Briefings in Bioinformatics, 22(4), 2021, 1–15

https://doi.org/10.1093/bib/bbaa344
Problem Solving Protocol

PharmKG: a dedicated knowledge graph benchmark for bomedical data mining

Shuangjia Zheng†, Jiahua Rao†, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang and Zhangming Niu

Corresponding authors: Yuedong Yang, School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510000, China. Tel.: +86 020-37106020; Fax.: +86 020-37106020. E-mail: yangyd25@mail.sysu.edu.cn; Zhangming Niu, Aladdin Healthcare Technologies Ltd., 25 City Rd, Shoreditch, London EC1Y 1AA, United Kingdom. Tel.: +44 7934215894; Fax. +49 30 700 140 150. E-mail: zhangming@aladdinid.com
†These authors contributed equally to this work.

Abstract

Biomedical knowledge graphs (KGs), which can help with the understanding of complex biological systems and pathologies, have begun to play a critical role in medical practice and research. However, challenges remain in their embedding and use due to their complex nature and the specific demands of their construction. Existing studies often suffer from problems such as sparse and noisy datasets, insufficient modeling methods and non-uniform evaluation metrics. In this work, we established a comprehensive KG system for the biomedical field in an attempt to bridge the gap. Here, we introduced PharmKG, a multi-relational, attributed biomedical KG, composed of more than 500 000 individual interconnections between genes, drugs and diseases, with 29 relation types over a vocabulary of ∼8000 disambiguated entities. Each entity in PharmKG is attached with heterogeneous, domain-specific information obtained from multi-omics data, i.e. gene expression, chemical structure and disease word embedding, while preserving the semantic and biomedical features. For baselines, we offered nine state-of-the-art KG embedding (KGE) approaches and a new biological, intuitive, graph neural network-based KGE method that uses a combination of both global network structure and heterogeneous domain features. Based on the proposed benchmark, we conducted extensive experiments to assess these KGE models using multiple evaluation metrics.

Shuangjia Zheng is currently a PhD student in the School of Data and Computer Science at the Sun Yat-Sen University. His research interests lie in the deep learning, knowledge graph embedding, drug discovery and computational biology.
Jiahua Rao is a PhD student in the School of Data and Computer Science at the Sun Yat-Sen University. His research interests include deep learning, multi-omics integration, knowledge graph and computational biology.
Ying Song is a master student in the School of Systems Science and Engineering at the Sun Yat-Sen University. His research interests lie in the computer vision and knowledge graph.
Jixian Zhang is a Senior Data Scientist in the Aladdin Healthcare Technologies Ltd. He received his master degree from the University of Science and Technology of China, and his research interests include data mining, knowledge graph construction, computational biology and machine learning.
Xianglu Xiao is a Senior Data Scientist in the Aladdin Healthcare Technologies Ltd. He received his master degree from the Imperial College London, and his research interests include knowledge graph construction, medical imaging and machine learning.
Evandro Fei Fang is an Associate Professor in the Department of Clinical Molecular Biology, University of Oslo and Akershus University Hospital, Lørenskog, Norway. He is a Molecular Gerontologist and runs a research lab aiming at understanding the molecular mechanisms of human aging. His team uses the bench-top knowledge to guide the development of novel interventional strategies towards human aging, with a final goal to improve the quality of life in the elderly.
Yuedong Yang is a Professor in the School of Data and Computer Science and the National Super Computer Center at Guangzhou, Sun Yat-sen University, China. Currently, his research group emphasizes on developing HPC and AI algorithms for multi-omics data integration and intelligent drug design. He is also responsible for constructing the HPC platform for biomedical applications based on the Tianhe-2 supercomputer.
Zhangming Niu is CTO of the Aladdin Healthcare Technologies Ltd. He is a recognized leader in AI development with 10 years of experience in building novel products for blue chip companies, governments and institutions. He is ranked in the top 10 of multiple global AI competitions for healthcare and industry applications.
Submitted: 12 August 2020; **Received (in revised form):** 12 October 2020

Finally, we discussed our observations across various downstream biological tasks and provide insights and guidelines for how to use a KG in biomedicine. We hope that the unprecedented quality and diversity of PharmKG will lead to advances in biomedical KG construction, embedding and application.

Key words: knowledge graph; knowledge graph embedding; computational prediction model; drug repositioning; Alzheimer's disease

Introduction

The complex interplay between biomedical entities—i.e. genes, chemicals and diseases—has long been intriguing to biomedical researchers. Understanding these interconnections is the key to illuminating the underlying mechanisms behind different biological functions (e.g. agonism, metabolism, mutation, etc.) and can thus greatly benefit various biomedical studies, such as drug repositioning [1], adverse drug reaction analysis [2], proteomics data analysis [3], etc. This has encouraged the development of numerous physical and computational strategies to evaluate, analyze and infer different types of these associations.

For a decade, basic networks such as undirected and uni-relational graphs were used to model intricate interactions in biomedical systems [4–8]. Despite the impressive performances of these models, these networks failed to capture the semantics within different types of relationships between biomedical entities. For example, drug–protein interactions modeled with basic networks cannot distinguish between different kinds of interactions such as inhibition, activation, binding, etc. Because of this, many recent works have since switched to using multi-relational networks, i.e. knowledge graphs (KGs), where KG embedding (KGE) [9–12] approaches were utilized to map graphs into a low-dimensional space while maximally preserving its topological properties. As such, downstream tasks such as relation prediction, clustering and visualization can be done by typical non-network-based models [13–15]. Although existing methods show the capacity of processing KGs and demonstrate great promises, the usage of KGs in biomedical applications has suffered from the lack of standard evaluation benchmarks and expressly designed biomedical KGE models [16, 17].

Unlike other areas where data are clean and well-structured, knowledge bases derived from the biomedical domain are usually sparse, redundant and incomplete. Though many manually curated knowledge bases such as PharmGKB [18], OMIM [19], CTD [20] and DrugBank [21] offer high-quality data sources, they often record only simple connections between entity pairs while losing the essential mechanisms behind the associations, making it challenging to construct multi-relational graphs. Meanwhile, many recently developed, preliminary biomedical KGs such as Bio2RDF [22], OpenBiolink [23] and Triple [24], contain either a significant number of metadata relations or trivial biomedical entities that can interfere with the performance of KGE algorithms [17]. Therefore, it is necessary to construct a comprehensive, high-quality benchmark.

In recent years, KGE models have experienced rapid developments that have allowed them to make accurate predictions of relations [10, 12, 25–27]. Advances in methodology are both evaluated and steered by some established general-domain benchmarks, such as the FB15K benchmark derived from Freebase and the WN18 benchmark derived from WordNet [9]. Unfortunately, most of the existing methodologies struggle to reflect the domain-specific properties of heterogeneous biomedical KGs due to either their well-structured knowledge networks (FB15K) or hierarchical taxonomies (WN18). In fact, biomedical KGs often combine richly structured ontological hierarchies with complex interconnections, thus making it hard to predict relations based on simple or straightforward rules. Likewise, biomedical KGs usually contain abundant non-topological domain information, e.g. gene expression [28], chemical structure and disease description [29], requiring a deep combination of network embedding methods and feature learning techniques to make full use of the information.

Another problem worth noting is that the existing studies often benchmark proposed methods on disjointed dataset collections for specific downstream tasks, such as drug repurposing [7, 30, 31] and adverse drug reactions [32, 33]. As such, the evaluation metrics and setting are not standardized across different works, making it a challenge to judge whether a proposed method does, in fact, improve performance.

In this study, we attempted to bridge the gap by establishing a comprehensive KG system, PharmKG, for the biomedical field. As shown in Figure 1, the PharmKG is a multi-relational, attributed biomedical KG composed of more than 500 000 individual interconnections between genes, drugs and diseases of 29 relation types, annotated by a vocabulary of ∼8000 disambiguated entities. Each entity in the PharmKG is attached with heterogeneous, domain-specific information obtained from multi-omics data, including gene expression, chemical structure and disease word embedding, while preserving both semantic and biomedical information. For baselines, we offer nine state-of-the-art embedding approaches and a novel, biological, intuitive graph neural network-based KGE method that integrates both global network structure and heterogeneous domain features. We have also conducted extensive experiments with various state-of-the-art KGE models in the same evaluation standard using the proposed benchmark. Furthermore, we discussed our observations across various downstream biological tasks and provided some insights and guidelines for how to use KGs in for biomedicine tasks and applications.

In summary, our contributions are 3-fold as follows.

1. Establishment of a dedicated, high-quality and trustworthy benchmark optimized for evaluating multi-relation prediction methods in large attributed biomedical KGs. By integrating six professional public resources and text-mined knowledge bases, this dedicated biomedical KG, PharmKG, contains thousands of nodes containing genes, chemical compounds and diseases, connected by a set of semantic relationships derived from the abstracts of biomedical literature. Each entity in PharmKG is labeled with domain-specific information, preserving semantic and biomedical features of the data.
2. Introduction of a new baseline through a novel biological intuitive graph neural network-based KGE method meant to alleviate issues found in existing methods and to capture heterogeneous information embedded in complex, biomedical KGs.

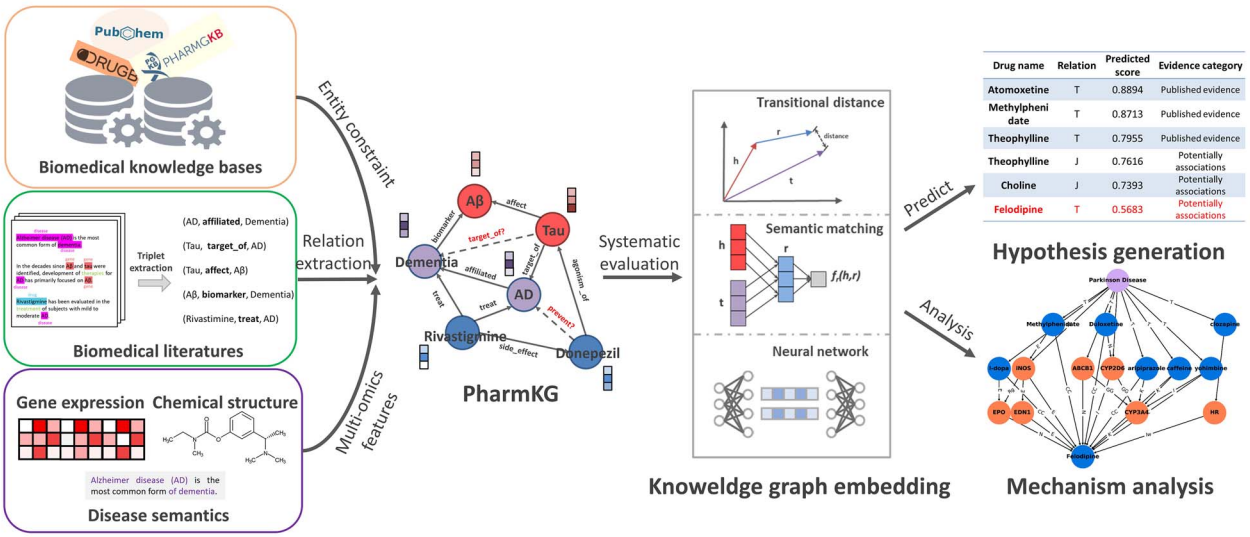

Figure 1. Pipeline for PharmKG construction, modeling and applications. PharmKG depends on biological knowledge bases and includes a larger set of biomedical facts derived from the research literature to fill in missing semantics. Each entity in PharmKG was labeled with domain-specific information, persevering semantic and biomedical features. Low-dimensional entity and relation representations are first learned from PharmKG by KGE approaches and then used to build specific systems for hypothesis generation and mechanism analysis.

3. Validation of the PharmKG through extensive experiments with various KGE models using multiple evaluation metrics. One external biomedical KG was also used to compare the quality of the benchmark and to evaluate KGE methods systematically. We discussed our observations across various downstream biological tasks and applications to provide insights and guidelines for how to use PharmKG in the biomedical domain.

The rest of the paper is organized as follows: Background Section—A review of the most relevant works on methods to construct and embed biomedical KGs. We also discuss flaws in current embedding methods and in KG dataset construction. Materials Section—A detailed construction and analysis of PharmKG. Modeling and application of PharmKG Section—An exploration of the difficulties encountered in the PharmKG dataset upon using several baselines and a discussion of the predictive and analytical capabilities of KGE models on constructed KG datasets novel and possible novel solutions. Downstream apllications Section—Analysis and discussion of the downstream applications of PharmaKG.

Background

In this section, we introduced the related works in two areas. First, we briefly reviewed current progress in the biomedical KG field. Second, we summarized current approaches for KG embedding.

KGs in biomedical research

KGs are multi-relational, directed graphs in which nodes represent entities and edges represent their relations. Starting from a number of high-quality, manually curated biomedicine knowledge bases such as TTD [34], PharmGKB [18], OMIM [19] and DrugBank [21], knowledge network-based studies have rapidly advanced to utilizing larger datasets for the next generation of network analysis algorithms. For more explorations of drug knowledge bases, we refer readers to an extensive survey [35]. However, most of them only identify the existence of a relationship between entity pairs without containing specific semantic relation types, thus cannot be treated as a KG in the strict sense [16].

The first major biomedical KG work was published by Belleau *et al.* [22], where semantic web technologies were applied to convert publicly available bioinformatics databases into RDF formats. From the processed RDF file, the biomedical triplets (entity, relation and entity) could be subsequently obtained to construct a biomedical KG. Unfortunately, this kind of KG contains a significant number of metadata relations that can interfere with the performance of link prediction algorithms, and special care was needed to exclude trivially inferable statements from the test set [23].

Since then, efforts have been focused on constructing task-oriented KGs and applying them to downstream biomedical applications, such as drug repositioning [15, 24, 31], with only a few KGs focused on the construction of annotated, clarified biomedical knowledge networks. For example, Percha *et al.* [36] compiled a rough KG known as the Global Network of Biomedical Relationships (GNBR) from large-scale biomedical literature with unsupervised techniques and used it to generate drug repurposing hypotheses [31]. Though this work first provided specific themes for each kind of interaction between two entities, it suffered from name ambiguity and a high false-positive rate, and additionally no benchmark dataset was provided. Himmelstein *et al.* [37] constructed an integrative network encoding data from 29 public resources and provided a basic model to study drug repurposing. Alshahrani [16] and Breit *et al.* [23] also benchmarked large-scale KGs by integrating several public databases with annotation by Bio2RDF. While these works are ideologically consistent with the larger applications of KGs, they are restricted to ambiguous relationship types and uncommon entities that include a large number of trivial knowledge. Furthermore, the embedding methods used in these works treated the entities as simple homogenous instances, ignoring the heterogeneous biological information embedded in the entities themselves.

Table 1. A comparison between some of the existing biological KGs in terms of the properties and coverage of different types of biological entities. The abbreviation M represents manually curated data, A represents automated curated data and MA represents the combination of these two kinds of sources

Category	Name	Properties				Coverage			
		construction	Semantic relation	Graph structure	Domain feature	Ontology	Chemical	Disease	Gene
Knowledge base	TTD [33]	M				√	√		√
	PharmGKB [17]	MA				√	√		√
	SIDER [34],	A					√		
	OMIM [18]	M				√		√	√
	Drugbank [20]	M					√		√
	CTD [19]	A				√	√	√	√
KG	Hetionet [35]	M	√	√		√	√	√	√
	GNBR [36]	A	√				√	√	√
	OpenBioLINK [22]	A	√	√			√	√	√
	Triple [23]	M		√		√	√	√	√
	Polypharm [31]	MA		√	√		√		√
	Biokeen [37]	A		√			√	√	√
	PharmKG	MA	√	√	√	√	√	√	√

To overcome these limitations, PharmKG depends on recent versions of biological knowledge bases and includes a broader set of biomedical facts derived from the research literature to fulfill the missing semantics in the knowledge bases. To deal with synonymy and ambiguity, each entity in the PharmKG was allocated manually curated synonymous tables of chemicals, genes and diseases. More detailed cleaning processes are shown in Materials Section. Table 1 summarizes the specializations and the different types of covered biological entities of a set of popular biological knowledge bases.

KG embedding methods

The key issue with embedding KGs is learning to create a low-dimensional distributed representations of entities and relations. Once learned, representations can then be processed using various scoring functions to give probability scores for all triplets. As many survey articles [17, 27] have reviewed KGE approaches on general benchmarking settings, we provide only a brief description of several commonly used KGE methods that have been adopted as baselines in this work. Generally, these methods can be split into the following three categories.

Distance-based scoring function

The key idea of the translational distance-based models is that, for each triplet (h, r, t), the relation r is treated as a translation from head entity h to tail entity t, namely $h + r \simeq t$ in vector space. Bordes *et al.* [9] first proposed TransE by assuming that the added embedding of h and r should be close to the one of t. A drawback of TransE is that it struggles with N-to-1, 1-to-N and N-to-N structures. To address this issue, TransR [38] extends TransE by introducing separate latent spaces for entities and relations. These translational models are fast, require few parameters, but result in less expressive KGEs.

Semantic matching scoring function

The semantic matching models exploit similarity-based energy functions by matching latent semantics of entities and relations in the embedding spaces. RESCAL [11] was proposed based on the idea that entities are similar if connected to similar entities via similar relations. The similarity was calculated through a bilinear model by associating each relation r with a matrix M_r. Later, DistMult [12] was proposed by simplifying the bilinear formulation through using diagonal matrix M_r to model relation R. *complex* [39] further generalized DistMult by using complex embeddings and Hermitian dot products.

Neural network-based scoring function

Deep learning has been increasingly popular as these methods can outperform common machine learning methods. Recently, two convolutional neural network-based models have been proposed for relation prediction, namely ConvE [10] and ConvKB [40]. They both concatenate embeddings of entities and relations into a 2D feature map and use convolution operations to extract information. These models are parameter-efficient but learn each triplet independently without taking global relationships between the triplets into account. A graph-based neural network R-GCN [41] was introduced for learning connectivity structure under an encoder–decoder framework. It applies a graph convolution operation to the neighborhood of each entity and assigns them equal weights without considering heterogeneous information.

Though the abovementioned methods have impressive and expressive performances in general homogeneous KG datasets, they cannot capture the structured ontological hierarchies and heterogeneous features embedded in biomedical KGs. To this end, we have developed a novel heterogeneous graph attention neural network (HRGAT), which makes uses of not only the global information surrounding the target triplet but also the rich heterogeneous node attributes obtained from multi-omics resources. Detailed information about this model is presented in Modeling and application of PharmKG Section.

Table 2 summarizes the scoring functions of a set of popular KG embedding methods. In this study, we focus on embedding methods that operate on multi-relational graphs, as mentioned in the introduction of the paper. We did not include uni-relational KGE methods (i.e. DeepWalk [42], Node2Vec [43], etc.), and other methods specifically designed for a single dataset (Triple [24], Dragon [32], etc.).

Table 2. A summary of used KG embedding models

Category	Model	Scoring function
Transitional distance	TransE	$\|\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|\|_{L1/L2}$
	TransR	$-\|\|M_r\mathbf{h}+\mathbf{r}-M_r\mathbf{t}\|\|_2^2$
Semantic matching	Distmult	$\mathbf{h}^T\mathrm{diag}(M_r)\mathbf{t}$
	ComplEx	$\mathrm{Re}(<\mathbf{h},\mathbf{r},\mathbf{t}>)$
	RESCAL	$\mathbf{h}^T M_r\mathbf{t}$
Neural network	ConvE	$\sigma(W(\sigma([M_h,M_t])*\omega))\mathbf{t}$
	ConvKB	$concat(\sigma([\mathbf{h},\mathbf{r},\mathbf{t}])*\omega))w$
	RGCN	$\mathbf{h}^T\mathrm{diag}(M_r)\mathbf{t}$
	HRGAT	$concat(\sigma([\mathbf{h},\mathbf{r},\mathbf{t}])*\omega))w$

Materials

PharmKG was built from publicly available databases and text-mined knowledge bases and naturally combined the domain knowledge embedding of entities from different resources. In this section, we illustrate the details of our constructions and results. We also provide a detailed data analysis of our KG.

Construction of PharmKG

Integration of public knowledge bases

Our KG was constructed based on six public databases that offered high-quality structured information, including OMIM [19], DrugBank [21], PharmGKB [18], Therapeutic Target Database (TTD) [34], SIDER [44] and HumanNet [45]. These databases provided raw data files in various formats, including CSV, XML, TXT and TSV. The raw data files were parsed based on the data structure and then organized into structured interactions between entity pairs (e.g. gene–chemical interaction, chemical–disease interaction). To integrate these databases, we unified the gene name with the Entrez Gene ID as it was commonly used in the OMIM, DrugBank, PharmGKB and HumanNet. As the disease and chemical names are poorly standardized, we unified the names according to the Medical Subject Headings (MeSH) [46] and PubChem, respectively, and mapped names used in other databases accordingly. The unified entity names in our KG prevented the duplicate entities due to synonyms in different resources. It should be noted that the relation information extracted from these public resources is undirected and non-attributed. This preliminary network was referred to as the Interaction Network.

Combining Interaction Network with GNBR

To obtain an information-enriched KG, we further combined the Interaction Network with GNBR [36]. GNBR contains millions of noisy triplets extracted from large-scale biomedical literature with unsupervised techniques. We first performed entity disambiguation for GNBR with similar procedures described in the last section so that we can map millions of relationships obtained from GNBR to the Interaction Network. During the fusion process, the trivial entities that did not exist in the Interaction Network were removed to ensure the quality of the KG. We noticed that in most cases, relationships from GNBR could not be found in the interaction network obtained from public databases, which means that GNBR can largely enrich the coverage of public knowledge bases. The semantic themes in GNBR were inherited as the final relation types in the integrated KG. Herein, most of the interactions from public resources could be assigned directionality and attribute according to the theme of GNBR. We referred to this preliminary KG as the Raw PharmKG.

Entity filtering and heterogenous feature extraction

To obtain a high-quality benchmark, the Raw PharmKG was further polished by filtering trivial entities and attaching initial features for each entity using domain resources. Concretely, we selected 1497 FDA-approved drugs with molecular masses lower than 900 Da and extracted their chemical features, including extended-connectivity fingerprints [47] and physiochemical features as generated by Rdkit [48]. Then, we focused on the diseases above the fifth hierarchical levels in MeSH tree structure. Since most of the symptoms have only a few relationships, thousands of symptoms were merged into the corresponding fifth-level diseases according to the MeSH tree structure. We extracted disease semantic features from the biomedical language representations by applying pre-trained word embeddings obtained from BioBERT [29]. As for genes, we selected those expressive genes in BioGPS [49] and Connectivity Map [50] that are often studied by researchers, resulting in 4759 genes with their expression levels from different tissue cell types. The expression matrix constituted the feature embeddings of those genes. To eliminate the redundancy and reduce the dimensionality, we performed principal component analysis [51] on the features of the three types of entities and utilized the top 100 eigenvectors as their feature representation.

Finally, PharmKG contained a total of 29 types of 500 958 relationships between 7603 entities of 3 types.

Data analysis

As shown in Table 3, the 7603 entities consist of 1497 FDA-approved chemicals, 1347 complex diseases and 4759 genes. The 500 958 relationships were grouped into 29 types in 4 top-level categories including 'Interactions', 'Disease-Gene', 'Disease-Chemical' and 'Chemical-Gene' based on the entities they link. Figure 2 shows the semantic descriptions of all relationships and the percentage of each relationship in the categories. For example, the 'Ecg' relationship in the chemical-gene category occupies 2.28% (11, 421/500, 958) in PharmKG, which describes chemicals' effects on the gene expression level.

As shown in Figure 2B, each category has a variety of relationship types and the relationships within them are integrated from curated biomedical knowledge. The 'interaction' category made up 59.78%, the largest percentage of total relations since it involves many relationships between two identical entities. For example, gene–gene interaction has six different sub relations, including 'Q', 'Ra', 'Rg', 'Bgg', 'Egg' and 'GG', which conveys differences in meaning. The categories associated with chemical entities are relatively few, as only a small library of the more important and informative 1497 FDA-approved drug molecules were included. Note that the relation types in PharmKG are slightly different from the ones in GNBR as several new types were added, such as 'GG', 'CC' in the 'Interaction' category, and some semantically similar relationships were merged based on a clustering dendrogram to alleviate the low precision caused by unsupervised learning techniques. More detailed information of this PharmKG was shown in Tables S1 and S2.

Table 3. The type-wise distribution of the entities in PharmKG and their original data-source(s)

Type	Drugbank	TTD	OMIM	PharmGKB	GNBR	PharmKG
Chemical	1, 208	1, 347	–	615	1, 442	1, 497
Disease	–	399	987	419	1, 001	1, 347
Gene	1, 166	741	2, 320	1, 674	4716	4, 759

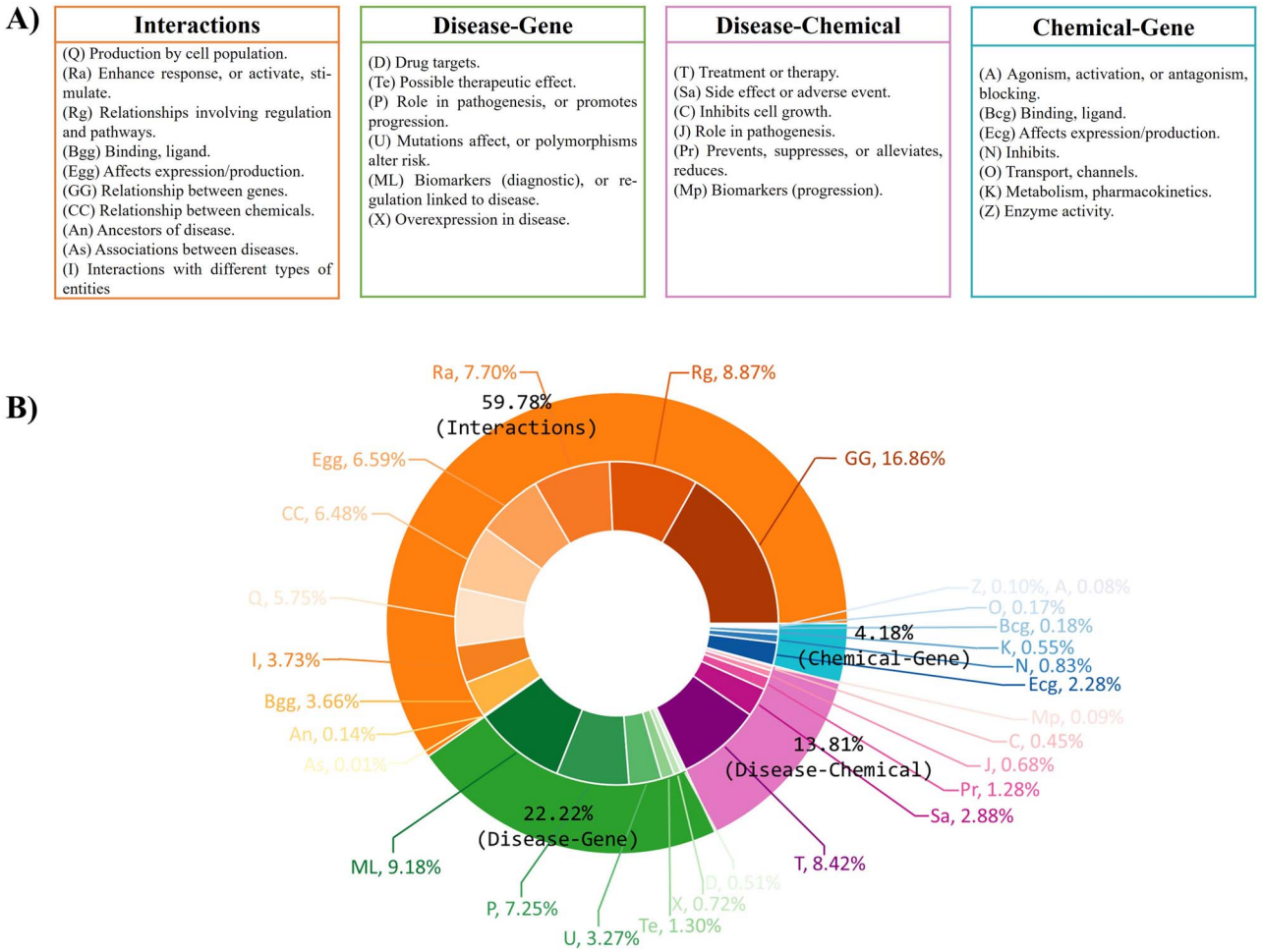

Figure 2. Summary of (A) relationship themes and (B) their distributions in PharmKG.

Modeling and application of PharmKG

In this section, we explore the difficulty of the PharmKG dataset for the relation prediction task and its usages for drug repurposing and target identification with several baselines and the proposed novel method.

Baselines

We used Pykeen v1 [52], an open-source Python package for KGE, including TransE, TransR, Distmult, ComplEx and RESCAL. We also implemented several neural network-based methods, including ConvE, ConvKB and RGCN. The details and limitations of these models have been introduced in Background Section.

Heterogeneous graph attention neural network

For our method, we propose a heterogeneous graph attention network (HRGAT) that obtains the optimally weighted combination of the biomedical triplet embedding by making efficient use of information shared across regions in the graph. As shown in Figure 3, the architecture of the model is built as an extension of the GAT [53] with attentive knowledge embedding [54] for the biomedical KG. More specifically, it allows the encoder–decoder paradigm to be trained in an end-to-end fashion. We formulated the encoding and decoding part of the HRGAT in the following sections.

Encoder

The encoder layers are input with two feature matrices. The first one is the entity feature matrix $H \in \mathbb{R}^{N_e \times T}$, where N_e is the total number of entities and T is the feature dimension of each entity embedding. Compared with the random homogenous initial embedding adopted in previous works [23, 24, 37], we used heterogeneous feature embedding obtained from the entity itself, as introduced in Construction of PharmKG Section. The second represents the embeddings of relations and is represented by the matrix $G \in \mathbb{R}^{N_r \times Q}$, where N_r is the total number of relations and

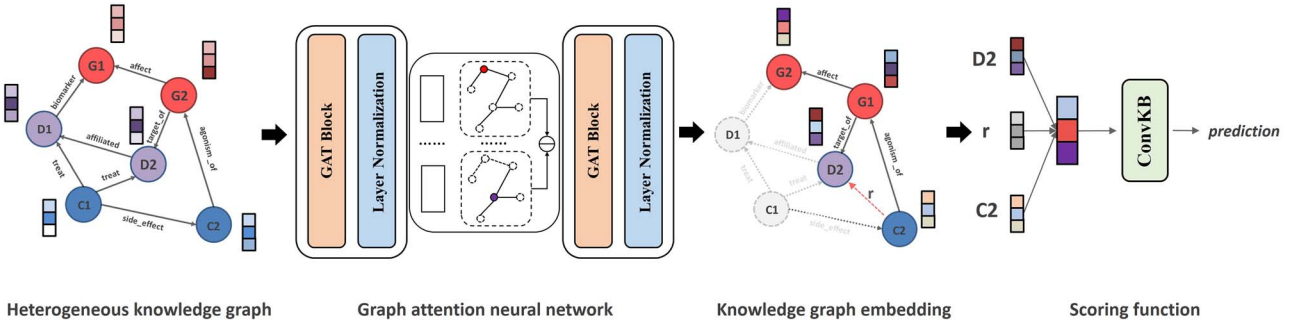

Figure 3. The overall framework of HRGAT. The model first captures the global network structure and heterogeneous domain features by several graph attention neural network blocks and then scores the triple with ConvKB.

Q is the feature dimension of each relation embedding. We used TransE to initialize the relation embeddings based on previous work [54].

This task aims to obtain the new embeddings of each entity e_i and relation r_k. The embedding for a triplet $t_{ijk} = (e_i, r_k, e_j)$ was learned as

$$h_{ijk} = \mathbf{W_1}\left[h_i \parallel g_k \parallel h_j\right],$$

where g_k, h_i and h_j are the initial feature embeddings of the relation $r_k$and entities e_i and e_j, respectively, and $\mathbf{W_1}$ is the linear transformation matrix. Following [53], the normalized attention value of each triplet was learned by

$$c_{ijk} = \mathrm{Relu}\left(\mathbf{W_2} h_{ijk}\right)$$

$$a_{ijk} = \mathrm{softmax}_{jk}\left(c_{ijk}\right) = \frac{\exp\left(c_{ijk}\right)}{\sum_{n\in N_i}\sum_{r\in R_{in}}\exp\left(c_{inr}\right)},$$

where $\mathbf{W_2}$ represents the corresponding linear transformation, N_i denotes the neighbors set of entity e_i and R_{in} is the relations set linking entity e_i and e_n. Attention weights for each adjacent triplet are the importance for a source entity e_i. The updated embedding of the entity e_i is the sum of each adjacent triplet representation weighted by their attention values as shown in

$$h_i' = \sigma\left(\sum_{j\in N_i}\sum_{k\in R_{ij}} a_{ijk} h_{ijk}\right),$$

where $\sigma$is the sigmoid activation function and $a_{ijk}$is the normalized attention coefficient of triplet t_{ijk}.

In addition, multi-head attention was introduced to stabilize the learning process and to encapsulate more information about the neighborhood [53]. To alleviate the computing cost, the output embedding in the final layer is calculated using averaging instead of a concatenation operation by

$$\hat{h}_i' = \sigma\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{j\in N_i}\sum_{k\in R_{ij}} a_{ijk}^m h_{ijk}^m\right),$$

where M is the number of attention head.

For the update of relation embedding G, we also performed a linear transformation with a weight matrix W_3

$$\hat{G} = \mathbf{W_3} G.$$

Finally, to avoid the loss of initial biological features, a widely used shortcut strategy was employed [55] by adding initial entity embedding $H$to the output hidden entity embedding H^f as shown in

$$\hat{H} = \mathbf{W_4} H^f + H.$$

Training objective

Following an idea from Bordes *et al.* [9], we optimized hinge loss by a similar translational scoring function as

$$\mathrm{Loss}_{\mathrm{encoder}} = \sum_{t_{ij}\in T}\sum_{t_{ij}'\in T'} \max\left\{d_{t_{ij}'} - d_{t_{ij}} + \gamma, 0\right\},$$

where $\gamma > 0$ is a margin hyperparameter, T and T' are the sets of valid and invalid triplets, respectively, and $d_{t_{ij}} = |h_i + g_k - h_j|$ is the distance for the triplet $t_{ijk} = (e_i, r_k, e_j)$ with relation r_k considered as a translation from head entity e_i to tail entity e_j, namely $e_i + r_k = e_j$ in embedding space.

Decoder

To extract the latent features inside the triplets and to analyze the global embedding properties of a triplet across each dimension, ConvKB [40] was used as a decoder. The scoring function with multiple feature maps can be written formally as

$$f\left(t_{ij}^k\right) = concat\left(\sigma\left(\left[h_i, g_k, h_j\right]) * \omega\right)\right)\mathbf{W},$$

where ω represents convolutional filter, '$*$' is a convolution operator and $\mathbf{W}$ is a transformation matrix used to calculate the final score for a given triplet t_{ij}^k.

The model is trained using soft-margin loss as

$$\mathrm{Loss}_{\mathrm{decoder}} = \sum_{t_{ij}^k\in T\cup T'} \log\left(1 + \exp\left(v_{t_{ij}^k} . f\left(t_{ij}^k\right)\right)\right) + \frac{1}{2}\|\mathbf{W}\|_2^2,$$

$$\text{where } v_{t_{ij}^k} = \begin{cases} 1, t_{ij}^k \in T \\ -1, t_{ij}^k \in T' \end{cases}.$$

Evaluation protocols

The most impressive ability of KGs is their ability to deduce new relations between biomedical entity pairs. For evaluation, we used ranking procedures as suggested by the KG community [9, 12]. For each test triplet, the head h_i is removed and replaced

by every other entity $e_i' \in E \backslash h_i$ in turn. We first computed a score for each triplet and then sorted these scores in ascending order to get the rank of the correct triplet (*h, r, t*). We report the mean reciprocal rank (MRR) and the proportion of correct entities in the top N ranks (Hits@N) for $N = 1, 3, 10$ and 100. One-sample *t*-tests were implemented to compare the HRGAT with the strongest baseline and *P*-values < 0.05 indicate that the improvements of HRGAT over the strongest baseline were statistically significant.

These ranking metrics proved to be suitable for the general evaluation of standard KGE methods but proved to be deficient in evaluating incomplete KGs, where so-called 'corrupt triplets' were more likely to be valid. For downstream tasks such as drug repurposing and target identification, we also introduced two widely used metrics: the areas under the receiver operating characteristic (AUROC) curve and the precision-recall curve (AUPRC).

We divided the triplets into a training set, a validation set and a test set in an 8:1:1 manner. For KGE baselines, we set the dimensionality of the initial embedding to 100. All the baselines were trained for 100–1000 epochs using margin ranking loss with learning rate [0.005, 0.01, 0.1] and batch size [256, 512, 1024]. Other hyper-parameters for each approach were set at their default settings, as recommended by the Pykeen package [52]. We implemented rough grid search for parameter optimization as we found that the model performance was not sensitive to reasonable settings.

In order to provide a comprehensive comparison of these baselines and the new method, we further made evaluations on Hetionet, a manually curated biomedical KG dataset [37]. The dataset originally contained 47 031 nodes of 11 types and 2 250 197 relationships of 24 types. To keep the data consistent, we kept only the gene, chemical and disease entities and the interconnections between them. The pruned Hetionet included 22 634 nodes of 3 types and 562 106 relationships of 13 types. Note that one portion of the relation types in Hetionet is naturally directed and asymmetric, such as Compound–treats–Disease (CtD) and Compound–downregulates–Gene (CdG), while several other relations are undirected such as Gene–interacts–Gene (GiG) and Gene–covaries–Gene (GcG). To match the paradigm of commonly used KG embedding methods, the undirected relations were treated as directed ones according to the original order of two entities in the Hetionet list. The statistical information of this dataset was shown in Figures S2B and S3.

Evaluation of KGs with different embedding methods

We tested different embedding methods on PharmKG and the pruned Hetionet. As shown in Table 4, the HRGAT model outperforms all other models in terms of Hit@N and MRR on two benchmarking datasets. Generally, compared with traditional techniques (e.g. Distmult, RESCAL and RGCN), the proposed embedding methods have largely improved the relation reasoning performance. For example, ConvKB achieves a 31.2% improvement in terms of Hit@100 value compared with RGCN. ComplEx obtains a 9.1% increment in the Hit@100 when compared with RESCAL. Considering the heterogeneous representation and multi-hop neighborhood features, our HRGAT method outperforms the best neural network baseline ConvKB by 4.8% (MRR), 2.3% (Hit@1), 10.6% (Hit@10) and 10.1% (Hit@100). It also improves over ComplEx with 4.7% on MRR (*P*-value = 2.42e−7) and 9.7% on Hit@100 (*P*-value = 4.19e−5). These results demonstrate that HRGAT is more effective and could be used on biological relation prediction tasks to improve prediction performance. Furthermore, we conducted an ablation study by omitting heterogeneous features. The ablated model (HRGAT-w/o) was found to cause a significant drop in the results by decreasing the MRR and Hit 100 with 0.016 (*P*-value = 1.84e−5) and 6.3% (*P*-value = 2.36e−4), respectively. These significant decreases suggest that heterogeneous features play a pivotal role in relation prediction.

A similar trend seen in evaluating the test Heitonet suggests that positive results derive from the method, and not from the dataset: our HRGAT-w/o is 4.0% better than the ComplEx (*P*-value = 2.34e−8) and 1.6% better than ConvKB (*P*-value = 9.13e−6) on MRR. HRGAT was not shown as the Hetionet does not have heterogeneous features for the structure of chemical entities and expression information of gene entities.

We further found that the results of the embedding algorithms in PharmKG were generally higher than those of Hetionet. Compared with Hetionet, PharmKG has 10-fold more hierarchical disease entities and disease-related edge types that can form more relations between genes and drugs, which facilitate the modeling of the KG. Additionally, the distribution of different associations in Hetionet is extremely unbalanced, where gene–gene interactions make up more than 85% of the total links, while disease–chemical interactions occupy less than 0.3%.

Evaluation of PharmKG's capacity for drug repurposing and target identification

Drug repositioning and target identification are the two most widely used applications of biological networks. To assess the ability of our model to carry out these two tasks, we calculated areas under receiver operating characteristic (ROC) and precision-recall (PR) curves for the task-related relationships based on the predicted scores given in Evaluation protocols Section. We retrieved the most drug repurposing-relevant types, 'C' (Cell inhibits), 'T' (Treatment) and 'J' (Role in pathogenesis) in the Drug Disease themes and the most target identification-relevant types, 'Te' (Possible therapeutic effect), 'D' (Drug targets), 'X' (Overexpression in disease) and 'ML' (Biomarkers) in the Disease Gene themes. As shown in Figure 4, our HRGAT model performs well in discriminating positive and negative pairs in drug repurposing tasks, achieving an AUROC of 0.912 and an AUPR of 0.911, significantly outperforming that of ConvKB (AUROC = 0.807, AUPR = 0.813), TransE (AUROC = 0.788, AUPR = 0.774) and ComplEx (AUROC = 0.794, AUPR = 0.781). The ablated model (HRGAT-w/o) was found to cause a significant drop in the results by decreasing the AUC and AUPR by 2.7 and 1.9%, respectively. Similarly, we found that HRGAT achieved the best performance in target identification tasks. The superior performances of HRGAT likely result from the full use of global information available in KGs and domain-knowledge-associated embedding. Therefore, the implementation of network embedding on such heterogeneous frameworks effectively integrates chemical, genomic, pharmacological and phenotypic information and, hence, is useful for providing accurate drug repositioning predictions and provides new insights into target identifications.

Pharmacological interpretation of HRGAT

It is interesting to know whether HRGAT can capture the biomedical semantics embedded in relationships. To this aim, we took the initial embedding h_{ijk} and final embedding $\hat{h}_{ijk}$ of triplets in different relations and embed these hidden vector

Table 4. Overall relation prediction performance on the two compiled biomedical KGs

Category	Model	PharmKG					HetioNet				
			Hits@N					Hits@N			
		MRR	$N=1$	$N=3$	$N=10$	$N=100$	MRR	$N=1$	$N=3$	$N=10$	$N=100$
Distance-based semantic matching	TransE	0.091	0.034	0.092	0.198	0.524	0.027	0.002	0.022	0.070	0.300
	TransR	0.075	0.030	0.071	0.155	0.510	0.040	0.013	0.036	0.088	0.317
	RESCAL	0.064	0.023	0.057	0.122	0.413	0.032	0.017	0.031	0.059	0.231
	ComplEx	0.107	0.046	0.110	0.225	0.552	0.070	0.029	0.069	0.148	0.382
	Distmult	0.063	0.024	0.058	0.133	0.461	0.037	0.012	0.034	0.087	0.286
Neural network	ConvE	0.086	0.038	0.087	0.169	0.425	0.075	0.032	0.071	0.155	0.408
	ConvKB	0.106	0.052	0.107	0.209	0.548	0.094	0.045	0.090	0.186	0.442
	RGCN	0.067	0.027	0.062	0.139	0.236	0.030	0.011	0.021	0.052	0.209
Proposed	HRGAT–w/o	0.138	0.068	0.148	0.275	0.586	**0.110**	**0.055**	**0.105**	**0.210**	**0.483**
	HRGAT	**0.154**	**0.075**	**0.172**	**0.315**	**0.649**	–	–	–	–	–

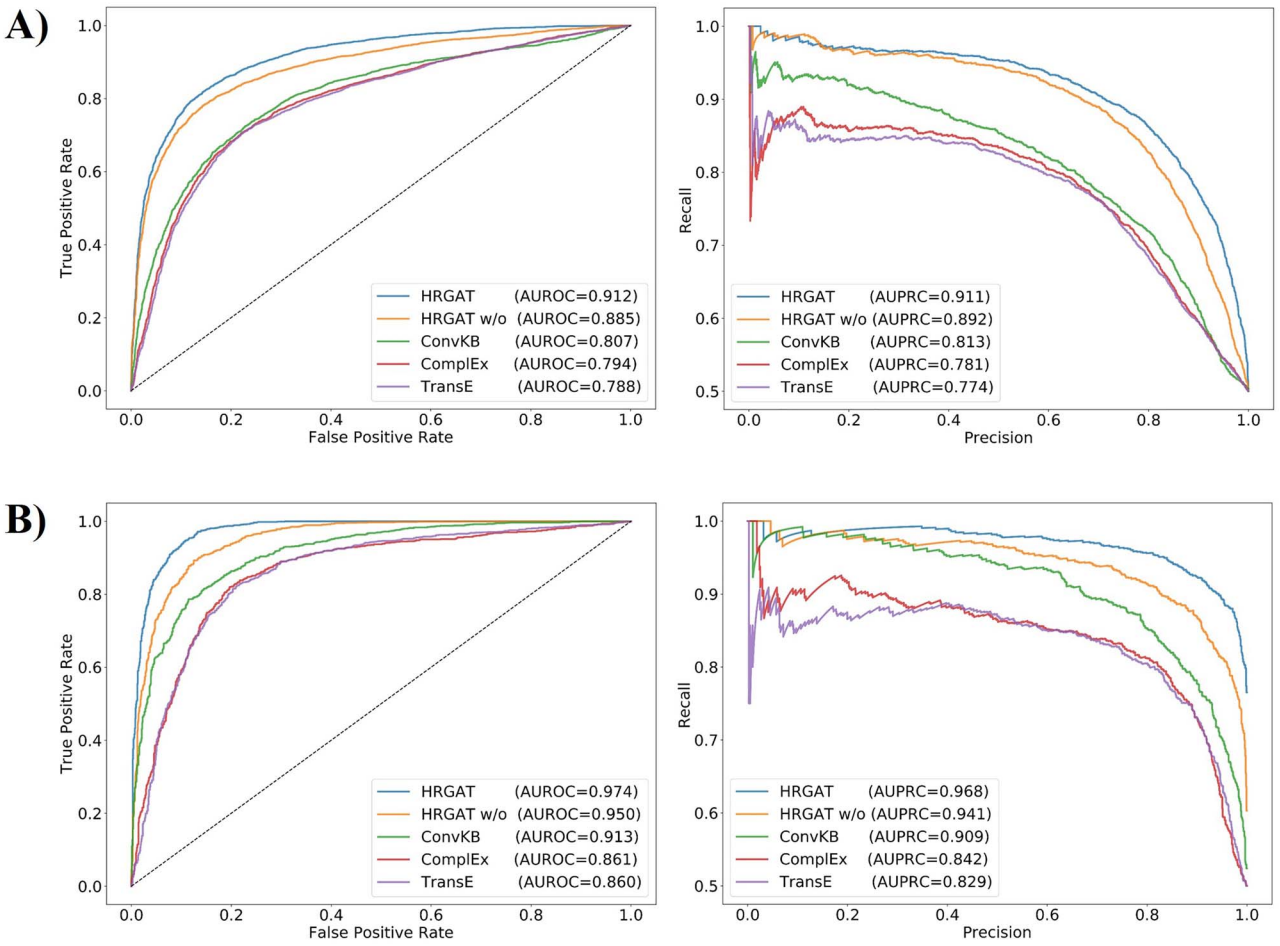

Figure 4. A summary of the results of an evaluation of the predictive accuracy of KG embedding models compared with other models on two biological inference tasks: drug repurposing (A) and target identification (B). The reported results represent the score percentage of the area under the ROC and PR curves for the left and right side bars, respectively.

representation into a 2D space using t-SNE [56]. As shown in Figure 5, the initial feature of triplets in different relationships is disentangled because of disjointed random embeddings of the relation r, while relations in the same theme cluster together after the global structure information and heterogeneous features embedded in the biomedical KG are learned. The results demonstrate that our model is capable of learning biomedical semantic information embedded in such relationships.

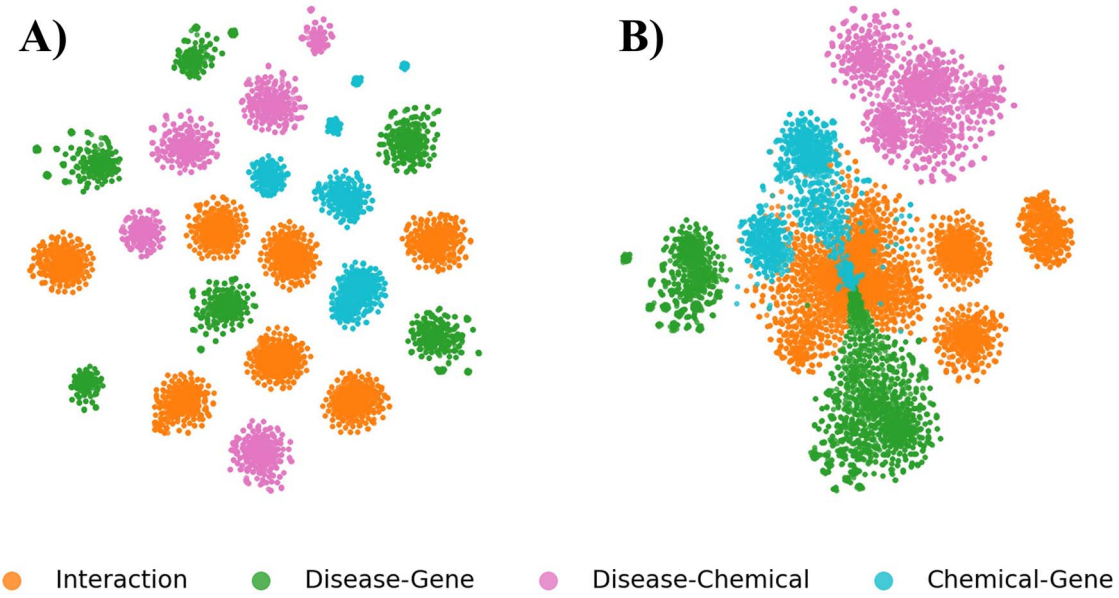

Figure 5. Visualization of the triples in different themes in the 2D space using the t-SNE package for the (A) initial triplet embedding and (B) final embedding following learning by HRAGT.

Downstream applications

To further validate the efficacies of our model, we conducted case studies to infer novel drug repurposing and target identification candidates for two types of neurodegenerative diseases, Alzheimer's disease (AD) and Parkinson's disease (PD). AD and PD are the most common neurodegenerative diseases in the world, and both are currently without a cure; PharmKG may help to propel mechanistic exploration and the identification of novel drug candidates [57–60]. We performed a detailed survey from various evidence such as literature evidence and validations of the prediction using three categories: (i) published evidence, where there is literature evidence indicating the use of the drug or target to influence human disease; (ii) potential associations, where the drug or target may produce different physiological effects than those expected and (iii) unknown/no effect, where there is no literature evidence to support the drug-relation-disease or target-relation-disease combination. We focused on predicted relationships that do not exist in our training, validation and test sets.

Case study of drug repositioning: computationally identifying approved drugs for AD and PD

In accordance with the learned model, we selected the top-scored candidates to evaluate the validity of the prediction. Supplementary Table S3 shows the top 10 highest-scoring novel drug repurposing candidates for AD with the canonical name of the drug, predicted relation, disease name, predicted score, evidence category and PMID (literature reference supporting interpretation). In total, among the top 10 predicted drug candidates, four drugs (40%) are validated for treating AD by literature evidence and five candidates (50%) have a potential relationship with AD. For example, enalapril is a drug used to reduce high blood pressure and to prevent or treat heart failure [61]. This prediction was supported by a previous study indicating that enalapril pretreatment could be used as a therapeutic approach for Alzheimer's patients [62]. Likewise, imatinib, marketed as Gleevec, has been proven effective for gastrointestinal stromal tumors [63] and might also be considered and may have the basis to be a potential novel therapy for AD [64]. Desipramine and isoprenaline were predicted to play a role in pathogenesis, which were also supported by literature evidence [65–67]. Trifluoperazine, while predicted to be linked, has not been previously reported to be associated with AD. We noticed that predicted relations might be inconsistent with actual ones. For instance, etoposide was predicted to be associated by inhibiting cell growth but recent studies show its ability to induce cellular senescence that may have negative implications in brain aging and neurodegenerative conditions [68].

Table S4 lists the top 10 drug repurposing candidates for PD, among which four candidates (40% success rate) were validated by various evidence from literature and another four candidates (40%) were proved to be associated. For example, atomoxetine is a medication approved for the treatment of attention deficit hyperactivity disorder (ADHD) [69]. Here, atomoxetine is the top predicted candidate for repurposing to treat PD, a potential that is supported by previous studies [70–72]. Methylphenidate, a stimulant medication also used to treat ADHD along with narcolepsy, was predicted by our model to be a potential treatment for PD, which is evidenced in the literature [73–75]. Two predicted candidates (everolimus and neostigmine) have not been reported to associate with PD.

Case study of target identification: computationally identified druggable targets for AD and PD

Table S5 lists the top 10 highest-scoring candidate targets for AD. We found that 8 of 10 candidates (80% success rate) were found to be associated with AD and supported by evidence from the literature. For instance, the top predicted target, FASLG (Fas ligand), is a protein belonging to the tumor necrosis factor (TNF) family. FASLG was found to be associated with neurotic degeneration in the AD brain and to participate in β-amyloid-induced neuronal death [76]. Likewise, CYP2E1, predicted by our model

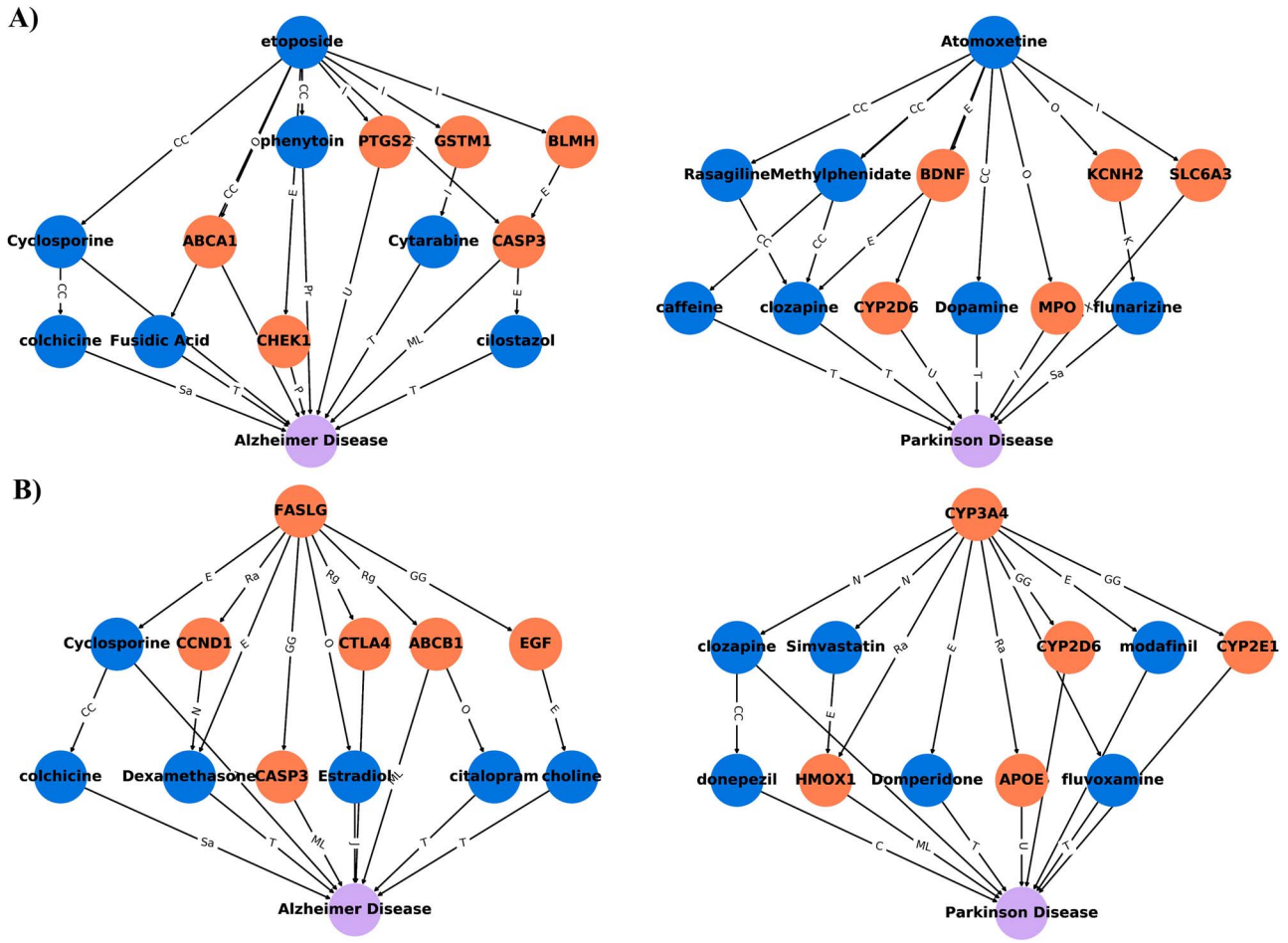

Figure 6. Mechanism analysis of four top-scoring triplets for (A) drug repurposing and (B) target identification in case study sections, respectively. Each figure shows the 10 most supportive paths in the KG.

to be involved in increased disease risk, is a broadly expressed enzyme involved in the metabolism of xenobiotics in the brain and liver and has also been found to generate reactive oxygen species, capable of contributing to many diseases, including AD [77]. Furthermore, SERPINC1, a protein encoded by the gene antithrombin III, was predicted by our model as a target for AD, with its involvement supported by previous studies [78].

For PD (Table S6), seven targets (70% hit rate) were validated by various evidence. For example, our model found that mutations in PSEN1 and PSEN2 may be risk factors for PD, a theory that is supported by several studies [79, 80]. Furthermore, Sirtuin 2 is an enzyme encoded by the SIRT2 gene; inhibition of SIRT2 has been shown to be protective in PD [81, 82]. Our model suggests that improper regulation of the Sirtuin 2 gene may be responsible for its links to the disease. In addition, TNFRSF1A, a receptor that binds tumor necrosis factor-alpha (TNF-α), was predicted by our model to be a target implicated in PD pathogenesis, a prediction supported by evidence in the literature [83–85].

In summary, PharmKG offers a useful tool to identify potential drugs for repurposing and to suggest novel targets for diseases, such as in AD and PD.

KG visualization of top-predicted candidates

Compared with a basic network, another advantage of KGs is their interpretability. To exemplify this, we selected four top-predicted drug repositioning and target identification candidates to analyze their mechanism in PharmKG. Figure 6 shows the top paths supporting the generated hypotheses. By depicting the 10 shortest paths in the KG, we can see that our model is learning information from neighbor nodes when inferring new relationships. For instance, the CYP2D6 gene, a member of the cytochrome P450 gene family, is believed to associate with PD because of its highly polymorphic expression [86, 87], and atomoxetine is traditionally believed to be metabolized through the cytochrome P450 2D6 (CYP2D6) enzyme pathway in its treatment of ADHD [88]. Therefore, our method hypothesizes that atomoxetine is a potential drug for PD, as shown in Figure 6. This suggests the ability of our method to capture heterogeneous information in the KG and repurpose it to infer potential therapeutic drugs and novel targets for diseases.

Discussion and conclusion

In this paper, a novel biomedical KG, PharmKG, was presented. We provided a multi-relational attributed dataset containing over 500 000 interconnections between gene, drug and disease, including 29 relation types annotated with a vocabulary of ~8000 high-quality entities. We demonstrated the wide variety of biomedical information embedded in our dataset. We have also introduced a novel neural network-based embedding approach,

HRGAT, to alleviate the drawbacks of existing methods and to optimize the capture of heterogeneous information embedded in biomedical KGs. We demonstrated the high performance of HRGAT on PharmKG and Hetionet and highlight its ability to make drug repurposing predictions and to generate new potential drug targets. We further validate HRGAT's ability to identify valid connections by finding supporting evidence using independent sources and by identifying the meta-pathways in the original KG that help explain the prediction results.

Compared with recently published works that either used basic networks [8, 30] or existing biomedical knowledge bases [37, 89] to generate hypotheses, PharmKG integrated both the curated knowledge bases and semantic triplets derived from the abstracts of vast biomedical literature. It is a semantically enriched KG benchmark that can be applied in not only drug repurposing and target identification but also protein–protein interaction prediction, adverse drug reaction analysis, etc. Various hypotheses can be generated by targeting different relational semantics embedded in PharmKG. Moreover, it can be used as a high-quality training corpus to perform relation extraction tasks. By integration with the disease-specific context, PharmKG can be further utilized to accelerate drug repurposing and mechanism analysis for emerging human diseases, such as COVID-19 [90]. We believe that the dataset and benchmark described in this work constitute an important step for biomedical KG construction, modeling and application. Based on PharmKG, a wide range of current and upcoming relation prediction models can be evaluated and utilized for biomedical cases. Pharmaceutical scientists and biologists could also benefit from this by analyzing the positions of interested entities and top-scored hypotheses.

There are several potential limitations of PharmKG under the current deep learning framework. First of all, although we made efforts to assemble large-scale, literally reported interactions from a number of publicly available databases and literature, the quality and integrity of the metadata cannot be fully assured. For example, due to the inherent lack of negative drug-disease pairs in the publicly available databases and published literature, it is challenging to obtain truly negative samples in our KG. Secondly, the final version of PharmKG does not have large-scale of entities compared with previous KGs [23, 37]. We noticed that there are 180 000 entities with 1 million relations in the raw version that kept trivial entities like less studied compounds and genes (we also released the raw version in the Github). However, over 90% vertices contain less than two links in this version. By such KG, the assessment of models might be biased due to the large number of missing relations between the less studied entities. As this study emphasized on building a high-quality benchmark to assess KG models, we have made a trade-off between the quality and the scale of entities. It will be a long-term effort for us to increase the scale while maintaining the quality of PharmKG. In the future, we will further expand PharmKG by merging multi-omics sources, i.e. transcriptomics and clinical data, and explore novel automated relation extraction methods. In addition, we would like to exploit the inductive-based knowledge embedding method, alleviating the problems caused by the addition of new information.

Availability and implementation

The datasets and code are available at available on https://github.com/MindRank-Biotech/PharmKG.

Key Points

- We established a dedicated, high-quality and highly challenging benchmark optimized for the task of evaluating multi-relation pre-diction methods in large attributed biomedical knowledge graphs (KGs). By integrating six representative public resources and text-mined knowledge bases, this biomedical-attributed KG, PharmKG, contains thousands of nodes of gene, chemical com-pound and disease, connected by a set of semantic relationships derived from the abstracts of biomedical literature. Each entity in the PharmKG was labeled with domain-specific information, persevering the semantic and biomedical features.
- A novel biological intuitive graph neural network-based KGE method is introduced as a new baseline to alleviate the drawbacks of existing methods and to capture the heterogeneous information embedded in the biomedical knowledge.
- We conduct extensive experiments on the PharmKG with various KGE models using various evaluation metrics. We discuss our observations across various down-stream biological tasks to provide insights and guidelines for how to use the KG in the biomedical area.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

We thank Brian C. Gilmour (NO-Age, University of Oslo) for language editing and scientific inputs of the manuscript.

Funding

Innovative Medicines Initiative Program – IMI2-RIA (#101005122); National Natural Science Foundation of China (#62041209, #61772566, #81971327); Helse Sør-Øst (#2017056 to E.F.F.); Research Council of Norway (#262175, #277813 to E.F.F.); Akershus University Hospital Strategic grant (#269901 to E.F.F.); Rosa Sløyfe grant (#207819 to E.F.F.) from the Norwegian Cancer Society.

Conflict of Interest

E.F.F. has a CRADA arrangement with ChromaDex and is a consultant to the Aladdin Healthcare Technologies and the Vancouver Dementia Prevention Centre.

References

1. Li J, Zheng S, Chen B, *et al*. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;**17**:2–12.

2. Abdelaziz I, Fokoue A, Hassanzadeh O, *et al*. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. *J Web Semant* 2017;**44**:104–17.
3. Gavin A-C, Bösche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;**415**:141–7.
4. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
5. Cohen JD, Servan-Schreiber D. Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol Rev* 1992;**99**:45.
6. Janjić V, Pržulj N. Biological function through network topology: a survey of the human diseasome. *Brief Funct Genomics* 2012;**11**:522–32.
7. Luo Y, Zhao X, Zhou J, *et al*. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat Commun* 2017;**8**:1–13.
8. Zhou Y, Hou Y, Shen J, *et al*. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 2020;**6**:14.
9. Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2. Lake Tahoe, Nevada, 2013, p. 2787–2795. Curran Associates Inc.
10. Dettmers T, Minervini P, Stenetorp P, *et al*. Convolutional 2d knowledge graph embeddings. In: Thirty-Second AAAI Conference on. *Artificial Intelligence* 2018.
11. Nickel M, Tresp V, Kriegel H-P. *A three-way model for collective learning on multi-relational data*. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. Bellevue, Washington, USA, 2011, p. 809–816. Omnipress.
12. Yang B, Yih W-T, He X, *et al*. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:14126575. 2014.
13. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;**290**:2323–6.
14. Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia, 2015, p. 891–900. Association for Computing Machinery.
15. Ou M, Cui P, Pei J, *et al*. Asymmetric transitivity preserving graph embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2016, p. 1105–1114. Association for Computing Machinery.
16. Alshahrani M, Khan MA, Maddouri O, *et al*. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* 2017;**33**:2723–30.
17. Su C, Tong J, Zhu Y, *et al*. Network embedding in biomedical data science. *Brief Bioinform* 2020;**21**:182–97.
18. Hewett M, Oliver DE, Rubin DL, *et al*. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res* 2002;**30**: 163–5.
19. Hamosh A, Scott AF, Amberger J, *et al*. Online Mendelian inheritance in man (OMIM). *Hum Mutat* 2000;**15**: 57–61.
20. Davis AP, Grondin CJ, Johnson RJ, *et al*. The comparative Toxicogenomics database: update 2019. *Nucleic Acids Res* 2019;**47**:D948–54.
21. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
22. Belleau F, Nolin M-A, Tourigny N, *et al*. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;**41**:706–16.
23. Breit A, Ott S, Agibetov A, *et al*. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics* 2020;**36**:4097–4098.
24. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 2020;**36**:603–10.
25. Wang Q, Mao Z, Wang B, *et al*. Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans Knowl Data Eng* 2017;**29**:2724–43.
26. Wang Z, Zhang J, Feng J, *et al*. *Knowledge graph embedding by translating on hyperplanes*. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, Québec, Canada, 2014, p. 1112–1119. AAAI Press.
27. Nickel M, Murphy K, Tresp V, *et al*. A review of relational machine learning for knowledge graphs. *Proc IEEE* 2015;**104**:11–33.
28. Consortium GO. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;**47**:D330–8.
29. Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.
30. Zeng X, Zhu S, Liu X, *et al*. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019;**35**:5191–8.
31. Sosa DN, Derry A, Guo M, *et al*. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *bioRxiv* 2019;727925.
32. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018;**34**:i457–66.
33. Xiao C, Zhang P, Chaovalitwongse WA, *et al*. Adverse drug reaction prediction with symbolic latent dirichlet allocation.. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA, 2017, p. 1590–1596. AAAI Press.3
34. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res* 2002;**30**:412–5.
35. Zhu Y, Elemento O, Pathak J, *et al*. Drug knowledge bases and their applications in biomedical informatics research. *Brief Bioinform* 2019;**20**:1308–21.
36. Percha B, Altman RB. A global network of biomedical relationships derived from text. *Bioinformatics* 2018;**34**:2614–24.
37. Himmelstein DS, Lizee A, Hessler C, *et al*. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 2017;**6**:e26726.
38. Lin Y, Liu Z, Sun M, *et al*. Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas, 2015, p. 2181–2187. AAAI Press.
39. Trouillon T, Welbl J, Riedel S, *et al*. Complex embeddings for simple link prediction. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning – Volume 48. New York, NY, USA, 2016, p. 2071–2080. JMLR.org.
40. Nguyen DQ, Nguyen TD, Nguyen DQ, *et al*. A novel embedding model for knowledge base completion based on convolutional neural network. arXiv preprint arXiv:171202121. 2017.

41. Schlichtkrull M, Kipf TN, Bloem P, *et al*. *Modeling relational data with graph convolutional networks*. European Semantic Web Conference: Springer, 2018.
42. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, New York, USA, 2014, p. 701–710. Association for Computing Machinery.
43. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2016, p. 855–864. Association for Computing Machinery.
44. Kuhn M, Letunic I, Jensen LJ, *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–9.
45. Hwang S, Kim CY, Yang S, *et al*. HumanNet v2: human gene networks for disease research. *Nucleic Acids Res* 2019;**47**:D573–80.
46. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc* 2000;**88**:265.
47. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.
48. Landrum G. *RDKit: Open-source cheminformatics*, 2006.
49. Wu C, Orozco C, Boyer J, *et al*. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009;**10**:1–8.
50. Lamb J, Crawford ED, Peck D, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–35.
51. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intel Lab Syst* 1987;**2**:37–52.
52. Ali M, Hoyt CT, Domingo-Fernández D, *et al*. BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics* 2019;**35**:3538–40.
53. Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. arXiv preprint arXiv:171010903. 2017.
54. Nathani D, Chauhan J, Sharma C, *et al*. Learning attention-based embeddings for relation prediction in knowledge graphs. arXiv preprint arXiv:190601195. 2019.
55. He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, 2016, p. 855–864. Association for Computing Machinery.
56. Lvd M, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
57. Aman Y, Frank J, Lautrup SH, *et al*. The NAD(+)-mitophagy axis in healthy longevity and in artificial intelligence-based clinical applications. *Mech Ageing Dev* 2020;**185**: 111194.
58. Fang EF, Hou Y, Palikaras K, *et al*. Mitophagy inhibits amyloid-beta and tau pathology and reverses cognitive deficits in models of Alzheimer's disease. *Nat Neurosci* 2019;**22**: 401–12.
59. Gilmour BC, Gudmundsrud R, Frank J, *et al*. Targeting NAD(+) in translational research to relieve diseases and conditions of metabolic stress and ageing. *Mech Ageing Dev* 2020;**186**:111208.
60. Lautrup S, Sinclair DA, Mattson MP, *et al*. NAD(+) in brain aging and neurodegenerative disorders. *Cell Metab* 2019;**30**:630–55.
61. Todd PA, Heel RC. Enalapril. *Drugs* 1986;**31**:198–248.
62. Meamar R, Dehghani L, Ghasemi M, *et al*. Enalapril protects endothelial cells against induced apoptosis in Alzheimer's disease. *J Res Med Sci* 2013;**18**:S1.
63. Druker BJ. STI571 (Gleevec™) as a paradigm for cancer therapy. *Trends Mol Med* 2002;**8**:S14–8.
64. Eisele YS, Baumann M, Klebl B, *et al*. Gleevec increases levels of the amyloid precursor protein intracellular domain and of the amyloid-β–degrading enzyme neprilysin. *Mol Biol Cell* 2007;**18**:3591–600.
65. Wang D-D, Li J, Yu L-P, *et al*. Desipramine improves depression-like behavior and working memory by up-regulating p-CREB in Alzheimer's disease associated mice. *J Integr Neurosci* 2016;**15**:247–60.
66. Ohm TG, Bohl J, Lemmer B. Reduced basal and stimulated (isoprenaline, Gpp (NH) p, forskolin) adenylate cyclase activity in Alzheimer's disease correlated with histopathological changes. *Brain Res* 1991;**540**:229–36.
67. Cowburn RF, Vestling M, Fowler CJ, *et al*. Disrupted β1-adrenoceptor—G protein coupling in the temporal cortex of patients with Alzheimer's disease. *Neurosci Lett* 1993;**155**:163–6.
68. Bang M, Do Gyeong Kim ELG, Kwon KJ, *et al*. Etoposide induces mitochondrial dysfunction and cellular senescence in primary cultured rat astrocytes. *Biomol Ther (Seoul)* 2019;**27**:530.
69. Garnock-Jones KP, Keating GM. Atomoxetine. *Pediatric Drugs* 2009;**11**:203–26.
70. Weintraub D, Mavandadi S, Mamikonyan E, *et al*. Atomoxetine for depression and other neuropsychiatric symptoms in Parkinson disease. *Neurology* 2010;**75**:448–55.
71. Jankovic J. Atomoxetine for freezing of gait in Parkinson disease. *J Neurol Sci* 2009;**284**:177–8.
72. Marsh L, Biglan K, Gerstenhaber M, *et al*. Atomoxetine for the treatment of executive dysfunction in Parkinson's disease: a pilot open-label study. *Mov Disord* 2009;**24**: 277–82.
73. Chatterjee A, Fahn S. Methylphenidate treats apathy in Parkinson's disease. *J Neuropsychiatry Clin Neurosci* 2002;**14**:461–2.
74. Auriel E, Hausdorff JM, Giladi N. Methylphenidate for the treatment of Parkinson disease and other neurological disorders. *Clin Neuropharmacol* 2009;**32**:75–81.
75. Mendonça DA, Menezes K, Jog MS. Methylphenidate improves fatigue scores in Parkinson disease: a randomized controlled trial. *Mov Disord* 2007;**22**:2070–6.
76. Su JH, Anderson AJ, Cribbs DH, *et al*. Fas and Fas ligand are associated with neuritic degeneration in the AD brain and participate in β-amyloid-induced neuronal death. *Neurobiol Dis* 2003;**12**:182–93.
77. Basaran R, Ozdamar ED, Can-Eke B. CYP2E1 and Parkinson's disease in a MPTP-induced C57BL/6 mouse model. *Mol Neurodegener* 2013;**8**:P9.
78. Kalaria R, Golde T, Kroon S, *et al*. Serine protease inhibitor antithrombin III and its messenger RNA in the pathogenesis of Alzheimer's disease. *Am J Pathol* 1993;**143**:886.
79. Cai Y, An SSA, Kim S. Mutations in presenilin 2 and its implications in Alzheimer's disease and other dementia-associated disorders. *Clin Interv Aging* 2015;**10**: 1163.
80. Ibanez L, Dube U, Davis AA, *et al*. Pleiotropic effects of variants in dementia genes in Parkinson disease. *Front Neurosci* 2018;**12**:230.

81. Outeiro TF, Kontopoulos E, Altmann SM, *et al*. Sirtuin 2 inhibitors rescue α-synuclein-mediated toxicity in models of Parkinson's disease. *Science* 2007;**317**:516–9.
82. Liu Y, Zhang Y, Zhu K, *et al*. Emerging role of Sirtuin 2 in Parkinson's disease. *Front Aging Neurosci* 2020;**11**:372.
83. Tansey MG, McCoy MK, Frank-Cannon TC. Neuroinflammatory mechanisms in Parkinson's disease: potential environmental triggers, pathways, and targets for early therapeutic intervention. *Exp Neurol* 2007;**208**:1–25.
84. Mogi M, Togari A, Kondo T, *et al*. Caspase activities and tumor necrosis factor receptor R1 (p55) level are elevated in the substantia nigra from parkinsonian brain. *J Neural Transm* 2000;**107**:335–41.
85. Boka G, Anglade P, Wallach D, *et al*. Immunocytochemical analysis of tumor necrosis factor and its receptors in Parkinson's disease. *Neurosci Lett* 1994;**172**:151–4.
86. Aslam M, Badshah M, Abbasi R, *et al*. Further evidence for the association of CYP2D6* 4 gene polymorphism with Parkinson's disease: a case control study. *Genes Environ* 2017;**39**: 1–6.
87. Lu Y, Peng Q, Zeng Z, *et al*. CYP2D6 phenotypes and Parkinson's disease risk: a meta-analysis. *J Neurol Sci* 2014;**336**:161–8.
88. Michelson D, Read HA, Ruff DD, *et al*. CYP2D6 and clinical response to atomoxetine in children and adolescents with ADHD. *J Am Acad Child Adolesc Psychiatry* 2007;**46**: 242–51.
89. Zhu Y, Che C, Jin B, *et al*. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics J* 2020; 1460458220937101.
90. Zhou Y, Wang F, Tang J, *et al*. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit Health* 2020.