Repurposing Non-pharmacological Interventions for Alzheimer's Diseases through Link Prediction on Biomedical Literature

Yongkang Xiao[1,#], Yu Hou[2,#], Huixue Zhou[1], Gayo Diallo[3], Marcelo Fiszman[4,5], Julian Wolfson[6], Halil Kilicoglu[7], You Chen[8], Chang Su[9], Hua Xu[10], William G. Mantyh[11], Rui Zhang[2,*]

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA
[2]Department of Surgery, University of Minnesota, Minneapolis, MN, USA
[3]INRIA SISTM, Team AHeaD - INSERM 1219 Bordeaux Population Health, University of Bordeaux, F-33000, France
[4]NITES - Núcleo de Inovação e Tecnologia Em Saúde, Pontifical Catholic University of Rio de Janeiro, Brazil
[5]Semedy Inc, Needham, Massachusetts, USA
[6]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA
[7]School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA
[8]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[9]Department of Health Service Administration and Policy, Temple University, Philadelphia, PA, USA
[10]Section of Biomedical Informatics and Data Science, Yale University, New Haven, Connecticut, USA
[11]Department of Neurology, University of Minnesota, Minneapolis, MN, USA

# Authors contributed equally

* Corresponding author:
Rui Zhang, PhD
University of Minnesota
zhan1386@umn.edu

Abstract

Recently, computational drug repurposing has emerged as a promising method for identifying new pharmaceutical interventions (PI) for Alzheimer's Disease (AD). Non-pharmaceutical interventions (NPI), such as Vitamin E and Music therapy, have great potential to improve cognitive function and slow the progression of AD, but have largely been unexplored. This study predicts novel NPIs for AD through link prediction on our developed biomedical knowledge graph. We constructed a comprehensive knowledge graph containing AD concepts and various potential interventions, called ADInt, by integrating a dietary supplement domain knowledge graph, SuppKG, with semantic relations from SemMedDB database. Four knowledge graph embedding models (TransE, RotatE, DistMult and ComplEX) and two graph convolutional network models (R-GCN and CompGCN) were compared to learn the representation of ADInt. R-GCN outperformed other models by evaluating on the time slice test set and the clinical trial test set and was used to generate the score tables of the link prediction task. Discovery patterns were applied to generate mechanism pathways for high scoring triples. Our ADInt had 162,213 nodes and 1,017,319 edges. The graph convolutional network model, R-GCN, performed best in both the Time Slicing test set (MR = 7.099, MRR = 0.5007, Hits@1 = 0.4112, Hits@3 = 0.5058, Hits@10 = 0.6804) and the Clinical Trials test set (MR = 1.731, MRR = 0.8582, Hits@1 = 0.7906, Hits@3 = 0.9033, Hits@10 = 0.9848). Among high scoring triples in the link prediction results, we found the plausible mechanism pathways of (Photodynamic therapy, PREVENTS, Alzheimer's Disease) and (Choerospondias axillaris, PREVENTS, Alzheimer's Disease) by discovery patterns and discussed them further. In conclusion, we presented a novel methodology to extend an existing knowledge graph and discover NPIs (dietary supplements (DS) and complementary and integrative health (CIH)) for AD. We used discovery patterns to find mechanisms for predicted triples to solve the poor interpretability of artificial neural networks. Our method can potentially be applied to other clinical problems, such as discovering drug adverse reactions and drug-drug interactions.

Keywords: Alzheimer’s disease, drug repurposing, non-pharmaceutical interventions, biomedical knowledge graph, link prediction, graph embeddings

Introduction

Alzheimer's disease (AD) and related dementias (ADRD) are chronic and multifactorial neurodegenerative disorders that affect cognition, behavior, functional ability and memory of affected individuals [1]. As of 2020, the worldwide prevalence of ADRD was approximately 50 million, and this number is expected to increase to 152 million by 2050, representing a significant and growing public health challenge [2]. The high prevalence of ADRD has significant economic, medical, and social consequences for society. In 2019, the global economic burden of ADRD was estimated to be $2.8 trillion, and this burden is projected to increase to $16.9 trillion by 2050 [3]. Despite significant advances in our understanding of the etiology and drug targets of AD/ADRD, effective prevention and treatment of these conditions remain elusive. Several medications, including lecanemab [4] and aducanumab [5], have been developed based on well-defined concepts and hypotheses about the etiology and drug targets of AD/ADRD. These medications are thought to reduce the pathological progression of the disease; however, their treatment effect is limited [6]. This suggests that our understanding of the pathogenesis of Alzheimer's disease is incomplete, and that novel unbiased approaches are needed to discover new therapies.

AD is a complex and multifactorial disorder that poses significant challenges to drug discovery research. Despite significant progress in this field, there remains an unmet need for effective treatments, prevention, or interventions to slow down the progression of AD [7]. Pharmacological interventions (PI) have demonstrated improvements in cognitive function, albeit with adverse side effects such as nausea, weight loss, leg cramps, and increased mortality risk [8,9]. On the other hand, non-pharmacological interventions (NPI) including sleep [10,11], diet [12], dietary supplements [13], aerobic exercise [14], aromatherapy [15], light therapy [16] and cognitive training [17] are widely used by healthcare consumers to enhance their well-being and manage diseases. Thus, NPIs represent a promising, versatile, and potentially cost-effective approach to improve outcomes and quality of life for patients with dementia [18]. Recent studies have demonstrated that certain NPIs may be protective against cognitive decline in individuals with positive biomarkers and cognitive impairment [19]. For example, aerobic exercise has been shown to benefit various aspects of cognition, including the stabilization of Mini-Mental State Examination (MMSE) scores, as well as improvements in attention, memory, and recognition [20,21]. Cognitive decline may also be attenuated by factors such as improved nutrition, appropriate dietary supplements, mental exercise, and social activities [22]. Notably, multimodal NPIs have shown promise in improving cognitive function [23,24]. However, a comprehensive understanding of the effects of NPI, as well as the potential synergistic effects of PI and NPI for AD/ADRD, remains lacking.

In recent years, the analysis of existing data on drugs and diseases has emerged as a promising approach for discovering new therapeutic potentials of existing drugs and identifying treatments for refractory diseases, a practice commonly referred to as drug repurposing [25]. Text mining is a popular data mining approach for drug repurposing due to the rapidly increasing volume of biomedical and pharmaceutical research literature. A vast number of semantic relations between biomedical entities can now be extracted from this literature. Knowledge graphs (KGs), which are heterogeneous networks, can be utilized to store, manage and represent these semantic relations. KGs can be tools used to model entities and their relationships, and the network structure of KGs can be leveraged to generate hypotheses by utilizing graph theory concepts and methods [25]. In biomedical knowledge graphs (BKGs), nodes signify biomedical entities, and edges represent the relationships between two entities [26]. BKGs can provide solutions to practical problems in the biomedical domain. For instance, the SuppKG, a Dietary Supplement domain knowledge graph, can identify interactions between drugs and dietary supplements through discovery patterns [27]. Link Prediction (LP) for knowledge graphs (also known as knowledge graph

completion) is the task of inferring missing or potential relations between entities in a knowledge graph [28].

The LP for Semantic MEDLINE Database (SemMedDB) [29] has been found to be effective for drug repurposing for COVID-19 [30]. To address the current lack of research exploring novel NPIs for AD, we first trained and evaluated various LP strategies (e.g., embedding-based, neural network based models). The best-performing model was further utilized to predict NPIs that may have the potential to prevent AD. The NPIs include natural products (e.g., dietary supplements (DS)) and complementary and integrative health (CIH), which are identified using self-contained information in SuppKG and a CIHLex we previously created [31], respectively. Subsequently, discovery patterns [32] are employed to generate mechanism pathways for candidates with high scores (i.e., high likelihood), and these pathways are evaluated by domain experts. Our contribution includes creating NPI resources and developing an innovative framework to predict NPIs that may potentially be repurposed for AD. To our best of knowledge, this is the first study to discover NPIs for AD. The developed framework can be applied to NPI discovery for other diseases.

Results

ADInt Statistics

The comprehensive AD Intervention knowledge graph (called ADInt) encompasses 162,213 entities across 113 UMLS semantic types, which after further identification include 25,604 Drugs, 16,474 Diseases, 46,060 Genes and Proteins, 2,525 DS, and 128 CIH. Furthermore, ADInt comprises 1,017,319 triples, capturing 15 distinct relation types such as INTERACTS_WITH, AFFECTS and TREATS. Detailed statistics can be found in Table 1.

Performance of LP models

Table 2 presents the performance obtained by various LP methods using the metrics Mean Rank (MR), Mean Reciprocal Rank (MRR), and Hits@k (k = 1, 3, and 10) [33]. A well-performing model should exhibit a low MR score and high MRR and Hit@k scores. The results demonstrate that the R-GCN model outperforms the other models in all metrics, followed by the TransE model. Notably, the CompGCN model performs the lowest performance across all metrics.

Additionally, Table 3 reports evaluation results of the trained models on the Clinical Trials dataset. The findings show that the R-GCN model performs the best across all metrics. In this case, some metrics of the RotatE model (Hits@3 = 0.6320, Hits@10=0.8107, MR=5.228) are better than TransE (Hits@3 = 0.6294, Hits@10=0.7621, MR=5.417). Collectively, from both evaluation results presented in Table 2 and Table 3, the R-GCN model exhibits the best performance, with the lowest MR and the highest MRR, Hits@1, Hits@3, and Hits@10 among the considered models. Thus, we used the R-GCN for further knowledge discovery of NPIs on AD prevention.

Embedding representation of knowledge graph

Subsequently, we utilized t-SNE (t-distributed stochastic neighbor embedding) [34] to obtain

two-dimensional projection of the learned node representations. t-SNE is a technique that reduces high-dimensional data to low-dimensional data while preserving the distribution properties of the original data. Moreover, it expresses the similarity between concepts through the proximity between nodes. As depicted in Figure 1, nodes with similar types tend to be grouped together, particularly the DS nodes.

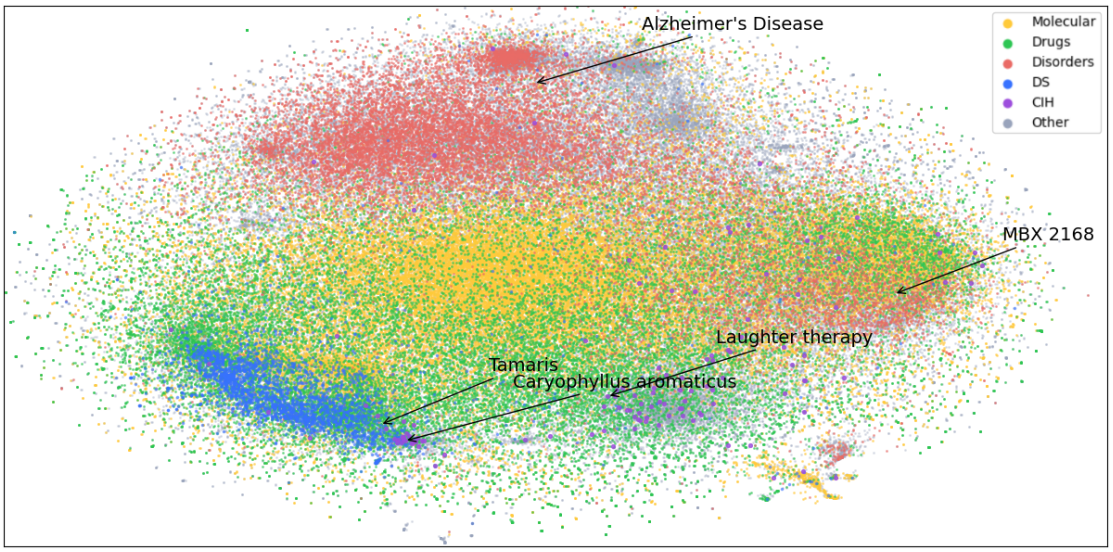

Figure 1: Visualization of nodes in ADInt dimensionally reduced by t-SNE algorithm and shown in a two-dimensional space. Different types of nodes are represented by different colors. Yellow: Molecular. Green: Drugs. Red: Disorders. Blue: DS (dietary supplement). Purple: CIH (complementary and integrative health). Gray: others.

Discovered NPI list for AD prevention

We utilized the embedding information obtained from R-GCN to compute the score of each candidate triple. Specifically, we designated the tail node of these corrupted triples as C0002395 (Alzheimer's Disease) and the edge as {PREVENTS, TREATS}. We then attempted to construct different triples by using all nodes in the graph as head nodes and calculated their score using the R-GCN model. Our focus was solely on the discovery of novel triples; thus, we excluded triples that already existed in ADInt. For novel triples, a higher score indicated a higher probability of being closely related to the true relationship. We categorized the triples into two groups based on the type of the head node, including DS and CIH, to discover novel NPIs for Alzheimer's disease. The top 10 predicted novel candidates for Alzheimer's disease are presented in Table 4

Figure 2 displays the network structure of the top-ranked predicted results. The network highlights three pathways that include a set of interesting findings, which will be further discussed in the following sections. Specifically, this pathway reveals potential mechanisms through which CIH and DS may influence the risk of AD, and suggests potential targets for therapeutic interventions. The identified associations and pathways represent a promising direction for future research into the prevention and treatment of AD.

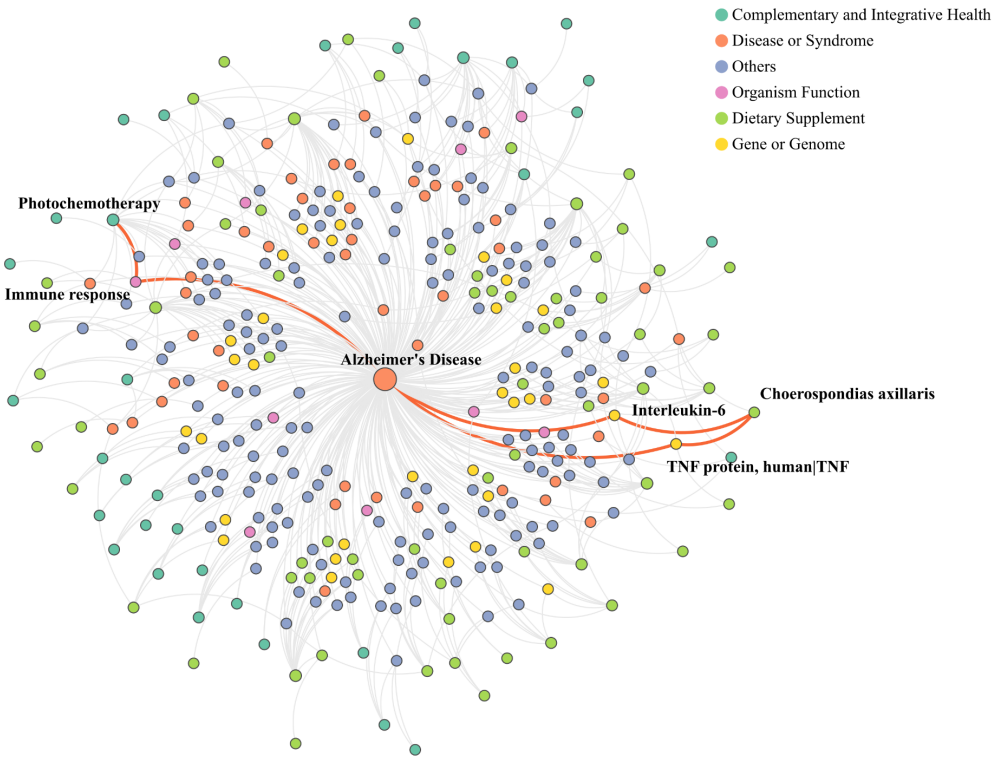

Figure 2: Top-ranked predicted results of ADInt-based exploration.

Discussion

In this study, we compared various LP methods on the task of knowledge discovery. The R-GCN model has demonstrated superior performance over other models on both the Time Slicing and Clinical Trials test sets (see Table 2 and Table 3). Notably, TransE exhibited the second-best overall performance, which is consistent with our prior work [30] demonstrating that relatively simple TransE outperformed other knowledge graph embedding methods (RotatE, DistMult, ComplEX) on the extended SemMedDB. We speculate that the poor performance of DistMult and ComplEX is due to their preference for high-degree entities, which we removed during the data preprocessing stage [35]. We believe that the reason for RotatE's underperformance is similar, as our filtered knowledge graph emphasizes simple relations. Although RotatE addresses some of the limitations of TransE in handling multiple and symmetric relations by introducing complex spaces [36], our findings suggest that this approach may not be appropriate for our knowledge graph. The superior performance of R-GCN suggests that the neighborhood aggregation operation of the graph convolution network is useful for learning graph representations [37]. However, we found that another graph convolutional network-based model, CompGCN, had a mediocre performance. We hypothesize that CompGCN's reliance on linear transformations for relation embeddings does not suit our knowledge graph [38]. Additionally, our evaluation of R-GCN on the Clinical Trials dataset, which mainly focuses on PREVENTS and TREATS relations, outperformed its performance on the Time Slicing evaluation. These results demonstrate that R-GCN is adept at distinguishing which subjects are feasible for treating or preventing AD. It is worth noting that while our experiments confirm R-GCN as the optimal LP model, metrics such as MR, MRR, and Hits@ratio only reflect the model's ability to predict interventions being trialed or known interventions. Indeed, models with low metrics may still produce valuable results [30]. Nevertheless, these metrics can inform model selection for NPI repurposing.

We used discovery patterns to generate mechanistic pathways for high-scoring triples predicted by the R-GCN model through the Neo4J platform. Photodynamic therapy (PDT) is a clinically used approach for treating various medical conditions, ranging from age-related macular degeneration to malignant tumors such as prostate cancer patients. PDT involves the use of light and a photosensitizing chemical substance along with molecular oxygen to elicit cell death [39]. Recently, PDT has been proposed as a potential therapeutic option for AD [39]. The precise mechanism of how PDT can provide therapeutic benefits for Alzheimer's disease remains elusive, and the practical use of PDT for treatment of AD is basically non-existent given that tissue must be directly exposed to light, which is not feasible when dealing with the entire brain. However, this finding provides theoretical support for treating AD through modulation of the immune system. For instance, a study evaluating the use of PDT with 5-aminolevulinic acid on mice has reported that it affects the immune response [40]. The study found that there was a significant reduction in the mRNA expression of interleukin-22 (IL-22), a cytokine produced by several immune cells that is associated with inflammation. Converging evidence has demonstrated that immune/inflammation response plays a crucial role in the initiation and regulation of Alzheimer's disease [41]. Thus, our PDT finding, while based on a therapy that has major practical limitations for treating AD, highlights immune mechanisms for preventing and treating AD. It should be noted that this is a preliminary finding based on a limited number of studies, and more research is needed to confirm these results.

Choerospondias axillaris, commonly known as Nepali hog plum, is a fruit that is approximately three centimeters long with sour flesh and yellow skin. Plums and other yellow-skinned fruits, such as papayas, tangerines, and oranges, are high in ß-cryptoxanthin, an antioxidant. A recent study [42] found an inverse association between serum β-cryptoxanthin levels and the incidence of Alzheimer’s Disease and all-cause dementias in individuals who consumed yellow-skinned fruits. Specifically, an increase of 8.6 micromole/liter in serum β-cryptoxanthin levels was associated with a 14% decreased risk of Alzheimer's disease. To propose a potential mechanism for this protection, we examined the patterns between Choerospondias axillaris and Alzheimer's disease. In a study [43], it was found that Choerospondias axillaris inhibits both TNF protein and interleukin-6. These two inflammation mediators are well-known inducers of Alzheimer's disease, as demonstrated in previous studies [44,45]. Specifically, interleukin-6 has been linked to the pathogenesis of Alzheimer's disease, while tumor necrosis factor-α has been proposed as a potent therapeutic target for Alzheimer's disease. Lutein, a carotenoid also found in Choerospondias axillas, we also found as a protective intervention. This finding corroborates prior reports that demonstrated an inverse association between lutein intake and dementia occurrence [45]. Furthermore, increased lutein intake has been associated with lower levels of AD neuropathology postmortem [46]. Overall, Choerospondias axillaris and other yellow-pigmented fruits may act as protectors by reducing the levels of pro-inflammatory cytokines crucially implicated in Alzheimer's disease.

There are several possibilities for future improvements to our approach. Firstly, we augmented SuppKG with triples extracted from the SemMedDB database, indicating that all triples in our ADInt were obtained through literature-based discovery. In order to further enhance our knowledge graph, we can merge it with other comprehensive biomedical databases and biological networks, such as DrugBank and KEGG [47]. This will enable us to expand the scope of our analysis and identify additional relevant interventions. Secondly, in addition to knowledge graph embedding and graph neural network models, other methods such as rule-based and reinforcement learning techniques have also demonstrated promising results on LP tasks. These methods could also be explored in future studies on drug repurposing. Lastly, since the determination of the plausibility of an intervention and its pathways to Alzheimer’s disease is a labor-intensive process, only the top 10 of each scoring

table were evaluated by experts. However, in future work, larger samples could be considered if the necessary resources are available.

Our analysis emphasizes the growing importance and popularity of studying NPIs in the context of disease management. By demonstrating the efficacy of our approach in revealing intricate relationships between biomedical entities, particularly NPI entities, and diseases of interest, we provide plausible mechanistic explanations for these associations. Notably, our contributions in this field include creating valuable NPI resources and developing an innovative framework to predict NPIs that may potentially be repurposed for AD. To the best of our knowledge, this is the first study that specifically aims to discover NPIs for AD. Furthermore, the versatility and adaptability of our approach enable its application to NPI discovery for a wide range of other diseases, including COVID-19. Our proposed approach also holds significant potential in addressing various clinical questions, such as the discovery of drug adverse reactions and drug-drug interactions, further emphasizing the importance and applicability of our research in the broader biomedical field.

Methods

The complete workflow is depicted in Figure 3. In order to investigate the association between PIs and NPIs and AD, we initially conducted preprocessing and integration of triples extracted from SemMedDB and SuppKG. Subsequently, we employed several graph representation models to derive the embedding information of ADInt, which included four knowledge graph embedding models (TransE [33], RotatE [36], DistMult [48] and ComplEX [49]) and two graph convolutional network models (R-GCN [50] and CompGCN [51]). Ultimately, we selected the most effective model for generating hypotheses regarding and NPIs for AD.

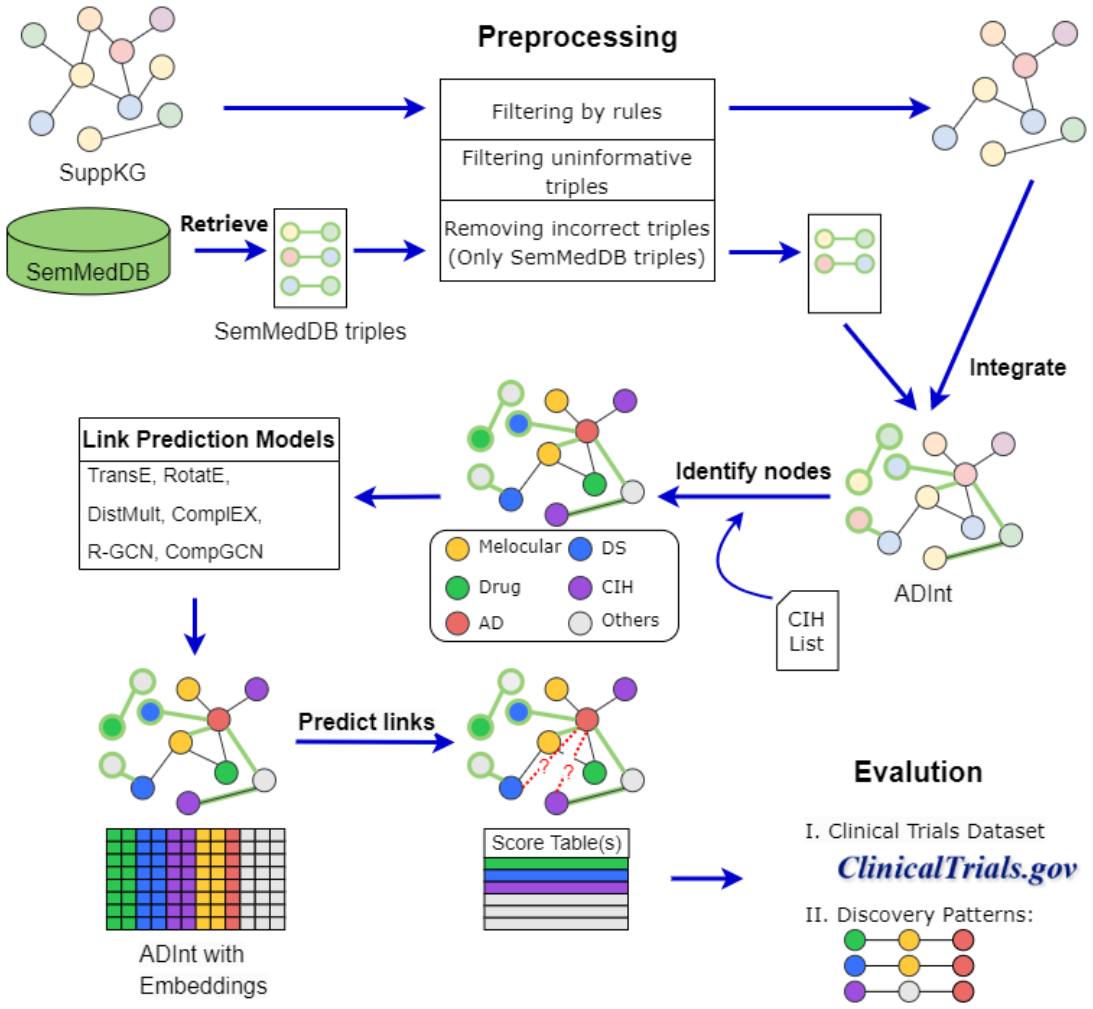

Figure 3: Diagram illustrating the workflow of the methodology.

Materials

SuppKG [27] is a knowledge graph that focuses on DS and has been shown to be useful in discovering potential drug-supplement interactions (DSI) through discovery patterns. In this study, we utilized SuppKG to explore DS for AD. To create SuppKG, the domain of the conceptual space of MetaMap [52] used in SemRep [53] was extended by incorporating dietary supplement terminologies and relations contained in the Integrated Dietary Supplement Knowledge base (iDISK) [54]. Subsequently, the extended SemRep was employed to extract semantic relations (triples) related to dietary supplements from PubMed abstracts retrieved using terms contained in iDISK. During the process of extracting semantic relations, some extracted semantic meanings were found to be opposite to the intended purpose of the corresponding text, resulting in the extraction of triples with opposite meanings. Due to the poor performance of SemRep (0.69 precision and 0.42 recall [53]), a fine-tuned PubMedBERT model [55], which is a pre-trained Bert model with abstracts from PubMed, was utilized to eliminate incorrect triples. SuppKG comprises 56,635 nodes and 595,222 directed edges, including 2,928 DS-specific nodes and 164,738 edges. The nodes in SuppKG are identified by unique UMLS CUIs, while the predicates in UMLS Semantic Network label the edges. To easily distinguish the DS-specific nodes, a letter "D" was added before the CUI representing the concept of dietary supplements. For example, "DC0633482" was used to indicate that "myrtol" is a dietary supplement concept.

SuppKG is a valuable resource for discovering potential drug-supplement interactions through pattern discovery, but due to its focus on the dietary supplement domain and the limitations of its source data, it may not include all pathway information related to AD or other CIH approaches beyond dietary supplements. To address this limitation, we extended SuppKG with additional triples extracted from the SemMedDB database [29]. SemMedDB is a repository of semantic triples extracted from PubMed abstracts and titles using the SemRep program [53,56], which provides detailed information such as the source text and PMID of the source article. We obtained triples from the PREDICTION table of SemMedDB and filtered them based on the correctness of the source sentences and text, which we extracted from the SENTENCE and PREDICATION_AUX tables. This allowed us to supplement SuppKG with a broader range of information related to interventions for AD beyond the dietary supplement domain.

Preprocessing and Integration

In this study, an enhanced representation of nodes and relations in the knowledge graph is proposed by filtering out low-quality triples. Low-quality triples often describe generic facts that are not meaningful for the study. For instance, the triple (disease, AFFECTS, patient) with "disease" and "patient" being generic concepts is not useful in this study. Additionally, since the triples in SuppKG and entries in SemMedDB database are extracted from text using the SemRep text mining tool, some of the semantic relations expressed by the triples may not align with the intended meaning of the source text. Therefore, preprocessing is necessary to integrate the information from SuppKG and SemMedDB. The preprocessing steps can be classified into three categories [30]:

1) **Filtering triples by rules.** First, we removed nodes in the graph that represented generic concepts, which was done by referencing the GENERIC_CONCEPT table provided by the SemMedDB database. This table contained concepts such as "Disease" and "Cells," which are known to be too broad to be useful for knowledge discovery. Additionally, concepts with semantic groups that were not likely to be useful for predicting interventions for ADRD were eliminated, including "Activities & Behaviors," "Concepts & Ideas," "Objects," "Occupations," "Organizations," and "Phenomena." Finally, only edges that were deemed relevant for LP were kept, specifically those with predicate types of AFFECTS, ASSOCIATED WITH, AUGMENTS, CAUSES, COEXISTS WITH, COMPLICATES, DISRUPTS, INHIBITS, INTERACTS WITH, MANIFESTATION OF, PREDISPOSES, PREVENTS, PRODUCES, STIMULATES, and TREATS.

2) **Removing high-degree concepts and uninformative semantic relations.** High-degree High-degree concepts in the KG may be too general to be useful for knowledge discovery due to their broad associations with many other concepts. To address this issue, we first computed the out-degree (k_i^{in}) and in-degree (k_i^{out}) of each node in the KG. Next, we calculated a log likelihood measure known as G^2 [57] for each triple, which quantifies the strength of the relationship between the items in the triple. The G^2 formula is given by:

$$G^2 = 2\sum_{i,j,k} n_{ijk} \times log(\frac{n_{ijk}}{m_{ijk}}),\ m_{ijk} = \frac{\sum_i n_{jk} \times \sum_j n_{ik} \times \sum_k n_{ij}}{T^2}$$

where n_{ijk} is the item *i,j,k* in the observation table (OT) containing observed frequencies of a triple, m_{ijk} is the item *i,j,k* in the expectation table (ET) describing the expected values assuming independence of terms in triples, and $T = \sum n_{ijk}$. Finally, we normalized k_i^{in}, k_i^{out}

and G^2 and summed them up together to get a final score for each triple. A higher score indicates that the triple is less specific and informative. For instance, the triple (Pharmaceutical Preparations, AFFECTS, Sleep) has a higher score and is more general compared to (DZIP1 gene, AFFECTS, heart valve development). Consequently, we filtered out some triples with high scores to manage the size of the knowledge graph for computational efficiency.

3) **Further removing incorrect triples by a trained PubMedBert model.** The triples extracted from the SemMedDB database through SemRep may contain false positives, as the semantics expressed by the triples may differ or be contrary to the content of their source sentences. To address this issue, we utilized a PubMedBert binary classification model that was fine-tuned in our previous work to evaluate the correctness of the triples by referencing their source sentence [30]. The F1 score of this model was 0.854, with a recall of 0.895 and a precision of 0.816.

For both SemMedDB and SuppKG triples, we applied steps 1) and 2) described above, but only applied step 3) to SemMedDB triples, as similar processing had been done during the generation of SuppKG. After filtering, we integrated the resulting triples from both sources, with DS concept nodes in SemMedDB triples identified by adding the letter D before their CUIs to match the identifiers in SuppKG. As the subject and object entities of the integrated triples are identified by UMLS CUIs and their predicates come from the UMLS Semantic Network, we added new triples to SuppKG that did not overlap with its existing triples, without mapping concepts or integrating ontologies. The resulting integrated knowledge graph, named ADInt, was obtained.

NPI nodes identification

We employed multiple approaches to identify nodes representing drugs, DS, and CIH concepts in our knowledge graph for analysis and repurposing efforts for AD. For drug nodes, we can directly utilize the semantic types provided in the UMLS Metathesaurus. Specifically, we identify a node as a drug node if its semantic type is Pharmacologic Substance (phsu) or Organic Chemical (orch). However, identifying DS concept nodes based on their semantic type is not feasible. Nonetheless, in SuppKG, DS concept nodes are denoted by a special mark, a letter D added before their CUI. This mark was retained during the integration of SuppKG and SemMedDB triples, allowing us to easily identify these nodes as DS concepts. Unlike drug and DS nodes, nodes describing CIH concepts cannot be identified directly from the knowledge graph. To overcome this limitation, we utilized an external list of CIH concepts provided by a graduate student of informatics with a background in Medicine. This list, known as the CAM concepts list or CIHLex, was compiled based on a review of the literature, as described in our previous work [58] and Natural Medicines [59].

Given that DS and CIH nodes may have semantic types of phsu or orch, which are also associated with drug concepts, it is possible for overlap to occur. To address this issue, we prioritize the identification of DS or CIH concepts over drug concepts. If a node has been identified as either DS or CIH, it is considered as such, regardless of its semantic type of phsu or orch. This approach ensures that there is no ambiguity in the identification of nodes within the knowledge graph.

Link prediction models training

A knowledge graph can be represented as a labeled directed multi-graph $KG = (E, R, G)$,

where *E* denotes the set of nodes representing entities, *R* denotes the set of edges representing relations, and $G \subseteq E \times R \times E$ is a set of triples$\langle h, r, t \rangle$, where h represents the head entity, *r* represents the relation, and *t* represents the tail entity. Link prediction (LP) is an essential task in knowledge graph completion, which aims to infer missing facts or relationships from the existing ones. Despite the vast amounts of information contained in knowledge graphs, they are often incomplete due to various factors, such as noise, missing data, and sparsity. Thus, LP methods seek to infer new triples that may not be explicitly represented in the knowledge graph, but which can be logically deduced from the existing ones. The objective of LP aims to predict the most probable entity or relation that completes (*h*, *r*, *?*) (tail prediction), (*h*, *?*, *t*) (edge prediction), or (*?*, *r*, *t*) (head prediction). Although new triples (*h'*, *r'*, *t'*) that describe additional facts may also exist in our knowledge graph, they are not present for some reason. LP for knowledge graphs can be represented as a ranking task, which aims to learn a prediction function that assigns higher scores to true triples and lower scores to false triples. To perform LP on our knowledge graph, we explored four knowledge graph embedding models (TransE, Rotate, DisMult and ComplEX) and two graph convolutional network models (R-GCN and CompGCN).

TransE [33] is a simple and effective model for LP, particularly for modeling one-to-one relations. In TransE, a triple (h, r, t) is represented as a translation from the embedding of the head entity h to the embedding of the tail entity t, with the relation r acting as the translation vector in the embedding space. This formulation implies that if a triple (h, r, t) exists, the embedding of entity h plus the representation of relation r should be close to the embedding of entity t. The TransE score function measures the plausibility of a triple and is defined as follows

$$s(h, r, t) = ||\boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t}||$$

where $\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t} \in \mathbb{R}^d$ is the embedding of *h*, *r* and *t*. Unlike TransE, The RotatE [36] model converts each relation to a rotation from a head entity to a tail entity in a complex vector space and the score function can be defined as

$$s(h, r, t) = ||\boldsymbol{h} \circ \boldsymbol{r} - \boldsymbol{t}||$$

where ○ is a Hadamard product.

DistMult [48] is the most basic semantic matching models, and its scoring function can be defined as

$$s(h, r, t) = \boldsymbol{h}^T \boldsymbol{r} \boldsymbol{t}$$

The drawback of DistMult is that it only works on symmetric relations, that is, the scores of (*h*,*r*,*t*) and (*t*,*r*,*h*) calculated by DistMult are the same. It may cause problems in our knowledge graph, for example the triple (Bariatric Surgery, TREATS, Alzheimer's) and the triple (Alzheimer's, TREATS, Bariatric Surgery) should have inconsistent scores. To address this limitation, ComplEX has been proposed as an extension of DistMult [49]. ComplEX uses a complex vector space and is capable of modeling asymmetric relations. Specifically, head and tail embeddings of the same entity are represented as complex conjugates, which enables (*h*, *r*, *t*) and (*t*, *r*, *h*) to be distinguished. This allows ComplEX to provide consistent scores for both symmetric and asymmetric relations. The scoring function of ComplEX can be defined as follows

$$s(h, r, t) = Re(\boldsymbol{h}^T \boldsymbol{r} \boldsymbol{t})$$

Where *Re*(·) is a real part of a complex vector.

GCNs are a neural network approach for processing graph-structured data [60]. However, most existing GCNs are designed for simple undirected graphs and cannot handle the multiple types of nodes and directed links that exist in our knowledge graph. To address this challenge, we explored special graph convolutional neural network models that can handle heterogeneous graphs. Specifically, we evaluated two models: Relational Graph Convolutional Network (R-GCN) [50] and CompGCN [51]. Based on the architectures of GCNs, R-GCNs [50] consider each different relation and perform feature fusion to participate in

updating the hidden states of nodes. The propagation model for calculating the forward-pass update of a node in R-GCNs can be defined as

$$\boldsymbol{x}_i^{(l+1)} = \sigma\Big(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^{(l)}\boldsymbol{x}_j^{(l)} + \boldsymbol{W}_0^{(l)}\boldsymbol{x}_i^{(l)}\Big),$$

where $\boldsymbol{x}_i^{(l)}\in\mathbb{R}^{d^{(l)}}$ is the hidden state of *i*-th nodes in the *l*-th layer of the neural network; $\mathcal{R}$ is the set of relations and $\mathcal{N}_i^r$ denotes the neighbor set of *i*-th node under relation $r\in\mathcal{R}$; $\boldsymbol{W}_r^{(l)}$ and $\boldsymbol{W}_0^{(l)}$ are the learnable weight matrix under relation r and self-loop weight matrix in the *l*-th layer respectively; $c_{i,r}$ is a normalization constant that can either be learned or chosen in advance. Using R-GCNs for LP tasks can be regarded as a process of encoding and decoding: an R-GCN producing latent feature vectors of entities and a tensor factorization model exploiting these vectors to predict edges. Taking the DistMult decomposition as an example, the score of a triple (*h*, *r*, *t*) is calculated as [50]

$$s(h,r,t) = \boldsymbol{h}^T\boldsymbol{r}\boldsymbol{t}$$

Thus, to make the model score observable triples higher than negative triples, the loss function can be defined as [50]:

$$\mathcal{L} = -\frac{1}{(1+\omega)|\hat{\varepsilon}|}\sum_{(h,r,t,y)\in\mathcal{T}} ylogl(s(h,r,t)) + (1-y)log(1-l(s(h,r,t))),$$

where $\mathcal{T}$ is the set of all triples (including positive and negative triples); ω is the number of negative triples; $|\hat{\varepsilon}|$ is the number of edges; $l(.)$ is the logistic sigmoid function; and y is an indicator, where $y = 1$ means triple is positive, otherwise negative.

CompGCN [51] is another extended version of GCN for heterogeneous graphs, which systematically leverages entity-relation composition operations and jointly learning latent feature vector representations for both nodes and edges in the graph. Different from R-GCNs, CompGCN performs a composition operation Φ over each edge in the neighbor of central node through the embedding of edges and nodes. The update equation of nodes embedding in CompGCN can be defined as

$$\boldsymbol{x}_i^{(l+1)} = f\Big(\sum_{(j,k)\in\mathcal{N}_i^r}\boldsymbol{W}_{\lambda(k)}^{(l)}\phi(\boldsymbol{x}_j^{(l)},\boldsymbol{y}_k^{(l)})\Big),$$

where $\boldsymbol{x}_j^{(l)}$and $\boldsymbol{y}_k^{(l)}$ are the hidden state of neighboring *j*-th node and its *k*-th relation respectively in the *l*-th layer, and $\boldsymbol{W}_{\lambda(k)}^{(l)}$ is a relation-type specific parameter, which can be used for direction specific weights. According to whether the edge is the original edge, inverse edge or self-loop edge, $\boldsymbol{W}_{\lambda(k)}^{(l)}$ will correspond to different weight matrices. Φ(.) is used to aggregate two vectors of the same size, which can be Subtraction [33], Multiplication [48], or Circular-correlation [61]. After updating the node embeddings, we can also update the relation embedding as follows [51]

$$\boldsymbol{y}_k^{(l+1)} = \boldsymbol{W}_{rel}^k\boldsymbol{y}_k^{(l)},$$

where $\boldsymbol{W}_{rel}^k$ is a weight matrix that projects all relations to the same embedding space as nodes, which allows them to be used in the next layer. Similar to R-GCNs LP model, we select a tensor factorization model (convE) to calculate the score of triples. And the same standard binary cross entropy loss function is applied to training the convolutional networks.

All work was conducted using Python scripts. The implementation of the TransE, RotatE, DistMult, and ComplEX models was carried out with the DGL-KE 0.1.0.dev0 package [62]

package. Both R-GCN and CompGCN models were constructed using the torch 1.13.1 [63] and DGL 1.0.1 [64] packages.

Evaluations

Open LBD task: The open discovery approach is specifically aimed at generating innovative hypotheses; given a head node, the system produces associated tail nodes, thereby facilitating the identification of previously unexplored triple relationships. [65]. In order to evaluate the effectiveness of our LP model, we utilized two evaluation methods. The first one is called Time Slicing [66]. This evaluation approach involves partitioning the KG at a specific time and using the data prior to this time to train the model, and subsequently testing the model on the data following this time to determine if the links formed after the partitioning time can be accurately predicted. Specifically, in our work, we ordered the triples chronologically and divided the knowledge graph into training, validation, and testing sets in an 8:1:1 ratio, with earlier triples used for training and more recent ones for testing, where the date of publication of the paper mentioning the triple is used as its time, and the partitioning times were set as April 2020 and April 2021 respectively. To evaluate the model performance, we computed three metrics for each model: MR, MRR, and Hits@k (k = 1, 3, and 10). Specifically, for each true triple in the testing set, we generated a batch of negative samples by randomly replacing the head or tail nodes while ensuring that these negative samples do not exist in our graph, i.e., we employed corruption with filtering. We then used the trained model to calculate the scores for the true triple and its negative samples, and obtained the ranks of the true triples to obtain the metrics of MR, MRR, and Hits@k. MR represents the average rank assigned to the true relations in the test set, MRR is the average inverse rank of all true triples in the test set, and Hits@k is the percentage of relations in which the true triple appears in the top k ranked triples [33].

In the second evaluation approach, we utilized clinical trial data from ClinicalTrials.gov as a benchmark for predicting potential interventions for Alzheimer's disease. Our approach was based on the assumption that interventions under investigation for AD have the potential to be repurposed for other indications. Specifically, we obtained a list of interventions utilized in AD clinical trials registered after April 21, 2020, by conducting a search for the term "Alzheimer" and restricting the results to interventional studies as of November 4, 2022. We excluded control interventions labeled as "placebo," resulting in a total of 671 interventions. We processed these interventions using MetaMap with the UMLS 2022AA release to identify relevant UMLS concepts, resulting in 1606 concepts. These concepts were subsequently used as head nodes, with "TREATS" and "PREVENTS" serving as the relationships, and "Alzheimer's disease" concepts as tail nodes, creating a series of new triples. Finally, we employed these newly generated triples based on clinical trial data as a test set to calculate MR, MRR, and Hits@k for each model.

Close LBD task evaluation: The closed discovery method strives to identify the connections between the given head and tail nodes in order to evaluate a specific hypothesis [65]. Although the knowledge graph embedding and graph neural network models only provide node and edge representations, patterns from closed discovery were used to infer possible mechanisms for the repurposed interventions. To uncover potential logical connections between concepts in a network, we employed a closed discovery approach by combining sequences of relation types, such as "drug x INHIBITS substance y, substance y CAUSES disease z" [32]. This method was used to identify possible pathways between nodes in the knowledge graph. For DS, The discovery patterns we focused on were:

InterventionA-INHIBITS|INTERACTS_WITH-**ConceptB** AND **ConceptB**-AFFECTS|CAUSES|PREDISPOSES|ASSOCIATED-**Alzheimer's disease** AND NOT (**InterventionA**-TREATS|PREVENTS-**Alzheimer's disease**)

where InterventionA is a node whose type is DS; ConceptB can be any concept; | indicates logical OR; and for Alzheimer's disease, we focus on the node with identifier C0002395. To analyze the repurposing potential of Complementary and Integrative Health (CIH) interventions, we encountered a challenge due to the UMLS semantic types of most CIHs being “topp” (Therapeutic or Preventive Procedure) or “dora” (Daily or Recreational Activity). As these types do not have INHIBIT or INTERACT_WITH relationships to other concepts in the UMLS Semantic Network, and the number of possible paths is not extensive, we did not constrain the predicates in the patterns. The discovery patterns for CIH were:

```
InterventionB—(any predicate)-ConceptB AND
ConceptB-(any predicate)-Alzheimer's disease AND
NOT (InterventionB-TREATS|PREVENTS-Alzheimer's disease
```

where InterventionB is a node whose type is CIH. We visualized the network structure using ChiPlot (https://www.chiplot.online/).

Data Availability

ADInt knowledge graph data is available in the following google drive: https://drive.google.com/drive/folders/187HnI2d-RRFeYk_C7MYSCHtVX-6luNVS?usp=sharing. The complete SemMedDB database can be accessed directly on https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR.html.

Code Availability

The code used for data preprocessing, model training, result evaluation and visualization in this study is available in the following repositories: https://github.com/YKXia0/LBD_AD.

Funding

Research reported in this publication was supported by the National Institutes of Health (NIH)/National Institute On Aging (NIA) under Award Number R01AG078154 (PI: RZ) and the NIH/National Center For Complementary & Integrative Health (NCCIH) under Award Number R01AT00945 (PI: RZ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. GD has received financial support for his mobility to UoM from the French State in the frame of the “Investments for the future” Programme IdEx Bordeaux, reference ANR-10-IDEX-03-02.

Acknowledgement

We would like to acknowledge the AWS Public Sector Cloud Credit for Research program to partially support this research.

Reference

1. Srivastava, S., Ahmad, R. & Khare, S. K. Alzheimer's disease and its treatment by different approaches: A review. *Eur. J. Med. Chem.* **216**, 113320 (2021).
2. Hampel, H. *et al.* Designing the next-generation clinical care pathway for Alzheimer's disease. *Nat. Aging* **2**, 692–703 (2022).
3. Nandi, A. *et al.* Global and regional projections of the economic burden of Alzheimer's disease and related dementias from 2019 to 2050: A value of statistical life approach. *EClinicalMedicine* **51**, 101580 (2022).
4. Swanson, C. J. *et al.* A randomized, double-blind, phase 2b proof-of-concept clinical trial in early Alzheimer's disease with lecanemab, an anti-Aβ protofibril antibody. *Alzheimers Res. Ther.* **13**, 1–14 (2021).
5. Selkoe, D. J. Alzheimer disease and aducanumab: adjusting our approach. *Nat. Rev. Neurol.* **15**, 365–366 (2019).
6. Scales, K., Zimmerman, S. & Miller, S. J. Evidence-based nonpharmacological practices to address behavioral and psychological symptoms of dementia. *The Gerontologist* **58**, S88–S102 (2018).
7. Hebert, L. E., Weuve, J., Scherr, P. A. & Evans, D. A. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology* **80**, 1778–1783 (2013).
8. Tariot, P. N. *et al.* A randomized, double-blind, placebo-controlled study of the efficacy and safety of donepezil in patients with Alzheimer's disease in the nursing home setting. *J. Am. Geriatr. Soc.* **49**, 1590–1599 (2001).
9. Loy, C. & Schneider, L. Galantamine for Alzheimer's disease and mild cognitive

impairment. *Cochrane Database Syst. Rev.* (2006).

10. Shi, L. *et al.* Sleep disturbances increase the risk of dementia: a systematic review and meta-analysis. *Sleep Med. Rev.* **40**, 4–16 (2018).
11. Borges, C. R., Poyares, D., Piovezan, R., Nitrini, R. & Brucki, S. Doença de Alzheimer e distúrbios do sono: uma revisão. *Arq. Neuropsiquiatr.* **77**, 815–824 (2019).
12. Liyanage, S. I., Vilekar, P. & Weaver, D. F. Nutrients in Alzheimer's disease: the interaction of diet, drugs and disease. *Can. J. Neurol. Sci.* **46**, 23–34 (2019).
13. Olivera-Pueyo, J. & Pelegrín-Valero, C. Dietary supplements for cognitive impairment. *Actas Esp. Psiquiatr.* **45**, (2017).
14. Cui, M. Y., Lin, Y., Sheng, J. Y., Zhang, X. & Cui, R. J. Exercise intervention associated with cognitive improvement in Alzheimer's disease. *Neural Plast.* (2018).
15. Jimbo, D., Kimura, Y., Taniguchi, M., Inoue, M. & Urakami, K. Effect of aromatherapy on patients with Alzheimer's disease. *Psychogeriatrics* **9**, 173–179 (2009).
16. Hanford, N. & Figueiro, M. Light therapy and Alzheimer's disease and related dementia: past, present, and future. *J. Alzheimers Dis.* **33**, 913–922 (2013).
17. Giovagnoli, A. R. *et al.* Cognitive training in Alzheimer's disease: a controlled randomized study. *Neurol. Sci.* **38**, 1485–1493 (2017).
18. Olazarán, J. *et al.* Nonpharmacological therapies in Alzheimer's disease: a systematic review of efficacy. *Dement. Geriatr. Cogn. Disord.* **30**, 161–178 (2010).
19. Andrieu, S., Coley, N., Lovestone, S., Aisen, P. S. & Vellas, B. Prevention of sporadic Alzheimer's disease: lessons learned from clinical trials and future directions. *Lancet Neurol.* **14**, 926–944 (2015).
20. Lee, J. The relationship between physical activity and dementia: a systematic review and meta-analysis of prospective cohort studies. *J. Gerontol. Nurs.* **44**, 22–29 (2018).
21. Groot, C. *et al.* The effect of physical activity on cognitive function in patients with dementia: a meta-analysis of randomized control trials. *Ageing Res. Rev.* **25**, 13–23 (2016).
22. Miquel, S. *et al.* Poor cognitive ageing: Vulnerabilities, mechanisms and the impact of

nutritional interventions. *Ageing Res. Rev.* **42**, 40–55 (2018).

23. Yorozuya, K., Kubo, Y., Tomiyama, N., Yamane, S. & Hanaoka, H. A systematic review of multimodal non-pharmacological interventions for cognitive function in older people with dementia in nursing homes. *Dement. Geriatr. Cogn. Disord.* **48**, 1–16 (2019).
24. Chalfont, G., Milligan, C. & Simpson, J. A mixed methods systematic review of multimodal non-pharmacological interventions to improve cognition for people with dementia. *Dementia* **19**, 1086–1130 (2020).
25. Jarada, T. N., Rokne, J. G. & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J. Cheminformatics* **12**, 1–23 (2020).
26. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **18**, 1414–1428 (2020).
27. Schutte, D. *et al.* Discovering novel drug-supplement interactions using SuppKG generated from the biomedical literature. *J. Biomed. Inform.* **131**, 104120 (2022).
28. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Phys. Stat. Mech. Its Appl.* **390**, 1150–1170 (2011).
29. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G. & Rindflesch, T. C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **28**, 3158–3160 (2012).
30. Zhang, R. *et al.* Drug repurposing for COVID-19 via knowledge graph completion. *J. Biomed. Inform.* **115**, 103696 (2021).
31. Zhou, H., Austin, R., Kilicoglu, H., Lu, S.-C. & Zhang, R. CIHLex: Complementary and Integrative Health Lexicon. in *American Medical Informatics Association Annual Symposium* (2022).
32. Hristovski, D., Friedman, C., Rindflesch, T. C. & Peterlin, B. Exploiting semantic relations for literature-based discovery. in *AMIA annual symposium proceedings* vol. 2006 349 (American Medical Informatics Association, 2006).
33. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. & Yakhnenko, O. Translating

embeddings for modeling multi-relational data. *Adv. Neural Inf. Process. Syst.* **26**, (2013).

34. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).
35. Vu, T., Nguyen, T. D., Nguyen, D. Q. & Phung, D. A capsule network-based embedding model for knowledge graph completion and search personalization. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 2180–2189 (2019).
36. Sun, Z., Deng, Z.-H., Nie, J.-Y. & Tang, J. Rotate: Knowledge graph embedding by relational rotation in complex space. *ArXiv Prepr. ArXiv190210197* (2019).
37. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *ArXiv Prepr. ArXiv181000826* (2018).
38. Zhang, Z., Wang, J., Ye, J. & Wu, F. Rethinking Graph Convolutional Networks in Knowledge Graph Completion. in *Proceedings of the ACM Web Conference 2022* 798–807 (2022).
39. Xu, Y. *et al.* Photodynamic Alzheimer's disease therapy: From molecular catalysis to photo-nanomedicine. *Coord. Chem. Rev.* **470**, 214726 (2022).
40. Souza, D. M. *et al.* 5-ALA-mediated photodynamic therapy reduces the parasite load in mice infected with Leishmania braziliensis. *Parasite Immunol.* **39**, e12403 (2017).
41. Zhou, C. *et al.* Genomic deletion of TLR2 induces aggravated white matter damage and deteriorated neurobehavioral functions in mouse models of Alzheimer's disease. *Aging* **11**, 7257 (2019).
42. Beydoun, M. A. *et al.* Association of serum antioxidant vitamins and carotenoids with incident Alzheimer disease and all-cause dementia among US adults. *Neurology* **98**, e2150–e2162 (2022).
43. Sun, B., Xia, Q. & Gao, Z. Total flavones of Choerospondias axillaris attenuate cardiac dysfunction and myocardial interstitial fibrosis by modulating NF-κB signaling pathway. *Cardiovasc. Toxicol.* **15**, 283–289 (2015).

44. Sawkulycz, X. *et al.* Regulation of interleukin 6 by a polymorphic CpG within the frontal cortex in Alzheimer's disease. *Neurobiol. Aging* **92**, 75–81 (2020).

45. Paouri, E., Tzara, O., Zenelak, S. & Georgopoulos, S. Genetic deletion of tumor necrosis factor-α attenuates amyloid-β production and decreases amyloid plaque formation and glial response in the 5xfad model of Alzheimer's disease. *J. Alzheimers Dis.* **60**, 165–181 (2017).

46. Yuan, C. *et al.* Dietary carotenoids related to risk of incident Alzheimer dementia (AD) and brain AD neuropathology: a community-based cohort of older adults. *Am. J. Clin. Nutr.* **113**, 200–208 (2021).

47. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

48. Yang, B., Yih, W., He, X., Gao, J. & Deng, L. Embedding entities and relations for learning and inference in knowledge bases. *ArXiv Prepr. ArXiv14126575* (2014).

49. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É. & Bouchard, G. Complex embeddings for simple link prediction. in *International conference on machine learning* 2071–2080 (PMLR, 2016).

50. Schlichtkrull, M. *et al.* Modeling relational data with graph convolutional networks. in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15* 593–607 (Springer, 2018).

51. Vashishth, S., Sanyal, S., Nitin, V. & Talukdar, P. Composition-based multi-relational graph convolutional networks. *ArXiv Prepr. ArXiv191103082* (2019).

52. Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. in *Proceedings of the AMIA Symposium* 17 (American Medical Informatics Association, 2001).

53. Kilicoglu, H., Rosemblat, G., Fiszman, M. & Shin, D. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinformatics* **21**, 1–28 (2020).

54. Rizvi, R. F. *et al.* iDISK: the integrated DIetary Supplements Knowledge base. *J. Am. Med. Inform. Assoc.* **27**, 539–548 (2020).

55. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. Health* **3**, 1–23 (2021).

56. Rindflesch, T. C. & Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.* **36**, 462–477 (2003).

57. McInnes, B. T. Extending the Log Likelihood Measure to Improve Collection Identification. (University of Minnesota, Duluth, 2004).

58. Austin, R. R. *et al.* Evaluating systemized nomenclature of medicine clinical terms coverage of complementary and integrative health therapy approaches used within integrative nursing, health, and medicine. *CIN Comput. Inform. Nurs.* **39**, 1000–1006 (2021).

59. Natural Medicines. Natural Medicines - Health & Wellness https://naturalmedicines-therapeuticresearch-com.ezp2.lib.umn.edu/databases/health-wellness.aspx.

60. Zhang, S., Tong, H., Xu, J. & Maciejewski, R. Graph convolutional networks: a comprehensive review. *Comput. Soc. Netw.* **6**, 1–23 (2019).

61. Nickel, M., Rosasco, L. & Poggio, T. Holographic embeddings of knowledge graphs. in *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 30 (2016).

62. Zheng, D. *et al.* Dgl-ke: Training knowledge graph embeddings at scale. in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* 739–748 (2020).

63. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, (2019).

64. Wang, M. Y. Deep graph library: Towards efficient and scalable deep learning on graphs. in *ICLR workshop on representation learning on graphs and manifolds* (2019).

65. Weeber, M., Klein, H., de Jong-van den Berg, L. T. & Vos, R. Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.* **52**, 548–557 (2001).

66. Henry, S. & McInnes, B. T. Literature based discovery: models, methods, and trends. *J. Biomed. Inform.* **74**, 20–32 (2017).

Tables

Table 1: The frequency and the proportion of relation types in ADInt.

Relations	**Counts (%)**	**Relations**	**Counts (%)**
COEXISTS_WITH	332,441 (32.68)	DISRUPTS	23239 (2.28)
INTERACTS_WITH	209,458 (20.59)	AUGMENTS	21,912 (2.15)
AFFECTS	96,804 (9.52)	PRODUCES	21,828 (2.14)
TREATS	90,812 (8.93)	PREDISPOSES	13,509 (1.33)
CAUSES	76,236 (7.49)	PREVENTS	12,258 (1.20)
ASSOCIATED_WITH	46,126 (4.53)	COMPLICATES	3,519 (0.35)
INHIBITS	39,158 (3.85)	MANIFESTATION_OF	1,926 (0.19)
STIMULATES	28,093 (2.76)		
TOTAL		1,017,319	

Table 2: The metrics of link prediction results for different models on integrated knowledge graph, ADInt, by time slicing evaluation.

	TransE	RotatE	DistMult	ComplEX	RGCN	CompGCN
Hits@1	0.1770	0.1786	0.1109	0.1062	**0.4112**	0.0906
Hits@3	0.3242	0.3055	0.2586	0.2467	**0.5058**	0.1509
Hits@10	0.5996	0.5340	0.5921	0.5854	**0.6804**	0.4267
MRR	0.3109	0.2987	0.2547	0.2479	**0.5007**	0.1973
MR	8.861	10.11	9.278	9.380	**7.099**	10.26

Table 3: The metrics of link prediction results for different models on integrated knowledge graph, ADInt, by clinical trials dataset evaluation.

	TransE	RotatE	DistMult	ComplEX	RGCN	CompGCN
Hits@1	0.5580	0.4545	0.2405	0.2143	**0.7906**	0.4826
Hits@3	0.6294	0.6320	0.3752	0.3058	**0.9033**	0.5497
Hits@10	0.7621	0.8107	0.5391	0.4537	**0.9848**	0.7030
MRR	0.6258	0.5768	0.3543	0.3084	**0.8582**	0.5535

MR	5.417	5.228	9.991	11.57	**1.731**	6.475

Table 4: Top 10 proposed entities for different categories with predicate PREVENTS/TREATS.

	Dietary Supplement	**Complementary and Integrated Health**
1	Caryophyllus aromaticus	Photodynamic therapy
2	Tamaris	Interpersonal psychotherapy
3	Shark Liver Oil	Guided imagery
4	Glucomannan	Laughter therapy
5	Desmodii herba	Cold therapy
6	bidens pilosa	Massage Therapy
7	Lutein	Manual lymphatic drainage
8	Artichoke	Myofascial release
9	Millet (as Grain, fiber)	Mindfulness Relaxation
10	Damask rose	Art Therapy