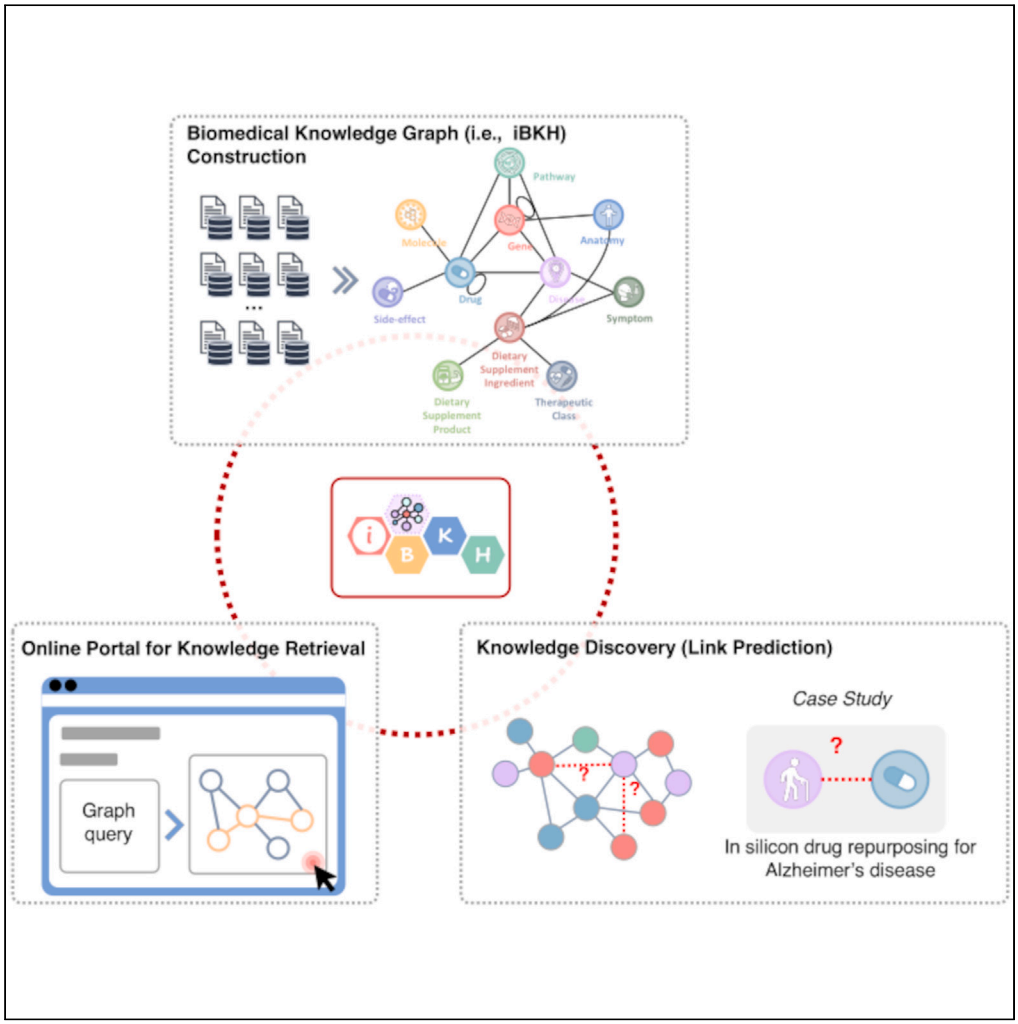
iScience

CellPress
OPEN ACCESS

Article

Biomedical discovery through the integrative biomedical knowledge hub (iBKH)



Chang Su, Yu Hou, Manqi Zhou, ..., Steven T. DeKosky, Jiang Bian, Fei Wang

few2001@med.cornell.edu

Highlights

We build integrative biomedical knowledge Hub (iBKH), a comprehensive biomedical knowledge graph (BKG)

iBKH enables fast biomedical knowledge retrieval

We build a knowledge discovery pipeline based on machine learning and iBKH

A case study: iBKH-based knowledge discovery for Alzheimer's disease drug repurposing

Su et al., iScience 26, 106460
April 21, 2023
https://doi.org/10.1016/j.isci.2023.106460

iScience

CellPress
OPEN ACCESS

Article

Biomedical discovery through the integrative biomedical knowledge hub (iBKH)

Chang Su,[1,17] Yu Hou,[2,3,17] Manqi Zhou,[4] Suraj Rajendran,[5] Jacqueline R.M. A. Maasch,[6] Zehra Abedi,[2] Haotan Zhang,[7] Zilong Bai,[2] Anthony Cuturrufo,[8] Winston Guo,[9] Fayzan F. Chaudhry,[7] Gregory Ghahramani,[7] Jian Tang,[10] Feixiong Cheng,[11,12,13] Yue Li,[14] Rui Zhang,[3] Steven T. DeKosky,[15] Jiang Bian,[16] and Fei Wang[2,18,*]

SUMMARY

The abundance of biomedical knowledge gained from biological experiments and clinical practices is an invaluable resource 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 this study, we created a comprehensive BKG called the integrative Biomedical Knowledge Hub (iBKH) by harmonizing and integrating information from diverse biomedical resources. To make iBKH easily accessible for biomedical research, we developed a web-based, user-friendly graphical portal that allows fast and interactive knowledge retrieval. Additionally, we also implemented an efficient and scalable graph learning pipeline for discovering novel biomedical knowledge 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.

INTRODUCTION

Biomedicine is a discipline with enormous volume of highly specialized biomedical knowledge accumulated from biological experiments and clinical practice. In the past decade, efforts have been drawn to collect and manage the abundant biomedical knowledge. BKG (BKG) has emerged as a novel paradigm for better management of large scale and heterogeneous biomedical knowledge and attracted significant interests in biomedicine recently.[1–6] Typically, a BKG is a multi-relational graph or network that integrates, harmonizes, and stores biomedical knowledge collected from single or multiple expert-derived knowledge sources.[1,2,4,7] A BKG contains a set of nodes that correspond to biomedical entities (e.g., diseases, drugs, genes, biological processes, etc.) and a set of edges that are relations linking the biomedical entities (e.g., drug-treats-disease, disease-associates-gene, and drug-interacts-drug relations.).[1,2,4,7] In the past decade, large amounts of efforts have been made to construct BKGs by integrating diverse expert curated knowledge bases[2,4,7–9] and extracting knowledge from literature using natural language processing techniques.[10–12] As a result, many different BKGs have been built.[13–16]

Despite the promising results achieved from existing BKG efforts, there are still limitations that hinder their utility in modern biomedical research and clinical practice. First, most of the current BKGs focus on one or a few sub-domains of biomedicine; hence they cannot characterize the human health holistically and comprehensively.[15,16] This makes it challenging for efficient exploration of cross-domain biomedical knowledge to provide system-level understanding of diseases. Second, existing BKGs are mostly publicly available as raw text information of the nodes and edges therein,[1,15] which requires informatics training for the end users to make full utilization of them. Thus, there remains a need for a publicly available and easy-to-use user interface (UI) to facilitate knowledge exploration on these BKGs. Third, the reasoning and inference capabilities available on existing BKGs are limited. With the revolution of deep learning technologies in NLP[17] and reasoning in general domain knowledge graphs,[18] there is a huge potential of making high-quality reasoning/hypothesis generation with evidence support as an addition functionality of BKG to accelerate new biomedical knowledge discovery.

To fill in the above gaps, this present study built a comprehensive BKG, termed the iBKH, through integrating information from 18 high-quality and well-curated knowledge sources. We developed a

[1]Department of Health Service Administration and Policy, College of Public Health, Temple University, Philadelphia, PA 19122, USA

[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065, USA

[3]Department of Surgery, University of Minnesota, Minneapolis, MN 55455, USA

[4]Department of Computational Biology, Cornell University, Ithaca, NY 14850, USA

[5]Tri-Institutional Computational Biology & Medicine Program, Cornell University, New York, NY 10065, USA

[6]Department of Computer Science, Cornell Tech, New York, NY 10044, USA

[7]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA

[8]Computer Science, Cornell University, Ithaca, NY 14850, USA

[9]Department of Medicine, Weill Cornell Medicine, New York, NY 10021, USA

[10]Mila-Quebec AI Institute and HEC Montreal, Montreal, QC H2S 3H1, Canada

[11]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[12]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA

[13]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

[14]School of Computer Science, McGill University,

Continued

web-based, easy-to-use, intelligent graphical portal for iBKH that facilitates fast and interactive knowledge retrieval.

To enable high-quality knowledge discovery, we further integrated modern graph learning pipelines in iBKH. In general, graph learning is an emerging branch of machine learning that aims at exploring knowledge from graph structured data.[19,20] In recent years, they have been applied in BKG to accelerate biomedical hypothesis generation such as drug repurposing[7,21–23] and disease risk gene prioritization.[24,25] In iBKH, we have implemented a knowledge discovery module based on DGL-KE (Deep Graph Library – Knowledge Embedding),[26] the python package for efficient and scalable graph learning. To demonstrate its potentials, we conducted two proof-of-concept studies: 1) in-silico hypothesis generation for Alzheimer's disease (AD) drug repurposing and 2) enhancing data analysis of a patient cohort of older adults with Apolipoprotein E (APOE) ε4 genotype, which is a significant genetic risk factor of AD.

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

By collecting, harmonizing, and integrating data from 18 publicly available biomedical knowledge sources (see Table 1), we curated a comprehensive BKG, named the iBKH. The knowledge sources include biomedical ontologies such as the BRENDA Tissue Ontology,[27] the Cell Ontology[28] the Disease Ontology,[29] and the Uberon[30]; manually curated biomedical knowledge bases for biomedical entity and relation data such as the Bgee,[31] the Comparative Toxicogenomics Database,[32] the DrugBank,[33] the Kyoto Encyclopedia of Genes and Genomes (KEGG),[34] the Pharmacogenetics Knowledge Base (PharmGKB),[35] the Reactome,[36] the Side effect resource,[37] and the TISSUE[38]; existing BKGs curated by integrating multiple knowledge bases such as the drug repurposing knowledge graph (DRKG, https://github.com/gnn4dr/DRKG),[39] the Hetionet,[4] the Integrated Dietary Supplement Knowledge Base (integrated Dietary Supplements Knowledge (iDISK)),[40] our curated knowledge graph that covers a variety of dietary supplements, including vitamins, herbs, minerals, etc.; and other biomedical sources such as Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC),[41] Chemical Biology Information Resource from EMBL [European Molecular Biology Laboratory] (ChEMBL),[42] and Chemical Entities of Biological Interest (ChEBI).[43] More details of the sources can be found in Table 1.

After data management and necessary data cleaning, we integrated data from different sources through biomedical entity term normalization and knowledge integration (more details can be found in the STAR Methods 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, and 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 an 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

Montreal, QC H3A 0C6, Canada

[15]Department of Neurology, College of Medicine, University of Florida, Gainesville, FL 32610, USA

[16]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL 32610, USA

[17]These authors contributed equally

[18]Lead contact

*Correspondence: few2001@med.cornell.edu

https://doi.org/10.1016/j.isci.2023.106460

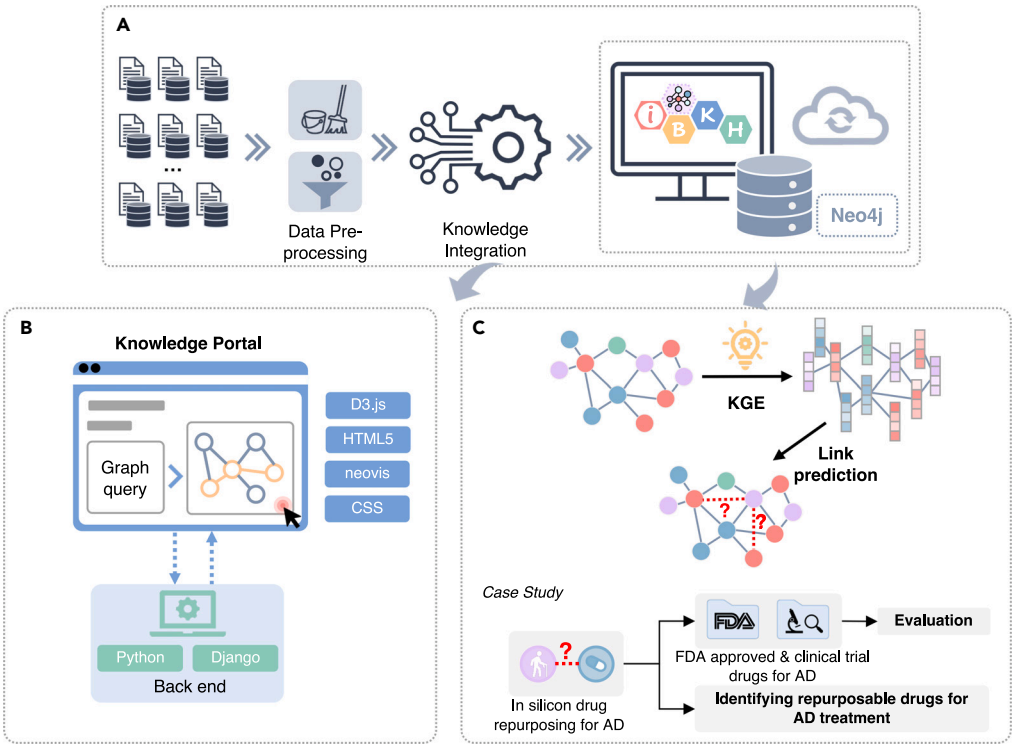


Figure 1. An illustration of study pipeline

(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, KGE 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.

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 comma-separated values (CSV) format, available at: https://github.com/wcm-wanglab/iBKH. Of note, the deployed version of iBKH excluded data from KEGG due to restriction.

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 information 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

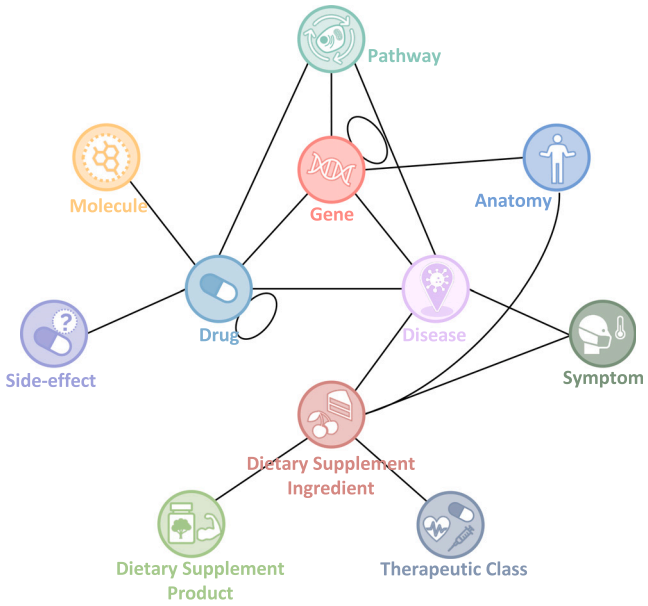


Figure 2. Schema of iBKH

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," and "Text_Semantic_Relation" can exist between a pair of disease and gene entities.

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 gene, which produces APOE protein and is the known major risk gene for AD.[44,45] By choosing DrugBank and PharmGKB 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. For instance, besides AD, APOE is also associated with diseases such depressive disorders, hyperlipidemia, atrial fibrillation, and hypertriglyceridemia, which have been reported as comorbidities of 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.[46,47]

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). The retrieved information was illustrated in Figure 3B, where we visualized 100 retrieved triplets (by setting Limit of Triplet as 100 in the portal). Centered around the disease entity AD, genes associated with it were retrieved first. 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.[48]

A machine learning pipeline for iBKH-based biomedical knowledge discovery

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.[1,19] Our pipeline contains two steps (see STAR Methods section and Figure 1C). First, we utilized the knowledge graph embedding (KGE) algorithms which calculate machine-readable embedding vectors for entities and/or relations in iBKH while preserving the graph structure.[19,20,49] Here we utilized an efficient

Table 1. Data sources integrated for constructing iBKH

		Entity		Relation			
Source	Description	Types	Number	Types	Number	URL	License
Bgee[31]	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[27]	A tissue-specific ontology.	Tissue (Anatomy)	6,478	–	–	https://www.brenda-enzymes.org/index.php	https://creativecommons.org/licenses/by/4.0/
Cell Ontology[28]	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)[32]	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 e-mail.
ChEMBL[42]	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)[43]	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/

(*Continued on next page*)

Table 1. *Continued*

Source	Description	Entity		Relation		URL	License
		Types	Number	Types	Number		
Drug Repurposing Knowledge Graph (DRKG)[39]	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[29]	Standardized ontology for human disease.	Disease	10,648	–	–	https://disease-ontology.org/	https://creativecommons.org/publicdomain/zero/1.0/
DrugBank[33]	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/

(*Continued on next page*)

Table 1. *Continued*

		Entity		Relation			
Source	Description	Types	Number	Types	Number	URL	License
Hetionet[4]	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/

(*Continued on next page*)

Table 1. *Continued*

Source	Description	Entity		Relation		URL	License
		Types	Number	Types	Number		
HUGO Gene Nomenclature Committee (HGNC)[41]	The resource for approved human gene nomenclature	Gene	41,439	–	–	https://www.genenames.org/	No restriction
Integrated Dietary Supplement Knowledge Base (iDISK)[40]	Our curated knowledge graph that covers a variety of dietary supplements, including vitamins, herbs, minerals, etc.	Dietary Supplement Ingredient, Dietary Supplement Product, Disease, Drug, Anatomy, Symptom, Therapeutic Class	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/
Kyoto Encyclopedia of Genes and Genomes (KEGG)[34]	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)[35]	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[36]	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/

(*Continued on next page*)

Table 1. *Continued*

		Entity		Relation			
Source	Description	Types	Number	Types	Number	URL	License
Side effect resource (SIDER)[37]	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[38]	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[30]	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[a]
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

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.
[a]The identifiers used for entity term normalization.

python package for graph learning, Deep Graph Library – Knowledge Embedding (DGL-KE).[26] We used four advanced KGE algorithms in DGL-KE including TransE,[50] TransR,[51] ComplEx,[52] and DistMult.[53] Second, link prediction (predicting potential relations between a pair of entities) was performed based on the learned embedding vectors calculated by each KGE algorithm. We split iBKH into 90% training and 10% testing sets, where the training set was used to train KGE models, and the testing set was used for evaluating link prediction performance of the models based on multiple metrics (see STAR Methods section). Table 4 shows that the four KGE models can achieve desirable performance in link prediction in iBKH. After that, we retrained KGE models using the entire iBKH to obtain entity and relation embeddings and applied our iBKH-based knowledge discovery pipeline for in-silico hypothesis generation as detailed below.

In-silico hypothesis generation: a case study of Alzheimer's disease drug repurposing

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 (see STAR Methods section and Figure 1.).[54–57] Such analysis has been used to identify repurposable drug candidates for COVID-19 in our previous study.[57] In order to assess the effectiveness of our approach for predicting repurposable drugs for AD, we used a ground truth consisting of FDA-approved drugs and drugs currently being tested in clinical trials for AD treatment. This included a total of 10 FDA-approved drugs and 215 drugs in various stages of clinical trials (30 in Phase IV, 43 in Phase III, 95 in Phase II, and 47 in Phase I). To prevent any potential data leakage during the prediction process, all connections between the AD entity and any drug in the ground truth list were removed from the iBKH (see STAR Methods section). Figure 4 provides an overview of the performance of our method, which involved generating predictions based on embedding vectors produced by four different KGE algorithms (TransE, TransR, ComplEx, and DistMult), as well as an ensemble model that combined all four algorithms (see the STAR Methods section for more details). Our approach achieved strong prediction performance, with an AUC score over 0.83 for all methods in predicting FDA-approved AD drugs and an AUC over 0.75 in predicting both FDA-approved drugs and drugs in Phase IV clinical trials (n = 40). This suggests that our approach is particularly effective at ranking FDA-approved and Phase IV clinical trial drugs for AD. Furthermore, our ensemble model achieved even better performance (e.g., AUC = 0.9 for FDA-approved drugs, AUC = 0.79 for FDA-approved plus Phase IV clinical trial drugs for AD), indicating that it benefits from the use of multiple 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 5).

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.

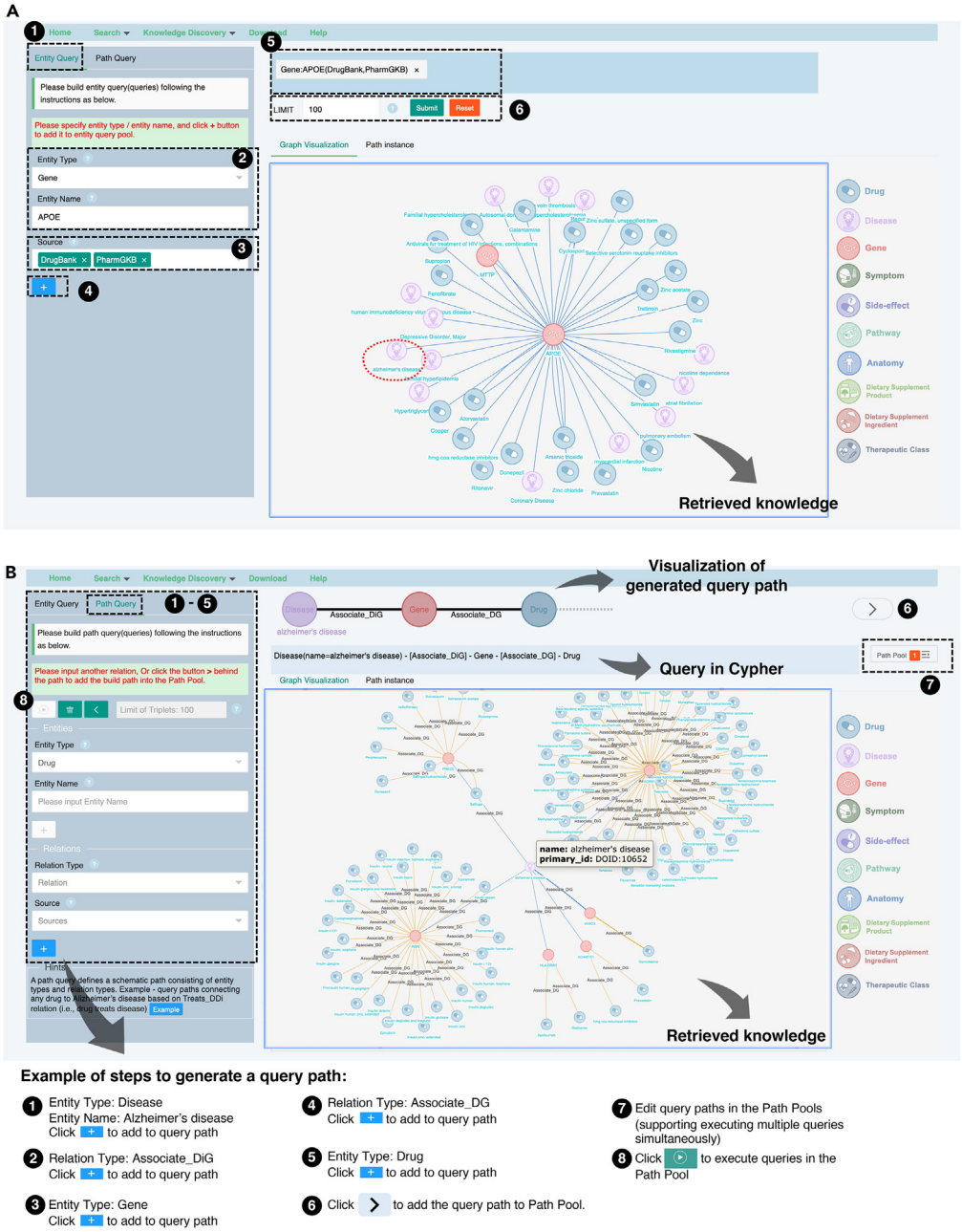


Figure 3. Examples of knowledge retrieval

(A) An example of entity query—retrieving neighborhood context of APOE 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.

First, approach identified three **anti-hypertensive drugs** that ranked high as potential drug candidates for AD treatment, including *labetalol* (DrugBank ID: DB00598), *phenoxybenzamine* (DrugBank ID: DB00925), and *mibefradil* (DrugBank ID: DB01388). Labetalol belongs to the class of β-blockers and there is evidence suggesting that β-blockers may enhance cerebrospinal fluid flow, resulting in increased brain clearance of certain metabolites. Recent studies have also reported that the use of β-blockers is associated with a reduced risk of AD onset[61] and functional decline in AD.[62] Phenoxybenzamine is an α-blocker, which has been reported to have neuroprotective activity.[63] Additionally, recent drug repurposing studies have

Table 4. Link prediction performance

	Model			
Measurement	TransE	TransR	DistMult	ComplEx
Hit@1	0.74	0.81	0.39	0.42
Hit@3	0.88	0.92	0.62	0.64
Hit@10	0.95	0.98	0.80	0.82
MR	3.55	2.64	10.87	9.49
MRR	0.82	0.87	0.53	0.56

For Hit@k (k = 1, 3, or 10) and MRR, a value close to 1 indicates good link prediction performance, otherwise close to 0. For MR, a smaller value, i.e., close to 1, indicates good link prediction performance.
Abbreviations: MR = Mean Rank; MRR = Mean Reciprocal Rank.

also suggested that phenoxybenzamine could be a promising candidate for AD treatment.[64,65] Although mibefradil was withdrawn from the market in 1998 due to harmful interactions with other drugs, our findings suggest that CCBs could be potential candidates for AD treatment because calcium dysregulation has been implicated in AD[66] and CCBs have demonstrated multiple beneficial effects in cell culture and animal models of AD.[67,68]

Second, our analysis also identified 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.[65] Flupentixol, on the other hand, is a 5-hydroxytryptamine receptor antagonist, which has been suggested as a potential treatment for cognitive deficits in AD.[75,76]

We also identified other drugs as potential candidates for AD treatment, including *loperamide* (DrugBank ID: DB00836), *cyproheptadine* (DrugBank ID: DB00434), *peginterferon alfa-2b* (DrugBank ID: DB00022), *apomorphine* (DrugBank ID: DB00714), and *enoxacin* (DrugBank ID: DB00467). Loperamide, commonly used to treat diarrhea, has been shown to target opioid receptors, which may be linked to AD pathology,[58,59] which has been suggested to be potentially linked to AD pathology.[60] Cyproheptadine, a histamine antagonist, has been demonstrated to reduce cognitive symptoms in AD.[69] 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 (Janus Kinase/Signal Transducer and Activator of Transcription) pathway, which has been suggested as a potential target for AD.[70,71] Apomorphine, a dopamine receptor agonist for Parkinson's disease, has been shown to protect against oxidative stress, which plays a role in AD pathology[72] and improve memory function in AD.[73,74] Enoxacin, a fluoroquinolone used to treat bacterial infections, has been suggested to potentially decrease the risk of developing AD when used appropriately with other antibiotics, such as macrolides and fluoroquinolones.[77]

DISCUSSIONS

In this study, we built a comprehensive BKG called iBKH, through collecting, cleaning and normalizing raw data from diverse information sources. To date, iBKH has incorporated biomedical knowledge from 18 diverse information sources. In addition to the entity types that are popular in existing BKGs, such as genes, diseases, drugs, pathways, etc., iBKH also involves other complementary sources such as iDISK,[40] the supplement knowledge base we curated recently. We have made iBKH publicly available in both tabular format as CSV files for sophisticated users who can work with these source files, as well as Neo4j based on which we developed a web-based graphical portal to allow user-friendly knowledge retrieval and exploration. We would continuously enrich the content of iBKH and improve its graphical user interface (GUI) in the future.

In addition, we have also implemented a graph inference engine based on DGL-KE (Deep Graph Library - Knowledge Embedding)[26] in iBKH to facilitate novel biomedical knowledge discovery. As a proof of concept, we demonstrated the application of iBKH for in-silico hypothesis generation for AD drug repurposing. We observed good quantitative performance of iBKH on drugs that have already been approved and on clinical trial for treating AD. We have also identified novel potentially repurposable drugs for AD with evidence

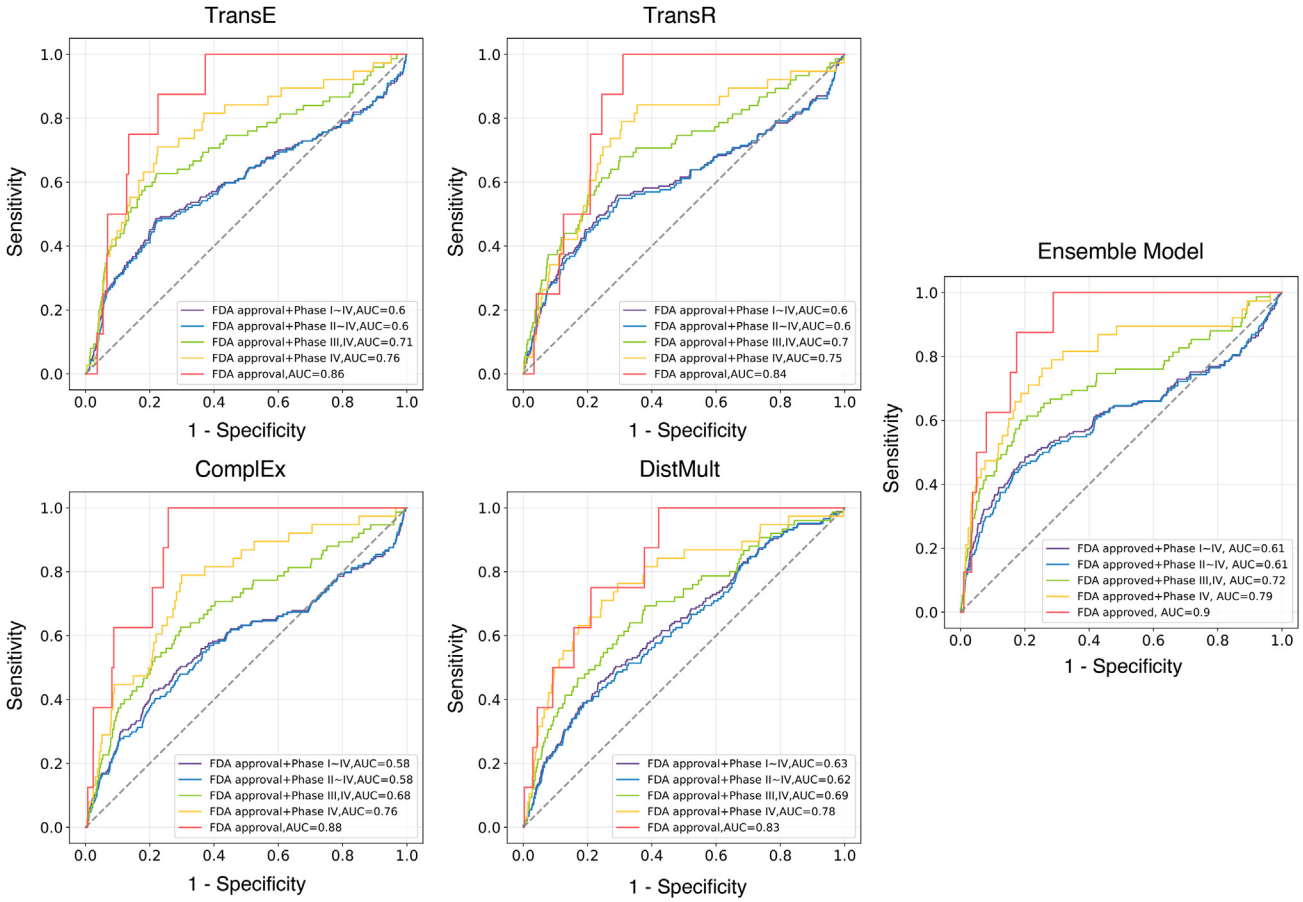


Figure 4. Model performance of in-silicon Alzheimer's disease drug repurposing

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.

supported by existing literature. It is worthy of mentioning that iBKH can be flexibly extended to drug repurposing for other diseases, as well as predicting other types of biomedical relations, 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 extensive efforts of raw data file extraction and pre-processing, data annotation, as well as terminology normalization, which is not error free. To maximally reduce the probability of error in iBKH, we leveraged the well-curated biomedical vocabularies such as the unified medical language system (UMLS) to facilitate entity term normalization and conducted multiple rounds of manual review based on random sampling with replacement. In addition, we will also conduct graph learning-based knowledge graph refinement/completion to address this issue.[78]

Second, although iBKH has collected and integrated data from diverse sources, the information contained therein can still be incomplete due to the volume and speed of the new biomedical knowledge that has been generated day by day. In this context, efforts on deriving knowledge from biomedical literature[79–81] or real-world data such as the EHR (electronic health records)[6,82] would be critical. Moreover, we will make curating and adding new information into iBKH a continuous effort.

Third, like many existing BKGs, iBKH mainly focused on the general biomedical knowledge, which means it may lack fine-grained knowledge for studying particular diseases. On this aspect, there has been research

Table 5. 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,[58,59] which has been suggested to be potentially linked to AD pathology.[60]
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 cerebrospinal fluid (CSF) flow. Recent studies have demonstrated that the use of β-blockers is associated with reduced risk of AD onset[61] and functional decline in AD.[62]
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.[63] Recent drug repurposing studies have also suggested phenoxybenzamine as repurposable drug candidate to treat AD.[64,65]
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.[66] Though the usefulness of CCBs in AD remains controversial, it has shown multiple beneficial effects cell culture and animal models of AD.[67,68]
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.[69]
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.[70,71]
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.[72] Emerging evidence showed that Apomorphine has a significant impact on improving memory function in AD.[73,74]
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.[65]

(*Continued on next page*)

Table 5. ***Continued***

Rank	DrugBank ID	Drug Name	Category	Description	Notes
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.[75,76]
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.[77]

on building disease specific BKGs. For instance, Coronavirus Disease-knowledge graph (COVID-KG)[83] contained knowledge on COVID-19; knowledge graph for Hepatocellular Carcinoma (KGHC)[9] is constructed focusing on hepatocellular carcinoma. In the future, we will further enhance iBKH by incorporating more detailed knowledge on specific diseases.

Last but not the least, it is important to validate the novel knowledge discovered from iBKH, which is not supported in the current portal. As a related effort, we have built a biomedical evidence generation engine based on literature mining,[84] which can retrieve and synthesize evidence supporting particular hypotheses from state-of-the-art scientific publications. 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 computational clinical trial emulation approach based on real-world clinical data.[85,86]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Overview
  - Raw data processing
  - Term harmonization
  - Knowledge integration
  - Integrative biomedical knowledge Hub deployment with Neo4j
  - Graphical portal for fast knowledge retrieval
  - iBKH-based knowledge discovery
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluation of link prediction for KGE models
  - Evaluation of AD drug repurposing performance

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from National Science Foundation (NSF) 1750326, 2212175 and National Institutes of Health (NIH) R01AG076234, R01AG076448, RF1AG072449, R01AG080991, R01AG080624, R01AG078154, R56AG069880, and R01AT009457 for this research.

AUTHOR CONTRIBUTIONS

F.W. contributed for conceptualization, investigation, writing, reviewing, and editing of the manuscript. C.S. contributed for investigation, drafting, editing, and reviewing manuscript. Y.H. led the effort on data preparation, knowledge graph construction, data analysis and web interface implementation. S.R., J.M., Z.A. contributed for improving data standardization and organization, efficiency of UI, and language of the manuscript. M.Z., H.Z., F.F.C., and G.G. contributed for data collection and data preparation. A.C. contributed for KGE implementation. Z.B. contributed for critical discussion on constructing iBKH. W.G. contributed for knowledge graph quality check. J.T. and Y.L. contributed for critical discussion on KGE algorithms. F.C., R.Z., S.D., and J.B. contributed 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 INTERESTS

The authors declare no competing interests.

Received: June 16, 2022
Revised: September 20, 2022
Accepted: March 16, 2023
Published: March 21, 2023

REFERENCES

1. Nicholson, D.N., and Greene, C.S. (2020). Constructing knowledge graphs and their biomedical applications. Comput. Struct. Biotechnol. J. *18*, 1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017.

2. 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. Nat. Biotechnol. *40*, 692–702. https://doi.org/10.1038/s41587-021-01145-6.

3. 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. Nat. Commun. *10*, 3045. https://doi.org/10.1038/s41467-019-11069-0.

4. 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. https://doi.org/10.7554/eLife.26726.

5. 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. Sci. Data *6*, 151. https://doi.org/10.1038/s41597-019-0152-0.

6. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., and Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. Sci. Rep. *7*, 5994. https://doi.org/10.1038/s41598-017-05778-z.

7. 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 J. *26*, 2737–2750. https://doi.org/10.1177/1460458220937101.

8. 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 Res. *47*, D1090–D1101. https://doi.org/10.1093/nar/gky1042.

9. Li, N., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., and Wang, J. (2020). KGHC: a knowledge graph for hepatocellular carcinoma. BMC Med. Inform. Decis. Mak. *20*, 135. https://doi.org/10.1186/s12911-020-1112-5.

10. Percha, B., and Altman, R.B. (2018). A global network of biomedical relationships derived from text. Bioinformatics *34*, 2614–2624. https://doi.org/10.1093/bioinformatics/bty114.

11. Ernst, P., Siu, A., and Weikum, G. (2015). KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinf. *16*, 157. https://doi.org/10.1186/s12859-015-0549-5.

12. 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. Knowl. Inf. Syst. *62*, 317–336. https://doi.org/10.1007/s10115-019-01351-4.

13. Rubin, D.L., Shah, N.H., and Noy, N.F. (2008). Biomedical ontologies: a functional perspective. Brief. Bioinform. *9*, 75–90. https://doi.org/10.1093/bib/bbm059.

14. 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 Biol. *6*, R46. https://doi.org/10.1186/gb-2005-6-5-r46.

15. Zhu, Y., Elemento, O., Pathak, J., and Wang, F. (2019). Drug knowledge bases and their applications in biomedical informatics research. Brief. Bioinform. *20*, 1308–1321. https://doi.org/10.1093/bib/bbx169.

16. Callahan, T.J., Tripodi, I.J., Pielke-Lombardo, H., and Hunter, L.E. (2020). Knowledge-based biomedical data science. Annu. Rev. Biomed. Data Sci. *3*, 23–41. https://doi.org/10.1146/annurev-biodatasci-010820-091627.

17. Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. *13*, 55–75. https://doi.org/10.1109/MCI.2018.2840738.

18. Chen, X., Jia, S., and Xiang, Y. (2020). A review: knowledge reasoning over knowledge graph. Expert Syst. Appl. *141*, 112948. https://doi.org/10.1016/j.eswa.2019.112948.

19. Su, C., Tong, J., Zhu, Y., Cui, P., and Wang, F. (2018). Network embedding in biomedical data science. Brief. Bioinform. *21*, 182–197. https://doi.org/10.1093/bib/bby117.

20. Mohamed, S.K., Nounu, A., and Nováček, V. (2021). Biological applications of knowledge graph embedding models. Brief. Bioinform. *22*, 1679–1693. https://doi.org/10.1093/bib/bbaa012.

21. Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., and Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. J. Biomed. Inform. *115*, 103696. https://doi.org/10.1016/j.jbi.2021.103696.

22. Zhou, Y., Wang, F., Tang, J., Nussinov, R., and Cheng, F. (2020). Artificial intelligence in COVID-19 drug repurposing. Lancet. Digit. Health *2*, e667–e676. https://doi.org/10.1016/S2589-7500(20)30192-8.

23. 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. https://doi.org/10.1007/978-981-16-6054-2_24.

24. 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 Genom. Bioinform. *3*, lqab078. https://doi.org/10.1093/nargab/lqab078.

25. 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 Res. *49*, W153–W161. https://doi.org/10.1093/nar/gkab449.

26. 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. https://doi.org/10.1145/3397271.3401172.

27. 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 Res. *49*, D498–D508. https://doi.org/10.1093/nar/gkaa1025.

28. 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. J. Biomed. Semantics *7*, 44. https://doi.org/10.1186/s13326-016-0088-7.

29. 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 Res. *40*, D940–D946. https://doi.org/10.1093/nar/gkr972.

30. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., and Haendel, M.A. (2012). Uberon, an integrative multi-species anatomy ontology. Genome Biol. *13*, R5. https://doi.org/10.1186/gb-2012-13-1-r5.

31. Bastian, F.B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S.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 Res. *49*, D831–D847. https://doi.org/10.1093/nar/gkaa793.

32. 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 Res. *47*, D948–D954. https://doi.org/10.1093/nar/gky868.

33. 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 Res. *46*, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

34. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

35. 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 Res. *30*, 163–165. https://doi.org/10.1093/nar/30.1.163.

36. 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 Res. *46*, D649–D655. https://doi.org/10.1093/nar/gkx1132.

37. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The SIDER database of drugs and side effects. Nucleic Acids Res. *44*, D1075–D1079. https://doi.org/10.1093/nar/gkv1075.

38. 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. https://doi.org/10.1093/database/bay003.

39. 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/.

40. 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. J. Am. Med. Inform. Assoc. *27*, 539–548. https://doi.org/10.1093/jamia/ocz216.

41. 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 Res. *47*, D786–D792. https://doi.org/10.1093/nar/gky930.

42. 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 Res. *40*, D1100–D1107. https://doi.org/10.1093/nar/gkr777.

43. 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 Res. *38*, D249–D254. https://doi.org/10.1093/nar/gkp886.

44. Liu, C.-C., Liu, C.C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat. Rev. Neurol. *9*, 106–118. https://doi.org/10.1038/nrneurol.2012.263.

45. Strittmatter, W.J., and Roses, A.D. (1995). Apolipoprotein E and alzheimer disease. Proc. Natl. Acad. Sci. USA *92*, 4725–4727. https://doi.org/10.1073/pnas.92.11.4725.

46. 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*, 1164. https://doi.org/10.3390/biom10081164.

47. 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. J. Neurosci. *41*, 3025–3038. https://doi.org/10.1523/JNEUROSCI.1980-20.2020.

48. Aisen, P.S. (2002). The potential of anti-inflammatory drugs for the treatment of Alzheimer's disease. Lancet Neurol. *1*, 279–284. https://doi.org/10.1016/S1474-4422(02)00133-3.

49. Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: a survey of approaches and applications. IEEE Trans. Knowl. Data Eng. *29*, 2724–2743. https://doi.org/10.1109/TKDE.2017.2754499.

50. Bordes A., Usunier N., Garcia-Duran A., Weston J., Yakhnenko O. Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems (NIPS 2013); Curran Associates, Inc., 2013. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

51. Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In Twenty-ninth AAAI conference on artificial intelligence. https://ojs.aaai.org/index.php/AAAI/article/view/9491.

52. Théo, T., Johannes, W., Sebastian, R., Eric, G., and Guillaume, B. (2016). Complex Embeddings for Simple Link Prediction. Proceedings of The 33rd International Conference on Machine Learning, held in New York, USA, 2016/06/11. (JMLR.org), pp. 2071–2080. http://proceedings.mlr.press/v48/trouillon16.html?ref=https://githubhelp.com.

53. 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. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6575.

54. 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 Res. Ther. *14*, 7–23. https://doi.org/10.1186/s13195-021-00951-z.

55. 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 Res. Ther. *13*, 24. https://doi.org/10.1186/s13195-020-00760-w.

56. 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. Nat. Aging *1*, 1175–1188. https://doi.org/10.1038/s43587-021-00138-z.

57. 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. J. Proteome Res. *19*, 4624–4636. https://doi.org/10.1021/acs.jproteome.0c00316.

58. 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. J. Pharmacol. Exp. Ther. *289*, 494–502.

59. 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 Sci. *33*, 315–318. https://doi.org/10.1016/0024-3205(83)90506-4.

60. Cai, Z., and Ratka, A. (2012). Opioid system and Alzheimer's disease. NeuroMolecular Med. *14*, 91–111. https://doi.org/10.1007/s12017-012-8180-3.

61. Beaman, E.E., Bonde, A.N., Larsen, S.M.U., Ozenne, B., Lohela, T.J., Nedergaard, M., Gíslason, G.H., Knudsen, G.M., and Holst, S.C. (2023). Blood–brain barrier permeable β-blockers linked to lower risk of Alzheimer's disease in hypertension. Brain *146*, 1141–1151. https://doi.org/10.1093/brain/awac076.

62. 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. Am. J. Geriatr. Psychiatry *16*, 883–892. https://doi.org/10.1097/JGP.0b013e318181276a.

63. 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. Int. J. Mol. Sci. *15*, 1402–1417. https://doi.org/10.3390/ijms15011402.

64. 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. Transl. Psychiatry *9*, 220. https://doi.org/10.1038/s41398-019-0555-x.

65. 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. https://doi.org/10.1093/bioinformatics/btaa563.

66. Bojarski, L., Herms, J., and Kuznicki, J. (2008). Calcium dysregulation in Alzheimer's disease. Neurochem. Int. *52*, 621–633. https://doi.org/10.1016/j.neuint.2007.10.002.

67. Anekonda, T.S., and Quinn, J.F. (2011). Calcium channel blocking as a therapeutic strategy for Alzheimer's disease: the case for isradipine. Biochim. Biophys. Acta *1812*, 1584–1590. https://doi.org/10.1016/j.bbadis.2011.08.013.

68. 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. Rev. Neurosci. *25*, 231–246. https://doi.org/10.1515/revneuro-2013-0056.

69. 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.

70. 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*, 425. https://doi.org/10.3390/cells8050425.

71. 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. Ann. Neurosci. *28*, 191–200. https://doi.org/10.1177/09727531211070532.

72. Perry, G., Cash, A.D., and Smith, M.A. (2002). Alzheimer disease and oxidative stress. J. Biomed. Biotechnol. *2*, 120–123. https://doi.org/10.1155/S1110724302203010.

73. 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. J. Alzheimer's Dis. *58*, 1151–1161. https://doi.org/10.3233/JAD-160344.

74. Himeno, E., Ohyagi, Y., Ma, L., Nakamura, N., Miyoshi, K., Sakae, N., Motomura, K., Soejima, N., Yamasaki, R., Hashimoto, T., et al. (2011). Apomorphine treatment in Alzheimer mice promoting amyloid-β degradation. Ann. Neurol. *69*, 248–256. https://doi.org/10.1002/ana.22319.

75. 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.

76. 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. J. Med. Chem. *57*, 7160–7181. https://doi.org/10.1021/jm5003952.

77. 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. Front. Aging Neurosci. *13*, 701899. https://doi.org/10.3389/fnagi.2021.701899.

78. Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., and Duan, Z. (2020). Knowledge graph completion: a review. IEEE Access *8*, 192435–192456. https://doi.org/10.1109/ACCESS.2020.3030076.

79. 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. https://doi.org/10.1093/bioinformatics/btt359.

80. 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. https://doi.org/10.1093/bioinformatics/btx659.

81. Zhao, S., Su, C., Lu, Z., and Wang, F. (2021). Recent advances in biomedical literature mining. Brief. Bioinform. *22*, bbaa057. https://doi.org/10.1093/bib/bbaa057.

82. 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. https://doi.org/10.1142/9789811215636_0003.

83. Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W., et al. (2021). COVID-19 literature knowledge graph construction and drug repurposing report generation. In 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2007.00576.

84. Zhao, S., Wang, A., Qin, B., and Wang, F. (2022). Biomedical evidence engineering for data-driven discovery. Bioinformatics *38*, 5270–5278. https://doi.org/10.1093/bioinformatics/btac675.

85. Ozery-Flato, M., Goldschmidt, Y., Shaham, O., Ravid, S., and Yanover, C. (2020). Framework for identifying drug repurposing candidates from observational healthcare data. JAMIA Open *3*, 536–544. https://doi.org/10.1093/jamiaopen/ooaa048.

86. Liu, R., Wei, L., and Zhang, P. (2021). A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. Nat. Mach. Intell. *3*, 68–75. https://doi.org/10.1038/s42256-020-00276-w.

87. Bodenreider, O. (2004). The unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. *32*, D267–D270. https://doi.org/10.1093/nar/gkh061.

88. Goyal, P., and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: a survey. Knowl. Based. Syst. *151*, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022.

89. 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. Brief. Bioinform. *22*, bbaa344. https://doi.org/10.1093/bib/bbaa344.

90. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., and Sun, J. (2017). Gram: graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax, NS, Canada; Association for Computing Machinery). https://doi.org/10.1145/3097983.3098126.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Bgee	Bastian et al., 2021[31]	https://bgee.org/
Brenda Tissue Ontology	Chang et al., 2021[27]	https://www.brenda-enzymes.org/index.php
Cell Ontology	Diehl et al., 2016[28]	http://obofoundry.org/ontology/cl.html
Comparative Toxicogenomics Database (CTD)	Davis et al., 2019[32]	http://ctdbase.org/
ChEMBL	Gaulton et al., 2012[42]	https://www.ebi.ac.uk/chembl/
Chemical Entities of Biological Interest (ChEBI)	de Matos et al., 2010[43]	https://www.ebi.ac.uk/chebi/init.do
Drug Repurposing Knowledge Graph (DRKG)	Ioannidis et al., 2020[39]	https://github.com/gnn4dr/DRKG
Disease Ontology	Schriml et al., 2012[29]	https://disease-ontology.org/
DrugBank	Wishart et al., 2018[33]	https://go.drugbank.com/
Hetionet	Himmelstein et al., 2017[4]	https://github.com/hetio/hetionet
HUGO Gene Nomenclature Committee (HGNC)	Braschi et al., 2019[41]	https://www.genenames.org/
Integrated Dietary Supplement Knowledge Base (iDISK)	Rizvi et al., 2020[40]	https://conservancy.umn.edu/handle/11299/204783
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Kanehisa and Goto, 2000[34]	https://www.kegg.jp/
Pharmacogenetics Knowledge Base (PharmGKB)	Hewett et al., 2002[35]	https://www.pharmgkb.org/
Reactome	Fabregat et al., 2018[36]	https://reactome.org/
Side effect resource (SIDER)	Kuhn et al., 2016	http://sideeffects.embl.de/
TISSUE	Palasca et al., 2018[38]	https://tissues.jensenlab.org/
Uberon	Mungall et al., 2012[30]	https://www.ebi.ac.uk/ols/ontologies/uberon
Unified Medical Language System (UMLS)	Bodenreider, 2004[87]	https://www.nlm.nih.gov/research/umls/index.html
iBKH source files	This paper	https://github.com/wcm-wanglab/iBKH/tree/main/iBKH
iBKH portal	This paper	http://ibkh.ai/
Software and algorithms		
Neo4j	Neo4j, Inc.	https://neo4j.com
Python	Python Software Foundation	https://www.python.org
Django	Django Software Foundation	https://www.djangoproject.com
neovis.js	Neo4j Contrib	https://github.com/neo4j-contrib/neovis.js
D3.js	Mike Bostock	https://d3js.org
DGL-KE (Deep Graph Library – Knowledge Graph Embedding)	Zheng et al., 2020[26]	https://github.com/awslabs/dgl-ke
iBKH construction and iBKH-based knowledge discovery	This paper	https://github.com/wcm-wanglab/iBKH

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 format are publicly available online at https://github.com/wcm-wanglab/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.
- 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 BKG 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 CSV, tab-separated values, TXT, EXCEL tablet, Hypertext Markup Language (HTML), Resource Description Framework, 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. 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 improve the entity vocabulary one by one. The process of constructing the linkage pool for each entity type primarily depended on two procedures: 1) The term dictionary of existing sources was utilized. For example, Disease Ontology[29] provides Disease Ontology ID, Medical Subject Heading (Medical Subject Headings (MeSH)) ID, UMLS[87] Concept Unique Identifier (CUI) for each disease entity. 2) The UMLS Application Program Interfaces served as the bridge for term normalization for terms that could not be mapped with the existing term dictionary. To ensure data quality, synonyms were obtained for each term with strict term types such as "preferred" and "abbreviation." Finally, multiple rounds of manual quality checks were conducted.

For **gene** entity type, we used the HGNC gene repository[41] 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[33] 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, MeSH terms, MeSH term IDs, UMLS[87] Concept Unique Identifiers (CUIs), and the drug names in UMLS.

For **molecule** entity type, we used the ChEMBL,[42] 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[37] and described them by using the UMLS CUIs.

For **disease** entity type, we used the Disease Ontology[29] 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[4] and iDISK,[40] 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,[36] 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[30] 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 **DSI**, **DSP**, and **TC** entities, data were collected from our previous curated iDISK.[40] 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.

Integrative biomedical knowledge Hub deployment with Neo4j

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. 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

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

Step 1, 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.[19,49,88] 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,[1,7,21,89] multi-omics data analysis,[1,2] clinical data (e.g., electronic healthcare record) analysis,[3,90] and knowledge extraction from biomedical literature.[83] In this work, we used the Deep Graph Library - Knowledge Embedding (DGL-KE) (https://github.com/awslabs/dgl-ke),[26] a Python-based implementation for the advanced KGE algorithms, such as TransE,[50] TransR,[51] ComplEx,[52] and DistMult.[53] 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 score 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)

(B) **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 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,[50] TransR,[51] ComplEx,[52] and DistMult.[53] 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluation of link prediction for KGE models

Model evaluation. We randomly split all triplets of iBKH into 90% training set and 10% testing set. The training set was used to train the KGE algorithm and the testing set was used to evaluate model performance. We assessed model performance in link prediction using the standard metrics including:

- $Hit@k = \frac{1}{Q}\sum_{1}^{Q}\mathbb{I}_{rank_i \leq k}$ (k = 1, 3, or 10), which measures the average number of times the positive triplet is among the k highest ranked triplets;
- $MR = \frac{1}{Q}\sum_{1}^{Q} rank_i$, i.e., Mean Rank, is the average rank of the positive triplets;
- $MRR = \frac{1}{Q}\sum_{1}^{Q}\frac{1}{rank_i}$, i.e., Mean Reciprocal Rank, is the average reciprocal rank of the positive instances.

where, Q is the total number of positive triplets and $\mathbb{I}_{rank_i \leq k}$ is 1 if $rank_i \leq k$, otherwise it is 0. Higher values of $Hit@k$ and MRR and a lower value of MR indicate good performance, and vice versa.

After that, we re-trained KGE algorithms using all triplets in iBKH.

Evaluation of AD drug repurposing performance

To evaluate performance of AD drug repurposing, we used FDA-approved drugs as ground truth and produce the receiver operating characteristic curve (ROC) and the area under ROC (AUC) score.