A review of biomedical datasets relating to drug discovery: a knowledge graph perspective

OXFORD

Briefings in Bioinformatics, 2022, 1–19

https://doi.org/10.1093/bib/bbac404

Review

A review of biomedical datasets relating to drug discovery: a knowledge graph perspective

Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt and William L Hamilton

Corresponding author: Stephen Bonner, E-mail: stephen.bonner1@astrazeneca.com

Abstract

Drug discovery and development is a complex and costly process. Machine learning approaches are being investigated to help improve the effectiveness and speed of multiple stages of the drug discovery pipeline. Of these, those that use Knowledge Graphs (KG) have promise in many tasks, including drug repurposing, drug toxicity prediction and target gene–disease prioritization. In a drug discovery KG, crucial elements including genes, diseases and drugs are represented as entities, while relationships between them indicate an interaction. However, to construct high-quality KGs, suitable data are required. In this review, we detail publicly available sources suitable for use in constructing drug discovery focused KGs. We aim to help guide machine learning and KG practitioners who are interested in applying new techniques to the drug discovery field, but who may be unfamiliar with the relevant data sources. The datasets are selected via strict criteria, categorized according to the primary type of information contained within and are considered based upon what information could be extracted to build a KG. We then present a comparative analysis of existing public drug discovery KGs and an evaluation of selected motivating case studies from the literature. Additionally, we raise numerous and unique challenges and issues associated with the domain and its datasets, while also highlighting key future research directions. We hope this review will motivate KGs use in solving key and emerging questions in the drug discovery domain.

Keywords: knowledge graph embeddings, disease–gene prediction, drug–target discovery

Introduction

Discovering new drugs is a complex task, requiring knowledge from numerous biological and chemical domains. Due to this, the process of developing a new drug and bringing it to market is expensive and has a high chance of failure [1]. Hence, researchers are striving to increase the probability of success for the drug discovery process. As part of this, the field is increasingly looking towards computational [2] and machine learning approaches to help in various tasks within the drug discovery process [3], with the success of approaches like AlphaFold2 being a clear example of the trend [4]. Related to this, graphs[1] have long been used for representing data in the life sciences as they are well suited to the complex interconnected systems often studied in the domain [5–7].

[1] Also commonly known as networks within the biological domain. In this review we use the term graph interchangeably with network and without loss of generality.

Recently Knowledge Graphs (KGs) have begun to be utilized to model various aspects of the drug discovery domain as they offer a way to integrate vast and disparate data sources into a single unified resource, which can enable the discovery of hidden patterns and relationships [8]. KGs are heterogeneous data representations and build upon the linked open data and semantic web principles [9]. In a KG, both the vertices and edges can be of multiple different types, allowing for more complex and nuanced relationships to be captured [8]. In the context of drug discovery, the entities represent key elements such as genes, disease or drugs, with the edge types capturing different categories of interaction between them (An example drug discovery KG schema is displayed in Figure 3). As an example of where having distinct edge types could be crucial, an edge between a drug and disease entity could indicate that the drug has been clinically successful in treating the disease. Conversely, an edge between the same two entities could mean the drug was assessed but ultimately proved unsuccessful. This distinction in the precise meaning of the relationship between

Stephen Bonner is a post-doctoral research fellow within Data Sciences and Quantitative Biology, Discovery Sciences, AstraZeneca Cambridge. His work focuses on using knowledge graphs and machine learning to predict novel drug targets.

Ian P Barrett is a Senior Director in the Data Sciences and Quantitative Biology department in Discovery Sciences, AstraZeneca Cambridge. In this role Ian has led teams spanning a range of computational disciplines, to further drug discovery efforts across multiple disease areas.

Cheng Ye is a Senior Data Scientist in the Data Sciences and Quantitative Biology department at R&D AstraZeneca, working on applying AI techniques such as knowledge graphs and machine learning to advance drug discovery across multiple therapy areas.

Rowan Swiers is a Senior Data Scientist at AstraZeneca working on applying AI, machine learning and knowledge graphs to improve target selection in early stage drug discovery.

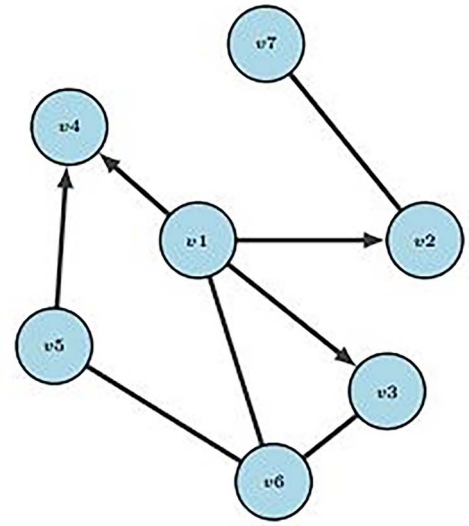
Ola Engkvist is head of the Molecular AI Department in Discovery Sciences, AstraZeneca Gothenburg and Professor of Machine Learning and AI for molecular design at Chalmers University of Technology. His main interest is to speed up the identification of novel clinical candidates with ML/AI.

Andreas Bender is a Professor for Molecular Informatics at Cambridge University, working on data analysis methods related to compound safety and efficacy. Previously he was a Director for Digitial Life Sciences at Nuvisan in Berlin, as well as as an Associate Director for Data Science and AI in the Clinical Pharmacology & Safety Sciences group at AstraZeneca.

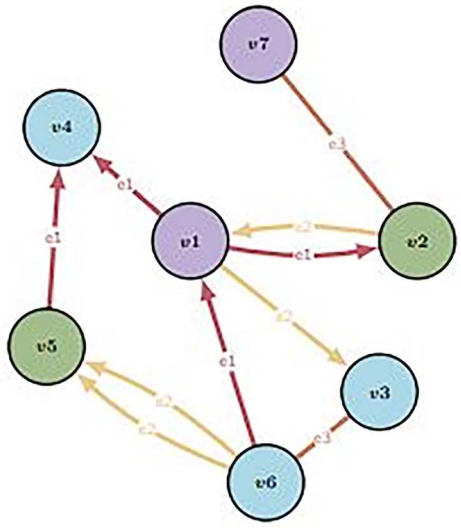
Charles Tapley Hoyt is a Postdoctoral Research Fellow in the Laboratory of Systems Pharmacology at Harvard Medical School based out of Germany. His research interests cover the interface of biocuration, knowledge graphs, and machine learning with systems biology, networks biology, and drug discovery.

William L Hamilton is an Adjunct Professor of Computer Science at McGill University and a Senior Quantitative Researcher at Citadel LLC. He develops machine learning models that can reason about our complex, interconnected world.

Received: May 12, 2022. **Revised:** July 14, 2022. **Accepted:** August 20, 2022



(a) A Homogeneous Graph.



(b) A Heterogeneous Graph.

Figure 1. A homogeneous and heterogeneous graph. In homogeneous graphs, vertices and edges are typically of only a single type. In comparison, heterogeneous graphs allow both vertices and edges to be of different type (Indicated by colour in this figure).

the two entities would not truly be captured in the simple binary option offered by homogeneous graphs, whereas a KG representation would preserve this important difference and enable that knowledge to be used to inform better predictions. As a topical concrete application, KGs have been utilized to address various tasks in helping to combat the coronavirus disease 2019 (COVID-19) pandemic [10–15]. Additionally, considering the domain as a KG has the potential to enable recent advances in graph-specific machine learning to be exploited [16].

However, constructing a suitable and informative KG requires that the correct primary data are captured in the process. An interesting aspect of the drug discovery domain, and perhaps in contrast to others, is that there is a wealth of well-curated, publicly available data sources, many of which can be represented as, or used to construct, KGs [17]. Many of these are maintained by government and international level agencies and are regularly updated with new results [17]. Indeed, one could argue that there is sometimes too much data available, rather than too

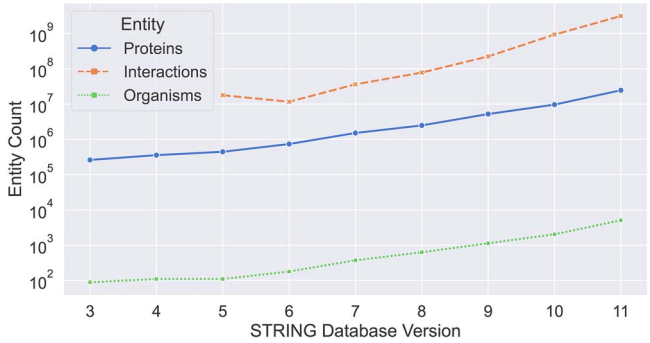


Figure 2. The evolution of the STRING database over major versions showing the increase in Organisms, Interactions and Proteins.

little, and researchers working in drug discovery must instead consider other issues when looking to use these data resources with graph analytics. Such issues include assessing how reliable the underlying information is, how best to integrate disparate and heterogeneous resources, how to deal with the uncertainty inherent in the domain, how best to translate key drug discovery objectives into machine learning training objectives and how to model and express data that are often quantitative and contextual in nature. Despite these complications, an increasing level of interest suggests that KGs could play a crucial role in enabling machine learning approaches for drug discovery [16, 18, 19].

We present a review of the publicly available data sources for drug discovery; we detail how they could be utilized in a KG setting and analyse the existing pre-constructed graphs. To the best of our knowledge, this is the first time these resources have been compared and evaluated in the literature. The primary contributions of this review are as follows:

- We detail the numerous unique research challenges posed by the use of KGs in the drug discovery domain.
- We review key data sources within drug discovery, present a taxonomy based on their primary biomedical area and consider how amenable they are for use in KGs by detailing what type of information could be extracted from them (relational versus entity features).
- We perform a comparative analysis of existing public drug discovery KGs based on their underlying data sources and graph composition decisions.
- We detail motivating case studies of KG use within drug discovery.
- We outline the key directions for future research and open problems within the domain.

Our hope is that this review will enable greater, easier and more effective use of KGs in drug discovery by signposting key resources in the field and highlighting some of the primary challenges. We aim to help foster a multidisciplinary and collaborative outlook that we believe will be critical in considering graph composition and construction in concert with analytical approaches and clarity of purpose.

An open-source collection of the resources detailed in this review has also been released. (https://github.com/AstraZeneca/awesome-drug-discovery-knowledge-graphs)

Review organization

In the Background Section, we introduce the required background information and detail existing work; in the Biomedical Ontologies Section, we introduce major ontologies from the domain which are often incorporated into KGs; the Exemplar Drug Discovery

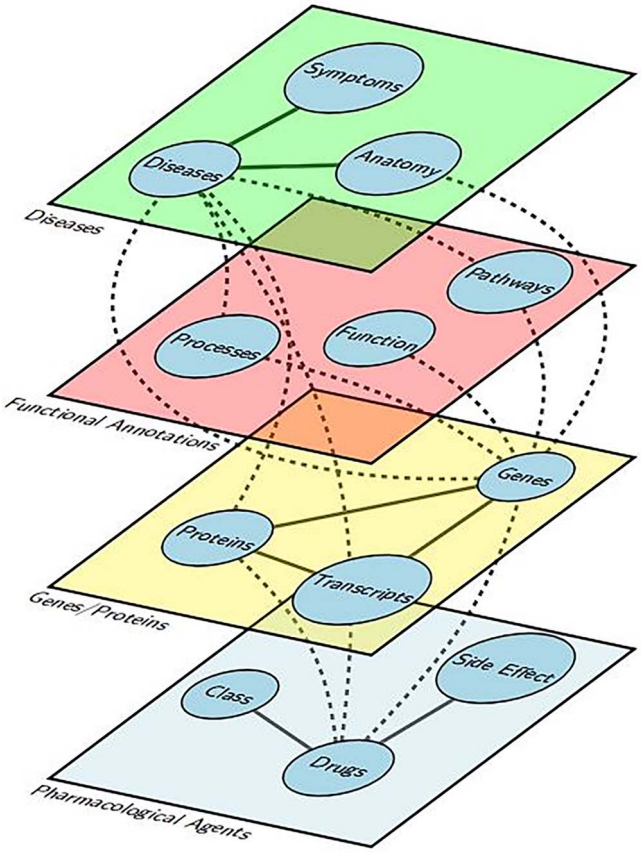


Figure 3. A simplified hierarchical view of a drug discovery knowledge graph schema.

Datasets Section details key sources which could be mined to populate a KG; the Existing Biomedical Knowledge Graphs Section details the existing public graphs and performs a comparative analysis of them; the Future Challenges AND Key Issues Section explores some of the key issues still facing the domain and the Conclusion Section concludes the work.

Background

In this section, we introduce key background concepts for the rest of the review including subtasks within drug discovery and KGs. We then detail prior work and explore some of the research challenges in the domain.

Subtasks within drug discovery

It can be helpful to consider partitioning the drug discovery process up into smaller subtasks, some of the most common being:

- *Drug Repositioning*—Which drugs previously tested in clinical trials could be ascribed new indications?
- *Disease Target Identification*—Which molecular entities (genes and proteins) are implicated in causing or maintaining a disease? Also known as Target Identification and Gene–Disease Prioritization.
- *Drug–Target Interaction*—Given a drug with unknown interactions, what proteins may it interact with in a cell? Also known as Target Binding and Target Activity.
- *Drug Combinations*—What are the beneficial or toxicity consequences of more than one drug being present and interacting with the biological system?
- *Drug Toxicity Predictions*—What toxicities may be produced by a drug, and in turn which of those are elicited by modulating the intended target of the drug, and which are from other properties of the drug? Also known as Toxicity Prediction.

Knowledge graphs

Knowledge graphs contain distinct different types of both vertices and edges, which can be defined as $G = (V, E, R)$ [20, 21]. Here, V is a set of vertices, R a set of relations and each edge is defined by its relation type $r \in R$, meaning that edges are represented as triplet values $(u, r, v) \in E, u, v \in V$ [22]. The vertices are often known as entities, with the first entity in the triple called the head entity, connected via a relation to the tail entity. Two vertices can be linked by more than one edge type, or even multiples of the same type.

An example KG is presented in Figure 1b and contains some key differences with a homogeneous graph: there are three types of vertex (blue, purple and green) and these are linked through a mix of directed and undirected edges of three relation types ($e1, e2, e3 \in R$). An example triple from this graph could be $(v_1, e_1, v_2) \in E$, stating that the entity v_1 is linked to v_2 via the e_1 relation type.

Prior work

The area of drug repurposing has been addressed in several reviews [23–26]. A recent work has detailed over 100 relevant drug repurposing databases, as well as appropriate methods [23]. In [24], a review of repurposing from the point of view of machine learning is presented, covering methods and over 20 datasets. KG-specific approaches for repurposing have been reviewed, with the authors detailing suitable datasets and then choosing six to form the KG used in their experimental evaluation [25]. In [26], the authors review and then partition the available drug databases into four categories based upon the type of information contained within: raw data, target-based, area specific and drug design.

The area of drug–target interaction has been reviewed [27, 28], both focusing upon the various methods for predicting interactions, however potential data sources are also presented. Conversely, machine learning based approaches for predicting drug–drug interaction have been detailed, with comparative evaluations conducted [29]. The authors construct a drug–drug interaction KG from a subset of Bio2RDF [30]. A review of 13 drug-related databases has been presented [31], covering a broad range of databases detailing drugs and drug–target interactions.

One study reviews both datasets and approaches for biological KG embeddings [32]. Although the review focuses upon the evaluation of different methodologies, 16 relevant databases are also discussed. However, as the work is experimentally driven, only a limited dataset discussion is undertaken. A different survey of the wider biomedical area and KG use within it has been presented [33]. Finally, a recent study presents a detailed overview of the application of graph-based machine learning in drug discovery [16]. The review is wide-ranging but makes no mention of suitable public datasets. We do however feel that it strongly complements our own review and serves as a method-focused counterpart to our dataset overview.

Knowledge graph use in drug discovery: research challenges

There are many challenges that arise when constructing a KG suitable for drug discovery tasks. Some of the most interesting research challenges are detailed below:

- *Graph Composition*—Strategies are needed to define how to convert data into information for modelling in a graph (e.g. instantiating a node or edge versus a feature on those entities), and what scale and composition of graph(s) may be optimal for a given task. Is a single large graph best, or should task-specific graphs be constructed? In addition, which type of analytical approach to use—reasoning-based, network/-graph theoretical, machine learning or hybrid approaches.
- *Heterogeneous & Uncertain*—In biomedical graphs, the data types are heterogeneous and have differing levels of confidence (e.g. well-characterized and curated findings versus NLP-derived assertions), and much of the data will be dependent on numerous factors—both time and the dose of drug used as well as the genetic background in the study. Overall, this means edges are much less certain, and thus less trustworthy, than in other domains.
- *Evolving Data*—The underlying data sources integrated and used in suitable KGs are also often changing over time as the field develops, requiring attention to versioning and other reproducible research practices. As an example of this, the evolution of the frequently used STRING dataset is demonstrated in Figure 2[2].
- *Bias*—There are various biases evident in different data sources, for example negative data remain underrepresented in some sources, including the primary scientific literature, and some areas have been studied more than others, introducing ascertainment bias in the graphs [34].
- *Fair Evaluation*—Several works have shown promise in applying machine learning techniques on a KG of drug discovery data. However, ensuring a fair data split is used for evaluation is perhaps more complicated than other domains, as it is easy for biologically or chemically meaningful data to leak across train/test splits. Thus, care should be taken to construct more meaningful data splits, as well as considering if replicated knowledge has been incorporated in the graph and could potentially leak across from the train/test split. For example, proteochemometrics approaches often employ a clustering-based splitting of chemicals to reduce leakage of similar chemicals between the training and testing sets [35].
- *Meaningful Evaluation*—While most practical applications of link prediction only focus on a single relation type (e.g. chemical modulates protein), metrics are often reported as an aggregate over all relation types. Because bias could be introduced by a large number of relations of other types either scoring much better or worse than the target relation type on average, leading to an inaccurate evaluation, metrics should be reported broken down by relation type.
- *Beware of Metrics*—Because common metrics used in link prediction tasks like mean rank (MR), mean reciprocal rank (MRR) and Hits at K are not comparable on results from KGs of different sizes, alternative metrics like the adjusted mean rank (AMR) should be employed [36]. Different implementations of link prediction evaluation calculate metrics very differently and caution should be observed when comparing results from different packages. Further, link prediction models built on biological KGs often influence real-world experimentation, so discussion on evaluation metrics should be considered with respect to how it can help achieve real-world goals.

[2] This data has been collected from https://string-db.org/cgi/access.pl?sessionId=dbw44gRWU7Xo&footer_active_subpage=archive

Ultimately we feel there is now an interesting opportunity to experiment at the intersection of various research fields spanning graph theoretic and other network analysis approaches for molecule networks [37, 38], machine learning approaches [19] and quantitative systems pharmacology [39].

Biomedical ontologies

This section details key biomedical ontologies which after often incorporated in KGs to help establish relations such as links between different disease subtypes and links between genes and a description of their function. An ontology is a set of controlled terms that defines and categorizes objects in a specific subject area. Modern biomedical ontologies are usually human constructed representations of a domain, capturing key entities and relationships and distilling the knowledge into a concise machine readable format [40]. There is a need for consistency when discussing concepts like diseases and gene functions which can be interpreted in multiple ways.

Ontology overviews

This section details the major ontologies which are relevant for drug discovery tasks, detailed in Table 1. Note that a full review is beyond the scope of this work and interested readers are referred to a dedicated review [41].

Disease ontologies

Due to the complexities associated with properly defining, categorizing and linking diseases, a large number of ontologies have been developed. Prominent examples include the Medical Subject Headings (MeSH) [43], Human Disease Ontology (DO) [45], Human Phenotype Ontology (HPO) [44] and Monarch Disease Ontology (Mondo) [42]. These typically differ in their intended use-case, for example DO was designed to help in the linking of different datasets, MeSH was created to aid in the indexing of MEDLINE/PubMed articles, HPO describes the phenotypes (the observable traits) of disease and Mondo was designed to harmonize disease definitions between other ontologies.

Gene-related ontologies

The function of genes and associated products is also frequently captured in ontologies, with common ones used in the construction of biomedical KGs such as Gene Ontology (GO) [46]. GO focuses on defining gene activity on the molecular level, linking genes to locations in the body where its function is performed and establishing links between genes and biological processes. In contrast, DTO focuses on linking gene products in relation to drug discovery considerations such as druggability.

Integrator ontologies

The Experimental Factor Ontology (EFO) was created by the European Bioinformatics Institute to provide a systematic description of experimental variables available in its databases including disease, anatomy, cell type, cell lines, chemical compounds and assay [47]. The Open Targets Platform uses EFO to provide the description, phenotypes, cross-references, synonyms, ontology and classification for annotating disease entities.

Exemplar drug discovery datasets

In this section, we introduce some of the key resources providing information on crucial entities within drug discovery: genes and gene products, disease and drugs, as well as sources capturing the

Table 1. An overview of ontologies suitable for use in drug discovery

Ontology name	Primary domain	Classes	Number of properties	Max depth	Licence	Update frequency
Monarch Disease Ontology (Mondo) [42]	Diseases	24K	25	16	CC BY 4.0	Monthly
Medical Subject Headings (MeSH) [43]	Medical terms	300K	38	15	Custom	Annually
Human Phentoype Ontology (HPO) [44]	Disease phenotype	19K	0	16	Custom	Bimonthly
Disease Ontology (DO) [45]	Diseases	19K	4	33	CC0 1.0	Monthly
Gene Ontology (GO) [46]	Genes	44K	11	–	CC BY 4.0	Monthly
Experimental Factor Ontology (EFO) [47]	Integrator	28K	66	20	Apache 2.0	Monthly

Table 2. Overview of integrated drug discovery resources

Dataset	First released	Update frequency	ELIXIR core	Data access	Commercial use	Summary
Open Targets Platform [49, 50]	2016	> Quarterly	✗	GraphQL, REST, Python, Flat files	✓	Resource focused for target discovery. Contains information from 20 data sources including UniProt, Reactome and ChEMBL.
Pharos [51]	2014	*approx* Monthly	✗	GraphQL	✓	Front end for the TCRD database for the drugable genome. Contains information from ChEMBL, STRING and UniProt.

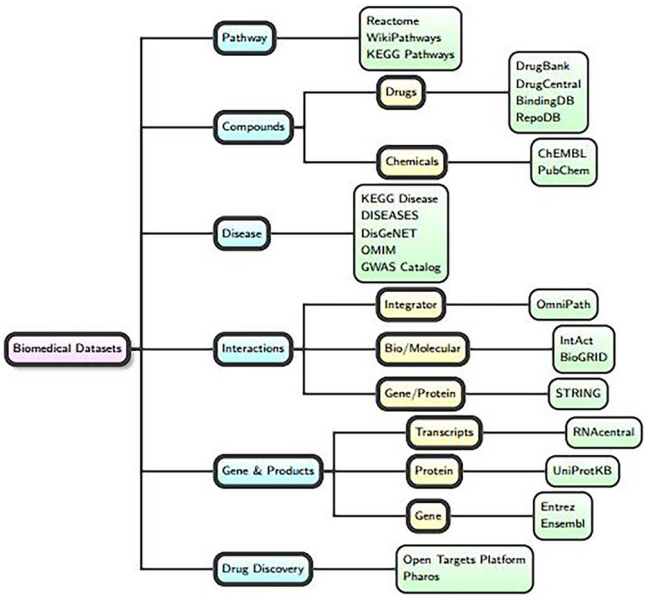


Figure 4. Dataset taxonomy.

relationships between them via interactions, pathways and processes. A taxonomy of these datasets is presented in Figure 4 and a description of each is provided in the Supplementary Material S2. We now detail key resources for each of these entities and explore what information could be extracted from the datasets for use in biomedical KGs.

Note on tables: Tables 2, 3, 5, 7, 9 and 11 compare datasets on when they were first released, how regularly they are updated, ELIXIR core resource [48] status (ELIXIR is a pan-European organization dedicated to detailing best practices for biomedical data and enabling stable funding. Core resources are datasets identified as crucial to the life science industry and of high scientific quality [48], thus we indicate those resources given this status.) and if free commercial use is allowed.

Integrated drug discovery resources

Table 2 outlines resources which are tailored specifically for the drug discovery field. Typically, these resources combine entity-specific data sources and add additional information useful for the domain. These resources can also be useful as a reference point for some best practices with regards to data handling and integration.

Gene and gene products

Genes and gene products (i.e. transcripts, proteins) are the key entities related to drug discovery and as such there are numerous rich public resources related to them. The gene and gene product datasets are summarized in Table 3.

Table 4 summarizes the potential types of relations and features which could be extracted from the gene and gene product resources. The table highlights that many of these resources contain rich information which could be mined for gene level features, be that from the gene or protein sequence (the sequence of nucleic or amino acids represented as base pair letters which can be mined to form a representation [56]), structure (the structure the protein forms once folded) or expression level (to what level is the gene expressed in different tissue types). The table also shows these resources to be good for extracting gene or protein interactions, as well as links to functional annotations via links to Gene Ontology (GO).

Interactions, pathways and biological processes

We now detail the resources specializing in the linking of the entities through interaction, processes and pathways. The interaction resources are presented in Table 5, while the processes and pathways resources are detailed in Table 7.

Table 3. Primary data sources relating to genes and proteins.

Dataset	First Released	Update Frequency	ELIXIR Core	Data Access	Commercial Use	Summary
UniProtKB [52]	2003	8 Weeks	✓	REST, Python, Java, SPARQL	✓	Primary protein resource. Can be mined for protein–protein interactions and protein features.
Ensembl [53]	1999	3 Months	✓	REST, MySQL dump	✓	Primary source for gene and transcripts. Gene–gene and gene–disease relationships can be extracted, as well as many gene-based features.
RNAcentral [54]	2014	3–6 Months	✗	REST, Flat file	✓	One of the primary sources of non-coding transcript data.
Entrez Gene [55]	2003	Daily	✗	Flat file	✓	Another primary gene data resource. Used in existing KGs for gene entity annotations.

Table 4. Comparing gene (G) and gene product (GP) resources on what relational information and entity-level features they provide.

	Potential Relations			Potential Features		
Dataset	G/GP-G/GP	Gene–Protein	G/GP-GO	Sequence	Structure	Expression
UniProtKB [52]	✓	–	✓	✓	✓	✓
Ensembl [53]	–	✓	✓	✓	✓	✓
RNAcentral [54]	–	–	✓	✓	✓	–
Entrez Gene [55]	✓	–	✓	✓	✓	✓

Table 5. Primary data sources relating to interactions.

Dataset	First Released	Update Frequency	ELIXIR Core	Data Access	Commercial Use	Summary
STRING [57]	2003	Monthly	✓	REST, Flat file, edgelist	✓	One of the most commonly used sources for physical and functional protein–protein interactions in existing KGs.
BioGRID [58]	2003	Monthly	✗	REST, Flat file, edgelist, Cytoscape	✓	Contains interactions between gene, protein and chemical entities with could be included directly in a KG.
IntAct [59]	2003	Monthly	✓	Flat file	✓	Contains molecular reactions between gene, protein and chemical entities. Uses UniProt for identifiers.
OmniPath [60]	2016	> Annually	✗	REST, Flat file, Cytoscape, Python, R	✓	An integrator of interaction resources that could be included in a KG via its RDF version.

Table 6. Comparing interaction resources on what relational information and features they provide.

	Potential Relations					Potential Features	
Dataset	Gene–Gene	Protein–Protein	Gene–Protein	Gene–Drug	Protein–Drug	Types	Weightings
STRING	–	✓	–	–	–	✓	✓
BioGRID	✓	–	–	✓	–	✓	✓
IntAct	–	✓	✓	–	✓	✓	✓
OmniPath	✓	✓	✓	✓	✓	✓	✓

Table 6 summarizes the potential relations types and features which could be extracted from the interaction resources. The table shows these data sources to be a rich potential for mining gene–gene or protein–protein interactions, with resources like IntAct and BioGrid also being suitable for extracting relations between gene products and compounds. All of the resources are also suitable for extracting features from the included weightings between the interactions or by providing different types or levels of interaction.

Pathways

Pathway resources comprise expert-curated subsets of interactions that are relevant for a given biological processes (e.g. apoptosis) or pathogenic mechanisms that lead to disease. There

Table 7. Primary data sources relating to pathways and processes.

Dataset	First Released	Update Frequency	ELIXIR Core	Data Access	Commercial Use	Summary
Reactome [62]	2003	> Annually	✓	Neo4J, Flat files	✓	A core resource for pathways and reactions. Amiable for graph representation and already included in several KGs.
WikiPathways [63]	2008	Monthly	✗	REST, SPARQL, RDF, Python, R, Java	✓	A crowdsourced collection of pathway resources. Also provided in graph amiable formats.
KEGG Pathways [64]	1995	Bi-Annually	✗	REST, R, Python	✗	A highly influential resource for pathways. Free use is limited to academic work only.

Table 8. Comparing pathway resources on what relational information and features they provide.

	Potential Relations			Potential Features	
Dataset	Protein–Protein	Gene–Pathway	Drug–Pathway	Graph Representation	Text Description
Reactome	✓	✓	✓	✓	✓
WikiPathways	–	✓	–	✓	✓
KEGG Pathways	–	✓	✓	✓	✓

Table 9. Primary data sources relating to disease.

Dataset	First Released	Update Frequency	ELIXIR Core	Data Access	Commercial Use	Summary
KEGG DISEASE [65]	2008	Monthly	✗	REST, Flat file	✗	A comprehensive disease resource for viewing disease as part of a biological system. Access is restricted for industrial use.
DISEASES [66]	2015	Daily	✗	Flat file	✓	A resource detailing links between genes and diseases. Already commonly used in drug discovery KGs.
DisGeNET [67]	2010	Annually	✗	REST, SPARQL, SQL, Flat tile, R, Cytoscape	✗	One of the most frequently used disease sources in existing KGs. Contains a mix of resources including experimental and text-mined data.
OMIM [68]	1987	Daily	✗	REST, Flat file	✗	One of the oldest disease databases, focusing upon mendelian disorders. Can provide gene-disease relationships.
GWAS Catalog [69]	2008	Biweekly	✗	REST, Flat file	✓	Contains the results from GWAS studies which could be used to provide less studied links between genes and diseases into a KG.

are implicit biases in the definitions which have been shown to be mitigated by harmonizing and combining their definitions when possible [61].

Table 8 summarizes the potential types of relations and features which could be extracted from the pathway resources. The table shows all resources can be mined for gene-pathway links, with Reactome and KEGG Pathway also providing links from drugs to affected pathways. The table also highlights how all of the resources contain text descriptions of the pathways, as well as a graph-based representation which could further be mined for features.

Diseases

Key resources containing information on diseases are detailed in Table 9.

Table 10 summarizes the potential relations and features which could be extracted from the disease resources. The table shows that unsurprisingly establishing gene to disease links is the primary focus of these resources. However, KEGG DISEASE could also be used to extract links from disease to both drugs and pathways, while DisGeNET also provides disease–disease similarity links. All of the resources provide some level of evidence for the links, while KEGG DISEASE and OMIM contain text descriptions which could be mined for features.

Drugs and compounds

Key datasets containing information relating to drugs and compounds are detailed in Table 11.

Table 12 summarizes potential relations and features which could be extracted from the drug resources. The table shows

Table 10. Comparing disease resources on what relational information and entity-level features they provide.

	Potential Relations				Potential Features	
Dataset	**Disease–Disease**	**Disease–Gene**	**Disease–Drug**	**Disease–Pathway**	**Text Description**	**Evidence**
KEGG DISEASE	–	✓	✓	✓	✓	✓
DISEASES	–	✓	–	–	–	✓
DisGeNET	✓	✓	–	–	–	✓
OMIM	–	✓	–	–	✓	✓
GWAS Catalog	–	✓	–	–	–	✓

Table 11. Primary data sources relating to drugs.

Dataset	First Released	Update Frequency	ELIXIR Core	Data Access	Commercial Use	Summary
ChEMBL [70]	2009	> Annually	✓	REST, SQL dump, SPARQL	✓	One of the primary resources for drug-like molecules. Could provide relational information between gene and drugs.
PubChem [71]	2004	As Sources Are	✗	REST, Flat file, SPARQL	✓	A comprehensive integrator of other chemical resources provided in RDF format, enabling easy incorporation into a KG.
DrugBank [72]	2006	> Annually	✗	REST, Flat file	✗	A rich source of drug, disease and gene information. Free use is limited to academic work only.
DrugCentral [73]	2016	Annually	✗	SQL, Flat file	✓	A collection of drug information extracted from literature and other sources. A potential source of drug features.
BindingDB [74]	1995	Weekly	✗	REST, Flat file	✓	A data resource of target protein and compound information. Already incorporated in existing KGs.
RepoDB [75]	2017	No Set Schedule	✗	Flat file	✓	A resources of drug to disease links containing both successful and failed examples. A rare source of negative information.

Table 12. Comparing drug resources on what relational information and entity-level features they provide.

	Potential Relations				Potential Features		
Dataset	**Drug–Drug**	**Drug–Gene**	**Drug–Disease**	**Drug–Pathway**	**Text Description**	**Structure**	**Attributes**
ChEMBL	–	✓	–	–	–	✓	✓
PubChem	✓	✓	–	✓	✓	✓	✓
DrugBank	✓	✓	–	–	✓	✓	✓
DrugCentral	–	✓	✓	–	–	✓	✓
BindingDB	–	✓	–	–	✓	✓	✓
RepoDB	–	✓	–	–	–	–	–

all the resources focus on providing links between drugs and genes, with PubChem and DrugBank being sources of drug–drug interactions and DrugCentral providing potential drug–disease linkages. Almost all of the resources provide compand structure information (Usually in SMILES format which can be used to learn a representation [76]) and numerical attributes (molecular weight for example). KEGG DISEASE and OMIM also provides text-based descriptions of the drugs which could be mined.

Dataset evaluation

We now summarize the key comparison points we have identified in our consideration of the highlighted datasets.

Data trust

Table 13 highlights the different types of information in the resources, in addition to information pertaining to the level of annotation available. Resources are compared as to whether they are curated by human experts, if information is taken from some form of experimental evidence or predicted and automated pipelines, and if the dataset contains information extracted from other primary resources. Resources are also compared if the province of the information is available (linking to the original manuscript or source), if any form of confidence weight is provided on the information and the directionality of potential edges that could be mined. The table shows that many of the covered resources have some level of human curation but it should be noted that this does not guarantee the accuracy of the information, as human bias and error can still be a factor. The table also highlights that predicted and automatically derived data are contained within many key resources such as STRING and DisGeNET, something to be cognizant of when including these in a KG. There are also various integrator resources available, like Omnipath and PubChem, which aggregate other primary datasets. While caution is needed around potential replicated knowledge, they offer a way for KGs to incorporate diverse information from a single resource.

Table 13. Comparing sources and annotations for the primary resources.

	Data Sources				Annotations		
Dataset	**Expert Curated**	**Experimental Evidence**	**Predicted and Automated**	**Integrator Resource**	**Provenance**	**Confidence**	**Directionality**
UniProtKB	✓	✓	✓	–	✓	✓	Undirected
Ensembl	–	✓	✓	–	✓	–	Undirected
RNAcentral	–	–	–	✓	✓	–	Mix
Entrez Gene	✓	✓	–	–	✓	–	Undirected
STRING	✓	✓	✓	–	✓	✓	Undirected
BioGRID	✓	✓	–	–	✓	✓	Mix
IntAct	✓	✓	–	–	✓	✓	Undirected
OmniPath	–	–	–	✓	✓	–	Mix
Reactome	✓	✓	✓	–	✓	✓	Mix
WikiPathways	✓	✓	–	–	✓	–	Mix
KEGG Pathways	✓	✓	–	–	✓	–	Mix
KEGG DISEASE	✓	✓	–	–	✓	–	Mix
DISEASES	✓	✓	✓	–	✓	✓	Undirected
DisGeNET	✓	✓	✓	–	✓	✓	Undirected
OMIM	✓	✓	–	–	✓	–	Undirected
GWAS Catalog	✓	✓	–	–	✓	✓	Undirected
ChEMBL	✓	✓	✓	–	✓	✓	Undirected
PubChem	✓	✓	–	✓	✓	–	Mix
DrugBank	✓	✓	–	–	✓	–	Undirected
DrugCentral	✓	✓	✓	–	✓	–	Undirected
BindingDB	✓	✓	–	✓	✓	✓	Undirected
RepoDB	–	✓	–	–	✓	–	Undirected

Relation mining

Figure 5 shows how the different resources covered in this review could be used to link key entities within a KG. The figure highlights how certain relation types are overrepresented by the datasets, with Gene–Gene and Gene–Drug having many potential sources. Care should be taken to avoid duplicated edges if many of these resources are used in graph composition. The figure also highlights where information is lacking, with Disease–Pathway links only being present in one source. It is also interesting to note that many of the resources detailed here are already provided in some form that is amenable for ingestion into a KG—either as edgelists or by providing RDF versions. This reduces the complexity of incorporating the resources as any issues arising from parsing and formatting process are avoided.

Graph enrichment

There are many primary data sources which capture more information about key entities within drug discovery than just relational interactions. UniProtKB, for example, details numerous sequence and functional properties of proteins which may not be captured by relations alone. However, thus far, this wealth of information is under-explored and could be used to greatly enrich a KG with more domain knowledge. Of course this would come at the potential cost of some level of manual feature engineering being required—an often complicated, domain-specific and iterative process by itself, and one that much of the research into representation learning is attempting to avoid [77, 78].

Untapped resources

Finally, there are resources specific to drug discovery, such as OpenTargets and Pharos, which have thus far not been incorporated into any public KG. However, they are not currently provided in a format enabling easy incorporation into a KG, meaning that some manual conversion process is required. Yet, they still hold great potential as a way to create a more drug discovery focused resource.

Existing biomedical knowledge graphs

This section highlights the few existing KGs covering various aspects of the drug discovery process. A selection of the most relevant resources is summarized in Table 14. A selection of other biomedical graphs and construction resources is detailed in Supplementary Material S3.

Biomedical knowledge graphs overviews

Hetionet v1.0. One of the first attempts to create a holistic KG suitable for various tasks within drug discovery was Hetionet [18]. Hetionet was developed as part of project Rephetio, a study looking at drug purposing through the use of KG-based approaches. The graph is publicly available[3] and is provided as a Neo4j [83] dump, as well as in JSON and edge list. The underlying data are mined from sources including Entrez Gene [55], DrugBank [84], DisGeNET [85], Reactome [62] and Gene Ontology [86]. The thresholds for the edges are not included in the graph, instead the preselected values are detailed in the accompanying paper [18].

From the time of writing, Hetionet has not been updated since 2017, although a project called the Scalable Precision Medicine Oriented Knowledge Engine (SPOKE) [87] looks to update Hetionet with extra data sources. However, to date, this updated resource has not been made publicly available, thus it has been excluded from our review.

Drug Repurposing Knowledge Graph. The Drug Repurposing Knowledge Graph (DRKG) [13] is a resource which builds upon Hetionet by integrating several additional data resources and was originally developed as part of a project for drug repurposing to

[3] https://het.io/

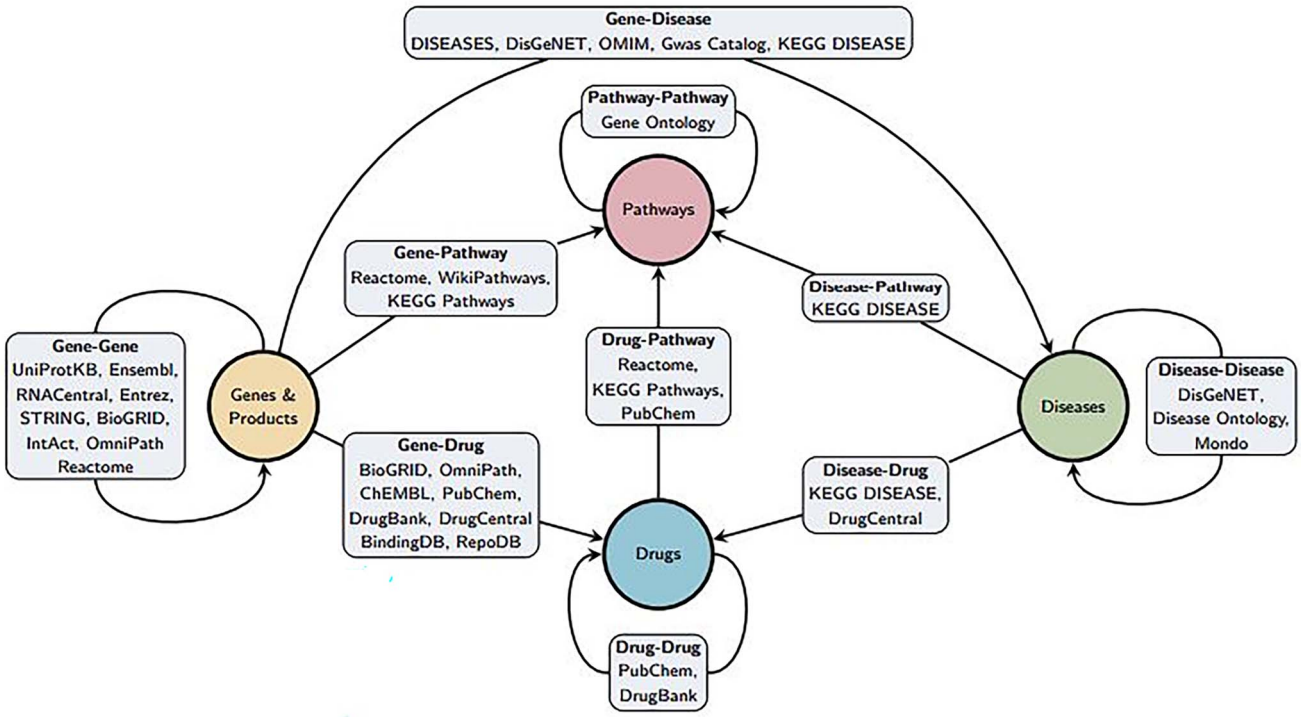


Figure 5. Dataset usage for relations to link entity types in a simplified drug discovery knowledge graph schema.

Table 14. Pre-existing knowledge graphs suitable for use in various drug discovery applications.

KG Dataset	Design Usecase	Entities	Triples	Entity Types	Relation Types	Contains Features	Constituent Datasets	Version Info	Last Update
Hetionet [18]	Repurposing	47K	2.2M	11	24	✗	29	✗	2017
DRKG [13]	Repurposing	97K	5.7M	13	107	molecular embeddings	34	✗	2020
BioKG [79]	General	105K	2M	10	17	categorical	13	✗	2020
PharmKG [80]	Repurposing/Target Prediction	7.6K	500K	3	29	continuous	7	✗	2020
OpenBioLink [81]	Benchmark	184K	4.7M	7	30	✗	17	✗	2020
Clinical Knowledge Graph [82]	Personalized Medicine	16M	220M	35	57	✗	35	✗	2020

target COVID-19 [88]. The dataset is aligned with the Deep Graph Library (DGL) package for graph-based machine learning [89], with pre-trained embeddings being provided from the package with the dataset. The data are publicly available (https://github.com/gnn4dr/DRKG.) and provided in edgelist format.

DRKG has enriched Hetionet with recent COVID-19 related data from STRING [57], DrugBank [84] and GNBR [90]. DRKG also includes pre-computed GNN-based embeddings for molecules, however no other entities have associated features.

BioKG. BioKG is a project for integrating various biomedical resources and creating a KG from them [79]. As part of the project, various tools are provided to enable a simplified KG construction process. A public pre-made version of the graph is available (https://github.com/dsi-bdi/biokg) , as well as the code for building it.

The data which make up BioKG is taken from 13 different data sources, including UniProt [52], Reactome [62], OMIM [68] and Gene Ontology [86]. One interesting aspect of BioKG is that a small number of categorical features are provided with some of the entities. For example, drug entities are enriched with information pertaining to any associated negative side effects.

PharmKG. The PharmKG project had the goal of designing a high-quality general purpose KG and associated GNN-based model for use within the drug discovery domain [80]. Table 14 shows that compared with others highlighted in this section, PharmKG is compact, containing entities of just three types: chemical, gene and disease.

The data are integrated from seven sources including OMIM [68], DrugBank [72], PharmGKB [91], Therapeutic Target Database (TTD) [92], SIDER [93], HumanNet [94] and GNBR [90]. A filtering process is then applied to ensure that only high-quality knowledge is kept. One unique aspect is that numerical features are provided with all the entities. Such features include chemical connectivity and other physiochemical features for the chemical entities, the use of BioBERT [95] to create features for the disease entities and a reduced expression matrix to create a feature vector for gene entities. The unfiltered PharmKG graph is available to download (https://github.com/MindRank-Biotech/

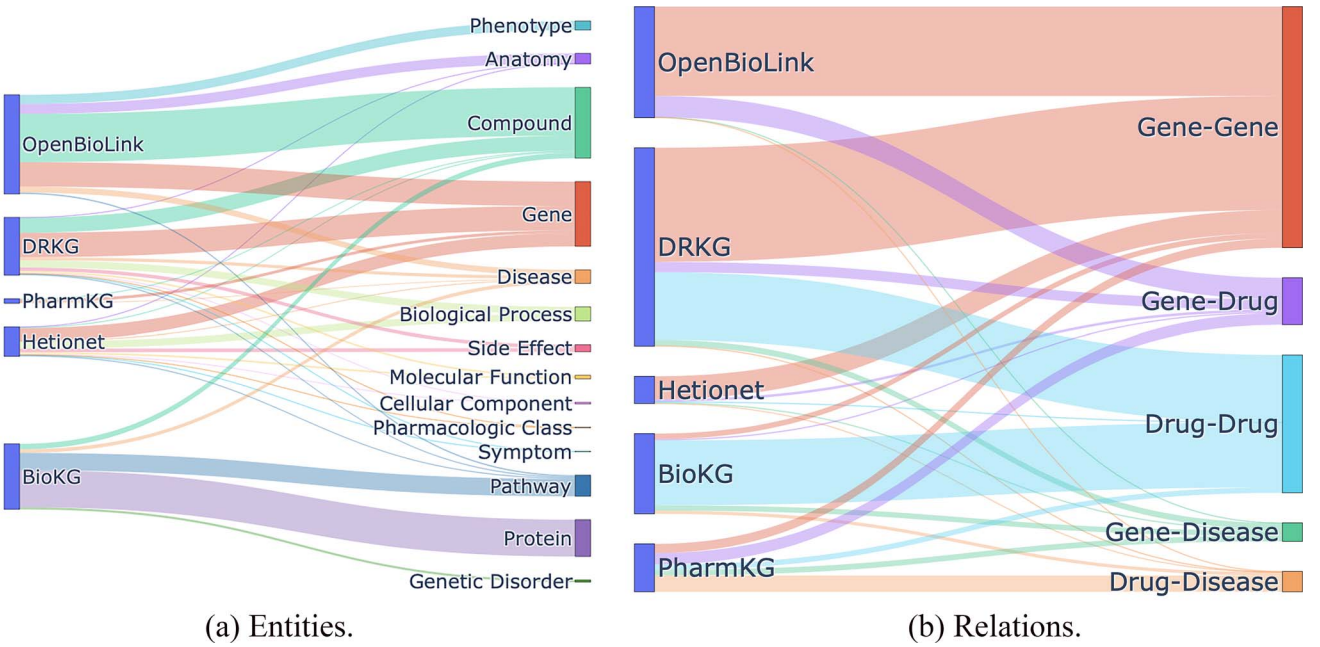


Figure 6. Sankey diagrams showing relationship between entity and relations in the KGs. Line thickness equates to entity volume. Note that for the relations, the value is the sum of all relation types between the two entities.

PharmKG), however at the time of writing, the entity features vectors have been released.

OpenBioLink. OpenBioLink (OBL) is a project to allow for easier and fairer comparison of KG completion approaches for the biomedical domain [81]. As part of the project, a benchmark KG has been created covering aspects of the drug discovery landscape. The dataset is publicly available (https://zenodo.org/record/3834052) and is provided in edgelist and RDF formats.

Data are taken from 17 datasets including STRING [57], DisGeNET [85], Gene Ontology [86], CTD [96], Human Phenotype Ontology [97], SIDER [93] and KEGG [98], among other resources. Of interest is that OpenBioLink contains additional *true negatives* for a selection of relation types, meaning that this relation was explicitly detailed not to exist. This can be used to avoid the issues inherent with the choice of negative sampling strategy when training KG embedding models [99].

Clinical Knowledge Graph. The Clinical Knowledge Graph (CKG) builds upon previous benchmark KGs but with additional focus on *-omics* data [82]. Its relations come from 25 databases and 10 ontologies, many of which overlap with previous examples but notably include protein state information such as post-translational modifications from PhosphoSite [100]. The CKG GitHub repository (https://github.com/MannLabs/CKG/) not only provides code for rebuilding the graph, but also tools for uploading it into Neo4J as well as visualization and exploration in Jupyter Notebooks. However, CKG cannot redistribute many of its constituent datasets because of licensing restrictions.

Comparative analysis of KG resources

We now present a comparative analysis of the KGs by considering graph composition choices, dataset usage and documentation levels. (Note that we exclude CKG from much of this analysis due to licensing limitations.) This analysis is undertaken to better understand the types of drug discovery problems each graph is suitable for addressing, as well as allowing interpretation of the level of trust that can be placed in each graph through exploration of dataset province. We believe this is the first time these resources have been compared and contrasted in the literature.

Graph composition: entities

Table 15 highlights which entity types are included in the KGs as well as offering a fine-grained view of how larger concepts like gene-products are modelled, while Figure 6a shows the amount of these different entities present across the KGs. These show that the KGs take differing approaches to how entities are modelled and the volumes included, which in turn could determine for which task they are best suited. Overall, the different KGs share only three common entities: gene products (be that genes or proteins), compounds and disease. As these are the core entities involved in drug discovery, this is no surprise. Pathways are also frequently included, with only PharmKG leaving them absent. BioKG and CKG are the only resources to model at the level of proteins instead of genes, while BioKG is the only resource to split genetic disorders from diseases. It can also be seen that Hetionet (and DRKG by virtue of it being an expanded Hetionet) captures more ancillary information compared with other KGs in the form of entities such as drug side effects, disease symptoms and various gene-level annotations, and thus might be well suited for tasks which could benefit from more fine-grained information.

Table 16 highlights the identifiers used by the KGs for key entities. Typically, entities of a certain type are represented using one of the two choices of identifier, with pathways having the lowest level of consistency. Knowing which identifiers are used allows additional sources of information to be joined onto the graphs with greater ease.

Graph composition—relations

Table 17 shows the number of different relationship types in the KGs between key entity pairs. The table highlights the different nuance with which the relationships are modelled. It is clear that there is a large variation in what an edge between entity pairs is actually representing. However, note that the values for DRKG are

Table 15. Comparison of a subset of entity types named across the knowledge graphs.

	Gene Products			Compounds		Disease					
KG Dataset	**Gene**	**Proteins**	**Transcripts**	**Drugs**	**Chemicals**	**Disease**	**Genetic Disorder**	**Anatomy**	**Pathways**	**Side Effect**	**Symptoms**
Hetionet	✓	–	–	✓	–	✓	–	✓	✓	✓	✓
DRKG	✓	–	–	✓	–	✓	–	✓	✓	✓	✓
BioKG	–	✓	–	✓	–	✓	✓	–	✓	–	–
PharmKG	✓	–	–		✓	✓	–	–		–	–
OpenBioLink	✓	–	–	✓	–	✓	–	✓	✓	–	–
CKG	✓	✓	✓	✓	–	✓	–	–	✓	–	–

Table 16. Entity identifiers used in the different knowledge graphs. Multiple IDs being present means all are used as entity identifiers within the graph.

KG Dataset	**Gene/Products**	**Compound**	**Disease**	**Pathways**
Hetionet	Entrez GeneID	DrugBank AN	DOID	Custom
DRKG	Entrez GeneID	DrugBank AN	MeSH	Custom
BioKG	UniProt	DrugBank AN	MeSH	Reactome/KEGG
PharmKG	Entrez GeneID	PubChem ID	MeSH	✗
OpenBioLink	Entrez GeneID	PubChem ID	DOID	Reactome/KEGG

Table 17. The number of relation types between entities across the KGs.

KG Dataset	**Gene–Gene**	**Gene–Disease**	**Gene–Drug**	**Drug–Drug**	**Drug–Disease**
Hetionet	3	3	3	1	2
DRKG	32	15	34	2	10
BioKG	1	1	5	1	1
PharmKG	1	6	7	1	6
OpenBioLink	10	1	10	✗	1

inflated as the data source is captured in the relationship name. Considering the relation granularity can further help guide on KG task suitability. For example Hetionet and OpenBioLink both have multiple relationship types between gene entities, while other KGs have only one, perhaps indicating these to be good choices if a complex understanding of gene interaction is required, whereas OpenBioLink would not be the graph to use if interaction between drugs was crucial to the task as it has no drug–drug edges. Overall, some general trends are observable regarding relation modelling choices. For example, BioKG tends to use only a single relation type, while PharmKG, despite its smaller overall size, often chooses to have multiple types. Additionally, it can be seen in Table 17 that drug entity pairs are consistently modelled as only a single relation type across the graphs.

Figure 6b displays the volume of each relation category contained within the graphs and shows a marked difference between the KGs. For example, DRKG has a large number of both Gene–Gene and Drug–Drug edges in comparison with other types, while OpenBioLink chooses to include more gene interactions and BioKG has a large number of drug interactions. Overall, Gene–Gene and/or Drug–Drug relations form the majority in many of the KGs. This has the potential to cause issues for tasks like target discovery, which relies on gene to disease connections, as there tend to be fewer examples in the graph. Thus, any model training on top of these graphs will have fewer to learn from, potentially leading to suboptimal predictive performance.

Underlying dataset use

Figure 7 represents dataset use in the KGs for a series of key relationship types, where the typed edges indicate relations are taken from that dataset. PharmKG and CKG are missing as it was not possible to determine the data sources used for the relations. The figure shows that the KGs utilize many of the same underlying datasets, with DrugBank for example being used in the majority of the graphs, with often multiple relationship types being extracted from it.

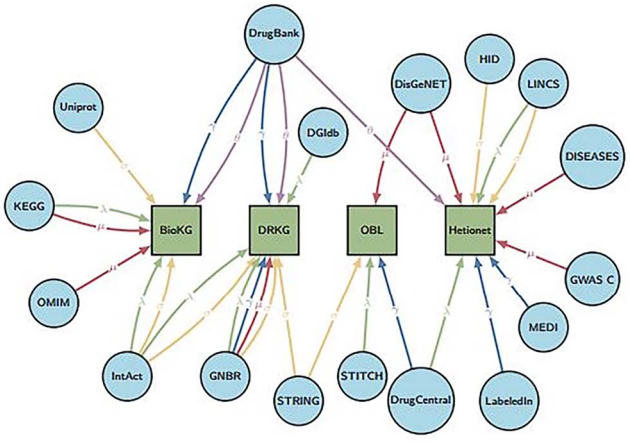


Figure 7. The relationship between drug discovery knowledge graphs and underlying data sources. Relationships are presented for five major relation categories: Gene–Gene (σ), Gene–Disease (μ), Gene–Drug (λ), Drug–Drug (θ) and Drug–Disease (γ).

Considering the difference in dataset usage, we can see some of the choices made during the composition pertaining to the

KGs intended use cases. For example, DRKG extracts four different relationship types from the text mining-based GNBR [90] dataset, while no other KG uses any. The creators of DRKG must have deemed these lower confidence edges useful for discovering potential repurposing candidates for COVID-19. It can also be seen that Hetionet tends to use multiple smaller datasets to build a single relation type, the gene–disease edges use three for example. Hetionet also differs in that its drug–disease edges do not come from larger aggregator resources DrugBank or DrugCentral used by the other KGs. In contrast, OpenBioLink has a one-to-one mapping between dataset and relation type and makes use of larger resources like STRING and STITCH, perhaps showing its intended benchmark use.

Figure 7 also highlights some of the pitfalls of mining multiple resources for relations of the same type. DRKG extracts gene–gene interaction edges from both IntAct and STRING, which could result in duplicated edges being present as both datasets contain many of the same interactions. Without care during the composition and evaluation process, this could lead to situations where training edges used in a model could also potentially be used for evaluation.

Evaluation of documentation quality and reproducibility

Table 18 presents our evaluation of the documentation quality (be that from the original paper, supplementary material, code repository or website) and overall reproducibility of the KGs. The graphs are evaluated using the following criteria, where the documentation quality categories are scored from one to three—*Schema Overview:* Is the graph schema design well explained and justified? A score of three means all entity and relation types detailed in full, two means the schema is outlined but not fully justified and one means only basic details are provided. *Dataset Filtering:* Is there a clear description of how the underlying datasets were filtered? A score of three indicates filtering thresholds detailed enough for reproducibility, two means that some description is provided but not enough to reproduce the work and one means that only a limited amount of information is provided. *Relation Explanation:* Is the meaning behind relations well explained? A score of three indicates each relation type is fully explained and linked to the source dataset, two means that either a full relation explanation or source dataset mapping was missing and one means that no mappings were provided. *Updates:* Are any future planned updates detailed? *Data-Relation Mappings:* Is it possible to map edges directly back to the underlying data sources? *Construction Code:* Is code available to construct the graph? *Licence Info:* Are underling dataset licences detailed?

Table 18 highlights that, despite being the oldest resource, Hetionet remains the KG with the highest overall level of documentation quality. Regarding reproducibility, two of the KGs did not provide code to recreate the graphs from the source datasets. It was also interesting to note that none of the resources provided any details on whether they would be updated going forward. Overall, it is clear that further work needs to be undertaken to improve documentation and reproducibility which will aid in both increasing trust and also ease of use of future KG resources.

Shortcomings of existing KG resources

When looking at these existing KG as a whole, we can identify the following shortcomings:

- *Lack of Updates*—None of the detailed KGs have any form of maintenance or update schedule in place. This means they will become increasingly out of date as the underlying data resources continue to evolve.
- *Lack of Detailed Documentation*—Some of the resources are not properly documented, missing clear justifications for some of the design choices, not including crucial information for reproducibility such as threshold information and lacking clear mappings back to the source datasets. This makes the graphs more challenging to use in tackling real-world problems.
- *No Dataset Version Information*—Many of the resources do not detail from which version or year of a certain dataset the information has been collected.

Case studies

In this section, we highlight case studies from the literature, detailed in Table 19, where KGs have been successfully exploited in drug discovery. We detail the successes, as well as analyse areas for further improvement. Note that an extra study is given in Supplementary Matetial section 4.

Polypharmacy prediction

The problem of adverse side effects that arises through the use of Polypharmacy (the use of more than one drug simultaneously to treat one or multiple conditions) has been modelled through the use of a KG and a novel GNN-based model entitled Decagon [19].

Graph Composition. The KG constructed was actually bipartite, containing only drug (over 900 unique entities) and protein (over 19K unique entities) entities [19]. These are linked through 964 unique edge types between drug–drug pairs, representing the various types of adverse side effects and a single edge type used to represent drug–protein and protein–protein interactions. Compared with the existing public KGs (Section 5), this graph places a lot of its complexity in the relation types, which has the potential drawback of limiting the amount of each seen during model training. Also, unlike the existing public KGs, the graph is limited to just two entity types, suggesting disease or pathway information was deemed unimportant.

Underlying Datasets. Data were extracted from protein-centric databases like BioGRID [103], STRING [57], STITCH [104], as well as drug-centric resources like SIDER [93], OFFSIDES and TWOSIDES [105], with much of the processed data being available in the BioSNAP project. Additionally, the graph is enriched with features on only the drug vertices containing descriptive single drug side-effect information.

Model. The model encoder, similar to a Relational Graph Convolutional Network (R-GCN) [106], uses a separate parameter matrix for each edge type to learn relational aware vertex level embeddings. These embeddings are then input into a tensor factorization-based decoder to directly predict potential negative drug–drug interactions via link prediction. The presented results show that compared with non-graph specific and homogeneous models, Decagon is better able to predict existing, and even propose novel, drug–drug interactions.

Gene–disease prioritization

The task of gene prioritization has been addressed via the use of a KG [102]. The overall approach, entitled Rosalind, details the construction of a knowledge graph and the choosing of a suitable model. The work proposes that the disease–target identification problem can be modelled as a link prediction task where the prediction of an edge between a disease and a gene entity would indicate possible association between the two.

Table 18. Comparing documentation levels and reproducibility across the KGs. Documentation quality is scored on scale of 1 to 3.

	Documentation					Reproducibility		
KG Dataset	**Schema Overview**	**Dataset Filtering**	**Relation Explanation**	**Updates**	**Overvierw**	**Data-Relation Mappings**	**Construction Code**	**Licence Info**
Hetionet	✓✓✓	✓✓✓	✓✓✓	✗	Graph creation process well documented and design choices well explained.	✓	✓	✓
DRKG	✓✓	✓	✓✓	✗	Dataset usage well documented but graph creation details lacking.	✓	✗	✓
BioKG	✓✓	✓	✓✓	✗	Overall good, could be improved by more details on construction choices	✓	✓	✓
PharmKG	✓	✓✓	✓✓	✗	Dataset to relation mapping not detailed but graph creation process explained.	✗	✗	✗
OpenBioLink	✓✓✓	✓✓	✓✓	✗	Well documented but relations could be better explained.	✓	✓	✓

Table 19. An overview of drug discovery related approaches in the literature employing the use of KGs.

						Dataset Information		
Approach	**Domain**	**Model**	**Prediction Task**	**Entites**	**Relations**	**Entity Types**	**Relation Types**	**Num Datasets in Graph**
Decagon [19]	Drug–Drug Interactions	Relational GCN with tensor factorization decoder	Link Prediction	19.6K	5.3M	2	964	≈7
TriModel [101]	Drug–Target Interactions	Tensor factorization	Link Prediction	5K	12K	11	26	3
Rosalind [102]	Disease–Gene Prioritization	Tensor factorization	Link Prediction	319K	2.6M	5	11	≈15

Graph Composition. The Rosalind KG comprises five entity type (genes, compounds, diseases, bio-processes and pathways) linked via 11 relation types. As such, Rosalind most closely resembles existing KGs like Hetionet and BioKG in structure. Of note is that it captures some of the subtlety around disease–gene prioritization, as ideally the model would predict which genes have some causal effect on the disease, not just an association. In Rosalind, they use two different types of edge between disease and gene entities—one indicating association and the other therapeutic links (a drug exists targeting the gene to help alleviate disease). However, to date, the authors have not released the KG, making reproducibility challenging.

Underlying Datasets. The Rosalind KG is constructed from many of the datasets detailed in this review. For example, the graph incorporates disease information from resources like DisGeNET, OMIM and GWAS Catalog, interaction information from BioGRID, pathway information from Reactome and compound information from ChEMBL.

Model. The model chosen for the work is the ComplEx tensor factorization approach [107]. The evaluation of the approach demonstrates that it outperforms competing methods, including OpenTarget [50], by as much as an extra 20% of recall when predicting potential gene–disease relationships over 198 diseases. Model performance is evaluated only on this therapeutic edge type. Additionally, results are presented on a time-slices graph, where the model is trained on historical data and predictions are made on future edges. This is attempting to replicate the task we would ideally want performed—using the currently available knowledge to predict currently unknown information, in this case, unknown relationships between genes and diseases.

Evaluative summary

These case studies have highlighted the potential and successes of KGs aiding in a diverse set of drug discovery tasks. However, there are still areas for improvement regarding aspects of underlying data use and graph composition. Thus, we make the following observations:

- *Composition*. The composition of the KGs in these studies varies dramatically regarding entity and relation type quantities. This suggests there is not yet a consensus on the optimal way to compose drug discovery KGs for use in ML pipelines.

- *Dataset Usage*. These studies use many of the datasets covered in this review to build their KGs. However there is still variety in where common relationship types are extracted from and usually no justification as to why a certain source was chosen.
- *One graph to rule them all?* It is striking that none of the approaches utilize an existing KG. Instead, custom task-specific graphs are still typically created, perhaps highlighting the challenge in creating a single KG to address all possible tasks within drug discovery.

Future challenges and key issues

While there has been significant progress made in the field, there are still numerous open challenges and issues to be addressed. In this section, we detail major areas still needing improvement, which could help produce better drug discovery KGs.

Graph Composition. Constructing a useful KG for use in the drug discovery domain is still a challenging problem, especially when performed by non-domain experts. Many choices must be made when transforming a data source into a graph, especially if it is not relational by nature. Here, there is however great scope for interdisciplinary collaborations between domain scientists and KG and machine learning researchers. Additionally, we would like to see more high-quality pre-constructed KGs, designed and validated by domain scientists, be made available for use by researchers. Further, creating graph construction toolkits, in which source datasets can be parsed in a unified and reproducible manner, would enable simpler creation of bespoke KGs. Additionally, the field should consider establishing common composition workflows with associated quality metrics. Such metrics could include a measure of triple uniqueness to help assess the impact for information bleed and adherence to included ontology hierarchies.

New Data Modalities. The continued evolution of computational methods means that new types of data modalities are being generated which could potentially be incorporated into a drug discovery KG. For example, predictions resulting from AlphaFold2 [2] could be converted into relational data and included in a graph. How best to combine predictions from these new modalities with existing relational resources is an open research question however.

Data Value. The availability of massive datasets has been partially credited with enabling the success of recent neural network models in areas such as computer vision [108]. It might be tempting then to incorporate as much data as possible into a drug discovery KG. However, much work still needs to be done in assessing the benefit of incorporating different data modalities. The consideration of value can also be extended to a financial view point: data collection, storage and processing can be expensive, especially if larger datasets do not improve performance in the task of interest. Another question is whether a single *super* graph should be created, which captures all knowledge around drug discovery, or whether smaller, more task-specific, projections enable better predictions overall. Ultimately, the value in the use of a particular dataset will be driven by its contribution to some downstream objective and as such will require careful experimentation by practitioners.

Better Metadata. As highlighted throughout the review, many of the core data resources are typically updated and refined at frequent intervals. However, many of the pre-existing KGs do not capture which dataset versions were used during its construction. Storing this information might allow for better reproducibility, as well as measure any change in predictive performance as the underlying knowledge is updated over time. Further, the incorporation of a time stamp on each edge capturing the year the particular interaction was established could allow for interesting opportunities in model validation and trend prediction. Improved metadata could also capture if the relationship was taken from an expert curated, or automated pipeline. Additionally, graphs could provide common alternative identifiers (for example including both Entrez and Ensembl identifiers for gene entities) as properties to enable easier incorporation of additional resources. More generally, KGs should be managed in accordance with FAIR principles [109] to help enable wider sharing and reuse of existing resource.

Graph Compatibility. In addition to the underlying data sources adhering to FAIR principles, steps could be taken to increase their utility for KG practitioners. For example, primary data sources could expose any relational data in RDF or other amiable open-source graph formats. Where this is not possible, example instructions and code could be provided to demonstrate how the data can be parsed into such formats. Good documentation, clearly describing the resource and any associated schemas, is also an invaluable tool in aiding simpler incorporation into KGs. As previously mentioned, mappings between any internal identifies and commonly used ones should be provided. Finally, clear descriptions of changes between dataset versions can aid practitioners when updating KGs.

Incorporation of Features. Typically many existing KGs are provided as little more than edge lists, with models trying to make predictions using this relational information alone. Throughout this review, we have attempted to highlight where data resources may be used to add additional features for entities and relations. However, it is easier to imagine suitable features for certain entities (proteins and chemicals for example, where structural information could be incorporated) than others. Additionally, any potential benefits of incorporating these extra features would need to be assessed fairly. Nevertheless, we feel that there is scope for the incorporation of features to enable graph-specific neural models to be better exploited in the domain, with some recent promising work being demonstrated in the literature [80].

Addressing Bias. Many biases will be present in a drug discovery KG and any model being trained upon it may have its predictive performance skewed away from underrepresented, but potentially crucial relationships. Even manually curated resources may incur bias from the person performing the curation. Practitioners should be aware of these issues and steps could be taken to mitigate them by, for example, reweighing the model training process. Additionally, users could consider removing over represented entities if they are confident that they are not required in the area of study. The lack of true negative samples in many graphs also means that the negative sampling strategy employed can bias the results. Recent inclusion of true negative samples in a benchmark graph [81] is encouraging, however where they are not possible to collect, more domain-aware sampling strategies should be investigated.

Careful Evaluation. Due to the combinatorial ingestion process used to construct KGs, it is common for edges to be duplicated if the relationship is captured in more than one source. This can cause obvious issues when it comes to creating train/test splits for evaluation if the issues are not considered. Further, the presence of trivial inverse relationships, many of which may be present, can also skew performance metrics [110]. It may also be more useful to assess model performance on more *biologically meaningful* data splits, for example by splitting on disease or protein family.

It could help move the field forward if meaningful splits for key tasks within drug discovery could be created by experts and made available for public use.

Uncertainty. So much of the data represented in a biological KG is uncertain, either due to the nature of the experiment that generated it, or because it has been automatically mined from the literature. Yet, this uncertainty is rarely represented in the graph itself, perhaps leading to a false sense of trust being created by the presence of certain relationships. We feel that more should be done to incorporate any uncertainty directly inside the KG. This could allow methods to directly learn from this information, thus creating better and more robust predictions.

Reproducibility. As in many areas of machine learning [111–113], reproducibility of results is still a major issue in the KG field [114]. It is common for many papers to publish results without also providing the exact graph constructed to generate them. We believe further improvements in this area are essential for continued development in the field.

Conclusion

The use of KGs, combined with machine learning techniques, has the potential to help address key challenges in the field of drug discovery, with promising early applications already being demonstrated in the tasks of drug repositioning, drug–drug interactions and gene prioritization. In this review, we have presented an overview of the various key related datasets which could provide some of the fundamental building blocks for a hypothetical drug discovery KG. The review has also detailed and evaluated the range of pre-existing public KGs in the drug discovery domain. Additionally, we have highlighted the many pitfalls and challenges of working with drug discovery-based data and signposted key issues practitioners should consider when choosing suitable sources.

Our hope is that this review of suitable data sources, combined with recent works evaluating graph-specific machine learning models in the context of drug discovery [16], can help guide researchers from across the KG mining and machine learning fields in applying state-of-the-art techniques in the field. Overall, we hope this review can serve as a catalyst in making the drug discovery domain more accessible, sparking new thought and innovation, while allowing researchers to more easily address key tasks within the domain, ultimately helping to improve and extend human life through new medicines.

Key Points

- Knowledge graphs offer a unified way to exploit the inherent interconnected nature of the drug discovery domain and can be used to help inform predictions about diverse drug discovery tasks including drug repurposing and target prioritization.
- However for these knowledge graphs to be truly useful, they need to be populated with high-quality information and constructed in a task-specific and suitable manner.
- We detail key data sources, suggest how these could be utilized to construct high-quality graphs and highlight potential pitfalls unique to the domain.
- We also present a comparative analysis of the existing public drug discovery KG resources along with motivating case studies.
- We conclude with highlighting promising future research directions for the field.

Acknowledgement

We would like to thank Manasa Ramakrishna, Ufuk Kirik, Benedek Rozemberczki, Natalie Kurbatova, Elizaveta Semenova and Claus Bendtsen for help and feedback throughout the preparation of this manuscript. Stephen Bonner is a fellow of the AstraZeneca postdoctoral program.

References

1. Morgan P, Brown DG, Lennard S, *et al.* Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat Rev Drug Discov* 2018;**17**(3):167.
2. Terstappen GC, Reggiani A. In silico research in drug discovery. *Trends Pharmacol Sci* 2001;**22**(1):23–6.
3. Vamathevan J, Clark D, Czodrowski P, *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019;**18**(6):463–77.
4. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**(7873):583–9.
5. Chen H, Ding L, Wu Z, *et al.* Semantic web for integrated network analysis in biomedicine. *Brief Bioinform* 2009;**10**(2):177–92.
6. Li X, Li W, Zeng M, *et al.* Network-based methods for predicting essential genes or proteins: a survey. *Brief Bioinform* 2020;**21**(2):566–83.
7. Rintala TJ, Ghosh A, Fortino V. Network approaches for modeling the effect of drugs and diseases. *Brief Bioinform* 2022;06.
8. Hogan A, Blomqvist E, Cochez M, *et al.* Knowledge graphs. *ACM Computing Surveys (CSUR)* 2021;**54**(4):1–37.
9. Jupp S, Malone J, Bolleman J, *et al.* The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* 2014;**30**(9):1338–9.
10. Bettencourt-Silva JH, Mulligan N, Jochim C, *et al.* Exploring the Social Drivers of Health During a Pandemic: Leveraging Knowledge Graphs and Population Trends in COVID-19. *Stud Health Technol Inform* 2020;**275**:6–11.
11. Cernile G, Heritage T, Sebire NJ, *et al.* Network graph representation of COVID-19 scientific publications to aid knowledge discovery. *BMJ Health & Care Informatics* 2020;**28**(1). https://informatics.bmj.com/content/28/1/e100254.citation-tools.
12. Domingo-Fernandez D, Baksi S, Schultz B, *et al.* COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* 2020;**37**(9):09.
13. Ioannidis VN, Song X, Manchanda S, *et al.* DRKG - Drug Repurposing Knowledge Graph for Covid-19; 2020. https://github.com/gnn4dr/DRKG/.
14. Reese JT, Unni D, Callahan TJ, *et al.* KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns* 2020;**2**:100155.
15. Wise C, Calvo MR, Bhatia P, *et al.* COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature. In: *Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP*, 2020, 1–10.
16. Gaudelet T, Day B, Jamasb AR, *et al.* Utilizing graph machine learning within drug discovery and development. *Brief Bioinform* 2021;**22**(6):05.

17. Rigden DJ, Fernández XM. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Res* 2020;**48**(D1):D1–8.
18. Himmelstein DS, Lizee A, Hessler C, *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife* 2017;**6**:e26726.
19. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018;**34**(13):i457–66.
20. Zhang C, Song D, Huang C, *et al.* Heterogeneous graph neural network. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, 793–803.
21. Hamilton WL. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning* 2020;**14**(3):1–159.
22. Lee B, Zhang S, Poleksic A, *et al.* Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. *Front Genet* 2020;**10**:1381.
23. Tanoli Z, Seemab U, Scherer A, *et al.* Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief Bioinform* 2020;**22**:1656–78.
24. Luo H, Li M, Yang M, *et al.* Biomedical data and computational models for drug repositioning: a comprehensive review. *Brief Bioinform* 2020;**22**:1604–19.
25. Zhu Y, Che C, Jin B, *et al.* Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics J* 2020;**26**:2737–50.
26. Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M, *et al.* Drug databases and their contributions to drug repurposing. *Genomics* 2020;**112**(2):1087–95.
27. Bagherian M, Sabeti E, Wang K, *et al.* Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Brief Bioinform* 2020;**22**:247–69.
28. Chen R, Liu X, Jin S, *et al.* Machine learning for drug-target interaction prediction. *Molecules* 2018;**23**(9):2208.
29. Celebi R, Uyar H, Yasar E, *et al.* Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC bioinformatics* 2019;**20**(1):1–14.
30. Belleau F, Nolin MA, Tourigny N, *et al.* Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;**41**(5):706–16.
31. Zhu Y, Elemento O, Pathak J, *et al.* Drug knowledge bases and their applications in biomedical informatics research. *Brief Bioinform* 2019;**20**(4):1308–21.
32. Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models. *Brief Bioinform* 2020;**22**:1679–93.
33. Callahan TJ, Tripodi IJ, Pielke-Lombardo H, *et al.* Knowledge-Based Biomedical Data Science. Annual Review of Biomedical Data. *Science* 2020;**3**:23–41.
34. Oprea TI, Bologa CG, Brunak S, *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* 2018;**17**(5):317.
35. Lopez-Del Rio A, Nonell-Canals A, Vidal D, *et al.* Evaluation of Cross-Validation Strategies in Sequence-Based Binding Prediction Using Deep Learning. *J Chem Inf Model* 2019;**59**(4):1645–57.
36. Berrendorf M, Faerman E, Vermue L, *et al.* On the Ambiguity of Rank-Based Evaluation of Entity Alignment or Link Prediction MethodsarXiv preprint arXiv:200206914. 2020.
37. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**(1):56–68.
38. Choobdar S, Ahsen ME, Crawford J, *et al.* Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**(9):843–52.
39. Sorger PK, Allerheiligen SR, Abernethy DR, *et al.* Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. In: *An NIH white paper by the QSP workshop group*, Vol. **48**. NIH Bethesda Bethesda, MD, 2011.
40. en Schulze-Kremer S. Ontologies for molecular biology. Computer and Information. *Science* 2001;**6**(21).
41. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. *Brief Bioinform* 2008;**9**(1):75–90.
42. Vasilevsky NA, Matentzoglu NA, Toro S, *et al.* Mondo: Unifying diseases for the world, by the world. *medRxiv* 2022.
43. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc* 2000;**88**(3):265.
44. Robinson PN, Köhler S, Bauer S, *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics* 2008;**83**(5):610–5.
45. Schriml LM, Mitraka E, Munro J, *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res* 2019;**47**(D1):D955–62.
46. Consortium GO. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**(suppl_1):D258–61.
47. Malone J, Holloway E, Adamusiak T, *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010;**26**(8):1112–8.
48. Durinx C, McEntyre J, Appel R, *et al.* Identifying ELIXIR core data resources. *F1000Research* 2016;**5**. https://f1000research.com/articles/5-2422/v2.
49. Koscielny G, An P, Carvalho-Silva D, *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* 2017;**45**(D1):D985–94.
50. Carvalho-Silva D, Pierleoni A, Pignatelli M, *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* 2019;**47**(D1):D1056–65.
51. Nguyen DT, Mathias S, Bologa C, *et al.* Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res* 2017;**45**(D1):D995–1002.
52. Apweiler R, Bairoch A, Wu CH, *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**(suppl_1):D115–9.
53. Yates AD, Achuthan P, Akanni W, *et al.* Ensembl 2020. *Nucleic Acids Res* 2020;**48**(D1):D682–8.
54. Sweeney BA, Petrov AI, Burkov B, *et al.* RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res* 2019;**47**(D1):D1250–1.
55. Maglott D, Ostell J, Pruitt KD, *et al.* Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005;**33**(suppl_1):D54–8.
56. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**(15):123–35.
57. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**(D1):D607–13.
58. Stark C, Breitkreutz BJ, Reguly T, *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**(suppl_1):D535–9.

59. Hermjakob H, Montecchi-Palazzi L, Lewington C, *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res* 2004;**32**(suppl_1):D452–5.
60. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 2016;**13**(12):966–7.
61. Mubeen S, Hoyt CT, Gemünd A, *et al.* The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front Genet* 2019;**10**:1203.
62. Jassal B, Matthews L, Viteri G, *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**(D1):D498–503.
63. Slenter DN, Kutmon M, Hanspers K, *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 2018;**46**(D1):D661–7.
64. Kanehisa M, Furumichi M, Tanabe M, *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1):D353–61.
65. Kanehisa M, Araki M, Goto S, *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 2007;**36**(suppl_1):D480–4.
66. Pletscher-Frankild S, Pallejà A, Tsafou K, *et al.* DISEASES: Text mining and data integration of disease–gene associations. *Methods* 2015;**74**:83–9.
67. Piñero J, Queralt-Rosinach N, Bravo A, *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015;**2015**. https://academic.oup.com/database/article/doi/10.1093/database/bav028/2433160.
68. Hamosh A, Scott AF, Amberger J, *et al.* Online Mendelian inheritance in man (OMIM). *Hum Mutat* 2000;**15**(1):57–61.
69. Buniello A, MacArthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**(D1):D1005–12.
70. Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**(D1):D930–40.
71. Kim S, Thiessen PA, Bolton EE, *et al.* PubChem substance and compound databases. *Nucleic Acids Res* 2016;**44**(D1):D1202–13.
72. Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**(suppl_1):D901–6.
73. Ursu O, Holmes J, Knockel J, *et al.* DrugCentral: online drug compendium. *Nucleic Acids Res* 2016;**45**:gkw993.
74. Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen* 2001;**4**(8):719–25.
75. Brown AS, Patel CJ. A standard database for drug repositioning. *Scientific data* 2017;**4**(1):1–7.
76. Hirohara M, Saito Y, Koda Y, *et al.* Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC bioinformatics* 2018;**19**(19):83–94.
77. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 2013;**35**(8):1798–828.
78. Mikolov T, Sutskever I, Chen K, *et al.* Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 2013;**26**:3111–9.
79. Walsh B, Mohamed SK, Nováček V. BioKG: A Knowledge Graph for Relational Learning On Biological Data. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, 3173–80.
80. Zheng S, Rao J, Song Y, *et al.* PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Brief Bioinform* 2021;**22**(4):bbaa344. https://doi.org/10.1093/bib/bbaa344.
81. Breit A, Ott S, Agibetov A, *et al.* OpenBioLink: A benchmarking framework for large-scale biomedical link prediction. *Bioinformatics* 2020;**36**:4097–98.
82. Santos A, Colaço AR, Nielsen AB, *et al.* A knowledge graph to interpret clinical proteomics data. *Nat Biotechnol* 2022;**45**:1–11.
83. Have CT, Jensen LJ. Are graph databases ready for bioinformatics? *Bioinformatics* 2013;**29**(24):3107.
84. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**(D1):D1074–82.
85. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**(D1):D845–55.
86. Consortium GO. The gene ontology project in 2008. *Nucleic Acids Res* 2008;**36**(suppl_1):D440–4.
87. Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat Commun* 2019;**10**(1):1–10.
88. Ioannidis VN, Zheng D, Karypis G. Few-shot link prediction via graph neural networks for Covid-19 drug-repurposingarXiv preprint arXiv:200710261. 2020.
89. Zheng D, Wang M, Gan Q, *et al.* Scalable graph neural networks with deep graph library. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, 1141–2.
90. Percha B, Altman RB. A global network of biomedical relationships derived from text. *Bioinformatics* 2018;**34**(15):2614–24.
91. Whirl-Carrillo M, McDonagh EM, Hebert J, *et al.* Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* 2012;**92**(4):414–7.
92. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res* 2002;**30**(1):412–5.
93. Kuhn M, Letunic I, Jensen LJ, *et al.* The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**(D1):D1075–9.
94. Hwang S, Kim CY, Yang S, *et al.* HumanNet v2: human gene networks for disease research. *Nucleic Acids Res* 2019;**47**(D1):D573–80.
95. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**(4):1234–40.
96. Davis AP, Grondin CJ, Johnson RJ, *et al.* The comparative toxicogenomics database: update 2019. *Nucleic Acids Res* 2019;**47**(D1):D948–54.
97. Köhler S, Carmody L, Vasilevsky N, *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 2019;**47**(D1):D1018–27.
98. Kanehisa M, Goto S, Furumichi M, *et al.* KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;**38**(suppl_1):D355–60.
99. Zhang Y, Yao Q, Shao Y, *et al.* NSCaching: simple and efficient negative sampling for knowledge graph embedding. In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, 614–25.
100. Hornbeck PV, Zhang B, Murray B, *et al.* PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Res* 2015;**43**(D1):D512–20.
101. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 2020;**36**(2):603–10.

102. Paliwal S, de Giorgio A, Neil D, *et al*. Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. *Sci Rep* 2020;**10**(1):1–19.
103. Oughtred R, Stark C, Breitkreutz BJ, *et al*. The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019;**47**(D1): D529–41.
104. Szklarczyk D, Santos A, von Mering C, *et al*. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* 2016;**44**(D1): D380–4.
105. Tatonetti NP, Patrick PY, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;**4**(125):125ra31–1.
106. Schlichtkrull M, Kipf TN, Bloem P, *et al*. Modeling relational data with graph convolutional networks. In: *European Semantic Web Conference*. Springer, 2018, 593–607.
107. Trouillon T, Welbl J, Riedel S, *et al*. Complex embeddings for simple link prediction. In: *International Conference on Machine Learning (ICML)*, 2016.
108. Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, 248–55.
109. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 2016;**3**(1):1–9.
110. Toutanova K, Chen D. Observed versus latent features for knowledge base and text inference. In: *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, 2015, 57–66.
111. Dacrema MF, Boglio S, Cremonesi P, *et al*. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 2021;**39**(2):1–49.
112. Errica F, Podda M, Bacciu D, *et al*. A fair comparison of graph neural networks for graph classificationarXiv preprint arXiv:191209893. 2019.
113. Lipton ZC, Steinhardt J. Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue* 2019;**17**(1):45–77.
114. Ali M, Berrendorf M, Hoyt CT, *et al*. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Trans Pattern Anal Mach Intell* 2021. https://ieeexplore.ieee.org/document/9601281.