JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 19, No. 1, March 2025, pp. 1-15

Survey on AI-Drug Discovery with Knowledge Graphs: Data, Algorithm, and Application

Daeun Kong†
AIGENDRUG Co. Ltd., Seoul, Korea
b.daeun113@gmail.com

Yoojin Ha†
Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea
yoojinha@snu.ac.kr

HaEun Yoo†
Department of Computer Science and Engineering, Seoul National University, Seoul, Korea
yhhen111@snu.ac.kr

Dongmin Bang
AIGENDRUG Co. Ltd., Seoul, Korea
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea
eugenomics@snu.ac.kr

Sun Kim*
AIGENDRUG Co. Ltd., Seoul, Korea
Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea
Department of Computer Science and Engineering, Seoul National University, Seoul, Korea
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea
sunkim.bioinfo@snu.ac.kr

Abstract

Drug discovery is a complex, costly, and high-risk endeavor reliant on extensive experiments and clinical trials. Recent advances in artificial intelligence (AI) are transforming biomedical research by modeling complex biological relationships and accelerating therapeutic discovery. Central to these innovations are biomedical knowledge graphs (KGs), which systematically integrate diverse, heterogeneous data, from molecular interactions and genetic profiles to drug-disease associations. In particular, heterogeneous knowledge graphs (HKGs) capture complex biological phenomena through interconnected multi-modal data sources. This survey provides a comprehensive overview of AI-driven drug discovery via HKGs, detailing their definitions, construction methodologies, and evaluation criteria. We further review state-of-the-art AI algorithms from graph representation learning to hybrid reasoning approaches, and examine their applications in key drug discovery tasks such as drug-target identification, drug repurposing, combination therapies, and integration with large language models. Through this investigation, we highlight emerging opportunities and future directions that aim to guide researchers in harnessing the full potential of KGs for novel therapeutic development.

Open Access **http://dx.doi.org/10.5626/JCSE.2025.19.1.1** **http://jcse.kiise.org**

Received 05 April 2025; **Accepted** 11 May 2025
*Corresponding Author

†These authors contributed equally to this work.

 pISSN: 1976-4677 eISSN: 2093-8020

Category: Bioinformatics

Keywords: Knowledge graph; Artificial intelligence; Drug discovery; Graph learning; Survey

I. INTRODUCTION

Drug discovery is a complex, costly, and high-risk endeavor, traditionally relying on extensive experiments and clinical trials. In recent years, artificial intelligence (AI) is reshaping biomedical research, offering powerful approaches to model intricate biological relationships, identify promising therapeutic candidates, and predict clinical outcomes with unprecedented accuracy.

However, drug discovery is a very complex task that involves various biological and medical entities, such as drugs, genes and diseases, and their relationships, e.g., drug-target interactions (DTIs) and disease-gene associations. Thus application of single or a few AI tools is not powerful enough to address the complexity of drug discovery. To leverage the accumulated knowledge in chemical, biological, and medical domains, scientists worked hard to put together such knowledge as single integrated resources, which is known as biomedical knowledge graphs (KGs). KGs enable the systematic integration and representation of vast, heterogeneous biomedical data, ranging from molecular interactions, genetic and epigenetic profiles, chemical structures, clinical observations, to literature-derived evidence. Specifically, heterogeneous knowledge graphs (HKGs), which integrate multi-source, multi-modality data, play a critical role by capturing complex biological phenomena through interconnected entities and relationships, thereby offering more comprehensive biological insights compared to homogeneous or simpler graph structures.

While the promise of KGs in drug discovery is widely recognized, the field is rapidly evolving, characterized by diverse approaches and applications. This survey paper is to provide a unified and comprehensive overview of the entire spectrum of AI-driven drug discovery (AIDD) through HKGs by systematically reviewing state-of-the-art methodologies, data resources, and practical applications of knowledge graphs in AI-enabled drug discovery (Fig. 1).

First, we provide a comprehensive overview of heterogeneous biomedical KGs, in terms of definitions, construction methodologies, and evaluation criteria (Section II). Second, we review various AI algorithms specifically tailored for learning from knowledge graphs, including graph representation learning, graph neural networks, and hybrid reasoning approaches (Section III). Third, we compile a broad spectrum of drug discovery tasks that can be addressed with knowledge graph methods: drug-target identification and drug response prediction, to advanced applications such as drug repurposing, combination therapies, multimodal data alignment, and the integration of knowledge graphs with large language models (LLMs) (Section IV).

Through this survey of state-of-the-art technologies in AI-drug discovery with KGs, we hope to help researchers find more effective use of knowledge graphs for accelerating the development of novel therapeutics.

II. HETEROGENEOUS KNOWLEDGE GRAPHS

A biomedical KG represents biological entities such as drugs, genes, and diseases, along with their relationships [1]. A HKG integrates multiple node and relationship types, unlike a homogeneous knowledge graph, which consists of a single type [2]. Since biomedical research requires integrating diverse data sources, HKGs are particularly important [3]. Fig. 2 illustrates a conceptual model of HKGs.

The key advantage of HKGs is their ability to infer

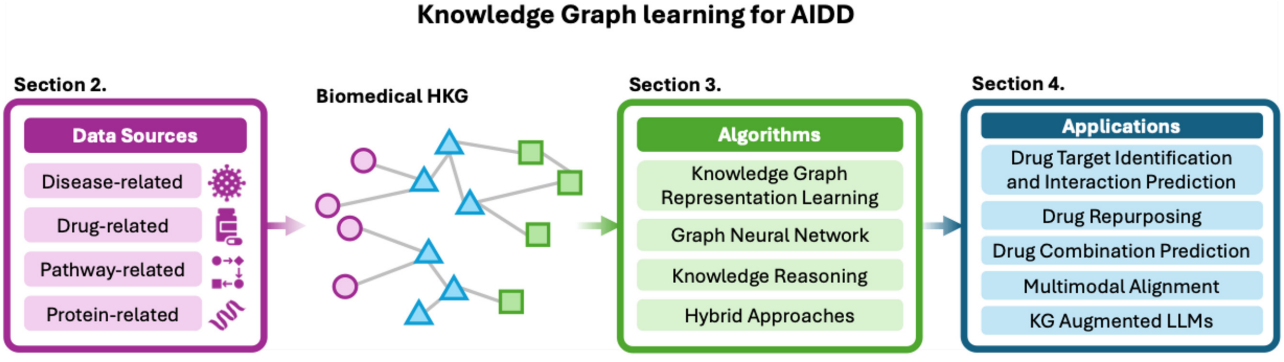


Fig. 1. Overview of AI-drug discovery with knowledge graphs (KGs). This survey integrates data sources (Section 2), algorithms (Section 3), and applications (Section 4) for KG-based AI-drug discovery.

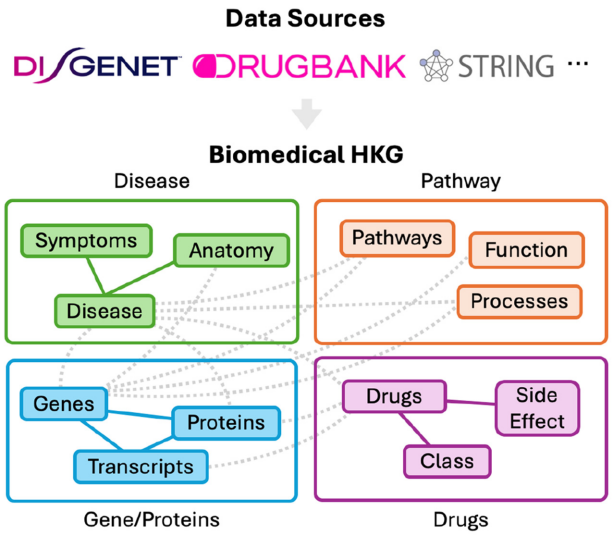


Fig. 2. Structure of biomedical heterogeneous knowledge graphs (HKGs). Biomedical HKGs aggregate diverse data sources into a single graph.

knowledge by linking diverse data types. While traditional research relied on single datasets [4, 5], HKGs facilitate an integrative approach, improving applications such as drug-target prediction and drug repurposing [1, 2, 3, 6].

A. Construction of Biomedical HKGs

The construction of an HKG requires defining key entities and relationships, integrating data from multiple sources, and ensuring a reliable graph structure [3, 6]. This section defines the key components of biomedical HKGs, discusses the challenges encountered during their construction, and introduces prominent existing biomedical KGs [5].

1) Key Components of Biomedical HKGs

Biomedical data is not merely structured data but consists of heterogeneous information, including drugs, genes, proteins, and diseases, which are intricately interconnected [4, 7]. Therefore, effectively integrating these diverse entities and relationships is essential. To achieve this, an HKG defines key nodes and relationships that reflect biological significance, typically comprising the elements outlined in Table 1.

The primary nodes in a HKG are classified into four types: drugs, genes/proteins, diseases, and biological pathways. Drug nodes include FDA-approved drugs and clinical candidates, while gene/protein nodes is to denote genetic variations and protein-protein interactions. Disease nodes provide information on specific diseases and genetic disorders, and biological pathway nodes represent metabolic and signaling pathways.

These nodes are interconnected through various relationships, including drug–target, gene–disease, drug–disease, and protein–protein interactions. Each relationship conveys biologically meaningful information. Beyond serving as a knowledge repository, HKGs enhance biomedical research and drug discovery by data source to AI-based analysis to uncover novel patterns and hidden associations.

2) Data Sources for Biomedical HKGs

The core aspect of HKG construction lies in the integration of diverse biomedical data sources. Table 2 compiles major data sources commonly used in biomedical HKGs [8-16]. These data sources provide distinct types of information and are utilized to define relationships between entities within a HKG.

Table 1. Nodes and edges of biomedical heterogeneous knowledge graph

Type	Category	Example
Node	Drug	FDA-approved drugs, clinical candidates
	Gene/Protein	Target genes, protein interactions
	Disease	Diseases, phenotype data
	Pathway	Metabolic pathways, signaling pathways
Edge	Drug-Target	Drug-target protein interactions
	Gene-Disease	Gene-disease associations
	Drug-Disease	Drug-disease therapeutic relations
	Protein-Protein	Protein-protein interactions
	Disease-Disease	Disease similarity or comorbidity relations
	Pathway-Gene	Gene involvement in biological pathways
	Pathway-Disease	Disease-related pathways
	Drug-Drug	Drug-drug interactions (DDI)

Table 2. List of data sources commonly utilized for construction of Biomedical heterogeneous knowledge graphs

Dataset	Relations Provided	Features	First released	Update frequency	Data access
Disease-related					
DisGeNET [8]	Disease-Disease, Disease-Gene	Evidence	2010	Annually	REST API, SPARQL, SQL
OMIM [9]	Disease-Gene	Text description, Evidence	1987	Daily	Flat file, REST API
GWAS Catalog [10]	Disease-Gene, Disease-Trait	Evidence	2008	Biweekly	REST API, Flat file
Drug-related					
DrugBank [11]	Drug-Drug, Drug-Gene, Drug-Disease	Text description, Structure, Attributes	2006	Annually	Flat file, REST API
PubChem [12]	Drug-Drug, Drug-Gene, Drug-Structure	Text description, Structure, Attributes	2004	As sources update	REST API, SPARQL, Flat file
Pathway-related					
KEGG [13]	Protein-Protein, Gene-Pathway, Drug-Pathway	Graph representation, Text description	1995	Bi-annually	REST API, Python, R
Reactome [14]	Protein-Protein, Gene-Pathway, Drug-Pathway	Graph representation, Text description	2003	Annually	Neo4J, Flat file
Protein-related					
STRING [15]	Protein-Protein	Types, Weightings	2003	Monthly	Flat file, REST API
BioGRID [16]	Gene-Gene, Protein-Protein	Types, Weightings	2003	Monthly	REST API, Flat file

B. Existing Knowledge Graphs

Various KGs have been developed to support drug discovery, differing in data sources, entity types, and relationship structures. Some representative KGs are as follows:

- HetioNet [6]: A large-scale KG for drug repurposing, integrating multiple datasets such as DrugBank, Gene Ontology, and DisGeNET.
- PrimeKG [17]: A precision medicine KG that incorporates clinical guideline texts to support AI-driven analysis.
- CKG (clinical knowledge graph) [18]: A KG for personalized medicine that integrates multi-omics data and includes post-translational modification information.
- PharMeBINet [19]: A pharmaceutical knowledge network that facilitates drug discovery and pharmaceutical research.
- BioKG [20]: A KG integrating biomedical entities from 13 data sources, including UniProt, Reactome, and OMIM.
- DRKG (drug repurposing knowledge graph) [21]: A KG integrating biomedical entities from 13 data sources, including UniProt, Reactome, and OMIM.
- PharmKG [22]: A KG benchmark for drug discovery and repurposing, including drug–disease and drug–target interactions.
- OpenBioLink [23]: A benchmark KG for evaluating biomedical KG completion models, incorporating data from multiple biological databases.

Further statistics and characteristics of each KGs are summarized in Table 3.

C. Challenges in Constructing Biomedical HKGs

The construction of a biomedical HKG faces several challenges, including data heterogeneity, reliability, updating difficulties, and sparsity [17, 24].

Data heterogeneity arises from the varied formats of biomedical data. Genomic data are often continuous (e.g., sequence information), drug data are typically discrete (e.g., molecular structures), and clinical data are frequently in natural language, requiring different processing techniques. Integrating these diverse data types into a unified framework remains a key challenge [17].

Data reliability varies across sources [17, 25]. Experimentally validated data are highly trustworthy but limited in volume due to cost and time constraints. In contrast, automatically extracted data from literature using natural language processing (NLP) techniques are abundant but prone to inaccuracies. Balancing data quality and quantity is crucial for building a reliable HKG.

KG update stems from the rapid pace of biomedical research [17, 24]. Static KGs struggle to incorporate new discoveries, leading to outdated relationships. For example,

Table 3. Statistics and use-cases of existing biomedical heterogeneous knowledge graphs utilized for drug discovery

HKG database	Nodes	Edges	Node types	Edge types	Design usecase	Last update
HetioNet [6]	47,031	2,250,197	11	24	Repurposing	2017
PrimeKG [17]	129,375	4,050,249	10	30	Repurposing	2022
CKG [18]	16 M	220 M	19	57	Personalized medicine	2021
PharMeBINet [19]	2,869,407	15,883,653	66	208	Drug discovery	2024
BioKG [20]	11,479,285	42,504,072	10	17	General	2024
DRKG [21]	97 K	5.7 M	13	107	Repurposing	2020
PharmKG [22]	7.6 K	500 K	3	29	Repurposing/target prediction	2021
OpenBioLink [23]	184 K	4.7 M	7	30	Benchmark	2020

failing to add newly identified drug–target interactions can reduce the predictive performance of AI models. Thus the KG update frequency is the key issue in choosing KG for drug discovery projects.

Graph sparsity is a common issue due to missing relationships [17, 26]. Rare disease targets are often underrepresented, and drug interactions for novel compounds may not yet be recorded. This incompleteness complicates relationship prediction and impacts AI model performance in drug discovery and disease research.

Addressing these challenges is critical to ensuring that biomedical HKGs remain comprehensive, reliable, and up-to-date for AI-driven applications.

III. ALGORITHMS FOR LEARNING KNOWLEDGE GRAPHS

A. Graph Learning AI for Knowledge Graphs

Graph AI learning technologies learn and infer patterns from graph-structured data, making it highly relevant to KGs. A KG represents entities (nodes) and relationships (edges) in a structured graph format, offering a more effective way to capture relationships than traditional tabular data. However, simple query-based approaches often fail to extract meaningful insights from KGs. Graph AI enables deeper learning of hidden patterns, prediction of new relationships, and more sophisticated knowledge inference [27].

KGs are structured representation of relationships between entities, modeled as graph-structured data. A KG is typically represented as a directed graph and formally defined as: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$.

In this definition, $\mathcal{V}$ represents the set of entities (nodes) in the graph, while $\mathcal{E}$ denotes the set of relations (edges) that connect these entities. The set of triples, $\mathcal{T}$, consists of directed edges that capture relationships between entities. Each triple (h, r, t) represents a directed edge from a head entity h to a tail entity t via a relation r, formally defined as $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{V}, r \in \mathcal{E}\}$. For example, a KG can represent information such as "Drug A inhibits Protein B," which can be encoded as the triple $(A, \text{inhibits}, B)$ [28].

To leverage KGs in machine learning models, graph-structured data must be converted into numerical representations. Typically, a KG is transformed into a node feature matrix X and an adjacency matrix A.

The node feature matrix X represents the attributes of entities within a KG. Each entity (node) $v \in \mathcal{V}$ has distinct properties that can be expressed as a d-dimensional feature vector $x_v \in \mathbb{R}^d$. The complete node feature matrix is given as $X \in \mathbb{R}^{|\mathcal{V}| \times d}$. This transformation allows machine learning models to process entity characteristics numerically.

The adjacency matrix A captures the relationships between entities by defining their connections. This matrix A has dimensions $|\mathcal{V}| \times |\mathcal{V}|$, where each entry indicates the existence of a relationship between two entities. For KGs containing multiple types of relations, multi-relational adjacency matrices are used, where each relation type r is associated with an independent matrix A_r.

Structuring KG data into X and A transforms entity attributes and relationships into numerical matrix representations, enabling machine learning models to effectively learn patterns and structural information within the KG.

1) Prediction on Graph AI

Graph AI performs various predictive tasks in KGs, primarily node prediction, edge prediction, and graph prediction [29].

Node-level prediction: This task involves predicting attributes of a node in a graph $f(v_i) = y$, where v_i is a specific node in the graph, and y represents the predicted attribute of that node. For instance, this approach can be used to predict the function y of an unknown protein v_i.

Edge-level prediction: This involves predicting whether an edge exists between two nodes, formulated as $f(v_i, v_j) = y$, $y \in \{0, 1\}$, where v_i and v_j are nodes in the KG, and y represents the presence ($y = 1$) or absence ($y = 0$) of an interaction between them. In drug discovery, this can be

used to predict whether a candidate drug v_i will bind to a target protein v_j.

Graph-level prediction: This task involves predicting properties of an entire graph, mathematically represented as $f(\mathcal{G}) = y$, where $\mathcal{G}$ is the entire graph, and y represents a specific property of the graph. In molecular graph analysis, the structure of a chemical compound can be modeled as a graph, and the prediction task can determine whether the compound exhibits toxicity, where $y \in \{0, 1\}$.

These methods allow KGs to be utilized for relationship prediction and data analysis in drug discovery.

2) Approaches of Graph AI

The main approaches in Graph AI evolve progressively from (1) knowledge graph embeddings (KGEs) → (2) graph neural networks (GNNs) → (3) knowledge reasoning. KGE transforms nodes and relationships in a KG into vector representations, enabling efficient computation but potentially losing structural information [30]. To address this, GNN learns patterns by incorporating neighboring node information, preserving graph structure but lacking logical reasoning capabilities [31]. Finally, knowledge reasoning enhances AI's ability to perform logical inference within a KG using rule-based and neural-symbolic reasoning approaches [32, 33]. In the following sections of this paper, each approach will be discussed individually.

B. Knowledge Graph Representation Learning

Relationship among nodes in KG is very complex and it is very difficult to make inference on KG for specific questions. One common technique is to make node representations that include neighbor nodes information or contextual information of a node under consideration [34]. This task, known as knowledge graph representation learning, can be broadly categorized into two approaches: shallow embedding and KGE.

1) Shallow Embedding

Shallow embedding is a technique for representing nodes in a KG as low-dimensional vectors, learning individual node representations by leveraging statistical and structural properties of the graph. Based on the learning approach, it can be categorized into structure-based methods and stochastic approaches.

The structure-based approach in shallow embedding learns node embeddings by analyzing the topological structure of the graph and mathematically modeling relationships between nodes.

High-order proximity embedding (HOPE) [35] is a model that preserves node similarity in a graph when learning embeddings. It generates a similarity-based matrix using the Katz Index, which considers all possible paths between nodes while assigning higher weights to shorter paths, and then applies singular value decomposition (SVD) to obtain low-dimensional vectors. Global graph representation learning (GraRep) [36] is a model that captures long-range relationships between distant nodes to learn the global structure of the graph. It computes a k-step transition probability matrix and applies SVD to transform it into low-dimensional vectors. This approach not only considers directly connected nodes but also learns relationships between frequently co-occurring nodes over multiple steps.

However, shallow embedding techniques learn only individual node features (e.g., drugs, proteins) and struggle to capture hidden relationships. Additionally, since it learns fixed patterns, it fails to account for environmental variations affecting drug responses or side effects. Furthermore, generating embeddings for all nodes in large-scale drug discovery KGs incurs high computational costs and lacks scalability.

The stochastic approach in Shallow Embedding learns node embeddings through probabilistic exploration (Random Walk) techniques, rather than directly modeling the entire graph structure. By simulating the exploration process, this approach efficiently scales to large KGs in drug discovery.

DeepWalk [37] generates random node sequences and applies the Skip-gram model to predict neighboring nodes, optimizing the objective, $\max_\theta \sum_{v \in V} \sum_{u \in N_s(v)} \log P(u|v; \theta)$, where $N_s(v)$ is the set of neighboring nodes of v. Node2vec [38] improves DeepWalk by balancing breadth-first search (BFS; local) and depth-first search (DFS; global) exploration to capture diverse relationships.

Personalized PageRank (PPR) [39] extends random walks by prioritizing specific nodes through a teleport probability α, defined as $\pi = (1-\alpha)A\pi + \alpha v$, where π is the node importance vector, A is the adjacency matrix, and v is the preference vector. This ensures frequent returns to key nodes, improving embedding quality.

While stochastic approaches are scalable and flexible, they do not explicitly capture relationships between nodes, limiting their ability to infer new connections in complex biological networks. KGE techniques address this by jointly learning node and relationship representations, which will be discussed in the next section.

2) Knowledge Graph Embedding

A KG represents entities and relationships in a structured format, but its high-dimensional and sparse nature makes effective utilization challenging. Traditional shallow embedding methods learn node representations based on structural patterns but fail to explicitly capture relationships, limiting their ability to model relational dependencies [40]. In contrast, KGE jointly learns entity and relationship representations, preserving both structural and semantic information, thereby enabling more precise reasoning and prediction within the KG. KGE methods can be categorized into translational distance-based and semantic matching-based approaches, depending on how they model relationships.

Translational distance-based models represent relationships as translation operations in a continuous vector space,

mapping both entities and relations into the same embedding space. Relations act as transformations between entity vectors, with TransE [41], TransH [28], and TransR [42] being the most well-known models.

TransE [41] enforces $h+r\approx t$, meaning that applying relation r to entity h should result in entity t. It optimizes a margin-based ranking loss to distinguish correct triples from incorrect ones: $\mathcal{L}=\sum_{(h,r,t)\in\mathcal{D}}\sum_{(h',r,t')\in\mathcal{D}'}\max(0,\gamma+d(h+r,t)-d(h'+r,t'))$, where $d(x, y)$ is the distance metric, γ is the margin hyperparameter, and $\mathcal{D}, \mathcal{D}'$ represent valid and corrupted triples. While efficient, TransE struggles with one-to-many, many-to-one, and many-to-many relations, as it assumes a simple translation.

TransH [28] mitigates this by projecting entity vectors onto relation-specific hyperplanes, allowing for better multi-relational modeling. TransR [42] further extends this by separating entity and relation spaces unlike previous models that represent both in the same space, using a relation-specific transformation matrix M_r to map entities before applying translation.

Semantic matching-based models represent entities and relations using matrix operations or tensor decomposition, directly modeling relationships through mathematical operations rather than minimizing distances like translational distance-based models. This enables more precise modeling of entity interactions. Key models in this category include DistMult [43], ComplEx [44], and RotatE [45].

DistMult [43] represents relations as diagonal matrices and scores triples (h, r, t) using the inner product: $f(h, r, t) = h^{\top}Rt$. While computationally efficient, DistMult assumes all relationships are symmetric, which limits its applicability in real-world KGs. This means that relations such as "Drug A inhibits Drug B" and "Drug B inhibits Drug A" are treated as identical.

ComplEx [44] addresses this by introducing complex-valued representations, using the real part of the following computation: $f(h, r, t) = \mathrm{Re}(h^{\top}R\bar{t})$, where $\bar{t}$ is the complex conjugate of entity vector t. By leveraging imaginary components, ComplEx can model both symmetric and asymmetric relations but at a higher computational cost.

RotatE [45] extends ComplEx by modeling relationships as rotations in complex space, defining the transformation as: $t = h \circ e^{ir}$, where e^{ir} is a complex rotation matrix defined by Euler's formula. This allows RotatE to naturally capture relational directionality, making it effective for cyclic and transitive relationships. However, it still struggles with high-dimensional, complex relations.

Semantic matching-based models capture complex relationships using structured mathematical operations but rely on static embeddings. To address their limitations, GNNs dynamically aggregate relational information, which will be explored next.

C. Graph Neural Networks

Traditional KGE methods are limited in fully capturing the structural properties and dynamic changes of the graph. They also require retraining whenever new entities or relationships are introduced [46]. To address these limitations, GNNs have been introduced. GNNs update embeddings dynamically through neighborhood aggregation, capturing both direct and indirect dependencies within the graph [47, 48]. Additionally, they can adapt to newly introduced entities and relationships, overcoming the rigidity of KGE models [49]. This section explores GNN models including GCN [50], GAT [51], and heterogeneous GNN (HGNN) [52].

1) Basic Concept of GNN

GNNs dynamically update entity embeddings by aggregating neighboring node information, which is effective in capturing both direct and indirect dependencies. The general update rule is defined as follows [48]: $h_v^{(k)}=\sigma\left(W^{(k)}\cdot\sum_{u\in N(v)}\frac{1}{|N(v)|}h_u^{(k-1)}\right)$. In this equation, $h_v^{(k)}$ represents the node embedding at layer k. The trainable weight matrix $W^{(k)}$ is responsible for transforming node information at each layer. Neighboring nodes $N(v)$ are aggregated to update v's embedding. The activation function σ, such as ReLU or sigmoid, adds nonlinearity to enhance expressiveness.

This formula explains how neighboring node information propagates sequentially across layers, progressively refining node representations. The deeper the GNN model, the farther each node can receive information from its neighbors, ultimately incorporating global structural properties of the entire graph.

2) Existing GNN Architectures

GNNs have several variations that allow them to learn different types of graph structures, making them applicable to KGs. While traditional GNNs are optimized for homogeneous graphs, KGs contain various entity types and relation types, making them heterogeneous graphs. Therefore, specialized GNN models have been developed to better capture the structural complexity of KGs. This section describes key GNN models and their mechanisms in the context of KGs.

The graph convolutional network (GCN) [50] extends the concept of local filters in convolutional neural networks (CNNs) [53] to graph data, making it an efficient neural network model for incorporating information from neighboring nodes. While traditional KGE models represent relationships between entities as static vectors, GCN leverages the structural properties of graphs to learn more refined relationships between nodes. The core concept of GCN is spectral graph convolution, where each node aggregates information from its connected neighbors to update its own embedding.

GCNs aggregate features from neighboring nodes to update node embeddings following update rules:

$$H^{(k+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(k)}W^{(k)}\right). \quad (1)$$

In Eq. (1), $H^{(k)}$ represents the node representation at layer k. The term $\tilde{A} = A + I$ denotes the adjacency matrix with self-loops. The matrix $\tilde{D}$ is the degree normalization matrix.

Specifically, GCNs employ a normalized degree matrix $\tilde{D}$, which adjusts the information flow within the graph and ensures balanced message passing.

GCNs are structurally simple and computationally efficient, but they are primarily designed for standard graphs and require further extensions for graphs with diverse relationships, such as KGs. To address this limitation, the relational GCN (R-GCN) [54] was introduced as an extension GCN for multi-relational graphs, applying different weight matrices for each relationship type and thus enabling the propagation of information while distinguishing between various relationships.

Graph attention network (GAT) [51] is a GNN model that enhances node feature aggregation by incorporating an attention mechanism. Unlike traditional models such as GCN, which assign uniform weights to neighboring nodes, GAT dynamically adjusts the importance of each neighbor using self-attention. This allows the model to prioritize more relevant nodes in real-world graphs where neighbor significance varies.

GAT computes attention scores between node pairs and normalizes them using a softmax function to determine attention weights.

A key advantage of GAT is its adaptability to different graph structures. Unlike GCN, which relies on a fixed adjacency matrix, GAT learns an optimal neighborhood weighting dynamically. This flexibility makes it particularly effective for a wide range of graph-based learning tasks.

HGNNs [52] are designed to learn from heterogeneous graphs, which consist of multiple types of nodes (e.g., drugs, proteins, diseases) and multiple types of relationships (e.g., binding, inhibition, expression). Unlike traditional GNN models such as GCN and GAT, which treat all nodes and edges uniformly, HGNN incorporates meta-paths to capture relational transitions and assigns different importance to each type of relationship, enabling more refined graph learning.

While R-GCN handles multi-relational graphs by learning separate weight matrices for each relation type, HGNN goes beyond simple relation-specific weight assignments. Instead, it models relational transitions, utilizes meta-paths for information aggregation, and applies attention mechanisms to extract meaningful insights. This allows HGNN to effectively learn the complex interactions among drugs, proteins, genes, and diseases in KGs for drug discovery, where multiple entity types and complex relationships exist.

D. Knowledge Reasoning and Hybrid Approaches

Knowledge reasoning enables AI to infer implicit relationships and complete incomplete KGs, going beyond KGE and GNN for deeper knowledge understanding. For instance, even if a drug-protein interaction is missing in a KG, AI can infer it by analyzing similar drugs, biological pathways, and relational patterns. However, a single approach (KGE or GNN) is insufficient to fully capture KG complexity. This is mainly due to the heterogeneous graph nature of KG, i.e., multiple node types and complex interactions or edge types among nodes. Hybrid approaches combine multiple techniques to enhance reasoning and improve accuracy. This section explores knowledge reasoning methods and hybrid AI approaches.

1) Knowledge Reasoning Approaches

Knowledge reasoning is the process of inferring hidden relationships and patterns within a KG to generate new knowledge. A KG is a structured dataset composed of entities (nodes) and relations (edges), making it more effective at representing structural relationships compared to traditional tabular data. However, simple query-based retrieval methods often fail to extract meaningful hidden insights. By leveraging knowledge reasoning, it is possible to perform logical inference based on the structural information of the KG and predict new relationships through data-driven learning. Knowledge reasoning approaches can be broadly categorized into rule-based reasoning and neural-based reasoning.

The first approach, rule-based reasoning, applies explicit logical rules to generate new relationships within a KG. These methods perform deductive reasoning, where conclusions are logically derived based on predefined rules. First-order logic (FOL) [55] represents relationships between entities using logical rules to facilitate inference, while rule-based inference engines [56] perform inference based on predefined rules specified by users. Ontology-based reasoning (OWL, RDF schema) [57] utilizes hierarchical relationships to expand knowledge. Rule-based reasoning provides high explainability and follows explicit logical rules, making it a reliable inference meth in automatically learning complex patterns.

The second approach, neural-based reasoning, utilizes machine learning and deep learning models to automatically learn patterns within the KG and predict hidden relationships. While rule-based reasoning relies on human-defined logic, neural-based reasoning derives new relationships by analyzing data. This approach employs inductive reasoning and probabilistic reasoning, making it more scalable and effective for large-scale KGs.

2) Hybrid Approaches for Knowledge Graphs

Hybrid approaches combine graph AI with deep learning models to enable more sophisticated reasoning and prediction. While standalone models such as KGE and

GNN each offer distinct advantages, they also exhibit limitations when used independently. To overcome these challenges, recent research has introduced various hybrid approaches, including graph-based hybrid models that combine GNN and KGE, contextual graph models that integrate GNN with Transformer, and knowledge-augmented AI that fuses LLM with KG.

Two primary methods for learning structural information within a KG are KGE and GNN. By combining these two approaches, GNN and KGE models, such as CompGCN [58], can jointly leverage both structural (GNN) and semantic (KGE) information, leading to more powerful relationship prediction.

While GNN excels at learning relationships between entities in a graph, Transformer-based models are optimized for understanding natural language context. However, KGs lack contextual information, making it difficult to extract nuanced meanings, whereas Transformers, despite their strong contextual understanding, struggle to learn structural relationships directly. By integrating GNN with Transformer, models such as heterogeneous graph transformer (HGT) [59], which dynamically learns relation-specific weights, and Graphormer [60], which enhances graph relational representation using Transformer mechanisms, make it possible to learn both the structural information of a KG and the contextual meaning of natural language, enabling more refined reasoning.

Among hybrid approaches, integrating LLM with KG is emerging as the most powerful knowledge reasoning model. KGs provide structured relational information that serves as a reliable knowledge base, but they lack a natural language interface for question-answering. Conversely, LLMs possess robust capabilities for understanding and generating natural language but often suffer from factual inconsistencies and hallucinations. By integrating KG with LLMs, logically curated information can be incorporated into model training, allowing LLMs to generate more accurate responses while preserving contextual coherence.

One prominent method for achieving LLM-KG integration is retrieval-augmented generation for graphs (Graph RAG) [61]. In this approach, a user query is processed as follows:

First, the LLM converts the user's natural language query into a graph query (e.g., Cypher, SPARQL), then the query retrieves structured information from the KG. Lastly, the retrieved structured data is fed into the LLM, which synthesizes the information and generates response.

By leveraging structured knowledge from the KG, Graph RAG enhances response reliability and reduces hallucination issues, making it a far superior approach compared to conventional retrieval-based methods that rely solely on textual data. Ultimately, it enables more powerful reasoning than simply searching documents without utilizing a KG.

Overall, the introduced hybrid approaches bridge the limitations of individual techniques by combining KGs with AI models to enable more sophisticated reasoning.

IV. KNOWLEDGE GRAPH LEARNING TASKS FOR DRUG DISCOVERY

While previous sections examined the fundamental structure of heterogeneous biomedical networks and the technical approaches to learning from networks, this section explores the practical applications of KGs for AIDD. Each subsection explores a different pharmaceutical task where the KG paradigm offers unique advantages beyond traditional methods. Unlike traditional methods, KG based approaches are capable of integrating diverse data types—from molecular structures to genetic expressions to clinical outcomes—while preserving the natural relationships between biological entities. The key focus areas includes: drug target identification and interaction prediction, drug repurposing, drug combination prediction, multimodal alignment, and integration with large language models.

A. Drug Target Identification and Interaction Prediction

Drug target identification and interaction prediction is the process of discovering and validating specific biological molecules (like proteins, genes, or receptors) that can be modulated by drugs to produce therapeutic effects, and then predicting how potential drug compounds will bind to and interact with these targets to help and establish drug design and development strategies.

KGs enhance molecular interaction predictions by integrating protein-protein interaction networks with drug chemical structures. They improve accuracy by incorporating both structural and functional protein information, uncover binding mechanisms through graph patterns, and enhance off-target predictions via network proximity analysis.

Several models leverage KGs for DTI prediction. DTINet [62] applies a random walk with restart on heterogeneous KGs, using network diffusion patterns to identify functionally similar drugs and targets. KGENFM [63] learns drug and target embeddings from a heterogeneous KG and combines them with neural factorization machines, improving interaction prediction for novel targets. TriModel [64] treats DTI prediction as a direct link prediction task, using tensor factorization with three embedding vectors per entity to capture complex interaction patterns and handle sparse data. NeoDTI [65] utilizes neural message passing across different edge types, updating node embeddings by aggregating neighborhood information while preserving network topology through reconstruction objectives. These models demonstrate the power of KGs in improving DTI prediction, offering more effective strategies for drug discovery.

B. Drug Repurposing

Drug repurposing identifies new therapeutic applications for approved drugs to treat diseases that are not approved when drug was approved. If successful, drug repositioning can accelerate drug development timelines and reduce drug development costs because many characteristics of existing drugs, e.g., toxicity and PK/PD, are already well supported. Thus, this approach is particularly valuable for rare diseases, pandemic responses, and conditions with limited treatment options.

KGs enhance drug repurposing by structuring complex drug-target-disease interactions into an interconnected network. Traditional methods struggle with the vast combinatorial space of genes, pathways, and drugs, making systematic evaluation challenging. KGs overcome this by capturing biomedical relationships in an interpretable framework, allowing algorithms to uncover hidden associations and efficiently traverse molecular-to-clinical interactions.

Several KG-based models have demonstrated success. For example, KG-Predict [66] integrates genotypic and phenotypic data to predict novel drug-disease associations. DREAMwalk [67] employs a multi-layer semantic strategy to analyze similarity patterns in drug-gene-disease networks. Project Rephetio [6] systematically analyzes network patterns from 29 public datasets to identify treatment candidates. Zhang et al. [68] applied KG completion techniques to prioritize potential COVID-19 treatments, showcasing the method's ability to rapidly navigate complex biomedical landscapes. These approaches highlight the potential of KGs in accelerating drug repurposing by systematically leveraging existing biomedical knowledge.

C. Drug Combination Prediction

Drug combination prediction identifies effective drug pairs that enhance therapeutic outcomes beyond single-drug treatments. This approach is crucial for overcoming treatment resistance and addressing complex diseases while minimizing side effects and reducing the time and cost of traditional trial-and-error experiments. Use of already approved drugs has advantages for avoiding toxicity and better PK/PD results, compared to traditional trial-and-error experimental approaches for developing new drugs.

KGs offer a powerful framework for predicting drug interactions by capturing the complex relationships among drugs, targets, pathways, and side effects. Unlike traditional models that treat drug combinations as isolated pairs, KGs provide a holistic view of the biomedical landscape, enabling the identification of synergistic effects and potential adverse interactions. By representing biological entities as interconnected nodes, KGs allow researchers to analyze shared or complementary pathways, revealing mechanisms that drive drug synergy or antagonism.

Several innovative approaches support the effectiveness of KG-based drug combination prediction. Decagon [69] integrates protein-protein interactions and drug-protein targets to improve polypharmacy side effect prediction by 69% over baseline methods. Gu et al. [70] employ supervised contrastive learning to enhance drug-drug interaction modeling, particularly in addressing the challenge of negative sample scarcity. KG2ECapsule [71] expands beyond binary classification by modeling diverse drug-drug relationships, while SimVec [72] mitigates the "cold start" problem for unseen drugs by leveraging structure-aware node initialization and weighted similarity edges. Other notable approaches include Bean et al.'s model [73] for adverse reaction prediction, which examines molecular targets and pathway effects, and tumor-biomarker KG [74], which uncovers mechanistic insights into drug interactions. These methods not only predict effective or harmful drug combinations but also provide interpretable insights through graph-based reasoning, making them highly valuable for clinical decision-making.

D. Multimodal Alignment for Harmonized Data Integration

Multimodal alignment in drug discovery refers to the harmonization of diverse data sources, including genomic, protein structural, clinical, and chemical information, to enhance drug efficacy and toxicity predictions. This comprehensive approach reduces development time and costs while improving success rates by capturing complex biological interactions that single-modal methods often overlook.

Traditional models struggle to bridge the gap between molecular structures and functional properties while preserving chemical validity. KGs address this challenge by structuring chemical domain knowledge, acting as a semantic bridge between different modalities. For example, KANO [75] employs KG-guided graph augmentation during contrastive learning, linking atomic structures to functional properties through KG-derived prompts, thereby improving both predictive performance and interpretability. Expanding on this idea, KEDD [76] integrates structured and unstructured knowledge sources using feature fusion techniques, effectively handling missing modality issues and creating a more robust drug discovery framework.

E. Knowledge Graph-Augmented LLMs

Drug discovery begins to benefit from the remarkable advances in LLMs. Integration of KGs with LLMs has become a new powerful approach to improving molecular understanding and biomedical reasoning. KGs provide structured representations of biomedical relationships, while LLMs process and interpret unstructured textual knowledge. Their integration bridges the gap between relational data and natural language, enabling more

comprehensive drug discovery insights.

A hybrid KG-LLM systems, BioBRIDGE, aligns independently trained biomedical foundation models across proteins, molecules, and text using transformation modules learned from relational triplets in PrimeKG, facilitating DTI prediction and disease-driven drug discovery without fine-tuning.

CLADD, in contrast, employs RAG to dynamically retrieve biomedical knowledge from KGs, molecular databases, and predictive tools, generating contextualized insights for drug-target prediction, molecular annotation, and toxicity assessment. These models illustrate the potential of KG-LLM integration to improve interpretability, scalability, and adaptability in AI-driven biomedical research.

V. CONCLUSION

Biomedical KGs, alongside AI frameworks, are powerful and promising tools for drug discovery by systematically harnessing diverse biological data sources and uncovering complex biomedical relationships that traditional methods alone cannot reveal. In this survey, we provided a comprehensive overview of biomedical HKGs, detailed key methodologies for their construction, evaluation, and representation learning, and thoroughly reviewed state-of-the-art AI algorithms and drug discovery applications.

Through KG methodologies, including graph representation learning, GNNs, knowledge reasoning, and hybrid approaches, researchers have substantially advanced crucial drug discovery tasks such as DTI prediction, drug repurposing, drug response analysis, and multimodal alignment for harmonized data integration. The integration of HKGs with emerging LLMs further extends the potential of KG, enabling richer insights, more accurate predictions, and innovative discoveries.

However, significant challenges remain. Issues concerning data quality, completeness, interpretability, computational scalability, and standardized evaluation frameworks must be addressed to fully realize the promise of KGs. Future research directions should focus on enhancing HKG methodologies, developing more robust multimodal learning strategies, improving transparency and interpretability of AI models, and expanding the integration of KGs with advanced language models and reasoning frameworks.

In conclusion, the combination of KGs and AI is transforming biomedical research and accelerating development of therapeutics. Continued interdisciplinary collaboration among bioinformatics researchers, AI practitioners, domain experts, and clinical professionals will be essential to overcoming current challenges and fully exploiting the remarkable opportunities KGs present for the future of drug discovery.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ACKNOWLEDGEMENTS

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF), funded by the Ministry of Science and ICT, Republic of Korea (Grant No. RS-2022-NR067933), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program; Seoul National University) and AIGENDRUG Co. Ltd.

REFERENCES

1. A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, et al., "Knowledge graphs," *ACM Computing Surveys (Csur)*, viol. 54, no. 4, article no. 71, 2021. https://doi.org/10.1145/3447772
2. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, et al., "The EBI RDF platform: linked open data for the life sciences," *Bioinformatics*, vol. 30, no. 9, pp. 1338-1339, 2014. https://doi.org/10.1093/bioinformatics/btt765
3. T. Gaudelet, B. Day, A. R. Jamasb, J. Soman, C. Regep, G. Liu, et al., "Utilizing graph machine learning within drug discovery and development," *Briefings in Bioinformatics*, vol. 22, no. 6, article no. bbab159, 2021. https://doi.org/10.1093/bib/bbab159
4. T. J. Rintala, A. Ghosh, and V. Fortino, "Network approaches for modeling the effect of drugs and diseases," *Briefings in Bioinformatics*, vol. 23, no. 4, article no. bbac229, 2022. https://doi.org/10.1093/bib/bbac229
5. D. J. Rigden and X. M. Fernandez, "The 27th annual Nucleic Acids Research database issue and molecular biology database collection," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1-D8, 2020. https://doi.org/10.1093/nar/gkz1161
6. D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *eLife*, vol. 6, article no. e26726, 2017. https://doi.org/10.7554/eLife.26726
7. H. Chen, L. Ding, Z. Wu, T. Yu, L. Dhanapalan, and J. Y. Chen, "Semantic web for integrated network analysis in biomedicine," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 177-192, 2009. https://doi.org/10.1093/bib/bbp002
8. J. Pinero, J. M. Ramirez-Anguita, J. Sauch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, no. D1, pp. D845-D855, 2020. https://doi.org/10.1093/nar/gkz1021
9. J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "OMIM.org: leveraging knowledge across phenotype–gene relationships," *Nucleic Acids Research*, vol. 47, no. D1, pp.

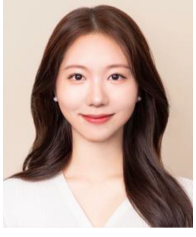
D1038-D1043, 2019. https://doi.org/10.1093/nar/gky1151
10. A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, et al., "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1005-D1012, 2019. https://doi.org/10.1093/nar/gky1120
11. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, et al., "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074-D1082, 2018. https://doi.org/10.1093/nar/gkx1037
12. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, et al., "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1388-D1395, 2021. https://doi.org/10.1093/nar/gkaa971
13. M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: integrating viruses and cellular organisms," *Nucleic Acids Research*, vol. 49, no. D1, pp. D545-D551, 2021. https://doi.org/10.1093/nar/gkaa970
14. B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, et al., "The reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 48, no. D1, pp. D498-D503, 2020. https://doi.org/10.1093/nar/gkz1031
15. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, et al., "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607-D613, 2019. https://doi.org/10.1093/nar/gky1131
16. C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34(suppl_1), pp. D535-D539, 2006. https://doi.org/10.1093/nar/gkj109
17. P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, article no. 67, 2023. https://doi.org/10.1038/s41597-023-01960-3
18. A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, P. E. Geyer, F. Coscia, et al., "Clinical knowledge graph integrates proteomics data into clinical decision-making," *bioRxiv*, 2020 [Online]. Available: https://doi.org/10.1101/2020.05.09.084897.
19. C. Konigs, M. Friedrichs, and T. Dietrich, "The heterogeneous pharmacological medical biochemical network PharMeBINet," *Scientific Data*, vol. 9, article no. 393, 2022. https://doi.org/10.1038/s41597-022-01510-3
20. Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, et al., "COVID-19 literature knowledge graph construction and drug repurposing report generation," 2020 [Online]. Available: https://arxiv.org/abs/2007.00576v1.
21. R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, and H. Kilicoglu, "Drug repurposing for COVID-19 via knowledge graph completion," 2020 [Online]. Available: https://arxiv.org/abs/2010.09600v1.
22. S. Zheng, J. Rao, Y. Song, J. Zhang, X. Xiao, E. F. Fang, Y. Yang and Z. Niu, "PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in Bioinformatics*, vol. 22, no. 4, article no. bbaa344, 2021. https://doi.org/10.1093/bib/bbaa344
23. A. Breit, S. Ott, A. Agibetov, and M. Samwald, "OpenBioLink: a benchmarking framework for large-scale biomedical link prediction," *Bioinformatics*, vol. 36, no. 13, pp. 4097-4098, 2020. https://doi.org/10.1093/bioinformatics/btaa274
24. K. Hansel, S. N. Dudgeon, K. H. Cheung, T. J. Durant, and W. L. Schulz, "From data to wisdom: biomedical knowledge graphs for real-world data insights," *Journal of Medical Systems*, vol. 47, no. 1, article no. 65, 2023. https://doi.org/10.1007/s10916-023-01951-2
25. C. Su, Y. Hou, M. Zhou, S. Rajendran, J. R. Maasch, Z. Abedi, et al., "Biomedical discovery through the integrative biomedical knowledge hub (iBKH)," *Iscience*, vol. 26, no. 4, article no. 106460, 2023. https://doi.org/10.1016/j.isci.2023.106460
26. A. Jimenez, M. J. Merino, J. Parras, and S. Zazo, "Explainable drug repurposing via path based knowledge graph completion," *Scientific Reports*, vol. 14, no. 1, article no. 16587, 2024. https://doi.org/10.1038/s41598-024-67163-x
27. C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: opportunities and challenges," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 13071-13102, 2023. https://doi.org/10.1007/s10462-023-10465-9
28. Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, pp. 1112-1119, 2014. https://doi.org/10.1609/aaai.v28i1.8870
29. R. Johnson, M. M. Li, A. Noori, O. Queen, and M. Zitnik, "Graph artificial intelligence in medicine," *Annual Review of Biomedical Data Science*, vol. 7, pp. 345-368, 2024. https://doi.org/10.1146/annurev-biodatasci-110723-024625
30. C. Wang, Y. Yang, J. Song, and X. Nan, "Research progresses and applications of knowledge graph embedding technique in chemistry," *Journal of Chemical Information and Modeling*, vol. 64, no. 19, pp. 7189-7213, 2024. https://doi.org/10.1021/acs.jcim.4c00791
31. Z. Zhang, P. Cui, J. Pei, X. Wang, and W. Zhu, "Eigen-GNN: a graph structure preserving plug-in for GNNs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2544-2555, 2023. https://doi.org/10.1109/TKDE.2021.3112746
32. Z. Zeng, Q. Cheng, and Y. Si, "Logical rule-based knowledge graph reasoning: a comprehensive survey," *Mathematics*, vol. 11, no. 21, article no. 4486, 2023. https://doi.org/10.3390/math11214486
33. L. N. DeLong, R. F. Mir, and J. D. Fleuriot, "Neurosymbolic AI for reasoning over knowledge graphs: a survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 7822-7842, 2025. https://doi.org/10.1109/TNNLS.2024.3420218
34. G. A. Gesese, R. Biswas, M. Alam, and H. Sack, "A survey on knowledge graph embeddings with literals: Which model links better literal-ly?," *Semantic Web*, vol. 12, no. 4, pp. 617-647, 2020. https://doi.org/10.3233/SW-200404
35. M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1105-1114. https://doi.org/10.1145/2939672.2939751
36. S. Cao, W. Lu, and Q. Xu, "GraRep: learning graph representations with global structural information," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Melbourne, Australia, 2015, pp. 891-900. https://doi.org/10.1145/2806416.2806512

37. B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 701-710. https://doi.org/10.1145/2623330.2623732
38. A. Grover and J. Leskovec, "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 855-864. https://doi.org/10.1145/2939672.2939754
39. S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117, 1998. https://doi.org/10.1016/S0169-7552(98)00110-X
40. L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "struc2vec: learning node representations from structural identity," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 385-394. https://doi.org/10.1145/3097983.3098061
41. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2787-2795, 2013.
42. Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, pp. 2181-2187, 2015. https://doi.org/10.1609/aaai.v29i1.9491
43. B. Yang, W. T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," 2014 [Online]. Available: https://arxiv.org/abs/1412.6575v1.
44. T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," *Proceedings of Machine Learning Research*, vol. 48, pp. 2071-2080, 2016.
45. Z. Sun, Z. H. Deng, J. Y. Nie, and J. Tang, "RotatE: knowledge graph embedding by relational rotation in complex space," 2019 [Online]. Available: https://arxiv.org/abs/1902.10197.
46. R. Biswas, L. A. Kaffee, M. Cochez, S. Dumbrava, T. E. Jendal, M. Lissandrini, M., ... & De Melo, G. (2023). "Knowledge graph embeddings: open challenges and opportunities," *Transactions on Graph Data and Knowledge*, vol. 1, no. 1, article no. 4, 2023. https://dx.doi.org/10.4230/TGDK.1.1.4
47. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *Proceedings of Machine Learning Research*, vol. 70, pp. 1263-1272, 2017.
48. K. Xu, C. Li, Y. Tian, T. Sonobe, K. I. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," *Proceedings of Machine Learning Research*, vol. 80, pp. 5453-5462, 2018.
49. T. Hamaguchi, H. Oiwa, M. Shimbo, and Y. Matsumoto, "Knowledge transfer for out-of-knowledge-base entities: a graph neural network approach," 2017 [Online]. Available: https://arxiv.org/abs/1706.05674.
50. T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016 [Online]. Available: https://arxiv.org/abs/1609.02907v1.
51. P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," 2017 [Online]. Available: https://arxiv.org/abs/1710.10903v1.
52. C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, USA, 2019, pp. 793-803. https://doi.org/10.1145/3292500.3330961
53. K. O'shea and R. Nash, "An introduction to convolutional neural networks," 2015 [Online]. Available: https://arxiv.org/abs/1511.08458.
54. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The Semantic Web*. Cham, Switzerland: Springer, 2018, pp. 593-607. https://doi.org/10.1007/978-3-319-93417-4_38
55. S. N. Artemov and T. Yavorskaya, "First-order logic of proofs," 2011 [Online]. Available: https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1354&context=gc_cs_tr.
56. D. Carral, I. Dragoste, L. Gonzalez, C. Jacobs, M. Krotzsch, and J. Urbani, "VLog: a rule engine for knowledge graphs," in *The Semantic Web*. Cham, Switzerland: Springer, 2019, pp. 19-35. https://doi.org/10.1007/978-3-030-30796-7_2
57. A. Polleres, A. Hogan, R. Delbru, and J. Umbrich, "RDFS and OWL reasoning for linked data," in *Reasoning Web International Summer School*. Heidelberg, Germany: Springer, 2013, pp. 91-149. https://doi.org/10.1007/978-3-642-39784-4_2
58. S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," 2019 [Online]. Available: https://arxiv.org/abs/1911.03082v1.
59. Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of the Web Conference 2020*, Taipei, Taiwan, 2020, pp. 2704-2710. https://doi.org/10.1145/3366423.3380027
60. C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T. Y. Liu, "Do transformers really perform badly for graph representation?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28877-28888, 2021.
61. D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," 2024 [Online]. Available: https://arxiv.org/abs/2404.16130v1.
62. Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Communications*, vol. 8, article no. 573, 2017. https://doi.org/10.1038/s41467-017-00680-8
63. Q. Ye, C. Y. Hsieh, Z. Yang, Y. Kang, J. Chen, D. Cao, S He, and T. Hou, "A unified drug–target interaction prediction framework based on knowledge graph and recommendation system," *Nature Communications*, vol. 12, article no. 6775, 2021. https://doi.org/10.1038/s41467-021-27137-3
64. S. K. Mohamed, V. Novacek, and A. Nounu, "Discovering protein drug targets using knowledge graph embeddings," *Bioinformatics*, vol. 36, no. 2, pp. 603-610, 2020. https://doi.org/10.1093/bioinformatics/btz600
65. F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "NeoDTI: neural integration of neighbor information from a

heterogeneous network for discovering new drug–target interactions," *Bioinformatics*, vol. 35, no. 1, pp. 104-111, 2019. https://doi.org/10.1093/bioinformatics/bty543
66. Z. Gao, P. Ding, and R. Xu, "KG-Predict: a knowledge graph computational framework for drug repurposing," *Journal of Biomedical Informatics*, vol. 132, article no. 104133, 2022. https://doi.org/10.1016/j.jbi.2022.104133
67. D. Bang, S. Lim, S. Lee, and S. Kim, "Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers," *Nature Communications*, vol. 14, article no. 3570, 2023. https://doi.org/10.1038/s41467-023-39301-y
68. R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, and H. Kilicoglu, "Drug repurposing for COVID-19 via knowledge graph completion," *Journal of Biomedical Informatics*, vol. 115, article no. 103696, 2021. https://doi.org/10.1016/j.jbi.2021.103696
69. M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457-i466, 2018. https://doi.org/10.1093/bioinformatics/bty294
70. J. Gu, D. Bang, J. Yi, S. Lee, D. K. Kim, and S. Kim, "A model-agnostic framework to enhance knowledge graph-based drug combination prediction with drug–drug interaction data and supervised contrastive learning," *Briefings in Bioinformatics*, vol. 24, no. 5, article no. bbad285, 2023. https://doi.org/10.1093/bib/bbad285
71. X. Su, Z. You, D. Huang, L. Wang, L. Wong, B. Ji, and B. Zhao, "Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 5640-5651, 2023. https://doi.org/10.1109/TKDE.2022.3154792
72. N. Lukashina, E. Kartysheva, O. Spjuth, E. Virko, and A. Shpilman, "SimVec: predicting polypharmacy side effects for new drugs," *Journal of Cheminformatics*, vol. 14, article no. 49, 2022. https://doi.org/10.1186/s13321-022-00632-5
73. D. M. Bean, H. Wu, E. Iqbal, O. Dzahini, Z. M. Ibrahim, M. Broadbent, R. Stewart, and R. J. B. Dobson, "Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records," *Scientific Reports*, vol. 7, article no. 16416, 2017. https://doi.org/10.1038/s41598-017-16674-x
74. M. Wang, X. Ma, J. Si, H. Tang, H. Wang, T. Li, et al., "Adverse drug reaction discovery using a tumor-biomarker knowledge graph," *Frontiers in Genetics*, vol. 11, article no. 625659, 2021. https://doi.org/10.3389/fgene.2020.625659
75. Y. Fang, Q. Zhang, N. Zhang, Z. Chen, X. Zhuang, X. Shao, X. Fan, and H. Chen, "Knowledge graph-enhanced molecular contrastive learning with functional prompt," *Nature Machine Intelligence*, vol. 5, no. 5, pp. 542-553, 2023. https://doi.org/10.1038/s42256-023-00654-0
76. Y. Luo, X. Y. Liu, K. Yang, K. Huang, M. Hong, J. Zhang, Y. Wu, and Z. Nie, "Toward unified AI drug discovery with multimodal knowledge," *Health Data Science*, vol. 4, article no. 0113, 2024. https://doi.org/10.34133/hds.0113

Daeun Kong

Daeun Kong received her bachelor's degrees in Mathematics and Software & Computer Engineering from Ajou University in 2025. Since 2024, she has been working at AIGENDRUG Co. Ltd., where she has been involved in research related to the use of AI in drug development.



Yoojin Ha https://orcid.org/0009-0008-2976-9792

Yoojin Ha received her bachelor's degree in Computer Education from Sungkyunkwan University in 2025. She is currently pursuing a master's degree in the Interdisciplinary Program in Artificial Intelligence at Seoul National University. Her research interests include bioinformatics and machine learning.

HaEun Yoo

HaEun Yoo is currently an undergraduate student in the Department of Computer Science and Engineering at Seoul National University and a research intern at the BHI lab under Professor Sun Kim. Her research interests include bioinformatics, biomedical engineering, and machine learning.

Dongmin Bang https://orcid.org/0000-0001-9217-8380

Dongmin Bang is currently a PhD candidate in Bioinformatics at Seoul National University advised by Prof. Sun Kim and a senior research scientist at AIGENDRUG Co., Ltd. He received his received his Pharm.D. degree from Chung-Ang University. His research interests include precision medicine, intelligent knowledge integration and AI-augmented drug discovery.

Sun Kim https://orcid.org/0000-0001-5385-9546

Sun Kim received the Ph.D. degree in computer science from University of Iowa, in 1997. He is currently a professor with the School of Computer Science and Engineering, and an affiliated faculty of Interdisciplinary Program in bioinformatics with Seoul National University. Prof. Kim also is the co-founder of AIGENDRUG Co. Ltd. His research interests include bioinformatics, computational biology, machine learning, and data mining.