



A geometric deep learning framework for drug repositioning over heterogeneous information networks

Bo-Wei Zhao , Xiao-Rui Su , Peng-Wei Hu, Yu-Peng Ma, Xi Zhou and Lun Hu 

Corresponding author: Lun Hu, The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China.

E-mail: hulun@ms.xjb.ac.cn

Abstract

Drug repositioning (DR) is a promising strategy to discover new indicators of approved drugs with artificial intelligence techniques, thus improving traditional drug discovery and development. However, most of DR computational methods fall short of taking into account the non-Euclidean nature of biomedical network data. To overcome this problem, a deep learning framework, namely DDAGDL, is proposed to predict drug-drug associations (DDAs) by using geometric deep learning (GDL) over heterogeneous information network (HIN). Incorporating complex biological information into the topological structure of HIN, DDAGDL effectively learns the smoothed representations of drugs and diseases with an attention mechanism. Experiment results demonstrate the superior performance of DDAGDL on three real-world datasets under 10-fold cross-validation when compared with state-of-the-art DR methods in terms of several evaluation metrics. Our case studies and molecular docking experiments indicate that DDAGDL is a promising DR tool that gains new insights into exploiting the geometric prior knowledge for improved efficacy.

Keywords: drug repositioning, geometric deep learning, heterogeneous information network, drug-disease association prediction, artificial intelligence

Introduction

Due to the high risk of failure, traditional drug development process is expensive and time-consuming [1]. It has been reported that developing a novel drug from scratch costs around USD 1.24 billion through traditional drug development process [2]. Nevertheless, the ever-increasing demands on efficacy and safety are the main reasons contributing to the low success rate (< 10%) in bringing new drugs to the market [3]. Taking Alzheimer's disease as an example, it is the most common form of dementia, which affects more than 40 million people around the world with an increasing trend, but there are no pharmacological treatments that have been licensed for use in individuals with mild cognitive impairment [4]. Hence, finding more efficacious and safer drugs still presents a huge challenge to the scientific community.

As artificial intelligence techniques have been undergoing a rapid development, drug repositioning (DR), or drug repurposing, has attracted much attention as an alternative yet complementary strategy to discover new indicators for approved or experimental drugs, thus offering significant advantages to accelerate the drug development process by saving a lot of time and labor [5]. In recent years, a variety of computational-based DR methods

have been developed for discovering novel drug-disease associations (DDAs) from a biomedical data point of view [6]. These methods normally extract desired features from the biomedical data related to drugs and diseases, and then incorporate the features into well-established classifiers for achieving the DR task. In light of feature extraction, they can be broadly classified into the categories of either machine learning (ML)-based or deep learning (DL)-based.

ML-based DR methods target to apply different ML techniques, such as matrix factorization [7, 8], support vector machines (SVM) [9] and neural networks [10], to predict unknown DDAs with extracted shallow features. For instance, DTINet [8] adopts matrix factorization to decompose the high-dimensional DDA matrix into the product of two low-dimensional matrices, where the feature vectors of drugs and diseases are extracted in a more compact form for accurately discovering unknown DDAs. However, shallow features used by ML-based methods are limited in their ability to represent drugs and diseases at a highly abstract level. To overcome this issue, several DL-based DR methods [11–17] have been developed by taking advantage of the powerful representation learning ability of DL. We note from [13] that deep learning-

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.

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.

Peng-Wei Hu is a professor in Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, China. His research interests include machine learning, big data analysis and its applications in bioinformatics.

Yu-Peng Ma is a professor in Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, China. His research interests include internet of Things application technology and big data analysis.

Xi Zhou is a professor in Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, China. His research interests include machine learning and big data analysis.

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.

Received: June 5, 2022. **Revised:** August 1, 2022. **Accepted:** August 9, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

based methods are believed to have excellent advantages over traditional machine learning-based methods in addressing drug repositioning tasks, particularly in processing different types of input data involved in drug repositioning according to the nature of their frameworks. As a representative work in this field, deepDR [14] first integrates different kinds of drug-related associations by modeling them with matrices, and then applies a multi-modal deep autoencoder to learn the feature representations of drugs and diseases, with which a variational autoencoder is constructed to infer novel indicators for approved drugs. SKCNN [15] combines a convolutional neural network with sigmoid kernel to effectively learn the features of drugs and diseases for DDA prediction. CBPreD [16] integrates a convolutional neural network (CNN) and a bidirectional long short-term memory (BiLSTM) to predict DDAs, where the CNN is used to learn the original representation of drugs and diseases base on their similarities and DDAs, and the BiLSTM is used to learn the path representations of DDAs. SkipGNN [17] first initializes embeddings by using a node2vec algorithm, and then the modified graph convolutional networks are applied to obtain final embeddings to complete the molecular interactions predictions. DeepR2cov [10] constructs multiple meta-paths to automatically learn low-dimensional vectors of drugs by deep neural networks, and then successfully discovers anti-inflammatory agents for COVID-19. Though effective, these methods have only demonstrated their success on curated biomedical data where an underlying Euclidean structure is observed, as they model DDAs as the functions related to the points of drugs and targets in the Euclidean space.

Recently, there has been an increasing interest in applying representation learning of drugs and diseases over heterogeneous information networks (HINs) in which there is a non-Euclidean geometric property. In particular, HINs of interest are graph models composed of not only drugs, diseases and related associations, but also their biological information denoted as the signals of corresponding nodes. Obviously, the structure of HINs is of great significance to reveal certain properties on their non-Euclidean domains, but both ML-based and DL-based DR methods fall short of capturing such information, and thereby lead to their unsatisfactory performance when applied to HINs [18]. Hence, certain effort has been devoted to apply geometric deep learning (GDL) for better learning the representations of drugs and diseases over HIN [19, 20]. As an emerging technique, GDL attempts to generalize deep neural networks to graph data in the non-Euclidean domains. For instance, DRHGCN [20] adopts multiple graph convolutional layers to learn the embedding representations of drugs and diseases by integrating three kinds of networks, including drug-disease, drug similarity and disease similarity networks. However, the over-smoothing issue resulted from the aggregation of neighborhood information within n -hops diminishes the discriminative ability of drug and disease representations learned by GDL [21]. Besides, for most GDL-based DR methods, equal contributions are taken for granted in aggregating neighbor representations to compose the final representations of drugs and diseases, but such an operation fails to capture significant features that are more representative [22, 23]. Consequently, the representation quality is negatively affected.

In this work, we propose a new framework, namely DDAGDL, to address these problems by proposing an attention-based GDL network. Toward this end, DDAGDL first integrates three kinds of drug-related networks, including drug-disease network, drug-protein network and protein-disease network, to compose a heterogeneous biomedical network, and a HIN is thus generated by further incorporating the biological knowledge of drugs, diseases and proteins. Second, DDAGDL takes advantage of complicated

biological information to learn smoothed feature representations of drugs and diseases with the geometric prior knowledge in the non-Euclidean domain. Moreover, an attention mechanism is adopted by DDADRL to distinguish the significance of features when it learns the final representations of drugs and diseases. Last, a Gradient Boosting Decision Tree (GBDT) classifier, i.e. XGBoost [24], is employed to complete the DR task. Experimental results show that DDAGDL yields a promising performance across all the three benchmark datasets under classical 10-fold cross-validation when compared with several state-of-the-art DR methods. To further demonstrate the advantage of DDAGDL in the DR task, we have also conducted comparative case studies on the top-ranked drug candidates predicted by each comparing method for Alzheimer's Disease and Breast Cancer. Our findings indicate that DDAGDL is able to identify high-quality DDAs that have already been reported by previously published studies, and some of them are not even identified by other methods. In addition to Alzheimer's Disease and Breast Cancer, DDAGDL is also a useful DR tool for newly discovered diseases according to the results of molecular docking experiments for COVID-19. In conclusion, the key reason for the success of DDAGDL is its ability to leverage GDL for better handling the HIN data, in which there is an underlying non-Euclidean structure. Hence, our work opens a new avenue in drug repositioning with new insights gained from GDL.

Methods

Datasets

To evaluate the performance of DDAGDL, three benchmark datasets, i.e. B-dataset, C-dataset and F-dataset, are adopted to construct different HINs. Each dataset contains three kinds of biological networks, i.e. a DDA network, a drug-protein association network and a protein-disease association network. B-dataset and F-dataset are collected from previous studies [25–27] while C-dataset is constructed by following the instruction of Luo et al [28]. Drug-protein associations and protein-disease associations are downloaded from the DrugBank database [29] and the DisGeNET database [30], respectively. The details of these three datasets are presented in Table 1.

Construction of HIN

A HIN is denoted as a three-element tuple, i.e. $\text{HIN}(\mathbf{V}, \mathbf{C}, \mathbf{E})$, where $\mathbf{V} = \{V^{dr}, V^{pr}, V^{di}\}$ is a total of $|\mathbf{V}|$ biomolecules including drugs (V^{dr}), proteins (V^{pr}) and diseases (V^{di}), $\mathbf{C} = [\mathbf{C}^{dr}, \mathbf{C}^{di}, \mathbf{C}^{pr}]^T \in \mathbb{R}^{|\mathbf{V}| \times d}$ is a matrix representing the biological knowledge of all nodes in \mathbf{V} , $\mathbf{E} = \{E^{dd}, E^{dp}, E^{pd}\}$ represents all DDAs (E^{dd}), drug-protein associations (E^{dp}), protein-disease associations (E^{pd}). Moreover, N , K and M are used to denote the respective numbers of drugs, proteins and diseases, and the adjacency matrix of HIN is defined as $\mathbf{A} \in \mathbb{R}^{|\mathbf{V}| \times |\mathbf{V}|}$.

Representation learning from biological knowledge

Regarding the biological knowledge of drugs, we take advantage of the Simplified Molecular Input Line Entry System (SMILES) [31], which is a linear symbol to represent molecular reactions, to construct \mathbf{C}^{dr} . Specifically, we first collect the SMILES data of each drug from the DrugBank database [29], and then process it with the RDkit tool [32] to obtain the feature vector c_i^{dr} of the drug $c_i^{dr} \in V^{dr}$. Since c_i^{dr} is high-dimensional, an auto-encoder model [33] is applied to obtain a more compact form by reducing its dimension to d , whose value is set as 64 in our work. As last, with all c_i^{dr} , we are able to obtain $\mathbf{C}^{dr} = [c_1^{dr}; c_2^{dr}; \dots; c_N^{dr}]^T$.

Table 1. Summary of three benchmark datasets

Dataset	DDAs	Drug-protein associations	Protein-disease associations	Drugs	Diseases	Proteins	Density
B-dataset	18,416	3110	5898	269	598	1021	0.1144
C-dataset	2532	3773	10,734	663	409	993	0.0093
F-dataset	1933	3243	54,265	593	313	2741	0.0104

Similarly, we extract the biological knowledge of diseases in light of medical subject descriptors collected from the Medical Subject Headings (MeSH) thesaurus, and then construct \mathbf{C}^{di} by calculating the semantic similarity between diseases by following the instruction of Guo et al [34]. After that, an auto-encoder model [33] is also used to reduce the dimension of \mathbf{C}^{di} to $M \times d$. Since the biological knowledge of drugs and diseases in the C-dataset are not found in related databases, we explicitly use the processed matrices provided by Van et al. [35] and Luo et al. [28] as \mathbf{C}^{di} and \mathbf{C}^{dr} , respectively. An auto-encoder model is also applied to reduce the size of vectors in \mathbf{C}^{di} and \mathbf{C}^{dr} to d .

Regarding proteins in V^{pr} , we utilize their sequence information to construct \mathbf{C}^{pr} . In particular, each amino acid is first divided into four classes, i.e. (Ala, Val, Leu, Ile, Met, Phe, Trp, Pro), (Gly, Ser, Thr, Cys, Asn, Gln, Tyr), (Arg, Lys, His) and (Asp, Glu), according to the nature of the side chain. A 3-mer algorithm [36] is then applied to obtain c_i^{pr} for each protein. Given all c_i^{pr} , we are able to obtain $\mathbf{C}^{pr} = [c_1^{pr}; c_2^{pr}; \dots; c_k^{pr}]^T$.

So far, we are able to compose \mathbf{C} with \mathbf{C}^{dr} , \mathbf{C}^{di} and \mathbf{C}^{pr} , and a HIN can thus be constructed. One thing about the biological knowledge should be noted. For an drug v , all elements of its corresponding vector in \mathbf{C} are set as 0 if we could not obtain its biological knowledge from relevant databases. Similar operations are also applied to diseases and proteins without biological knowledge.

Attention-based GDL network

Existing GDL-based methods generally learn the feature representations of drugs and diseases by simultaneously considering network structure with non-Euclidean data and biological knowledge available in HIN, and different neural network models are adopted to achieve this purpose, such as graph convolutional network [37] used by DRHGCN [20] and graph attention network [22]. However, a common disadvantage of these well-established GDL networks is the over-smoothing issue, which diminishes the discriminative ability of feature representations of drugs and diseases. Following a general GDL model, the feature representations of drugs and diseases at the l th layer, denoted as $\mathbf{X}^{(l+1)}$, can be obtained with an aggregation operation described as:

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{L}\mathbf{X}^{(l)}\mathbf{W}_1^{(l)}) \quad (1)$$

where $\mathbf{L} = \mathbf{D}^{-1/2} \tilde{\mathbf{A}} \mathbf{D}^{-1/2}$ denotes the normalized Laplacian matrix, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is an adjacency matrix with added self-loops, $\tilde{\mathbf{D}}_{ij} = \sum_j \tilde{\mathbf{A}}_{ij}$ represents a degree matrix, \mathbf{W}_1 is a trainable weight matrix, and $\sigma(\cdot)$ is an activation function. As has been pointed out by Wu et al. [38] and Li et al. [39], the over-smoothing issue is mainly caused by the multiplication of \mathbf{L} and $\mathbf{X}^{(l)}$ in Equation (1). To facilitate the derivation of smoothed $\mathbf{X}^{(l+1)}$, we assume \mathbf{W}_1 is the identity matrix and $\sigma(\cdot)$ is an identity function. Hence, Eq. (1) can be reconstructed as the following form.

$$\mathbf{X}^{(l+1)} = \mathbf{L}\mathbf{X}^{(l)} \quad (2)$$

When the number of layers, or aggregation depth, l is large enough, the Eq. (2) can be rewritten as:

$$\mathbf{X}^{(l)} = \mathbf{L}^{(l)} \mathbf{X}^{(0)} \quad (3)$$

where $\mathbf{X}^{(0)}$ is the initial representation matrix equivalent to \mathbf{C} . Following a node-dependent local smoothing (NDLS) strategy [40], we calculate the node-specific minimal value of l with a distance parameter δ , which is a small constant to control the smoothing effect. The definition of NDLS for node v_i ($v_i \in \mathbf{V}$) is given as:

$$\text{NDLS}(v_i, \delta) = \min \{l : \|\mathbf{L}_{v_i}^{(\infty)} - \mathbf{L}_{v_i}^{(l)}\|_2 < \delta\} \quad (4)$$

where $\|\cdot\|_2$ denotes the Fresenius norm, and $\text{NDLS}(v_i, \delta) > 0$. To aggregate useful features by propagating neighborhood information through l layers, we design an attention mechanism when learning the representation of v_i . Consequently, the update rule of $\mathbf{X}_{v_i}^{(l)}$ is given as:

$$\mathbf{X}_{v_i}^{(l)} = \frac{\exp(\mathbf{h}^T \text{ReLU}(\mathbf{W}_2 \mathbf{X}_{v_i} + b))}{\sum \exp(\mathbf{h}^T \text{ReLU}(\mathbf{W}_2 \mathbf{X}_{v_i} + b))} \quad (5)$$

$$\mathbf{X}_{v_i} = [\mathbf{L}_{v_i}^{(0)} \mathbf{C}_{v_i}; \mathbf{L}_{v_i}^{(1)} \mathbf{X}_{v_i}^{(1)}; \dots; \mathbf{L}_{v_i}^{(l)} \mathbf{X}_{v_i}^{(l)}], l = \text{NDLS}(v_i, \delta) \quad (6)$$

where \mathbf{W}_2 is an $l \times l$ trainable weight matrix, b is bias and \mathbf{h} is a trainable parameter.

Finally, an $(N + M) \times d$ matrix \mathbf{X} can be constructed to denote the feature representations of drugs and diseases.

Algorithm 1. The complete procedure of DDAGDL.

Input: graph $\text{HIN}(\mathbf{V}, \mathbf{C}, \mathbf{E})$.

representation size: d

the number of regression trees: T

Output: the prediction matrix \mathbf{R}

1: Initialization: \mathbf{R}

2: Obtaining biological knowledge matrix of drug \mathbf{C}^{dr}

3: Obtaining biological knowledge matrix of disease \mathbf{C}^{di}

4: Obtaining biological knowledge matrix of protein \mathbf{C}^{pr}

5: Reducing dimensions by the autoencoder

6: $\mathbf{C} = [\mathbf{C}^{dr}; \mathbf{C}^{di}; \mathbf{C}^{pr}]^T \in \mathbb{R}^{|\mathbf{V}| \times d}$

7: **for** each $v_i \in \mathbf{V}$ **do**

8: $\text{NDLS}(v_i, \delta) = \min\{l : \|\mathbf{L}_{v_i}^{(\infty)} - \mathbf{L}_{v_i}^{(l)}\|_2 < \delta\}$

9: $\mathbf{X}_{v_i}^{(l)} = \frac{\exp(\mathbf{h}^T \text{ReLU}(\mathbf{W}_2 \mathbf{X}_{v_i} + b))}{\sum \exp(\mathbf{h}^T \text{ReLU}(\mathbf{W}_2 \mathbf{X}_{v_i} + b))}$

10: **end for**

11: **for** each $e_{ij} = \langle v_i, v_j \rangle \in E^{dd}$ **do**

12: the concatenated feature set $H = \{(H_i, y_i)\} (1 \leq i \leq |H|)$

13: $\mathbf{R} = \text{XGBoost}(H, T)$

14: **end for**

15: **return** \mathbf{R}

DDA prediction

Since the task of DDA prediction is normally regressed as a binary classification problem, DDAGDL adopts a well-established ensemble learning classifier, XGBoost [24], to complete the prediction task with \mathbf{X} . In particular, XGBoost performs its classification task via multiple decision trees each of which contributes to the final prediction result. To train XGBoost, we first compose a set of drug-target pairs denoted as $H = \{(H_i, y_i)\} (1 \leq i \leq |H|)$, where H_i denotes the concatenated feature vector of the i th drug-disease pair, y_i is its label to indicate the existence of an interaction, and $|H|$ is the size of H . Assuming that $v_{dr} \in V^{dr}$ and $v_{di} \in V^{di}$ consist of the i th drug-disease pair in H . H_i is the concatenation of $\mathbf{X}_{v_{dr}}$ and $\mathbf{X}_{v_{di}}$, which are the respective representation vector of v_{dr} and v_{di} obtained with Equation (5), and the value of y_i is 1 if $e_{dd} \in E^{dd}$ and 0 otherwise.

Assuming that $F(x)$ is a linear combination of weak classifiers, which are regression trees in our work, and $L(y_i, F(x))$ is loss function, the purpose of XGBoost is to optimize $F(x)$ such that the minimization of $\sum_{i=1}^{|H|} L(y_i, F(H_i))$ can be achieved. In particular, for the i th drug-disease pair in H , $F(H_i)$ is the classification result of H_i , and $L(y_i, F(H_i))$ is the loss between y_i and $F(H_i)$. The negative loglikelihood is used to calculate $L(y_i, F(H_i))$, and its definition is given as below.

$$L(y_i, F(H_i)) = \log(1 + \exp(-2y_i F(H_i))) \quad (7)$$

To minimize $\sum_{i=1}^{|H|} L(y_i, F(H_i))$, XGBoost first builds an initial decision tree, and then iteratively constructs new regression trees to reduce the residues computed with the loss function $L(y_i, F(H_i))$ till convergence. Given a query drug-disease pair, DDAGDL applies the XGBoost classifier trained with aforementioned steps to predict a probability score indicating the likelihood of an interaction existed between the query drug and disease. The complete procedure of DDAGDL is described in Algorithm 1.

Before training a DDAGDL model, users are required to formulate their own dataset in accordance with the input configurations of DDAGDL. After that, a customized DDAGDL model can be obtained by running the scripts available at <https://github.com/stevejobws/DDAGDL>. For a query drug-disease pair, a prerequisite to calculate their association score with DDAGDL is that the corresponding drug and disease nodes should be existed in the HIN constructed from the training dataset. If that is the case, DDAGDL first obtains the representations of drug and disease nodes, and then takes them as input to the trained XGBoost classifier, with which the association score of the query drug-disease pairs can be calculated.

Results

Overview of DDAGDL

DDAGDL is composed of three steps, and its overall framework is presented in Figure 1. Given a HIN, DDAGDL first employs an autoencoder to obtain the initial representations of drugs and diseases from their biological knowledge. Second, an attention-based GDL network is developed to avoid the over-smoothing issue by adaptively adjusting the range of neighborhood information during aggregation, and its attention mechanism allows it to extract useful features for high-quality representation learning. With smoothed representations of drugs and diseases, DDAGDL infers new DDAs according to the scores predicted by the XGBoost classifier.

Performance comparison with state-of-the-art DR methods

To accurately evaluate the performance of DDAGDL, we adopt a 10-fold cross-validation (CV) scheme by dividing a benchmark dataset into 10-fold, each of which is alternatively taken as a testing set while the rest are used as the training set. The performance of DDAGDL on each fold has been evaluated with Accuracy, MCC and F1-score, and the results are presented in Supplementary Material. Regarding the generation of negative samples, we randomly pair up drugs and diseases whose associations are not found in the benchmark dataset, and the number of negative samples is equal to that of positive ones. As one of the most important GDL operations, neighborhood aggregation makes node representations less distinguishable if more layers are stacked to enlarge receptive fields, leading to the over-smoothing issue. Several recent studies [41, 42] have shown that dense graphs with sufficient connectivity and label information could alleviate this issue by ensuring effective aggregation. It is noted from Table 1 that the density of B-Dataset is considerably larger than that of F-Dataset. In this regard, the performance deterioration caused by the over-smoothing issue is less significant in B-Dataset, and accordingly DDAGDL has only slightly better performance on B-Dataset.

We have compared the overall performance of DDAGDL with five state-of-the-art baseline methods, including SKCNN [15], DeepR2Cov [10], deepDR [14], DTINet [8] and DRHGCN [20]. The details of these comparing methods are presented in the section of Introduction. Regarding their parameter settings used for training, we explicitly adopt the default parameter values recommended in their original work for conducting a fair comparison.

The experimental results of 10-fold CV on three benchmark datasets, including B-dataset, C-dataset and F-dataset, are presented in Figures 2 and 3. The details of these three datasets are presented in the section of Methods. We note that DDAGDL yields the best performance across all the benchmark datasets, as on average it gives 11.39%, 18.22% and 21.54% relative improvement in Accuracy, MCC and F1-score, respectively, over all baseline methods. Another point worth noting is that two GDL-based methods, i.e. DDAGDL and DRHGCN, considerably outperform both SKCNN, DeepR2cov, deepDR and DTINet that only handle the data with an underlying Euclidean structure. This could be a strong indicator that the consideration of GDL enhances the representation learning ability of computational DR methods over HIN. A further improvement achieved by DDAGDL is mainly due to its capability of avoiding the over-smoothing issue by adaptively adjusting the aggregation depth for each node in HIN. Hence, we have reason to believe that DDAGDL is preferred as a promising DR tool when applied to discover new indicators for existing drugs.

In addition to its superiority in identifying novel DDAs, DDAGDL is also more robust against noisy data without sacrificing the accuracy. Taking DRHGCN, which is the second-best method, as an example, DDAGDL performs better by 3.57%, 4.60% and 6.78% than it by averaging all evaluation scores over B-dataset, C-dataset and F-dataset respectively. Moreover, similar behaviors are observed on the evaluation metrics of Recall and Precision for all baseline methods. In particular, we note that the Precision scores obtained by deepDR, DTINet and DRHGCN are much larger than their Recall scores, as they are prone to predict known DDAs as negative. But for DDAGDL, the incorporation of an attention mechanism alleviates the impact of noisy data by concentrating on the most representative neighborhood information during aggregation. Consequently, its Recall and Precision scores are

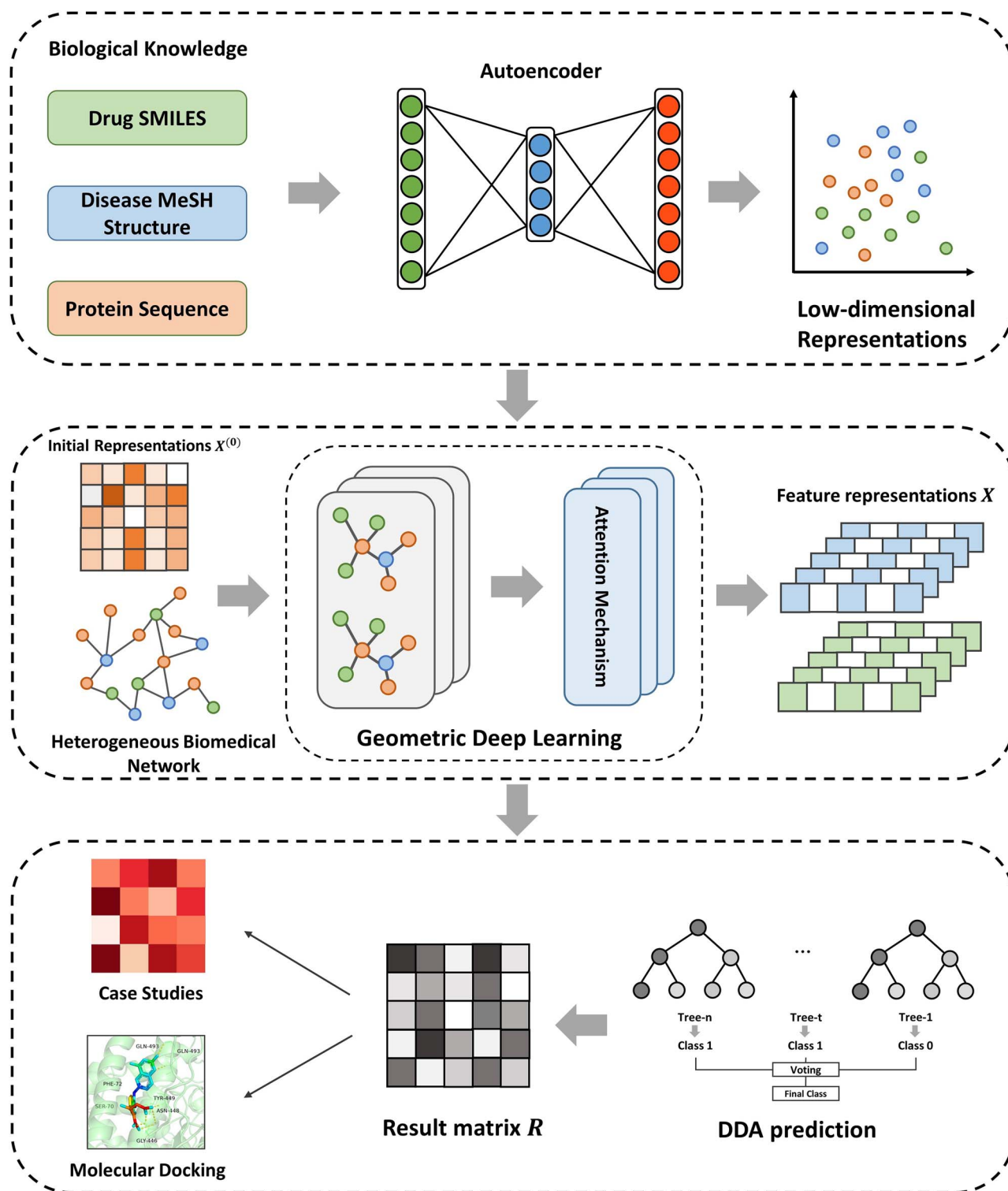


Figure 1. The overall workflow of DDAGDL.

much closer to each other. Hence, DDAGDL is granted a stronger discriminative ability in distinguishing between known DDAs and those randomly paired up when compared with baseline methods.

Regarding the unsatisfactory performance of SKCNN, DeepR2Cov, deepDR and DTINet for the DR task, their operations conducted in the Euclidean domain have a 2-fold effect. First, they normally formalize the drug-related network data as matrices such that the statistical properties of the data can be exploited

on the Euclidean space. However, different kinds of drug-related networks may present contradicting properties on their own Euclidean spaces, thus confusing the computational methods in predicting novel DDAs. Second, they fail to characterize the structure of HIN constructed in our work, and thereby miss certain geometric prior knowledge for better learning the representations of drugs and diseases in the context of HIN. A possible reason for the unsatisfactory performance of DeepR2cov

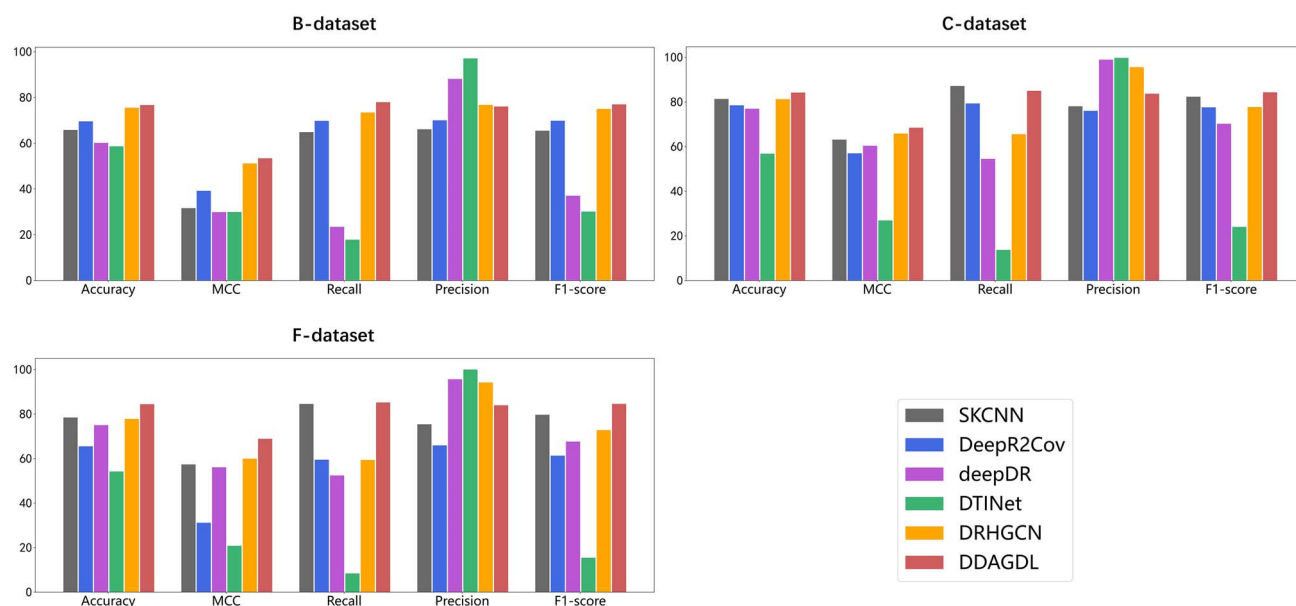


Figure 2. The experimental results of all comparing models on three benchmark datasets, and they are presented in subfigures, respectively.

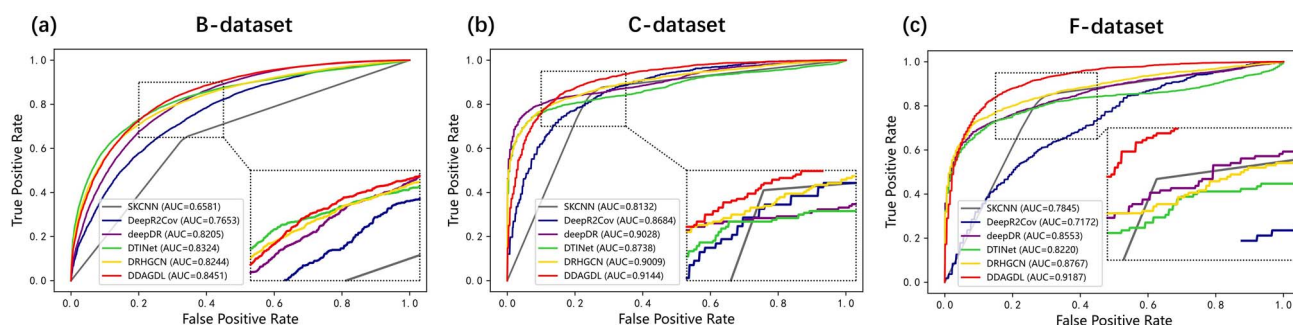


Figure 3. The ROC curves w.r.t. the overall performance of all comparing models on three benchmark datasets, and they are presented in subfigures (A–C), respectively.

could be ascribed to the fact that manually selected meta-paths may not be able to appropriately capture the characteristics of DDAs for improved prediction accuracy. Although DRHGCN achieves the second-best performance on all three benchmark datasets, it suffers from the over-smoothing disadvantage, as the resulting features tend to follow a uniform distribution, which in return constraints its predictive ability for the DR task.

In summary, the geometric prior knowledge of HIN is of benefit for DDAGDL to correctly capture its full complexity and structural richness in a non-Euclidean domain, and the proposed attention-based GDL network allows DDAGDL to seamlessly incorporate such knowledge for learning high-quality smoothed representations of drugs and diseases. Consequently, DDAGDL yields a promising performance in identifying novel DDAs.

Ablation study of DDAGDL

To better investigate the influence of GDL on the performance of DDAGDL, we have also constructed an ablation study by developing three variants of DDAGDL, i.e. DDAGDL-A, DDAGDL-N and DDAGDL-G. The main difference lying between them is how to obtain the representations of drugs and diseases. In particular, DDAGDL-A learns the representations of drugs and diseases from their biological knowledge by following an autoencoder scheme, but DDAGDL-N adopts the proposed attention-based GDL network for representation learning. When compared with DDAGDL,

DDAGDL-N simply uses a one-hot encoding method to initialize the representations of nodes in HIN. To evaluate and testify the capability of avoiding the over-smoothing issue for DDAGDL, we develop a variant of DDAGDL, namely DDAGDL-G, to learn the feature representations of drugs and diseases with a traditional graph convolutional network [37], which has the over-smoothing issue during the aggregation of neighborhood information. Although the results of our ablation study indicate that DDAGDL-G achieves the second-best performance on all three benchmark datasets, yet its performance is constrained by the over-smoothing issue, as the resulting features tend to follow a uniform distribution with less distinguishable difference. When compared with DDAGDL-G, DDAGDL further improves the accuracy of DDA prediction by avoiding the over-smoothing issue with a node-dependent local smoothing (NDLS) strategy. The XGBoost classifiers used by these three variants share the same parameter setting as DDAGDL. Their experimental results of 10-fold CV on three benchmark datasets are presented in Figures 4 and 5, where several things can be noted.

First, DDAGDL-A achieves the worst performance among DDAGDL and its variants. In this regard, only relying on the biological knowledge of drugs and diseases may not be sufficiently enough to achieve desired DR performance. Particularly, for newly discovered diseases, lack of sufficient biological knowledge increases the difficulty of accurately identifying their unknown

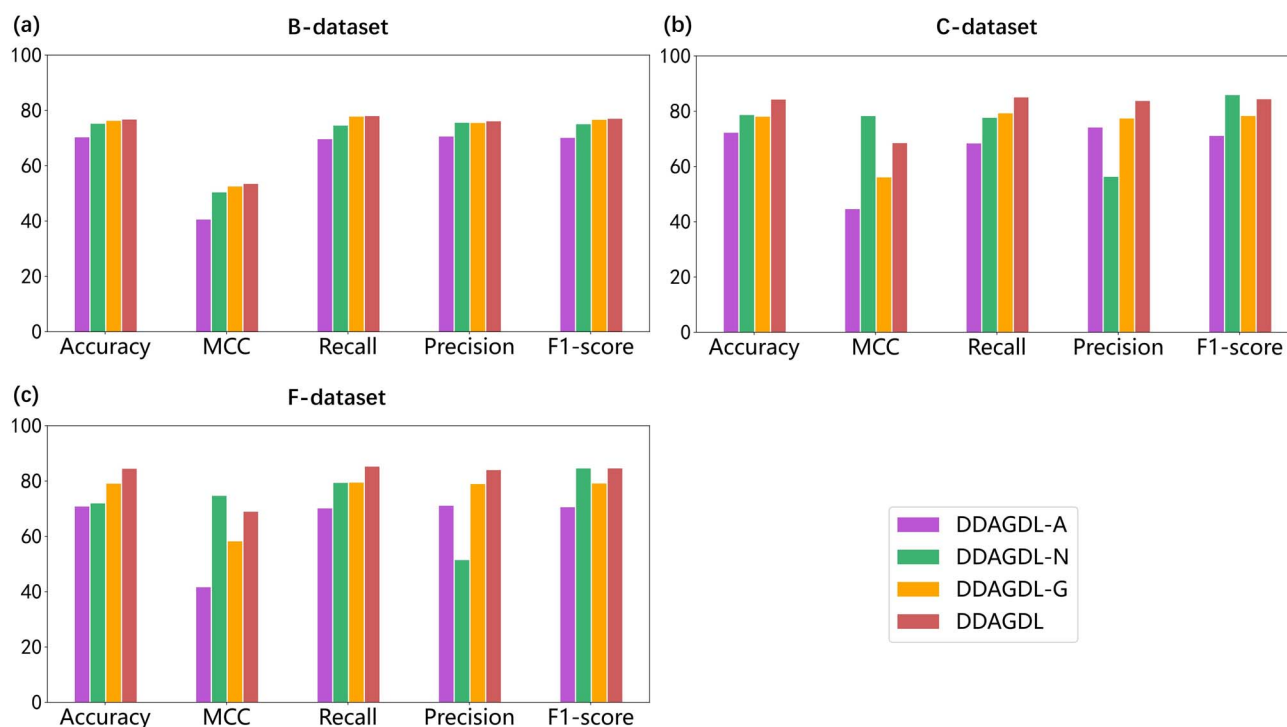


Figure 4. The performance of DDAGDL-A, DDAGDL-N and DDAGDL-G on the three benchmark datasets in the ablation study, and they are presented in subfigures (A–C), respectively.

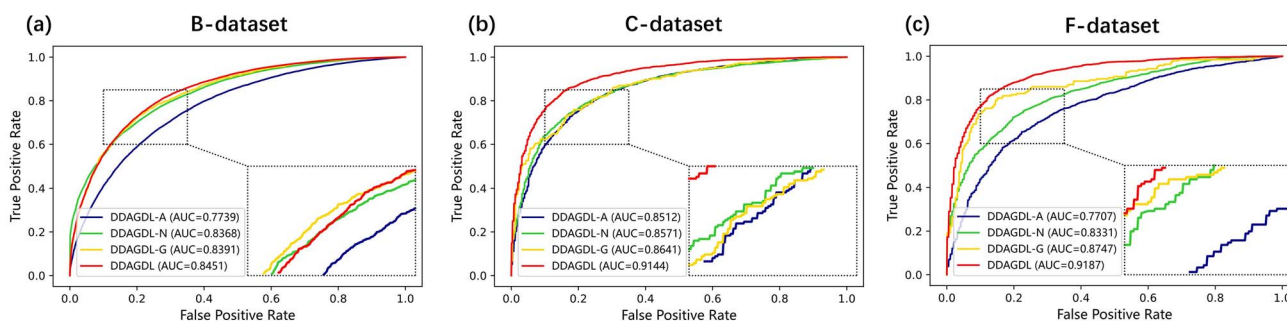


Figure 5. The ROC and PR curves are obtained by three variants of DDAGDL over three benchmark datasets in the ablation study, and they are presented in subfigures (A–C), respectively.

DDAs with DDAGDL-A. Second, DDAGDL-N shows a bigger margin in performance against DDAGDL-A. On average, DDAGDL-N performs better by 6.67%, 5.33%, 9.67% and 5.33% than DDAGDL-A in terms of AUC, ACC, MCC and F1-score, respectively, across all the benchmark datasets. Thus, the geometric prior knowledge of network structure allows DDAGDL-N to better capture the characteristics of drugs and diseases in a non-Euclidean domain. Last, a further improvement is observed from DDAGDL by combining the advantages of DDAGDL-A and DDAGDL-N. Comparing the performance of DDAGDL-A with that of DDAGDL-N, we reason that it is the integration of GDL that contributes the most to the performance of DDAGDL.

Classifier selection of DDAGDL

In particular, we apply many well-established classifiers, i.e. Logistic regression (LR), SVM, Random Forest Classifier (RF), GDBT and XGBoost, to perform the task of drug repositioning with the representations of drugs and diseases learned by DDAGDL. Our purpose is to select the classifier with which the best performance of DDAGDL can be achieved. To achieve the purpose, all classi-

fiers are trained under the same 10-fold CV over all the three benchmark datasets and their experimental results are presented in Figures 6 and 7. We note that DDAGDL yields the best performance when using XGBoost as its classifier. The main reason for that phenomenon is due to the robust ensemble learning ability of XGBoost. Hence, we decide to incorporate XGBoost into DDAGDL for DDA prediction.

In addition, there are two points worth further commentary. On the one hand, among all classifiers, LR and SVM obtain the worst performance in terms of Accuracy, MCC, F1-score and AUC, as they have low fitting ability to heterogeneous information networks and are not applicable for the task of drug repositioning. On the other hand, the performances of RF and GDBT are only worse than XGBoost, but better than LR and SVM. This could be a strong indicator that nonlinear classifiers are better to characterize the associations between drugs and diseases.

Generalization ability of DDAGDL

In particular, we have conducted cross-data validation by taking C-dataset and F-dataset as the training and testing datasets

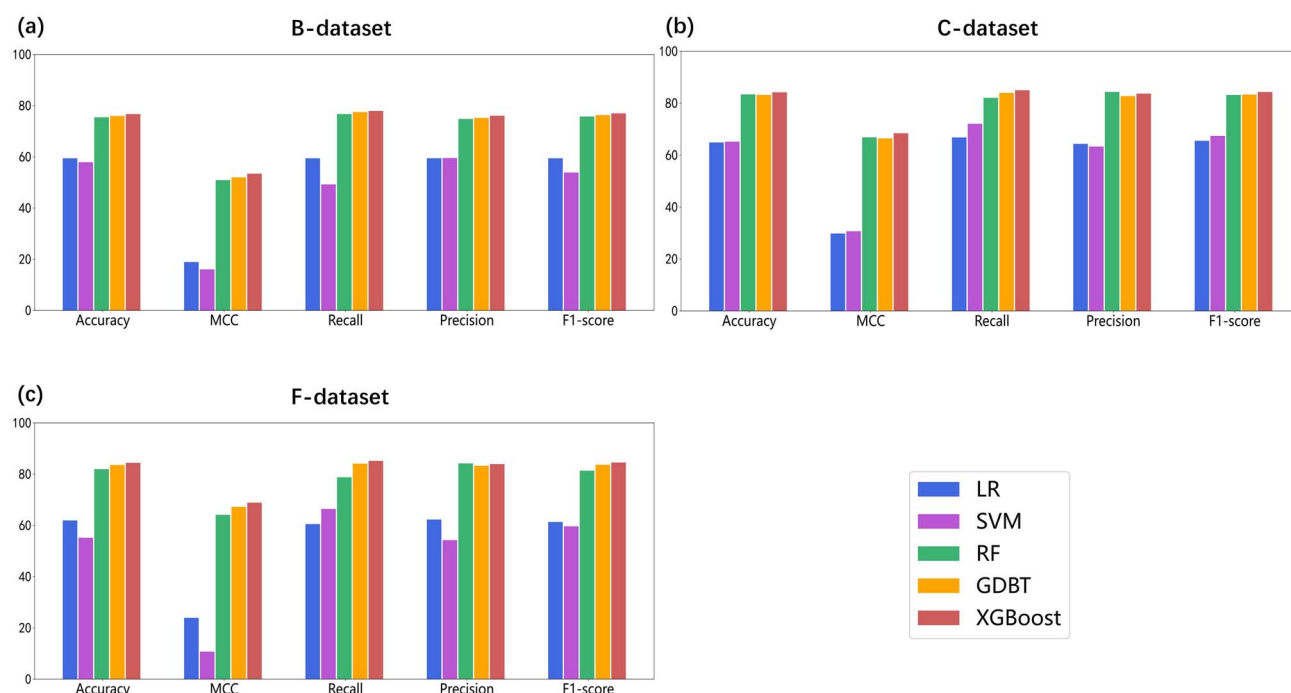


Figure 6. The performance of different classifiers on the three benchmark datasets, and they are presented in subfigures (A–C), respectively.

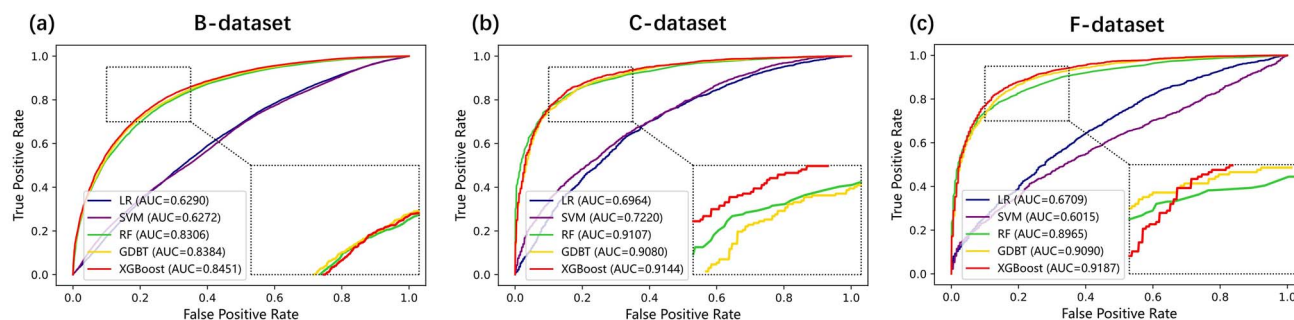


Figure 7. The ROC and PR curves are obtained by different classifiers over three benchmark datasets in the ablation study, and they are presented in subfigures (A–C), respectively.

respectively. To this end, all known DDAs in the C-dataset are regarded as the positive samples in the training dataset, while those in the F-dataset are used to compose the testing dataset. The strategy of selecting negative samples for both training and testing datasets is the same as we adopt for 10-fold CV. The experimental results indicate that the Accuracy, MCC, F1-score and AUC scores obtained by DDAGDL are 90.14%, 81.15%, 90.81% and 96.58%, respectively. In this regard, DDAGDL still yields a promising performance under cross-dataset validation. One should note that a prerequisite for DDAGDL to predict well out-of-sample is that both training and external testing datasets should share many common drugs and diseases. Otherwise, DDAGDL may not be able to accurately predict novel DDAs for drugs and diseases that are not found in the training dataset.

Case studies on Alzheimer's disease and breast cancer

To demonstrate the capability of DDAGDL in practically discovering potential DDAs, we have conducted additional experiments on the B-dataset. In particular, all DDAs in the B-dataset are used to construct the training dataset, and our purpose is to predict new candidate drugs for two diseases, i.e. Alzheimer's Disease

and Breast Cancer, as the case studies. Moreover, the reasons why we select B-dataset are 2-fold. On the one hand, although DDAGDL performs worse on B-dataset than on F-dataset, it is more meaningful for us to discover novel DDAs from B-dataset so as to strongly indicate the superior ability of DDAGDL in drug repositioning. On the other hand, there are more diseases included in the B-dataset, and accordingly the generalization ability of DDAGDL can thus be verified by applying it to many different diseases. Since Alzheimer's Disease and Breast Cancer are two popular diseases that have been well studied by researchers, approved drugs for these two diseases are much more complete than those of the other diseases. It is for this reason that we take Alzheimer's Disease and Breast Cancer as our case studies in this work. Besides, we also report the drug candidates of other diseases discovered by DDAGDL from B-dataset in Supplementary materials.

In Table 2, we list the top 10 candidates discovered by DDAGDL as the potential drugs of Alzheimer's Disease, and among them six candidates are verified with evidence collected from relevant literature. Taking chlorpromazine as an example, it is already known that a structural analog of chlorpromazine can treat early cognitive deficit by reducing the levels of amyloid beta

Table 2. The top 10 candidate drugs predicted by DDAGDL for AD

Disease	Drugs	Scores	Evidence (PMID)
Alzheimer's disease	Phenytoin	0.89	16781825
	Valproic acid	0.88	19748552
	Risperidone	0.87	33176899
	Chlorpromazine	0.86	N/A
	Carbamazepine	0.86	28193995
	Fluoxetine	0.84	30592045
	Cocaine	0.82	N/A
	Methotrexate	0.81	32423175
	Diazepam	0.81	N/A
	Diphenhydramine	0.80	N/A

Table 3. The top 10 candidate drugs predicted by DDAGDL for breast cancer

Disease	Drugs	Scores	Evidence (PMID)
Breast cancer	Methylprednisolone	0.94	12884026
	Valproic acid	0.92	30075223
	Nifedipine	0.88	25436889
	Phenytoin	0.86	22678159
	Simvastatin	0.86	33705623
	Amiodarone	0.86	26515726
	Sirolimus	0.84	32335491
	Ethinyl estradiol	0.83	N/A
	Betamethasone	0.83	N/A
	Acetaminophen	0.83	N/A

(A β) [43]. Since pathological proteins of AD mainly contain A β [44], we have reason to believe that chlorpromazine has a pharmacological effect on the treatment of Alzheimer's Disease. We also investigate the prediction results obtained by deepDR and DRHGCN, and find that none of them is able to discover the association between chlorpromazine and Alzheimer's Disease. Hence, this phenomenon could be a strong indicator for the superior ability of DDAGDL in discovering new potential drugs for diseases.

Regarding the case study of Breast Cancer, the top 10 candidates of potential drugs predicted by DDAGDL are shown in Table 3, and seven out of them have been verified to be effective for the treatment of Breast Cancer according to our literature review. Among all unverified drugs, ethinyl estradiol obtains the largest prediction score, and an in-depth analysis is given after a systematic literature review. As has been pointed out by Iwase et al [45], ethinyl estradiol is of benefit for metastatic breast cancer after prior aromatase inhibitor treatment. In this regard, our findings indicate a possible treatment for breast cancer by ethinyl estradiol from the perspective of artificial intelligence.

Besides, we have performed a detailed analysis to explain why DDAGDL successfully discover verified drug candidates whose associations with their corresponding diseases are unknown in the B-dataset. In particular, the representations of verified drug candidates are compared with those of drugs whose associations with Alzheimer's Disease and Breast Cancer are known in the B-dataset, and the Pearson coefficients between these two kinds of drugs are calculated to indicate the similarity between them. The results are presented in Figures 8 and 9 for Alzheimer's Disease and Breast Cancer, respectively. We note that each verified drug candidate is highly similar to at least one of the known drugs according to the distribution of blocks with dark color. This again demonstrates the rationality of DDAGDL to assign higher scores to these drug candidates.

Regarding the performance of deepDR and DRHGCN in our case studies, their experiment results are presented in Supplementary Material. When compared with DDAGDL, both deepDR and DRHGCN yield poor performance when discovering new drugs for Alzheimer's Disease and Breast Cancer. With deepDR, only 2 out of top 10 drug candidates predicted for Alzheimer's Disease have been verified by relevant literature, while that number for Breast Cancer is only 1. DRHGCN performs slightly better than deepDR, as three drug candidates and two candidate drugs are verified for Alzheimer's Disease and Breast Cancer respectively among top 10 results. Moreover, the prediction scores of top 10 drug candidates yielded by DRHGCN are much lower than those of DDAGDL. In other words, DDAGDL is more confident about its prediction results. The main reason is that the attention-based GDL network used by DDAGDL allows it to focus on significant features for representation learning over HIN. Hence, DDAGDL could be a useful DR tool due to its promising performance.

Molecular docking experiments for COVID-19

The purpose of molecular docking experiments is to evaluate the performance of DDAGDL for newly discovered diseases, such as COVID-19, thus further demonstrating the generalization ability of DDAGDL. In particular, docking-based drug repositioning is a kind of structure-based methods that aim to simulate the binding process of drugs to their target proteins by predicting the structures of receptor-ligand complexes, thereby identifying new indications for approved drugs [46, 47]. However, docking-based drug repositioning suffers the disadvantages of high false-positive rate and being time-consuming. To address these issues, DDAGDL simultaneously considers the underlying non-Euclidean structure of HIN and the biological knowledge of drugs and diseases to improve the quality of feature representations of drugs and diseases from different perspectives. By incorporating these feature representations with XGBoost, DDAGDL is able to predict large-scale DDAs in a reasonable time. Toward this end, we first collect a total of 55 DDAs related to COVID-19 from HDVD [48] by following the instruction of Su et al [49], and add them into B-dataset. Following the same training procedure as we use for case studies, we select the top five drug candidates predicted by DDAGDL for COVID-19, and conduct molecular docking experiments to evaluate their binding energies with SARS-CoV-2 spike protein or human angiotensin-converting enzyme 2 (ACE2) [50], which are important functional receptors for COVID-19. The chemical structures of candidate drugs are downloaded from DrugBank [29], and the coordinate information of binding sites is obtained from RCSB [51]. The molecular docking experiments are performed by using the AutoDockTools and AutoDock software [52], where SARS-CoV-2 spike protein and ACE2 are taken as receptors and each candidate drug is considered as a ligand of interest. In particular, when conducting the molecular docking experiments with AutoDock software, we explicitly set the area of SARS-Cov-2 spike receptor-binding domain bounding with ACE2 with the coordinate, i.e. $x = -36.884$, $y = 29.245$, and $z = -0.005$ as reported by [53]. The binding energies of top five drug candidates are shown in Table 4, and their molecular docking results are presented in Figure 10.

We note that the binding energies of top five drug candidates are positioned at a low level as indicated by Table 4, demonstrating that the molecules of these candidates have strong binding affinities with the receptors of COVID-19. To further indicate their eligibility in treating COVID-19, we also perform additional molecular docking experiments on two approved drugs, i.e. Remdesivir

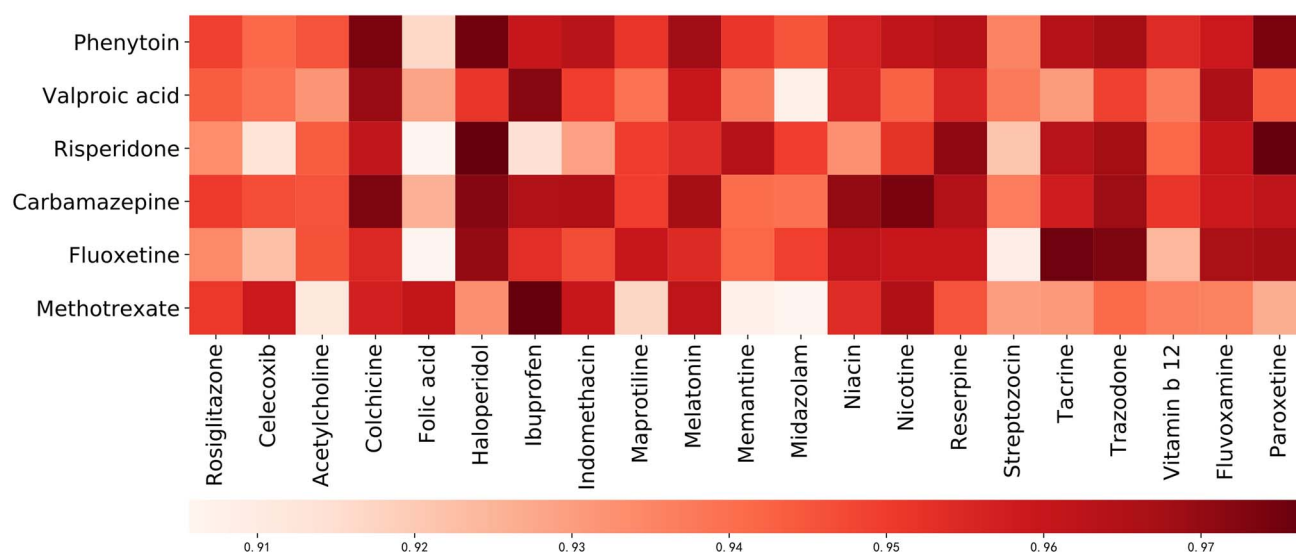


Figure 8. The similarity of feature representations between predicted drugs and approved drugs for AD. The vertical axis denotes the predicted drugs, whereas the horizontal axis denotes the approved drugs.

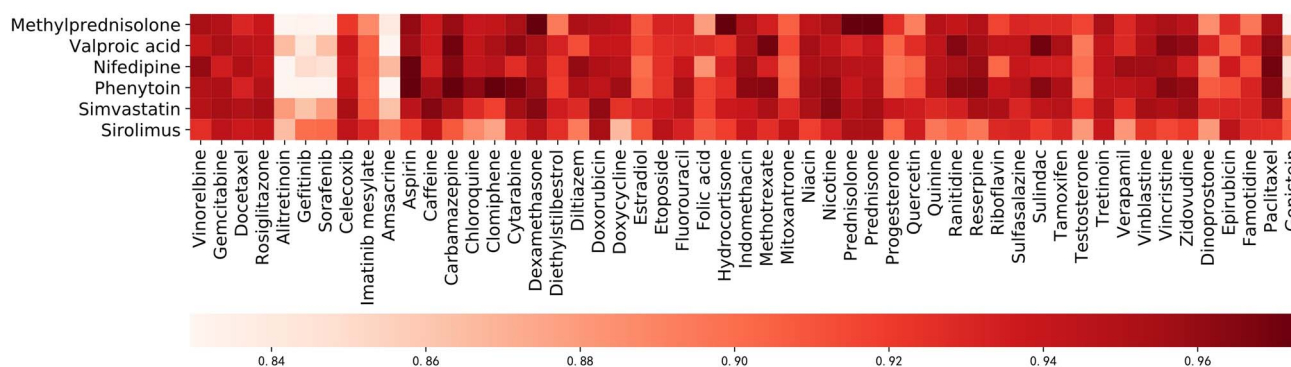


Figure 9. The similarity of feature representations between predicted drugs and approved drugs for Breast Cancer. The vertical axis denotes the predicted drugs, whereas the horizontal axis denotes the approved drugs.

Table 4. Binding energies between top five drug candidates and SARS-CoV-2 spike protein/ACE2

Rank	Drug name	Drug bank ID	Score	Binding energy (kcal/mol)
1	Methotrexate	DB00563	0.97	-6.05
2	Clozapine	DB00363	0.96	-6.80
3	Olanzapine	DB00334	0.96	-6.85
4	Morphine	DB00295	0.96	-7.83
5	Clonidine	DB00575	0.95	-6.79

and Ribavirin, and find that the binding energies of Remdesivir and Ribavirin are -7.25 kcal/mol and -6.87 kcal/mol, respectively. It is observed that among top five drug candidates, the binding energies of Morphine are smaller than those of Remdesivir and Ribavirin. Hence, Morphine is likely to have therapeutic effect against SARS-Cov-2, and thereby it could be considered an alternative treatment for COVID-19. Moreover, we note from Table 4 that Methotrexate obtains the largest prediction score, and it means that DDAGDL has the most confidence in predicting the association between Methotrexate and COVID-19. According to a

literature review, we find that Methotrexate can affect the SARS-CoV-2 virus by disrupting the specific protein interactions of the targeted hub protein DDX39B [54]. One should note that due to the high false-positive rate, the results of molecular docking only indicate a therapeutic possibility for the antiviral drugs newly discovered by DDAGDL, and in-depth follow-up laboratory-based experiments are required to verify the practical therapeutic effect of these drugs for the treatment of related diseases. Overall, we believe that these five drug candidates are increasingly likely to have therapeutic effects against SARS-CoV-2 for the treatment of COVID-19.

Discussion and conclusion

Regarding the DR task, although a variety of ML-based and DL-based computational methods have been proposed, few of them can consider the non-Euclidean nature of biomedical data that are modeled with graphs, and thereby limiting their accuracy in identifying novel DDAs. To overcome this problem, we leverage the learning ability of GDL to improve the quality of feature representations of drugs and diseases by incorporating the geometric prior knowledge of HIN, and propose an efficient DR framework, namely DDAGDL, in this work to accomplish the DR task over

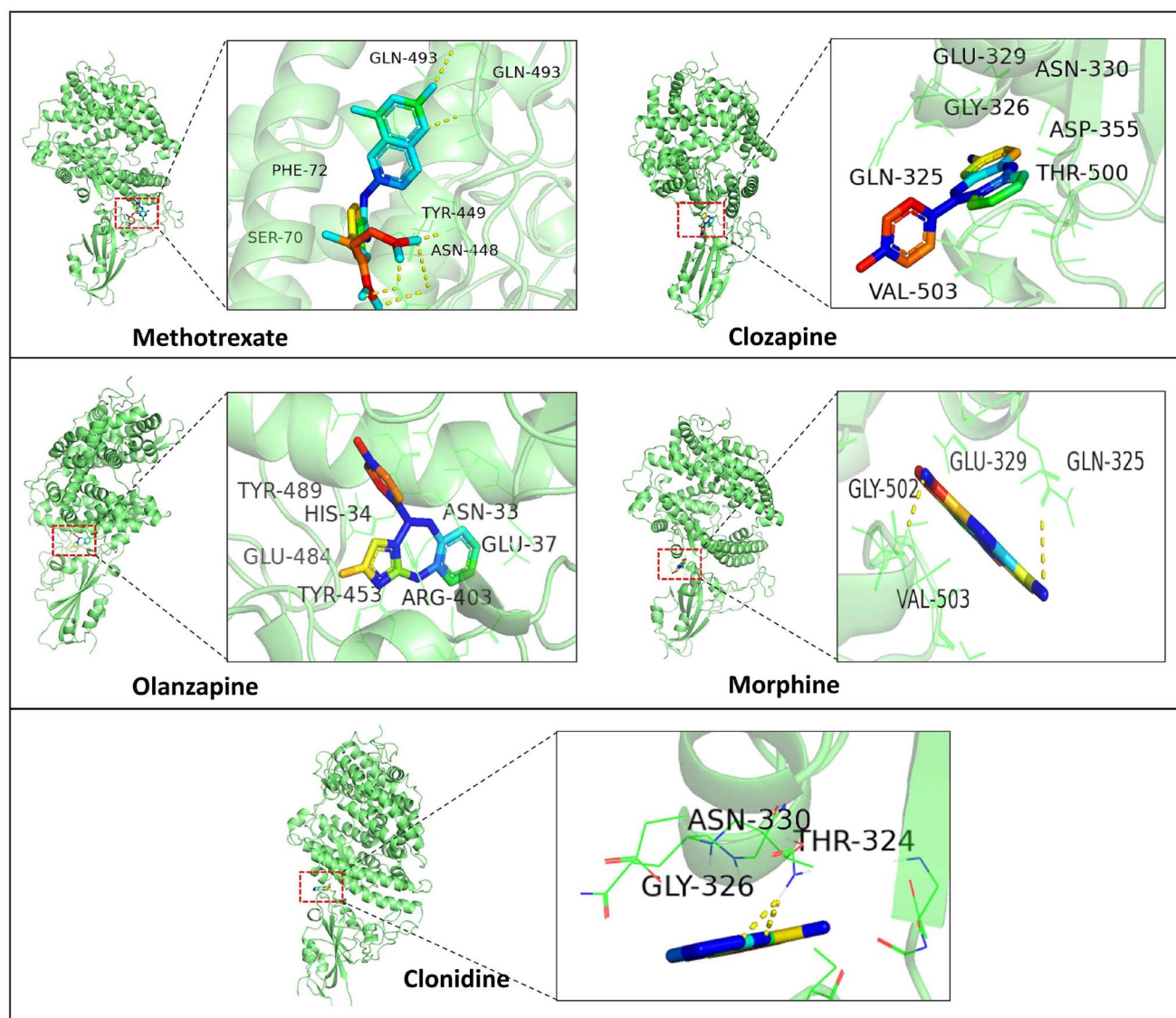


Figure 10. Molecular docking results for top five drug candidates bound with SARS-CoV-2 spike protein/ACE2.

HIN. Experimental results demonstrate that DDAGDL yields a superior performance across all the three benchmark datasets under 10-fold CV when compared with several state-of-the-art baseline methods in terms of Accuracy, MCC, F1-score and AUC. This could be a strong indicator that DDAGDL can effectively learn the smoothed feature representations of drugs and diseases by projecting complicated biological information, characterized by its non-Euclidean nature, onto a latent space with GDL. Furthermore, we have also conducted the case studies to show the usefulness of DDAGDL in predicting novel DDAs by validating the top-ranked drug candidates for Alzheimer's disease and Breast Cancer. Our findings indicate that most of the drug candidates are of high quality, as they have already been reported by previously published studies, and some of them are not even found in the prediction results of the other compared methods. Moreover, the results of molecular docking experiments indicate that DDAGDL is also efficacious for newly discovered diseases. Obviously, leveraging GDL provides us an alternative view to address the DR task by properly handling the non-Euclidean nature of HIN, which has been ignored by most existing DR methods. In conclusion, DDAGDL is a promising DR tool for identifying novel DDAs, and the consideration of GDL gains new insight into the representation

learning of drugs and diseases over HIN by fully exploiting its geometric prior knowledge.

There are several reasons contributing to the superior performance of DDAGDL in the DR task. First, we introduce a HIN model composed of not only the biological knowledge of drugs and diseases, but also three kinds of association networks, i.e. a drug-disease network, a drug-protein network and a protein-disease network. It provides us an opportunity to address the DR task from an integrated perspective. Second, the non-Euclidean nature of HIN raises new challenges, as the basic operations of most existing DR computational methods are particularly designed in the Euclidean case. In this regard, an improved GDL network is incorporated into DDAGDL such that the smoothed representations of drugs and diseases are obtained by adaptively adjusting the aggregation depth. Last, DDAGDL adopts an attention mechanism to aggregate the most significance neighborhood information for representation learning. By doing so, its robustness against noisy data in HIN can be enhanced as indicated by the experimental results.

The limitations of DDAGDL are discussed from three aspects. First, according to the ablation study, we note that the consideration of our attention-based GDL network plays a critical role in

improving the accuracy of DR. However, for new diseases without any known DDAs, DDAGDL has limited ability in learning their representations through network structure. Second, the issue of classifier selection matters the performance of DDAGDL, and currently we could only adopt the trial-and-test method to determine the classifier from a set of well-established classifiers. Last, the efficiency of DDAGDL is constrained due to the extra cost of computing the aggregation depth for each node in HIN.

In addition to proposing specific solutions to address the above limitations, we also would like to unfold our future work from another two aspects. First, we would like to integrate more different association networks, such as to increase the richness of HIN, and expect that DDAGDL is able to learn more expressive representations of drugs and diseases. Second, we are interested in evaluating the generalization ability of DDAGDL by applying it to address other kinds of association prediction problems, such as drug–drug association prediction [55] and protein–protein interaction prediction [56]. Understanding under what circumstances and to what extent DDAGDL generalize across different tasks should be studied as well.

Key Points

- We develop a novel attention-based GDL framework, i.e. DDAGDL, to learn smoothed feature representations of drugs and diseases with geometric prior knowledge in the non-Euclidean domain for improved performance of DDA prediction.
- We leverage the learning ability of GDL to improve the quality of feature representations of drugs and diseases by adaptively adjusting the aggregation depth and then an attention mechanism is constructed to distinguish the significance of features when DDAGDL learns the final representations of drugs and diseases.
- Experimental results on all benchmark datasets demonstrate that DDAGDL performs better than several state-of-the-art baseline methods under 10-fold cross-validation. Furthermore, we have conducted the case studies and molecular docking experiments to indicate that DDAGDL is a promising DR tool that gains new insights into exploiting the geometric prior knowledge for improved efficacy.

Data availability

The dataset and source code can be freely downloaded from <https://github.com/stevejobws/DDAGDL>.

Authors' contributions

B.-W.Z., L.H. and X.-R.S. contributed to the conception, and design of the study and performed the statistical analysis. P.-W.H. and Y.-P.M. organized the database. X.Z., B.-W.Z. and L.H. wrote the first draft of the manuscript. All authors have read and agreed to the published version of the manuscript.

Supplementary data

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

Funding

This work has been supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under grant (2021D01D05), in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, in part by the Tianshan Youth Project–Outstanding Youth Science and Technology Talents of Xinjiang under grant (2020Q005).

References

1. Whitebread S, Hamon J, Bojanic D, et al. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today* 2005;**10**:1421–33.
2. Rudrapal M, Khairnar SJ, Jadhav AG. Drug repurposing (DR): an emerging approach in drug discovery, drug repurposing hypothesis. *Mol Asp Ther Appl* 2020. <http://dx.doi.org/10.5772/intechopen.83082>.
3. Hay M, Thomas DW, Craighead JL, et al. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014;**32**: 40–51.
4. Ballard C, Aarsland D, Cummings J, et al. Drug repositioning and repurposing for Alzheimer disease. *Nat Rev Neurol* 2020;**16**: 661–73.
5. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* 2019;**18**: 41–58.
6. Bagherian M, Sabeti E, Wang K, et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Brief Bioinform* 2021;**22**:247–69.
7. Luo H, Li M, Wang S, et al. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**:1904–12.
8. 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.
9. Ding Y, Tang J, Guo F. Identification of drug–target interactions via fuzzy bipartite local model. *Neural Comput Applic* 2020;**32**: 10303–19.
10. Wang X, Xin B, Tan W, et al. DeepR2cov: deep representation learning on heterogeneous drug networks to discover anti-inflammatory agents for COVID-19. *Brief Bioinform* 2021;**22**:bbab226.
11. Yue X, Wang Z, Huang J, et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 2020;**36**:1241–51.
12. Su X, Hu L, You Z, et al. Attention-based knowledge graph representation learning for predicting drug–drug interactions. *Brief Bioinform* 2022;**23**:bbac140.
13. Yu J-L, Dai Q-Q, Li G-B. Deep learning in target prediction and drug repositioning: recent advances and challenges. *Drug Discov Today* 2021;**27**:1796–1814.
14. 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.
15. Jiang H-J, You Z-H, Huang Y-A. Predicting drug–disease associations via sigmoid kernel-based convolutional neural networks. *J Transl Med* 2019;**17**:382.
16. Xuan P, Ye Y, Zhang T, et al. Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations. *Cell* 2019;**8**:705.

17. Huang K, Xiao C, Glass LM, et al. SkipGNN: predicting molecular interactions with skip-graph networks. *Sci Rep* 2020;**10**:1–16.
18. Bronstein MM, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 2017;**34**:18–42.
19. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. *Nat Mach Intell* 2021;**3**:1023–32.
20. Cai L, Lu C, Xu J, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Brief Bioinform* 2021;**22**:bbab319.
21. Mavi V, Jangra A, Jatowt A. A survey on multi-hop question answering and generation. arXiv preprint arXiv:2204.09140 2022. <https://doi.org/10.48550/arXiv.2204.09140>.
22. Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv preprint arXiv:1710.10903 2017. <https://doi.org/10.48550/arXiv.1710.10903>.
23. Hu L, Pan X, Tan Z, et al. A fast fuzzy clustering algorithm for complex networks via a generalized momentum method. *IEEE Trans Fuzzy Syst* 2021;**1**. <https://doi.org/10.1109/TFUZZ.2021.3117442>.
24. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–94.
25. Zhang W, Yue X, Lin W, et al. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics* 2018;**19**:1–12.
26. Zhao B-W, Hu L, You Z-H, et al. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. *Brief Bioinform* 2022;**23**:bbab515.
27. 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.
28. 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.
29. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2017;**46**:D1074–82.
30. 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**:D833–D839.
31. Weininger DSMILES. A chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6.
32. Landrum G. Rdkit documentation. Release 2013;**1**:1–79.
33. Liou C-Y, Cheng W-C, Liou J-W, et al. Autoencoder for words. *Neurocomputing* 2014;**139**:84–96.
34. 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.
35. Van Driel MA, Bruggeman J, Vriend G, et al. A text-mining analysis of the human phenome. *Eur J Hum Genet* 2006;**14**:535–42.
36. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8.
37. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 2016. <https://doi.org/10.48550/arXiv.1609.02907>.
38. Wu F, Souza A, Zhang T et al. Simplifying graph convolutional networks. In: *International Conference on Machine Learning*. 2019, p. 6861–71. PMLR, Copenhagen, Denmark. <https://proceedings.mlr.press/v97/wu19e.html>.
39. Li Q, Han Z, Wu X-M. Deeper insights into graph convolutional networks for semi-supervised learning. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018, **32**.
40. Zhang W, Yang M, Sheng Z, et al. Node dependent local smoothing for scalable graph learning. *Advances in Neural Information Processing Systems* 2021;**34**:7460–71.
41. Liu M, Gao H, Ji S. Towards deeper graph neural networks. In: *Proceedings of the 26th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2020, p. 338–48.
42. Zhang W, Jiang Y, Li Y et al. ROD: reception-aware online distillation for sparse graphs. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery, 2021, p. 2232–42.
43. Njomen E, Tepe JJ. Proteasome activation as a new therapeutic approach to target proteotoxic disorders. *J Med Chem* 2019;**62**:6469–81.
44. Pickett EK, Herrmann AG, McQueen J, et al. Amyloid beta and tau cooperate to cause reversible behavioral and transcriptional deficits in a model of Alzheimer's disease. *Cell Rep* 2019;**29**:3592, e3595–3604.e5.
45. Iwase H, Yamamoto Y, Yamamoto-Ibusuki M, et al. Ethinylestradiol is beneficial for postmenopausal patients with heavily pre-treated metastatic breast cancer after prior aromatase inhibitor treatment: a prospective study. *Br J Cancer* 2013;**109**:1537–42.
46. Gao KY, Fokoue A, Luo H, et al. Interpretable drug target prediction using deep neural representation. In: *IJCAI* 2018; 3371–7.
47. Khanjiwala Z, Khale A, Prabhu A. Docking structurally similar analogues: dealing with the false-positive. *J Mol Graph Model* 2019;**93**:107451.
48. Meng Y, Jin M, Tang X, et al. Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Appl Soft Comput* 2021;**103**:107135.
49. Su X, Hu L, You Z, et al. A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. *Brief Bioinform* 2022;**23**:bbab526.
50. Tada T, Fan C, Chen JS, et al. An ACE2 microbody containing a single immunoglobulin fc domain is a potent inhibitor of SARS-CoV-2. *Cell Rep* 2020;**33**:108528.
51. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
52. Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 2009;**30**:2785–91.
53. Peng L, Shen L, Xu J, et al. Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures. *Sci Rep* 2021;**11**:1–11.
54. Liu X, Huuskonen S, Laitinen T, et al. SARS-CoV-2-host proteome interactions for antiviral drug discovery. *Mol Syst Biol* 2021;**17**:e10396.
55. Su X-R, Huang D-S, Wang L, et al. Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction. *IEEE Transactions on Knowledge and Data Engineering* 2022;**1**. <http://doi.org/10.1109/TKDE.2022.3154792>.
56. Hu L, Wang X, Huang Y-A, et al. A survey on computational models for predicting protein-protein interactions. *Brief Bioinform* 2021;**22**:bbab036.