A Review on Knowledge Graphs for Healthcare: Resources, Applications, and Promises

CARL YANG*, HEJIE CUI, JIAYING LU, SHIYU WANG†, and RAN XU†, Emory University, USA
WENJING MA†, University of Michigan, USA
YUE YU†, Georgia Institute of Technology, USA
SHAOJUN YU†, XUAN KAN†, and CHEN LING, Emory University, USA
TIANFAN FU, Rensselaer Polytechnic Institute, USA
LIANG ZHAO and JOYCE HO, Emory University, USA
FEI WANG, Cornell University, USA

Healthcare knowledge graphs (HKGs) are valuable tools for organizing biomedical concepts and their relationships. The recent advance of large language models (LLMs) has paved the way for building more comprehensive and accurate HKGs. This, in turn, can improve the reliability and evaluation of LLMs. However, the challenges and opportunities of HKGs are not fully understood, highlighting the need for detailed reviews. This work provides the first comprehensive review of HKGs, summarizing the pipeline and key techniques for HKG construction and successful HKG utilization in various health-related applications. Lastly, we highlight the opportunities for HKGs in the era of LLMs.

CCS Concepts: • **Applied computing → Health informatics**; *Health care information systems*; *Bioinformatics*; • **Information systems → Graph-based database models**; • **Computing methodologies → Knowledge representation and reasoning**.

Additional Key Words and Phrases: knowledge graph, healthcare, language models, multimodality, interpretability, trustworthy AI

ACM Reference Format:
Carl Yang, Hejie Cui, Jiaying Lu, Shiyu Wang, Ran Xu, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Tianfan Fu, Liang Zhao, Joyce Ho, and Fei Wang. 2024. A Review on Knowledge Graphs for Healthcare: Resources, Applications, and Promises. 1, 1 (August 2024), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

A knowledge graph (KG) is a data structure that captures the relationships between different entities and their attributes [73, 115]. KG models and integrates data from various sources, including structured and unstructured data, and has been studied to support a wide range of applications such as search engines [153], recommendation systems [160, 199],

*Corresponding author: Carl Yang, j.carlyang@emory.edu.
†These authors contributed equally to this research.

Authors' Contact Information: Carl Yang; Hejie Cui; Jiaying Lu; Shiyu Wang; Ran Xu, Emory University, Atlanta, GA, USA; Wenjing Ma, University of Michigan, Ann Arbor, MI, USA; Yue Yu, Georgia Institute of Technology, Atlanta, GA, USA; Shaojun Yu; Xuan Kan; Chen Ling, Emory University, Atlanta, GA, USA; Tianfan Fu, Rensselaer Polytechnic Institute, Troy, NY, USA; Liang Zhao; Joyce Ho, Emory University, Atlanta, GA, USA; Fei Wang, Cornell University, New York, NY, USA.

Manuscript submitted to ACM

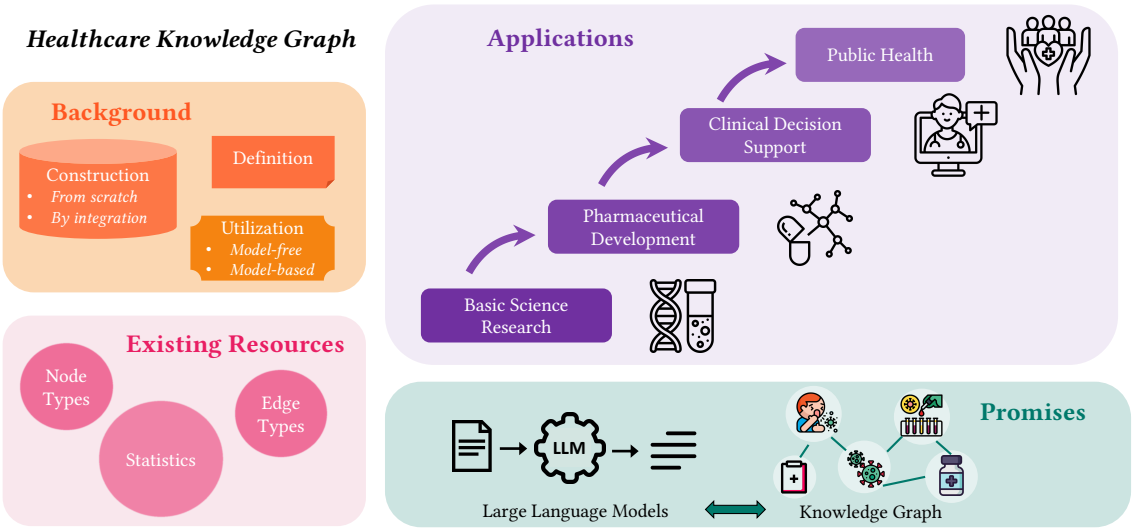

Fig. 1. The content overview of this review on healthcare knowledge graph.

and question answering [76, 88, 174, 176]. Healthcare Knowledge Graph (HKG) facilitates an interpretable representation of medical concepts, e.g., drugs and disease, as well as the relations among those medical concepts. This data structure enables the connection of contexts and enhances clinical research and decision-making [17, 128].

On the data side, HKG is usually built on complex medical systems such as electronic health records, medical literature, clinical guidelines, and patient-generated data [12, 123]. However, these data resources are often heterogeneous and distributed, which makes it challenging to integrate and analyze them effectively [107]. This data heterogeneity can also lead to incomplete or inconsistent data representations, limiting their usefulness for downstream healthcare tasks [24]. Additionally, the current use of domain-specific knowledge graphs may result in limited coverage and granularity of the knowledge captured across different levels. This hinders identifying correlations and relationships between medical concepts from multiple domains. These challenges highlight the need for continued research on HKGs to realize their full potential.

On the modeling side, the construction of HKGs can be done either from scratch or by integrating existing dataset resources. Many crucial steps, such as entity and relation extraction, can be optimized with natural language processing tools and algorithms. Recently, there have been significant advancements in general domain knowledge extraction, thanks to pre-trained large language models (LLMs) such as BERT [26], GPT Series [13], and others. These models revolutionize the field and make it possible to integrate heterogeneous medical data from various sources effectively. The use of pre-trained models has also led to the development of more accurate and comprehensive medical ontologies and taxonomies [159, 171, 184, 187, 190]. This allows for the evaluation of generated contents from LLMs and reduces LLM hallucination.

A comprehensive healthcare knowledge graph has the potential to contribute to health research across various levels [57, 85, 128]. At the micro-scientific level, HKGs can help researchers identify new phenotypic and genotypic correlations and understand the underlying mechanisms of disease [60], leading to more targeted and effective treatments [17, 131]. At the clinical care level, HKGs can be used to develop clinical decision support systems that provide

clinicians with relevant information, improving clinical workflows and patient outcomes [16, 35]. Therefore, a thorough review of existing literature on HKGs becomes an essential roadmap and invaluable resource to drive transformative advancements in the field.

This review is the first comprehensive overview of HKGs, covering contents shown in Figure 1. Specifically, in Section 2, we delve into the construction pipelines of HKGs, including both building from scratch and integration approaches, and highlight the key techniques employed in HKG construction. Additionally, we explore two standard utilization methods of HKGs, namely model-free and model-based approaches. In Section 3, we compile a comprehensive summary of existing HKG resources with their scopes and applications. Furthermore, Section 4 investigates the literature on mainstream health applications, offering an in-depth overview of the diverse use cases of HKGs in healthcare. Finally, we address promising research opportunities in the era of LLMs in Section 5.

2 Backgrounds

2.1 Knowledge Graphs for Healthcare (HKG)

Definition. A healthcare knowledge graph (HKG) is a domain-specific knowledge graph designed to capture medical concepts such as drugs, diseases, genes, phenotypes, and so on, and their relationships in a structured and semantic way.

Terminology. Our focus is on healthcare knowledge graphs (HKGs), which are structured, semantic representations of medical concepts and relationships. We also include ontologies and knowledge bases, which are commonly used in constructing HKGs. Ontologies define a set of concepts and categories in a domain, as well as the relationships between them, while knowledge bases store factual information about entities and their attributes. By covering these categories of terminology, we provide a comprehensive overview of the different types of resources available for organizing and representing medical knowledge in a structured and semantically rich manner.

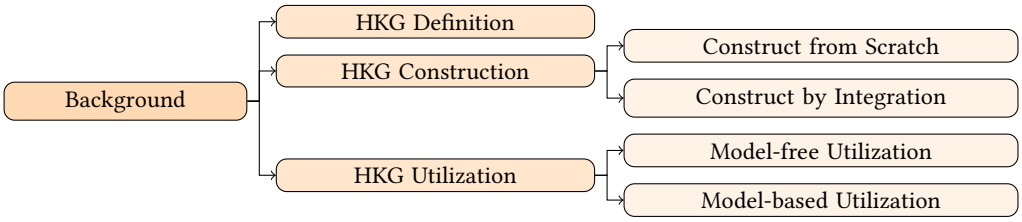

Fig. 2. Detailed taxonomy of the background section of healthcare knowledge graphs.

2.2 HKG Construction

Healthcare knowledge graphs can be constructed from scratch or through the integration of existing data resources.

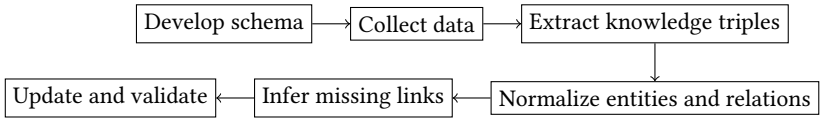

Fig. 3. The pipeline of constructing HKGs from scratch.

Constructing HKGs from Scratch. A multi-step pipeline, as in Figure 3, is used to construct HKGs from scratch.

(1) The first step is to identify the scope and objectives. In most cases, researchers develop a schema [9, 55] or use existing schemas [4, 6, 55, 130] to serve as the formal and explicit specification of a domain, thus ensuring consistent, coherent, and aligned domain knowledge. Unlike the general domain KG, utilizing schemas is a common practice in HKG construction.
(2) Secondly, researchers gather data from various sources, including medical literature, clinical trials, and patient-generated data. It is essential to ensure the quality and consistency of the data and to remove identifiable information for patient privacy.
(3) The third step is to extract and transform the data into a structured format. This step involves identifying medical entities and creating relationships between them via specialized biomedical Natural Language Processing (NLP) tools [58, 141, 169].
(4) Next, researchers map the entities and relationships to the chosen ontologies with the help of thesauruses [10] or terminologies [32, 65]. This step ensures that the knowledge graph is interoperable with other healthcare systems and facilitates data integration.
(5) Until now, an initial KG has been built. The next step is to populate the KG to infer missing links between entities. This inference can be done using graph databases [154] or link prediction models [11, 99].
(6) The final step is to continuously update and validate the KG to ensure accuracy and relevance. This step involves incorporating new data and knowledge, refining the schema, and evaluating the quality of the KG.

Constructing HKGs by Integration. Considering significant efforts have been paid to construct and curate HKGs from scratch, it is promising to integrate these data resources to avoid repetitive work. Healthcare KG integration (also called Healthcare KG fusion) refers to the processing of merging two or more HKGs into a single, more comprehensive graph [64, 142, 181]. The integration process is challenging because different HKGs may use different terminologies, schemas, or data formats. To address these challenges, researchers have developed various techniques and algorithms for knowledge graph fusion, including ontology matching [41, 61], schema alignment [103, 144], entity resolution [5, 68], and conflicts resolution [102]. These methods aim to identify and reconcile the differences between KGs.

Techniques for HKG Constructions. Traditionally, each step of HKG construction involves one specially designated model. For instance, Hidden Markov Models and Recurrent Neural Networks are widely used for healthcare named entity recognition, relation extraction, and other sequence tagging tasks, while Translational Models and Graph Neural Networks are used for HKG completion and conflicts resolution tasks. Recently, large language models (LLMs) have shown great utility to serve as a uniform tool for constructing KGs [177]. Several key steps of constructing KGs, such as named entity recognition [18, 70, 87, 92], relation extraction [100, 175, 203], entity linking [20, 25, 111], and KG completion [129, 136, 168**?**], have been successfully tackled by these large foundation models. Early explorations of construction HKG with large foundation models show that healthcare entity normalization [2, 191], healthcare entity recognition [45, 67], healthcare entity linking [200], and healthcare knowledge fusion [98] can also be performed, without extensive training on expensive healthcare annotated corpus. On the other hand, researchers start to construct KGs under the open-world assumption [23, 86, 99, 118, 137], thus getting rid of the dependency on pre-defined schemas and exhaustive entity&relation normalization. Although open-world KGs greatly increase the coverage, ensuring the quality of extracted knowledge is still an open research challenge, especially for explainable and trustworthy HKGs.

Table 1. Resource of existing healthcare knowledge graph (HKG).

Name	Node Types	Edge Types	Statistic	Application
HetioNet [63]	11 (e.g., drug, disease)	24 (e.g., drug-disease)	#N: 47.0 K, #E: 2.3 M	Medicinal Chemistry
DrKG [71]	13 (e.g., disease, gene)	107 (e.g., disease-gene)	#N: 97 K, #E: 5.8 M	Medicinal Chemistry
PrimeKG [17]	10 (e.g., phenotypes)	30 (e.g., disease-phenotype)	#N: 129.4 K, #E: 8.1 M	Medicinal Chemistry
Gene Ontology[a] [4]	3 (e.g., biological process)	4 (e.g., partOf)	#N: 43 K, #E: 7544.6K	Bioinformatics
KEGG[b] [77]	16 (e.g., pathway)	4 (e.g., partOf)	#N: 48.5 M, #E: unknown	Bioinformatics
STRING[c] [145]	1 (e.g., protein)	4 (e.g., interactions)	#N: 67.6 M, #E: 20 B	Bioinformatics
Cell Ontology[d] [27]	1 (i.e., cell type)	2 (e.g, subClassOf)	#N: 2.7 K, #E: 15.9 K	Bioinformatics
GEFA [124]	510 (e.g., kinases)	2 (e.g., drug-drug)	#N: 0.5 K, #E: 30.1 K	Drug Development
Reaction [83]	2 (e.g., reactant & normal)	19 (e.g., reaction paths)	#N: 2192.7 K, #E: 932.2 K	Drug Development
ASICS [72]	2 (e.g., reactant & product)	1 (e.g., reactions)	#N: 1674.9 K, #E: 923.8 K	Drug Development
Hetionet [72]	11 (e.g., biological process)	24 (e.g., disease–associates–gene)	#N: 47.0 K, #E: 2250.2 K	Drug Development
LBD-COVID [192]	1 (i.e., concept)	1 (i.e., SemMedDB relation)	#N: 131.4 K, #E: 1016.1 K	Drug Development
GP-KG [51]	7 (e.g., drug)	9 (e.g., disease–gene)	#N: 61.1 K, #E: 1246.7 K	Drug Development
DRKF [194]	4 (e.g., drug)	43 (e.g., drug-disease)	#N: 12.5 K, #E: 165.9 K	Drug Development
DDKG [53]	2 (i.e., drug & disease)	1 (e.g., drug-disease)	#N: 551, #E: 2.7 K	Drug Development
Disease Ontology[e] [130]	1 (i.e., disease)	2 (e.g., subClassOf)	#N: 11.2 K, #E: 8.8 K	Clinical Decision Support
DrugBank [162]	4 (e.g., drug, pathway)	4 (e.g., drug-target)	#N: 7.4 K, #E: 366.0 K	Clinical Decision Support
KnowLife [38]	6 (e.g., genes)	14 (e.g., gene-diseases)	#N: 2.9 M, #E: 11.4 M	Clinical Decision Support
PharmKG [198]	3 (e.g., diseases)	3 (e.g., chemical-diseases)	#N: 7601, #E: 500958	Clinical Decision Support
ROBOKOP[f] [8]	54 (e.g., genes, drugs)	1064 (e.g., biolink, CHEBI)	#N: 8.6M , #E: 130.4 M	Clinical Decision Support
iBKH[g] [142]	11 (e.g., anatomy, disease)	18 (e.g., anatomy-gene)	#N: 2.4 M, #E: 48.2 M	Clinical Decision Support

[a]http://geneontology.org/
[b]https://www.genome.jp/kegg/
[c]https://string-db.org/
[d]https://www.ebi.ac.uk/ols4/ontologies/cl
[e]https://disease-ontology.org/
[f]https://robokop.renci.org/
[g]https://github.com/wcm-wanglab/iBKH

2.3 HKG Utilization

Model-free Utilization. Various query languages can be used for KGs, such as SPARQL, Cypher, and GraphQL [154]. These query languages allow users to query healthcare KGs using a standardized syntax, thus enabling users to retrieve, manipulate, and analyze data in a structured and consistent way. More complex applications can be further supported by graph queries. For instance, automatic healthcare question answering can be tackled by Natural Language Question-to-Query (NLQ2Query) approach [80], where natural language questions are first translated into executable graph queries and then answered by the query responses. HKGs can also be utilized as an up-to-date and trustworthy augmentation to large language models (LLMs) for many applications. Some pioneering studies [56, 94, 138, 172] show that retrieved knowledge triples can improve the reliability of LLMs in various knowledge-intensive tasks, by addressing the nonsensical or unfaithful generation. Moreover, KGs can be a useful tool for fact-checking [106, 147, 151] as they provide a structured representation of information that can be used to quickly and efficiently verify the accuracy of claims. Researchers have explored the utility of HKGs in identifying ingredient substitutions of food [139], COVID-19 fact-checking [108], etc.

Model-based Utilization. Utilizing HKGs in complex reasoning tasks often involves utilizing machine learning models. HKG embeddings [143, 183] have shown great potential to tackle these tasks. In particular, HKG embedding models are a class of machine learning models that aim to learn low-dimensional vector representations, or embeddings, of the entities and relations in a knowledge graph. After obtaining HGK embeddings, they can be plugged into any kind of deep neural network and further fine-tuned toward downstream objectives. On the other hand, symbolic logic models represent another prominent approach for KG reasoning due to their interpretability. More specifically, symbolic reasoning models first mine logical rules from existing knowledge by inductive logic programming [112], association rule mining [49], or Markov logic networks [82]. These minded rules are used to infer new facts, make logical deductions and answer complex queries. Recently, researchers start to explore combining logical rules into KG embedding to further improve the generalization and performance of HKG reasoning [3, 202].

3 Resources

In this section, we compile a detailed resource overview of existing healthcare knowledge graphs to assist researchers and healthcare professionals in constructing and utilizing HKGs, organized in Table 1. We present key attribute information, including HKG name, node types, edge types, statistics, and their applications. Details and external links can be referred to at our public repository[1].

4 Applications

4.1 Basic Science Research

Several previous biological terms can also be considered knowledge graphs, such as ontology (gene ontology, cell ontology, disease ontology), network (gene regulatory network), etc. We use the original biological terms as they are more popular according to historical reasons.

4.1.1 Medicinal Chemistry. Topics related to medicinal chemistry involve drug-drug interactions (DDIs) and drug-target interactions (DTIs), which will be discussed in this section.

[1]Resource: https://github.com/lujiaying/Awesome-HealthCare-KnowledgeBase

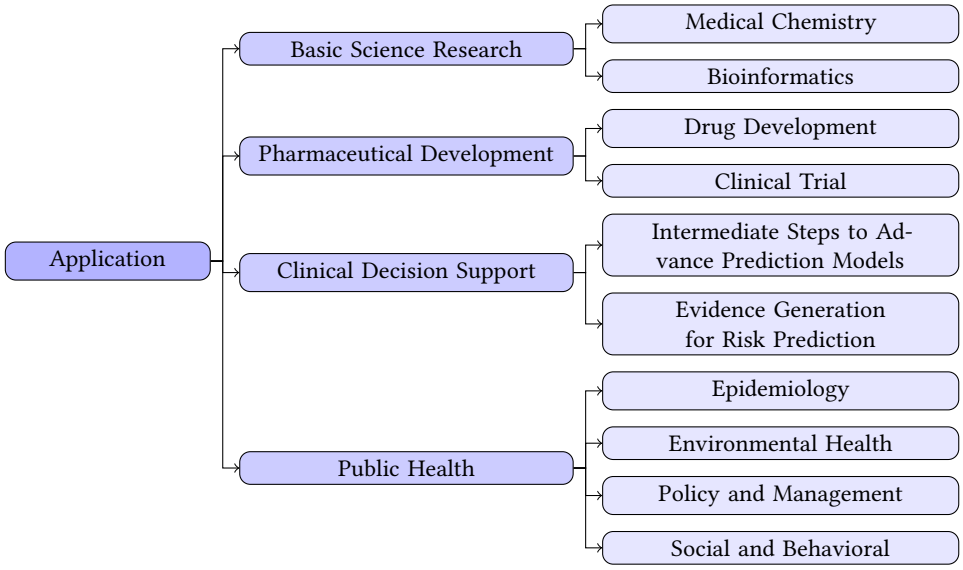

Fig. 4. Detailed taxonomy of the application section of healthcare knowledge graphs.

Drug-drug interactions (DDIs) refer to changes in the actions, or side effects, of drugs when they are taken at the same time or successively [54]. In general, DDIs are a significant contributor to life-threatening adverse events [143], and their identification is one of the key tasks in public health and drug development. The existence of diverse datasets on drug-drug interactions (DDIs) and biomedical KGs has enabled the development of machine-learning models that can accurately predict DDIs. Yu et al. [183] develop SumGNN, a model that includes a subgraph extraction module to efficiently extract relevant subgraphs from a KG, a self-attention-based summarization scheme to generate reasoning paths within the subgraph, and a multichannel module for integrating knowledge and data, resulting in significantly improved predictions of multi-typed DDIs. Su et al. [143] propose DDKG, an attention-based KG representation learning framework that involves an encoder-decoder layer to learn the initial embeddings of drug nodes from their attributes in the KG. Karim et al. [78] compare various techniques for generating KG embeddings with different settings and conclude that a combined convolutional neural network and LSTM network yields the highest accuracy when predicting drug-drug interactions (DDIs). Dai et al. [22] propose a new KG embedding framework by introducing adversarial autoencoders based on Wasserstein distances and Gumbel-Softmax relaxation for DDI tasks. Lin et al. [89] develop KGNN that resolves the DDI prediction by capturing drugs and their potential neighborhoods by mining associated relations in KG.

Drug-target interactions (DTIs) is just as important as DDIs [19]. Machine learning models can leverage knowledge graphs constructed from various types of interactions, such as drug-drug, drug-disease, protein-disease, and protein-protein interactions, to aid in predicting DTIs. For instance, Li et al. [84] utilize the KG transfer probability matrix to redefine the drug-drug and target-target similarity matrix, thus constructing the final graph adjacent matrix to learn node representations by VGAE and augmenting them by utilizing dual Wasserstein Generative Adversarial Network with gradient penalty. Zhang et al. [193] propose a new hybrid method for DTI prediction by first constructing DTI-related KGs and then employing graph representation learning model to obtain feature vectors of the KG. Wang et al. [156] construct a knowledge graph of 29,607 positive drug-target pairs by DistMult embedding strategy, and propose

a Conv-Conv module to extract features of drug-target pairs. Ye et al. [179] learn a low-dimensional representation for various entities in the KG, and then integrate the multimodal information via neural factorization machine.

4.1.2 Bioinformatics Research. In bioinformatics settings, a knowledge graph is a graph-based representation where nodes are biomedical entities (such as mutations, genes, proteins, metabolites, diseases, and biological pathways), and edges are their relationships (such as associations, interactions, and regulations). Through the integration, researchers can gain a more comprehensive understanding of complex biological processes and diseases.

Multi-Omics Applications: In recent years, the field of multi-omics data analysis has become increasingly important for understanding complex biological systems. With the advancement of high-throughput technologies, researchers can generate large-scale and high-dimensional data from different omics fields, such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics. More KG applications based on multi-omics data integration have emerged, aiming to provide new research methods to uncover the complex relationships between different omics layers and reveal biological systems' underlying mechanisms.

Knowledge graphs have been used to identify disease-associated mutations, genes, proteins, and metabolites by integrating multi-omics data with existing biological knowledge. This approach has led to the discovery of novel biomarkers and therapeutic targets for various diseases and the interpretation of the functional effects of genetic elements [189]. Quan et al. built a comprehensive multi-relational knowledge graph called AIMedGraph, providing an interpretation of the impact of genetic variants on disease or treatment [122]. They curated detailed information about diseases, drugs, genetic variants, and the impact of genetic variations on disease development and drug treatment from multiple data resources. The entities integrated into AIMedGraph are connected by evidence-based relations and form a comprehensive gene–variant–disease–drug–trial–reference knowledge network. AIMedGraph uses the Adamic-Adar algorithm to predict new relations between entities based on shared neighbors and their proximity in the knowledge network. GenomicsKG is a knowledge graph to analyze and visualize multi-omics data. GenomicsKG can be used to improve drug development based on clinical genomics correlations and personalized drug customization in the extended version based on interactive relationships. It also provides multi-dimensional visualization, linked functional knowledge graphs, and reporting for clinical genomics.

Single-Cell Analysis: Cells are fundamental and essential units of living organisms. With high-throughput sequencing technologies advancing to measure genomic profiles in a single-cell resolution, cell functions (inside cells) and cell-cell interactions (between cells) are revealed [91]. When diving into the functions of cells, gene regulatory mechanisms can be a critical factor. Gene regulatory mechanisms control the expression of genes, which can affect cell differentiation, response to stimuli, disease progression, etc. It can be visualized as gene regulatory networks (GRNs) which depict the interactions between genes and their regulators. Traditional approaches to constructing the GRN are based on gene knockdown or knockout experiments, which can be time-consuming, labor-intensive, and costly. In addition, these experiments are limited to only one or a few genes which neglects gene-gene interactions. In contrast, single-cell sequencing data could provide whole-genome scale measurements in each individual cell with comprehensive information such as gene expression, transcription factors (TF) binding sites, DNA methylation, epigenetic modifications, etc. By mining these publicly available data, it is possible to reveal the underlying and universal GRN biomedical researchers to understand biological processes better. For example, GRNdb provides detailed regulon and TF-target pairs information from different human and mouse tissues under different conditions by analyzing existing sequencing data [40]. GenomicKB integrates existing datasets and genome annotations and formulates the data into a knowledge graph to emphasize the relationships among genomic entries [44]. We can foresee that with more data being generated

and collected, one generalized GRN (gene regulatory knowledge graph) along with cell-type-specific GRNs can be reconstructed and used to help reveal the underlying mechanisms of gene expression, biological processes, and disease progression.

Besides cell-type-specific GRN, single-cell sequencing data can also be used to infer cell-cell interactions, which play a critical role in understanding cell cycles, cell fate decisions, tissue development, immune response, etc. Among several approaches through which cells can interact, cell-cell communication or cell signaling is of the most interest [37] as the physical distance between cells does not limit it. Cells can communicate with each other by sending signaling molecules called ligands and receiving through receptors located on cell surfaces. Several methods have been developed to implicitly measure cell-cell communication by modeling and scoring ligand-receptor interactions (LRIs) from single-cell sequencing data [36, 74]. However, these methods do not utilize the spatial information of cells. Recent advancements in spatial sequencing techniques provide not only genomic measurements but also complementary information on the spatial coordinates of each cell, where the spatial coordinates indicate the probabilities of cells communicating with each other as cells with closer distances tend to interact more. A recent benchmark study on evaluating cell-cell communication inference methods based on spatial distances [96] while another benchmark study also considers other biological factors such as cytokine activities and receptor protein abundance [29]. With the availability of spatial transcriptomics data, many methods have been developed to leverage colocalization information to infer LRIs. Although many LRIs databases [36, 135] have been constructed and applied to infer cell-cell communication, only recently, SpaTalk [134] integrates CellTalkDB, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Reactome and TFs from AnimalTFDB to construct a ligand-receptor-target knowledge graph to help improve the inference of cell-cell communication. The success of SpaTalk provides an unprecedented view of integrating existing networks into a knowledge graph and shows that correctly utilizing the knowledge graph could provide helpful information in cell-cell communication tasks.

4.2 Pharmaceutical Research Development

4.2.1 Drug Development. Drug development is the process of identifying novel chemical compounds that can effectively treat or alleviate human diseases. Before the drug can be designated as a final product for clinical use, several critical steps need to be undertaken from the initial target identification, chemical synthesis and clinical trials. The whole process typically spans over a decade and involves expenses of approximately one billion dollars [1], yet it is characterized by a low success rate for clinical approval [28].

With advancements in chemistry and the availability of vast chemical libraries, it is possible to reduce the costs of drug development by extracting and integrating valuable insights from existing data and using computer algorithms to accelerate the drug design process [33, 116, 117, 119, 155, 157, 188]. Despite the growing trend of computer-assisted drug discovery, a key question remains regarding how to effectively integrate data and extract valuable insights from the vast chemical dataset.

To approach this question, knowledge graphs (KGs) have been employed for drug discovery due to their various advantages [188]: (1) In contrast to traditional methods that capture only one type of relationship, KGs are capable of providing heterogeneous information that includes diverse entities (*e.g.*, scaffolds, proteins, and genes); (2) In addition, KGs are capable of handling multiple types of relationships between various types of entities, such as drug-target pairs; and (3) KGs can provide unstructured semantic relationships between entities. In such graphs, entities are represented as nodes while their relationships are represented as edges, by which complex relations in biochemical systems can be easily handled.

In general, the field of drug development encompasses two main areas: *drug design* and *drug repurposing* (also known as drug repositioning). Drug design is the process of creating novel and diverse drug molecules with desirable pharmaceutical properties [46, 48, 75], whereas drug repurposing identifies new uses for existing approved drugs that were originally developed for a different indication [69, 120].

Drug Design. KGs are widely employed in drug design, particularly in generating novel molecules that are promising drug candidates for various diseases [83, 124]. Ranjan et al. [124] utilize Gated Graph Neural Network (GGNN) to generate novel molecules that target the coronavirus (i.e., SARS-CoV-2) [59] and integrate KGs into their approach to reduce the search space. Specifically, KGs were leveraged to discard non-binding molecules before inputting them into the Early Fusion model, thus optimizing the efficiency of the drug design process. In addition to employing deep learning for direct structure design, KGs are also utilized in the analysis of chemical synthesis. Quantitative estimation of molecular synthetic accessibility is critical in prioritizing the molecules generated from generative models. For instance, Li et al. [83] utilize reaction KGs to construct classification models for compound synthetic accessibility. By leveraging KGs that capture information about reactions, including reaction types, substrates, and reaction conditions, they can train machine learning models that could predict the synthetic accessibility of compounds. Jeong et al. [72] introduce an intelligent system that integrates generative exploration and exploitation of reaction knowledge base to support synthetic path design.

Drug Repurposing. Compared to drug design, KGs are more commonly utilized to expedite the drug re-purposing process [51, 53, 64, 104, 170, 192, 194, 201]. Many applications on drug re-purposing that utilize KGs are primarily focused on link prediction tasks [104]. To re-purpose promising drug candidates for new indications, many methods employ predictive models that focus on predicting drug-treats-disease relationships within pharmacological knowledge graphs KGs. Himmelstein et al. [64] use a degree-normalized pathway model on the hetionet KG, which includes genes, diseases, tissues, pathophysiologies, and multimodal edges, to identify potentially repurposable drugs for epilepsy. Xu et al. [170] develop a multi-path random walk model on a network that incorporates gene-phenotype associations, protein-protein interactions, and phenotypic similarities for training and prediction purposes. Zhang et al. [192] introduce an integrative and literature-based discovery model for identifying potential drug candidates from COVID-19-focused research literature, including PubMed and other relevant sources. Gao et al. [51] construct a knowledge graph (KG) by integrating multiple genotypic and phenotypic databases. They then learn low-dimensional representations of the KG and utilize these representations to infer new drug-disease interactions, providing insights into potential drug repurposing opportunities. Zhang and Che [194] introduce a model for drug re-purposing in Parkinson's disease that leverages a local medical knowledge base incorporating accurate knowledge along with medical literature containing novel information. Ghorbanali et al. [53] present the DrugRep-KG method, which utilizes a KG embedding approach for representing drugs and diseases in the process of drug repurposing.

4.2.2 Clinical Trial. The major goal of clinical trials is to assess the safety and effectiveness of drug molecules on human bodies. A novel drug molecule needs to pass three phases of clinical trials before it is approved by the Food and Drug Administration (FDA) and enters the drug market. The whole process is prohibitively time-consuming and expensive, costing 7-11 years and two billion dollars on average [105].

Clinical Trial Optimization targets identifying eligible patients for clinical trials based on their medical history and health conditions [62, 126]. Recently, with massive electronic health records (EHR) data and trial eligibility criteria (EC), data-driven methods have been studied to automatically assign appropriate patients for clinical trials [95, 148, 185]. However, it is often hard to fully capture and represent the complex knowledge present in unstructured ECs and EHR

data, as ECs may only provide general disease concepts. In contrast, patient EHR data contain more specific medical codes to represent patient conditions. To better capture the interactions among different medical concepts from EHR records and ECs, Gao et al. [50] enhance patient records with hierarchical taxonomies to align medical concepts of varying granularity between EHR codes and ECs. Besides, Fu et al. [47] leverage additional knowledge-embedding modules along with drug pharmacokinetic and historical trial data to improve the patient trial optimization process, and Wang et al. [161] leverage the knowledge graphs to learn static trial embedding and further designed meta-learning module to generalize well over the imbalanced clinical trial distribution.

4.3 Clinical Decision Support

Nowadays, abundant Electronic Health Record (EHR) data enables better computational models for accurate diagnoses and treatments. EHR contains essential patient information such as disease diagnoses, prescribed medications, and test results. Due to this valuable information, EHRs are extensively utilized to identify patterns in patient health and assist healthcare providers in making informed clinical decisions.

However, the sparsity of EHR data typically allows for only a small fraction of medical codes to be learned effectively, thereby restricting the ability of deep learning approaches. To overcome this drawback, knowledge graphs have been applied to incorporate prior medical knowledge for these deep learning models, which augment the representation of medical codes to better support the downstream prediction tasks.

4.3.1 Intermediate Steps to Advance Prediction Models. **ICD Coding** aims to extract diagnosis and procedure codes from clinical notes, which often consisted of raw text [31, 113, 152, 197]. It is often challenging, as the size of the candidate target codes can be large and the distribution of the codes is often long-tailed [79]. To overcome this, Xie et al. [167] and Cao et al. [15] propose to leverage knowledge graphs as *distant supervision* [87, 109], and inject the label information via structured *knowledge graph propagation* by leveraging graph convolution networks [81] to learn the correlations among medical codes. Besides, Lu et al. [97] propose to leverage knowledge graphs and the co-occurrence graph among clinical nodes simultaneously with a knowledge aggregation module to boost the ICD coding performance. Ren et al. [125] design a learning curriculum based on the hierarchical structure of the code to address the highly imbalanced label distribution issue and balance between frequent and rare labels. Overall, injecting additional knowledge with graph neural networks offers a way to mitigate the imbalanced label distribution issue and thus better.

Entity and Relation Extraction from Health Records. Health records contain rich unstructured or semi-structured data, making it difficult for clinicians to access and analyze relevant information. Entity and relation extraction helps convert this unstructured text into structured data that can be more easily processed, understood, and utilized. Specifically, *entity extraction* aims to identify entity mentions from clinical-free texts. There are two key steps for entity extraction, i.e., named entity recognition (NER) and disambiguation (NED). By leveraging additional knowledge graphs, Varma et al. [150] transfer structural knowledge from the knowledge base to the medical domain, improving rare entities' disambiguation accuracy. Yuan et al. [186] inject additional knowledge from the knowledge graphs for entity linking and proposed two additional strategies, namely Post-pruning and Thresholding, to improve the efficiency and remove the effect of unlinkable entity mention. Fries et al. [45] leverage clinical ontologies to provide *weak supervision* sources to create additional training data for clinical entity disambiguation. Besides, *relation extraction* aims to identify and classify relationships between entities in unstructured text, which facilitates understanding complex biological processes, drug interactions, and disease mechanisms. To incorporate the external knowledge graph, several works [43, 127] proposed additional post-training steps to align the language models with biomedical knowledge. Hong et al. [66] construct

embeddings for a wide range of codified concepts from EHRs to identify relevant features related to a disease of interest, and Lin et al. [90] design a co-training scheme to jointly learn from text and knowledge graphs for extracting and classifying disease-disease relations. In summary, fusing knowledge graphs with language models can flexibly accommodate missing data types and bring additional performance gains, especially for those rare entities and relations.

4.3.2 Evidence Generation for Risk Prediction Models. **Disease Prediction** aims to predict the potential diseases of a given patient with his past clinical records. To assist the diagnosis with additional knowledge, GRAM [21] and KAME [101] utilize a medical ontology [34] where the leaf nodes are the medical codes found in EHR data, and their ancestors are more general categories. By incorporating information from medical ontologies into deep learning models via neural attention, these approaches learn better embeddings for different medical concepts to alleviate the data scarcity bottleneck. [180, 195] further consider the domain-specific knowledge graph KnowLife [39] to enrich the embeddings of medical entities with their neighbors on the knowledge graph. These approaches mainly directly update the embeddings of different concepts to improve the feature learning, but may be at the risk of ignoring the high-level order information from the knowledge graph. To tackle this drawback, Ye et al. [178] explicitly exploit *paths* in KG from the observed symptoms to the target disease to model the personalized information for diverse patients with a relational-guided attention mechanism. Xu et al. [173] design a self-supervised learning approach to pre-train a graph attention network for learning the embedding of medical concepts and completing the knowledge graph simultaneously. These approaches better harness the structure information, and often lead to better performance than the pure embedding-based knowledge integration techniques.

Treatment Recommendation aims to recommend personalized medications to patients based on their individual health conditions, which can help physicians select the most effective medications for their patients, and improve treatment outcomes [7, 133, 196]. To effectively exploit external knowledge, Shang et al. [132] use drug ontologies to design additional pretraining loss and directly improve the representation of drugs, and several studies [146, 165] attempt to extract the additional drug interaction graphs to model the negative side effects of specific drug pairs and reduce the possibility of recommending negative drug-drug interaction combinations. Besides, Wu et al. [166] leveraged ontologies to improve the drug representations, and facilitates drug recommendation under a more challenging few-shot setting.

4.4 Public Health

Public Health research can significantly benefit from HKGs. Knowledge graphs can help organize, structure, and formalize extensive information from diverse and heterogeneous sources. This allows researchers to analyze data, reason about factors, and make decisions on a larger scale.

Epidemiology. The field of epidemiology has seen an increase in the use of knowledge graphs to analyze and understand the spread of diseases. A study conducted by Gao et al. [52] analyzes the research hotspots and development trends of wastewater-based epidemiology (WBE) using knowledge graphs constructed from nearly 900 papers. Domingo-Fernández et al. [30] create the COVID-19 Knowledge Graph, a comprehensive cause-and-effect network constructed from the scientific literature on the coronavirus. Additionally, Turki et al. [149] use knowledge graphs to assess and validate the portion of Wikidata related to COVID-19 epidemiology using an automatable task set. Pressat Laffouilhère et al. [121] develop OntoBioStat, a domain ontology related to covariate selection and bias in biostatistics, which can help interpret significant statistical associations between variables.

Environmental Health. Fecho et al. [42] develop ROBOKOP, a biomedical knowledge graph-based system, to validate associations between workplace chemical exposures and immune-mediated diseases. Wolffe et al. [163] propose using knowledge graphs in systematic evidence mapping in environmental health. This approach overcomes the limitations of rigid data tables by offering a more suitable model for handling the highly connected and complex nature of environmental health data.

Health Policy and Management. Wu et al. [164] have analyzed the COVID-19 epidemic situation using a knowledge graph of patient activity. This method enables in-depth study of the transmission process, analysis of key nodes, and tracing of activity tracks. Meanwhile, Yu et al. [182] develop a chronic management system, which combines knowledge graphs and big data to optimize the management of chronic diseases in children. This system enhances treatment and resource utilization while conforming to the requirements of the Chronic Care Model.

Social and Behavioral Health. There have been several interesting research studies. Cao et al. [14] build a high-level suicide-oriented knowledge graph combined with deep neural networks for detecting suicidal ideation on social media platforms. Also, Liu et al. [93] conduct a bibliometric analysis of driver behavior research. Additionally, Wang et al. [158] create an analysis framework for interpreting causal associations in emotional logic. They introduce a knowledge graph into appraisal theories, improving human emotional inference.

5 Promise and Outlook

The potential impact of comprehensive and fine-grained HKGs on biomedical research and clinical practice is significant. By integrating vast amounts of biomedical knowledge from multiple domains, HKGs can facilitate the discovery of new disease mechanisms and the identification of novel drug targets. They also help to enable personalized medicine by identifying patient subgroups with shared disease mechanisms. The recent success of large language models (*a.k.a.* foundation models) such as ChatGPT offers promising opportunities in capturing such semantics from the biomedical context [2, 110, 114, 140], enabling the construction of unprecedentedly comprehensive HKGs. In turn, HKGs also help improve LLMs by providing accurate and contextualized knowledge to regularize the generated content. This is particularly useful in evaluating LLMs in biomedical applications and addressing the problem of hallucination in critical areas.

6 Conclusion

Healthcare knowledge graphs have emerged as a promising approach for capturing and organizing medical knowledge in a structured and interpretable way. This comprehensive review paper provides an overview of the current state of HKGs, including their construction, utilization models, and applications in healthcare. Furthermore, the paper discusses the potential future developments of HKGs. In conclusion, HKGs have played a significant role in advancing health research. With the advent of LLMs, there are even more opportunities to combine HKGs and LLMs to reduce the generation of false or unreliable content. We hope that our comprehensive review of this field offers a helpful perspective for future reference.

References

[1] Christopher P Adams and Van V Brantner. 2006. Estimating the cost of new drug development: is it really $802 million? *Health affairs* 25, 2 (2006), 420–428.

[2] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 1998–2022.

[3] Mona Alshahrani, Mohammad Asif Khan, Omar Maddouri, Akira R Kinjo, Núria Queralt-Rosinach, and Robert Hoehndorf. 2017. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* 33, 17 (2017), 2723–2730.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[5] John A Bachman, Benjamin M Gyori, and Peter K Sorger. 2018. FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. *BMC bioinformatics* 19 (2018), 1–14.

[6] Jonathan Bard, Seung Y Rhee, and Michael Ashburner. 2005. An ontology for cell types. *Genome biology* 6, 2 (2005), 1–5.

[7] Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, and Ngiap Chuan Tan. 2021. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems (TOIS)* 40, 3 (2021), 1–23.

[8] Chris Bizon, Steven Cox, James Balhoff, Yaphet Kebede, Patrick Wang, Kenneth Morton, Karamarie Fecho, and Alexander Tropsha. 2019. ROBOKOP KG and KGB: integrated knowledge graphs from federated sources. *Journal of chemical information and modeling* 59, 12 (2019), 4968–4973.

[9] Kathrin Blagec, Adriano Barbosa-Silva, Simon Ott, and Matthias Samwald. 2022. A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. *Scientific Data* 9, 1 (2022), 322.

[10] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl_1 (2004), D267–D270.

[11] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[12] Lina Bouayad, Anna Ialynytchev, and Balaji Padmanabhan. 2017. Patient health record systems scope and functionalities: literature review and future directions. *Journal of medical Internet research* 19, 11 (2017), e388.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[14] Lei Cao, Huijun Zhang, and Ling Feng. 2020. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia* 24 (2020), 87–102.

[15] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3105–3114.

[16] Christian Castaneda, Kip Nalley, Ciaran Mannion, Pritish Bhattacharyya, Patrick Blake, Andrew Pecora, Andre Goy, and K Stephen Suh. 2015. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics* 5, 1 (2015), 1–16.

[17] Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Nature Scientific Data* (2023). https://doi.org/10.1038/s41597-023-01960-3

[18] Xiang Chen, Lei Li, Qiaoshuo Fei, Ningyu Zhang, Chuanqi Tan, Yong Jiang, Fei Huang, and Huajun Chen. 2023. One Model for All Domains: Collaborative Domain-Prefix Tuning for Cross-Domain NER. *arXiv preprint arXiv:2301.10410* (2023).

[19] Xing Chen, Chenggang Clarence Yan, Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin, and Yongdong Zhang. 2016. Drug–target interaction prediction: databases, web servers and computational models. *Briefings in bioinformatics* 17, 4 (2016), 696–712.

[20] Young Min Cho, Li Zhang, and Chris Callison-Burch. 2022. Unsupervised Entity Linking with Guided Summarization and Multiple-Choice Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 9394–9401.

[21] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 787–795.

[22] Yuanfei Dai, Chenhao Guo, Wenzhong Guo, and Carsten Eickhoff. 2021. Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Briefings in bioinformatics* 22, 4 (2021), bbaa256.

[23] Rajarshi Das, Ameya Godbole, Nicholas Monath, Manzil Zaheer, and Andrew McCallum. 2020. Probabilistic Case-based Reasoning for Open-World Knowledge Graph Completion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4752–4765.

[24] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. 2019. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* 6, 1 (2019), 1–25.

[25] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. [n. d.]. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics* 10 ([n. d.]), 274–290.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[27] Alexander D Diehl, Terrence F Meehan, Yvonne M Bradford, Matthew H Brush, Wasila M Dahdul, David S Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, et al. 2016. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics* 7 (2016), 1–10.

[28] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. 2015. The cost of drug development. *New England Journal of Medicine* 372, 20 (2015), 1972–1972.

[29] Daniel Dimitrov, Dénes Türei, Martin Garrido-Rodriguez, Paul L Burmedi, James S Nagai, Charlotte Boys, Ricardo O Ramirez Flores, Hyojin Kim, Bence Szalai, Ivan G Costa, et al. 2022. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications* 13, 1 (2022), 3224.

[30] Daniel Domingo-Fernández, Shounak Baksi, Bruce Schultz, Yojana Gadiya, Reagon Karki, Tamara Raschka, Christian Ebeling, Martin Hofmann-Apitius, and Alpha Tom Kodamullil. 2021. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* 37, 9 (2021), 1332–1334.

[31] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine* 5, 1 (2022), 159.

[32] Kevin Donnelly et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.

[33] Yuanqi Du, Shiyu Wang, Xiaojie Guo, Hengning Cao, Shujie Hu, Junji Jiang, Aishwarya Varala, Abhinav Angirekula, and Liang Zhao. 2021. Graphgt: Machine learning datasets for graph generation and transformation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[34] Erik R Dubberke, Kimberly A Reske, L Clifford McDonald, and Victoria J Fraser. 2006. ICD-9 codes and surveillance for Clostridium difficile–associated disease. *Emerging infectious diseases* 12, 10 (2006), 1576.

[35] John Eberhardt, Anton Bilchik, and Alexander Stojadinovic. 2012. Clinical decision support systems: potential with pitfalls. *Journal of Surgical Oncology* 105, 5 (2012), 502–510.

[36] Mirjana Efremova, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo. 2020. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols* 15, 4 (2020), 1484–1506.

[37] Holger K Eltzschig, Thomas Weissmüller, Alice Mager, and Tobias Eckle. 2006. Nucleotide metabolism and cell-cell interactions. *Cell-Cell Interactions: Methods and Protocols* (2006), 73–87.

[38] Patrick Ernst, Cynthia Meng, Amy Siu, and Gerhard Weikum. 2014. Knowlife: a knowledge graph for health and life sciences. In *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 1254–1257.

[39] Patrick Ernst, Amy Siu, and Gerhard Weikum. 2015. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC bioinformatics* 16 (2015), 1–13.

[40] Li Fang, Yunjin Li, Lu Ma, Qiyue Xu, Fei Tan, and Geng Chen. 2021. GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic acids research* 49, D1 (2021), D97–D103.

[41] Daniel Faria, Ernesto Jiménez-Ruiz, Catia Pesquita, Emanuel Santos, and Francisco M Couto. 2014. Towards annotating potential incoherences in bioportal mappings. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13*. Springer, 17–32.

[42] Karamarie Fecho, Chris Bizon, Frederick Miller, Shepherd Schurman, Charles Schmitt, William Xue, Kenneth Morton, Patrick Wang, Alexander Tropsha, et al. 2021. A biomedical knowledge graph system to propose mechanistic hypotheses for real-world environmental health observations: cohort study and informatics application. *JMIR Medical Informatics* 9, 7 (2021), e26714.

[43] Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics* 22, 3 (2021), bbaa110.

[44] Fan Feng, Feitong Tang, Yijia Gao, Dongyu Zhu, Tianjun Li, Shuyuan Yang, Yuan Yao, Yuanhao Huang, and Jie Liu. 2023. GenomicKB: a knowledge graph for the human genome. *Nucleic Acids Research* 51, D1 (2023), D950–D956.

[45] Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications* 12, 1 (2021), 2017.

[46] Tianfan Fu, Wenhao Gao, Cao Xiao, Jacob Yasonik, Connor W Coley, and Jimeng Sun. 2021. Differentiable scaffolding tree for molecular optimization. *arXiv preprint arXiv:2109.10469* (2021).

[47] Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2022. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns* 3, 4 (2022), 100445.

[48] Tianfan Fu and Jimeng Sun. 2022. Antibody complementarity determining regions (cdrs) design using constrained energy model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 389–399.

[49] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*. 413–422.

[50] Junyi Gao, Cao Xiao, Lucas M Glass, and Jimeng Sun. 2020. COMPOSE: cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 803–812.

[51] Zhenxiang Gao, Pingjian Ding, and Rong Xu. 2022. Kg-predict: a knowledge graph computational framework for drug repurposing. *Journal of biomedical informatics* 132 (2022), 104133.

[52] Zhihan Gao, Min Gao, Chun-hua Chen, Yifan Zhou, Zhi-Hui Zhan, and Yuan Ren. 2023. Knowledge graph of wastewater-based epidemiology development: A data-driven analysis based on research topics and trends. *Environmental Science and Pollution Research* 30, 11 (2023), 28373–28382.

[53] Zahra Ghorbanali, Fatemeh Zare-Mirakabad, Mohammad Akbari, Najmeh Salehi, and Ali Masoudi-Nejad. 2023. DrugRep-KG: Toward Learning a Unified Latent Space for Drug Repurposing Using Knowledge Graphs. *Journal of Chemical Information and Modeling* (2023).

[54] Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. 2007. When good drugs go bad. *Nature* 446, 7139 (2007), 975–977.

[55] Nicola Guarino, Daniel Oberle, and Steffen Staab. 2009. What is an ontology? *Handbook on ontologies* (2009), 1–17.

[56] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

[57] Amelie Gyrard, Manas Gaur, Saeedeh Shekarpour, Krishnaprasad Thirunarayan, and Amit Sheth. 2018. Personalized health knowledge graph. In *CEUR workshop proceedings*, Vol. 2317.

[58] Udo Hahn and Michel Oleynik. 2020. Medical information extraction in the age of deep learning. *Yearbook of medical informatics* 29, 01 (2020), 208–220.

[59] Mustafa Hasöksüz, Selcuk Kilic, and Fahriye Sarac. 2020. Coronaviruses and sars-cov-2. *Turkish journal of medical sciences* 50, 9 (2020), 549–556.

[60] Keywan Hassani-Pak and Christopher Rawlings. 2017. Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. *Journal of integrative bioinformatics* 14, 1 (2017).

[61] Yuan He, Jiaoyan Chen, Hang Dong, Ernesto Jiménez-Ruiz, Ali Hadian, and Ian Horrocks. 2022. Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching. In *The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*. Springer, 575–591.

[62] Zhe He, Xiang Tang, Xi Yang, Yi Guo, Thomas J George, Neil Charness, Kelsa Bartley Quan Hem, William Hogan, and Jiang Bian. 2020. Clinical trial generalizability assessment in the big data era: a review. *Clinical and translational science* 13, 4 (2020), 675–684.

[63] Daniel S Himmelstein and Sergio E Baranzini. 2015. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS computational biology* 11, 7 (2015), e1004259.

[64] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 6 (2017), e26726.

[65] JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. 2016. ICD-10: history and context. *American Journal of Neuroradiology* 37, 4 (2016), 596–599.

[66] Chuan Hong, Everett Rush, Molei Liu, Doudou Zhou, Jiehuan Sun, Aaron Sonabend, Victor M Castro, Petra Schubert, Vidul A Panickan, Tianrun Cai, et al. 2021. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ digital medicine* 4, 1 (2021), 151.

[67] Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot Clinical Entity Recognition using ChatGPT. *arXiv preprint arXiv:2303.16416* (2023).

[68] Yu Hu, Tiezheng Nie, Derong Shen, Yue Kou, and Ge Yu. 2021. An integrated pipeline model for biomedical entity alignment. *Frontiers of Computer Science* 15 (2021), 1–15.

[69] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. 2020. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* 36, 22-23 (2020), 5545–5547.

[70] Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International conference on computational linguistics*. 2515–2527.

[71] Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis. 2020. DRKG - Drug Repurposing Knowledge Graph for Covid-19. https://github.com/gnn4dr/DRKG/.

[72] Joonsoo Jeong, Nagyeong Lee, Yongbeom Shin, and Dongil Shin. 2022. 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* 130 (2022), 103982.

[73] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems* 33, 2 (2021), 494–514.

[74] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. 2021. Inference and analysis of cell-cell communication using CellChat. *Nature communications* 12, 1 (2021), 1088.

[75] Yankang Jing, Yuemin Bian, Ziheng Hu, Lirong Wang, and Xiang-Qun Sean Xie. 2018. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *The AAPS journal* 20 (2018), 1–10.

[76] Xuan Kan, Hejie Cui, and Carl Yang. 2021. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. Springer, 466–482.

[77] Minoru Kanehisa and Susumu Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28, 1 (2000), 27–30.

[78] Md Rezaul Karim, Michael Cochez, Joao Bosco Jares, Mamtaz Uddin, Oya Beyan, and Stefan Decker. 2019. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*. 113–123.

[79] Byung-Hak Kim and Varun Ganapathi. 2021. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In *Proceedings of the 6th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 149)*. 196–208.

[80] Daeyoung Kim, Seongsu Bae, Seungho Kim, and Edward Choi. 2022. Uncertainty-Aware Text-to-Program for Question Answering on Structured Electronic Health Records. In *Proceedings of the Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 174)*, Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (Eds.). PMLR, 138–151.

[81] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *The International Conference on Learning Representations (ICLR)* (2016).

[82] Stanley Kok and Pedro Domingos. 2005. Learning the structure of Markov logic networks. In *Proceedings of the 22nd international conference on Machine learning*. 441–448.

[83] Baiqing Li and Hongming Chen. 2022. Prediction of compound synthesis accessibility based on reaction knowledge graph. *Molecules* 27, 3 (2022), 1039.

[84] Guodong Li, Weicheng Sun, Jinsheng Xu, Lun Hu, Weihan Zhang, and Ping Zhang. 2023. GA-ENs: A novel drug–target interactions prediction method by incorporating prior Knowledge Graph into dual Wasserstein Generative Adversarial Network with gradient penalty. *Applied Soft Computing* 139 (2023), 110151.

[85] Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, et al. 2020. Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine* 103 (2020), 101817.

[86] Qian Li, Daling Wang, Shi Feng Kaisong Song, Yifei Zhang, and Ge Yu. 2022. OERL: Enhanced Representation Learning Via Open Knowledge Graphs. *IEEE Transactions on Knowledge and Data Engineering* (2022), 1–14. https://doi.org/10.1109/TKDE.2022.3218850

[87] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1054–1064.

[88] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2829–2839.

[89] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. 2020. KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction.. In *IJCAI*, Vol. 380. 2739–2745.

[90] Yucong Lin, Keming Lu, Sheng Yu, Tianxi Cai, and Marinka Zitnik. 2022. Multimodal Learning on Graphs for Disease Relation Extraction. *arXiv preprint arXiv:2203.08893* (2022).

[91] Sten Linnarsson and Sarah A Teichmann. 2016. Single-cell genomics: coming of age. , 3 pages.

[92] Andy T Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. QaNER: Prompting question answering models for few-shot named entity recognition. *arXiv preprint arXiv:2203.01543* (2022).

[93] Hui Liu, Yifan Li, Rui Hong, Zhenming Li, Ming Li, Wei Pan, Adam Glowacz, and Hao He. 2020. Knowledge graph analysis and visualization of research trends on driver behavior. *Journal of Intelligent & Fuzzy Systems* 38, 1 (2020), 495–511.

[94] Qi Liu, Dani Yogatama, and Phil Blunsom. 2022. Relational Memory-Augmented Language Models. *Transactions of the Association for Computational Linguistics* 10 (2022), 555–572.

[95] Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping, et al. 2021. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* 592, 7855 (2021), 629–633.

[96] Zhaoyang Liu, Dongqing Sun, and Chenfei Wang. 2022. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biology* 23, 1 (2022), 1–38.

[97] Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. 2020. Multi-label Few/Zero-shot Learning with Knowledge Aggregated from Multiple Label Graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2935–2943.

[98] Jiaying Lu, Jiaming Shen, Bo Xiong, Wengjing Ma, Staab Steffen, and Carl Yang. 2023. HiPrompt: Few-Shot Biomedical Knowledge Fusion via Hierarchy-Oriented Prompting. In *Proceedings of The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval - Short Paper (SIGIR 2023)*.

[99] Jiaying Lu and Carl Yang. 2022. Open-World Taxonomy and Knowledge Graph Co-Learning. In *4th Conference on Automated Knowledge Base Construction (AKBC 2022)*.

[100] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5755–5772.

[101] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 743–752.

[102] Jiangtao Ma, Chenyu Zhou, Yonggang Chen, Yanjun Wang, Guangwu Hu, and Yaqiong Qiao. 2023. TeCre: A Novel Temporal Conflict Resolution Method Based on Temporal Knowledge Graph Embedding. *Information* 14, 3 (2023), 155.

[103] Meriem Maaroufi, Rémy Choquet, Paul Landais, and Marie-Christine Jaulent. 2014. Formalizing mappings to optimize automated schema alignment: application to rare diseases. In *e-Health–For Continuity of Care*. IOS Press, 283–287.

[104] Finlay MacLean. 2021. Knowledge graphs and their applications in drug discovery. *Expert opinion on drug discovery* 16, 9 (2021), 1057–1069.

[105] Linda Martin, Melissa Hutchens, Conrad Hawkins, and Alaina Radnov. 2017. How much do clinical trials cost. *Nat Rev Drug Discov* 16, 6 (2017), 381–382.

[106] Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. 2022. DEAP-FAKED: Knowledge graph based approach for fake news detection. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 47–51.

[107] Nishita Mehta and Anil Pandit. 2018. Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics* 114 (2018), 57–65.

[108] Paolo Mengoni and Jinyu Yang. 2022. Empowering COVID-19 Fact-Checking with Extended Knowledge Graphs. In *Computational Science and Its Applications–ICCSA 2022 Workshops: Malaga, Spain, July 4–7, 2022, Proceedings, Part I*. Springer, 138–150.

[109] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 777–782.

[110] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616, 7956 (2023), 259–265.

[111] Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi, and Hamed Firooz. 2022. Detection, Disambiguation, Re-ranking: Autoregressive Entity Linking as a Multi-Task Problem. In *Findings of the Association for Computational Linguistics: ACL 2022*. 1972–1983.

[112] Stephen Muggleton. 1992. *Inductive logic programming*. Number 38. Morgan Kaufmann.

[113] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1101–1111.

[114] Siddharth Nath, Abdullah Marie, Simon Ellershaw, Edward Korot, and Pearse A Keane. 2022. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *British Journal of Ophthalmology* 106, 7 (2022), 889–892.

[115] David N Nicholson and Casey S Greene. 2020. Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal* 18 (2020), 1414–1428.

[116] Christos A Nicolaou, Nathan Brown, and Constantinos S Pattichis. 2007. Molecular optimization using computational multi-objective methods. *Current Opinion in Drug Discovery and Development* 10, 3 (2007), 316.

[117] Christos A Nicolaou and Constantinos S Pattichis. 2006. Molecular substructure mining approaches for computer-aided drug discovery: A review. *Proc. of ITAB* (2006), 26–28.

[118] Lei Niu, Chenpeng Fu, Qiang Yang, Zhixu Li, Zhigang Chen, Qingsheng Liu, and Kai Zheng. 2021. Open-world knowledge graph completion with multiple interaction attention. *World Wide Web* 24 (2021), 419–439.

[119] Bo Pan, Yinkai Wang, Xuanyang Lin, Muran Qin, Yuanqi Du, Shiva Ghaemi, Aowei Ding, Shiyu Wang, Saleh Alkhalifa, Kevin Minbiole, et al. 2022. Property-Controllable Generation of Quaternary Ammonium Compounds. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 3462–3469.

[120] Xiaoqin Pan, Xuan Lin, Dongsheng Cao, Xiangxiang Zeng, Philip S Yu, Lifang He, Ruth Nussinov, and Feixiong Cheng. 2022. Deep learning for drug repurposing: Methods, databases, and applications. *Wiley interdisciplinary reviews: Computational molecular science* 12, 4 (2022), e1597.

[121] Thibaut Pressat Laffouilhère, Julien Grosjean, Jean Pinson, Stéfan J Darmoni, Emilie Leveque, Emilie Lanoy, Jacques Bénichou, and Lina F Soualmia. 2022. Ontological Representation of Causal Relations for a Deep Understanding of Associations Between Variables in Epidemiology. In *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings*. Springer, 47–56.

[122] Xueping Quan, Weijing Cai, Chenghang Xi, Chunxiao Wang, and Linghua Yan. [n. d.]. AIMedGraph: A Comprehensive Multi-Relational Knowledge Graph for Precision Medicine. 2023 ([n. d.]), baad006.

[123] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine* 1, 1 (2018), 18.

[124] Amit Ranjan, Shivansh Shukla, Deepanjan Datta, and Rajiv Misra. 2022. 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* 11 (2022), 1–11.

[125] Weiming Ren, Ruijing Zeng, Tongzi Wu, Tianshu Zhu, and Rahul G. Krishnan. 2022. HiCu: Leveraging Hierarchy for Curriculum Learning in Automated ICD Coding. In *Proceedings of the 7th Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 182)*. 198–223.

[126] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, Melanie J Calvert, Hutan Ashrafian, Andrew L Beam, Gary S Collins, Ara Darzi, Jonathan J Deeks, et al. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health* 2, 10 (2020), e549–e560.

[127] Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 5357–5366.

[128] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, et al. 2022. A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology* 40, 5 (2022), 692–702.

[129] Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-Sequence Knowledge Graph Completion and Question Answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2814–2828.

[130] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* 40, D1 (2012), D940–D946.

[131] Oshani Seneviratne, Jonathan Harris, Ching-Hua Chen, and Deborah L McGuinness. 2021. Personal health knowledge graph for clinically relevant diet recommendations. *arXiv preprint arXiv:2110.10131* (2021).

[132] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 5953–5959.

[133] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1126–1133.

[134] Xin Shao, Chengyu Li, Haihong Yang, Xiaoyan Lu, Jie Liao, Jingyang Qian, Kai Wang, Junyun Cheng, Penghui Yang, Huajun Chen, et al. 2022. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk. *Nature Communications* 13, 1 (2022), 4429.

[135] Xin Shao, Jie Liao, Chengyu Li, Xiaoyan Lu, Junyun Cheng, and Xiaohui Fan. 2021. CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. *Briefings in bioinformatics* 22, 4 (2021), bbaa269.

[136] Jianhao Shen, Chenguang Wang, Ye Yuan, Jiawei Han, Heng Ji, Koushik Sen, Ming Zhang, and Dawn Song. 2022. PALT: Parameter-Lite Transfer of Language Models for Knowledge Graph Completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 3833–3847.

[137] Baoxu Shi and Tim Weninger. 2018. Open-world knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[138] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. *arXiv preprint arXiv:2301.12652* (2023).

[139] Sola S Shirai, Oshani Seneviratne, Minor E Gordon, Ching-Hua Chen, and Deborah L McGuinness. 2021. Identifying ingredient substitutions using a knowledge graph of food. *Frontiers in Artificial Intelligence* 3 (2021), 621766.

[140] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138* (2022).

[141] Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. 2021. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics* 22, 6 (2021), bbab282.

[142] Chang Su, Yu Hou, Manqi Zhou, Suraj Rajendran, Jacqueline RMA Maasch, Zehra Abedi, Haotan Zhang, Zilong Bai, Anthony Cuturrufo, Winston Guo, et al. 2023. Biomedical discovery through the integrative biomedical knowledge hub (iBKH). *Iscience* (2023).

[143] Xiaorui Su, Lun Hu, Zhuhong You, Pengwei Hu, and Bowei Zhao. 2022. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Briefings in bioinformatics* 23, 3 (2022), bbac140.

[144] Fabian M Suchanek, Serge Abiteboul, and Pierre Senellart. 2011. PARIS: probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment* 5, 3 (2011), 157–168.

[145] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. 2023. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* 51, D1 (2023), D638–D646.

[146] Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S Hertzberg, and Carl Yang. 2022. 4SDrug: Symptom-based Set-to-set Small and Safe Drug Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3970–3980.

[147] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A knowledge graph of fact-checked claims. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*. Springer, 309–324.

[148] Yitong Tseo, MI Salkola, Ahmed Mohamed, Anuj Kumar, and Freddy Abnousi. 2020. Information extraction of clinical trial eligibility criteria. *arXiv preprint arXiv:2006.07296* (2020).

[149] Houcemeddine Turki, Dariusz Jemielniak, Mohamed A Hadj Taieb, Jose E Labra Gayo, Mohamed Ben Aouicha, Mus'ab Banat, Thomas Shafee, Eric Prud'hommeaux, Tiago Lubiana, Diptanshu Das, et al. 2022. Using logical constraints to validate statistical information about disease outbreaks in collaborative knowledge graphs: the case of COVID-19 epidemiology in Wikidata. *PeerJ Computer Science* 8 (2022), e1085.

[150] Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4566–4575.

[151] Nikhita Vedula and Srinivasan Parthasarathy. 2021. Face-keg: Fact checking explained using knowledge graphs. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 526–534.

[152] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conferences on Artificial Intelligence*. 3335–3341.

[153] Peilu Wang, Hao Jiang, Jingfang Xu, and Qi Zhang. 2019. Knowledge graph construction and applications for Web search and beyond. *Data Intelligence* 1, 4 (2019), 333–349.

[154] Ping Wang, Tian Shi, and Chandan K Reddy. 2020. Text-to-SQL generation for question answering on electronic medical records. In *Proceedings of The Web Conference 2020*. 350–361.

[155] Shiyu Wang, Yuanqi Du, Xiaojie Guo, Bo Pan, Zhaohui Qin, and Liang Zhao. 2022. Controllable data generation by deep learning: A review. *arXiv preprint arXiv:2207.09542* (2022).

[156] Shudong Wang, Zhenzhen Du, Mao Ding, Alfonso Rodriguez-Paton, and Tao Song. 2022. KG-DTI: a knowledge graph based deep learning method for drug-target interaction predictions and Alzheimer's disease drug repositions. *Applied Intelligence* 52, 1 (2022), 846–857.

[157] Shiyu Wang, Xiaojie Guo, Xuanyang Lin, Bo Pan, Yuanqi Du, Yinkai Wang, Yanfang Ye, Ashley Petersen, Austin Leitgeb, Saleh AlKhalifa, et al. 2022. Multi-objective Deep Data Generation with Correlated Property Control. *Advances in Neural Information Processing Systems* 35 (2022), 28889–28901.

[158] Shuo Wang, Yifei Zhang, Bochen Lin, and Boxun Li. 2022. Interpretable emotion analysis based on knowledge graph and OCC model. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2038–2045.

[159] Suyuchen Wang, Ruihui Zhao, Xi Chen, Yefeng Zheng, and Bang Liu. 2021. Enquire one's parent and child before decision: Fully exploit hierarchical structure for self-supervised taxonomy expansion. In *Proceedings of the Web Conference 2021*. 3291–3304.

[160] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 950–958.

[161] Zifeng Wang, Cao Xiao, and Jimeng Sun. 2023. SPOT: Sequential Predictive Modeling of Clinical Trial Outcome with Meta-Learning. *arXiv preprint arXiv:2304.05352* (2023).

[162] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* 46, D1 (2018), D1074–D1082.

[163] Taylor AM Wolffe, John Vidler, Crispin Halsall, Neil Hunt, and Paul Whaley. 2020. A survey of systematic evidence mapping practice and the case for knowledge graphs in environmental health and toxicology. *Toxicological Sciences* 175, 1 (2020), 35–49.

[164] Jiajing Wu. 2021. Construct a knowledge graph for China Coronavirus (COVID-19) patient information tracking. *Risk Management and Healthcare Policy* (2021), 4321–4337.

[165] Jialun Wu, Buyue Qian, Yang Li, Zeyu Gao, Meizhi Ju, Yifan Yang, Yefeng Zheng, Tieliang Gong, Chen Li, and Xianli Zhang. 2022. Leveraging multiple types of domain knowledge for safe and effective drug recommendation. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 2169–2178.

[166] Zhenbang Wu, Huaxiu Yao, Zhe Su, David M Liebovitz, Lucas M Glass, James Zou, Chelsea Finn, and Jimeng Sun. 2022. Knowledge-Driven New Drug Recommendation. *arXiv preprint arXiv:2210.05572* (2022).

[167] Xiancheng Xie, Yun Xiong, Philip S Yu, and Yangyong Zhu. 2019. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 649–658.

[168] Xin Xie, Ningyu Zhang, Zhoubo Li, Shumin Deng, Hui Chen, Feiyu Xiong, Mosha Chen, and Huajun Chen. 2022. From discrimination to generation: knowledge graph completion with generative transformer. In *Companion Proceedings of the Web Conference 2022*. 162–165.

[169] Rui Xing, Jie Luo, and Tengwei Song. 2020. BioRel: towards large-scale biomedical relation extraction. *BMC bioinformatics* 21 (2020), 1–13.

[170] Bo Xu, Yu Liu, Shuo Yu, Lei Wang, Jie Dong, Hongfei Lin, Zhihao Yang, Jian Wang, and Feng Xia. 2019. A network embedding model for pathogenic genes prediction by multi-path random walking on heterogeneous network. *BMC Medical Genomics* 12 (2019), 1–12.

[171] Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F Rousseau, Xin Li, Weijia Xu, Vetle I Torvik, et al. 2020. Building a PubMed knowledge graph. *Scientific data* 7, 1 (2020), 205.

[172] Ran Xu, Yue Yu, Joyce C Ho, and Carl Yang. 2023. Weakly-Supervised Scientific Document Classification via Retrieval-Augmented Multi-Stage Training. In *the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[173] Xiao Xu, Xian Xu, Yuyao Sun, Xiaoshuang Liu, Xiang Li, Guotong Xie, and Fei Wang. 2021. Predictive Modeling of Clinical Events with Mutual Enhancement Between Longitudinal Patient Records and Medical Knowledge Graph. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 777–786.

[174] Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2021. Learning Contextualized Knowledge Structures for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4038–4051.

[175] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *npj Digital Medicine* 5, 1 (2022), 194.

[176] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 535–546.

[177] Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative Knowledge Graph Construction: A Review. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 1–17.

[178] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021*. 1397–1409.

[179] Qing Ye, Chang-Yu Hsieh, Ziyi Yang, Yu Kang, Jiming Chen, Dongsheng Cao, Shibo He, and Tingjun Hou. 2021. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature communications* 12, 1 (2021), 6775.

[180] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. 2019. Domain knowledge guided deep learning with electronic health records. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 738–747.

[181] Jason Youn, Navneet Rai, and Ilias Tagkopoulos. 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. *Nature Communications* 13, 1 (2022), 2360.

[182] Gang Yu, Mohammad Tabatabaei, József Mezei, Qianhui Zhong, Siyu Chen, Zheming Li, Jing Li, LiQi Shu, and Qiang Shu. 2022. Improving chronic disease management for children with knowledge graphs and artificial intelligence. *Expert Systems with Applications* 201 (2022), 117026.

[183] Yue Yu, Kexin Huang, Chao Zhang, Lucas M Glass, Jimeng Sun, and Cao Xiao. 2021. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics* 37, 18 (2021), 2988–2995.

[184] Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. 2020. Steam: Self-supervised taxonomy expansion with mini-paths. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1026–1035.

[185] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association* 26, 4 (2019), 294–305.

[186] Hongyi Yuan, Keming Lu, and Zheng Yuan. 2023. Exploring Partial Knowledge Base Inference in Biomedical Entity Linking. *arXiv preprint arXiv:2303.10330* (2023).

[187] Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. 2021. Enhancing taxonomy completion with concept generation via fusing relational representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2104–2113.

[188] Xiangxiang Zeng, Xinqi Tu, Yuansheng Liu, Xiangzheng Fu, and Yansen Su. 2022. Toward better drug discovery with knowledge graph. *Current opinion in structural biology* 72 (2022), 114–126.

[189] Bai Zhang, Yi Fu, Yingzhou Lu, Zhen Zhang, Robert Clarke, Jennifer E Van Eyk, David M Herrington, and Yue Wang. 2021. DDN2.0: R and Python packages for differential dependency network analysis of biological systems. *bioRxiv* (2021), 2021–04.

[190] Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. 2021. Taxonomy completion via triplet matching network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4662–4670.

[191] Jiayou Zhang, Zhirui Wang, Shizhuo Zhang, Megh Manoj Bhalerao, Yucong Liu, Dawei Zhu, and Sheng Wang. 2023. GraphPrompt: Biomedical Entity Normalization Using Graph-based Prompt Templates. In *The 37th AAAI Conference on Artificial Intelligence*.

[192] Rui Zhang, Dimitar Hristovski, Dalton Schutte, Andrej Kastrin, Marcelo Fiszman, and Halil Kilicoglu. 2021. Drug repurposing for COVID-19 via knowledge graph completion. *Journal of biomedical informatics* 115 (2021), 103696.

[193] Shuo Zhang, Xiaoli Lin, and Xiaolong Zhang. 2021. Discovering DTI and DDI by knowledge graph with MHRW and improved neural network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 588–593.

[194] Xiaolin Zhang and Chao Che. 2021. Drug repurposing for Parkinson's disease by integrating knowledge graph completion model and knowledge fusion of medical literature. *Future Internet* 13, 1 (2021), 14.

[195] Xianli Zhang, Buyue Qian, Yang Li, Changchang Yin, Xudong Wang, and Qinghua Zheng. 2019. KnowRisk: an interpretable knowledge-guided model for disease risk prediction. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1492–1497.

[196] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*. 1315–1324.

[197] Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 24–34.

[198] Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2021. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics* 22, 4 (2021), bbaa344.

[199] Sijin Zhou, Xinyi Dai, Haokun Chen, Weinan Zhang, Kan Ren, Ruiming Tang, Xiuqiang He, and Yong Yu. 2020. Interactive recommender system via knowledge graph-enhanced reinforcement learning. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 179–188.

[200] Tiantian Zhu, Yang Qin, Qingcai Chen, Baotian Hu, and Yang Xiang. 2022. Enhancing Entity Representations with Prompt Learning for Biomedical Entity Linking. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*. 4036–4042.

[201] Yongjun Zhu, Chao Che, Bo Jin, Ningrui Zhang, Chang Su, and Fei Wang. 2020. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal* 26, 4 (2020), 2737–2750.

[202] Zhaocheng Zhu, Mikhail Galkin, Zuobai Zhang, and Jian Tang. 2022. Neural-symbolic models for logical queries on knowledge graphs. In *International Conference on Machine Learning*. PMLR, 27454–27478.

[203] Yuchen Zhuang, Yinghao Li, Junyang Zhang, Yue Yu, Yingjun Mou, Xiang Chen, Le Song, and Chao Zhang. 2022. ReSel: N-ary Relation Extraction from Scientific Text and Tables by Learning to Retrieve and Select. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 730–744.

Received 30 July 2024; revised –; accepted –