OXFORD

Briefings in Bioinformatics, 2022, **23**, 1–15

https://doi.org/10.1093/bib/bbab515

Problem Solving Protocol

HINGRL: predicting drug–disease associations with graph representation learning on heterogeneous information networks

Bo-Wei Zhao, Lun Hu, Zhu-Hong You, Lei Wang and Xiao-Rui Su

Corresponding author. Lun Hu, The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China. Fax: +86 991-3838957; E-mail: hulun@ms.xjb.ac.cn

Abstract

Identifying new indications for drugs plays an essential role at many phases of drug research and development. Computational methods are regarded as an effective way to associate drugs with new indications. However, most of them complete their tasks by constructing a variety of heterogeneous networks without considering the biological knowledge of drugs and diseases, which are believed to be useful for improving the accuracy of drug repositioning. To this end, a novel heterogeneous information network (HIN) based model, namely HINGRL, is proposed to precisely identify new indications for drugs based on graph representation learning techniques. More specifically, HINGRL first constructs a HIN by integrating drug–disease, drug–protein and protein–disease biological networks with the biological knowledge of drugs and diseases. Then, different representation strategies are applied to learn the features of nodes in the HIN from the topological and biological perspectives. Finally, HINGRL adopts a Random Forest classifier to predict unknown drug–disease associations based on the integrated features of drugs and diseases obtained in the previous step. Experimental results demonstrate that HINGRL achieves the best performance on two real datasets when compared with state-of-the-art models. Besides, our case studies indicate that the simultaneous consideration of network topology and biological knowledge of drugs and diseases allows HINGRL to precisely predict drug–disease associations from a more comprehensive perspective. The promising performance of HINGRL also reveals that the utilization of rich heterogeneous information provides an alternative view for HINGRL to identify novel drug–disease associations especially for new diseases.

Keywords: drug–disease associations, prediction, heterogeneous information network, graph representation learning, drug repositioning

Introduction

The traditional process of drug discovery suffers from the disadvantages of being labor-intensive, time-consuming and high-risk. Discovering a new drug normally takes more than 10 years from development to clinical use, and the corresponding cost is between $500 million and $2 billion, or more [1]. Nevertheless, only less than 10% of new drugs have been approved for clinical use [2, 3]. In this regard, drug repositioning has attracted increasing attention in the pharmaceutical industry, and has achieved successful applications over the past years. For example, sildenafil was originally utilized to treat the cardiovascular disease, but later it was found to have an effect on the erectile function of male patients [4].

Traditional drug repositioning approaches target to find abnormal clinical manifestations by manually screening clinical drug databases, and they require a large number of testing experiments on the targeted drugs. Recently, due to the increased accumulation of high-throughput genomics and proteomics data, much more attention has been given to develop different computational methods based on data mining techniques [5].

Bo-Wei Zhao is a PhD candidate at the University of Chinese Academy of Sciences and the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences.

Lun Hu received the B.Eng. degree from the Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2006, and the M.Sc. and Ph.D. degrees from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2008 and 2015, respectively. He joined the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, China, in 2020 as a professor of computer science. His research interests include machine learning, complex network analytics and their applications in bioinformatics.

Zhu-Hong You received his B.E. degree in Electronic Information Science and Engineering from Hunan Normal University, Changsha, China, in 2005. He obtained his Ph.D. degree in control science and engineering from University of Science and Technology of China (USTC), Hefei, China, in 2010. From June 2008 to November 2009, he was a visiting research fellow at the Center of Biotechnology and Information, Cornell University. He is currently a professor with Northwestern Polytechnical University, Xi'an, China. His current research interests include neural networks, intelligent information processing, sparse representation, and its applications in bioinformatics.

Lei Wang received the Ph.D. degree from the School of Computer Science Technology, China University of Mining and Technology, Jiangsu, China, in 2018. He is currently a Professor with Guangxi Academy of Science, Nanning, China. His research interests include data mining, pattern recognition, machine learning, deep learning, computational biology, and bioinformatics.

Xiao-Rui Su is a PhD candidate at the University of Chinese Academy of Sciences and the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences.

Received: September 9, 2021. **Revised:** November 8, 2021. **Accepted:** November 9, 2021

The main reason for the popularity of computational methods is because of its advantages of low-cost and high-efficiency.

At present, existing computational methods proposed for drug repositioning are classified into four categories, including recommender system-based methods, machine learning-based methods, deep learning-based methods and network-based methods [6]. Recommender system-based methods consider the identification of potential drug indications as a recommendation task and mainly adopt the matrix factorization approach to complete their tasks [7–10]. Although effective, these methods are not applicable to make an accurate prediction for new drugs or diseases. Machine learning-based methods are widely applied to predict associations between drugs and diseases [11, 12]. However, they heavily rely on the input data that is assumed to well represent the characteristics of drugs and diseases, and such assumption is difficult to satisfy in practical applications. Taking advantage of its powerful learning ability, deep learning-based methods can directly transform the original data into abstract feature representation [13, 14]. Although they are able to address the incompleteness problem of manually curated features [15], a large amount of training data is required for them to obtain high accuracy. In other words, deep learning-based methods are prone to over-fitting if the input drug–disease association network is sparse.

Network-based methods are widely applied for drug repositioning [16–19]. Their performances have been verified to be better than those in the other three categories, as they improve the accuracy of drug repositioning by capturing similar information across different kinds of biological networks as the features of drugs and diseases [20]. To do so, heterogeneous networks are introduced to represent the integration of different kinds of biological networks, and the similarities preserved across different biological networks gain new insight into the prediction of unobserved associations between drugs and diseases. However, network-based methods concentrate on constructing various heterogeneous networks while ignoring the intrinsic characteristics of different kinds of molecules, thus making it difficult to fully exploit the potential knowledge of biological networks for accurate drug repositioning. Previous studies have shown that the additional consideration of node attributes is of great significance in conducting an accuracy analysis for complex networks [21–25], but few attempts have been made in drug repositioning by simultaneously considering network topology and biological knowledge of drugs and diseases in the same heterogeneous network. A major reason for that phenomenon is the lack of a general model that possesses the ability of properly handling these two kinds of information for predicting the associations between drugs and diseases.

Furthermore, most of existing drug repositioning methods ignore the critical role of proteins when discovering novel associations between drugs and diseases. As has been pointed out by [26], proteins are an active macromolecule in biological cells. The change in protein expressions is directly related to disease manifestation and drug action. Specifically, drugs improve disease symptoms by acting on enzymes in living organisms. Taking valproic acid as an example, the expression of histone proteins is affected in cells, thus changing the life cycle of breast cancer cells [27]. In this regard, it is of great significance to introduce proteins to predict the relationship between drugs and diseases. Moreover, giving the fact that biological networks composed of drugs and diseases are normally sparse, the connectivity between drugs and diseases can thus be enhanced if protein–drug and protein–diseases association are integrated into these networks.

To address these challenges, a novel model, namely HINGRL, is proposed to integrate network topology and biological knowledge of drugs and diseases for drug repositioning. To distinguish from existing network-based methods that focus on heterogeneous networks, a heterogeneous information network (HIN) is introduced for the additional consideration of biological knowledge. More specifically, HINGRL first integrates three kinds of biological networks including drug–disease, drug–protein and protein–disease networks, to obtain a HIN with the biological information of drugs and diseases collected from drug structures and semantic knowledge graphs of disease, respectively. After that, different representation learning techniques are adopted by HINGRL to learn the features of nodes in the HIN from the topological and biological perspectives. In particular, the biological knowledge of drugs and diseases is processed by using different metrics in order to obtain similarity matrices and then autoencoders are applied to construct the biological feature vectors of drugs and diseases in a more concise manner. To properly handling the information of network topology, a well-established graph representation learning algorithm, i.e. DeepWalk, is adopted such that the network representations of drugs and diseases can be learned from the topological perspective. After that, the biological and topological representations of drugs and diseases obtained from the given HIN are concatenated together to compose integrated feature vectors of drugs and diseases, which are then considered as the input of a Random Forest (RF) classifier to complete the task of predicting potential drug–disease associations. Experimental results demonstrate that HINGRL performs better in terms of several independent metrics on two real datasets when compared with state-of-the-art prediction models proposed for drug repositioning. The overall workflow of HINGRL is presented in Figure 1. The main contributions of this work are summarized as:

- Rich heterogeneous information, i.e. protein-related associations and biological knowledge of drugs and diseases, is integrated to capture the representations of drugs and diseases from a comprehensive perspective.

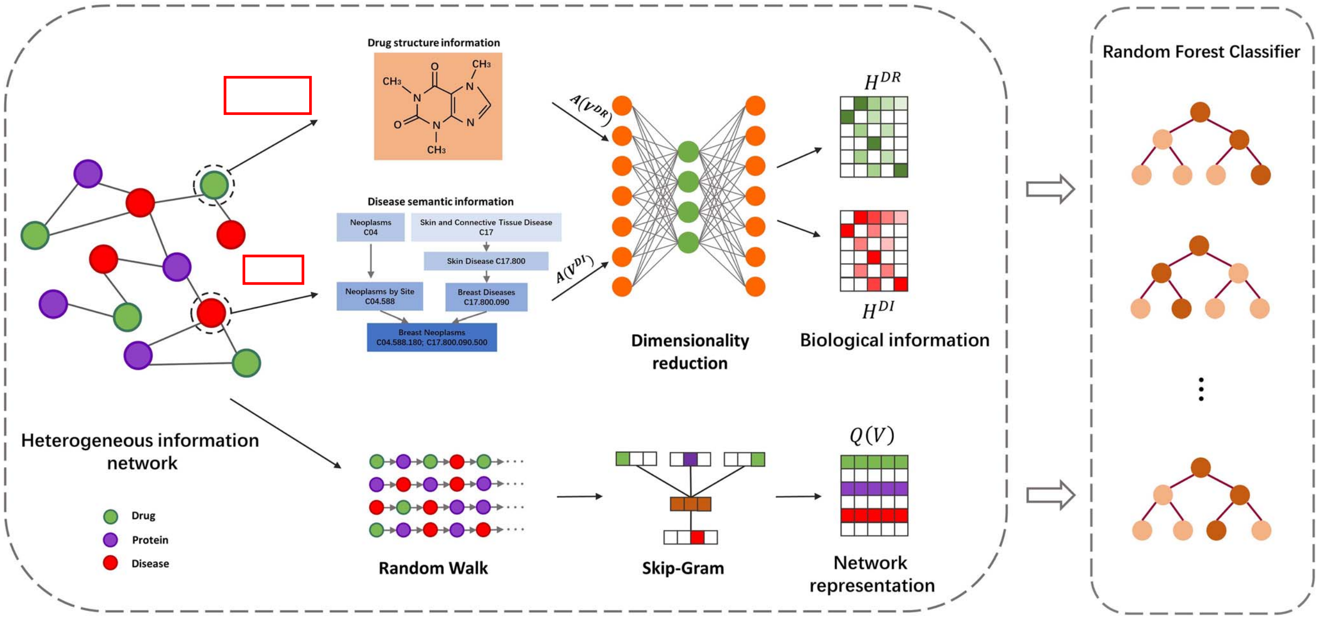
Drug structure information
CH3
CH3
O
N
N
O
N
N
CH3
Disease semantic information
Neoplasms C04
Skin and Connective Tissue Disease C17
Skin Disease C17.800
Neoplasms by Site C04.588
Breast Diseases C17.800.090
Breast Neoplasms C04.588.180; C17.800.090.500
$A(V^{DR})$
$A(V^{DI})$
Dimensionality reduction
H^{DR}
H^{DI}
Biological information
Heterogeneous information network
Drug
Protein
Disease
Random Walk
Skip-Gram
$Q(V)$
Network representation
Random Forest Classifier

Figure 1. The overall workflow of HINGRL.

- A novel HIN-based model, namely HINGRL, is proposed to precisely identify new indications for drugs. Different graph representation learning techniques are adopted by HINGRL to better learn the integrated features of drugs and diseases by simultaneous considering network topology and biological knowledge of drugs and diseases.
- Experimental results demonstrate that HINGRL outperforms several state-of-the-art algorithms on two benchmark datasets of drug repositioning. The promising performance of HINGRL also reveals that the utilization of rich heterogeneous information allows HINGRL to identify novel drug indications especially for new diseases without any known associations.

Materials and methods

Dataset

To construct a HIN for performance evaluation, we adopt a benchmark dataset, namely B-dataset, composed of three kinds of biological association networks, including drug–disease, drug–protein and protein–disease associations. Among them, the drug–disease association network is obtained from the CTD database [28] by Zhang *et al.* [29], and it contains 269 drugs, 598 diseases and 18 416 known drug–disease associations. The drug–protein association network is collected from the DrugBank database [30], and it is composed of 969 drugs, 613 proteins and 11 107 verified drug–protein associations. The protein–disease association network is derived from the DisGeNET database [31], and there are 832 proteins, 692 diseases and 25 087 protein–disease associations in it.

Moreover, to better demonstrate the generalization ability of HINGRL, we also evaluate its performance on another benchmark dataset, namely F-dataset, obtained from Gottlieb *et al.* [11]. F-dataset is much sparser than B-dataset in terms of the amount of drug–disease associations, as it only includes 593 drugs, 313 diseases and 1933 drug–disease interactions. Regarding the drug–protein and protein–diseases associations in F-dataset, we download them from the DrugBank and DisGeNET databases, respectively. By scanning these two databases, a total of 3243 drug–protein associations and 71 840 protein–disease associations are collected to compose the drug–protein and protein–disease association networks. To construct the set of negative samples from B-dataset and F-dataset, HINGRL randomly pairs up drugs and diseases whose associations are not found in the positive samples, and moreover the number of negative samples is equal to that of positive samples to avoid the unbalanced issue.

HIN modeling

As mentioned before, a HIN of interest is composed by drug–disease, drug–protein and protein–disease association networks. Obviously, there are two kinds of information available in the HIN, one is the biological knowledge of drugs and diseases and the other is the network topology. To model a HIN, we introduce a three-element tuple, i.e. HIN $= \{\mathbf{V}, \mathbf{A}, \mathbf{E}\}$, where $\mathbf{V} = \{V^{DR}, V^{DI}, V^{PR}\}$ denotes all $|V|$ nodes including drugs (V^{DR}), diseases (V^{DI}) and proteins (V^{PR}), $\mathbf{A} = \{A^{DR}, A^{DI}\}$ is the biological information of drugs and diseases and $\mathbf{E} = \{E^{DD}, E^{DP}, E^{PD}\}$ represents the collection of all drug–disease associations (E^{DD}), drug–protein associations (E^{DP}) and protein–disease associations (E^{PD}) in a HIN. Assuming that N is the number of drugs, K is the number of biological attributes of drugs and M is the number of diseases, we have $\mathbf{A}^{DR} \in \mathbb{R}^{N\times K}$ as a $N \times K$ matrix and $\mathbf{A}^{DI} \in \mathbb{R}^{M\times M}$ as a $M \times M$ matrix.

Biological knowledge extraction for drugs and diseases

Regarding the biological knowledge of drugs, since drug molecules with similar chemical structures are normally involved in the same biological activities [32], we make use of such chemical information as the biological attributes of drugs. To determine the chemical structures of each drug in a HIN, we first obtain its chemical descriptors from the Simplified Molecular Input Line Entry System (SMILES) [33] that can be downloaded in the DrugBank database (https://go.drugbank.com/). After that, we adopt the RDKit [34] tool to examine the existence of a particular chemical structure in drug molecules. Applying the same process to all drugs, we can obtain A^{DR}, each element of which has the value of 1 or 0 to indicate the existence of corresponding chemical structure. Note that there is a total of K chemical structures considered for A^{DR}.

Motivated by the observation that diseases are similar if the drugs they are associated with are also similar [35], we extract the biological information of diseases in light of medical subject descriptors collected from the Medical Subject Headings (MeSH) thesaurus [36]. In particular, the relationships among diseases are described by the MeSH tree structure and computed as representation vectors [37]. To do so, each disease is first described with a directed acyclic graph (DAG) by the MeSH descriptors, and then the similarity between two diseases, i.e. V_a^{DI} and V_b^{DI} $(a, b = 1, 2, \cdots, M)$, is calculated by the generalized Jaccard formula. Assuming that $\mathbf{DAG}_{V_a^{DI}} = (V_a^{DI}, F(V_a^{DI}), E(V_a^{DI}))$, where $F(V_a^{DI})$ denotes all ancestor nodes of disease V_a^{DI} and $E(V_a^{DI})$ is the set of all links of V_a^{DI}, the contribution of V_t^{DI} to V_a^{DI} in $\mathbf{DAG}_{V_a^{DI}}$ is $D(V_a^{DI})$ defined as below.

$$\begin{cases} D_{V_a^{DI}}\left(V_t^{DI}\right) = 1 \; \mathit{if} \; V_a^{DI} = V_t^{DI} \\ D_{V_a^{DI}}\left(V_t^{DI}\right) = \mathbf{max}\left\{\gamma \times D_{V_a^{DI}}\left(V_t^{DI\prime}\right) | V_t^{DI\prime} \in \mathbf{children\ of}\ V_t^{DI}\right\} \; \mathit{if} \; V_a^{DI} \neq V_t^{DI} \end{cases} \quad (1)$$

where γ is the factor of semantic contribution. Obviously, the contribution of V_t^{DI} is mainly driven by the distance between V_t^{DI} and V_a^{DI}. By summing up the contributions of all ancestors in $F(V_a^{DI})$, the semantic value of V_a^{DI} can be obtained with Equation (2).

$$DV\left(V_a^{DI}\right) = \sum_{V_t^{DI} \in F\left(V_a^{DI}\right)} D_{V_a^{DI}}\left(V_t^{DI}\right) \quad (2)$$

Combining Equations (1) and (2), the semantic similarity between V_a^{DI} and V_b^{DI} is calculated as:

$$\mathbf{Sim}\left(V_a^{DI}, V_b^{DI}\right) = \frac{\sum_{V_t^{DI} \in F\left(V_a^{DI}\right) \cap F\left(V_b^{DI}\right)} \left(D_{V_a^{DI}}\left(V_t^{DI}\right) + D_{V_b^{DI}}\left(V_t^{DI}\right)\right)}{DV\left(V_a^{DI}\right) + DV\left(V_b^{DI}\right)} \quad (3)$$

where the contributions of V_t^{DI} made to V_a^{DI} and V_b^{DI} are denoted as $D_{V_a^{DI}}(V_t^{DI})$ and $D_{V_b^{DI}}(V_t^{DI})$, respectively.

Given V_a^{DI}, its attribute information is defined as the semantic similarities between it and the other diseases in the HIN. Assuming that A_a^{DI} is the corresponding row of V_a^{DI} in A^{DI}, we have that $A_a^{DI} = [\mathbf{Sim}(V_a^{DI}, V_b^{DI})]$ $(1 \leq b \leq M)$. Accordingly, A^{DI} is defined as follows.

$$A^{DI} = \left[A_1^{DI}, A_2^{DI}, \cdots, A_M^{DI}\right]^{\mathrm{T}} \quad (4)$$

It is worth noting that in the F-dataset, since the identifiers of diseases are not consistent with those used by MeSH, we could not able to obtain their MeSH descriptors. In this regard, each element in A^{DI} is set as 0 when we apply this step to the F-dataset.

Autoencoder-based dimension reduction

After obtaining A^{DR} and A^{DI}, HINGRL applies an unsupervised learning neural network model, i.e. autoencoder [38], to reduce the dimensions of A^{DR} and A^{DI} into a more concise representation. The advantage of using autoencoder is that it solves the problem of redundancy and sparsity in the original data. In this regard, it is anticipated to not only improve the generalization ability of HINGRL but also avoid the overfitting during training. In autoencoder, there are three layers including input layer, hidden layer and output layer. Specifically, the input and output layers denote the original and new feature spaces, respectively, whereas the hidden layer is to ensure that the loss in the conversion from the original space to the new one is minimized.

When we incorporate autoencoder into HINGRL for dimension reduction, the biological information of drugs and diseases, i.e. A^{DR} and A^{DI}, is considered as the input for the input layer. Since the dimensions of A^{DR} and A^{DI} are reduced with the same process, we take A^{DR} as an example to demonstrate the details of how to apply autoencoder. Assuming that d_1 is the number of neurons in the hidden layer, the weight matrix from the input layer to the hidden layer is defined as $W \in \mathbb{R}^{d_1 \times K}$. In our work, d_1 is set as 64. $H^{DR} \in \mathbb{R}^{d_1 \times N}$ is represented as the mapping result encoded in the new feature space with Equation (5).

$$H^{DR} = \sigma\left(WA\left(V^{DR}\right) + b\right) \quad (5)$$

In the above equation, b is the bias, W is the weight matrix from the input layer to the hidden layer and $\sigma(\bullet)$ is the activation function of neurons.

The purpose of using a decoder is to map the encoded feature $h \in H^{DR}$ back to the original space so as to reconstruct A^{DR}. Assuming that $(A^{DR})'$ is the reconstruction result of A^{DR}, we can obtain it as:

$$\left(A^{DR}\right)' = \sigma\left(W'H^{DR} + b'\right) \quad (6)$$

where b' is the bias and W' is the weight matrix from the hidden layer to the output layer.

During the learning of new encoded features, the autoencoder model is trained by continuously minimizing the loss between A^{DR} and $(A^{DR})'$. The weight matrices, i.e. W and W', are alternatively optimized by using a gradient descent algorithm. The loss function of autoencoder used by HINGRL is defined as follows.

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}\left\|A\left(V^{DR}\right)_i - A\left(V^{DR}\right)'_i\right\|^2 \tag{7}$$

After the optimization process, H^{DR} is considered as the reduced biological information of drugs. Similarly, we can also obtain H^{DI} as the reduced biological information of diseases derived from A^{DI}. At the end of this step, a $(N+M)\times d_1$ matrix $\mathbf{H} = [H^{DR}, H^{DI}]^{\mathrm{T}}$ is obtained to represent the biological information of drugs and diseases in a more concise manner.

Heterogeneous network representation of drugs and diseases

Unlike the biological information that only involves individual drugs and diseases, the network topology information observed in a HIN is more complicated, as it represents the relationship between pairwise nodes. Hence, it is essential to incorporate such information into HINGRL in light of network structure. To do so, HINGRL extracts the network representations of drugs and diseases from a given HIN with DeepWalk [39], which is an effective graph representation learning algorithm. DeepWalk takes pairwise nodes as input and learns the sequence representation of each node by following the random walk theory. The output of DeepWalk are the corresponding representation vectors of nodes obtained from a skip-gram model. In a HIN, assuming that a random walk sequence from v_0 to v_{i-1} $(1 \le i \le |V|)$ is denoted as $\{v_0, v_1, \cdots, v_{i-1}\}$, the probability that the next node to arrive is v_i is defined as:

$$\Pr\left(v_i|\left(v_0, \cdots, v_{i-1}\right)\right) \tag{11}$$

We aim to obtain a vector representation for each node in V, and a mapping function $\Phi : v \in V \to \mathbb{R}^{|V|\times d_2}$ is introduced for this purpose. More specifically, a $|V| \times d_2$ matrix is denoted as the potential representation of each drug (disease) in a d_2-dimensional space. Here, $d_2 = 64$. In this way, the above equation can be rewritten as:

$$\Pr\left(v_i|\left(\Phi\left(v_0\right), \cdots, \Phi\left(v_{i-1}\right)\right)\right) \tag{12}$$

Finally, a skip-gram model is adopted to calculate Equation (12) as indicated by the following equation.

$$\underset{\Phi}{\text{minimize}} \quad -\log\Pr(\{v_{i-w}, \cdots, v_{i-1}, v_{i+1}, \cdots, v_{i+w}\}$$

$$|\Phi\left(v_i\right)) = \prod_{\substack{j=i-w,\\ j\neq i}}^{i+w} \Pr\left(v_j|\Phi(i)\right) \tag{13}$$

In Equation (13), w is the scope for determining the neighbor nodes of v_i. By solving the minimization problem of Equation (13), we could obtain $\Phi(V) \in \mathbb{R}^{|V|\times d_2}$ as the network representations for all nodes in V. In the rest of this paper, a $(N+M)\times d_2$ matrix $\mathbf{Q}$ is used to denote the representation vectors of drugs and diseases derived from $\Phi(V)$, and hence we have $\mathbf{Q} = \Phi(\{V^{DR}, V^{DI}\}) \in \mathbb{R}^{(N+M)\times d_2}$. Moreover, the loss functions of biological knowledge extraction and network representations of drugs and diseases are presented as Equations (7) and (13), respectively.

Drug repositioning via random forest classifier

According to the previous steps, HINGRL is able to extract two kinds of features for drugs and diseases from a given HIN, one is the biological representation denoted as $\mathbf{H}$ and the other is the network representation denoted as $\mathbf{Q}$. Hence, HINGRL concatenates these two matrices to compose an integrated matrix $\mathbf{X} \in \mathbb{R}^{(N+M)\times(d_1+d_2)}$, which is then used as the input to train a classifier for predicting unknown drug-disease associations. In particular, given an arbitrary node $v \in \{V^{DR}, V^{DI}\}$, its corresponding representation vectors in $\mathbf{H}$ and $\mathbf{Q}$ are denoted as $\mathbf{H}_v$ and $\mathbf{Q}_v$ respectively, and the final representation vector of v in $\mathbf{X}$ is $\mathbf{X}_v = [\mathbf{H}_v, \mathbf{Q}_v]$.

To complete the task of drug repositioning, HINGRL adopts the RF classifier. During the training phase, pairs of drugs and diseases compose the training dataset. For each pair, the representation vectors of its drug and disease are combined as the input of RF. Regarding the output, we introduce the matrix $\mathbf{P}$ to represent the prediction results between drugs and diseases whose associations are unknown in advance. The value of each element $\mathbf{P}$ is either 1 or 0, indicating that the association between the corresponding drug and disease is existed or not. A complete description about the procedure of HINGRL is presented in Algorithm 1.

Algorithm 1: The complete procedure of HINGRL.

```
Input: graph $HG(V, A, E)$.
    representation sizes: $d_1, d_2$
    the number of random walks: $n$
    random walk length $k$
    context size: $w$
    the number of trees: $t$
Output: the relationships matrix $\mathbf{P} \in \mathbb{R}^{E^{DD}}$ of node $v_i$ and node $v_j$, $v_i, v_j \in V$
Initialization: $\mathbf{P}$
Calculate the attribute similarity information of drugs $A(V^{DR})$
Calculate the attribute similarity information of diseases $A(V^{DI})$
Dimensionality reduction for $A(V^{DR})$ and $A(V^{DI})$
$H^{DR} = \textbf{\textit{AutoEncoder}}(A(V^{DR}), d_1)$
$H^{DI} = \textbf{\textit{AutoEncoder}}(A(V^{DI}), d_1)$
$\mathbf{H} = \begin{bmatrix} H^{DR} \\ H^{DI} \end{bmatrix}$
```

```
Learned the network representation of nodes
$\mathbf{Q} = \textbf{\textit{DeepWalk}}(E, d_2, n, k, w)$
Trained the prediction model by RF classifier
for each $e_{ij} =< v_i, v_j >\in E^{DD}$ do
    the features matrix of nodes $\mathbf{X} = [\mathbf{H}(V)\ \ \mathbf{Q}(V)]$
    $\mathbf{P} = \textbf{\textit{Random Forest Classifier}}([\ \ \mathbf{X}(v_i)\quad \mathbf{X}(v_j)\ \ ], t)$
end for
Predicted unknown drug-disease associations in P
```

Results and discussion

Evaluation metrics

To evaluate the accuracy of HINGRL, the receiver operating characteristic (ROC) curve is used. It is plotted by two variables including false positive rate and true positive rate. Considering the biased performance of arear under the curve (AUC) for imbalanced datasets, we also make use of the precision–recall (PR) curve to precisely reflect the actual performance of prediction models. AUC and AUPR are the areas under ROC and PR curves respectively, and they are used to quantitatively indicate the performance in terms of AUC and PR. Another two indicators, i.e. Matthews correlation coefficient (MCC) and F1-score, are also used to evaluate the overall performance of prediction models from different perspectives. In the experiments, the performance of HINGRL is evaluated by following a 10-fold cross-validation (CV) scheme. More specifically, we have performed an independent 10-fold CV to evaluate the performance of HINGRL on each of B-dataset and F-dataset. Taking B-dataset as an example, we first split it into split into 10-folds. For each fold, HINGRL is trained using the other 9-folds as training data, and then the resulting HINGRL model is validated on that fold. This procedure is repeated for 10 times by alternatively taking each fold as testing data.

Comparison with state-of-the-art algorithms

For the purpose of performance evaluation, we compare HINGRL with three state-of-the-art algorithms proposed for drug repositioning, i.e. LAGCN [14], DTINet [17] and deepDR [18]. Among them, LAGCN learns the embeddings of drugs and diseases from multiple networks through a graph convolution algorithm, and then adopts attention mechanisms to integrate these embeddings for predicting new associations. DTINet obtains the characteristic representations of drugs and proteins from different biological networks, and then searches for an optimal projection to force the feature vectors of drugs close to the known interacting proteins in the space. For deepDR, multiple drug-related heterogeneous networks are constructed to extract the features of drugs during repurposing, and then utilizes the random walk with restart algorithm to infer the potential indications of drugs by capturing the representations of these networks. One should note that all these three competing algorithms make use of drug–disease associations, but LAGCN additionally integrates the biological knowledge of drugs and diseases during repurposing.

Regarding the setting of parameters involved when running these algorithms, we adopt the default parameter settings for the competing models, i.e. LAGCN, deepDR and DTINet, as recommended in their original works for a fair comparison. Meanwhile, we conduct several trials with different settings and take the parameter values that obtain the best performance of HINGRL as the recommended setting. One should also note that all competing models are re-trained on each dataset by using the default parameter settings.

The experimental results of 10-fold CV on B-dataset and F-dataset are presented in Table 1 and Figures 2 and 3. We note that among all algorithms, HINGRL outperforms the other three algorithms across all datasets in terms of AUC, AUPR, MCC and F1-score. This could be a strong indicator that HINGRL is preferred over state-of-the-art algorithms when applied to drug repositioning. For HINGRL, its detailed results of 10-fold CV on B-dataset and F-dataset are shown in Supplementary material.

In addition to its superior accuracy, HINGRL is also more robust than the other algorithms as indicated by their evaluation scores. Taking deepDR as an example, its scores of AUC and AUPR are much larger than those of MCC and F1-score. Similar observations can also be made for DTINet and LAGCN. It is worth noting that deepDR and DTINet have higher precision scores, the reason for that phenomenon is that the number of positive samples correctly predicted by competing models is much less than that of HINGRL. In other words, HINGRL is preferred over competing models in terms of the ability of discovering novel drug–disease association as indicated by its superior performance in terms of Recall. But for HINGRL, its performance fluctuation across all the evaluation metrics is much less than the other three algorithms. There are two reasons accounting for the robustness of HINGRL. First, the introduction of heterogeneous information allows HINGRL to predict unknown drug–disease associations from different perspectives. Second, as an effective ensemble model, RF is adopted by HINGRL to complete the binary classification task, thus improving the robustness and generalization ability of HINGRL [40].

When compared with LAGCN that also makes use of the biological knowledge of drugs and diseases, HINGRL again demonstrates its advantage in drug repositioning. On average, HINGRL performs better by 0.45%, 73.20%, 40.95% and 67.64% than LAGCN in terms of AUC, AUPR, MCC and F1-score. Although the difference in AUC between HINGRL and LAGCN is moderate, HINGRL shows a bigger margin in AUPR, MCC and F1-score against LAGCN. The main reason for that phenomenon is due to the imbalance in our benchmark datasets, where the number of positive samples is much less than that of negative samples. Regarding the poor performance of LAGCN in terms of AUPR, we also perform an in-depth investigation into the experimental results obtained by LAGCN and find that LAGCN intends to assign smaller prediction probabilities to both positive and negative

Table 1. Experimental results of performance comparison on two benchmark datasets

Dataset	Methods	AUC	AUPR	MCC	F1-score		
					Precision	Recall	F1-score
B-dataset	deepDR	0.8205	0.8043	0.2987	0.8814	0.2345	0.3704
	DTINet	0.8324	0.8472	0.2994	**0.9710**	0.1783	0.3012
	LAGCN	0.8790	0.1448	0.1917	0.0689	0.6931	0.1253
	HINGRL	**0.8835**	**0.8768**	**0.6012**	0.7971	**0.8063**	**0.8017**
F-dataset	deepDR	0.8553	0.8871	0.5609	0.9564	0.5241	0.6762
	DTINet	0.8220	0.8721	0.2081	**1.0000**	0.0841	0.1545
	LAGCN	0.8462	0.0068	0.0542	0.0058	0.6653	0.0115
	HINGRL	**0.9363**	**0.9446**	**0.7340**	0.8868	**0.8402**	**0.8625**

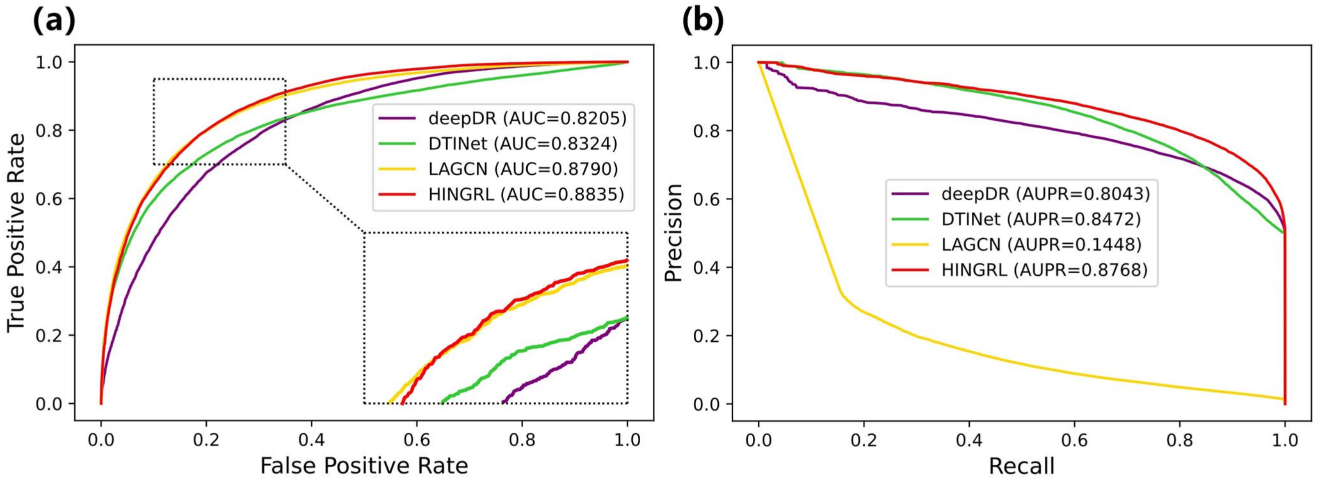

Figure 2. The ROC and PR curves of all algorithms on B-dataset, and they are presented in subfigures (**A**) and (**B**), respectively.

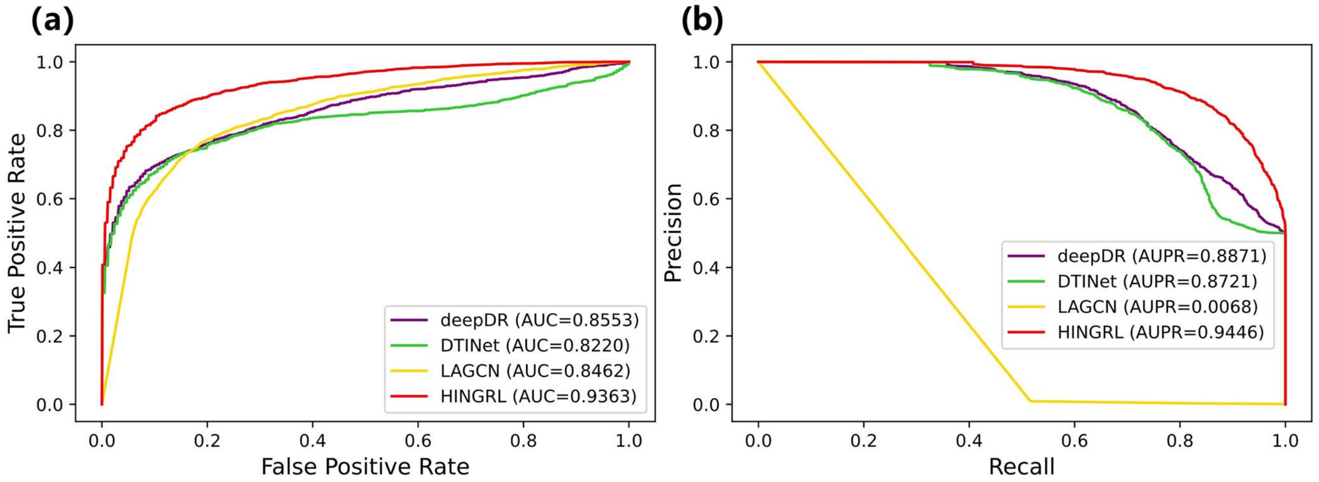

Figure 3. The ROC and PR curves of all algorithms on F-dataset, and they are presented in subfigures (**A**) and (**B**), respectively.

samples. It is for this reason that the AUC performance of LAGCN is much better than its AUPR performance especially for imbalanced datasets. This finding is consistent with the original work of LAGCN [14], where its AUPR performance is also poor. Moreover, the sparsity of HIN also accounts for the unsatisfactory performance of LAGCN in all the evaluation metrics except AUC, and accordingly the graph convolutional network used by LAGCN tends to over smooth when learning the representation from drug–disease association networks. But for HINGRL, the influence of sparsity is alleviated by using graph embedding, which is able to learn the representation of drugs and diseases from the perspective of network topology in a more effective way.

We also note that the scores of different evaluation metrics obtained by HINGRL from F-dataset are larger than those from B-dataset. The reasons for that phenomenon are two-fold: (1) the HIN constructed from F-dataset is much sparser than that from B-dataset, and accordingly fewer overlapping nodes are observed in the

Table 2. Experimental results of HINGRL-A, HINGRL-B and HINGRL on B-dataset

Feature	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision	Recall	F1-score
HINGRL-A	83.06 ± 0.55	82.20 ± 0.53	50.33 ± 1.21	74.58 ± 0.59	76.33 ± 1.10	75.44 ± 0.67
HINGRL-B	87.65 ± 0.45	86.68 ± 0.55	58.94 ± 1.06	79.38 ± 0.34	79.60 ± 1.21	79.49 ± 0.65
HINGRL	**88.35** ± **0.41**	**87.68** ± **0.51**	**60.12** ± **1.02**	**79.71** ± **0.53**	**80.63** ± **1.33**	**80.17** ± **0.62**

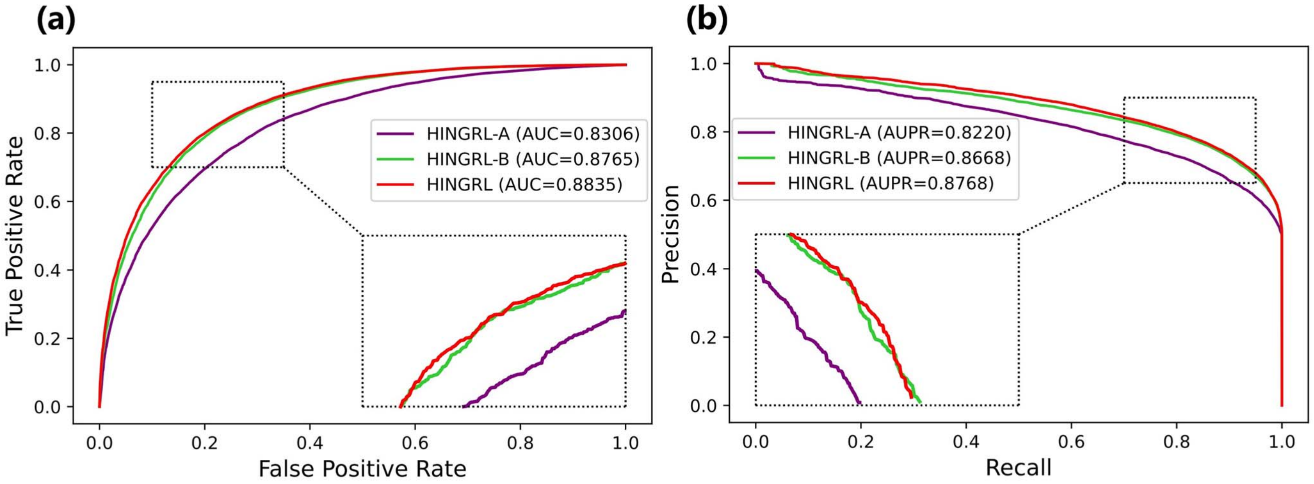

Figure 4. The ROC and PR curves of HINGRL-A, HINGRL-B and HINGRL on B-dataset.

random walk sequences involved in F-dataset; (2) after visualizing both B-dataset and F-dataset, we find that the modularity of the HIN constructed from F-dataset is better than that from B-dataset, thus making HINGRL able to learn the topological representation of nodes in a more effective manner. For F-dataset, each element in A^{DI} is set as 0, as the biological knowledge of diseases in the F-dataset is unavailable. Nevertheless, the introduction of heterogeneous information provides us an alternative view to complete the task of drug repositioning even some information is missed, thus enhancing the robustness of HINGRL.

Heterogeneous information influence on the performance of HINGRL

To better study the influence of heterogeneous information, we also implement two variants of HINGRL, i.e. HINGRL-A and HINGRL-B. In particular, HINGRL-A only considers the biological knowledge of drugs and diseases, whereas HINGRL-B additionally integrates the drug–disease association network on the basis of HINGRL-A. The RF classifiers used by these two variants are configured with the same parameters and their performances are also evaluated under 10-fold CV. Since HINGRL-A and HINGRL-B yield similar performances on B-dataset and F-dataset, we take the experimental results obtained from B-dataset as an example for analysis and present them in Table 2 and Figure 4, where several things can be noted.

First, the performance of HINGRL-A is the worst among HINGRL and its variants. In other words, only relying on the biological knowledge of drugs and diseases may not be sufficiently enough to achieve a promising performance for drug repositioning. Nevertheless, the consideration of biological knowledge provides a solid basis for the prediction accuracy of HINGRL. Since new diseases often encounter the situation that no associations are verified with existing drugs, this could be a strong indicator that HINGRL is particularly useful to identify novel indications for new drugs by only making use of their biological information. Second, after incorporating the drug–disease associations, HINGRL-B shows a bigger margin in performance against HINGRL-A in each evaluation metric. In particular, HINGRL-B performs better by 4.59%, 4.48%, 8.61% and 4.05% than HINGRL-A in terms of AUC, AUPR, MCC and F1-score, respectively. Hence, the network topology information represented by drug–disease associations allows HINGRL-B to better capture the characteristics of drugs and diseases when training the RF classifier. Lastly, a further improvement is observed from HINGRL by taking into account more heterogeneous association information, i.e. drug–protein and protein–disease associations, as HINGRL obtains the best performance across all evaluation metrics. In other words, protein-related associations enrich the heterogeneous information from the topological perspective, thus improving the network representations of drugs and diseases in determining **Q**.

Table 3. The performance of HINGRL by using different classifiers

Classifier	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision	Recall	F1-score
Gaussian NB	74.94 ± 0.71	71.65 ± 1.35	38.33 ± 1.32	69.07 ± 0.79	69.41 ± 1.00	69.24 ± 0.66
SVM	78.04 ± 0.72	76.80 ± 0.79	42.19 ± 1.48	70.83 ± 0.68	71.73 ± 1.39	71.27 ± 0.86
LR	78.69 ± 0.69	77.73 ± 0.56	42.75 ± 1.54	70.94 ± 0.82	72.41 ± 1.14	71.66 ± 0.79
KNN	80.17 ± 0.75	76.05 ± 1.01	45.19 ± 1.10	66.65 ± 0.57	**86.38 ± 0.74**	75.25 ± 0.44
RF	**88.35 ± 0.41**	**87.68 ± 0.51**	**60.12 ± 1.02**	**79.71 ± 0.53**	80.63 ± 1.33	**80.17 ± 0.62**

Classifier selection of HINGRL

Since there are many well-established classifiers, such as Gaussian Naïve Bayes (Gaussian NB), support vector machine (SVM), logistic regression (LR), K nearest neighbor (KNN) and RF, it is critical for us to select a proper classifier such that the best performance of HINGRL can be achieved. To this end, experiments have been conducted by comparing the performance of HINGRL with the use of different classifiers.

Regarding the hyperparameters setting of each machine learning algorithm, taking the KNN classifier as an example, the number of neighbors is of great significance to tune the performance of KNN and hence we conduct several trials by varying its value from 1 to 14 at a step size of 1 on B-dataset and F-dataset. The experimental results are shown in Supplementary Figure S1. We note that for the F-dataset, the best AUC performance of KNN is obtained when the number of neighbors is set as 9. Regarding B-dataset, the AUC performance of KNN is gradually improved when the number of neighbors becomes larger, but the increase in AUC is much smaller when the number of neighbors is larger than 9. Considering the AUC performance of KNN obtained on B-dataset and F-dataset, we reckon that the best performance of KNN is obtained when the number of neighbors is set as 9. By applying the similar tuning process to the other classifiers, we could also obtain their parameter settings with the best performance. In particular, the hyperparameters of all classifiers as shown in Supplementary Table S3.

The experimental results are presented in Table 3 and Figure 5. In general, HINGRL yields the best performance when using RF as its classifier. It is for this reason that we decide to incorporate RF into HINGRL for predicting novel drug–disease associations. Besides, there are several points worth further commentary.

First, among all classifiers, the performance of Gaussian NB is the worst. The main reason for its unsatisfactory performance is that Gaussian NB assumes the independence of features, which is difficult to be satisfied for the application of drug repositioning. Second, the performances of SVM and LR are fair, and thus the degree of nonlinearity in our datasets is yet to be verified. Third, although KNN is the second-best classifier, its ability of fault tolerance tends to become less efficient when the number of features increases. Lastly, as an efficient technique in ensemble learning, RF is preferred over the other classifiers due to its enhanced ability in processing high-dimensional data, which is the case of our datasets.

Graph representation learning selection of HINGRL

As we know, there are many graph representation learning methods that can well learn the network representation of biomolecules in biological information networks. To investigate their performance when integrating them with HINGRL, we compare five well-known graph representation learning methods, including graph convolution network (GCN) [41], LINE [42], SDNE [43], Node2vec [44] and DeepWalk on B-dataset and present the experimental results in Table 4 and Figure 6, where we note that DeepWalk yields a better performance than the other methods, thus indicating that DeepWalk is more suitable for learning the network representations of drugs and diseases in a HIN. Moreover, the performance of GCN is moderate because of its excessive smoothness, and the difference in the performance between LINE and SDNE is rather small due to the fact that they share similar ideas of learning network representations for nodes.

Generalization ability of HINGRL

Since B-dataset and F-dataset are two different datasets, the promising performance of HINGRL on them could be, to some extent, an indicator to demonstrate its generalization ability. To further investigate the generalization ability of HINGRL, we have conducted additional experiments. Rather than applying the HINGRL model trained on B-dataset to prediction the drug–disease associations in F-dataset, we adopt a different strategy by following [45, 46], which proposes to analyze the generalization ability on HINs with different sparsity by removing a certain proportion of drug–disease associations. The reason for this is due to the crucial constraint of DeepWalk, which requires DeepWalk to be retrained for learning the representations of new nodes in a given network [47]. In doing so, we expect that the generalization ability of HINGRL can be appreciated from an alternative perspective.

In our experiment, the proportion of drug–disease associations removed from the HINs of B-dataset and F-dataset is varied from 10% to 90% at a step size of 10%. The results obtained by HINGRL are shown in Tables 5 and 6. It is noted that the performance of HINGRL

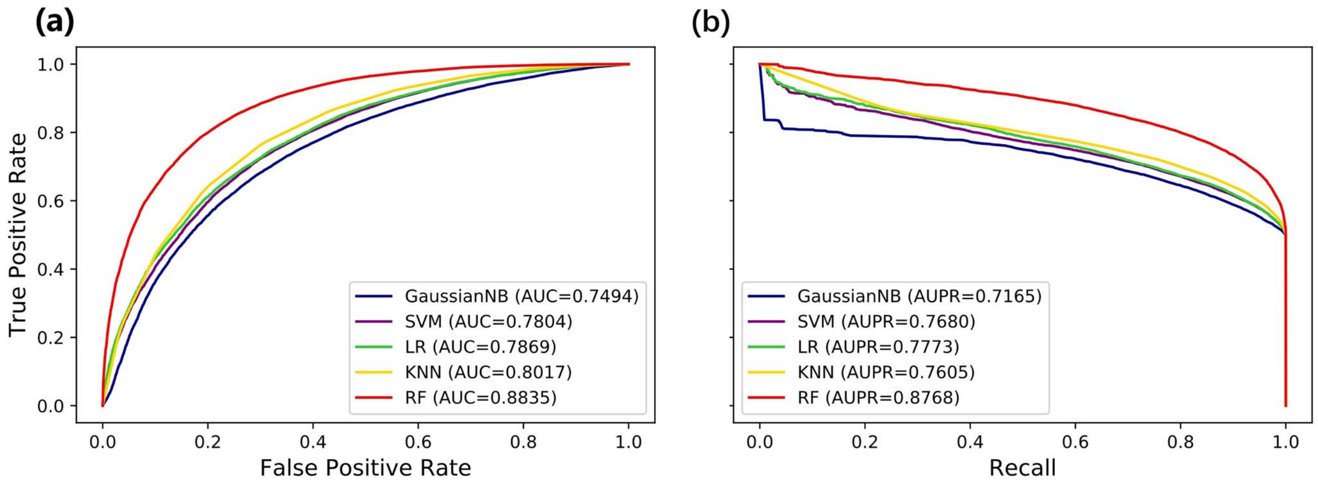

Figure 5. The ROC and PR curves of HINGRL by using different classifiers on B-dataset, and they are presented in subfigures (**A**) and (**B**), respectively.

Table 4. The performance of different graph representation learning of HINGRL on B-dataset

Classifier	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision	Recall	F1-score
HINGRL-GCN	83.52 ± 0.72	82.84 ± 0.79	50.88 ± 2.05	74.47 ± 0.86	77.34 ± 1.43	75.88 ± 1.09
HINGRL-LINE	86.28 ± 0.52	71.79 ± 0.55	56.13 ± 1.08	77.67 ± 0.62	78.76 ± 1.10	78.20 ± 0.60
HINGRL-SDNE	86.74 ± 0.51	72.48 ± 0.48	57.37 ± 0.97	78.40 ± 0.54	79.17 ± 1.15	78.78 ± 0.57
HINGRL-Node2vec	88.10 ± 0.47	73.65 ± 0.41	59.55 ± 0.86	79.88 ± 0.55	80.34 ± 1.25	79.88 ± 0.53
HINGRL-DeepWalk	88.35 ± 0.41	87.68 ± 0.51	60.12 ± 1.02	79.71 ± 0.53	80.63 ± 1.33	80.17 ± 0.62

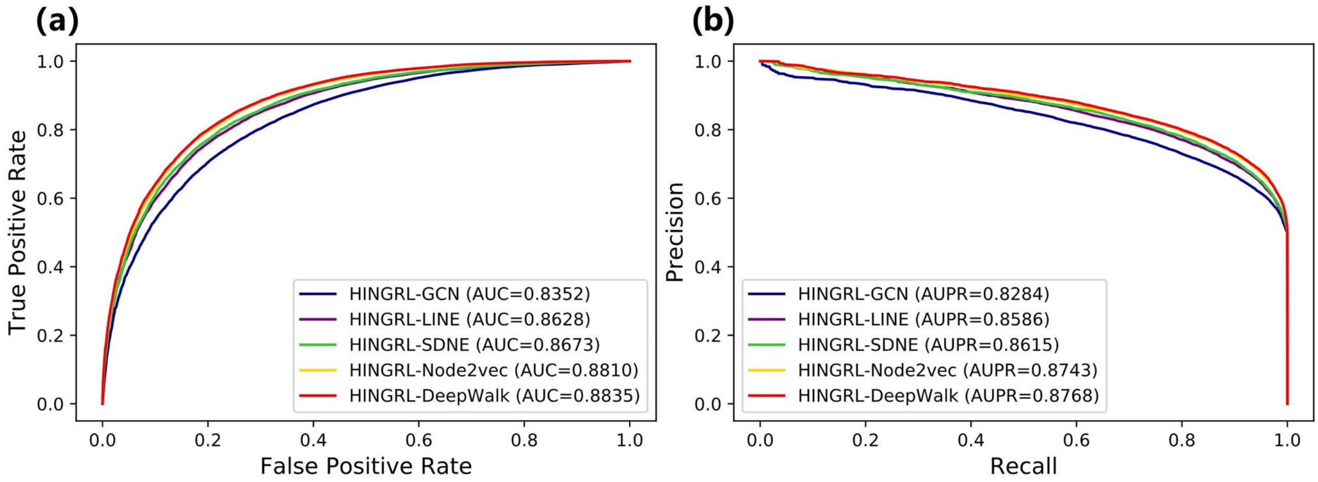

Figure 6. The ROC and PR curves of different graph representation learning of HINGRL on B-dataset, and they are presented in subfigures (**A**) and (**B**), respectively.

is improved when more drug–disease associations are involved in training. The main reason for that phenomenon is that the network representations of drugs and diseases can be enhanced by HINGRL if more heterogenous information about them are observed in training data. Moreover, when the proportions of removed drug–disease associations increases from 10% to 20%, the results of AUC, AUPR, MCC and F1-score only reduce on average by 1.16%, 1.855%, 3.87% and 2.215%, respectively, which verifies the generalization ability of HINGRL. In summary, although the generalization ability of HINGRL is heavily dependent on the size of common drugs and diseases shared by two datasets used for training and testing respectively, the consideration of heterogenous information alleviates the effect resulted from the constraint of DeepWalk.

Case study

To demonstrate the ability of HINGRL in discovering novel drug–disease associations, we have conducted additional experiments on the B-dataset. In particular,

Table 5. The performance comparison achieved of HINGRL by training different proportions on B-dataset

Fold	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision (%)	Recall (%)	F1-score (%)
10%	82.42	68.44	49.27	74.86	74.18	74.52
20%	84.56	70.20	52.87	76.43	76.45	76.44
30%	85.68	71.27	54.93	77.43	77.54	77.48
40%	86.46	72.05	56.4	78.21	78.18	78.20
50%	87.24	72.75	57.86	78.64	79.44	79.04
60%	87.72	73.13	58.62	78.91	79.99	79.45
70%	88.27	73.40	59.24	79.03	80.62	79.82
80%	88.77	74.28	60.90	79.77	81.57	80.66
90%	89.12	74.68	62.19	79.48	83.71	81.54

Table 6. The performance comparison achieved of HINGRL by training different proportions on F-dataset

Fold	AUC (%)	AUPR (%)	MCC (%)	F1-score (%)		
				Precision (%)	Recall (%)	F1-score (%)
10%	75.97	62.28	36.69	67.09	71.82	68.38
20%	83.40	67.63	48.16	73.27	75.77	74.50
30%	87.23	71.69	56.05	77.39	79.16	78.27
40%	89.84	74.58	61.17	80.38	80.93	80.65
50%	91.71	77.49	66.08	83.21	82.78	83.00
60%	92.59	79.89	69.62	85.90	83.25	84.55
70%	92.73	79.33	68.45	85.79	81.96	83.83
80%	92.85	79.63	69.35	85.49	83.51	84.49
90%	94.82	82.94	75.80	86.93	89.18	88.04

all known associations between drugs and diseases are used to compose the training dataset and then HINGRL is applied to verify unknown associations. An in-depth investigation into the experimental results is performed and several case studies are selected for further discussion as follows.

As one of the drugs for the treatment of schizophrenia, clozapine has been deeply studied by many pharmacological scientists because of its remarkable clinical efficacy [48]. In Table 7, the top 10 disease candidates are predicted by HINGRL to have associations with clozapine, and 5 of them have already been experimentally confirmed by the relevant literature. In order to verify the rationality behind the prediction results, we take anxiety disorders as an example to explain why it is a potential disease that can be cured by clozapine in theory. As has been pointed out by [49], anxiety disorders often occur as a common complication with schizophrenia due to the relationship between anxiety and the abnormal regulation of serotonin observed in the patients. Since clozapine can reduce the increase in serotonin caused by a noncompetitive antagonist of N-methyl-D-aspartate receptors [50], we have reason to believe that clozapine is likely to produce a pharmacological effect for anxiety disorders. More evidences can be found in relevant databases. First, anxiety disorders and pain are two similar diseases as indicated by the DisGeNET database, and a known association between pain and clozapine has existed in the HIN of B-dataset. Second, according to the DrugBank database, the chemical structures of olanzapine and clozapine are similar as their cosine similarity is as large as 0.7, and furthermore olanzapine and anxiety disorders are known to be associated in B-dataset. After investigating the prediction results obtained from deepDR, DTINet and LAGCN, none of them is able to identify this novel association. It could also a strong indicator for the ability of HINGRL in discovering novel associations for drugs and diseases.

Breast neoplasms are the most common symptom in the female population. The top 10 candidates of potential drugs predicted by HINGRL are shown in Tables 8 and 6 of them have been recorded in literature to be effective when used to treat breast neoplasms. Cocaine obtains the largest prediction score among all unverified drugs, and an in-depth analysis is given after a systematic literature review. As indicated by [51], celecoxib has an inhibitory effect on the growth of breast cancer cells containing cyclooxygenase-2, and it also has a verified association with breast neoplasms in B-dataset. According to the DrugBank database, celecoxib is associated with cocaine due to the fact that the combination of celecoxib and cocaine is able to slow down the metabolism of cells [30]. In this regard, our findings indicate a possible treatment for breast neoplasms by the collaboration of celecoxib and cocaine. For HINGRL, its reasons regarding the discovery of the association between cocaine and breast neoplasms are 2-fold: (1) there are many neighboring nodes, i.e. 45 diseases and 3 proteins, shared by cocaine

Table 7. The top 10 candidate drugs predicted by HINGRL for clozapine

Drug	Disease	MESH ID	Score	Evidence (PMID)
Clozapine	Headache	D006261	0.9919	16804270
	Ataxia	D001259	0.9829	31673444
	Anxiety disorders	D001008	0.9439	N/A
	Atrial fibrillation	D001281	0.9319	9555602
	Status epilepticus	D013226	0.9239	28632525
	Memory disorders	D008569	0.9219	N/A
	Sleep initiation and maintenance disorders	D007319	0.9109	N/A
	Peripheral nervous system diseases	D010523	0.9059	N/A
	Tachycardia, ventricular	D017180	0.9019	12503253
	Child behavior disorders	D002653	0.8978	N/A

Table 8. The top 10 candidate drugs predicted by HINGRL for breast neoplasms

Disease	Drug	DrugBank ID	Score	Evidence (PMID)
Breast neoplasms	Valproic acid	DB00313	0.9079	30075223
	Phenytoin	DB00252	0.8868	22678159
	Cocaine	DB00907	0.8458	N/A
	Methylprednisolone	DB00959	0.8388	12884026
	Phenobarbital	DB01174	0.8128	N/A
	Melatonin	DB01065	0.7737	19193248
	Streptozocin	DB00428	0.7597	N/A
	Acetaminophen	DB00316	0.7447	10048744
	Daunorubicin	DB00694	0.7397	18406070
	Diclofenac	DB00586	0.7177	N/A

and celecoxib in the HIN of B-dataset; and (2) celecoxib is associated with breast neoplasms. Since more network paths are existed between them during random walk, the representations of cocaine and celecoxib are more similar from the perspective of network topology. Furthermore, the introduction of protein-related associations also strengthens the connectivity between cocaine and celecoxib.

To explain why HINGRL successfully identify six verified drugs whose associations with breast neoplasms are unknown in the B-dataset, we compare their chemical structures with known drugs whose associations with breast neoplasms are already existed in the B-dataset, and adopt their Pearson coefficients in H^{DR} to indicate the similarities between verified drugs and known ones. The experimental results are presented in Figure 7, and we note that each of the verified drugs is highly similar to some of the known drugs according to the distribution of blocks with dark color. Moreover, we have also examined the experimental results of HINGRL-A, which is a variant of HINGRL that only utilizes the biological knowledge of diseases and drugs, and found that all verified drugs except valproic acid are predicted by HINGRL-A to have associations with breast neoplasms. In other words, HINGRL is able to identify these verified drugs for breast neoplasms solely from the perspective of biological knowledge.

In sum, these case studies again demonstrate the promising accuracy of HINGRL in drug repositioning, and hence it is believed that HINGRL could be a useful tool to discover novel drug–disease associations especially for new diseases without any known associations.

Conclusion

In this work, a novel HIN-based model, namely HINGRL, is proposed to predict potential drug–disease association based on graph representation learning techniques. To capture the features of drugs and disease from a more comprehensive perspective, HINGRL first integrates protein-related associations and the biological knowledge of drugs and diseases into the original drug–diseases association network, thus composing a complicated HIN. After that, different graph representation learning techniques are utilized by HINGRL to capture the targeted features of drugs and diseases from the perspectives of network topology and biological knowledge. HINGRL finally completes its prediction task by making use of the RF classifier. Experimental results on two benchmark datasets demonstrate that HINGRL yields a better performance than state-of-the-art drug repositioning algorithms in terms of accuracy and robustness. Our in-depth analysis of case study is also a strong indicator that HINGRL could be a useful tool to discover novel drug–disease associations especially for new diseases without any known associations. On the other hand, the promising performance of HINGRL reveals that the utilization of rich heterogeneous

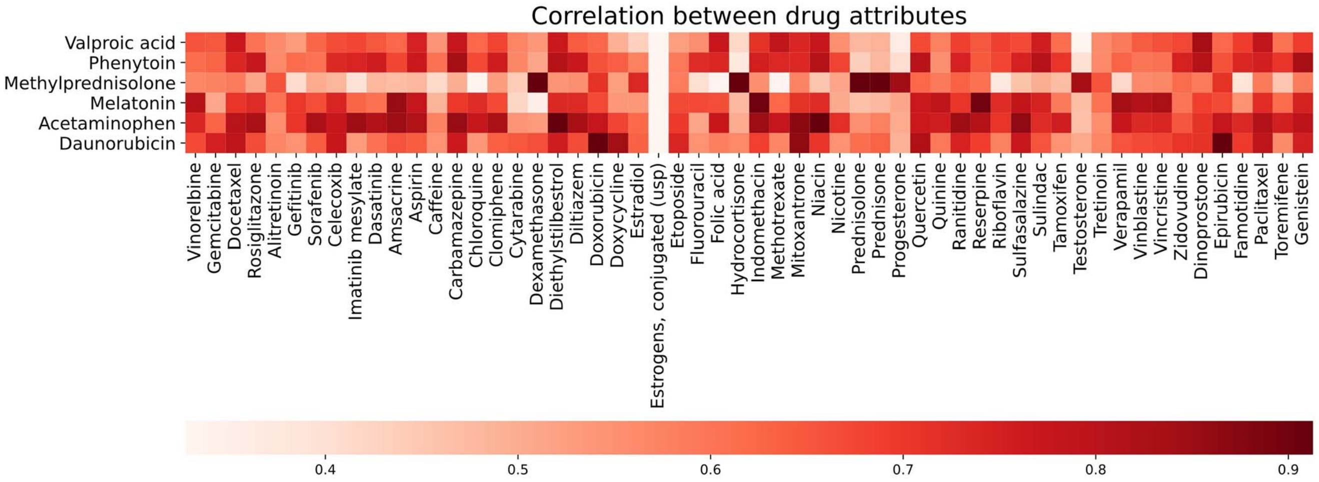

Figure 7. The similarity of attribute information between verified drugs and known drugs for breast neoplasms in B-dataset. The horizontal axis represents the known drugs, whereas the vertical axis represents the verified ones.

information allows HINGRL to achieve the goal of drug repositioning in a more effective manner.

Regarding the future work, we would like to extend our research from four aspects. First, we are interested in exploring the possibility of applying HINGRL to other relevant applications, such as protein–protein interaction prediction [52, 53], and miRNA–disease association prediction [54]. Second, regarding the construction of HIN, we intend to incorporate more specific information originated from the molecular mechanism of diseases and evaluate the importance of these heterogeneous information in drug repositioning. Third, we would like to improve the generalization ability of HINGRL by addressing the constraint of DeepWalk. Last, since there are many other kinds of biological network, we aim to explore the possibility of proposing a better model that can adaptively learn the representations of drugs and diseases in a more complicated HIN.

Key Points

- We integrate rich heterogeneous information, i.e. protein-related associations and biological knowledge of drugs and diseases, into a drug-disease association network and compose a HIN, where the representations of drugs and diseases can be captured from a comprehensive perspective.
- We propose a novel HIN-based model, namely HINGRL, is proposed to precisely identify new indications for drugs. Different graph representation learning techniques are adopted by HINGRL to better learn the integrated features of drugs and diseases by simultaneous considering network topology and biological knowledge of drugs and diseases.
- Experimental results demonstrate that HINGRL outperforms several state-of-the-art algorithms on two benchmark datasets of drug repositioning. The promising performance of HINGRL also reveals that the utilization of rich heterogeneous information allows HINGRL to identify novel drug indications especially for new diseases without any known associations.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors would like to thank all anonymous reviewers for their constructive advice.

Data availability

The data sets and source code can be freely downloaded from: https://github.com/stevejobws/HINGRL.

Funding

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (grant 2021D01D05), in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, in part by the National Natural Science Foundation of China (grants 62172355, 61702444), in part by Awardee of the NSFC Excellent Young Scholars Program (grants 61722212, 61902342), in part by the Tianshan youth – Excellent Youth (grant 2019Q029) and in part by the Qingtan scholar talent project of Zaozhuang University.

References

1. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really $802 million? *Health Aff* 2006;**25**:420–8.
2. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**: 673–83.
3. Li J, Zheng S, Chen B, *et al.* A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;**17**:2–12.
4. Goldstein I, Lue TF, Padma-Nathan H, *et al.* Oral sildenafil in the treatment of erectile dysfunction. *N Engl J Med* 1998;**338**: 1397–404.
5. Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Chem* 2020;**12**:1–23.
6. Luo H, Li M, Yang M, *et al.* Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief Bioinform* 2019;**22**(2):1604–19.
7. Dai W, Liu X, Gao Y, *et al.* Matrix factorization-based prediction of novel drug indications by integrating genomic space. *Comput Math Methods Med* 2015;**2015**:9. http://dx.doi.org/10.1155/2015/275045.
8. Zhang W, Zou H, Luo L, *et al.* Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* 2016;**173**:979–87.
9. Huang F, Qiu Y, Li Q, *et al.* Predicting drug-disease associations via multi-task learning based on collective matrix factorization. *Front Bioeng Biotechnol* 2020;**8**:218. doi: 10.3389/fbioe.2020.00218.
10. Luo H, Li M, Wang S, *et al.* Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**:1904–12.
11. Gottlieb A, Stein GY, Ruppin E, *et al.* PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**:496.
12. Wang Y, Chen S, Deng N, *et al.* Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* 2013;**8**:e78518.
13. Li Z, Huang Q, Chen X, *et al.* Identification of drug-disease associations using information of molecular structures and clinical symptoms via deep convolutional neural network. *Front Chem* 2020;**7**:924.
14. Yu Z, Huang F, Zhao X, *et al.* Predicting drug–disease associations through layer attention graph convolutional network. *Brief Bioinform* 2020;**22**(4). https://doi.org/10.1093/bib/bbaa243.
15. Zeng X, Zhu S, Lu W, *et al.* Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 2020;**11**:1775–97.
16. Luo H, Wang J, Li M, *et al.* Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* 2016;**32**:2664–71.
17. Luo Y, Zhao X, Zhou J, *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;**8**:1–13.
18. Zeng X, Zhu S, Liu X, *et al.* deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019;**35**:5191–8.
19. Chu Y, Wang X, Dai Q, *et al.* MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. *Brief Bioinform* 2021;**22**(6). https://doi.org/10.1093/bib/bbab165.
20. Yang M, Wu G, Zhao Q, *et al.* Computational drug repositioning based on multi-similarities bilinear matrix factorization. *Brief Bioinform* 2020;**22**(4). https://doi.org/10.1093/bib/bbaa267.
21. Hu L, Chan KC. Fuzzy clustering in a complex network based on content relevance and link structures. *IEEE Trans Fuzzy Syst* 2015;**24**:456–70.
22. Hu L, Chan KC, Yuan X, *et al.* A variational Bayesian framework for cluster analysis in a complex network. *IEEE Trans Knowl Data Eng* 2019;**32**:2115–28.
23. Hu L, Zhang J, Pan X, *et al.* HiSCF: leveraging higher-order structures for clustering analysis in biological networks. *Bioinformatics* 2021;**37**:542–50.
24. Chu Y, Kaushik AC, Wang X, *et al.* DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief Bioinform* 2021;**22**:451–62.
25. Dai Q, Chu Y, Li Z, *et al.* MDA-CF: predicting MiRNA-disease associations based on a cascade forest model by fusing multi-source information. *Comput Biol Med* 2021;**136**:104706.
26. Hu L, Wang X, Huang Y-A, *et al.* A survey on computational models for predicting protein–protein interactions. *Brief Bioinform* 2021;**22**(5). https://doi.org/10.1093/bib/bbab036.
27. Aztopal N, Erkisa M, Erturk E, *et al.* Valproic acid, a histone deacetylase inhibitor, induces apoptosis in breast cancer stem cells. *Chem Biol Interact* 2018;**280**:51–8.
28. Davis AP, Grondin CJ, Johnson RJ, *et al.* The comparative toxicogenomics database: update 2017. *Nucleic Acids Res* 2017;**45**: D972–8.
29. Zhang W, Yue X, Lin W, *et al.* Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 2018;**19**:1–12.
30. Wishart DS, Feunang YD, Guo AC, *et al.* Drug Bank 5.0: a major update to the Drug Bank database for 2018. *Nucleic Acids Res* 2017;**46**:D1074–82.
31. Piñero J, Bravo À, Queralt-Rosinach N, *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2016;**45**(D1):D833–D839.
32. Huang L, Luo H, Li S, *et al.* Drug–drug similarity measure and its applications. *Brief Bioinform* 2021;**22**. doi: 10.1093/bib/bbaa265.
33. Weininger DSMILES. A chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6.
34. Landrum G. Rdkit documentation. *Release* 2013;**1**:1–79.
35. Yan S, Yang A, Kong S, *et al.* Predictive intelligence powered attentional stacking matrix factorization algorithm for the computational drug repositioning. *Appl Soft Comput* 2021;**110**:107633.
36. Guo Z-H, You Z-H, Huang D-S, *et al.* MeSHHeading2vec: a new method for representing MeSH headings as vectors based on graph embedding algorithm. *Brief Bioinform* 2021;**22**:2085–95.
37. Wang L, You Z-H, Huang D-S, *et al.* MGRCDA: metagraph recommendation method for predicting CircRNA-disease association, IEEE transactions on. *Cybernetics* 2021;1–9. doi: 10.1109/TCYB.2021.3090756.
38. Liou C-Y, Cheng W-C, Liou J-W, *et al.* Autoencoder for words. *Neurocomputing* 2014;**139**:84–96.
39. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014, p. 701–10. ACM.
40. Hu L, Chan KC. Extracting coevolutionary features from protein sequences for predicting protein-protein interactions. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**14**:155–66.
41. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv e-prints. 2016, arXiv:1609. 02907.
42. Tang J, Qu M, Wang M, *et al.* Line: large-scale information network embedding. In: *Proceedings of the 24th International Conference*

on World Wide Web. International World Wide Web Conferences Steering Committee, 2015, 1067–77.

43. Wang D, Cui P, Zhu W. Structural deep network embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, p. 1225–34.
44. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, p. 855–64.
45. Wang X, Xin B, Tan W, *et al.* DeepR2cov: Deep Representation Learning on Heterogeneous Drug Networks to Discover Anti-inflammatory Agents for COVID-19. *Briefings in Bioinformatics*. 2021;**22**(6). https://doi.org/10.1093/bib/bbab226.
46. Zhou S, Yue X, Xu X, *et al. LncRNA-miRNA interaction prediction from the heterogeneous network through graph embedding ensemble learning*. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, p. 622–7. IEEE.
47. Yu B, Zhang Y, Xie Y, *et al.* Influence-aware graph neural networks. *Applied Soft Computing* 2021;**104**:107169. https://doi.org/10.1016/j.asoc.2021.107169.
48. Konte B, Walters JT, Rujescu D, *et al.* HLA-DQB1 6672G> C (rs113332494) is associated with clozapine-induced neutropenia and agranulocytosis in individuals of European ancestry. *Transl Psychiatry* 2021;**11**:1–10.
49. Muller JE, Koen L, Seedat S, *et al.* Anxiety disorders and schizophrenia. *Curr Psychiatry Rep* 2004;**6**:255–61.
50. López-Gil X, Babot Z, Amargós-Bosch M, *et al.* Clozapine and haloperidol differently suppress the MK-801-increased glutamatergic and serotonergic transmission in the medial prefrontal cortex of the rat. *Neuropsychopharmacology* 2007;**32**: 2087–97.
51. Arun B, Zhang H, Mirza N, *et al.* Growth inhibition of breast cancer cells by celecoxib. *Breast Cancer Res Treat* 2001;**69**(3):234. http://www.scopus.com/inward/citedby.url?
52. Pan X, Hu L, Hu P, *et al.* Identifying protein complexes from protein-protein interaction networks based on fuzzy clustering and GO semantic information. *IEEE/ACM Trans Comput Biol Bioinform* 2021;1–1. doi: 10.1109/TCBB.2021.3095947.
53. Hu L, Yang S, Luo X, *et al.* A distributed framework for large-scale protein-protein interaction data analysis and prediction using map reduce. *IEEE/CAA J Autom Sin* 2021;**9**:160–72.
54. Huang Y-A, Hu P, Chan KC, *et al.* Graph convolution for predicting associations between miRNA and drug resistance. *Bioinformatics* 2020;**36**:851–8.