Title: Biomedical Discovery through the integrative Biomedical Knowledge Hub (iBKH)

Authors: Chang Su[1#], Yu Hou[2#], Suraj Rajendran[3], Jacqueline R. M. A. Maasch[4], Zehra Abedi[2], Haotan Zhang[5], Zilong Bai[2], Anthony Cuturrufo[6], Winston Guo[7], Fayzan F. Chaudhry[5], Gregory Ghahramani[5], Jian Tang[8], Feixiong Cheng[9,10,11], Yue Li[12], Rui Zhang[13], Jiang Bian[14], Fei Wang[2,15]*

[1]Department of Health Service Administration and Policy, College of Public Health, Temple University, Philadelphia, PA, USA.
[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.
[3]Tri-Institutional Computational Biology & Medicine Program, Cornell University, NY, USA
[4]Department of Computer Science, Cornell Tech, New York, NY, USA
[5]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA.
[6]Computer Science, Cornell University, Ithaca, NY, USA.
[7]Department of Medicine, Weill Cornell Medicine, New York, NY, USA.
[8]Mila-Quebec AI Institute and HEC Montreal, Montreal, Canada.
[9]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA.
[10]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA
[11]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA
[12]School of Computer Science, McGill University, Montreal, Canada.
[13]Department of Surgery, University of Minnesota, Minneapolis, MN, USA.
[14]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, FL, USA.
[15]Lead contact
[#]Equal Contribution
*Corresponding author

Corresponding Author:
Fei Wang,
425 E 61 St. New York City. NY 10065. USA
few2001@med.cornell.edu

Summary
The massive and continuously increasing volume of biomedical knowledge derived from biological experiments or gained from healthcare practices has become an invaluable treasure for biomedicine. The emerging biomedical knowledge graphs (BKGs) provide an efficient and effective way to manage the abundant knowledge in biomedical and life science. In the present study, we harmonized and integrated data from diverse biomedical resources to curate a comprehensive BKG, named the integrative Biomedical Knowledge Hub (iBKH). To facilitate the usage of iBKH in biomedical research, we developed a web-based, easy-to-use, publicly available graphical portal that allows fast, interactive, and visualized knowledge retrieval in iBKH. Furthermore, an efficient and scalable graph learning pipeline was developed for novel knowledge discovery in iBKH. As a proof of concept, we performed our iBKH-based method for computational in silico drug repurposing for Alzheimer's disease. The iBKH is publicly available at: http://ibkh.ai/.

Introduction

Biomedicine is a discipline with enormous volume of highly specialized biomedical knowledge accumulated from biological experiments and healthcare practices. In the past decades, efforts have been drawn to collect and manage the abundant biomedical knowledge and have resulted in diverse biomedical knowledge sources. For example, the biomedical ontologies (Rubin et al., 2008; Smith et al., 2005) store hierarchical relationship-based descriptions for biomedical entities, and the manually curated biomedical knowledge bases (Callahan et al., 2020; Zhu et al., 2019) store biomedical relational data. However, each knowledge source typically focuses on a sub-domain in biomedicine, and hence cannot provide a comprehensive perspective of life science. This hinders the efficient usage of cross-domain biomedical knowledge to provide system-level understanding of human diseases.

At this point, the biomedical knowledge graph (BKG) has become a novel paradigm for better management of the massive volume, sophisticated biomedical knowledge and has attracted significant attentions in recent years(Himmelstein et al., 2017; Nelson et al., 2019; Nicholson and Greene, 2020; Rotmensch et al., 2017; Santos et al., 2022; Sügis et al., 2019). Typically, a BKG is a graph or network that integrates, harmonizes, and stores biomedical knowledge collected from single or multiple expert-derived knowledge sources, where nodes are a set of biomedical entities (e.g., diseases, drugs, genes, biological processes, etc.) and edges between nodes/entities are relations linking the biomedical entities (e.g., drug-treats-disease, disease-associates-gene, drug-interacts-drug, etc.).(Himmelstein *et al.*, 2017; Nicholson and Greene, 2020; Santos *et al.*, 2022; Zhu et al., 2020)

In this context, although efforts have been made to construct BKGs by integrating diverse expert curated knowledge bases (Himmelstein *et al.*, 2017; Li et al., 2020; Santos *et al.*, 2022; Yu et al., 2019; Zhu *et al.*, 2020) or by extracting knowledge from literature using natural language processing (NLP) techniques (Ernst et al., 2015; Percha and Altman, 2018; Yuan et al., 2020), they are not perfect and there remains the space to build a more comprehensive BKG to support advanced biomedical research. In addition, though these BKGs are publicly available, there remains a need for an open, easy-to-use user interface (UI) to fill the gap between the BKG and biomedical researchers and healthcare providers. To this end, this present study proposed a comprehensive BKG, termed the integrative Biomedical Knowledge Hub (iBKH), which was curated by integrating data from 18 high-quality and well-known knowledge sources, including biomedical ontologies, manually curated biomedical knowledge bases, existing BKGs, and NLP-extracted biomedical knowledge sources. To further demonstrate the use of iBKH in biomedical research, we developed a web-based, easy-to-use, intelligent graphical portal that allows fast, interactive knowledge retrieval in iBKH and visualization of the retrieved knowledge.

In addition, we introduced advanced graph learning approaches to the iBKH for computational knowledge discovery. Current graph learning techniques (Mohamed et al., 2021; Su et al., 2020), an emerging branch of machine learning and deep learning that can learn underlying knowledge from graph structure data, have advanced the application of BKG in accelerating novel biomedical knowledge discovery such as drug repurposing (Su et al., 2022; Zhang et al., 2021; Zhou et al., 2020; Zhu *et al.*, 2020) and disease risk gene prioritization(Hu et al., 2021; Peng et al., 2021). In this context, we introduced the advanced graph learning approaches to the iBKH for computational knowledge discovery. Specifically, we designed a knowledge discovery module based on the DGL-KE (Deep Graph Library - Knowledge Embedding) software (Zheng et al., 2020) for efficient and scalable graph learning in iBKH. As a proof of concept, we demonstrated a use case of iBKH, armed with the graph learning algorithms, for in

silico hypothesis generation for Alzheimer's disease (AD) drug repurposing – one of the grand challenges in current biomedical research.

Results

Figure 1 illustrates overall pipeline of the present study, which includes the following modules including: 1) iBKH construction through biomedical knowledge integration, 2) development of graphical portal for fast knowledge retrieval based on iBKH, and 3) iBKH-based computational knowledge discovery through deep graph learning. **Figure 2** illustrates the schema of our BKG, i.e., iBKH. The iBKH is publicly available at: http://ibkh.ai/.

The integrative Biomedical Knowledge Hub (iBKH)

By collecting, harmonizing, and integrating data from 18 publicly available biomedical knowledge sources (see **Table 1**), we curated a comprehensive biomedical knowledge graph, named the integrative Biomedical Knowledge Hub (iBKH). The knowledge sources include biomedical ontologies such as the Brenda Tissue Ontology (Chang et al., 2021), the Cell Ontology (Diehl et al., 2016) the Disease Ontology (Schriml et al., 2012), and the Uberon (Mungall et al., 2012); manually curated biomedical knowledge bases for biomedical entity and relation data such as the Bgee (Bastian et al., 2021), the Comparative Toxicogenomics Database (CTD) (Davis et al., 2019), the DrugBank,(Wishart et al., 2018) the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), the Pharmacogenetics Knowledge Base (PharmGKB) (Hewett et al., 2002), the Reactome (Fabregat et al., 2018), the Side effect resource (SIDER)(Kuhn et al., 2016), and the TISSUE (Palasca et al., 2018); existing BKGs curated by integrating multiple knowledge bases such as the Drug Repurposing Knowledge Graph (DRKG, https://github.com/gnn4dr/DRKG) (Ioannidis et al., 2020), the Hetionet (Himmelstein *et al.*, 2017), the Integrated Dietary Supplement

Knowledge Base (iDISK) (Rizvi et al., 2020), our curated knowledge graph that covers a variety of dietary supplements, including vitamins, herbs, minerals, etc.; and other biomedical sources such as HUGO Gene Nomenclature Committee (HGNC) (Braschi et al., 2019), ChEMBL(Gaulton et al., 2012), and Chemical Entities of Biological Interest (ChEBI) (de Matos et al., 2010). More details of the sources can be found in **Table 1**.

After data management and necessary data cleaning, we integrated data from the diverse sources through biomedical entity term normalization and knowledge integration (more details can be found in the **Method** section). Current version of the resulted iBKH contains a total of 2,384,501 entities of 11 types, including 23,003 anatomy entities, 19,236 disease entities, 37,997 drug entities, 88,376 gene entities (including human and other species), 2,065,015 molecule entities, 1,361 symptom entities, 2,988 pathway entities, 4,251 side-effect entities, 4,101 dietary supplement ingredient (DSI) entities, 137,568 dietary supplement product (DSP) entities, 605 dietary's therapeutic class (TC) entities (see **Figure 2** and **Table 2**). In addition, there are 45 relation types within 18 kinds of entity pairs, including Anatomy-Gene, Drug-Disease, Drug-Drug, Drug-Gene, Disease-Disease, Disease-Gene, Disease-Symptom, Gene-Gene, DSI-Disease, DSI-Symptom, DSI-Drug, DSI-Anatomy, DSI-DSP, DSI-TC, Disease-Pathway, Drug-Pathway, Gene-Pathway and Drug-Side Effect, which means multiple types of relations can exist between a pair of biomedical entities (see **Table 3**). Specifically, 2 types of potential relations can exist between a Anatomy-Gene pair, including 'Expresses' and 'Absent'; 6 relation types between a Drug-Disease pair, such as 'Treats' and 'Effects'; 2 relation types between a Drug-Drug pair including 'Interaction' and 'Resembles'; 10 relation types between a Drug-Gene pair, such as 'Targets', 'Upregulates', and 'Downregulates'; 2 relation types between a Disease-Disease pair including 'Is_A' and 'Resembles'; 5 relation types between a Disease-Gene pair, such as 'Associates', 'Upregulates', and 'Downregulates'; the 'Presents' relation type between a Disease-Symptom pair; and 5 relation types between a Gene-Gene pair, such as

'Covaries', 'Interacts', and 'Regulates'; the 'Has_Adverse Reaction' relation between a DSI-Symptom pair; the 'Is_Effective_For' relation type between a DSI-Disease pair; the 'Interacts' relation type between a DSI-Drug pair; the 'Has_Adverse_Effect_On' relation type between a DSI-Anatomy pair; the 'Has_Ingredient' relation type between a DSI-DSP pair; the 'Has_Therapeutic_Class' relation type between a DSI-TC pair; the 'Reaction' and 'Associates' relation types between a Gene-Pathway pair; the 'Associates' relation between a Disease-Pathway pair; the 'Associates' relation between a Drug-pathway pair; the 'Causes' relation type between Drug-Side Effect pair.

We deployed our iBKH using Neo4j (https://neo4j.com), a robust graph database platform. We also released entity and relation source files of iBKH in CSV (comma-separated values) format, available at: https://github.com/wcm-wanglab/iBKH. Of note, the deployed version of iBKH excluded data from KEGG, as it forbids data redistribution.

An easy-to-use interactive online portal for fast knowledge retrieval

Knowledge retrieval is the most common application scenario for a BKG like iBKH in biomedical research. In contrast to knowledge query in the traditional databases, knowledge retrieval in the iBKH needs to match the logical and structural patterns of entities and relations. This can be done by defining graph-based queries.

To fill the gap between the iBKH and biomedical and clinical researchers to facilitate its usage, we developed a web-based graphical portal that allows users to design graph-based queries for fast knowledge retrieval in a flexible, interactive manner and visualize the retrieved knowledge immediately (see **Figure 1**). Specifically, our portal has two functional modules for knowledge retrieval, i.e., biomedical entity query and path query. First, the biomedical entity query allows to

retrieval information of the queried entity and its one-hop context in the iBKH, i.e., neighboring entities that directly link to the queried entity. **Figure 3a** illustrates an example of exploring biomedical context of the APOE (Apolipoprotein E) gene, which produces APOE protein and is the known major risk gene for AD (Liu et al., 2013; Strittmatter and Roses, 1995). By choosing KEGG and DrugBank in the Source section, we narrow down the query to explore entities that has relations connecting to APOE based on knowledge from the two knowledge sources. Specifically, besides AD, APOE is also associated with diseases including Sea-blue histiocytosis, Hypertriglyceridemia, Hyperlipoproteinemia type iii, and Lipoprotein glomerulopathy, which can be comorbidities of AD. APOE is associated with the AD pathway and cholesterol metabolism pathway that play a role in AD. APOE also has relations with drugs like Zinc medications (Zinc, Zinc sulfate, Zinc chloride, and Zinc acetate) that target APOE to affect progression of AD (Rivers-Auty et al., 2021; Squitti et al., 2020).

In addition, there is also a need for more sophisticated queries to retrieve multi-hop context information of the queried entity, which may help discover inconspicuous but meaningful knowledge from iBKH. **Figure 3b** illustrates an example of discovering drugs that connect to AD through the path $disease - [Associates\_DiG] - gene - [Associats\_DG] - drug,$ where $Associates\_DiG$ and $Associats\_DG$ denote relations in terms of the "association" between a pair of disease and gene and the "association" between a pair of gene and drug, respectively. Such a query path can be generated by iteratively defining entities and relations, combined with constraints, in our portal (see **Figure 3b**). **Figure 3b** shows the retrieved knowledge. For simplicity, we only visualized 100 retrieved triplets (by setting Limit of Triplet as 100 in the portal). Centered on the AD entity, genes that have been associated with AD were first retrieved. Then, drugs that had been associated with these genes were retrieved, which can be considered as potential repurposable drugs for AD treatment. For instance, Cyclophosphamide, a medication used as chemotherapy and to suppress the immune system, is connected to the

AD through the shared neighbor INSR (Insulin Receptor) gene. This is in line with previous evidence that Cyclophosphamide may help reduce cognitive decline in AD (Aisen, 2002).

In silico hypothesis generation for Alzheimer's disease drug repurposing

Another important application scenario for iBKH is the discovery of unknown knowledge, e.g., missing relations among entities, based on the existing, incomplete knowledge graph. In this study, we utilized a computational method for knowledge discovery in iBKH based on the advanced graph learning approaches (Nicholson and Greene, 2020; Su *et al.*, 2020). As a proof of concept, we performed in silico hypothesis generation for AD drug repurposing, i.e., predicting drugs that potentially connect to the AD entity (Fang et al., 2022; Fang et al., 2021; Zeng et al., 2020; Zhou et al., 2021). We utilized knowledge graph embedding (KGE) algorithms to calculate machine-readable embedding vectors for entities and/or relations in iBKH, while preserving the graph structure (Mohamed *et al.*, 2021; Su *et al.*, 2020; Wang et al., 2017), using Deep Graph Library - Knowledge Embedding (DGL-KE) (Zheng *et al.*, 2020). We used four advanced KGE algorithms in DGL-KE including TransE (Bordes et al., 2013), TransR (Lin et al., 2015), ComplEx (Théo et al., 2016), and DistMult (Yang et al., 2015). Then, a possibility score was calculated for each candidate drug entity based on the learned embedding vectors to measure the possibility that the drug can link to the AD via any relation(s). Such analysis has been used to identify repurposable drug candidates for COVID-19 in our previous study (Zeng *et al.*, 2020). More details can be found in the **Method** section and **Figure 1**.

To evaluate performance of our method in predicting repurposable drugs for AD, we considered the FDA (Food and Drug Administration) approved drugs and drugs being tested in clinical trials for AD treatment as the ground truth, which include 10 FDA-approved, 30 in Phase IV trials, 43 in Phase III trials, 95 in Phase II trials, and 47 in Phase I trials. To avoid information leaking in

prediction, all relations between the AD entity and any drug in the grand truth drug list in the iBKH were removed (see **Method** section). **Figure 4** illustrates the performance of our method. Specifically, predictions were made based on embedding vectors produced by four different KGE algorithms, i.e., TransE, TransR, ComplEx, and DistMult, respectively. We also proposed an ensemble model based on the four methods (see **Methods** section). Overall, our methods achieved desirable prediction performances, with an area under the receiver operating characteristic curve (AUC) score over 0.83 for all methods in predicting the FDA approved AD drugs, and an AUC over 0.75 in predicting FDA approved drugs and drugs in Phase IV clinical trials (n=40). In other words, the FDA approved drugs and Phase IV clinical trial drugs for AD rank higher based on our approach. In addition, the ensemble model shows a higher performance (e.g., AUC = 0.9 for FDA approved drugs, AUC = 0.79 for FDA approved plus Phase IV clinical trial drugs for AD) compared to other models, indicating that it may benefit from different KGE algorithms.

Our model can also suggest potential drug candidates for AD, which have not been approved or involved in clinical trials for AD treatment. As a proof of concept, we highlighted the top-10 ranked potential drugs for AD treatment based on the ensemble model and iBKH (see **Table 4**).

First, we found three **Anti-hypertensive drugs** ranking high based on our approach, including *Labetalol* (DrugBank ID: DB00598), *Phenoxybenzamine* (DrugBank ID: DB00925), and *Mibefradil* (DrugBank ID: DB01388). Specifically, Labetalol is a type of β-blockers. There has been evidence suggesting that β-blockers increase brain clearance of these metabolites by enhancing cerebrospinal fluid flow. Recent studies have demonstrated that the use of β-blockers is associated with reduced risk of AD onset (Beaman et al., 2022) and functional decline in AD (Rosenberg et al., 2008). Phenoxybenzamine is an α-blocker, which has been reported to have neuroprotective activity (Rau et al., 2014). Recent drug repurposing studies

have also suggested phenoxybenzamine as a repurposable drug candidate to treat AD (Peng et al., 2020; Williams et al., 2019). Mibefradil is a calcium channel blocker (CCB). Though Mibefradil was withdrawn from the market in 1998 due to harmful interactions with other drugs, our findings may suggest CCB as potential candidate for AD because that calcium dysregulation has been reported to play a role in AD (Bojarski et al., 2008) and CCB has shown multiple beneficial effects cell culture and animal models of AD (Anekonda and Quinn, 2011; Saravanaraman et al., 2014).

Second, we found two **Antipsychotic drugs** as candidates for AD treatment: *Fluphenazine* (DrugBank ID: DB00623) and *Flupentixol* (DrugBank ID: DB00875). Fluphenazine has been reported as a drug candidate in a recent AD drug repurposing study based on integrated network and transcriptome analysis (Zhao et al., 2020b). On the other hand, Flupentixol is a 5-hydroxytryptamine receptor antagonist, which has been reported as potential treatment for cognitive deficiency in AD (Benhamú et al., 2014; Upton et al., 2008).

We also found other candidate drugs for AD, including *Loperamide* (DrugBank ID: DB00836), *Cyproheptadine* (DrugBank ID: DB00434), *Peginterferon alfa-2b* (DrugBank ID: DB00022), *Apomorphine* (DrugBank ID: DB00714), and *Enoxacin* (DrugBank ID: DB00467). In particular, Loperamide is used to treat diarrhea. Previous studies reported that Loperamide targets opioid receptors (DeHaven-Hudkins et al., 1999; Giagnoni et al., 1983), which has been suggested to be potentially linked to AD pathology(Cai and Ratka, 2012). Cyproheptadine belongs to the histamine antagonists, which have been demonstrated to reduce cognitive symptoms in AD (Zlomuzica et al., 2016). Peginterferon alfa-2b is a recombinant interferon, which is used in the treatment of hepatitis B and C, genital warts, and some cancers. Peginterferon alfa-2b has been reported to bind to and activate human type 1 interferon receptors. Such a procedure activates the JAK/STAT pathway, which has been suggested as a potential target for AD (Jain et al.,

2021; Nevado-Holgado et al., 2019). Apomorphine is a dopamine receptor agonist for Parkinson's disease (PD). It can protect against oxidative stress, which plays a role in AD pathology (Perry et al., 2002). Emerging evidence showed that Apomorphine has a significant impact on improving memory function in AD (Himeno et al., 2011; Nakamura et al., 2017). Enoxacin belongs to the fluoroquinolones, which is used for treatment of bacterial infections. A recent study reported that appropriate use of antibiotics with macrolides and fluoroquinolones may decrease the risk of developing AD (Ou et al., 2021).

Discussions

Due to the unparalleled development rate of novel techniques in biomedical research and healthcare, a massive and continuously increasing volume of biomedical knowledge has been produced in the past decades. In addition, the biomedical knowledge captured from different sub-domains of biomedicine is typically stored in different types of databases. These include biomedical ontologies (Rubin *et al.*, 2008; Smith *et al.*, 2005) that provide hierarchical relationship-based descriptions for entities from a specific entity type and manually curated knowledge bases that focuses on a specific sub-domain in biomedicine (Callahan *et al.*, 2020; Zhu *et al.*, 2019). To better use the rich biomedical knowledge while overcoming the massive volume and data heterogeneity, there is the need for harmonizing and integrating the biomedical knowledge from data sources across diverse sub-domains in biomedicine. To this end, we curated a comprehensive BKG, the iBKH, through collecting, managing, and cleaning raw data from diverse sources, creating standardized vocabularies for biomedical entity normalization, as well as knowledge integration. To date, our iBKH has collected biomedical knowledge from 18 sources, including not only the biomedical ontologies and biomedical knowledge bases, based on which most existing BKGs have been built, but also existing BKGs that have integrated diverse sources. In addition to the general entity types that are commonly studied in

biomedicine, such as genes, diseases, drugs, pathways, etc., our iBKH also involves our previously curated dietary supplement knowledge base, the iDISK (Rizvi *et al.*, 2020). Research studies have demonstrated that the dietary supplements play a role in human diseases, such as AD (Luchsinger and Mayeux, 2004; Luchsinger et al., 2007), cancers (Williams and Hord, 2005), diabetes (van Dam et al., 2002), etc. We believe that involvement of the dietary supplement knowledge will provide complementary knowledge for better human healthcare.

We deployed the iBKH publicly available in both tabular format (CSV files) and Neo4j, which allows fast knowledge retrieval through creating Cypher queries. Though Cypher, inspired by the SQL, is relatively easy to learn, there remains a gap between the iBKH in Neo4j and biomedical researchers and healthcare providers, as knowledge retrieval in a graph needs to define the queries by matching the logical and structural patterns of entities and relations, which are more complex than the SQL queries. As a result, we developed a web-based graphical portal, which allows users to design the desired graph query using a graphical UI. The query is translated to Cypher query in the back end for fast knowledge retrieval and the retrieved knowledge (typically a sub-graph from the iBKH) is visualized in the portal immediately. In this way, the iBKH is more user-friendly, providing users who have no Cypher programming experience the large flexibility to implement efficient biomedical knowledge retrieval.

In addition, we implemented the advanced graph learning approaches to the iBKH for novel biomedical knowledge discovery. The graph learning approaches is a branch of machine learning and artificial intelligence (AI), which devote to building learning algorithms to model graph structure for discovering unobserved knowledge. In this study, we utilized the DGL-KE (Deep Graph Library - Knowledge Embedding) (Zheng *et al.*, 2020), a Python-based implementation for advanced knowledge graph embedding algorithms, for efficient and scalable graph learning in iBKH. As a proof of concept, we demonstrated a use case of iBKH, armed with

the graph learning algorithms, for in silico hypothesis generation for AD drug repurposing. We not only observed good performance of our method, using the FDA-approved drugs and clinical trial drugs for AD as ground truth, but also identified repurposable drugs for AD treatment. By manual literature review, we found evidence supporting potential of the identified candidate drugs to treat AD. More importantly, our iBKH-based knowledge discovery pipeline is flexible and feasible, and can be applied to more diseases of interest beyond AD, by predicting potential relations between the disease of interest and drugs in iBKH. Our pipeline can also adapt to other biomedical application scenarios, such as prioritizing risk genes of disease (gene-disease relation prediction), predicting candidate target protein for drugs (drug-gene relation prediction), identifying potential drug-drug interactions (drug-drug relation prediction), etc.

Limitations of the study

Our iBKH has a few limitations. First, the procedures of constructing and curating iBKH rely on sophisticated efforts of raw data file extraction and pre-processing, data annotation, as well as terminology normalization, which may lead to **incorrectness**, referring to facts in the iBKH that is inconsistent with real-world evidence. To address this, we utilized the well-designed biomedical vocabularies such as the Unified Medical Language System (UMLS) to enhance entity term normalization, which can help reduce the risk of errors caused by the ambiguous biomedical entities. We also performed manual review to reduce incorrectness. More specifically, the integrated file for each entity type or relation type underwent multiple rounds of manual review based on random sampling with replacement. Even so, due to the massive volume of iBKH, there remains the need for a more efficient way to address incorrect facts in iBKH. Graph learning algorithms for knowledge graph refinement is a potential solution in this context. For instance, our early effort in graph learning-based knowledge graph refinement could be extended to address this issue (Zhao *et al.*, 2020b).

Another issue is knowledge **incompleteness**. We built the iBKH by collecting and integrating data from diverse sources. It includes knowledge in a board range of sub-domains of human health. However, incompleteness is still inevitable. On one hand, abundant knowledge sources have been online available and there are good sources that are not involved in iBKH yet. On the other hand, there remains massive biomedical knowledge that has not yet been discovered or is deep buried in the noisy biomedical and health data and literature. In this context, some studies have been focused on deriving knowledge from biomedical literature (Xu et al., 2013; Zhang et al., 2018; Zhao et al., 2021) or human healthcare data such as the EHR (electronic health records) (Chen et al., 2019; Rotmensch *et al.*, 2017). The derived knowledge could be a good complement for our iBKH. In addition, the use of graph learning algorithm to discover hidden knowledge based on the existing iBKH graph structure is another solution and needs more attention in our future work.

Like most existing BKGs, e.g., Hetionet and CKG, our iBKH focuses on the general biomedical knowledge. For the sake of precision medicine on some specific human diseases or health conditions, there is the need for more fine-grained knowledge with a specific focus on them. For instance, COVID-KG (Wang et al., 2020) included biomedical knowledge with a specific focus on COVID-19; KGHC (Li *et al.*, 2020) is a knowledge graph constructed focusing on addressing hepatocellular carcinoma. Following this idea, we will adapt our iBKH to address problems in specific diseases and health conditions like AD, Parkinson's disease, and mental illness. For example, we plan to collect the fine-grained data, such as genotype-phenotype associations and brain region atrophy-phenotype associations and incorporate them to enrich iBKH, for the specific usage of these diseases.

Last, there is the need of further validation for the discovered novel knowledge from iBKH. To this end, our future work will also focus on knowledge validation by leveraging advanced data

science techniques. On one hand, we plan to build a knowledge validation system based on biomedical text mining (Zhao *et al.*, 2021). For instance, leveraging our previous biomedical literature retrieval/matching algorithms (Zhao et al., 2020a; Zhao et al., 2019), we will be able to identify evidence from massive biomedical literature resource, supporting the identified novel triplets (i.e., knowledge) in iBKH. We plan to add this new functionality to iBKH portal. On the other hand, for drug repurposing hypothesis generation, we will validate treatment efficiency of the identified repurposable drug candidates for target disease, such as AD, using our computational clinical trial emulation approach (Zang et al., 2022) based on real-world clinical data.

Author Contributions

FW for conceptualization, investigation, writing, reviewing, and editing of the manuscript. CS for investigation, drafting, editing, and reviewing manuscript. YH led the effort on data preparation, knowledge graph construction, data analysis and web interface implementation. SR, JM, ZA for improving data standardization and organization, efficiency of user interface, and language of the manuscript. HZ, FFC, and GG for data collection and data preparation. AC for knowledge graph embedding implementation. ZB for critical discussion on constructing iBKH. WG for knowledge graph quality check. JT and YL for critical discussion on knowledge graph embedding algorithms. FC, RZ and JB for discussion, design, and interpretation of the case study on AD. All authors have given approval to the final version of the manuscript.

Declaration of Interest

The authors declare no competing interests.

Figures

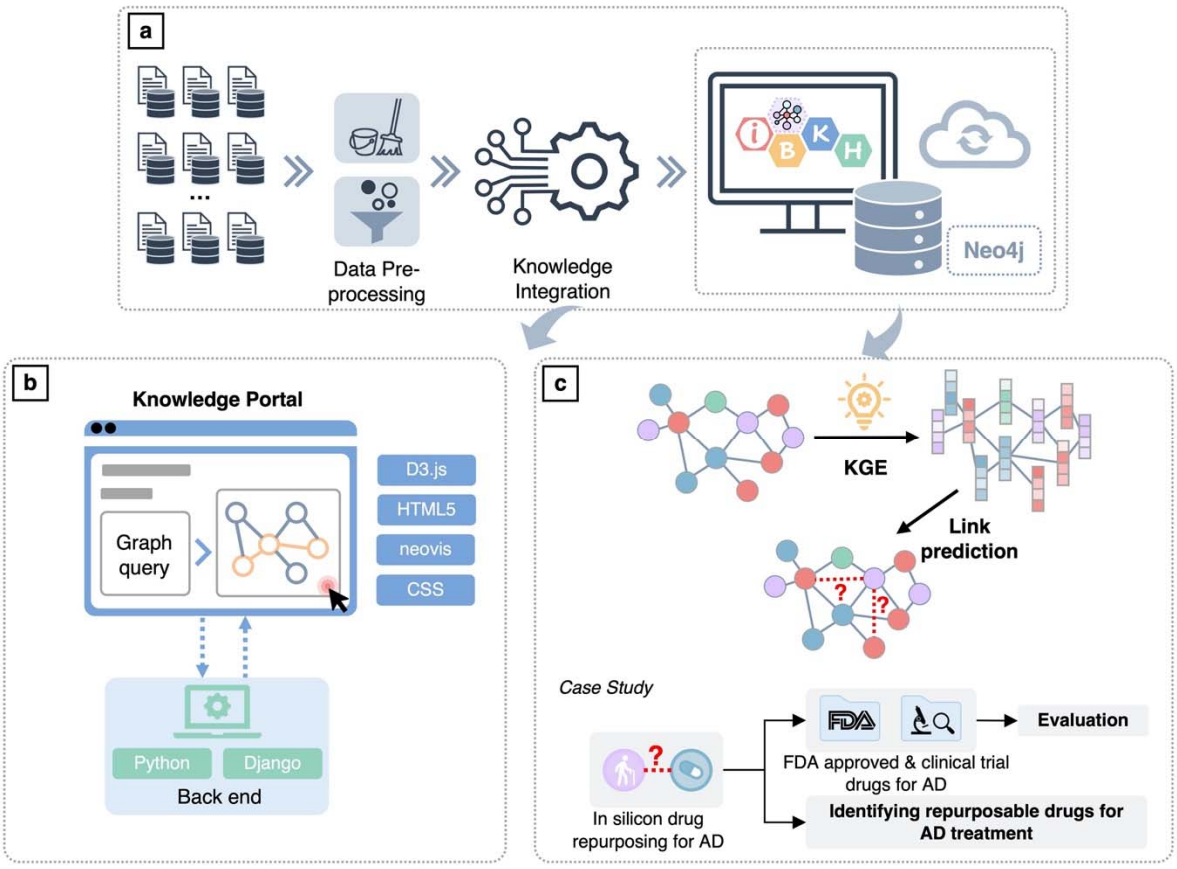

Figure 1. An illustration of study pipeline. a. Steps for curating iBKH. We first collected data from diverse biomedical data sources. Next, necessary data pre-processing, such as data cleaning and data filtering were performed. After that, knowledge from diverse sources were integrated to build an integrative knowledge graph, i.e., iBKH, which was deployed using Neo4j graph database. **b.** A web-based, easy-to-use graphical portal was developed for fast knowledge retrieval. **c.** A graph learning module was introduced to iBKH for novel knowledge discovery. Specifically, knowledge graph embedding was conducted to learn compressed vector representations for entities and relations in iBKH, which were further used for link prediction. As a proof of concept, we performed in silicon drug repurposing for Alzheimer's disease.

Abbreviations: AD = Alzheimer's disease; CSS = Cascading Style Sheets; HTML5 = HyperText Markup Language Version 5; iBKH = integrative Biomedical Knowledge Hub; KGE = knowledge graph embedding.

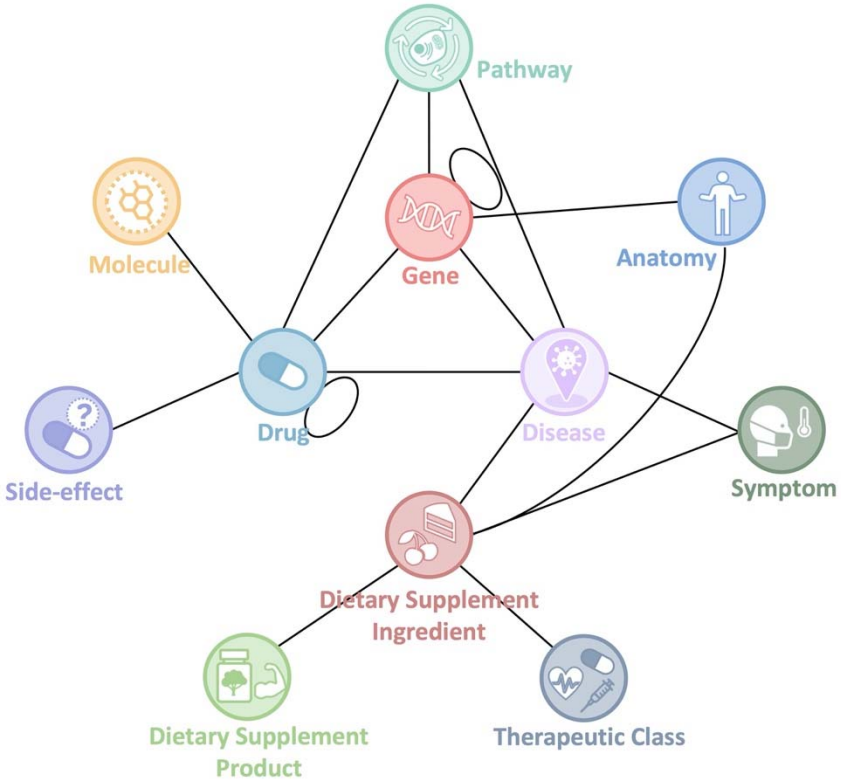

Figure 2. Schema of iBKH. Each circle denotes an entity type, and each link denotes a meta relation between a pair of entities. Of note, a meta relation can represent multiple types of relations between a specific pair of entities. For example, five potential relations including ‘Associates’, ‘Downregulates’, ‘Upregulates’, ‘Inferred_Relation’, ‘Text_Semantic_Relation’ can exist between a pair of disease and gene entities.

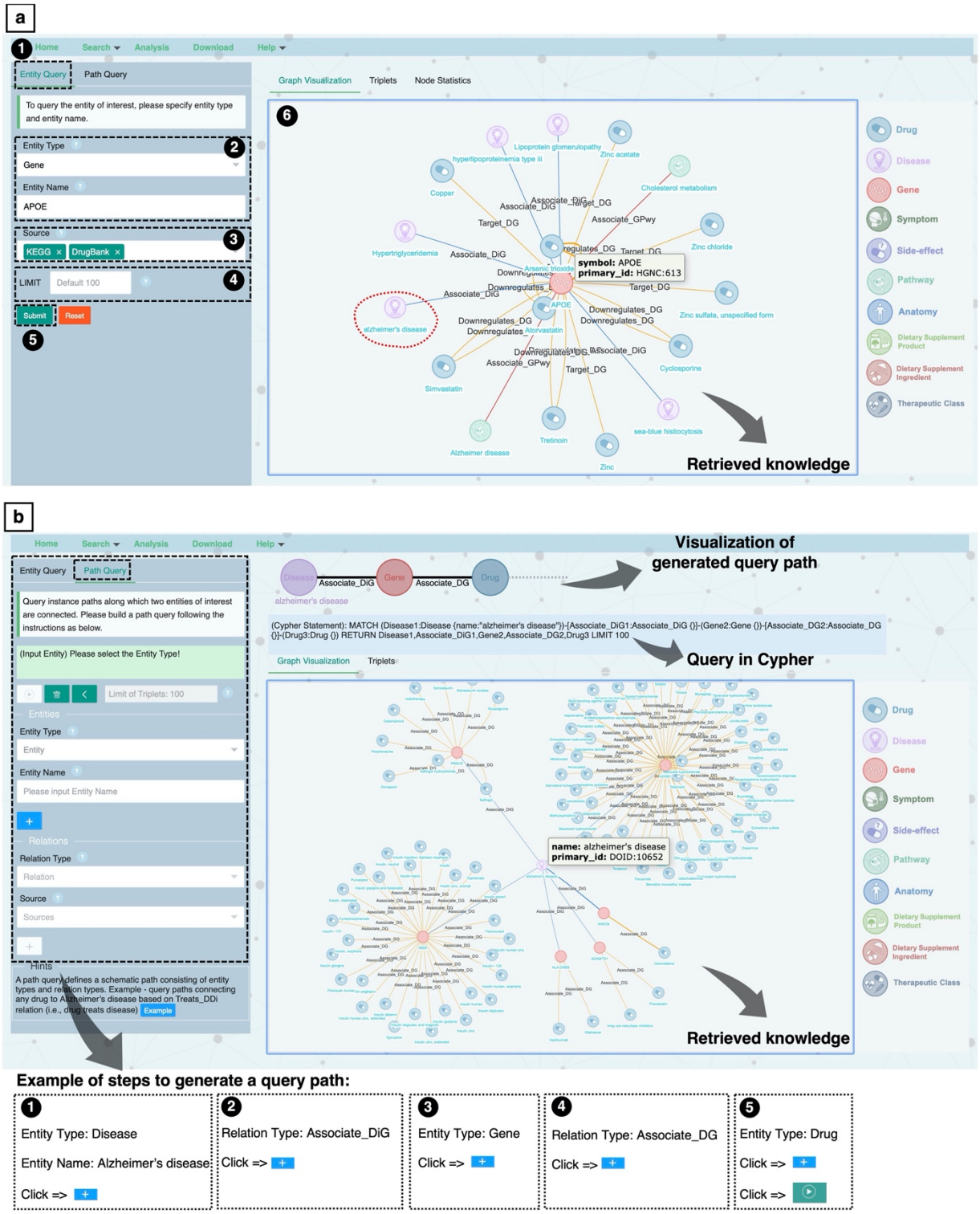

Figure 3. Examples of knowledge retrieval. a. An example of entity query – retrieving neighborhood context of APOE (Apolipoprotein E) gene in iBKH. **b.** An example of path query, retrieving drugs that connect to Alzheimer's disease through the path $disease - [Associates\_DiG] - gene - [Associats\_DG] - drug$, where $Associates\_DiG$ and

$Associats\_DG$ denote relation types in terms of the association between a pair of disease and gene as well as the association between a gene and a drug.

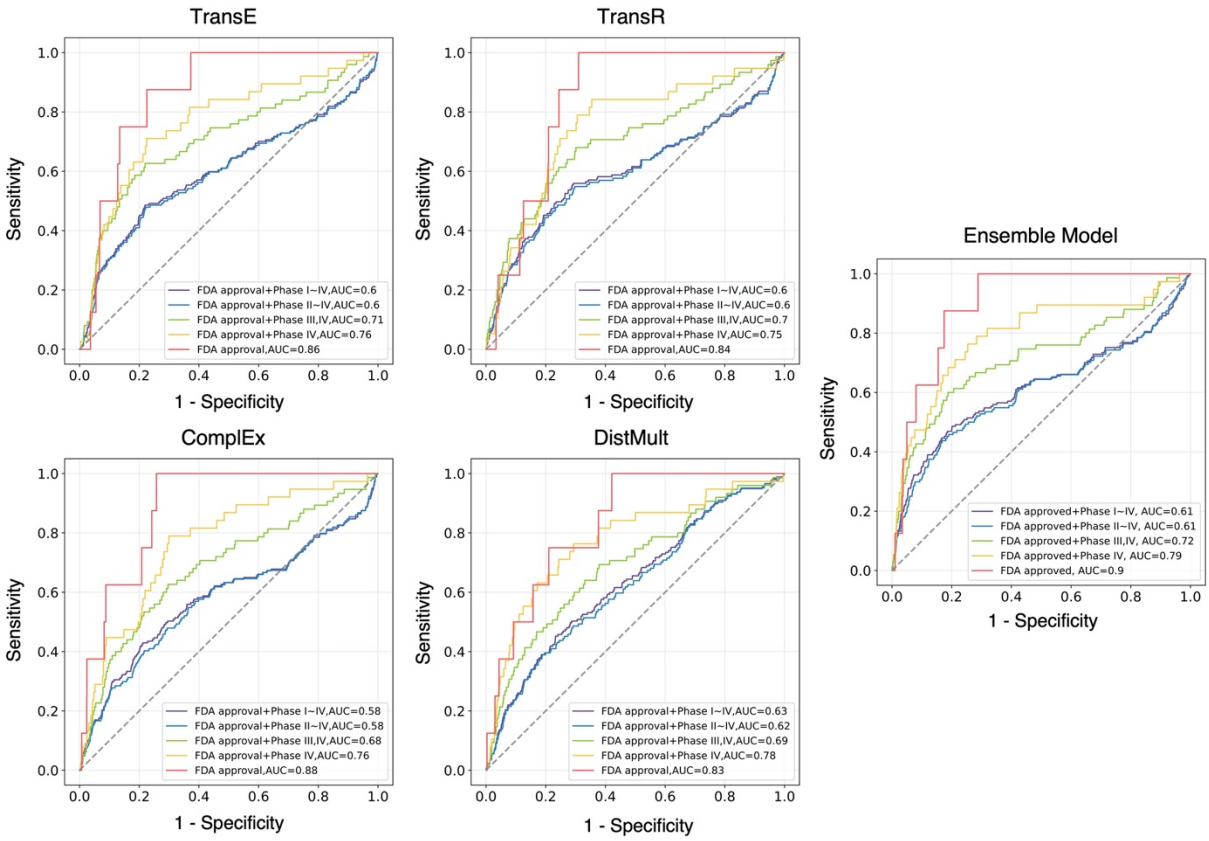

Figure 4. Model performance of in silicon Alzheimer's disease drug repurposing. We used the FDA approved and clinical trial drugs for Alzheimer's disease as ground truth.

Abbreviations: AUC = area under the receiver operating characteristic curve; FDA = Food and Drug Administration.

Table 1. Data sources integrated for constructing iBKH

Source	Description	Entity		Relation		URL	License
		Types	Number	Types	Number		
Bgee (Bastian *et al.*, 2021)	A database for retrieval and comparison of gene expression patterns across multiple animal species.	Anatomy, Gene	60,072	Anatomy-Express Present-Gene, Anatomy-Express Absent-Gene	11,731,369	https://bgee.org/	https://creativecommons.org/publicdomain/zero/1.0/
Brenda Tissue Ontology (Chang *et al.*, 2021)	A tissue-specific ontology.	Tissue (Anatomy)	6,478	-	-	https://www.brenda-enzymes.org/index.php	https://creativecommons.org/licenses/by/4.0/
Cell Ontology (Diehl *et al.*, 2016)	A structured controlled vocabulary for cell types in animals.	Cells (Anatomy)	2,200	-	-	http://obofoundry.org/ontology/cl.html	https://creativecommons.org/licenses/by/4.0/
Comparative Toxicogenomics Database (CTD) (Davis *et al.*, 2019)	A knowledge base that relates toxicological information for chemicals, genes, phenotypes, and diseases, as well as literature-based and manually curated interactions	Disease, Gene, Drug, Pathway	73,922	Chemical-Gene, Chemical-Disease, Chemical-Pathway, Gene-Disease, Gene-Pathway, Disease-Pathway	38,344,568	http://ctdbase.org/	Confirmed via email.
ChEMBL(Gaulton *et al.*, 2012)	A manually curated database of bioactive molecules with drug-like properties.	Molecular	1,940,733	-	-	https://www.ebi.ac.uk/chembl/	https://creativecommons.org/licenses/by-sa/3.0/
Chemical Entities of Biological Interest (ChEBI) (de Matos *et al.*, 2010)	A freely available dictionary of molecular entities focused on 'small' chemical compounds	Molecular	155,342	-	-	https://www.ebi.ac.uk/chebi/init.do	https://creativecommons.org/licenses/by/4.0/
Drug Repurposing Knowledge Graph (DRKG) (Ioannidis *et al.*, 2020)	A biological knowledge graph.	Anatomy, Pathway, Compound (Drug), Disease, Gene, Molecular function, Pathway, Pharmacologic class, Side effect, Symptom	97,238	Gene-Gene, Compound-Gene, Disease-Gene, Atc-Compound, Compound-Compound, Compound-Disease, Gene-Tax, Biological process-Gene, Disease-Symptom, Anatomy-Disease, Disease-Disease, Anatomy-Gene, Gene-Molecular function, Compound-Pharmacologic class, Cellular component-Gene, Gene-Pathway, Compound-Side effect	5,874,261	https://github.com/gnn4dr/DRKG	https://www.apache.org/licenses/LICENSE-2.0

Disease Ontology (Schriml *et al.*, 2012)	Standardized ontology for human disease.	Disease	10,648	-	-	https://disease-ontology.org/	https://creativecommons.org/publicdomain/zero/1.0/
DrugBank (Wishart *et al.*, 2018)	A web-enabled database containing comprehensive molecular information about drugs, their mechanisms, their interactions, and their targets.	Drug	15,128	Drug-Target, Drug-Enzyme, Drug-Carrier, Drug-Transporter	28,014	https://go.drugbank.com/	http://creativecommons.org/licenses/by-nc/4.0/
Hetionet (Himmelstein *et al.*, 2017)	A biomedical knowledge graph for drug repurposing.	Anatomy, Biological process, Cellular component, Compound (Drug), Disease, Gene, Molecular function, Pathway, Pharmacologic class, Side effect, Symptom	47,031	Anatomy–downregulates–Gene, Anatomy–expresses–Gene, Anatomy–upregulates–Gene, Compound–binds–Gene, Compound–causes–Side Effect, Compound–downregulates–Gene, Compound–palliates–Disease, Compound–resembles–Compound, Compound–treats–Disease, Compound–upregulates–Gene, Disease–associates–Gene, Disease–downregulates–Gene, Disease–localizes–Anatomy, Disease–presents–Symptom, Disease–resembles–Disease, Disease–upregulates–Gene, Gene–covaries–Gene, Gene–interacts–Gene, Gene–participates–Biological Process, Gene–participates–Cellular Component, Gene–participates–Molecular Function, Gene–participates–Pathway, Gene–regulates–Gene, Pharmacologic Class–includes–Compound	2,250,197	https://github.com/hetio/hetionet	https://creativecommons.org/publicdomain/zero/1.0/
HUGO Gene Nomenclature Committee (HGNC) (Braschi *et al.*, 2019)	The resource for approved human gene nomenclature	Gene	41,439	-	-	https://www.genenames.org/	No restriction.
Integrated Dietary Supplement Knowledge Base (iDISK) (Rizvi *et al.*, 2020)	Our curated knowledge graph that covers a variety of dietary supplements, including vitamins, herbs, minerals,	Dietary Supplement Ingredient, Dietary Supplement Product, Disease, Drug, Anatomy, Symptom, Therapeutic	144,536	DSI-Anatomy, DSI-Symptom, DSI-Disease, DSI-Drug, DSI-DSP, DSI-TC	705,075	https://conservancy.umn.edu/handle/11299/204783	Our copyright. https://creativecommons.org/licenses/by-sa/3.0/us/

	etc.	Class					
Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000)	A biomedical knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information.	Drug, Disease, Gene, Pathway	42,181	Drug-Gene, Disease-Gene, Gene-Pathway, Drug-Disease, Drug-Pathway, Disease-Pathway	65,505	https://www.kegg.jp/	KEGG forbids data redistribution. The deployed version of iBKH excluded KEGG data.
Pharmacogenetics Knowledge Base (PharmGKB) (Hewett *et al.*, 2002)	A biomedical knowledge base containing genomic, phenotype and clinical information collected from ongoing pharmacogenetic studies.	Genes, Variant, Drug, Phenotype	43,112	Disease-Gene, Drug/Chemical-Gene, Gene-Gene, Gene-Variant, Disease-Variant, Drug/Chemical-Variant	61,616	https://www.pharmgkb.org/	https://creativecommons.org/licenses/by-sa/4.0/
Reactome (Fabregat *et al.*, 2018)	A knowledge base of molecular details of signal transduction, transport, DNA replication, metabolism, and other cellular processes.	Genes, Pathways (H. sapiens)	13,589	Gene-Pathway	13,732	https://reactome.org/	https://creativecommons.org/licenses/by/4.0/
Side effect resource (SIDER) (Kuhn *et al.*, 2016)	A data resource of public information on drug side effects.	Drugs, Side effects	5,681	Drug-Side effect	163,206	http://sideeffects.embl.de/	https://creativecommons.org/licenses/by-nc-sa/4.0/
TISSUE (Palasca *et al.*, 2018)	A public resource that integrates evidence on tissue expression from manually curated literature, proteomics and transcriptomics screens, and automatic text mining.	Genes, Tissues	26,260	Tissue-Express-Gene	6,788,697	https://tissues.jensenlab.org/	https://creativecommons.org/licenses/by/4.0/
Uberon (Mungall *et al.*, 2012)	A cross-species anatomy ontology.	Anatomy	14,944	-	-	https://www.ebi.ac.uk/ols/ontologies/uberon	http://creativecommons.org/licenses/by/3.0/

Table 2. Statistics of biomedical entities in iBKH

Entity Type	Number	Included Identifiers[1]
Anatomy	23,003	Uberon ID, BTO ID, MeSH ID, Cell Ontology ID
Disease	19,236	Disease Ontology ID, KEGG ID, PharmGKB ID, MeSH ID, OMIM ID
Drug	37,997	DrugBank ID, KEGG ID, PharmGKB ID, MeSH ID
Gene	88,376	HGNC ID, NCBI ID, PharmGKB ID
Molecule	2,065,015	CHEMBL ID, CHEBI ID
Symptom	1,361	MeSH ID
Pathway	2,988	Reactome ID, KEGG ID, Gene Ontology ID
Side-effect	4,251	UMLS CUI
Dietary Supplement Ingredient	4,101	iDISK ID
Dietary Supplement Product	137,568	iDISK ID
(Dietary) Therapeutic Class	605	iDISK ID, UMLS CUI

[1] The identifiers used for entity term normalization.
Abbreviations: BTO = BRENDA Tissue Ontology; ChEBI = Chemical Entities of Biological Interest; HGNC = HUGO Gene Nomenclature Committee; ID = identifier; KEGG = Kyoto encyclopedia of genes and genomes; iDISK = integrated dietary supplement knowledge base; MeSH = Medical Subject Headings; NCBI = National Center for Biotechnology Information; OMIM = Online Mendelian Inheritance in Man; UMLS CUI= Unified Medical Language System - Concept Unique Identifiers.

Table 3. Statistics of relations among entities in iBKH

Entity pair	Relation type	Number of relations of the specific type	Total Number
Anatomy-gene relation	Anatomy-Expresses-Gene	10,388,168	12,171,021
	Anatomy-Absent-Gene	2,837,741	
Anatomy-DSI relation	DSI-Has_Adverse_Effect_On-Anatomy	3,121	4,334
Drug-disease relation	Drug-Palliates-Disease	390	2,717,947
	Drug-Treats-Disease	5,492	
	Drug-Effects-Disease	5,136	
	Drug-Associates -Disease	96,458	
	Drug-Inferred_Relation-Disease	2,589,522	
	Drug-Text_Semantic_Relation-Disease	50,653	
Drug-Drug	Drug-Interacts-Drug	2,682,157	2,684,682
	Drug-Resembles -Drug	6,486	
Drug-Gene	Drug-Targets-Gene	16,518	1,303,747
	Drug-Transporter-Gene	3,066	
	Drug-Enzyme-Gene	5,241	
	Drug-Carrier-Gene	853	
	Drug-Downregulates-Gene	66,994	
	Drug-Upregulates-Gene	72,361	
	Drug-Associates-Gene	19,434	
	Drug-Binds-Gene	11,571	
	Drug-Interacts-Gene	1,181,492	
	Drug-Text_Semantic_Relation -Gene	68,429	
Drug-Pathway	Drug-Associates-Pathway	3,231	3,231
Drug-Side effect	Drug-Causes-side-effect	163,206	163,206
Drug-molecule	Molecule-Is_A-Drug	8,757	8,757
Drug-DSI	DSI-Interacts-Drug	3,057	3,057
Disease-Disease	Disease-Is_A-Disease	10,529	11,072
	Disease-Resembles-Disease	543	
Disease-Gene	Disease-Associates-Gene	47,965	27,538,774
	Disease-Downregulates-Gene	7,623	
	Disease-Upregulates -Gene	7,731	
	Disease-Inferred_Relation-Gene	27,454,631	
	Disease-Text_Semantic_Relation -Gene	94,759	
Disease-Symptom	Disease-Presents-Symptom	3,357	3,357
Disease-Pathway	Disease-Associates-Pathway	1,941	1,941

Disease-DSI relation	DSI-Is_Effective_For-Disease	5,134	5,134
Gene-Gene	Gene-Covaries-Gene	61,690	735,156
	Gene-Interacts-Gene	147,164	
	Gene-Regulates-Gene	265,672	
	Gene-Associates-Gene	2,602	
	Gene-Text_Semantic_Relation -Gene	301,752	
Gene-Pathway	Gene-Reaction-Pathway	118,480	152,243
	Gene-Associates-Pathway	47,742	
Symptom-DSI	DSI-Has_Adverse_Reaction-Symptom	2,093	2,093
DSI-DSP	DSP-Has_ingredient-DSI	689,297	689,297
DSI-TC	DSI-Has_therapeutic_class-TC	5,430	5,430

Abbreviations: DSI = Dietary Supplement Ingredient; DSP = Dietary Supplement Product; TC = Therapeutic Class

Table 4. List of the top ten drugs repurposable for Alzheimer's disease treatment

Rank	DrugBank ID	Drug Name	Category	Description	Notes
1	DB00836	Loperamide	Diarrhea medication	Loperamide is used to treat diarrhea. It is often used for this purpose in inflammatory bowel disease.	Loperamide targets opioid receptors (DeHaven-Hudkins *et al.*, 1999; Giagnoni *et al.*, 1983), which has been suggested to be potentially linked to AD pathology(Cai and Ratka, 2012).
2	DB00598	Labetalol	Anti-hypertensive drug, β-blocker	Labetalol is one of the medications called β-blockers, which is used to treat cardiovascular diseases like hypertension.	There has been evidence suggesting that β-blockers increase brain clearance of these metabolites by enhancing CSF flow. Recent studies have demonstrated that the use of β-blockers is associated with reduced risk of AD onset (Beaman *et al.*, 2022) and functional decline in AD (Rosenberg *et al.*, 2008).
3	DB00925	Phenoxybenzamine	Anti-hypertensive drug, α-blocker	Phenoxybenzamine is an α-blocker for treating hypertension, specifically that caused by pheochromocytoma.	Phenoxybenzamine has been reported to have neuroprotective activity(Rau *et al.*, 2014). Recent drug repurposing studies have also suggested phenoxybenzamine as repurposable drug candidate to treat AD (Peng *et al.*, 2020; Williams *et al.*, 2019)..
4	DB01388	Mibefradil	Calcium channel blocker (CCB)	Mibefradil is CCB, which was used for the treatment of hypertension and chronic angina pectoris. Mibefradil was withdrawn from the market in 1998 due to potentially harmful interactions with other drugs.	Previous studies have demonstrated that calcium dysregulation plays an important role in AD (Bojarski *et al.*, 2008). Though the usefulness of CCBs in AD remains controversial, it has shown multiple beneficial effects cell culture and animal models of AD (Anekonda and Quinn, 2011; Saravanaraman *et al.*, 2014).
5	DB00434	Cyproheptadine	Antihistamine	Cyproheptadine is used in the treatment of allergic symptoms.	Cyproheptadine is a histamine antagonist, which has been demonstrated to reduce cognitive symptoms in AD (Zlomuzica *et al.*, 2016).
6	DB00022	Peginterferon alfa-2b	Recombinant interferon	Peginterferon alfa-2b is used in the treatment of hepatitis B and C, genital warts, and some cancers	Peginterferon alfa-2b binds to and activates human type 1 interferon receptors, activating the JAK/STAT pathway, which has been suggested as a potential target for AD (Jain *et al.*, 2021; Nevado-Holgado *et al.*, 2019)..
7	DB00714	Apomorphine	Dopaminergic agonist	Apomorphineis a type of dopaminergic agonist medication used for Parkinson's disease (PD)	Apomorphine is a dopamine receptor agonist for Parkinson disease and also protects against oxidative stress, which plays a role in AD (Perry *et*

					al., 2002). Emerging evidence showed that Apomorphine has a significant impact on improving memory function in AD (Himeno *et al.*, 2011; Nakamura *et al.*, 2017).
8	DB00623	Fluphenazine	Antipsychotic	Fluphenazine is a phenothiazine antipsychotic medication used for treatment of psychotic disorders.	Fluphenazine is reported as a drug candidate in a recent AD drug repurposing study based on integrated network and transcriptome analysis (Peng *et al.*, 2020).
9	DB00875	Flupentixol	Antipsychotic drug	Flupentixol is a thioxanthene neuroleptic used to treat psychotic disorders such as schizophrenia and depression.	Flupentixol is a 5-hydroxytryptamine receptor antagonist which has been reported as potential treatment for cognitive deficiency in AD (Benhamú *et al.*, 2014; Upton *et al.*, 2008).
10	DB00467	Enoxacin	Fluoroquinolones	Enoxacin is a fluoroquinolone used for treatment of bacterial infections.	A recent study reported that appropriate use of antibiotics with macrolides and fluoroquinolones may decrease the risk of developing AD (Ou *et al.*, 2021).

STAR*METHODS

Resource Availability

Lead contact

Further information should be directed to and will be fulfilled by the lead contact, Dr. Fei Wang, (few2001@med.cornell.edu)

Materials Availability

- The harmonized entity and relation source files for iBKH in CSV (comma-separated values) format are publicly available online at https://github.com/wcm-wanglab/iBKH/tree/main/iBKH.
- The iBKH online portal is publicly available at http://ibkh.ai/.

The deployed version of iBKH excluded data from KEGG, as it forbids data redistribution.

Data and Code Availability

- This paper integrates publicly available biomedical knowledge bases. These accession URLs for the knowledge bases are listed in the key resources table.
- The computer codes for iBKH construction and iBKH-based knowledge discovery are publicly available online at https://github.com/wcm-wanglab/iBKH/tree/main/Codes.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Overview

Our ultimate goal was to build a biomedical knowledge graph via comprehensively incorporating biomedical knowledge as much as possible. To date, we have collected and integrated 18 publicly available data sources, harmonized and consolidated them into a comprehensive data compendium. Details of the used data sources were listed in **Table 1**.

Raw data processing

Given the data sources, the first step was to pre-process the raw files of them and extract knowledge, including entity information and relation information. Generally, the databases release their raw data files in various formats, such as comma-separated values (CSV), tab-separated values (TSV), TXT, EXCEL tablet, Hypertext Markup Language (HTML), Resource Description Framework (RDF), and Web Ontology Language (OWL). To address this, for each database, we parsed the raw files and extracted structured data, i.e., the descriptive files for each type of biomedical entity and the files of each type of relation. Such procedure varies by databases or even by files within the same database.

Term harmonization

To integrate data from diverse sources, there is a need for harmonizing the entity terms. To achieve this, we utilized a greedy strategy. Specifically, for a specific entity type, we first chose a database to initialize the entity vocabulary. Next, we built a linkage pool, containing multiple identifiers of the given entity type, to map and integrate entities from all databases to enrich the entity vocabulary one by one.

For **gene** entity type, we used the HUGO Gene Nomenclature Committee (HGNC) gene repository (Braschi *et al.*, 2019) as the initial vocabulary of gene entities, as it defines a standard nomenclature for human the genes. The linkage pool for normalization included HGNC IDs, HGNC symbols, and National Center for Biotechnology Information (NCBI) IDs.

For **drug** entity type, we initialized our vocabulary using DrugBank (Wishart *et al.*, 2018) as it provides the up-to-date list of approved drugs and investigational drugs under clinical trials. The linkage pool for drug entity normalization included DrubBank IDs, Medical Subject Heading (MeSH) terms, MeSH term IDs, Unified Medical Language System (UMLS)(Bodenreider, 2004) Concept Unique Identifiers (CUIs), and the drug names in UMLS.

For **molecule** entity type, we used the ChEMBL(Gaulton *et al.*, 2012), a manually curated database of molecules with drug properties, for initializing the vocabulary. The linkage pool for the molecule entities normalization included ChEMBL IDs and International Chemical Identifier (InChi).

For **Side-Effect** entity type, we collected the side-effect entities from the SIDER(Kuhn *et al.*, 2016) and described them by using the UMLS CUIs.

For **disease** entity type, we used the Disease Ontology (Schriml *et al.*, 2012) for initializing the vocabulary, as it is a structured database of diseases based on etiological classification. The linkage pool we used for the disease entity normalization included Disease Ontology IDs, MeSH terms, MeSH term IDs, UMLS CUIs, and the disease names in UMLS.

For **symptom** entity type, we collected the symptom entities from the Hetionet (Himmelstein *et al.*, 2017) and integrated Dietary Supplements Knowledge (iDISK) (Rizvi *et al.*, 2020), and

described them by using the MeSH term and MeSH term ID. We used UMLS CUI as the linkage for symptom entities normalization.

For **Pathway** entity type, we used the Reactome (Fabregat *et al.*, 2018), a manually curated and peer-reviewed pathway database, for initializing the vocabulary. The linkage pool for the pathway entities normalization contained the Reactome IDs, Gene Ontology IDs, and KEGG IDs.

For **anatomy** entity type, we used the Uberon (Mungall *et al.*, 2012) for initializing the vocabulary, as it is a cross-species anatomical ontology based on traditional anatomical classification. The linkage pool for the anatomy entities harmonization included Uberon IDs, MeSH terms, MeSH term IDs, UMLS CUIs, and the anatomy names in UMLS.

For **Dietary Supplement Ingredient (DSI)**, **Dietary Supplement Product (DSP)**, and **Therapeutic Class (TC)** entities, data were collected from our previous curated iDISK (integrated Dietary Supplements Knowledge) (Rizvi *et al.*, 2020). We used iDISK concept IDs and UMLS CUIs (for TCs) to describe them.

Knowledge integration

After the above normalization procedures, we obtained a CSV file for each entity type, storing all normalized entity terms of the specific entity type followed by their synonyms and detailed descriptions. We were then able to integrate knowledge extracted from different knowledge bases to build iBKH. Specifically, in a BKG, a basic knowledge unit is a triplet, typically defined as <***head entity***, ***relation***, ***tail entity***>, which indicates that there exists a relation from the ***head entity*** to the ***tail entity*** in iBKH. Of note, for each pair of head entity and tail entity, there can be multiple types of relations. For instance, we stored “targets”, “Transporter”, “Enzyme”, “Carrier”,

"downregulates", "upregulates", "associates", "binds", "interacts", and "text_semantic" relations between drugs and genes. We also stored the data source information, indicating from which data source(s) we acquired the specific triplet.

iBKH deployment based on graph database

We deployed our curated BKG, i.e., the iBKH, using Neo4j (https://neo4j.com), a well-designed graph database platform that allows structured queries in a grap. Specifically, Neo4j can take the CSV files of entities and relations we generated above as input and automatically created a KG instance. In this way, the iBKH can be updated efficiently and flexibly.

Graphical portal for fast knowledge retrieval

We developed a web-based graphical portal, which allows the users to design graph query paths visually and flexibly and translates them into Cypher queries (query language provided by Neo4j) automatically in the back end. Specifically, we built the back end (i.e., the server side) using Django (https://www.djangoproject.com/), a high-level Python-based web framework. The iBKH, stored in Neo4j, was linked to the back end. The front end (i.e., the web application side) was built based on HyperText Markup Language Version 5 (HTML5), and Cascading Style Sheets (CSS). JavaScript-based software, the neovis (https://github.com/neo4j-contrib/neovis.js/) and D3.js (https://d3js.org/), were used for graph visualization and data exploration and visualization, respectively.

iBKH-based knowledge discovery

We developed a machine learning pipeline for knowledge discovery in the iBKH, which contains two steps as follows.

Step 1, knowledge graph embedding (KGE) learning. The goal of KGE is to learn embeddings, i.e., meaningful and machine-readable vector-based representations for entities and/or relations in iBKH, while preserve the graph structure (Goyal and Ferrara, 2018; Su *et al.*, 2020; Wang *et al.*, 2017). In biomedicine, the learned embeddings (i.e., vector representations) of biomedical entities and relations can be used in accelerating diverse down-stream research tasks, such as drug implication discovery (Nicholson and Greene, 2020; Zhang *et al.*, 2021; Zheng et al., 2021; Zhu *et al.*, 2020), multi-omics data analysis (Nicholson and Greene, 2020; Santos *et al.*, 2022), clinical data (e.g., electronic healthcare record) analysis (Choi et al., 2017; Nelson *et al.*, 2019), and knowledge extraction from biomedical literature (Wang *et al.*, 2020). In this work, we used the Deep Graph Library - Knowledge Embedding (DGL-KE) (https://github.com/awslabs/dgl-ke)(Zheng *et al.*, 2020), a Python-based implementation for the advanced KGE algorithms, such as TransE (Bordes *et al.*, 2013), TransR (Lin *et al.*, 2015), ComplEx (Théo *et al.*, 2016), and DistMult (Yang *et al.*, 2015). Using the advanced multi-processing and multi-GPU (graphics processor unit) techniques, the DGL-KE accelerates the learning procedures in large-scale graphs like iBKH.

Step 2, link prediction. The task can be formulated as predicting the probability that an unobserved triplet $< h, r, t >$ exists in the iBKH, where h and t are the head and tail entities, and r is the potential relation, respectively. Specifically, we defined a possibility score of a candidate triplet $< h, r, t >$ as $PS(< h, r, t >) = sigmoid(f(h, r, t))$. The sigmoid function is defined as

$sigmoid(a) = 1/(1 + \exp(-a))$. $f(\cdot)$ is the scoring function of the KGE algorithm we used to calculate the embedding vectors.

- TransE, $f(h, r, t) = -\|\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}\|_p$, where $\boldsymbol{h}$, $\boldsymbol{r}$, $\boldsymbol{t}$ are the embedding vectors of h, r, t, respectively.
- TransR, $f(h, r, t) = -\|\mathbf{M}_r\boldsymbol{h} + \boldsymbol{r} - \mathbf{M}_r\boldsymbol{t}\|_p^2$, where $\mathbf{M}_r$ is a projection matrix for each relation r that project entities h and t to semantic space of the relation.
- ComplEx, $f(h, r, t) = < Re(\boldsymbol{h}), Re(\boldsymbol{r}), Re(\boldsymbol{t}) > + < Im(\boldsymbol{h}), Im(\boldsymbol{r}), Im(\boldsymbol{t}) > + < Re(\boldsymbol{h}), Im(\boldsymbol{r}), Im(\boldsymbol{t}) > - < Im(\boldsymbol{h}), Im(\boldsymbol{r}), Re(\boldsymbol{t}) >$, where $Re(\boldsymbol{x})$ and $Im(\boldsymbol{x})$ are the real and imaginary parts of the complex valued vector $\boldsymbol{x}$, respectively.
- DistMult, $f(h, r, t) = \boldsymbol{h}^T\mathbf{W}_r\boldsymbol{t}^T$, where $\mathbf{W}_r$ is relation matrix, which is restricted to a diagonal matrix.

Summarized details of the KGE algorithms can be found elsewhere (https://dglke.dgl.ai/doc/kg.html).

In silico hypothesis generation for Alzheimer's disease drug repurposing. As a proof of concept, we performed in silico hypothesis generation for Alzheimer's disease (AD) drug repurposing, which is to predict potential drug entities that can be linked to the AD entity with a 'treats' relation in the iBKH. To this end, we first downloaded all Food and Drug Administration (FDA) approved drugs and drugs in clinical trials (Phases I-IV) for AD from the DrugBank (https://go.drugbank.com/), constructing the grand truth drug list. Specifically, we obtained a total of 10 FDA-approved drugs, 30 drugs in Phase IV trials, 43 drugs in Phase III trials, 95 drugs in Phase II trials, and 47 drugs in Phase I trials for AD treatment. Next, to avoid information leaking in prediction, all relations between the AD entity and any drug in the grand truth drug list in the iBKH were removed. Then, entity and relation embedding vectors were calculated using the KGE algorithms. After that, we calculated possibility scores for potential all

$< e_d, r, e_{AD} >$ triplets, where e_d indicates any drug entity, e_{AD} indicates the AD entity, and r indicates a relation between them. The drugs were ranked based on the possibility scores. In this study, we calculated the possibility scores based on four KGE algorithms, i.e., TransE (Bordes *et al.*, 2013), TransR (Lin *et al.*, 2015), ComplEx (Théo *et al.*, 2016), and DistMult (Yang *et al.*, 2015). To enhance prediction, we also proposed an ensemble model. Specifically, the rank of drug e_d in the ensemble model was defined as $PS^{ensemble}(< e_d, r, e_{AD} >) = \sum_i (N^{Dr} - Rank^i(< e_d, r, e_{AD} >))$ where i indicates the i-th KGE algorithm and N^{Dr} indicates total number of drugs in iBKH.

To evaluate prediction performance, we compared the top K ranked drugs with the ground truth drugs. By sliding the value of K, we were able to produce the receiver operating characteristic curve (ROC) and the area under ROC (AUC) score.

Finally, we re-trained the KGE models without removing known relations between AD and drug entities and used the embeddings to predict novel repurposable drug candidates for AD treatment. For the predicted drugs that potentially link to AD, we performed manual literature review to identify supporting evidence of the prediction.

References:

1. Aisen, P.S. (2002). The potential of anti-inflammatory drugs for the treatment of Alzheimer's disease. The Lancet Neurology *1*, 279-284. https://doi.org/10.1016/S1474-4422(02)00133-3.
2. Anekonda, T.S., and Quinn, J.F. (2011). Calcium channel blocking as a therapeutic strategy for Alzheimer's disease: The case for isradipine. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease *1812*, 1584-1590. https://doi.org/10.1016/j.bbadis.2011.08.013.
3. Bastian, F.B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, Sara S., de Farias, T.M., Moretti, S., Parmentier, G., de Laval, V.R., Rosikiewicz, M., et al. (2021). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Research *49*, D831-D847. 10.1093/nar/gkaa793.
4. Beaman, E.E., Bonde, A.N., Ulv Larsen, S.M., Ozenne, B., Lohela, T.J., Nedergaard, M., Gíslason, G.H., Knudsen, G.M., and Holst, S.C. (2022). Blood–brain barrier permeable β-blockers linked to lower risk of Alzheimer's disease in hypertension. Brain.
5. Benhamú, B., Martín-Fontecha, M., Vázquez-Villa, H., Pardo, L., and López-Rodríguez, M.L. (2014). Serotonin 5-HT6 Receptor Antagonists for the Treatment of Cognitive Deficiency in Alzheimer's Disease. Journal of Medicinal Chemistry *57*, 7160-7181. 10.1021/jm5003952.
6. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research *32*, D267-D270. 10.1093/nar/gkh061.
7. Bojarski, L., Herms, J., and Kuznicki, J. (2008). Calcium dysregulation in Alzheimer's disease. Neurochemistry International *52*, 621-633. https://doi.org/10.1016/j.neuint.2007.10.002.
8. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems *26*.
9. Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., and Bruford, E. (2019). Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Research *47*, D786-D792. 10.1093/nar/gky930.
10. Cai, Z., and Ratka, A. (2012). Opioid System and Alzheimer's Disease. NeuroMolecular Medicine *14*, 91-111. 10.1007/s12017-012-8180-3.
11. Callahan, T.J., Tripodi, I.J., Pielke-Lombardo, H., and Hunter, L.E. (2020). Knowledge-Based Biomedical Data Science. Annual Review of Biomedical Data Science *3*, 23-41. 10.1146/annurev-biodatasci-010820-091627.
12. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., and Schomburg, D. (2021). BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Research *49*, D498-D508. 10.1093/nar/gkaa1025.
13. Chen, I.Y., Agrawal, M., Horng, S., and Sontag, D. (2019). Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. In Biocomputing 2020, (WORLD SCIENTIFIC), pp. 19-30. doi:10.1142/9789811215636_0003
14. 10.1142/9789811215636_0003.
15. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., and Sun, J. (2017). GRAM: Graph-based Attention Model for Healthcare Representation Learning. Proceedings of the 23rd

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery.

16. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2019). The Comparative Toxicogenomics Database: update 2019. Nucleic Acids Research *47*, D948-D954. 10.1093/nar/gky868.
17. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical Entities of Biological Interest: an update. Nucleic Acids Research *38*, D249-D254. 10.1093/nar/gkp886.
18. DeHaven-Hudkins, D.L., Burgos, L.C., Cassel, J.A., Daubert, J.D., DeHaven, R.N., Mansson, E., Nagasaka, H., Yu, G., and Yaksh, T. (1999). Loperamide (ADL 2-1294), an Opioid Antihyperalgesic Agent with Peripheral Selectivity. Journal of Pharmacology and Experimental Therapeutics *289*, 494.
19. Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., et al. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. Journal of Biomedical Semantics *7*, 44. 10.1186/s13326-016-0088-7.
20. Ernst, P., Siu, A., and Weikum, G. (2015). KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics *16*, 157. 10.1186/s12859-015-0549-5.
21. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Research *46*, D649-D655. 10.1093/nar/gkx1132.
22. Fang, J., Zhang, P., Wang, Q., Chiang, C.-W., Zhou, Y., Hou, Y., Xu, J., Chen, R., Zhang, B., Lewis, S.J., et al. (2022). Artificial intelligence framework identifies candidate targets for drug repurposing in Alzheimer's disease. Alzheimer's Research & Therapy *14*, 7. 10.1186/s13195-021-00951-z.
23. Fang, J., Zhang, P., Zhou, Y., Chiang, C.-W., Tan, J., Hou, Y., Stauffer, S., Li, L., Pieper, A.A., Cummings, J., and Cheng, F. (2021). Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease. Nature Aging *1*, 1175-1188. 10.1038/s43587-021-00138-z.
24. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J.P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research *40*, D1100-D1107. 10.1093/nar/gkr777.
25. Giagnoni, G., Casiraghi, L., Senini, R., Revel, L., Parolaro, D., Sala, M., and Gori, E. (1983). Loperamide: Evidence of interaction with μ and δ opioid receptors. Life Sciences *33*, 315-318. https://doi.org/10.1016/0024-3205(83)90506-4.
26. Goyal, P., and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems *151*, 78-94. https://doi.org/10.1016/j.knosys.2018.03.022.
27. Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B., and Klein, T.E. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Research *30*, 163-165. 10.1093/nar/30.1.163.
28. Himeno, E., Ohyagi, Y., Ma, L., Nakamura, N., Miyoshi, K., Sakae, N., Motomura, K., Soejima, N., Yamasaki, R., and Hashimoto, T. (2011). Apomorphine treatment in Alzheimer mice promoting amyloid-β degradation. Annals of neurology *69*, 248-256.
29. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S.E. (2017). Systematic integration of

biomedical knowledge prioritizes drugs for repurposing. eLife *6*, e26726. 10.7554/eLife.26726.

30. Hu, J., Lepore, R., Dobson, R.J.B., Al-Chalabi, A., M. Bean, D., and Iacoangeli, A. (2021). DGLinker: flexible knowledge-graph prediction of disease–gene associations. Nucleic Acids Research *49*, W153-W161. 10.1093/nar/gkab449.
31. Ioannidis, V.N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., and Karypis, G. (2020). DRKG - Drug Repurposing Knowledge Graph for Covid-19. https://github.com/gnn4dr/DRKG/.
32. Jain, M., Singh, M.K., Shyam, H., Mishra, A., Kumar, S., Kumar, A., and Kushwaha, J. (2021). Role of JAK/STAT in the Neuroinflammation and its Association with Neurological Disorders. Annals of Neurosciences *28*, 191-200. 10.1177/09727531211070532.
33. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research *28*, 27-30. 10.1093/nar/28.1.27.
34. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The SIDER database of drugs and side effects. Nucleic Acids Research *44*, D1075-D1079. 10.1093/nar/gkv1075.
35. Li, N., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., and Wang, J. (2020). KGHC: a knowledge graph for hepatocellular carcinoma. BMC Medical Informatics and Decision Making *20*, 135. 10.1186/s12911-020-1112-5.
36. Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. Twenty-ninth AAAI conference on artificial intelligence.
37. Liu, C.-C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nature Reviews Neurology *9*, 106-118. 10.1038/nrneurol.2012.263.
38. Luchsinger, J.A., and Mayeux, R. (2004). Dietary factors and Alzheimer's disease. The Lancet Neurology *3*, 579-587. https://doi.org/10.1016/S1474-4422(04)00878-6.
39. Luchsinger, J.A., Noble, J.M., and Scarmeas, N. (2007). Diet and Alzheimer's disease. Current Neurology and Neuroscience Reports *7*, 366-372. 10.1007/s11910-007-0057-8.
40. Mohamed, S.K., Nounu, A., and Nováček, V. (2021). Biological applications of knowledge graph embedding models. Briefings in Bioinformatics *22*, 1679-1693. 10.1093/bib/bbaa012.
41. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., and Haendel, M.A. (2012). Uberon, an integrative multi-species anatomy ontology. Genome Biology *13*, R5. 10.1186/gb-2012-13-1-r5.
42. Nakamura, N., Ohyagi, Y., Imamura, T., Yanagihara, Y.T., Iinuma, K.M., Soejima, N., Murai, H., Yamasaki, R., and Kira, J.-i. (2017). Apomorphine Therapy for Neuronal Insulin Resistance in a Mouse Model of Alzheimer's Disease. Journal of Alzheimer's Disease *58*, 1151-1161. 10.3233/JAD-160344.
43. Nelson, C.A., Butte, A.J., and Baranzini, S.E. (2019). Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. Nature Communications *10*, 3045. 10.1038/s41467-019-11069-0.
44. Nevado-Holgado, A.J., Ribe, E., Thei, L., Furlong, L., Mayer, M.-A., Quan, J., Richardson, J.C., Cavanagh, J., Consortium, N., and Lovestone, S. (2019). Genetic and Real-World Clinical Data, Combined with Empirical Validation, Nominate Jak-Stat Signaling as a Target for Alzheimer's Disease Therapeutic Development. Cells *8*. 10.3390/cells8050425.
45. Nicholson, D.N., and Greene, C.S. (2020). Constructing knowledge graphs and their biomedical applications. Computational and Structural Biotechnology Journal *18*, 1414-1428. https://doi.org/10.1016/j.csbj.2020.05.017.

46. Ou, H., Chien, W.-C., Chung, C.-H., Chang, H.-A., Kao, Y.-C., Wu, P.-C., and Tzeng, N.-S. (2021). Association Between Antibiotic Treatment of Chlamydia pneumoniae and Reduced Risk of Alzheimer Dementia: A Nationwide Cohort Study in Taiwan. Frontiers in Aging Neuroscience *13*. 10.3389/fnagi.2021.701899.
47. Palasca, O., Santos, A., Stolte, C., Gorodkin, J., and Jensen, L.J. (2018). TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database *2018*, bay003. 10.1093/database/bay003.
48. Peng, C., Dieck, S., Schmid, A., Ahmad, A., Knaus, A., Wenzel, M., Mehnert, L., Zirn, B., Haack, T., Ossowski, S., et al. (2021). CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genomics and Bioinformatics *3*, lqab078. 10.1093/nargab/lqab078.
49. Peng, Y., Yuan, M., Xin, J., Liu, X., and Wang, J. (2020). Screening novel drug candidates for Alzheimer's disease by an integrated network and transcriptome analysis. Bioinformatics *36*, 4626-4632. 10.1093/bioinformatics/btaa563.
50. Percha, B., and Altman, R.B. (2018). A global network of biomedical relationships derived from text. Bioinformatics *34*, 2614-2624. 10.1093/bioinformatics/bty114.
51. Perry, G., Cash, A.D., and Smith, M.A. (2002). Alzheimer Disease and Oxidative Stress. Journal of Biomedicine and Biotechnology *2*, 542340. 10.1155/S1110724302203010.
52. Rau, T.F., Kothiwal, A., Rova, A., Rhoderick, J.F., and Poulsen, D.J. (2014). Phenoxybenzamine Is Neuroprotective in a Rat Model of Severe Traumatic Brain Injury. International Journal of Molecular Sciences *15*. 10.3390/ijms15011402.
53. Rivers-Auty, J., Tapia, V.S., White, C.S., Daniels, M.J.D., Drinkall, S., Kennedy, P.T., Spence, H.G., Yu, S., Green, J.P., Hoyle, C., et al. (2021). Zinc Status Alters Alzheimer's Disease Progression through NLRP3-Dependent Inflammation. The Journal of Neuroscience *41*, 3025. 10.1523/JNEUROSCI.1980-20.2020.
54. Rizvi, R.F., Vasilakes, J., Adam, T.J., Melton, G.B., Bishop, J.R., Bian, J., Tao, C., and Zhang, R. (2020). iDISK: the integrated DIetary Supplements Knowledge base. Journal of the American Medical Informatics Association *27*, 539-548. 10.1093/jamia/ocz216.
55. Rosenberg, P.B., Mielke, M.M., Tschanz, J., Cook, L., Corcoran, C., Hayden, K.M., Norton, M., Rabins, P.V., Green, R.C., Welsh-Bohmer, K.A., et al. (2008). Effects of Cardiovascular Medications on Rate of Functional Decline in Alzheimer Disease. The American Journal of Geriatric Psychiatry *16*, 883-892. https://doi.org/10.1097/JGP.0b013e318181276a.
56. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., and Sontag, D. (2017). Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports *7*, 5994. 10.1038/s41598-017-05778-z.
57. Rubin, D.L., Shah, N.H., and Noy, N.F. (2008). Biomedical ontologies: a functional perspective. Briefings in Bioinformatics *9*, 75-90. 10.1093/bib/bbm059.
58. Santos, A., Colaço, A.R., Nielsen, A.B., Niu, L., Strauss, M., Geyer, P.E., Coscia, F., Albrechtsen, N.J.W., Mundt, F., Jensen, L.J., and Mann, M. (2022). A knowledge graph to interpret clinical proteomics data. Nature Biotechnology. 10.1038/s41587-021-01145-6.
59. Saravanaraman, P., Chinnadurai, R.K., and Boopathy, R. (2014). Why calcium channel blockers could be an elite choice in the treatment of Alzheimer's disease: a comprehensive review of evidences. Reviews in the Neurosciences *25*, 231-246. doi:10.1515/revneuro-2013-0056.
60. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research *40*, D940-D946. 10.1093/nar/gkr972.

61. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., and Rosse, C. (2005). Relations in biomedical ontologies. Genome Biology *6*, R46. 10.1186/gb-2005-6-5-r46.
62. Squitti, R., Pal, A., Picozza, M., Avan, A., Ventriglia, M., Rongioletti, M.C., and Hoogenraad, T. (2020). Zinc Therapy in Early Alzheimer's Disease: Safety and Potential Therapeutic Efficacy. Biomolecules *10*. 10.3390/biom10081164.
63. Strittmatter, W.J., and Roses, A.D. (1995). Apolipoprotein E and Alzheimer disease. Proceedings of the National Academy of Sciences *92*, 4725-4727. 10.1073/pnas.92.11.4725.
64. Su, C., Hou, Y., and Wang, F. (2022). GNN-based Biomedical Knowledge Graph Mining in Drug Development. In Graph Neural Networks: Foundations, Frontiers, and Applications, L. Wu, P. Cui, J. Pei, and L. Zhao, eds. (Springer Nature Singapore), pp. 517-540. 10.1007/978-981-16-6054-2_24.
65. Su, C., Tong, J., Zhu, Y., Cui, P., and Wang, F. (2020). Network embedding in biomedical data science. Briefings in Bioinformatics *21*, 182-197. 10.1093/bib/bby117.
66. Sügis, E., Dauvillier, J., Leontjeva, A., Adler, P., Hindie, V., Moncion, T., Collura, V., Daudin, R., Loe-Mie, Y., Herault, Y., et al. (2019). HENA, heterogeneous network-based data set for Alzheimer's disease. Scientific Data *6*, 151. 10.1038/s41597-019-0152-0.
67. Théo, T., Johannes, W., Sebastian, R., Eric, G., and Guillaume, B. (2016). Complex Embeddings for Simple Link Prediction. International conference on machine learning. PMLR.
68. Upton, N., Chuang, T.T., Hunter, A.J., and Virley, D.J. (2008). 5-HT6 Receptor Antagonists as Novel Cognitive Enhancing Agents for Alzheimer's Disease. Neurotherapeutics *5*, 458-469. https://doi.org/10.1016/j.nurt.2008.05.008.
69. van Dam, R.M., Rimm, E.B., Willett, W.C., Stampfer, M.J., and Hu, F.B. (2002). Dietary Patterns and Risk for Type 2 Diabetes Mellitus in U.S. Men. Annals of Internal Medicine *136*, 201-209. 10.7326/0003-4819-136-3-200202050-00008.
70. Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., and Liu, W. (2020). COVID-19 literature knowledge graph construction and drug repurposing report generation. arXiv preprint arXiv:2007.00576.
71. Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Transactions on Knowledge and Data Engineering *29*, 2724-2743. 10.1109/TKDE.2017.2754499.
72. Williams, G., Gatt, A., Clarke, E., Corcoran, J., Doherty, P., Chambers, D., and Ballard, C. (2019). Drug repurposing for Alzheimer's disease based on transcriptional profiling of human iPSC-derived cortical neurons. Translational Psychiatry *9*, 220. 10.1038/s41398-019-0555-x.
73. Williams, M.T., and Hord, N.G. (2005). The role of dietary factors in cancer prevention: beyond fruits and vegetables. Nutrition in clinical practice *20*, 451-459.
74. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research *46*, D1074-D1082. 10.1093/nar/gkx1037.
75. Xu, R., Li, L., and Wang, Q. (2013). Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. Bioinformatics *29*, 2186-2194. 10.1093/bioinformatics/btt359.
76. Yang, B., Yih, S.W.-t., He, X., Gao, J., and Deng, L. (2015). Embedding Entities and Relations for Learning and Inference in Knowledge Bases. International Conference on Learning Representations (ICLR).
77. Yu, Y., Wang, Y., Xia, Z., Zhang, X., Jin, K., Yang, J., Ren, L., Zhou, Z., Yu, D., Qing, T., et al. (2019). PreMedKB: an integrated precision medicine knowledgebase for

interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Research *47*, D1090-D1101. 10.1093/nar/gky1042.

78. Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., and Luo, J. (2020). Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowledge and Information Systems *62*, 317-336. 10.1007/s10115-019-01351-4.
79. Zang, C., Zhang, H., Xu, J., Zhang, H., Fouladvand, S., Havaldar, S., Cheng, F., Chen, K., Chen, Y., and Glicksberg, B.S. (2022). High-Throughput Clinical Trial Emulation with Real World Data and Machine Learning: A Case Study of Drug Repurposing for Alzheimer's Disease. medRxiv.
80. Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., Zhang, Z., Li, K., Karypis, G., and Cheng, F. (2020). Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. Journal of Proteome Research *19*, 4624-4636. 10.1021/acs.jproteome.0c00316.
81. Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., and Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. Journal of Biomedical Informatics *115*, 103696. https://doi.org/10.1016/j.jbi.2021.103696.
82. Zhang, Y., Zheng, W., Lin, H., Wang, J., Yang, Z., and Dumontier, M. (2018). Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics *34*, 828-835. 10.1093/bioinformatics/btx659.
83. Zhao, S., Huang, Y., Su, C., Li, Y., and Wang, F. (2020a). Interactive attention networks for semantic text matching. (IEEE), pp. 861-870.
84. Zhao, S., Qin, B., Liu, T., and Wang, F. (2020b). Biomedical knowledge graph refinement with embedding and logic rules. arXiv preprint arXiv:2012.01031.
85. Zhao, S., Su, C., Lu, Z., and Wang, F. (2021). Recent advances in biomedical literature mining. Briefings in Bioinformatics *22*, bbaa057. 10.1093/bib/bbaa057.
86. Zhao, S., Su, C., Sboner, A., and Wang, F. (2019). Graphene: A precise biomedical literature retrieval engine with graph augmented deep learning and external knowledge empowerment. pp. 149-158.
87. Zheng, D., Song, X., Ma, C., Tan, Z., Ye, Z., Dong, J., Xiong, H., Zhang, Z., and Karypis, G. (2020). DGL-KE: Training Knowledge Graph Embeddings at Scale. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, (Association for Computing Machinery), pp. 739–748. 10.1145/3397271.3401172.
88. Zheng, S., Rao, J., Song, Y., Zhang, J., Xiao, X., Fang, E.F., Yang, Y., and Niu, Z. (2021). PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in Bioinformatics *22*, bbaa344. 10.1093/bib/bbaa344.
89. Zhou, Y., Fang, J., Bekris, L.M., Kim, Y.H., Pieper, A.A., Leverenz, J.B., Cummings, J., and Cheng, F. (2021). AlzGPS: a genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery. Alzheimer's Research & Therapy *13*, 24. 10.1186/s13195-020-00760-w.
90. Zhou, Y., Wang, F., Tang, J., Nussinov, R., and Cheng, F. (2020). Artificial intelligence in COVID-19 drug repurposing. The Lancet Digital Health *2*, e667-e676. https://doi.org/10.1016/S2589-7500(20)30192-8.
91. Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C., and Wang, F. (2020). Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. Health Informatics Journal *26*, 2737-2750.
92. Zhu, Y., Elemento, O., Pathak, J., and Wang, F. (2019). Drug knowledge bases and their applications in biomedical informatics research. Briefings in Bioinformatics *20*, 1308-1321. 10.1093/bib/bbx169.

93. Zlomuzica, A., Dere, D., Binder, S., De Souza Silva, M.A., Huston, J.P., and Dere, E. (2016). Neuronal histamine and cognitive symptoms in Alzheimer's disease. Neuropharmacology *106*, 135-145. https://doi.org/10.1016/j.neuropharm.2015.05.007.