JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Review

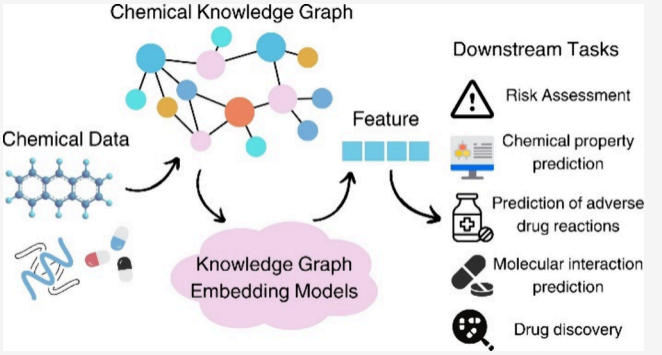
Research Progresses and Applications of Knowledge Graph Embedding Technique in Chemistry

Chuanghui Wang, Yunqing Yang, Jinshuai Song, and Xiaofei Nan*

Cite This: https://doi.org/10.1021/acs.jcim.4c00791 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A knowledge graph (KG) is a technique for modeling entities and their interrelations. Knowledge graph embedding (KGE) translates these entities and relationships into a continuous vector space to facilitate dense and efficient representations. In the domain of chemistry, applying KG and KGE techniques integrates heterogeneous chemical information into a coherent and user-friendly framework, enhances the representation of chemical data features, and is beneficial for downstream tasks, such as chemical property prediction. This paper begins with a comprehensive review of classical and contemporary KGE methodologies, including distance-based models, semantic matching models, and neural network-based approaches. We then catalogue the primary databases employed in chemistry and biochemistry that furnish the KGs with essential chemical data. Subsequently, we explore the latest applications of KG and KGE in chemistry, focusing on risk assessment, property prediction, and drug discovery. Finally, we discuss the current challenges to KG and KGE techniques and provide a perspective on their potential future developments.



KEYWORDS: *Knowledge Graph, Knowledge Graph Embedding, Chemical Database, Chemical Risk Assessment, Adverse Drug Reaction, Drug−Drug Interaction Prediction, Drug−Target Interaction Prediction, Drug Discovery*

1. INTRODUCTION

In recent years, knowledge graphs (KGs), as structured representations of human knowledge, have been extensively employed in knowledge representation and reasoning. They have become essential tools for intelligent systems in solving complex problems and have garnered widespread attention because of their integration of human knowledge with artificial intelligence (AI) technology. Prominent KGs, such as DBpedia,[1] Freebase,[2] Wikidata,[3] NELL,[4] YAGO,[5] etc., have been applied in various fields, such as information extraction, entity disambiguation, semantic analysis, recommendation systems, and question-answering systems.

A KG is a directed graph that structurally represents facts through entities, relationships and semantic descriptions. Entities represent existing objects or abstract concepts, while relationships define the connections between these entities or their semantic descriptions. In a KG, head and tail entities, along with their relations are represented by nodes and edges, forming a triple structure, i.e., (head, relation, tail). With the continuous expansion of data in chemistry, KGs have been successfully applied in data integration and information extraction. Unlike molecular graph structures, KGs can incorporate richer information beyond molecular structure, such as functional groups, molecular properties, and other important prior knowledge. Additionally, KGs facilitate heterogeneous information integration, encompassing various entity types (e.g., chemical elements, compounds, drugs and proteins) and relationships (e.g., chemical reactions between compounds). KG can also provide unstructured semantic relationships between entities. This heterogeneity allows for modeling complex entity relationships due to the diverse graph structure of KGs. Although KGs are effective in representing structured chemical data, the basic symbolic properties of these triples lead to difficulties in manipulation of KGs. Furthermore, due to the extensive scale of KGs and their incorporation of substantial external information, the representation of entities and relations often exhibits high dimensionality, which poses challenges for subsequent feature processing in technological research. To address this issue, recent research on KGs has mainly focused on knowledge graph embedding (KGE), also known as knowledge representation learning (KRL), which maps entities and relations to a low-dimensional space while simultaneously capturing their semantics.[6] The objective of

Received: May 14, 2024
Revised: August 28, 2024
Accepted: August 28, 2024

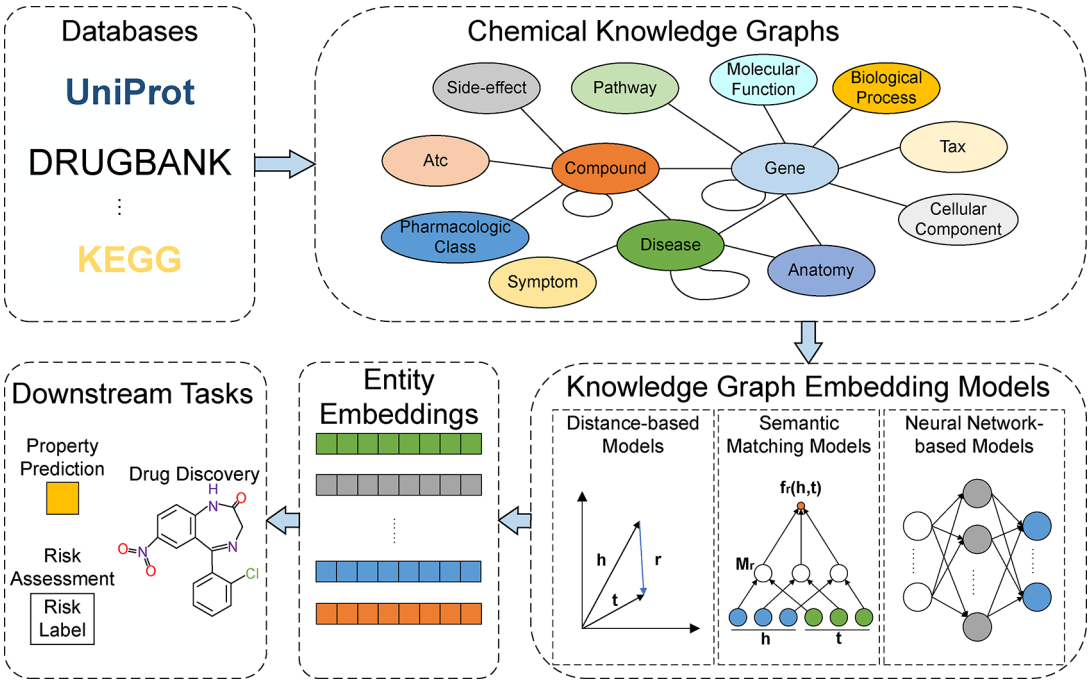


Figure 1. General process of applying knowledge graph embedding models in chemical tasks. A chemical KG is constructed based on resources from chemical databases with additional chemical knowledge. Subsequently, the KG and external knowledge are input into a KGE model to obtain embeddings for chemical entities or other chemical information. Finally, these embeddings are applied to various downstream chemical tasks.

KGE techniques is to depict entities and relationships within KGs as dense, low-dimensional real-valued vectors, maintaining the structural integrity of the KG while minimizing computational complexity. By employing KGE techniques, researchers can extract low-dimensional embeddings that represent the desired chemical entities and relationships from KGs, subsequently applying them across various chemical applications, as shown in Figure 1. These embeddings, obtained through KGE techniques, capture essential chemical properties and latent semantics, thereby enhancing the outcomes of downstream tasks. In summary, the enhancements by KGE techniques facilitate the integration of prior knowledge into traditional chemical applications, thereby further promoting advancements in the field of chemistry applications. This paper reviews recent progress in KGE techniques and their applications in chemistry, categorized according to KG and KGE approaches, chemical databases, and detailed chemical tasks.

2. KNOWLEDGE GRAPH EMBEDDING MODELS

KGs typically encompass rich and intricate knowledge information, particularly in domains such as chemistry and bioinformatics. To effectively extract features from KGs, a model capable of handling heterogeneous graph information and performing efficient knowledge reasoning is essential. However, the high dimensionality and sparsity of chemical KGs result in high computational demands for conventional graph analysis methods. Additionally, the heterogeneity of chemical graph data poses substantial analytical challenges in KG. Unlike conventional graph analysis methods, KGE techniques optimize the mapping of node embeddings into a low-dimensional vector space, preserving the original graph information and capturing the latent semantics of entities and relationships. Furthermore, KGE technique retains the structural information on the KG while reducing computational complexity compared to traditional methods. Ongoing research aims to develop more efficient KGE techniques to enhance embedding quality and refine their application in downstream tasks.

In order to further understand knowledge graph representation learning, we mainly introduce and classify several representative KGE models and novel embedding models emerged in recent years. In general, **h**, **t**, **r** are used to represent the head entity, tail entity and relation of the KG, respectively. The triple (**h**,**r**,**t**) typically represents a fact, such as (Gas, isStateOf, O), which is a chemical element fact indicating that the state of the element oxygen is gas. Among these models, distance-based models and semantic matching models are classical models; neural network-based models and some other models have attracted more and more attention from researchers in recent years. Due to different mathematical principles or neural network structures, each KGE embedding model has different learning emphases in terms of semantic relationships and topological structures, so the learned embeddings are also different. It is crucial to choose an appropriate KGE model in different application scenarios.

2.1. Distance-Based Models. The distance-based model mainly uses a distance-based scoring function to measure the rationality of the fact triples. In this model, the rationality of a fact is regarded as the distance between the head entity and the tail entity in the vector space. According to the translation principle $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, models utilize the distance between $\mathbf{h} + \mathbf{r}$ and **t** of each triplet to calculate its score, and the fact triples observed in KG often have higher scores than unobserved ones.

2.1.1. Translation-Based Models. TransE[7] is the most classical translation-based model, where entities and relations are represented into the same vector space. The relation is interpreted as a translation between the head entity and tail entity, as shown in Figure 2(a). TransE is simple and efficient,

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

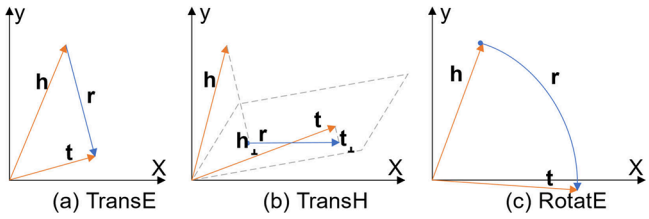


Figure 2. Illustration of TransE, TransH, and RotatE, where **h**, **r**, and **t** denote head entity, relation, and tail entity, respectively. (a) TransE links **h** and **t** using **r** in the same vector space. (b) TransH represents **r** as a translation from **h** to **t** on a relation-specific hyperplane. (c) RotatE represents **r** as a rotation from **h** to **t** in a complex vector space.

but cannot handle noninjective relations (1-to-N, N-to-1 and N-to-N) very well. Wang et al. proposed TransH,[8] in Figure 2(b), which can capture the semantic difference of the same entity in different relations. TransH can effectively handle noninjective relations.

TransE and TransH assume that entities and relations are in the same semantic space, but for complex relations, an entity may have multiple aspects, and different relations pay different attention to each aspect of the entity. To address this issue, TransR[9] makes use of relation-specific spaces to deal with complex relations efficiently, as shown in Figure 3(a).

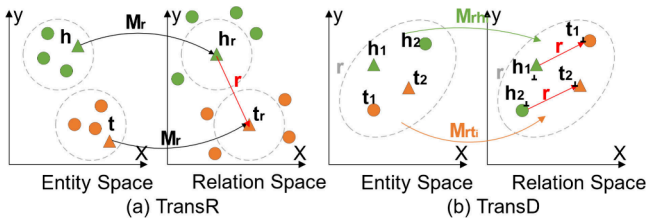


Figure 3. Illustration of TransR and TransD, where **M** denote the projection matrix. (a) TransR models the diversity of relations by defining a separate space for each relation. (b) TransD further refines the projection matrix to model complex entities and relationships.

Although TransR has improved its ability to handle complex relations, it introduces a projection matrix for each relation, which increases the number of parameters and makes the model more complex than TransE/TransH.

Translation-based models like TransE, TransH, and TransR share the same translation principle $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, but they differ in their definitions of relation-specific vector spaces. This translation principle dictates that the tail entity should be the one closest to **h** + **r** in these models. Consequently, these models apply spherical equipotential hyper-surfaces, in Figure 4(a), making it difficult to distinguish the correct matching entity among those in close proximity within the vector space. Additionally, these translation-based models oversimplify the loss metric by treating different dimensions of entity vectors the same way, without considering the significance of features in various dimensions and the relationships between these dimensions. As a result, these models struggle to effectively model complex entities and relationships.

To address these issues, Xiao et al.[10] proposed a more flexible embedding model, TransA. They argue that a relation is influenced only by certain specific dimensions, while other irrelevant dimensions are noisy. Therefore, treating all dimensions equally introduces a significant amount of noise, negatively impacting the model's performance. TransA

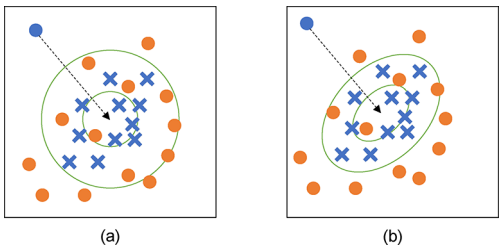


Figure 4. Illustration of equipotential surfaces for target entities. The blue circle represents the head entity, while the dashed lines and arrows indicate the process of matching the tail entity using the translation principle. The blue cross marks the correct target entity, and the orange circle represents the incorrect target entity. Entities on equipotential surfaces closer to the center are better matched to the head entity. In other words, the equipotential surfaces indicate the approximate distribution range of the target entities. (a) Spherical equipotential hyper-surfaces in vector spaces. (b) Elliptical equipotential hyper-surfaces in vector spaces.

adaptively weights the dimensions of entity features, applying elliptical equipotential hyper-surfaces to the model, as shown in Figure 4(b). Additionally, TransA introduces adaptive weights in the loss metric, in Figure 5, making the scoring

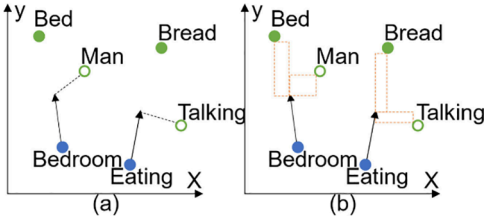


Figure 5. Illustration of TransA. The solid dots are actual tail entities, while the rings are not. (a) The tail entities in spherical equipotential hyper-surfaces are incorrectly matched because of the isotropic Euclidean distance. (b) TransA adaptively weights *x*-axis component and *y*-axis component of loss, thus achieving an exact match.

function more adaptive. TransA can be seen as weighting transformed feature dimensions, significantly improving the translation-based model's ability to handle complex entities and relations.

TransR constructs a single mapping matrix for each typical relation, causing all entities to share this matrix. However, entities linked by a relation are likely to belong to different types, necessitating distinct mapping matrices for different entities. Furthermore, this mapping matrix represents the interaction between entities and relations, and should thus be determined by both entities and relations. To improve upon TransR, Ji et al.[11] proposed the TransD model, as shown in Figure 3(b). In TransD, each entity and relation are defined by two vectors: one vector represents the definition of the entity or relation, and the other vector serves as a projection vector indicating the projection of the entity into the relation vector space. The mapping matrix for a relation is jointly constructed by the projection vectors of both the relation and the entity, resulting in a unique dynamic projection matrix for each entity-relation pair. Therefore, TransD takes into account the diversity of both entities and relations, embedding entities into the relation vector space in a more flexible manner. This approach not only improves the quality of embeddings but also reduces the number of model parameters. TransA and TransD have made improvements to translation-based models in two

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

distinct ways, achieving similar and highly commendable results in handling complex relationships.

d'Amato et al. proposed TransOWL,[12] which enhances the simple but effective entity-based embedding models, such as the TransE model, by better utilizing existing background knowledge. They also improved TransR in a similar way and proposed TransROWL.[12] Specifically, TransOWL introduces specific constraints to improve the learning of embeddings, and injects background knowledge into the base model during the training phase, further improving the quality of embeddings.

Yu et al. proposed MQuadE,[13] which allows for a better capture of complex associations between entities and relations. It uses a matrix pair to represent the relation, and projects the head entity and tail entity into a hidden space respectively to measure the rationality of the triple, as shown in Figure 6. MQuadE is theoretically proved that the method can model various types of relations such as symmetric/asymmetric, inverse, and composite relations.

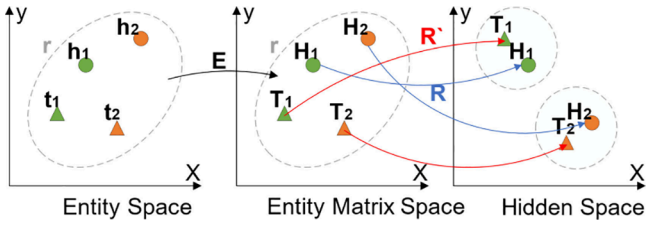


Figure 6. Illustration of MQuadE, where the matrix pair ⟨**R**,**R**′⟩ represents the relation.

2.1.2. Gaussian-Based Models. Models such as TransE ensure that positive triples in a KG have higher scores than negative triples by optimizing a global loss function. Nevertheless, these models overlook the (un)certainty of entities and relationships.[14] In fact, different entities and relations may possess varying degrees of (un)certainty.[14] For example, the more relations an entity is related to, the more specific that entity becomes, so an entity containing fewer relations has higher uncertainty. Similarly, a relation linking more entities has higher uncertainty. To address this issue, some research has considered modeling entities and relations as random variables. KG2E[14] utilizes Gaussian distributions to explicitly models the (un)certainty of entities and relations. The model can model 1-to-N, N-to-1, and reflexive relations, and performs well in link prediction and triple classification.

Xiao et al. pointed out the issue of multiple semantic relationships in KGEs, where a relation in a KG may have multiple semantics for the corresponding entity pairs and needs to be represented as a mixture of Gaussian distributions. To address this issue, they proposed TransG.[15] TransG can automatically learn the latent semantics of a relation and embed factual triples. As shown in Figure 7, in the traditional model illustrated in Figure 7(a), the semantics of all relationships are assumed to be the same, making it difficult to effectively distinguish between correct and incorrect triples. However, in the TransG model depicted in Figure 7(b), which takes into account multiple semantic relationships, these relationships can be distinguished from each other, allowing for more accurate and effective embedding of the triples.

2.1.3. Rotational Models. Inspired by Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, Sun et al. proposed a rotational model called RotatE.[16] In a complex vector space, the model represents relationships as rotations from the head entity to the tail entity,

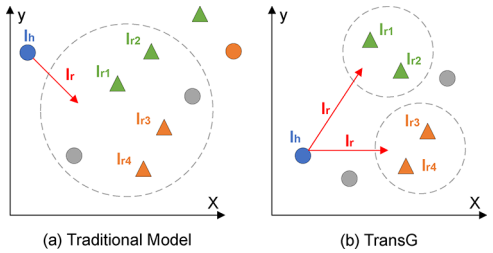


Figure 7. Illustration of traditional model and TransG, where $\mathbf{I}_h$ denotes the head entities, $\mathbf{I}_r$ denotes the relations, and $\mathbf{I}_r\bullet$ denotes the result of the head entity mapped by relation vectors with different semantics. The triangles represent correct tail entities, and the circles represent incorrect tail entities. (a) Traditional models cannot distinguish between relational semantics. (b) TransG maps head entities to different regions for different semantics of relations.

as shown in Figure 2(c). When entities and relationships are mapped to a complex vector space, the model can capture directional and phase information on entities and relationships, which enables the model to better handle complex relations, such as relations with directionality and temporality. RotatE can effectively handle symmetric, antisymmetric, inverse, and composite relations and demonstrates advanced performance in many applications.

However, one issue with RotatE is that it learns entity and relation embeddings from individual triplets at each step, without leveraging additional local information in the graph to improve the embeddings. To address this problem, Ma et al. proposed a RotatE-based embedding model called RotatSAGE.[17] Compared to RotatE, RotatSAGE can leverage local structural information in the graph to enhance embedding learning, thereby improving the performance of RotatE. Ma et al. also proposed a sampling strategy to further eliminate redundant entity information and simplify the proposed model.

QuatE[18] extends the space of complex vectors into a space of hypercomplex values, capturing the potential interdependencies between head entities and relations in the four-dimensional space. QuatE offers greater expressive rotation capabilities compared to RotatE, and can effectively model symmetric, antisymmetric, and inverse relations. Moreover, this framework is an extension of ComplEx (introduced in Section 2.2.1) in the hypercomplex space, providing better geometric interpretation.

However, existing hypercomplex embedding models such as QuatE only utilize embeddings of head entities, tail entities, and relationships to compute the score of a triplet, failing to fully capture the associations between head and tail entities. Distance-based models like TransR partially address this issue by associating each relationship with a translation matrix, but this significantly increases the number of model parameters. To tackle these problems, Nguyen et al. proposed a simple yet effective embedding model called QuatRE.[19] QuatRE further utilizes two relationship-aware rotations to obtain embeddings of head and tail entities, and simplifies the commonly used translation matrices in distance-based models, greatly reducing the computational overhead. QuatRE extends the QuartE model, capturing the associations between head and tail entities more comprehensively.

Previous methods like TransH aimed at solving the problem of complex relation representation, but they could only model symmetric and antisymmetric relations. RotatE achieved excellent performance in modeling symmetric, antisymmetric,

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

inverse, and composite relations, but predicting complex relations remains a challenge. Therefore, Chao et al. proposed PairRE,[20] as shown in Figure 8, which can simultaneously

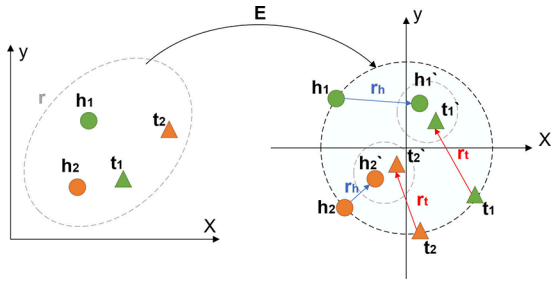


Figure 8. Illustration of PairRE. All entities are projected onto the unit circle in entity space and then are projected to different regions by relation vectors.

encode complex relations and multiple relation patterns. PairRE uses a pairwise relation representation method to enable the loss function to adapt to different complex relations. Additionally, this model can effectively capture the semantic relationships between relation vectors, enabling modeling of important relation patterns like symmetric, antisymmetric, inverse, and composite relations. By adding simple constraints on relation representations, relation pairs can further encode subrelations.

It should be noted that although QuatE and RotatE both utilize the concept of geometric rotation, QuatE belongs to the semantic matching model, and RotatE belongs to the translation model. For convenience of classification, they are all referred to as rotation models in this paper.

2.2. Semantic Matching Models. Semantic matching models mainly use similarity-based scoring functions to measure the plausibility of a fact by performing semantic matching. This process typically employs a multiplication formula $\mathbf{h}^T \mathbf{M}_r \approx \mathbf{t}$, where $\mathbf{M}_r$ is the mapping matrix of the relation. The geometric meaning of multiplying a vector with a matrix is to transform a vector into another vector by a linear transformation, so the head entity **h** is transformed and close to the tail entity **t**.

2.2.1. Linear/Bilinear Models. Linear/bilinear models project head entities into the representation vector space close to tail entities and represent the relation as a linear/bilinear mapping.

Latent factor model (LFM)[21] applies a relation-specific bilinear transformation to measure the correlation between entities and relations, which simply implements the distributed representation of head and tail entities. However, LFM requires a great number of parameters to participate in modeling, which limits its performance. Therefore, Yang et al. proposed the DistMult,[22] as shown in Figure 9(b), which reduces the parameters in LFM by restricting the mapping matrix to a diagonal matrix. Although DistMult can handle symmetric relations effectively, it cannot express antisymmetric relations.

ComplEx[23] introduces a complex vector space to extend DistMult, and can model symmetric as well as antisymmetric relations. In this model, the embeddings of entities and relations are represented into the complex space.

Given that most traditional factorization models face the problem of large memory consumption, Kishimoto et al.[24] proposed a model to binarize the parameters of CANDE-

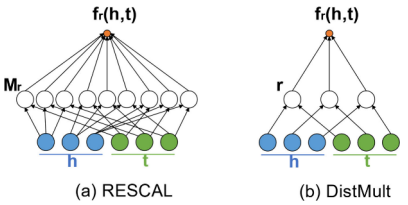


Figure 9. Illustration of RESCAL and DistMult. (a) RESCAL computes the score function using matrix multiplication. (b) DistMult computes the score function using the Hadamard product.

COMP/PARAFAC (CP) tensor decomposition to significantly reduce the memory usage. While ensuring the performance of Knowledge Graph Completion (KGC) tasks, the model size was successfully reduced by an order of magnitude. Kazemi et al. proposed SimplE,[25] an enhanced CP model, which solved the independence problem between the two embedding vectors of entities and demonstrated excellent performance while simplifying the CP decomposition method. As shown in Figure 10, SimplE can capture symmetric, antisymmetric, and inverse relations.

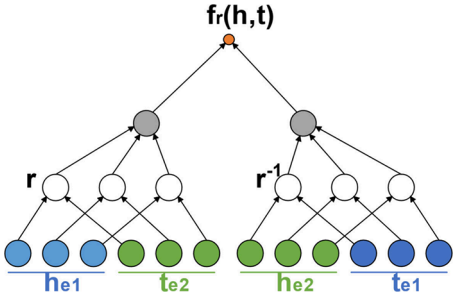


Figure 10. Illustration of SimplE.

The aforementioned DistMult, ComplEx, and SimplE models are canonical methods of linear/bilinear encoding and can also be regarded as extensions of some tensor decomposition models.

2.2.2. Tensor Decomposition-Based Models. In the tensor decomposition model, a triple is represented as a third-order tensor. The primary objective of tensor decomposition is to decompose the tensor into the product of three low-rank matrices. The three matrices correspond individually to the embedding representation of the head entity, relation, and tail entity. This approach can capture high-order relationships and complex structures in the data, especially for large-scale high-dimensional data. As shown in Figure 9(a), RESCAL[26] is a classical model based on tensor decomposition, which can naturally model high-order relations and has strong generalization ability.

TuckER[27] employs element-wise tensor products to capture complex interactions between entities and relations, as illustrated in Figure 11, and embraces a parameter-sharing strategy to enhance the efficiency of the learning process. TuckER is fully expressive, and the previous mentioned ComplEx, DistMult, SimplE, and RESCAL can be viewed as particular cases based on the TuckER decomposition.

Luo et al. proposed BTDE,[28] in Figure 12, which improves the Tucker model to get more accurate entity and relation embedding vectors. The model is also fully expressive and can generate low-dimensional interpretable embeddings. BTDE

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

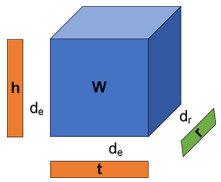


Figure 11. Illustration of TuckER.

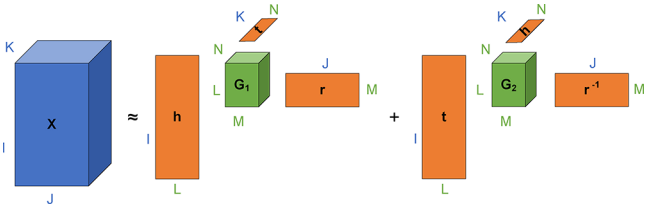


Figure 12. Illustration of BTDE.

performs well in link prediction tasks and can model symmetric, asymmetric, and transitive relations.

2.3. Neural Network-Based Models. The neural network-based model uses embeddings of entity and relation as inputs to measure the plausibility of factual triples by calculating and outputting the probability of the factual triple. The neural network utilizes nonlinear activation functions and more complicated network structures to encode relational data.

2.3.1. Convolutional Neural Networks. Convolutional Neural Networks (CNNs) learn nonlinear features and capture complex relationships with fewer parameters, as illustrated in Figure 13, while also being able to learn deep representations.

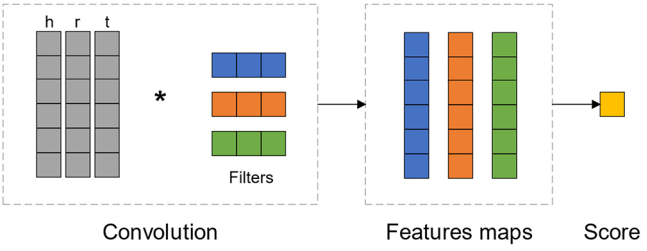


Figure 13. Illustration of CNN-based models.

ConvE[29] is the first neural link prediction model to use two-dimensional convolutional layers, which uses convolutional layers and fully connected layers to model the interactions between entities and relations. ConvE has high parameter efficiency, achieving the similar performance as DistMult and R-GCN while reducing parameters by 8 and 17 times, respectively. Additionally, ConvE is particularly effective in modeling complex KG.

Nguyen et al.[30] argued that ConvE ignores the global relationships between entities and relations in the same dimension. Therefore, they designed ConvKB,[30] which uses CNN to encode the concatenation of entities and relations, to retain global relationships and transitional features as well as enhance the learning ability of potential features. Compared to TransE, it further models the global relationship.

Using convolutional networks can potentially enhance interactions and capture implicit and latent features, but deeper structures may lead to the loss of surface knowledge and an excessive number of parameters. To strike a balance between performance and cost, Zhou et al. proposed a new convolution-based KGE model called JointE.[31] The model not only significantly reduces redundant parameters but also maximizes interactions, thereby thoroughly capturing latent and implicit knowledge. JointE demonstrates remarkable generalization capability and robustness.

2.3.2. Graph Neural Networks. Distance-based methods such as TransE, DistMult, and RotatE, as well as semantic matching methods, are generally referred to as KGC methods. Triples in KGC methods are independently processed in the objective function. As a result, these KGC methods lack the ability to strengthen local or global smoothness in the embedding space using graph structures. To address this issue, embedding models can use nonlinear transformations to encode the representation of KGs into vector spaces by introducing neural network architectures. Graph neural networks (GNNs) can discover potential long-range correlations between nodes, making them a suitable architecture for KRL, as shown in Figure 14(a). Lin et al. proposed KGNN,[32] an end-to-end model that extends space-based GNN methods to KGs by selectively aggregating multiple neighborhood

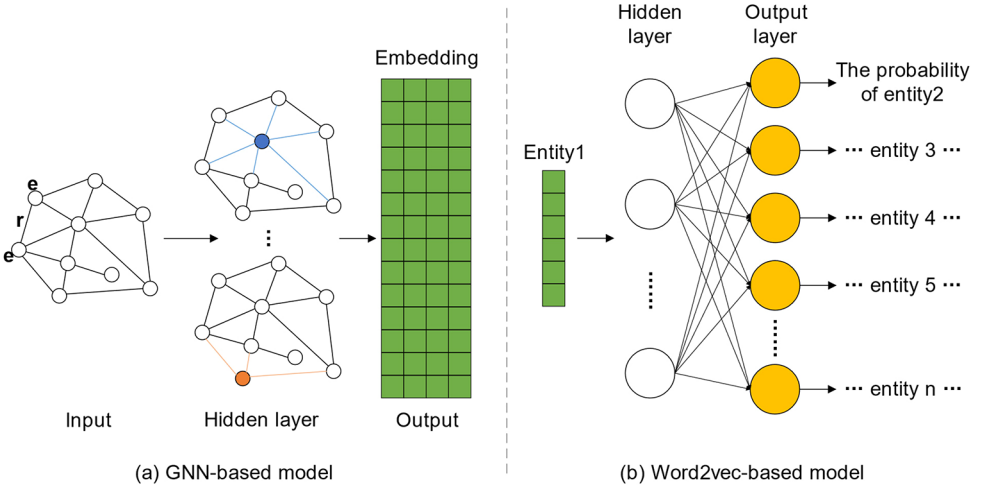


Figure 14. Architectures of GNN-based model and Word2vec-based model. (a) GNN-based models capture local features and global information by aggregating information from neighbor nodes. (b) Word2vec-based models exploit the co-occurrence relationship of neighboring nodes to capture local semantic similarity.

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

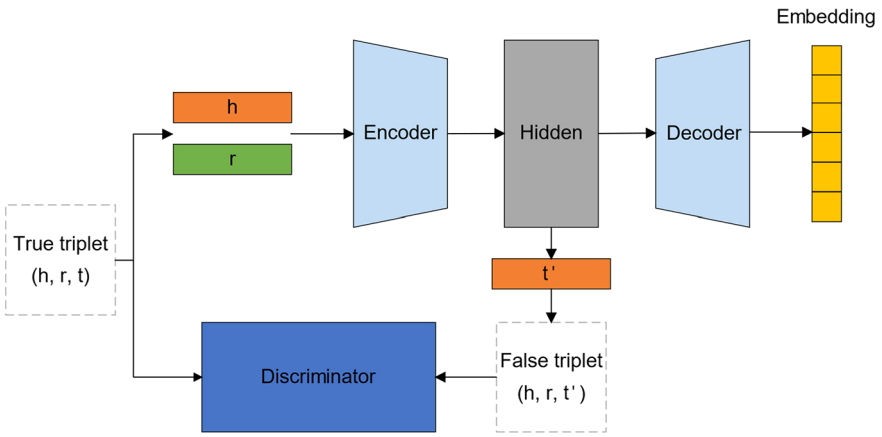


Figure 15. Architecture of adversarial autoencoder.

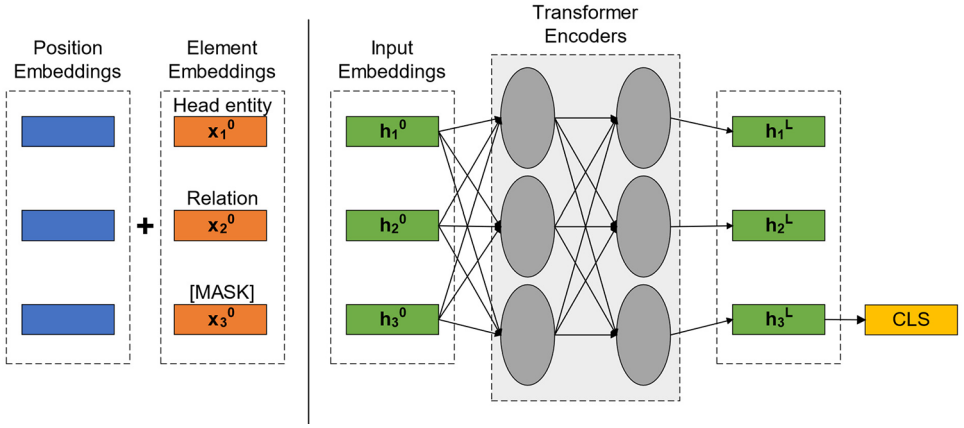


Figure 16. Architecture of transformer-based model.

information. It can learn both the topological structure information and semantic relationships of the KG.

Schlichtkrull et al. first demonstrated that GCN can be applied to relationship data modeling, especially for entity classification and link prediction tasks, and proposed R-GCN.[33] They applied R-GCNs to multiple graphs with a large number of relationships by introducing parameter sharing and enforcing sparse constraints. In entity classification tasks, R-GCN can be served as an efficient encoder. In link prediction tasks, R-GCN utilizes DistMult decomposition for decoding and outperforms the directly optimized decomposition models.

To take full advantage of GCN models and KGC methods, Yu et al. proposed the KE-GCN[34] framework, which utilizes various KGE techniques to update embeddings of entities and relations through graph convolution operations. It performs well in entity classification and KG alignment tasks. Yu et al. also demonstrated that R-GCN can be viewed as a special case of KE-GCN.

However, existing GNN-based approaches have overlooked the directionality of relationships. Additionally, they have not taken into account the influence of neighborhood on relationship representation learning, thereby limiting the ability to represent relationships. D-AEN[35] aggregates neighborhood information to simultaneously learn representations of both relationships and entities, while utilizing entity and relationship representations to facilitate each other's learning. Specifically, the model employs a bidirectional attention mechanism to incorporate the directionality of relationships into entity embedding learning, capturing the semantic information on neighboring entities when they are respectively acting as head or tail entities. Furthermore, the model utilizes relationship-specific attention mechanisms to fully consider the influence of the neighborhood on relationship representation learning. D-AEN demonstrates superior performance in link prediction tasks.

2.3.3. Other Models. Word2vec[36] is a classical word embedding method, as shown in Figure 14(b), which uses one entity as input and outputs multiple entities. Vectorizing words can solve some sparsity problems and capture the contextual information on words. Word2vec can be used directly as a prediction model, or it can be applied to obtain the word vector matrix generated during model training, which contains the contextual information on words.

Dai et al.[37] first introduced adversarial autoencoders into knowledge graph embedding learning. As shown in Figure 15, the adversarial autoencoder utilizes the generated latent vectors as negative samples, i.e., the representation of negative triples. The discriminator uses positive samples and generated negative samples to train various KGE models. ComplEx, SimplE, RotatE, and other models have shown better results after being trained with adversarial autoencoders.

Considering that entities and relations may contain different properties in different contexts, Wang et al. proposed Contextualized Knowledge Graph Embedding (CoKE).[38] CoKE learns dynamic and contextualized entity and relation embeddings, which is a new paradigm that considers

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

contextual properties. The model takes a sequence as input and uses the Transformer[39] to obtain contextual embeddings, as shown in Figure 16, so the embeddings can capture the contextual meanings of entities and relations.

KG-BERT[40] is a novel method to using pretrained language models for KGC tasks. In this model, entities, relations and triples are considered as text sequences, so it converts KGC tasks into sequence classification problems. Then, the BERT model is fine-tuned on these sequences to predict the plausibility of relations or triples. KG-BERT performs outstandingly in link prediction, relation prediction, and triple classification tasks.

In addition to the triple format, KGs can also be represented in the form of web ontologies using the Web Ontology Language (OWL). Web ontologies contain richer semantic information and are widely utilized in the field of cheminformatics. To capture the semantic information on web ontologies, Chen et al. proposed the OWL2Vec[41] model. OWL2Vec is a robust semantic embedding framework specifically designed for OWL ontologies. It extracts three types of documents from the ontology that capture the ontology's structural information, logical information, and lexical information. These documents are then used to learn word embeddings using the Word2Vec model, resulting in the embedding representation of the web ontology.

2.4. Model Summary. This section introduces some classical KGE models and state-of-the-art KGE models, including distance-based models, semantic matching models, neural network-based models, and other models. Table 1 lists all the KGE models mentioned above and summarizes their respective model characteristics. For the first two types of KGE models, we primarily focus on comparing their capacity for relation modeling. As for the latter two types of models, we focus on understanding their characteristics and their proficiency in handling specific downstream tasks.

3. COMMON CHEMICAL DATABASES

In this section, widely used chemical databases were collected, providing key entity or relationship information for various chemical application scenarios: chemical substances, drugs, genes, proteins, interactions, and side effects, as shown in Figure 17. Some databases are used in more than one scenario, so we categorized them broadly and made statistics on their data information, as shown in Table 2.

3.1. Chemical Database. ChEMBL[94] is an open-access repository of bioactive drug-like small molecules. The database integrates information from literature, patents, and other public sources, encompassing drug bioactivities, drug metabolism, ADME (absorption, distribution, metabolism, and excretion) properties, and drug toxicity data. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. ChEMBL has evolved into a repository for numerous data sources and types, offering information pertinent to chemical biology and all phases of drug discovery.

PubChem[95] is a widely used and freely accessible chemical substance database that provides comprehensive literature reports and rich physical and chemical properties of molecular materials. It provides comprehensive details including chemical and physical properties, biological activities, safety and toxicity data, patents, and literature citations. While predominantly focusing on small molecules, PubChem also includes larger compounds like nucleotides, carbohydrates, lipids, peptides, and chemically modified macromolecules. The database compiles extensive information on chemical structures, identifiers, properties, biological activities, patents, health effects, safety profiles, toxicity data, and more.

Reaxys[96] is a comprehensive database designed for chemical research and discovery. It provides access to a vast collection of experimentally validated chemical substance and reaction data, including chemical substance structures, properties, chemical reaction data, compound synthesis pathways, biological activity data as well as medicinal chemical and pharmacological data.

CCCBDB (Computational Chemistry Comparison and Benchmark DataBase)[97] provides experimental and quantum mechanics-computed thermochemical data for a curated selection of 2186 gas-phase atoms and small molecules. It offers tools to compare ideal-gas thermochemical properties between experimental and computational results. The database includes enthalpies of formation, entropies, integrated heat capacities, and essential data for computing thermochemical properties such as molecular geometries, rotational constants, vibrational frequencies, barriers to internal rotation, and electronic energy levels. Additionally, CCCBDB provides computed properties like atomic charges, electric dipole moments, quadrupole moments, polarizabilities, and HOMO−LUMO gaps, among others.

ZINC[98] serves as a comprehensive compound repository designed for virtual screening and drug discovery, aiming to provide scientists and researchers with an extensive array of compound information to support biomedical research. Users can freely access and retrieve structural information on these compounds through the ZINC database Web site or related interfaces. ZINC also permits users to purchase specific compounds, enabling more in-depth research and experimentation.

QMx is A database composed of a number of data sets used in different chemical tasks. QM7[99] provides the Coulomb matrix representation of molecules, their atomization energies, and a large variety of molecular structures. QM7b is an extension of the QM7 data set for multitask. QM8[100] is a data set for modeling the electronic spectra and excited state energies of small molecules. QM9[100] contains spatial information, energy, electronic and thermodynamic properties of molecules. It is widely used to test and compare various data-driven molecular property prediction methods.

3.2. Drug Database. DrugBank[89] is a unique resource in the field of bioinformatics and cheminformatics, combining detailed drug data with comprehensive drug target information. It systematically collects and manages extensive drug-related information from various sources, including literature, FDA-approved drug labels, and clinical trial data. DrugBank provides drug information, drug target data, drug−drug interactions, drug-food interactions, drug indications, drug metabolism pathways, and spectral and chromatographic data, etc. DrugBank continuously expands its data repository and enhances data quality, making it a widely utilized tool in biomedical research and clinical applications.

DrugCentral[90] is a pharmaceutical information database that provides information on active ingredients, chemical entities, drugs, indications, and other drug-related information. The database was updated in 2023 to expand its data content. In addition to enhancing existing data, it also added 396 veterinary drugs and related data, opening new directions and opportunities for preclinical research.

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Summary of KGE Models

category	model	description	modeling capability	year
Distance-based model	TransE[7]	Classical and simple.	Injective, Inverse, Composite	2013
	TransH[8]	Extending TransE, it can capture the multiple semantics of entities.	Injective, Noninjective, Reflexive	2014
	TransR[9]	Extending TransE, it can capture complex relations.	Injective, Noninjective, Reflexive	2015
	TransA[10]	Improving the loss metric of translation-based model, it enhances the embedding quality	Injective, Noninjective, Reflexive, Symmetric/Antisymmetric	2015
	TransD[11]	Extending TransR, it achieves higher embedding quality with fewer parameters.	Injective, Noninjective, Reflexive, Symmetric/Antisymmetric	2015
	TrasnsOWL/ TransROWL[12]	Improving TransE/R, it enhances the embeddings.	Injective, Noninjective, Reflexive	2021
	MQuadE[13]	It can capture more complex relations.	Injective, Noninjective, Symmetric/Antisymmetric, Inverse, Composite	2021
	KG2E[14]	It can capture (un)certainty of entities and relations.	Injective, Noninjective, Reflexive	2015
	TransG[15]	It can capture the multiple semantic relations.	Injective, Noninjective, Reflexive	2015
	RotatE[16]	It can capture the directivity and phase information of relations.	Injective, Symmetric/Antisymmetric, Inverse, Composite	2019
	RotatSAGE[17]	Improving RotatE, it reduces the computational overhead.	Injective, Noninjective, Symmetric/Antisymmetric, Inverse, Composite	2022
	QuatE[18]	Extending RotatE, it can capture the latent relations.	Injective, Noninjective, Symmetric/Antisymmetric, Inverse	2019
	QuatRE[19]	Improving QuatE, it reduces the computational overhead.	Injective, Noninjective, Symmetric/Antisymmetric, Inverse	2022
	PairRE[20]	It can capture the relationships between relations, and can model more relation patterns.	Injective, Noninjective, Symmetric/Antisymmetric, Inverse, Composite	2020
Semantic matching model	LFM[21]	Classical bilinear model.	Injective, Noninjective	2012
	DistMult[22]	It reduces the number of parameters, and can model symmetric relations.	Injective, Symmetric, Composite	2014
	ComplEx[23]	Extending DistMult, it can model antisymmetric relations.	Injective, Noninjective, Symmetric/Antisymmetric, Composite	2016
	Kishimoto et al.[24]	It reduces the computational overhead of the CP decomposition model.	Injective, Noninjective, Symmetric, Composite	2019
	SimplE[25]	Improving CP decomposition model, it can capture complex relations.	Fully Expressive	2018
	RESCAL[26]	It is the classical tensor factorization model, and can capture higher-order relations.	Injective, Noninjective, Reflexive	2011
	TuckER[27]	Extending RESCAL, it can capture more complex relational interactions.	Fully Expressive	2019
	BTDE[28]	Improving TuckER, it enhances the embeddings.	Fully Expressive	2020
Neural network-based model	ConvE[29]	It can model complex KG with high parameter efficiency.	* Neural network models, as well as other models such as large language models, perform excellently in many downstream tasks. However, due to the black-box nature of these models, there are no mathematical formulas that can prove their capacity for modeling relationships.	2018
	ConvKB[30]	It can capture global relational features and latent features.		2018
	JointE[31]	It reduces the number of parameters, and can capture latent features.		2022
	KGNN[32]	It can capture topological structure information, and enhances the embeddings.		2020

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. continued

category	model	description	modeling capability	year
Other models	R-GCN[33]	It can model complex KG, and improves the computational efficiency.		2018
	KE-GCN[34]	Extending R-GCN, it enhances the embeddings.		2021
	D-AEN[35]	It aggregates neighborhood information and can capture the directionality of relations.		2023
	Word2vec[36]	It can solve part of the sparsity problem and capture the context information.		2013
	Dai et al.[37]	It can enhance the embeddings after fusing other KGE models.		2021
	CoKE[38]	It can capture the differences between entities or relations in different contexts.		2019
	KG-BERT[40]	It enhances the embeddings with rich text information.		2019
	OWL2Vec[41]	It can capture the semantic information of the web ontology.		2021

BindingDB[91] is a widely used biomolecular interaction database that focuses on the interactions of proteins considered to be candidate drug-targets with ligands that are small, drug-like molecules. It offers measured binding affinities, facilitating medicinal chemistry and drug discovery through literature awareness and the development of structure−activity relationships (SAR and QSAR). BindingDB also assists in validating computational chemistry and molecular modeling techniques, including docking, scoring, and free energy methods. Additionally, it supports chemical biology, chemical genomics, and the study of molecular recognition's physical chemistry. The database features a collection of host−guest binding data valuable to researchers studying supramolecular systems.

repoDB[92] is a drug repositioning database that includes a set of standards for successful and unsuccessful drug repositioning. This means it contains information on both approved and failed drugs, along with their indications. The database is designed to benchmark and test computational drug repurposing methods in a fair and reproducible manner. Approved indications in repoDB are sourced from DrugCentral, while failed indications are obtained from the AACT database.[108]

SuperTarget[93] is a biopharmaceutical target database that focuses on studying the interactions between drugs and protein targets. It integrates drug-related data, including indications, adverse drug reactions, metabolism, pathways, and target protein gene ontology terms. The core data set currently consists of over 6,000 targets, 19,000 drugs and presumed drugs (about 2,500 of which are approved drugs), and 33,000 drug-target interactions.

3.3. Gene and Protein Database. Uniprot[79] is a protein database that offers comprehensive, high-quality protein sequence information accessible to users for free. It collects extensive protein sequence and functional annotation information. The database continually updates and supplements its entries by gathering knowledge from literature. Currently, it includes information on over 2.27 million protein sequences. Additionally, Uniprot integrates related biological data resources by indexing them, making it a central hub of biological knowledge.

Ensembl[80] is a freely accessible genome browser and annotation platform that offers high-quality annotations, tools, and services for vertebrate animals and model organisms. Currently, it supports over 1,700 annotations for more than 1,000 different species. The platform provides valuable data on genes, transcripts, variations, and more.

RNAcentral[81] is a free public resource that provides access to comprehensive and up-to-date collections of noncoding RNA sequences. Currently, the database offers 44 RNA resources and serves as a single access point for over 18 million ncRNA sequences from a wide range of organisms and RNA types. It also includes structural information for over 13 million RNA sequences. In addition to gene sequences, RNAcentral provides extensive annotation types such as genomic coordinates, microRNA-target interactions, gene ontology, homologues, and orthologs.

NCBI Gene[82] integrates various gene-related data resources and enables quick access to specific gene information, including gene-related sequences, structures, gene expression, and other related information. It consolidates diverse types of gene-related data into single entries, including gene sequences,

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

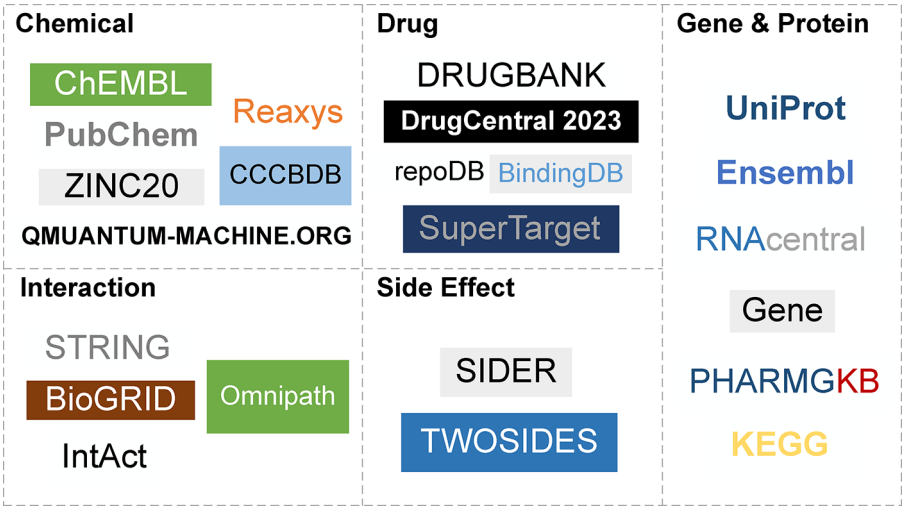


Figure 17. Widely used databases and their application scenarios.

Table 2. Common Databases in the Field of Chemistry[a]

category	database	statistics	open source	year	update frequency
Chemical	ChEMBL	1.6 M assays, 2.4 M compounds, 89K documents, 15K targets, 6.9K mechanisms, 48K indications, 15K drugs.	√	2009	1–6 Weeks
	PubChem	118 M compounds, 319 M substances, 295 M bioactivities, 41 M literature, 51 M patents.	√	2004	Daily
	Reaxys	275 M substances, 64 M reactions, 109 M documents, 41 M patents, 45 M bioactivities.	×	2009	-
	CCCBDB	2.1K gas-phase atoms and small molecules.	√	1998	-
	ZINC	1.4B compounds.	√	2005	-
	QMx	QM7:7K molecules made up of 23 atoms. QM7b: 7K molecules. QM9:134 K stable small organic molecules made up of CHONF. QM8:20K synthetically feasible small organic molecules.	√	2013–2016	-
Drug	DrugBank	26K publications, 500 K drug and drug products.	*	2006	3 Months
	DrugCentral	5K drugs, 152 K pharmaceuticals.	√	2016	Annually
	BindingDB	2.9 M data for 1.3 M compounds and 9.3K targets.	√	1995	Weekly
	repoDB	1.5K drugs, 2K diseases.	√	2017	-
	SuperTarget	6K targets, 19K compounds, 330 K interactions.	√	2008	-
Gene and Protein	Uniprot	570 K reviewed sequence entries and 244 M unreviewed sequence entries.	√	2003	8 Weeks
	Ensembl	Genome information and annotations for 324 species.	√	1999	3 Months
	RNAcentral	35 M sequence entries from 53 expert databases.	√	2014	3–6 Months
	NCBI Gene	50K taxa, 53 M genes.	√	2003	Daily
	PharmGKB	1K drug label annotations, 5K clinical annotations, 27K variant annotations, 0.8K annotated drugs.	√	2000	Monthly
	KEGG	1.1 M pathways, 54 M genes, 19K compounds, 12K drugs, 56K drug lables.	√	1995	1–3 Months
Interaction	STRING	12K organisms, 59MM proteins, 27B interactions.	√	2003	Annually
	BioGRID	85K publications, 2.7 M protein and genetic interactions, 31K chemical interactions, 1.1 M post translational modifications.	√	2003	Monthly
	IntAct	23K publications, 143 K interactors, 844 K interactions, 1.5 M binary interactions.	√	2003	Monthly
	OmniPath	Over 100 resources from 5 databases.	√	2016	1–6 Months
Side Effect	SIDER	1.4K drugs, 139 K drug-side effect pairs.	√	2010	-
	TWOSIDES	Over 0.8 M significant associations between 59K pairs of drugs and 1.3K adverse events.	√	2013	-

[a]"-" indicates that the update has stopped or the update frequency is uncertain. "*" indicates that the database is partially open source. Note that these databases also contain other data not mentioned because of incomplete data statistics.

structures, gene expression, nomenclature, publications, interactions, and other pertinent information.

PharmGKB[83] is a resource dedicated to pharmacogenomics knowledge, providing information on drug metabolism, drug responses, and related genes. Notably, PharmGKB includes extensive clinical information such as clinical guidelines and drug labels, highlighting potential clinically actionable gene-drug associations and genotype-phenotype relationships.

KEGG[84] is a comprehensive bioinformatics database resource aimed at understanding advanced functions and utilities of biological systems (such as cells, organisms, and ecosystems) from molecular-level information. It provides information on genes, proteins, metabolic pathways, diseases, drugs, and more, particularly encompassing large-scale molecular data sets generated from genome sequencing and other high-throughput experimental technologies.

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

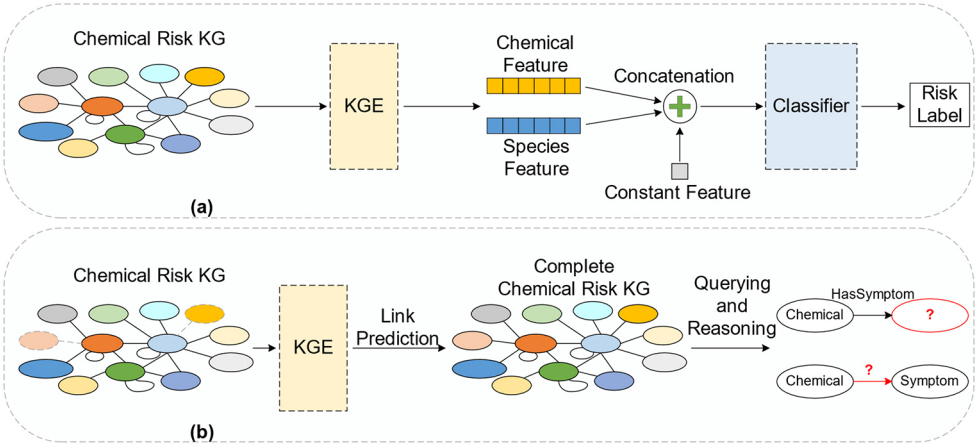


Figure 18. Frameworks for the application of KG and KGE technique to chemical risk assessment tasks. (a) A chemical hazard prediction framework utilizing the KGE model where constant feature can be specific concentrations of chemicals. (b) A framework to retrieve chemicals based on symptoms.

3.4. Interaction Database. STRING[85] primarily focuses on information regarding protein−protein interactions. These interaction data are systematically integrated and summarized from various sources, including scientific literature, experimental databases, and interaction prediction results. The database encompasses both known and predicted interactions, covering both physical and functional interactions. For known proteins within the database, STRING also provides predictions of interaction partners, serving as a valuable reference for further research.

BioGRID[86] is a widely used biological interaction database designed to store, integrate, and present data on gene and protein interactions. Its goal is to facilitate biomedical discoveries, particularly those related to human health. The interaction data in BioGRID are manually curated from experimental evidence in biomedical literature, encompassing both focused low-throughput studies and extensive high-throughput data sets. Additionally, BioGRID also collects interactions between proteins or genes and small molecules.

IntAct[87] is a comprehensive protein interaction database. It primarily focuses on analyzing interactions between proteins and other omics entities. IntAct primarily collects information from scientific literature and direct experimental data, aiming to provide relevant experimental details for the involved interactions. Notably, in addition to general protein data, IntAct also collects a mutation data set that describes the impact of minimal sequence variations on protein interactions.

OmniPath[88] is a comprehensive signaling and metabolic pathway database that stores, integrates, and presents molecular interaction relationships within cellular signaling networks. OmniPath collects data from over 100 resources, encompassing protein interactions involved in intercellular and intracellular signaling, as well as transcriptional and post-transcriptional regulation. The resources within OmniPath primarily include molecular interactions, enzyme-PTM relationships, protein complexes, and molecular annotations.

3.5. Side Effect Database. SIDER[101] focuses on marketed medicines and their recorded adverse drug reactions. The data in SIDER is sourced from publicly available documents and drug packaging. The information includes details on drugs, the frequency of adverse effects, classifications of these effects, and indexed links to additional information such as drug-target relationships.

TWOSIDES[102] is a resource focused on various drug side effects, including information on the associations between certain drug pairs and adverse reactions. Notably, TWOSIDES emphasizes multidrug side effects, thus the collected association information is limited to those that cannot be attributed to any single drug.

Although we have categorized these databases by their application domains, in reality, these databases are highly integrated, with some databases covering data from multiple fields. The significant data types included in each database are organized in this chapter, which can be used to determine the various application domains of the databases. According to incomplete statistics, there is overlap between the data in some databases. For example, part of the data in DrugBank comes from PubChem, KEGG, and UniProt, while some data in DrugCentral is sourced from Uniprot and ChEMBL. Besides overlapping data, databases often include cross-references to other databases in their entries to create more comprehensive and rich data entries. Users can click these links to access more detailed data in the corresponding databases. Additionally, most databases have their own data identifiers for indexing different data. However, some databases, such as rerpoDB, STRING, BioGRID, IntAct, OmniPath, TWOSIDES, and QM, do not have unique identifiers and instead use identifiers from other databases. Most databases allow data retrieval using data names or unique identifiers on their homepages, and some databases also support direct data retrieval using identifiers from other databases. We have compiled statistics on the data overlap and cross-referencing among these databases, which are detailed in the Supporting Information.

4. APPLICATIONS OF CHEMICAL KNOWLEDGE GRAPHS AND KNOWLEDGE GRAPH EMBEDDING MODELS

In this section, we primarily focus on introducing application cases of KGs and KGE models in the field of chemistry. We summarize the KGE models and chemical KGs constructed or utilized in the presented cases at the end of this section.

4.1. Chemical Risk Assessment. Assessing the toxicity of compounds can effectively reduce the time and money costs for in vitro or in vivo studies. Additionally, risk assessment can screen out potentially toxic compounds at an early stage. The KG is a framework that allows for rapid screening and

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

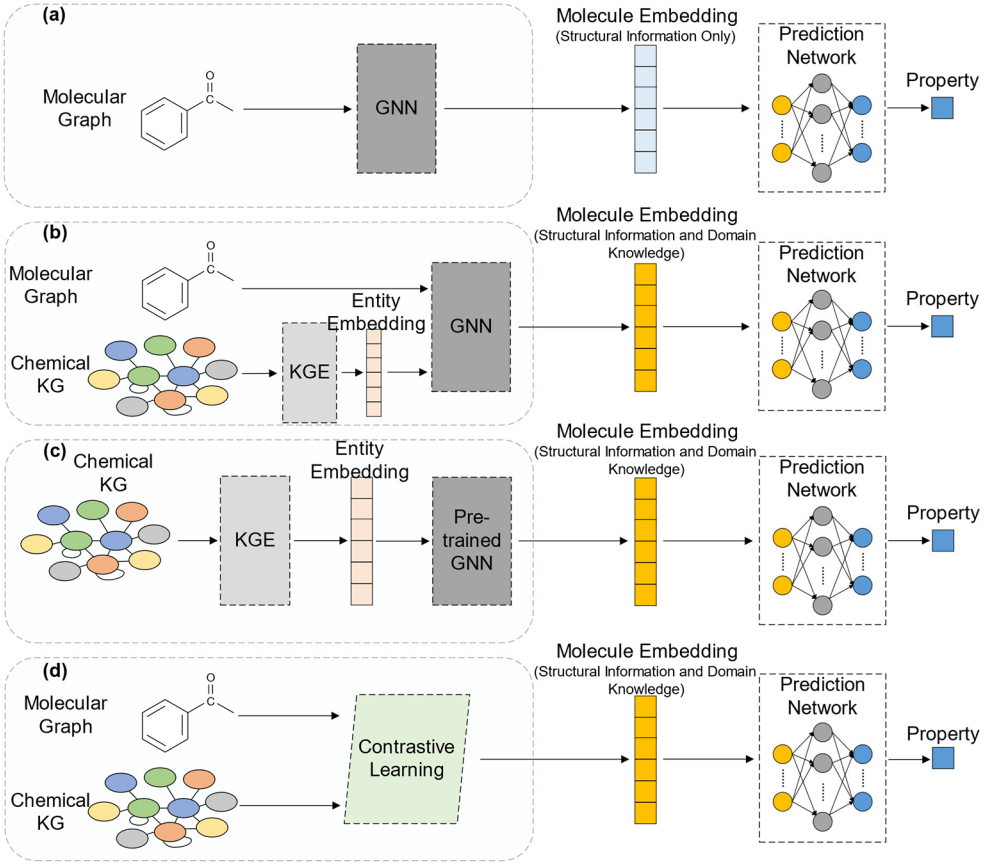


Figure 19. Frameworks for the application of KG and KGE technique to compound property prediction tasks. (a) A general GNN-based framework for property prediction, which can only capture the structural information on molecules. (b) Utilizing KGE methods, domain knowledge of molecules can be effectively captured, which is used to guide GNN-based models to learn chemical semantics. (c) For any pretrained GNN models, KGE methods can be used for fine-tuning to further enhance the learning and representation of chemical semantics. (d) A contrastive learning framework utilizing KGE method can be constructed by treating a molecular graph and its KGE-enhanced counterpart as positive samples.

assessment of the potential toxicity of compounds to organisms and environment. Additional information about specific compounds can be also implemented in the KG.[42] Therefore, compared with traditional databases, KG is more integrated, provides better retrieval efficiency, and is gradually applied to chemical risk assessment tasks. The general application framework of KG is shown in Figure 18.

Based on the primary data information commonly used in ecotoxicological risk assessments tasks, Myklebust et al.[42] created a toxicology effects and risk assessment KG (TERA KG). They designed a prediction model on the KG as a baseline model. The node embeddings outputted from KGE were then used to assess chemical toxicity and infer adverse biological effects on organism on the baseline. Nine node embedding models were evaluated on the KG, and the results showed that using KGE can improve the accuracy of model predictions on a simple baseline.

Zheng et al.[43] developed a deep learning-based entity recognition system to collect hazardous chemical data in unstructured documents and integrated the data in a KG, which links knowledge and information in isolated databases. The unified system they created improved the hazardous chemical management.

Shin et al.[44] selected 1001 chemicals with high risk and collected knowledge about their symptoms and structures. They then systematically integrated this complex connected knowledge to construct SEARCH-KG for chemical substance diagnosis. Three models include TransE, ComplEx, and ConvKB were used to verify the stored knowledge for knowledge graph completion. Furthermore, they proposed the SEARCH system to implement symptom-based chemical substance judgment using KG querying and reasoning techniques.

4.2. Prediction of Compound Properties. Accurate prediction of molecular properties can help identify compounds with inadequate physical and chemical properties early on. Predicting the properties of lead compounds could decrease the chances of clinical trial failures and improve the probability of successful drug development. Traditional methods of predicting molecular properties typically require a significant amount of experimental data, which are costly and time-consuming. Numerous researches have demonstrated the enormous potential of machine learning techniques, especially deep learning, for predicting molecular properties. To enhance drug development efficiency and reduce costs, these studies adopt sequence or graph structures to represent molecules, and employ sequence modeling or GNNs to predict molecular properties. Due to the vast and complex nature of chemical space, KGs have become excellent carriers for storing chemical data. The research on KGE method is beneficial for improving the representation of chemical space and facilitating the modeling of chemical rules. Therefore, as shown in Figure 19,

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

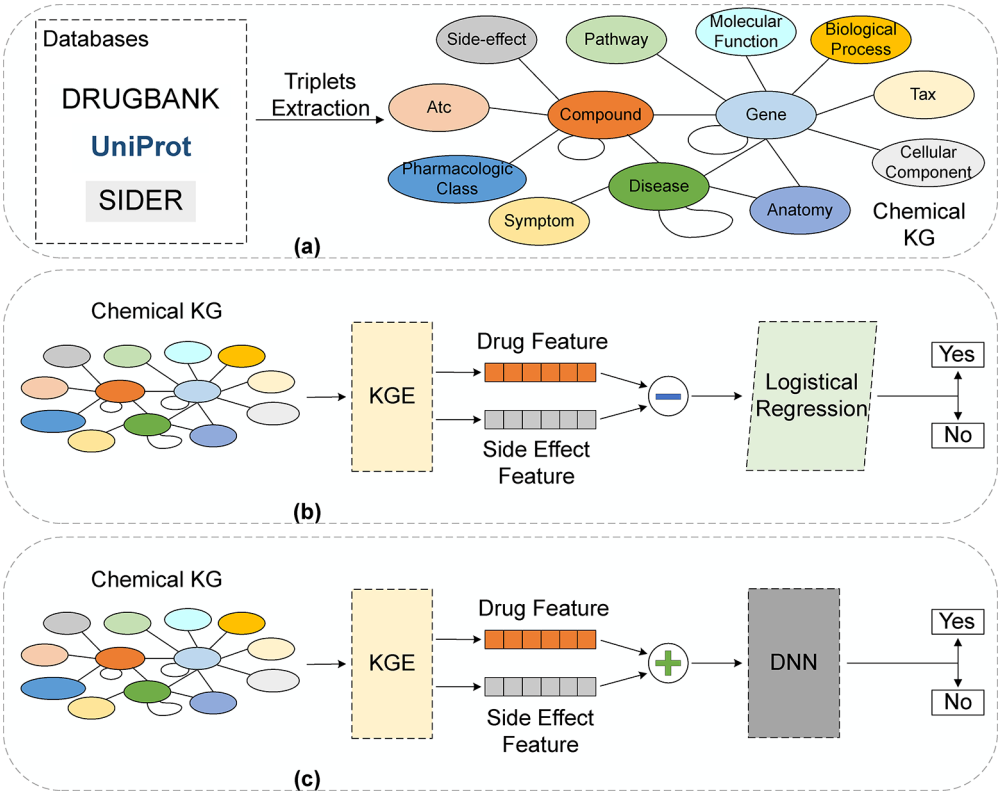


Figure 20. Frameworks for the application of KG and KGE technique to ADR prediction tasks. (a) Data are extracted from the chemical database to construct the required KG. (b,c) KGE is utilized to learn the representations of drug and side effect entities, respectively, which are then input to different prediction modules for prediction.

KGE technique has broad application prospects in the field of chemical property prediction.

Some supervised models have been used to learn molecular representations for predicting molecular properties, but these methods require a significant amount of labeled data. Graph contrastive learning can alleviate this problem, but current graph contrastive learning methods do not take into account fundamental domain knowledge and the correlations between atoms. To overcome these problems, Fang et al.[6] incorporated domain knowledge into molecular graph representation learning by constructing an element KG to model the microscopic relationships between chemical elements, and proposed a novel knowledge-enhanced contrastive learning (KCL) framework. KCL uses RotatE to capture the connections between chemical elements and obtain the molecular embedding representation. The model demonstrated excellent performance and is beneficial for downstream tasks such as molecular property prediction.

Hua et al.[45] proposed a multiview molecular property prediction framework that employs a knowledge-guided graph transformer pretraining algorithm named KPGT. To improve the accuracy of molecular property prediction, KPGT integrates chemical domain knowledge and further incorporates functional group information into the KG. From molecular graph to functional group to atom, the multiview perspective enhances the molecular representation from coarse to fine. Additionally, to accumulate chemical domain knowledge, Hua et al. proposed a novel BiLSTM-based recurrent module. By considering functional group information, molecular graph, and atomic physical and chemical properties in molecular property prediction, KPGT exhibits excellent performance.

Many existing methods for predictive tasks in the field of chemistry are data-driven and primarily focus on the molecular topology information. These methods lack chemical prior knowledge and rely heavily on data, in which hinders further generalization and application. To address this issue, Fang et al.[46] introduced chemical functional group information into a chemical element KG, and proposed KANO,[46] which enhances the KG using functional group information. They used OWL2vec as an embedded model to obtain embedded representations of entities and relationships. In addition, they incorporated functional prompts in the fine-tuning process. These prompts guide the model to acquire task-related knowledge for downstream tasks. The fusion of functional group information and element KG enables the model to capture chemical entities and relationships more effectively. KANO demonstrates superior performance in various chemistry prediction tasks and provides chemically plausible explanations for its predictions.

Xie et al.[47] constructed a knowledge graph of chemical reactions by treating reactants and products as nodes and reaction rules as edges, and further proposed a chemical reaction knowledge embedding framework ReaKE. ReaKE conducts constructive learning at both reaction and molecular levels, and jointly optimizes these two types of constructive losses to learn reaction-aware molecular embedding representations. Experimental results demonstrated that guided by prior knowledge of chemical reactions, this framework could extract high-quality reaction embeddings and molecular embeddings, exhibiting outstanding performance in downstream tasks such

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

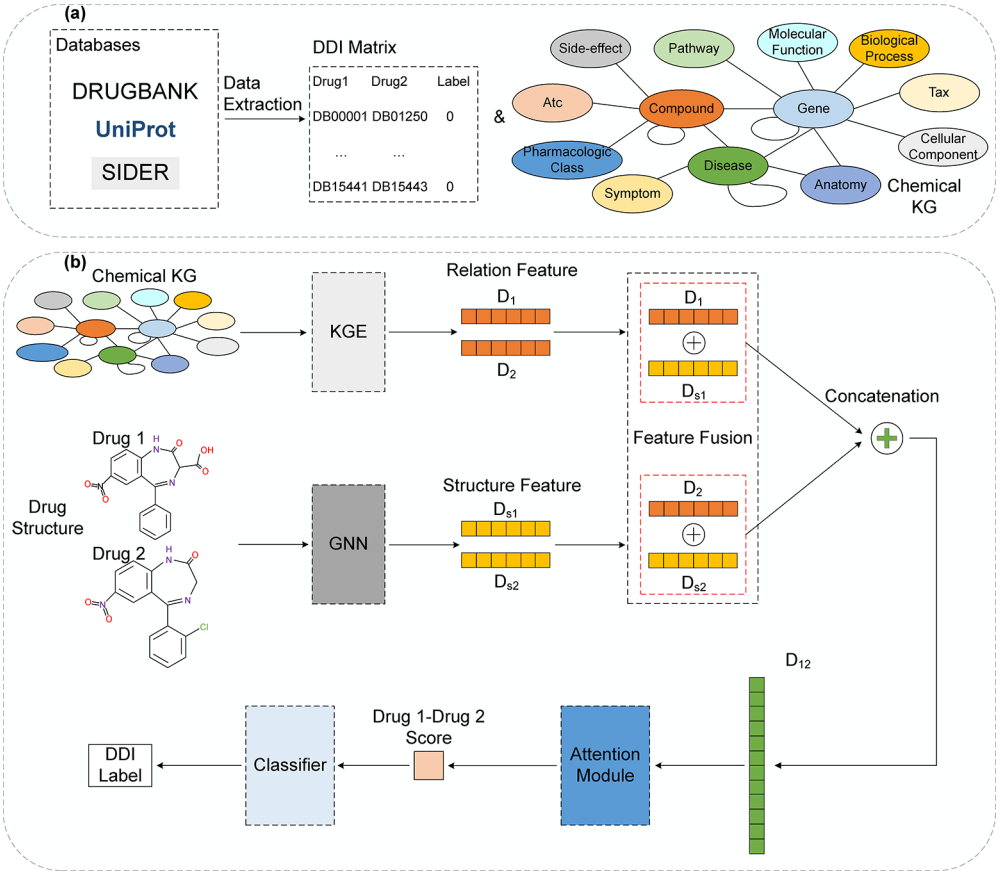


Figure 21. Frameworks for the application of KG and KGE technique to DDI prediction tasks. (a) Data are extracted from the chemical database to construct the required KG and a DDI matrix. (b) Utilizing KGE method, representations of both drugs in a drug pair can be learned separately and then fed into a prediction module. Extra drug information not present in the KG can be integrated to further improve feature representation.

as reaction classification, product prediction, yield prediction, and molecular property prediction.

4.3. Prediction of Adverse Drug Reactions. Adverse drug reactions (ADRs) are unforeseen or unintentional effects in patient taking drugs at a normal dose. Detecting ADRs during early drug development can help drug developers reevaluate the process and save costs. While machine learning-based ADR prediction models have been designed in the past, the accuracy could be improved furtherly. By incorporating knowledge graph-related methods in ADR prediction, in Figure 20, more effective but underutilized information that existing models may have overlooked can be integrated, experimental data can be expanded, and the accuracy of prediction methods can be enhanced.

By utilizing simple enrichment tests and a KG they constructed, Bean et al.[48] designed a machine learning algorithm for predicting ADRs. The KG consisted of four types of nodes: drugs, protein targets, indications, and ADRs. They vectorized the adjacent matrix of drug nodes and proposed a classifier similar to logistic regression for predicting adverse reactions. The method was shown to perform very well in classifying known adverse reaction causes.

Zhang et al.[49] proposed a uniform model combined ADR prediction tasks with KGE and predicted potential ADRs for marketed drugs. They constructed a KG (with ADRs as side effects) containing four types of nodes: drugs, indications, targets, and side effects, and developed a new KGE model. The model utilized the Word2vec model to create multidimensional vectors that capture the complex relationships between drugs, indications, targets, and side effects. They then build a vectorized drug and side effect classification model to predict adverse reactions. Their results showed that KGE method is effective in encoding drugs and ADRs, as well as in predicting ADRs.

Previous studies have utilized KG to develop ADR prediction models, but either separately or with limited entities such as target proteins and indications. However, these studies failed to integrate important information such as gene interactions and pathways. Joshi et al.[50] constructed a KG with six node types and five edge types, integrating pathway and gene information in addition to target and indication information. They used a deep neural network to propose a better-performing ADR prediction method. The node types in the KG included drugs, adverse reactions, indications, target proteins, genes, and pathways. They used Node2vec[51] and CBOW algorithms to embed the nodes of the graph into an 800-dimensional space. They then designed a deep neural network for predicting ADRs. The model has achieved good efficiency in the task of predicting common adverse reactions, but it is difficult to detect rare adverse reactions. To further enhance the efficiency of the model in the future, alternative entities and more advanced embedding techniques could be explored.

4.4. Molecular Interaction Prediction. *4.4.1. Prediction of Drug−Drug Interactions (DDIs).* When one drug is used in combination with another drug or multiple drugs, ADRs frequently occur as a result of drug−drug interactions (DDIs). Patients with complex conditions are often find themselves in a

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

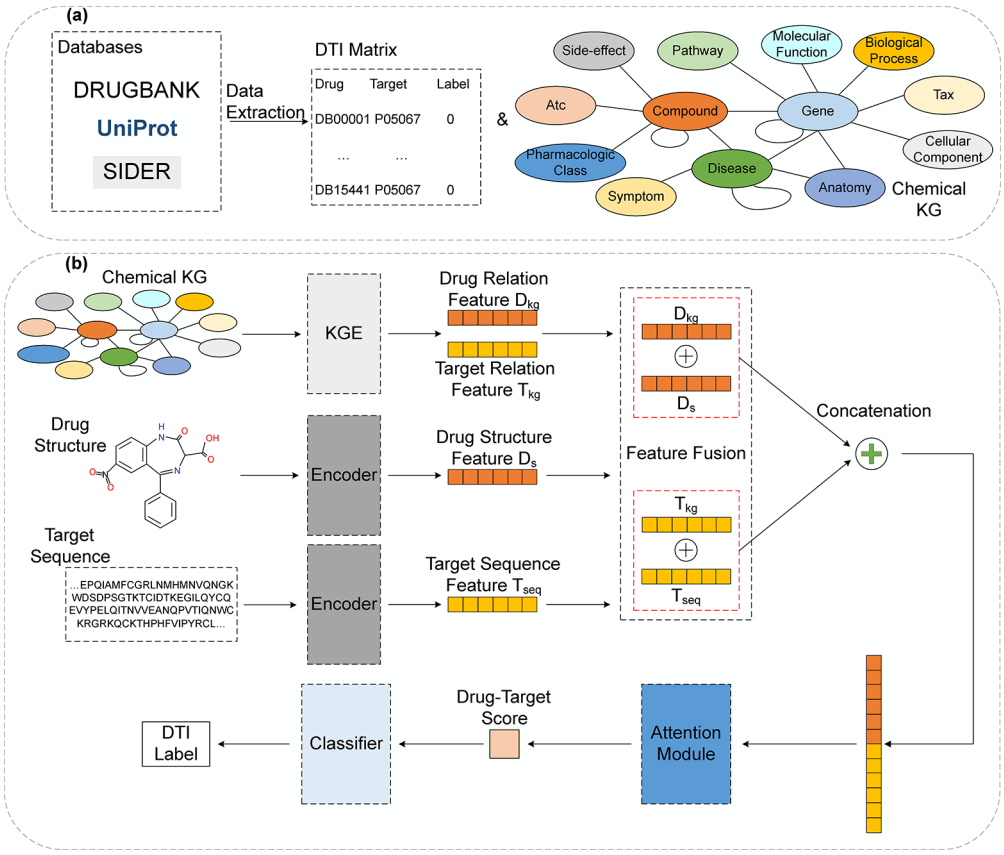


Figure 22. Frameworks for the application of KG and KGE technique to DTI prediction tasks. (a) Data are extracted from the chemical database to construct the required KG and a DTI matrix. (b) Utilizing KGE method, representations of the drug and the target in a drug-target pair can be learned separately and then fed into a prediction module. Extra drug information and target information not present in the KG can be integrated to further improve feature representation.

dangerous situation due to the occurrence of ADRs. Undetected ADRs can pose a significant threat to patient health, making accurate prediction of DDIs an essential clinical task for reducing associated risks and costs. Early methods for predicting potential DDIs mostly involved integrating multiple data sources or combining popular embedding methods. However, in recent years, KGs and GNNs have become increasingly popular in DDI prediction. The general application framework of KG is shown in Figure 21.

Malone et al.[52] utilized a publicly available preprocessed version of a KG, DistMult embedding technology, and the KBLRN framework[53] to demonstrate the exceptional performance of a multirelation KG in predicting multiple drug side effects. The KG consists of two primary components: a drug−drug interaction network, and a protein−protein interaction network, with known drug−protein-target relationships connect these different components. Notably, each drug−drug link is associated with a specific multidrug side effect. This method is also applicable to any multirelation knowledge graph link prediction task.

Lin et al.[32] first proposed the knowledge graph neural network (KGNN) in an end-to-end framework by exploring drug topologies to predict DDIs. The model extends spatial GNN methods to KGs, learning both the topological information and semantic information on KGs, as well as the neighborhood information on drugs and related entities. Due to the captured neighborhood information, KGNN gives excellent performance in DDI prediction. However, KGNN primarily attends to DDI information, disregarding other types of entities and relationships in KG, as well as the multiplicity of tasks. In general, DDI prediction is approached as a binary classification task; in comparison, the prediction of multiple types of drug interactions is more significant. To address this issue, Yu et al.[54] developed SumGNN, which employs a multichannel neural encoder for achieving multitype DDI prediction. This model extracts a subgraph containing the neighborhood entities of a specific drug pair, and integrates diversified information used to generate drug pairs by generating GNN-based pathways to provide mechanisms for drug interactions.

In addition, to address the sparsity issue of KG and enhance the performance of KGE models, Zhang et al.[55] integrated molecular structure information into KGE models and proposed MKGE (KG embedding with Molecular). Graph-based strategies and text-based strategies were utilized for extracting molecular structure information. These two vectors were concatenated as the initial embedding of KG. They constructed the KCCR KG with typical sparsity, and conducted entity prediction and relationship prediction experiments on this KG and DeepDDI,[56] verifying the effectiveness of molecular structure information for KGE. MKGE gives excellent results in multiple tasks, such as entity prediction, relationship prediction, and DDI prediction.

Chen et al.[57] attempted to further integrate the structural information on drug molecules with the topology and semantic information on KGs, proposing the MSKG-DDI framework

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

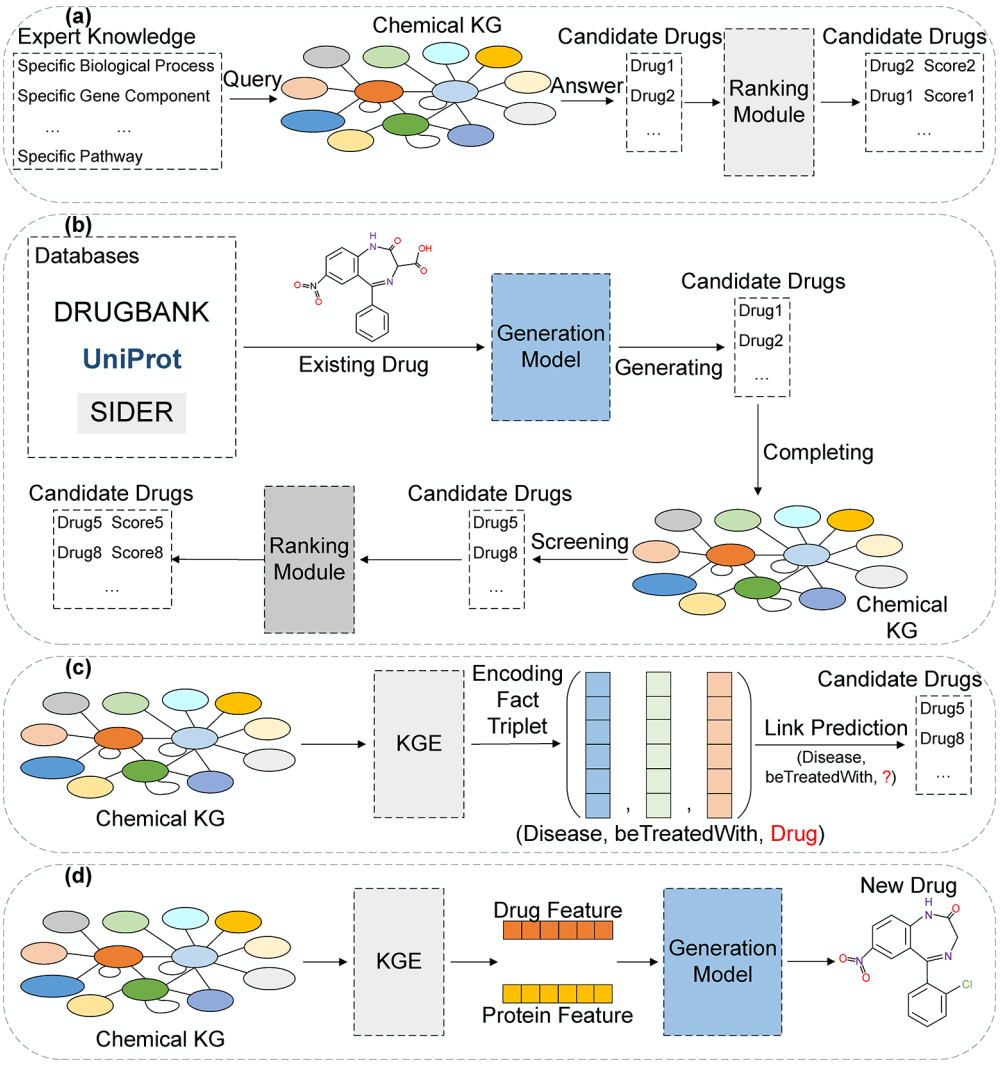


Figure 23. Frameworks for the application of KG and KGE technique to drug discovery tasks. (a) Utilizing KG querying and reasoning capabilities, drug repurposing can be achieved. (b) Utilizing KG can enable molecular filtering within the generative space of general generative models, significantly reducing the number of candidate molecules and lowering the ranking cost. (c) By employing KGE technique for link prediction, drug repurposing can be effectively achieved. (d) KGE method introduces additional domain knowledge into general generative models, thereby enhancing the generation outcomes.

composed of a molecular graph structure module and a KG module. The molecular graph structure module utilized message passing mechanisms to extract rich and effective features of drug molecules. The KG module employed a GNN to extract embeddings of the KG, thus capturing both the topological information and semantic relationships within the KG. Finally, a multimodal fusion network was utilized to complementarily integrate the two types of features of drug molecules, thereby achieving effective DDI prediction. This framework could also be extended to other interaction prediction tasks.

4.4.2. Prediction of Drug–Target Interactions (DTIs). Traditional experimental methods for detecting DTIs can be expensive and time-consuming. However, DTI plays a crucial role in drug development applications, such as lead compound discovery, drug repurposing, and identifying potential side effects. To address this challenge, deep learning-based end-to-end methods and network-based approaches are developed, such as NeoDTI.[58] These methods enable large-scale DTI prediction in a more efficient and cost-effective manner. In recent years, knowledge graph-based machine learning models have rapidly developed, as shown in Figure 22. These models have been successfully applied to address real-world challenges in biomedical and biochemical development. By extracting fine-grained multimodal knowledge elements from diverse data sources, these methods can express them as link predictions within the KG.

Mohamed et al.[59] designed a multistage DTI prediction method based on KGE, i.e., TriModel. They first constructed a KG that is relevant to drugs and targets, utilizing KEGG, UniProt, and DrugBank knowledge bases. Then, they trained the model to efficiently learn embeddings of drugs and targets within the KG. The TriModel is a tensor factorization-based KGE model that effectively extends the DistMult and ComplEx models. Their results indicate that the TriModel can efficiently predict new drug target interactions, and outperforms state-of-the-art models at that time.

Despite the significant efforts devoted to enhancing DTI prediction, many models still encounter challenges related to the sparsity, and suffer cold start problems inherent in DTI

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Application of KGEMs

domain	use case	KGEM	downstream task	year
Chemical risk assessment	Myklebust et al.[42]	DistMult, ComplEx, HolE, TransE, RotatE, pRotatE, HAKE, ConvE, ConvKB	Prediction of Adverse Biological Effects of Chemicals (Link Prediction)	2022
	Shin et al.[44]	TransE, ComplEx, ConvKB	Knowledge Graph Completion (Link Prediction)	2022
Prediction of Compound Properties	Fang et al.[6]	RotatE	Feature Enhancement	2022
	Hua et al.[45]	RotatE, ComplEx	Link Prediction	2022
	Fang et al.[46]	OWL2Vec, Word2vec	Feature Enhancement	2023
	Xie et al.[47]	TransE	Property Prediction	2024
Prediction of ADR	Zhang et al.[49]	Word2vec	Entity Property Prediction	2021
	Joshi et al.[50]	Node2vec	Entity Classification	2021
Prediction of Molecule Interaction	Malone et al.[52]	DistMult	Prediction of DDI (Link Prediction)	2019
	Lin et al.[32]	KGNN	Prediction of DDI (Relation Prediction)	2020
	Yu et al.[54]	TransE	Prediction of DDI (Relation Prediction)	2021
	Zhang et al.[55]	Improved TransE, RotatE, DistMult, ComplEx and SimplE	Prediction of DDI (Relation Prediction)	2022
	Mohamed et al.[59]	TriModel	Prediction of DTI (Link Prediction)	2020
	Ye et al.[60]	DistMult	Prediction of DTI (Link Prediction)	2021
	Li et al.[61]	RotatE	Prediction of DTI (Link Prediction)	2024
	Ma et al.[63]	RGCN	Feature Enhancement	2023
	Zahra et al.[64]	Word2vec	Prediction of DTI (Link Prediction)	2023
Drug Discovery	Chen et al.[69]	RESCAL	Feature Enhancement	2023

data sets. Ye et al.[60] combined KG with recommendation systems and proposed KGE_NFM, a highly scalable unified framework for DTI prediction. It uses the DistMult embedding model to learn entity embeddings in the KG. Then, the dimensionality of the embeddings was reduced through Principal Component Analysis (PCA). Finally, this model integrates multimodal information through the Neural Factorization Machine (NFM). Using integrated multimodal data, e.g., structural information on biomolecules and association information from biochemical networks, leading the DTI prediction more reliable and highly competitive.

TTModel[61] proposed by Li et al. leverages structural information from knowledge graphs (KGs), biomedical texts, and entity type information simultaneously, effectively alleviating the issue of data sparsity inherent in KGs while endowing the model with the capability to address long-tail problems. Moreover, the mechanism for learning biomedical texts and entity types in this model serves as a plug-and-play module, facilitating easy integration into other KGEMs to enhance the model's representation learning capacity, thereby yielding improved performance in downstream tasks.

KG-based DTI predictions often use graph embedding methods to generate embedding vectors of drugs and targets. Despite these methods are effective in learning the latent embeddings of nodes, they have limitations in capturing rich neighborhood information on KG entities. Wu et al.[62] proposed KGAT (knowledge graph attention network), which selectively aggregates neighboring nodes with attention weights to learn high-order topological and semantic features of the KG. It inputs the feature vectors for drugs and targets into the prediction model. Because the high-order topological structure captures more useful features, it improves embedding quality of entities for high accuracy of DTI prediction. KGAT model outperforms the previous methods and could be used in future DTI predictions.

Ma et al.[63] designed a new framework named KG-MTL, which proposed a new shared unit based on multitask learning principles. This shared unit captures semantic information from both the compound molecule graph and the semantic relationships among entities in the KG. KG-MTL provides new insights for DTI prediction by leveraging both types of information simultaneously. It was used to predict interaction between COVID-19 drug candidates and the active protein.

Zahra et al.[64] extracted the features of drugs and diseases from some databases and constructed a knowledge graph DrugRep-KG. They used the Word2vec model to extract entity features and obtained embeddings for drugs and diseases. The embeddings of drug-disease pairs are concatenated and fed into a logistic regression model to predict the interaction between drug and disease. The results show that the use of KG can effectively address drug repurposing.

4.5. Drug Discovery. The discovery and development of drugs is a very complex process aimed at finding drugs that can effectively treat certain diseases. To enhance the success rate of new drug clinical trials and reduce the time it takes, effective methods are needed to predict the validity of candidate drugs in the clinic. The introduction of KGs into the drug development field provides a clear structure for integrating heterogeneous data, providing structured relationships between multiple entities and semantic relationships between entities. The general application framework of KG in the field of drug discovery is illustrated in Figure 23.

The COVID-19 pandemic has prompted researchers worldwide to intensively explore and develop effective drugs and therapies for treating this disease. Because the characteristics of COVID-19 (officially named SARS-CoV-2 in 2020) are affected by multiple interconnected physiological systems, including lung inflammation, severe lung damage, coagulation disorders, kidney and neurological problems, as well as cell pathways that make up these systems, KG methods have valuable applications in identifying connections between physiological systems and potential treatment methods. Jacob et al.[65] used a century of knowledge in CAS scientific information management to construct a CAS biomedical KG. This model consists of 6 million nodes and 18 million relationships, which integrate numerous small molecules with

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 4. Summary of Open-Source KGs[a]

KG	downstream task	entities	relation	triplets	information	year
TERA[42]	Chemical risk assessment	-	-	-	Chemicals, Species, Chemical Toxicity	2020
Chemical ElementKG[6]	Prediction of Compound Properties	118 elements and 107 attributes	- (17 types)	1643	Chemical Elements, Attributes	2022
Chemical Synthesis KG[47]	Prediction of Compound Properties	819604 (products and reactants)	103339 (reaction templates)	587403	Reactants, Products, Reactions	2024
Drug Knowledge Graph[48]	Prediction of ADR	524 drugs and 5304 other nodes	70382 (clinical indications, targets, ADRs)	-	Drugs, Clinical Indications, Protein Targets, ADRs	2017
-[49]	Prediction of ADR	3632 drugs, 2598 indications, 4286 targets, 5589 side effects	154239 (side effects, targets, indications)	-	Drugs, Targets, Indications, Side Effects	2021
-[52]	Prediction of Molecule Interaction	645 drugs and 19085 proteins	5385339 (protein−protein edges, drug−drug edges, drug- protein edges)	-	Drugs, Proteins, Indications, Side Effects	2018
KCCR[55]	Prediction of Molecule Interaction	6839 compounds	2928 (reactions)	8809	Compounds, Molecular Structure Information	2022
DeepDDI[56]	Prediction of Molecule Interaction	1710 drugs	-(86 interactions)	192284	Drugs, DDIs	2018
-[59]	Prediction of Molecule Interaction	4284 drugs and 945 targets	12112 (DTIs)	-	Drugs, Targets, DTIs	2020
DKG4RS[71]	Drug Discovery	1229 diseases and 1509 drugs	-(28 interactions)	1441310	Diseases, Drugs, DTIs	2024

[a]"-" indicates that the data is not being counted.

external databases containing diseases, molecular processes, human genes, and pathways. Using this approach, the researchers identified 1,350 small molecules that have the potential to treat COVID-19, thus facilitating drug repurposing efforts and new drug discovery research.

Ranjan et al.[66] developed a new framework combining gated graph neural networks (GGNN),[67] KG, and early fusion methods[68] for SARS-CoV-2 inhibitor generation. To improve the efficiency of the early fusion model, the molecules generated by GGNN were first screened using the KG to remove nonbinding molecules. The early fusion method was used to predict the binding affinity score of the generated molecules. This framework successfully generated a potent candidate molecule active to inhibit the SARS-CoV-2 with a high binding score. Therefore, this approach could serve as a critical component of AI-based drug discovery efforts.

Chen et al.[69] initially constructed a KG using drug and disease data and employed the RESCAL model to extract embedding features for transport proteins and drugs. They integrated these embedding features with sequence features and utilized a multihead attention mechanism to predict transport proteins that might interact with drugs. Subsequently, the fused features of these transport proteins were employed as input conditions for optimization in the MolGPT[70] model, aiming to generate small molecule drugs specifically targeting these transport proteins. Experimental results indicate that features extracted by the KGE model can optimize the MolGPT framework, resulting in the generation of novel and effective small molecules.

Inspired by recommender system tasks, Tayebi et al.[71] proposed an end-to-end drug recommendation method called EKGDR, leveraging a drug KG as side information for disease-drug interactions, thus recommending therapeutic drugs for specific diseases. EKGDR assumes that the latent connections between diseases and drugs are determined by intents. Therefore, it models intents as combinations of existing relations in the KG, and further extends the disease-drug bipartite graph into disease-intent-drug triplets. The model utilizes GNN aggregation layers to extract features for these triples and the drug KG separately, thereby learning embeddings for diseases and drugs. Using these representations, the model recommends drugs for specific diseases based on the results of link prediction.

4.6. Summary of KGE Models and Open-Source KGs Used in Chemistry. This section focuses on the KGE models and open-source KGs used in the aforementioned case studies. The application domains and downstream tasks of various KGE models are summarized in Table 3. The application domains, data overview, and the included chemical information on some open-source KGs are summarized in Table 4.

5. CHALLENGES AND PROSPECTS

5.1. Knowledge Graph. As a novel approach, the application of KGs in the field of chemistry has achieved tremendous prospects for development. Integrating chemical knowledge into structured data is crucial for effective chemical data analysis, leading to enhanced standardization and usability. KGs organize chemical knowledge in a structured manner, visually depicting connections between various chemical entities through graphs, thus facilitating knowledge integration.[42−44] For instance, chemical KGs established from public chemical databases and scientific literature can extract precise associations between chemical substances, diseases, or drug interactions. By linking these chemical entities, such as molecular structures, properties, and reaction conditions, KGs offer a more comprehensive repository of chemical information, aiding in the discovery of potential relationships and chemical patterns.[49,50]

Furthermore, KGs provide a practical framework for chemical data analysis. Through querying and inference of KGs, rapid retrieval and analysis of chemical data become possible, uncovering latent semantics and patterns to assist downstream tasks like chemical property prediction and drug design.[65,66] For example, by representing a complex network of chemical reactions as a KG, all chemical reactions corresponding to certain compounds can be organized together for exploration. This approach enables the discovery of new synthesis routes or reaction patterns rapidly and efficiently.[103]

KGs provide a potent tool for addressing chemical tasks, enhancing both efficiency and creativity in their execution, thereby introducing numerous advantages and opportunities to

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the field of chemistry. However, several challenges need to be addressed:

5.1.1. Incomplete Information in KGs. Across diverse chemical applications like risk assessment, property prediction, interaction prediction, and drug design, distinct chemical entities play varying roles. It is necessary to construct tailored KGs specific to each chemical application, encompassing the pertinent chemical entity information. Some existing KGs may not fully capture the intricacies of relevant chemical knowledge or might overlook certain interconnections between chemical entities. Furthermore, when novel chemical substances continue to emerge, chemical KGs must be regularly updated to incorporate new chemical entities and uncover potential new entity relationships. This adaptability ensures that the KG remains current and reflective of the evolving landscape of chemical science.

5.1.2. Quality of KGs. The effectiveness of utilizing KGs applied in chemistry is significantly contingent upon its quality. When constructing KGs, apart from utilizing data from existing databases, the entities and relationships extracted from chemical literature are also significant additional information. As a result, KGs often encompass a substantial amount of chemical data with unavoidably noise data, leading poor quality of the KG.

To address issues related to the data integrity and data quality of KGs, future development may encompass various aspects:

5.1.3. Data Quality. When constructing KGs for specific domains, integrating information from diverse data sources and focusing on the development of error detection and correction methods can enhance the quality and credibility of KGs at the data level.[72]

5.1.4. Semi-Automated Construction. Building KGs in the field of chemistry requires substantial human effort and time, given the complexity of chemical entity features. In the future, exploring semiautomated methods for KG construction could assist in constructing large-scale KGs.[73]

5.1.5. Incremental Updates. As new chemical compounds, reactions, experimental data, and other chemical information emerge, timely incremental updates to the chemical KG are necessary.[74]

5.1.6. Knowledge Fusion. KGs should ideally incorporate valuable features of chemical entities.[75] Through multimodal information fusion, additional information such as molecular structure data and other prior knowledge can be integrated into the KG to enhance its richness with high-quality external chemical knowledge. Additionally, merging multiple existing KGs, even from different domains, can expand the cross-domain information within the KG, offering solutions to complex chemical problems.

Some efforts have utilized ontology modeling and RDF triples to integrate chemical resources, resulting in the creation of large ontologies and semantic databases. While these resources have increased data integration and achieved small-scale data normalization, they still face challenges such as manual maintenance of resources and incomplete query results. Recently, an attempt to address data management and integration challenges through dynamic knowledge graphs has been made. This research aims to construct a highly integrated, standardized, and efficiently queryable large dynamic knowledge graph called The World Avatar KG (TWA KG),[104] and it has already achieved preliminary results. For example, the chemical species ontology OntoSpecies[105] within the TWA KG uses ontology networks to model the structure of chemical data, employs an automatically computed agent for data updates, and provides data access via a SPARQL end point. The goal is to manage chemical data through standardized unified representation and enable complex query functions. Compared to traditional large databases, the TWA KG built using dynamic knowledge graphs offers advantages in terms of integrability, updatability, and manageability. Therefore, KGs are expected to become extremely important data management tools in the future, especially for complex data domains such as chemistry.

5.2. Knowledge Graph Embedding. The concept of knowledge graph completion has been introduced to address the issue of missing content within KGs, and one crucial approach to solving this completion problem is KGE. KGE techniques involve embedding the entities and relationships of a KG into a continuous vector space, allowing for the discovery of latent relationships between entities. This approach captures underlying semantics and structural information, mitigating the sparsity of the KG. Among existing KGE models, distance-based models and semantic matching models are effective in modeling complex relationship patterns. These models are well-suited for addressing the complexity of entity properties and the diversity of relationship types present in the field of chemistry.

Conversely, neural network-based models and other embedding methods that integrate large language models might lack interpretability, but they are capable of effectively capturing high-order topological structures within the KG and neighborhood information on chemical entities. These approaches have demonstrated impressive performance and efficiency in tasks such as chemical prediction. However, due to the intricate nature of entity and relationship features in chemical KGs, which are rich in chemical information, higher demands are placed on KG techniques to meet the unique challenges posed by the domain of chemistry.

5.2.1. Expressive Capabilities of KGE Models. While existing KGE models can alleviate the sparsity of chemical data to some extent, they may not completely solve the issue of data sparsity, particularly for rare compounds or chemical reactions where obtaining sufficient experimental data for embedding representation is challenging. Furthermore, mapping complex chemical entities and relationships to a low-dimensional vector space might lead to the loss of certain semantic information. Preserving this semantic information poses a challenge that needs to be addressed in the future.

5.2.2. Interpretability. Although neural network-based embedding models and some other models effectively enhance the accuracy of chemical prediction tasks, they may struggle to provide explanations for their prediction results from a mathematical principles standpoint. In the field of chemistry, a reliable understanding of the model's prediction process is crucial.

5.2.3. High Computational Cost. Training KGE models, especially those based on neural networks and large language models, is very computationally intensive. The complexity of chemical data and the relationships between chemical entities puts forward higher requirements for computing resources. Therefore, the training of KGE models generally requires high hardware cost and time cost, which makes it challenging to develop and maintain large-scale chemical knowledge graphs.

5.2.4. Frequent Updates. As the field of chemistry continues to evolve, new compounds, reactions, and

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

experimental data are constantly being discovered. KG will maintain continuous updates to ensure the integrity of the data. Therefore, the KGE model also needs to supplement the learning of new embeddings at the same time as the update of KG. This process can be resource intensive as it can involve retraining or model fine-tuning, which further increases the computational cost.

Given the strong performance of KGE technology in the field of chemistry, the future development of this technique holds the potential to introduce further innovations and breakthroughs in both chemical research and applications. In this context, we will briefly explore some potential directions for the future evolution of KGE models:

5.2.5. Optimized Training Algorithms. Investigating more efficient KGE training algorithms can help reduce the computational cost. The model can try to use techniques such as stochastic optimization, distributed computing, and parallel processing to improve the efficiency of model training, and use model compression and pruning methods to save computing resources without significantly affecting model performance.

5.2.6. Incremental Learning. Developing incremental learning methods for KGE is perhaps a future direction to address the high cost of training KGE models. Incremental learning allows the KGE model to update its embedding based on new data in the KG while retaining previously learned information. The proposed method can significantly reduce the new computational overhead when the KG is updated frequently.

5.2.7. Multimodal Knowledge Graph Embedding Techniques. Chemical data encompasses various data types such as chemical structures, textual descriptions, and experimental data. Integrating these data types into a single embedding vector space allows models to comprehensively capture connections between chemical entities, explore potential chemical patterns, and acquire richer chemical information to support downstream tasks.[76]

5.2.8. Temporal Knowledge Graph Embedding. Research and applications concerning temporal KGs have emerged.[77] Given the dynamic nature of chemical reaction processes, the field of chemical KGE should explore how to incorporate the temporal and sequential relationships of chemical reactions into prediction models, providing a more precise representation of the dynamic changes of molecules during chemical reactions.

5.2.9. Integration of Chemical Domain Knowledge. Chemical structural information, including atom types, bonds, and molecular formulas, aids in capturing intermolecular relationships. Information about chemical reactions that describe the relationships between reactants and products contributes to understanding chemical reaction principles. Molecular properties, interactions, and biomedical data play vital roles in chemical prediction and design tasks.[72] While it is challenging for a chemical KG to encompass all relevant features, selectively integrating specific aspects of chemical domain information can enhance the practicality of embedding models. For instance, fusing functional group information with chemical structural data can facilitate the exploration of reaction mechanisms and prediction of reaction products.[78]

5.2.10. Incorporation of Expert Knowledge. To effectively enhance the ability of artificial intelligence methods to provide solutions to chemical problems, it is necessary to supply computers with various forms of chemical knowledge. While KGs can integrate complex chemical data and KGE models can capture intricate chemical semantics, these tools may not accurately discern subtle chemical rules. Although chemical experts may not remember vast amounts of chemical data, their extensive experience and deep understanding of chemical rules guide them in providing more accurate solutions. Therefore, the design of KG and KGE models should establish closer connections with experts, involving them in the development process to assist in the design of AI methods.[106]

5.2.11. Hybrid Knowledge Graph Embeddings. In large-scale knowledge graphs like the TWA KG, heterogeneous information from multiple domains is often integrated. Handling these heterogeneous and cross-domain entities and relationships with a single KGE model is highly complex and challenging. Recently, a study[107] designed a hybrid KGE system that employs different embedding methods for different ontologies, achieving embedding learning for the TWA KG. This hybrid KGE approach can be considered a practical method for handling complex embeddings in large-scale knowledge graphs.

By addressing these challenges and exploring these prospects, KGE techniques can significantly advance the field of chemistry even when considering the training and updating costs. Employing advanced techniques and strategic resource management can address these challenges, making KGE models a viable solution for integrating and analyzing complex chemical data.

6. SUMMARIES

The KG related techniques have received widespread attention and have good development prospects in various fields. Research in the chemistry field relies on large amounts of domain-specific data sets, and leveraging KGs can significantly improve data integration and data understanding. By embedding the entities and relations of KG into a continuous vector space, KGE technique preserves the inherent structure of KG while streamlining operations. It also captures the similarity between entities and relationships by measuring the low-dimensional embedding of entities and relationships. The application of KGE technology in the field of chemistry spans multiple directions, aimed at exploring chemical patterns and addressing complex chemical challenges.

This review offers an overview of KGE technique and its applications in the chemistry field. It also introduces successful applications of KGs in recent years, such as chemical risk assessment, compound property prediction, adverse drug reaction prediction, molecular interaction prediction, and drug discovery. We have summarized the current challenges and development potential of KG in the chemistry field, and we hope that this survey can assist in the future applications of KG.

■ ASSOCIATED CONTENT

Data Availability Statement

The data are given in the Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00791.

Summary of studies on predicting compound properties, adverse drug reactions, and molecular interactions (PDF)

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

Xiaofei Nan − *School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China*; Email: iexfnan@zzu.edu.cn

Authors

Chuanghui Wang − *School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China*; orcid.org/0009-0008-7430-2251

Yunqing Yang − *School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China*

Jinshuai Song − *Green Catalysis Center, College of Chemistry, Zhengzhou University, Zhengzhou 450001, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00791

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, Busan, Korea, November 11−15, 2007; Lecture Notes in Computer Science, Vol. *4825*; pp 722−735.

(2) Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*; Association for Computing Machinery, 2008; pp 1247−1250.

(3) Vrandečić, D.; Krötzsch, M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* **2014**, *57*, 78−85.

(4) Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka, E.; Mitchell, T. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence, 2010; pp 1306−1313.

(5) Suchanek, F. M.; Kasneci, G.; Weikum, G. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*; Association for Computing Machinery, 2007; pp 697−706.

(6) Fang, Y.; Zhang, Q.; Yang, H.; Zhuang, X.; Deng, S.; Zhang, W.; Chen, Z.; Fan, X.; Chen, H. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence, 2022; pp 3968−3976.

(7) Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, Vol. *26*; Curran Associates, Inc., 2013.

(8) Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge graph embedding by translating on hyperplanes. *AAAI* **2014**, *28* (1), 8870.

(9) Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning entity and relation embeddings for knowledge graph completion. *AAAI* **2015**, *29* (1), 9491.

(10) Xiao, H.; Huang, M.; Hao, Y.; Zhu, X. TransA: An adaptive approach for knowledge graph embedding. *arXiv*, September 28, 2015, 1509.05490, ver. 2. DOI: 10.48550/arXiv.1509.05490.

(11) Ji, G.; He, S.; Xu, L.; Liu, K.; Zhao, J. Knowledge graph embedding via dynamic mapping matrix. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* **2015**, *1*, 687−696.

(12) d'Amato, C.; Quatraro, N. F.; Fanizzi, N. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In *The Semantic Web: 18th International Conference ESWC 2021*; Virtual Event, June 6−10, 2021; pp 441−457.

(13) Yu, J.; Cai, Y.; Sun, M.; Li, P. Mquade: a unified model for knowledge fact embedding. *Proceedings of the Web Conference 2021* **2021**, 3442−3452.

(14) He, S.; Liu, K.; Ji, G.; Zhao, J. Learning to represent knowledge graphs with gaussian embedding. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* **2015**, 623−632.

(15) Xiao, H.; Huang, M.; Hao, Y.; Zhu, X. TransG: A generative mixture model for knowledge graph embedding. *arXiv*, December 27, 2015, 1509.05488, ver. *4*. DOI: 10.48550/arXiv.1509.05488.

(16) Sun, Z.; Deng, Z. H.; Nie, J. Y.; Tang, J. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv*, February 26, 2019, 1902.10197, ver. *1*. DOI: 10.48550/arXiv.1902.10197.

(17) Ma, Y.; Ding, Y.; Wang, G. RotatSAGE: A Scalable Knowledge Graph Embedding Model Based on Translation Assumptions and Graph Neural Networks. In *International Conference on Data Mining and Big Data*; Springer Nature Singapore: Singapore, 2022, 93−104.

(18) Zhang, S.; Tay, Y.; Yao, L.; Liu, Q. Quaternion knowledge graph embeddings. In *Advances in neural information processing systems*, Vol. *32*; Curran Associates, Inc., 2019.

(19) Nguyen, D. Q.; Vu, T.; Nguyen, T. D. Quatre: Relation-aware quaternions for knowledge graph embeddings. *Companion Proceedings of the Web Conference 2022* **2022**, 189−192.

(20) Chao, L.; He, J.; Wang, T.; Chu, W. Pairre: Knowledge graph embeddings via paired relation vectors. *arXiv*, November 7, 2020, 2011.03798, ver. *1*. DOI: 10.48550/arXiv.2011.03798.

(21) Jenatton, R.; Roux, N.; Bordes, A.; Obozinski, G. R. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems*, Vol. *25*; Curran Associates, Inc., 2012.

(22) Yang, B.; Yih, W. T.; He, X.; Gao, J.; Deng, L. Embedding entities and relations for learning and inference in knowledge bases. *arXiv*, December 27, 2014, 1412.6575, ver. *2*. DOI: 10.48550/arXiv.1412.6575.

(23) Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; Bouchard, G. Complex Embeddings for Simple Link Prediction. *Proceedings of The 33rd International Conference on Machine Learning* **2016**, *48*, 2071−2080.

(24) Kishimoto, K.; Hayashi, K.; Akai, G.; Shimbo, M.; Komatani, K. Binarized knowledge graph embeddings. *Advances in Information Retrieval* **2019**, *11437*, 181−196.

(25) Kazemi, S. M.; Poole, D. SimplE Embedding for Link Prediction in Knowledge Graphs. In *Advances in Neural Information Processing Systems*, Vol. *31*; 2018

(26) Nickel, M.; Tresp, V.; Kriegel, H. P. A three-way model for collective learning on multi-relational data. In *ICML '11: Proceedings of the 28th International Conference on International Conference on Machine Learning*, Bellevue, WA, June 28−July 2, 2011; pp 809−816.

(27) Balažević, I.; Allen, C.; Hospedales, T. M. TuckER: Tensor Factorization for Knowledge Graph Completion. *arXiv*, August 24, 2019, 1901.09590, ver. *2*. DOI: 10.48550/arXiv.1901.09590.

(28) Luo, T.; Wei, T.; 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; pp 817−824.

(29) Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, LA, February 2−7, 2018; pp 1811−1818.

(30) Nguyen, D. Q.; Nguyen, T. D.; Nguyen, D. Q.; Phung, D. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv*, December 6, 2017, 1712.02121, ver. *1*. DOI: 10.48550/arXiv.1712.02121.

(31) Zhou, Z.; Wang, C.; Feng, Y.; Chen, D. JointE: Jointly utilizing 1D and 2D convolution for knowledge graph embedding. *Knowledge-Based Systems* **2022**, *240*, 108100.
(32) Lin, X.; Quan, Z.; Wang, Z. J.; Ma, T.; Zeng, X. KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* **2020**, *380*, 2739−2745.
(33) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference ESWC 2018*, Heraklion, Crete, Greece, June 3−7, 2018; Lecture Notes in Computer Science, Vol. *10843*; pp 593−607.
(34) Yu, D.; Yang, Y.; Zhang, R.; Wu, Y. Knowledge embedding based graph convolutional network. In *Proceedings of the Web Conference 2021, April*, 2021, 1619−1628.
(35) Fang, H.; Wang, Y.; Tian, Z.; Ye, Y. Learning knowledge graph embedding with a dual-attention embedding network. *Expert Systems with Applications* **2023**, *212*, 118806.
(36) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv*, September 7, 2013, 1301.3781, ver. 3. DOI: 10.48550/arXiv.1301.3781.
(37) Dai, Y.; Guo, C.; Guo, W.; Eickhoff, C. Drug-drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Briefings Bioinf.* **2021**, *22*, bbaa256.
(38) Wang, Q.; Huang, P.; Wang, H.; Dai, S.; Jiang, W.; Liu, J.; Lyu, Y.; Zhu, Y.; Wu, H. CoKE: Contextualized Knowledge Graph Embedding. *arXiv*, November 6, 2019, 1911.02168, ver. *1*. DOI: 10.48550/arXiv.1911.02168.
(39) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. *30*; Curran Associates, Inc., 2017.
(40) Yao, L.; Mao, C.; Luo, Y. KG-BERT: BERT for knowledge graph completion. *arXiv*, September 11, 2019, 1909.03193, ver. *2*. DOI: 10.48550/arXiv.1909.03193.
(41) Chen, J.; Hu, P.; Jimenez-Ruiz, E.; Holter, O. M.; Antonyrajah, D.; Horrocks, I. OWL2Vec*: Embedding of OWL ontologies. *Machine Learning* **2021**, *110*, 1813−1845.
(42) Myklebust, E. B.; Jiménez-Ruiz, E.; Chen, J.; Wolf, R.; Tollefsen, K. E. Prediction of adverse biological effects of chemicals using knowledge graph embeddings. *Semantic Web* **2022**, *13*, 299−338.
(43) Zheng, X.; Wang, B.; Zhao, Y.; Mao, S.; Tang, Y. A knowledge graph method for hazardous chemical management: Ontology design and entity identification. *Neurocomputing* **2021**, *430*, 104−111.
(44) Shin, E.; Yoo, S.; Ju, Y.; Shin, D. Knowledge graph embedding and reasoning for real-time analytics support of chemical diagnosis from exposure symptoms. *Process Safety and Environmental Protection* **2022**, *157*, 92−105.
(45) Hua, R.; Wang, X.; Cheng, C.; Zhu, Q.; Zhou, X. A Chemical Domain Knowledge-Aware Framework for Multi-view Molecular Property Prediction. In *CCKS 2022-Evaluation Track: 7th China Conference on Knowledge Graph and Semantic Computing Evaluations*, Qinhuangdao, China, August 24−27; Communications in Computer and Information Science, Vol. *1711*; Springer Nature Singapore, 2022; pp 1−11.
(46) Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; Chen, H. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence* **2023**, *5*, 542−553.
(47) Xie, J.; Wang, Y.; Rao, J.; Zheng, S.; Yang, Y. Self-Supervised Contrastive Molecular Representation Learning with a Chemical Synthesis Knowledge Graph. *J. Chem. Inf. Model.* **2024**, *64*, 1945.
(48) Bean, D. M.; Wu, H.; Iqbal, E.; Dzahini, O.; Ibrahim, Z. M.; Broadbent, M.; Stewart, R.; Dobson, R. J. B. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.* **2017**, *7*, 16416.
(49) Zhang, F.; Sun, B.; Diao, X.; Zhao, W.; Shu, T. Prediction of adverse drug reactions based on knowledge graph embedding. *BMC Med. Inf. Decis. Making* **2021**, *21*, 38.
(50) Joshi, P.; V, M.; Mukherjee, A. A knowledge graph embedding based approach to predict the adverse drug reactions using a deep neural network. *J. Biomed. Inf.* **2022**, *132*, 104122.
(51) Grover, A.; Leskovec, J. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 855−864.
(52) Malone, B.; García-Durán, A.; Niepert, M. Knowledge graph completion to predict polypharmacy side effects. In *Data Integration in the Life Sciences: 13th International Conference DILS*; Hannover, Germany, November 20−21, 2018, Springer; pp 144−149.
(53) García-Durán, A.; Niepert, M. KBLRN: End-to-End Learning of Knowledge Base Representations with Latent, Relational, and Numerical Features. *arXiv*, September 14, 2017, 1709.04676, ver. *1*. DOI: 10.48550/arXiv.1709.04676.
(54) Yu, Y.; Huang, K.; Zhang, C.; Glass, L. M.; Sun, J.; Xiao, C. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics* **2021**, *37*, 2988−2995.
(55) Zhang, Y.; Li, Z.; Duan, B.; Qin, L.; Peng, J. MKGE: Knowledge graph embedding with molecular structure information. *Computational Biology and Chemistry* **2022**, *100*, 107730.
(56) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep learning improves prediction of drug-drug and drug-food interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E4304−E4311.
(57) Chen, S.; Semenov, I.; Zhang, F.; Yang, Y.; Geng, J.; Feng, X.; Meng, Q.; Lei, K. An effective framework for predicting drug-drug interactions based on molecular substructures and knowledge graph neural network. *Computers in Biology and Medicine* **2024**, *169*, 107900.
(58) Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; Zeng, J. NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics* **2019**, *35*, 104−111.
(59) Mohamed, S. K.; Nováček, V.; Nounu, A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* **2020**, *36*, 603−610.
(60) Ye, Q.; Hsieh, C. Y.; Yang, Z.; Kang, Y.; Chen, J.; Cao, D.; He, S.; Hou, T. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. *Nat. Commun.* **2021**, *12*, 6775.
(61) Li, N.; Yang, Z.; Wang, J.; Lin, H. Drug-target interaction prediction using knowledge graph embedding. *iScience* **2024**, *27*, 109393.
(62) Wu, Z.; Zhang, X.; Lin, X. KGAT: Predicting Drug-Target Interaction Based on Knowledge Graph Attention Network. In *Intelligent Computing Theories and Application. ICIC 2022. Lecture Notes in Computer Science*, Vol. *13394*; Springer: Cham, Switzerland, 2022; pp 438−450.
(63) Ma, T.; Lin, X.; Song, B.; Yu, P. S.; Zeng, X. KG-MTL: Knowledge Graph Enhanced Multi-Task Learning for Molecular Interaction. *IEEE Transactions on Knowledge & Data Engineering* **2022**, *35*, 7068−7081.
(64) Ghorbanali, Z.; Zare-Mirakabad, F.; Akbari, M.; Salehi, N.; Masoudi-Nejad, A. Drugrep-kg: Toward learning a unified latent space for drug repurposing using knowledge graphs. *J. Chem. Inf. Model.* **2023**, *63*, 2532−2545.
(65) Al-Saleem, J.; Granet, R.; Ramakrishnan, S.; Ciancetta, N. A.; Saveson, C.; Gessner, C.; Zhou, Q. Knowledge graph-based approaches to drug repurposing for COVID-19. *J. Chem. Inf. Model.* **2021**, *61*, 4058−4067.
(66) Ranjan, A.; Shukla, S.; Datta, D.; Misra, R. Generating novel molecule for target protein (SARS-CoV-2) using drug-target interaction based on graph neural network. *Network Modeling Analysis in Health Informatics and Bioinformatics* **2022**, *11*, 6.
(67) Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated Graph Sequence Neural Networks. *arXiv*, November 19, 2015, 1511.05493, ver. *2*. DOI: 10.48550/arXiv.1511.05493.

(68) Nguyen, T. M.; Nguyen, T.; Le, T. M.; Tran, T. Gefa: early fusion approach in drug-target affinity prediction. *IEEE/ACM transactions on computational biology and bioinformatics* **2022**, *19*, 718−728.
(69) Chen, X.; Ruan, Y.; Liu, Y.; Duan, X.; Jiang, F.; Tang, H.; Zhang, H.; Zhang, Q. Transporter proteins knowledge graph construction and its application in drug development. *Computational and Structural Biotechnology Journal* **2023**, *21*, 2973−2984.
(70) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **2022**, *62*, 2064−2076.
(71) Tayebi, J.; BabaAli, B. EKGDR: An End-to-End Knowledge Graph-Based Method for Computational Drug Repurposing. *J. Chem. Inf. Model.* **2024**, *64*, 1868.
(72) Chandak, P.; Huang, K.; Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data* **2023**, *10*, 67.
(73) Zhang, H.; Wang, X.; Pan, J.; Wang, H. SAKA: an intelligent platform for semi-automated knowledge graph construction and application. *SOCA* **2023**, *17*, 201−212.
(74) Wei, Y.; Chen, W.; Li, Z.; Zhao, L. Incremental update of knowledge graph embedding by rotating on hyperplanes. *2021 IEEE International Conference on Web Services (ICWS)* **2021**, 516−524.
(75) Xia, X.; Zhu, C.; Zhong, F.; Liu, L. MDTips: A Multimodal-data based Drug-Target interaction prediction system fusing knowledge, gene expression profile, and structural data. *Bioinformatics* **2023**, *39*, btad411.
(76) Krix, S.; DeLong, L. N.; Madan, S. MultiGML: Multimodal Graph Machine Learning for Prediction of Adverse Drug Events. *bioRxiv*, December 21, 2022, 520738. DOI: 10.1101/2022.12.16.520738.
(77) Cai, B.; Xiang, Y.; Gao, L. Temporal Knowledge Graph Completion: A Survey. *arXiv*, January 16, 2022, 2201.08236, ver. *1*. DOI: 10.24963/ijcai.2023/734.
(78) Vilela, J.; Asif, M.; Marques, A. R.; Santos, J. X.; Rasga, C.; Vicente, A.; Martiniano, H. Biomedical knowledge graph embeddings for personalized medicine: Predicting disease-gene associations. *Expert Systems* **2023**, *40*, No. e13181.
(79) The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523−D531.
(80) Harrison, P. W.; Amode, M. R.; Austine-Orimoloye, O.; Azov, A. G.; Barba, M.; Barnes, I.; Becker, A.; Bennett, R.; Berry, A.; Bhai, J.; et al. Ensembl 2024. *Nucleic Acids Res.* **2024**, *52*, D891−D899.
(81) RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.* **2021**, *49*, D212−D220.
(82) Brown, G. R.; Hem, V.; Katz, K. S.; Ovetsky, M.; Wallin, C.; Ermolaeva, O.; Tolstoy, I.; Tatusova, T.; Pruitt, K. D.; Maglott, D. R.; et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **2015**, *43*, D36−D42.
(83) Whirl-Carrillo, M.; Huddart, R.; Gong, L.; Sangkuhl, K.; Thorn, C. F.; Whaley, R.; Klein, T. E. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* **2021**, *110*, 563−572.
(84) Jin, Z.; Sato, Y.; Kawashima, M.; Kanehisa, M. KEGG tools for classification and analysis of viral proteins. *Protein Sci.* **2023**, *32*, No. e4820.
(85) Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A. L.; Fang, T.; Doncheva, N. T.; Pyysalo, S.; et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **2023**, *51*, D638−D646.
(86) Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B. J.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F.; et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **2021**, *30*, 187−200.
(87) Del Toro, N.; Shrivastava, A.; Ragueneau, E.; Meldal, B.; Combe, C.; Barrera, E.; Perfetto, L.; How, K.; Ratan, P.; Shirodkar, G.; et al. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res.* **2022**, *50*, D648−D653.
(88) Türei, D.; Valdeolivas, A.; Gul, L.; Palacio-Escat, N.; Klein, M.; Ivanova, O.; Ölbei, M.; Gábor, A.; Theis, F.; Módos, D.; et al. Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **2021**, *17*, No. e9923.
(89) Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E.; Strawbridge, S. A.; et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Res.* **2024**, *52*, D1265−D1275.
(90) Avram, S.; Wilson, T. B.; Curpan, R.; Halip, L.; Borota, A.; Bora, A.; Bologa, C. G.; Holmes, J.; Knockel, J.; Yang, J. J.; et al. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Res.* **2023**, *51*, D1276−D1287.
(91) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.
(92) Brown, A. S.; Patel, C. J. A standard database for drug repositioning. *Sci. Data* **2017**, *4*, 170029.
(93) Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E. G.; Gewiess, A.; Jensen, L. J.; et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **2007**, *36*, D919−D922.
(94) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **2024**, *52*, D1180−D1192.
(95) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373−D1380.
(96) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The making of reaxys-towards unobstructed access to relevant chemistry information, The Future of the History of Chemical Information. *American Chemical Society* **2014**, *1164*, 127−148.
(97) NIST Computational Chemistry Comparison and Benchmark Database. *NIST Standard Reference Database Number 101*, Release *22*; Johnson III, R. D., Ed.; 2020..
(98) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065−6073.
(99) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(100) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(101) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44*, D1075−D1079.
(102) Tatonetti, N. P.; Ye, P. P.; Daneshjou, R.; Altman, R. B. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* **2012**, *4* (125), 125ra31.
(103) Jeong, J.; Lee, N.; Shin, Y.; Shin, D. Intelligent generation of optimal synthetic pathways based on knowledge graph inference and retrosynthetic predictions using reaction big data. *Journal of the Taiwan Institute of Chemical Engineers* **2022**, *130*, 103982.
(104) Lim, M. Q.; Wang, X.; Inderwildi, O.; Kraft, M. The World Avatar—a World Model for Facilitating Interoperability. In *Intelligent decarbonisation: can artificial intelligence and cyber-physical systems help achieve climate mitigation targets?*; Springer International Publishing: Cham, Switzerland, 2022; pp 39−53.
(105) Pascazio, L.; Rihm, S.; Naseri, A.; Mosbach, S.; Akroyd, J.; Kraft, M. Chemical species ontology for data integration and knowledge discovery. *J. Chem. Inf. Model.* **2023**, *63*, 6569−6586.
(106) Strieth-Kalthoff, F.; Szymkuć, S.; Molga, K.; Aspuru-Guzik, A.; Glorius, F.; Grzybowski, B. A. Artificial Intelligence for Retrosynthetic

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Planning Needs Both Data and Expert Knowledge. *J. Am. Chem. Soc.* **2024**, *146*, 11005−11017.
(107) Zhou, X.; Zhang, S.; Agarwal, M.; Akroyd, J.; Mosbach, S.; Kraft, M. Marie and BERT - A Knowledge Graph Embedding Based Question Answering System for Chemistry. *ACS omega* **2023**, *8*, 33039−33057.
(108) Tasneem, A.; Aberle, L.; Ananth, H.; Chakraborty, S.; Chiswell, K.; McCourt, B. J.; Pietrobon, R. The database for aggregate analysis of ClinicalTrials. gov (AACT) and subsequent regrouping by clinical specialty. *PloS one* **2012**, *7*, No. e33677.

https://doi.org/10.1021/acs.jcim.4c00791
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX