Journal Title Here, 2024, pp. 1–15

doi: DOI HERE
Advance Access Publication Date: Day Month Year

Know2BIO: A Comprehensive and Evolving Multi-View Knowledge Graph for Biomedical Discovery

Yijia Xiao,[1,†] Dylan Steinecke,[2,3,†] Alexander R. Pelletier,[1] Yushi Bai,[4] Peipei Ping[2] and Wei Wang*[1]

[1]Department of Computer Science, UCLA, [2]Department of Physiology, David Geffen School Of Medicine, UCLA, [3]Medical Informatics Home Area, UCLA and [4]Department of Computer Science, Tsinghua University
[†]These authors contributed equally to this work.
*Corresponding author. weiwang@cs.ucla.edu
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Biomedical science has the potential to unlock new frontiers in human health by diagnosing, treating, and preventing conditions that plague billions of human beings. However, progress is hindered because of the siloed nature of our fragmented biomedical knowledge. Furthermore, its unfathomably complexity renders it prohibitive to be efficiently leveraged for knowledge discovery by human efforts alone. This necessitates automated data integration for powerful predictive methods which can propose biomedical hypotheses, prioritizing research directions. To capitalize on these opportunities, we present a biomedical knowledge graph, Know2BIO, which automatically integrates the latest biomedical information from 30 data sources into a multi-view and multi-modal network of 219,000 nodes and 6,180,000 edges—and counting. Its rich data covers biomedicine more comprehensively and in an updating manner not present in most other KGs. As a multi-view graph with multi-modal node features, Know2BIO is suited for cutting edge machine learning methods. To demonstrate, we apply 13 KG embedding methods to the graph, integrating this pipeline alongside the KG. Know2BIO is poised to serve the biomedical research community by providing first-rate knowledge to inform computational predictions of new biomarkers, drug indications, and disease processes.

Key words: Biomedical Knowledge Graph, Knowledge Discovery, Multi-View Learning, Knowledge Graph Embeddings

Introduction

The proliferation of biomedical data coupled with a concomitant revolution in computational methods is poised to continually enhance all facets of biomedical science. Though rapidly growing, this data from domains spanning genomics, transcriptomics, proteomics, pharmacology, healthcare, and pathology have unrealized potential. If computational methods are to leverage this data to discover new biomedical knowledge, the data must be effectively integrated into structures suited for predictive analyses. For optimal predictive analyses, data must richly represent biomedical information most pertinent to the undiscovered information.

Often, this data can be represented in knowledge graphs (KGs), collections of nodes and edges, i.e., triples, head nodes connected by a relation to a tail node. KGs are increasingly employed to flexibly encode interconnected concepts and facts, both within and across biomedical domains [33, 37, 87, 8, 83, 69, 64, 34, 12, 3, 16]. Moreover, predictive machine learning methods (e.g., knowledge graph embeddings and graph neural networks), can act upon the KG to discover new knowledge such as interactions between proteins and drugs, associations between genes and diseases, and drug repurposing predictions. Because this representation is upper bounded by the totality of current biomedical knowledge, the growth of knowledge over time elevates this upper bound. The primary domain-specific databases used to create these KGs often keep pace with this growth; this includes domains such as genomics [18, 67, 65], proteomics [7, 21, 71], metabolomics [60, 40, 82], pharmacology (e.g., drug design [81, 44, 19], drug targets [81, 84], adverse effects [23, 48]), physiology (e.g., biological processes [44, 21], and anatomy [26, 57, 53, 6]. However, the KGs themselves tend to remain static [33, 37, 87, 8, 83, 69, 34, 12, 3]. These KGs inevitably lack critical new information, warranting the creation of new KGs. Thus, either old efforts are put to waste (e.g., static KGs) or unnecessarily burdensome future efforts are required (e.g., new KG construction frameworks). Instead—as is only done scarcely and recently [64, 16]—to facilitate the proliferation of such resources and their biomedical usefulness, these KGs should be updateable in an automated fashion, evolving alongside the biomedical field. Reflecting the biomedical field also entails a comprehensive representation of biomedicine, a representation that is not confined to particular use cases [37] and does not ignore key data types available at the time of construction (Table 1).

Richer KG representations can be achieved by exploiting multi-modal and multi-view information of the same fact or concept. These modalities and views may include natural

language descriptions, molecular structures, atomic sequences, or semantically separate neighborhoods within the KG (e.g., instance view, ontology view). Though the tandem use of these modalities and views is often neglected, they each have great potential to augment representations and thus predictions.

Therefore, we propose a comprehensive, evolving, multi-modal KG for biomedicine, Know2BIO (Knowledge Graph Based on Biomedical Instances and Ontologies). Know2BIO represents a more comprehensive representation of the biomedical domain compared to popular biomedical knowledge graphs (Table 1); it is larger (219,000 nodes, 6,180,000 edges), integrates data from more sources (30 sources), represents 11 biomedical categories, and includes biomedically-relevant edge types absent in many other KGs (e.g., anatomy-specific gene expression, transcription factor regulation of genes). Not only is its data currently more up-to-date, but unlike most others, it can be automatically updated to reflect the most recent biomedical knowledge obtained from its primary data sources. By representing the latest scientific knowledge, Know2BIO defines a better real world learning task for graph learning methods and provides a greater opportunity for biomedical knowledge discovery. Additionally, Know2BIO enables methods development at the forefront of graph learning: its instance and ontology views enable multi-view KG learning tasks; its multi-modal node features (e.g., natural language descriptors, chemical sequences, protein structures) enable multi-modal and multi-view learning including advanced NLP techniques for biology and medicine (e.g., language models [35, 50, 62, 31]), as well as other data integration strategies [76, 86, 54, 36]. By providing a holistic KG that can reflect—in perpetuity—the latest biomedical knowledge from multiple perspectives, Know2BIO serves as an excellent resource to evaluate KG representation learning models for particular biomedical purposes (e.g., biomarker discovery, phenotyping, drug discovery, diagnostics).

To establish a framework for applying these methods to biomedical knowledge discovery, we evaluate 13 KG representation models from 3 categories on a KG-wide link prediction task, predicting missing nodes in triples. To elaborate upon the methods, these models learn low-dimensional embeddings that capture the contextual information of entities and their relationships. The 3 categories (Section S8) represent entities and relations in embedding spaces, either as translations in Euclidean space [11, 77, 52, 41], in complex space [73, 70, 15], or in hyperbolic space [5, 15].

- Know2BIO is a general purpose biomedical KG that covers a wide array of biomedical categories
- Know2BIO can be automatically updated to reflect the latest biomedical knowledge
- Know2BIO enables multi-modal learning with natural language, molecular sequences, and molecular structures.
- Know2BIO is a multi-view KG with a specified instance and ontology view
- We apply 13 KG representation learning models to Know2BIO, a pipeline integrated with the graph

Related Works

Biomedical Knowledge Graphs

Several biomedical KGs have been released in recent years. **Hetionet** [33] has been applied to predict disease-associated genes and for drug repurposing but is now relatively small and less up-to-date. Amazon's **DRKG** [37], has twice as many nodes, though it has has a narrow focus on COVID-19 drug repurposing. The Mayo Clinic's **BETA** [87] is a benchmark for predicting drug targets, but it is largely composed of older data from Bio2RDF [8], and its reported size is inflated due to unaligned nodes. **PharmKG** [83] includes non-graph data modalities for node features (e.g., gene expression, disease word embeddings), but it is relatively small and only has 3 node types [83]. The **iBKH** KG [69] represents the general biomedical domain, and although it is larger, over 90% of its nodes are molecule nodes linked to drug compounds. **CKG** [64] is a massive updating KG for clinical decision support, integrating experimental data, publications, and biomedical KBs; though impressive, its text mined data potentially introduces additional uncertainty, compared to carefully curated findings from biomedical KBs and its size may be intractable. **Open Graph Benchmark (OGB)** [34], a collection of KG benchmarks has a biomedical KG, ogbl-biokg, but it only includes 5 biomedical categories and is limited in size. Although **OpenBioLink** [12], is large, high-quality, and was intended to be updated, but like most other such KGs, it has not been continually updated. The **COVID-19 KG** [80] introduces a heterogeneous graph for visualizing complex relations in COVID-19 literature. **HKG-DDIE** [3] proposes a method to extract drug-drug interactions by integrating diverse pharmaceutical KG data with corpus text and drug information. **PrimeKG**, a recent updateable KG for enabling precision medicine, has diverse and rich data, broadly covering biomedicine and including natural language descriptions of nodes.

In sum, many biomedical KGs have been created and used for biomedical discovery. However, often their utility for biomedical discovery is hampered by old or static data, incomplete entity alignment, restricted focuses, substandard biomedical coverage, uni-modal/single-view data, and/or a lack of a provided analysis pipeline. A comparison of these KGs and our proposed KG, Know2BIO, is provided in Table 1.

Table 1. An overview of biomedical KGs and Know2BIO: the number of nodes, edges, node types, edge types, and data source; the presence of multi-modal information (natural language, molecular sequences & structures) provided with the KG; whether the KG can update.

KG	Nodes (mil.)	Edges (mil.)	Node tp.	Edge tp.	Data src.	Nat. Lng.	Mol. Seq.	Mol. Str.	Can Updt.
BETA	0.95	2.56	3	9	9		✓	✓	
CKG	16.0	220.0	36	47	15	✓			✓
DRKG	0.097	5.87	13	17	6			✓	
Hetionet	0.047	2.25	11	24	29	✓	✓		
iBKH	2.380	48.19	11	18	17	✓			
OGB:biokg	0.093	5.09	5	6	-				
OpenBioLink	0.184	9.30	7	30	16				
PharmKG	0.188	1.09	3	29	6	✓	✓	✓	
COVID-19 KG	0.336	3.33	5	5	1	✓			
HKG-DDIE	0.021	2.75	5	8	5	✓	✓	✓	
PrimeKG	0.162	4.60	10	29	20	✓			✓
Know2BIO	0.219	6.18	16	108	30	✓	✓	✓	✓

Know2BIO Knowledge Graph

We propose a general-purpose biomedical KG, Know2BIO, which represents 11 biomedical categories across 16 node types, totaling 219,169 nodes, 6,181,160 edges, and 30 unique pairings of node types across 108 unique edge types (Table 2). Most node pairs have 1-2 edge types, while compound-to-protein edges have 51 unique edge types. The most numerous edge types are compound-to-compound, at 2,902,659 edges.

Table 2. Scale and average degree of each biomedical category in Know2BIO.

Category	Nodes	Edges	Avg. node degree
Anatomy	4,960	226,630	45.7
Bio. Process	27,991	209,959	7.5
Cell. Component	4,096	96,239	23.5
Disease	21,842	419,338	19.2
Compound	26,549	3,561,235	134.1
Drug Class	5,721	10,859	1.9
Gene	28,476	1,757,428	61.7
Mol. Func.	11,272	85,779	7.6
Pathway	52,215	467,420	9.0
Protein	21,879	1,937,114	88.5
Reaction	14,168	236,113	16.7
Total	219,169	6,181,160	28.2

Node features, i.e., multi-modal data, are provided alongside the KG, enabling users to integrate and embed such data with models and feature fusion strategies of their choosing. These node features include DNA sequences for 22,000 gene nodes, amino acid sequences for 21,000 protein nodes, the SMILES sequence of 7,200 compound nodes (sequences which can be turned into graphs/structures), graph structures for 21,000 protein nodes, and natural language names for 208,500 nodes.

Knowledge Graph Construction

To construct our KG, we integrate data from 30 data sources spanning several biomedical disciplines (Table 3, Fig 2). Data sources often choose different vocabularies for the entities or concepts they store such as Ensembl, HGNC, or Entrez/NCBI identifiers (IDs) for genes. To connect the biological information from these data sources, we mapped IDs to a standard vocabulary for each node type, such as Entrez for all gene IDs. However, this process can be circuitous, requiring the identification and utilization of one or more intermediary resources. For example, to map the compound-targets-protein information from the Therapeutic Target Database (TTD) to the standard compound and protein vocabularies, DrugBank and UniProt, the following relationships are aligned. To map the protein from its TTD ID to its UniProt ID, we first use UniProt's API to map the TTD ID to the UniProt name, and then the UniProt name to the UniProt ID. To map the compound from its TTD ID to its DrugBank ID, we work backwards, first using DrugBank to map its IDs to external IDs: CAS, PubChem, and ChEBI. Then, we use TTD-provided mappings from CAS, PubChem, and ChEBI IDs to the new TTD IDs. Finally, with the TTD compound ID mapped to the DrugBank ID and the TTD protein ID mapped to the UniProt ID, the final standard relationship presented in the KG is produced, a DrugBank Compound ID targeting a UniProt Protein ID. This solution is often not immediately apparent, and the tedious nature of mapping these identifiers often disincentivizes the inclusion of extra databases, a hurdle we sought to overcome, including more databases than many other biomedical KGs (Table 1).

The alignment process was successful and complete for the majority of nodes. However, some explanation is needed for the discrepancy between the number of biomedical categories, 11, and the node types, 16. Some of the compounds could not be mapped from one of 10 vocabularies to both DrugBank and MeSH IDs. Thus, our chosen default in Know2BIO is to preserve these unaligned nodes at the risk of potential duplicate compounds represented by 2 nodes, one for DrugBank ID and one for MeSH ID. This decision is debatable, and therefore, if a user desires, the KG can be pruned to only preserve one vocabulary. Additionally, biological pathways are derived from 3 vocabularies. These 3 vocabularies are preserved in the KG because liberal attempts to map pathways did not yield any identity relationships between vocabularies. This is understandable because pathway definitions are subjective and based on the focus of biocurators such as small molecule drug pathways for SMPDB. Furthermore, MeSH IDs represent instances of diseases and anatomies. MeSH tree numbers represent the locations within a disease or anatomy ontology which an instance is linked to. Because merging these two IDs would collapse the disease and anatomy ontologies, we preserved both vocabularies.

The process to map the latest available version of this data can be executed through scripts we provide on our GitHub[1]. These scripts automatically construct part or all of the KG, although DrugBank, UMLS, and DisGeNET require free accounts and for two files to be manually downloaded into the input folder.

Relationships are also backed by varying levels of evidence (e.g., protein-protein associations from STRING and gene-disease associations from DisGeNET). We favored high confidence data, and our criteria for integrating data was based on how confidence scores are calculated, what past researchers have selected, KB author recommendations, and resulting data availability[2]. Many manually-curated sources did not provide confidence scores (e.g., GO, DrugBank, Reactome) and are ostensibly high-confidence and were not filtered. Appendix 9 and our GitHub[3] provide details on Know2BIO's unique relations between entity types.

Dual View Knowledge Graph

Often, a node in a KG may represent an entity in the *instance view* (e.g., a specific compound such as ibuprofen) or a concept in the *ontology view* (e.g., cardiovascular system drugs). Instance view relations include interactions or associations (e.g., a drug targeting a protein). Ontology view relations are typically hierarchical, relating concept to sub-concept (e.g., the mitochondria is a cellular component). Most KG representation learning methods fail to take advantage of the performance gains from jointly learning embeddings of each KG view [28, 38, 29]. Both views are specified in Know2BIO to enable multi-view learning. Together, the two views and the connecting bridge nodes form the whole view of the KG (Figure 1).

Benchmarking Know2BIO

Datasets

We comprehensively evaluate Know2BIO from the ontology, instance, and whole views. Bridge nodes are only evaluated as part of the whole view. The KG was split into a train, test, and validation set KG using GraPE [13]. Splits included all connected components with >10 nodes. To ensure connectivity of the train KG, the minimum spanning tree of each component was included, evenly splitting up to 20% of the remaining edges between test and validation Table S2 in the appendix.

[1] `https://github.com/Yijia-Xiao/Know2BIO/dataset`

[2] `https://github.com/Yijia-Xiao/Know2BIO/dataset/create_edge_files_utils/README.md`

[3] `https://github.com/Yijia-Xiao/Know2BIO/dataset/create_edge_files_utils`

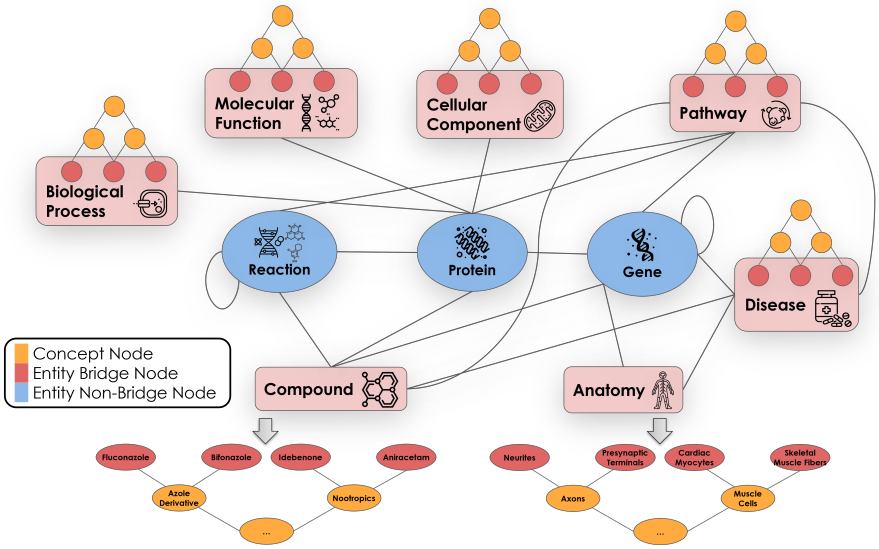

Fig. 1. Schema of Know2BIO. Blue nodes are for instances, red for bridge entities, and orange for concepts. Concepts are in a hierarchy.

Experiments

Evaluation Tasks

We evaluate KG embedding models on Know2BIO under a link prediction task: predicting a missing node (`h`/`t`) in a triple (`h`, `r`, `t`). Given a triple missing a node, the model scores candidates for the missing nodes. These candidate nodes are then ranked. Evaluation metrics include Hits@k and MRR (mean reciprocal rank). Hits@k quantifies the proportion of correct entities that are present within the initial k entities of the ranked list. MRR computes the arithmetic mean of the reciprocal ranks.

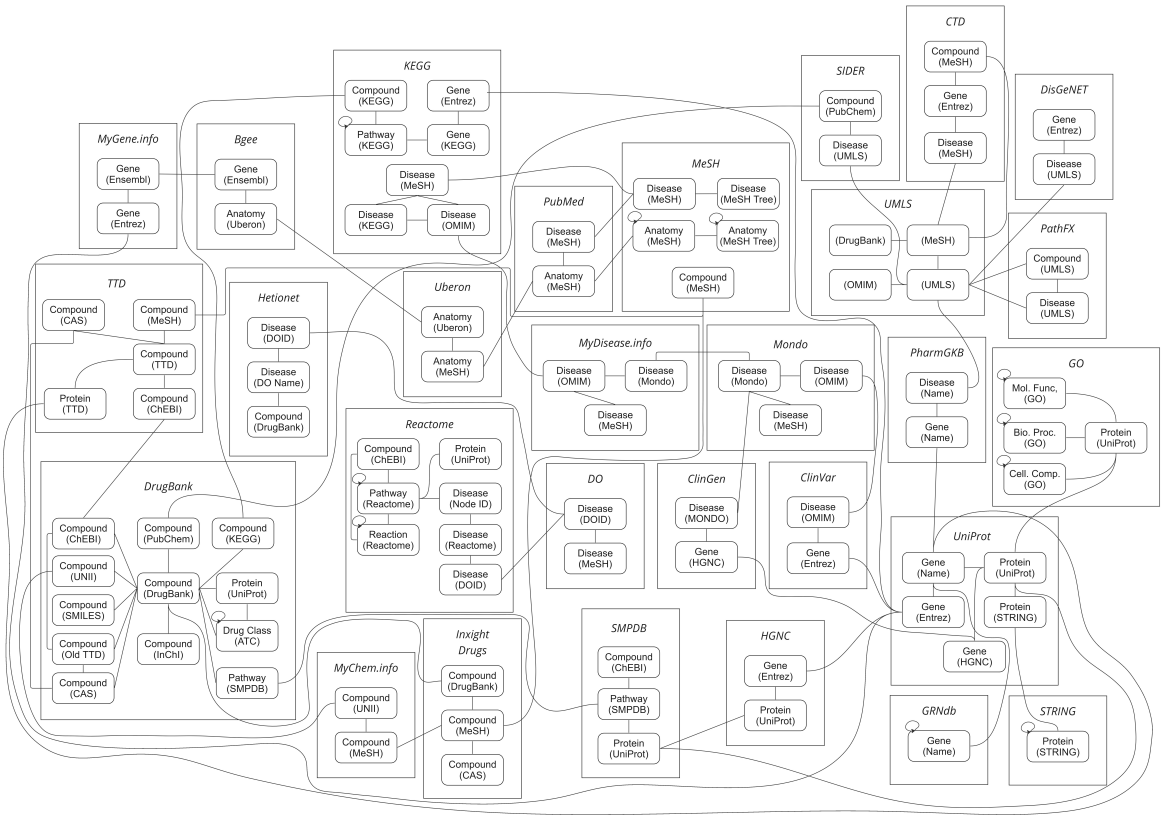

Fig. 2. Database schema of Know2BIO.

Table 3. Data Sources for Know2BIO's Biomedical Categories

Category	# Srcs.	Srcs.	Original Identifiers	Identifier(s) Aligned To
Anatomy	4	Bgee [6], PubMed, MeSH[53], Uberon [26, 57]	MeSH ID & tree number	MeSH ID & tree number
Bio. process	1	GO [14, 4]	GO	GO
Cell. component	1	GO	GO	GO
Compounds	11	DrugBank [81], MeSH, CTD [19], UMLS [9], KEGG [44], TTD [84], Inxight Drugs [68], Hetionet [85], PathFX [79], SIDER [48], MyChem.info [51]	DrugBank, MeSH ID, UMLS, UNII, ATC, KEGG Drug, KEGG Compound, PubChem Substance [46] & Compound [46], CAS [39], InChI [32], SMILES [78], ChEBI [30], TTD (two versions)	DrugBank, MeSH ID
Disease	14	PubMed, MeSH, DisGeNET[59], SIDER, ClinVar[49], ClinGen [61], PharmGKB[25], MyDisease.info[51] PathFX, UMLS, OMIM, Mondo, DOID[66], KEGG	MeSH ID & MeSH tree number, UMLS, DOID, KEGG, OMIM[2], Mondo[75]	MeSH ID, MeSH tree number
Drug Class	1	ATC	ATC	ATC
Genes	9	HGNC, GRNdb [22], KEGG, ClinVar, ClinGen, SMPDB [40] DisGeNET [59], PharmGKB [25], MyGene.info [51]	Entrez, Ensembl [18], HGNC [67],Gene name	Entrez
Molecular func.	1	GO	GO	GO
Pathways	3	Reactome[21], KEGG, SMPDB	Reactome, KEGG, SMPDB	Reactome, KEGG, SMPDB
Proteins	6	UniProt [20, 7], Reactome, TTD SMPDB, STRING, HGNC	UniProt, STRING [71], TTD	UniProt
Reactions	1	Reactome	Reactome	Reactome

Experiment Setup

As hyperparameter tuning has been demonstrated to strongly impact model performance and enable fair comparisons between models, [10]. we performed hyperparameter tuning with beam search on the batch size (512, 1024, 2048), learning rate ($1e^{-4}$,$5e^{-4}$,$1e^{-3}$,$1e^{-2}$,$1e^{-1}$), and negative sampling ratio (None, 5, 25, 50, 100, 125, 150, 250). We fix the maximum training epoch to be 1000 and early stopping patients to 5 epochs. Negative samples are constructed by replacing a positive triple's tail with a random node from the entire knowledge graph. We utilize the Adam optimizer [47] for Euclidean and hyperbolic models and SparseAdam for complex space models. SparseAdam is an Adam variant designed to handle sparse gradients and work efficiently with dense parameters. For testing, metrics are calculated by averaging the prediction performances on both heads and tails. Predictions are filtered by edge types' respective node types. All the models' hidden sizes are set to 512 to ensure a fair comparison. All experiments are performed on 2 servers with AMD EPYC 7543 Processor (128 cores), 503 GB RAM, and 4 NVIDIA A100-SXM4-80GB GPUs. Complete training, validation, and testing configuration files are available in Know2BIO's repository[4].

Results

We benchmark Know2BIO's ontology, instance, and whole views—not just a single view [17]—with models from Euclidean, complex, and hyperbolic spaces. Models are categorized into complex space, hyperbolic space, and Euclidean space models. Euclidean space models are further categorized into distance-based (Euclidean distance similarity) and semantic-based (dot similarity) models. For researchers unfamiliar with models' mechanisms, in Table S6 we detail their scoring functions [58, 42].

Ontology View

The ontology view of Know2BIO is characterized by a tree-like structure with 5 relation types, much less than the 76 types in the more densely connected instance view (Table S2). This scarcity of neighboring information makes modeling Know2BIO's ontology view non-trivial. Such properties enable researchers to better evaluate their models' capacity to capture biomedical knowledge in a hierarchical manner. Such hierarchical relations are best modeled by hyperbolic space models[15] which outperform Euclidean and complex space models on average (Table S3). Benchmark comparisons are shown in Figure 3.

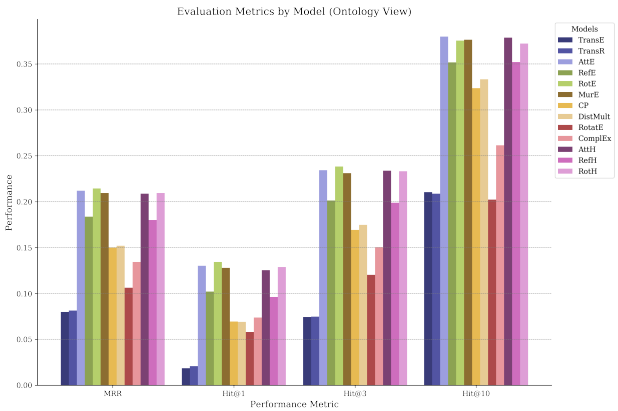

Fig. 3. Benchmark comparisons for the Ontology View of Know2BIO. Evaluation metrics for 13 models are shown, describing their Mean Reciprocal Rank (MRR), and Hits @ k.

Instance View

Know2BIO's instance view is more densely connected than the ontology view, providing more information for a node's embedding. However, enhanced context advantages also come at a price: the KG models need to represent more types of relations. Such properties enable researchers to better evaluate their models' capacity to capture the complex relations and structures in biomedical knowledge graphs. Such relations are best modeled by the complex space models which outperform Euclidean and hyperbolic space models on average (Table S4). Benchmark comparisons are shown in Figure 4.

[4] `https://github.com/Yijia-Xiao/Know2BIO/benchmark/configs`

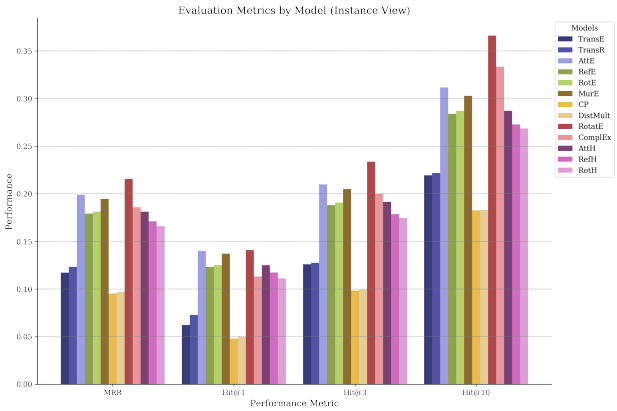

Fig. 4. Benchmark comparisons for the Instance View of Know2BIO. Evaluation metrics for 13 models are shown, describing their Mean Reciprocal Rank (MRR), and Hits @ k.

Whole View

In the ontology view, hyperbolic models perform best. In the instance view, complex models perform best. Euclidean models lie in the middle on average, depending on the embedding transformation strategy. To provide a balanced benchmarking scheme, we have created the whole view by adding bridge nodes (Table S2), entities that connect the instance view to the ontology view nodes. Table S5 show the evaluation of the whole view. For researchers using Know2BIO, we recommend evaluation on at least the whole view, since it measures models' capacities to capture both conceptual knowledge (ontology view) and factual knowledge (instance view). Benchmark comparisons are shown in Figure 5.

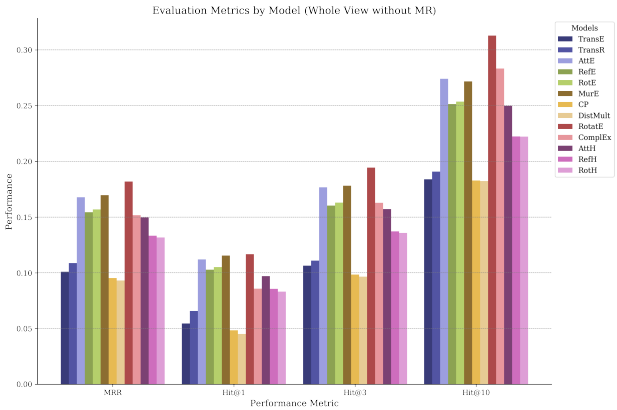

Fig. 5. Benchmark comparisons for the Whole View of Know2BIO. Evaluation metrics for 13 models are shown, describing their Mean Reciprocal Rank (MRR), and Hits @ k.

Conclusions

Know2BIO is aptly suited to drive biomedical discovery. This is because it provides usable biomedically salient information ready for predictive analyses. Know2BIO richly represents biomedicine: it is large, comprehensive, updating, and multi-modal/multi-view. Know2BIO is easy to use for a researcher's own scientific questions: its data construction pipeline affords the ability to incorporate biomedical data categories pertinent to a scientific investigation. The dataset can also be optimized for biomedical relevance because it is scalable: not only does it allow the addition or removal of automatically integrated data subsets, but it automatically integrates the most recent releases from 30 primary data sources and permits users to augment Know2BIO with their own data. These customized datasets can be effectively handled by the KG embedding pipeline to propose biological or medical hypotheses.

Limitations

Because biological systems are only partially understood, biomedical data will inevitably fail to capture all relevant biomedical realities. Know2BIO is not an exception, though great effort has been made to ensure that biomedical coverage is broad and that new discoveries are swiftly incorporated into the KG. It is also expected that updating the datasets will affect downstream analyses [72]. The philosophy driving our data curation process was to identify the biomedical data types related to molecular processes underlying disease phenotypes as well as how those affect the human body and how they can be modulated through therapeutic mechanisms. In doing so, we sought reputable datasets that could integrate with each other, filtering datasets for high confidence data (Section 3.1. Despite our best efforts, this process introduces some subjectivity into the curation.

In our application of 13 KG embedding methods, the unweighted version of the graph was used, (e.g., considering all partially similar diseases as equally similar). This likely hinders performance relative to the weighted version. Ultimately though, the purpose of this KG-wide predictive analysis is not as the final authoritative measure of each model, but as a display and setup for researchers who can benefit from biomedical hypothesis generation in their own use cases—broad or specific.

Future Work

Our primary goal for Know2BIO is to drive biomedical research. This is accomplished by a secondary goal, engineering effective machine learning pipelines. These pipelines can experiment with different KG models (e.g., KG embedding or GNN models), multi-modal learning strategies including feature fusion approaches and model selection (e.g., transformer models for text, graph learning methods for molecular structures), multi-view learning (e.g., dual view KG embeddings) [29, 38], pre-training strategies for KGs and the other modalities (e.g., protein language models), and training strategies (e.g., parameter initialization, curriculum learning).

Beyond these more canonical extensions of our link prediction pipeline, the KG can be leveraged for alternative purposes by emerging technologies. For example, large language models can utilize the up-to-date information and structure of Know2BIO to facilitate predictive analyses (e.g., in-context learning) and information retrieval (e.g., retrieval augmented generation). We plan to explore each of these areas in future works.

We also hope that this work influences the development of other biomedical KGs to include automated updates, multi-modal and multi-view data, and integrated predictive pipelines. These attributes will improve the biomedical information and user uptake of not only Know2BIO, but biomedical KGs in general.

References

1. M. Ali et al. Pykeen 1.0: A python library for training and evaluating knowledge graph embeddings. *J. Mach. Learn. Res.*, 22:82:1–82:6, 2020.
2. J. S. Amberger et al. Omim.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47:D1038 – D1043, 2018.
3. M. Asada, M. Miwa, and Y. Sasaki. Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature. *Bioinformatics*, 39(1):btac754, 11 2022.
4. M. Ashburner and C. A. B. et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
5. I. Balazevic, C. Allen, and T. M. Hospedales. Multi-relational poincaré graph embeddings. *ArXiv*, abs/1905.09791, 2019.
6. F. B. Bastian and J. R. et al. The bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Research*, 49:D831 – D847, 2020.
7. A. Bateman and M. J. M. et al. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51:D523 – D531, 2022.
8. F. Belleau et al. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41 5:706–16, 2008.
9. O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70, 2004.
10. S. Bonner et al. Understanding the performance of knowledge graph embeddings in drug discovery. *ArXiv*, abs/2105.10488, 2021.
11. A. Bordes et al. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.
12. A. Breit and S. O. et al. Openbiolink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, 2019.
13. L. Cappelletti and T. F. et al. Grape: fast and scalable graph processing and embedding. *ArXiv*, abs/2110.06196, 2021.
14. S. Carbon et al. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49:D325 – D334, 2020.
15. I. Chami et al. Low-dimensional hyperbolic knowledge graph embeddings. *ArXiv*, abs/2005.00545, 2020.
16. P. Chandak et al. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10, 2022.
17. D. Chang et al. Benchmark and best practices for biomedical knowledge graph embeddings. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2020:167–176, 2020.
18. F. Cunningham, J. E. Allen, et al. Ensembl 2022. *Nucleic Acids Research*, 50:D988 – D995, 2021.
19. A. P. Davis and T. C. W. et al. Comparative toxicogenomics database (ctd): update 2023. *Nucleic Acids Research*, 51:D1257 – D1262, 2022.
20. T. Dogan. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47:D506 – D515, 2018.
21. A. Fabregat, S. Jupe, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 42:D472 – D477, 2013.
22. L. Fang and Y. L. et al. Grndb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Research*, 49:D97 – D103, 2020.
23. N. P. Giangreco and N. P. Tatonetti. A database of pediatric drug effects to evaluate ontogenic mechanisms from child growth and development. *Med*, 2021.
24. M. E. Gillespie and B. J. et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50:D687 – D692, 2021.
25. L. Gong, M. Whirl-Carrillo, and T. E. Klein. Pharmgkb, an integrated resource of pharmacogenomic knowledge. *Current Protocols*, 1, 2021.
26. M. A. Haendel and J. P. B. et al. Unification of multi-species vertebrate anatomy ontologies for comparative biology in uberon. *Journal of Biomedical Semantics*, 5:21 – 21, 2014.
27. X. Han et al. Openke: An open toolkit for knowledge embedding. In *Conference on Empirical Methods in Natural Language Processing*, 2018.
28. J. Hao and M. C. al. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.
29. J. Hao and C. J.-T. J. et al. Bio-joie: Joint representation learning of biological knowledge bases. *bioRxiv*, 2020.
30. J. Hastings et al. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44:D1214 – D1219, 2015.
31. M. Heinzinger and A. E. et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20, 2019.
32. S. R. Heller and A. M. et al. Inchi, the iupac international chemical identifier. *Journal of Cheminformatics*, 7, 2015.
33. D. S. Himmelstein and A. L. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6, 2017.
34. W. Hu et al. Open graph benchmark: Datasets for machine learning on graphs. *ArXiv*, abs/2005.00687, 2020.
35. K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019.
36. K. Huang et al. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36:5545 – 5547, 2020.
37. V. N. Ioannidis and X. e. a. Song. Drkg - drug repurposing knowledge graph for covid-19. `https://github.com/gnn4dr/DRKG/`, 2020.
38. R. G. Iyer and Y. B. et al. Dual-geometric space embedding model for two-view knowledge graphs. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.
39. A. Jacobs et al. Cas common chemistry in 2021: Expanding access to trusted chemical information for the scientific community. *Journal of Chemical Information and Modeling*, 62:2737 – 2743, 2022.
40. T. Jewison and Y. S. et al. Smpdb 2.0: Big improvements to the small molecule pathway database. *Nucleic Acids Research*, 42:D478 – D484, 2013.
41. G. Ji et al. Knowledge graph embedding via dynamic mapping matrix. In *Annual Meeting of the Association for Computational Linguistics*, 2015.
42. S. Ji and S. P. et al. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33:494–514, 2020.
43. J. M. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

44. M. Kanehisa et al. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45:D353 – D361, 2016.
45. M. Kanehisa et al. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51:D587 – D592, 2022.
46. S. Kim et al. Pubchem 2023 update. *Nucleic acids research*, 2022.
47. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
48. M. Kuhn et al. The sider database of drugs and side effects. *Nucleic Acids Research*, 44:D1075 – D1079, 2015.
49. M. J. Landrum et al. Clinvar: improvements to accessing data. *Nucleic acids research*, 2019.
50. J. Lee and W. Y. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2019.
51. S. Lelong et al. Biothings sdk: a toolkit for building high-performance data apis in biomedical research. *Bioinformatics*, 38:2077 – 2079, 2021.
52. Y. Lin et al. Learning entity and relation embeddings for knowledge graph completion. In *AAAI Conference on Artificial Intelligence*, 2015.
53. C. E. Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88 3:265–6, 2000.
54. Y. Luo et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8, 2017.
55. D. Mendez et al. Chembl: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47:D930 – D940, 2018.
56. P. Mozzicato. Meddra. *Pharmaceutical Medicine*, 23:65–75, 2020.
57. C. J. Mungall and C. T. et al. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13:R5 – R5, 2012.
58. D. Q. Nguyen. A survey of embedding models of entities and relationships for knowledge graph completion. *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, 2020.
59. J. Piñero and J. S. et al. The disgenet cytoscape app: Exploring and visualizing disease genomics data. *Computational and Structural Biotechnology Journal*, 19:2960 – 2967, 2021.
60. C. D. Powell and H. N. B. Moseley. The metabolomics workbench file status website: A metadata repository promoting fair principles of metabolomics data. *bioRxiv*, 2022.
61. H. L. Rehm, J. S. Berg, et al. Clingen–the clinical genome resource. *The New England journal of medicine*, 372 23:2235 – 42, 2015.
62. A. Rives and S. G. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118, 2019.
63. A. Sadeghi et al. Benchembedd: A fair benchmarking tool for knowledge graph embeddings. In *International Conference on Semantic Systems*, 2021.
64. A. Santos et al. Clinical knowledge graph integrates proteomics data into clinical decision-making. *bioRxiv*, 2020.
65. E. W. Sayers et al. Genbank. *Nucleic Acids Research*, 50:D161 – D164, 2021.
66. L. M. Schriml et al. The human disease ontology 2022 update. *Nucleic Acids Research*, 50:D1255 – D1261, 2021.
67. R. L. Seal and B. B. et al. Genenames.org: the hgnc resources in 2023. *Nucleic Acids Research*, 51:D1003 – D1009, 2022.
68. V. B. Siramshetty and I. G. et al. Ncats inxight drugs: a comprehensive and curated portal for translational research. *Nucleic acids research*, 2021.
69. C. Su et al. Biomedical discovery through the integrative biomedical knowledge hub (ibkh). *iScience*, 26 4:106460, 2023.
70. Z. Sun et al. Rotate: Knowledge graph embedding by relational rotation in complex space. *ArXiv*, abs/1902.10197, 2018.
71. D. Szklarczyk and A. L. G. et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47:D607 – D613, 2018.
72. A. Tomczak et al. Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. *Scientific Reports*, 8, 2018.
73. T. Trouillon et al. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, 2016.
74. M. Váradi et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50:D439 – D444, 2021.
75. N. A. Vasilevsky et al. Mondo: Unifying diseases for the world, by the world. In *medRxiv*, 2022.
76. F. Wan et al. Neodti: Neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *bioRxiv*, 2018.
77. Z. Wang et al. Knowledge graph embedding by translating on hyperplanes. In *AAAI Conference on Artificial Intelligence*, 2014.
78. D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.
79. J. L. Wilson and R. R. et al. Pathfx provides mechanistic insights into drug efficacy and safety for regulatory review and therapeutic development. *PLoS Computational Biology*, 14, 2018.
80. C. Wise et al. Covid-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature, 2020.
81. D. S. Wishart et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46:D1074 – D1082, 2017.
82. D. S. Wishart et al. Hmdb 5.0: the human metabolome database for 2022. *Nucleic Acids Research*, 50:D622 – D631, 2021.
83. S. Zheng et al. Pharmkg: a dedicated knowledge graph benchmark for biomedical data mining. *Briefings in bioinformatics*, 2020.
84. Y. Zhou and Y. Z. et al. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Research*, 50:D1398 – D1407, 2021.
85. Y. Zhu, O. Elemento, J. Pathak, and F. Wang. Drug knowledge bases and their applications in biomedical informatics research. *Briefings in bioinformatics*, 2019.
86. N. Zong et al. Drug-target prediction utilizing heterogeneous bio-linked network embeddings. *Briefings in bioinformatics*, 2019.

87. N. Zong et al. Beta: a comprehensive benchmark for computational drug–target prediction. *Briefings in Bioinformatics*, 23, 2022.

Knowledge Graph Schema

Figure 2 illustrates the organization of Know2BIO. White rectangles represent different source databases, within which the smaller rectangles with round corners represent different node types. The lines linking them represent the relationships between various node types and source databases. The figure shows the various biomedical relationships and prerequisite node identifier mappings/alignments needed to construct Know2BIO. The italicized text at the top of a database rectangle is the database name. The text without parentheses in a node type rectangles is the node type, and the text in parentheses is the identifier vocabulary used.[5]

Here we provide details on the biomedical categories and data sources in Table 4. Know2BIO integrates data of 11 biomedical types represented by 16 data types using 32 identifiers extracted from 30 sources. Biomedical types are anatomy, biological process, cellular component, compounds/drugs, disease, drug class, genes, molecular function, pathways, proteins, and reactions. Each biomedical type has at least one data type/identifier in Know2BIO. Due to unalignable/disjoint sets of pathways across pathway databases, 3 pathway identifiers are used (Reactome, KEGG, SMPDB). Because we need to represent both the ontological structure of the anatomy data and MeSH disease, the anatomy and disease have 2 identifiers, one for unique MeSH IDs pointing to the potentially multiple MeSH tree numbers in the ontology; and the other for compounds due to incomplete alignment between DrugBank and MeSH identifiers. The remaining data types have 1 identifier to which all other identifiers are aligned.

The identifiers used include those of DrugBank, Medical Subject Headings IDs (MeSH), MeSH tree numbers, the old Therapeutic Target Database (TTD), the current TTD, PubChem Substance, PubChem Compound, Chemical Entities of Biological Interest (ChEBI), ChEMBL[55], Simplified Molecular Input Line Entry System (SMILES), Unique Ingredient Identifier (UNII), International Chemical Identifier (InChI), Anatomical Therapeutic Chemical Classification System (ATC), Chemical Abstracts Service (CAS), Disease Ontology, Online Mendelian Inheritance in Man (OMIM), Monarch Disease Ontology (Mondo), Gene Ontology, Small Molecule Pathway Database (SMPDB), Reactome, Kyoto Encyclopedia of Genes and Genomes (KEGG), Bgee, Uberon, SIDER [56], Comparative Toxicogenomics Database (CTD), PharmGKB, Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), UniProt, Gene Regulatory Network database (GRNdb), HUGO Gene Nomenclature Committee (HGNC), and Entrez, and Unified Medical Language System (UMLS).

The data sources include various databases, knowledge bases, API services, and knowledge graphs: MyGene.info, MyChem.info, MyDisease.info, Bgee, KEGG, PubMed, MeSH, SIDER, UMLS, CTD, PathFX, DisGeNET, TTD, Hetionet, Uberon, Mondo, PharmGKB, DrugBank, Reactome, DO, ClinGen, ClinVar, UniProt, GO, STRING, InxightDrugs, SMPDB, HGNC, and GRNdb.

We used these for different edges: Bgee for gene-anatomy edges; CTD for compound-gene and gene-disease; ClinGen for gene-disease; ClinVar for gene-disease; Disease Ontology for disease-disease alignments; DisGeNET for gene-disease; DrugBank for compound-compound (interactions and alignment), protein-compound, and pathway-compound; Gene Ontology for GO term ontology edges of molecular function, biological process, and cellular component, as well as edges between the GO terms and proteins; GRNdb for transcription factor to regulon edges, i.e., protein-gene; HGNC for gene-protein; Hetionet for compound-disease; Inxight Drugs for compound-compound alignments; KEGG for compound-pathway, pathway-pathway, pathway-gene, and alignments for disease-disease and gene-gene; MeSH for disease-disease, anatomy-anatomy, and compound-compound alignments, as well as disease-disease and anatomy-anatomy ontology edges; Mondo for disease-disease alignments; MyChem.info for compound-compound alignments; MyDisease.info for compound-compound alignments; MyGene.info for gene-gene alignments; PathFX for compound-disease; PharmGKB for gene-disease; PubMed for disease-anatomy; Reactome for reaction-reaction, compound-reaction, pathway-reaction, pathway-pathway, disease-pathway, and pathway-pathway, as well as alignments for disease-disease; SIDER for compound-disease (i.e., side effect / adverse drug event); SMPDB for protein-pathway and compound-pathway; STRING for protein-protein; TTD for compound-compound and protein-protein alignments, as well as compound-protein; Uberon for anatomy-anatomy alignments; UMLS for disease-disease and compound-compound alignments; and UniProt for protein-protein and gene-gene alignments.

The way in which the data and the identifiers were mapped to each other and merged into the same node is shown in Figure 2 and in the source code on GitHub, with provided documentation in the notebooks and README file. Except for the additional 5 of the 16 main identifiers discussed above, all other identifiers were mapped/aligned (often circuitously) to the main identifier types. In Know2BIO, these entities/concepts are represented by a unique node, not duplicating for the different identifiers as this would be computationally counterproductive and not biomedically insightful.

Node feature data is also included. DNA sequences were obtained from Ensembl and UniProt. Protein sequences were obtained from UniProt. Compound sequences were obtained from DrugBank. Protein structures were obtained from EBI DeepMind. Natural language names were obtained from the nodes' respective data sources.

Biomedical knowledge graphs are often very large. Therefore, we follow the common graph practice of subdividing the data to be benchmarked on multiple basic models. Here, we separately benchmark the ontology and instance views and then benchmark the whole dataset. Various toolkits have been developed to expedite the repetitive and time-consuming task of adapting models to datasets [27, 63, 13, 1]. We use the OpenKE [27] toolkit as it provides base models and tasks needed.

Below, we summarize the mapping process in more detail for the scripts that create the edge files / triple files[22]:

anatomy_to_anatomy The official xml file from MeSH was used to map anatomy MeSH IDs and MeSH tree numbers to each other, as well as MeSH tree numbers to each other to form the hierarchical relationships in the ontology. MeSH IDs were aligned to Uberon IDs via the official Uberon obo file (used in gene-to-anatomy)

[5] Although very recent versions of the data were used, the data used in this KG do not necessarily reflect the most current data from each source at the time of publication (e.g., PubMed, GO).

[22] https://anonymous.4open.science/r/Know2BIO/dataset/create_edge_files_utils

Table 4. Data Source for Know2BIO

Data Source	License
AlphaFold [43, 74]	CC-BY 4.0
Bgee [6]	CC0
CTD [19]	Custom[6]
ClinGen [61]	CC0[7]
ClinVar [49]	Custom[8]
DO [66]	CC0
DisGeNET [59]	CC BY-NC-SA 4.0
DrugBank [81]	CC BY-NC 4.0 International[9]
GO [14, 4]	CC Attribution 4.0[10] Unported
GRNdb [22]	Custom [22][11]
HGNC [67]	CC0
Hetionet [33]	CC0
Inxight Drugs [68]	None provided
KEGG [45]	Custom[12]
MeSH [53]	Custom[13]
Mondo [75]	CC-BY 4.0
MyChem.info [51]	Custom[14]
MyDisease.info [51]	Custom[15]
MyGene.info [51]	Custom[16]
PathFX [79]	CC0/CC-BY 4.0
PharmGKB [25]	CC-BY 4.0[17]
PubMed [18]	Custom[19]
Reactome [24]	CC0
SIDER [48]	CC-BY-NC-SA 4.0
SMPDB [40]	None provided
STRING [71]	CC-BY
TTD [84]	None provided[20]
Uberon [26, 57]	CC-BY 3.0
UMLS [9]	Custom[21]
UniProt [7]	CC-BY 4.0

compound_to_compound The compound_to_compound_alignment script aligned numerous compound identifiers in order to align DrugBank and MeSH IDs, two of the most prevalent IDs from data sources for different relationships in the scripts here. To produce this file, numerous resources were used to directly map DrugBank to MeSH IDs or to indirectly align the IDs (e.g., via DrugBank to UNII from DrugBank, then UNII to MeSH via MyChem.info). Resources include UMLS's MRCONSO.RRF file, DrugBank, MeSH, MyChem.info, the NIH's Inxight Drugs, KEGG, and TTD. In other scripts, DrugBank and MeSH compounds are mapped to one another via this mapping file.

Compound interactions were extracted from DrugBank.

compound_to_disease The majority of the compound-treats-disease and compound-biomarker_of-disease edges were from the Comparative Toxicogenomics Database. Additional edges were from PathFX (i.e., from repoDB) and Hetionet (reviewed by 3 physicians).

compound_to_drug_class Mappings from compounds to drug classes (ATC) were provided by DrugBank.

compound_to_gene Mapping compound to gene largely relies on CTD, though some relationships come from KEGG. Like many other compound mappings, this relies on the DrugBank-to-MeSH alignments from compound_to_compound_alignment.

compound_to_pathway Mapping compounds to SMPDB pathways relies on DrugBank. Mapping compounds to Reactome pathways relies on Reactome, plus alignments to ChEBI compounds. Mapping compounds to KEGG pathways relies on KEGG.

compound_to_protein Most compound-to-protein relationships are from DrugBank. Some are taken from TTD, relying on mappings provided by TTD and aligning identifiers based on DrugBank- and TTD-provided identifiers.

disease_to_disease The official xml file from MeSH was used to map disease MeSH IDs and MeSH tree numbers to each other, as well as MeSH tree numbers to each other to form the hierarchical relationships in the ontology.

To measure disease similarity, edges were obtained from DisGeNET's curated data. The UMLS-to-MeSH alignment was used (from compound_to_compound_alignment).

Disease Ontology was used to align Disease Ontology to MeSH. Mondo and MyDisease.info were relied on to align Mondo to MeSH, DOID, OMIM, and UMLS. These alignments were used to align relationships from other scripts to the MeSH disease identifiers.

compound_to_side_effect Mappings from compounds to the side effects they are associated with were provided by SIDER. This required alignments from PubChem to DrugBank (provided by DrugBank) and UMLS to MeSH (provided in compound_to_compound_alignment.py).

disease_to_anatomy Disease and anatomy association mappings rely on MeSH for aligning the MeSH IDs and MeSH tree numbers and rely on the disease-anatomy coocurrences in PubMed articles' MeSH annotations.

disease_to_pathway KEGG was used to map KEGG pathways to disease. Reactome was used to map Reactome pathways to diseases, relying on the DOID-to-MeSH alignments for disease.

gene_to_anatomy Gene expression in anatomy was derived from Bgee. To align the Bgee-provided Ensembl gene IDs to Entrez, MyGene.info was used. To align the Bgee-provided Uberon anatomy IDs to MeSH, Uberon was used (see anatomy_to_anatomy)
gene_to_disease Virtually all gene-disease associations were obtained from DisGeNET's entire dataset. Additional associations—many of which were already present in DisGeNET—were obtained from ClinVar, ClinGen, and PharmGKB. (Users may be interested in only using the curated evidence from DisGeNET or increasing the confidence score threshold for DisGeNET gene-disease association. We chose a threshold of 0.06 based on what a lead DisGeNET author mentioned to the Hetionet creator in a forum.
gene_to_protein We relied on UniProt and HGNC to map proteins to the genes that encode them. Notably, there is a very large overlap between these sources (95%). HGNC currently broke, so only UniProt is being used.
go_to_go The source of the Gene Ontology ontologies is Gene Ontology itself.
go_to_protein The source of the mappings between proteins and their GO terms is Gene Ontology.
pathway_to_pathway The source of pathway hierarchy mappings for KEGG is KEGG and for Reactome is Reactome. (SMPDB does not have a hierarchy)
protein_and_compound_to_reaction The source of mappings from proteins and compounds to reactions is Reactome. This file relies on alignments from ChEBI to DrugBank.
protein_and_gene_to_pathway To map proteins and genes to pathways, KEGG was used for KEGG pathways (genes), Reactome for Reactome pathways (proteins and genes), and SMPDB for SMPDB pathways (proteins).
protein_to_gene_ie_transcription_factor_edges To map the proteins (i.e., transcription factors) to their targeted genes (i.e., the proteins that affect expression of particular genes), GRNdb's high confidence relationships virtually all derived from GTEx, were used. This also required aligning gene names to Entrez gene IDs through MyGene.info
protein_to_protein Protein-protein interactions (i.e., functional associations) were derived from STRING. To map the STRING protein identifiers to UniProt, the UniProt API was used. A confidence threshold of 0.7 was used. (Users may adjust this in the script)
reaction_to_pathway To map reactions to the pathways they participate in, Reactome was used.
reaction_to_reaction To map reactions to reactions that precede them, Reactome was used.

Knowledge Graph Models Benchmarked in Experiments

The KG representation learning models used for experiments can be classified into five categories based on their mechanism (scoring function, etc.): translation-based models, bilinear models, neural network models, complex vector models, and hyperbolic space models. Generally, the neural network models' scoring functions are very flexible and can include various spatial transformations; while most translation-based and bilinear models are models in Euclidean space.

Translation-based models, also known as Trans-X models, conceptualize relations as translation operations on the representations of entities. For example, TransE perceives each relation type as a translation operator that moves from the head entity to the tail entity. The principle of this movement can be represented mathematically as $v_h + v_r \approx v_t$. TransE is particularly suited for capturing 1-to-1 relationships, where each head entity is linked to a maximum of 1 tail entity for a given relation type. Later, TransH, TransR, and TransD extended the core idea of translation-based representation.

Bilinear models such as DistMult represent each relation as a diagonal matrix, facilitating interactions between entity pairs. SimplE is an extension of DistMult, allowing for the learning of two dependent embeddings for each entity.

Neural network models leverage neural networks (e.g. convolutional neural networks) for knowledge graph embedding. ConvE and ConvKB are prime examples. ConvE employs a convolution layer directly on the 2D reshaping of the embeddings of the head entity and relation. ConvKB applies a convolution layer over embedding triples. Each of these triples is represented as a 3-column matrix, where each column vector represents one element of the triple.

Complex vector models use vectors from Complex or Euclidean space to expand their expressive capacity. Notable examples include ComplEx, RotatE, and AttE.

Hyperbolic space models take advantage of hyperbolic space's ability to represent hierarchical structures with minimal distortion. In Euclidean space, distances between points are measured using the Euclidean metric, which assumes a flat space. However, in hyperbolic space, distances are measured using the hyperbolic metric, which takes into account the negative curvature of the space. This property allows hyperbolic space models to capture long-range dependencies more efficiently than Euclidean space models. Models like RefH and AttH enhance the quality of KG embedding by incorporating hyperbolic geometry and attention mechanisms to model complex relational patterns.

Table 5. Summary statistics of Know2BIO's different views: number of nodes, relation types, training set triples, validation set triples, test set triples, and total triples

	Nodes	Edge Types	Train	Valid	Test	Total
Ontology	68,314	5	93,056	8,368	8,367	109,827
Bridge	102,111	29	366,780	45,748	45,748	475,779
Instance	145,445	76	3,320,385	415,050	415,050	5,595,554
Whole	219,169	108	3,780,221	469,166	469,165	6,181,160

Table 6. Ontology View: Model Performance (Condensed)

Ontology View						
Category	**Model**	**Performance**				
		MR	MRR	Hit@1	Hit@3	Hit@10
Distance	TransE	1323.58	0.0799	0.0186	0.0743	0.2103
	TransR	1804.35	0.0813	0.0208	0.0746	0.2086
	AttE	2038.19	0.2120	0.1302	0.2344	0.3799
	RefE	1417.40	0.1836	0.1020	0.2013	0.3517
	RotE	2174.07	0.2143	0.1343	0.2382	0.3755
	MurE	1684.75	0.2094	0.1279	0.2310	0.3765
Semantic	CP	6658.02	0.1499	0.0693	0.1692	0.3237
	DistMult	6706.65	0.1520	0.0690	0.1747	0.3334
Complex	RotatE	8703.68	0.1061	0.0580	0.1202	0.2022
	ComplEx	9395.01	0.1342	0.0738	0.1504	0.2615
Hyperbolic	AttH	2151.64	0.2087	0.1253	0.2337	0.3788
	RefH	1372.03	0.1801	0.0962	0.1989	0.3522
	RotH	2272.63	0.2095	0.1287	0.2332	0.3722

Table 7. Instance View: Model Performance

Instance View						
Category	**Model**	**Performance**				
		MR	MRR	Hit@1	Hit@3	Hit@10
Distance	TransE	1316.30	0.1171	0.0621	0.1259	0.2194
	TransR	1299.94	0.1233	0.0728	0.1275	0.2218
	AttE	725.61	0.1989	0.1400	0.2099	0.3116
	RefE	792.09	0.1792	0.1233	0.1881	0.2841
	RotE	794.64	0.1812	0.1250	0.1907	0.2871
	MurE	783.26	0.1946	0.1372	0.2050	0.3028
Semantic	CP	1427.38	0.0953	0.0481	0.0983	0.1827
	DistMult	1434.64	0.0968	0.0499	0.0995	0.1832
Complex	RotatE	1178.16	0.2157	0.1410	0.2337	0.3662
	ComplEx	1601.89	0.1859	0.1131	0.2000	0.3335
Hyperbolic	AttH	841.83	0.1813	0.1250	0.1915	0.2872
	RefH	859.59	0.1712	0.1173	0.1787	0.2728
	RotH	874.41	0.1661	0.1112	0.1747	0.2687

[7] https://ctdbase.org/about/legal.jsp

[8] Its sources CGI & PharmGKB are CC0 https://clinicalgenome.org/tools/clingen-website/attribution/

[9] https://www.ncbi.nlm.nih.gov/clinvar/docs/maintenance_use/

[10] https://go.drugbank.com/about

[11] http://geneontology.org/docs/go-citation-policy/

[12] freely accessible for non-commercial use

[13] https://www.kegg.jp/kegg/legal.html

[14] https://www.nlm.nih.gov/databases/download/terms_and_conditions_mesh.html

[15] https://mychem.info/terms

[16] https://mychem.info/terms

[17] https://mygene.info/terms

[18] https://creativecommons.org/licenses/by-sa/4.0/

[19] https://pubmed.ncbi.nlm.nih.gov/

[20] "Terms and Condition" in https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/README.txt

[21] https://db.idrblab.net/ttd/

[22] https://www.nlm.nih.gov/databases/umls.html, https://www.nlm.nih.gov/databases/umls.html

Table 8. Whole View: Model Performance

Whole View						
Category	**Model**	**Performance**				
		MR	MRR	Hit@1	Hit@3	Hit@10
Distance	TransE	1508.11	0.1008	0.0545	0.1063	0.1839
	TransR	1542.63	0.1087	0.0657	0.1108	0.1907
	AttE	805.07	0.1677	0.1119	0.1766	0.2741
	RefE	854.55	0.1543	0.1027	0.1602	0.2513
	RotE	857.33	0.1568	0.1051	0.1629	0.2535
	MurE	846.02	0.1697	0.1154	0.1781	0.2717
Semantic	CP	1594.40	0.0952	0.0483	0.0983	0.1827
	DistMult	1584.33	0.0930	0.0451	0.0965	0.1823
Complex	RotatE	2639.09	0.1818	0.1166	0.1943	0.3128
	ComplEx	3419.68	0.1516	0.0857	0.1627	0.2832
Hyperbolic	AttH	973.13	0.1497	0.0969	0.1571	0.2498
	RefH	1012.54	0.1333	0.0855	0.1372	0.2223
	RotH	1004.64	0.1316	0.0830	0.1358	0.2221

Table 9. Model categorization and scoring functions

Model		**Scoring function** $f(h,r,t)$
Translation	TransE [11]	$-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{1/2}$ where $\mathbf{r}\in\mathbb{R}^k$
	TransR [52]	$-\|\mathbf{M}_r\mathbf{h}+\mathbf{r}-\mathbf{M}_r\mathbf{t}\|_{1/2}$ where $\mathbf{M}_r\in\mathbb{R}^{n\times k}$, $\mathbf{r}\in\mathbb{R}^n$
	RotE [70]	$-\|\boldsymbol{c}_h\circ\boldsymbol{c}_r-\boldsymbol{c}_t\|_{1/2}$ where $\boldsymbol{c}_h,\boldsymbol{c}_r,\boldsymbol{c}_t\in\mathbb{C}^k$; $\circ$ denotes the element-wise product
Complex	ComplEx [73]	$\mathrm{Re}\left(\boldsymbol{c}_h^\top\mathbf{C}_r\bar{\boldsymbol{c}}_t\right)$ where $\mathrm{Re}(c)$ denotes the real part of the complex value $c\in\mathbb{C}$
	RotatE [70]	$-\|\boldsymbol{c}_h\circ\boldsymbol{c}_r-\boldsymbol{c}_t\|_{1/2}$ where $\boldsymbol{c}_h,\boldsymbol{c}_r,\boldsymbol{c}_t\in\mathbb{C}^k$; $\circ$ denotes the element-wise product
Hyperbolic	RefH [15]	$-d_{\mathbb{B}}^{c_r}\left(\mathbf{q}_{\mathrm{Ref}}^H,\mathbf{e}_t^H\right)^2+b_h+b_t$ where $\mathbf{h},\mathbf{t}\in\mathbb{B}_c^d, b_h,b_t\in\mathbb{R}$, $\mathbf{r}\in\mathbb{B}_c^d$, $\mathbf{q}_{\mathrm{Ref}}^H=\mathrm{Ref}(\Theta_r)\mathbf{e}_h^H$
	RotH [15]	$-d_{\mathbb{B}}^{c_r}\left(\mathbf{q}_{\mathrm{Rot}}^H,\mathbf{e}_t^H\right)^2+b_h+b_t$ where $\mathbf{h},\mathbf{t}\in\mathbb{B}_c^d, b_h,b_t\in\mathbb{R}$, $\mathbf{r}\in\mathbb{B}_c^d$, $\mathbf{q}_{\mathrm{Rot}}^H=\mathrm{Rot}(\Theta_r)\mathbf{e}_h^H$
	AttH [15]	$-d_{\mathbb{B}}^{c_r}\left(\mathrm{Att}\left(\mathbf{q}_{\mathrm{Rot}}^H,\mathbf{q}_{\mathrm{Ref}}^H;\mathbf{a}_r\right)\oplus^{c_r}\mathbf{r}_r^H,\mathbf{e}_t^H\right)^2+b_h+b_t$ where $\mathbf{h},\mathbf{t}\in\mathbb{B}_c^d, b_h,b_t\in\mathbb{R}$, $\mathbf{r}\in\mathbb{B}_c^d$

Relation Table in Know2BIO

Out of the 6.18 million edges, there are 108 unique edge types. While most edges (i.e., relations) are between only one pair of biomedical categories, some relations exist across multiple pairs (e.g., the -is_a- edge connects drug classes to drug classes, diseases to diseases, anatomies to anatomies, pathways to pathways, and GO terms to GO terms for the ontology edges). Detailed in Tables 10 & 11, there are 30 unique pairs of biomedical category nodes, with the number of unique relationships between each pair of biomedical categories and the names of relations between them. Compound- compound is the node pair with the highest number of relations, with over 2.9 million edges across two types of relations: '-is-' and '-interacts_with->' indicating an alignment between two identical drugs and interaction between two compounds, respectively. While most pairs of biomedical concepts consist of one or two types of relations, the pair with the largest number of relation types is between protein and compound with 51 different relations, shown separately in Table 11 for practical purposes. These relations describe specifically how a protein interacts with a compound.

Dataset Accessibility and Maintenance

The intended use of this dataset is as a general-use biomedical KG. We note that many other biomedical KGs were constructed with a single use-case in mind and were often assembled in a one-time effort and have not been updated continuously. Source codes used to generate and update this dataset as well as the accompanying software codes to process and model this KG are available at https://anonymous.4open.science/r/Know2BIO. Datasheet describing the dataset and accompanying metadata is

Table 10. Unique Relations Between Entity Types [1]

Head Type	Tail Type	# Type of Relations	# Triple	Relations
Gene	Compound	2	546	-decreases->, -increases->
Disease	Pathway	1	751	-disease_involves->
Pathway	Pathway	2	3025	-pathway_is_parent_of->, isa
Compound	Drug Class	1	5152	-is-
Drug Class	Drug Class	1	5707	isa
Anatomy	Anatomy	2	6299	-is-, isa
Cellular Component	Cellular Component	1	6498	isa
Compound	Reaction	1	11934	-participates_in->
Molecular Function	Molecular Function	1	13747	isa
Reaction	Pathway	1	14925	-involved_in->
Compound	Pathway	3	17401	-compound_participates_in->, -drug_participates_in_pathway->, -drug_participates_in->
Gene	Protein	1	21330	-encodes->
Biological Process	Biological Process	4	64560	-negatively_regulates->, isa, -positively_regulates->, -regulates->
Compound	Disease	1	67715	-treats->
Protein	Molecular Function	4	72032	NOT|enables, NOT|contributes_to, enables, contributes_to
Gene	Pathway	1	80486	-may_participate_in->
Protein	Cellular Component	8	89741	is_active_in, colocalizes_with, NOT|colocalizes_with, NOT|part_of, NOT|is_active_in, located_in, NOT|located_in, part_of
Disease	Disease	4	136406	-diseases_share_variants-, -is-, isa, -diseases_share_genes-
Protein	Biological Process	10	139399	NOT|acts_upstream_of_or_within_negative_effect, acts_upstream_of, acts_upstream_of_or_within_negative_effect, acts_upstream_of_positive_effect, acts_upstream_of_negative_effect, acts_upstream_of_or_within_positive_effect, acts_upstream_of_or_within, NOT|involved_in, NOT|acts_upstream_of_or_within, involved_in
Gene	Disease	2	201336	-not_associated_with-, -associated_with-
Protein	Reaction	5	209254	-output->, -entityFunctionalStatus->, -regulatedBy->, -input->, -catalystActivity->
Gene	Anatomy	2	217166	-overexpressed_in->, -underexpressed_in->
Protein	Protein	1	245958	-ppi-
Protein	Pathway	2	350832	-participates_in->, -may_participate_in->
Compound	Gene	6	487733	increases, -decreases->, -associated_with->, -affects->, -increases->, decreases
Protein	Gene	1	748831	-transcription_factor_targets->
Compound	Compound	2	2902659	-is-, -interacts_with->

also included in the GitHub repository. The licenses for all datasets are detailed in Table 4. We acknowledge that we bear responsibility in case of violation of license and rights for data included in our KG. We release the data available under the respective licenses of the data sources (See Table 9) license publicly; the remainder are available upon request with the appropriate easily-requestable academic credentials from DrugBank. Some resources require free accounts to access and use the data (e.g., UMLS). The source code to obtain the data is released under MIT license and the data are released under the respective license of the data sources. The dataset will be updated periodically as new biomedical knowledge are updated and made available. The dataset is currently not yet released and will be released upon acceptance of the manuscript, through the GitHub repository. The dataset is available in three formats: 1) as raw input files (.csv) detailing individually extracted biomedical knowledge via API and downloads. These files also include intermediate files for mapping between ontologies as well as node features (e.g., text descriptions, sequence data, structure data), and edge weights which were not included in the combined dataset as they were not included in the model evaluation. A folder also contains only the final edges to be used in the KG. 2) a combined KG following the head-relation-tail (h,r,t) convention, as a comma-separated text file. These KGs are released for the ontology view, instance view, and bridge view, as well as a combined whole KG. 3) To facilitate comparison between different KG embedding models, we also release the train, validation, and test split KGs. Long-term

Table 11. Unique Relations Between Entity Types [2]

Head Type	Tail Type	# Type of Relations	# Triple	Relations
Drug	Protein	51	59737	-binder->, -inhibitor->, -translocation_inhibitor->, -drug_targets_protein->, -chelator->, -inhibitory_allosteric_modulator->, -inverse_agonist->, -allosteric_modulator->, -antagonist->, -unknown->, -product_of->, -inactivator->, -cofactor->, -regulator->, -chaperone->, -partial_antagonist->, -other/unknown->, -cleavage->, -inhibits_downstream_inflammation_cascades-> -neutralizer->, -gene_replacement->, -blocker->, -drug_uses_protein_as_carriers-, -partial_agonist->, -incorporation_into_and_destabilization->, -suppressor->, -drug_uses_protein_as_enzymes-, -drug_uses_protein_as_transporters-, -multitarget->, -potentiator->, -inducer->, -binding->, -degradation->, -stimulator->, -antisense_oligonucleotide->, -modulator->, -component_of->, -substrate->, -positive_allosteric_modulator->, -downregulator->, -weak_inhibitor->, -activator->, -other->, -stabilization->, -inhibition_of_synthesis->, -agonist->, -ligand->, -negative_modulator->, -antibody->, -oxidizer->, -nucleotide_exchange_blocker->

preservation of the dataset will be done through versioning as the data are updated and the source codes are run to construct the updated KG. The construction of this KG uses the API available through numerous APIs and biomedical research knowledge sources. Therefore, the source codes to construct the KG may deprecate when these resources update their APIs. However the functionality will be restored upon the next update of the dataset.