IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 2, 2018, accepted November 28, 2018, date of publication December 12, 2018,
date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886311

GrEDeL: A Knowledge Graph Embedding Based Method for Drug Discovery From Biomedical Literatures

SHENGTIAN SANG[1], ZHIHAO YANG[1], XIAOXIA LIU[1], LEI WANG[2], HONGFEI LIN[1], JIAN WANG[1], AND MICHEL DUMONTIER[3]

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China
[2]Beijing Institute of Health Administration and Medical Information, Beijing 100191, China
[3]Institute of Data Science, Maastricht University, 6229 ER Maastricht, The Netherlands

Corresponding authors: Zhihao Yang (yangzh@dlut.edu.cn) and Lei Wang (wangleibihami@gmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0901902, in part by the Natural Science Foundation of China under Grant 61272373, Grant 61572102, and Grant 61572098, and in part by the Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China under Grant NCET-13-0084.

ABSTRACT Drug discovery is the process by which new candidate medications are discovered. Developing a new drug is a lengthy, complex, and expensive process. Here, in this paper, we propose a biomedical knowledge graph embedding-based recurrent neural network method called GrEDeL, which discovers potential drugs for diseases by mining published biomedical literature. GrEDeL first builds a biomedical knowledge graph by exploiting the relations extracted from biomedical abstracts. Then, the graph data are converted into a low dimensional space by leveraging the knowledge graph embedding methods. After that, a recurrent neural network model is trained by the known drug therapies which are represented by graph embeddings. Finally, it uses the learned model to discover candidate drugs for diseases of interest from biomedical literature. The experimental results show that our method could not only effectively discover new drugs by mining literature, but also could provide the corresponding mechanism of actions for the candidate drugs. It could be a supplementary method for the current traditional drug discovery methods.

INDEX TERMS Drug discovery, biomedical knowledge graph, recurrent neural network, deep learning.

I. INTRODUCTION

Drug discovery is defined as the process whereby a drug candidate or lead compound is identified and partially validated for the treatment of a specific disease [1]. It is a lengthy and expensive process, which is estimated to take 14 years and cost approximately $1.8 billion for developing one drug [2]. In contrast, literature based discovery (LBD) is a much less time-consuming and expensive approach to identify new drugs for indications [3]. It has been successfully applied in the field of biomedicine [4]. The LBD was pioneered by Don R. Swanson (1924-2012) who found dietary fish oils (A) might be used to treat Raynaud's disease (C) based on their shared connections to blood viscosity (B) in literature [5]. This hypothesis was clinically confirmed by DiGiacomo *et al.* two years later [6]. Swanson's method is called the ABC model which hypothesizes the combination of two separately published premises "A implies B" and "B implies C" indicates a relationship between A and C. Since then, a series of automatic ABC model based methods have been introduced to discover drugs from literature [7]–[9]. Co-occurrence methods are the basic ABC model based literature mining techniques which directly use co-occurrences in text as relationships between terms [9]. Directly using co-occurrences could capture all possible relations in text. However, the main issue of co-occurrence methods is that the extracted relationships lack logical explanations [10]. Furthermore, some extracted pairs of entities with high co-occurrence frequency could be completely uncorrelated actually [11]. In order to solve the problem, many sophisticated semantic models have been developed, which employ natural language processing methods to determine what constitutes a relationship [12], [13]. In addition, Hristovski *et al.* [14] introduced discovery patterns which serve as an effective filtering method that reduces the number of false positive discoveries and also supports explanation of discoveries. Although semantic models increase the precision of linking,

the major limitation of above semantic models is that more complex associations will go undetected due to semantic models are restricted to the ABC paradigm [3]. More recently, a series of literature mining methods have utilized knowledge graph to discover complex associations. Cameron *et al.* [15] introduced an automatic subgraph model which first clusters semantic paths in a semantic graph and then elucidates the latent associations between biomedical entities by the corresponding clusters. Malas *et al.* [16] leverage knowledge graph features such as the total number of intermediate concepts, the number of different semantic categories, and the predicates connecting a drug-disease pair to predict novel drug-disease associations. Bakalb and Talari [17] exploit simple paths connecting biomedical entities as features of logistic regression model to discover drugs. In our previous work, we introduced a biomedical knowledge graph based inference method - SemaTyP - which exploits the distribution of semantic types of entities to discover drug therapies [18]. The limitations of above knowledge graph based methods are: Cameron's method is mainly used to explain the associations between drug and disease, rather than discover new drugs. Malas' method can not find complex associations between drugs and diseases. Bakalb's method and SemaTyP can not capture the dependencies of entities in the drug-target-disease associations due to the logistic regression model does not consider the order of entities in the associations. In addition, the two methods can not provide the detailed drug mechanism of action. Besides the above methods, recently, significant amount of research attentions have been drawn to leverage various deep learning based approaches in the field of drug discovery [19]–[21]. However, these methods focus on identifying interactions between known drugs and targets from existing biomedical databases without considering the known knowledge contained in the huge amount of biomedical literature. In conclusion, although literature-based discovery is a powerful paradigm with potential to complement traditional drug discovery methods, there is still considerable room for improvement in mining literature for new drug therapies.

In this paper, we propose a biomedical knowledge **Gr**aph **E**mbedding based **D**eep **L**earning method - GrEDeL - to discover potential drugs from literature. Firstly, a biomedical knowledge graph was constructed with the relations extracted from PubMed abstracts. Compared with our previous work [18], the biomedical knowledge graph constructed in this work differentiates semantic types of entities. Secondly, we proposed to use the knowledge graph embedding method to convert the knowledge graph into low dimensional vector space. The embeddings of entities and relations could not only preserve the structures of the knowledge graph but also capture the semantic information of entities and relations. After that a Long Short-Term Memory Networks (LSTM) model was trained by known drug therapies from Therapeutic Target Database. Finally, the trained model was used to discover potential drugs from literature. The experimental results show that our method could not only discover drugs for diseases of interest, but also could provide corresponding potential mechanism of actions for the candidate drugs.

Our contributions are two-fold:

- We are the first to consider the process of literature-based discovery as a series analysis problem.
- We propose a knowledge graph based deep learning framework for LBD. To the best of our knowledge, this is the first method that employs deep learning method combined with knowledge graph for drug discovery. Additionally, we demonstrate the usefulness of graph embedding-based features for predicting potential drug-disease associations.

The rest of this paper is organized as follows. Section II introduces the related data and tools used in our work. In Section III, we present the details of the proposed method. Subsequently, we describe different evaluation metrics used in this paper and the experimental results in Section IV. Section V is the discussion part and Section VI presents our conclusion.

II. RELATED MATERIALS

A. DATABASE

1) PubMed DATABASE

PubMed comprises citations for biomedical literature from MEDLINE and life science journals. Currently, PubMed contains over 26 million biomedical abstracts, which represents an enormous corpus that could be used for drug discovery [3]. The knowledge graph used in this work was constructed with the relations extracted from the PubMed abstracts.

2) UMLS SEMANTIC NETWORK

The Unified Medical Language System (UMLS) semantic network consists of 133 semantic types that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and 54 semantic relations that exist between semantic types [22], [23].

3) THERAPEUTIC TARGET DATABASE

Therapeutic Target Database (TTD) is a database which provides information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs directed at each of these targets [24]. In this work, we constructed the training and test data sets by utilizing the drug-disease associations in TTD.

B. RELATED TOOLS AND TECHNIQUES

1) SemRep

SemRep is a program that extracts semantic predications from biomedical free text [25]. Predications consist of a subject argument, an object argument, and the relation that binds them. For example, from the sentence "Hydrocortisone increased slow wave sleep activity.", SemRep extracts a predication:

- Hydrocortisone|**phsu** increase|**AUGMENTS** Sleep, Slow-Wave|**phsu**

SemRep assigns a UMLS semantic type to the entity and relation (the black bold abbreviation on the right of '|'). For example, 'phsu' represents 'Pharmacologic Substance'. In this paper, the abbreviations are used to represent UMLS semantic types.

2) KNOWLEDGE GRAPH

Knowledge graphs (KGs) about entities, their properties, and the relationships between entities, have become an important asset for semantic search, analytics, and smart recommendations over Web contents and other kinds of big data. Notable knowledge graph systems include Freebase [26], DBpedia [27], YAGO [28] and many others. In the biomedical domain, KG such as the Gene Ontology and the Disease Ontology are prominent examples of the rich knowledge that are digitally available. In our previous work, we constructed a biomedical knowledge graph - SemKG - covering a wide range of terminology in multiple biomedical domains [18]. However, in SemKG, the same entity with different semantic types is considered to be the same one. In this work, we constructed a biomedical knowledge graph which differentiates semantic types of entity and relations.

3) KNOWLEDGE GRAPH EMBEDDING

Graph embedding methods convert the graph data into a low dimensional space in which the graph structural information and graph properties are maximally preserved [29]. Let h, r, t denote head, tail and relation of one edge in knowledge graph, knowledge graph embedding methods follow a common assumption $\mathbf{h}_r+\mathbf{r} \approx \mathbf{t}_r$, where $\mathbf{h}_r$ and $\mathbf{t}_r$ are either the original vectors of h and t, or the transformed vectors under a certain transformation with respect to r. The forerunner TransE [30] is adopted in this work as the knowledge graph embedding method for converting entities and relations of knowledge graph into vectors.

4) RECURRENT NEURAL NETWORK

Recurrent Neural Networks (RNNs) are, in general, good at capturing temporal dependencies in data and hence are effective in many time-series analysis applications [31]. However, RNNs have trouble learning time-dependencies more than a few time steps long [32] and suffer from severe overfitting problems [33]. To learn long-term dependencies, an alternative RNN architecture, Long Short Term Memory (LSTM), has been proposed to solve the long term dependency problem [34]. In addition, dropout technique which drops out units (hidden and visible) in a neural network was used to solve the overfitting problems of RNNs [33]. In this paper, we propose a LSTM-based RNN model which incorporates the drug-target-disease sequential data and the structures of knowledge graph to discover drugs from literature.

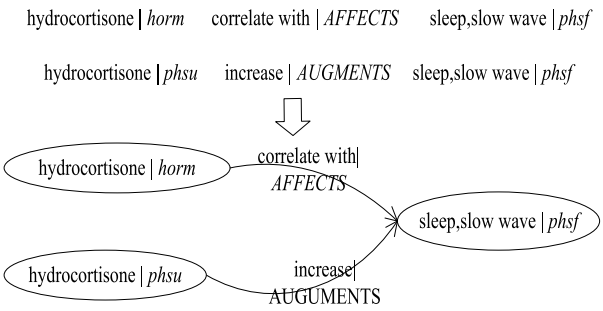


FIGURE 1. An illustration of constructing knowledge graph. There are two relations extracted by SemRep on the top of the figure. The figure shows the same entity (hydrocortisone) with different UMLS semantic types (horm and phsu) is considered as different nodes in the graph.

III. METHOD

Here, we consider the process of drug discovery as a drug-target-disease sequential data analysis problem. For example, the process by which chlorpromazine functions to produce a pharmacological effect on cardiac hypertrophy is as follows [35]:

chlorpromazine→**INHIBITS**→calmodulin
calmodulin→**STIMULATES**→calcineurin
calcineurin→**CAUSES**→ cardiac hypertrophy

The chlorpromazine acts on cardiac hypertrophy through a series of entities. Since RNNs are well-suited for analyzing sequential data, we proposed a LSTM-based RNN model to predict drug-target-disease associations.

In this section, we first present the biomedical knowledge graph constructed in this study. Then, we introduce a novel approach called GrEDeL which integrates knowledge graph embeddings and LSTM model to score potential associations between drugs and diseases. After that, the trained drug discovery model is implemented to discover potential drugs for diseases.

A. CONSTRUCTION OF BIOMEDICAL KNOWLEDGE GRAPH

For constructing a biomedical knowledge graph, we first employed SemRep to extract predications from PubMed abstracts, then the predications were used to build the knowledge graph. Figure 1 is an illustration of constructing the knowledge graph, it shows that the same entity with different semantic types is considered as different nodes in the knowledge graph. In this paper, $E = \{e_1, e_2, \ldots, e_N\}$ denotes entities (an entity is a UMLS Metathesaurus concept) of the knowledge graph, $R = \{r_1, r_2, \ldots, r_M\}$ denotes the relations between entities and $T = \{t_1, t_2, \ldots, t_K\}$ is the set of semantic types of entities and relations. The elements of T are all from the UMLS semantic network.

B. PREPARATION OF TRAINING DATA

Given a knowledge graph KG, a path π is defined as a sequence of predications $e_0r_0e_1r_1e_2r_2\ldots$. In this work, a gold standard case is represented as *drug-target-disease* triple, which means the *drug* can treat the *disease* by

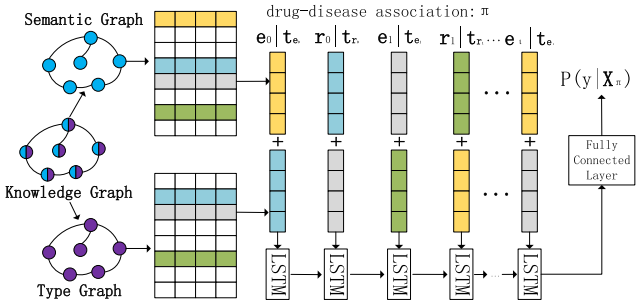


FIGURE 2. The framework of GrEDeL. The blue circle denotes entity and the purple node denotes the entity's corresponding semantic type.

acting on the *target*. We first employed the path exploring method described in our previous work to construct training data [18]. More concretely, we obtained $\pi^{\ell} = \rho(drug_i \rightarrow disease_i; target_i, \ell)$, which encodes all the paths of length ℓ reaching node disease$_i$ from source node drug$_i$ and crossing node target$_i$. Then $\pi^{\ell} = \{\pi_1^{\ell}, \pi_2^{\ell}, \pi_3^{\ell}, \pi_4^{\ell} \ldots\}$ is the set of all ℓ length paths. In addition, the paths in π^{ℓ} containing broad concept entities [10] are discarded. An entity e_i is considered as a broad concept entity when the type of e_i belongs to broad semantic types, because that the path containing broad concept entities can not express the drug mechanism of action for the particular disease. For example, the type of e_i in $\pi = e_0 r_0 \ldots e_i \ldots r_{\ell-1} e_{\ell}$ is "Animal (anim)" or "Manufactured Object (mnob)", then path is filtered out. After that, all paths in $\P = \{\pi^2, \pi^3 \ldots \pi^{\ell}\}$ are considered as positive training data. Similarly, we constructed negative data set by exploring false cases $drug_j'$-$target_j'$-$disease_j'$. Each false case denotes that the $drug_j'$ has no therapeutic effect on the $disease_j'$ or the corresponding drug target is not the $target_j'$.

C. GRAPH EMBEDDING-BASED LSTM DRUG DISCOVERY MODEL

Given a path $\pi_i^{\ell} = e_0 r_0 e_1 r_1 \ldots r_{\ell-1} e_{\ell}$ where e_0 indicates a drug and e_{ℓ} indicates a disease. The objective of our model is to predict the probability of the association between a potential drug and the disease of interest:

$$p(y|\pi_i^{\ell}) = D(g(\pi_i^{\ell}), \theta) \tag{1}$$

where $p(y = 1|\pi_i^{\ell})$ is the probability of the candidate drug for treating the disease and $D(.)$ represents any kind of discriminative model with trainable parameter θ. $g(.)$ is a function for graph embedding feature extraction. Figure 2 is an illustration of our model.

1) GRAPH EMBEDDING BASED FEATURE EXTRACTION

As input π_i^l is represented as a series of entity names, it is difficult to investigate the relation between entities. Thus, we applied TransE to the input π_i^l to learn more dense representations. In TransE, if a relationship (e_{head}, r, e_{tail}) holds, then the embedding of e_{tail} should be close to the embedding of e_{head} plus the embedding of r. To obtain both the graph structural information and the relations between semantic types, the biomedical knowledge graph was transformed into two graphs - Semantic Graph (*SG*) and Type Graph (*TG*). Semantic Graph contains entities and relations. Type Graph contains semantic types of entities and relations. In addition, x_i represents element of $\pi_i^{\ell} = e_0 r_0 e_1 r_1 \ldots r_{\ell-1} e_{\ell}$ (x_i could be either an entity e or relation r), then each element x_i of π_i^l is embedded as follows:

$$\mathbf{x}_i = g(x_i)_{SG} \bowtie g(x_i)_{TG} \tag{2}$$

where $g(.)$ is a function for graph embedding based feature extraction and symbol $\bowtie$ is concatenation of two vectors. To learn the vector embeddings of the entities and relations in the Semantic Graph (the process of $g(x_i)_{SG}$), we minimize the loss function L over the training set S and S':

$$L = \sum_{(e_1,r,e_2)\in S} \sum_{(e_1',r,e_2')\in S'} [\gamma + d(\mathbf{e_1} + \mathbf{r}, \mathbf{e_2}) - d(\mathbf{e_1'} + \mathbf{r}, \mathbf{e_2'})]_+ \tag{3}$$

where bold font indicates vector embedding of the corresponding element (For example, $\mathbf{e_1}$ is the embedding of e_1). In addition, $[x]_+$ denotes the positive part of x, d is L_1-norm and $\gamma > 0$ is a margin hyperparameter. The positive training set $S_{(e_1,r,e_2)}$ contains all the triplets (e_1,r,e_2) in Semantic Graph, and the negative training set S' is constructed by replacing e_1 or e_2 of triplets in S with a random entity (each triplet of S' does not appear in S):

$$S'_{(e_1,r,e_2)} = \{(e_1', r, e_2)|e_1' \in E\} \cup \{(e_1 + r, e_2')|e_2' \in E\} \tag{4}$$

The optimization procedure is carried out by stochastic gradient descent and the process of embedding the Type Graph $(g(x_i)_{TG})$ is same as $g(x_i)_{SG}$. The theoretical number of parameters for training TransE is $O(n_e k + n_r k)$, where n_e and n_r is the number of entities and relations, respectively, and k is the dimension of graph embedding vector.

After obtaining embedding $\mathbf{x}_i$ of x_i, new input matrix $\mathbf{X}$ for π_i^l is given as follows:

$$\mathbf{X}_{\pi_i^l} = \bigcup_{x_i \in \pi_i^l} \mathbf{x}_i \tag{5}$$

The π_i^{ℓ} is converted into a $(L_{SG} + L_{TG}) * \ell$ matrix, where L_{SG} and L_{TG} are the length of SG and TG embedding vector, respectively. The left part of Figure 2 illustrates the process of graph embedding based feature extraction. We found that concatenation of the embedding vectors of an entity and its semantic type could give a slight improvement of the performance for drug discovery, this will be further discussed in the result section.

2) MODEL DESCRIPTION

We employed a recurrent neural network with long short term memory model to capture the dependencies between entities of drug-disease associations. Considering a single hidden layer network in which x_t, h_t and y_t denote the input, hidden and output layer neuron outputs, respectively, a general

recurrent network can be described as:

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (6)$$

where, W_{xh}, W_{hh} and b_h are the weight matrices across different connections and σ is a basic sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$). Note that $h_0 = e_0$ for our task. The dimension of fully connected layer in Figure 2 is T*1, where T is the hidden layer dimension of RNN model. The probability for the $drug_i$-$target_i$-$disease_i$ association π_i^ℓ is given as follows:

$$p(y_j = 1|\mathbf{X}) = \sigma(V_{h\ell}h_\ell) \quad (7)$$

where $\mathbf{X}$ represents the embeddings of input π_i^ℓ. $V_{h\ell}$ is the fully connected output layer, the dimension of $V_{h\ell}$ is 1*H, H is the dimension of hidden layer, respectively. Since RNNs have difficulties learning long-range dependencies, we adopted a LSTM as the drug discovery model in our experiment. LSTMs are explicitly designed to avoid the long-term dependency problem. It solves the gradient vanishing and exploding problem by introducing memory cell and forget gate. The LSTM is constructed as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (8)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (9)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (10)$$
$$g_t = tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \quad (11)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (12)$$
$$h_t = o_t \odot tanh(c_t) \quad (13)$$

where i_t, f_t and o_t are input gate, forget gate and output gate, respectively. The current cell state c_t will be generated by calculating the weighted sum using both previous cell state and current information generated by the cell. Trainable parameters are the weight matrices W_* and b_*. $\odot$ denotes element-wise multiplication. After obtaining the hidden state of LSTM, probability for drug-disease association is calculated as Equation 7.

We added a dropout layer to the non recurrent part of the LSTM to mitigate overfitting problem when training our model. Finally, we defined and optimized a cross entropy loss function $L(\theta)$ as follows:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n} yln(p(y|\mathbf{X}_i)) + (1-y)ln(1-p(y|\mathbf{X}_i)) \quad (14)$$

where y is 1 or 0. Our model was trained with back propagation through time [36].

D. IMPLEMENTATION FOR DRUG DISCOVERY

The trained LSTM model was used for discovering potential drugs for the disease of interest. To evaluate a potential treatment $drug_{potential}$-$target_{potential}$-$disease$, first a set of paths $\P_{potential} = \{\rho(drug_{potential} \rightarrow disease; target_{potential}, 2 \ldots \ell)\}$ were obtained. Then the score of the $drug_{potential}$ is calculated as:

$$score(drug_{potential}) = \mathbf{max}_{\pi_i \in \P_{potential}} D(g(\pi_i), \theta) \quad (15)$$

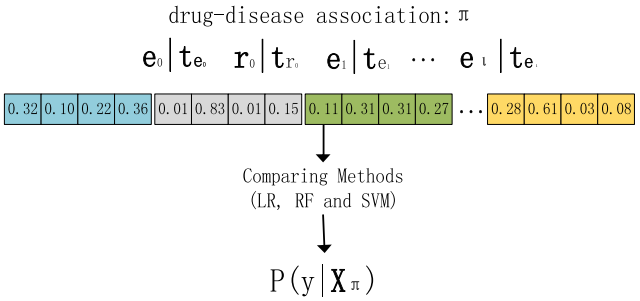


FIGURE 3. The features used in the comparing methods is the concatenation of the graph embedding vectors of training case.

Since the treatment of the *disease* is unknown, all pharmaceuticals could be potential drugs for one specific disease. Then all possible drugs were tested as candidate drugs for treating *disease*. Finally, we ranked all the candidate drugs by their scores.

E. BASELINE METHODS

To evaluate the performance of our method, we conducted ten-fold cross-validation and drug rediscovery test.

1) BASELINE METHODS FOR CROSS VALIDATION

Recently, numerous machine learning methods have been applied to predict drug target interactions [37]. The commonly used machine learning methods take drug target pairs as input, and the output of the methods is whether there is an interaction between a drug target pair. The most applied and successful machine learning models are binary classifiers such as logistic regression (LR), random forest (RF) [38] and support vector machine (SVM) [39]. Here, we used the most applied models as our baseline models for cross validation. The baseline models and our model used for the evaluations are as follows:

- Logistic Regression (LR).
- Random Forest (RF).
- Support Vector Machine (SVM).
- GrEDeL: our proposed graph embedding based LSTM model.

Features for the competitive methods were constructed in the same way as our model. The features of one element in *drug-disease* association is the concatenation of Semantic Graph embedding and Type Graph embedding. Figure 3 is an illustration of features used in the competitive methods. As shown in Figure 3, the input vector of the methods (LR, RF and SVM) is the concatenation of the graph embedding vectors of the training case.

2) BASELINE METHODS FOR DRUG REDISCOVERY

In this test, we compared our method with basic random walk method, two graph-based drug repositioning methods - NRWRH [40] and TP-NRWRH [41] - and three state-of-the-art knowledge graph based drug discovery methods Malas's method [16], Bakalb's method [17] and SemaTyP [18].

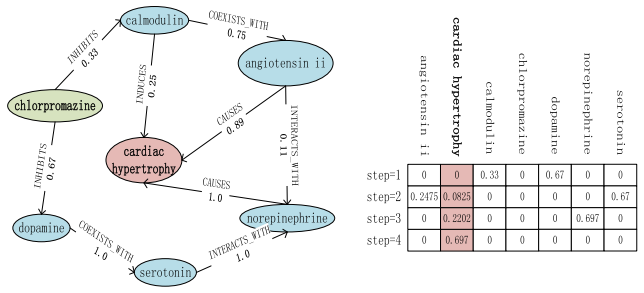


FIGURE 4. An illustration of RWA-based methods for drug discovery.

Specifically, NRWRH and TP-NRWRH are both graph-based random walk with restart algorithms [42]. NRWRH is a network-based random walk with restart method for inferring potential drug-target interactions. TP-NRWRH improves NRWRH by introducing a two-pass random walk algorithm. The process of discovering candidate drugs for $disease_i$ by RWA-based methods is as follows: First, the RWA algorithm starts from the $drug_{potential}$ point and the expected number of steps is set to n. Then, the $drug_{potential}$-$disease_i$ association could be scored by the RWA-based methods. After that, for each $disease_i$, RWA-based methods score all drugs for the disease of interest. At last, all candidate drugs could be ranked by their scores. Figure 4 illustrates an example of evaluating "*chlorpromazine*" to be the treatment of "*cardiachypertrophy*". The left part of Figure 4 is a weighted semantic graph. The right part of Figure 4 presents the results of basic RWA method with starting point "*chlorpromazine*". It shows that when the expected number of step is 1, the score of "*chlorpromazine*" is 0 due to "*chlorpromazine*" can not reach "*cardiachypertrophy*" in one step. Similarly, the score of "*chlorpromazine*"-"*cardiachypertrophy*" association is 0.0825 which is assigned by step_2 RWA. As shown in Figure 4, the highest score of "*chlorpromazine*" for treating "*cardiachypertrophy*" is 0.697 when the step of RWA is set to 4.

In addition, Malas's method leverages knowledge graph features such as the total number of intermediate concepts, the number of different semantic categories, and the predicates connecting a drug-disease pair to predict novel drug-disease associations [16]. Bakalb's method exploits all the paths connecting biomedical entities as features of logistic regression model to discover drugs [17]. SemaTyP is a knowledge graph based drug discovery method which exploits the distribution of semantic types of entities to score $drug_{potential}$-$disease_i$ association [18].

IV. RESULTS

In this section, we firstly introduce the biomedical knowledge graph and training data. Then we evaluate the performance of graph embedding method on link predication test. After that, our method was evaluated on cross-validation and drug rediscovery test separately. In addition, case studies are conducted to confirm the ability of our model to find potential drugs for diseases.

TABLE 1. The description of biomedical knowledge graph.

Materials	Number
PubMed abstracts	22,769,789
predications	39,133,975
selected predications	17,651,279
entities of knowledge graph	1,067,092
relations of knowledge graph	14,419,744
entity types	133
relation types	52

TABLE 2. The broad concept entity list for filtering.

Broad concept entities						
plnt	alga	fngs	virs	rich	bact	arch
anim	invt	vtbt	amph	bird	fish	rept
mamm	humn	anst	emst	cgab	acab	ffas
bdsy	evnt	acty	bhvr	socb	inbe	dora
ocac	hlca	lbpr	diap	topp	resa	mbrt
gora	edac	mcha	phpr	hcpp	eehu	npop
phob	mnob	medd	resd	cnce	idcn	tmco
qlco	qnco	spco	geoa	ocdi	bmod	orgt
hcro	pros	shro	grup	prog	popg	famg
aggp	podg	grpa	food	sosy	anab	neop

A. THE BIOMEIDCAL KNOWLEDGE GRAPH AND TRAINING DATA

The biomedical knowledge graph is constructed by extracting predications from the abstracts published in PubMed before June 1, 2013. In addition, in order to ensure the accuracy of extracted predication, we filtered out the predications that only appear once. Table 1 presents the details of the biomedical knowledge graph constructed in our work.

For building training set, on one side, we selected 7,144 *drug-target-disease* triplets from TTD as true cases. Then following the process in Section III-B, we obtained 6,188,265 positive training data. The ℓ was set to 4 as described in our previous work. The broad concept entities used are listed in Table 2. On the other side, we randomly constructed a set of $drug_{random}$-$target_{random}$-$disease_{random}$ cases for building false training data. Specifically, we kept the $drug_{random}$-$target_{random}$-$disease_{random}$ triplets that do not exist in TTD as false cases. In order to balance the positive and negative training data, 6,188,265 negative training data were obtained by the the same process of constructing the positive training data.

B. KNOWLEDGE GRAPH EMBEDDING

For evaluating the performance of knowledge graph embedding, we use the same evaluation protocols as in [30]. We selected subset of 'predications' which constructs the knowledge graph to evaluate graph embedding model. For each subset, 90% of the data were randomly selected as training data and the other 10% were test data. For each test triplet (e_{head}, r, e_{tail}), the e_{tail} is removed and replaced by each of the entities of test set. The dissimilarities of new triplets are first computed and then sorted by ascending order; the rank of the correct entity e_{tail} is finally stored. Then the mean of those predicted ranks and hits@10 are reported.

TABLE 3. The performance of different models.

Methods	Precision	Recall	F-score
LR	0.715	0.889	0.791
RF	0.742	0.750	0.743
SVM	0.622	0.766	0.635
GrEDeL$_{MLP_2}$	0.699	0.896	0.785
GrEDeL$_{MLP_3}$	0.707	0.901	0.792
GrEDeL$_{MLP_4}$	0.715	0.889	0.793
GrEDeL$_{MLP_5}$	0.711	0.903	0.796
GrEDeL$_{RNN}$	0.819	0.938	0.874
GrEDeL	**0.881**	**0.971**	**0.924**

TABLE 4. Performance of knowledge graph embedding method.

Dataset	5,000	10,000	20,000	40,000	100,000
Mean Ranks	305	261	176	122	108
Hits@10 (%)	9.4	15.7	22.6	27.1	29.8

TABLE 5. Example of link prediction results.

INPUT (e_{head} and r)	PREDICTED e_{tail}
Cysteine STIMULATE	beta-Lactams, bathocuproine disulfonate, hydrogen peroxide, **Pheomelanin**, N-Methylaspartate, aspartic acid receptor, deoxyhemoglobin, IL2, Cyclic GMP, Antineoplastic Agents
Alteplase INHIBITS	Cycloheximide, Urokinase, Plasminogen Activator Inhibitor 1, Thromboxane-A Synthase, **Alteplase**, alpha 1-Antitrypsin, FASTK Gene, Thromboxane A2 Receptor, Plasminogen Inactivators, Antithrombin III
Heparin AUGMENTS	Megakaryocytopoiesis, thrombin activity, Antiinflammatory Effect, growth factor binding, anti-toxin activity, IgG binding, Hemostatic function, Osteoclastic resorption, Nitrous oxide concentration, **Tyrosine Phosphorylation**

The dimension of embedding was set to 100 and γ =1. Table 4 shows the results of embeddings of Semantic Graph, the 'Dataset' row represents the number of triplets selected from Semantic Graph. Table 5 presents three examples of link predication results: given the entity e_{head} and the relation r, graph embedding method predicts the entity e_{tail}. The 'PREDICATED e_{tail}' column shows the top predicated e_{tail} entities, the entity in bold is the known e_{tail} entity. The results shows that graph embedding method could capture dependencies between entities and relations.

C. TEN-FOLD CROSS VALIDATION

We conducted ten-fold cross validations to evaluate the performance of the proposed GrEDeL against other three baseline methods in terms of commonly used measure: Precision, Recall and F-score. The dataset were split into ten subsets with equal size, and the positive and negative data in each subset is balanced. In particular, each subset does not share the same drug-disease pairs as other subsets in order to avoid overestimation of the training accuracy. Then each subset was taken in turn as a test set and other nine subsets were taken as input to run the methods. The average prediction accuracies over the test subsets were regarded as overall performance measures.

TABLE 6. The performance of GrEDeL with different knowledge graph embedding based features.

Methods	Precision	Recall	F-score
GrEDeL$_{random}$	0.608	0.759	0.68
GrEDeL$_{TG}$	0.773	0.824	0.797
GrEDeL$_{SG}$	0.819	**0.992**	0.897
GrEDeL$_{SG+OG}$	**0.881**	0.971	**0.924**

Table 3 shows that GrEDeL outperforms LR, RF and SVM models. We argue that our method can effectively capture temporal dependencies in drug-disease sequences. To further investigate the impact of each components in GrEDeL model, performances of GrEDeL with various graph embedding-based features are reported in Table 3. Random embedding (GrEDeL$_{random}$), Semantic Graph embedding only (GrEDeL$_{SG}$) and Type Graph embedding only (GrEDeL$_{TG}$) based models were proposed to evaluate the performance of GrEDeL. Specifically, GrEDeL$_{random}$ model uses the random embeddings for entities and relations, GrEDeL$_{SG}$ only uses semantic graph embedding for entities and relations (where the L$_{TG}$ =0) and similarly, L$_{SG}$ of GrEDeL$_{TG}$ is 0. Table 3 shows that although GrEDeL$_{TG}$ model only adopts type graph embeddings for the representation of knowledge graph, it still has reasonable performances. The reason is that the embedding of type graph could learn the rules of relations between entity types. This is similar as defining a semantic type pattern - such as *drug-INHIBITES-protein-STIMULATES-disease* - which could be used to select drug-disease associations based on their semantic types. However, there are only 133 entity types and 52 relation types in our knowledge graph (Table 1), which results in that GrEDeL$_{TG}$ model can not differentiate different entities with the same semantic type. In addition, Table 3 shows that the semantic graph embedding based GrEDeL model (GrEDeL$_{SG}$) has significantly boosted the performance. What's more, knowledge graph embedding (GrEDeL$_{SG+TG}$) outperforms other methods, the reason may be that knowledge graph embedding not only considers structure of knowledge graph but also preserves the information of semantic types. The graph embedding features capture rich information about the graph and entity-entity relations.

In additon, we replaced the LSTM of GrEDeL with other deep learning methods such as MLP (Multi-Layer Perceptron) and vanilla RNN in order to explore the influence of deep learning parts for GrEDeL. Table 3 presents the results and the N in $GrEDeL_{MLP\_N}$ is the number of hidden layers of MLP. The number of parameters of MLP and RNN is $N * n^2+nk+nm$ and $n^2+nk+nm$, respectively. Where k, n and m is the dimension of input layer, hidden layer and output layer, respectively. Table 7 shows the best performing dimension of hidden layer of MLP and RNN is 100 and 50, respectively. Although the total number of parameters of MLP is more than four times that of RNN, Table 3 shows the vanilla RNN based GrEDeL outperforms MLP based GrEDeL.The reason is that

TABLE 7. Validation hyper parameters. Steps of each range and selected values are shown.

Methods	Range	Step	Optimal
LSTM cell dimension	[25-150]	25	100
LSTM dropout	[0.3-0.8]	0.05	0.5
Hidden layer dimension of GrEDeL	[10-100]	10	30
L_{SG}	[25-100]	25	100
L_{TG}	[10-50]	10	10
Hidden layer dimension of MLP	[10-100]	100	100
Hidden layer dimension of RNN	[10-100]	10	50

GrEDeL works well not because of deep learning structure but because of sequential characteristic. The performance improvement could come from sequential characteristics of RNN-based methods. In addition, LSTM based GrEDeL achieves better performances than vanilla RNN based method due to LSTM could learn long-term dependent relationships in the *drug-target-disease* associations.

We validated 5 different hyper parameters of our model (as shown in Table 7). Best performing parameter set on validation phase was used for the GrEDeL model. The Adam optimizer [43] was used for optimizing GrEDeL and the learning rate decay was set to 0.99 for every 100 iterations. Additionally, mini batch of size 500 was used. Implementations of LR, RF and SVM is based on Sklearn library.

D. DRUG REDISCOVERY TEST

We conducted drug rediscovery test to evaluate the performance of our method in discovering drugs for diseases of interest. Here, we adopted the same *drug-disease* relationships as used in our previous work for drug rediscovery test. In particular, in the path obtaining process of this work, we filtered out the paths between *drug* and *disease* that containing broad concepts. After the path exploring process, there are only 115 of the 360 standard relationships used in our previous work have paths in the knowledge graph. Then we used the 115 *drug-disease* relationships as gold standard cases for drug rediscovery test. For each gold standard cases, TTD has reported that the drug could treat the disease, but the corresponding drug targets are not clear. For evaluation, we used the same ranking procedure as described in [18]. Specifically, for each gold standard $drug_i$-$disease_i$, we randomly selected other 100 drugs (including chemical entities and new biologic entities) from TTD as potential drugs for treating $disease_i$. Then drug discovery models scored all the 101 drugs for treating the $disease_i$. Lastly, the average ranking of standard drugs (mean ranks) among all the standard cases and the proportion of known drugs ranked in the top 10 (Hits@10) were reported to evaluate the performance of the models. For RWA-based methods (NRWRH and TP-NRWRH), if the standard $drug_i$ of $disease_i$ is not found by the drug discovery model, then the $drug_i$ is scored 0 and the corresponding ranking is 101. What's more, for SemaTyP and GrEDeL, we selected 5,785 targets from TTD as candidates drug targets for constructing $drug_i$-$target_{candidate}$-$disease_i$ associations.

TABLE 8. Performance of discovering drugs for disease.

Method	Not Found	Mean Ranks	Hits@10 (%)
RWA_1	93	90.82	17.39
RWA_2	52	44.31	24.46
RWA_3	4	30.33	23.48
RWA_4	0	33.84	18.26
RWA_5	0	35.27	14.78
RWA_6	0	39.14	11.30
Malas's Method [16]	52	37.68	24.34
NRWRH [40]	26	32.17	26.09
TP-NRWRH [41]	19	31.13	27.83
Bakalb's Method [17]	4	29.44	31.30
SemaTyP [18]	0	29.87	30.58
GrEDeL	0	**27.05**	**33.04**

In our experiments, all competitive methods (NRWRH, TP-NRWRH, Malas's method, Bakalb's method and SemaTyP) follow the recommended settings reported in their papers and the expected number of step of the basic RWA method was set from 1 to 6. The overall results of drug rediscovery are presented in Table 8. The "Not Found" denotes the number of gold standard drugs which are not discovered by the corresponding method. "RWA_n" denotes the basic random walk algorithm with n steps.

For "Not found", we find that increasing the number of steps improves the performance of basic RWA method. The reason is that the RWA with more steps could cover more entities. For instance, as shown in Table 8, 93 golden standard drugs are not discovered by RWA_1, this is because there are only 22 (115-93) golden standard drugs directly connect to their corresponding diseases in the knowledge graph. As the number of steps increases, basic RWA method could discover more drugs. Table 8 shows that the basic RWA methods could find all drugs when the number of steps exceeds 3. It is due to the fact that in the knowledge graph of our experiments, all golden standard drugs and their corresponding diseases can be connected by a path of length 4. As we can see from Table 8, there are 52 and 4 drugs were missed by Malas's method and Bakalb's method, respectively. This reason is Malas's method considers one intermediate between drug and disease and Bakalb's method just considers all paths of length ≤ 3. Table 8 shows NRWRH and TP-NRWRH achieve poor performance than some basic RWA methods with respect to "Not found" metric. The primary reason is that, NRWRH and TP-NRWRH are both random walk with restart algorithms, which may result in that some disease nodes can not be reached by the golden standard drugs within desired steps. Moreover, Table 8 shows SemaTyP and GrEDeL rediscover all drugs for the diseases, as both methods could explore all paths of lengths 3 to 6.

Table 8 shows RWA_1 achieves the worst result (90.82) in terms of "*Mean Ranks*". For the reason that most of the drugs (93 drugs) have not been found by this method. In addition, RWA_2 dramatically improves the performance as RWA with 2 steps could discover more drugs than that of RWA_1 method. As the number of steps increases, the

TABLE 9. Case studies of drug discovery.

Disease	Drug	Ranking	Score	Mechanism of Action
cardiovascular disease	ioxaglate	1	0.57	ioxaglate *DISRUPTS* platelet aggregation *AFFECTS* signal transduction pathway *AFFECTS* cardiovascular disease
inflammatory disease	sb-612111	1	0.69	sb-612111 *NEG_AFFECTS* consumption *AFFECTS* bone metabolism *AFFECTS* rheumatoid arthritis *PROCESS_OF* inflammatory disease
dementia	rx-77368	9	0.51	rx-77368 *NEG_AFFECTS* blood flow *AFFECTS* hsf1 *AFFECTS* disease *COEXISTS_WITH* dementia
mood disorder	mcpp	4	0.28	mcpp *COEXISTS_WITH* cortisol *DISRUPTS* telomere *LOCATION_OF* disease *COEXISTS_WITH* mood disorder
pain	dpi-3290	9	0.44	pi-3290 *DISRUPTS* tension *COEXISTS_WITH* pe *CAUSES* symptom *PROCESS_OF* pain
cancer	tr-2	10	0.13	tr-2 *INTERACTS_WITH* cyclosporine *ISA* p-glycoprotein *AFFECTS* tissue *LOCATION_OF* cancer

performance in terms of "*Mean Ranks*" can be further improved. For the basic RWA methods, Table 8 shows RWA_3 reaches the best performance (30.33). However, the performance decreases as the number of steps continues to grow. This is because the larger number of steps allows the RWA method to find more candidate drugs, which in turn leads to a lower ranking of the golden standard drugs. Table 8 shows that NRWRH and TP-NRWRH outperform all basic RWA methods. This is because both NRWRH and TP-NRWRH incorporate semantic types of entity besides the graph structure information. Similarly, Table 8 shows SemaTyP further improves the performance by making full use of semantic information of knowledge graph. In addition, Malas's method achieves poor performance (37.68) in terms of "*Mean Ranks*", this is due to there are 52 drugs were missed by the method.

For "*Hits*@10", RWA_2 outperforms other basic RWA methods (24.46%). As shown in Table 8, although the performance decreases slightly when the step increases to 3 (23.48%), continuously increasing the number of steps of RWA will cause significantly performance degradation. Furthermore, Table 8 shows NRWRH and TP-NRWRH achieve better performances than all basic RWA methods. Compared with the results obtained by RWA-based methods, Bakalb's method and SemaTyP achieve better performance in terms of "*Hits*@10" metric.

Table 8 shows our method, GrEDeL, outperforms all counterparts on all metrics (The "*Mean Ranks*" and "*Hits*@10" is 27.05 and 33.04%, respectively). The main reasons are: 1) GrEDeL makes full use of both semantic information and structure of graph: the graph embedding features capture more information of the biomedical knowledge graph than other competitive methods. 2) GrEDeL considers the process of literature based discovery as a series analysis problem. By using recurrent neural network, GrEDeL can learn the dependencies among entities of drug-disease associations. We are the first to consider the process of literature based discovery as a series analysis problem.

E. CASE STUDY

In this section, we conducted six case studies to show the efficacy of our approach (Table 9). The scores in the table are the probability indicating whether there is a relationship between a drug and a disease. Since the drug mechanism of actions are unknown, the associations obtained by our model was adopted to verify the hypotheses. For each disease of interest, GrEDeL predicts both the potential drugs and the corresponding drug targets simultaneously. For example, TTD has reported that ioxaglate is one known drug for cardiovascular disease, but the drug mechanism is still unknown. GrEDeL ranks ioxaglate as the 1st potential drug for treating cardiovascular disease. What's more, our method also provides corresponding mechanism of action of ioxaglate, Table 9 shows that ioxaglate acts on cardiovascular disease by disrupting platelet aggregation which affects the signal transduction pathway in cardiovascular disease. Other examples: rx-77368 is predicated to treat dementia by acting on hsf. Tr-2 is predicated to treat cancer by acting on p-glycoprotein, etc. The cases in Table 9 show our method has the potential to discover drugs as well as their corresponding drug targets. However, the drug mechanism of actions generated by LBD must then be verified by human judgment and with experimental methods or clinical studies, depending on the nature of the discovery [44].

F. COMPLEXITY

Our experiments were conducted on a PC with 4 Intel(R) Xeon(R) CPU E5-2609 of 2.4 GHz and 8GB internal memory, the LSTM was implemented in TensorFlow 1.1.0 with GPU (CUDA 8.0.61) support. The source code of our implementation was released at.[1] Table 10 shows the time needed

[1] https://github.com/ShengtianSang/GrEDeL

TABLE 10. Running time (in hours), the number in left column is the length of graph embedding.

graph embedding dimension	running time
$SG_{25} + TG_{10}$	4.6
$SG_{50} + TG_{20}$	11.2
$SG_{75} + TG_{30}$	15.3
$SG_{100} + TG_{10}$	16.1
$SG_{100} + TG_{50}$	27.3

for training of the GrEDeL in terms of the different graph embedding dimensions (the sum of the length of SG embedding and TG embedding). The first column in Table 10 is the dimension of graph embedding and the second column show the total running time for training GrEDeL. The dimension of hidden layer in LSTM was adopted the optimal setting in Table 7.

The running time for drug discovery depends on the total number of candidate drugs and corresponding drug targets. In this work, 100 candidate drugs and 5,785 targets (a protein, peptide or nucleic acid) extracted from TTD were used to construct candidate *drug-target-disease* associations for a given disease of interest. The average running time for processing one *drug-target-disease* association is 12 ms.

V. DISCUSSION

As far as we know, this is the first method that incorporates biomedical knowledge graph, knowledge graph embedding, as well as deep learning methods to discover drugs from literature. Our overall approach however, has several limitations: 1) The construction of our biomedical knowledge graph relies heavily on effective natural language processing tool (SemRep). Although we filtered out all isolated predications in order to improve the quality of the biomedical knowledge graph, there are still a large number of false positive relations existing in the knowledge graph, which in turn leads to our method inferring lower-quality results. 2) The positive training data constructed in our work consist of instances corresponding to paths extracted from the KG. However, the instances may correspond to overlapping paths. This could introduce bias to the ten-fold cross validation evaluation as an instance appearing on the training set could be very similar (due to the overlap) to another instance that is used in the test set, and thus lead to an overestimation of the performance. 3) The TransE is adopted in our method for knowledge graph embedding. It only achieves promising results in handling 1-to-1 relations. However, the biomedical knowledge graph also contains some 1-to-n and n-to-n relations.

In future work, we would like to develop high-quality NLP tools, in particular, biomedical NLP tools, to improve the quality of the biomedical knowledge graph. Additionally, other graph embedding methods could be used for capturing multi-mapping relations between entities.

VI. CONCLUSION

In this paper we have introduced a framework to jointly utilize knowledge graph and deep learning methods for discovering drugs from literature. The experimental results show that our method can effectively discover potential drugs and their corresponding mechanism of actions. It could be a supplementary method for current drug discovery methods, which could improve the successfulness in discovering new medicine for currently incurable diseases.

REFERENCES

[1] R. Goulding and E. Marden, "An overview of drug discovery and drug development," *OHI Del.*, vol. 1, 2009.

[2] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: A systematic review," *Health Policy*, vol. 100, no. 1, pp. 4–17, 2011.

[3] A. Korhonen *et al.*, "Improving literature-based discovery with advanced text mining," in *Computational Intelligence Methods for Bioinformatics and Biostatistics* (Lecture Notes in Computer Science), vol. 8623. Springer, 2014, pp. 89–98.

[4] D. Gubiani, I. Petrič, E. Fabbretti, and T. Urbančič, "Mining scientific literature about ageing to support better understanding and treatment of degenerative diseases," Tech. Rep., 2015.

[5] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986.

[6] R. A. Digiacomo, J. M. Kremer, and D. M. Shah, "Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study," *Amer. J. Med.*, vol. 86, no. 2, pp. 158–164, 1989.

[7] R. K. Lindsay and M. D. Gordon, "Literature-based discovery by lexical statistics," *J. Assoc. Inf. Sci. Technol.*, vol. 50, no. 7, pp. 574–587, 1999.

[8] M. D. Gordon and R. K. Lindsay, "Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil," *J. Assoc. Inf. Sci. Technol.*, vol. 47, no. 2, pp. 116–128, 1996.

[9] M. Weeber, H. Klein, L. T. W. de Jong-van den Berg, and R. Vos, "Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries," *J. Assoc. Inf. Sci. Technol.*, vol. 52, no. 7, pp. 548–557, 2001.

[10] S. Sang, Z. Yang, Z. Li, and H. Lin, "Supervised learning based hypothesis generation from biomedical literature," *BioMed Res. Int.*, vol. 2015, May 2015, Art. no. 698527.

[11] S. Henry and B. T. Mcinnes, "Literature based discovery: Models, methods, and trends," *J. Biomed. Inform.*, vol. 74, pp. 20–32, Oct. 2017.

[12] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin, "Exploiting semantic relations for literature-based discovery," in *Proc. Annu. Symp. AMIA*. Bethesda, MD, USA: American Medical Informatics Association, 2006, p. 349.

[13] M. Rastegar-Mojarad, R. K. Elayavilli, D. Li, R. Prasad, and H. Liu, "A new method for prioritizing drug repositioning candidates extracted by literature-based discovery," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Nov. 2015, pp. 669–674.

[14] C. B. Ahlers, D. Hristovski, H. Kilicoglu, and T. C. Rindflesch, "Using the literature-based discovery paradigm to investigate drug mechanisms," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2007, pp. 6–10.

[15] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider, "Context-driven automatic subgraph creation for literature-based discovery," *J. Biomed. Inform.*, vol. 54, pp. 141–157, Apr. 2015.

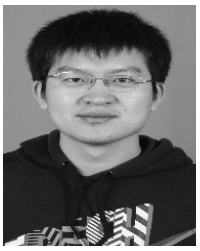
[16] T. B. Malas *et al.*, "Drug repurposing using a semantic knowledge graph," Tech. Rep.

[17] G. Bakal, G. Bakal, E. V. Kakani, and R. Kavuluru, "Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations," *J. Biomed. Inform.*, vol. 82, pp. 189–199, Jun. 2018.

[18] S. Sang, Z. Yang, L. Wang, X. Liu, H. Lin, and J. Wang, "Sematyp: A knowledge graph based literature mining method for drug discovery," *BMC Bioinf.*, vol. 19, no. 1, p. 193, 2018.

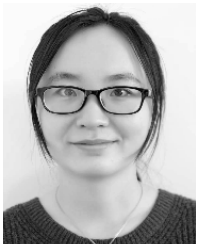
[19] I. I. Baskin, D. Winkler, and I. V. Tetko, "A renaissance of neural networks in drug discovery," *Expert Opinion Drug Discovery*, vol. 11, no. 8, pp. 785–795, 2016.

[20] P.-W. Keith, C. C. Chan, and Z.-H. You, "Large-scale prediction of drug-target interactions from deep representations," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 1236–1243.
[21] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, no. 6, pp. 1241–1250, 2018.
[22] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system," *Methods Inf. Med.*, vol. 32, no. 04, pp. 281–291, 1993.
[23] A. T. McCray, "The UMLS semantic network," in *Proc. Annu. Symp. Comput. Appl. Med. Care*. Bethesda, MD, USA: American Medical Informatics Association, 1989, pp. 503–507.
[24] X. Chen, Z. L. Ji, and Y. Z. Chen, "TTD: Therapeutic target database," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 412–415, 2002.
[25] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text," *J. Biomed. Inform.*, vol. 36, no. 6, pp. 462–477, 2003.
[26] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.
[27] C. Bizer *et al.*, "DBpedia—A crystallization point for the Web of data," *J. Web Semantics*, vol. 7, no. 3, pp. 154–165, 2009.
[28] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. De Melo, and G. Weikum, "Yago2: Exploring and querying world knowledge in time, space, context, and many languages," in *Proc. ACM 20th Int. Conf. Companion World Wide Web*, 2011, pp. 229–232.
[29] H. Cai, V. W. Zheng, and K. C.-C. Chang. (2017). "A comprehensive survey of graph embedding: Problems, techniques and applications." [Online]. Available: https://arxiv.org/abs/1709.07604
[30] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.
[31] M. Hüsken and P. Stagge, "Recurrent neural networks for time series classification," *Neurocomputing*, vol. 50, pp. 223–235, Jan. 2003.
[32] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[33] W. Zaremba, I. Sutskever, and O. Vinyals. (2014). "Recurrent neural network regularization." [Online]. Available: https://arxiv.org/abs/1409.2329
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[35] J. D. Wren, R. Bekeredjian, J. A. Stewart, R. V. Shohet, and H. R. Garner, "Knowledge discovery by automated identification and ranking of implicit relationships," *Bioinformatics*, vol. 20, no. 3, pp. 389–398, 2004.
[36] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.
[37] M. Wen *et al.*, "Deep-learning-based drug–target interaction prediction," *J. Proteome Res.*, vol. 16, no. 4, pp. 1401–1409, 2017.
[38] D.-S. Cao *et al.*, "Computational prediction of drug–target interactions using chemical, biological, and network features," *Mol. Inform.*, vol. 33, no. 10, pp. 669–681, 2014.
[39] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification," *J. Chem. Inf. Model.*, vol. 43, no. 6, pp. 1882–1889, 2003.
[40] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Mol. BioSyst.*, vol. 8, no. 7, pp. 1970–1978, 2012.
[41] H. Liu, Y. Song, J. Guan, L. Luo, and Z. Zhuang, "Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks," *BMC Bioinf.*, vol. 17, no. 17, p. 539, 2016.
[42] M. Gori, A. Pucci, V. Roma, and I. Siena, "ItemRank: A random-walk based scoring algorithm for recommender engines," in *Proc. IJCAI*, vol. 7, Jan. 2007, pp. 2766–2771.
[43] D. P. Kingma and L. J. Ba, "A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, p. 13.
[44] D. Hristovski, T. Rindflesch, and B. Peterlin, "Using literature-based discovery to identify novel therapeutic approaches," *Cardiovascular Hematolog. Agents Medicinal Chem. (Formerly Current Medicinal Chem.-Cardiovascular Hematolog. Agents)*, vol. 11, no. 1, pp. 14–24, 2013.



SHENGTIAN SANG is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Dalian University of Technology, Dalian, China. His research interests include literature-based discovery, knowledge graph, and data mining.

ZHIHAO YANG received the B.Sc., M.Sc., and Ph.D. degrees from the Dalian University of Technology, China, in 1997, 2003, and 2008, respectively. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. He has published over 30 research papers on topics in biomedical literature data mining. His current research interests include biomedical literature data mining, natural language processing, and machine learning. His research projects are funded by the National Natural Science Foundation of China and the Major State Research Development Program of China.



XIAOXIA LIU is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include network science, data mining, and bioinformatics.

LEI WANG received the Ph.D. degree from the Beijing Institute of Health Administration and Medical Information, China, in 2006. She is currently a Professor with the Beijing Institute of Health Administration and Medical Information. Her current research interests include biomedical literature data mining, medical science and technology information, and science and technology development strategy. She has published over 20 research papers and four monographs. Her research projects are funded by the National Natural Science Foundation of China and the Major State Research Development Program of China.

HONGFEI LIN received the B.Sc. degree from Northeastern Normal University, in 1983, the M.Sc. degree from the Dalian University of Technology, in 1992, and the Ph.D. degree from Northeastern University, in 2000. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. He has published over 100 research papers in various journals, conferences, and books. His research interests include information retrieval, text mining, natural language processing, and effective computing. In recent years, he has focused on text mining for biomedical literatures.

JIAN WANG received the B.Sc., M.Sc., and Ph.D. degrees from the Dalian University of Technology, China, in 1997, 2003, and 2008, respectively. She is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. She has published over 30 research papers on topics in biomedical literature data mining. Her current research interests include biomedical literature data mining, natural language processing, and machine learning.

MICHEL DUMONTIER was an Associate Professor of medicine (biomedical informatics) with the Stanford University School of Medicine, and also an Associate Professor of bioinformatics with Carleton University. He is currently a Distinguished Professor of data science with Maastricht University. He is best known for his work in biomedical ontologies, linked data, and biomedical knowledge discovery. His research has been funded by the Natural Sciences and Engineering Research Council, the Canada Foundation for Innovation, Mitacs Canada, the Ontario Ministry of Research, Innovation and Science, CANARIE, and the US National Institutes of Health. He has an h-index of over 30 and has authored over 125 scientific publications in journals and conferences. His research focuses on methods to represent knowledge on the web, with applications for drug discovery and personalized medicine.