Neurocomputing 511 (2022) 247–258

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An explainable framework for drug repositioning from disease information network

Chengxin He [a,b], Lei Duan [a,b,*], Huiru Zheng [c,*], Linlin Song [d,e], Menglin Huang [a]

[a] School of Computer Science, Sichuan University, Chengdu 610065, China
[b] Med-X Center for Informatics, Sichuan University, Chengdu 610065, China
[c] School of Computing, Ulster University, Northern Ireland BT37 0QB, United Kingdom
[d] Department of Ultrasound, West China Hospital of Sichuan University, Chengdu 610041, China
[e] Frontiers Science Center for Disease-related Molecular Network, West China Hospital of Sichuan University, Chengdu 610041, China

ARTICLE INFO

Article history:
Received 9 April 2022
Revised 7 August 2022
Accepted 4 September 2022
Available online 13 September 2022
Communicated by Zidong Wang

Keywords:
Drug repositioning
Graph neural network
Interpretable prediction
Disease information network

ABSTRACT

Exploring efficient and high-accuracy computational drug repositioning methods has become a popular and attractive topic in drug development. This technology can systematically identify potential drug-disease interactions, which could greatly alleviate the pressures from the high cost and long period taken by traditional drug research and discovery. However, plenty of current computational drug repositioning approaches lack interpretability in predicting drug-disease associations, which will not be friendly to their subsequent in-depth research.

To this end, we hereby propose a novel computational framework, called *EDEN*, for exploring explainable drug repositioning from the disease information network (DIN). *EDEN* is a graph neural network framework that learns the local semantics and global structure of the DIN, and models the drug-disease associations into the DIN by maximizing the mutual information of both and an end-to-end manner. In this way, the learned biomedical entity and link embeddings are enabled to retain the ability to drug repositioning with the semantical structure of external knowledge, thereby making interpretation possible. Meanwhile, we also propose a matching score based on the final embeddings to generate the predictive drug repositioning explanation. Empirical results on the real-world dataset show that *EDEN* outperforms other state-of-the-art baselines on most of the metrics. Further studies reveal the effectiveness of the explainability of our approach.

1. Introduction

The development of new drugs is a lengthy process with a slow pace, high attrition rates, and substantial costs, entering them into the market successfully needs more effort. Indeed, the majority of drug candidates are eliminated during their phase I clinical trials [1]. Therefore, exploring efficient ways to improve the success rate of drug research and discovery is pressing and significant. In recent years, an attractive proposition in the field of drug development, *drug repositioning*, attracting increasing interest from both the pharmaceutical industry and research community.

Drug repositioning, or drug repurposing, aims to identify new therapeutic opportunities for existing drugs, and to reduce the time, cost and risk of conventional drug development [2]. The most straightforward way to find new indications for existing drugs is through biological experiments to perform target- or cell-based screens for thousands of medications. While this activity-based strategy can directly detect potential indications for drugs, it is still a time-consuming and labor-intensive process, and testing drugs in assays based on some available comprehensive clinical compound databases is also extremely challenging [3]. Fortunately, the rapid advances in multi-omics have provided a great opportunity to exploit drug repositioning by computational approaches with a much faster repositioning process at a lower cost.

Many computational methods for drug repositioning have been developed [4–28], involving techniques ranging from traditional machine learning, matrix factorization, to network analysis and deep learning. For instance, Wang et al. [4] proposed a clustering-based framework, named GS4CDRSC, to identify gene signatures for cancer drug repositioning by grouping samples into several clusters based on their gene expression profiles. DrPOCS [5]

[*] Corresponding authors at: School of Computer Science, Sichuan University, Chengdu 610065, China (L. Duan). School of Computing, Ulster University, Northern Ireland BT37 0QB, United Kingdom (H. Zheng).

E-mail addresses: leiduan@scu.edu.cn (L. Duan), h.zheng@ulster.ac.uk (H. Zheng).

https://doi.org/10.1016/j.neucom.2022.09.063
0925-2312/© 2022 The Author(s). Published by Elsevier B.V.

is a matrix completion-based method that integrated drug structure and disease phenotype information through the idea of projection onto convex sets (POCS) to predict potential associations between drugs and diseases. Moreover, Yang et al. [6] developed a network analysis-based approach, called HED, to infer drug repositioning by constructing a drug-disease association heterogeneous network and then applying network embedding technology. Similarly, this method can be extended to network pharmacology. PINA [7] applied this extension to predict potential indications of Traditional Chinese Medicines with Liuwei-Dihuang-Wan as a case study. Yan et al. [8] devised a deep learning-based method, named MLMC, to determine drug indications through multi-view learning with matrix completion.

However, these methods have mainly focused on how to better fuse high-throughput data related to drugs at various levels, such as genomic data, protein structures and phenotypes, to improve the predictive performance of drug repositioning, rarely exploring which aspects of such multi-source data enable the model to predict the association between drug and disease, in other words, these methods lack interpretability for this prediction. Citing *valproic acid* as an example, it can be used to treat *bipolar disorder* and *seizures* because of its ability to bind to the *mitochondrial enzymes succinate-semialdehyde dehydrogenase (ALDH5A1)* and *4-aminobutyrate aminotransferase (ABAT)* [29]. This is in terms of its impact on proteins. Furthermore, *valproic acid* could induce *differentiation, growth arrest, and apoptosis in cancer cells*, leading to its repositioning to the treatment of neoplastic conditions such as *familial adenomatous polyposis* [30], which is reflected in its action on pathways.

Through the above observations, we can see that when current computational models of drug repositioning give a prediction result, it seldom further explains from which aspects such the associations arise (e.g., from the aforementioned proteins or pathways). Instead, these methods are more concerned with designing reasonable ideas to integrate these aspects to improve the prediction accuracy of drug repositioning. It is obvious that the interpretable prediction is very important when assessing the performances of a computational model and for better understanding the underlying mechanisms of drug repositioning, as well as providing researchers with relevant insights and decision support in subsequent in-depth studies based on the prediction results. We are inspired by the recent developments of explainable recommendation [31–33], which have the potential of achieving the goal but have not been explored much for drug repositioning. This paper will focus on the interpretable prediction of the computational models in the drug repositioning.

The following questions should be considered in the design of explainable drug repositioning computational model:

- How could the interpretable prediction be incorporated into the drug repositioning model?
- How to reflect the interpretability of drug repositioning prediction results?
- How to ensure the validity of drug repositioning prediction results while introducing interpretability?

To tackle the above challenges, we propose a novel computational framework called **EDEN** (short for explainable drug repositioning) for exploring explainable drug repositioning. The characteristics of EDEN include: (1) it projects multiple types of biomedical entities and relations in the constructed disease information network into a unified low-dimensional space; (2) it aggregates neighbor messages into the final embeddings based on different semantic connections to preserve the internal structure of the knowledge; (3) it equips drugs and diseases with structured knowledge in an end-to-end way by maximizing the mutual information between global structure and local semantic to generate explanations for drug repositioning and improve the predictive performance; and (4) it designs a matching score to construct explanations regarding the prediction results by searching over the paths in the embedding space.

In summary, the main contributions of this work include:

- Highlighting the importance of capturing different aspects of action in drug repositioning to provide better interpretable prediction results for subsequent studies;
- Proposing a new framework, EDEN, which learns over the heterogeneous knowledge from the constructed disease information network for explainable drug repositioning under the graph neural network paradigm;
- Conducting empirical studies on real-world datasets to demonstrate the effectiveness of EDEN.

The rest of this paper is organized as follows. The related work is reviewed in Section 2. Some preliminaries and the overall design of EDEN are depicted in Section 3. Then, the experimental settings and results are discussed in detail in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

We review previous studies relevant to this work in two areas: the computational methods for drug repositioning and the prediction tasks for interpretability.

2.1. Computational Drug Repositioning

Depending on the implementation technique, most of the current computational approaches to drug repositioning can be roughly divided into the following three main categories: matrix factorization-based, network propagation-based and machine learning-based.

2.1.1. Matrix factorization-based

Dai et al. [9] proposed a matrix factorization model based on known drug-disease associations to predict new drug indications. Meanwhile, the authors integrated genomic space into the model, which is to provide molecular biological information for exploring drug-disease associations. Similarly, Zhang et al. [10] presented a method called SNNMF, which used a non-negative matrix factorization to fuse the different effects of drug-disease associations to improve the prediction performance. Considering that many of the non-occurring edges in the drug-disease associations are actually unknown or missing cases, Ezzat et al. [11] designed two matrix factorization methods that utilized the graph regularization to enhance the prediction of new drugs and new target cases. To combine multiple side information with the idea of matrix factorization, recent work adopted the singular value decomposition [12] or the tensor decomposition [13] for drug repositioning. Moreover, Bagherian et al. [14] proposed a matrix factorization-based method termed CMMC to capture potential associations between drugs and diseases by the idea of coupled matrix completion. Meanwhile, this work also extended CMMC to “coupled tensor-matrix completion” in order to merge multiple types of information provided in different databases.

2.1.2. Network propagation-based

In recent years, network-based approaches have been widely employed, due to their powerful advantage in being able to organize the relationship between biomedical entities well. For example, Luo et al. developed a novel drug repositioning model, RWHNDR, which extended the random walk method to the constructed drug-target-disease network [15], and a bi-random walk algorithm to predict potential drug-disease associations on drug and disease similarity networks they built [16]. Shahreza et al. [17] proposed a method called Heter-LP to identify interactions between drugs and targets by propagating the label information across the constructed heterogeneous network. Wang et al. [18] presented a novel heterogeneous network model which integrated drug-disease and drug-target interactions prediction into a unified computational framework. Ji et al. [19] developed a network propagation-based method, named DTINet, to predict novel disease targets. This approach identifies novel disease targets for drugs based on the obtained network topological similarities among known diseases and drugs associations through the idea of induction matrix completion. Besides, NEDD [20] is a meta-path-based computational method to predict new associations between drugs and diseases using meta paths of different lengths to explicitly capture the indirect relationships, or high order proximity, within drugs and diseases. Zhao et al. [21] designed a novel heterogeneous information network-based model, named HINGRL, to learn the features of nodes in the network constructed with biological knowledge from the topological and biological perspectives by applying different representation strategies. HINGRL then employed a Random Forest classifier to predict unknown drug-disease associations based on the features obtained in the previous step.

2.1.3. Machine learning-based

The continuous development of machine learning techniques has provided various effective and efficient solutions for drug repositioning. Wang et al. [22] proposed a framework, called PreDR, for drug repositioning, which calculated the similarity between drug-disease pairs through the constructed kernel function, and then trained a support vector machine with the defined kernel to find the novel effects between drugs and diseases. Olayan et al. [23] developed a model, named DDR, to improve the drug-target interaction prediction accuracy. The DDR applied the random forest method to extract features from the constructed heterogeneous graph and predict the relationship between drug and target. In addition to classical machine learning approaches, many deep learning-based frameworks (e.g. convolutional neural network, graph neural network) have recently demonstrated their excellence in exploring drug repositioning, such as GNDD [24] and DR-HGCN [25]. In addition, Jarada et al. [26] developed a deep learning-based framework, called SNF-NN, by using similarity selection, similarity network fusion, and a highly tuned novel neural network to predict new drug-disease interactions. On this basis, Jarada et al. [27] also proposed SNF-CVAE, which integrated similarity network fusion and collective variational auto-encoder to conduct a non-linear analysis and improved the drug-disease interaction prediction accuracy. Meng et al. [28] proposed a neighborhood interaction-based neural collaborative filtering method, called DRWBNCF, to infer new potential drugs for diseases. DRWBNCF integrated known drug-disease associations into a unified representation through a weighted bilinear graph convolution operation. And then DRWBNCF utilized the multi-layer perceptron to predict drug-disease associations.

2.2. Explainable Prediction Model

At present, the majority of work on introducing interpretability into model is in the recommender systems. These studies leverage knowledge graph (KG), which is rich in semantic information, to make explainable decisions for recommender systems. For example, Ai et al. [31] proposed to learn knowledge-based embeddings for the explainable recommendation. Wang et al. [32] designed an RNN(Recurrent Neural Network)-based model to reason over KG for recommendation. Also, Xian et al. [33] presented a method called PGPR, which generates causal inference through reinforcement learning to enhance the interpretability of recommendation.

In addition, in the field of biomedical information, there are some computational models that have made efforts on the interpretability. Fout et al. [34] used graph neural networks to learn effective latent representations of the 3D structure of proteins and visualized predicted protein interface. Gao et al. [35] presented an interpretable framework to predict drug-target interactions, which introduced a two-way attention mechanism to track the likelihood of drug atoms interacting with each amino acid component.

However, these approaches only utilize the values generated by the attention mechanism to evaluate the impact of an input feature on the prediction results or to visualize them, and do not yet have a high quality of interpretability, that is, it should contain rich semantic messages.

3. Methodology

In this paper, we focus on *explainable drug repositioning*, where the objective of EDEN is to predict novel potential associations between drugs and diseases, and explain why the connection between them could produce.

Firstly, we introduce some basic concepts and formalize the problem definition. Then we present EDEN based on the graph neural network over the constructed disease information network to solve the problem.

3.1. Preliminaries

In the scenario of exploring drug repositioning, we typically have known interactions between drugs and diseases. Here, we use a drug-disease bipartite network to denote the associations.

Definition 1 (Drug-Disease Bipartite Network): A drug-disease bipartite network $\mathscr{G}_b$ is defined as $\{(u, y_{ud}, d) | u \in \mathscr{U}, d \in \mathscr{D}\}$, where $\mathscr{U}$ and $\mathscr{D}$ respectively represent the drug and disease sets, and a link $y_{ud} = 1$ indicates that there is a known or approved association between drug u and disease d; otherwise $y_{ud} = 0$.

To model explainable drug repositioning, the key point of EDEN is to equip it with rich semantic information. We hereby build a disease information network from multiple resources to organize these auxiliary data.

Definition 2 (Disease Information Network): A disease information network (DIN), which is a typical heterogeneous graph, is defined as $\mathscr{G}_d = \{\mathscr{V}, \mathscr{E}\}$, where $\mathscr{V}$ and $\mathscr{E}$ represent the biomedical entity set and the link set, respectively. Each entity $v \in \mathscr{V}$ and each link $\varepsilon \in \mathscr{E}$ are associated with their mapping function $\phi(v) : \mathscr{V} \rightarrow \mathscr{A}$ and $\varphi(\varepsilon) : \mathscr{E} \rightarrow \mathscr{R}$, in which $\mathscr{A}$ and $\mathscr{R}$ denote the biomedical entity types and link types, respectively.

It is worth noting that the alignment of a drug-disease bipartite network $\mathscr{G}_b$ and a disease information network $\mathscr{G}_d$ can construct a unified graph $\mathscr{G}$. We also use a triple to encode the relationship between biomedical entities in the unified graph. Formally, it is presented as $\mathscr{G} = \{(h, r, t) | h, t \in \mathscr{V}\prime, r \in \mathscr{E}\prime\}$, where $\mathscr{V}\prime = \mathscr{V} \cup \mathscr{U}$ ($\mathscr{D} \subset \mathscr{V}$) and $\mathscr{E}\prime = \mathscr{E} \cup \{treat\}$, where *treat* represents the semantic relation between drug u and disease d, e.g., $y_{ud} = 1$.

Now, we formulate the task to be addressed in this paper:

- **Input:** the drug-disease bipartite network $\mathscr{G}_b$ and a disease information network $\mathscr{G}_d$.

- **Output:** a drug repositioning prediction function that predicts the probability $\widetilde{y}_{ud}$ that drug u would treat disease d, and a corresponding set of explanations E_{ud}.

3.2. Design of EDEN

Fig. 1 shows the EDEN architecture, which consists of four main components: 1) *local embedding layer*, which learns local representations of entities and links in the graph $\mathscr{G}$ by aggregating the information from neighbors, and adjust the weight of each neighbor according to the semantics of link during a propagation; 2) *global embedding layer*, which parameterizes each entity as an embedding by preserving the structure of $\mathscr{G}$; 3) *optimization layer*, which maximizes the mutual information between local embeddings and global embeddings, and then refines the entity and link embeddings from training the objective function; 4) *explanation extraction*, which generates a set of sound logical inference paths from the query drug to predicted disease based on the final embeddings.

3.2.1. Local embedding layer

Here we employ TransE [36], an effective way to model relation as entity translation, on $\mathscr{G}$ to initialize entity and link embeddings. Specifically, for a given triple (h, r, t), if it exists in $\mathscr{G}$, the translation principle $e_h + e_r \approx e_t$ is used to learn each entity and link embeddings, which is formalized as:

$$trans(h,r,t) = \exp((e_h + e_r) \cdot e_t) \tag{1}$$

where e_h, e_r and $e_t \in \mathbb{R}^k$ are the embedding for h, r and t, respectively; k represents the dimension of embedding and a higher score of $trans(h, r, t)$ suggests that the triple is more likely to be true, and vice versa.

To train the TransE, we adopt a pairwise ranking loss to discriminate valid triples and broken ones:

$$\mathscr{L}_1 = \sum_{(h,r,t,t\prime)\in T} -\ln \mu(trans(h,r,t) - trans(h,r,t\prime)) \tag{2}$$

where $T = \{(h,r,t,t\prime)|(h,r,t) \in \mathscr{G}, (h,r,t\prime) \notin \mathscr{G}\}$, and $(h,r,t\prime)$ is a broken triple constructed by replacing one entity in a valid triple randomly; $\mu(\cdot)$ is the Sigmoid function.

Next, similar to our previous work [37], we build upon the graph convolution operations to recursively aggregate neighbor's messages; moreover, by exploiting the idea of attention mechanism [38], we introduce attentive weights to links to reveal the importance of different semantic relations when gathering neighbor embeddings.

Considering an entity h, we use N_h^r to denote the set of neighbors of h, which are connected by link r, and use π_r to represent the attentive weight. Thus, after the l-step convolutional operations, we recursively formulate the local embedding of an entity as:

$$loc(e_h) = e_h^{(l)} = \mathrm{ReLU}\left(W\left(e_h^{(l-1)} + e_{N_h^r}^{(l-1)}\right)\right) \tag{3}$$

$$e_{N_h^r}^{(l-1)} = \sum_{r\in\mathscr{E}\prime}\sum_{t\in N_h^r}\pi_r e_t^{(l-1)} \tag{4}$$

$$\pi_r = \frac{\tanh((e_h + e_r) \cdot e_t)}{\sum_{(h,r\prime,t\prime)\in\mathscr{G}} \tanh((e_h + e_{r\prime}) \cdot e_{t\prime})} \tag{5}$$

where $W \in \mathbb{R}^{k\prime \times k}$ is the trainable parameter, $k\prime$ is the transformation size, *tanh* and *ReLU* are the activation function. In particular, $e_h^{(0)} = e_h$.

3.2.2. Global embedding layer

In order to obtain the entity embedding while preserving the global structure of the graph $\mathscr{G}$, here we use the DeepWalk [39] method to generate the global embedding of the entities. DeepWalk is a deep unsupervised learning model, which can vectorize all entities in the unified graph $\mathscr{G}$ by utilizing truncated random walks and SkipGram [40]. This way integrates information about the global structure of the graph into the entity embedding based on maximizing the probability of observing entity in view of all entities previously visited up to the current point in the random walk. Compared with other representation learning methods that capture the global structure of the graph (e.g., struc2vec [41], metapath2vec [42]), the advantage of this way is that it does not require introduction of pre-defined or domain knowledge to guide.

For starters, truncated random walks that the number of random walks ρ of length ζ to start at each entity are performed. Then, the SkipGram model is implemented to learn the entity embedding for each random walk. SkipGram maximally compute the co-occurrence likelihood among the entities that come into view within a path of window w, and its objective function is as follows:

$$min_{\Phi} - \log P_r(\{e_{h-w}, \cdots, e_{h+w}\} \quad e_h|\Phi(e_h)) \tag{6}$$

$$glo(e_h) = \Phi(e_h) \tag{7}$$

Here min_{Φ} means the global embedding-based representations Φ related to each entity e_h derived from this objective function make the value solved by the subsequent formula as minimum as possible. And SkipGram further approximates the above conditional probability using the assumption as follows:

$$P_r(\{e_{h-w}, \cdots, e_{h+w}\} \quad e_h|\Phi(e_h)) = \prod_{m=h-w, m\neq h}^{h+w} P_r(e_h|\Phi(e_h)) \tag{8}$$

More details about DeepWalk are sketched in [39].

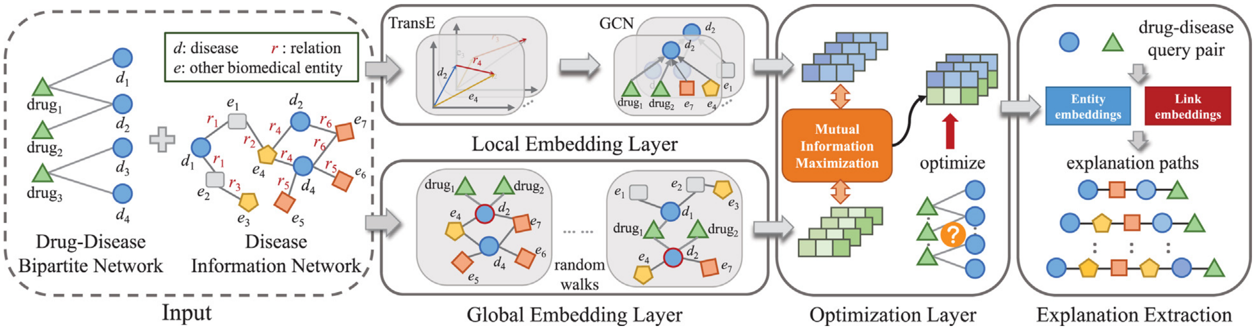


Fig. 1. Illustration of the proposed EDEN, which includes four components: local embedding layer, global embedding layer, optimization layer and explanation extraction.

3.2.3. Optimization layer

Hereafter, we maximize the mutual information between the local embeddings and global embeddings of entities, following the idea of Graph Infomax [43]. Formally, the objective function is:

$$\mathcal{L}_2 = \sum_{h\in\mathcal{V}'} -\log\tau \frac{s(loc(e_h), glo(e_h))}{\sum_{h'\in\mathcal{V}'} s(loc(e_h), glo(e_{h'}))} \tag{9}$$

where $s(\cdot)$ is the function measuring the affinity of any local embedding and global embedding of an entity, which is set as cosine similarity function here; and τ is the hyper-parameter to prevent overfitting.

After performing the above series of operations, we can obtain the representation for drug entity $e_u \in \{e_h|\forall h \in \mathcal{V}'\}$; analogous to disease entity e_d are obtained. Finally, we conduct the inner product on the drug and disease embeddings to predict their association score:

$$\widetilde{y}_{ud} = e_u^{\top} e_d \tag{10}$$

Algorithm 1: *EDEN*$(\mathcal{G}_b, \mathcal{G}_d, N, z, u, d)$

```
Input: Drug-disease bipartite network $\mathcal{G}_b$, disease
  information network $\mathcal{G}_d$, epoch $N$, query drug and disease
  pair $u, d \in \mathcal{G}_b$ and the maximum explanation path depth $z$
Output: The top-$K$ explanation paths $E_{ud}$
Initializes $e_h, e_t$ and $e_r$
for $n \leftarrow 1$ to $N$ do
    $B_1 \leftarrow$ Sampling a set of training sample batches from
      $\mathcal{G}_b \cup \mathcal{G}_d$
    for each batch of $B_1$ do
        calculates $loc(e_h)$ and $glo(e_h)$ via Eq. (3)–(8)
        updates $e_h, e_t$ and $e_r$ with Eq. (2) and Eq. (9)
    end for
    $B_2 \leftarrow$ Sampling a set of training sample batches from $\mathcal{G}_b$
    for each batch of $B_2$ do
        $e_u, e_d \leftarrow e_h \cup e_t$ via mapping
        updates $e_h, e_t$ and $e_r$ with Eq. (10) and Eq. (11)
    end for
end for
$S \leftarrow \{(e_h, e_r, e_t)\}$ // The set of embeddings $S$ for all entities
  and relations.
$V_u, R_u, P_u \leftarrow$BFS$(S, u, z)$
$V_d, R_d, P_d \leftarrow$BFS$(S, d, z)$
$P \leftarrow \varnothing$ // The set of paths $P$.
for $m \in V_u \cap V_d$ do
    $P[m] \leftarrow P_u(m) + P_d(m)$ // $P[m]$ represents all paths from
      $u$ to $d$ connected by entity $m$ and the corresponding scores
      for these paths.
end for
Pick up the set of paths $E_{ud}$ with the top-$K$ largest $P[m]$.
return $E_{ud}$

function BFS $(S, i, z)$
    $V_i \leftarrow$ all entities in the set $S$ within $z$ hops from $i$
    $R_i \leftarrow$ the paths from $i$ to the space of each entity in
      $\mathcal{G}_b \cup \mathcal{G}_d$
    $P_i \leftarrow$ the score of each combined path from $i$ computed
      by Eq. (13)
    return $V_i, R_i, P_i$
end function
```

To optimize EDEN, we also employ the pairwise ranking loss, based on the assumption that the observed interactions from the drug-disease bipartite network $\mathcal{G}_b$, which indicates more likely to provide guidance for drug repositioning, should be assigned higher values in the prediction than unobserved ones:

$$\mathcal{L}_3 = \sum_{(u,d_i,d_j)\in\mathcal{O}} -\ln\mu\left(\widetilde{y}_{ud_i} - \widetilde{y}_{ud_j}\right) \tag{11}$$

where $\mathcal{O} = \{(u, d_i, d_j)|(u, d_i) \in \mathcal{G}_b, (u, d_j) \notin \mathcal{G}_b\}$ denotes the training set; $\mu(\cdot)$ is the Sigmoid function.

Finally, we have the total objective function is as follows:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \lambda||\Theta||_2^2 \tag{12}$$

where Θ is the set of model parameters; and λ controls the L2 regularization strength to prevent overfitting. Recall that in our designs of $\mathcal{L}_1$ (Eq. 2), $\mathcal{L}_2$ (Eq. 9), and $\mathcal{L}_3$ (Eq. 11), since the Sigmoid function takes values in the range $(0, 1)$, which will make the difference in the loss during training not obvious and is not conducive to convergence, we here adopt the "ln" function to map the value ranges of $\mathcal{L}_1$ and $\mathcal{L}_3$ to the same value range as $\mathcal{L}_2$ $(0, +\infty)$ in order to facilitate the combination of $\mathcal{L}_1, \mathcal{L}_2$, and $\mathcal{L}_3$ to obtain the total objective function and enable it to converge during training. Then we use the Adam optimization method [44] to optimize the model and update the model parameters.

3.2.4. Explanation extraction

Now, we describe how to generate explanations of the drug repositioning prediction with EDEN. Similar to the work [31], given the predicted drug and disease pair, we utilize the entity and link embeddings obtained after executing the optimization layer to match the optimal paths connecting them, so as to create explanations. Since the obtained entity and link embeddings are from the disease information network we constructed, in order for the resulting explanation path to have reasonable semantics, the explanation path should be from the disease information network [45].

Technically, we conduct breadth first search (BFS) with maximum depth z from the drug u and the disease d to find paths than can potentially link them. Also, for each path $p =< u, \cdots + r_\alpha + h_\gamma + r_\beta + \cdots, d > (r_\alpha, r_\beta \in \mathcal{E}', h_\gamma \in \mathcal{V}')$, we devise a measurement to calculate its score for matching u and d:

$$score(p) = \frac{\sum_{(h,r',t')\in p} \tanh((e_h + e_{r'}) \cdot e_{t'})}{|p|} \tag{13}$$

where $|p|$ represents the number of links in the path p.

At last, EDEN ranks these paths by their matching scores and returns K paths with the highest score as explanations E_{ud} for predicting the result between drug u and disease d. The pseudo-code of EDEN is summarized in Algorithm 1.

4. Experiments and Result Discussion

In this section, we conduct experiments on real-world datasets to evaluate EDEN. We first illustrate the details of experimental settings including dataset, baseline methods, metrics and setup. Then we discuss the experimental results by answering the following research questions:

- **RQ1:** Compared with the state-of-the-art computational drug repositioning methods, how does EDEN perform?
- **RQ2:** How do different components and parameters influence EDEN?

- **RQ3:** Can EDEN provide reasonable explanations about the prediction results of drug repositioning?

4.1. Experimental Settings

4.1.1. Datasets

To construct the drug-disease bipartite network, we merged two data sources: CTD [46] and repoDB [47]. Specifically, the links selected data that is annotated as "therapeutic" in the CTD and status is "approved" in the repoDB. We also collected the following data sources to build a disease information network with rich semantic relations: CTD, DrugBank [48], DisGeNET [49], Gene Ontology(GO) [50], HGNC [51], BioGRID [52] and MedGen[1]. Table 1 lists the statistics of entities and links in these datasets and their detailed descriptions. It is worth noting that the disease information network constructed in EDEN can preserve the original and rich biological knowledge in these databases. For example, the BioGRID database stores two relationships about protein–protein interactions (corresponding to the two types of links between proteins and proteins in the disease information network), one is generated by the reaction between proteins through physical contact (physical), and the other is generated by functional association (genetic). Similarly, the Gene Ontology database describes genes from three aspects: molecular function, cellular component, and biological process. And thus the types of links between genes and GO terms in the disease information network include: function, component, and process.

4.1.2. Baseline methods

In order to demonstrate the effectiveness of EDEN, we compared it with ten baseline methods: SNNMF [10], TS-SVD [12], RWHNDR [15], Heter-LP [17], DTINet [19], GNDD [24], DR-HGCN [25], DRWBNCF [28], the work of Ezzat et al. [11] and Wang et al. [13]. We applied the ideas of these methods to the datasets we constructed and then evaluated the effectiveness of EDEN by these experimental results. The detailed characteristics of these methods can be referred to Section 2.

4.1.3. Evaluation metrics

As the drug repositioning is essentially a prediction task, we employed several metrics (i.e. the area under the receiver operating characteristic (ROC) curve (**AUC**), the area under the precision-recall (PR) curve (**AUPR**), F1-score (**F1**) and Hits Ratio (**HR**)), which are widely used in prediction tasks, to evaluate the performance of EDEN as well as the baseline methods.

4.1.4. Experimental setup

Throughout each experiment, several running parameters in EDEN are set as in default: the dimension of each entity representations as 64, the epoch as 200, the step of the convolution operation (i.e., l) as 3 and the learning rate as 0.001 (this rate also denotes the ratio of learning data available in global embedding layer). Specifically, why these parameters are set in this way, we will describe in detail in Section 4.3. Moreover, the global embedding layer contains three parameters: ρ, ζ and w. The default values of them are 20, 40, and 10. Likewise, if there are similar parameters involved in the baseline methods we set them to the same standard as EDEN. Besides, to analyze the influence of the different proportion of training dataset on the predictive ability of EDEN and baseline methods, we randomly select four groups of training set and test set from the drug-disease bipartite network, and their distribution ratios are: 6:4, 7:3, 8:2 and 9:1, respectively.

All experiments were conducted on a PC with four GeForce RTX 2080 Ti GPU and 512 GB main memory, running the Ubuntu 20.04. All algorithms were implemented in Tensorflow and compiled by Python 3.7. We have released the codes at https://github.com/AbernHE/EDEN.

4.2. Effectiveness Evaluation (RQ1)

We first report the performance of all methods on the metrics AUPR, AUC, and F1, and then investigate their hits ratios in the case of randomly generated negative samples and true negative samples in the test set. In order to reduce occasionality, the experiments were repeated 50 times (i.e., these experiments were performed on randomly constructed training and testing sets for 50 times) and then the average results were obtained as our eventual reports.

The performance comparison results of EDEN and other baseline methods are presented in Table 2. Based on such results, we have the following observations:

In general, EDEN consistently yields outstanding performance on these four groups of datasets with different training ratios under the metrics AUC, AUPR and F1. Specifically, EDEN attains AUC scores of 0.8686, 0.8977, 0.9002 and 0.9032, along with AUPR scores of 0.9436, 0.9586, 0.964 and 0.9698 and F1 scores of 0.8679, 0.8712, 0.8961 and 0.8972 when the ratio of training set to test set is 6:4, 7:3, 8:2 and 9:1, respectively. In particular, EDEN obviously improves over the strongest baseline methods *w.r.t.* AUC, AUPR and F1 by 0.84%, 0.2% and 0.45% in the training ratio of 6:4, respectively. In other cases, compared with the best baseline, the improvement effect of AUPR is 0.34% and 0.67% as well as of F1 is 0.56%, 2.94% and 3.2%, respectively. We can also find that as the training ratio increases, the performance of EDEN under the three evaluation metrics of AUPR, AUC and F1 also increases steadily. Although EDEN *w.r.t.* AUPR only does not reach the best in the training ratio of 9:1 and *w.r.t.* AUC does not reach the best in the training ratio of 7:3, 8:2, and 9:1, it has decent scores. In these results, we also found that while the baseline DRWBNCF outperformed EDEN in terms of the AUC metric for three different training ratio datasets, DRWBNCF significantly underperformed in terms of AUPR and F1 metrics. This is because DRWBNCF can be good at improving its accuracy and precision under the task of drug repositioning, but there is a clear limitation in its performance for recall, whereas EDEN could perform well under all of these evaluation metrics. In addition, we note that although the two evaluation metrics, AUPR and F1, are both obtained by comprehensively considering precision and recall, there are obvious differences in the scores of these two metrics in our some experimental results. This is because the precision-recall curve formed by the AUPR metric would gradually become smooth over iterations so that the area under the curve (i.e. the AUPR score) would be higher than the F1 score which is a metric of purely numerical calculation. (In other words, the metric calculates a harmonic mean of precision and recall at each iteration, and then averages them at the end so that the score is balanced and not too high.).

Besides, we also conduct experiments to investigate the hits ratio performance of EDEN and baseline methods in the negative sample of the test set generated from random and real cases under different training ratios. The results are shown in Fig. 2. We can observe that whether the negative samples in the test set come from the randomly generated (Fig. 2 (a)) or the real ones, which are from the repoDB database except for the data whose status is "approved" (Fig. 2 (b)), EDEN outperforms the other baseline methods on the whole. In these experimental results, only one baseline, DRWBNCF, outperformed EDEN under the HR metric when the negative sample of the test set come from the real ones and when the ratio of training to test set is 9:1.

[1] https://www.ncbi.nlm.nih.gov/medgen/

Table 1
Descriptions and statistics of the datasets.

Entity Type	# of Entities	Link Type	# of Links (Type)	Density/Max	Specific Semantic	Data Source
		Drug — Disease	30305 (2)	3.63/273	therapeutic, approved	CTD & repoDB
Drug	8352	Drug — Protein	417781 (1)	55.41/1939	target	DrugBank
Disease	17093	Disease — Gene	84038 (1)	0.73/340	curated_associate	DisGeNET
Gene	114472	Gene — Protein	20135 (11)	0.17/5	gene_with_protein_product, protocadherin, immunoglobulin, etc.	HGNC
Protein	22096					
GO term	24093	Protein — Protein	572512 (2)	135.39/2940	physical, genetic	BioGRID
Pathway	2363	GO — Gene	1048575 (3)	3.85/261	component, process, function	Gene Ontology
Phenotype	8299	Gene — Pathway	135814 (1)	1.19/357	participation	CTD
		Disease — Phenotype	158640 (35)	7.27/186	has_manifestation, related_to, etc.	MedGen

# of Links (Type) indicates the number of links (outside the parenthesis) and the number of link types (inside the parenthesis); Density refers to the average number of links of this type per instance entity; Max represents the maximum number of links owned by an entity among the instance entities involved in this type of link; Specific Semantic refers to the specific semantics of each link type under the relationship.

Table 2
The prediction performance comparison of EDEN with baseline methods using different training and testing set ratios.

Method	train: test | 6: 4			train: test | 7: 3			train: test | 8: 2			train: test | 9: 1		
	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1	AUC	AUPR	F1
Ezzat et al. [11]	0.5253	0.6504	0.8020	0.8542	0.9252	0.8256	0.8788	0.9391	0.8286	0.8916	0.9435	0.8272
Wang et al. [13]	0.7034	0.8201	0.8139	0.7065	0.8186	0.8178	0.7299	0.8429	0.8166	0.7601	0.8705	0.8171
Heter-LP [17]	0.7514	0.8544	0.8144	0.7929	0.8874	0.8163	0.8100	0.8924	0.8156	0.8136	0.8964	0.8160
SNNMF [10]	0.8241	0.9280	0.8634	0.8372	0.9325	0.8653	0.8403	0.9321	0.8667	0.8349	0.9274	0.8652
RWHNDR [15]	0.8322	0.9315	0.8508	0.8366	0.9332	0.8522	0.8149	0.9224	0.8513	0.8448	0.9368	0.8506
DR-HGCN [25]	0.8331	0.9173	0.8258	0.8661	0.9325	0.8267	0.8902	0.9452	0.8294	0.8993	0.9479	0.8283
GNDD [24]	0.8576	0.9416	0.8099	0.8968	0.9552	0.8656	0.8933	0.9573	0.8652	0.8977	**0.9729**	0.8641
TS-SVD [12]	0.8599	0.9337	0.8257	0.8755	0.9415	0.8312	0.8789	0.9414	0.8286	0.8791	0.9416	0.8296
DTINet [19]	0.8602	0.9336	0.8142	0.8888	0.9476	0.8165	0.8787	0.9414	0.8161	0.8794	0.9417	0.8166
DRWBNCF [28]	0.8497	0.4404	0.2890	**0.9266**	0.2853	0.3381	**0.9323**	0.3126	0.3567	**0.9390**	0.3297	0.3732
EDEN	**0.8686**	**0.9436**	**0.8679**	0.8977	**0.9586**	**0.8712**	0.9002	**0.9640**	**0.8961**	0.9032	0.9698	**0.8972**

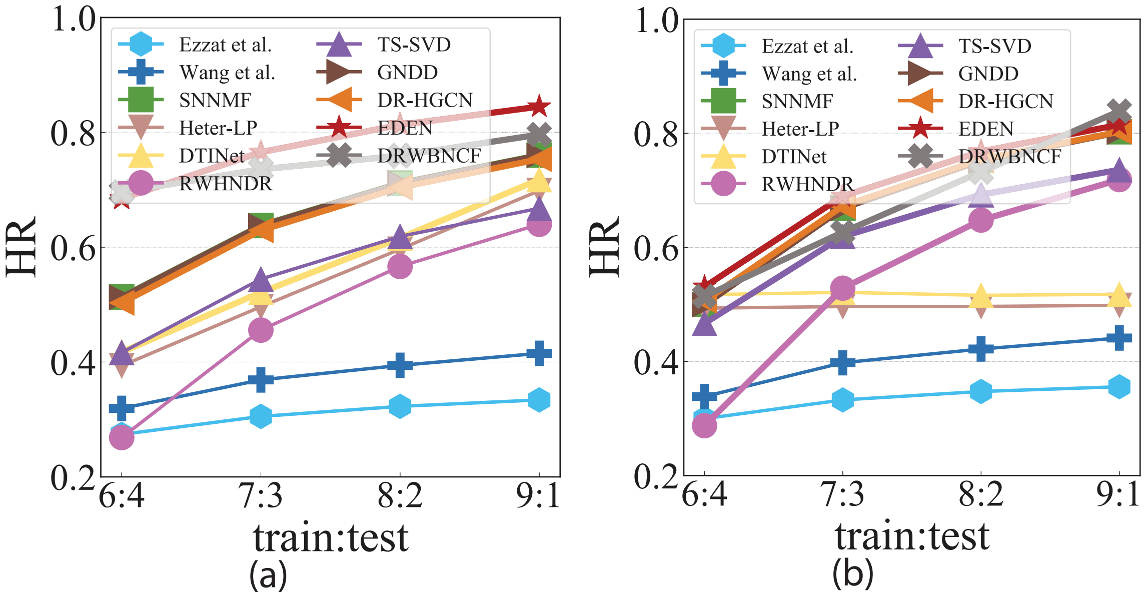
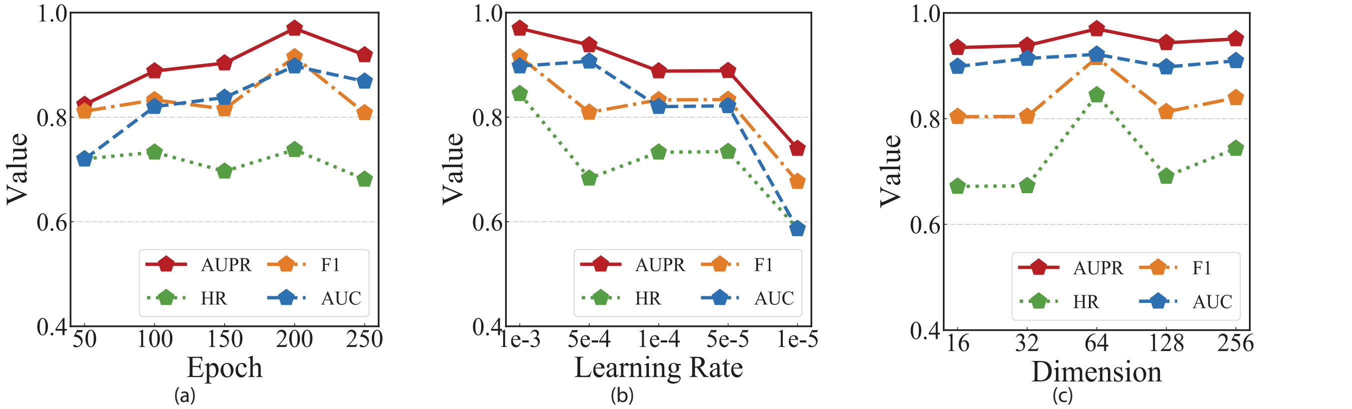


Fig. 2. Performance comparison of proposed EDEN with ten baseline methods over the HR on different training set ratios. (a) Results where the negative samples are randomly generated. (b) Results where the negative samples are obtained from real ones.

4.3. Parameter Sensitivity & Ablation Study (RQ2)

To get deep insights on each component and parameter settings of EDEN, we investigate their impact.

EDEN contains several hyper-parameters, which have been tested to evaluate their impacts on EDEN measured by AUPR, AUC, F1 and HR scores, including the epoch and learning rate in the optimization process, and the dimension of embeddings. In this part, all experiments are performed on our dataset with the ratio of training set to test set is 9:1. When comparing a parameter, we keep the other parameters unchanged. Their performances are presented in Fig. 3. To verify the effect of epoch on EDEN, we change the epoch in {50, 100, 150, 200, 250}. From Fig. 3 (a), it can be seen that with the increase of epoch, the scores of all evaluation metrics are increasing. The four evaluation scores reach the best when the epoch is set to 200. Similarly, to analyze the influence of the learn-



(a)

(b)

(c)

Fig. 3. Analysis of impact of parameters on the performance: (a) epoch, (b) learning rate and (c) dimension using four evaluation metrics: AUPR, AUC, F1 and HR.

ing rates on the EDEN, we vary the learning rate from {0.001, 0.0005, 0.0001, 0.00005, 0.00001}. It is observed from Fig. 3 (b) that when the learning rate is set to 0.001, EDEN obtains the best performance. We also examine the sensitivity of the dimension when it changes in {16, 32, 64, 128, 256}. As shown in Fig. 3 (c), when the dimension is set to 64, EDEN has the best scores.

Next, we change the step of the convolution operation l in EDEN to investigate the efficiency of gathering multi-step neighbors' messages. In particular, the step number is searched in the range of {1, 2, 3, 4}; we use EDEN-1 to indicate the model with one step, and similar notations for others. We summarize the results in Table 3, and have the following observations: increasing the step of the convolution operation in EDEN is capable of improving the performance. Clearly, EDEN-3 achieves consistent improvement over others across all the board. To verify the impact of *local embedding layer*, *global embedding layer* and mutual information maximization from *optimization layer*, we do ablation study by considering three variants of EDEN. In particular, we only use the local embedding layer, termed EDEN-loc. Similarly, only using the global embedding layer termed EDEN-glo. Based on using these two components, the information maximization operation is not used, and the embedding obtained by the two layers is simply concatenate termed EDEN-g&l. We also summarize the experimental results in Table 3 and have the following finding: removing any of these three components degrades the model's performance. Especially the removal of mutual information maximization and local embedding layer has the greatest impact.

Table 3
Effect of different convolution steps and components.

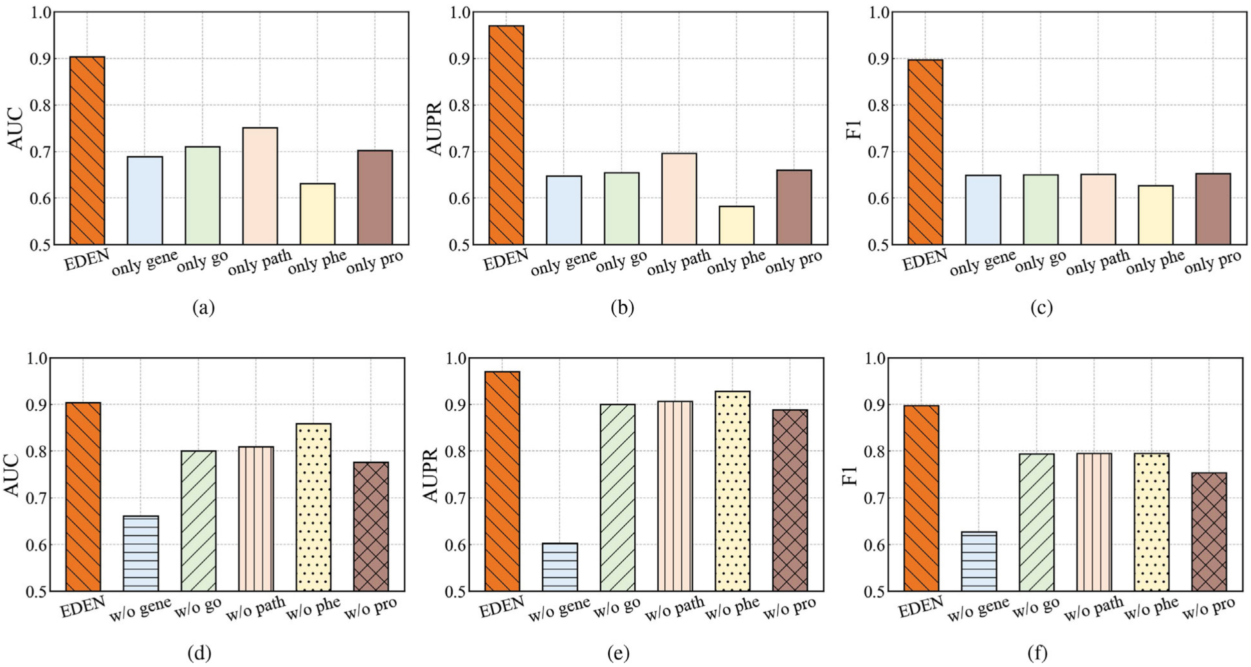
	AUPR	AUC	F1	HR
EDEN-1	0.9080	0.8686	0.8093	0.6834
EDEN-2	0.9276	0.8897	0.8448	0.7877
EDEN-3	**0.9698**	**0.9032**	**0.8972**	**0.8288**
EDEN-4	0.9393	0.8217	0.8194	0.6931
EDEN-glo	0.7770	0.7061	0.6347	0.5363
EDEN-loc	0.8429	0.7147	0.7282	0.7478
EDEN-g&l	0.8508	0.7278	0.7285	0.7582

Furthermore, the definition of a disease information network is also an important part of our approach. Therefore, inspired by [53], we change the input of EDEN, that is, use the disease information network of different scales, to explore the influence of the disease information network constructed by different types of biomedical entities on EDEN. Specifically, here we generated ten inputs of different scales based on the five types of biomedical entities included in the disease information network constructed by datasets we collected. Inputs of "only gene" and "only phe" respectively represent that the entities contained in the disease information network are only diseases and genes or phenotypes. Since the construction of the network needs to ensure its connectivity, gene entities need to be added when only entities GO term, pathway, or protein are considered in this construction. Therefore, "only go", "only path", and "only pro" indicate that the only entities included in the disease information network are GO terms, pathways or proteins in addition to diseases and genes, respectively. These entities then form five different disease information networks through the links between them. Similarly, "w/o go", "w/o path", "w/o phe", and "w/o pro" represent the network generated by only removing entities GO terms, pathways, phenotypes or proteins from the entire disease information network (the original input of EDEN), respectively. We note that the network generated by "w/o gene" is the same as that of "only phe". Because the entity gene was removed from the entire disease information network constructed by EDEN, the other three types of entities it links cannot be connected in the network.

In this way, we generated ten different disease information networks to analyze the impact on EDEN by the three metrics AUC, AUPR, and F1 under the default experimental settings. From Fig. 4, it can be found that when the disease information network constructed in our model contains less information, the prediction performance of EDEN on drug repositioning would be weaker. In particular, the information contained in the entity genes and proteins has the most significant impact on the performance of EDEN. This can be observed from Figs. 4 (d), (e), and (f) that the performance of "w/o gene" and "w/o pro", i.e. the network with the entity genes and proteins removed, is low compared to the performance produced by other disease information networks. Besides, it can also be observed from Fig. 4 that the performance is worst when the network "only phe" contains only entity phenotypes, and the performance of EDEN is least affected ("w/o phe") when only entity phenotypes are removed from the whole network. This reflects that the entity phenotypes in the disease information network have the least impact on EDEN.

4.4. Case Study (RQ3)

To further testify the effectiveness of EDEN and generated explanation paths, we selected four drugs *Piroxicam*, *Triamcinolone*, *Atorvastatin* and *Methylphenidate* as query objects. Firstly, through EDEN, we respectively predicted ten diseases that are potentially associated with *Piroxicam*, *Triamcinolone* and *Atorvastatin* these three drugs, which are in the top ten of prediction results. The results are shown in Table 4. It can be seen that the prediction results are supported by corresponding evidence reported in rele-



(a) (b) (c)

(d) (e) (f)

Fig. 4. Analysis of the impact of constructing different disease information networks on EDEN performance.

Table 4
The top-10 prediction results and their support evidence for the query drugs Piroxicam, Triamcinolone and Atorvastatin.

Query	Top-10 prediction results	Evidence*
Piroxicam	Sexual Dysfunction, Physiological	PMID:18726914
	Hypertension	PMID:15199296
	Migraine	PMID:30219683
	Hypertrophy	PMID:27652271
	Subarachnoid Hemorrhage	PMID:27157545
	Melanoma	PMID:24495407
	Breast Neoplasms	PMID:15802278
	Seizures	PMID:19488739
	Movement Disorders	PMID:26526685
	Acute Kidney Injury	PMID:12185885
Triamcinolone	Seizures	PMID:14714756
	Neoplasms	PMID:18494554
	Campylobacter Infection	-
	Edema	PMID: 1459535
	Hypertension	PMID:20667508
	Carcinoma	PMID:15637090
	Proteinuria	PMID: 4368615
	Bradycardia	PMID: 1676337
	Pain	PMID:20133530
	Chancroids	-
Atorvastatin	Kidney Diseases	PMID:11682445
	Seizures	PMID:18096215
	Edema	PMID: 4327920
	Chemical and Drug Induced Liver Injury	PMID:20623750
	Inflammation	PMID:16025360
	Hypertension	PMID:16620303
	Breast Neoplasms	PMID:16322251
	Chronic Myeloproliferative Disorder	-
	Arrhythmias, Cardiac	PMID: 1654493
	Pulmonary Hypertension	-

PMID represents literature's PubMed (https://pubmed.ncbi.nlm.nih.gov/) ID.

vant literatures. We then selected *Methylphenidate* and *Hyperalgesia* as the query pair, and the top ten explanation paths extracted by EDEN are shown in Table 5. For the third path in Table 5, we can observe from the relevant literatures:

Methylphenidate (MPH) is the first-line treatment of choice for attention-deficit hyperactive disorder (ADHD) [54]. However, a considerable interindividual variability exists in clinical outcome, which may reflect underlying genetic influences. The presumed mechanism of MPH activity is in blockade of the dopamine transporter (DAT), inhibiting the reuptake of monoamine, such as dopamine and noradrenaline, leading to increased synaptic catecholamines [55]. Thus, the DAT1 gene has long been considered a prime candidate that may contribute to the effectiveness and safety of MPH [56].

Both SLC6A3 and Catechol-O-methyltransferase (COMT) are well-known genes that play important roles in the pathophysiology of different psychiatric illnesses including ADHD. SLC6A3 and COMT are also two well characterized polymorphisms in dopamine-related genes. Two particular polymorphisms in SLC6A3 and COMT may alter the response function of dopamine and, in turn, activation underlying response inhibition. SLC6A3 encodes DAT, which is responsible for removing dopamine from the synapse [57]. COMT is the major catecholamine-degrading enzyme involved in the degradation of catecholamines in synapses, preferentially affecting prefrontal cortical dopamine metabolism [58].

Hyperalgesia is characterized by decreased pain threshold, increased pain to normally painful stimuli, and spontaneous pain. Sensitization of peripheral nociceptors or central pain-encoding neurons leading to hyperalgesia [59]. More and more literature demonstrated that catecholamines and pathways regulating their bioavailability influence pain [60]. Here, based on the theory of pharmacogenetics and pharmacogenomics, we could infer that MPH is the potential treatment option for hyperalgesia patients with genetic polymorphism in COMT, SLC6A3, or other catecholamines system related genes. This shows that the third explanation path extracted by EDEN is reasonable.

Similarly, we also generated explainable paths for some of the prediction results in Table 4. Due to space limitations, we only list the paths that are supported by evidence in Table 6. We can observe that EDEN demonstrates a good interpretability for these prediction results.

Table 5
The explanation paths for the query pair Methylphenidate and Hyperalgesia.

Query pair	Top-10 explanation paths
Methylphenidate - Hyperalgesia	1.Methylphenidate -*target*- 5-hydroxytryptamine receptor 1A -*gene_with_protein_product*- HTR1A -*participation*- Serotonin receptors -*participation*- HTR2A -*curated_associate*- Hyperalgesia
	2.Methylphenidate -*target*- 5-hydroxytryptamine receptor 1A -*gene_with_protein_product*- HTR1A -*participation*- Amine ligand-binding receptors -*participation*- DRD3 -*curated_associate*- Hyperalgesia
	3.Methylphenidate -*target***- Sodium- dependent dopamine transporter -***gene_with_protein_product***- SLC6A3 -***participation***- Dopamine clearance from the synaptic cleft -***participation***- COMT -***curated_associate***- Hyperalgesia**
	4.Methylphenidate -*target*- Sodium- dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Cocaine addiction -*participation*- PDYN -*curated_associate*- Hyperalgesia
	5.Methylphenidate -*target*- Sodium- dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Transmission across Chemical Synapses -*participation*- COMT -*curated_associate*- Hyperalgesia
	6.Methylphenidate -*target*- Sodium- dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Neurotransmitter Clearance In The Synaptic Cleft -*participation*- COMT -*curated_associate*- Hyperalgesia
	7.Methylphenidate -*target*- Sodium- dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Amphetamine addiction -*participation*- PRKCG -*curated_associate*- Hyperalgesia
	8.Methylphenidate -*target*- 5-hydroxytryptamine receptor 1A -*gene_with_protein_product*- HTR1A -*participation*- cAMP signaling pathway -*participation*- MAPK8 -*curated_associate*- Hyperalgesia
	9.Methylphenidate -*target*- 5-hydroxytryptamine receptor 1A -*gene_with_protein_product*- HTR1A -*participation*- Taste transduction -*participation*- P2RX3 -*curated_associate*- Hyperalgesia
	10.Methylphenidate -*target*- Sodium- dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Neuronal System -*participation*- COMT -*curated_associate*- Hyperalgesia

5. Conclusion

Drug repositioning is a good option for reducing the cost of new drugs research and shortening its developing time. The rapid development of high-throughput technologies and the explosion of various biomedical data have provided lots of reliable and abundant resources for identifying drug-target and drug-disease interactions through computational approaches. Recently, a variety of computer techniques such as matrix factorization and completion, machine learning and deep neural network, have been used to develop computational drug repositioning models. Different from these studies, this paper mainly focuses on the interpretability of predicted drug repositioning results, which has not been considered in previous methods, yet is an important factor in the follow-up in-depth understanding of the prediction results.

Hence, in this work, we propose a novel framework named EDEN, which leverages the idea of the graph neural network to capture the features of local semantics and global structures from the unified graph of the DIN and the drug-disease bipartite network, and iteratively aggregates the heterogeneous information of neighbors with attentive weights to update all biomedical entities and links embedding based on maximizing the mutual information of these two features. Finally, we can not only predict potential drug-disease associations, but also generate explanation paths for them by the total embeddings, which are retrieved through the proposed matching score. Extensive experiments on real-world datasets demonstrate the rationality and effectiveness of EDEN. Further case studies also show the validity of the explanation paths generated by EDEN.

For the future work, improvements can be made in considering expanding the rich information of the drug (scuh as the molecular structure and side effects) into the model, integrating literature knowledge to introduce the decision process into explanations generation, and developing a system for updating the results in real time when new data sources are available. In addition, making the explanation paths generated for the prediction results more natural language, i.e. generating texts that can be directly understood and read, is also an interesting issue for further research.

Table 6
The explanation paths generated from the prediction results shown in Table 4, with evidence supported by literatures.

Explanation paths	Evidence*
Piroxicam -*target*- Prostaglandin G/H synthase 2 -*gene_with_protein_product*- PTGS2 -*process*- inflammatory response -*process*- PTGER1 -*curated_associate*- Sexual Dysfunction, Physiological	PMID:24996777
Piroxicam -*target*- Prostaglandin G/H synthase 2 -*gene_with_protein_product*- PTGS2 -*process*- positive regulation of vasoconstriction -*process*- HTR2A -*curated_associate*- Migraine	PMID:30219683 & 27489378
Piroxicam -*target*- Prostaglandin G/H synthase 2 -*gene_with_protein_product*- PTGS2 -*process*- positive regulation of prostaglandin biosynthetic process -*process*- PTGS2 -*curated_associate*- Seizures	PMID: 9642033
Triamcinolone -*target*- Glucocorticoid receptor -*gene_with_protein_product*- NR3C1 -*participation*- Gene Expression -*participation*- VDR -*curated_associate*- Neoplasms	PMID:24128352 & 27768599
Triamcinolone -*target*- Glucocorticoid receptor -*gene_with_protein_product*- NR3C1 -*participation*- Circadian Clock -*participation*- AVP -*curated_associate*- Pain	PMID:31895268 & 32761684
Triamcinolone -*target*- Glucocorticoid receptor -*gene_with_protein_product*- NR3C1 -*participation*- Gene Expression -*participation*- APOE -*curated_associate*- Proteinuria	PMID:24128352 & 31019291
Atorvastatin -*target*- Histone deacetylase 2 -*physical*- Serum albumin -*gene_with_protein_product*- ALB -*curated_associate*- Kidney Diseases	PMID:34154367
Atorvastatin -*target*- Histone deacetylase 2 -*physical*- Signal transducer and activator of transcription 3 -*gene_with_protein_product*- STAT3 -*curated_associate*- Inflammation	PMID:22197944 & 33952812
Atorvastatin -*target*- Histone deacetylase 2 -*physical*- Hypoxia-inducible factor 1-alpha -*gene_with_protein_product*- HIF1A -*curated_associate*- Hypertension	PMID:33908728
Methylphenidate -*target*- 5-hydroxytryptamine receptor 1A -*gene_with_protein_product*- HTR1A -*participation*- cAMP signaling pathway -*participation*- MAPK8 -*curated_associate*- Hyperalgesia	PMID:31371490 & 33744339
Methylphenidate -*target*- Sodium-dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Transmission across Chemical Synapses -*participation*- MAPK1 -*curated_associate*- Hypertrophy	PMID:28955722 & 32103377
Methylphenidate -*target*- Sodium-dependent dopamine transporter -*gene_with_protein_product*- SLC6A3 -*participation*- Neuronal System -*participation*- SLC22A2 -*curated_associate*- Kidney Diseases	PMID:32927790 & 23958595

PMID represents literature's PubMed (https://pubmed.ncbi.nlm.nih.gov/) ID.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (61972268), the Sichuan Science and Technology Program (2020YFG0034), and the Med-X Center for Informatics Funding Project of SCU (YGJC001).

References

[1] N. Berdigaliyev, M. Aljofan, An overview of drug discovery and development, Future, Medicinal Chemistry 12 (2020) 939–947.

[2] J.-P. Jourdan, R. Bureau, C. Rochais, P. Dallemagne, Drug repositioning: A brief overview, The Journal of Pharmacy and Pharmacology 72 (2020) 1145–1151.

[3] H. Luo, M. Li, M. Yang, F.-X. Wu, Y. Li, J. Wang, Biomedical data and computational models for drug repositioning: A comprehensive review, Briefings in Bioinformatics 22 (2021) 1604–1619.

[4] F. Wang, Y. Ding, X. Lei, B. Liao, F.-X. Wu, Identifying gene signatures for cancer drug repositioning based on sample clustering, IEEE ACM Transactions on Computational Biology and Bioinformatics 19 (2022) 953–965.

[5] Y.-Y. Wang, C. Cui, L. Qi, H. Yan, X.-M. Zhao, DrPOCS: Drug repositioning based on projection onto convex sets, IEEE ACM Transactions on Computational Biology and Bioinformatics 16 (2019) 154–162.

[6] K. Yang, X. Zhao, D. Waxman, X.-M. Zhao, Predicting drug-disease associations with heterogeneous network embedding, Chaos 29 (2019) 123109.

[7] Y.-Y. Wang, H. Bai, R.-Z. Zhang, H. Yan, K. Ning, X.-M. Zhao, Predicting new indications of compounds with a network pharmacology approach: Liuwei Dihuang Wan as a case study, Oncotarget 8 (2017) 93957–93968.

[8] Y. Yan, M. Yang, H. Zhao, G. Duan, X. Peng, J. Wang, Drug repositioning based on multi-view learning with matrix completion, Briefings in Bioinformatics 23 (2022).

[9] W. Dai, X. Liu, Y. Gao, L. Chen, J. Song, D. Chen, K. Gao, Y. Jiang, Y. Yang, J. Chen, P. Lu, Matrix factorization-based prediction of novel drug indications by integrating genomic space, Computational and Mathematical Methods in Medicine 2015 275045 (1–275045) (2015) 9.

[10] W. Zhang, F. Huang, X. Yue, X. Lu, W. Yang, Z. Li, F. Liu, Prediction of drug-disease associations and their effects by signed network-based nonnegative matrix factorization, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, December 3–6, 2018, pp. 798–802.

[11] A. Ezzat, P. Zhao, M. Wu, X. Li, C.K. Kwoh, Drug-target interaction prediction with graph regularized matrix factorization, IEEE ACM Transactions on Computational Biology and Bioinformatics 14 (2017) 646–656.

[12] G. Wu, J. Liu, Predicting drug-disease treatment associations based on topological similarity and singular value decomposition, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18–21, 2019, pp. 153–158.

[13] R. Wang, S. Li, M.H. Wong, K.-S. Leung, Drug-protein-disease association prediction and drug repositioning based on tensor decomposition, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, Madrid, Spain, December 3–6, 2018, pp. 305–312.

[14] M. Bagherian, R.B. Kim, C. Jiang, M.A. Sartor, H. Derksen, K. Najarian, Coupled matrix-matrix and coupled tensor-matrix completion methods for predicting drug-target interactions, Briefings in Bioinformatics 22 (2021) 2161–2171.

[15] H. Luo, J. Wang, M. Li, J. Luo, P. Ni, K. Zhao, F.-X. Wu, Y. Pan, Computational drug repositioning with random walk on a heterogeneous network, IEEE ACM Transactions on Computational Biology and Bioinformatics 16 (2019) 1890–1900.

[16] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F.-X. Wu, Y. Pan, Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm, Bioinformatics 32 (2016) 2664–2671.

[17] M.L. Shahreza, N. Ghadiri, S.R. Mousavi, J. Varshosaz, J.R. Green, Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning, Journal of Biomedical Informatics 68 (2017) 167–183.

[18] W. Wang, S. Yang, X. Zhang, J. Li, Drug repositioning by integrating target information through a heterogeneous network model, Bioinformatics 30 (2014) 2923–2930.

[19] X. Ji, J.M. Freudenberg, P. Agarwal, Integrating biological networks for drug target prediction and prioritization, Methods in Molecular Biology 2019 (1903) 203–218.

[20] R. Zhou, Z. Lu, H. Luo, J. Xiang, M. Zeng, M. Li, NEDD: A network embedding based method for predicting drug-disease associations, BMC Bioinformatics 21-S (2020) 387.

[21] B. Zhao, L. Hu, Z. You, L. Wang, X. Su, HINGRL: Predicting drug-disease associations with graph representation learning on heterogeneous information networks, Briefings in Bioinformatics 23 (2022).

[22] Y. Wang, S. Chen, N. Deng, Y. Wang, Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data, PLoS One 8 (2013) e78518.

[23] R.S. Olayan, H. Ashoor, V.B. Bajic, DDR: Efficient computational method to predict drug-target interactions using graph mining and machine learning approaches, Bioinformatics 34 (2018) 1164–1173.

[24] B. Wang, X. Lyu, J. Qu, H. Sun, Z. Pan, Z. Tang, GNDD: A graph neural network-based method for drug-disease association prediction, in Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, 18–21, San Diego, CA, USA, November 2019, pp. 1253–1255.

[25] H. Sun, X. Lyu, B. Wang, Y. Wang, Z. Tang, Z. Liu, An enhanced LRMC method for drug repositioning via gcn-based HIN embedding, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, Virtual Event, South Korea, December 16–19, 2020, pp. 1137–1141.

[26] T.N. Jarada, J.G. Rokne, R. Alhajj, SNF-NN: Computational method to predict drug-disease interactions using similarity network fusion and neural networks, BMC Bioinformatics 22 (2021) 28.

[27] T.N. Jarada, J.G. Rokne, R. Alhajj, SNF-CVAE: Computational method to predict drug-disease interactions using similarity network fusion and collective variational autoencoder, Knowledge-Based Systems 212 (2021) 106585.

[28] Y. Meng, C. Lu, M. Jin, J. Xu, X. Zeng, J. Yang, A weighted bilinear neural collaborative filtering approach for drug repositioning, Briefings in Bioinformatics 23 (2022).

[29] D. Parisi, M.F. Adasme, A. Sveshnikova, S.N. Bolz, Y. Moreau, M. Schroeder, Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs, Computational and Structural, Biotechnology Journal 18 (2020) 1043–1055.

[30] X. Huang, B. Guo, Adenomatous polyposis coli determines sensitivity to histone deacetylase inhibitor-induced apoptosis in colon cancer cells, Cancer Research 66 (2006) 9245–9251.

[31] Q. Ai, V. Azizi, X. Chen, Y. Zhang, Learning heterogeneous knowledge base embeddings for explainable recommendation, Algorithms 11 (2018) 137.

[32] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, T.-S. Chua, Explainable reasoning over knowledge graphs for recommendation, in: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 5329–5336.

[33] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang, Reinforcement knowledge graph reasoning for explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25, 2019, pp. 285–294.

[34] A. Fout, J. Byrd, B. Shariat, A. Ben-Hur, Protein interface prediction using graph convolutional networks, in: Proceedings of the Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, December 4–9, 2017, pp. 6530–6539.

[35] K.Y. Gao, A. Fokoue, H. Luo, A. Iyengar, S. Dey, P. Zhang, Interpretable drug target prediction using deep neural representation, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13–19, 2018, pp. 3371–3377.

[36] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Lake Tahoe, Nevada, United States, December 5–8, 2013, pp. 2787–2795.

[37] C. He, L. Duan, H. Zheng, J. Li-Ling, L. Song, L. Li, Graph convolutional network approach to discovering disease-related circRNA-miRNA-mRNA axes, Methods 198 (2022) 45–55.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, December 4–9, 2017, pp. 5998–6008.

[39] B. Perozzi, R. Al-Rfou, S. Skiena, DeepWalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014, New York, USA, August 24–27, 2014, pp. 701–710.

[40] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, Workshop Track Proceedings, 2013, p. (poster).

[41] L.F.R. Ribeiro, P.H.P. Saverese, D.R. Figueiredo, struc2vec" Learning node representations from structural identity, in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17, Halifax, NS, Canada, August, 2017, pp. 385–394.

[42] Y. Dong, N.V. Chawla, A. Swami, metapath2vec: Scalable representation learning for heterogeneous networks, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17, 2017, pp. 135–144.

[43] P. Velickovic, W. Fedus, W.L. Hamilton, P. Liò, Y. Bengio, R.D. Hjelm, Deep graph infomax, in: Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019, p. (poster).

[44] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, p. (poster).

[45] Y. Zhang, L. Duan, H. Zheng, J. Li-Ling, R. Qin, Z. Chen, C. He, T. Wang, Mining similar aspects for gene similarity explanation based on gene information network, IEEE ACM Transactions on Computational Biology and Bioinformatics 19 (2022) 1734–1746.

[46] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, J. Wiegers, T.C. Wiegers, C.J. Mattingly, Comparative toxicogenomics database (CTD): Update 2021, Nucleic Acids Research 49 (2021) D1138–D1143.

[47] A.S. Brown, C.J. Patel, A standard database for drug repositioning, Scientific Data 4 (2017) 170029.

[48] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: A major update to the DrugBank database for 2018, Nucleic Acids Research 46 (2018) D1074–D1082.

[49] J. Piñero, J.M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L.I. Furlong, The disgenet knowledge platform for disease genomics, update, Nucleic Acids Research 48 (2020) (2019) D845–D855.

[50] Gene Ontology Consortium, The gene ontology resource: Enriching a GOld mine, Nucleic Acids Research 49 (2021) D325–D334.

[51] S. Tweedie, B. Braschi, K. Gray, T.E.M. Jones, R.L. Seal, B. Yates, E.A. Bruford, Genenames.org: the HGNC and VGNC resources in 2021, Nucleic Acids Research 49 (2021) D939–D946.

[52] R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, M. Tyers, The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions, Protein Science: a Publication of the Protein Society 30 (2021) 187–200.

[53] G. Dong, J. Feng, F. Sun, J. Chen, X.-M. Zhao, A global overview of genetically interpretable multimorbidities among common diseases in the UK Biobank, Genome Medicine 13 (2021) 110.

[54] M.A. Katzman, T. Sternat, A review of OROS methylphenidate (concertaö) in the treatment of attention-deficit/hyperactivity disorder, CNS Drugs 11 (2014) 1005–1033.

[55] N.D. Volkow, G.-J. Wang, J.S. Fowler, F. Telang, L. Maynard, J. Logan, S.J. Gatley, N. Pappas, C. Wong, P. Vaska, W. Zhu, J.M. Swanson, Evidence that methylphenidate enhances the saliency of a mathematical task by increasing dopamine in the human brain, American Journal of Psychiatry 161 (2004) 1173–1180.

[56] S.H. VanNess, M.J. Owens, C.D. Kilts, The variable number of tandem repeats element in dat1 regulates in vitro dopamine transporter density, BMC Genetics 6 (2005) 55.

[57] M.J. Bannon, S.K. Michelhaugh, J. Wang, P. Sacchetti, The human dopamine transporter gene: gene organization, transcriptional regulation, and potential involvement in neuropsychiatric disorders, The Journal of the European College of, Neuropsychopharmacology 11 (2001) 449–455.

[58] T. Lotta, J. Vidgren, C. Tilgmann, I. Ulmanen, K. Melén, I. Julkunen, J. Taskinen, Kinetics of human soluble and membrane-bound catechol o-methyltransferase: A revised mechanism and description of the thermolabile variant of the enzyme, Biochemistry 34 (1995) 4202–4210.

[59] R.D. Treede, R.A. Meyer, S.N. Raja, J.N. Campbell, Peripheral and central mechanisms of cutaneous hyperalgesia, Progress in Neurobiology 38 (1992) 397–421.

[60] J.E. Hartung, B.P. Ciszek, A.G. Nackley, β2- and β3-adrenergic receptors drive COMT-dependent pain by increasing production of nitric oxide and cytokines, Pain 155 (2014) 1346–1355.

Chengxin He received his B.S. degree in computer science from Sichuan University, China, in 2018. He is currently working towards the Ph.D. degree in the School of Computer Science, Sichuan University. His research interests include data mining, bioinformatics, and graph representation learning.

Lei Duan received his B.S. and Ph.D. degrees both in computer science from Sichuan University, China, in 2003 and 2008, respectively. He was a visiting Ph.D. student in the Department of Computer Science and Engineering, Wright State University, Dayton, Ohio from 2007 to 2008, and was a visiting scholar in the School of Computing Science, Simon Fraser University, Canada, from 2012 to 2013. He is currently a professor in the School of Computer Science, Sichuan University, China. His research interests include data mining, knowledge management, evolutionary computation, bioinformatics, and health informatics.

Huiru Zheng received her B.S., M.S., and Ph.D. degrees from Zhejiang University, Fuzhou University, and the University of Ulster in 1989, 1992, and 2003, respectively. Her research area includes data mining, machine learning, artificial intelligence, and their applications. She has published over 330 research papers in peer-reviewed international journals and conferences. Dr. Zheng is currently a professor of Computer Science and AI Research Centre Theme Lead with the School of Computing at Ulster University, UK.

Linlin Song received her B.M. degree from Fudan University in 2017, and the M.S. degree from Sichuan University in 2020. She is currently studying for Ph.D. in Medical Imaging, West China Hospital, Sichuan University. Her research focuses on application of nano materials in cancer therapy.

Menglin Huang received her B.S. degree from Sichuan University in 2020. She is currently studying for her M.S. degree in the School of Computer Science, Sichuan University, China. Her research interests include data mining and natural language processing.