Original Paper

The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research

Joseph D Romano[1,2,3], MA, MPhil, PhD; Van Truong[1,4,5], MS; Rachit Kumar[1,4,5,6], BS; Mythreye Venkatesan[7], BE, MS; Britney E Graham[7], PhD; Yun Hao[1,4], PhD; Nick Matsumoto[7], BA; Xi Li[7], MS; Zhiping Wang[7], MS, PhD; Marylyn D Ritchie[1,3,5], PhD; Li Shen[1,3], PhD; Jason H Moore[7], PhD

[1]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[2]Center of Excellence in Environmental Toxicology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[3]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[4]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[5]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[6]Medical Scientist Training Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States
[7]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States

Corresponding Author:
Joseph D Romano, MA, MPhil, PhD
Institute for Biomedical Informatics
Perelman School of Medicine
University of Pennsylvania
403 Blockley Hall
423 Guardian Drive
Philadelphia, PA, 19104
United States
Phone: 1 2155735571
Email: joseph.romano@pennmedicine.upenn.edu

Abstract

Background: As global populations age and become susceptible to neurodegenerative illnesses, new therapies for Alzheimer disease (AD) are urgently needed. Existing data resources for drug discovery and repurposing fail to capture relationships central to the disease's etiology and response to drugs.

Objective: We designed the Alzheimer's Knowledge Base (AlzKB) to alleviate this need by providing a comprehensive knowledge representation of AD etiology and candidate therapeutics.

Methods: We designed the AlzKB as a large, heterogeneous graph knowledge base assembled using 22 diverse external data sources describing biological and pharmaceutical entities at different levels of organization (eg, chemicals, genes, anatomy, and diseases). AlzKB uses a Web Ontology Language 2 ontology to enforce semantic consistency and allow for ontological inference. We provide a public version of AlzKB and allow users to run and modify local versions of the knowledge base.

Results: AlzKB is freely available on the web and currently contains 118,902 entities with 1,309,527 relationships between those entities. To demonstrate its value, we used graph data science and machine learning to (1) propose new therapeutic targets based on similarities of AD to Parkinson disease and (2) repurpose existing drugs that may treat AD. For each use case, AlzKB recovers known therapeutic associations while proposing biologically plausible new ones.

Conclusions: AlzKB is a new, publicly available knowledge resource that enables researchers to discover complex translational associations for AD drug discovery. Through 2 use cases, we show that it is a valuable tool for proposing novel therapeutic hypotheses based on public biomedical knowledge.

(J Med Internet Res 2024;26:e46777) doi: 10.2196/46777

KEYWORDS

Alzheimer disease; knowledge graph; knowledge base; artificial intelligence; drug repurposing; drug discovery; open source; Alzheimer; etiology; heterogeneous graph; therapeutic targets; machine learning; therapeutic discovery

Introduction

Background

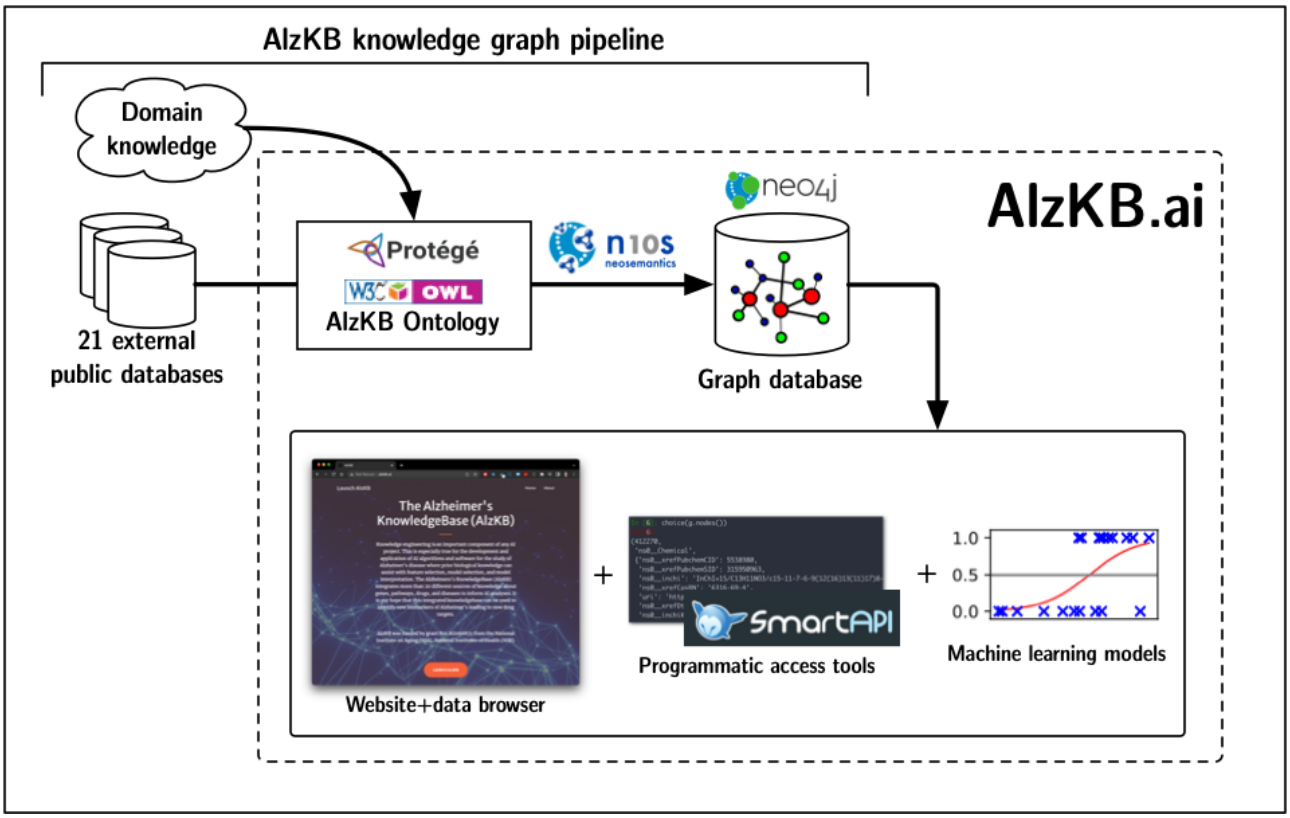
Alzheimer disease (AD) is a progressive, neurodegenerative disease affecting an estimated 6.5 million Americans aged ≥65 years and represents a significant clinical, economic, and emotional burden worldwide [1]. AD is often cited as one of the greatest health care problems of the 21st century, particularly in high-income nations with an increasing proportion of older adults. Despite its societal impact, effective pharmaceutical treatments for AD remain notoriously elusive. The US Food and Drug Administration has approved 5 drugs for the treatment of AD, 4 of which (donepezil, rivastigmine, galantamine, and memantine) only temporarily treat symptoms but do not alter the overall progression of the disease [2], whereas the fifth (aducanumab) is highly controversial in terms of evidence of effectiveness and its safety profile [3]. AD researchers have prioritized the discovery and approval of new therapies for the disease both in terms of newly discovered compounds and by repurposing drugs that are already approved to treat other (non-AD) human diseases.

AD is associated with substantial changes in pathology, including the presence of neuritic plaques associated with the amyloid-β protein, extracellular deposition of amyloid-β, and neurofibrillary tangles. Previous research has shown that these neuropathological changes begin to occur years before clinical symptoms are apparent [4,5]. Despite decades of research, why this pathology begins to develop remains largely unknown [6]. Current consensus is that AD risk is multifactorial. The most well-established risk factors include age; family history; and certain genetic factors, especially the presence of the σ4 allele of the *apolipoprotein E* gene, which is involved in fat metabolism and cholesterol transport. However, the exact mechanism through which these factors—including *APOE*-σ4 presence—cause or contribute to AD risk is unknown [7].

Of the many techniques used in AD therapeutics research, there is a wealth of computer-aided approaches that leverage recent advances in bioinformatics, epidemiology, artificial intelligence (AI), and machine learning (ML). For example, Rodriguez et al [8] developed an ML framework to assess gene lists constructed by differential gene expression data in response to drug treatment to determine whether those drugs would be candidates for repurposing in AD. Tsuji et al [9] used an autoencoder neural network to perform dimensionality reduction of a high-density protein interaction network to identify new possible drug targets and then found drugs associated with those targets. Genome-wide association studies have long been used for the identification of genes that confer AD risk, particularly for rare genes or genes with small (but statistically significant) contributions to disease risk [10].

In this paper, we describe the design and deployment of a major new knowledge resource for computational AD research—named The Alzheimer's Knowledge Base (AlzKB) [11]—with a particular focus on drug discovery and drug repurposing. The overall structure and contents of AlzKB are summarized in Figure 1. At its core, AlzKB consists of a large, heterogeneous graph database describing entities related to AD at multiple levels of biological organization, with rich semantic relationships describing how those entities are linked to one another. To demonstrate its value, we present two data-driven analyses involving ML on AlzKB's knowledge graph: (1) predicting Parkinson disease (PD) genes that may also be associated with AD and (2) generating and explaining drug repurposing hypotheses for treating AD, both of which replicate existing knowledge while proposing entirely novel directions for future experimental validation. AlzKB is free, open source, and publicly available [11] and consists entirely of publicly sourced knowledge integrated from 22 diverse web-based biomedical databases. We hypothesized that the relationships and entities in AlzKB contain valuable knowledge that cannot be effectively captured in existing data resources, with the additional advantage of improving the explainability of new predictions.

Figure 1. Schematic overview of the Alzheimer's Knowledge Base (AlzKB).



Existing Graph-Based Approaches to AD Research

Due to the increased popularity and success of analyses using integrated knowledge, previous efforts have used knowledge graphs in AD research for a variety of purposes, including drug repurposing [12-14] and gene identification [15] and as general informational resources [16]. Similar to AlzKB, these bodies of work draw from a variety of sources to construct the underlying knowledge graphs, including scientific literature and formally structured biomedical databases. Some, including the Alzheimer Disease Knowledge Graph [14] and the Heterogeneous network-based data set for AD [16], have been released as publicly accessible resources similar to AlzKB. Other studies have used existing resources not specifically intended for AD research (such as the Semantic MEDLINE Database [13]) to answer questions related to AD. To our knowledge, AlzKB is the largest graph-based knowledge representation that focuses solely on AD and draws from the greatest number of source databases. For comparison, the next largest AD-specific knowledge graph that we are aware of is AD-KG, which contains 30,729 nodes and 398,544 edges (compared to AlzKB's 118,902 nodes and 1,309,527 edges). Our emphasis on merging similar nodes or edges and cleaning the graph structure using an underlying biomedical ontology reduces the amount of noise that tends to be associated with many different node or edge types in a single graph, enabling more robust inference about relationships in AD, especially when used with emerging graph ML algorithms. Furthermore, AlzKB offers a public, web interface that allows for easy access and application to new research questions, whereas existing resources have either restricted access or are entirely unavailable for reuse. Given the challenge of identifying new or repurposed drugs for etiologically complex diseases such as AD, AlzKB represents a major step forward by improving both quantitatively and structurally on existing resources.

Methods

AlzKB Ontology

Graph databases are renowned for their flexibility in representing data that do not conform to a rigid, tabular structure, but this comes at the expense of implicitly enforcing consistency and semantic standardization [17]. To mitigate this issue, we designed a Web Ontology Language (OWL) 2 ontology—describing the types of entities relevant to AD and treatment of AD, as well as the types of relationships that link those entities—that serves as a *template* for nodes and edges in the knowledge graph. Ontologies (including OWL 2 ontologies) are formal representations of knowledge that are frequently used in biomedicine to computationally structure, retrieve, and make inferences about knowledge within a domain of interest [18]. Briefly, as many of the components of a graph database have a 1-to-1 correspondence with components of an OWL 2 ontology (eg, OWL 2 classes are equivalent to graph database node labels, and OWL 2 object properties are equivalent to edge types in a graph database), it is possible to populate the ontology using biomedical knowledge and translate the contents of the populated ontology into an equivalent graph database. Therefore, enforcing consistency in the ontology becomes equivalent to enforcing consistency in the graph database.

We constructed the ontology manually using the Protégé ontology editor (version 5.5.0; Stanford Center for Biomedical Informatics Research) [19] following an iterative process guided

by expert domain knowledge. First, we prototyped a class hierarchy containing the types of nodes (eg, gene, disease, pathway, and drug) desired in the knowledge base. We then annotated these classes with data properties (eg, drugs can be assigned a property value corresponding to molecular weight) and object properties (relationship types that link 2 entities, such as "drug treats disease"). A thorough description of the components of OWL 2 ontologies is provided by Hitzler et al [20]. Finally, we placed restrictions on the ontology to reflect biology and clinical practice. For example, we specified restrictions stating that all pathways must contain one or more genes or that all drugs in the knowledge base must have a valid DrugBank ID. We repeated these steps several times, making revisions on previous iterations until several domain experts agreed that the semantic contents of the ontology were consistent with current AD knowledge and systems biology processes involved in AD etiology. After collecting the data sources used to populate the ontology (see the following section), we included additional data properties corresponding to identifiers in those source databases, enabling data provenance and facilitating both interoperability and validation. The final ontology structure consists of entity types involved in AD etiology (modeled as OWL 2 classes), types of semantic relationships that can link those entity types (modeled as OWL 2 object properties), and properties that can be annotated onto entities of specific types (modeled as OWL 2 data properties). Both before and after populating the ontology with individuals (see the *Implementing AlzKB* section), we validated its contents and structure by running FaCT++—an ontology inference engine that identifies errors by evaluating all assertions in the ontology against the ontology's class or property hierarchy and other restrictions [21].

Collecting and Assembling Third-Party Data Sources

Using the AlzKB ontology's class hierarchy as a starting point, we determined a set of the most important entity types to include in the first release of the knowledge base. For example, we prioritized inclusion of entities representing diseases (specifically AD and its various subtypes), genes, and drugs, among others. Similarly, we identified important relationship types (eg, "DRUG_BINDS_GENE" or "GENE_ASSOCIATED_WITH_DISEASE") to include in the knowledge base. For each of these entity and relationship types, we identified a third-party, public data source that would serve as a collection of "ground truth knowledge" for that entity or relationship type. In the assembled knowledge base, there is roughly a 1-to-1 correspondence between a data record in the original "ground truth" data source and its corresponding entity or relationship in AlzKB, with some important exceptions. For example, we made the decision to only include neurological diseases in AlzKB rather than all diseases described in the "ground truth" data source (in this case, the Disease Ontology). We also identified instances in which properties from additional data sources could be used to augment the "ground truth" entities. For example, while DrugBank is used to specify the drugs described in AlzKB, we also used fields from Distributed Structure-Searchable Toxicity and PubChem to augment the properties annotated onto drugs (such as molecular weight, chemical fingerprint, and synonyms).

Implementing AlzKB

We populated the ontology by sequentially carrying out the following steps:

1. Import distinct entities from each data source corresponding to the corresponding ontology class and define those entities as ontology individuals (ie, instances of that class). For example, the drug memantine is defined as an instance of the ontology class Drug.
2. Populate data properties for all instances of each ontology class using data from relevant sources. For example, memantine is annotated with the Chemical Abstracts Service Registry number 19982-08-2.
3. Populate object properties as the semantic relationships linking pairs of entities using the appropriate data source. For example, an object property of type "DRUG_TREATS_DISEASE" links memantine to the instance of Disease named Alzheimer's Disease.

After populating the AlzKB ontology with entities, relationships, and data properties, we serialized the ontology into the Resource Description Framework (RDF) or XML graph data format, which is compatible with modern graph database software as an input format. A complete list of the data sources used in AlzKB at the time of writing is provided in Table 1. We then populated a Neo4j graph database (version 4.4.5; Neo4j, Inc) [22] with the contents of the RDF or XML file using the neosemantics library [23], which parses the RDF data, inserting semantic triples into the graph database corresponding to each entity or relationship. Finally, we stripped the newly populated graph database of unnecessary artifacts that are components of the OWL 2 standard, leaving only nodes, relationships, and properties defined within the hierarchy. For the publicly hosted version of AlzKB, we created a web server that hosts both the static AlzKB website (containing information, documentation, and use details) and the Neo4j graph database, which is available by navigating to a subdomain [24] of the main website [11]. For reproducibility, this entire pipeline (including mappings to source databases) is provided as a single Python script available on GitHub (the most recent version) [25] or Zenodo (an archived version of the code at the time of publication) [26].

Table 1. Third-party public data sources used in the Alzheimer's Knowledge Base (AlzKB), which data elements were used from them, and total size of the data source (counts of entities of relevant data types only)[a].

Data source	Use in AlzKB	Size (number of entities)
AOP-DB[b] [27]	Adverse outcome pathways and chemical-gene associations	1,207,456
Bgee[c] [28]	Tissue-specific gene expression data; only human gene expression data were used in AlzKB	9,093,494
Disease Ontology[c] [29]	Human diseases—only AD[d], subtypes of AD, and related neurodegenerative diseases were included in AlzKB	8043
DisGeNET [30]	Diseases, genes, and disease-gene associations with scores representing levels of evidence; only AD and related neurodegenerative terms were used for diseases	51,841
DrugBank [31]	On-market and experimental pharmaceutical drugs	15,550
EPA[e] DSSTox[f] [32]	Chemical toxicity data—filtered in AlzKB to drugs contained in DrugBank	1,200,059
EPA ACToR[g] [33]	Chemical toxicity data—filtered in AlzKB to drugs contained in DrugBank	504,871
Gene Ontology[c] [34,35]	Biological processes, molecular functions, and cellular components	42,950
GWAS[h] Catalog[c] [36]	Gene-disease associations	60,071
Hetionet [37]	Graph modeling and entity resolution (for data sources marked with footnote indicator "c")	47,031
Human Reference Protein Interactome Mapping Project[c] [38]	Human protein-protein interactions (modeled as gene-gene interactions)	9094
LINCS[i] L1000[c] [39]	Human differential gene expression data	7467 genes
NCBI[j] MeSH[k] [40]	Clinical and biomedical concepts (annotated to various node types)	Approximately 27,000
NCBI Entrez Gene [41]	Human genes and gene synonyms	62,407
Pathway Interaction Database[c] [42]	Pathways and gene-pathway membership	223
PharmacotherapyDB[c] [43]	Drug indications for human diseases	698
PubChem [44]	Chemical structures and identifiers—only chemicals in DrugBank were included in AlzKB	115,067,800
Reactome pathway database[c] [45]	Pathways and gene-pathway membership	1341
SIDER[c,l] [46]	Drug side effects (modeled as diseases)	5868
TISSUES[c] [47]	Tissue-specific gene expression data	__[m]
Uberon[c] [48]	Human anatomical structures	402
WikiPathways[c] [49]	Pathways and gene-pathway membership	298

[a]As source data elements do not correspond in a 1-to-1 manner with entities in the graph (eg, entities may be merged, filtered, or used as edges rather than nodes), actual counts for entities in AlzKB stratified by source are not available. The sizes are the best available estimates at the time of publication. Table 2 and Table S1 in Multimedia Appendix 1 [50-56] provide actual node and edge type counts in AlzKB.

[b]AOP-DB: Adverse Outcome Pathway Database.

[c]The derived data are structured in part using Hetionet.

[d]AD: Alzheimer disease.

[e]EPA: Environmental Protection Agency.

[f]DSSTox: Distributed Structure-Searchable Toxicity.

[g]ACToR: Aggregated Computational Toxicology Resource.

[h]GWAS: genome-wide association studies.

[i]LINCS: Library of Integrated Network-Based Cellular Signatures.

[j]NCBI: National Center for Biotechnology Information.

[k]MeSH: Medical Subject Headings.

[l]SIDER: Side Effect Resource.

[m]Counts not applicable (TISSUES associations map to edges rather than nodes in the graph).

Table 2. Node types and counts in the Alzheimer's Knowledge Base listed in descending order by prevalence. Additional node types will be added over time, and counts will increase as new data sources are incorporated or existing sources are updated to newer versions.

Node label	Total nodes, N
Gene	62,407
Drug	35,063
BiologicalProcess	11,381
Pathway	4570
MolecularFunction	2884
CellularComponent	1391
Symptom	438
BodyPart	402
DrugClass	345
Disease	20

Validating AlzKB Using Real-World Use Cases

After building AlzKB's knowledge graph, we designed two ML-based use cases that resemble real-world tasks for which AlzKB was originally designed: (1) proposing genetic targets for new drugs based on disease similarity and topological graph features and (2) predicting new edges in the knowledge graph linking AD to repurposed drugs via a graph completion model. These 2 use cases are intended to assess the external validity of AlzKB—for the ML models to perform well on tasks defined using real-world evaluation end points (eg, effective drugs or etiologically important genes), the informative patterns and phenomena underlying those end points need to be adequately captured in the knowledge graph.

In the first use case (identifying genetic targets via graph topology measures), we trained a random forest (RF) classifier (implemented in the scikit-learn library [Python Software Foundation] for the Python programming language) using the following topological graph features, which are computed for every node pair in the graph (regardless of whether an edge does or does not exist between them): common neighbors, total neighbors, preferential attachment, Adamic-Adar, and resource allocation [57-60]. Each feature gives a different measure of network "relatedness" for a pair of nodes, which are then used as predictive features in the RF model. For a given node pair (n_1, n_2), these metrics are defined as follows:

$$\mathrm{CN}(n_1, n_2) = |\mathrm{N}(n_1) \cap \mathrm{N}(n_2)|$$

$$\mathrm{TN}(n_1, n_2) = |\mathrm{N}(n_1) \cup \mathrm{N}(n_2)|$$

$$\mathrm{PA}(n_1, n_2) = |\mathrm{N}(n_1) \times \mathrm{N}(n_2)|$$

$$\mathrm{AA}(n_1, n_2) = \sum_{x \in \mathcal{N}(n_1) \cap \mathcal{N}(n_2)} \frac{1}{\log|\mathcal{N}(x)|}$$

$$\mathrm{RA}(n_1, n_2) = \sum_{x \in \mathcal{N}(n_1) \cap \mathcal{N}(n_2)} \frac{1}{|\mathcal{N}(x)|}$$

where $N(n_1)$ is the set of neighbor (adjacent) nodes of node *i*. Our training procedure for the RF model included 3-fold grid search cross-validation to optimize hyperparameters, an 80%/20% train/test split, and repeating the procedure 10 times with random sampling.

To accomplish the second use case (drug repurposing via graph completion models), we implemented and compared the performance of 5 graph completion algorithms applied to the entire AlzKB knowledge graph. These models learn low-dimensional representations of graph nodes as vector embeddings. The embeddings are then combined to propose all possible triples in the graph (source node, edge, and target node), and scores are generated to indicate the plausibility of the triple. The 5 models we evaluated are TransE, RotatE, DistMult, ComplEx, and ConvE [60].

We implemented the 5 models using PyKEEN—a Python library for knowledge graph embeddings [50]. We randomly split the data set of all triples into 80/10/10 training/validation/testing sets and used grid search to empirically set embedding dimensions to 256 and the number of epochs to 100 with early stopping allowed. All remaining hyperparameters were set to the PyKEEN defaults. We trained the models on Google Colab using a single Tesla T4 graphics processing unit and evaluated the results using the rank-based evaluation metrics hits@k (*k*=1, 3, and 10) and mean reciprocal rank (MRR) [61]. Ranking-based evaluation sorts the scores of triples in descending order and sets their rank as the index in the sorted list. In the case of multiple "true" triples having an equal score, we used the average of the most optimistic (best) and pessimistic (worst) ranks across the metrics. Briefly, hits@k is the ratio of true triples in the test set that have been ranked within the top *k* predictions of the model. Higher values indicate better performance. The MRR, also known as inverse harmonic mean rank, is the arithmetic mean of the inverse rank of the true

triples. We performed evaluation on both left- and right-side predictions (ie, how well they can predict missing entities in partial triples without either the head [source] or tail [target] entities).

Ethical Considerations

No human participants were involved in this research. All data used to build and evaluate AlzKB were derived from publicly available biomedical knowledge retrieved from open access databases. None of the data included were derived from individual human participants. Similarly, AlzKB is entirely open source and publicly available and complies with the licensing terms of all 22 source databases used to build the knowledge base.

Results

Knowledge Base Description

The first release of AlzKB (version 1.0) [26] contains 118,902 distinct nodes (representing biomedical entities) and 1,309,527 relationships linking those nodes. A full summary of node and relationship types with counts, respectively, is provided in Table 2 and Table S1 in Multimedia Appendix 1. Users can interact with AlzKB in their web browser using the built-in Neo4j interface or programmatically by connecting to the graph database over the internet. We also provide instructions for installing a local copy of the graph database as well as how to build the database from its original data sources.

Proposing New Therapeutic Targets for AD

As a proof of concept, we performed an analysis to predict whether known PD genes are also linked to AD etiology. PD is a chronic, progressive neurological disorder characterized by uncontrollable movements and possible mental and behavioral changes. Similar to AD, the precise etiology of PD is not fully understood, but the disease is characterized by the death or dysfunction of basal ganglia neurons. A growing body of work has established physiological and genetic similarities between PD and AD [62], and it has been proposed that drugs targeting PD genes could potentially treat AD as well. To approach this hypothesis computationally, we defined a binary classification task to predict whether gene nodes in the AlzKB knowledge graph are or are not AD genes [63]. To assemble the data set, we considered all gene nodes adjacent to AD as positive (n=101) and all gene nodes not adjacent to AD as negative (n=62,306). The negative samples are assumed to contain a mixture of true negatives and false negatives; in link prediction tasks, the goal is to recover the false negatives. We further filtered the negative nodes to omit PD genes (n=73) and orphan gene nodes (n=43,032) and down sampled the remaining genes to 303 (ie, 3 times the number of positive samples). To evaluate the performance, we used accuracy, balanced accuracy, precision, recall, F_1-score, area under the receiver operating characteristic curve, and area under the precision-recall curve, as shown in Figure 2.

The RF model predicted gene-disease relationships with an average balanced accuracy of 96.2% (precision=0.88; recall=0.98). We applied the trained models to predict PD genes that are likely to also be AD genes. Of the 73 PD genes in AlzKB, 8 (11%; *FYN*, *DCTN1*, *SNCA*, *SYNJ1*, *RSP12*, *ATXN2*, *KCNIP3*, and *CHRNB1*; described in Table 3) were predicted to be AD genes. A total of 10% (7/73) of the genes were predicted to be AD genes in all 10 models, whereas *CHRNB1* was predicted in 7 of the 10 models.

Figure 2. Random forest classifier performance (over 10 independent training runs) on the task of predicting whether Parkinson disease genes are also Alzheimer disease genes based on patterns of graph connectivity in the Alzheimer's Knowledge Base's heterogeneous knowledge graph. Across all metrics, a score of 1.00 represents the best possible performance. AUCPR: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

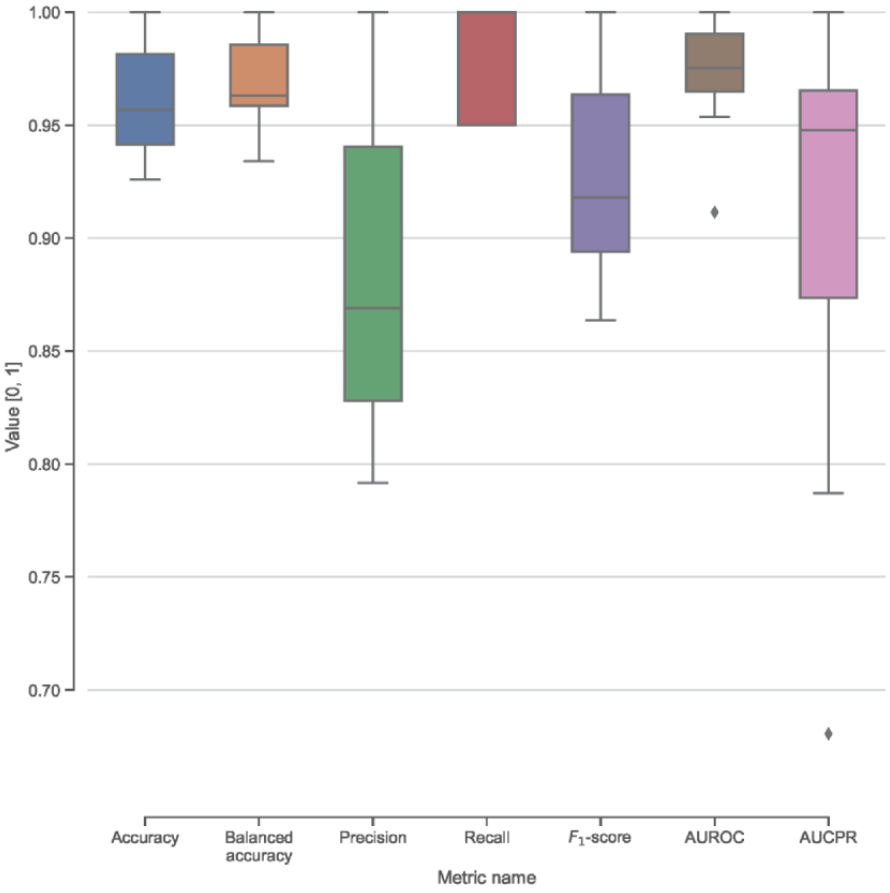


Table 3. Parkinson disease genes predicted by a graph-augmented random forest model to also be associated with Alzheimer disease.

Gene symbol	Gene name	Notes (from the Entrez Gene summary)
ATXN2	Ataxin-2	Modulates endocytosis, ribosomal translation, and mitochondrial function; aberrations are linked to diverse neurodegenerative diseases, diabetes, and obesity
CHRNB1	Cholinergic receptor nicotinic β1 subunit	Beta subunit of muscle acetylcholine receptor involved in transmitting signals at neuromuscular junction
DCTN1	Dynactin subunit 1	Dynactin is a macromolecular complex involved in many cellular functions, including the formation of neuronal pathways
FYN	FYN proto-oncogene, Src tyrosine kinase family	Membrane-associated tyrosine kinase involved in control of cell growth; highly expressed in brain tissue
KCNIP3	Potassium voltage-gated channel interacting protein 3	Voltage-gated potassium channel–interacting protein that is critical to neuronal excitability

Drug Repurposing via Graph Data Science

As a second use case, we considered the task of repurposing existing drugs—currently used to treat other diseases—based on patterns in the knowledge graph that suggest that they may also treat AD. To do this, we trained 5 state-of-the-art knowledge graph completion methods (TransE, RotatE, DistMult, ComplEx, and ConvE) [51] on AlzKB and selected the highest-performing one to predict links between drugs and AD. Additional details about the differences between these methods are provided in Multimedia Appendix 1.

The performance of the 5 different knowledge graph completion models is shown in Table 4. Among them, RotatE performed best, with the highest MRR and hits@k values. Therefore, we used RotateE to make predictions on the test set to obtain missing head entities with the template ([*drug*], DRUG_TREATS_DISEASE, AD). The top 10 predicted drugs are listed in Table 5 along with their current approved use and relevant clinical trial status pertaining to AD efficacy. Of the

top 10 predictions, 3 (30%) have been investigated in clinical trials to treat symptoms of AD. To further explore these predictions, we generated visualizations of a minimum spanning tree linking the 10 drugs to AD in AlzKB's knowledge graph, as shown in Figure 3. The visualization shows that the shortest paths between the drugs and AD are mediated by a small set of AD-associated genes, each of which is associated with one or more of the proposed drugs. The visualization is suggestive of interpretable biological mechanisms through which the drugs could act on AD etiology and provides hypotheses to further explore their validity.

Table 4. Ranking-based evaluation metrics of 5 embedding-based link prediction models on the Alzheimer's Knowledge Base knowledge graph. Metrics are derived from the likelihood of existing (known) links being predicted by the models. Higher scores indicate better performance.

Model name	Hits@1	Hits@3	Hits@10	MRR[a]
RotatE	*0.126* [b]	*0.220*	*0.358*	*0.202*
TransE	0.046	0.097	0.198	0.097
DistMult	0.027	0.056	0.126	0.061
ComplEx	0.074	0.142	0.263	0.136
ConvE	0.002	0.005	0.013	0.006

[a]MRR: mean reciprocal rank.

[b]Italicized values indicate maximum scores within a given column.

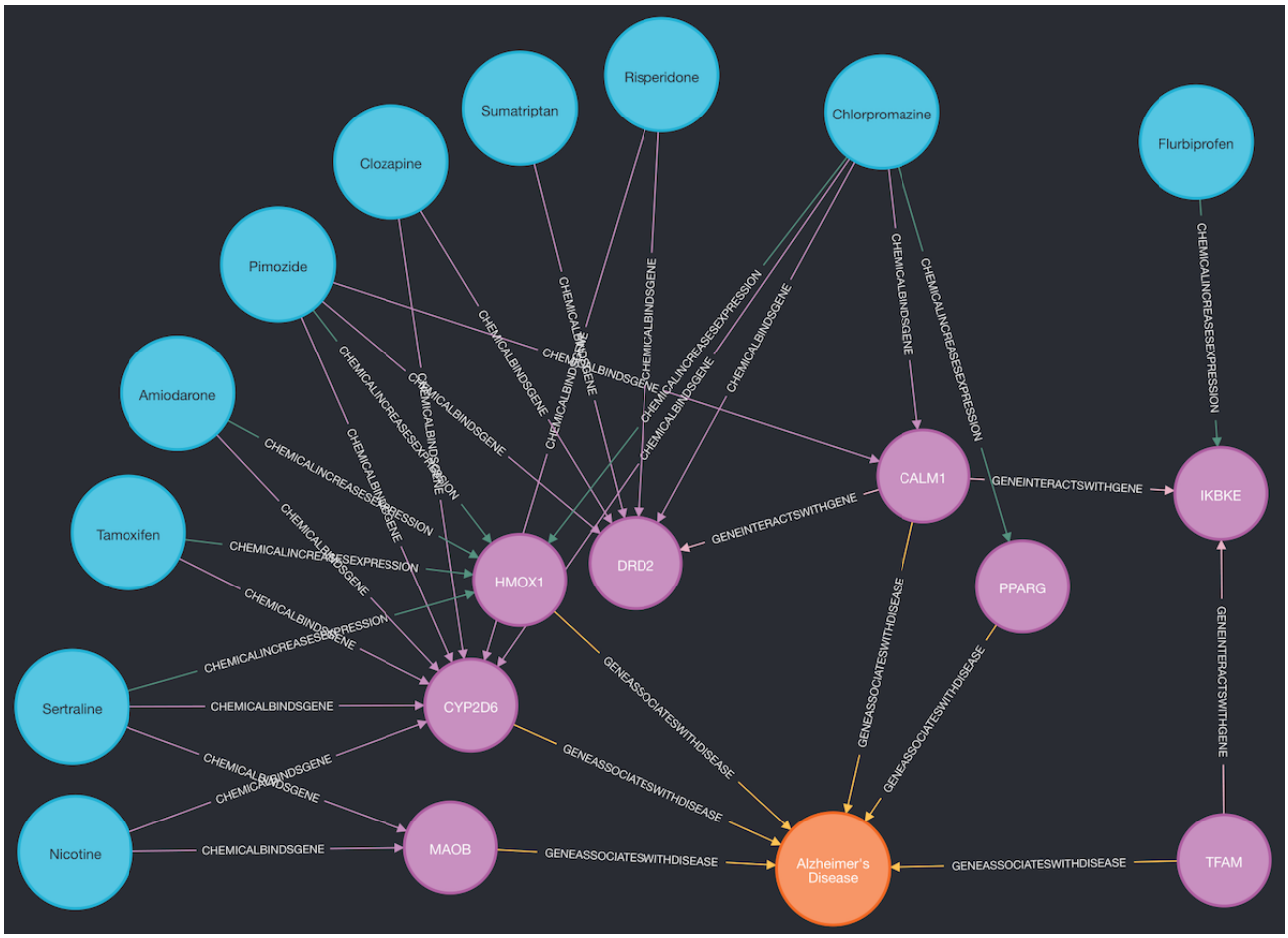
Table 5. Drug repurposing predictions made by the best-performing topological link prediction model (RotatE). Also shown are current approved indications and (if available) clinical trials investigating the efficacy of the drug for treating Alzheimer disease (AD).

Drug name	Approved indications	AD-related clinical trials
Sumatriptan	Migraines and cluster headaches	—[a]
Nicotine	Nicotine withdrawal symptoms	NCT00018278 (completed)
Pimozide	Tourette disorder	—
Risperidone	Schizophrenia, bipolar mania, and psychosis	NCT00034762 (completed)
Flurbiprofen	Osteoarthritis and rheumatoid arthritis	—
Sertraline	Depressive disorder and social anxiety disorder	NCT00086138 (completed); NCT00009191 (completed)
Clozapine	Schizophrenia	—
Tamoxifen	ER+[b] breast cancer	—
Amiodarone	Recurrent hemodynamically unstable ventricular tachycardia and recurrent ventricular fibrillation	—
Chlorpromazine	Nausea, vomiting, preoperative anxiety, schizophrenia, bipolar disorder, and severe behavioral problems in children	—

[a]No known AD-related clinical trials for the given drug.

[b]ER+: estrogen-receptor positive.

Figure 3. Spanning tree linking the 10 highest-scoring Alzheimer disease (AD) drug predictions (listed in Table 5) to AD. Blue nodes are drugs, pink nodes are genes, and the orange node is AD. Genes on the shortest path between a drug and AD can be considered putative mechanistic explanations for how the drug may act on AD etiology.



Discussion

Principal Findings

AlzKB is a freely available resource for the biomedical research community, with the primary goal of expanding the repertoire of therapies for AD via drug repurposing. In the previous sections, we described the current contents of AlzKB, the process of constructing it, and 2 specific data-driven use cases that illustrate how it can be applied to drug repurposing tasks. These use cases consisted of predicting the shared genetic architecture of AD and PD (potentially allowing for PD therapies to be repurposed for AD) and directly proposing drugs to repurpose for treating AD by predicting new links between drug and disease nodes in the knowledge graph. In both cases, the results are both biologically plausible and supported by quantitative metrics, yielding new hypotheses that merit experimental validation. AlzKB is a flexible resource that is not limited to these analyses, and we encourage other research teams to use it for different and complementary knowledge discovery tasks.

The Role of AlzKB in Biomedical Knowledge Discovery

AD and other neurodegenerative diseases present one of the greatest challenges in modern biomedicine. AD is by and large a disease of old age, and as improvements to health care continue to increase the overall global life expectancy, we can expect the number of people with various forms of dementia to also increase. As the etiology and pathophysiology of AD are highly multifactorial, there is likely no single “cure” for the disease. Instead, researchers and public health officials have shifted much of their focus toward finding therapies that reduce risk, slow the progression of the disease, or reverse neuronal damage. In addition, as there are various subtypes of AD with underlying mechanisms, any therapy might be effective for only some patients with AD. Therefore, an essential step for reducing global disease burden is to propose many new therapeutic agents that target various aspects of AD pathology. This is precisely the motivating use case for AlzKB. As we have demonstrated, AlzKB provides a rich representation of existing knowledge about AD and the biological context in which it acts. The 2 ML-based use cases we presented previously use real-world end points to demonstrate that the knowledge captured in AlzKB is meaningful and representative of the biological processes underlying the disease. AlzKB stands to become a major resource in the AD research community, where pattern analysis and integration with observational data can be used to propose a diverse array of new therapeutic hypotheses along with interpretable mechanistic explanations of how those therapies may act in the human body.

Building the initial release of AlzKB was a highly interdisciplinary effort involving contributions from experts in translational bioinformatics, data science, and clinical

informatics as well as medical scientists. Although each of these domains was essential in delivering a knowledge base that reflects important biomedical patterns describing AD etiology and treatment, a key need during the design and implementation phases was data literacy. To support future work in this and related areas, we encourage the inclusion of informatics and data analysis techniques in all types of biomedical curricula. Beyond AlzKB, our approach for building the knowledge graph is generalizable to practically any domain and depends on (1) defining an ontology using expert knowledge that formally describes the domain of interest and (2) identifying source databases that provide the entities and relationships described in the ontology. We are directly involved in the ongoing development of other knowledge bases using this same approach, including ComptoxAI—a knowledge base that supports AI research in toxicology [64]. As both knowledge bases share many of the same "core" entities (genes, diseases, pathways, and anatomical structures), the knowledge graphs are already semantically harmonized and ready for integration in larger, cross-disciplinary biomedical knowledge applications.

Discovering Putative Therapies Through Graph Data Science

Of the PD genes predicted to also be AD genes (see the *Proposing New Therapeutic Targets for AD* section; Table 3), some are involved in neuronal signaling and structure, and some are known to be involved in a wide range of neurological disorders. *FYN* has seen recent attention and investigation into its possible link to AD due to its broad expression in brain tissue and known interactions with tau proteins [65,66]. Among the other identified genes, one (*CHRNB1*) is known to be involved in acetylcholine signaling [67,68], and another (*KCNIP3*) codes a protein that interacts with presenilin, and mutations in presenilin are causal for hereditary AD [69,70]. Some of these gene hits (*ATXN2* and *DCTN1*) have limited or no current research directly linking them to AD but are biologically plausible. As such, they may represent novel therapeutic targets or targets for further research and investigation [71]. For example, *DCTN1* encodes the dynactin-1 protein, and deficits in dynactin are connected to several neurodegenerative diseases; however, there is limited research linking this gene to AD [72,73].

Among the drug repurposing predictions (see the *Drug Repurposing via Graph Data Science* section; Table 5) are some agents that have previously been proposed for the treatment of AD (risperidone and sertraline) or for symptoms associated with AD (nicotine). Sumatriptan has been the subject of several studies focused on AD [74] and is connected to a strong comorbidity of migraine headaches and dementia in women [75]. Pimozide has been shown to reduce the aggregation of tau protein in mice [76] and is linked to AD in a number of unrelated *in silico* models [77]. The inclusion of nicotine is also noteworthy as it has seen recent interest among AD researchers and is the subject of an ongoing clinical trial to improve memory [78]. Other drugs listed in Table 5 have not yet been identified as AD treatments and represent novel repurposing candidates. Each can be considered a testable hypothesis meriting further investigation, giving credence to the increased detective power of AlzKB's knowledge graph approach over existing AD data resources. It should be noted that this approach can only propose new indications for existing drugs and is based on existing knowledge and derived from known biological associations with those drugs. Other approaches (including emerging techniques in graph ML) could be used to propose entirely new drugs that could treat AD.

Future Directions With AlzKB

AlzKB is a growing resource, and we have plans for adding new features and data types that are in various stages of implementation. As a central hypothesis of AD pathogenesis revolves around the atypical accumulation of proteins within and around brain cells, an important step will be to adequately distinguish and differentiate genes from the proteins that those genes code for. Existing data resources available for inclusion in AlzKB largely fail to make this distinction in a way that is accepted by the scientific community, so we are currently evaluating options to use either postprocessing of existing knowledge sources or synthesis of new knowledge to achieve a good representation of genes, proteins, and functional or structural variants that are key to understanding AD.

Current ML models often do not generalize well to heterogeneous graphs such as the one that constitutes AlzKB's knowledge graph. This is largely because traditional models cannot use the network structure and heterogeneous nature of different entity types. Several promising algorithms can be used for prediction on heterogeneous graphs—including GraphSAGE [79] and metapath2vec [80]—but most fail to scale effectively when the number of node or edge types increases. As any effective therapy must be accompanied by a mechanistic understanding of how it functions, we also need to ensure that new heterogeneous graph ML models are *explainable*. With this in mind, we are using AlzKB as a motivating resource for designing new, cutting-edge algorithms that produce interpretable predictions from highly heterogeneous knowledge graphs. Furthermore, the increasing popularity of large language models (LLMs; such as GPT-4) presents a wealth of opportunities for incorporating knowledge graphs such as AlzKB into diverse AI applications [81]. One application we are considering is using AlzKB to provide LLMs with formalized knowledge about AD that allows them to more effectively produce informative outputs about AD etiology. Currently, LLMs can perform poorly on technically complex or poorly understood domains due to a scarcity of relevant content in their training corpora, and augmenting their performance using domain-specific knowledge graphs is an emerging strategy for fixing that issue. As we do so, these will be released alongside AlzKB with educational resources that facilitate ease of use and adoptability by various stakeholders.

Knowledge graphs—including AlzKB—come with several important limitations that will be crucial to address in coming years. One of these is the subjective nature of determining what does and does not constitute "knowledge," implying broad acceptance by the scientific community (as opposed to "data," which consist of individual observations). Currently, we use expert domain knowledge and careful screening of source databases to accomplish this, but with the advent of broadly accessible generative AI tools, there may be emerging strategies

that minimize sources of human bias [82]. Furthermore, new predictions made using knowledge graphs still necessitate costly and time-consuming experimental or observational follow-up studies to validate those predictions. This is due in part to the absence of negative samples for training predictive models. While the presence of an edge between 2 nodes in a knowledge graph is interpreted as a "positive sample" for model training, the absence of an edge simply means that we do not know whether a relationship does or does not exist, and therefore, it may not in fact be a negative sample. New methods, including self-supervised contrastive learning, show promise in alleviating this issue [83], but further work is needed to determine whether these generalize well to AlzKB and similar highly heterogeneous biomedical knowledge graphs. Nonetheless, these are active areas of research in the AI, informatics, and computer science communities, and in spite of them, our results are still robust enough to provide compelling evidence demonstrating AlzKB's scientific value.

Ultimately, we aim to provide AlzKB as a robust resource that helps unravel the etiology of AD. It is already a large, high-quality knowledge base from which graph-based AI or ML approaches can be developed for drug repurposing and drug discovery. As we and the rest of the biomedical research community make these discoveries in the coming years, they will be included and publicized on the AlzKB website as a public resource to drive innovation and scientific progress.

Obtaining AlzKB for Local Use and Extending the Knowledge Graph

As it is a public and open-source resource for scientific discovery, we provide AlzKB through a variety of interfaces with distinct advantages for different use cases and user types. Casual users who wish to browse the knowledge base or perform simple analyses can do so directly through the Neo4j browser interface [24]. However, for more advanced use cases (or when computational needs exceed those available on the public version of the knowledge base), AlzKB can be either downloaded and populated locally into a Neo4j installation or built from the original source data files via the tools included on the AlzKB GitHub repository [25]. The latter of these options also allows users to extend the knowledge base to include additional data sources, entity types, or relationships beyond those provided in the official knowledge base distribution. We also encourage users who make modifications to the knowledge base to submit their changes for review to be included in the main code distribution. Instructions for how to contribute to AlzKB are also available on the GitHub repository.

As the data sources included in AlzKB are all, themselves, from open-source databases, we urge users to ensure that any new data sources they merge into AlzKB similarly comply with open-source standards. In brief, AlzKB can only be maintained under the most restrictive license terms of its included third-party sources, so restrictive license terms in a database being considered decrease that database's suitability for inclusion. We hope for AlzKB to be recognized as a community effort for aggregating and democratizing the discovery of new AD therapeutics and, therefore, encourage public discussion of new methods and data sources to be included.

Conclusions

In this work, we introduced AlzKB as a free, publicly available toolkit and data resource for novel discoveries in AD research, with a particular focus on therapeutic approaches to treating AD. AlzKB is both new and continually growing, and we aim to cultivate a community of researchers to collaboratively increase the impact, speed, and throughput of AD research, along with rapid dissemination to health care, academia, and the pharmaceutical industry. In the future, we will develop new AI and data science methods to continually extract knowledge from AlzKB, but in this study, we already demonstrate through graph data science that AlzKB can both replicate existing AD knowledge and generate entirely new, testable hypotheses to drive the future of drug repurposing and drug discovery.

Acknowledgments

The Alzheimer's Knowledge Base is supported by US National Institutes of Health grants U01-AG066833, R01-LM010098, R01-LM013463 (principal investigator [PI]: JHM), and R00-LM013646 (PI: JDR).

Data Availability

The data sets generated during and analyzed during this study are available in the GitHub and Zenodo repositories [25,26].

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplemental information providing expanded details on the knowledge graph completion methods used to validate Alzheimer's Knowledge Base, as well as counts for relationship types in the knowledge graph.
[DOCX File , 23 KB-Multimedia Appendix 1]

References

1. 2022 Alzheimer's disease facts and figures. Alzheimers Dement. Apr 2022;18(4):700-789. [doi: 10.1002/alz.12638] [Medline: 35289055]

2. Yiannopoulou KG, Papageorgiou SG. Current and future treatments in Alzheimer disease: an update. J Cent Nerv Syst Dis. Feb 29, 2020;12:1179573520907397. [FREE Full text] [doi: 10.1177/1179573520907397] [Medline: 32165850]
3. Rabinovici GD. Controversy and progress in Alzheimer's disease - FDA approval of aducanumab. N Engl J Med. Aug 26, 2021;385(9):771-774. [doi: 10.1056/NEJMp2111320] [Medline: 34320284]
4. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. Mol Neurodegener. Aug 02, 2019;14(1):32. [FREE Full text] [doi: 10.1186/s13024-019-0333-5] [Medline: 31375134]
5. Aisen PS, Cummings J, Jack CRJ, Morris JC, Sperling R, Frölich L, et al. On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimers Res Ther. Aug 09, 2017;9(1):60. [FREE Full text] [doi: 10.1186/s13195-017-0283-5] [Medline: 28793924]
6. Fan L, Mao C, Hu X, Zhang S, Yang Z, Hu Z, et al. New insights into the pathogenesis of Alzheimer's disease. Front Neurol. 2019;10:1312. [FREE Full text] [doi: 10.3389/fneur.2019.01312] [Medline: 31998208]
7. Silva MV, de Mello Gomide Loures C, Alves LC, de Souza LC, Borges KB, Carvalho MD. Alzheimer's disease: risk factors and potentially protective measures. J Biomed Sci. May 09, 2019;26(1):33. [FREE Full text] [doi: 10.1186/s12929-019-0524-y] [Medline: 31072403]
8. Rodriguez S, Hug C, Todorov P, Moret N, Boswell SA, Evans K, et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nat Commun. Feb 15, 2021;12(1):1033. [FREE Full text] [doi: 10.1038/s41467-021-21330-0] [Medline: 33589615]
9. Tsuji S, Hase T, Yachie-Kinoshita A, Nishino T, Ghosh S, Kikuchi M, et al. Artificial intelligence-based computational framework for drug-target prioritization and inference of novel repositionable drugs for Alzheimer's disease. Alzheimers Res Ther. May 03, 2021;13(1):92. [FREE Full text] [doi: 10.1186/s13195-021-00826-3] [Medline: 33941241]
10. Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet. Apr 15, 2007;16(8):865-873. [doi: 10.1093/hmg/ddm031] [Medline: 17317784]
11. The Alzheimer's KnowledgeBase (AlzKB). AlzKB. URL: https://alzkb.ai/ [accessed 2023-02-24]
12. Daluwatumulle G, Wijesinghe R, Weerasinghe R. In silico drug repurposing using knowledge graph embeddings for Alzheimer's disease. In: Proceedings of the 9th International Conference on Bioinformatics Research and Applications. 2022. Presented at: ICBRA '22; September 18-20, 2022; Berlin, Germany. [doi: 10.1145/3569192.3569203]
13. Nian Y, Hu X, Zhang R, Feng J, Du J, Li F, et al. Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing. BMC Bioinformatics. Sep 30, 2022;23(Suppl 6):407. [FREE Full text] [doi: 10.1186/s12859-022-04934-1] [Medline: 36180861]
14. Hsieh KL, Plascencia-Villa G, Lin KH, Perry G, Jiang X, Kim Y. Synthesize heterogeneous biological knowledge via representation learning for Alzheimer's disease drug repurposing. iScience. Nov 26, 2022;26(1):105678. [FREE Full text] [doi: 10.1016/j.isci.2022.105678] [Medline: 36594024]
15. Binder J, Ursu O, Bologa C, Jiang S, Maphis N, Dadras S, et al. Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Commun Biol. Feb 11, 2022;5(1):125. [FREE Full text] [doi: 10.1038/s42003-022-03068-7] [Medline: 35149761]
16. Sügis E, Dauvillier J, Leontjeva A, Adler P, Hindie V, Moncion T, et al. HENA, heterogeneous network-based data set for Alzheimer's disease. Sci Data. Aug 14, 2019;6(1):151. [doi: 10.1038/s41597-019-0152-0] [Medline: 31413325]
17. Robinson I, Webber J, Eifrem E. Graph Databases: New Opportunities for Connected Data. Sebastopol, CA. O'Reilly Media; 2015.
18. Davis R, Shrobe H, Szolovits P. What is a knowledge representation? AI Mag. 1993;14(1):17. [doi: 10.1609/aimag.v14i1.1029]
19. Musen MA, Protégé Team. The Protégé project: a look back and a look forward. AI Matters. Jun 2015;1(4):4-12. [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
20. Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S. OWL 2 Web ontology language primer. World Wide Web Consortium. Apr 21, 2009. URL: https://www.w3.org/TR/2009/WD-owl2-primer-20090421/ [accessed 2024-03-25]
21. Tsarkov D, Horrocks I. FaCT++ description logic reasoner: system description. In: Proceedings of the International Joint Conference on Automated Reasoning. 2006. Presented at: IJCAR 2006; August 17-20, 2006; Seattle, WA. [doi: 10.1007/11814771_26]
22. Neo4j. URL: https://neo4j.com/ [accessed 2022-10-25]
23. Barrasa J, Cowley A. neosemantics (n10s): Neo4j RDF and semantics toolkit. Neo4j. URL: https://neo4j.com/labs/neosemantics/ [accessed 2022-10-25]
24. Neo4j browser. Neo4j. URL: http://neo4j.alzkb.ai/browser/ [accessed 2023-02-24]
25. EpistasisLab/AlzKB. GitHub. URL: https://github.com/EpistasisLab/AlzKB [accessed 2023-02-24]
26. Romano J, Wang P. EpistasisLab/AlzKB: AlzKB first DOI release. Zenodo. Aug 22, 2022. URL: https://zenodo.org/records/7015728 [accessed 2024-03-27]
27. Mortensen HM, Senn J, Levey T, Langley P, Williams AJ. The 2021 update of the EPA's adverse outcome pathway database. Sci Data. Jul 12, 2021;8(1):169. [FREE Full text] [doi: 10.1038/s41597-021-00962-3] [Medline: 34253739]

28. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. Bgee: integrating and comparing heterogeneous transcriptome data among species. In: Proceedings of the Data Integration in the Life Sciences. 2008. Presented at: DILS 2008; June 25-27, 2008; Evry, France. [doi: 10.1007/978-3-540-69828-9_12]
29. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res. Jan 08, 2019;47(D1):D955-D962. [FREE Full text] [doi: 10.1093/nar/gky1032] [Medline: 30407550]
30. Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015;2015:bav028. [FREE Full text] [doi: 10.1093/database/bav028] [Medline: 25877637]
31. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. Jan 2008;36(Database issue):D901-D906. [FREE Full text] [doi: 10.1093/nar/gkm958] [Medline: 18048412]
32. Grulke CM, Williams AJ, Thillanadarajah I, Richard AM. EPA's DSSTox database: history of development of a curated chemistry resource supporting computational toxicology research. Comput Toxicol. Nov 01, 2019;12:100096. [FREE Full text] [doi: 10.1016/j.comtox.2019.100096] [Medline: 33426407]
33. Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, et al. ACToR--Aggregated computational toxicology resource. Toxicol Appl Pharmacol. Nov 15, 2008;233(1):7-13. [doi: 10.1016/j.taap.2007.12.037] [Medline: 18671997]
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. May 2000;25(1):25-29. [FREE Full text] [doi: 10.1038/75556] [Medline: 10802651]
35. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Jan 08, 2021;49(D1):D325-D334. [FREE Full text] [doi: 10.1093/nar/gkaa1113] [Medline: 33290552]
36. Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. Jan 08, 2019;47(D1):D1005-D1012. [FREE Full text] [doi: 10.1093/nar/gky1120] [Medline: 30445434]
37. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. Sep 22, 2017;6:e26726. [FREE Full text] [doi: 10.7554/eLife.26726] [Medline: 28936969]
38. The human reference protein interactome mapping project. The Human Reference Interactome. URL: http://www.interactome-atlas.org/ [accessed 2023-02-24]
39. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl. Aug 04, 2016;2(1):16015. [FREE Full text] [doi: 10.1038/npjsba.2016.15] [Medline: 28413689]
40. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc. Jul 2000;88(3):265-266. [FREE Full text] [Medline: 10928714]
41. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. Jan 2011;39(Database issue):D52-D57. [FREE Full text] [doi: 10.1093/nar/gkq1237] [Medline: 21115458]
42. Schaefer C, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nat Prec. Aug 29, 2008. [doi: 10.1038/npre.2008.2243.1]
43. Himmelstein D, Pouya K, Hessler CS, Green AJ, Baranzini S. PharmacotherapyDB 1.0: the open catalog of drug therapies for disease. Figshare. 2016. URL: https://tinyurl.com/mv8k46em [accessed 2024-03-25]
44. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. Jan 08, 2021;49(D1):D1388-D1395. [FREE Full text] [doi: 10.1093/nar/gkaa971] [Medline: 33151290]
45. Wu G, Haw R. Functional interaction network construction and analysis for disease discovery. Methods Mol Biol. 2017;1558:235-253. [doi: 10.1007/978-1-4939-6783-4_11] [Medline: 28150241]
46. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. Jan 04, 2016;44(D1):D1075-D1079. [FREE Full text] [doi: 10.1093/nar/gkv1075] [Medline: 26481350]
47. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ. TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database (Oxford). Jan 01, 2018;2018:2. [FREE Full text] [doi: 10.1093/database/bay028] [Medline: 30403794]
48. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. Jan 31, 2012;13(1):R5. [FREE Full text] [doi: 10.1186/gb-2012-13-1-r5] [Medline: 22293552]
49. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers KA, et al. WikiPathways: connecting communities. Nucleic Acids Res. Jan 08, 2021;49(D1):D613-D621. [FREE Full text] [doi: 10.1093/nar/gkaa1024] [Medline: 33211851]
50. Ali M, Berrendorf M, Hoyt CT, Vermue L, Sharifzadeh S, Tresp V, et al. PyKEEN 1.0: a python library for training and evaluating knowledge graph embeddings. J Mach Learn Res. 2021;22(82):1-6. [FREE Full text]
51. Zamini M, Reza H, Rabiei M. A review of knowledge graph completion. Information. Aug 21, 2022;13(8):396. [doi: 10.3390/info13080396]

52. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. 2013. Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, Nevada.
53. Sun Z, Deng ZH, Nie JY, Tang J. RotatE: knowledge graph embedding by relational rotation in complex space. arXiv. Preprint posted online February 26, 2019. [FREE Full text]
54. Yang B, Yih WT, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. arXiv. Preprint posted online December 20, 2014. [FREE Full text]
55. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. arXiv. Preprint posted online June 20, 2016. [FREE Full text]
56. Dettmers T, Minervini P, Stenetorp P, Riedel S. Convolutional 2D knowledge graph embeddings. arXiv. Preprint posted online July 5, 2017. [FREE Full text] [doi: 10.1609/aaai.v32i1.11573]
57. Newman ME. Clustering and preferential attachment in growing networks. Phys Rev E. Jul 26, 2001;64(2):025102. [doi: 10.1103/physreve.64.025102]
58. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. Oct 15, 1999;286(5439):509-512. [doi: 10.1126/science.286.5439.509] [Medline: 10521342]
59. Adamic LA, Adar E. Friends and neighbors on the web. Soc Netw. Jul 2003;25(3):211-230. [doi: 10.1016/s0378-8733(03)00009-1]
60. Zhou T, Lü L, Zhang YC. Predicting missing links via local information. Eur Phys J B. Oct 10, 2009;71(4):623-630. [doi: 10.1140/epjb/e2009-00335-8]
61. Gao Z, Ding P, Xu R. KG-Predict: a knowledge graph computational framework for drug repurposing. J Biomed Inform. Aug 2022;132:104133. [FREE Full text] [doi: 10.1016/j.jbi.2022.104133] [Medline: 35840060]
62. Nussbaum RL, Ellis CE. Alzheimer's disease and Parkinson's disease. N Engl J Med. Apr 03, 2003;348(14):1356-1364. [doi: 10.1056/NEJM2003ra020003] [Medline: 12672864]
63. Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, et al. Application of network link prediction in drug discovery. BMC Bioinformatics. Apr 12, 2021;22(1):187. [FREE Full text] [doi: 10.1186/s12859-021-04082-y] [Medline: 33845763]
64. Romano JD, Hao Y, Moore JH, Penning TM. Automating predictive toxicology using ComptoxAI. Chem Res Toxicol. Aug 15, 2022;35(8):1370-1382. [FREE Full text] [doi: 10.1021/acs.chemrestox.2c00074] [Medline: 35819939]
65. Iannuzzi F, Sirabella R, Canu N, Maier TJ, Annunziato L, Matrone C. Fyn tyrosine kinase elicits amyloid precursor protein Tyr682 phosphorylation in neurons from Alzheimer's disease patients. Cells. Jul 30, 2020;9(8):1807. [FREE Full text] [doi: 10.3390/cells9081807] [Medline: 32751526]
66. Nygaard HB, van Dyck CH, Strittmatter SM. Fyn kinase inhibition as a novel therapy for Alzheimer's disease. Alzheimers Res Ther. Feb 5, 2014;6(1):8. [FREE Full text] [doi: 10.1186/alzrt238] [Medline: 24495408]
67. Lardenoije R, Roubroeks JA, Pishva E, Leber M, Wagner H, Iatrou A, et al. Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood. Clin Epigenetics. Nov 27, 2019;11(1):164. [FREE Full text] [doi: 10.1186/s13148-019-0755-5] [Medline: 31775875]
68. Lombardo S, Maskos U. Role of the nicotinic acetylcholine receptor in Alzheimer's disease pathology and treatment. Neuropharmacology. Sep 2015;96(Pt B):255-262. [FREE Full text] [doi: 10.1016/j.neuropharm.2014.11.018] [Medline: 25514383]
69. Jo DG, Lee JY, Hong YM, Song S, Mook-Jung I, Koh JY, et al. Induction of pro-apoptotic calsenilin/DREAM/KChIP3 in Alzheimer's disease and cultured neurons after amyloid-beta exposure. J Neurochem. Feb 2004;88(3):604-611. [FREE Full text] [doi: 10.1111/j.1471-4159.2004.02159.x] [Medline: 14720210]
70. Jin JK, Choi JK, Wasco W, Buxbaum JD, Kozlowski PB, Carp RI, et al. Expression of calsenilin in neurons and astrocytes in the Alzheimer's disease brain. Neuroreport. Apr 04, 2005;16(5):451-455. [doi: 10.1097/00001756-200504040-00007] [Medline: 15770150]
71. Rosas I, Martínez C, Clarimón J, Lleó A, Illán-Gala I, Dols-Icardo O, et al. Role for ATXN1, ATXN2, and HTT intermediate repeats in frontotemporal dementia and Alzheimer's disease. Neurobiol Aging. Mar 2020;87:139.e1-139.e7. [doi: 10.1016/j.neurobiolaging.2019.10.017] [Medline: 31810584]
72. Aboud O, Parcon PA, DeWall KM, Liu L, Mrak RE, Griffin WS. Aging, Alzheimer's, and APOE genotype influence the expression and neuronal distribution patterns of microtubule motor protein dynactin-P50. Front Cell Neurosci. Mar 25, 2015;9:103. [FREE Full text] [doi: 10.3389/fncel.2015.00103] [Medline: 25859183]
73. Caroppo P, Le Ber I, Clot F, Rivaud-Péchoux S, Camuzat A, De Septenville A, et al. DCTN1 mutation analysis in families with progressive supranuclear palsy-like phenotypes. JAMA Neurol. Feb 2014;71(2):208-215. [FREE Full text] [doi: 10.1001/jamaneurol.2013.5100] [Medline: 24343258]
74. Zochodne DW, Ho LT. Sumatriptan blocks neurogenic inflammation in the peripheral nerve trunk. Neurology. Jan 1994;44(1):161-163. [doi: 10.1212/wnl.44.1.161] [Medline: 8290056]
75. Liu CT, Wu BY, Hung YC, Wang LY, Lee YY, Lin TK, et al. Decreased risk of dementia in migraine patients with traditional Chinese medicine use: a population-based cohort study. Oncotarget. Oct 03, 2017;8(45):79680-79692. [FREE Full text] [doi: 10.18632/oncotarget.19094] [Medline: 29108348]

76. Kim YD, Jeong EI, Nah J, Yoo SM, Lee WJ, Kim Y, et al. Pimozide reduces toxic forms of tau in TauC3 mice via 5' adenosine monophosphate-activated protein kinase-mediated autophagy. J Neurochem. Sep 11, 2017;142(5):734-746. [FREE Full text] [doi: 10.1111/jnc.14109] [Medline: 28632947]
77. Kumar S, Chowdhury S, Kumar S. In silico repurposing of antipsychotic drugs for Alzheimer's disease. BMC Neurosci. Oct 27, 2017;18(1):76. [FREE Full text] [doi: 10.1186/s12868-017-0394-8] [Medline: 29078760]
78. van Duijn CM, Hofman A. Relation between nicotine intake and Alzheimer's disease. BMJ. Jun 22, 1991;302(6791):1491-1494. [FREE Full text] [doi: 10.1136/bmj.302.6791.1491] [Medline: 1855016]
79. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. arXiv. Preprint posted online June 7, 2017. [FREE Full text]
80. Dong Y, Chawla NV, Swami A. metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017. Presented at: KDD '17; August 13-17, 2017; Halifax, NS. URL: https://dl.acm.org/doi/10.1145/3097983.3098036
81. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. arXiv. Preprint posted online June 14, 2023. [FREE Full text] [doi: 10.48550/arXiv.2306.08302]
82. Zhu Y, Wang X, Chen J, Qiao S, Ou Y, Yao Y, et al. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. arXiv. Preprint posted online May 22, 2023. [FREE Full text]
83. Kefato ZT, Girdzijauskas S. Self-supervised Graph Neural Networks without explicit negative sampling. arXiv. Preprint posted online March 27, 2021. [FREE Full text]

Abbreviations

AD: Alzheimer disease
AI: artificial intelligence
AlzKB: Alzheimer's Knowledge Base
LLM: large language model
ML: machine learning
MRR: mean reciprocal rank
OWL: Web Ontology Language
PD: Parkinson disease
RDF: Resource Description Framework
RF: random forest

Edited by T de Azevedo Cardoso; submitted 24.02.23; peer-reviewed by P Dabas, N Mungoli, B Xie, C Sun; comments to author 21.04.23; revised version received 23.06.23; accepted 07.11.23; published 18.04.24

Please cite as:
Romano JD, Truong V, Kumar R, Venkatesan M, Graham BE, Hao Y, Matsumoto N, Li X, Wang Z, Ritchie MD, Shen L, Moore JH
The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research
J Med Internet Res 2024;26:e46777
URL: https://www.jmir.org/2024/1/e46777
doi: 10.2196/46777
PMID: