Computers in Biology and Medicine 192 (2025) 110285

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Alzheimer's disease knowledge graph enhances knowledge discovery and disease prediction

Yue Yang [a], Kaixian Yu [b], Shan Gao [c], Sheng Yu [d], Di Xiong [e], Chuanyang Qin [c], Huiyuan Chen [c], Jiarui Tang [a], Niansheng Tang [c], Hongtu Zhu [a,*]

[a] Department of Biostatistics, University of North Carolina at Chapel Hill, USA
[b] Insilicom LLC, Tallahassee FL, USA
[c] Department of Mathematics and Statistics, Yunnan University, China
[d] Center for Statistics Science, Tsinghua University, China
[e] Department of Mathematics, Shanghai University, China

ARTICLE INFO

Keywords:
Alzheimer's disease
Disease prediction
Knowledge graph construction
Link prediction

ABSTRACT

Objective: To construct an Alzheimer's Disease Knowledge Graph (ADKG) by extracting and integrating relationships among Alzheimer's disease (AD), genes, variants, chemicals, drugs, and other diseases from biomedical literature, aiming to identify existing treatments, potential targets, and diagnostic methods for AD.
Methods: We annotated 800 PubMed abstracts (ADERC corpus) with 20,886 entities and 4935 relationships, augmented via GPT-4. A SpERT model (SciBERT-based) trained on this data extracted relations from PubMed abstracts, supported by biomedical databases and entity linking refined via abbreviation resolution/string matching. The resulting knowledge graph trained embedding models to predict novel relationships. ADKG's utility was validated by integrating it with UK Biobank data for predictive modeling.
Results: The ADKG contained 3,199,276 entity mentions and 633,733 triplets, linking >5K unique entities and capturing complex AD-related interactions. Its graph embedding models produced evidence-supported predictions, enabling testable hypotheses. In UK Biobank predictive modeling, ADKG-enhanced models achieved higher AUROC of 0.928 comparing to 0.903 without ADKG enhancement.
Conclusion: By synthesizing literature-derived insights into a computable framework, ADKG bridges molecular mechanisms to clinical phenotypes, advancing precision medicine in Alzheimer's research. Its structured data and predictive utility underscore its potential to accelerate therapeutic discovery and risk stratification.

1. Introduction

Alzheimer's disease (AD), a progressive neurodegenerative disorder affecting over 55 million individuals globally, is characterized by amyloid-β plaques, tau tangles, and cognitive decline [1]. Despite advances in understanding multifactorial mechanisms—neuroinflammation, vascular dysfunction, and genetic risks like APOE-ε4 [2,3]—effective therapies remain elusive [2,3]. Concurrently, innovations in blood-based biomarkers, AI-driven neuroimaging, and digital health tools are reshaping early diagnosis and personalized interventions [4–6]. The controversial approval of anti-amyloid therapies like Aducanumab underscores the urgent need for robust frameworks to identify mechanistically validated therapeutic targets [7,8]. Innovations in blood-based biomarkers (e.g., p-tau217) and AI-driven diagnostics highlight the need for integrative frameworks to translate discoveries into clinical solutions [9–16].

Knowledge graphs (KGs) have emerged as pivotal tools for synthesizing fragmented AD research. Early KGs like the Gene Ontology and UMLS relied on manual curation [17–19], while later efforts (e.g., Hetionet [20], BioGrakn [21]) combined existing databases but lacked disease-specific granularity. In the AD domain, specialized ontologies such as AlzPathway [22,23] (signaling pathways) and the Alzheimer's Disease Ontology (ADO) [24] (clinical classifications) provided foundational frameworks but were limited by static, expert-driven curation. Recent initiatives like AlzGPS [25], TACA [26], and the Alzheimer's Knowledge Base [27] integrated multi-omics data but underutilized literature-derived interactions, leaving critical gaps in capturing emerging hypotheses.

Advances in data-driven KG construction are addressing these gaps.

* Corresponding author.
E-mail address: htzhu@email.unc.edu (H. Zhu).

https://doi.org/10.1016/j.compbiomed.2025.110285
Received 27 October 2024; Received in revised form 26 March 2025; Accepted 24 April 2025
Available online 29 April 2025

Abbreviation

KG	knowledge graph
AD	Alzheimer's disease
ADKG	Alzheimer's disease knowledge graph
NER	Named Entity Recognition
KGE	Knowledge Graph Embedding

Zhu et al. [28], established AD-specific KGs using Att-BiLSTM-CRF models to improve named entity recognition (NER) over rule-based systems, though their relation extraction remained constrained by limited semantic scope. Nian et al. [29] leveraged SemMedDB [30,31] to infer broader relationships (e.g., drug-pathway interactions), yet persistent challenges in entity disambiguation (e.g., distinguishing gene isoforms) hindered clinical applicability. Innovations in KG embedding—heterogeneous relation attention networks [32] (asymmetric relationships), recalibration convolutional networks [33] (noise resilience), and multi-granularity relational augmentation [34] (hierarchical interactions)—have enhanced link prediction accuracy. Frameworks like KEMA++ [35], TransModE [36], and DREAMwalk [37] further demonstrate the translational potential of these embeddings, prioritizing drug candidates (e.g., nilotinib) with validated preclinical efficacy. Despite progress, critical gaps persist: (1) incomplete entity normalization (e.g., ambiguous gene symbols), (2) limited handling of domain-specific abbreviations (e.g., "NFT" as neurofibrillary tangle vs. non-fungible token), and (3) reliance on expert curation rather than scalable, crowd-validated annotations.

Modern KG construction relies on two complementary information extraction (IE) paradigms: open IE (e.g., e.g., ReVerb [38], OLLIE [39], Stanford OpenIE [40], ClausIE [41], and SemRep [31]) for unsupervised relation discovery and supervised IE (e.g., SCIIE [42], SpERT [43]) for high-precision, domain-specific extraction. While open IE autonomously detects generic relationships, supervised methods like SpERT—trained on annotated corpora—achieve superior accuracy in biomedical contexts. Equally critical is entity linking, where tools like QuickUMLS [44] and TaggerOne [45] map textual mentions to standardized entries (e.g., "Tau" → MAPT gene in UniProt), resolving ambiguities that plague AD literature (e.g., "AD" denoting Alzheimer's disease vs. atopic dermatitis). Despite these advancements, no existing AD KG systematically integrates modern NLP (e.g., GPT-4 [46] –augmented training), crowdsourced validation, and multi-database entity resolution.

In this paper, we introduce a new AD knowledge graph (ADKG) built from a large corpus of PubMed abstracts. Our pipeline employs SpERT [43] for supervised entity and relation extraction, enhanced by GPT-4 [46] –driven data augmentation and robust abbreviation resolution. We then integrate these triplets with external databases (e.g., NCBI Gene [47], ChEBI [48]) via entity linking to form a comprehensive ADKG. Our key contributions are.

1. A human-annotated benchmark dataset (ADERC) focused on AD-specific NER and relation classification,
2. A domain-specific pipeline incorporating both advanced NLP and reference database linkage for KG construction, and
3. A demonstration of how the ADKG facilitates novel relationship prediction and improves AD risk modeling using data from the UK Biobank.

We conclude by discussing current limitations, including entity coverage and potential biases, and outline future steps to continually update and refine the ADKG.

2. Methods

The construction of the Alzheimer's Disease Knowledge Graph (ADKG) followed a multi-stage pipeline comprising corpus generation, triplet extraction, entity linking, knowledge fusion, and embedding (Fig. 1). We begin by outlining the development of the Alzheimer's Disease Entity-Relation Corpus (ADERC) and then discuss how this corpus was used to train an information extraction model for triplet detection. Finally, we detail the methods applied to construct the ADKG.

2.1. Corpus generation

To create the Alzheimer's Disease Entity-Relation Corpus (ADERC), 800 PubMed abstracts were selected from a larger set of 169,630 abstracts retrieved using the keyword "Alzheimer" via the Entrez function of the Bio package (accessed on February 5, 2021). Initial entity tagging was conducted with BERN [49], which automatically recognized genes, diseases, drugs, and mutations. Human annotators then broadened the entity categories to include Method (e.g., "18F-FDG-PET") and Other for mentions outside the predefined types.

Eight relationship types—treatment_for, treatment_target_for, help_diagnose, risk_factor_of, characteristic_for, hyponym_of, associated_with, and abbreviation_for—were identified as particularly relevant to AD research and were manually annotated using BRAT [50]. This manual process with 8 annotators led to a comprehensively labeled corpus encompassing both biomedical entities and their interrelations across 800 abstracts.

To address the limited availability of annotated data, GPT-4 [46] was employed to generate synthetic variations of sentences that included relationships. Techniques such as synonym substitution and paraphrasing were used to diversify the training examples. Expert reviewers evaluated these automatically produced sentences to ensure consistency, accuracy, and to reduce potential biases.

2.2. Information triplet extraction

A key strategy for managing the complexity of AD research is to extract interaction triplets, each composed of a head entity, a tail entity, and a relationship linking them [51]. For example, from the statement "PPARgamma may be a potential target for AD," one can derive the triplet ⟨PPARgamma, potential target for, AD⟩. Collectively, these triplets—of the form ⟨head entity, relationship, tail entity⟩—offer a machine-readable representation amenable to large-scale data mining and knowledge graph construction.

Using SpERT [43] framework with SciBERT [52] embeddings, we trained a unified model that simultaneously performs NER and relation extraction on ADERC, with the input of tokenized sentences. For negative sampling, we augmented the usual practice of including non-entity text and irrelevant entity pairs by introducing hand-engineered negative instances; here, positive examples were systematically altered by substituting either entities or relations to create realistic negative samples. This strategy balanced the dataset with both positive and negative cases, mitigating bias and enhancing precision in entity–relation detection.

After annotating the corpus, we divided the data into training, validation, and testing sets by randomization over relationship types. We trained the model across various parameter combinations—such as different relationship filtering thresholds, embedding dimensions, negative sampling rates, and dropout values—and selected the best-performing parameters based on validation set results. The finalized model, employing these optimal settings, was then used to extract triplets from the remaining unannotated abstracts.

We evaluated performance using precision, recall, and F1 scores for three tasks: entity recognition, relation extraction, and the joint model. Precision is defined as the proportion of correctly predicted positive instances among all instances predicted as positive. Recall is the

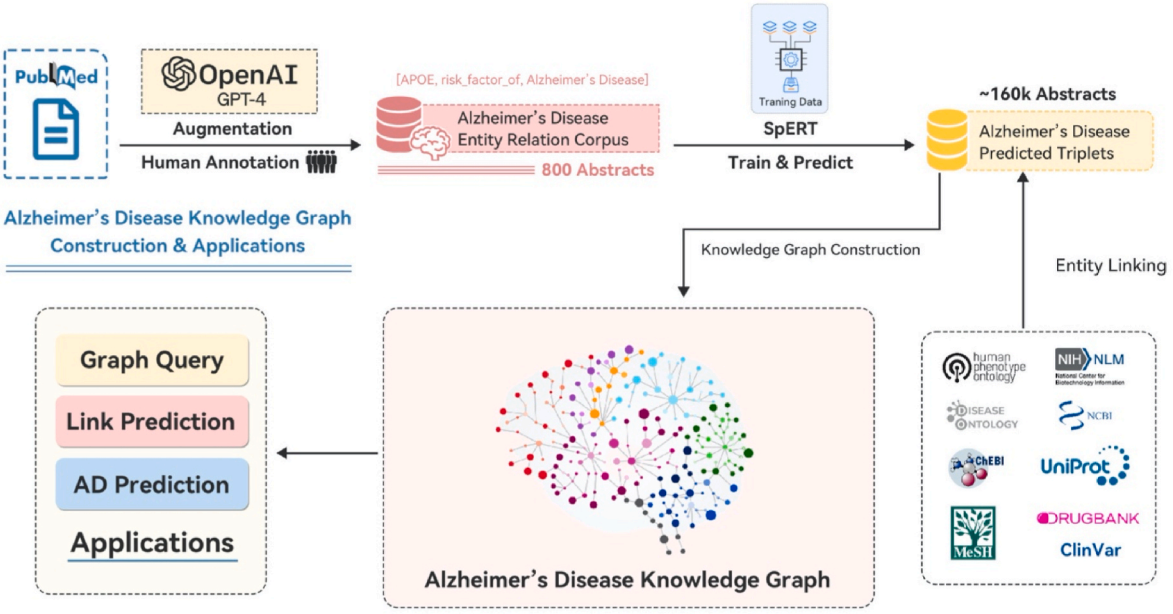


Fig. 1. General Pipeline: corpus generation, model building, entity linking, ADKG construction, and applications.

proportion of correctly predicted positive instances among all actual positive instances. Micro F1 Score is computed by aggregating precision and recall over all instances before calculating the harmonic mean, useful for addressing class imbalance by weighting each instance equally. Macro F1 Score is calculated separately for each class and then averaged, ensuring equal emphasis across distinct entity and relation types.

2.3. Entity linking

Consistency in entity representation is critical for coherent knowledge graph construction. To address inconsistent entity mentions—both within a single abstract and across multiple documents—we implemented a multi-database entity linking procedure. This approach mapped extracted entities to their canonical identifiers in relevant biomedical databases: NCBI Gene [47], UniProt [53], ChEBI [48], DrugBank [54], HPO [55], Disease Ontology [56], ClinVar [57,58], and MeSH [59]. We used simstring [60] to perform approximate string matching between extracted entity mentions and standard database names or synonyms. Each successfully linked entity was assigned a unique identifier (ID), accompanied by descriptive metadata to facilitate consistent entity resolution throughout the corpus.

2.3.1. Abbreviation resolution

Abbreviations in biomedical text often introduce ambiguity, as multiple long forms can map to a single short form. For example, "ASD" might refer to either "autism spectrum disorder" or "atrial septal defect". To address this issue in a context specific manner, we included an "abbreviation_for" relationship in our annotation schema. During the annotation process, abbreviation–long form pairs were explicitly recorded when both appeared in the same abstract, leveraging local context to improve disambiguation. By linking each short form to its expanded version, we reduced ambiguity and enhanced the precision of subsequent entity recognition and relation extraction steps.

2.4. Knowledge graph construction and the confidence

A knowledge graph (KG), $G(\mathbf{X}, \mathbf{E})$ consists of nodes $\{X_1, X_2, \ldots, X_N\} \in \boldsymbol{X}$ and edges, $\{E_1, E_2, \ldots, E_K\} \in \boldsymbol{E}$ between nodes. In this study, to build a knowledge graph, $G(\mathbf{X}, \mathbf{E})$, for AD from the existing literature, we extract entities (nodes) and relationships (edges) from abstracts related to AD.

In the node creation phase, abbreviations were resolved to their long forms, and entities were linked to standard databases. Subsequently, for each extracted triplet, we added a directed edge annotated with the relationship type, PubMed ID, text span of the head and tail entities, and the associated matching scores. Situations with multiple edges connecting the same node pair were resolved by prioritizing the most frequently observed relationship type.

2.5. Knowledge fusion

To increase coverage and reliability, we integrated external resources into the ADKG, a process often referred to as knowledge fusion [61]. Aligning entities and relationships from multiple databases helps ensure that the ADKG remains comprehensive and up to date. External data sources included DisGeNET [62], The Human Phenotype Ontology (HPO) [55], DrugBank [54], PharmGKB (Pharmacogenomics Knowledgebase) [63], OMIM (Online Mendelian Inheritance in Man) [64], and STRING [65].

2.6. Knowledge graph embedding

To facilitate novel knowledge discovery within ADKG, we generated knowledge graph embeddings. We split the ADKG into training (80 %), validation (10 %), and test (10 %) subsets, ensuring balanced distribution of relationships and comprehensive representation of entities.

We tested a range of methods to encode our knowledge graph, including: distance-based models like TransE [66], TransH [67], and TransR [68], the semantic matching-based ComplEx model [69], and the ConvKB model [70], which incorporates convolutional neural networks. We performed an extensive hyperparameter search over embedding dimensions (32, 64, 128, 256), learning rates (0.05, 0.005, 0.0005), batch sizes (128, 256, 512, 1024), and other model-specific parameters. A margin-based ranking loss with Bernoulli negative sampling [67] was used, and training was capped at 1000 epochs. We evaluated model performance using Hits@10 and Mean Rank under a filtered setting [66], while Hits@10 is the proportion of valid entities ranked in the top 10, reflecting high-confidence retrieval and mean Rank is the average ranking position of the correct entities, with a lower mean rank

indicating superior performance.

The best-performing model was selected based on Mean Rank, supplemented by Hits@10 to provide an additional measure of accuracy for top-ranked predictions. This final embedding model was then applied to the complete ADKG for link prediction and the identification of new, potentially impactful associations related to Alzheimer's disease.

2.6.1. Alzheimer's disease prediction using UK biobank data

We evaluated the Alzheimer's Disease Knowledge Graph (ADKG) using the UK Biobank cohort [71], which includes genetic, proteomic, and clinical data from ~500,000 participants. As of March 2023, we extracted protein expression profiles, lifestyle factors (diet, physical activity), and medical histories for 54,705 individuals. AD cases (N = 541) were identified via ICD-9/ICD-10 codes (e.g., G30) from hospital records or death registers, while controls (N = 52,164) excluded all dementia-related diagnoses (F00-F03, G30-G32).

ADKG-derived predictors were selected based on: (1) literature support (>2 PubMed references), (2) availability in UK Biobank (>50 % participant coverage), and (3) biological relevance to AD pathogenesis. This yielded 130 proteins (e.g., *GFAP*, *NEFL*) overlapping with UK Biobank assays, alongside demographic (age, APOE-ε4 status), lifestyle (sleep duration, alcohol use), and clinical variables (memory scores). Missing proteomic data were imputed via mean substitution [72], and class imbalance was addressed using ROSE oversampling [73]. Predictive models (logistic regression, XGBoost [74]) incorporated 214 variables, including ADKG-curated features and conventional predictors.

3. Results

3.1. Statistics for ADERC and ADKG

The Alzheimer's Disease Entity Relation Corpus (ADERC) comprises 800 PubMed abstracts annotated with 20,886 biomedical entities and 4935 relationships (Fig. 2A–B). The original predictions (based on SpERT model with SciBERT embeddings on GPT augmented annotations) on all the abstracts contains in total 3,199,276 entity mentions and 633,733 triplets, among which we identified 45,277 unique entities between mapped entities after entity linking (Fig. 2C–D).

To confirm the ADKG's accuracy and completeness, we randomly sampled 200 sentences from AD-related literature and compared the extracted triplets to the ground truth. Each triplet's faithful representation of the source sentence was validated by domain specialists following an initial GPT-4 screening. Precision, defined as the proportion of correctly represented triplets among all extracted triplets, reached 72 % (172 out of 239). Common errors at the KG level mirrored the model-level error patterns—primarily stemming from ambiguous abbreviations and closely related biomedical terms—underscoring the need for more robust disambiguation strategies. Nonetheless, these results demonstrate the ADKG's efficacy in capturing pertinent information.

Entity linking via SimString achieved 64 % precision (the proportion of correctly linked entities out of all entities predicted by the model) and 53 % recall (the proportion of correctly linked entities out of all ground truth entities.) (n-gram = 3, similarity threshold = 0.6) in a 100-mention validation set. While effective for approximate matches, performance gaps highlight opportunities to integrate contextual embeddings or domain-specific synonym dictionaries in future work.

3.2. Ablation study of joint model

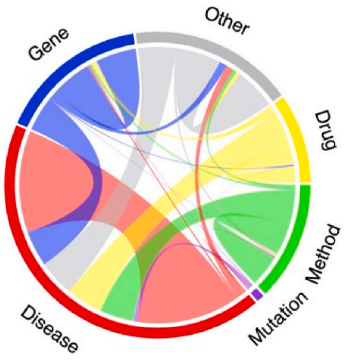
The Joint model, designed to concurrently perform entity recognition and relation extraction, achieved micro and macro F1 scores of 61.4 % and 61.5 %, respectively. While these scores are comparatively lower than those of models optimized for individual tasks (e.g., standalone entity recognition or relation extraction), this gap reflects the inherent complexity of jointly optimizing interdependent tasks. Specifically, errors in entity recognition propagate to downstream relation extraction, amplifying challenges in achieving high precision for both tasks simultaneously.

To dissect these results, we conducted an ablation study evaluating

A

Entity type	#Entity	Relation type	#Relation
Disease	8581	abbreviation_for	1864
Gene	4912	associated_with	1270
Mutation	221	treatment_for	595
Drug	2849	treatment_target_for	48
Method	2728	help_diagnose	348
Other	1595	risk_factor_of	456
		characteristic_of	354

B



C

Entity type	#Entity	Relation type	#Relation
Disease	1331854	abbreviation_for	330695
Gene	879646	associated_with	121603
Mutation	30976	treatment_for	61192
Drug	451034	treatment_target_for	3956
Method	336865	help_diagnose	31319
Other	168901	risk_factor_of	47384
		characteristic_of	37584

D

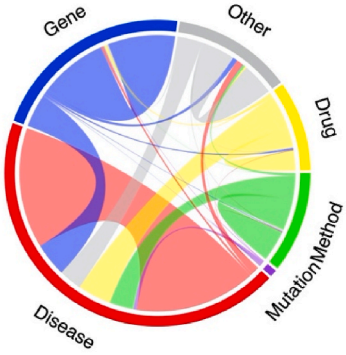


Fig. 2. Comparative Visualization of Biomedical Entity and Relationship Distribution for ADERC (A and B) and ADKG (C and D). (Not all the tail entities are Alzheimer's disease.)

the impact of two critical factors: pretrained embeddings (SciBERT vs. PubMedBERT) and GPT-4-derived data augmentation. On the ADERC dataset, SciBERT achieved an entity recognition micro F1 of 86.4 % without augmentation, improving to 88.9 % with augmentation, while PubMedBERT scored 85.7 % (unaugmented) and 87.8 % (augmented). For relation extraction, SciBERT's micro F1 rose from 66.2 % to 69.1 % with augmentation, and PubMedBERT improved from 64.8 % to 67.4 %. These results highlight two key trends. First, SciBERT consistently outperformed PubMedBERT across tasks, likely due to its domain-specific pretraining on scientific text. Second, GPT-4 augmentation enhanced performance for both tasks, mitigating data scarcity and improving generalizability.

3.3. ADKG featured important relationships for AD

The Alzheimer's Disease Knowledge Graph (ADKG) elucidates critical molecular and clinical associations, beginning with the well-established role of APOE as a genetic risk factor. Interrogating ADKG's genetic landscape revealed 5932 interactions linking AD to 1030 genes/proteins, 89 % of which overlap with ADSP Gene Verification Committee–curated risk/protective genes [75]. Pharmacological analysis identified 5665 AD–drug/chemical interactions, including repurposable candidates (e.g., sildenafil) and compounds targeting amyloid-β clearance. Beyond molecular mechanisms, ADKG maps disease comorbidities, connecting AD to 248 conditions—most prominently mild cognitive impairment (DOID:0081292), diabetes mellitus (DOID:9351), and obesity (DOID:9970). These associations, derived from 5130 literature-supported interactions, underscore AD's multifactorial etiology and inform strategies for managing comorbid conditions in clinical practice. Together, these findings position ADKG as a dynamic resource for hypothesis generation, bridging molecular pathways to patient-centered outcomes in Alzheimer's research.

3.4. Results of knowledge graph embedding

Systematic evaluation of knowledge graph embedding (KGE) models revealed ConvKB as the top performer, achieving a mean rank of 312 and Hits@10 of 27.8 %, outperforming TransE (mean rank: 387), TransH (373), TransR (377), and ComplEx (340) (Table 1). This superiority stems from ConvKB's convolutional neural network architecture, which captures intricate interaction patterns between entities.

3.5. Link prediction results reveal interesting findings

The application of ConvKB (a knowledge graph embedding model) to the full Alzheimer's Disease Knowledge Graph (ADKG) facilitated the identification of novel gene-disease associations which do not present in the original training corpus. Among the high-confidence predictions (Table 2), *CHI3L1* was linked to hippocampal atrophy (PubMed ID: 35234337) and *MFN2* to mitochondrial dysfunction (PubMed ID: 30649465)—relationships corroborated by recent independent studies excluded from the training data. Further, *APOE* was implicated in contexts, such as amyloidosis and tauopathy, while inflammatory biomarkers (*CRP*, *IL6*) showed novel associations with gastrointestinal and neuropsychiatric comorbidities in Alzheimer's patients. These results highlight the ADKG's ability to infer biologically plausible connections beyond its initial literature scope, generating actionable hypotheses for mechanistic validation.

Table 1
Knowledge Graph Embedding performance of the best setting on test set for different KGE models.

Model	Mean Rank	Hits@10
TransE	387	0.1646
TransH	373	0.1973
TransR	377	0.1941
ComplEx	340	0.2125
ConvKB	**312**	0.2781

3.6. ADKG empowers Alzheimer's disease prediction using UK biobank data

To evaluate the predictive utility of the Alzheimer's Disease Knowledge Graph (ADKG), we leveraged multimodal data from the UK Biobank [71], a population-scale cohort comprising genetic, proteomic, demographic, and clinical records from ~500,000 participants. From the ADKG, we extracted 130 proteins (e.g., *GFAP*, *NEFL*) overlapping with the UK Biobank's proteomic assays, alongside demographic (age, APOE-ε4 carrier status), lifestyle (sleep duration, alcohol consumption), and clinical variables (memory assessment scores). These variables were integrated into a predictive framework to model Alzheimer's disease (AD) risk, with clinical diagnoses defined by ICD-9/ICD-10 codes (e.g., G30 in ICD-10) and controls comprising participants without dementia-related diagnoses.

We contrasted two variable selection approaches. The first, ADKG-informed selection, prioritized genes, proteins, and comorbidities (e.g., neuropsychiatric symptoms) directly linked to AD through the knowledge graph. The second, conventional marginal screening, applied increasingly stringent p-value thresholds (from 0.05 to 0.0000005) to identify statistically significant predictors without domain-specific context. The ADKG-driven approach yielded a curated set of 214 predictors, including established biomarkers (e.g., *APOE-ε4*), proteins (e.g., *CHI3L1*), and lifestyle factors. In contrast, marginal screening produced unstable variable sets sensitive to thresholding, often excluding biologically relevant but statistically modest signals.

The ADKG-enhanced model achieved an AUC of 0.9137 (95 % CI: 0.907–0.920), outperforming the baseline model (AUC = 0.9025) that relied solely on traditional screening. Further refinement using XGBoost with ADKG-derived predictors elevated performance to AUC = 0.928, demonstrating the synergy between domain knowledge and machine learning. Crucially, noise perturbation experiments (Supplementary Material) revealed that ADKG-informed predictors exhibited lower performance degradation compared to marginally screened variables under simulated data corruption, underscoring their robustness. Fig. 3 summarizes this workflow, highlighting how structured biomedical knowledge bridges gaps in data-driven discovery. These results affirm that domain-specific frameworks like the ADKG enhance both the accuracy and generalizability of predictive models in complex diseases like Alzheimer's.

4. Discussion

Our study presents the Alzheimer's Disease Knowledge Graph (ADKG), a structured resource derived from 800 PubMed abstracts, and the Alzheimer's Disease Entity-Relation Corpus (ADERC), a benchmark dataset for AD-specific entity and relationship annotation. The ADKG integrates relationships among genes, chemicals, and diseases, enabling drug repurposing and biomarker discovery. By applying knowledge graph embedding (KGE) models, we identified novel associations and demonstrated its utility in predictive modeling using UK Biobank data (AUROC: 0.928). This underscores ADKG's capacity to bridge mechanistic insights with clinical risk prediction.

The construction of ADKG relied solely on publicly available abstracts, ensuring that patient privacy was maintained by excluding any identifiable information. Future expansions that incorporate patient-linked data, such as genomic or electronic health record datasets, will necessitate rigorous adherence to GDPR and HIPAA protocols through de-identification, informed consent, and ethical oversight. We also recognize that biases, such as the underrepresentation of non-European populations in genetic studies, may be present in our source data. To mitigate these issues, we plan to conduct proactive bias audits, strive for

Table 2
Top inferred triplets inferred from ADKG using ConvKB (red PubMed evidence indicates that the source is not included in our corpus).

Type	head	tail	rank	score	Pubmed Evidence
gene-disease	*APOE*	amyloidosis	38	44.80191	Many
	CHI3L1	Neurodegeneration	45	44.66664	35234337
	MFN2	Abnormality of mitochondrial metabolism	46	44.66243	30649465
	APOE	tauopathy	59	44.33801	Many
	CRP	Gastrointestinal inflammation	64	44.21387	Many
	HMOX1	progressive supranuclear palsy	74	43.9729	
	IL6	major depressive disorder	95	43.41697	Many
	NEFL	Neurodegeneration	99	43.35546	Many
	UBB	neurodegenerative disease	101	43.32749	Many
	CHI3L1	Hippocampal atrophy	109	43.21873	35234337

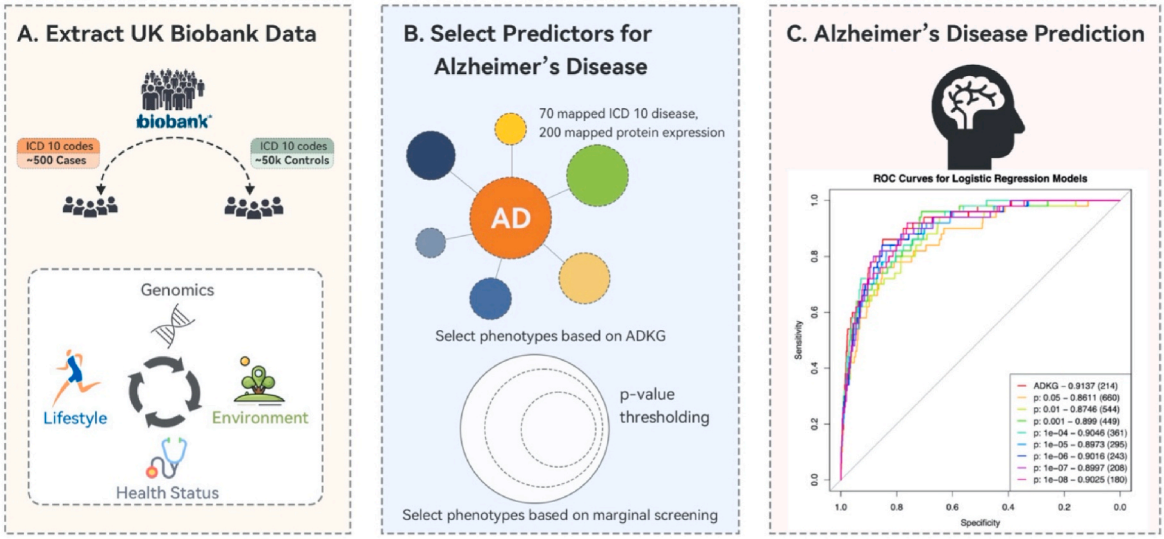


Fig. 3. Efficacy of ADKG in AD prediction with UK Biobank Data.

demographic balance, and maintain transparent reporting in future iterations.

To keep pace with the rapidly expanding biomedical literature, the ADKG pipeline is designed for modular updates. It features semi-automated annotation using previously trained SpERT model, incremental entity linking through federated databases, and periodic retraining of KGE models. This adaptable framework also facilitates the integration of emerging data types, including non-coding RNAs and proteome, without necessitating a complete graph reconstruction.

Although ADKG performs well in modeling one-to-one relationships, such as gene–disease relationship, it currently does not support direct extraction of more complex relational patterns like N-ary interactions (e.g., drug–gene–disease combinatorial effects), instead the methods will break N-ary into a series of binaries. This limitation arises from both the binary focus of SpERT and the scarcity of annotated N-ary examples in ADERC. To address this shortcoming, future work will explore hyper-relational KG frameworks and expand ADERC to incorporate composite relationships.

In our comparative analysis, ConvKB achieved a Hits@10 score of 27.8 %, outperforming TransE (19.1 %), TransH (21.3 %), and ComplEx (18.7 %). This performance gap stems from architectural limitations in the alternative models. TransE, which maps relationships as linear translations in embedding space, fails to capture asymmetric or hierarchical interactions, as its simplicity restricts representational flexibility. TransH partially addresses this by projecting entities onto relation-specific hyperplanes but still struggles with multi-hop relational patterns common in biomedical knowledge graphs (KGs). ComplEx, while theoretically capable of modeling asymmetric relations through complex-valued embeddings, underperformed in capturing hierarchical dependencies due to its reliance on multiplicative score functions, which dilute structural signals in sparse biomedical KGs. ConvKB circumvents these limitations by employing convolutional neural networks (CNNs) to model localized interactions between entities and relations. This allows it to detect nuanced patterns while maintaining scalability.

Beyond the applications demonstrated here, the structured data provided by ADKG holds potential for enhancing large language models for public-facing Alzheimer's disease Q&A systems. However, expanding this scope will require the integration of full-text articles and supplementary materials, such as tables and clinical notes, to capture more nuanced relationships, including negative associations and protective effects. Current limitations include sparse coverage of non-core entities like non-coding RNAs and an underrepresentation of diverse populations in genetic studies. Future iterations will focus on multi-omics integration, multilingual text mining, and more refined entity classification to distinguish between related biomolecules, such as proteins and their isoforms.

5. Conclusion

This study constructs the Alzheimer's Disease Knowledge Graph (ADKG) through advanced NLP techniques, including GPT-4-augmented text mining and entity linking with biomedical databases. ADKG's integration with UK Biobank data improved predictive modeling (AUROC: 0.928), highlighting its value in identifying risk factors like APOE-ε4 and *GFAP*. By transforming unstructured literature into a computable network, ADKG accelerates hypothesis generation for therapeutic discovery. Future work will focus on enriching data sources, addressing model limitations in complex relationships, and ensuring

ethical scalability. This framework exemplifies how domain-specific knowledge graphs can advance precision medicine in neurodegenerative diseases.

CRediT authorship contribution statement

Yue Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kaixian Yu:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Shan Gao:** Writing – review & editing, Visualization, Data curation. **Sheng Yu:** Writing – review & editing, Methodology. **Di Xiong:** Writing – review & editing, Data curation. **Chuanyang Qin:** Writing – review & editing, Data curation. **Huiyuan Chen:** Writing – review & editing, Data curation. **Jiarui Tang:** Writing – review & editing, Data curation. **Niansheng Tang:** Writing – review & editing, Supervision. **Hongtu Zhu:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

6. Statement of significance

6.1. Problem

Alzheimer's disease (AD) is a progressive neurodegenerative disorder with increasing prevalence and currently lacks effective treatments. There is an urgent need for innovative research tools to enhance our understanding of AD and to facilitate the discovery of new therapeutic targets and biomarkers.

6.2. What is already known

Knowledge graphs (KGs) have emerged as valuable tools in biomedical research, enabling the analysis of complex interactions among biological entities and supporting drug repurposing and biomarker discovery. Previous efforts have developed KGs and ontologies related to AD, but they often lack comprehensive integration of literature-derived data using advanced natural language processing (NLP) techniques.

6.3. What this paper adds

This study introduces the Alzheimer's Disease Knowledge Graph (ADKG), a comprehensive, literature-derived resource encapsulating interactions among genes, proteins, chemicals, drugs, and diseases related to AD. By creating a novel annotated dataset (ADERC) and employing advanced NLP techniques—including GPT-4 for data augmentation—we significantly enhance the accuracy of named entity recognition and relation extraction in.

AD literature. The ADKG enables the discovery of novel associations through knowledge graph embedding and link prediction methods. Notably, integrating the ADKG into predictive models using UK Biobank data improved performance in identifying individuals at risk of AD. Furthermore, we have developed an accessible website (biomedkg.com) where researchers can query and explore the ADKG, facilitating wider use and collaboration in the AD research community. This resource has the potential to advance understanding of disease mechanisms and inform the development of new treatment strategies for Alzheimer's disease.

Availability of data and material

The datasets generated during and/or analyzed during the current study are available in the Zenodo repository https://doi.org/10.5281/zenodo.5770100. The knowledge graph ADKG is accessible via our developed website https://biomedkg.com/for easy query and visualization.

Ethics statement

This research complies with all relevant ethical regulations and institutional guidelines.

Ethical approval

For studies involving human data from the **UK Biobank** (Application ID: 22783), ethical approval was granted by the UK Biobank Research Ethics Committee.

Human annotations

The creation of the **Alzheimer's Disease Entity Relation Corpus (ADERC)** involved human annotators (domain experts and trained reviewers). All annotators provided informed consent, and their contributions were anonymized to protect privacy.

Data privacy

No personally identifiable information (PII) was accessed, stored, or analyzed. All datasets, including those derived from PubMed abstracts and the UK Biobank, were processed in de-identified form on secure, encrypted servers with restricted access.

Open Science.

The datasets generated during and/or analyzed during the current study are available in the Zenodo repository https://doi.org/10.5281/zenodo.5770100. The knowledge graph ADKG is accessible via our developed website https://biomedkg.com/for easy query and visualization.

Animal Research.

This study did not involve animal experimentation.

Funding sources

Ms Yang and Dr. Zhu was partially supported by the National Institute On Aging (NIA) of the National Institutes of Health (NIH) under Award Numbers RF1AG082938 and 1R01AG085581 and the Gillings Generative AI Award to Dr. Zhu.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hongtu Zhu reports financial support was provided by National Institute On Aging (NIA) of the National Institutes of Health (NIH) under Award Numbers RF1AG082938 and 1R01AG085581. Hongtu Zhu reports financial support was provided by Gillings Generative AI Award. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.110285.

References

[1] Alzheimer's disease facts and figures, Alzheimer's Dement. 19 (2023) 1598–1695, https://doi.org/10.1002/alz.13016, 2023.

[2] D.J. Selkoe, J. Hardy, The amyloid hypothesis of Alzheimer's disease at 25 years, EMBO Mol. Med. 8 (2016) 595–608, https://doi.org/10.15252/emmm.201606210.

[3] E.E. Congdon, E.M. Sigurdsson, Tau-targeting therapies for Alzheimer disease, Nat. Rev. Neurol. 14 (2018) 399–415, https://doi.org/10.1038/s41582-018-0013-z.
[4] A. Nakamura, N. Kaneko, V.L. Villemagne, T. Kato, J. Doecke, V. Doré, C. Fowler, Q.-X. Li, R. Martins, C. Rowe, T. Tomita, K. Matsuzaki, K. Ishii, K. Ishii, Y. Arahata, S. Iwamoto, K. Ito, K. Tanaka, C.L. Masters, K. Yanagisawa, High performance plasma amyloid-β biomarkers for Alzheimer's disease, Nature 554 (2018) 249–254, https://doi.org/10.1038/nature25456.
[5] C.R. Jack, D.A. Bennett, K. Blennow, M.C. Carrillo, B. Dunn, S.B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J.L. Molinuevo, T. Montine, C. Phelps, K.P. Rankin, C.C. Rowe, P. Scheltens, E. Siemers, H.M. Snyder, R. Sperling, Contributors, NIA-AA research framework: toward a biological definition of Alzheimer's disease, alzheimers dement, J. Alzheimers Assoc. 14 (2018) 535–562, https://doi.org/10.1016/j.jalz.2018.02.018.
[6] M.A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review, Comput. Methods Progr. Biomed. 187 (2020) 105242, https://doi.org/10.1016/j.cmpb.2019.105242.
[7] D.S. Knopman, D.T. Jones, M.D. Greicius, Failure to demonstrate efficacy of aducanumab: an analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019, Alzheimers Dement. J. Alzheimers Assoc. 17 (2021) 696–701, https://doi.org/10.1002/alz.12213.
[8] J. Cummings, Y. Zhou, G. Lee, K. Zhong, J. Fonseca, F. Cheng, Alzheimer's disease drug development pipeline: 2023, Alzheimers Dement, Transl. Res. Clin. Interv. 9 (2023) e12385, https://doi.org/10.1002/trc2.12385.
[9] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S.G. Costafreda, A. Dias, N. Fox, L.N. Gitlin, R. Howard, H.C. Kales, M. Kivimäki, E.B. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E.L. Sampson, Q. Samus, L.S. Schneider, G. Selbæk, L. Teri, N. Mukadam, Dementia prevention, intervention, and care: 2020 report of the Lancet Commission, Lancet Lond. Engl. 396 (2020) 413–446, https://doi.org/10.1016/S0140-6736(20)30367-6.
[10] M.D. Sweeney, A. Montagne, A.P. Sagare, D.A. Nation, L.S. Schneider, H.C. Chui, M.G. Harrington, J. Pa, M. Law, D.J.J. Wang, R.E. Jacobs, F.N. Doubal, J. Ramirez, S.E. Black, M. Nedergaard, H. Benveniste, M. Dichgans, C. Iadecola, S. Love, P. M. Bath, H.S. Markus, R.A. Salman, S.M. Allan, T.J. Quinn, R.N. Kalaria, D. J. Werring, R.O. Carare, R.M. Touyz, S.C.R. Williams, M.A. Moskowitz, Z. S. Katusic, S.E. Lutz, O. Lazarov, R.D. Minshall, J. Rehman, T.P. Davis, C. L. Wellington, H.M. Gonzalez, C. Yuan, S.N. Lockhart, T.M. Hughes, C.L.H. Chen, P. Sachdev, J.T. O'Brien, I. Skoog, L. Pantoni, D.R. Gustafson, G.J. Biessels, A. Wallin, E.E. Smith, V. Mok, A. Wong, P. Passmore, F. Barkof, M. Muller, M.M. B. Breteler, G.C. Roman, E. Hamel, S. Seshadri, R.F. Gottesman, M.A. van Buchem, Z. Arvanitakis, J.A. Schneider, L.R. Drewes, V. Hachinski, C.E. Finch, A.W. Toga, J. M. Wardlaw, B.V. Zlokovic, Vascular dysfunction – the disregarded partner of Alzheimer's disease, Alzheimers Dement, J. Alzheimers Assoc. 15 (2019) 158–167, https://doi.org/10.1016/j.jalz.2018.07.222.
[11] R. Sims, S.J. van der Lee, A.C. Naj, C. Bellenguez, N. Badarinarayan, J. Jakobsdottir, B.W. Kunkle, A. Boland, R. Raybould, J.C. Bis, E.R. Martin, B. Grenier-Boley, S. Heilmann-Heimbach, V. Chouraki, A.B. Kuzma, K. Sleegers, M. Vronskaya, A. Ruiz, R.R. Graham, R. Olaso, P. Hoffmann, M.L. Grove, B. N. Vardarajan, M. Hiltunen, M.M. Nöthen, C.C. White, K.L. Hamilton-Nelson, J. Epelbaum, W. Maier, S.-H. Choi, G.W. Beecham, C. Dulary, S. Herms, A.V. Smith, C.C. Funk, C. Derbois, A.J. Forstner, S. Ahmad, H. Li, D. Bacq, D. Harold, C. L. Satizabal, O. Valladares, A. Squassina, R. Thomas, J.A. Brody, L. Qu, P. Sánchez-Juan, T. Morgan, F.J. Wolters, Y. Zhao, F.S. Garcia, N. Denning, M. Fornage, J. Malamon, M.C.D. Naranjo, E. Majounie, T.H. Mosley, B. Dombroski, D. Wallon, M.K. Lupton, J. Dupuis, P. Whitehead, L. Fratiglioni, C. Medway, X. Jian, S. Mukherjee, L. Keller, K. Brown, H. Lin, L.B. Cantwell, F. Panza, B. McGuinness, S. Moreno-Grau, J.D. Burgess, V. Solfrizzi, P. Proitsi, H.H. Adams, M. Allen, D. Seripa, P. Pastor, L.A. Cupples, N.D. Price, D. Hannequin, A. Frank-García, D. Levy, P. Chakrabarty, P. Caffarra, I. Giegling, A.S. Beiser, V. Giedraitis, H. Hampel, M.E. Garcia, X. Wang, L. Lannfelt, P. Mecocci, G. Eiriksdottir, P. K. Crane, F. Pasquier, V. Boccardi, I. Henández, R.C. Barber, M. Scherer, L. Tarraga, P.M. Adams, M. Leber, Y. Chen, M.S. Albert, S. Riedel-Heller, V. Emilsson, D. Beekly, A. Braae, R. Schmidt, D. Blacker, C. Masullo, H. Schmidt, R.S. Doody, G. Spalletta, W.T. Longstreth, T.J. Fairchild, P. Bossù, O.L. Lopez, M.P. Frosch, E. Sacchinelli, B. Ghetti, Q. Yang, R.M. Huebinger, F. Jessen, S. Li, M.I. Kamboh, J. Morris, O. Sotolongo-Grau, M.J. Katz, C. Corcoran, M. Dunstan, A. Braddel, C. Thomas, A. Meggy, R. Marshall, A. Gerrish, J. Chapman, M. Aguilar, S. Taylor, M. Hill, M.D. Fairén, A. Hodges, B. Vellas, H. Soininen, I. Kloszewska, M. Daniilidou, J. Uphill, Y. Patel, J.T. Hughes, J. Lord, J. Turton, A.M. Hartmann, R. Cecchetti, C. Fenoglio, M. Serpente, M. Arcaro, C. Caltagirone, M.D. Orfei, A. Ciaramella, S. Pichler, M. Mayhaus, W. Gu, A. Lleó, J. Fortea, R. Blesa, I. S. Barber, K. Brookes, C. Cupidi, R.G. Maletta, D. Carrell, S. Sorbi, S. Moebus, M. Urbano, A. Pilotto, J. Kornhuber, P. Bosco, S. Todd, D. Craig, J. Johnston, M. Gill, B. Lawlor, A. Lynch, N.C. Fox, J. Hardy, A.R.U.K. Consortium, R.L. Albin, L. G. Apostolova, S.E. Arnold, S. Asthana, C.S. Atwood, C.T. Baldwin, L.L. Barnes, S. Barral, T.G. Beach, J.T. Becker, E.H. Bigio, T.D. Bird, B.F. Boeve, J.D. Bowen, A. Boxer, J.R. Burke, J.M. Burns, J.D. Buxbaum, N.J. Cairns, C. Cao, C.S. Carlson, C. M. Carlsson, R.M. Carney, M.M. Carrasquillo, S.L. Carroll, C.C. Diaz, H.C. Chui, D. G. Clark, D.H. Cribbs, E.A. Crocco, C. DeCarli, M. Dick, R. Duara, D.A. Evans, K. M. Faber, K.B. Fallon, D.W. Fardo, M.R. Farlow, S. Ferris, T.M. Foroud, D. R. Galasko, M. Gearing, D.H. Geschwind, J.R. Gilbert, N.R. Graff-Radford, R. C. Green, J.H. Growdon, R.L. Hamilton, L.E. Harrell, L.S. Honig, M.J. Huentelman, C.M. Hulette, B.T. Hyman, G.P. Jarvik, E. Abner, L.-W. Jin, G. Jun, A. Karydas, J. A. Kaye, R. Kim, N.W. Kowall, J.H. Kramer, F.M. LaFerla, J.J. Lah, J.B. Leverenz, A. I. Levey, G. Li, A.P. Lieberman, K.L. Lunetta, C.G. Lyketsos, D.C. Marson, F. Martiniuk, D.C. Mash, E. Masliah, W.C. McCormick, S.M. McCurry, A. N. McDavid, A.C. McKee, M. Mesulam, B.L. Miller, C.A. Miller, J.W. Miller, J. C. Morris, J.R. Murrell, A.J. Myers, S. O'Bryant, J.M. Olichney, V.S. Pankratz, J. E. Parisi, H.L. Paulson, W. Perry, E. Peskind, A. Pierce, W.W. Poon, H. Potter, J. F. Quinn, A. Raj, M. Raskind, B. Reisberg, C. Reitz, J.M. Ringman, E.D. Roberson, E. Rogaeva, H.J. Rosen, R.N. Rosenberg, M.A. Sager, A.J. Saykin, J.A. Schneider, L. S. Schneider, W.W. Seeley, A.G. Smith, J.A. Sonnen, S. Spina, R.A. Stern, R. H. Swerdlow, R.E. Tanzi, T.A. Thornton-Wells, J.Q. Trojanowski, J.C. Troncoso, V. M. Van Deerlin, L.J. Van Eldik, H.V. Vinters, J.P. Vonsattel, S. Weintraub, K. A. Welsh-Bohmer, K.C. Wilhelmsen, J. Williamson, T.S. Wingo, R.L. Woltjer, C. B. Wright, C.-E. Yu, L. Yu, F. Garzia, F. Golamaully, G. Septier, S. Engelborghs, R. Vandenberghe, P.P. De Deyn, C.M. Fernadez, Y.A. Benito, H. Thonberg, C. Forsell, L. Lilius, A. Kinhult-Stählbom, L. Kilander, R. Brundin, L. Concari, S. Helisalmi, A.M. Koivisto, A. Haapasalo, V. Dermecourt, N. Fievet, O. Hanon, C. Dufouil, A. Brice, K. Ritchie, B. Dubois, J.J. Himali, C.D. Keene, J. Tschanz, A. L. Fitzpatrick, W.A. Kukull, M. Norton, T. Aspelund, E.B. Larson, R. Munger, J. I. Rotter, R.B. Lipton, M.J. Bullido, A. Hofman, T.J. Montine, E. Coto, E. Boerwinkle, R.C. Petersen, V. Alvarez, F. Rivadeneira, E.M. Reiman, M. Gallo, C. J. O'Donnell, J.S. Reisch, A.C. Bruni, D.R. Royall, M. Dichgans, M. Sano, D. Galimberti, P. St George-Hyslop, E. Scarpini, D.W. Tsuang, M. Mancuso, U. Bonuccelli, A.R. Winslow, A. Daniele, C.-K. Wu, G.E.R.A.D. Perades, Charge, Adgc, Eadi, O. Peters, B. Nacmias, M. Riemenschneider, R. Heun, C. Brayne, D. C. Rubinsztein, J. Bras, R. Guerreiro, A. Al-Chalabi, C.E. Shaw, J. Collinge, D. Mann, M. Tsolaki, J. Clarimón, R. Sussams, S. Lovestone, M.C. O'Donovan, M. J. Owen, T.W. Behrens, S. Mead, A.M. Goate, A.G. Uitterlinden, C. Holmes, C. Cruchaga, M. Ingelsson, D.A. Bennett, J. Powell, T.E. Golde, C. Graff, P.L. De Jager, K. Morgan, N. Ertekin-Taner, O. Combarros, B.M. Psaty, P. Passmore, S. G. Younkin, C. Berr, V. Gudnason, D. Rujescu, D.W. Dickson, J.-F. Dartigues, A. L. DeStefano, S. Ortega-Cubero, H. Hakonarson, D. Campion, M. Boada, J. K. Kauwe, L.A. Farrer, C. Van Broeckhoven, M.A. Ikram, L. Jones, J.L. Haines, C. Tzourio, L.J. Launer, V. Escott-Price, R. Mayeux, J.-F. Deleuze, N. Amin, P. A. Holmans, M.A. Pericak-Vance, P. Amouyel, C.M. van Duijn, A. Ramirez, L.-S. Wang, J.-C. Lambert, S. Seshadri, J. Williams, G.D. Schellenberg, Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease, Nat. Genet. 49 (2017) 1373–1384, https://doi.org/10.1038/ng.3916.
[12] D. Ferreira, C. Verhagen, J.A. Hernández-Cabrera, L. Cavallin, C.-J. Guo, U. Ekman, J.-S. Muehlboeck, A. Simmons, J. Barroso, L.-O. Wahlund, E. Westman, Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications, Sci. Rep. 7 (2017) 46263, https://doi.org/10.1038/srep46263.
[13] R.A. Sperling, D.M. Rentz, K.A. Johnson, J. Karlawish, M. Donohue, D.P. Salmon, P. Aisen, The A4 study: stopping AD before symptoms begin? Sci. Transl. Med. 6 (2014) 228fs13, https://doi.org/10.1126/scitranslmed.3007941.
[14] C.W. Ritchie, J.L. Molinuevo, L. Truyen, A. Satlin, S. Van der Geyten, S. Lovestone, European prevention of Alzheimer's dementia (EPAD) consortium, development of interventions for the secondary prevention of Alzheimer's dementia: the European prevention of Alzheimer's dementia (EPAD) project, Lancet Psychiatry 3 (2016) 179–186, https://doi.org/10.1016/S2215-0366(15)00454-X.
[15] D.R. Bateman, B. Srinivas, T.W. Emmett, T.K. Schleyer, R.J. Holden, H.C. Hendrie, C.M. Callahan, Categorizing health outcomes and efficacy of mHealth apps for persons with cognitive impairment: a systematic review, J. Med. Internet Res. 19 (2017) e301, https://doi.org/10.2196/jmir.7814.
[16] H.C. Kales, L.N. Gitlin, C.G. Lyketsos, Assessment and management of behavioral and psychological symptoms of dementia, Br. Med. J. 350 (2015) h369, https://doi.org/10.1136/bmj.h369.
[17] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.
[18] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906, https://doi.org/10.1093/nar/gkm958.
[19] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) D267–D270, https://doi.org/10.1093/nar/gkh061.
[20] D.S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S.L. Chen, D. Hadley, A. Green, P. Khankhanian, S.E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, Elife 6 (2017) e26726, https://doi.org/10.7554/eLife.26726.
[21] A. Messina, H. Pribadi, J. Stichbury, M. Bucci, S. Klarman, A. Urso, BioGrakn: a knowledge graph-based semantic database for biomedical sciences, in: L. Barolli, O. Terzo (Eds.), Complex Intell. Softw. Intensive Syst., Springer International Publishing, Cham, 2018, pp. 299–309, https://doi.org/10.1007/978-3-319-61566-0_28.
[22] S. Mizuno, R. Iijima, S. Ogishima, M. Kikuchi, Y. Matsuoka, S. Ghosh, T. Miyamoto, A. Miyashita, R. Kuwano, H. Tanaka, AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease, BMC Syst. Biol. 6 (2012) 52, https://doi.org/10.1186/1752-0509-6-52.
[23] S. Ogishima, S. Mizuno, M. Kikuchi, A. Miyashita, R. Kuwano, H. Tanaka, J. Nakaya, AlzPathway, an updated map of curated signaling pathways: towards deciphering Alzheimer's disease pathogenesis, Methods Mol. Biol. Clifton NJ 1303 (2016) 423–432, https://doi.org/10.1007/978-1-4939-2627-5_25.
[24] 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, Alzheimers Dement, J. Alzheimers Assoc. 10 (2014) 238–246, https://doi.org/10.1016/j.jalz.2013.02.009.
[25] Y. Zhou, J. Fang, L.M. Bekris, Y.H. Kim, A.A. Pieper, J.B. Leverenz, J. Cummings, F. Cheng, AlzGPS: a genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery, Alzheimers Res. Ther. 13 (2021) 24, https://doi.org/10.1186/s13195-020-00760-w.
[26] Y. Zhou, J. Xu, Y. Hou, L. Bekris, J.B. Leverenz, A.A. Pieper, J. Cummings, F. Cheng, The Alzheimer's Cell Atlas (TACA): a single-cell molecular map for translational therapeutics accelerator in Alzheimer's disease, Alzheimers Dement, Transl. Res. Clin. Interv. 8 (2022) e12350, https://doi.org/10.1002/trc2.12350.
[27] J.D. Romano, V. Truong, R. Kumar, M. Venkatesan, B.E. Graham, Y. Hao, N. Matsumoto, X. Li, Z. Wang, M.D. Ritchie, L. Shen, J.H. Moore, The Alzheimer's knowledge base: a knowledge graph for alzheimer disease research, J. Med. Internet Res. 26 (2024) e46777, https://doi.org/10.2196/46777.
[28] C. Zhu, Z. Yang, X. Xia, N. Li, F. Zhong, L. Liu, Multimodal reasoning based on knowledge graph embedding for specific diseases, Bioinformatics 38 (2022) 2235–2245, https://doi.org/10.1093/bioinformatics/btac085.
[29] Y. Nian, X. Hu, R. Zhang, J. Feng, J. Du, F. Li, L. Bu, Y. Zhang, Y. Chen, C. Tao, Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing, BMC Bioinf. 23 (2022) 407, https://doi.org/10.1186/s12859-022-04934-1.
[30] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T.C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (2012) 3158–3160, https://doi.org/10.1093/bioinformatics/bts591.
[31] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinf. 21 (2020) 188, https://doi.org/10.1186/s12859-020-3517-7.
[32] Z. Li, H. Liu, Z. Zhang, T. Liu, N.N. Xiong, Learning knowledge graph embedding with heterogeneous relation attention networks, IEEE Trans. Neural Netw. Learn. Syst. 33 (2022) 3961–3973, https://doi.org/10.1109/TNNLS.2021.3055147.
[33] Z. Li, H. Liu, Z. Zhang, T. Liu, J. Shu, Recalibration convolutional networks for learning interaction knowledge graph embedding, Neurocomputing 427 (2021) 118–130, https://doi.org/10.1016/j.neucom.2020.07.137.
[34] Z. Xue, Z. Zhang, H. Liu, S. Yang, S. Han, Learning knowledge graph embedding with multi-granularity relational augmentation network, Expert Syst. Appl. 233 (2023), https://doi.org/10.1016/j.eswa.2023.120953.
[35] H. Baalbaki, H. Hazimeh, H. Harb, R. Angarita, KEMA++: a full representative knowledge-graph embedding model (036), Int. J. Software Eng. Knowl. Eng. 32 (2022) 1619–1641, https://doi.org/10.1142/S0218194022500760.
[36] H. Baalbaki, H. Hazimeh, H. Harb, R. Angarita, TransModE: translational knowledge graph embedding using modular arithmetic, Procedia Comput. Sci. 207 (2022) 1154–1163, https://doi.org/10.1016/j.procs.2022.09.171.
[37] D. Bang, S. Lim, S. Lee, S. Kim, Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers, Nat. Commun. 14 (2023) 3570, https://doi.org/10.1038/s41467-023-39301-y.
[38] A. Fader, S. Soderland, O. Etzioni, Identifying relations for open information extraction, in: R. Barzilay, M. Johnson (Eds.), Proc. 2011 Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 1535–1545. https://aclanthology.org/D11-1142. (Accessed 14 November 2023).
[39] Mausam, M. Schmitz, S. Soderland, R. Bart, O. Etzioni, Open Language learning for information extraction, in: J. Tsujii, J. Henderson, M. Paşca (Eds.), Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn., Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 523–534. https://aclanthology.org/D12-1048. (Accessed 14 November 2023).
[40] G. Angeli, M.J. Johnson Premkumar, C.D. Manning, Leveraging linguistic structure for open domain information extraction, in: C. Zong, M. Strube (Eds.), Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Vol. 1 Long Pap., Association for Computational Linguistics, Beijing, China, 2015, pp. 344–354, https://doi.org/10.3115/v1/P15-1034.
[41] L. Del Corro, R. Gemulla, ClausIE: clause-based open information extraction, in: Proc. 22nd Int. Conf. World Wide Web, Association for Computing Machinery, New York, NY, USA, 2013, pp. 355–366, https://doi.org/10.1145/2488388.2488420.
[42] Y. Luan, L. He, M. Ostendorf, H. Hajishirzi, Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proc. 2018 Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3219–3232, https://doi.org/10.18653/v1/D18-1360.
[43] M. Eberts, A. Ulges, Span-based joint entity and relation extraction with transformer pre-training, Santiago Compost (2020).
[44] L. Soldaini, N. Goharian, QuickUMLS: a Fast, Unsupervised Approach for Medical Concept Extraction, (n.d.)..
[45] R. Leaman, Z. Lu, TaggerOne: joint named entity recognition and normalization with semi-Markov Models, Bioinforma. Oxf. Engl. 32 (2016) 2839–2846, https://doi.org/10.1093/bioinformatics/btw343.
[46] OpenAI, GPT-4 technical report. http://arxiv.org/abs/2303.08774, 2023. (Accessed 21 November 2023).
[47] G.R. Brown, V. Hem, K.S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K.D. Pruitt, D.R. Maglott, T.D. Murphy, Gene: a gene-centered information resource at NCBI, Nucleic Acids Res. 43 (2015) D36–D42, https://doi.org/10.1093/nar/gku1055.
[48] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, ChEBI in 2016: improved services and an expanding collection of metabolites, Nucleic Acids Res. 44 (2016) D1214–D1219, https://doi.org/10.1093/nar/gkv1031.
[49] D. Kim, J. Lee, C.H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, J. Kang, A neural named entity recognition and multi-type normalization tool for biomedical text mining, IEEE Access 7 (2019) 73729–73740, https://doi.org/10.1109/ACCESS.2019.2920708.
[50] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for NLP-assisted text annotation, in: F. Segond (Ed.), Proc. Demonstr. 13th Conf. Eur. Chapter Assoc. Comput. Linguist., Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107. https://aclanthology.org/E12-2021. (Accessed 14 November 2023).
[51] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A survey on knowledge graphs: representation, acquisition and applications, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–21, https://doi.org/10.1109/TNNLS.2021.3070843.
[52] I. Beltagy, K. Lo, A. Cohan, SciBERT: a pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proc. 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620, https://doi.org/10.18653/v1/D19-1371.
[53] UniProt Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Res. 51 (2023) D523–D531, https://doi.org/10.1093/nar/gkac1052.
[54] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082, https://doi.org/10.1093/nar/gkx1037.
[55] S. Köhler, L. Carmody, N. Vasilevsky, J.O.B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N.L. Harris, N. Matentzoglu, J.A. McMurry, D. Osumi-Sutherland, V. Cipriani, J.P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A.C. Jansen, A. Muaz, W.H. Chang, J. Bergerson, S.J. F. Lauledeerkind, Z. Yüksel, S. Beltran, A.F. Freeman, P.I. Sergouniotis, D. Durkin, A. L. Storm, M. Hanauer, M. Brudno, S.M. Bello, M. Sincan, K. Rageth, M.T. Wheeler, R. Oegema, H. Lourghi, M.G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R.C. Lovering, C. Hajek, A. Olry, L. Notarangelo, M. Similuk, X.A. Zhang, D. Gómez-Andrés, H. Lochmüller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J.D. Milner, D. Leroux, C.F. Boerkoel, A. Klion, M.C. Carter, T. Groza, D. Smedley, M. A. Haendel, C. Mungall, P.N. Robinson, Expansion of the human phenotype ontology (HPO) knowledge base and resources, Nucleic Acids Res. 47 (2019) D1018–D1027, https://doi.org/10.1093/nar/gky1105.
[56] L.M. Schriml, R. Lichenstein, K. Bisordi, C. Bearer, J.A. Baron, C. Greene, Modeling the enigma of complex disease etiology, J. Transl. Med. 21 (2023) 148, https://doi.org/10.1186/s12967-023-03987-x.
[57] M.J. Landrum, J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, D. R. Maglott, ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Res. 42 (2014) D980–D985, https://doi.org/10.1093/nar/gkt1113.
[58] M.J. Landrum, J.M. Lee, M. Benson, G.R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J.B. Holmes, B. L. Kattman, D.R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (2018) D1062–D1067, https://doi.org/10.1093/nar/gkx1153.
[59] F.B. Rogers, Medical subject headings, Bull. Med. Libr. Assoc. 51 (1963) 114–116.
[60] N. Okazaki, J. Tsujii, Simple and Efficient Algorithm for Approximate Dictionary Matching, (n.d.)..
[61] P. Cimiano, H. Paulheim, Knowledge graph refinement: a survey of approaches and evaluation methods, Semant. Web 8 (2017) 489–508, https://doi.org/10.3233/SW-160218.
[62] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L.I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. 45 (2017) D833–D839, https://doi.org/10.1093/nar/gkw943.
[63] L. Gong, M. Whirl-Carrillo, T.E. Klein, PharmGKB, an integrated resource of pharmacogenomic knowledge, Curr. Protoc. 1 (2021) e226, https://doi.org/10.1002/cpz1.226.
[64] J.S. Amberger, C.A. Bocchini, F. Schiettecatte, A.F. Scott, A. Hamosh, OMIM.org: online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders, Nucleic Acids Res. 43 (2015) D789–D798, https://doi.org/10.1093/nar/gku1205.
[65] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N.T. Doncheva, S. Pyysalo, P. Bork, L.J. Jensen, C. von Mering, The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (2022) D638–D646, https://doi.org/10.1093/nar/gkac1000.
[66] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 2013, in: https://proceedings.neurips.cc/paper_files/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html. (Accessed 14 November 2023).
[67] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, Proc. AAAI Conf. Artif. Intell. 28 (2014), https://doi.org/10.1609/aaai.v28i1.8870.
[68] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion, Proc. AAAI Conf. Artif. Intell. 29 (2015), https://doi.org/10.1609/aaai.v29i1.9491.
[69] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: Proc. 33rd Int. Conf. Mach. Learn., PMLR, 2016,

pp. 2071–2080, in: https://proceedings.mlr.press/v48/trouillon16.html. (Accessed 21 November 2023).

[70] D.Q. Nguyen, T.D. Nguyen, D.Q. Nguyen, D. Phung, A novel embedding model for knowledge base completion based on convolutional neural network, in: M. Walker, H. Ji, A. Stent (Eds.), Proc. 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Vol. 2 Short Pap, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 327–333, https://doi.org/10.18653/v1/N18-2053.

[71] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, PLoS Med. 12 (2015) e1001779, https://doi.org/10.1371/journal.pmed.1001779.

[72] B.B. Sun, J. Chiou, M. Traylor, C. Benner, Y.-H. Hsu, T.G. Richardson, P. Surendran, A. Mahajan, C. Robins, S.G. Vasquez-Grinnell, L. Hou, E.M. Kvikstad, O.S. Burren, J. Davitte, K.L. Ferber, C.E. Gillies, Å.K. Hedman, S. Hu, T. Lin, R. Mikkilineni, R. K. Pendergrass, C. Pickering, B. Prins, D. Baird, C.-Y. Chen, L.D. Ward, A. M. Deaton, S. Welsh, C.M. Willis, N. Lehner, M. Arnold, M.A. Wörheide, K. Suhre, G. Kastenmüller, A. Sethi, M. Cule, A. Raj, Alnylam Human Genetics, AstraZeneca Genomics Initiative, Biogen Biobank Team, Bristol Myers Squibb, Genentech Human Genetics, GlaxoSmithKline Genomic Sciences, Pfizer Integrative Biology, Population Analytics of Janssen Data Sciences, Regeneron Genetics Center, L. Burkitt-Gray, E. Melamud, M.H. Black, E.B. Fauman, J.M.M. Howson, H.M. Kang, M.I. McCarthy, P. Nioi, S. Petrovski, R.A. Scott, E.N. Smith, S. Szalma, D.M. Waterworth, L.J. Mitnaul, J.D. Szustakowski, B.W. Gibson, M.R. Miller, C.D. Whelan, Plasma proteomic associations with genetics and health in the UK Biobank, Nature 622 (2023) 329–338, https://doi.org/10.1038/s41586-023-06592-6.

[73] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Min. Knowl. Discov. 28 (2014) 92–122, https://doi.org/10.1007/s10618-012-0295-5.

[74] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[75] List of AD Loci and Genes with Genetic Evidence Compiled by ADSP Gene Verification Committee – ADSP, (n.d.). https://adsp.niagads.org/gvc-top-hits-list/(accessed November 21, 2023).