The Alzheimer's Knowledge Base – A knowledge graph for therapeutic discovery in Alzheimer's Disease research

Joseph D Romano, Van Truong, Rachit Kumar, Mythreye Venkatesan, Britney E. Graham, Yun Hao, Nick Matsumoto, Xi Li, Zhiping Wang, Marylyn Ritchie, Li Shen, Jason H. Moore

Submitted to: Journal of Medical Internet Research
on: February 24, 2023

Table of Contents

The Alzheimer's Knowledge Base – A knowledge graph for therapeutic discovery in Alzheimer's Disease research

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

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

Corresponding Author:
Joseph D Romano BS, MA, MPhil, PhD
Institute for Biomedical Informatics
Perelman School of Medicine
University of Pennsylvania
403 Blockley Hall
423 Guardian Drive
Philadelphia
US

Abstract

Background: As global populations age and become susceptible to neurodegenerative illnesses, new therapies for Alzheimer's Disease (AD) are urgently needed. Existing data resources for drug discovery and repurposing fail to capture heterogeneous biomedical knowledge that is central to the disease's etiology and response to drugs. We designed the Alzheimer's Knowledge Base (AlzKB) to alleviate this need by providing a comprehensive knowledge representation of AD etiology and candidate therapeutics.

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 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 (chemicals, genes, anatomy, diseases, etc.). AlzKB uses an OWL 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 at http://alzkb.ai, and currently contains 118,902 entities with 1,309,527 relationships between those entities. To demonstrate its value, we use graph data science and machine learning to (a.) propose new therapeutic targets based on similarities of AD to Parkinson Disease and (b.) 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 two use-cases, we show that it is a valuable tool for proposing novel therapeutic hypotheses based on public biomedical knowledge.

(JMIR Preprints 24/02/2023:46777)
DOI: https://doi.org/10.2196/preprints.46777

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The Alzheimer's Knowledge Base – A knowledge graph for therapeutic discovery in Alzheimer's Disease research

Authors: Joseph D. Romano[1,2,3,*], Van Truong[1,4,5], Rachit Kumar[1,4,5,6], Mythreye Venkatesan[7], Britney E. Graham[7], Yun Hao[1,4], Nick Matsumoto[7], Xi Li[7], Zhiping Wang[7], Marylyn D. Ritchie[1,3,5], Li Shen[1,3], Jason H. Moore[7,†]

Author Affiliations:
[1] Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA
[2] Center of Excellence in Environmental Toxicology, University of Pennsylvania, Philadelphia, PA
[3] Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania, Philadelphia, PA
[4] Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA
[5] Department of Genetics, University of Pennsylvania, Philadelphia, PA
[6] Medical Scientist Training Program, University of Pennsylvania, Philadelphia, PA
[7] Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA.
* Corresponding Author: Joseph.Romano@pennmedicine.upenn.edu
† Corresponding Author: Jason.Moore@csmc.edu

Abstract

Background: As global populations age and become susceptible to neurodegenerative illnesses, new therapies for Alzheimer's 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. 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 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 (chemicals, genes, anatomy, diseases, etc.). AlzKB uses an OWL 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 at http://alzkb.ai, and currently contains 118,902 entities with 1,309,527 relationships between those entities. To demonstrate its value, we use graph data science and machine learning to (a.) propose new therapeutic targets based on similarities of AD to Parkinson Disease and (b.) repurpose existing drugs that may treat AD. For each use case, AlzKB recovers known therapeutic associations while proposing biologically plausible new ones.

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

Introduction

Alzheimer's Disease (AD) is a progressive, neurodegenerative disease affecting an estimated 6.5 million Americans aged 65 and older, and represents a significant clinical, economic, and emotional burden worldwide[1]. AD is often cited as one of the greatest healthcare problems of the 21st century, particularly in developed 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 (FDA) has approved 5 drugs for the treatment of AD, 4 of which (donepezil, rivastigmine, galantamine, memantine) only temporarily treat symptoms but do not alter overall progression of the disease[2], while the fifth (aducanumab) is highly controversial in terms of evidence for 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 amyloid-beta protein, extracellular deposition of amyloid-beta, 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 ε4 allele of the *APOE* gene, which is involved in fat metabolism and cholesterol transport. However, the exact mechanism by which these factors—including *APOE*-ε4 presence—cause or contribute to AD risk is unknown[7].

Of the many techniques employed in AD therapeutics research, there is a wealth of computer-aided approaches that leverage recent advances in bioinformatics, epidemiology, artificial intelligence, and machine learning. For example, Rodriguez *et al*. developed a machine learning 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 Alzheimer's disease[8]. Tsuji et al. used an autoencoder neural network to perform dimensionality reduction of a high-density protein interaction network in order to identify new possible drug targets and then found drugs associated with those targets[9]. Genome-wide association studies (GWAS) 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 study, we describe the design and deployment of a major new knowledge resource for computational AD research—named The Alzheimer's Knowledge Base (AlzKB)—with a particular focus on drug discovery and drug repurposing. 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 related to one another. To demonstrate its value, we present two data driven analyses involving machine learning on AlzKB's knowledge graph: 1.) Predicting Parkinson's Disease 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 online[11], and consists entirely of publicly-sourced knowledge integrated from 22 diverse online biomedical databases. We hypothesize 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 explainability of new predictions.

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], gene identification[15], and as general informational resources[16]. Like 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 AD-KG[14] and HENA[16], have been released as publicly accessible resources like AlzKB. Other studies use existing resources not specifically intended for AD research (such as SemMedDB, in [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 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/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 machine learning algorithms. Furthermore, AlzKB offers a public, online web interface that allows 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 like AD, AlzKB represents a major step forward by improving both quantitatively and structurally on existing resources.

Methods

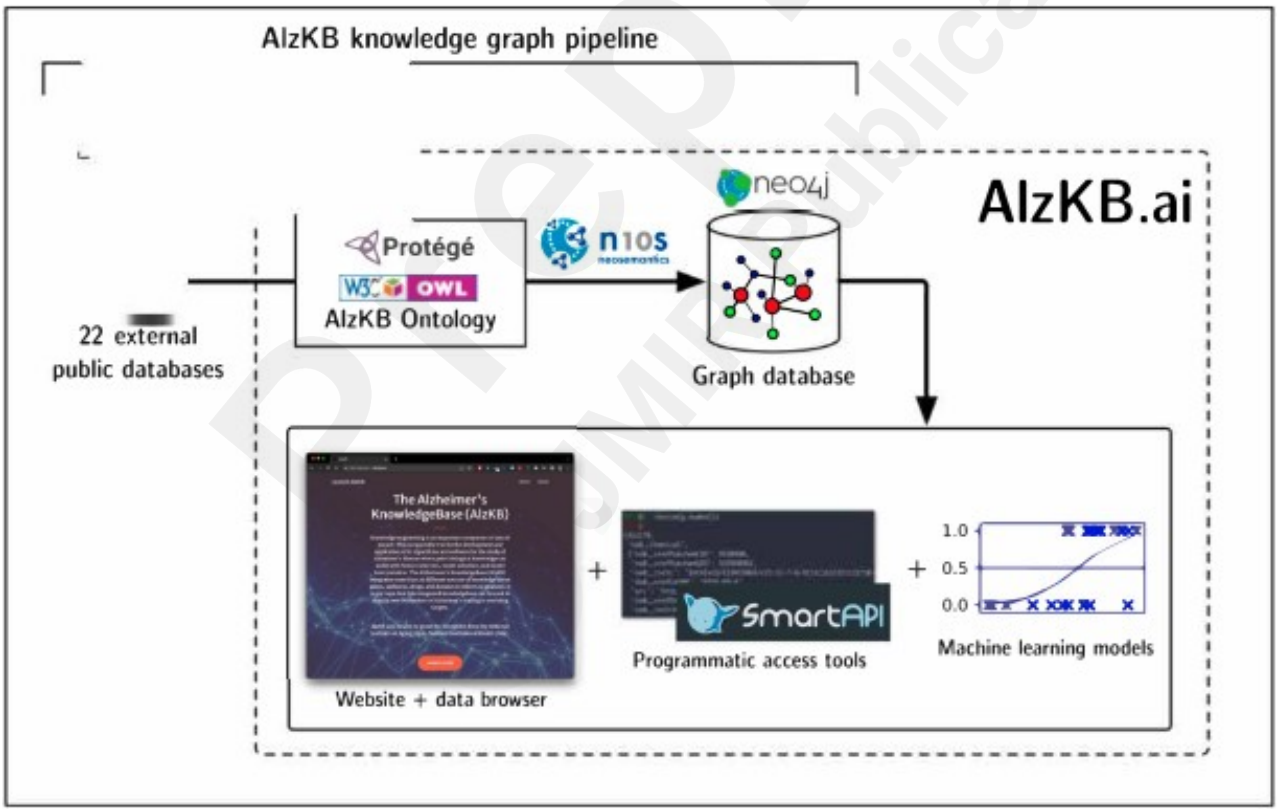

Figure 1. Schematic overview of the Alzheimer's Knowledge Base (AlzKB).

AlzKB Ontology

Graph databases are renowned for their flexibility in representing data that does 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 an 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 over knowledge within a domain of interest[18]. Briefly, since many of the components of a graph database have a one-to-one correspondence with components of an OWL 2 ontology (e.g., OWL 2 classes are equivalent to graph database node labels, OWL 2 object properties are equivalent to edge types in a graph database, etc.), 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 (v.5.5.0) [19], following an iterative process guided by expert domain knowledge. First, we prototyped a class hierarchy containing the types of nodes (e.g., Gene, Disease, Pathway, Drug) desired in the knowledge base. We then annotated these classes with data properties (e.g., drugs can be assigned a property value corresponding to molecular weight) and object properties (relationship types that link two entities, such as "drug treats disease"). A thorough description of the components of OWL 2 ontologies is given 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 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 the semantic contents of the ontology are consistent with current AD knowledge and systems biology processes involved in AD etiology. After collecting the data sources used to populate the ontology (see below), we added 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 §) 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/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 (e.g., "DRUG_BINDS_GENE" or "GENE_ASSOCIATED_WITH_DISEASE") to include in the knowledge base. For each of these entity types 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/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 where 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 use fields from DSSTox and PubChem to augment the properties annotated onto drugs (such as molecular weight, chemical fingerprint, synonyms, etc.).

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* (i.e., 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 data sources. For example, *Memantine* is annotated with the CAS 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 RDF/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 (v4.4.5)[22] with the contents of the RDF/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 usage details) and the Neo4j graph database 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 AlzKB, what data elements are used from them, and total size of the data source (counts of entities of relevant data types only). Since source data elements do not correspond in a 1-to-1 manner with entities in the graph (e.g., entities may be merged, filtered, used as edges rather than nodes, etc.), actual counts for entities in AlzKB stratified by source are not available. Sizes are best available estimates at the time of publication. For actual node and edge type counts in AlzKB, see Table 2 and Supplemental Table S1. * indicates that derived data is structured in part using Hetionet.

Data Source	Use in AlzKB	Size (# entities)
AOP-DB (Adverse Outcome Pathway Database) [27]	Adverse outcome pathways and chemical-gene associations.	1,207,456
Bgee* [28]	Tissue-specific gene expression data. Only human gene expression data are used in AlzKB.	9,093,494
Disease Ontology* [29]	Human diseases – only AD, subtypes of AD, and related neurodegenerative diseases are included in AlzKB.	8,043
DisGeNET [30]	Diseases, genes, and disease-gene associations with scores representing levels of evidence. Only AD and related neurodegenerative terms used for diseases.	51,841
DrugBank [31]	On-market and experimental pharmaceutical drugs.	15,550

EPA Distributed Structure-Searchable Toxicity (DSSTox) [32]	Chemical toxicity data – filtered in AlzKB for drugs contained in DrugBank.	1,200,059
EPA Aggregated Computational Toxicology Resource (AcTOR) [33]	Chemical toxicity data – filtered in AlzKB for drugs contained in DrugBank.	504,871
Gene Ontology* [34,35]	Biological processes, molecular functions, and cellular components.	42,950
GWAS Catalog* [36]	Gene-Disease associations	60071
Hetionet [37]	Graph modeling and entity resolution (for data sources marked with "*")	47,031
Human Reference Protein Interactome Mapping Project* [38]	Human protein-protein interactions (modeled as gene-gene interactions)	9,094
LINCS L1000* [39]	Human differential gene expression data.	7,467 genes
MeSH (NCBI Medical Subject Headings) [40]	Clinical and biomedical concepts (annotated to various node types).	~27,000
NCBI Gene (Entrez Gene) [41]	Human genes and gene synonyms.	62,407
Pathway Interaction Database* [42]	Pathways and gene-pathway membership.	223
PharmacotherapyDB* [43]	Drug indications for human diseases.	698
PubChem [44]	Chemical structures and identifiers – only chemicals in DrugBank are included in AlzKB.	115,067,800
Reactome Pathway Database* [45]	Pathways and gene-pathway membership.	1,341
SIDER* [46]	Drug side effects (modeled as Diseases).	5,868
TISSUES* [47]	Tissue-specific gene expression data.	
Uberon* [48]	Human anatomical structures.	402
WikiPathways* [49]	Pathways and gene-pathway membership.	298

Validating AlzKB using Real World Use Cases

After building AlzKB's knowledge graph, we designed two machine learning (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 two 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 endpoints (e.g., effective drugs or etiologically important genes), the informative patterns and phenomena underlying those endpoints 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 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 (CN), Total Neighbors (TN), Preferential Attachment (PA), Adamic-Adar (AA), and Resource Allocation (RA) [50–53]. 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:

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

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

$$PA(n_1, n_2) = |N(n_1)| * |N(n_2)|$$

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

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

where $N(n_i)$ is the set of neighbor (adjacent) nodes of node *i*. Our training procedure for the random forest model included 3-fold grid search cross validation to optimize hyperparameters, 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, target node) and scores are generated to indicate plausibility of the triple. The 5 models we evaluated are TransE, RotatE, DistMult, ComplEx, and ConvE[53].

We implemented the 5 models using PyKEEN – a Python library for knowledge graph embeddings[54]. We randomly split the dataset 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 GPU and evaluated the results using the rank-based evaluation metrics Hits@k (k=1, 3, and 10) and Mean Reciprocal Rank (MRR) [55]. Ranking-based evaluation sorts the scores of triples in descending orders 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. 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, i.e., how well they can predict missing entities in partial triples without either the head (source) or tail (target) entities.

Ethics Statement

No human subjects were involved in this research. All data used to build and evaluate AlzKB are derived from publicly available biomedical knowledge retrieved from open-access databases. None of the data included are derived from individual human subjects. 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 (v1.0)[26] contains 118,902 distinct nodes (representing biomedical entities) and 1,309,527 relationships linking those nodes. A full summary of node types and relationship types with counts, respectively, are given in **Table 2** and **Supplemental Table S1**. 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.

Table 2: Node types and counts in AlzKB, 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
Gene	62,407
Drug	35,063
BiologicalProcess	11,381
Pathway	4,570
MolecularFunction	2,884
CellularComponent	1,391
Symptom	438
BodyPart	402
DrugClass	345
Disease	20

Proposing new therapeutic targets for AD

As a proof of concept, we performed an analysis to predict whether known Parkinson Disease (PD) genes are also linked to AD etiology. PD is chronic, progressive neurological disorder characterized by uncontrollable movements and possible mental and behavioral changes. Like 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[56], 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[57]. To assemble the dataset, we considered all gene nodes adjacent to AD positive (n=101) and all gene nodes not adjacent to AD 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 (i.e., 3 times the number of positive samples). To evaluate the performance, we used accuracy, balanced accuracy, precision, recall, F1-score, AUROC, and AUCPR, as shown in **Figure 2**.

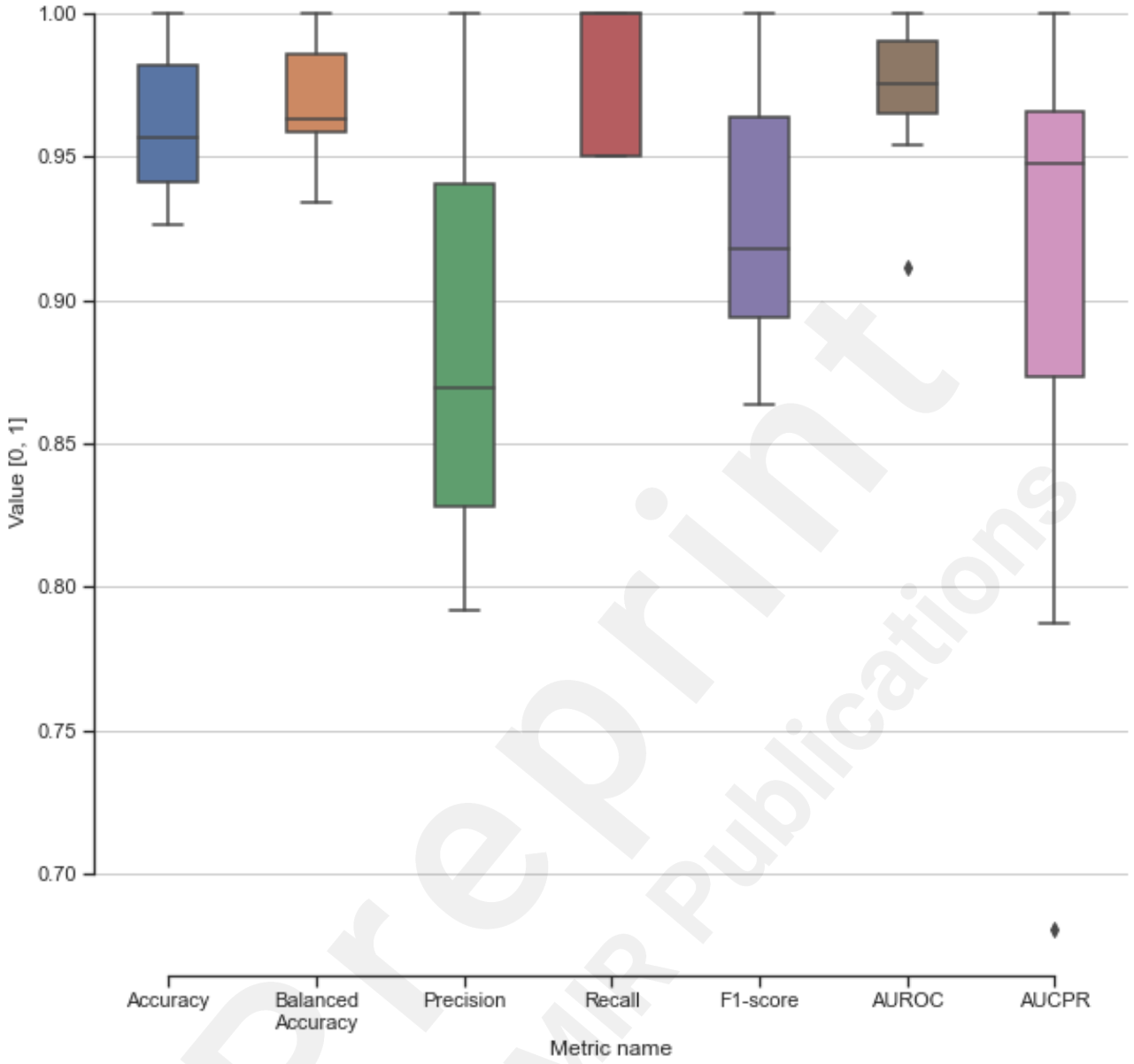

Figure 2. Random forest classifier performance (over 10 independent training runs) on the task of predicting whether PD genes are also AD genes based on patterns of graph connectivity in AlzKB's heterogeneous knowledge graph. Across all metrics, a score of 1.00 represents the maximum possible performance.

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. Among the 73 PD genes in AlzKB, 8 genes (*FYN*, *DCTN1*, *SNCA*, *SYNJ1*, *RSP12*, *ATXN2*, *KCNIP3*, and *CHRNB1*; described in **Table 3**) were predicted to be AD genes. 7 of the genes were predicted to be AD genes in all 10 models, while *CHRNB1* was predicted in 7 of the 10 models.

Table 3: Parkinson's Disease genes predicted by graph-augmented Random Forest model to also be associated with Alzheimer's Disease.

Gene Symbol	Gene Name	Notes (From 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 Beta 1 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 Family Tyrosine Kinase	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 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)[58] on AlzKB and selected the highest performing of them to predict links between drugs and AD. Additional detail about the difference between these methods is provided in **Supplemental Information**.

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

Model name	Hits@1	Hits@3	Hits@10	MRR
RotatE	0.126	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

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. We therefore used RotateE to make predictions on the test set to obtain missing head entities with the template ([*drug*], DRUG_TREATS_DISEASE, Alzheimer's Disease). The top 10 predicted drugs are listed in **Table 5** along with their current approved usage and relevant clinical trial status pertaining to AD efficacy. Among the top 10 predictions, 3 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 by which the diseases could act on AD etiology and provides hypotheses to further explore their validity.

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

Drug Name	**Approved Indication(s)**	**AD-related clinical trial(s)**
Sumatriptan	Migraines and cluster headaches	
Nicotine	Nicotine withdrawal symptoms	NCT00018278 (Completed)
Pimozide	Tourette's Disorder	
Risperidone	Schizophrenia, bipolar mania, psychosis	NCT00034762 (Completed)
Flurbiprofen	Osteoarthritis and rheumatoid arthritis	
Sertraline	Depressive disorder, social anxiety disorder	NCT00086138 (Completed), NCT00009191 (Completed)
Clozapine	Schizophrenia	
Tamoxifen	ER+ 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	

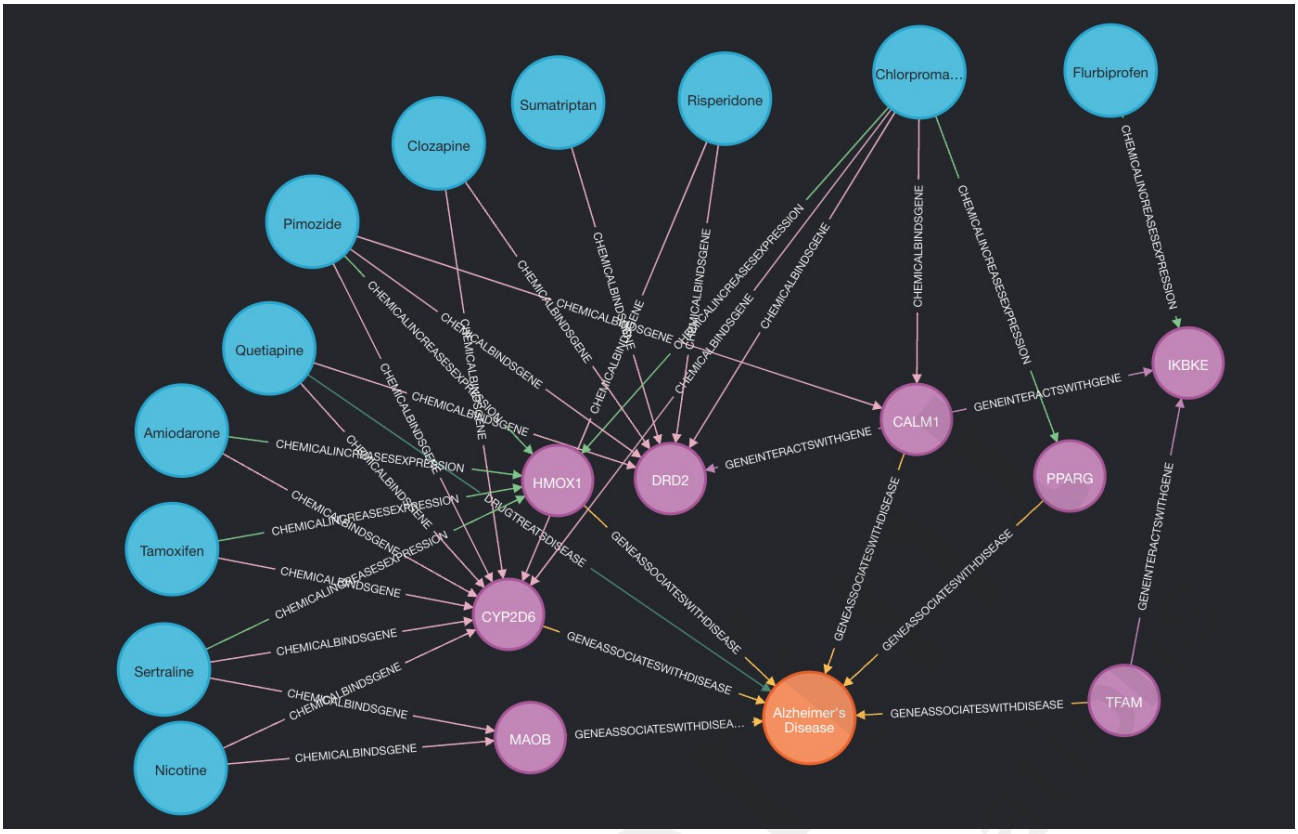

Figure 3: Spanning tree linking the 10 highest scoring AD drug predictions (listed in **Table 5**) to AD. Blue nodes are drugs, pink nodes are genes, and the orange node is AD. Genes on the shortest path between a drug and AD can be considered putative mechanistic explanations for how the drug may act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

Discussion

AlzKB is a freely available resource for the biomedical research community, with a primary goal of expanding the repertoire of therapies for AD via drug repurposing. Above, we described the current contents of AlzKB, the process of how it was constructed, and two 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 only 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 healthcare continue to increase the overall global life expectancy, we can expect the number of people with various forms of dementia to also increase. Since 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 towards finding therapies that reduce risk, slow the progression of disease, and/or reverse neuronal damage. Additionally, since there are various subtypes of AD with underlying mechanisms, any therapy might be effective for only some AD patients. Therefore, an essential step of 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

demonstrate, AlzKB provides a rich representation of existing knowledge about AD and the biological context in which it acts. The two ML-based use cases we presented above use real-world endpoints 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, clinical informatics, and 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. In order 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 virtually any domain and depends on (a.) defining an ontology using expert knowledge that formally describes the domain of interest and (b.) 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[59]. Since both knowledge bases share many of the same ‘core’ entities (genes, diseases, pathways, anatomical structures, etc.), 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 “**Proposing new therapeutic targets for AD”; 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.[60,61] Among the other identified genes, one (*CHRNB1*) is known to be involved in acetylcholine signaling[62,63]; another (*KCNIP3*) codes a protein that interacts with presenilin, and mutations in presenilin are causal for hereditary AD[64,65]. Some of these gene hits (*ATXN2*, *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[66]. For example, *DCTN1* encodes the dynactin-1 protein, and deficits in dynactin are connected to several neurodegenerative diseases; however, there is limited research linking it to AD[67,68].

Among the drug repurposing predictions (see **“Drug repurposing via graph data science”; 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 [69] and is connected to a strong comorbidity of migraine headaches and dementia in women[70]. Pimozide has been shown to reduce the aggregation of tau protein in mice [71] and is linked to AD in a number of unrelated *in silico* models[72]. 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[73]. 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 machine learning) 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. Since a central hypothesis of AD pathogenesis revolves around the abnormal 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 post-processing of existing knowledge sources or synthesis of new knowledge to achieve a good representation of genes, proteins, and functional/structural variants that are key to understanding AD.
Current machine learning models often do not generalize well to heterogeneous graphs, such as the one that comprises AlzKB's knowledge graph. This is largely because traditional models cannot utilize the network structure and heterogeneous nature of different entity types. Several promising algorithms can be used for prediction on heterogeneous graphs – including GraphSAGE [74] and metapath2vec [75] – but most fail to scale effectively when the number of node types or edge types grows. Since any effective therapy must be accompanied by a mechanistic understanding of how it functions, we also need to ensure that new heterogeneous graph machine learning models are *explainable*. With these in mind, we are using AlzKB as a motivating resource for designing new, cutting-edge algorithms that produce interpretable predictions over 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 like AlzKB into diverse artificial intelligence applications[76]. One application we are considering is using AlzKB to provide LLMs with formalized knowledge about AD that allow it 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 consists 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[77]. Furthermore, new predictions made using KGs 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 two nodes in a KG is interpreted as a "positive sample" for model training, absence of an edge simply means we do not know whether a relation 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[78], but further work is needed to determine whether these generalize well to AlzKB and similar highly heterogeneous biomedical KGs. 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 to unravel the etiology of AD. It is already a large, high-quality knowledge base from which graph-based AI/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 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 it can be 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 include in the main code distribution. Instructions for how to contribute to AlzKB are also available on the GitHub repository.
Since 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.

Software and Code Availability

AlzKB is publicly available online[11]. All pertinent source code and documentation is available on GitHub [25], or in an archived version on Zenodo[26]. Users have the option of hosting local and/or modified copies of AlzKB; see "**Obtaining AlzKB for local use and extending the knowledge graph"** for further information.

Conclusions

In this work, we introduced the Alzheimer's Knowledge Base 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 healthcare, 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 as well as generate entirely new, testable hypotheses to drive the future of drug repurposing and drug discovery.

Acknowledgements

AlzKB is supported by US National Institutes of Health grants U01-AG066833, R01-LM010098, R01-LM013463 (PI: Moore), and R00-LM013646 (PI: Romano).

Abbreviations

AA: Adamic-Adar

AD: Alzheimer's Disease
AI: Artificial Intelligence
AlzKB: Alzheimer's Knowledge Base
AUPRC: Area Under the Precision-Recall Curve
AUROC: Area Under the Receiver Operating Characteristic curve
CN: Common Neighbors
GWAS: Genome Wide Association Study
LLM: Large Language Model
ML: Machine Learning
MRR: Mean Reciprocal Rank
OWL 2: Web Ontology Language (version 2)
PA: Preferential Attachment
PD: Parkinson's Disease
RA: Resource Allocation
RDF: Resource Description Framework
TN: Total Neighbors
US FDA: United States Food and Drug Administration
XML: eXtensible Markup Language

References

1. 2022 Alzheimer's disease facts and figures. Alzheimers Dement 2022 Apr;18(4):700–789. PMID:35289055

2. Yiannopoulou KG, Papageorgiou SG. Current and Future Treatments in Alzheimer Disease: An Update. J Cent Nerv Syst Dis 2020;12:1179573520907397. PMID:32165850

3. Rabinovici GD. Controversy and Progress in Alzheimer's Disease - FDA Approval of Aducanumab. N Engl J Med 2021 Aug 26;385(9):771–774. PMID:34320284

4. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. Mol Neurodegener 2019 Aug 2;14(1):32. PMID:31375134

5. Aisen PS, Cummings J, Jack CR, Morris JC, Sperling R, Frölich L, Jones RW, Dowsett SA, Matthews BR, Raskin J, Scheltens P, Dubois B. On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimers Res Ther 2017 Aug 9;9(1):60. PMID:28793924

6. Fan L, Mao C, Hu X, Zhang S, Yang Z, Hu Z, Sun H, Fan Y, Dong Y, Yang J, Shi C, Xu Y. New Insights Into the Pathogenesis of Alzheimer's Disease. Front Neurol 2019;10:1312. PMID:31998208

7. Silva MVF, Loures C de MG, Alves LCV, de Souza LC, Borges KBG, Carvalho M das G. Alzheimer's disease: risk factors and potentially protective measures. J Biomed Sci 2019 May 9;26(1):33. PMID:31072403

8. Rodriguez S, Hug C, Todorov P, Moret N, Boswell SA, Evans K, Zhou G, Johnson NT, Hyman BT, Sorger PK, Albers MW, Sokolov A. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nat Commun 2021 Feb 15;12(1):1033. PMID:33589615

9. Tsuji S, Hase T, Yachie-Kinoshita A, Nishino T, Ghosh S, Kikuchi M, Shimokawa K, Aburatani H, Kitano H, Tanaka H. Artificial intelligence-based computational framework for drug-target prioritization and inference of novel repositionable drugs for Alzheimer's disease. Alzheimers Res Ther 2021 May 3;13(1):92. PMID:33941241

10. Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, Karnoub M, Nowotny P, Tacey K, Catanese J, Sninsky J, Brayne C, Rubinsztein D, Gill M, Lawlor B, Lovestone S, Holmans P, O'Donovan M, Morris JC, Thal L, Goate A, Owen MJ, Williams J. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet 2007 Apr 15;16(8):865–873. PMID:17317784

11. AlzKB - The Alzheimer's Knowledge Base. Available from: https://alzkb.ai/ [accessed Feb 24, 2023]

12. Daluwatumulle G, Wijesinghe R, Weerasinghe R. In Silico Drug Repurposing using Knowledge Graph Embeddings for Alzheimer's Disease. Proceedings of the 9th International Conference on Bioinformatics Research and Applications New York, NY, USA: Association for Computing Machinery; 2023. p. 61–66. doi: 10.1145/3569192.3569203

13. Nian Y, Hu X, Zhang R, Feng J, Du J, Li F, Bu L, Zhang Y, Chen Y, Tao C. Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing. BMC Bioinformatics 2022 Sep 30;23(6):407. doi: 10.1186/s12859-022-04934-1

14. Hsieh K-L, Plascencia-Villa G, Lin K-H, Perry G, Jiang X, Kim Y. Synthesize heterogeneous biological knowledge via representation learning for Alzheimer's disease drug repurposing. iScience 2023 Jan 20;26(1):105678. doi: 10.1016/j.isci.2022.105678

15. Binder J, Ursu O, Bologa C, Jiang S, Maphis N, Dadras S, Chisholm D, Weick J, Myers O, Kumar P, Yang JJ, Bhaskar K, Oprea TI. Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Commun Biol Nature Publishing Group; 2022 Feb 11;5(1):1–15. doi: 10.1038/s42003-022-03068-7

16. Sügis E, Dauvillier J, Leontjeva A, Adler P, Hindie V, Moncion T, Collura V, Daudin R, Loe-Mie Y, Herault Y, Lambert J-C, Hermjakob H, Pupko T, Rain J-C, Xenarios I, Vilo J, Simonneau M, Peterson H. HENA, heterogeneous network-based data set for Alzheimer's disease. Sci Data Nature Publishing Group; 2019 Aug 14;6(1):151. doi: 10.1038/s41597-019-0152-0

17. Robinson I, Webber J, Eifrem E. Graph Databases: New Opportunities for Connected Data. O'Reilly Media, Inc.; 2015. ISBN:978-1-4919-3086-1

18. Davis R, Shrobe H, Szolovits P. What Is a Knowledge Representation? AI Magazine 1993 Mar 15;14(1):17–17. doi: 10.1609/aimag.v14i1.1029

19. Musen MA. The protégé project: a look back and a look forward. AI matters ACM New York, NY, USA; 2015;1(4):4–12.

20. Hitzler P, Krötzsch M, Parsia B, others. OWL 2 web ontology language primer. 2009;

21. Tsarkov D, Horrocks I. FaCT++ Description Logic Reasoner: System Description. In: Furbach U, Shankar N, editors. Automated Reasoning Berlin, Heidelberg: Springer; 2006. p. 292–297. doi: 10.1007/11814771_26

22. Neo4j Community Edition. Neo4j; Available from: https://neo4j.com/ [accessed Oct 25, 2022]

23. Barrasa J, Cowley A. neosemantics (n10s): Neo4j RDF & Semantics toolkit. Available from: https://neo4j.com/labs/neosemantics/ [accessed Oct 25, 2022]

24. AlzKB Neo4j Browser. Available from: http://neo4j.alzkb.ai/browser/ [accessed Feb 24, 2023]

25. Romano JD, Truong V, Wang Z. AlzKB source code repository (GitHub). Epistasis Lab at Cedars Sinai; 2022. Available from: https://github.com/EpistasisLab/AlzKB [accessed Feb 24, 2023]

26. Romano J, Wang P. EpistasisLab/AlzKB: AlzKB first DOI release. Zenodo; 2022. doi: 10.5281/zenodo.7015728

27. Mortensen HM, Senn J, Levey T, Langley P, Williams AJ. The 2021 update of the EPA's adverse outcome pathway database. Sci Data 2021 Jul 12;8(1):169. PMID: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: Bairoch A, Cohen-Boulakia S, Froidevaux C, editors. Data Integration in the Life Sciences Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 124–131. doi: 10.1007/978-3-540-69828-9_12

29. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Research 2019 Jan 8;47(D1):D955–D962. doi: 10.1093/nar/gky1032

30. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015 Apr 15;2015(0):bav028–bav028. doi: 10.1093/database/bav028

31. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Research 2008 Jan 1;36(suppl_1):D901–D906. doi: 10.1093/nar/gkm958

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. Computational Toxicology 2019 Nov;12:100096. doi: 10.1016/j.comtox.2019.100096

33. Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, Cathey T, Transue TR, Spencer R, Wolf M. ACToR — Aggregated Computational Toxicology Resource. Toxicology and Applied Pharmacology 2008 Nov;233(1):7–13. doi: 10.1016/j.taap.2007.12.037

34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nat Genet 2000 May;25(1):25–29. doi: 10.1038/75556

35. The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E, Fey P, Thomas PD, Albou L-P, Ebert D, Kesling MJ, Mi H, Muruganujan A, Huang X, Mushayahama T, LaBonte SA, Siegele DA, Antonazzo G, Attrill H, Brown NH, Garapati P, Marygold SJ, Trovisco V, dos Santos G, Falls K, Tabone C, Zhou P, Goodman JL, Strelets VB, Thurmond J, Garmiri P, Ishtiaq R, Rodríguez-López M, Acencio ML, Kuiper M, Lægreid A, Logie C, Lovering RC, Kramarz B, Saverimuttu SCC, Pinheiro SM, Gunn H, Su R, Thurlow KE, Chibucos M, Giglio M, Nadendla S, Munro J, Jackson R, Duesbury MJ, Del-Toro N, Meldal BHM, Paneerselvam K, Perfetto L, Porras P, Orchard S, Shrivastava A, Chang H-Y, Finn RD, Mitchell AL, Rawlings ND, Richardson L, Sangrador-Vegas A, Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov DM, Harris MA, Oliver SG, Rutherford K, Wood V, Hayles J, Bähler J, Bolton ER, De Pons JL, Dwinell MR, Hayman GT, Kaldunski ML, Kwitek AE, Laulederkind SJF, Plasterer C, Tutaj MA, Vedi M, Wang S-J, D'Eustachio P, Matthews L, Balhoff JP, Aleksander SA, Alexander MJ, Cherry JM, Engel SR, Gondwe F, Karra K, Miyasato SR, Nash RS, Simison M, Skrzypek MS, Weng S, Wong ED, Feuermann M, Gaudet P, Morgat A, Bakker E, Berardini TZ, Reiser L, Subramaniam S, Huala E, Arighi CN, Auchincloss A, Axelsen K, Argoud-Puy G, Bateman A, Blatter M-C, Boutet E, Bowler E, Breuza L, Bridge A, Britto R, Bye-A-Jee H, Casas CC, Coudert E, Denny P, Estreicher A, Famiglietti ML, Georghiou G, Gos A, Gruaz-

Gumowski N, Hatton-Ellis E, Hulo C, Ignatchenko A, Jungo F, Laiho K, Le Mercier P, Lieberherr D, Lock A, Lussi Y, MacDougall A, Magrane M, Martin MJ, Masson P, Natale DA, Hyka-Nouspikel N, Orchard S, Pedruzzi I, Pourcel L, Poux S, Pundir S, Rivoire C, Speretta E, Sundaram S, Tyagi N, Warner K, Zaru R, Wu CH, Diehl AD, Chan JN, Grove C, Lee RYN, Muller H-M, Raciti D, Van Auken K, Sternberg PW, Berriman M, Paulini M, Howe K, Gao S, Wright A, Stein L, Howe DG, Toro S, Westerfield M, Jaiswal P, Cooper L, Elser J. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research 2021 Jan 8;49(D1):D325–D334. doi: 10.1093/nar/gkaa1113

36. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research 2019 Jan 8;47(D1):D1005–D1012. doi: 10.1093/nar/gky1120

37. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Valencia A, editor. eLife eLife Sciences Publications, Ltd; 2017 Sep 22;6:e26726. doi: 10.7554/eLife.26726

38. HuRI | Home. Available from: http://www.interactome-atlas.org/ [accessed Feb 24, 2023]

39. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, Readhead B, Tritsch SR, Hodos R, Hafner M, Niepel M, Sorger PK, Dudley JT, Bavari S, Panchal RG, Ma'ayan A. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. npj Syst Biol Appl 2016 Dec;2(1):16015. doi: 10.1038/npjsba.2016.15

40. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000 Jul;88(3):265–266. PMID:10928714

41. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 2011 Jan 1;39(Database):D52–D57. doi: 10.1093/nar/gkq1237

42. Schaefer C, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow K. PID: The Pathway Interaction Database. Nat Prec 2008 Aug 29; doi: 10.1038/npre.2008.2243.1

43. Himmelstein D, Pouya Khankhanian, Hessler CS, Green AJ, Baranzini S. PharmacotherapyDB 1.0: the open catalog of drug therapies for disease. figshare; 2016. p. 212598 Bytes. doi: 10.6084/M9.FIGSHARE.3103054.V1

44. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Research 2021 Jan 8;49(D1):D1388–D1395. doi: 10.1093/nar/gkaa971

45. Wu G, Haw R. Functional Interaction Network Construction and Analysis for Disease Discovery. In: Wu CH, Arighi CN, Ross KE, editors. Protein Bioinformatics New York, NY: Springer New York; 2017. p. 235–253. doi: 10.1007/978-1-4939-6783-4_11

46. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic

Acids Res 2016 Jan 4;44(D1):D1075–D1079. doi: 10.1093/nar/gkv1075

47. Palasca O, Santos A, Stolte C, Gorodkin J, Jensen LJ. TISSUES 2.0: an integrative web resource on mammalian tissue expression. Database 2018 Jan 1;2018. doi: 10.1093/database/bay003

48. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol 2012;13(1):R5. doi: 10.1186/gb-2012-13-1-r5

49. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, A. Miller R, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M. WikiPathways: connecting communities. Nucleic Acids Research 2021 Jan 8;49(D1):D613–D621. doi: 10.1093/nar/gkaa1024

50. Newman MEJ. Clustering and preferential attachment in growing networks. Phys Rev E 2001 Jul 26;64(2):025102. doi: 10.1103/PhysRevE.64.025102

51. Barabási A-L, Albert R. Emergence of Scaling in Random Networks. Science 1999 Oct 15;286(5439):509–512. doi: 10.1126/science.286.5439.509

52. Adamic LA, Adar E. Friends and neighbors on the Web. Social Networks 2003 Jul;25(3):211–230. doi: 10.1016/S0378-8733(03)00009-1

53. Zhou T, Lü L, Zhang Y-C. Predicting missing links via local information. Eur Phys J B 2009 Oct;71(4):623–630. doi: 10.1140/epjb/e2009-00335-8

54. Ali M, Berrendorf M, Hoyt CT, Vermue L, Sharifzadeh S, Tresp V, Lehmann J. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. Journal of Machine Learning Research 2021;22(82):1–6.

55. Gao Z, Ding P, Xu R. KG-Predict: A knowledge graph computational framework for drug repurposing. Journal of Biomedical Informatics 2022 Aug;132:104133. doi: 10.1016/j.jbi.2022.104133

56. Nussbaum RL, Ellis CE. Alzheimer's Disease and Parkinson's Disease. New England Journal of Medicine Massachusetts Medical Society; 2003 Apr 3;348(14):1356–1364. PMID:12672864

57. Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai S-M, Hasan Q. Application of network link prediction in drug discovery. BMC Bioinformatics 2021 Dec;22(1):187. doi: 10.1186/s12859-021-04082-y

58. Zamini M, Reza H, Rabiei M. A Review of Knowledge Graph Completion. Information 2022 Aug 21;13(8):396. doi: 10.3390/info13080396

59. Romano JD, Hao Y, Moore JH, Penning TM. Automating Predictive Toxicology Using ComptoxAI. Chem Res Toxicol 2022 Aug 15;35(8):1370–1382. PMID:35819939

60. 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 2020 Jul 30;9(8):E1807. PMID:32751526

61. Nygaard HB, van Dyck CH, Strittmatter SM. Fyn kinase inhibition as a novel therapy for Alzheimer's disease. Alzheimers Res Ther 2014;6(1):8. PMID:24495408

62. Lardenoije R, Roubroeks JAY, Pishva E, Leber M, Wagner H, Iatrou A, Smith AR, Smith RG, Eijssen LMT, Kleineidam L, Kawalia A, Hoffmann P, Luck T, Riedel-Heller S, Jessen F, Maier W, Wagner M, Hurlemann R, Kenis G, Ali M, Del Sol A, Mastroeni D, Delvaux E, Coleman PD, Mill J, Rutten BPF, Lunnon K, Ramirez A, van den Hove DLA. Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood. Clin Epigenetics 2019 Nov 27;11(1):164. PMID:31775875

63. Lombardo S, Maskos U. Role of the nicotinic acetylcholine receptor in Alzheimer's disease pathology and treatment. Neuropharmacology 2015 Sep;96(Pt B):255–262. PMID:25514383

64. Jo D-G, Lee J-Y, Hong Y-M, Song S, Mook-Jung I, Koh J-Y, Jung Y-K. Induction of pro-apoptotic calsenilin/DREAM/KChIP3 in Alzheimer's disease and cultured neurons after amyloid-beta exposure. J Neurochem 2004 Feb;88(3):604–611. PMID:14720210

65. Jin J-K, Choi J-K, Wasco W, Buxbaum JD, Kozlowski PB, Carp RI, Kim Y-S, Choi E-K. Expression of calsenilin in neurons and astrocytes in the Alzheimer's disease brain. Neuroreport 2005 Apr 4;16(5):451–455. PMID:15770150

66. Rosas I, Martínez C, Clarimón J, Lleó A, Illán-Gala I, Dols-Icardo O, Borroni B, Almeida MR, van der Zee J, Van Broeckhoven C, Bruni AC, Anfossi M, Bernardi L, Maletta R, Serpente M, Galimberti D, Scarpini E, Rossi G, Caroppo P, Benussi L, Ghidoni R, Binetti G, Nacmias B, Sorbi S, Piaceri I, Bagnoli S, Antonell A, Sánchez-Valle R, De la Casa-Fages B, Grandas F, Diez-Fairen M, Pastor P, Ferrari R, Álvarez V, Menéndez-González M. Role for ATXN1, ATXN2, and HTT intermediate repeats in frontotemporal dementia and Alzheimer's disease. Neurobiol Aging 2020 Mar;87:139.e1-139.e7. PMID:31810584

67. Aboud O, Parcon PA, DeWall KM, Liu L, Mrak RE, Griffin WST. Aging, Alzheimer's, and APOE genotype influence the expression and neuronal distribution patterns of microtubule motor protein dynactin-P50. Front Cell Neurosci 2015;9:103. PMID:25859183

68. Caroppo P, Le Ber I, Clot F, Rivaud-Péchoux S, Camuzat A, De Septenville A, Boutoleau-Bretonnière C, Mourlon V, Sauvée M, Lebouvier T, Bonnet A-M, Levy R, Vercelletto M, Brice A, French Clinical and Genetic Research Network on Frontotemporal Dementia/Frontotemporal Dementia–Amyotrophic Lateral Sclerosis. DCTN1 mutation analysis in families with progressive supranuclear palsy-like phenotypes. JAMA Neurol 2014 Feb;71(2):208–215. PMID:24343258

69. Zochodne DW, Ho LT. Sumatriptan blocks neurogenic inflammation in the peripheral nerve trunk. Neurology Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology; 1994 Jan 1;44(1):161–161. PMID:8290056

70. Liu C-T, Wu B-Y, Hung Y-C, Wang L-Y, Lee Y-Y, Lin T-K, Lin P-Y, Chen W-F, Chiang J-H, Hsu S-F, Hu W-L. Decreased risk of dementia in migraine patients with traditional Chinese medicine use: a population-based cohort study. Oncotarget 2017 Jul 8;8(45):79680–79692. PMID:29108348

71. Kim YD, Jeong EI, Nah J, Yoo S-M, Lee WJ, Kim Y, Moon S, Hong S-H, Jung Y-K. Pimozide reduces toxic forms of tau in TauC3 mice via 5′ adenosine monophosphate-activated protein kinase-

mediated autophagy. Journal of Neurochemistry 2017;142(5):734–746. doi: 10.1111/jnc.14109

72. Kumar S, Chowdhury S, Kumar S. In silico repurposing of antipsychotic drugs for Alzheimer's disease. BMC Neuroscience 2017 Oct 27;18(1):76. doi: 10.1186/s12868-017-0394-8

73. van Duijn CM, Hofman A. Relation between nicotine intake and Alzheimer's disease. BMJ 1991 Jun 22;302(6791):1491–1494. PMID:1855016

74. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. arXiv; 2018. doi: 10.48550/arXiv.1706.02216

75. metapath2vec | Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Available from: https://dl.acm.org/doi/10.1145/3097983.3098036 [accessed Oct 25, 2022]

76. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv; 2023. doi: 10.48550/arXiv.2306.08302

77. Zhu Y, Wang X, Chen J, Qiao S, Ou Y, Yao Y, Deng S, Chen H, Zhang N. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. arXiv.org. 2023. Available from: https://arxiv.org/abs/2305.13168v1 [accessed Nov 4, 2023]

78. Kefato ZT, Girdzijauskas S. Self-supervised Graph Neural Networks without explicit negative sampling. arXiv.org. 2021. Available from: https://arxiv.org/abs/2103.14958v4 [accessed Nov 4, 2023]

Supplementary Files

Multimedia Appendixes

Supplemental Information for "The Alzheimer’s Knowledge Base – A knowledge graph for therapeutic discovery in Alzheimer’s Disease research".
URL: http://asset.jmir.pub/assets/c1adfaf78704e5f18b7c6f8f6a164fda.docx