University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Pu, Y;Beck, D;Verspoor, K

Title:
Graph embedding-based link prediction for literature-based discovery in Alzheimer's Disease

Date:
2023-09-01

Citation:
Pu, Y., Beck, D. & Verspoor, K. (2023). Graph embedding-based link prediction for literature-based discovery in Alzheimer's Disease. Journal of Biomedical Informatics, 145, https://doi.org/10.1016/j.jbi.2023.104464.

Persistent Link:
https://hdl.handle.net/11343/337561

License:
CC BY

Journal of Biomedical Informatics 145 (2023) 104464

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Graph embedding-based link prediction for literature-based discovery in Alzheimer's Disease

Yiyuan Pu [a], Daniel Beck [a], Karin Verspoor [b,a,*]

[a] School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia
[b] School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia

ARTICLE INFO

Keywords:
Literature-based discovery
Alzheimer's Disease
Link prediction
Knowledge graph
Text mining
Graph embedding

ABSTRACT

Objective: We explore the framing of literature-based discovery (LBD) as link prediction and graph embedding learning, with Alzheimer's Disease (AD) as our focus disease context. The key link prediction setting of prediction window length is specifically examined in the context of a time-sliced evaluation methodology.
Methods: We propose a four-stage approach to explore literature-based discovery for Alzheimer's Disease, creating and analyzing a knowledge graph tailored to the AD context, and predicting and evaluating new knowledge based on time-sliced link prediction. The first stage is to collect an AD-specific corpus. The second stage involves constructing an AD knowledge graph with identified AD-specific concepts and relations from the corpus. In the third stage, 20 pairs of training and testing datasets are constructed with the time-slicing methodology. Finally, we infer new knowledge with graph embedding-based link prediction methods. We compare different link prediction methods in this context. The impact of limiting prediction evaluation of LBD models in the context of short-term and longer-term knowledge evolution for Alzheimer's Disease is assessed.
Results: We constructed an AD corpus of over 16 k papers published in 1977–2021, and automatically annotated it with concepts and relations covering 11 AD-specific semantic entity types. The knowledge graph of Alzheimer's Disease derived from this resource consisted of ~11 k nodes and ~394 k edges, among which 34% were genotype-phenotype relationships, 57% were genotype-genotype relationships, and 9% were phenotype-phenotype relationships. A Structural Deep Network Embedding (SDNE) model consistently showed the best performance in terms of returning the most confident set of link predictions as time progresses over 20 years. A huge improvement in model performance was observed when changing the link prediction evaluation setting to consider a more distant future, reflecting the time required for knowledge accumulation.
Conclusion: Neural network graph-embedding link prediction methods show promise for the literature-based discovery context, although the prediction setting is extremely challenging, with graph densities of less than 1%. Varying prediction window length on the time-sliced evaluation methodology leads to hugely different results and interpretations of LBD studies. Our approach can be generalized to enable knowledge discovery for other diseases.
Availability: Code, AD ontology, and data are available at https://github.com/READ-BioMed/readbiomed-lbd.

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia [1]. Symptoms include difficulty in remembering recent events, problems with language, disorientation, and loss of body functions. Lack of knowledge on precise molecular changes is one significant factor making the development of effective AD treatments difficult [2]. The prevalence and incidence of Alzheimer's disease have great public health impacts, including an increase in deaths, huge costs, and the burden of caring [3]. The biomedical literature is a key source of information about AD, especially on pathologic processes and clinical symptoms, facilitating insights into new research questions.

However, the considerable volume and the furious pace of publication of AD research findings make manual review of all relevant information challenging. For instance, a search for "Alzheimer's disease" on PubMed[1] retrieves more than 192k results at the time of writing this paper. Alzheimer's disease is also complicated on both molecular and clinical levels. Medical researchers might neglect implicit but

* Corresponding author at: School of Computing Technologies, RMIT University, Melbourne, Victoria, Australia.
E-mail addresses: yiyuanp1@student.unimelb.edu.au (Y. Pu), beck.d@unimelb.edu.au (D. Beck), karin.verspoor@rmit.edu.au (K. Verspoor).
[1] https://pubmed.ncbi.nlm.nih.gov/.

https://doi.org/10.1016/j.jbi.2023.104464
Received 22 January 2023; Received in revised form 29 July 2023; Accepted 30 July 2023
Available online 2 August 2023
1532-0464/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Statement of significance.

Summary	Description
Problem	The performance of current literature-based discovery methods (LBD) under the Alzheimer's Disease (AD) context is unknown. In the context of large-scale application of neural network link prediction methods, the experimental framework and evaluation methodology present key challenges.
What is already known	The latest neural network graph-embedding link prediction methods have shown good success in general recommendation tasks. Early efforts to frame LBD as link prediction with such methods have shown the potential to support knowledge discoveries; this was done in a specific disease context with targeted queries under the two-hop ABC (transitive closure) model of Swanson.
What this paper adds	This paper provides the first study of large-scale application of neural network graph embedding-based link prediction methods to LBD, going beyond the ABC paradigm to allow links to be predicted between any two nodes of the graph. We demonstrate an approach to establishing a knowledge graph specifically to support LBD in AD, and design an experimental framework based on time-slicing that provides us with labeled training data for link prediction at scale. The impact on link prediction evaluation of the prediction window length in time-slicing approaches is specifically examined.

vital connections when pursuing discoveries for AD. Literature-based discovery (LBD) methods aim to support researchers to automatically generate new hypotheses by combining information in separate papers [4], thereby helping to address information overload. LBD has been applied in the biomedical domain for different tasks, including drug discovery [5–8], drug repurposing [9–11], adverse drug reaction prediction [12], gene and protein function discovery [13–15], biomarker discovery [16], as well as cancer research discovery [17]. However, no prior study has specifically applied LBD for AD research discovery.

Link prediction approaches and graph-based approaches have been integrated into the LBD framework to allow scalable knowledge discovery [18,19]. These techniques, together with recent advances in LBD evaluation methodologies, improve the implicit knowledge discovery processes. For instance, employing time-sliced evaluation [20] for link prediction allows for the automatic construction of gold-standard datasets for training and testing and supports large-scale evaluation, addressing two of the biggest evaluation challenges in LBD [21]. Similarly, recently developed graph embedding models improve the prediction of distant and indirect discoveries. Although previous work [22] has explored how different input representations affect link prediction performance, no prior work has examined how different evaluation settings for link prediction affect the interpretations of knowledge discovery.

A key choice affecting the assessment of LBD models is the amount of time for an inferred discovery to be scientifically established. Different prediction window lengths may lead to different evaluation results and interpretations of LBD studies. We therefore examine the impact of limiting prediction evaluation of LBD models in the context of short-term versus longer-term knowledge evolution, based on a tailored AD knowledge graph.

More specifically, we apply natural language processing techniques for AD concept and entity recognition to a corpus of AD literature to build an entity co-occurrence graph, and then apply link prediction based on recent graph embedding methods to infer new connections representing putative new discoveries with varying prediction window lengths. This paper makes several key contributions, also summarized in Table 1:

- We demonstrate the adaptation of LBD specifically for the Alzheimer's Disease context, through incorporation of disease-specific semantic resources and construction of an AD-focused knowledge graph. We construct training and test datasets with both short-term and long-term contexts automatically from co-occurrences in the literature, comprehensively capturing the associations between entities. This represents the first large-scale resource for literature-based link prediction in a specific disease context.
- We provide a large-scale study of the application of graph embedding-based link prediction methods to LBD using the AD-focused annotated corpus and the derived AD knowledge graph, going beyond the original two-hop only ABC paradigm of Swanson [23] (see Section 2.1). Our work is the first in the LBD domain that does not undersample negatives from test sets, providing more accurately measured performance and correct ranking of link prediction methods.
- We examine the impact of varying the prediction window length in time-slicing approaches for the evaluation of LBD methods. We show that as the prediction window progresses over 20 years, the one-year prediction horizon we adopt in training hugely undercounts true positive predictions in performance evaluation, as compared to considering to a fixed future end date. This is the first study in LBD that highlights the importance of what we consider as given or new knowledge when evaluating link predictions.

The rest of the paper is organized as follows. In Section 2, we review related work in LBD. Section 3 presents our proposed four-stage approach and describes the key experimental configurations. In Section 4, we report the experiment results, followed by a discussion in Section 5, including the impact of reframing the prediction task and the limitations observed in the pipeline. Finally, Section 6 concludes the paper.

2. Related work

2.1. Literature-based discovery and link prediction

Literature-based discovery (LBD) was developed as a remedy for the biomedical literature overload issue more than three decades ago [19]. It aims at discovering hidden knowledge from scientific findings. Early work in LBD focused on manual methods, under a paradigm proposed by Swanson [23] known as the "ABC" approach. This approach takes advantage of transitive closure to infer relationships. Specifically, if two concepts A and B are known to be related, and concept B is also related to concept C, then it follows that A and C might be indirectly related. We can think of this as a "two-hop" (or transitive) model, traversing links from A to C via B. Early findings based on this model included asserting a connection between Fish Oils and Raynaud's Disease [23], hypothesizing how indomethacin relates to AD [24], and identifying 11 factors that are relevant to both migraine and magnesium [25]. More recent work has transitioned the manual hypothesis generation processes into automatic ones by developing computational techniques [19].

Starting in 2016, the LBD community has reframed LBD as a problem of link prediction [19,22,26]. Link prediction is the task of predicting potential links between two existing nodes in a network. In this approach, each pair of vertices that has not already been connected in a network is considered as a potential link and scored for linkage. Those above a given score threshold are considered to be *positive*, i.e. a link is predicted between the nodes, while those below are *negative*, i.e. no link between the nodes is predicted.

The links in a network can either be undirected [22,26,27] or directed [11,28], leading to two ways of indicating connections between the vertices — co-occurrence based and semantic relation-based [18]. For the co-occurrence based models, if two nodes appear in the same text window, an undirected edge is formed between the two nodes. For the semantic relation-based models, a semantic parser such as

SemRep [29] is used to point one node to another with a labeled type. In a semantic-based network, only pre-defined relations are captured; while a co-occurrence based network allows a more comprehensive capture of associations between nodes.

2.2. Graph embedding methods

Recent approaches in link prediction rely on graph embedding methods [30], which learn low-dimensional node representations as inputs for link prediction. Such methods include matrix factorization-based methods, random walk-based methods, and neural network-based methods.

Matrix factorization-based methods use adjacency matrix and matrix factorization to learn embeddings. Two well-known models under this category include HOPE [31] and GraRep [32]. HOPE preserves asymmetric transitivity by approximating high-order proximities of large-scale graphs, while GraRep integrates global structural information by capturing different k-step local relational information.

Random walk-based methods treat node sequences generated from random walks as word sequences and learn node representations through language modeling techniques. Example models include node2vec [33] and DeepWalk [34]. DeepWalk learns structural regularities through short random walks, while node2vec balances breadth-first search and depth-first search to capture graph structure. After generalizing a sequence of nodes with random walks, both DeepWalk and node2vec feed the sequence into the SkipGram model [35].

Another category of graph embedding method is neural network-based methods, such as LINE [36], GCN [37,38], GraphSAGE [39], and SDNE [40]. These models encode and preserve network structures with different neural network architectures. LINE preserves both the local and global network structures through first-order proximity and second-order proximity. GCN is able to encode both graph structures and node features. It uses a convolutional graph neural network and aims for semi-supervised node classification. A layer-wise propagation rule is formed from the first-order approximation of spectral convolutions on graphs. GraphSAGE learns node embeddings through sampling and aggregating features from the local neighborhood of each node. It uses a supervised graph neural network with a message passing learning paradigm. Both GCN and GraphSage can successfully work with heterogeneous graphs. SDNE is a semi-supervised deep model that is able to capture highly non-linear network structures through multiple layers of non-linear functions. First-order proximity and second-order proximity are exploited to preserve the network structure.

2.3. Evaluation in LBD

According to recent surveys [18,21], there are three research challenges when evaluating LBD systems: (1) creating a gold standard dataset, (2) automating the evaluation process, and (3) quantifying the evaluation outputs. Yetisgen-Yildiz and Pratt [20] first proposed a time-slicing based evaluation methodology to compare the performance of LBD systems on the task of discovering new discoveries for a single term under Swanson's ABC paradigm [23]. The proposed evaluation methodology picks a cut-off date for all the existing data. All the data before the cut-off date are viewed as past data, while all the data after the cut-off date are viewed as future data that are unknown. The past data are used as the prior knowledge to predict the implicit knowledge in the future. The data appearing in the future data but not in the past data are viewed as true discoveries. This setting makes it possible to have a gold standard dataset and automate the evaluation process. Information retrieval metrics such as Precision, Recall, F-measure, Precision at k, and Mean average precision are used in time-slicing quantification [20]. While this evaluation methodology has regularly been used to examine link prediction models [22,27,41], as pointed out by [42], the effect of temporal distance in the evaluation of link prediction systems should be highlighted.

3. Materials and methods

This section presents our four-stage proposed approach to exploring LBD for AD. The first stage involved collecting an AD-specific corpus (Section 3.1). In the second stage, a knowledge graph for AD was constructed by identifying AD-specific concepts and relations between them from this corpus (Section 3.2). In stage 3, we constructed 20 pairs of training and testing data with the time-slicing methodology, and apply and evaluate several link prediction methods in stage 4. Section 3.3 describes these link prediction experiments. The impact of varying prediction window lengths is also examined.

3.1. AD corpus collection

The AD corpus was based on an extensive bibliography of Alzheimer's Disease literature collected by an expert in the field, Professor Colin Masters,[2] who discovered the cause of Alzheimer's Disease — proteolytic neuronal origin of the A*beta* amyloid protein. The bibliography has been created and maintained with the following procedures. First, a literature search was conducted every 2 weeks on Web of Science, with the keywords *Alzheimer's disease (AD), A-beta (or Aβ), amyloid-beta (or amyloid-β), alpha-secretase (or α-secretase), beta-secretase (or β-secretase), amyloid precursor protein (APP), BACE (or BACE1), mild cognitive impairment (MCI), Notch, [C11]PiB, presenilin, and Apolipoprotein E (ApoE or ApoE4)*. Then, Professor Colin Masters selected articles with his expertise from the search results and added the new entries to the bibliography. This corpus represents an expert-curated and hence reliable source of Alzheimer's Disease knowledge.

We used the papers in the bibliography between 1977 and 2021 as the source data for this study. We preprocessed the bibliography to enable the collection of the corresponding text of the papers. Titles and abstracts were used as the document representation for each paper. PMID, the unique identifier for the paper in PubMed, was retrieved to access PubTator Central (PTC) [43], which includes a repository of annotations over PubMed abstracts. We used Esearch[3] to link the papers without PMID entries in the bibliography to corresponding PMIDs.

3.2. AD knowledge graph construction

To construct a knowledge graph based on the collected AD corpus, we annotated relevant concepts and entities in the corpus, and established links between them based on co-occurrence. The framework is shown in Fig. 1.

3.2.1. AD-related biomedical knowledge resources

To create a knowledge graph tailored to AD, we used two biomedical knowledge resources: the Neuropsychological Integrative Ontology (NIO) [44] and Pubtator Central [45]. This allows the entities in the knowledge graph reach the granularity level for AD.

Neurocognitive integrated ontology. The Neurocognitive Integrated Ontology (NIO) aims to model "the early detection of MCI (Mild Cognitive Impairment) with a high probability of conversion to AD" [44]. MCI is a transitional phase before dementia, so detecting it early helps intervene or detect Alzheimer's disease. The ontology covers concepts from four interrelated domains in Alzheimer's Disease: (1) neuropsychological testing, (2) cognitive processes, (3) brain areas, as well as (4) neurodegenerative disease. The NIO was built by integrating existing ontologies for the four domains. The corresponding coverage and the ontologies are shown in Table 2.

[2] Professor Colin Masters FRCPath, FRCPA, FFSc, FAA, FTSE, FAHMS is a distinguished neuropathology researcher at the Florey Institute in Melbourne, Australia. He is a collaborator in the Cognitive Computing for Medical Technologies training center that supported this work.

[3] https://www.ncbi.nlm.nih.gov/books/NBK25499/.

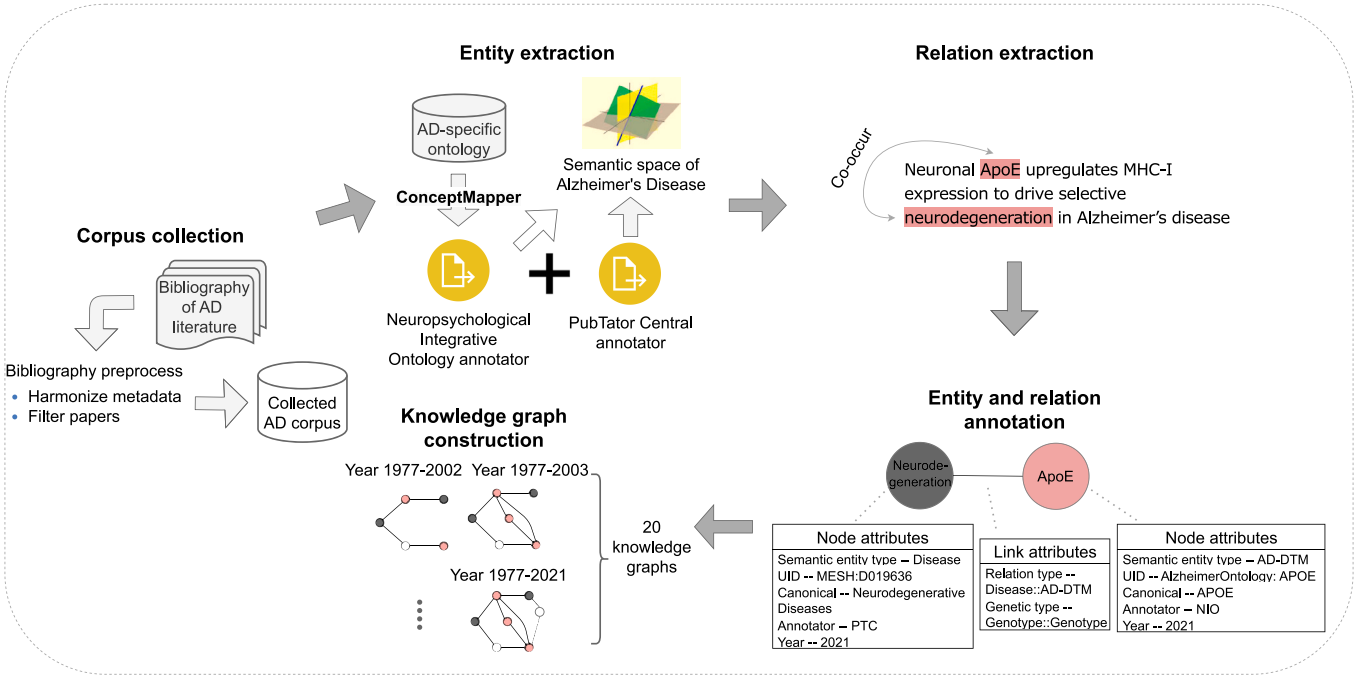


Fig. 1. The framework of Alzheimer's Disease knowledge graph construction. First, we collected an AD corpus from an extensive bibliography of AD literature. Then, we extracted entities/nodes by creating an AD-specific annotator and complementing it with an advanced biomedical entity recognition tool. Next, relations/links between the entities were extracted if two entities co-occur on the citation level. Entities and relations had node attributes and link attributes respectively. At last, the entity co-occurrence graph was constructed with entities as nodes and relations between the entities as links. The knowledge graph construction process was repeated 20 times by segmenting the AD corpus with 20 progressing years.

Table 2
Four domains, corresponding coverage, and ontologies in NIO.

Domain	Coverage	Ontologies[a]
Neurodegenerative disease	AD diagnosis, treatment, and molecular mechanisms	ADO
Brain areas	Brain anatomy and brain imaging	FMA, NDDO
Neuropsychological testing	Socio-demographic data, neuropsychological tests, and testing results	NPT
Cognitive processes	Mental processes, traits, and cognitive functions	MF

[a]ADO: Alzheimer Disease Ontology [46], FMA: Foundation Model of Anatomy [47], NDDO: Neurodegenerative Disease Data Ontology [48], NPT: Neuro-Psychological Testing Ontology [49], MF: Mental Functioning Ontology [50].

Through discussion of preliminary annotation results based on NIO with AD researchers in our research group, we identified that the vocabulary from NIO does not comprehensively capture all concepts relevant to describing Alzheimer's Disease. Analysis of the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) database suggested specific groups of concepts are missing in the NIO. The AIBL database includes four categories: lifestyle, biomarkers, neuroimaging, as well as clinical and cognitive characteristics. Through a comparison between the AIBL database and the NIO, we identified that genes, proteins, diseases, chemicals, and lifestyles are lacking in the NIO.

PubTator central. We filled in most of the concepts missing from the NIO with PubTator Central (PTC). PTC is a web-based repository of automatically derived entity annotations with over 29 million abstracts from PubMed and 3 million full-text papers from PMC Text Mining [43,45]. The PubTator tool is an integrated tagger and supports annotation of six concept types: genes/proteins, genetic variants, diseases, chemicals, and cell lines.

3.2.2. AD annotators

We created an AD-specific annotator with the Neuropsychological Integrative Ontology (NIO) as the source by adapting the ConceptMapper [55] annotation tool, utilizing a pipeline developed for pathogen annotation over PubMed citations [56]. ConceptMapper is a dictionary-based annotation tool that identifies named entities from a controlled vocabulary, that has been shown to have good performance in biomedical concept recognition tasks [57]. We built our **NIO annotator** tool with ConceptMapper coupled with a dictionary derived from the NIO.

We retrieved annotations of genes, proteins, diseases, and chemicals from the PTC repository for each article in the corpus. We refer to this process as the **PTC annotator**.

3.2.3. AD knowledge graph

The knowledge graph for Alzheimer's Disease can be denoted as $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of links (edges). We used the concept and entity annotations over the AD corpus to construct the knowledge graph of Alzheimer's Disease. The knowledge graph was constructed with AD concepts as nodes. Links represented relations between those concepts. We used *node*, *entity*, and *concept* interchangeably.

Nodes and their semantic entity types. Nodes were derived from entity annotations in a PubMed citation in the AD corpus. Each entity extracted from the corpus was mapped to a unique identifier (UID), which was then associated with a canonical name. For instance, for the entity with UID *NDDO:NDDO_10000767*, the associated canonical name is *Early Mild Cognitive Impairment*.

Each node in the knowledge graph was also assigned a semantic entity type as a node attribute. For nodes annotated by the NIO annotator, the semantic entity types came from prefixes of UIDs and the four domains. For nodes annotated by the PTC annotator, the semantic

Table 3
Six domains, corresponding coverage, and taggers in PTC.

Domain	Coverage	Tagger
Gene	Chromosomes, organelles, plasmids, viruses, transcripts, and proteins[a]	GNormPlus [51]
Variant	Single nucleotide variation, short multi-nucleotide changes, etc.[b]	tmVar 2.0 [52]
Disease	Infections, Neoplasms, Eye Diseases, etc.[c]	TaggerOne [53]
Chemical	Inorganic Chemicals, Organic Chemicals, Heterocyclic Compound, etc.[c]	TaggerOne [53]
Species	Homo sapiens, Mus musculus, etc.[d]	SR4GN [54]
Cell Line	Immortalized cell lines, naturally immortal cell lines, finite life cell lines, etc.[e]	TaggerOne [53]

[a] https://www.ncbi.nlm.nih.gov/books/NBK3841/#EntrezGene.Quick_Start.
[b] https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/.
[c] https://meshb.nlm.nih.gov/treeView.
[d] https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1.
[e] https://www.cellosaurus.org/description.html.

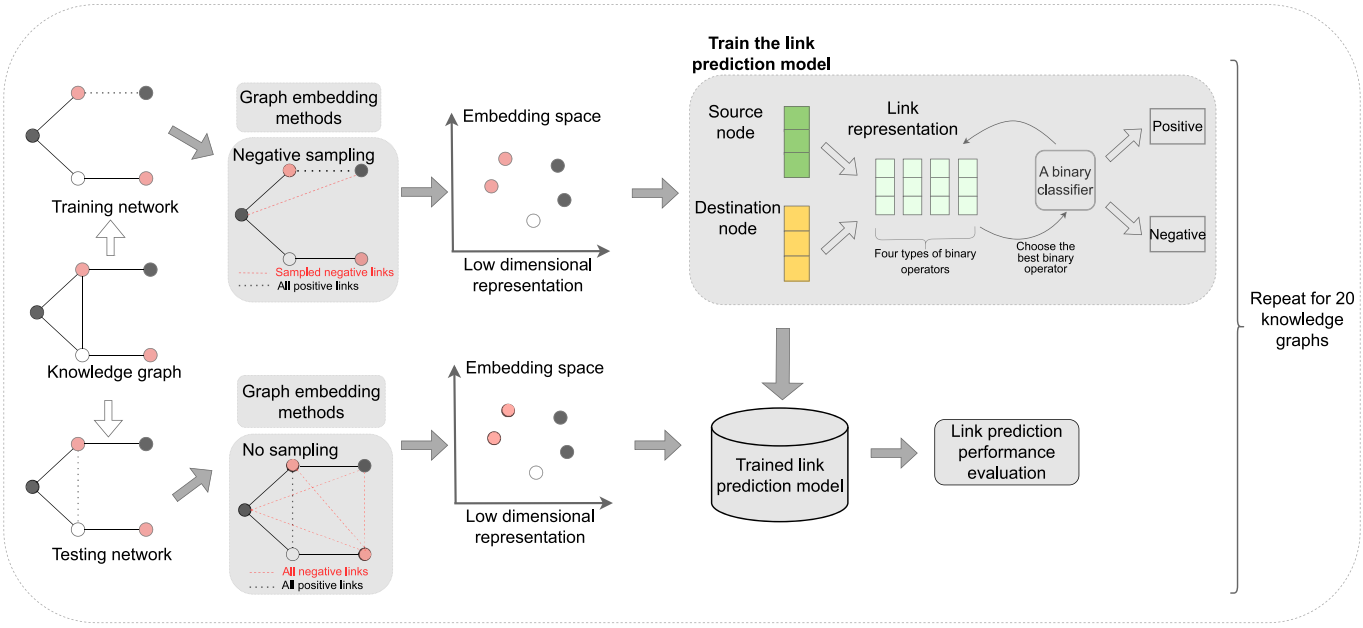


Fig. 2. The framework of knowledge inference with link prediction models for the Alzheimer's Disease knowledge graph. First, we split the knowledge graph into a training network and a testing network with time slicing. (Top): Training data were derived from the training network. Negative links in the training network were processed with negative sampling to maintain a 1:1 ratio of positive and negative links. Graph embedding methods were implemented to encode the network into a low-dimensional representation in an embedding space. Then, the embeddings of the source node and the destination node for each potential link were retrieved from the embedding space. Four types of binary operators were used to integrate the source node embedding and the destination node embedding. The link representations were used to train a binary classifier. (Bottom): The testing network went through the same graph embedding process. Yet, all negative links in the testing network were used, together with positive links, to evaluate the trained link prediction model. The knowledge inference process was repeated for 20 knowledge graphs.

entity types include Gene, Variant, Disease, Chemical, Species, and Cell Line [43]. The corresponding coverage and taggers are shown in Table 3. Each semantic entity type was further classified into genotype and phenotype as the genetic type. For instance, the nodes with the semantic entity type of *BrainArea* belong to *Phenotype*, while those of *Gene* belong to *Genotype*.

Link formation. In this study, we established relations between two entities by using co-occurrence on the citation level (title and abstract). If two entities appeared in the same citation, then a co-occurrence link was formed between them. We opted for using co-occurrence rather than semantic predicates as relations because initial extractions of semantic triples with SemRep [29] showed inaccuracies and low coverage of AD-related concepts and relations. A discussion with AD researchers in our group also suggested utilizing association rather than causation because the intricacies of relations between AD concepts (e.g., the relations between two AD-related genes may change after five years) require more clear definition by collaborating with experts. The strategy of co-occurrence also allows us to automatically generate labeled data at scale for LBD.

Link types. Following previous work [58], we defined a relation type by combining the semantic entity types of the two nodes that form the link. For instance, if two nodes belong to the Gene type and the Disease type respectively, then the relation type is Gene::Disease.

3.3. AD link prediction

To infer implicit links in the Alzheimer's Disease knowledge graph, we performed graph embedding-based link prediction on 20 time-sliced datasets. The workflow is illustrated in Fig. 2. We used *graph* and *network* interchangeably.

3.3.1. Task description and notation

We framed link prediction as a binary classification task: predicting whether a link forms between a pair of vertices that do not have a link between them by leveraging the history of link formation between the vertices. If the prediction is that a link forms, then the label for the classification is "positive"; otherwise, the label is "negative". Formally, we used $G_t = (V_t, E_t)$ to denote the network at the timestamp t and used $G_{t+1} = (V_{t+1}, E_{t+1})$ to denote the network at the next timestamp $t+1$. All links in a fully connected graph at G_t were defined as E_u^t. We defined the link prediction task as to predict whether the potential link E_{ij}^t at G_t that may be formed from the node pair (V_i, V_j), forms a link

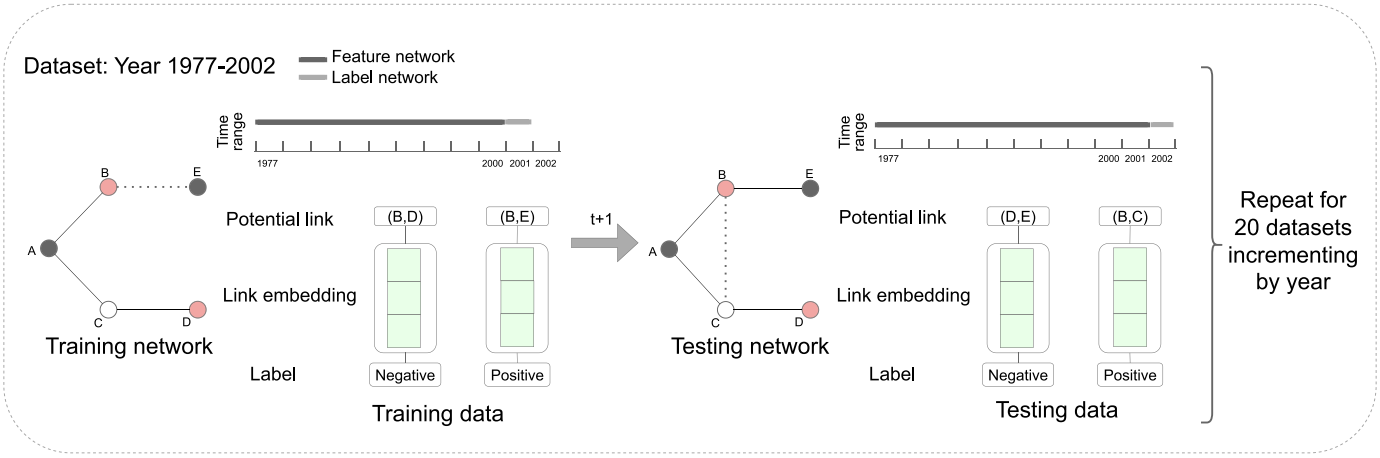


Fig. 3. The framework of training and testing data construction utilizing the time-slicing methodology. We used the annotations of relevant entities and links in the AD corpus between 1977 and 2002 as an example. (Left): The training data was constructed with link prediction features derived from the feature network (shown as the black period in the time range) and labels derived from the label network (shown as the grey period in the time range). Based on the feature network made from existing links (solid lines connecting nodes A,B,C,D), we achieved the link embedding for potential links connecting node pairs (B,D) and (B,E), respectively. Based on the label network made from the dashed lines, we knew that the link between B and D did not form, and the link between B and E formed. Then the time progressed to the next year, where we used the testing network to create the testing data. (Right): The testing data was constructed following the same process as the training data, shifted forward by one year. This construction process was repeated for 20 datasets based on progressing time slices.

at the timestamp G_{t+1} (i.e., $E^t_{ij} \in E_{t+1}$), where $V_i \in V_t$, $V_j \in V_t$, and $(V_i, V_j) \in E^t_u - E_t$.

3.3.2. Training and testing data construction

We used the time-slicing approach [11,20,26] to create training and testing data. Time slicing is a common technique in link prediction to split the dataset into the training and testing data with a cut-off date.

Following the previous study [42], we generated the training and testing data by constructing two pairs of feature and label networks. The framework is illustrated in Fig. 3. The key to constructing the training and testing data is to capture the link formation process between the feature network and the label network. The construction process can be thought of as taking snapshots: at the very beginning, all nodes are in the network, yet without any links. At each timestamp, links are formed and a snapshot of the network is taken. As time progresses, the network includes more and more links.

3.3.3. Key considerations in link prediction settings

The link prediction model is complex in terms of evaluation characteristics. Different experimental configurations lead to variations in results, which could result in entirely opposite interpretations [42]. Thus, we identified several considerations for link prediction.

Class imbalance. Class imbalance is an issue in link prediction dataset construction [26,59]. All potential links at any timestamp are ${}^{V}C_2$, where V corresponds to all possible nodes at that timestamp. This creates an enormous number of negative links compared to the number of positive links. Following previous studies [22,26,59], we used undersampling to reduce the number of negative links and maintain a 1:1 ratio of positive and negative links as training data for a binary classifier. Empirical results from [42] showed that undersampling negatives from testing data leads to an incorrect measure and ranking of link prediction model performance. Thus, we used all negative links in the testing data for a full picture of the performance of link prediction models.

Prediction window. To keep the freshness of data, implicit knowledge in the immediate following year of the cut-off year was used to evaluate the predictions based on the prior knowledge (Section 3.3.2). However, a predicted link that does not show up in the immediately following year could appear in the networks in the more distant future. This turns false positives into future true positives, and thus affects evaluation results. The first proposed evaluation methodology in LBD systems defined the prediction window as from the testing year of the dataset (or the cut-off year) to the end of the horizon [20]. We wondered what is the effect of varying how far in the future we allow the hypothesized links to emerge as discoveries. In other words, *What is the impact of limiting prediction evaluation of LBD models in the context of short-term vs longer-term knowledge evolution?*

We implemented the models on all datasets by changing the one-year window to the length of all the years from the cut-off date to 2021. For instance, for the dataset *Year 1977–2002*, the prior knowledge was derived between 1977 and 2001. We compared the evaluation results of (1) when the link prediction is evaluated only against the new knowledge added in 2002, and (2) when the link prediction is evaluated against the full network reflecting all knowledge through 2021. We only considered nodes from the original network. Since predicted links may appear in 2003 and later as knowledge grows, we expected a change of false positives at prediction.

Uninformative term removal. Previous LBD studies used uninformative term removal as one pre-processing step. The uninformative terms were defined as "providing no new or interesting information to the user" [21], such as terms that are general, obvious, uninteresting, or correlated with most terms [21,60]. To remove uninformative terms, we used node degrees to rank the nodes in the AD knowledge graph. At the 75% quartile, the node degree is 2000. This indicates that each of the top 25% connected nodes in the graph is connected with more than 2000 other nodes. We selected the 75% quartile of the degree distribution as the threshold for correlating with most terms. The distribution of the node degrees is in the supplementary material. Then, we examined the top-ranked nodes for whether they are general, obvious, or uninteresting. Among the 131 nodes with a degree more than 2000, 42 are general, representing canonical names such as *disease*, *protein*, *model*, and *intervention*; 27 are obvious, such as *Alzheimer Disease*, *brain*, and *amyloid beta protein*; and 12 are uninteresting, such as *role*, *other*, and *regulates*. The full list of uninformative terms is in the supplementary material.

New nodes. Another key consideration is the new nodes in the label network but not in the feature network. As time progresses, the network grows from the feature network to the label network. New nodes that have never appeared in previous years appear in the network. For instance, if using the annotations between 1977 and 2019 to build the

feature network and the annotations between 2020 and 2021 to build the label network, the label network has 236 new nodes and 21 193 new links compared to the feature network. Among the 21 193 new links, 7957 links are formed between *New node-New node* or *New node-Old node*, while 13 236 are new links formed between *Old node-Old node*. The complete statistics for all existing nodes and new nodes in each year are in the supplementary material. Previous work argued for removing the links that are known to be impossible for the models to predict from the feature network [42]. Apart from GraphSage, there is no way for the other 7 graph embedding models to predict the appearance of the 7957 links involving new nodes. We opted for ignoring the links involving new nodes and defining the knowledge inference task as returning the most confident set of link predictions only for existing nodes. Therefore, for each feature network and label network, we constrained the nodes to be in the feature network only.

3.3.4. Graph embedding learning

We followed previous studies [22,30] that defined the link prediction as a binary classification problem to choose graph embedding models. We explored three categories of graph embedding methods: HOPE [31] and GraRep [32] for matrix factorization-based models; DeepWalk [34] and node2vec [33] for random walk-based models; and LINE [36], SDNE [40], GCN [37], and GraphSage [39] for neural network-based models. These graph embedding methods differ in encoding the network into a low-dimensional representation in the embedding space. We refer to Section 2.2 for a detailed introduction of the graph embedding models. The learned node embeddings were used to train a binary classifier to perform link prediction as a downstream task. Since it is a link prediction task, the node embeddings need to be turned into edge embeddings as features for the predicted links. Following the previous study [33], we considered four types of binary operators to integrate two node embeddings into the link embedding: Hadamard, Weighted-L1, Weighted-L2, and Average.

3.3.5. Implementation of link prediction models

All preprocessing steps were done using Bash and Python scripts. HOPE, GraRep, DeepWalk, node2vec, and LINE graph embeddings were implemented in PyTorch using the CogDL package [61]. GCN and GraphSage were implemented in Tensorflow using the StellarGraph package [62]. For hyperparameters, we followed guidelines of previous work on biological networks [30]. The details of hyperparameters for each graph embedding method are in the supplementary material.

3.4. Evaluation

3.4.1. Precision–recall curve

We followed the evaluation metrics used in previous studies [20,22,41] that defined the link prediction task as a binary classification problem to evaluate the results. We adopted the standard metrics of precision and recall, with respect to the label network. Precision is defined as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

and recall as:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

In this study, a True Positive (TP) is a positive predicted link that exists in the label network. A False Positive (FP) refers to a positive prediction of a link that does not exist in the label network. A False Negative (FN) refers to a potential link predicted as negative (no link) that does in fact exist in the label network.

In our experiments, after applying link prediction models for the AD knowledge graph with each of the 20 pairs of training and testing data, one precision–recall curve was plotted at each threshold with different precision and recall values. AUPRC (Area under Precision–Recall curve) was used as a summary for the PR curve to interpret model performance.

3.4.2. Precision at k

Precision at k is a metric that uses a ranked list of predictions, and calculates the precision of only the top k predictions. This is useful to understand how many of those top k predictions are correct; the denominator for the precision calculation is exactly k while the numerator is the number of True Positives in that set of k. Here, we used the score produced for a prediction by the model to rank predictions, thereby only considering the top-scoring predictions.

We applied this metric in Section 5.1 to examine the impact of prediction window length in link prediction settings. Since the size of the knowledge graph for each dataset varies, we followed previous work [22] that chose the k values based on a fixed percentage of possible positive links, rather than a fixed absolute number. Since an expert who examines the predictions is more likely to focus on the top-returned results, we reported the experiment results based on a small k in the main manuscript. We set $k = 10\%$ of the positive links in the testing network of the dataset, to quantify the changes between the context of short-term knowledge evolution and longer-term evolution. Results on additional k values can be found in the supplementary material.

3.4.3. Baselines

To compare the performance of link prediction models, we used a dummy model that predicts every potential link as positive as the baseline for each pair of the training and testing datasets. The dummy model shows the worst case in the link prediction task on the AD knowledge graph.

4. Results

4.1. AD corpus collection

The bibliography of AD literature includes 28 084 papers published between 1977 and 2021. After excluding the papers without titles or abstracts, the number of papers remaining is 21 840. After linking the papers to PMIDs via Esearch,[4] we further excluded 5388 papers that could not be linked to PMIDs. The remaining 16 452 papers were used as the corpus for the literature-based discovery task in this study. The distributions of papers with and without abstracts and PMIDs for each year are illustrated in the supplementary material.

4.2. AD knowledge graph construction

4.2.1. Semantic entity types

The NIO annotator covers 5 semantic entity types in the AD knowledge graph, namely AD-DTM, BrainArea, NeuroTest, CogFunc, and ADPrep.[5] The first four semantic entity types correspond with the four domains described in Table 2. The *ADPrep* semantic entity type covers regions, entities, predicates, units, descriptors, and data types. A detailed breakdown of the construction of these 5 semantic entity types is shown in the supplementary material. The PTC annotator covers 6 semantic entity types in the AD knowledge graph, namely Gene, Variant, Disease, Chemical, Species, and Cell Line, corresponding to concept types in Table 3.

[4] https://www.ncbi.nlm.nih.gov/books/NBK25499/.

[5] AD-DTM: Alzheimer's Disease Diagnosis, Treatment, and Molecular mechanisms. BrainArea: Brain areas. NeuroTest: Neuropsychological testing. CogFunc: Cognitive processes. ADPrep: Alzheimer's Disease Prepositions.

Table 4
Statistics of entities in the knowledge graph of Alzheimer's Disease (Year 1977–2021).

Semantic entity type	Mention[a]	Entity[b]	Annotator	Genetic type
AD-DTM	70 375	787	NIO	Genotype
BrainArea	4948	117	NIO	Phenotype
NeuroTest	806	15	NIO	Phenotype
CogFunc	4635	131	NIO	Phenotype
ADPrep	9134	97	NIO	Phenotype
Gene	20 442	4097	PTC	Genotype
Variant	1976	1314	PTC	Genotype
Disease	16 209	1299	PTC	Phenotype
Chemical	20 443	2809	PTC	Genotype
Species	2025	462	PTC	Phenotype
Cell Line	1702	221	PTC	Genotype
Total	153 532	11 349	–	–

[a]Annotations in all papers in all years.
[b]Unique entities/nodes in the knowledge graph.

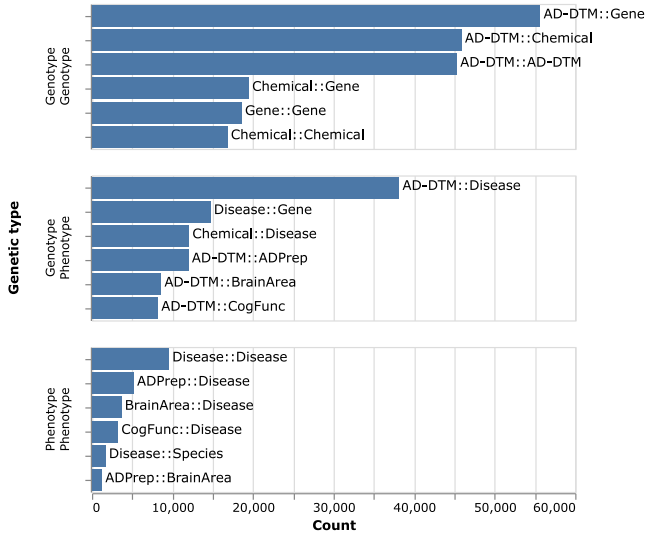


Fig. 4. Top 6 relation types among each genetic type. The text on the right side of each bar represents the relation type derived from the semantic entity type of the source node and the destination node of each link.

4.2.2. Descriptive statistics of the knowledge graph

As shown in Table 4, the AD knowledge graph has 11 349 nodes, including 153 532 mentions in the collected AD corpus. Statistics for the top 6 relation types grouped by the genetic types are shown in Fig. 4. Overall, there are 393 808 relationships between all edges, among which 34% relationships are between genotype and phenotype, 57% relationships are between genotype and genotype, and 9% relationships are between phenotype and phenotype. Among them, the relation types of AD-DTM::Disease, Disease::Gene, and Chemical::Disease are the top 3 link types among genotype and phenotype; the relation types of AD-DTM::Gene, AD-DTM::Chemical, and AD-DTM::AD-DTM are the top 3 link types among genotype and genotype; while Disease::Disease, ADPrep::Disease, and BrainArea::Disease are the top 3 link types among phenotype and phenotype.

4.3. AD link prediction

4.3.1. Training and testing data construction

We created 20 pairs of training and testing networks (Table 5). Each pair was split with time slicing from a knowledge graph built from entities and relations extracted from the AD corpus in a time range.

The data constructed from the training and testing networks are shown in Table 6. Each row corresponds to one row in Table 5. For instance, for the networks ranging from 1977 to 2002, the training data involves 4309 positive links. The positive links are the new links appearing in the label network of the training network. The negative links are potential links in the feature network that do not appear in the label network.

4.3.2. Evaluation results

Table 7 reports the AUPRC results of link prediction with eight graph embedding models and the baseline method on the 20 datasets. As time progresses, the link prediction task becomes harder. The number of nodes increases, and the density of the feature network decreases. The number of edges increases at a slower rate than all possible connections in the network. This indicates that the binary classifier has been trained with negative links sampled from a larger number of possible links. Training becomes more difficult and the change in density from the feature network to the label network also decreases. Class imbalance becomes more severe in the testing data: the rate increases from 10^{-2} in 2002 to 10^{-6} in 2021.

We observed that SDNE consistently outperforms other models for all 20 datasets in inferring new knowledge as the prediction task becomes more difficult. GCN also has consistently strong performance. This indicates that these link prediction models provide the most confident set of predictions as time progresses. The most difficult link prediction task is on the last pair of training and testing data (Dataset 1977–2021), where the density of the network is the smallest and the class imbalance rate is the highest.

Furthermore, we had the following key observations and analyses:

- For the Matrix factorization-based models, GraRep generally performs better than HOPE on the first 10 datasets and performs worse with the datasets from later years, where the link prediction tasks become harder. This indicates that capturing asymmetric transitivity helps capture structures of large-scale graphs. Although capturing global structural properties of the graph with k-step loss functions helps learn graph representations, it is not scalable.
- For the Random walk-based models, node2vec performs better than DeepWalk on all of the datasets excluding the first dataset, where the link prediction task is the easiest. This indicates that node neighborhoods constructed from biased random walk procedures with multiple sampling strategies lead to richer graph representations compared with the uniform sampling strategy. A diverse node neighborhood benefits link prediction tasks when the prediction setting becomes more challenging.
- For the Neural network-based models, SDNE consistently and significantly outperforms the other neural network-based models for all 20 datasets, with each dataset gradually becoming more difficult. This indicates that capturing highly non-linear network structure with deep learning and preserving the first-order and second-order proximity help learn useful network representations and is robust to sparse networks. LINE performs way worse than the SDNE model for all the datasets. This demonstrates that a shallow structure and the direct concatenation of the representations of first-order and second-order proximity are not enough to preserve graph structure information. GCN also consistently obtains better link prediction performance than other models, which indicates the capability of autoencoders to learn graph latent representations. GraphSage performs the worst among the neural network-based models, which indicates that learning embeddings by sampling and aggregating features from each node's local neighborhood is not enough for the downstream link prediction task.

Table 5
Details of the training and testing networks for the link prediction models. Each row refers to one entity co-occurrence graph, split into a training network and a testing network with time slicing. *Density* (D) is defined as $\frac{2*Links}{Nodes*(Nodes-1)}$.

Nodes	Training network						Testing network					
	Feature network			Label network			Feature network			Label network		
	Years	Links	D (%)	Year	New links[a]	ΔD (%)	Years	Links	D (%)	Year	New links	ΔD (%)
607	1977–2000[b]	6 176	3.36	2001	4 309	+2.34	1977–2001	10 485	5.70	2002	4 238	+2.30
1767	1977–2001	25 483	1.63	2002	12 661	+0.81	1977–2002	38 144	2.44	2003	9 511	+0.61
2675	1977–2002	48 836	1.37	2003	11 868	+0.33	1977–2003	60 704	1.70	2004	12 060	+0.34
3299	1977–2003	67 376	1.24	2004	13 082	+0.24	1977–2004	80 458	1.48	2005	14 309	+0.26
3905	1977–2004	87 042	1.14	2005	15 214	+0.20	1977–2005	102 256	1.34	2006	14 169	+0.19
4550	1977–2005	109 268	1.06	2006	14 957	+0.14	1977–2006	124 225	1.20	2007	5 117	+0.05
5168	1977–2006	130 749	0.98	2007	5 289	+0.04	1977–2007	136 038	1.02	2008	15 584	+0.12
5329	1977–2007	137 565	0.97	2008	15 810	+0.11	1977–2008	153 375	1.08	2009	16 195	+0.11
5900	1977–2008	159 503	0.92	2009	16 979	+0.10	1977–2009	176 482	1.01	2010	14 741	+0.08
6525	1977–2009	183 514	0.86	2010	15 413	+0.07	1977–2010	198 927	0.93	2011	11 725	+0.06
7161	1977–2010	205 806	0.80	2011	12 206	+0.05	1977–2011	218 012	0.85	2012	18 473	+0.07
7634	1977–2011	222 832	0.76	2012	18 924	+0.06	1977–2012	241 756	0.83	2013	16 618	+0.06
8248	1977–2012	248 875	0.73	2013	17 129	+0.05	1977–2013	266 004	0.78	2014	18 627	+0.05
8828	1977–2013	272 435	0.70	2014	19 164	+0.05	1977–2014	291 599	0.75	2015	16 030	+0.04
9527	1977–2014	299 221	0.66	2015	16 720	+0.04	1977–2015	315 941	0.70	2016	14 356	+0.03
10 066	1977–2015	321 833	0.64	2016	14 744	+0.03	1977–2016	336 577	0.66	2017	11 675	+0.02
10 491	1977–2016	341 502	0.62	2017	11 884	+0.02	1977–2017	353 386	0.64	2018	14 228	+0.03
10 815	1977–2017	356 891	0.61	2018	14 355	+0.02	1977–2018	371 246	0.63	2019	8 012	+0.01
11 140	1977–2018	375 121	0.60	2019	8 045	+0.01	1977–2019	383 166	0.62	2020	7 740	+0.01
11 349	1977–2019	385 558	0.60	2020	7 820	+0.01	1977–2020	393 378	0.61	2021	430	<0.01

[a]'New links' refers to links that appear in the label network that were not in the feature network.
[b]Feature network starts from 1977–2000 because the data before 2000 is not enough to perform link prediction under the experimental setting in this study.

Table 6
Details of training and testing data for the link prediction model. Each row represents one pair of training and testing datasets constructed from one pair of training and testing networks in Table 5.

Years	Training data		Testing data	
	#Positive	#Negative	#Positive	#Negative
1977–2002	4 309	4 309	4 238	169 198
1977–2003	12 661	12 661	9 511	1 512 606
1977–2004	11 868	11 868	12 060	3 503 711
1977–2005	13 082	13 082	14 309	5 345 284
1977–2006	15 214	15 214	14 169	7 506 135
1977–2007	14 957	14 957	5 117	10 219 633
1977–2008	5 289	5 289	15 584	13 199 906
1977–2009	15 810	15 810	16 195	14 026 886
1977–2010	16 979	16 979	14 741	17 210 827
1977–2011	15 413	15 413	11 725	21 073 898
1977–2012	12 206	12 206	18 473	25 399 895
1977–2013	18 924	18 924	16 618	28 876 787
1977–2014	17 129	17 129	18 627	33 725 997
1977–2015	19 164	19 164	16 030	38 654 749
1977–2016	16 720	16 720	14 356	45 046 804
1977–2017	14 744	14 744	11 675	50 308 893
1977–2018	11 884	11 884	14 228	54 657 681
1977–2019	14 355	14 355	8 012	58 097 447
1977–2020	8 045	8 045	7 740	61 653 324
1977–2021[a]	7 820	7 820	430	64 034 468

[a]Year 2021 is not a full year (until 2021/6/10).

5. Discussion

5.1. Impact of reframing the prediction task

The experimental framework reflected in Table 7 considers only a single year in the future as the prediction space for the evaluation. However, since knowledge grows continuously, predicted links missing in the evaluated label network may appear in subsequent years. In other words, a False Positive prediction with respect to a single year may turn out to be a True Positive if more years in the future are considered. In Table 8, the column labeled *#FP→ future TP* shows the number of FP in each dataset that appears in the networks beyond the immediate following year, and column *Proportion* calculates the proportion of the potential links that are considered as FPs but appear in the more distant future. It is clear that the prediction space leads to underestimated evaluation results as 11 out of 20 datasets have more than 50% false positive cases that are not actually false positives, in that they become true as more knowledge emerges. As the future time horizon becomes shorter, the possibility of the knowledge appearing in that future decreases.

We therefore changed the prediction evaluation context from short-term to longer-term. The result of this change appears in Fig. 5, reported in terms of precision@10% (see Section 3.4.2) in the main manuscript. The AUPRC scores for link prediction in the longer-term context and corresponding training and testing data are in the supplementary material. Under this evaluation setting, we considered all links that emerge between the test year of the dataset and 2021 as positive links. Hence we considered the full knowledge that appears in the literature until 2021 as the evaluation horizon.

We observed a huge improvement in model performance, demonstrating that the models have made meaningful predictions yet are not able to be given credit for those predictions due to the experimental framework, and lack of complete knowledge in the label network. Comparing different LBD systems is challenging: although previous work has proposed a time-sliced evaluation methodology [20], we demonstrated the importance of considering the prediction window length at evaluation.

In addition to the prediction space quantified in this section, other key experimental settings specified in Section 3.3.3, including the negative sampling strategy in the testing dataset, the stop word lists, as well as the new nodes may also impact the evaluation results and the model interpretations. This needs to be empirically demonstrated. At least, it is important for other LBD researchers to make explicit experimental configurations on which the evaluation results and the model interpretations are conditioned.

5.2. Limitations of KG construction process

5.2.1. Semantic entity types

Some node types in the AD knowledge graph lead consistently to errors. The most obvious error comes from the $ADPrep$ node type. The $ADPrep$ node type accounts for 13% false positive links after changing the prediction window to be longer-term. The nodes with the $ADPrep$

Table 7
AUPRC scores for link prediction. The scores are scaled to 100. Each row refers to one dataset, which is used to train and test the link prediction model with graph embedding methods. Results highlighted in bold are the best in each category.

Dataset	Baseline	Matrix factorization-based		Random walk-based		Neural network-based				Mean
	Dummy	GraRep	HOPE	node2vec	DeepWalk	LINE	GCN	GraphSage	SDNE	
1977–2002	2.44	**10.02**	6.46	9.42	**14.56**	7.55	14.00	2.67	**16.05**	9.24
1977–2003	0.62	**4.58**	3.99	**4.37**	2.61	1.04	8.86	0.74	**9.33**	4.02
1977–2004	0.34	**3.82**	3.62	**2.94**	1.86	0.45	7.04	0.39	**7.54**	3.11
1977–2005	0.27	3.70	**3.74**	**3.16**	1.99	0.36	6.56	0.29	**7.02**	3.01
1977–2006	0.19	3.54	**4.05**	**2.87**	1.91	0.28	5.14	0.21	**5.39**	2.62
1977–2007	0.05	**0.95**	0.92	**0.88**	0.63	0.08	1.37	0.06	**1.48**	0.71
1977–2008	0.12	2.69	**2.78**	**2.54**	1.74	0.17	4.05	0.12	**4.17**	2.04
1977–2009	0.12	3.51	**3.73**	**3.04**	2.27	0.17	4.54	0.13	**5.23**	2.53
1977–2010	0.09	**2.52**	2.46	**1.93**	1.61	0.13	3.44	0.10	**3.72**	1.77
1977–2011	0.06	2.01	**2.08**	**1.51**	1.25	0.08	2.94	0.07	**3.18**	1.46
1977–2012	0.07	**2.56**	2.17	**2.24**	1.92	0.10	3.74	0.08	**3.87**	1.86
1977–2013	0.06	**2.09**	2.03	**1.97**	1.69	0.08	3.20	0.06	**3.65**	1.65
1977–2014	0.06	**2.30**	**2.30**	**1.72**	1.54	0.07	3.12	0.06	**3.49**	1.63
1977–2015	0.04	1.87	**1.98**	**1.62**	1.39	0.06	2.55	0.05	**2.95**	1.39
1977–2016	0.03	1.67	**1.83**	**1.51**	1.28	0.04	2.29	0.04	**2.67**	1.26
1977–2017	0.02	1.22	**1.37**	**1.05**	0.92	0.07	1.70	0.03	**1.92**	0.92
1977–2018	0.03	1.44	**1.49**	**1.29**	1.07	0.08	2.01	0.03	**2.36**	1.09
1977–2019	0.01	0.89	**0.92**	**0.74**	0.60	0.04	1.23	0.02	**1.45**	0.66
1977–2020	0.01	0.86	**1.07**	**0.71**	0.62	0.04	1.13	0.02	**1.28**	0.64
1977–2021	<0.01	0.06	**0.08**	**0.04**	**0.04**	<0.01	0.15	<0.01	**0.16**	0.06
Mean	0.23	2.62	2.45	2.28	2.08	0.55	3.95	0.26	4.35	2.08

Table 8
The error related to false positives, due to the limited prediction horizon considered.

Dataset	#FP	#FP→future TP	Proportion
1977–2002	276	258	93.48%
1977–2003	679	626	92.19%
1977–2004	865	776	89.71%
1977–2005	1025	870	84.88%
1977–2006	1281	552	43.09%
1977–2007	480	417	86.88%
1977–2008	1204	920	76.41%
1977–2009	1329	916	68.92%
1977–2010	1289	796	61.75%
1977–2011	981	627	63.91%
1977–2012	1515	854	56.37%
1977–2013	1291	750	58.09%
1977–2014	1594	588	36.89%
1977–2015	1345	512	38.07%
1977–2016	1275	330	25.88%
1977–2017	1057	202	19.11%
1977–2018	1279	170	13.29%
1977–2019	752	51	6.78%
1977–2020	710	3	0.42%
1977–2021	43	0	0.00%

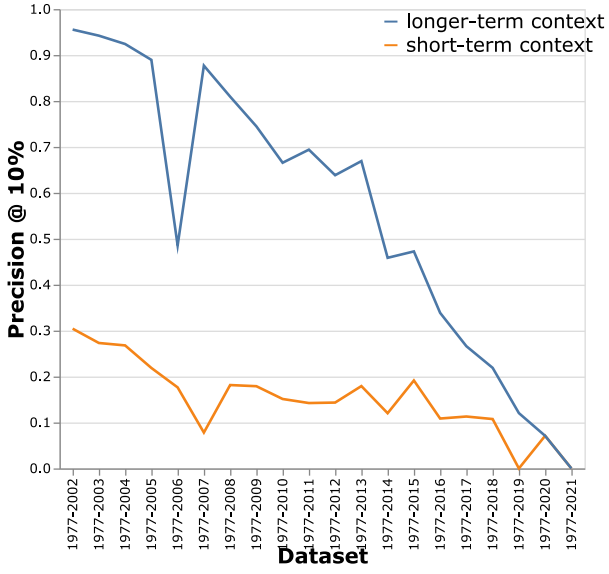


Fig. 5. Improvements of model performance after changing evaluation context from short term to longer term. Metric presented here is precision@10% (top 10%) of predictions. The orange line corresponds to the original evaluation setting considering a single next year to the training dataset, while the blue line considers the full future period from the end of the training set range until 2021. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

type include: minute (obo:UO_0000031), hour (obo:UO_0000032), second (obo:UO_0000010), week (obo:UO_0000034), source (dc:source), located in (obo:RO_0001025), is part of (obo:BFO_0000050), participate in (obo:RO_0000056), member of (obo:RO_0002350), and pH (obo:UO_0000196). It is understandable why the potential links with *ADPrep* as one of the node types are errors. The co-occurrence links with *ADPrep* nodes such as hour do not make sense to be a hypothesis independently of the concepts they modify. They serve as modifiers to make AD concepts more precise, e.g. by indicating that amyloid is amyloid located in a certain area of the brain. The nodes with the *ADPrep* types need to be combined with other nodes via a relation more specific than co-occurrence, which can be ignored for discovery.

Similarly, the AD-DTM semantic type needs further refinement. In this study, it is categorized as a genotype, yet a number of the concepts in AD-DTM relate to the diagnosis of Alzheimer's Disease, such as *Problems with memory or concentration*. These belong to the phenotype. Improved definitions of semantic entity types of Alzheimer's Disease will support more meaningful inferences from the knowledge graph.

We plan to collaborate with AD experts to refine the semantic entity types we considered, so that some may be entirely excluded from the graph. We further can experiment with ignoring certain link types for inference purposes.

5.2.2. Redundant entities

Combining multiple annotators leads to one limitation. Because we used two annotators – the PTC annotator and the NIO annotator – redundant entities may be introduced. For instance, for the

same text span *calcium*, the PTC annotator annotates it with the UID *MESH:D002118*, while the NIO annotator annotates it with the UID *AlzheimerOntology:calcium*. Both UIDs refer to the canonical *calcium*, yet they appear as two distinct nodes in the graph. Such redundancy also happens within the NIO annotator, due to its inclusion of multiple source vocabularies. For instance, both UIDs *AlzheimerOntology:pareital_lobe* and *obo:fma_61826* from the NIO annotator refer to the canonical *parietal lobe*. Without post-processing, the overlapping annotations lead to two distinct nodes being created in the network.

Redundant entities lead to inflated evaluation results. Overall, we observed 77 pairs of UIDs that are redundant across all 20 datasets. The complete list is in the Github repository.[6] We mapped the pairs of UIDs to the NIO UIDs if the ambiguity occurs across the annotators, and map the pairs of UIDs to the *AlzheimerOntology* if the ambiguity happens within the NIO annotator. The model performances improve very slightly (less than 1% for all 20 datasets) after removing redundant nodes, as compared to the original evaluation results.

5.2.3. Corpus collection

For this study, we excluded papers without titles or abstracts, and papers without PMIDs in the corpus collection process. For instance, more than 80% of the papers published prior to 1998 have no abstract. In 2007, around two-thirds of papers do not have abstracts. After excluding these papers, we also removed an additional 20% that could not be mapped to a PMID via Esearch.[7] However, we noticed a few limitations of Esearch: for example, presence of Greek letters can lead to the unsuccessful linking of titles to PMIDs. The exclusion of papers influences the knowledge graph construction process, leading to different entities and links in the graph, and different training and testing datasets. The exclusion of papers also influences the knowledge inference process, resulting in different predictions of genotype and phenotype relationships in Alzheimer's Disease. We aim to include these papers in future studies.

5.2.4. Annotator creation

The NIO annotator includes 4843 unique identifiers. The configuration parameters in ConceptMapper allow the annotator to catch the variation of the mentions, such as using different search strategies, excluding stopwords, and ignoring orders of words. However, the NIO annotator has three limitations.

First, the annotator cannot recognize biochemical notations that are typically realized with superscripts but represented in the ontology as a flat string. For instance, the UID *AlzheimerOntology:Psen2tm1Ber* refers to the canonical *Psen2tm1Ber*. "tm1Ber" is typically presented as a superscript that refers to a specifically targeted mutation of the Psen2 gene, e.g. "Psen2^{tm1Ber}". But since the annotator only recognizes the unbroken string "psen2tm1Ber", the correct annotation depends on how the paper presents the superscript. Second, the NIO annotator cannot catch all variations of concept names in the papers. For instance, the vocabulary of the NIO annotator includes "Abeta", but it cannot recognize "A beta", "A beta(40)", and "A beta(42)" in the text. Note that strategies for addressing this via text regularization have been proposed in prior work [63]. Third, the annotator has issues in lemmatizing abbreviations, leading to errors. For instance, the BioLemmatizer [64] incorporated into the annotator lemmatizes the abbreviation of "mass spectrometry" — "ms" — as "m". Then the annotator annotates the "M" in the author's name "Sun, C. M" as *AlzheimerOntology:mass_spectrometry*. Such limitations affect the quality of the Alzheimer's Disease knowledge graph and the link prediction model performance. We aim to reduce these errors in future studies.

[6] https://github.com/READ-BioMed/readbiomed-lbd/blob/main/00_data/metadata/entity_resolution.json.

[7] https://www.ncbi.nlm.nih.gov/books/NBK25499/.

5.2.5. Link representation

Links are included in the AD knowledge graph based on any co-occurrence of mentioned concepts, and represented as unweighted binary links (i.e., a link is present or absent). A more nuanced representation that weights links on the basis of aggregation of individual mentioned co-occurrences, such as the frequency of co-occurrence or a statistical aggregation method [17], may also be considered for the graph representation and for use in modeling. Another nuanced link representation is to consider different relations between two types of entities, which may not be generalized by co-occurrence. A better representation leads to a high-quality knowledge graph, which further helps generate more accurate link predictions.

Furthermore, while we have discarded the use of more specifically typed and/or directed links in our methods due to a lack of available relation extraction tools specific to the AD context, the ability to establish more focused links would result in more biologically meaningful analysis and predictions.

5.2.6. Knowledge evolution

Knowledge changes over time. New entities emerge, and new associations between them are established. Our current methods focus on predicting new associations, but with certain limitations.

When creating the training and testing data for knowledge inference, we excluded new nodes that occur in the label network but not in the feature network. This is largely because 7 out of 8 link prediction models used in this study are not designed for new nodes. However, inferring links that involve new nodes is a future direction that makes the LBD task more realistic.

Time is an important feature in predicting implicit knowledge hypotheses. In this study, we only considered the case when the knowledge is newly discovered each year or is repeatedly asserted. However, knowledge can also be negated or even disappear. The extent of this phenomenon for knowledge discovery in AD, and how to capture such knowledge evolution processes within the model design are potential directions for further research.

6. Conclusion

In this paper, we adapted LBD specifically for the Alzheimer's Disease context by incorporating AD-specific semantic resources and building an AD-focused knowledge graph. As the first study that applies link prediction at scale for LBD under the AD context, we inferred new knowledge through graph embedding-based models and comprehensively evaluated LBD on 20 time-sliced datasets. We found that the neural network-based method SDNE consistently outperforms other graph embedding models in returning the most confident set of AD knowledge discoveries over datasets spanning 20 years. This suggests the importance of capturing highly non-linear network structures in learning useful network representation for the link prediction downstream task.

We further examined the impact of different prediction window lengths on the time-sliced evaluation methodology. We observed that longer-term prediction context leads to hugely different results than a short-term prediction context does, highlighting the need for thoughtful consideration of evaluation strategies in LBD. Considering what is thought of as accurate link prediction leads to improvements in the LBD field. We also identified limitations of the LBD pipeline in corpus collection, annotator creation, redundant entities, semantic entity types, link representation, and knowledge evolution. Addressing these limitations will lead to more accurate and realistic AD research discoveries. These key learnings are not specific to Alzheimer's Disease, but can be generalized to other diseases.

As the most common form of dementia, Alzheimer's Disease continues to be a substantial public health burden. We hope that research support systems such as literature-based discovery will reduce the issue of information overload, as well as enable untangling of the intricacies of the disease to uncover the most confident implicit knowledge. Our work represents an important step in this direction.

CRediT authorship contribution statement

Yiyuan Pu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Daniel Beck:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. **Karin Verspoor:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yiyuan Pu reports financial support was provided by Australian Research Council. Karin Verspoor reports a relationship with BioGrid Australia Ltd that includes: board membership.

Acknowledgments

We thank Professor Colin Louis Masters for the generous sharing of the extensive bibliography of Alzheimer's Disease literature. We thank Dr. Antonio Jimeno Yepes for his assistance with adapting the ontology concept recognition tool to our context. We thank Professor Alistair Moffat for inspiring us to reconsider experimental configurations under the time-slicing setting.

Funding

This research was conducted by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030) and funded by the Australian Research Council. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104464.

References

[1] World Health Organization, URL https://www.who.int/en/news-room/fact-sheets/detail/dementia.

[2] Z. Longhe, 2020 Alzheimer's disease facts and figures, Alzheimer's Dement. 16 (2020) http://dx.doi.org/10.1002/alz.12068.

[3] K. Macklin, On the frontlines of the Alzheimer's crisis:, Del. J. Public Health 7 (2021) 20–23, http://dx.doi.org/10.32481/djph.2021.09.005.

[4] D.R. Swanson, Literature-based discovery? The very idea, in: P. Bruza, M. Weeber (Eds.), Literature-Based Discovery, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN: 978-3-540-68690-3, 2008, pp. 3–11, http://dx.doi.org/10.1007/978-3-540-68690-3_1.

[5] R. Kostoff, Literature-related discovery: Potential treatments and preventatives for SARS, Technol. Forecast. Soc. Change 78 (2011) 1164–1173, http://dx.doi.org/10.1016/j.techfore.2011.03.022.

[6] R. Kostoff, Literature-related discovery (LRD): Potential treatments for cataracts, Technol. Forecast. Soc. Change 75 (2008) 215–225, http://dx.doi.org/10.1016/j.techfore.2007.11.006.

[7] R. Kostoff, J. Block, J. Stump, D. Johnson, Literature-related discovery (LRD): Potential treatments for Raynaud's Phenomenon, Technol. Forecast. Soc. Change 75 (2008) 203–214, http://dx.doi.org/10.1016/j.techfore.2007.11.005.

[8] R. Kostoff, M. Briggs, Literature-related discovery (LRD): Potential treatments for Parkinson's Disease, Technol. Forecast. Soc. Change 75 (2008) 226–238, http://dx.doi.org/10.1016/j.techfore.2007.11.007.

[9] M. Tropmann-Frick, T. Schreier, Towards drug repurposing for COVID-19 treatment using literature-based discovery, Front. Artif. Intell. Appl. 343 (2022) http://dx.doi.org/10.3233/FAIA210488.

[10] A. Daowd, S. Abidi, S.S.R. Abidi, A knowledge graph completion method applied to literature-based discovery for predicting missing links targeting cancer drug repurposing, in: M. Michalowski, S.S.R. Abidi, S. Abidi (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham, ISBN: 978-3-031-09342-5, 2022, pp. 24–34.

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

[12] Q. Xie, K. Yang, G. Heo, M. Song, Literature based discovery of alternative TCM medicine for adverse reactions to depression drugs, BMC Bioinformatics 21 (2020) http://dx.doi.org/10.1186/s12859-020-03735-8.

[13] Y. Kim, S. Beak, A. Charidimou, M. Song, Discovering new genes in the pathways of common sporadic neurodegenerative diseases: A bioinformatics approach, J. Alzheimer's Dis. 51 (2016) http://dx.doi.org/10.3233/JAD-150769.

[14] Z. Dai, Q. Li, G. Yang, W. Yini, Y. Liu, Z. Zheng, Y. Tu, S. Yang, B. Yu, Using literature-based discovery to identify candidate genes for the interaction between myocardial infarction and depression, BMC Med. Genet. 20 (2019) http://dx.doi.org/10.1186/s12881-019-0841-8.

[15] J. Hur, K. Sullivan, Y. Hong, M. Pande, D. States, H. Jagadish, E. Feldman, Literature-based discovery of diabetes- and ROS-related targets, BMC Med. Genom. 3 (2010) 49, http://dx.doi.org/10.1186/1755-8794-3-49.

[16] D. Gubiani, E. Fabbretti, B. Cestnik, N. Lavrac, T. Urbančič, Outlier based literature exploration for cross-domain linking of Alzheimer's disease and gut microbiota, Expert Syst. Appl. 85 (2017) http://dx.doi.org/10.1016/j.eswa.2017.05.026.

[17] S. Pyysalo, S. Baker, I. Ali, S. Haselwimmer, T. Shah, A. Young, Y. Guo, J. Högberg, U. Stenius, M. Narita, A. Korhonen, LION LBD: A literature-based discovery system for cancer biology, Bioinformatics (ISSN: 1367-4803) 35 (9) (2018) 1553–1561, http://dx.doi.org/10.1093/bioinformatics/bty845.

[18] V. Gopalakrishnan, K. Jha, W. Jin, A. Zhang, A survey on literature based discovery approaches in biomedical domain, J. Biomed. Inform. 93 (2019) 103141, http://dx.doi.org/10.1016/j.jbi.2019.103141.

[19] M. Thilakaratne, K. Falkner, T. Atapattu, A systematic review on literature-based discovery, ACM Comput. Surv. (ISSN: 0360-0300) 52 (6) (2020) 1–34, http://dx.doi.org/10.1145/3365756, 1557-7341.

[20] M. Yetisgen-Yildiz, W. Pratt, A new evaluation methodology for literature-based discovery systems, J. Biomed. Inform. (ISSN: 1532-0464) 42 (4) (2009) 633–643, http://dx.doi.org/10.1016/j.jbi.2008.12.001, URL https://www.sciencedirect.com/science/article/pii/S1532046408001482.

[21] S. Henry, B. Mcinnes, Literature based discovery: Models, methods, and trends, J. Biomed. Inform. 74 (2017) http://dx.doi.org/10.1016/j.jbi.2017.08.011.

[22] G. Crichton, Y. Guo, S. Pyysalo, A.-L. Korhonen, Neural networks for link prediction in realistic biomedical graphs: A multi-dimensional evaluation of graph embedding-based approaches, BMC Bioinformatics 19 (2018) http://dx.doi.org/10.1186/s12859-018-2163-9.

[23] D.R. Swanson, Fish oil, Raynaud's Syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1986) 18–7.

[24] N.R. Smalheiser, D.R. Swanson, Indomethacin and Alzheimer's disease, Neurology (ISSN: 0028-3878) 46 (2) (1996) 583, http://dx.doi.org/10.1212/WNL.46.2.583, URL https://n.neurology.org/content/46/2/583.1.

[25] D.R. Swanson, Migraine and magnesium: Eleven neglected connections, Perspect. Biol. Med. 31 (4) (1987) 526–557, http://dx.doi.org/10.1353/pbm.1988.0009.

[26] T. Rindflesch, D. Hristovski, A. Kastrin, Link prediction on a network of co-occurring MeSH terms: Towards literature-based discovery, Methods Inf. Med. (ISSN: 0026-1270) 55 (04) (2016) 340–346, http://dx.doi.org/10.3414/me15-01-0108.

[27] G. Crichton, S. Baker, Y. Guo, A. Korhonen, Neural networks for open and closed Literature-based Discovery, PLoS One 15 (2020) e0232891, http://dx.doi.org/10.1371/journal.pone.0232891.

[28] S. Sang, X. Liu, X. Chen, D. Zhao, A scalable embedding based neural network method for discovering knowledge from biomedical literature, IEEE/ACM Trans. Comput. Biol. Bioinform. PP (2020) 1, http://dx.doi.org/10.1109/TCBB.2020.3003947.

[29] T. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text, J. Biomed. Inform. 36 (2004) 462–477, http://dx.doi.org/10.1016/j.jbi.2003.11.003.

[30] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S. Lin, W. Zhang, P. Zhang, H. Sun, Graph embedding on biomedical networks: Methods, applications, and evaluations, Bioinformatics 26 (4) (2019) 1241–1251.

[31] M. Ou, P. Cui, J. Pei, Z. Zhang, W. Zhu, Asymmetric transitivity preserving graph embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450342322, 2016, pp. 1105–1114, http://dx.doi.org/10.1145/2939672.2939751.

[32] S. Cao, W. Lu, Q. Xu, GraRep: Learning graph representations with global structural information, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450337946, 2015, pp. 891–900, http://dx.doi.org/10.1145/2806416.2806512.

[33] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings International Conference on Knowledge Discovery & Data Mining, Vol. 2016, 2016, pp. 855–864, http://dx.doi.org/10.1145/2939672.2939754.

[34] B. Perozzi, R. Al-Rfou, S. Skiena, DeepWalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450329569, 2014, pp. 701–710, http://dx.doi.org/10.1145/2623330.2623732.

[35] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013).

[36] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, LINE: large-scale information network embedding, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 1067–1077, http://dx.doi.org/10.1145/2736277.2741093.

[37] T. Kipf, M. Welling, Variational graph auto-encoders, in: Bayesian Deep Learning (NeurIPS Workshops), 2016, URL https://arxiv.org/abs/1611.07308v1.

[38] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, ICLR, 2017.

[39] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[40] D. Wang, P. Cui, W. Zhu, Structural deep network embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450342322, 2016, pp. 1225–1234, http://dx.doi.org/10.1145/2939672.2939753.

[41] A. Kastrin, T. Rindflesch, D. Hristovski, Link prediction on a network of co-occurring MeSH terms: Towards literature-based discovery, Methods Inf. Med. 55 (2016) http://dx.doi.org/10.3414/ME15-01-0108.

[42] Y. Yang, R. Lichtenwalter, N. Chawla, Evaluating link prediction methods, Knowl. Inf. Syst. 45 (2015) http://dx.doi.org/10.1007/s10115-014-0789-0.

[43] C.-H. Wei, A. Allot, R. Leaman, Z. Lu, PubTator Central: Automated concept annotation for biomedical full text articles, Nucleic Acids Res. (ISSN: 0305-1048) 47 (W1) (2019) W587–W593, http://dx.doi.org/10.1093/nar/gkz389.

[44] A. Gomez-Valades, R. Martinez-Tomas, M. Rincon, Integrative base ontology for the research analysis of Alzheimer's disease-related mild cognitive impairment, Front. Neuroinform. (ISSN: 1662-5196) 15 (2021) http://dx.doi.org/10.3389/fninf.2021.561691, URL https://www.frontiersin.org/articles/10.3389/fninf.2021.561691.

[45] C.-H. Wei, H.-Y. Kao, Z. lu, PubTator: A web-based text mining tool for assisting biocuration, Nucleic Acids Res. 41 (2013) http://dx.doi.org/10.1093/nar/gkt441.

[46] A. Malhotra, E. Younesi, M. Gündel, B. Müller, M.T. Heneka, M. Hofmann-Apitius, ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease, Alzheimer's Dement. (ISSN: 1552-5260) 10 (2) (2014) 238–246, http://dx.doi.org/10.1016/j.jalz.2013.02.009.

[47] C. Rosse, J. Mejino, The foundational model of anatomy ontology, in: Anatomy Ontologies for Bioinformatics: Principles and Practice, Vol. 6, ISBN: 978-1-84628-884-5, 2008, http://dx.doi.org/10.1007/978-1-84628-885-2_4.

[48] A. Kostovska, I. Tolovski, F.S. Maikore, L.N. Soldatova, P. Panov, Neurodegenerative disease data ontology, in: IFIP Working Conference on Database Semantics, 2019.

[49] A.P. Cox, M. Jensen, A. Ruttenberg, K. Szigeti, A.D. Diehl, Measuring cognitive functions: Hurdles in the development of the neuropsychological testing ontology, in: International Conference on Biomedical Ontology, ICBO, 2013.

[50] J. Hastings, W. Ceusters, M. Jensen, K. Mulligan, B. Smith, Representing mental functioning: Ontologies for mental health and disease, in: International Conference on Biomedical Ontology, ICBO, 2012.

[51] C.-H. Wei, H.-Y. Kao, Z. lu, GNormPlus: An integrative approach for tagging genes, gene families, and protein domains, BioMed Res. Int. 2015 (2015) 918710, http://dx.doi.org/10.1155/2015/918710.

[52] C.-H. Wei, L. Phan, J. Feltz, R. Maiti, T. Hefferon, Z. Lu, tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine, Bioinformatics 34 (2018) 80–87.

[53] R. Leaman, Z. Lu, TaggerOne: joint named entity recognition and normalization with Semi-Markov Models, Bioinformatics (ISSN: 1367-4803) 32 (18) (2016) 2839–2846, http://dx.doi.org/10.1093/bioinformatics/btw343.

[54] C.-H. Wei, H.-Y. Kao, Z. lu, SR4GN: A species recognition software tool for gene normalization, PLoS One 7 (2012) e38460, http://dx.doi.org/10.1371/journal.pone.0038460.

[55] M. Tanenblatt, A. Coden, I. Sominsky, The ConceptMapper approach to named entity recognition, in: 7th Language Resource and Evaluation Conference, 2010.

[56] A. Jimeno Yepes, K. Verspoor, Classifying literature mentions of biological pathogens as experimentally studied using natural language processing, J. Biomed. Semant. (ISSN: 2041-1480) (2022) http://dx.doi.org/10.1186/s13326-023-00282-y.

[57] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. Cohen, L. Hunter, K. Verspoor, Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters, BMC Bioinformatics (ISSN: 1471-2105) 15 (1) (2014) 59, http://dx.doi.org/10.1186/1471-2105-15-59, URL http://www.biomedcentral.com/1471-2105/15/59.

[58] Z. Chen, B. Peng, V.N. Ioannidis, M. Li, G. Karypis, X. Ning, A knowledge graph for clinical trials (CTKG), Sci. Rep. 12 (2022) 4724, http://dx.doi.org/10.1038/s41598-022-08454-z.

[59] G.J. de Bruin, C. Veenman, H. Herik, F. Takes, Supervised temporal link prediction in large-scale real-world networks, Soc. Netw. Anal. Min. 11 (2021) http://dx.doi.org/10.1007/s13278-021-00787-3.

[60] W. Pratt, M. Yetisgen-Yildiz, LitLinker: Capturing connections across the biomedical literature, in: Proceedings of the 2nd International Conference on Knowledge Capture, in: K-CAP '03, Association for Computing Machinery, New York, NY, USA, ISBN: 1581135831, 2003, pp. 105–112, http://dx.doi.org/10.1145/945645.945662.

[61] Y. Cen, Z. Hou, Y. Wang, Q. Chen, Y. Luo, Z. Yu, H. Zhang, X. Yao, A. Zeng, S. Guo, Y. Dong, Y. Yang, P. Zhang, G. Dai, Y. Wang, C. Zhou, H. Yang, J. Tang, CogDL: A toolkit for deep learning on graphs, in: Proceedings of the ACM Web Conference 2023 (WWW '23), 2023, http://dx.doi.org/10.1145/3543507.3583472.

[62] CSIRO's Data61, StellarGraph machine learning library, 2018, URL https://github.com/stellargraph/stellargraph.

[63] K. Verspoor, C. Roeder, H.L. Johnson, K.B. Cohen, W.A. Baumgartner, L.E. Hunter, Exploring species-based strategies for gene normalization, IEEE/ACM Trans. Comput. Biol. Bioinform. (ISSN: 1557-9964) 7 (2010) 462–471, http://dx.doi.org/10.1109/TCBB.2010.48.

[64] H. Liu, T. Christiansen, W.A. Baumgartner, K. Verspoor, BioLemmatizer: A lemmatization tool for morphological processing of biomedical text, J. Biomed. Semant. (ISSN: 2041-1480) 3 (2012) 3, http://dx.doi.org/10.1186/2041-1480-3-3.