idf\ (N_{drug}, j)$ results in lower weights for more common substructures and higher weights for less common substructures. This is consistent with an observation in the information theory that rarer events generally have higher entropy and are thus more informative. Then, the matrix $M_{drug-sub}$ was decomposed by singular value decomposition (SVD) into three matrices R, Σ, and Q, such that $M_{drug-sub} = R\Sigma Q$. $\Sigma \in \mathbb{R}^{N_{drug} \times N_{sub}}$ is a diagonal matrix with the eigenvalues of $M_{drug-sub}$, and R is an $N_{drug} \times N_{drug}$ matrix in which each column is an eigenvector $R_{.j}$ of $M_{drug-sub}$ corresponding to the eigenvalue Σ_{jj}. Afterward, in order to embed the features into the low-dimensional space $\mathbb{R}^{d_{drug}}$, we extracted the feature vectors corresponding to the top d_{drug} largest singular values to form a new drug attribute feature matrix F_{drug}.

Similar to drug attribute feature extraction, disease attribute features are also extracted. Diseases are often accompanied by a large number of symptoms when they occur. Zhou *et al.* established a disease–symptom network when studying the commonalities between diseases (Zhou et al., 2014). They gave 322 common symptoms for each disease, established a disease–symptom relationship matrix, and also used the term frequency-inverse document frequency method to weight. After that, we also used the SVD method to obtain a disease feature matrix F_{dis} in $\mathbb{R}^{d_{dis}}$ space. The feature vectors corresponding to the top d_{dis} largest singular values form the disease attribute feature matrix F_{dis}.

Extraction of Network Clustering Feature

In the previous section, we have obtained attribute features of drugs and diseases. However, the network features between drugs and diseases were not involved. On the other hand, numerous studies have confirmed the modularity that exists between biomolecules (Ni et al., 2020) (Groza et al., 2021). Matrix factorization, as a commonly used low-rank matrix approximation method, can achieve the goal by adding expectation constraints. Therefore, we aimed to use the matrix factorization method to measure the features of the relationship between drugs and diseases and consider the modularity of drugs and diseases. Two constraints were added to matrix factorization, one is sparsity and the other is clustering constraints. For sparsity, it is desirable to obtain a basis matrix with fewer parameters and be able to restore the original associations. It can be written as follows:

$$min\, J(U,V) = \min_{U,V}\left\{\frac{1-2\alpha}{2}\|P \odot (Y - UV)\|_F^2 + \frac{\alpha}{2}\|U\|_F^2 + \frac{\alpha}{2}\|V\|_F^2\right\}, \tag{5}$$

where $U \in R^{N_{drug} \times k}$, $V \in R^{k \times N_{dis}}$ are the feature matrices of drugs and diseases, k can be used as the embedded feature dimension, and P is the observation matrix. In this matrix, the elements corresponding to positive and negative samples are marked as 1, and the other elements are 0. $\odot$ is the Hadamard product. For clustering attributes, we first need to cluster nodes in the drug network and disease network. MCODE (Bader and Hogue, 2003) is a very mature network clustering method, which has been widely used in a variety of network analyses. We used it to cluster the reconstructed drug–drug interaction network and disease similarity network. When extracting features for drug and disease networks, the embedded features should satisfy the property that drugs or diseases of different clusters have greater distinguishability. Using Euclidean distance as the measure function of similarity between features, the matrix factorization subject to clustering constraints can be written as follows:

$$\begin{aligned} min\, J(U,V) = \min_{U,V} \Bigg\{ & \frac{1-2\alpha-2\beta}{2} \|P \odot (Y-UV)\|_F^2 + \frac{\alpha}{2}\|U\|_F^2 \\ & - \frac{\beta}{2}\sum_{i=1}^{c_{\mathrm{drug}}} \left\| \bar{U}^{(i)} - \bar{U}_{all} \right\|_2^2 + \frac{\alpha}{2}\|V\|_F^2 \\ & - \frac{\beta}{2}\sum_{i=1}^{c_{\mathrm{dis}}} \left\| \bar{V}^{(i)} - \bar{V}_{all} \right\|_2^2 \Bigg\}, \end{aligned} \tag{6}$$

where c_{drug} and $c_{disease}$ are the cluster number of drugs and diseases, respectively; $\bar{U}^{(i)}$ ($\bar{V}^{(i)}$) denotes the average vector of the drug (disease) feature vectors in the *i-th* cluster; $\bar{U}_{all}$, ($\bar{V}_{all}$) is the average vector of all drug (disease) feature vectors; and α and β are control parameters. We set s_i ($s_i^{'}$) to the node number of *i-th* drug (disease) cluster, $N_{drug} = s_1 + s_2 + \ldots + s_{c_{drug}}$, and $N_{dis} = s_1^{'} + s_2^{'} + \ldots + s_{c_{dis}}^{'}$To facilitate the solution, let $A_{drug}^{(i)} = [\frac{1}{s_i}, \frac{1}{s_i}, \ldots, \frac{1}{s_i}]_{1\times s_i}$ and $A_{dis}^{(i)} = [\frac{1}{s_i^{'}}, \frac{1}{s_i^{'}}, \ldots, \frac{1}{s_i^{'}}]^T_{s_i^{'}\times 1}$, so the average of the feature values of *i-th* cluster samples can be calculated as follows:

$$\begin{aligned} \bar{U}^{(i)} &= A_{drug}^{(i)}\left[U^{(i)}(1), U^{(i)}(2), \cdots, U^{(i)}(s_i)\right]^T \\ \bar{V}^{(i)} &= \left[V^{(i)}(1), V^{(i)}(2), \cdots, V^{(i)}\left(s_i^{'}\right)\right] A_{dis}^{(i)}, \end{aligned} \tag{7}$$

where $U^{(i)}(x)\,(V^{(i)}(x))$ is the *x-th* feature vector of *i-th* drug (disease) cluster. The matrix formed by the average vector of all clusters can be represented by

$$\begin{aligned} \bar{U} &= \left[\bar{U}^{(1)}, \bar{U}^{(2)}, \ldots, \bar{U}^{(c_{drug})}\right]^T = A_{z_{drug}} U \\ \bar{V} &= \left[\bar{V}^{(1)}, \bar{V}^{(2)}, \ldots, \bar{V}^{(c_{dis})}\right] = V A_{z_{dis}}, \end{aligned} \tag{8}$$

where

$$A_{z_{drug}} = \begin{bmatrix} A_{drug}^{(1)} & & & \\ & A_{drug}^{(2)} & & \\ & & \ddots & \\ & & & A_{drug}^{(c_{\mathrm{drug}})} \end{bmatrix}_{c_{\mathrm{drug}}\times N_{\mathrm{drug}}}$$

$$A_{z_{dis}} = \begin{bmatrix} A_{dis}^{(1)} & & & \\ & A_{dis}^{(2)} & & \\ & & \ddots & \\ & & & A_{dis}^{(c_{dis})} \end{bmatrix}_{N_{dis}\times c_{dis}} . \tag{9}$$

Then, we defined matrices B_{drug} and $B_{disease}$ as follows:

$$B_{drug} = \begin{bmatrix} 1/N_{drug} & 1/N_{drug} & \cdots & 1/N_{drug} \\ 1/N_{drug} & 1/N_{drug} & \cdots & 1/N_{drug} \\ \vdots & \vdots & \ddots & \vdots \\ 1/N_{drug} & 1/N_{drug} & \cdots & 1/N_{drug} \end{bmatrix}_{c_{drug}\times N_{drug}}$$

$$B_{dis} = \begin{bmatrix} 1/N_{dis} & 1/N_{dis} & \cdots & 1/N_{dis} \\ 1/N_{dis} & 1/N_{dis} & \cdots & 1/N_{dis} \\ \vdots & \vdots & \ddots & \vdots \\ 1/N_{dis} & 1/N_{dis} & \cdots & 1/N_{dis} \end{bmatrix}_{N_{dis}\times c_{dis}} . \tag{10}$$

$\bar{U}_{all}$, ($\bar{V}_{all}$) can be written in the following matrix form:

$$\begin{aligned} \left[\bar{U}_{all}, \bar{U}_{all}, \ldots, \bar{U}_{all}\right]^T c_{drug} \times k &= B_{drug} U \\ \left[\bar{V}_{all}, \bar{V}_{all}, \ldots, \bar{V}_{all}\right]_{k\times c_{dis}} &= V B_{dis}. \end{aligned} \tag{11}$$

Therefore, the constraint term of clustering can be expressed by **formula (12)**:

$$\begin{aligned} \sum_{i=1}^{c_{drug}} \left\|\bar{U}^{(i)} - \bar{U}_{all}\right\|_2^2 &= tr\Big(\left(A_{z_{drug}}U - B_{drug}U\right)\left(A_{z_{drug}}U - B_{drug}U\right)^T\Big) \\ \sum_{i=1}^{c_{dis}} \left\|\bar{V}^{(i)} - \bar{V}_{all}\right\|_2^2 &= tr\Big(\left(VA_{z_{dis}} - VB_{dis}\right)^T \left(VA_{z_{dis}} - VB_{dis}\right)\Big). \end{aligned} \tag{12}$$

As a result, the constraint matrix factorization in **formula (6)** has been transformed into

$$\begin{aligned} J(U,V) = & \frac{1-2\alpha-2\beta}{2} tr\left(\left(P^T \odot Y^T\right)(P \odot Y)\right) \\ & - (1-2\alpha-2\beta) tr\left((P \odot (UV))\left(P^T \odot Y^T\right)\right) \\ & + \frac{1-2\alpha-2\beta}{2} tr\left(P \odot (UV)\left(P^T \odot \left(V^T U^T\right)\right)\right) \\ & + \frac{\alpha}{2} tr\left(UU^T\right) + \frac{\alpha}{2} tr\left(VV^T\right) - \frac{\beta}{2} tr\Big(A_{z_{drug}} U U^T A_{z_{drug}}^T\Big) \\ & + \beta tr\Big(A_{z_{drug}} U U^T B_{drug}^T\Big) - \frac{\beta}{2} tr\Big(B_{drug} U U^T B_{drug}^T\Big) \\ & - \frac{\beta}{2} tr\Big(A_{z_{dis}}^T V^T V A_{z_{dis}}\Big) + \beta tr\Big(A_{z_{dis}}^T V^T V B_{dis}\Big) \\ & - \frac{\beta}{2} tr\left(B_{dis}^T V^T V B_{dis}\right). \end{aligned} \tag{13}$$

The partial derivatives of $J(U, V)$ with respect to U and V are calculated as follows:

$$\begin{aligned} \frac{\partial J(U,V)}{\partial(U)} = & -(1-2\alpha-2\beta)(P \odot Y)V^T + (1-2\alpha-2\beta)(P \odot (UV))V^T + \alpha U \\ & - \beta A_{z_{drug}}^T A_{z_{drug}} U + \beta B_{drug}^T A_{z_{drug}} U - \beta B_{drug}^T B_{drug} U + \beta A_{z_{drug}}^T B_{drug} U, \\ \frac{\partial J(U,V)}{\partial(V)} = & -(1-2\alpha-2\beta)U^T(P \odot Y) + (1-2\alpha-2\beta)U^T(P \odot (UV)) + \alpha V \\ & - \beta V A_{z_{dis}} A_{z_{dis}}^T + \beta V B_{dis} A_{z_{dis}}^T + \beta V A_{z_{dis}} B_{dis}^T - \beta V B_{dis} B_{dis}^T. \end{aligned} \tag{14}$$

After the initial U and V are randomly given, solution is solved as per the following iterative rules until the stopping condition is met. Drug network clustering feature U and disease network clustering feature V are obtained.

$$U_{ij} \leftarrow U_{ij}\frac{\Big((1-2\alpha-2\beta)(P\odot Y)V^{T}+\beta A_{z_{drug}}^{T}A_{z_{drug}}U+\beta B_{drug}^{T}B_{drug}U\Big)_{ij}}{\Big((1-2\alpha-2\beta)(P\odot(UV))V^{T}+\alpha U+\beta\Big(B_{drug}^{T}A_{z_{drug}}+A_{z_{drug}}^{T}B_{drug}\Big)U\Big)_{ij}},$$
$$V_{ij} \leftarrow V_{ij}\frac{\big((1-2\alpha-2\beta)U^{T}(P\odot Y)+\beta VA_{z_{dis}}A_{z_{dis}}^{T}+\beta VB_{dis}B_{dis}^{T}\big)_{ij}}{\big((1-2\alpha-2\beta)U^{T}(P\odot(UV))+\alpha V+\beta V\big(B_{dis}A_{z_{dis}}^{T}+A_{z_{dis}}B_{dis}^{T}\big)\big)_{ij}}. \tag{15}$$

Drug Repositioning Using GraphSAGE

GraphSAGE (SAmple and aggreGatE) (Hamilton et al., 2017) is a new graph convolutional neural (GCN) (Defferrard et al., 2016) model proposed, which has two improvements to the original GCN. On the one hand, it used the strategy of sampling neighbors to transform the GCN from a full graph training method to a node-centric small batch training method, which made large-scale data distributed training possible. On the other hand, the algorithm extended the operation of aggregating neighbors. In this study, we used the GraphSAGE model for the drug–drug interaction network and disease similarity network, respectively, to obtain their low dimensional embedding vectors and make predictions through a simple neural network. The feature x of each node v in these networks is marked as $x_v, v\mathcal{B}$, where $\mathcal{B}$ denotes a batch sample set. In each iteration, only the nodes in the batch set are trained. Assuming that the model has L layers when sampling the nodes in the batch set, a top–down sampling method is adopted. It collects n_k nodes from each layer at a time. Neighborhood sampling functions $\mathcal{H}_l$ of the l-th layer are defined by sampling the n_k most similar neighbors of the source node $\mathcal{B}$. $\mathcal{H}_l\ (v)$ represents the sampling set of nodes around the node v of the l-th layer. The sampling process is from $\mathcal{B}^L$ to $\mathcal{B}^0$ shown in the sampling section of Algorithm 1. Then we extract the feature h_u^0 of each node u in the $\mathcal{B}^0$ set as training features. First, each node v aggregates the representations of the nodes in its sampling neighborhood, $\{h_u^{l-1}, u \in \mathcal{H}_l(v)\}$ into a single vector $\mathcal{H}_{\mathcal{H}_l(v)}^l$. After aggregating the neighboring feature vectors, GraphSAGE concatenates the node's current representation, h_v^{l-1}, with the aggregated neighborhood vector, $\mathcal{H}_{\mathcal{H}_l(v)}^l$, and this concatenated vector is fed to a fully connected layer with a nonlinear activation function σ, which transforms the representations to be used at the next step of the algorithm for h_v^l. The embedding generation of a given drug node is shown in the embedding section of Algorithm 1. The different aggregator functions can be used in the aggregation steps:

Mean aggregator:

$$\begin{aligned} h_{\mathcal{H}_l(v)}^l &\leftarrow mean\left(\left\{h_u^{l-1},\ \forall u \in \mathcal{H}_l(v)\right\}\right) \\ h_v^l &\leftarrow \sigma\left(W^l concat\left(h_v^{l-1}, h_{\mathcal{H}_l(v)}^l\right) + b^l\right). \end{aligned} \tag{16}$$

MeanPool aggregator:

$$\begin{aligned} h_{\mathcal{H}_l(v)}^l &\leftarrow mean\left(\left\{\sigma\left(W^l h_u^{l-1} + b\right), \forall u \in \mathcal{H}_l(v)\right\}\right) \\ h_v^l &\leftarrow \sigma\left(W^l concat\left(h_v^{l-1}, h_{\mathcal{H}_l(v)}^l\right) + b^l\right). \end{aligned} \tag{17}$$

MaxPool aggregator:

$$\begin{aligned} h_{\mathcal{H}_l(v)}^l &\leftarrow max\left(\left\{\sigma\left(W^l h_u^{l-1} + b\right), \forall u \in \mathcal{H}_l(v)\right\}\right) \\ h_v^l &\leftarrow \sigma\left(W^l concat\left(h_v^{l-1}, h_{\mathcal{H}_l(v)}^l\right) + b^l\right). \end{aligned} \tag{18}$$

GCN aggregator:

$$h_v^l \leftarrow \sigma\left(W^l mean\left(\left\{h_v^{l-1}\right\} \cup \left\{h_u^{l-1},\ \forall u \in \mathcal{H}_l(v)\right\}\right) + b^l\right). \tag{19}$$

LSTM aggregator:

$$\begin{aligned} h_{\mathcal{H}_l(v)}^l &\leftarrow LSTM\left(random\_order\left\{h_u^{l-1},\ \forall u \in \mathcal{H}_l(v)\right\}\right) \\ h_v^l &\leftarrow \sigma\left(W^l concat\left(h_v^{l-1}, h_{\mathcal{H}_l(v)}^l\right) + b^l\right), \end{aligned} \tag{20}$$

where W^l and b^l are parameter matrix and bias of the l-th layer, respectively. The final model outputs a low dimensional embedding vector z_v of node v. Since **formula (19)** is a linear approximation of local spectral convolution, it is called a GCN aggregator. It is important to note that LSTM is not inherently symmetric because it processes inputs in a sequential manner. GraphSAGE adopts LSTM to operate on an unordered set by simply applying the LSTM to a random permutation. Unlike GCN, GraphSAGE can perform batch sampling and save the required neighbor features before the node feature aggregation operation. After training, GraphSAGE can perform feature embedding for newly added network nodes. In this way, the network model is actually formed into a subnetwork model according to the sampled nodes, which can increase the learning speed of the model and is suitable for processing larger samples. In this study, the relationship prediction of two types of nodes is involved, and the number of samples is $N_{drug} \times N_{dis}$, which is very large, so GraphSAGE has better performance. The GraphSAGE minibatch forward propagation is described in Algorithm 1.

Algorithm 1. GraphSAGE minibatch forward propagation in drug–drug interaction or disease similarity network.

```
Input: Graph $G_{RDDI}$ or $G_{RDS}$;
    node features $\{x_v, \forall v \in \mathcal{B}\}$;
    Layer $L$; weight matrices $W^l, \forall l \in \{1,2,\ldots,L\}$;
    neighborhood sampling functions $\mathcal{H}_l, \forall l \in \{1,2,\ldots,L\}$
Output: Vector embed representations $z_v$ for all $v \in \mathcal{B}$
Sampling:
  $\mathcal{B}^L \leftarrow \mathcal{B}$
  For $l$=$L$, … ,1 do
    $\mathcal{B}^{l-1} \leftarrow \mathcal{B}^l$;
    For $u \in \mathcal{B}^l$ do
      $\mathcal{B}^{l-1} \leftarrow \mathcal{B}^{l-1} \cup \mathcal{H}_l(u)$;
    End
  End
Embedding generation:
  $h_u^0 \leftarrow x_u, \forall u \in \mathcal{B}^0$;
  For $l$=1, …, $L$ do
    For $u \in \mathcal{B}^l$ do
      $h_u^l \leftarrow aggregator(h_u^{l-1}, \{h_{u'}^{l-1}, u' \in \mathcal{H}_l(u)\})$ via eq. (16)-eq. (20)
      $h_u^l \leftarrow \frac{h_u^l}{\|h_u^l\|_2}$
    End
  End
  $z_v \leftarrow h_u^l$
```

Specifically, we feed the drug attribute feature F_{drug} and drug network clustering feature U to the GraphSAGE to get the embedded features z_{drug}^F and z_{drug}^U and feed the disease attribute feature F_{dis} and disease network feature V to the GraphSAGE to get the embedded features z_{dis}^F and z_{dis}^V. Then we connected the drug embedding features with the disease embedding features to obtain the association features, so as to learn their low dimensional features and predict their relationships. For example, to predict the association between drug i and disease j, we connect z_{drug}^F, z_{drug}^U, z_{dis}^F and z_{dis}^V as *concat* ($z^F_{drug_i}$, $z^U_{drug_i}$, $z^F_{dis_j}$, $z^V_{dis_j}$), input it into a three-layer fully connected network, and finally use the SoftMax function to find its probability P_{ij}.

Optimization

GraphSAGE can perform unsupervised learning (Xu et al., 2020), but this objective function is completely based on the topological properties of the network, ignoring the original features of the nodes. If it is applied to this research, each training needs to use a different network. Its essence can reflect the relationship of the features between nodes very well, but it cannot predict the relationship very well. Therefore, we still used the cross-entropy function as the objective function. In order to prevent the over-fitting problem, an L2-regularization is also adopted:

$$Loss = -\sum_{i=1}^{N_{drug}} \sum_{j=1}^{N_{dis}} \left(Y_{ij} \, log \, P_{ij} + \left(1 - Y_{ij}\right) log\left(1 - P_{ij}\right) \right) + \frac{\lambda}{N} \sum_{l=1}^{L} \sum_{w \in W^l} w^2, \quad (21)$$

where P_{ij} represents the associated probability of drug i and disease j, $Y_{ij} \in \{0, 1\}$ is the known associations, and $N_{drug}\,(N_{dis})$ is the drug (disease) sample size. Since no negative samples are given in the two databases, extracting reliable negative samples is also an important part of the experiment. The usual operation is to randomly select the same number of negative samples as positive samples from unknown samples. But this will actually interfere with the model learning, so we used the network double random walk (Xie et al., 2012) method to determine the negative samples. After the random walk, the same samples with the smallest scores are regarded as negative samples.

EXPERIMENTAL RESULTS AND ANALYSIS

Based on previous works, we validate our method by answering the following questions:

- Are the features we extracted valid, and can network clustering features improve the performance of the method?
- Can DRGCC predict drug–disease associations with higher accuracy?
- Can we verify that the predicted repositioning drugs are effective, especially for COVID-19?

Experiment Setting

In our study, we used 5-fold cross-validation (5-fold CV) to evaluate the prediction performance of DRGCC and other competing methods. All samples were randomly divided into five equal-sized parts, four parts of them were used as training data, and the remaining one was used as test data. This process was repeated 5 times, with each part of the data tested once, and the average result of these 5 times was taken as the result of this cross-validation. After that, the samples were randomly divided again, cross-validation was also performed 5 times, and the results were averaged. We mainly used seven metrics: area under the receiver operating characteristic curve (AUC), area under the precision and recall curve (PRAUC), F1_SCORE, ACCURACY, SPECIFICITY, PRECISION, and RECALL (Yu et al., 2020), to comprehensively evaluate the performance of the method. We took the prediction threshold that maximizes the F1_SCORE and built two-layer GraphSAGE models for drugs and diseases separately. After further statistical analysis of drug and disease features, we set some default parameters. The attribute feature dimension d_{drug} of drugs was set to 300, while the attribute feature dimension d_{dis} of diseases was set to 100. The network clustering feature dimension k was set to 200. In GraphSAGE, the layer dimensions of drug attribute features were {128, 64}, the layer dimensions of disease attribute features were {64, 32}, and the layer dimensions of network clustering features were {128, 32}. The number of epochs was 30. The learning rate was 0.001. The value of λ in loss function was 0.01. The layer dimensions of a fully connected network were {64, 32, 2}. The dropout was set to 0.5. With the MCODE (Bader and Hogue, 2003) algorithm, the drug and disease (virus) networks in the CTD and HDVD databases were split into 8, 14, and 15, 4 subnetworks, respectively.

Parameter Sensitivity Analysis

In constraint matrix factorization, the regularization of parameters α and β has an important influence on the extraction of network clustering features. We tested all possible combinations of α and β, as shown in **Figure 2A**. We found that if $\alpha = 0.2$, $\beta = 0.1$, the method has the best AUC value on the CTD dataset. At the same time, since the DRGCC is sampled and trained in batches, the size of the batch is particularly important. If the batch is too small, it will be difficult to converge. If the batch is too large, it demands a large amount of computation. We tested the effect of different batch_size on the method, as shown in **Figure 2B**. The method has the best performance when the batch_size is equal to 128.

For the GraphSAGE, there are a total of five different aggregation methods. We performed comparisons on dataset CTD and dataset HDVD, respectively. We can find that the performance of the aggregation methods based on mean, meanPool, and maxPool are similar in **Figure 3** and are significantly higher than that of the aggregation based on the LSTM and GCN. This may illustrate that structural features between drugs and symptom features between diseases can be fused using linear methods. Finally, we used the mean method as the aggregation method of the DRGCC.

We also evaluated the sampling number of network neighbors. Similar to Cui et al. (2021), we tested 4 cases, where n_k is {3, 5, 10, 15} and finally determined that it is better to take the nearest 5 neighbor nodes as aggregation nodes. **Figure 4** shows the distribution of AUC values for a total of 25 times in 5 cross-validations. This test is run on the HDVD network because it is sparser, and the test on the CTD dataset has a similar effect.

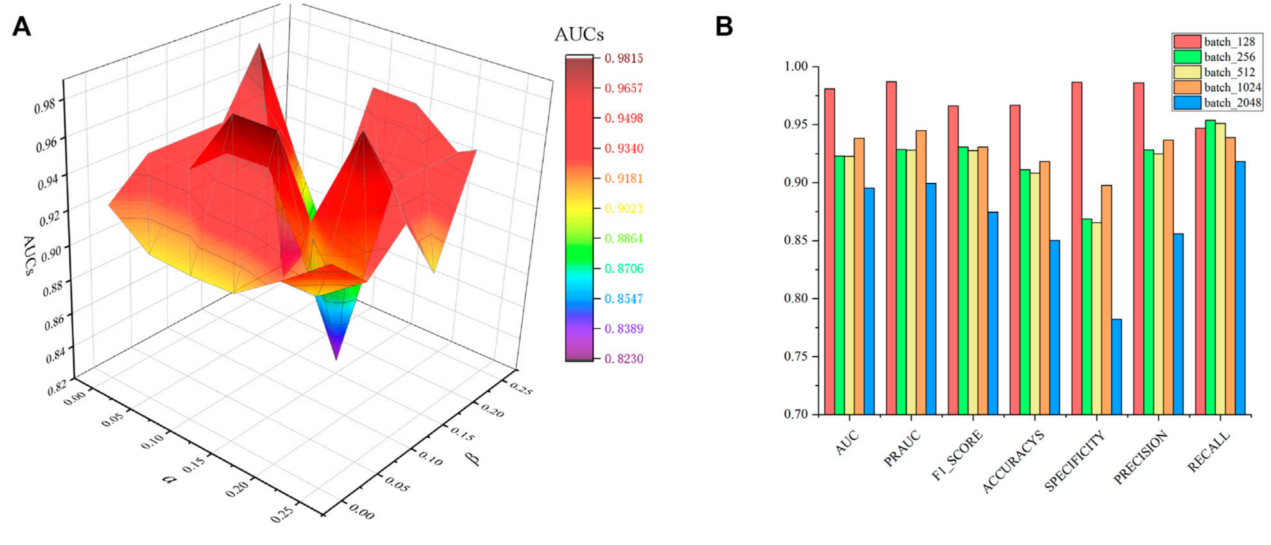


FIGURE 2 | Parametric analysis, **(A)** effects α and β on prediction accuracy, and **(B)** effect of batch size.

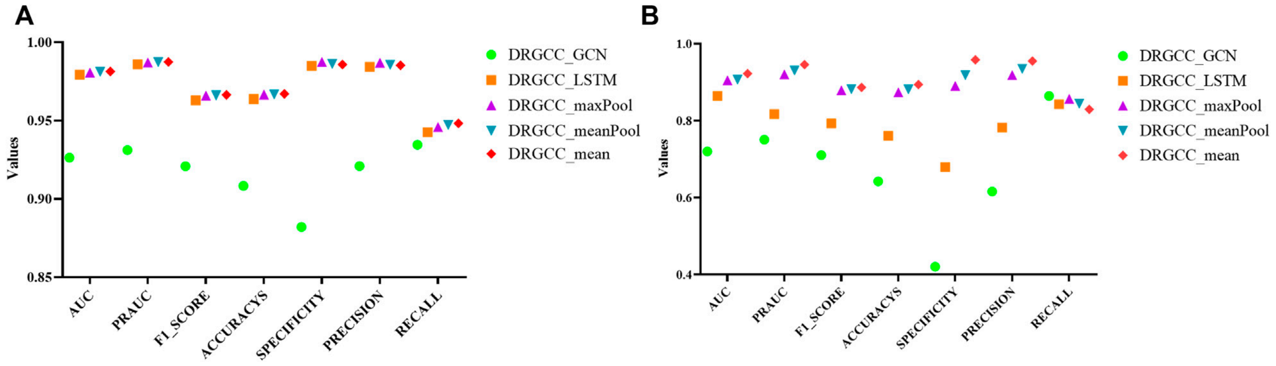


FIGURE 3 | Impact of different aggregation methods on performance. **(A)** Performance of different aggregation methods on the CTD dataset; **(B)** Performance of different aggregation methods on the HDVD dataset.

Effectiveness of Network Clustering Features

To answer the first question of the experiment, we conduct ablation experiments using only attribute features DRGCC_Attribute and only network clustering features DRGCC_Cluster for prediction, respectively. **Table 2** shows that the model with clustering features is slightly higher than the model with only attribute features, and the fusion of the two features has a prominent effect on the CTD database. In **Figure 5**, the ROC curve of a 5-fold cross-validation is depicted. The average of 5 times is also calculated. It is clear that the performance of applying two features to DRGCC at the same time is better than using a single one, and the AUC is as high as 0.9809. The network clustering feature has a better effect on improving the performance of the method.

Comparative Analysis With Other Methods

To answer the second question of the experiment, we compared DRGCC with six state-of-the-art drug repositioning methods in this section, such as MBiRW (Luo et al., 2016), DRRS (Luo et al., 2018), BNNR (Zhang W. et al., 2020), SCPMFDR (Meng et al., 2021), NIMCGCN (Li et al., 2020), and LAGCN (Yu et al., 2020) on CTD and HDVD datasets. These methods are mainly divided into three categories: methods based on network propagation, methods based on low-rank matrix approximation, and methods based on the GNN.

- MBiRW (Luo et al., 2016) integrates drug or disease feature information with known drug–disease associations, and the comprehensive similarity measures are developed to calculate similarity for drugs and diseases. They are incorporated into a heterogeneous network with known drug–disease

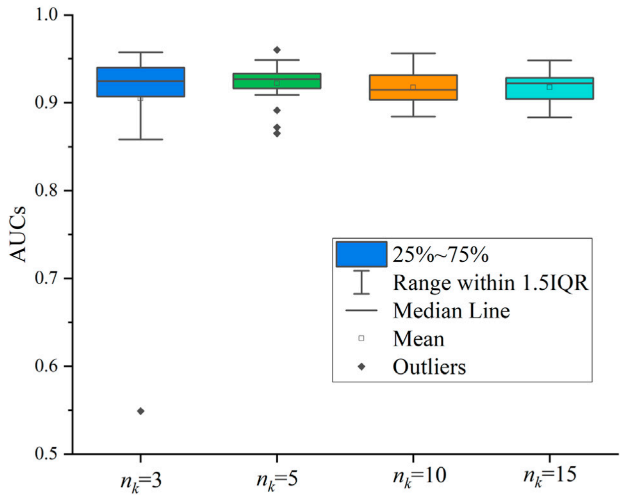


FIGURE 4 | Effect of the number n_k of neighbor samples on performance.

interactions. Based on the drug–disease heterogeneous network, the bi-random walk (BiRW) algorithm is used to identify potential novel indications for a given drug.

- DRRS (Luo et al., 2018) is a matrix completion-based recommendation system on a drug–disease heterogeneous network to predict drug–disease associations.
- BNNR (Zhang W. et al., 2020) is a bounded nuclear norm regularization method to complete a drug–disease heterogeneous network.
- SCPMFDR (Meng et al., 2021) is implemented on an adjacency matrix of a heterogeneous drug–virus network, which integrates the known drug–virus interactions, drug chemical structures, and virus genomic sequences. SCPMF projects the drug–virus interactions matrix into two latent feature matrices for the drugs and viruses, which reconstruct the drug–virus interactions matrix when multiplied together, and then introduces similarity constrained probabilistic matrix factorization to predict associations.
- NIMCGCN (Li et al., 2020) use GCNs to learn latent feature representations of miRNA and disease from the similarity networks and then put the learned features into a neural inductive matrix completion model to obtain a reconstructed association matrix. NIMCGCN is a GCN-based method proposed for the miRNA–disease association prediction, and we adopt it as the baseline method for the drug–disease association.
- LAGCN (Yu et al., 2020) integrates the known drug–disease associations, drug–drug similarities, and disease–disease similarities into a heterogeneous network and applies the graph convolution operation to the network to learn the embeddings of drugs and diseases. It combines the embeddings from multiple graph convolution layers using the attention mechanism.

In **Table 3**, the results show that our method outperforms other methods on all 7 metrics for the CTD database. In large networks, only considering the relationship between nodes in the network and ignoring the biochemical properties of the nodes themselves have poor prediction performance. For the drug–virus prediction, disease similarity networks based on amino acid sequences and structure-based on drug similarity networks have been provided in HDVD. Since there are no features of viruses in the original dataset, we used DRGCC_cluster for the prediction problem. The prediction results are slightly lower than the results on the CTD due to a large number of unknown relationships and the inclusion of the new virus COVID-19, as shown in **Table 4**. Except for the RECALL, the other evaluation values are the highest. The AUC reaches 0.9222, and the PRAUC reaches 0.9458. It can be seen that DRGCC has excellent performance.

Case Studies

To answer the third question of the experiment, we presented an analysis of the predicted repositioned drugs. The top 10 predicted drug–disease relationships were extracted, as shown in **Table 5**. Among the top 10 prediction results, we can find corroborations or explanations for 6 predictions from other studies. Early evidence in rats suggested that acetazolamide may inhibit sodium and water transport in the ileum in addition to inhibiting bicarbonate secretion (Sladen, 1973). It may have an influence on duodenal ulcer treatment. Rimonabant was shown to be safe and effective in treating the combined cardiovascular risk factors of smoking and obesity (Cleland et al., 2004). Hypoosmolar hyponatremia occurs in conditions of plasma volume depletion such as cirrhosis and heart failure and syndromes of inappropriate antidiuretic hormone secretion. Conventional proposals for euvolemic and hypervolemic hyponatremia consist of lithium carbonate (Gross, 2008). Peyrani *et al.* believed that therapeutics beyond antibiotics (e.g., heparin or aspirin) may be indicated during and after hospitalization for the patients with community-acquired pneumonia (Peyrani and Ramirez 2013). Newer antiemetic with prokinetic properties (cisapride) have also been introduced in the management of

TABLE 2 | Comparison of different features on prediction performance.

Method	AUC	PRAUC	F1_SCORE	ACCURACY	SPECIFICITY	PRECISION	RECALL
DRGCC_Attribute	0.9327 ± 0.0463	0.9379 ± 0.0467	0.9267 ± 0.0278	0.9141 ± 0.0442	0.8893 ± 0.0950	0.9283 ± 0.0459	0.9390 ± 0.0066
DRGCC_Cluster	0.9418 ± 0.0478	0.9477 ± 0.048	0.9430 ± 0.0295	0.9303 ± 0.0459	0.9083 ± 0.0964	0.9477 ± 0.0476	0.9524 ± 0.0050
DRGCC	0.9809 ± 0.0005	0.9871 ± 0.0003	0.9661 ± 0.0006	0.9668 ± 0.0006	0.9866 ± 0.0020	0.9861 ± 0.0020	0.9470 ± 0.0008

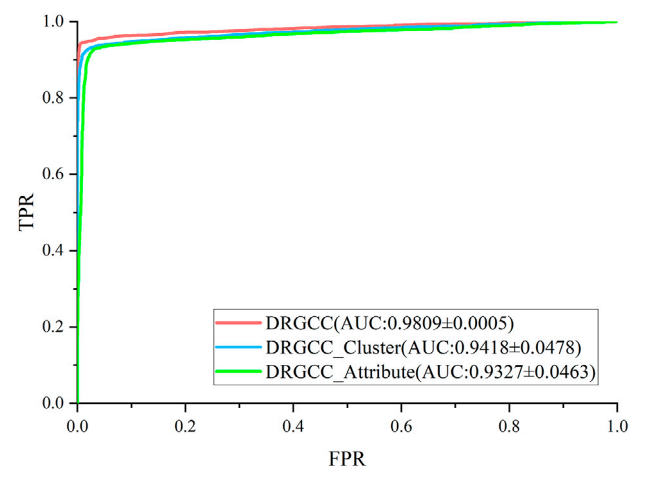


FIGURE 5 | ROC and AUC comparison of DRGCC with different features.

gastrointestinal motility disturbances and inflammatory bowel diseases. Some benzodiazepines have been shown to be effective in treating certain anxiety disorders (Swedish Council on Health Technology Assessment, 2005).

The novel coronavirus disease 2019 (COVID-19) pandemic has triggered a massive health crisis and upended economies across the globe. However, the research and development of traditional medicines for the new coronavirus is very expensive in terms of time, manpower, and funds. Drug repurposing emerged as a promising therapeutic strategy during the COVID-19 virus crisis. We also predicted the top 10 possible drugs for anti-COVID-19, as shown in **Table 6**. Excitingly, seven of them have been reported by medical researchers, such as, triazavirin is a guanine nucleotide analog antiviral that has shown efficacy against influenza A and B, including the H5N1 strain. Given the similarities between SARS-CoV-2 and H5N1, health scientists are investigating triazavirin as an option to combat COVID-19 (Shahab and Sheikhi, 2021)

TABLE 3 | Performance of comparison methods on CTD dataset.

Method	AUC	PRAUC	F1_SCORE	ACCURACY	SPECIFICITY	PRECISION	RECALL
MbiRW	0.8524±0.0006	0.8487±0.0004	0.7880±0.0016	0.7730±0.0026	0.7025±0.0086	0.7395±0.0047	0.8435±0.0046
DRRS	0.9647±0.0006	0.9655±0.0005	0.9020±0.0009	0.9010±0.0012	0.8909±0.0065	0.8933±0.0053	0.9111±0.0045
BNNR	0.9302±0.0007	0.9479±0.0004	0.8748±0.0012	0.8790±0.0009	0.9120±0.0052	0.9060±0.0045	0.8459±0.0055
SCPMFDR	0.9667±0.0003	0.9734±0.0002	0.9101±0.0011	0.9118±0.0011	0.9304±0.0036	0.9279±0.0032	0.8932±0.0029
NIMCGCN	0.7989±0.0130	0.7311±0.0221	0.8172±0.0081	0.7984±0.0093	0.71780±0.0194	0.7727±0.0173	0.8789±0.0054
LAGCN	0.9259±0.0044	0.7939±0.0054	0.8055±0.0052	0.8843±0.0035	0.8993±0.0091	0.7825±0.0061	0.8314±0.0119
DRGCC	0.9809±0.0005	0.9871±0.0003	0.9661±0.0006	0.9668±0.0006	0.9866±0.0020	0.9861±0.0020	0.9470±0.0008

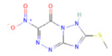
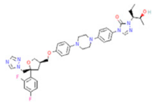
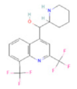
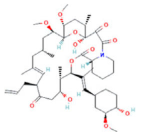
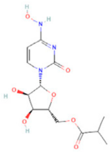
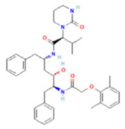
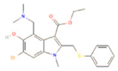
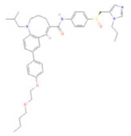
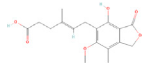
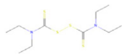
TABLE 4 | Performance of comparison methods on HDVD dataset.

Method	AUC	PRAUC	F1_SCORE	ACCURACY	SPECIFICITY	PRECISION	RECALL
MBiRW	0.9113±0.0059	0.9237±0.0052	0.8580±0.0061	0.8541±0.0070	0.8312±0.0149	0.8431±0.0119	0.8769±0.0084
DRRS	0.8936±0.0030	0.92539±0.0021	0.85451±0.0055	0.8664±0.0044	0.9477±0.0117	0.9439±0.0099	0.7851±0.0117
BNNR	0.9088±0.0086	0.93075±0.0062	0.8530±0.0103	0.8580±0.01120	0.8901±0.02428	0.8878±0.0202	0.8260±0.0174
SCPMFDR	0.8655±0.0073	0.8813±0.0064	0.8311±0.0089	0.8325±0.0082	0.8400 ±0.01992	0.8397±0.0141	0.8251±0.0194
NIMCGCN	0.6002±0.0103	0.5922±0.0108	0.7074±0.0034	0.6062±0.0125	0.2686±0.0448	0.5721±0.0167	0.9438±0.0202
LAGCN	0.7433±0.0164	0.5307±0.0097	0.6048±0.0060	0.6989±0.0165	0.6284±0.03212	0.4878±0.0156	0.8105±0.0315
DRGCC	0.9222±0.0080	0.9458±0.0042	0.8863±0.0052	0.8938±0.0048	0.9582±0.0171	0.9548±0.0166	0.8295±0.0165

TABLE 5 | Top 10 repositioned drugs predicted by the DRGCC.

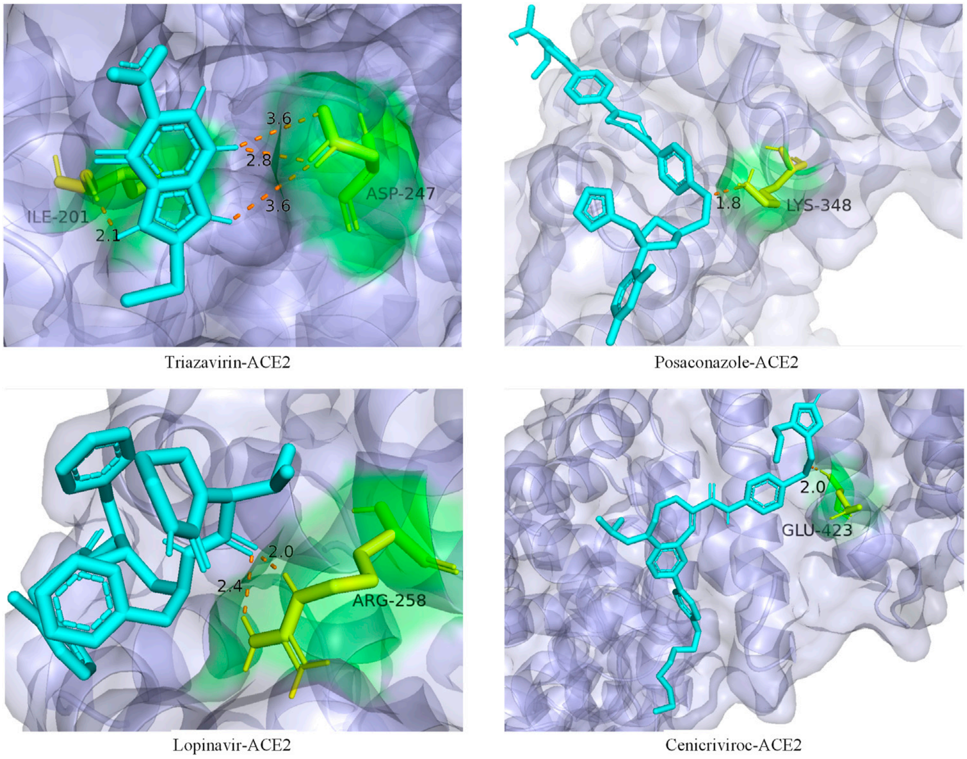
Rank	Drug name	Disease name	Evidence (PMID)
1	Acetazolamide	Duodenal ulcer	4360063, Sladen (1973)
2	Salinomycin	Stroke	*NA*
3	Rimonabant	Heart failure	15182777, Cleland *et al.* (2004)
4	Lithium carbonate	Liver cirrhosis and biliary cirrhosis	18480571, Gross (2008)
5	Acetylcarnitine	Hematologic neoplasms	*NA*
6	Heparin	Community-acquired infections	23398875, Peyrani and Ramirez (2013)
7	Icariin	Sialorrhea	*NA*
8	Cisapride	Inflammatory bowel diseases	1974182, Lauritsen *et al.* (1990)
9	Moxifloxacin	Insulin resistance	*NA*
10	Benzodiazepines	Stress disorders, post-traumatic	28876726, Swedish Council on Health Technology Assessment (2005)

TABLE 6 | Top 10 possible anti-COVID-19 drugs predicted by the DRGCC.

Rank	Accession number	Drug name	2D structure	Evidence (PMID)
1	DB15622	Triazavirin		32436829, Shahab and Sheikhi (2021) 33249050, Valiulin *et al.* (2021)
2	DB01263	Posaconazole		34016284, Cadena *et al.* (2021)
3	DB00358	Mefloquine		34126913, Uddin *et al.* (2021)
4	DB00864	Tacrolimus		33495742, Solanich *et al.* (2021)
5	DB15661	EIDD-2801		34271264, Painter *et al.* (2021)
6	DB01601	Lopinavir		*NA*
7	DB13609	Umifenovir		33336780, Trivedi *et al.* (2020)
8	DB11758	Cenicriviroc		*NA*
9	DB01024	Mycophenolic acid		32639598, Lai *et al.* (2020)
10	DB00822	Disulfiram		*NA*

(Valiulin et al., 2021). Aspergillus-producing diseases range from allergic syndromes to chronic lung disease and invasive infections and are frequently observed following COVID-19 infection. Posaconazole has better efficacy with less toxicity for extensive infection and severe immunosuppression (Cadena et al., 2021). In the reports on possible drugs for COVID-19, Uddin *et al.* mentioned that mefloquine may be one of the options (Uddin et al., 2021). In the research of Solanich *et al.*, methylprednisolone and tacrolimus were considered that might be beneficial to treat those COVID-19 patients progressing into severe pulmonary failure and systemic hyperinflammatory syndrome (Solanich et al., 2021). Molnupiravir (EIDD-2801) was originally designed for the treatment of alphavirus infections. Painter *et al.* described its evolution into a potential drug for the prevention and treatment of COVID-19 (Painter et al., 2021). Umifenovir was deemed one of the most hopeful antiviral agents for improving the health of COVID-19 patients (Trivedi et al., 2020). The studies of Lai *et al.* showed that the use of mycophenolic acid might be a strategy to reduce viral replication (Lai et al., 2020).

In addition, we also analyzed the docking state of unverifiable drugs and receptors. Angiotensin-converting enzyme 2 (ACE2) was considered an important functional



Triazavirin-ACE2

Posaconazole-ACE2

Lopinavir-ACE2

Cenicriviroc-ACE2

FIGURE 6 | Ligand–protein binding mode between the predicted drugs and the protein receptor ACE2. The purple part is the protein ACE2, the blue part is the drug compound, the yellow part is the amino acid residue, and the orange dotted line is the connecting hydrogen bond. The numbers represent atomic distances.

receptor for SARS and other coronaviruses (Li et al., 2003). Like SARS-CoV, SARS-CoV-2 infects human respiratory epithelial cells through invasion mediated by human cell surface s-protein and ACE2 protein receptors. Obstructing the combination of ACE2 and the virus has become one of the effective means to prevent the respiratory infection of the crown virus. The molecular docking technology allows us to clearly determine the binding sites and bond strengths between molecules (Meng et al., 2011). We examined the binding of 4 drug compounds triazavirin, posaconazole, lopinavir, and cenicriviroc to the receptor protein ACE2. As shown in **Figure 6**, triazavirin and ACE2 have 4 hydrogen bonds bound to amino acids ILE and ASP, respectively. Lopinavir has 2 hydrogen bonds bound to amino acid ARG in ACE2. Posaconazole and cenicriviroc also have binding sites to ACE2. It can be seen that only one of the 3 unreported drugs has not been corroborated. It can be seen that these drugs may provide some help in the treatment of COVID-19.

CONCLUSION

In this article, we have proposed a drug repositioning method DRGCC to predict potential relationships between existing drugs and new diseases. The method first reconstructed the drug–drug interaction network, established the disease semantic similarity network, then extracted the structural features of drugs and disease symptoms as attribute features, and obtained network clustering features through matrix factorization. Finally, all features were fed to the GraphSAGE model to obtain predictions of drug–disease associations. With experiments testing on two datasets, it is found that our method has better performance than other competing methods. Experiments also demonstrated the importance of network clustering features for accurate prediction. At the same time, DRGCC is suitable for training and predicting large-scale samples and can add new nodes to the network after training, such as the SARS-CoV-2 virus. After analyzing the predicted repositioning drugs, we gave several possible drug treatment combinations and recommended several anti-COVID-19 drugs. These predictions have been supported or discussed by other studies. It can be seen that DRGCC has certain reliability in drug repositioning studies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: http://ctdbase.org/; https://github.com/luckymengmeng/HDVD; https://go.drugbank.com/; https://pubchem.ncbi.nlm.nih.gov/.

AUTHOR CONTRIBUTIONS

YZ and XL proposed the concept and idea; YZ implemented the algorithm and wrote the draft manuscript; F-XW provided the method improvement strategy; XL, F-XW, and YP evaluated the results and revised the manuscript; and YP and F-XW supervised the whole study. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61972451, 61902230); the Shenzhen Science and Technology Program (No. KQTD20200820113106007); the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (2018TS079); and the program of China Scholarships Council (No. 202006870037).

REFERENCES

Bader, G. D., and Hogue, C. W. (2003). An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics* 4, 2. doi:10.1186/1471-2105-4-2

Bader, S., Kühner, S., and Gavin, A. C. (2008). Interaction Networks for Systems Biology. *FEBS Lett.* 582, 1220–1224. doi:10.1016/j.febslet.2008.02.015

Booth, B., and Zemmel, R. (2004). Prospects for Productivity. *Nat. Rev. Drug Discov.* 3, 451–456. doi:10.1038/nrd1384

Cadena, J., Thompson, G. R., 3rd, and Patterson, T. F. (2021). Aspergillosis: Epidemiology, Diagnosis, and Treatment. *Infect. Dis. Clin. North. Am.* 35, 415–434. doi:10.1016/j.idc.2021.03.008

Cai, L., Lu, C., Xu, J., Meng, Y., Wang, P., Fu, X., et al. (2021). Drug Repositioning Based on the Heterogeneous Information Fusion Graph Convolutional Network. *Brief Bioinform* 22, bbab319. doi:10.1093/bib/bbab319

Cheng, F., Lu, W., Liu, C., Fang, J., Hou, Y., Handy, D. E., et al. (2019). A Genome-wide Positioning Systems Network Algorithm for In Silico Drug Repurposing. *Nat. Commun.* 10, 3476. doi:10.1038/s41467-019-10744-6

Cleland, J. G., Ghosh, J., Freemantle, N., Kaye, G. C., Nasir, M., Clark, A. L., et al. (2004). Clinical Trials Update and Cumulative Meta-Analyses from the American College of Cardiology: WATCH, SCD-HeFT, DINAMIT, CASINO, INSPIRE, STRATUS-US, RIO-Lipids and Cardiac Resynchronisation Therapy in Heart Failure. *Eur. J. Heart Fail.* 6, 501–508. doi:10.1016/j.ejheart.2004.04.014

Cui, C., Ding, X., Wang, D., Chen, L., Xiao, F., Xu, T., et al. (2021). Drug Repurposing against Breast Cancer by Integrating Drug-Exposure Expression Profiles and Drug-Drug Links Based on Graph Neural Network. *Bioinformatics* 37, 2930–2937. doi:10.1093/bioinformatics/btab191

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., et al. (2021). Comparative Toxicogenomics Database (CTD): Update 2021. *Nucleic Acids Res.* 49, D1138–D1143. doi:10.1093/nar/gkaa891

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407. doi:10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," in 30th Conference on Neural Information Processing Systems, Barcelona, Spain. NY, United States: Curran Associates Inc., 3844–3852.

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011). Exploiting Drug-Disease Relationships for Computational Drug Repositioning. *Brief Bioinform* 12, 303–311. doi:10.1093/bib/bbr013

Fiscon, G., Conte, F., Farina, L., and Paci, P. (2021). SAveRUNNER: A Network-Based Algorithm for Drug Repurposing and its Application to COVID-19. *Plos Comput. Biol.* 17, e1008686. doi:10.1371/journal.pcbi.1008686

Gross, P. (2008). Treatment of Hyponatremia. *Intern. Med.* 47, 885–891. doi:10.2169/internalmedicine.47.0918

Groza, V., Udrescu, M., Bozdog, A., and Udrescu, L. (2021). Drug Repurposing Using Modularity Clustering in Drug-Drug Similarity Networks Based on Drug-Gene Interactions. *Pharmaceutics* 13, 2117. doi:10.3390/pharmaceutics13122117

Guan, N. N., Zhao, Y., Wang, C. C., Li, J. Q., Chen, X., and Piao, X. (2019). Anticancer Drug Response Prediction in Cell Lines Using Weighted Graph Regularized Matrix Factorization. *Mol. Ther. Nucleic Acids* 17, 164–174. doi:10.1016/j.omtn.2019.05.017

Hahn, M., and Roll, S. C. (2021). The Influence of Pharmacogenetics on the Clinical Relevance of Pharmacokinetic Drug-Drug Interactions: Drug-Gene, Drug-Gene-Gene and Drug-Drug-Gene Interactions. *Pharmaceuticals (Basel)* 14, 187. doi:10.3390/ph14050487

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). "Inductive Representation Learning on Large Graphs," in Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA: Curran Associates Inc.).

Huang, F., Qiu, Y., Li, Q., Liu, S., and Ni, F. (2020). Predicting Drug-Disease Associations via Multi-Task Learning Based on Collective Matrix Factorization. *Front. Bioeng. Biotechnol.* 8, 218. doi:10.3389/fbioe.2020.00218

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* 49, D1388–D1395. doi:10.1093/nar/gkaa971

Lai, Q., Spoletini, G., Bianco, G., Graceffa, D., Agnes, S., Rossi, M., et al. (2020). SARS-CoV2 and Immunosuppression: A Double-Edged Sword. *Transpl. Infect. Dis.* 22, e13404. doi:10.1111/tid.13404

Lauritsen, K., Laursen, L. S., and Rask-Madsen, J. (1990). Clinical Pharmacokinetics of Drugs Used in the Treatment of Gastrointestinal Diseases (Part I). *Clin. Pharmacokinet.* 19, 11–31. doi:10.2165/00003088-199019010-00002

Lei, X.-J., Bian, C., and Pan, Y. (2021). Predicting CircRNA-Disease Associations Based on Improved Weighted Biased Meta-Structure. *J. Comput. Sci. Technol.* 36, 288–298. doi:10.1007/s11390-021-0798-x

Li, J., Zhang, S., Liu, T., Ning, C., Zhang, Z., and Zhou, W. (2020). Neural Inductive Matrix Completion with Graph Convolutional Networks for miRNA-Disease Association Prediction. *Bioinformatics* 36, 2538–2546. doi:10.1093/bioinformatics/btz965

Li, J., Zhang, S., Wan, Y., Zhao, Y., Shi, J., Zhou, Y., et al. (2019). MISIM v2.0: a Web Server for Inferring microRNA Functional Similarity Based on microRNA-Disease Associations. *Nucleic Acids Res.* 47, W536–W541. doi:10.1093/nar/gkz328

Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., et al. (2003). Angiotensin-converting Enzyme 2 Is a Functional Receptor for the SARS Coronavirus. *Nature* 426, 450–454. doi:10.1038/nature02145

Li, Y., Wang, K., and Wang, G. (2021). Evaluating Disease Similarity Based on Gene Network Reconstruction and Representation. *Bioinformatics* 37, 3579–3587. doi:10.1093/bioinformatics/btab252

Li, Z., Huang, Q., Chen, X., Wang, Y., Li, J., Xie, Y., et al. (2019). Identification of Drug-Disease Associations Using Information of Molecular Structures and Clinical Symptoms via Deep Convolutional Neural Network. *Front. Chem.* 7, 924. doi:10.3389/fchem.2019.00924

Lu, L., and Yu, H. (2018). DR2DI: a Powerful Computational Tool for Predicting Novel Drug-Disease Associations. *J. Comput. Aided Mol. Des.* 32, 633–642. doi:10.1007/s10822-018-0117-y

Luo, H., Li, M., Wang, S., Liu, Q., Li, Y., and Wang, J. (2018). Computational Drug Repositioning Using Low-Rank Matrix Approximation and Randomized Algorithms. *Bioinformatics* 34, 1904–1912. doi:10.1093/bioinformatics/bty013

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., et al. (2016). Drug Repositioning Based on Comprehensive Similarity Measures and Bi-random Walk Algorithm. *Bioinformatics* 32, 2664–2671. doi:10.1093/bioinformatics/btw228

Madhukar, N. S., Khade, P. K., Huang, L., Gayvert, K., Galletti, G., Stogniew, M., et al. (2019). A Bayesian Machine Learning Approach for Drug Target Identification Using Diverse Data Types. *Nat. Commun.* 10, 5221. doi:10.1038/s41467-019-12928-6

Meng, X. Y., Zhang, H. X., Mezei, M., and Cui, M. (2011). Molecular Docking: a Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided Drug Des.* 7, 146–157. doi:10.2174/157340911795677602

Meng, Y., Jin, M., Tang, X., and Xu, J. (2021). Drug Repositioning Based on Similarity Constrained Probabilistic Matrix Factorization: COVID-19 as a Case Study. *Appl. Soft Comput.* 103, 107135. doi:10.1016/j.asoc.2021.107135

Ni, P., Wang, J., Zhong, P., Li, Y., Wu, F. X., and Pan, Y. (2020). Constructing Disease Similarity Networks Based on Disease Module Theory. *Ieee/acm Trans. Comput. Biol. Bioinform* 17, 906–915. doi:10.1109/TCBB.2018.2817624

Painter, G. R., Natchus, M. G., Cohen, O., Holman, W., and Painter, W. P. (2021). Developing a Direct Acting, Orally Available Antiviral Agent in a Pandemic: the Evolution of Molnupiravir as a Potential Treatment for COVID-19. *Curr. Opin. Virol.* 50, 17–22. doi:10.1016/j.coviro.2021.06.003

Peyrani, P., and Ramirez, J. (2013). What Is the Association of Cardiovascular Events with Clinical Failure in Patients with Community-Acquired Pneumonia? *Infect. Dis. Clin. North. Am.* 27, 205–210. doi:10.1016/j.idc.2012.11.010

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., et al. (2019). Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion. *Nucleic Acids Res.* 47, D955–D962. doi:10.1093/nar/gky1032

Shahab, S., and Sheikhi, M. (2021). Triazavirin - Potential Inhibitor for 2019-nCoV Coronavirus M Protease: A DFT Study. *Curr. Mol. Med.* 21, 645–654. doi:10.2174/1566524020666200521075848

Sladen, G. E. (1973). The Pathogenesis of Cholera and Some Wider Implications. *Gut* 14, 671–680. doi:10.1136/gut.14.8.671

Solanich, X., Antolí, A., Padullés, N., Fanlo-Maresma, M., Iriarte, A., Mitjavila, F., et al. (2021). Pragmatic, Open-Label, single-center, Randomized, Phase II Clinical Trial to Evaluate the Efficacy and Safety of Methylprednisolone Pulses and Tacrolimus in Patients with Severe Pneumonia Secondary to COVID-19: The TACROVID Trial Protocol. *Contemp. Clin. Trials Commun.* 21, 100716. doi:10.1016/j.conctc.2021.100716

Strating, J. R., Van Der Linden, L., Albulescu, L., Bigay, J., Arita, M., Delang, L., et al. (2015). Itraconazole Inhibits Enterovirus Replication by Targeting the Oxysterol-Binding Protein. *Cell Rep* 10, 600–615. doi:10.1016/j.celrep.2014.12.054

Swedish Council on Health Technology Assessment (2005). "SBU Systematic Review Summaries," in *Treatment of Anxiety Disorders: A Systematic Review* (Stockholm: Swedish Council on Health Technology Assessment (SBU) Copyright).

Trivedi, N., Verma, A., and Kumar, D. (2020). Possible Treatment and Strategies for COVID-19: Review and Assessment. *Eur. Rev. Med. Pharmacol. Sci.* 24, 12593–12608. doi:10.26355/eurrev_202012_24057

Uddin, E., Islam, R., AshrafuzzamanBitu, N. A., Bitu, N. A., Hossain, M. S., Islam, A. N., et al. (2021). Potential Drugs for the Treatment of COVID-19: Synthesis, Brief History and Application. *Curr. Drug Res. Rev.* 13, 184–202. doi:10.2174/2589977513666210611155426

Valiulin, S. V., Onischuk, A. A., Dubtsov, S. N., Baklanov, A. M., An'kov, S. V., Plokhotnichenko, M. E., et al. (2021). Aerosol Inhalation Delivery of Triazavirin in Mice: Outlooks for Advanced Therapy against Novel Viral Infections. *J. Pharm. Sci.* 110, 1316–1322. doi:10.1016/j.xphs.2020.11.016

Varothai, S., and Bergfeld, W. F. (2014). Androgenetic Alopecia: An Evidence-Based Treatment Update. *Am. J. Clin. Dermatol.* 15, 217–230. doi:10.1007/s40257-014-0077-5

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the Human microRNA Functional Similarity and Functional Network Based on microRNA-Associated Diseases. *Bioinformatics* 26, 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, Y., Lei, X., and Pan, Y. (2022). Predicting Microbe-Disease Association Based on Heterogeneous Network and Global Graph Feature Learning. *Chin. J. Electro.* 31, 1–9. doi:10.1049/cje.2020.00.212

Wang, Y. Y., Cui, C., Qi, L., Yan, H., and Zhao, X. M. (2019). DrPOCS: Drug Repositioning Based on Projection onto Convex Sets. *Ieee/acm Trans. Comput. Biol. Bioinform* 16, 154–162. doi:10.1109/TCBB.2018.2830384

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi:10.1093/nar/gkx1037

Wu, G., Liu, J., and Yue, X. (2019). Prediction of Drug-Disease Associations Based on Ensemble Meta Paths and Singular Value Decomposition. *BMC Bioinformatics* 20, 134. doi:10.1186/s12859-019-2644-5

Xie, M., Hwang, T., and Kuang, R. (2012). "Prioritizing Disease Genes by Bi-random Walk," in *Advances in Knowledge Discovery and Data Mining*. Editors P.-N. Tan, S. Chawla, C. K. Ho, and J. Bailey (Kuala Lumpur, Malaysia: Springer Berlin Heidelberg), 292–303. doi:10.1007/978-3-642-30220-6_25

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., and Achan, K. (2020). "Inductive Representation Learning on Temporal Graphs," in 2020 International Conference on Learning Representations, Barcelona, Spain. NY, United States: Curran Associates Inc., 1–17.

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug Repositioning Based on Bounded Nuclear Norm Regularization. *Bioinformatics* 35, i455–i463. doi:10.1093/bioinformatics/btz331

Yu, G. (2018). Using Meshes for MeSH Term Enrichment and Semantic Analyses. *Bioinformatics* 34, 3766–3767. doi:10.1093/bioinformatics/bty410

Yu, L., Huang, J., Ma, Z., Zhang, J., Zou, Y., and Gao, L. (2015). Inferring Drug-Disease Associations Based on Known Protein Complexes. *BMC Med. Genomics* 8, S2. doi:10.1186/1755-8794-8-S2-S2

Yu, L., Zhao, J., and Gao, L. (2018). Predicting Potential Drugs for Breast Cancer Based on miRNA and Tissue Specificity. *Int. J. Biol. Sci.* 14, 971–982. doi:10.7150/ijbs.23350

Yu, Z., Huang, F., Zhao, X., Xiao, W., and Zhang, W. (2020). Predicting Drug-Disease Associations through Layer Attention Graph Convolutional Network. *Brief Bioinform* 22 (4), bbaa243. doi:10.1093/bib/bbaa243

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a Network-Based Deep Learning Approach to In Silico Drug Repositioning. *Bioinformatics* 35, 5191–5198. doi:10.1093/bioinformatics/btz418

Zhang, W., Xu, H., Li, X., Gao, Q., and Wang, L. (2020). DRIMC: an Improved Drug Repositioning Approach Using Bayesian Inductive Matrix Completion. *Bioinformatics* 36, 2839–2847. doi:10.1093/bioinformatics/btaa062

Zhang, Y., Lei, X., Fang, Z., and Pan, Y. (2020). CircRNA-disease Associations Prediction Based on Metapath2vec++ and Matrix Factorization. *Big Data Min. Anal.* 3, 280–291. doi:10.26599/bdma.2020.9020025

Zhou, R., Lu, Z., Luo, H., Xiang, J., Zeng, M., and Li, M. (2020). NEDD: a Network Embedding Based Method for Predicting Drug-Disease Associations. *BMC Bioinformatics* 21, 387. doi:10.1186/s12859-020-03682-4

Zhou, X., Menche, J., Barabási, A. L., and Sharma, A. (2014). Human Symptoms-Disease Network. *Nat. Commun.* 5, 4212. doi:10.1038/ncomms5212

Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C., and Wang, F. (2020). Knowledge-driven Drug Repurposing Using a Comprehensive Drug Knowledge Graph. *Health Inform. J* 26, 2737–2750. doi:10.1177/1460458220937101

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

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