A Multimodal Framework for Improving *in Silico* Drug Repositioning With the Prior Knowledge From Knowledge Graphs

Zhankun Xiong, Feng Huang, Ziyan Wang, Shichao Liu, and Wen Zhang

Abstract—Drug repositioning/repurposing is a very important approach towards identifying novel treatments for diseases in drug discovery. Recently, large-scale biological datasets are increasingly available for pharmaceutical research and promote the development of drug repositioning, but efficiently utilizing these datasets remains challenging. In this paper, we develop a novel multimodal framework, termed GraphPK (Graph-based Prior Knowledge) for improving *in silico* drug repositioning via using the prior knowledge from a drug knowledge graph. First, we construct a knowledge graph by integrating relevant bio-entities (drugs, diseases, etc.) and associations/interactions among them, and apply the knowledge graph embedding technique to extract prior knowledge of drugs and diseases. Moreover, we make use of the known drug-disease association, and obtain known association-based features from an association bipartite graph through graph embedding, and also take into account biological domain features, i.e., drug chemical structures and disease semantic similarity. Finally, we design a multimodal neural network to combine three types of features from the knowledge graph, the known associations and the biological domain, and build the prediction model for predicting drug-disease associations. Massive experiments show that our method outperforms other state-of-the-art methods in terms of most metrics, and the ablation analysis regarding the three types of features reveals that prior knowledge from knowledge graphs can not only lift the predictive power of *in silico* drug repositioning, but also enhance the model's robustness to different scenarios. The results of case studies offer support that GraphPK has the potential for actual use.

Index Terms—Knowledge graph, graph convolutional network, drug-disease associations, multimodal

1 Introduction

The development of a new drug is a risky, time-consuming, high-cost, and low-efficiency process now that it costs about $2.5 billion and takes at least 12 years to bring a single drug to market [1]. Drug repositioning/ repurposing is a process of finding novel indications for drugs that have already been approved, and offers a relatively low-cost and high-efficiency substitution for drug discovery. There have been immense researches on drug repositioning since it was firstly proposed in 2004 [2]. Among them, drug repositioning through wet methods is still costly and clumsy [3]. Consequently, computational drug repositioning, especially the machine learning-based methods of predicting drug-disease associations, has caught much attention in the past decade. Existing computational methods can be roughly grouped into three categories [4], [5], including classification-based methods, matrix factorization-based methods and network-based methods.

Classification-based methods manually assemble feature vectors to characterize drug-disease pairs and train a binary classifier to discriminate the associated pairs from others [6], [7], [8]. The matrix factorization-based methods reconstruct the known drug-disease associations by the product of decomposed factors that are usually constrained by some side information of drugs and diseases [9], [10], [11]. For example, DisDrugPred [9] is a non-negative matrix factorization method that integrates prior information by coupled factorization of multiple similarity matrices and SCMFDD [10] imposed a similarity-based manifold regularizer into the matrix factorization model. Network-based methods construct complex biomedical networks and treat drug repositioning as a link prediction in the drug-disease association bipartite graph [12], [13], [14], [15]. For example, RWHNDR [12] prioritized drug candidates for diseases by using the random walk model on a heterogeneous network. Recently, network-based deep learning methods for drug repositioning have gained popularity. DeepDR [16] is a network-based deep-learning approach that fuses multiple drug-related networks for drug repositioning. LAGCN [5] presented a layer-attentive graph convolutional network model to absorb heterogeneous information for drug-disease association prediction. In general, the existing methods make use of the known drug-disease associations and biological features to develop prediction models.

Among these advances, it seems to be a consensus that incorporating multiple data sources can potentially boost the accuracy of *in silico* drug repositioning [16], [17]. After all, heterogeneous data sources such as associations or interactions among different bio-entities (e.g., drug-target interactions, miRNA-disease associations, etc.) indicating drug

- *The authors are with the College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, China. E-mail: {xiongzk, fhuang233, wangziyan}@webmail.hzau.edu.cn, {scliu, zhangwen}@mail.hzau.edu.cn.*

Manuscript received 21 February 2021; revised 6 July 2021; accepted 24 July 2021. Date of publication 10 August 2021; date of current version 7 October 2022.
(Corresponding author: Wen Zhang.)
Digital Object Identifier no. 10.1109/TCBB.2021.3103595

TABLE 1
Summary of Datasets

Dataset	Drugs	Diseases	Associations	Density*	Knowledge Graph			
					Entities	Entities types	Relations	Relation types
Zhang's dataset	267	570	17,414	0.1144	9,339	5	87,830	11
Zhang's therapeutic dataset	267	570	5,288	0.0347				
Amazon's dataset	3,302	2,917	44,856	0.0047	91,019	11	5,790,360	97

* *Density = the number of associations / the number of drug-disease pairs.*

chemistry and bioactivity, as well as disease pathogenesis and etiology, offer diverse knowledge and semantic information about complex pharmacology and biology. Chen *et al.* considered important relevant relations such as drug-target interactions and microRNA-small molecule associations for drug discovery [18], [19], [20], [21]. However, for drug repositioning, it is still full of challenges in extracting informative prior knowledge from the multi-resource data. The development of knowledge graph-related techniques, such as knowledge graph embedding (KGE) [22], [23], [24], provides a promising way to alleviate the challenge.

A knowledge graph is a graph-structured representation of multi-relational data, which contains rich semantic information and knowledge facts. The applications of knowledge graphs have been increasingly popular in many bioinformatics issues, such as drug-target interaction prediction [25] and drug-drug interaction prediction [26]. In terms of drug repositioning, Zhu *et al.* [27] introduced an approach to drug repositioning based on a drug knowledge graph. Ge *et al.* [28] applied both machine learning and statistical analysis approaches to integrate knowledge graphs, literature, and transcriptome data to discover the potential drug candidates against SARS-CoV-2. Zeng *et al.* [29] reported an integrative network-based deep-learning methodology to identify repurposable drugs for COVID-19. Although successful applications of the knowledge graphs have demonstrated their great potential in drug discovery, several critical issues need to be solved and clarified. One issue of wide interest is whether the knowledge graph-derived features can take place of classic features, e.g., known association-derived features and biological features. If knowledge graph-derived features perform differently from classic features, another issue is how to utilize all of them for building better models.

In this study, we develop a novel multimodal framework, termed GraphPK (***Graph***-based **P**rior **K**nowledge) for improving *in silico* drug repositioning (drug-disease association prediction) via using the prior knowledge extracted from the medical knowledge graph. First, we construct a knowledge graph by integrating relevant bio-entities (drugs, diseases, etc.) and associations/interactions among them, and apply the knowledge graph embedding technique to extract prior knowledge of drugs and diseases. Moreover, we make use of the known drug-disease associations, and obtain known association-based features from an association bipartite graph through graph embedding, and take into account biological domain features, i.e., drug chemical structures and disease semantic similarity. Finally, we design a multimodal neural network to combine three types of features from the knowledge graph, the known associations, and the biological domain, and build the prediction model for predicting drug-disease associations. Massive experiments show that our method outperforms other state-of-the-art methods in terms of most metrics, and the ablation analysis regarding the three types of features reveals that prior knowledge from knowledge graphs can not only lift the predictive power of *in silico* drug repositioning, but also enhance the model's robustness to different scenarios. The contributions of this paper can be summarized as follows:

(1) We construct the knowledge graph for the drug repositioning and extract prior knowledge from the knowledge graph, and detailedly elaborate how prior knowledge improves the drug-disease association prediction.
(2) We propose a multimodal framework that combines the prior knowledge with two types of classic features to improve *in silico* drug repositioning, and the multimodal model demonstrates superior performance and good generalization ability.
(3) We discuss the usefulness of biological features, known drug-disease association-based features, and knowledge graph-based features in our task, and illustrate their applicable scope and scenarios. The findings can provide a reference for other interaction/association prediction tasks.

2 Materials and Methods

2.1 Datasets

For the comprehensive study, we adopt two datasets to evaluate the prediction models. The summary of the datasets is displayed in Table 1.

Zhang's dataset was compiled in our previous work for drug-disease association prediction [10], where the drug-disease associations are classified into two types (therapeutic and others) based on the annotations in CTD [30]. Since predicting therapeutic effects helps discover novel treatment and draws much attention of biomedical researchers, we pick out the therapeutic associations from Zhang's dataset and construct another dataset, called Zhang's therapeutic dataset. Detailedly, there are 17414 drug-disease associations and 5288 therapeutic associations between 267 drugs and 570 diseases separately in Zhang's dataset and Zhang's therapeutic dataset. We construct a knowledge graph involving the drug entities and the disease entities in Zhang's dataset and Zhang's therapeutic dataset using multi-related data from [31]. Knowledge graph stores entities and their relations in triplet form (head entity, relation, tail entity), e.g., (hsa_circ_0000064, circRNA-disease, lung

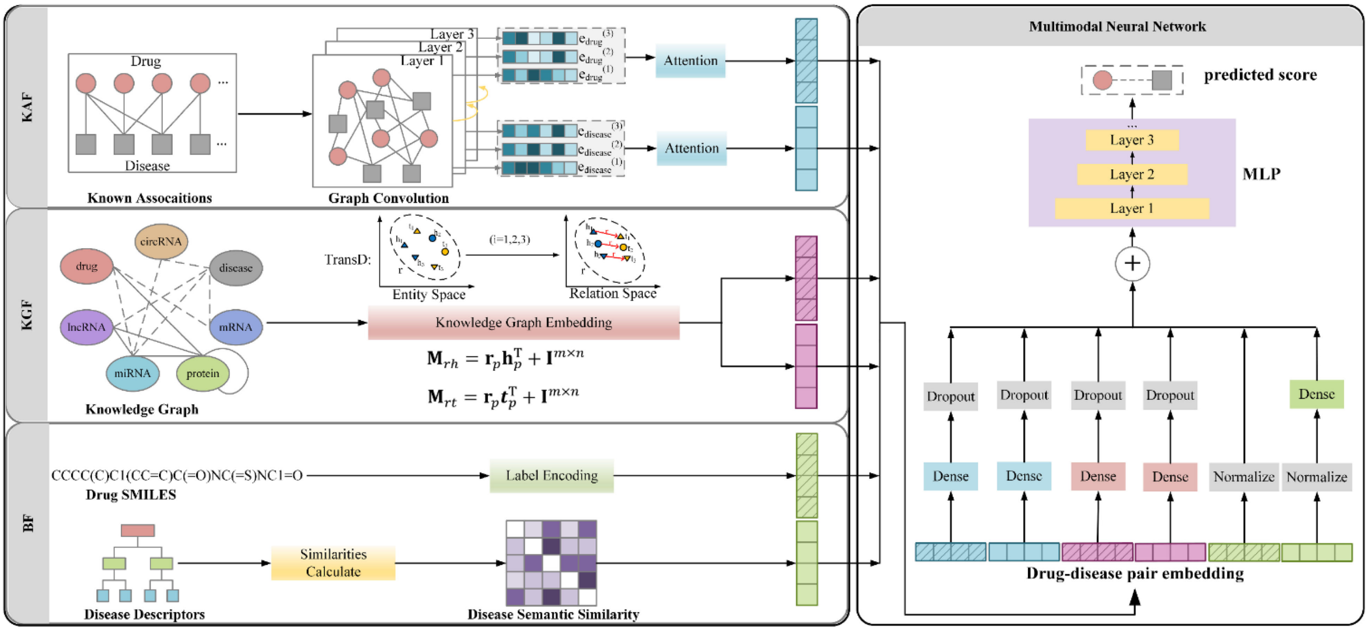


Fig. 1. The Pipeline of GraphPK. The dotted lines in the knowledge graph represent associations and the solid lines represent interactions.

neoplasms) where a head entity (circRNA) is connected to a tail entity (disease) through a predicate relation (circRNA-disease). Given a set of entities $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$ and a set of relation types $\mathcal{R} = \{r_1, r_2, \ldots, r_t\}$. We construct a knowledge graph $\mathcal{G}(\mathcal{E}, \mathcal{R})$ using bio-entity relations that are related to drug-disease associations which are extracted from [31]. A triplet (e_i, r_k, e_j) in the knowledge graph denotes that there is a relation belonging to r_k type between the entity e_i and the entity e_j. *More details about the knowledge graph we constructed are provided in the Supplemental Material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2021.3103595.*

We consider another knowledge graph named DRKG launched by Amazon AI (https://github.com/gnn4dr/DRKG). DRKG includes information from seven existing databases, including DrugBank [32], Hetionet [33], GNBR [34], String [35], IntAct [36], DGIdb [37] and data collected from recent publications particularly related to COVID-19. For DRKG, we remove drugs and diseases whose SMILES and MeSH are unavailable, and we obtain 44856 drug-disease associations between the remaining 3302 drugs and 2917 diseases, and name it Amazon's dataset. Correspondingly, DRKG is used as the knowledge graph in this study.

2.2 Architecture of GraphPK

As shown in Fig. 1, the proposed multimodal framework GraphPK requires three types of features: biological features (BF), known association-based features (KAF) and knowledge graph-based features (KGF), and utilizes a multimodal neural network to combine different features and build the prediction model.

2.2.1 Biological Features (BF)

The structures of drugs are considered as one of the most important features. Here, we extract SMILES [38] of drugs from PubChem [39], and then calculate label/integer encoding representations [40] of drugs from their SMILES. The label/integer encoding scheme represents each label in SMILES with a corresponding integer (e.g., "*CC*":1, "*OO*":2, "*NN*":3, " = ":4 etc.). For example, the SMILES [$CCNN = CC = OO$] can be transformed into [1, 3, 4, 1, 4, 2]. Since the length of drug canonical SMILES is varied, we set an appropriate length for all the drugs. As a result, a drug is represented by a vector of 128 dimensions.

For the disease features, we use disease-disease semantic similarities as disease features. Details to compute disease-disease semantic similarities can be found in the *Supplemental Material, available online.*

2.2.2 Known Association-Based Features (KAF)

The known associations usually contain important information for the drug-disease association prediction, and show good performances in the related studies [5], [10], [15]. The known associations naturally form a drug-disease association bipartite graph, which uses the drugs, diseases as nodes and uses the known associations as links. Thus, the graph embedding techniques can be applied to obtain the node representations for the downstream task.

Here, we adopt a variant of graph convolutional network [41] named LightGCN [42] to learn informative representations of drugs and diseases from the bipartite graph (more graph embedding methods are studied in the Section "Results and Discussion"). Let u denotes a drug, v denotes a disease, $\mathcal{N}_u$ denotes the set of nodes linked to u , and $\mathcal{N}_v$ denotes the set of nodes linked to v. The embedding update at k-th GCN layer is defined as:

$$\mathbf{e}_u^{(k+1)} = \sum_{v \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_v|}} \mathbf{e}_v^{(k)} \tag{1}$$

$$\mathbf{e}_v^{(k+1)} = \sum_{u \in \mathcal{N}_v} \frac{1}{\sqrt{|\mathcal{N}_v|}\sqrt{|\mathcal{N}_u|}} \mathbf{e}_u^{(k)} \tag{2}$$

To capture different information from different convolution layers, LightGCN allows an aggregation operation to form the final representation of a drug or a disease:

$$\mathbf{e}_u = \sum_{k=0}^{K} \alpha_k \mathbf{e}_u{}^{(k)} \ ; \quad \mathbf{e}_v = \sum_{k=0}^{K} \alpha_k \mathbf{e}_v{}^{(k)} \tag{3}$$

where $\alpha_k \geq 0$ is a trainable parameter as the adaptive contribution of the k-th layer embedding in constituting the final embedding. We used the way in [42] for initialization. The model prediction is defined as the inner product of drug and disease final representations:

$$\hat{y}_{uv} = \mathbf{e}_u \ \mathbf{W} \mathbf{e}_v{}^{\mathrm{T}} \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{\boldsymbol{k}\times\boldsymbol{k}}$ is a trainable matrix. Then, we adopt the focal loss [43] as the loss function. We use the Adam optimizer [44] to minimize the loss function and update the embeddings of drugs and diseases. Finally, $\mathbf{e}_u$ and $\mathbf{e}_v$ are used as known association-based features of drugs and diseases.

2.2.3 Knowledge Graph-Based Features(KGF)

Knowledge graph embedding methods learn the low-rank representations of entities and relations in the knowledge graph. Here, we use the knowledge graph embedding method TransD [24] to learn prior knowledge, i.e., embeddings of drugs and diseases (more knowledge graph embedding methods are discussed in the Section "Results and Discussion").

Given a triplet (*h, r, t*), *h, r, t* denote the head entity, the type of relation, and tail entity respectively. Their representation vectors are denoted by $\mathbf{h}, \mathbf{r}, \mathbf{t}$, which are trainable and initialized before training. TransD projects entities from entity space to relation space by using two mapping matrices $\mathbf{M}_{rh}, \ \mathbf{M}_{rt} \in \mathbb{R}^{m\times n}$, which can be decomposed as follows:

$$\mathbf{M}_{rh} = \mathbf{r}_p \ \mathbf{h}_p^{\mathrm{T}} + \mathbf{I}^{m\times n} \tag{5}$$

$$\mathbf{M}_{rt} = \mathbf{r}_p \ \boldsymbol{t}_p^{\mathrm{T}} + \mathbf{I}^{m\times n} \tag{6}$$

where $\mathbf{h}_p, \mathbf{r}_p, \mathbf{t}_p$ denote project vectors of *h, r, t* which are also trainable and initialized before training. After projecting, the projected vectors are defined as follows:

$$\mathbf{h}_\perp = \mathbf{M}_{rh} \ \mathbf{h}, \quad \mathbf{t}_\perp = \mathbf{M}_{rt} \ \mathbf{t} \tag{7}$$

Then these projected vectors are used to score the set of positive and negative training triplets using a scoring function defined as:

$$f_r \ (\mathbf{h}, \ \mathbf{t}) = \ - ||\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp||_2^2 \tag{8}$$

After scoring, the output scores are then passed to the margin-based ranking loss, which is defined as follows:

$$L \ = \sum_{\xi\in\Delta} \sum_{\xi'\in\Delta'} [\gamma + f_r(\xi') - f_r(\xi)]_+ \tag{9}$$

where $[x]_+ \overset{\Delta}{=} \max(0, x)$, and γ is the margin separating positive triplets and negative triplets. Δ and Δ' denotes the positive and negative triplets in the training set respectively.

Then, the margin-based ranking loss is used by optimizers Adam optimizer to update the initialized embeddings (i.e., **h, r, t, h**$_p$, **r**$_p$, **t**$_p$). Finally, **h, r, t** are used as knowledge graph-based features of *h, r, t*.

2.2.4 Multimodal Neural Network

Given a drug-disease pair $(u, \ v)$, we use three types of features for the drug and disease as the input, and construct a multimodal neural network to predict the score of the pair.

As shown in Fig. 1, the multimodal neural network is designed for effectively integrating three types of features. BF of the drug u and the disease v is firstly normalized using min-max normalization. Then, considering the generally accepted perspective that structure determines function, we believe that the biological features for drugs extracted from drug SMILES could have a heavy impact on drug-disease association prediction. Hence, we scale other features (KAF and KGF) into drug structure encoding (BF of drugs) space by a dense layer [45] with 128 output units. Next, three types of representation vectors are concatenated and fed into a four-layer MLP with 128 and 64 hidden units to yield the score for $(u, \ v)$. We use rectified linear unit (ReLU) [46] as the activation function at each layer. After each layer, we add dropout layers [47] to avoid over-fitting and enhance generalization ability.

Here, we use the binary cross-entropy as the loss function. We use a batch size of 6000 and use Adam optimizer with a learning rate of 0.001 as the optimization algorithm to train the networks. The number of training epochs is 300.

3 Results and Discussion

3.1 Performance Evaluation

Note that we have only reliable positive pairs (i.e., observed/known drug-disease associations) in all drug-disease pairs in a dataset, the others are unverified/unknown associations. Our goal is to computationally identify real positive pairs from the unverified associations. To evaluate predictive models, we use an evaluation procedure to simulate the actual situation. In the model training stage, we randomly choose 80 percent of known drug-disease associations as positive training samples, while the remaining 20 percent associations (supposed to be unknown) and all unknown associations are merged to serve as negative training samples. In the model testing stage, we take the supposed unknown associations as positive testing samples and a matching number of randomly sampled unknown associations as negative testing samples. The better prediction models are, the more correctly the supposed unknown associations are predicted. The procedure is repeated 50 times and the average of performances is considered. We calculate several evaluation metrics: the area under the precision-recall curve (AUPR), the area under the receiver-operating characteristic curve (AUC), recall (RE, also known as sensitivity), specificity (SP), accuracy (ACC), precision (PRE), and F1-measure (F1).

GraphPK has several hyperparameters: the dimensionality of embeddings k, the initial learning rate of optimizer lr, the training epochs of our knowledge graph embedding method α, the training epochs of LightGCN β, the training epochs of GraphPK γ, and the dropout rate δ, We considered different combinations of these parameters from the ranges $\alpha\epsilon\{1000, \ 2000, \ 3000\}$, $\beta\epsilon\{6000, \ 8000, \ 10000\}$, $\gamma\epsilon\{100, \ 200, \ 300\}$, $\delta\epsilon\{0.1, \ 0.2, \ 0.3, \ 0.4\}$. By adjusting the parameters

TABLE 2
Performances of GraphPK and Other Methods On Benchmark Datasets

Dataset	Methods	AUPR	AUC	F1	ACC	RE	SP	PRE
Zhang's dataset	KGML-RESCAL	0.8560	0.8594	0.7899	0.7739	0.8497	0.6982	0.7385
	KGML-TransD	0.7489	0.7679	0.7284	0.6790	**0.8606**	0.4975	0.6319
	GML-LightGCN	0.8645	0.8668	0.7954	0.7786	0.8601	0.6971	0.7401
	GML-GCN	0.8537	0.8585	0.7876	0.7696	0.8543	0.6847	0.7311
	SCMFDD	0.8706	0.8716	0.7990	0.7843	0.8574	0.7112	0.7484
	deepDR	0.8052	0.8206	0.7648	0.7362	0.8574	0.6149	0.6908
	GraphPK	**0.8710**	**0.8731**	**0.8005**	**0.7866**	0.8559	**0.7171**	**0.7523**
Zhang's therapeutic dataset	KGML-RESCAL	0.8563	0.8592	0.7889	0.7715	0.8536	0.6893	0.7338
	KGML-TransD	0.7866	0.8015	0.7499	0.7153	0.8529	0.5777	0.6705
	GML-LightGCN	0.6805	0.7438	0.7591	0.7224	0.8742	0.5705	0.6713
	GML-GCN	0.6458	0.7478	0.7730	0.7379	**0.8918**	0.5837	0.6826
	SCMFDD	0.8900	0.8886	0.8147	0.8040	0.8608	0.7472	0.7744
	deepDR	0.8492	0.8490	0.7856	0.7702	0.8410	0.6994	0.7381
	GraphPK	**0.9034**	**0.9024**	**0.8274**	**0.8196**	0.8641	**0.7751**	**0.7946**
Amazon's dataset	KGML-RESCAL	0.8709	0.8712	0.8011	0.7875	0.8558	0.7191	0.7532
	KGML-TransD	0.5812	0.6071	0.6771	0.5441	**0.9538**	0.1350	0.5260
	GML-LightGCN	0.7350	0.8374	0.8417	0.8317	0.8954	0.7679	0.7943
	GML-GCN	0.7082	0.8113	0.8397	0.8151	0.9322	0.6980	0.7699
	GraphPK	**0.9463**	**0.9450**	**0.8775**	**0.8751**	0.8944	**0.8558**	**0.8613**

empirically, we set the parameter $k = 128$, $lr = 0.001$, $\alpha = 3000$, $\beta = 10000$, $\gamma = 300$, and $\delta = 0.4$ in the following experiments.

3.2 Comparison With Other State-of-the-Art Methods

For comparison, we consider two state-of-the-art methods: SCMFDD [10] and deepDR [16], which were proposed for drug-disease association prediction or drug repositioning, and we also consider knowledge graph-based missing link prediction models (KGML) and graph-based missing link prediction models (GML) as baselines.

- *SCMFDD* [10] projects drug-disease associations into two low-rank spaces uncovering latent features for drugs and diseases, and then introduces similarity constraints to smooth the features.
- *DeepDR* [16] developed a multi-modal deep autoencoder for fusing multiple features and use a collective variational autoencoder for the drug-disease association prediction.
- *GML* builds the prediction models by using graph embedding methods. Here, we considered the graph embedding methods GCN [41] and LightGCN [42], and the constructed prediction models are named *GML-GCN, GML-LightGCN*.
- *KGML* builds the prediction models by directly using knowledge graph embeddings for drug-disease association prediction. Here, we use RESCAL [48] and TransD [24] to construct prediction models *KGML-RESCAL, KGML-TransD*.

First, we evaluate all prediction models on Zhang's datasets, and the results are shown in Table 2. GraphPK outperforms all compared methods in terms of most evaluation metrics, achieving AUC of 0.8731 and AUPR of 0.8710. Since the therapeutic effects are one kind of drug-disease associations that attract wide attention, we further conduct experiments on Zhang's therapeutic dataset, and GraphPK also performs better than compared methods. It is worth noting that GML-GCN and GML-LightGCN have significantly decreased performances (>10% lower AUC and >20% lower AUPR) on Zhang's therapeutic dataset than those on Zhang's dataset. That is because they heavily rely on graph convolutional operation on drug-disease association bipartite graph but Zhang's therapeutic dataset has fewer associations (i.e., lower density) for the bipartite graph. Other methods produce relatively stable performances on the two datasets, possibly because they absorb multi-source heterogeneous information to alleviate the over-reliance on single-source information. Specifically, GraphPK makes full use of all information by the multimodal neural network and thus achieves the best performances on both datasets.

Moreover, we evaluate prediction models on Amazon's dataset, which contains a larger knowledge graph. Note that SCMFDD and deepDR are not included here for their scalability issues. As shown in Table 2, GraphPK has more obvious superiority over the compared methods on the dataset, probably because the knowledge graph in Amazon's dataset contains more bio-entities and relations, and provides more abundant information to enhance the performance of GraphPK.

In general, our method performs best among these methods on three benchmark datasets, due to the multimodal framework with the prior knowledge from knowledge graphs and other two types of features.

3.3 Influence of Feature Extraction Methods

Among three types of features in GraphPK, KAF and KGF are extracted by the graph embedding method LightGCN and knowledge graph embedding method TransD respectively. However, many the graph embedding methods and knowledge graph embedding methods can be devised for the association bipartite graph and the knowledge graph to extract KAF and KGF. In this section, we build several variants of GraphPK, which utilize different feature exaction methods, to investigate how feature exaction methods influence the performance of GraphPK.

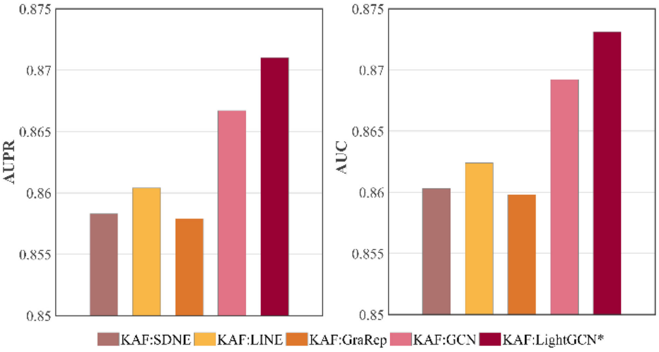


Fig. 2. Performances of GraphPK and variants of GraphPK with diverse graph embedding methods. * denotes the proposed GraphPK.

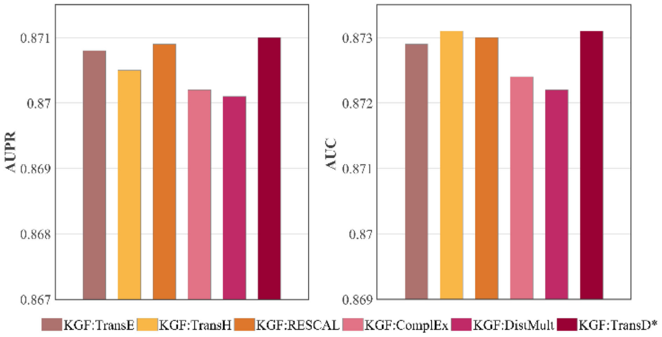


Fig. 3. Performances of GraphPK and variants of GraphPK with diverse knowledge graph embedding methods. * denotes the proposed GraphPK.

For KAF, we replace LightGCN with four graph embedding methods, i.e., SDNE [49], LINE [50], GraRep [51], GCN [41], and results on Zhang's dataset are shown in Fig. 2. According to the performances of variants, graph embedding methods perform differently, and the graph convolutional network methods, such as GCN and LightGCN, lead to better results than others; LightGCN is superior to GCN, because it is designed for bipartite graphs and stack multiple layers to alleviate the over-smoothing problem.

For KGF, we replace TransD with five knowledge graph embedding methods, i.e., TransE [23], TransH [52], RESCAL [48], DistMult [53], and ComplEx [54]. As displayed in Fig. 3, our method GraphPK achieves the best results, while other knowledge graph embedding methods also lead to the high-accuracy and similar results, indicating that leveraging prior knowledge from the knowledge graphs is useful for *in silico* drug repositioning and lead to robust performances regardless of knowledge graph embedding methods.

3.4 Ablation Analysis

GraphPK uses three types of features: KGF, KAF and BF to build the prediction model. Here, we consider all feature combinations to construct several variants of GraphPK, and compare GraphPK with variants on Zhang's dataset to explore the usefulness of these features. As shown in Table 3, GraphPK produces better results than variants with fewer features, especially those variants with an individual type of features, and more features lead to better results, indicating that combining diverse features can efficiently improve the drug-disease association prediction.

Noteworthily, the performances of variants (BF+KAF, KGF+KAF) are close to that of GraphPK. Although KAF seems to play a dominating role in GraphPK, its importance rapidly diminishing when known associations have low density. To test this viewpoint, we remove some known associations from Zhang's dataset at different ratios (10, 20, ..., 90 percent) to obtain datasets with lower association density, and build the GraphPK model and variants with other feature combinations, and then analyze the contributions of different types of features. As illustrated in Fig. 4, performances of GraphPK and variants decreased when reducing drug-disease association density. Especially, the performance of the model using KAF decreases greatly, and KAF performs worst among three types of features when removing 90 percent known associations. In contrast, the variants that combine KAF with KGF or BF can produce better results than the variant that uses only KAF, and the results show the use of KGF and BF can alleviate the reliance on the known drug-disease associations and increase the generalization ability of models.

The ablation analysis reveals: (1) the feature KAF can provide the most important information when known associations have high density; (2) the features BF and KGF provide information from other aspects, and they are very important when known associations are scarce. In the absence of known associations, GraphPK with BF and KGF can even work to predict drug-disease associations, which is discussed in the following subsection; (3) utilizing three types of features enhance the model's generalization ability and robustness to all scenarios.

3.5 The Capability of Prioritizing Drugs for New Diseases

The most important goal of drug-disease association prediction is to find treatments for newly emerged diseases timely, e.g., COVID-19. In this case, there is usually no known association between the new diseases and approved drugs.

TABLE 3
Performances of GraphPK With Different Feature Combinations on Zhang's Dataset

Methods	AUPR	AUC	F1	ACC	RE	SP	PRE
KGF	0.8491	0.8504	0.7804	0.7612	0.8484	0.6741	0.7227
KAF	0.8685	0.8701	0.7971	0.7813	0.8589	0.7038	0.7443
BF	0.8465	0.8478	0.7785	0.7593	0.8461	0.6726	0.7214
KGF+BF	0.8515	0.8537	0.7839	0.7648	0.8534	0.6762	0.7254
KGF+KAF	0.8685	0.8700	0.7966	0.7823	0.8524	0.7122	0.7480
BF+KAF	0.8708	0.8729	0.7999	0.7847	**0.8607**	0.7087	0.7475
ALL	**0.8710**	**0.8731**	**0.8005**	**0.7866**	0.8559	**0.7171**	**0.7523**

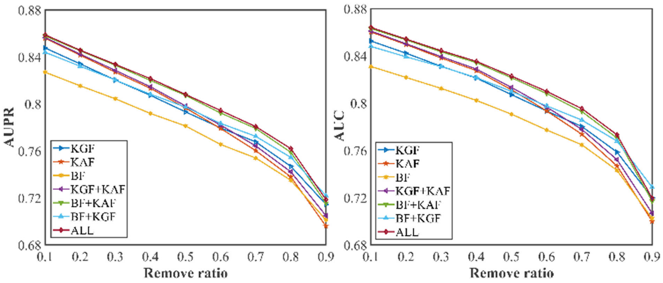


Fig. 4. Performances of GraphPK and different feature combinations with decreased drug-disease association density.

Here, we conduct experiments to evaluate the predictive ability of our method GraphPK for new diseases. We randomly split all diseases into 5 subsets of equal size. In each fold, we choose one subset of diseases and use them to simulate newly emerged diseases for testing, and other diseases are used as the training diseases. For comparison, we also construct two variants that only use KGF or BF. Under this setting, KAF is unapplicable for the testing diseases, because the associations between new diseases and approved drugs are absent. All models are trained on the data with training diseases and approved drugs, and are used to predict the associations between testing diseases and approved drugs. As displayed in Fig. 5, GraphPK produces satisfying performances under the experimental settings, achieving AUC of 0.791 and AUPR of 0.388, much better than variants based on KGF or BF, and incorporating KGF into the BF-based model can lead to great improvement in terms of AUC, AUPR and Recall@k. Moreover, we pay attention to the top predictions produced by GraphPK, and find out that almost 80 percent of real associations can be identified when only checking up on the top 45 percent of predictions.

The study demonstrates that two features KGF and BF can guarantee the capability of GraphPK in prioritizing drugs for new diseases, and KGF (i.e., prior knowledge from knowledge graphs) can provide abundant information to improve the performance in this scenario.

3.6 Case Study

In this section, we use case studies to demonstrate the capability of GraphPK for predicting novel drug-disease associations.

TABLE 4
Top 10 Novel Drug-Disease Therapeutic Associations Predicted on Zhang's Dataset

Rank	Drug Name	Disease Name	Evidence
1	Verapamil	Cocaine-related disorders	PMID: 15204172
2	Procainamide	Heart block	PMID:2462213
3	Lidocaine	Atrial flutter	N.A.
4	Etoposide	Leukemia	PMID: 1984829
5	Melatonin	Breast neoplasms	PMID: 29458781
6	Levodopa	Seizures	PMID: 6428968
7	Digoxin	Cardiomegaly	PMID: 12864722
8	Valproic acid	Catalepsy	N.A.
9	Naproxen	Hyperalgesia	PMID: 12411814
10	Losartan	Glomerulonephritis	PMID:17365932

Firstly, we build the GraphPK based on Zhang's therapeutic dataset and use it to predict novel drug-disease therapeutic associations. The top 10 predictions are shown in Table 4, and we have evidence to confirm eight therapeutic associations. For example, the approach to the treatment of Cocaine-induced myocardial infarction focused on medical combination treatment containing Verapamil [55]. In the study [56], it was proved that Losartan is effective for patients with Glomerulonephritis. In breast cancer, Melatonin is capable to disrupt estrogen-dependent cell signaling, resulting in a reduction of estrogen-stimulated cells [57].

Secondly, we conduct the case study using Amazon's dataset and list the top 10 drug-disease associations predicted by GraphPK in Table 5. All predicted associations can be confirmed by public literature. For example, the first liposomal encapsulated anticancer drug which received clinical approval against transplantable Leukemias was Doxorubicin HCl [58]. The study [59] indicated that Cisplatin-based chemotherapy is one of the most common treatments for Leukemia.

The case study on Zhang's dataset is given in *Supplemental Table S3, available online.*

4 Conclusion

Drug repositioning/repurposing is a very important approach towards identifying novel treatments for diseases in drug discovery. In previous studies, biological features and known drug-disease associations have been widely used, and have demonstrated great potential in drug repositioning. Recently, large-scale datasets related to drugs and diseases are increasingly available for drug repositioning, and information from these datasets may be useful for

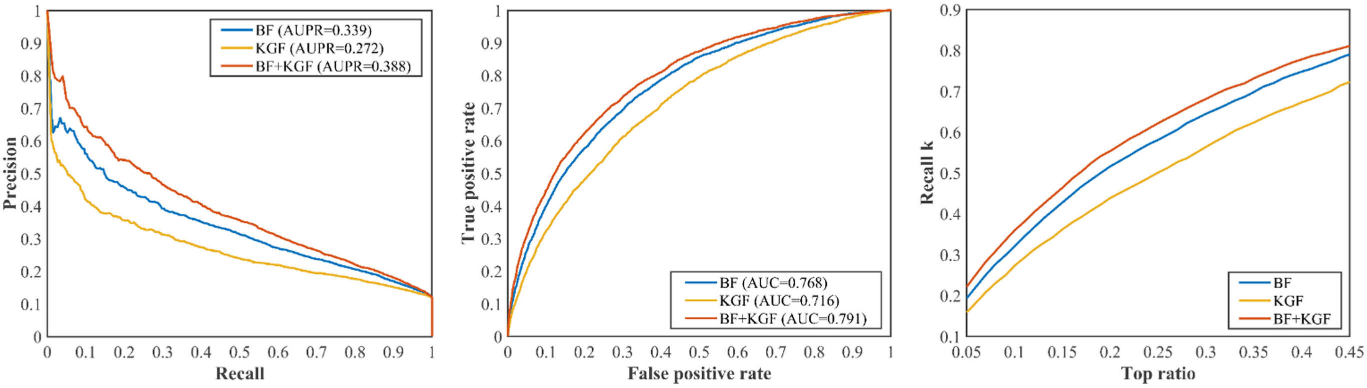


Fig. 5. Performances of GraphPK in prioritizing drugs for new diseases.

TABLE 5
Top 10 Novel Drug-Disease Associations Predicted on Amazon's Dataset

Rank	Drug Name	Disease Name	Evidence
1	Etoposide	Urinary Bladder Neoplasms	PMID: 10540713
2	Methotrexate	Multiple Myeloma	PMID: 23800093
3	Quinaprilat	Hypertension	PMID: 19761414
4	Doxorubicin	Leukemia	PMID:25579518
5	Cisplatin	Leukemia	PMID: 28677814
6	Mitomycin	Leukemia, Myeloid, Acute	PMID: 16753889
7	Hydrocortisone	Polymyositis	PMID: 72811
8	Ajmaline	Tachycardia, Supraventricular	PMID: 6488500
9	Cyclosporine	Lymphoma	PMID: 31393197
10	Prednisone	Choroiditis	PMID: 28418567

improving drug repositioning. Thus, we make use of relevant bio-entities (drugs, diseases, etc.) and associations/interactions among them to construct knowledge graphs and learn prior knowledge, and thus we develop a multimodal framework GraphPK that utilizes prior knowledge from knowledge graphs, biological features, and known drug-disease associations to improve *in silico* drug repositioning. Our studies show that the knowledge graph provides information distinct from biological features and known associations, and GraphPK takes advantage of diverse features and has better usability for different scenarios in drug repositioning.

In our future work, we have several directions to improve the drug-disease association prediction. Incorporating more biological entities and relations can enrich the knowledge graph, and thus provide more information for better performance. However, there may exist redundant biological entities and relations in knowledge graphs, which hinder learning accurate prior knowledge. Therefore, we will study how to distill or refine the knowledge graph by removing redundant information or noise. In GraphPK, the feature extraction is free from the multimodal network because of the computational efficiency, and a more efficient end-to-end learning frame is our future consideration. Drug combination can be regarded as a new way of drug repositioning, in the future, we can consider a drug combination as a node in the drug-disease association bipartite graph or the knowledge graph, and this may enhance the performance of finding novel drug therapy for diseases.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 62072206, 61772381, and 61572368 and in part by the Huazhong Agricultural University Scientific and Technological Self-innovation Foundation; Fundamental Research Funds for the Central Universities (2662021JC008). The funders have no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

REFERENCES

[1] J. K. Yella *et al.*, "Changing trends in computational drug repositioning," *Pharmaceuticals*, vol. 11, no. 2, 2018, Art. no. 57.

[2] T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–683, 2004.

[3] R. Santos *et al.*, "A comprehensive map of molecular drug targets," *Nat. Rev. Drug Discov.*, vol. 16, no. 1, pp. 19–34, 2017.

[4] H. Luo *et al.*, "Biomedical data and computational models for drug repositioning: A comprehensive review," *Brief. Bioinf.*, vol. 22, no. 2, pp. 1604–1619, 2020.

[5] Z. Yu *et al.*, "Predicting drug–disease associations through layer attention graph convolutional network," *Brief. Bioinf.*, vol. 22, 2020, Art. no. bbaa243.

[6] A. Gottlieb *et al.*, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, no. 1, 2011, Art. no. 496.

[7] M. Oh, J. Ahn, and Y. Yoon, "A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions," *PLoS One*, vol. 9, no. 10, 2014, Art. no. e111668.

[8] K. Yang *et al.*, "Predicting drug-disease associations with heterogeneous network embedding," *Chaos Interdiscipl. J. Nonlinear Sci.*, vol. 29, no. 12, 2019, Art. no. 123109.

[9] P. Xuan *et al.*, "Drug repositioning through integration of prior knowledge and projections of drugs and diseases," *Bioinformatics*, vol. 35, no. 20, pp. 4108–4119, 2019.

[10] W. Zhang *et al.*, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–12, 2018.

[11] P. Zhang, F. Wang, and J. Hu, "Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity," in *Proc. Annu. Symp. Proc.*, vol. 2014, pp. 1258–1267, 2014.

[12] H. Luo *et al.*, "Computational drug repositioning with random walk on a heterogeneous network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 6, pp. 1890–1900, Nov./Dec. 2018.

[13] W. Wang *et al.*, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.

[14] M. Yang *et al.*, "Heterogeneous graph inference with matrix completion for computational drug repositioning," *Bioinformatics*, vol. 36, no. 22-23, pp. 5456–5464, 2020.

[15] W. Zhang *et al.*, "Predicting drug-disease associations based on the known association bipartite network," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2017, pp. 503–509.

[16] X. Zeng *et al.*, "DeepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.

[17] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. Roy. Soc. Interface*, vol. 15, no. 141, 2018, Art. no. 20170387.

[18] X. Chen *et al.*, "Drug–target interaction prediction: Databases, web servers and computational models," *Brief. Bioinf.*, vol. 17, no. 4, pp. 696–712, 2016.

[19] X. Chen *et al.*, "NLLSS: Predicting synergistic drug combinations based on semi-supervised learning," *PLoS Comput. Biol*, vol. 12, no. 7, 2016, Art. no. e1004975.

[20] X. Chen *et al.*, "MicroRNA-small molecule association identification: From experimental results to computational models," *Brief. Bioinf.*, vol. 21, no. 1, pp. 47–61, 2020.

[21] J. Qu *et al.*, "Inferring potential small molecule–miRNA association based on triple layer heterogeneous network," *J. Cheminf.*, vol. 10, 2018, Art. no. 30.

[22] Q. Wang *et al.*, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.

[23] A. Bordes *et al.*, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[24] G. Ji *et al.*, "Knowledge graph embedding via dynamic mapping matrix," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Nat. Lang. Process. Asian Federation Nat. Lang. Process.*, 2015, pp. 687–696.

[25] S. K. Mohamed, V. Nováček, and A. Nounu, "Discovering protein drug targets using knowledge graph embeddings," *Bioinformatics*, vol. 36, no. 2, pp. 603–610, 2020.

[26] X. Lin *et al.*, "KGNN: Knowledge graph neural network for drug-drug interaction prediction," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 2739–2745.

[27] Y. Zhu *et al.*, "Knowledge-driven drug repurposing using a comprehensive drug knowledge graph," *Health Inform. J.*, vol. 26, no. 4, pp. 2737–2750, 2020.

[28] Y. Ge *et al.*, "An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19," *Signal Trans. Target. Ther.*, vol. 6, no. 1, 2021, Art. no. 165.

[29] X. Zeng *et al.*, "Repurpose open data to discover therapeutics for COVID-19 using deep learning," *J. Proteome Res.*, vol. 19, no. 11, pp. 4624–4636, 2020.

[30] A. P. Davis *et al.*, "The comparative toxicogenomics database: Update 2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D948–D954, 2019.

[31] Z.-H. Guo *et al.*, "A learning based framework for diverse biomolecule relationship prediction in molecular association network," *Commun. Biol.*, vol. 3, no. 1, pp. 1–9, 2020.

[32] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the drugbank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[33] D. S. Himmelstein *et al.*, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *Elife*, vol. 6, 2017, Art. no. e26726.

[34] B. Percha and R. B. Altman, "A global network of biomedical relationships derived from text," *Bioinformatics*, vol. 34, no. 15, pp. 2614–2624, 2018.

[35] D. Szklarczyk *et al.*, "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2019.

[36] S. Orchard *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D358–D363, 2014.

[37] K. C. Cotto *et al.*, "DGIdb 3.0: A redesign and expansion of the drug–gene interaction database," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1068–D1073, 2018.

[38] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[39] S. Kim *et al.*, "PubChem 2019 update: Improved access to chemical data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1102–D1109, 2019.

[40] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: Deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[41] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.

[42] X. He *et al.*, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.

[43] T.-Y. Lin *et al.*, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[45] X. Dong *et al.*, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, pp. 601–610.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[47] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] M. Nickel, V. Tresp, and H. -P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 809–816.

[49] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1225–1234.

[50] J. Tang *et al.*, "LINE: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1067–1077.

[51] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 891–900.

[52] Z. Wang *et al.*, "Knowledge graph embedding by translating on hyperplanes," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1112–1119.

[53] B. Yang *et al.*, "Embedding entities and relations for learning and inference in knowledge bases," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[54] T. Trouillon *et al.*, "Complex embeddings for simple link prediction," in *Proc. 33nd Int. Conf. Mach. Learn.*, 2016, pp. 2071–2080.

[55] H. Meltser, D. Bhakta, and V. Kalaria, "Multivessel coronary thrombosis secondary to cocaine use successfully treated with multivessel primary angioplasty," *Int. J. Cardiovasc. Intervent.*, vol. 6, no. 1, pp. 39–42, 2004.

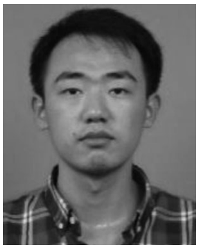
[56] S. Kahvecioglu *et al.*, "Comparison of higher dose of losartan treatment with losartan plus carvedilol and losartan plus ramipril in patients with glomerulonephritis and proteinuria," *Renal Fail*, vol. 29, no. 2, pp. 169–175, 2007.

[57] P. Kubatka *et al.*, "Melatonin and breast cancer: Evidences from preclinical and human studies," *Crit. Rev. Oncol. Hematol.*, vol. 122, pp. 133–143, Feb. 2018.

[58] S. Rivankar, "An overview of doxorubicin formulations in cancer therapy," *J. Cancer Res. Ther.*, vol. 10, no. 4, pp. 853–858, Oct.–Dec. 2014.

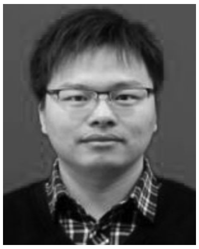
[59] X. J. Han *et al.*, "Involvement of mitochondrial dynamics in the antineoplastic activity of cisplatin in murine leukemia L1210 cells," *Oncol. Rep.*, vol. 38, no. 2, pp. 985–992, Aug. 2017.

Zhankun Xiong is currently working toward the graduation degree with the College of Informatics, Huazhong Agricultural University, China. His research interests include bioinformatics, complex network, and machine learning.



Feng Huang is currently working toward the PhD degree with the College of Informatics, Huazhong Agricultural University, China. His research interests include bioinformatics and machine learning.

Ziyan Wang is currently working toward the graduation degree with the College of Informatics, Huazhong Agricultural University, China. Her research interests include machine learning and data mining.



Shichao Liu is currently a lecturer with the Department of Data Science and Big Data Technology, Huazhong Agricultural University, China. His research interests include machine learning and data mining.

Wen Zhang received the bachelor's and master's degree in computational mathematics and the doctoral degree in computer science from Wuhan University in 2003, 2006, and 2009, respectively. He is currently a professor with the College of Informatics, Huazhong Agricultural University, China. His research interests include machine learning and bioinformatics.