Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Toward better drug discovery with knowledge graph

Xiangxiang Zeng[1], Xinqi Tu[1], Yuansheng Liu[1], Xiangzheng Fu[1] and Yansen Su[2]

Abstract

Drug discovery is the process of new drug identification. This process is driven by the increasing data from existing chemical libraries and data banks. The knowledge graph is introduced to the domain of drug discovery for imposing an explicit structure to integrate heterogeneous biomedical data. The graph can provide structured relations among multiple entities and unstructured semantic relations associated with entities. In this review, we summarize knowledge graph-based works that implement drug repurposing and adverse drug reaction prediction for drug discovery. As knowledge representation learning is a common way to explore knowledge graphs for prediction problems, we introduce several representative embedding models to provide a comprehensive understanding of knowledge representation learning.

Addresses

[1] College of Information Science and Engineering, Hunan University, Changsha, 410086, China

[2] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, 230601, China

Corresponding author: Liu, Yuansheng (yuanshengliu@hnu.edu.cn)

Current Opinion in Structural Biology 2022, **72**:114–126

This review comes from a themed issue on **Artificial Intelligence (AI) Methodologies in Structural Biology**

Edited by **Feixiong Cheng** and **Nurcan Tuncbag**

For complete overview of the section, please refer the article collection - Artificial Intelligence (AI) Methodologies in Structural Biology

Available online 11 October 2021

https://doi.org/10.1016/j.sbi.2021.09.003

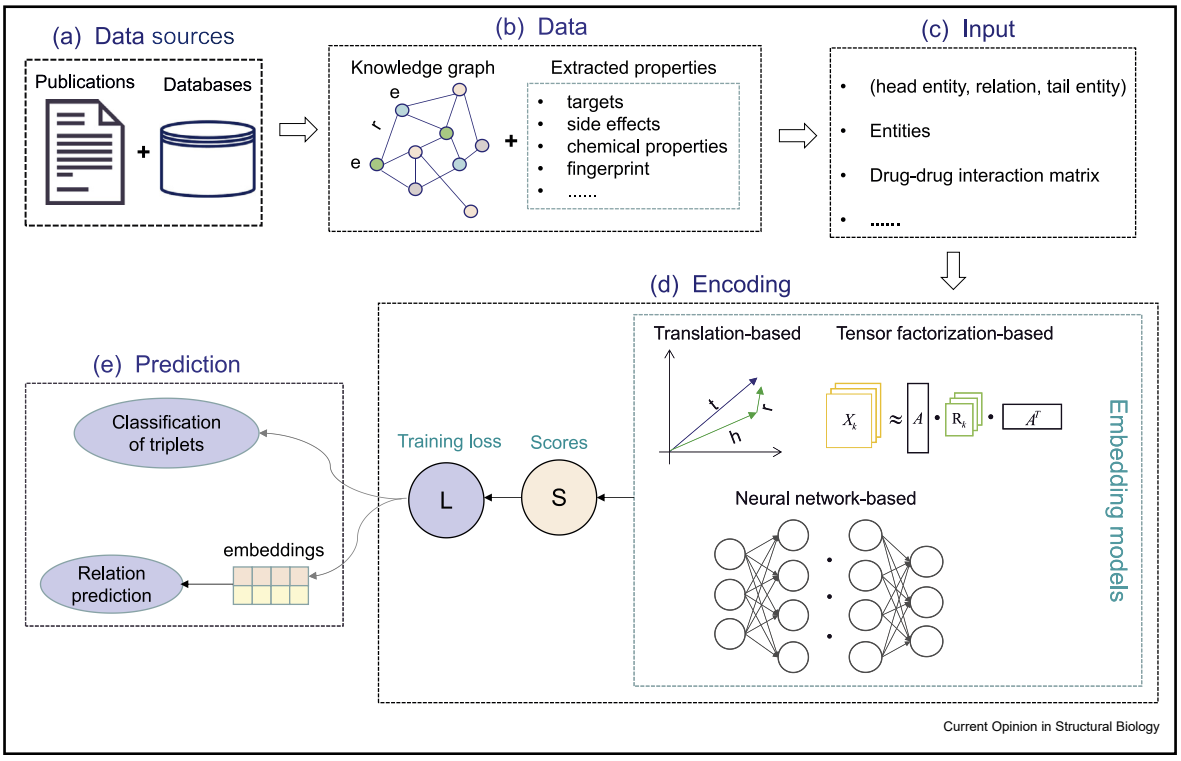
Introduction

Before the drug is identified as a final product, several key steps are conducted from initial target identification to final clinical trials [1,2], which usually takes 10–15 years and costs around 2 billion US dollars. For instance, researchers took 15.2 years to develop Eculizumab and 10.6 years for Brentuximab vedotin [3]. This process is also accompanied by a high risk of failure (>90%) [4]. For example, Solanezumab was developed with a huge investment but failed due to the poor performance in phase III clinical trial [5]. To reduce the cost and increase the success rate, researchers work on assisting and accelerating drug discovery by taking advantage of existing biomedical data that was generated during drug research and development.

With the increasing amount of biomedical data, implementing data integration and extracting useful information from data becomes a central question. To tackle this issue, the knowledge graph (KG) is used for drug discovery because of its advantages. Traditional graphs and networks used for integrating biomedical data only contain one type of relations (e.g. interactions between proteins), while the KG provides heterogeneous information, including various entities (e.g. proteins, targets, and drugs) and multiple types of relations (e.g. interactions between drugs or drug-target pairs) [6]. The KG can also provide unstructured semantic relations between each entity. In a KG, entities are represented as nodes, and relations between entities are represented as edges that connect various nodes [7]; complex relations between entities in biological systems can be easily modeled by a KG [8].

Because of the development of artificial intelligence, traditional drug discovery has been greatly improved by introdcing advanced deep learning methods to molecular property prediction [9] and *de nova* molecular design and optimization [10]. These breakthrough innovations shorten the cycle of drug discovery and reduce the cost. Opportunities in accelerating drug discovery have been increased by drug repurposing and adverse drug reaction (ADR) prediction based on biomedical KGs. Researchers commonly implement drug repurposing and ADR prediction by predicting relations between known entities. For instance, Thafar et al. [11] predicted drug-target interaction (DTI) to identify the repositioning of existing drugs. Yu et al. [12] predicted drug–drug interaction (DDI) to identify ADRs. Most existing works follow a common scheme shown in Figure 1. In this review, we summarize advanced works that assisting drug discovery with KGs. We present recent K-based works in the field of drug repurposing and ADRs prediction, and also discuss the future perspective of drug discovery.

Figure 1



Procedure of knowledge graph-based prediction for drug discovery. **(a)** Common data sources used for constructing biomedical knowledge graphs. **(b)** Data that are extracted from data sources. **(c)** Different inputs of the embedding model, including (head entity, relation, tail entity) triplets, drug–drug interaction matrix, and so on. **(d)** In the procedure of embedding, models with various structures are used to learn representations of knowledge graphs, and models are trained with specific score and loss functions. **(e)** Predictions can be implemented by identifying whether a triplet is a fact or using embeddings of entities to predict.

Construction and embedding of biomedical KGs

Biomedical KG construction relies on various data sources, including unstructured databases and structured ontologies. Table 1 summarizes commonly used databases with brief descriptions and specialty labels. Each data source has a specialty label that represents the main type of data in the database. For instance, DrugBank [13] and SuperTarget [14] mainly contain drug properties, and PubChem [15] and ChEMBL [16] provide chemical-related information, such as the functions and biological activities of compounds. A detailed overview of drug knowledge bases can be found in the study by Zhu et al. [17], wherein the authors provided valuable insights on applications of existing drug knowledge bases. Researchers also extract entities and relations from published literature with text mining technologies [18]. GNBR [19] and DRKG [20] are two publicly available KGs that consist of information extracted from biomedical publications. Hetionet [21] was constructed by integrating data from 29 publicly available data resources, and CBKH [22] is a comprehensive KG constructed by collecting and integrating data from diverse biomedical knowledge bases and KGs.

Researchers are exploring the utilization of knowledge representation learning methods that embed KGs into low-dimensional vectors for predictive tasks [32]. To provide a conceptual understanding of how to learn representations of KGs, we introduce several representative KG embedding models. Table 2 summarizes the representative KG embedding models and provides relevant links for interested researchers to follow up. Among these models, translation-based and tensor factorization-based models are classic; neural network-based models have become increasingly popular for knowledge representation learning, and the architectures of these models are illustrated in Figure 2. In this review, we tend to focus on KG representation learning and its applications of drug repurposing and ADR prediction. Li et al. [33] provided an observation on graph representation learning of biomedical networks from various levels, including molecular-level, genomics-level, therapeutics-level, and healthcare-level, which is complementary to the perspective of our article. And Su et al. [34] provided

Table 1

Commonly used databases for biomedical knowledge graph construction.

Databases	Specialty	Descriptions	Year	URL
DrugBank [13]	Drug	An unique resource consists of comprehensive drug-target information and drug properties.	2008	https://go.drugbank.com
Supertarget [14]	Drug	Containing a dataset of binding affinity between drugs and targets.	2012	http://bioinformatics.charite.de/supertarget/
PubChem [15]	Chemical	Providing information about chemical substances and biological activities.	2016	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL [16]	Chemical	Containing bioactivity measurement of compounds, functional information of protein targets, and drug-like bioactive compounds.	2012	https://www.ebi.ac.uk/chembldb
UniProt [23]	Protein	A comprehensive database that contains protein sequences and function information.	2019	https://www.uniprot.org/
TTD [24]	Protein	Providing information about the explored therapeutic targets that includes proteins and nucleic acid.	2002	http://db.idrblab.net/ttd/
PharmGKB [25]	Gene	A knowledge base that contains genomics, phenotype and clinical information.	2013	http://www.pharmgkb.org
GO [26]	Gene	Providing structured knowledge that is related to functions of genes and their products.	2019	http://geneontology.org
KEGG [27]	Gene	Linking genomics information with higher order functional information and provides systematic analysis of gene functions.	2000	http://www.genome.ad.jp/kegg/
Reactome [28]	Pathway	A database that contains pathways, reactions, and biological process.	2010	http://www.reactome.org
HPO [29]	Disease	Describing phenotypic abnormalities of human diseases by specific terms and providing a standardized vocabulary.	2017	www.human-phenotype-ontology.org
SIDER [30]	Side Effect	A database that contains drugs, adverse drug reactions and drug-adverse drug reaction pairs.	2016	http://sideeffects.embl.de
TWOSIDES [31]	Side Effect	A resource of polypharmacy side effects for drug pairs.	2012	http://tatonettilab.org/resources/tatonetti-stm.html

a specific perspective of network embedding-related works in the biomedical domain in their review.

Translation-based models

In translation-based models, the relation r is treated as a translation through which the head entity h can be projected to the embedding space close to the tail entity t [Figure 2(a)], and $r \in R$ (the set of relations) and h, $t \in E$ (the set of entities). A triplet (h, r, t) usually represents a fact. For example, (anticoagulant, hemorrhage, Celecoxib) is a DDI fact, indicating that anticoagulant and Celecoxib can be enhanced by each other and cause hemorrhage [58]. TransE [35] is the most classic translation-based model; it uses triplets (h, r, t) as positive samples and takes (h', r, t') as negative samples. A negative sample is generated by replacing an entity in a positive sample with a random entity. Following the assumption $h + r \approx t$, the authors trained the model to learn representations by minimizing the distance between $h + r$ and t and maximizing the distance between $h' + r$ and t'. However, TransE performs poorly on modeling relations with complex mapping properties (N-to-N, N-to-1 and 1-to-N). To tackle this issue, Wang et al. [36] proposed TransH to model each relation as a hyperplane on which the projections of h and t are required to be close. Aiming to further enhance the well-known basic models such as TransE, TransROWL [37] considers background knowledge and injects the background knowledge to embedding during the training process.

Considering that it is significant to model relations with composition, inversion and symmetry/anti-symmetry patterns, Sun et al. [38] proposed RotatE that takes the relations as rotations from head entities to tail entities. In RotatE, relations and entities are projected into the complex space. To handle relations with various mapping properties and patterns simultaneously, Chao et al. [39] proposed PairRE, wherein each relation is represented by paired vectors r^H and r^T, and entities are projected to the Euclidean space based on paired vectors. And Yu et al. [40] proposed MQuadE that represents a triplet with a matrix quadruple. The head entity is projected into one matrix of the relation and the tail entity is projected into another matrix of the relation, both of the two matrices of a relation are similar to positive samples and dissimilar to negative samples.

Tensor factorization-based models

Methods based on tensor factorization can model relationships as a three-mode tensor X without relying on any information about KGs. Two dimensions of the tensor denote entities, and one dimension refers to relations [Figure 2(b)]. X_k denotes a frontal slice of X in relation to k, and a tensor value X_{ijk} will be set as 1 if relation k exists between entity i and entity j, otherwise 0. RESCAL [41] is a typical tensor factorization-based

Table 2

Representative knowledge graph embedding models.

Category	Model	Description	Year	Source code
Translation	TransE [35]	A classic translation-based model that has advanced performance on handling relations with 1-to-1 mapping property.	2013	https://github.com/ZichaoHuang/TransE
	TransH [36]	A model that was proposed mainly for extending TransE in the aspect of modeling relations with more complex mapping properties, including N-to-N, N-to-1, and 1-to-N.	2014	https://github.com/thunlp/OpenKE
	TransROWL [37]	Injecting background knowledge to embedding during the process of training.	2021	https://github.com/Keehl-Mihael/TransROWL-HRS
	RotatE [38]	Aiming at modeling relations with symmetry/antisymmetry, inversion and composition patterns.	2019	https://github.com/DeepGraphLearning/KnowledgeGraphEmbedding
	PairRE [39]	A model that can handle relations with various mapping properties and patterns simultaneously.	2020	https://bit.ly/3Cjn2t9
	MQuadE [40]	A unified model for knowledge fact embedding, and it represents a triplet with a matrix quadruple.	2021	–
Tensor factorization	RESCAL [41]	A classic tensor factorization-based model that learns the KG representations by decompositing a three-mode binarized tensor.	2011	https://github.com/mnick/rescal.py
	DistMult [42]	Simplifying RESCAL by restricting the matrix of the k-th relation to a diagonal matrix.	2014	https://github.com/thunlp/OpenKE
	ComplEx [43]	Extending DistMult further in the field of complex space.	2016	https://github.com/ttrouill/complex
	TuckER [44]	A novel model based on Tucker decomposition.	2019	https://github.com/ibalazevic/TuckER
	BTDE [45]	A model that was designed based on Block term decomposition.	2020	–
	Kishimoto et al. [46]	A method based on CP tensor decomposition aims at reducing the memory requirement of the model.	2019	https://github.com/KokiKishimoto/cp_decomposition
	SimplE [47]	A model that enhanced the CP tensor decomposition and learned two embeddings of each entity dependently.	2018	https://github.com/Mehran-k/SimplE
	MEI [48]	A novel framework that learns the interaction mechanism automatically and divides the embedding of a relation or an entity into multiple partitions.	2020	https://github.com/tranhungnghiep/AnalyzeKGE
Neural network	KGNN [49]	It was designed based on GNN, and it can capture high-order structures and semantic relations in a knowledge graph.	2020	https://github.com/xzenglab/KGNN
	KE-GCN [50]	A GCN-based model that jointly propagates and updates the embeddings of both entities and relations in a recursive aggregation process.	2021	https://github.com/PlusRoss/KE-GCN
	RS-GCN [51]	A multi-relational GNN model for DDI prediction.	2021	–
	Alshahrani et al. [52]	A method that was designed based on Word2vec [53] and used for predicting drug-target interactions and drug−disease associations.	2018	https://github.com/bio-ontology-research-group/multi-drug-embedding
	Dai et al. [54]	A creative method that uses an adversarial autoencoder to generate knowledge graph representations.	2020	https://github.com/dyf0631/AAE_FOR_KG
	ConvE [55]	Introducing 2D convolutions to modeling and embedding interactions between entities and relationships.	2018	https://github.com/TimDettmers/ConvE
	TransMTL [56]		2021	https://github.com/daiquocnguyen/ConvKB

(*continued on next page*)

Table 2. (*continued*)

Category	Model	Description	Year	Source code
	ParamE [57]	A novel model that combines convolutional neural network and translation method to learn multiple embeddings of a KG simultaneously. Making use of the translational property as well as taking advantage of the non-linearity fitting ability of neural networks.	2020	–

model that learns the embedding of entities by decomposing X_k with the formula $X_k \approx AR_kA^T$. The matrix A contains the latent-component representation of entities, and R_k refers to the interaction of latent-component in relation to k. To simplify RESCAL, DistMult [42] restricts matrix R_k to be diagonal and mapping each relation to a fixed d-dimensional vector r. Because of that DistMult is more fit for symmetric KGs, ComplEx [43] was proposed to extends DistMult further in the field of complex space. These three models can be considered as special cases of TuckER [44] that factorizes binary tensors based on Tucker decomposition, and the decomposed tensor is represented as multiplication of factor matrices along with each mode and a core tensor. Luo et al. [45] designed a Block term decomposition-based model named BTDE and this model can generate interpretable embeddings with lower dimensions.

Most existing KG embedding models are facing the challenge of large memory consumption, especially when applied on a large-scale KG. To reduce memory usage, Kishimoto et al. [46] presented a novel method, wherein parameters of CANDECOMP/PARAFAC (CP) tensor decomposition are binarized for reducing model size drastically. CP decomposition-based models tend to learn two independent embeddings for each entity. Therefore, Kazemi et al. proposed [47] an enhancement CP called SimplE, and SimplE learns two embeddings of each entity dependently. Moreover, the embedding of each entity is usually considered as a whole, which might induce extensibility problems when applying embeddings with large sizes to model interactions. For this problem, Tran et al. [48] proposed a novel framework called MEI, wherein each embedding is divided into multiple parts for modeling interactions efficiently.

Neural network-based models

Embedding models can encode representations of KGs into vector space with non-linear transformations by introducing neural network architectures. For knowledge representation learning, graph neural network (GNN) [Figure 2(c)] is a favorable architecture due to its ability to obtain the potential long—distance correlation between nodes. Inspired by this, Lin et al. [49] presented an end-to-end model called KGNN, utilizing GNN to preserve the high-order topological neighborhood of drug pairs. This method aggregates information from neighborhood entities to learn the representation of the interest entity, and it captures the high-order structure and semantic relations. Yu et al. [50] proposed KE- graph convolution network (GCN) to learn KG representations by taking advantage of the GCN, and embeddings of both entities and relations are jointly propagated and updated in a recursive aggregation process. Feeney et al. [51] proposed a multi-relational GNN model for DDI prediction, named RS-GCN. In

Figure 2

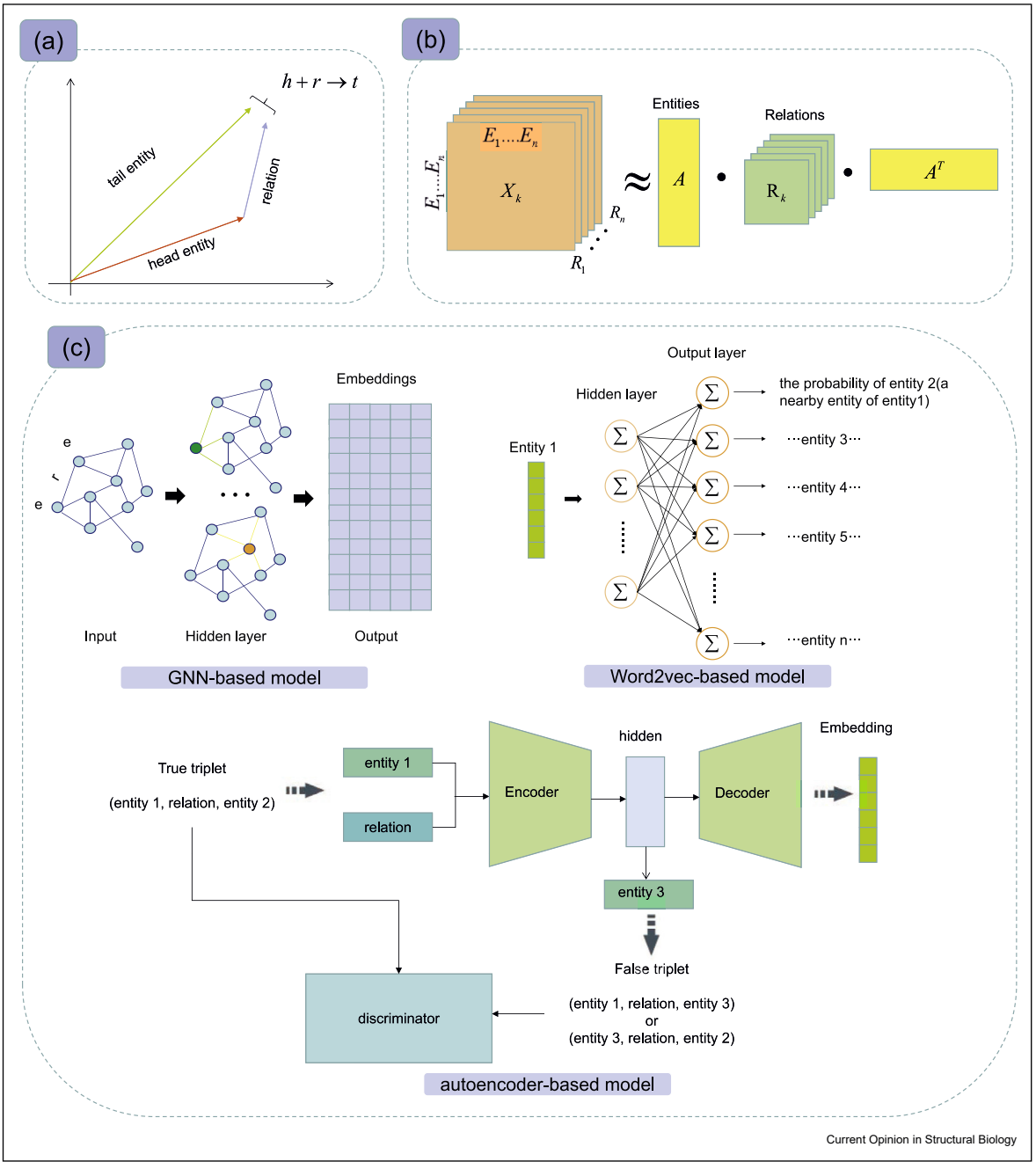


Illustration of embedding models with various concepts and architectures. **(a)** Translation-based embedding model. *h*, *r*, and *t* are embeddings of the head entity, relation, and tail entity, respectively. **(b)** Tensor factorization-based embedding model. $E_1 \ldots E_n$ denotes the entities and $R_1 \ldots R_n$ denotes relations. X_k is the *k*-th frontal slice of the three-mode tensor, and it can be decomposed to relation-latent-component R_k and matrix A that contains entity-latent-component. **(c)** Neural network-based embedding model, three types of network architectures are illustrated here. Graph neural network (GNN) is taken as an encoder to generate embeddings in a GNN-based model; word2vec-based model takes an entity as the input, and output probabilities of multiple entities being nearby; adversarial autoencoder-based model generates negative triplets by the encoder and improves the accuracy of embeddings by using a discriminator to distinguish positive and negative triplets.

RS-GCN, the importance of relation type that exists between sampled neighborhoods is modeled, and the learned probability of relationship type can reflect the frequency and importance, thereby providing a more reliable result.

Apart from GNN, Word2vec [53] is the most commonly used architecture for KG embedding. This kind of model takes an entity as input and the output corresponds to multiple entities. During the training process, entities are initialized as one-hot encoding and updated as multiplication of activated weights and the current representation [Figure 2(c)]. Alshahrani et al. [52] used DeepWalk [59] to generate a corpus from literature for each entity and extended Word2vec to learn the representations of entities based on the text corpus. In addition, a supervised model was trained to predict DTI and drug–disease association (DDA) based on the learned representations. More architectures of neural networks have been applied to learn KG representations for better performance. As an example, Dai et al. [54] constructed an adversarial autoencoder [Figure 2(c)] that consists of an autoencoder and a discriminator. The latent vectors generated by the encoder were used as representations of negative samples (i.e. false triplet (h', r, t')), and a decoder was used to reconstruct representations of inputs based on latent vectors. To optimize the autoencoder for generating more effective embeddings, a discriminator was trained by encouraging the encoder to generate latent vectors that are more close to representations of inputs. The convolutional neural network is capable of improving expression with multiple layers and keeping parameters efficient at the same time. Dettmers et al. [55] proposed ConvE that uses the 2D convolution with convolutional and fully connected layers to model interactions. More recent works are exploring the usage of translational properties, as well as taking advantage of the non-linearity fitting ability of neural networks. For example, Dou et al. [56] designed a model based on multi-task learning, named TransMTL. In this model, translational and neural network methods are combined to learn multiple embeddings of a KG simultaneously. And Che et al. [57] proposed ParamE that regards embeddings of relations as parameters of the model, and the model takes head entities and tail entities as the input and the output, respectively.

Application of KG in drug discovery

Before introducing existing works that apply KGs into drug discovery, we give a rough explanation about drug design and development to present a clearer understanding of drug discovery and stages that might be assisted by KG-based approaches.

Drug design and development

Target is the binding site of drugs in an organism. Bindings between targets and drugs can transform the functions of cells, thereby achieving therapies for diseases. Drug design begins with target identification, and targets are commonly selected from scientific literature or initially identified through experiments. Once the target is selected, researchers start to find lead compounds that might act on the target. Lead compounds can be derived from natural products or synthesized by computer simulation based on the spatial structure of the target. By further selection, lead compounds are determined and used to synthesize candidates of drugs through modification, optimization, structure–activity relationship, and multiple rounds of screening.

Designing a drug that is effective for certain diseases starts from target identification followed by a long cycle. To improve the efficiency of drug discovery, researchers explore the repurposing of 'old' drugs [60], which involves low-risk drugs and makes it possible to move to clinical phases earlier. Assessing ADRs in pharmacology analysis and clinical application is also a vital task, therefore, in the stages of safety pharmacology analysis and phase III clinical trials, researchers usually invest a lot to assess ADRs by identifying side effects and DDIs. From this perspective, taking a complementary technique that implements drug repurposing and ADR prediction based on the KG is an effective choice.

KG-based prediction

Prediction of ADR

ADRs are undesirable effects that do not meet the anticipated therapeutic effects and commonly cause injuries to patients [61]. Before a drug is available on the market, some clinical trials should have been arranged to test its safety. However, only 1000–5000 patients are likely to participate in phase III clinical trials, and most patients with different health conditions are not considered. As a result, safety detection often depends on the acumen of clinical trials. One factor of ADRs is polypharmacy side effects raised by interactions between drugs. A detailed survey shows that approximately 3%–5% of medication errors are caused by unexpected but preventable DDIs [62]. Because traditional experiments are time-consuming and cost a lot, computational approaches provide effective strategies to predict DDIs. For instance, works in the study by Ding et al. [63], Liu et al. [64], and Zhang et al. [65] made some positive contributions to predicting ADRs. Ryu et al. [66] developed a novel framework called DeepDDI, using names and structural information of drugs to predict DDI types, as well as drug–food interactions. Deng et al. [67] presented DDIMDL based on multi-modal learning, and the model combines four features of drugs, including chemical substructures, targets, enzymes and pathways, to learn representations. A drug is represented as a binary vector, and a value in the binary vector is set as 1 when an original feature exists, otherwise 0.

For sufficiently and accurately accessing ADRs of new drugs, existing public and proprietary sources of ADR-related data should be used. The KG is a popular choice used to integrate heterogeneous data and complete potential information for better ADR prediction. To decrease the risk of ADRs, researchers increased their interest in predicting DDIs based on KG techniques. Abdelaziz et al. [68] designed a similarity-based model, named Tiresias, which models DDIs as links in a KG and considers DDI prediction as link prediction. The authors constructed a biomedical KG with drug attributes and relations, and they inferred interactions by calculating the similarity between drugs. Tiresias was trained to predict potential DDIs with global and local features derived from the KG. Instead of computing similarity values between drugs for DDI prediction, Karim et al. [69] tried to learn embeddings of a KG and used the learned embeddings to predict DDIs. They constructed a KG by integrating 1200 drug features from various databases and trained ComplEx [43] to learn drug representations. The relations between drugs are represented by concatenated embeddings of each drug in a drug pair, and the authors trained the Conv-LSTM [70] to predict DDIs based on the learned relation representations. In this work, the authors used the long short-term memory to preserve effects between drugs with long-distance, which has the same purpose with the utilization of 'global features' in the study by Abdelaziz et al. [68].

Without any information about drug properties or structures, Dai et al. [54] used KGs, which only contain various drug names and interactions, to train a model for predicting potential interactions. An adversarial autoencoder was designed as the embedding model, and the decoder of this model is used to generate negative samples. Lin et al. [49] used a KG and a DDI matrix as input. They modeled DDI prediction as a binary classification problem and proposed an embedding model called KGNN, and the predicted values of KGNN indicate whether an interaction exists between the interest drug pair. The authors combined the KG and GNN to capture semantic relations and the high-order structure, which overcomes a challenge in previous works [54,69].

Most studies focus on the effectiveness and accuracy of embedding models but ignore the diversity of tasks. Universally, DDI prediction is modeled as a binary classification problem that indicates whether an interaction exists with a classification label, whereas prediction of multi-typed drug interaction is more meaningful. To fill the gaps, Yu et al. [12] used a multi-channel neural encoder to implement multi-typed DDI prediction. They extracted a subgraph that contains neighborhood entities of the given drug pair and provided the mechanism of drug interaction by generating a pathway based on GNN, thereby integrating diverse information to generate representations of drug pairs. Another key factor that also causes ADR is the side effect of a drug, for example, patients might be allergic to certain ingredients in a drug. Instead of focusing on predicting DDIs, Munoz et al. [71] used the multi-label ranking mechanism to predict side effects, and each label indicates a kind of side effect. In this work, various features are generated to train a multi-label ranking model.

Information contained in structures of molecules is important, as well as relations and topological structure information. To integrate information contained in drug structures and the KG, Chen et al. [72] presented a DDI prediction model, called MUFFIN, which learns representations of drug structures and corresponding knowledge in entities by a message-passing neural network [73] and TransE [35], respectively. The final representation of a drug is represented as a concatenation of the structure representation and the knowledge representation. Wang et al. [74] presented a DDI prediction model called MIRACLE. In MIRACLE, the information contained in drug structures and relations are modeled with a multi-view graph, wherein nodes are represented as molecular structures of drugs. These two works bring inspiration to the future development of KG-based approaches in drug discovery, and researchers can use existing data in a more efficacious way, that is, investing information from data by combining topological structures and semantic relations with molecular structures in future work.

Prediction for drug repurposing

Drug repurposing is a strategy used for finding a drug from approved drugs to treat a specific disease (i.e. identifying new uses of known drugs), and it is implemented by identifying associations between drugs and diseases or inferring interactions between drugs and targets. Therefore, predicting DDA and DTI is a major step in drug discovery. Commonly, researchers make efforts on developing machine learning-based approaches for drug repurposing, and these approaches take known DTIs as labels and use measured data (e.g. molecular structures, protein sequences, expression profiles, and molecular fingerprints) as input features. For instance, Chu et al. [75] presented a multi-label classification method called DTI-MLCD. In this method, drugs are described with molecular descriptors and fingerprints and targets are represented by three various types of sequence-derived features. Madhukar et al. [76] proposed a Bayesian-based approach called BANDIT, and the model implements DTI prediction with integrated multiple types of data, including growth inhibition data, gene expression data, and bioassays/chemical structures. Zhao et al. [77] presented a novel workflow to predict indications for a specific disease. They used expression profiles of drugs as features, and drugs with high repurposing probabilities are taken as candidates.

DTIs can help to identify the drug mechanism of action and unexpected effects of drugs. With reliable DTI prediction, candidate compounds can be designed by referring to the structure of the drug, and new uses of drugs can be found by applying a known drug to diseases that have related targets with the drug. For incorporating KG techniques to DTI prediction, Alshahrani et al. [78] combined the symbolic method and the neural network to generate embeddings of entities. Symbolic logic and automated reasoning are used for involving explicit and implicit information to embeddings, and Word2vec [53] is used to take the corpus generated by random walk as the input. The learned Representations are used to predict links that might indicate potential relations between genes and diseases, protein pairs, or drug-target pairs. Holding a similar concept with the study by Alshahrani et al. [78], the work in the study by Alshahrani and Hoehndorf [52] used the random walk to generate a corpus from a KG and learned representations with Word2Vec [53]. Furthermore, Mohamed et al. [79] proposed TriModel based on tensor factorization. In this work, a KG that consists of drugs and targets was constructed, and representations of entities are identified by TriModel and used for further interaction prediction.

A disease is considered to be similar to another when they have common therapies so that drugs used for one disease might be an efficacious treatment for the other disease in a disease pair [80]. Therefore, novel indications of a drug can be discovered through the other drug, which has similar indications [81]. Predicting DDAs help researchers to infer new indications of drugs, which is an effective method to find new therapies for rare diseases. Sang et al. [82] proposed GrEDel to learn representations of a biomedical KG, and a long short-term memory model was trained to predict probabilities of DDAs. Zhu et al. [83] developed a KG by integrating multiple knowledge bases that are related to drugs, and they presented a novel approach to implement drug repurposing. Sosa et al. [84] focused on rare diseases that have poor survival rates, aiming to use existing data to generate drug repurposing hypotheses for rare diseases. The authors used a heterogeneous KG called GNBR [19] to support repurposing hypotheses. In their work, not only was link prediction executed but also uncertainties of relationships were modeled as well, which is beneficial for generating convincing results. Himmelstein et al. [21] constructed an integrative KG called Hetionet and used a social network analysis algorithm to identify network patterns in the KG. Network patterns indicate that whether a drug is a treatment for a specific disease. For skin cancer without a known cure, Mc-cusker et al. [85] tried to use a probabilistic KG to find drug candidates, and a system ReDrugS was developed for examining drug-target-disease networks and finding supported drugs for melanoma. Notably, they announced that the accuracy of drug candidates is increased by computing the confidence scores of drug candidates.

COVID-19 has become the focal point of the entire world; there are growing opportunities of providing possible treatment strategies for COVID-19 based on drug repurposing. Zeng et al. [20] constructed a comprehensive KG, named DRKG, which contains expressions, genes, pathways, diseases, and drugs. The authors identified 41 drugs that are possible to treat COVID-19. DRKG was further used in the latest work [86], wherein a coupled tensor-matrix embedding framework was proposed to extract concise representations of KGs and implement drug repurposing in the fight against COVID-19. Moreover, Wang et al. [87] developed a knowledge-based framework called COVID-KG to derive knowledge elements from literature, and they constructed multimedia KGs for generating scientific reports and answering drug repurposing-related questions.

Future perspectives

Assisting drug discovery with KG-based approaches is a relatively new and valuable direction. Works in this field make it possible to effectively use existing data sources by extracting semantic information and structured relations from them. Notably, the KG brings new challenges, as well as benefits. In this section, we discuss current challenges and promising future directions of KG-based works in drug discovery from the following three aspects

- **Biomedical KG.** Drug discovery based on biomedical KGs relies heavily on the quality of KGs, which makes it crucial to construct comprehensive biomedical KGs with high quality. When constructing KGs, in addition to use data from databases, researchers take full advantage of entities and relations that are extracted from biomedical literature with natural language processing (NLP) techniques. However, noises might be brought due to unavoidable inaccuracy of NLP models, for example, nonexistent relations and inaccurately named entities might be generated [88]. In addition, biomedical KGs always cover a wide range of subject areas. For issues mentioned above, we take the view that developing powerful automated error detection methods that are specific to the biomedical field can help to improve the quality of KGs from the aspect of data sources. Furthermore, instead of limiting the gaze on building new KGs constantly, researchers should also pay attention to updates and maintenance of existing KGs. Knowledge learned from the latest published literature, new evidence that refutes the conclusion of previous studies, and new data generated by pharmaceutical companies and research institutions should be added to KGs for obtaining a more comprehensive one. Moreover,

multi-party constructing a KG in a unified format jointly might make it easier to maintain the KG. Last but not least, it is still hard to measure the quality of a KG, while quality estimation can help to further improve the quality of KGs. A revelation can be brought by the work in the study by Zhao et al. [89], wherein the authors estimated the probability of triplets in a KG with logic rules. Designing effective evaluation methods to assess the quality of KGs might be a valuable direction for future works.

- **Design of model.** Most existing models used for prediction in drug discovery are built based on embedding models, and they are proved to be effective to some extent. However, predictions generated by such models still lack reliable explanations that are critical to drug repurposing and ADR prediction. Commonly, existing models only predict whether interactions exist but do not interpret the mechanism about how a drug acts or provides pharmacological effects of treatments. Without interpretability, it is hard to convince people to trust results. From this aspect, models with the ability to generate explanations for predictions are in need. Notably, preliminary efforts have been made in some works. For instance, Kang et al. [90] used the most important fact to give an explanation for a prediction, and Zhang et al. [91] generated explanations by using embedding-based path searching. Moreover, embedding models are still facing challenges in learning representations effectively. For example, the negative sample is commonly generated by replacing an entity in the positive sample with a random entity. Because of the incompleteness of KGs, part of the negative samples might be positive in actuality. To tackle this issue, generating negative samples in a proper way is optimization with significance. Another optimization can be accomplished by introducing an evaluation mechanism to calculate the plausibility of predicted triplets, thereby providing more reliable results. In addition, owing to that training embedding models on large and sparse biomedical KGs is time-consuming and has high memory requirements, generic pre-trained embedding models in the field of biomedical are expected.
- **Strategies for better prediction.** Existing works mainly focus on implementing interaction prediction by predicting whether an interaction exists between the interest entity pair, that is, considering the prediction task as a binary classification problem. There are a diversity of relation types between two entities, but rare works make efforts to infer the type of predicted relations. Inferring types of relations might contribute to determining the effects of DDIs and finding new therapies for diseases. For example, the combination of Warfarin (a kind of anticoagulant) and Cefixme (a kind of antibiotics) can cause bleeding-related DDIs, including gastrointestinal bleeding and inhibition of clotting [58]. It is natural to infer that the interaction between Fluconazole (also a kind of antibiotics) and Warfarin is more likely to be bleeding-related. In light of that, inferring interaction with indicative labels is a potential way. In the future, we are also looking forward to making a transformation from single-modal learning into multi-modal learning. Inter-view information contained in molecular graphs and sequences is important as well as the intra-view relationships and semantic information contained in KGs. From this aspect, multi-modal learning is beneficial for learning representations with more comprehensive information, thereby achieving more accurate predictions. Therefore, designing novel approaches based on multi-modal learning to take full advantage of KGs, molecular images, sequences, and molecular graphs is a promising direction for better prediction in the drug discovery field.

Conclusions

The process of drug discovery can be accelerated by using KGs to assist data-driven pharmaceutical researches. Inspired by the effectiveness of such application, recent works introduced KGs into various prediction problems of drug discovery based on KG embedding models and biomedical KGs. In this review, we summarize commonly used databases for KG construction, and we also present an overview of representative knowledge embedding models and KG-based predictions in the drug discovery field. Finally, we conclude with caution that there are still some challenges that need to be solved in the future, and we also discuss these challenges in detail, aiming to give clear future perspectives in this field.

CRediT author statement

Xiangxiang Zeng: Conceptualization, Resources, Writing - Original Draft.

Xinqi Tu: Investigation, Writing - Original Draft.

Yuansheng Liu: Investigation, Supervision, Writing - Review & Editing.

Xiangzheng Fu: Visualization, Data Curation, Writing - Review & Editing.

Yansen Su: Writing - Review & Editing.

Funding

This work is supported in part by the Fundamental Research Funds for the Central Universities (531118010626), the National Natural Science Foundation of China (61872309) and the National Key Research and Development Program of China (2021YFE0102100).

Conflict of interest statement

Nothing declared.

Acknowledgements

We would like to thank Dr. Deepthi K from Cochin University of Science and Technology for English proofreading of the manuscript.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Mohs RC, Greig NH: **Drug discovery and development: role of basic biological research, Alzheimer's & Dementia**. *Translat Res Clin Intervent* 2017, **3**:651–657.

2. Irtaza RS, Drug discovery knowledge graph.

3. Prasad V, Mailankody S: **Research and development spending to bring a single cancer drug to market and revenues after approval**. *JAMA Int Med* 2017, **177**:1569–1575.

4. Berdigaliyev N, Aljofan M: **An overview of drug discovery and development**. *Future Med Chem* 2020, **12**:939–947.

5. Honig LS, Vellas B, Woodward M, Boada M, Bullock R, Borrie M, Hager K, Andreasen N, Scarpini E, Liu-Seifert H, *et al.*: **Trial of solanezumab for mild dementia due to Alzheimer's disease**. *N Engl J Med* 2018, **378**:321–330.

6. • Ji S, Pan S, Cambria E, Marttinen P, Yu PS: **A survey on knowledge graphs: representation, acquisition and applications**. *arXiv* 2020. preprint arXiv:2002.00388.

A comprehensive review in which the current development, limitations, and future perspectives of the knowledge graph were presented. Specially, this survey discussed current development of the knowledge graph from three scopes, including representation, acquisition, and application with detailed case studies.

7. A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, et al., Knowledge graphs, arXiv preprint arXiv:2003.02320.

8. Mohamed SK, Nounu A, Nováček V: **Biological applications of knowledge graph embedding models**. *Briefings Bioinf* 2021, **22**:1679–1693.

9. Jeon W, Kim D: **FP2VEC: a new molecular featurizer for learning molecular properties**. *Bioinformatics* 2019, **35**: 4979–4985.

10. Prykhodko O, Johansson SV, Kotsias P-C, Arús-Pous J, Bjerrum EJ, Engkvist O, Chen H: **A de novo molecular generation method using latent vector based generative adversarial network**. *J Cheminf* 2019, **11**:1–13.

11. Thafar MA, Olayan RS, Ashoor H, Albaradei S, Bajic VB, Gao X, Gojobori T, Essack M: **DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques**. *J Cheminf* 2020, **12**:1–17.

12. Y. Yu, K. Huang, C. Zhang, L. M. Glass, J. Sun, C. Xiao, SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization, arXiv preprint arXiv: 2010.01450.

13. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Res* 2008, **36**:D901–D906.

14. Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, Gilson MK, Bourne PE, Preissner R: **SuperTarget goes quantitative: update on drug–target interactions**. *Nucleic Acids Res* 2012, **40**:D1113–D1117.

15. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, *et al.*: **PubChem substance and compound databases**. *Nucleic Acids Res* 2016, **44**: D1202–D1213.

16. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res* 2012, **40**:D1100–D1107.

17. Zhu Y, Elemento O, Pathak J, Wang F: **Drug knowledge bases and their applications in biomedical informatics research**. *Briefings Bioinf* 2019, **20**:1308–1321.

18. Nicholson DN, Greene CS: **Constructing knowledge graphs and their biomedical applications**. *Comput Struct Biotechnol J* 2020, **18**:1414.

19. Percha B, Altman RB: **A global network of biomedical relationships derived from text**. *Bioinformatics* 2018, **34**: 2614–2624.

20. •• Zeng X, Song X, Ma T, Pan X, Zhou Y, Hou Y, Zhang Z, Li K, Karypis G, Cheng F: **Repurpose open data to discover therapeutics for COVID-19 using deep learning**. *J Proteome Res* 2020, **19**:4624–4636.

They built a comprehensive biomedical knowledge graph called DRKG. Based on DRKG, they proposed an integrative and network-based deep learning method to identify repurposable drugs for COVID-19. In this study, 41 drug candidates were identified to have therapeutic associations with COVID-19.

21. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE: **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**. *Elife* 2017, **6**, e26726.

22. C. Su, Y. Hou, W. Guo, F. Chaudhry, G. Ghahramani, H. Zhang, F. Wang, Cbkh: the cornell biomedical knowledge hub, medRxiv.

23. Consortium U: **UniProt: a worldwide hub of protein knowledge**. *Nucleic Acids Res* 2019, **47**:D506–D515.

24. Chen X, Ji ZL, Chen YZ: **TTD: therapeutic target database**. *Nucleic Acids Res* 2002, **30**:412–415.

25. Thorn CF, Klein TE, Altman RB: **PharmGKB: the pharmacogenomics knowledge base**. In *Pharmacogenomics*. Springer; 2013:311–320.

26. Consortium GO: **The gene ontology resource: 20 years and still GOing strong**. *Nucleic Acids Res* 2019, **47**:D330–D338.

27. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**:27–30.

28. Croft D, O'kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, *et al.*: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic Acids Res* 2010, **39**:D691–D697.

29. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, *et al.*: **The human phenotype ontology in 2017**. *Nucleic Acids Res* 2017, **45**:D865–D876.

30. Kuhn M, Letunic I, Jensen LJ, Bork P: **The SIDER database of drugs and side effects**. *Nucleic Acids Res* 2016, **44**: D1075–D1079.

31. Tatonetti NP, Patrick PY, Daneshjou R, Altman RB: **Data-driven prediction of drug effects and interactions**. *Sci Transl Med* 2012, **4**, 125ra31. 125ra31.

32. Dai Y, Wang S, Xiong NN, Guo W: **A survey on knowledge graph embedding: approaches, applications and benchmarks**. *Electronics* 2020, **9**:750.

33. • Li MM, Huang K, Zitnik M: **Representation learning for networks in biology and medicine: advancements, challenges, and opportunities**. *arXiv* 2021. preprint arXiv:2104.04883.

This work provided a complementary perspective of representation learning in the biomedical field, and the authors provided an observation on representation learning of biomedical networks at various levels, including molecular-level, genomic-level, therapeutics-level, and healthcare-level.

34. Su C, Tong J, Zhu Y, Cui P, Wang F: **Network embedding in biomedical data science**. *Briefings Bioinf* 2020, **21**:182–197.

35. •• Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O: **Translating embeddings for modeling multi-relational data**. In *Neural information processing systems (NIPS)*; 2013:1–9.

A classic translation-based embedding model was proposed in this work, named TransE, which performs well in handling relations with 1-to-1 mapping property.

36. Wang Z, Zhang J, Feng J, Chen Z: **Knowledge graph embedding by translating on hyperplanes**. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 28; 2014.

37. d'Amato C, Quatraro NF, Fanizzi N: **Injecting background knowledge into embedding models for predictive tasks on knowledge graphs**. In *European semantic Web conference*. Springer; 2021:441–457.

38. Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: knowledge graph embedding by relational rotation in complex space, arXiv preprint arXiv:1902.10197.

39. L. Chao, J. He, T. Wang, W. Chu, Pairre: knowledge graph embeddings via paired relation vectors, arXiv preprint arXiv: 2011.03798.

40. Yu J, Cai Y, Sun M, Li P: **Mquade: a unified model for knowledge fact embedding**. In *Proceedings of the Web conference 2021*; 2021:3442–3452.

41. •• Nickel M, Tresp V, Kriegel H-P: **A three-way model for collective learning on multi-relational data**. In *ICML*; 2011:809–816.
This work proposed a classic tensor factorization-based model called RESCAL, and the model learns the latent-component representations of entities and relations by factorizing a three-mode binarized tensor.

42. B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, arXiv preprint arXiv:1412.6575.

43. Trouillon T, Welbl J, Riedel S, Gaussier É, Bouchard G: **Complex embeddings for simple link prediction**. In *International conference on machine learning*. PMLR; 2016:2071–2080.

44. I. Balažević, C. Allen, T. M. Hospedales, Tucker: tensor factorization for knowledge graph completion, arXiv preprint arXiv: 1901.09590.

45. Luo T, Wei Y, Yu M, Li X, Zhao M, Xu T, Yu J, Gao J, Yu R: **Btde: block term decomposition embedding for link prediction in knowledge graph**. In *ECAI 2020*. IOS Press; 2020:817–824.

46. Kishimoto K, Hayashi K, Akai G, Shimbo M, Komatani K: **Binarized knowledge graph embeddings**. In *European conference on information Retrieval*. Springer; 2019:181–196.

47. S. M. Kazemi, D. Poole, Simple embedding for link prediction in knowledge graphs, arXiv preprint arXiv:1802.04868.

48. H. N. Tran, A. Takasu, Multi-partition embedding interaction with block term format for knowledge graph completion, arXiv preprint arXiv:2006.16365.

49. •• Lin X, Quan Z, Wang Z-J, Ma T, Zeng X: **KGNN: knowledge graph neural network for drug-drug interaction prediction**. In *Proceedings of the Twenty-Ninth International Joint conference on artificial intelligence*. IJCAI-20 (International Joint Conferences on Artificial Intelligence Organization); 2020:2739–2745.
An end-to-end framework for predicting drug–drug interaction. This framework adopts knowledge graphs to model correlations between different entities and combined the graph neural network to obtain the rich neighborhood information of each entity.

50. Yu D, Yang Y, Zhang R, Wu Y: **Knowledge embedding based graph convolutional network**. In *Proceedings of the Web conference 2021*; 2021:1619–1628.

51. A. Feeney, R. Gupta, V. Thost, R. Angell, G. Chandu, Y. Adhikari, T. Ma, Relation matters in sampling: a scalable multi-relational graph neural network for drug-drug interaction prediction, arXiv preprint arXiv:2105.13975.

52. M. Alshahrani, R. Hoehndorf, Drug repurposing through joint learning on knowledge graphs and literature.

53. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv: 1301.3781.

54. • Y. Dai, C. Guo, W. Guo, C. Eickhoff, Drug–drug interaction prediction with wasserstein adversarial autoencoder-based knowledge graph embeddings, Briefings Bioinf Doi: 10.1093/bib/bbaa256.
The first work that introduces adversarial autoencoder into knowledge graph embedding for predicting drug–drug interaction. This framework was developed based on wasserstein distance and Gumbel-Softmax relaxation, and it solved the challenge in generating high-quality negative samples.

55. Dettmers T, Minervini P, Stenetorp P, Riedel S: **Convolutional 2d knowledge graph embeddings**. In *Thirty-second AAAI conference on artificial intelligence*; 2018.

56. Dou J, Tian B, Zhang Y, Xing C: **A novel embedding model for knowledge graph completion based on multi-task learning**. *DASFAA* 2021, **1**:240–255.

57. Che F, Zhang D, Tao J, Niu M, Zhao B: **Parame: regarding neural network parameters as relation embeddings for knowledge graph completion**. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34; 2020:2774–2781.

58. T. Ma, J. Shang, C. Xiao, J. Sun, GENN: predicting correlated drug-drug interactions with graph energy neural networks, arXiv preprint arXiv:1910.02107.

59. 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*; 2014:701–710.

60. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, *et al.*: **Drug repurposing: progress, challenges and recommendations**. *Nat Rev Drug Discov* 2019, **18**:41–58.

61. Pirmohamed M, Breckenridge AM, Kitteringham NR, Park BK: **Adverse drug reactions**. *Br Med J* 1998, **316**:1295–1298.

62. D. Flockhart, P. Honig, S. Yasuda, C. Rosebraugh, Preventable adverse drug reactions: a focus on drug interactions, Centers for Education & Research on Therapeutics 452.

63. Ding Y, Tang J, Guo F: **Identification of drug-side effect association via multiple information integration with centered kernel alignment**. *Neurocomputing* 2019, **325**:211–224.

64. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X-w, Matheny ME, Xu H: **Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs**. *J Am Med Inf Assoc* 2012, **19**:e28–e35.

65. Zhang W, Liu F, Luo L, Zhang J: **Predicting drug side effects by multi-label learning and ensemble learning**. *BMC Bioinf* 2015, **16**:1–11.

66. Ryu JY, Kim HU, Lee SY: **Deep learning improves prediction of drug–drug and drug–food interactions**. *Proc Natl Acad Sci Unit States Am* 2018, **115**:E4304–E4311.

67. Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S: **A multimodal deep learning framework for predicting drug–drug interaction events**. *Bioinformatics* 2020, **36**:4316–4322.

68. Abdelaziz I, Fokoue A, Hassanzadeh O, Zhang P, Sadoghi M: **Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions**. *Journal of Web Semantics* 2017, **44**:104–117.

69. Karim MR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S: **Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network**. In *Proceedings of the 10th ACM International conference on Bioinformatics*. Computational Biology and Health Informatics; 2019: 113–123.

70. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, arXiv preprint arXiv:1506.04214.

71. • Muñoz E, Nováček V, Vandenbussche P-Y: **Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models**. *Briefings Bioinf* 2019, **20**: 190–202.
A representative work that casts adverse drug reaction prediction as a multi-label ranking problem and generates different feature sets via a multi-label learning model.

72. Y. Chen, T. Ma, X. Yang, J. Wang, B. Song, X. Zeng, Muffin: multi-scale feature fusion for drug–drug interaction prediction, Bioinformatics.

73. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE: **Neural message passing for quantum chemistry**. In *International conference on machine learning*. PMLR; 2017:1263–1272.

74. Y. Wang, Y. Min, X. Chen, J. Wu, Multi-view graph contrastive representation learning for drug-drug interaction prediction, arXiv preprint arXiv:2010.11711.

75. Chu Y, Shan X, Chen T, Jiang M, Wang Y, Wang Q, Salahub DR, Xiong Y, Wei D-Q: **Dti-mlcd: predicting drug-target interactions using multi-label learning with community detection method**. *Briefings Bioinf* 2021, **22**, bbaa205.

76. Madhukar NS, Khade PK, Huang L, Gayvert K, Galletti G, Stogniew M, Allen JE, Giannakakou P, Elemento O: **A bayesian machine learning approach for drug target identification using diverse data types**. *Nat Commun* 2019, **10**:1–14.

77. Zhao K, So H-C: **Drug repositioning for schizophrenia and depression/anxiety disorders: a machine learning approach leveraging expression data**. *IEEE journal of biomedical and health informatics* 2018, **23**:1304–1315.

78. Alshahrani M, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N, Hoehndorf R: **Neuro-symbolic representation learning on biological knowledge graphs**. *Bioinformatics* 2017, **33**:2723–2730.

79. • Mohamed SK, Nováček V, Nounu A: **Discovering protein drug targets using knowledge graph embeddings**. *Bioinformatics* 2020, **36**:603–610.
For predicting drug-target interaction in a multi-phase procedure, this work proposed a tensor factorization-based knowledge graph embedding model called TriModel.

80. Dudley JT, Deshpande T, Butte AJ: **Exploiting drug–disease relationships for computational drug repositioning**. *Briefings Bioinf* 2011, **12**:303–311.

81. Chiang AP, Butte AJ: **Systematic evaluation of drug–disease relationships to identify leads for novel drug uses**. *Clin Pharmacol Ther* 2009, **86**:507–510.

82. Sang S, Yang Z, Liu X, Wang L, Lin H, Wang J, Dumontier M, GrEDeL: **A knowledge graph embedding based method for drug discovery from biomedical literatures**. *IEEE Access* 2018, **7**:8404–8415.

83. Zhu Y, Che C, Jin B, Zhang N, Su C, Wang F: **Knowledge-driven drug repurposing using a comprehensive drug knowledge graph**. *Health Inf J* 2020, **26**:2737–2750.

84. • Sosa DN, Derry A, Guo M, Wei E, Brinton C, Altman RB: **A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases**. In *Pac symp biocomput*, vol. 25. World Scientific; 2020: 463–474.
This work concentrated on generating drug repurposing hypotheses for rare diseases. The authors utilized a knowledge graph embedding model to explicitly model the uncertainty of relationships derived from literature, and they generated drug repurposing hypotheses by link prediction.

85. McCusker JP, Dumontier M, Yan R, He S, Dordick JS, McGuinness DL: **Finding melanoma drugs through a probabilistic knowledge graph**. *PeerJ Computer Science* 2017, **3**: e106.

86. C. I. Kanatsoulis, N. D. Sidiropoulos, TeX-Graph: coupled tensor-matrix knowledge-graph embedding for COVID-19 drug repurposing, arXiv preprint arXiv:2010.11367.

87. Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, H. Zhang, W. Liu, et al., COVID-19 literature knowledge graph construction and drug repurposing report generation, arXiv preprint arXiv:2007.00576.

88. B. Song, F. Li, Y. Liu, X. Zeng, Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison, Briefings Bioinf.

89. S. Zhao, B. Qin, T. Liu, F. Wang, Biomedical knowledge graph refinement with embedding and logic rules, arXiv preprint arXiv: 2012.01031.

90. B. Kang, J. Lijffijt, T. De Bie, Explaine: an approach for explaining network embedding-based link predictions, arXiv preprint arXiv: 1904.12694.

91. Zhang W, Paudel B, Zhang W, Bernstein A, Chen H: **Interaction embeddings for prediction and explanation in knowledge graphs**. In *Proceedings of the Twelfth ACM International conference on Web Search and data mining*; 2019:96–104.