Article

A Knowledge Graph Framework for Dementia Research Data

Santiago Timón-Reina [1,2,*], Mariano Rincón [3] , Rafael Martínez-Tomás [3] , Bjørn-Eivind Kirsebom [4,5] and Tormod Fladby [2,6]

[1] Escuela Internacional de Doctorado—Doctorado en Sistemas Inteligentes,
Universidad Nacional de Educación a Distancia (UNED), 28015 Madrid, Spain
[2] Department of Neurology, Akershus University Hospital, 1478 Nordbyhagen, Norway
[3] Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia (UNED),
28015 Madrid, Spain
[4] Department of Neurology, University Hospital of North Norway, 9019 Tromsø, Norway
[5] Department of Psychology, Faculty of Health Sciences, UiT The Arctic University of Norway,
9019 Tromsø, Norway
[6] Institute of Clinical Medicine, Campus Ahus, University of Oslo, 0313 Oslo, Norway
* Correspondence: stimon3@alumno.uned.es

Featured Application: Applying knowledge graphs, graph analytics, and graph machine learning for integrating multi-modal dementia research data.

Abstract: Dementia disease research encompasses diverse data modalities, including advanced imaging, deep phenotyping, and multi-omics analysis. However, integrating these disparate data sources has historically posed a significant challenge, obstructing the unification and comprehensive analysis of collected information. In recent years, knowledge graphs have emerged as a powerful tool to address such integration issues by enabling the consolidation of heterogeneous data sources into a structured, interconnected network of knowledge. In this context, we introduce DemKG, an open-source framework designed to facilitate the construction of a knowledge graph integrating dementia research data, comprising three core components: a KG-builder that integrates diverse domain ontologies and data annotations, an extensions ontology providing necessary terms tailored for dementia research, and a versatile transformation module for incorporating study data. In contrast with other current solutions, our framework provides a stable foundation by leveraging established ontologies and community standards and simplifies study data integration while delivering solid ontology design patterns, broadening its usability. Furthermore, the modular approach of its components enhances flexibility and scalability. We showcase how DemKG might aid and improve multi-modal data investigations through a series of proof-of-concept scenarios focused on relevant Alzheimer's disease biomarkers.

Keywords: knowledge graphs; ontologies; graph databases; data modeling; dementia; omics

1. Introduction

The dawn of "omics" technologies, accompanied by advancements in imaging, clinical data collection, laboratory testing, and phenotyping, has profoundly influenced biomedical research [1–7]. This multi-modal setting has provided an unprecedented, comprehensive view of complex biological systems, thereby inspiring a shift towards a more integrated understanding of diseases. However, the introduction of data from diverse modalities also presents unique challenges. Effectively integrating and interpreting the sheer volume, complexity, and diversity of data generated by these sources requires sophisticated computational tools. Moreover, the data, which are often distributed across various databases, publications, and repositories, pose considerable barriers to seamless data integration. Even more daunting is the task of transforming multi-modal data into clinically actionable insights, requiring the ability to connect data from molecular to clinical scales, a feat

complicated by the enormous diversity and complexity of individual diseases. These hurdles highlight the need for innovative strategies and tools to harness the potential of multi-modal data in propelling the field of precision medicine.

Since biological reality is often modeled as a network or graph [8,9], one technological approach that has gained significant traction is the use of knowledge graphs (KGs) [10], which allow for the integration and organization of diverse biomedical data types, facilitating their analysis and interpretation.

After Google introduced the knowledge graph in 2012, highlighting the advantages of the approach [11], KGs have become increasingly popular, finding adoption in industry with subsequent launches by companies such as Microsoft, Amazon, Airbnb, and Facebook [12], as well as in academia [13,14]. Nonetheless, the definition of KGs can vary based on the application context. In biomedicine, they can be characterized as data structures meant to gather and disseminate real-world knowledge, where nodes depict significant biomedical entities and the edges delineate diverse relationships that could exist between these entities [15]. KGs embody a methodological transition toward a more comprehensive representation of reality, facilitating the integration of heterogeneous data types and providing an intuitive, graph-based structure for representing intricate relationships between diverse biomedical entities.

Constructing a KG entails a series of methodological and technological decisions that profoundly impact the utility and effectiveness of the resulting product. A pivotal consideration in this process is the selection of a graph paradigm, which provides the theoretical and practical foundation for the structure and function of the KG. There are two primary approaches in this regard: Resource Description Framework (RDF) and Labeled Property Graphs [16–18]. Both of these approaches offer robust technological solutions, but each has its own strengths and weaknesses. While RDF offers standardization and robustness ideal for semantic applications, it may suffer from verbosity and computational inefficiency. Conversely, LPGs excel in their flexibility and intuitive structure, which allow for the straightforward representation of complex relationships and properties on both nodes and edges, but they may struggle in scenarios demanding high interoperability and standardization. Thus, the choice often hinges on the specific project requirements and constraints.

In addition to choosing a graph paradigm, selecting a data model or graph schema is another critical decision for building a KG. This model dictates how entities of interest and their relationships are represented within the KG. This aspect can be approached in two main ways: using an ad hoc data model tailored to the project's specific needs or adopting a standard model such as ontologies. In particular, biomedical ontologies have emerged as essential tools in standardizing terminology, modeling biological realities [19], supporting data annotation [20–23], and facilitating biomedical text mining [24,25]. With ongoing concerted efforts from the scientific community, these ontologies have evolved to incorporate fine-grained knowledge across various biomedical subdomains, as exemplified by initiatives such as the Open Biological and Biomedical Ontologies (OBO) Foundry [26] and the National Center for Biomedical Ontology (NCBO) [27] and its BioPortal [28]. Moreover, using logical modeling and annotation, biomedical ontologies make assertions that span and connect levels of biological organization, from the molecular level to phenotype and disease definitions. This ability to traverse and link multiple scales of biological information makes ontologies an invaluable resource for the construction of KGs for biomedical research.

The biomedical field is rich in open databases that offer scientific knowledge from various subdomains, including molecular biology (genomics, proteomics, and pathways), drugs, and disease characterization. These sources hold the potential for a more comprehensive understanding of biomedical phenomena; however, their value is often hindered by their dispersal across different platforms. KGs have emerged as instrumental tools for integrating and exploiting these disparate sources, fostering a multitude of projects that aim to unify the spread-out biomedical knowledge.

A prime example of such an initiative is the Monarch Initiative [29], which integrates genetic, phenotypic, and disease-related data to facilitate the identification of disease genes and variants. Similarly, the Clinical Knowledge Graph (CKG) [30] is an open-source platform that integrates proteomics, public databases, and literature. It effectively utilizes KGs to augment and enrich biomedical data, thereby facilitating informed clinical decision-making. Likewise, PrimeKG [31] is a multimodal KG that integrates a multitude of high-quality resources, representing various biological scales, i.e., from genotypes to clinical phenotypes. The scalable precision medicine open knowledge engine (SPOKE) [32] also integrates multiple biological data sources to provide structured knowledge ranging from low-level molecular biology to pharmacology and clinical practice. Furthermore, the KG-COVID-19 [33] project responded to the COVID-19 crisis by building a unified KG from disparate biomedical information about SARS-CoV-2, illustrating how KGs can effectively drive knowledge synthesis, particularly in emergent health situations.

As the number of available KGs increases, it has become evident that social and technical limitations exist, especially the need for standardization in entity naming and graph representation approaches [34,35]. Regarding modeling standardization, the Biolink Model [36] has emerged as a high-level data model that provides standard terms and relations for describing biological entities and their relationships for organizing data in biomedical KGs. Biolink serves both as a map for bringing together data from different sources under one unified model and as a bridge between ontological domains. As a similar initiative to OBO, centered around KGs, the KG-Hub project [37] provides a collection of tools and libraries for building interoperable KGs and a mechanism for sharing them to foster their reuse.

In addition to their ability to model and query data, graph analytics and graph machine learning techniques have made notable advancements [38,39], supported by open-source libraries such as GRAPE [40] and KGTK [41]. One technique particularly relevant in the biomedical domain is graph embedding [42–47], which allows us to capture complex graph structures into lower-dimension vectors. Exploiting these features to integrate specific patient data with large biomedical KGs has already shown promising results in deriving actionable clinical outputs, as evidenced by advancements in understanding diseases such as multiple sclerosis [42] and Alzheimer's disease [48]. Recent dementia research uses multi-modal data to understand the condition from various aspects, including genomics, transcriptomics, metabolites, imaging, and clinical features. Having a framework that enables the systematic construction and instantiation of research and clinical data in a standardized manner offers significant benefits.

This paper introduces DemKG, a KG framework designed specifically for dementia research needs. The framework leverages reference ontologies from OBO, standard KG technologies from KG-Hub, and an instantiation tooling to transform source data into the KG following sound design patterns within the ontological model. DemKG reuses most of its knowledge sources, provides specific terminological extensions to cover gaps identified in the scope of dementia, and ingests biological databases of interest, resulting in an integrative KG that covers the multiple data modalities involved in the research, including genomics, proteomics, imaging, fine-grained phenotyping, and clinical tests. Thanks to its design, DemKG is easily extensible, delivering means to customize and deploy in modern graph databases for enhanced data querying and retrieval. The expressive knowledge model supports advanced analytics through graph and network algorithms, which play an active role in the progression of research and better patient care through the implementation of precision medicine.

2. Related Work

Advancements in storage and graph technologies, coupled with the increasing availability of open scientific data, have led to the emergence of multiple biological KGs [49]. Projects such as the Monarch Integrated Knowledge Graph, the Clinical Knowledge Graph

(CKG), PrimeKG, and the scalable precision medicine open knowledge engine (SPOKE), previously introduced in the introduction, bear similarities to our initiative.

The Monarch Integrated Knowledge Graph [29] is a notable example of biological KGs, which assimilates various data types (including genotype, phenotype, and disease) from multiple sources into a unified semantic graph model. The Monarch KG has been instrumental in our project, DemKG, as it not only serves as a primary data source but also offers an array of tools we utilize. Our philosophy aligns closely with that of the Monarch KG, emphasizing a robust semantic foundation while integrating data from a variety of external sources, including other ontologies and extensions. We build upon this work to extend it with dementia-related knowledge and provide means for integrating study data.

CKG [30] is an open-source platform designed to harmonize a wide range of "omics" data types into a coherent structure, including genomics, transcriptomics, proteomics, and metabolomics. CKG favors a custom data model formed from a selected set of concepts and relationships from specific ontologies. On top of the KG, CKG integrates statistical and machine learning algorithms to streamline the analysis and interpretation of typical proteomics workflows. DemKG resonates with CKG's mission to improve the modeling and integration of omics data. However, it deviates fundamentally from its approach to data modeling, wherein CKG employs a more circumscribed model.

PrimeKG [31] is a multimodal KG for precision medicine analyses. Like its counterparts, it integrates a plethora of resources to describe a broad spectrum of diseases with relationships across major biological scales. One of them is combining the entire range of approved drugs with their therapeutic action, distinguishing it from other systems. Moreover, unlike DemKG, PrimeKG employs a custom approach to its data model, incorporating ten types of nodes and thirty types of undirected edges extracted from reference ontologies. Furthermore, it lacks a systematic schema to integrate experimental and study data.

SPOKE [32] is a KG that connects information from 41 biological data sources, structured as 21 different node types and 55 edge types, ranging from low-level molecular biology to pharmacology and clinical practice. It uses 11 different ontologies to organize the data semantically meaningfully and, in its last iteration, also integrates the Biolink model whenever it is found to be practical. SPOKE is implemented as a Neo4j database built from a collection of Python scripts and provides a graphical user interface and a REST API for end-user access. Our method stands distinct from SPOKE in several crucial aspects. Primarily, it offers an open toolkit for KG construction and personalization, ensuring both platform and representational paradigm autonomy. Moreover, despite utilizing a comparable modeling approach, DemKG fosters a closer connection with a vast array of domain ontologies by preserving links to explicitly defined terms and relationships. Finally, our framework provides a flexible and robust module for research data integration.

In summary, our work distinguishes itself from similar efforts through a comprehensive approach that integrates a well-established terminological foundation and community standards, follows design patterns conducive to data integration, and defines terminological extensions specific to the dementia domain, facilitated through a dedicated low-code solution for seamless study data integration.

3. Materials and Methods

3.1. Terminological Foundation

In the construction of the knowledge graph, the initial and pivotal decision revolves around selecting an appropriate graph schema to provide a solid conceptual base that effectively captures data entities drawn from the array of biological subdomains pertinent to dementia research. This choice presents a dichotomy: one option involves creating a flexible, ad hoc schema tailored to the identified needs, while the alternative entails adopting a more structured strategy that employs standard terminologies and ontologies. Our methodology aligns with the latter approach, and a fundamental design principle in the construction of our KG is the utilization of domain reference ontologies to ensure the following:

1. The concept definitions are concise, accurate, and relevant;
2. There exists an active community keeping the ontology updated;
3. They are widely recognized, cross-referenced, and follow consistent design patterns.

The criteria set forth are congruent with the guiding principles of the OBO foundry. OBO endorses an extensive range of domain-specific ontologies that are distinguished by well-demarcated scopes, the reutilization of concepts across ontologies, and alignment with a unified upper-level model, specifically the Basic Formal Ontology (BFO) [50], and relations are defined in the Relations Ontology (RO). Given these attributes, we gave preferential consideration to OBO ontologies during our selection process.

As the KG must cater to a variety of domains, adopting this approach enables us to concentrate mainly on integration and only define new terms when detecting a gap. Some notable examples of the employed OBO ontologies include the Gene Ontology [51,52], Chemical Entities of Biological Interest (CHEBI) [53], and Protein Ontology (PR) [54] for the genetic and molecular domain. For the phenotype and disease domain, we utilize the Monarch Disease Ontology (MONDO) [55], Human Phenotype Ontology (HP) [56,57], and Phenotype And Trait Ontology (PATO) [58]. In the area of anatomy, we incorporate the Uber-Anatomy Ontology (UBERON) [59,60] and the Foundational Model of Anatomy (FMA) [61]. For neuropsychological tests and their relations, we include the Neuropsychological Testing Ontology (NPT) [62] and the Neurocognitive Integrated Ontology (NIO) [63]. For modeling experimental settings, the Ontology for Biomedical Investigations (OBI) [64,65] plays a central role.

These ontologies provide a significant level of detail, and reusing or referencing concepts between them expands the knowledge network, facilitating the exploitation of multi-domain and multi-level relations. For example, this interconnectedness simplifies navigation from HP phenotypes referenced in a disease definition in MONDO to specific genes in GO, proteins in PR, and molecular entities in CHEBI. Furthermore, we also include relevant Monarch data and annotation ingestions; specifically, gene and gene-phenotype annotations, filtered protein–protein interactions from the STRING database [66], and pathway knowledge from the Reactome pathway knowledgebase [67]. The complete list of knowledge sources and annotations is listed in Table 1.

Table 1. List of DemKG knowledge sources.

Source	Source Identifier	Reference
Basic Formal Ontology	BFO	[50]
Biolink model	biolink	[36]
Chemical Entities of Biological Interest	CHEBI	[53]
Cell Ontology	CL	[68]
Evidence and Conclusion Ontology	ECO	[69]
Environmental Factor Ontology	EFO	[70]
Gene Ontology	GO	[52]
Gene Ontology Annotations	GOA	-
Human Phenotype Ontology	HP	[57]
Human Phenotype Ontology Annotations	HPOA	-
Information Artifact Ontology	IAO	-

Table 1. *Cont.*

Source	Source Identifier	Reference
Mass Spectrometry Ontology	MS	[71]
Mondo Disease Ontology	MONDO	[55]
Monarch KG	Monarch	[29]
Neurocognitive Integrated Ontology	NIO	[63]
Neuropsychological Testing Ontology	NPT	[62]
Ontology of Biological Attributes	OBA	[72]
Ontology for Biomedical Investigations	OBI	[65]
Ontology for General Medical Science	OGMS	[73]
Ontology of Medically Related Social Entities	OMRSE	[74]
Phenotype And Trait Ontology	PATO	[58]
Phenomics Integrated Ontology	PHENIO	-
Protein Ontology	PR	[54]
Relations Ontology	RO	-
Reactome	Reactome	[67]
Scientific Evidence and Provenance Information Ontology	SEPIO	-
STRING database ingestion	STRING	[66]
Uber Anatomy Ontology	UBERON	[59]

While the standardization offered by domain ontologies is undoubtedly a strength, it can also impose limitations due to the inherent trade-off with flexibility. This high level of detail can complicate the integration of non-OBO ontologies and external datasets. Additionally, querying the graph requires a comprehensive understanding of the underlying model. We employ the Biolink model as our high-level data model to mitigate these issues. Biolink offers a means to utilize higher-level concepts from its "category" hierarchy while still allowing references to more specific ontology terms. The same versatility is available for relationships through the use of the "related_to" hierarchy, thus providing a balance between standardization and flexibility in our knowledge graph.

3.2. Terminological Extensions

OBO covers most of the conceptualization needs, but gaps remain relevant to the implementation. To overcome this issue, we implement an application ontology that is also one of the inputs of the merging process. The primary interventions relate to phenotypic normality, as well as to the necessary assay and platform definitions missing from OBI.

HP and MONDO thoroughly model disease states, conditions, and abnormal phenotypes, leaving out any reference to normal counterparts. To allow the categorization of instances of normal/healthy cases, we introduced a "Phenotypic normality" hierarchy. This new hierarchy is modeled as a sibling branch of the HP "Phenotypic abnormality", mirroring its hierarchy to allocate the "normality" concepts of interests.

In dementia research, the utilization of neuropsychological assessments such as the Mini-Mental State Examination (MMSE) [75], the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) wordlist memory test (WLT) [76], Visual Object and Space Perception (VOSP) battery [77], Trail Making Test (TMT) [78], Clock Drawing Test [79], and Controlled Oral Word Association Test (COWAT-FAS) [80] is instrumental in quantifying cognitive function domains and tracking disease progression. We have implemented the necessary concepts to cover CERAD, VOSP, and COWAT-FAS tests, with the primary classes

allocated under the "cognitive function assay" branch of NPT, while also relating to the mental and cognitive functions they assess.

The AT(N) classification system [81] is another tool of great importance for assessing the subject's biological state and understanding the intricate relationships between key biomarkers and their impact on disease evolution. AT(N) categorizes biomarkers according to their role in the disease progression, namely, Beta-amyloid deposition (A), pathologic tau (T), and neurodegeneration (N). Within each biomarker category, values can be positive or negative $(+/-)$, derived from defined normal or abnormal cut points, resulting in the creation of eight distinct AT(N) "biomarker profiles" (Table 2). To provide proper terminological coverage, we have defined new classes for each biomarker profile and phenotype terms related to abnormal CSF protein concentration phenotypes related to phosphorylated tau (P-tau) and total tau (T-tau) missing from HP. Each biomarker profile is defined under the "value specification" class from OBI, with asserted logical axioms to associate them with the specific phenotype.

Table 2. AT(N) biomarker profiles and categories as defined by the NIA-AA Research Framework. Each biomarker profile is modeled as a descendant of the "value specification" class defined in OBI.

AT(N) Profiles	Biomarker Category	
A-T-(N)-	Normal AD biomarkers	
A-+T-(N)-	Alzheimer's pathologic change	
A+T+(N)-	Alzheimer's disease	
A+T+(N)+	Alzheimer's disease	Alzheimer's continuum
A+T-(N)+	Alzheimer's and concomitant suspected non-Alzheimer's pathologic change	
A-T+(N)-	Non-AD pathologic change	
A-T-(N)+	Non-AD pathologic change	
A-T+(N)+	Non-AD pathologic change	

3.3. Technical Implementation

The implementation consists of three main software pieces covering different parts of the KG generation, integrated into a building pipeline: the extensions ontology builder, the KG-builder, and the data transformer module. To maximize effectiveness and reproducibility, in all three sub-projects, we employ state-of-the-art ontology and graph tooling maintained by the community and relevant projects such as Monarch and the "universal biomedical data translator" from the National Center for Advancing Translational Sciences (NCATS) [82].

The extensions ontology builder produces an OWL ontology using the Ontology Development Kit (ODK) v1.4.1 [83] as the building framework. The ODK provides a pre-configured, standardized environment with a set of tools that support all stages of the ontology lifecycle (creation and editing, building, and testing, and releasing with version control) and ensures a systematic approach to ontology maintenance. When possible, we define new classes that follow a pattern using the Dead Simple OWL Design Patterns (DOS-DP) v0.1.10 [84], reducing manual editing and consequently reducing errors and improving reproducibility. All the axioms are kept under OWL2 [85] DL profile.

The KG-builder is responsible for obtaining the different sources of knowledge and merging them into the terminological KG. Built upon the KG-Hub tooling ecosystem, the main configuration inputs are the merge and download YAML descriptor files, guiding the download and merge steps. When available, the ontologies are downloaded from the KG-Hub repository [86]. OBO ontologies are already maintained as Biolink-compliant graphs in the Knowledge eXchange Format (KGX) [87] in the KG-OBO project [88] and are directly merged from each specific release artifact. The merging step includes all downloaded sources and the extensions ontology to obtain a final KGX graph.

One challenge when converting OWL ontologies into a graph structure lies in the difficulty of accessing class relationships established through subclass and class equivalence axioms. These assertions hold significant value in capturing the biomedical knowledge outlined in the comprehensive OBO ontologies. To address this situation, both the ontology and builder modules materialize class equivalence axioms. In the context of the extensions ontology, we utilize the relation-graph [89] library during the later stages of the construction process. In the case of OBO ontologies, the KG-builder retrieves a subset of links from the materialization output within Ubergraph [90], which also employs relation-graph.

The transformer module is a Python solution that provides an accessible approach to generating graph data in KGX format from tabular source input. This module adopts a YAML-based transform definition schema, mirroring the approach of other tools in the pipeline. This schema adheres to a standardized structure wherein users can define mappings from columns to specific classes paired with various instantiation design patterns. The schema effectively models common research entities, including medical history, physical examination, and measurement assays, all aligned with dedicated instantiation patterns that are further elaborated upon in the subsequent subsection.

The builder pipeline integrates all steps and can be configured to generate two artifacts: solely the terminological graph or the terminological graph with data instantiation.

3.4. Data Transformation Design Patterns

One of the aims of the KG is to integrate raw research data to enable explicit connections with knowledge concepts. We propose a set of design patterns to support the data instantiation of patient/subject study visits, phenotype observations arising from these visits, measurements/analyses derived from samples collected from different specimens, and neuropsychological test results. In all these patterns, OBI is the central ontology employed to enable the relating of clinical and research concepts with specific entities of the biomedical domain. Figures 1–3 illustrate the main patterns through simplified concept map figures, depicting the main ontology classes and properties involved, identified with a pseudo-CURIE of the format PREFIX, namely, "class label", where prefix is the OBO ontology prefix.
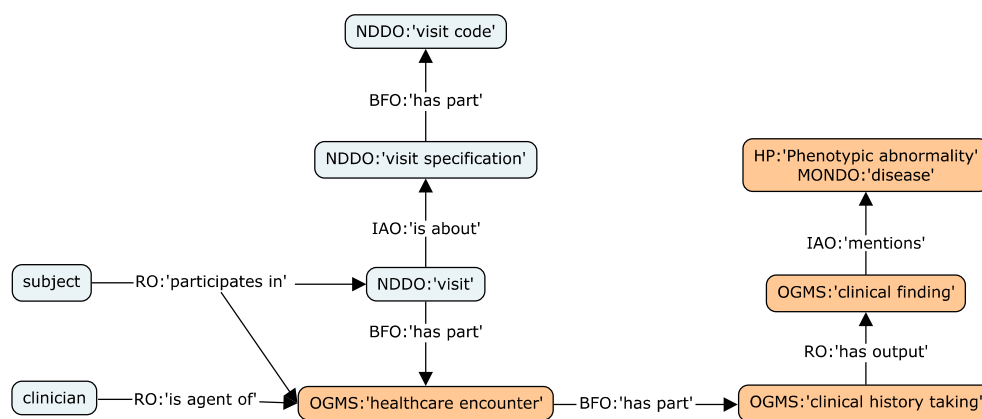
Figure 1. Concept map of the visits (light blue) and clinical (orange) design patterns, depicting the main ontology classes employed to model data entities.

The first pattern models the relations between study protocol/visit encounters, the agents involved, and the resulting outputs. The pattern mainly utilizes concepts defined in the Neurodegenerative Disease Data Ontology (NDDO) [91] (integrated in NIO) and the Ontology for General Medical Science (OGMS). The pattern supports a proper logical definition of longitudinal protocols, common in dementia research studies.
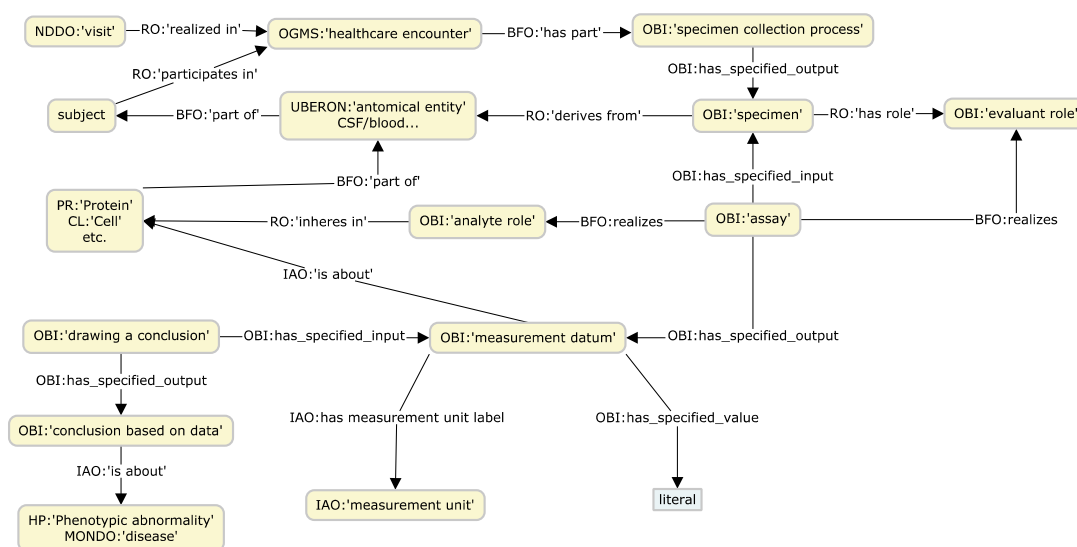
Figure 2. Concept map of the experimental measurements design pattern.
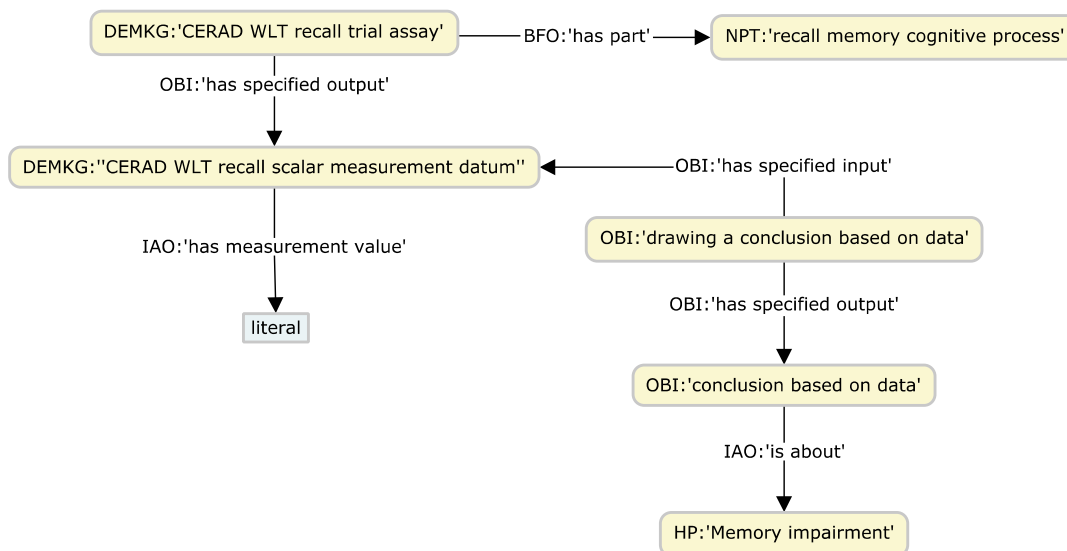

Figure 3. Exemplification of the neuropsychological test design pattern, through a CERAD recall test.

Clinical history phenotypes are characterized through observations at a study visit or from existing records. The framework leverages a pattern that relates visits with specific clinical administration, the finding, and the observed phenotype, usually a phenotype or disease concept from MONDO or HP. Relevant metadata can also be linked to the OGMS clinical entities, such as dates, agents involved, and locations. This pattern is shared across medical history, physical examination, and diagnosis processes. Figure 1 illustrates both the visits and clinical patterns.

A critical component of research data encompasses various assay measurements and proteomic datasets. We employ OBI's assay design patterns [92] to capture the multiple aspects involved in this process. These patterns enable the comprehensive integration of data pertaining to the assay, the specimen, and the molecule or material under examination, such as a protein or leukocyte count. Several relevant ontologies, including GO, PR, and Cell Ontology (CL), supply the necessary terminologies. We leverage entities from UBERON to denote the anatomical origin of the sample. This pattern facilitates the preservation of crucial metadata about processes, encompassing information about the type of assay, the specimen or sample employed, experimental conditions such as freeze–thaw cycles, and the date and time of collection. Such metadata is of considerable value for resource

management and can significantly aid research analyses. For instance, the type of tube in which a sample was collected could influence assay results and should be accounted for in linear models. Overall, it provides a more comprehensive context of the conditions under which experiments are conducted, enhancing the reproducibility and reliability of experimental outcomes.

Analyses derived from neuroimaging techniques, including segmentation measurements from tools such as Freesurfer [93] and Automatic Sub Hippocampal Segmentations (ASHS) [94], along with white matter evaluations from Diffusion Tensor Imaging (DTI) [95] and peak width of skeletonized mean diffusivity (PSMD) [96], play an indispensable role in dementia research. The pattern supporting this data modality follows a similar approach to the previous one, illustrated in Figure 2. To associate the measured anatomical entities, we utilize the FMA, which offers precise terms to align with the parcellation regions delineated by the widely used brain atlases in segmentation software, particularly for hemisphere-specific terms. More general terms from UBERON can be obtained using the "xref" property, employed for mapping concepts between different ontologies.

The last design pattern focuses on effectively relating the information content of a given test with the cognitive domain, providing means by which to stratify subjects via cognitive staging and the specific domain or phenotypic abnormality from HP at query time. This pattern exploits the axioms that connect cognitive tests with the evaluated domains.

4. Results

We have developed a KG framework that harmonizes biomedical knowledge and evidence from various sources, coupled with a transformation module designed to streamline the integration of multi-modal and omics data in dementia research. The core components of the framework encompass the extensions ontology builder, which provide ontological definitions to fill identified gaps from the domain ontologies; the KG-builder, in charge of obtaining, merging, and producing the KG; and the data transformer module, a low-code interface to transform source study data. All components are publicly accessible on GitHub (https://github.com/demkg-framework/, accessed on 30 August 2023). This trio of tools forms an intuitive building pipeline and also offers flexibility for customization, enabling users to construct the graph from scratch, adapt it to specific requirements, and deploy it on their preferred platform and graph database.

The backbone of our implementation is rooted in established community standards, technologies, and methodologies. The initial step involved the selection of a comprehensive array of domain reference biomedical ontologies, primarily from OBO, to form an expressive knowledge model for our primary KG. These ontologies offer a variety of well-defined concepts across varying levels of granularity, encapsulating intricate details of biological reality in the form of hierarchical relationships and concept networks.

To facilitate a consistent term mapping across various ontologies and mitigate computational demands, we utilized pre-built KGs from the KG-Hub initiative and the KG-OBO subset as our foundation, employing the KGX tool for the merging phase of the KG-builder pipeline. The KG-Hub initiative utilizes the Biolink model as its high-level data model, which we adopted to introduce greater flexibility and provide a comprehensive yet adaptable terminology overlay on the ontological model. The Biolink model facilitated the creation of both relaxed and detailed modeling and query capabilities, thereby enhancing the standardization and flexibility of our model. The default KG consists of 1.5 M nodes and 11.5 M edges.

To fill the identified gaps in the foundational model, we developed specific terminological extensions through the extensions ontology. We employed ODK to systematically introduce new terms, leveraging the OBO ecosystem to import and extend relevant external terms using DOS-DP whenever feasible.

Finally, the transformation module provides a low-code solution to transform tabular source data and generate necessary instance nodes and edges by following specific

design patterns that effectively depict study visits, phenotype observations, measurements/analyses derived from samples, and neuropsychological test results. These design patterns promote efficient data instantiation under the ontological model of the source research data, interconnecting various aspects of the study design outputs and providing a robust platform for data querying and network-oriented analyses. Figure 4 shows an overview of the framework components.
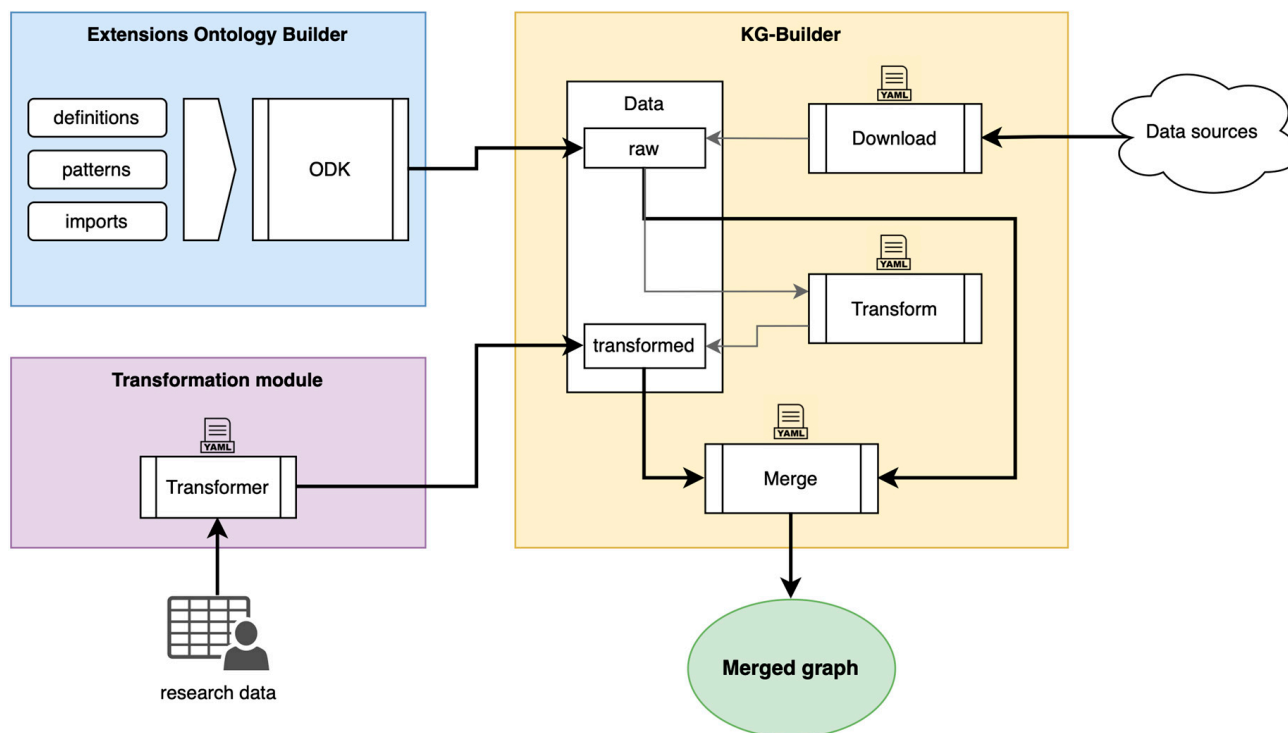
Figure 4. Overview of the DemKG framework components.

4.1. Use Case: Graph-Enabled Phenotype, Flow, and Protein Exploration from AT(N) Biomarker Profiles

To validate the DemKG framework, we applied it to the Dementia Disease Initiation (DDI) study data, a multi-site longitudinal observational study aimed at identifying early biomarkers for patients at risk of developing dementia [97]. The DDI dataset encompasses a range of clinical items, including medical history, standardized physical, neurological, and cognitive examinations, as well as laboratory and proteomic assays derived from blood and cerebrospinal fluid (CSF) samples, MRI, FDG-PET, and amyloid PET imaging, along with genomic analyses. We integrated these diverse data modalities and explored various aspects of the key biomarkers of the AD continuum, as categorized by the AT(N) classification.

4.1.1. Experimental Setup

The central DDI data platform is the XNAT archiving system [98], which is complemented by tailored customizations and data export functionalities, including automatic biomarker-based AT(N) classification, and population-adjusted norming for pertinent screening tests such as CERAD [99,100], VOSP [77], and TMT [78,101]. We implemented the transformation descriptor for the DDI data, involving direct mappings from clinical codes and rules to translate assay and experiment results into specific phenotype and disease entities. We then fed the descriptor along with the aggregated Comma-separated values (CSV) dump from XNAT to the transformation module to obtain the graph representation.

The DDI cohort graph comprises 96,939 nodes and 362,824 edges, whereas an average subject subgraph with four visits has 3469 nodes and 8284 edges. This transformed graph was merged into the final DDI-KG, which we ingested using the KGX module into a Neo4j Community instance deployed in a Podman container configured with eight cores and 16 GB of RAM, running on the secured servers of the TSD (Tjeneste for Sensitive Data) facilities managed by the University of Oslo. We opted for Neo4j due to its widespread adoption, the capabilities of its Cypher query language, and its reliable performance. Furthermore, KGX automatically creates node indices and constraints to improve loading and query performance for this platform.

Taking advantage of these features, the setup proves efficient with the resultant graph model, particularly for queries with clearly defined traversals and designated node labels. Figure 5 offers a preliminary analysis for estimating query performance, tracing the time consumed in navigating paths that extend from one to ten hops from subject nodes to various relevant node types in the graph. As anticipated, the number of target nodes considerably affects query performance, primarily driven by the increased number of edges to evaluate and traverse, coupled with the augmented data volume to handle. This scenario is especially pronounced in the most populated and interconnected node types, namely, proteins, genes, and diseases. Therefore, queries involving numerous or unrestricted quantities of such nodes require thoughtful design.
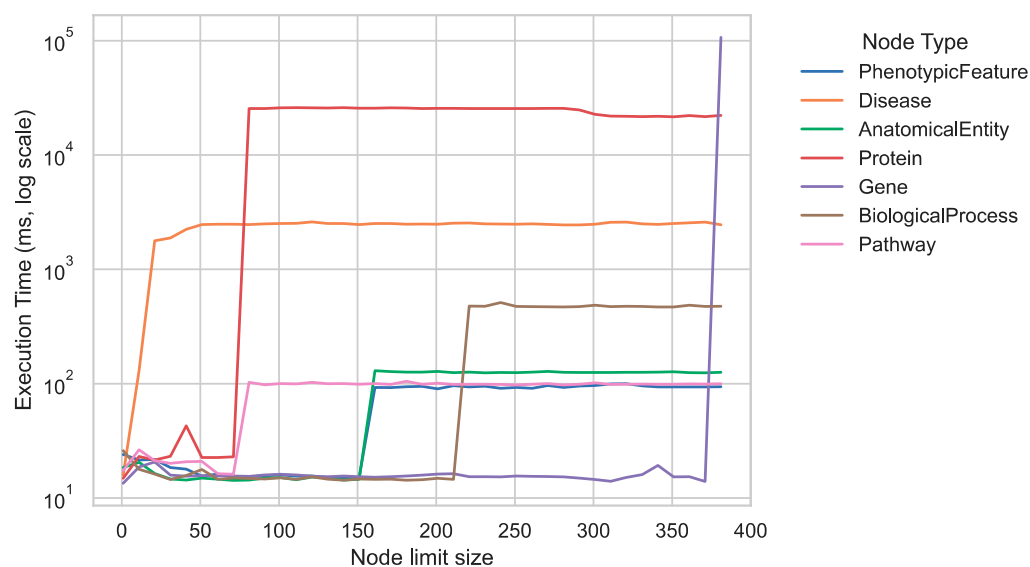
Figure 5. Mean execution times over ten runs for variable-length traversal queries between 1 and 10 connections, navigating from subject nodes to key Biolink categories.

4.1.2. Experimental Results

A key objective of the DDI study is to comprehend the evolution of subjects across different disease states within the biological reality, and the AT(N) classification system is a pivotal reference point. The developed design patterns facilitate connections at various levels, enabling the exploration of individual and group trajectories across visits and expediting the retrieval of relevant phenotypes using graph queries (Figure 6).

(**a**)

(**b**)
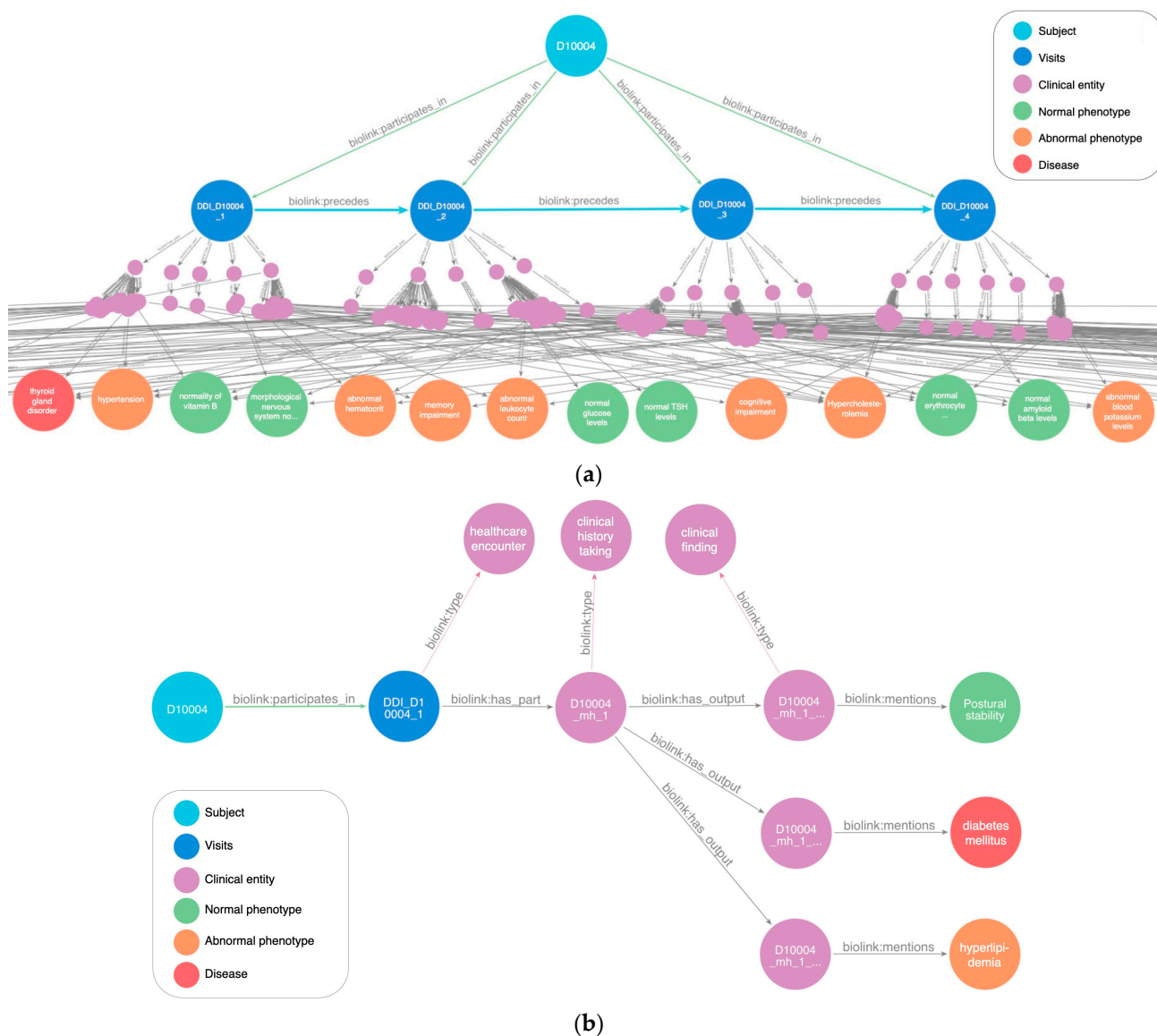
Figure 6. A DDI subject subgraph that illustrates study visits and associated phenotypes, visualized with Neo4j Bloom and further edited for readability. (**a**) An overview of longitudinal visits. Subjects are connected to each visit via the "biolink:participates_in" predicate. The logical sequencing of visits is established through the "biolink:precedes" predicate, facilitating query traversal. Clinical entity nodes represent associated medical processes (medical history, cognitive screenings, lab assays, and more), serving as the source of observations and conclusions while also supplying context and metadata for encounters and experimental setups. These nodes link to phenotype and disease entities to depict the outcomes of the clinical/research processes. (**b**) A specific visit branch tracing the path from the individual subject to the evaluated phenotypes and diseases noted during a medical history recording. Additional data from clinical entities are omitted to maintain clarity and uphold subject privacy.

Using the AT(N) entities defined in the extensions ontology, we queried the graph database to investigate the flow between the different biomarker profiles. This exploration helped unravel the transitions between them at the cohort level, aiding in data filtering for parallel research endeavors. Moreover, presented visually (Figure 7), the outcomes of these queries proved instrumental in quality control efforts by highlighting unlikely transitions from pathological to normal states. Such interventions are vital since AT(N) profiles derive

from biomarker measurements, where unexpected transitions may result from issues or errors in the respective assays.


(**a**)
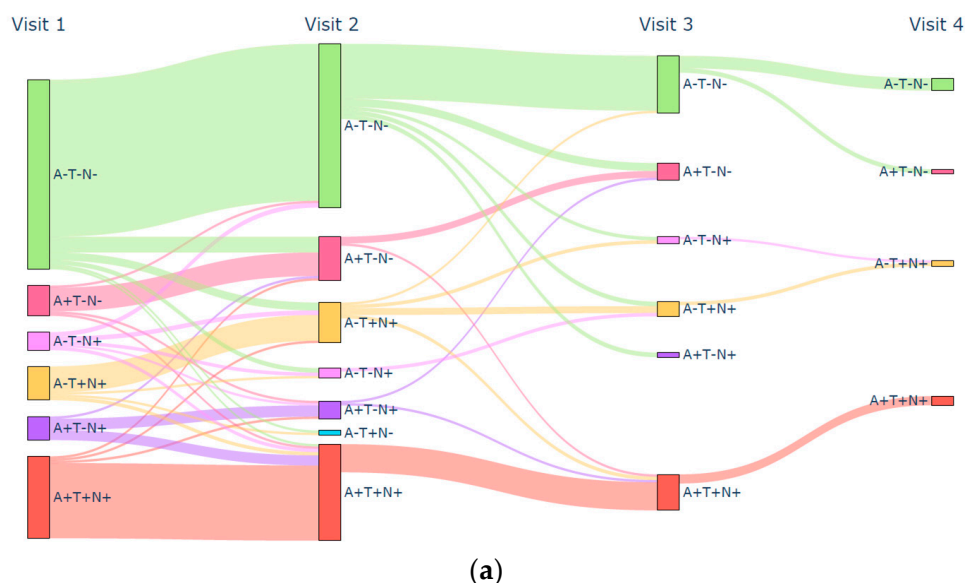
```
// AT(N) biomarker profiles are subclass of AT(N) value specifications in the extension ontology
MATCH
    (atn_val)-[:`biolink:subclass_of`]->(:`biolink:NamedThing` {id:'DEMKG:1900000'})
// source AT(N) values
MATCH
    (subject:`biolink:IndividualOrganism`)-[:`biolink:participates_in`]->(v1:`biolink:ClinicalEntity`)
    -[:`biolink:has_part`]->()-[:`biolink:has_output`]->()-[:`biolink:has_part`]->(atn_val)
// target ATNs values are collected from follow-up visits, using 'precedes' predicate
MATCH
    (v1)-[:`biolink:precedes`]->(vtarget:`biolink:ClinicalEntity`)-[:`biolink:has_part`]
    ->()-[:`biolink:has_output`]->()-[:`biolink:has_part`]->(target_atn_val)
// Prepare source AT(N), target AT(N), and count
WITH
    atn_val.name + '_' + v1.visit AS source,
    target_atn_val.name + '_' + vtarget.visit AS target,
    count(DISTINCT subject.name) AS count
WHERE count > 1
RETURN  source, target, count
ORDER BY source, target
```

(**b**)

Figure 7. Graph-based analysis illustrating the transitional flow among AT(N) biomarker profiles within the DDI cohort over successive protocol visits. (**a**) A Sankey diagram depicting the transitions in biomarker profiles. (**b**) The Cypher query utilized to calculate transition counts based on the predefined AT(N) biomarker profiles in the ontology.

As shown in Figure 7b, one of the valuable attributes of KGs that incorporate domain ontologies is richer semantic querying. Leveraging the hierarchical structure within phenotype and disease ontologies, we exploited semantic querying to gather phenotypes spanning different domains and visualized their prevalence across the AT(N) profiles. As depicted in Figure 8, we focused on phenotypes extracted from the "Abnormality of higher mental function" class within the HP ontology. Phenotypes related to memory, language, and executive function were referenced based on the rules established for the norming items in the cognitive screening section of the dataset descriptor.
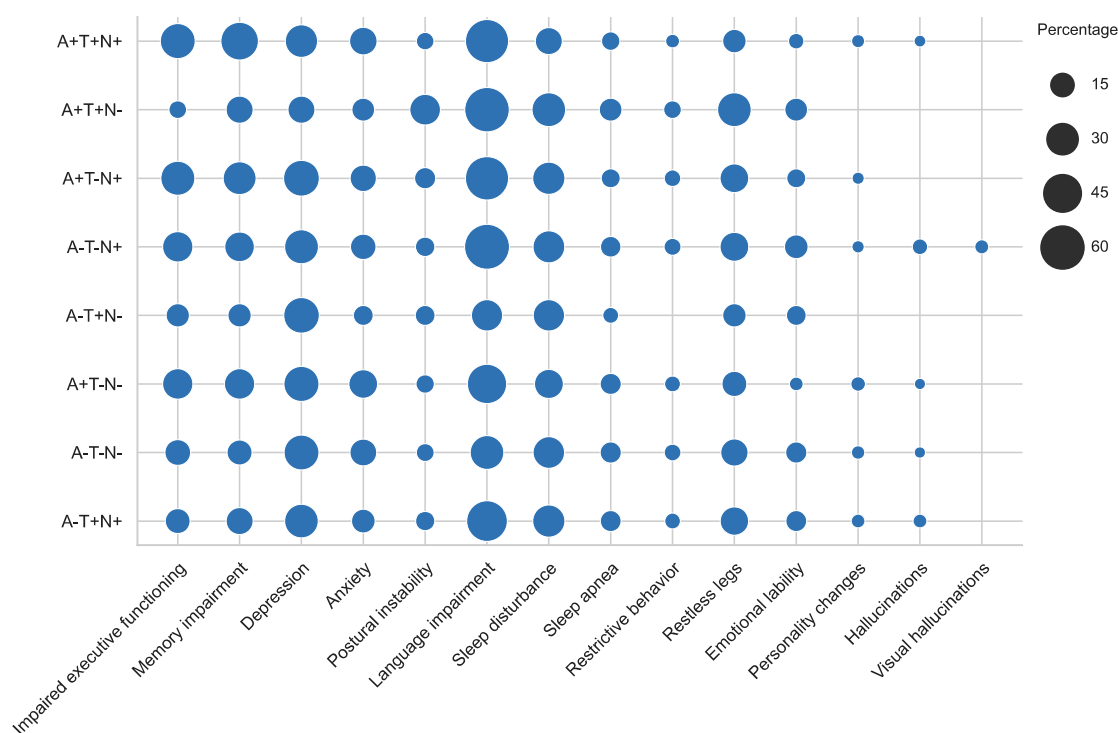
Figure 8. Dot plot from the collected phenotypes from subjects and their prevalence among the different AT(N) biomarker profiles.

To capture complex graph structures into low-dimensional vector space, we utilized the GRAPE library to create node embeddings using the node2Vec algorithm [102] with Skip Gram [103] and applied them to evaluate various aspects of the AT(N) biomarkers.

We conducted an interesting experiment to investigate if the embeddings of subject visits showed any patterns in the low-dimensional space or were influenced by specific AT(N) profiles. Using t-SNE [104] to reduce the embeddings to two dimensions, we observed a clear tendency for Tau pathology to group together in the embedding space, suggesting shared characteristics among the phenotypes assessed in those visits. The visit node embeddings are visualized in Figure 9, accompanied by a decision boundary computed through a logistic regression model.

Lastly, we combined the graph query capabilities, node embeddings, and topological metrics to obtain a broader overview of the relationships between assay proteins and the AT(N) protein biomarkers to assist in decision-making processes that could steer future analyses. Since the graph provides explicit links between available assays and the analytes being evaluated, we gathered CSF-derived ELISA and proteomics target proteins for comparison, focusing on the shared network encompassing GO biological processes (BPs).

For assessing protein relationships, we employed a simple pair-wise cosine similarity measure. This allowed us to quickly gauge how closely protein nodes were related and then rank the proteins that were most closely associated with the AT(N) panel (Figure 10).

To examine shared BPs between AT(N) and the assessed proteins, we employed a graph query to obtain the extensive network of protein activities. Given that proteins participate in thousands of such processes, to enhance navigability, we used GRAPE to calculate node betweenness and closeness centrality metrics, utilizing them as indicators of node relevance for prioritizing and narrowing down the pool of BPs to be investigated. A snapshot of this process is depicted in Figure 11.
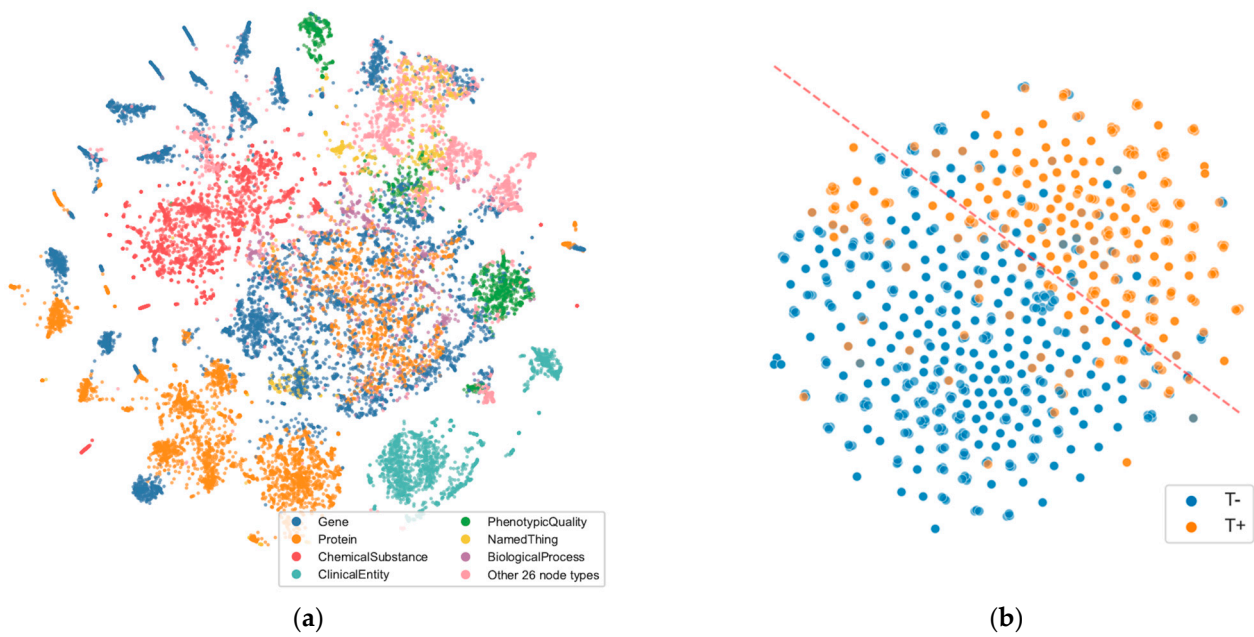
Figure 9. t-SNE visualizations of node embeddings. (**a**) Scatter plot output from GRAPE for all node embeddings from the KG representing the topological connectivity, colored by node type. It displays similarity and some possible clusters (Balanced accuracy: 60.32% $\pm$ 1.25%); separability consideration derives from evaluating a Decision Tree trained on five Monte Carlo holdouts, with a 70/30 split between training and test sets. (**b**) Visit node embeddings with nodes labeled by their associated T biomarker from AT(N) (pathologic tau). The dashed line marks the decision boundary between node types computed from a logistic regression model, with an accuracy of 0.831.
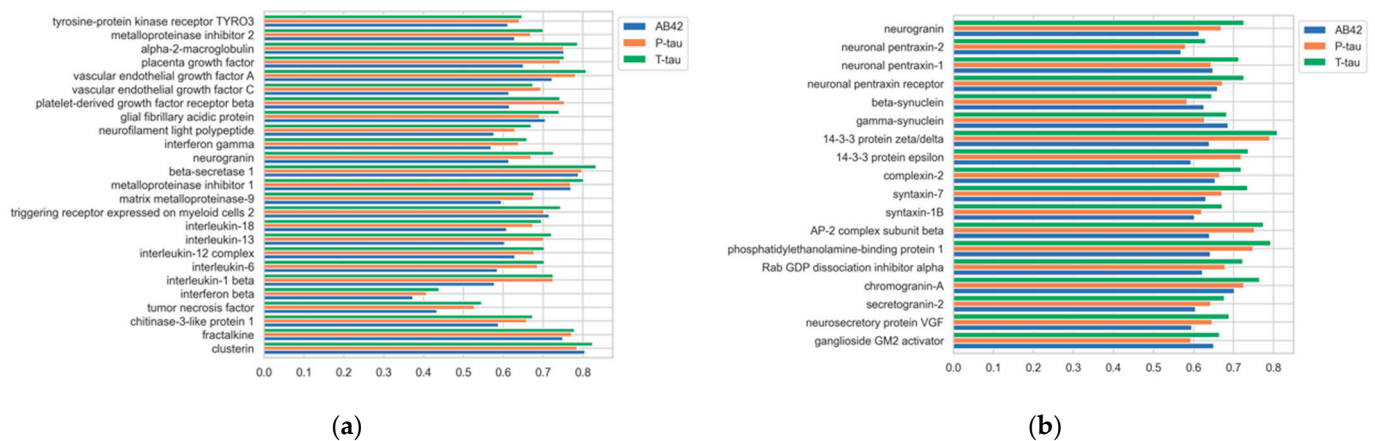
Figure 10. Cosine similarity of target proteins to AT(N) proteins. (**a**) CSF ELISA protein panel. (**b**) Synaptic protein panel from proteomics assays.
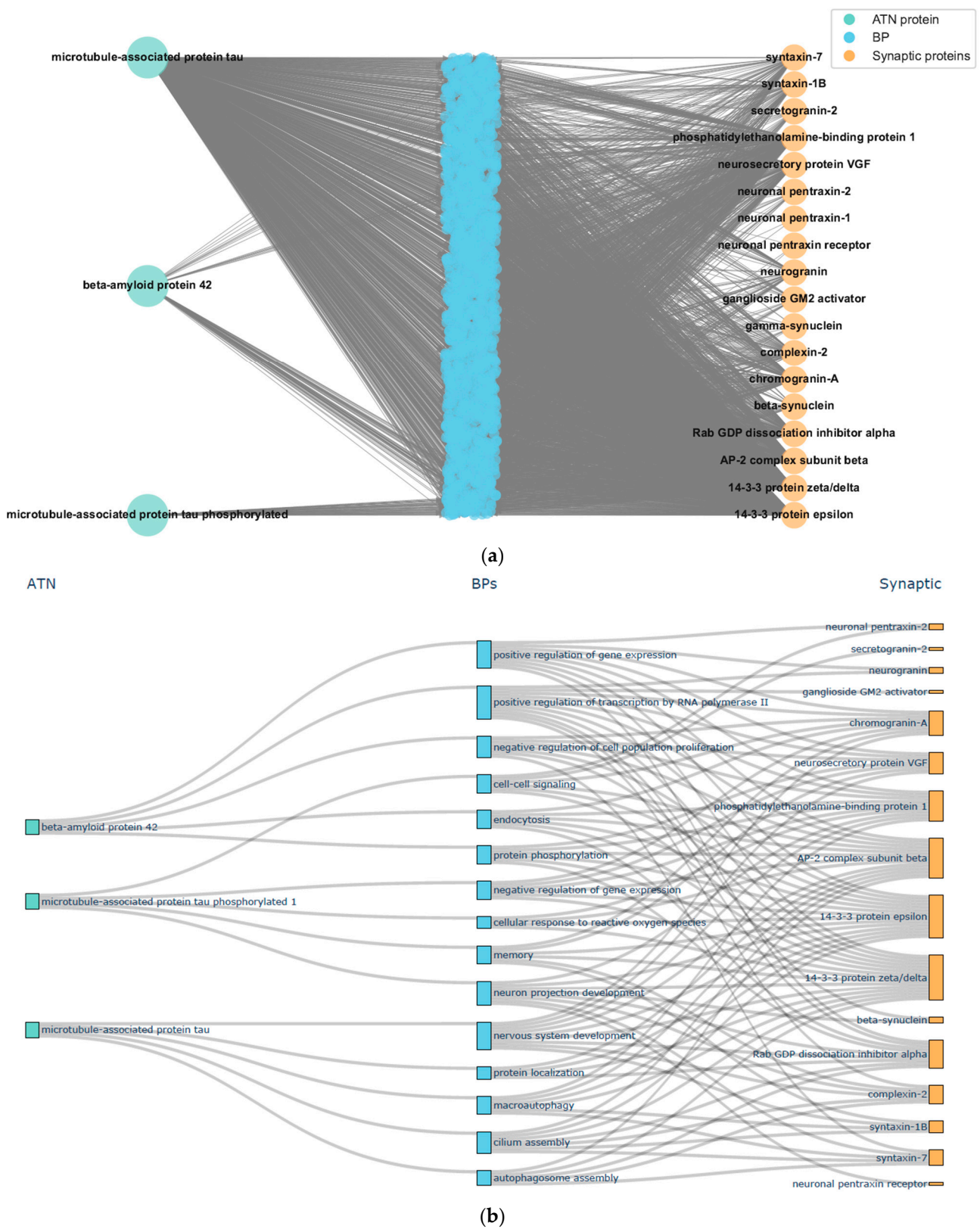
Figure 11. A snapshot of BP prioritization from node centrality. (**a**) Full subnetwork of shared BPs between AT(N) and synaptic panel proteins. (**b**) Sankey diagram with the top 10 BPs obtained from closeness centrality.

5. Discussion

In our work, we introduce DemKG, a KG framework designed to integrate various ontologies and knowledge sources to focus on dementia research data. This framework aims to cover terminological and design needs for multi-modal and omics data, with additional terminological extensions developed when necessary. We also followed specific patterns to cater to typical dementia research data outputs.

A key advantage of DemKG is its flexibility and ease of extension or customization to adapt to particular needs, made possible by the generalizable and pattern-based technologies employed in different components of the framework. Another relevant feature of DemKG is the friendly interface of the transformation module, which lowers the technical barrier to effectively integrating study research data in the KG.

However, there exists an important limitation in its implementation: once built, the KG does not support modifications without risking underlying integrity, forcing a complete build and possibly ingestion when new versions become available. This limitation, a consequence of using KGX as the backbone for merging and building operations, may ultimately limit projects with streamed or on-demand data ingestion needs.

Nevertheless, our implementations remain open-source, primarily based on open knowledge sources, and the building pipelines employ systematic approaches with templating engines that are easily customizable. While our focus is dementia research, the broad biomedical ontologies forming the foundation of our terminological model make our KG applicable to other biomedical research datasets as well.

Thus, the broader implications of our work extend beyond the application of the KG. Large biomedical KGs are proving to be an excellent tool for biomedical research, especially in domains requiring knowledge across different fields. The capacity to integrate disparate data and knowledge opens up opportunities for insights that were previously challenging to achieve. Approaches such as Precision Medicine greatly benefit from the implementation of KGs in their workflow.

This benefit is especially pronounced in dementia research, where the number of newly discovered biomarkers, phenotypes, and life conditions rapidly increases. These elements become part of the knowledge base that can be applied to the patient's biological signature. In this context, a KG like ours can play a crucial role in advancing our understanding of dementia and potentially informing patient care strategies.

6. Conclusions

In conclusion, DemKG presents a flexible and integrative approach to handle the ever-increasing complexity and multi-modality of dementia research data by leveraging a KG representation and relation capabilities.

The DemKG framework offers several distinct advantages over other solutions currently available. First, it is constructed based on well-established ontologies and adheres to recognized community standards, guaranteeing a solid and interoperable foundation. This is further enhanced by ontological extensions specifically crafted to facilitate detailed dementia research data analysis, filling a critical gap in the existing frameworks.

In addition to the above, DemKG integrates a low-code transformer module, simplifying the integration of study data and making the framework accessible to researchers with various levels of expertise. This module significantly reduces the time and technical know-how needed to merge study data, streamlining the data integration process considerably when compared to other solutions.

Furthermore, DemKG employs tooling to generate knowledge graphs in the platform-agnostic KGX format. This approach allows for easy deployment in a platform of the user's choice, offering flexibility in how and where the data can be used, and ensuring that the framework is adaptable to existing systems and future technological advancements. Enhancing its flexibility, the framework offers an open-source and customizable design, facilitating easy adoption and adaptation not only for dementia research but also potentially extending its utility to research into other diseases.

While there are limitations to the support for post-build modifications in its current iteration, addressing these in future work could broaden its applicability further. Despite these challenges, DemKG and similar KGs hold significant potential for propelling biomedical research and patient care advancements, extending from dementia to other medical conditions.

References

1. Manzoni, C.; Kia, D.A.; Vandrovcova, J.; Hardy, J.; Wood, N.W.; Lewis, P.A.; Ferrari, R. Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences. *Brief. Bioinform.* **2018**, *19*, 286–302. [CrossRef] [PubMed]
2. Misra, B.B.; Langefeld, C.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances and Future Approaches. *J. Mol. Endocrinol.* **2019**, *62*, R21–R45. [CrossRef]
3. Glaab, E.; Rauschenberger, A.; Banzi, R.; Gerardi, C.; Garcia, P.; Demotes, J. Biomarker Discovery Studies for Patient Stratification Using Machine Learning Analysis of Omics Data: A Scoping Review. *BMJ Open* **2021**, *11*, e053674. [CrossRef] [PubMed]
4. Sun, Z.; Ng, K.; Ramli, N. Biomedical Imaging Research: A Fast-Emerging Area for Interdisciplinary Collaboration. *Biomed. Imaging Interv. J.* **2011**, *7*, e21. [CrossRef]
5. Lussier, Y.A.; Liu, Y. Computational Approaches to Phenotyping: High-Throughput Phenomics. *Proc. Am. Thorac. Soc.* **2007**, *4*, 18–25. [CrossRef] [PubMed]
6. Che, Z.; Liu, Y. Deep Learning Solutions to Computational Phenotyping in Health Care. In Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 1100–1109.
7. Che, Z.; Kale, D.; Li, W.; Bahadori, M.T.; Liu, Y. Deep Computational Phenotyping. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 507–516.
8. Barabasi, A.-L.; Oltvai, Z.N. Network Biology: Understanding the Cell's Functional Organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113. [CrossRef] [PubMed]
9. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-Based Approach to Human Disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef]
10. Timón-Reina, S.; Rincón, M.; Martínez-Tomás, R. An Overview of Graph Databases and Their Applications in the Biomedical Domain. *Database* **2021**, *2021*, 26. [CrossRef]
11. Introducing the Knowledge Graph: Things, Not Strings. Available online: https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed on 13 September 2023).
12. Noy, N.; Gao, Y.; Jain, A.; Narayanan, A.; Patterson, A.; Taylor, J. Industry-Scale Knowledge Graphs: Lessons and Challenges: Five Diverse Technology Companies Show How It's Done. *Queue* **2019**, *17*, 48–75. [CrossRef]
13. Sheth, A.; Padhee, S.; Gyrard, A. Knowledge Graphs and Knowledge Networks: The Story in Brief. *IEEE Internet Comput.* **2019**, *23*, 67–75. [CrossRef]
14. Ehrlinger, L.; Wöß, W. Towards a definition of knowledge graphs. In *CEUR Workshop Proceedings*; CEUR-WS: Aachen, Germany, 2016; Volume 1695.
15. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S. Knowledge Graphs. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–37.

16. Besta, M.; Peter, E.; Gerstenberger, R.; Fischer, M.; Podstawski, M.; Barthels, C.; Alonso, G.; Hoefler, T. Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries. *ACM Comput. Surv.* **2019**, *56*, 1–40. [CrossRef]

17. Brandizi, M.; Singh, A.; Rawlings, C.; Hassani-Pak, K. Getting the Best of Linked Data and Property Graphs: Rdf2neo and the KnetMiner Use Case. In Proceedings of the CEUR Workshop Proceedings, Antwerp, Belgium, 3–6 December 2018; Volume 2275.

18. Alocci, D.; Mariethoz, J.; Horlacher, O.; Bolleman, J.T.; Campbell, M.P.; Lisacek, F. Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. *PLoS ONE* **2015**, *10*, e0144578. [CrossRef] [PubMed]

19. Hoehndorf, R.; Schofield, P.N.; Gkoutos, G.V. The Role of Ontologies in Biological and Biomedical Research: A Functional Perspective. *Brief. Bioinform.* **2015**, *16*, 1069–1080. [CrossRef]

20. Dovrolis, N.; Stefanut, T.; Dietze, S.; Yu, H.Q.; Valentine, C.; Kaldoudi, E. Semantic Annotation and Linking of Medical Educational Resources. In *5th European Conference of the International Federation for Medical and Biological Engineering 14–18 September 2011, Budapest, Hungary*; Jobbágy, Á., Ed.; IFMBE Proceedings; Springer: Berlin/Heidelberg, Germany, 2011; Volume 37, pp. 1400–1403.

21. Song, D.; Chute, C.G.; Tao, C. Semantator: Annotating Clinical Narratives with Semantic Web Ontologies. *AMIA Jt. Summits Transl. Sci. Proc.* **2012**, *2012*, 20–209. [PubMed]

22. Shah, N.H.; Bhatia, N.; Jonquet, C.; Rubin, D.; Chiang, A.P.; Musen, M.A. Comparison of concept recognizers for building the open biomedical annotator. In *BMC Bioinformatics*; BioMed Central: London, UK, 2009; Volume 10.

23. El-Haj, M.; Rutherford, N.; Coole, M.; Ezeani, I.; Prentice, S.; Ide, N.; Knight, J.; Piao, S.; Mariani, J.; Rayson, P.; et al. Infrastructure for Semantic Annotation in the Genomics Domain. In Proceedings of the LREC, Marseille, France, 20–25 June 2020.

24. Tan, H.; Lambrix, P. Selecting an ontology for biomedical text mining. In *Workshop on Current Trends in Biomedical Natural Language Processing*; Association for Computational Linguistics: Boulder, CO, USA, 2009; pp. 55–62.

25. Witte, R.; Kappler, T.; Baker, C.J.O. Ontology Design for Biomedical Text Mining. In *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*; Springer: Boston, MA, USA, 2007; Volume 9780387484, pp. 281–313. ISBN 978-0-387-48438-9.

26. Jackson, R.; Matentzoglu, N.; Overton, J.A.; Vita, R.; Balhoff, J.P.; Buttigieg, P.L.; Carbon, S.; Courtot, M.; Diehl, A.D.; Dooley, D.M.; et al. OBO Foundry in 2021: Operationalizing Open Data Principles to Evaluate Ontologies. *Database* **2021**, *2021*, baab069. [CrossRef] [PubMed]

27. Musen, M.A.; Noy, N.F.; Shah, N.H.; Whetzel, P.L.; Chute, C.G.; Story, M.A.; Smith, B. The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 190–195. [CrossRef]

28. Whetzel, P.L.; Noy, N.F.; Shah, N.H.; Alexander, P.R.; Nyulas, C.; Tudorache, T.; Musen, M.A. BioPortal: Enhanced Functionality via New Web Services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications. *Nucleic Acids Res.* **2011**, *39*, W541–W545. [CrossRef] [PubMed]

29. Mungall, C.J.; McMurry, J.A.; Köhler, S.; Balhoff, J.P.; Borromeo, C.; Brush, M.; Carbon, S.; Conlin, T.; Dunn, N.; Engelstad, M.; et al. The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species. *Nucleic Acids Res.* **2017**, *45*, D712–D722. [CrossRef] [PubMed]

30. Santos, A.; Colaço, A.R.; Nielsen, A.B.; Niu, L.; Strauss, M.; Geyer, P.E.; Coscia, F.; Albrechtsen, N.J.W.; Mundt, F.; Jensen, L.J.; et al. A Knowledge Graph to Interpret Clinical Proteomics Data. *Nat. Biotechnol.* **2022**, *40*, 692–702. [CrossRef] [PubMed]

31. Chandak, P.; Huang, K.; Zitnik, M. Building a Knowledge Graph to Enable Precision Medicine. *Sci. Data* **2023**, *10*, 67. [CrossRef] [PubMed]

32. Morris, J.H.; Soman, K.; Akbas, R.E.; Zhou, X.; Smith, B.; Meng, E.C.; Huang, C.C.; Cerono, G.; Schenk, G.; Rizk-Jackson, A. The Scalable Precision Medicine Open Knowledge Engine (SPOKE): A Massive Knowledge Graph of Biomedical Information. *Bioinformatics* **2023**, *39*, btad080. [CrossRef] [PubMed]

33. Reese, J.T.; Unni, D.; Callahan, T.J.; Cappelletti, L.; Ravanmehr, V.; Carbon, S.; Shefchek, K.A.; Good, B.M.; Balhoff, J.P.; Fontana, T.; et al. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns* **2021**, *2*, 100155. [CrossRef] [PubMed]

34. Badal, V.D.; Wright, D.; Katsis, Y.; Kim, H.-C.; Swafford, A.D.; Knight, R.; Hsu, C.-N. Challenges in the Construction of Knowledge Bases for Human Microbiome-Disease Associations. *Microbiome* **2019**, *7*, 129. [CrossRef] [PubMed]

35. Chaves-Fraga, D.; Endris, K.M.; Iglesias, E.; Corcho, O.; Vidal, M.-E. What Are the Parameters That Affect the Construction of a Knowledge Graph? In Proceedings of the On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, 21–25 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 695–713.

36. Unni, D.R.; Moxon, S.A.T.; Bada, M.; Brush, M.; Bruskiewich, R.; Caufield, J.H.; Clemons, P.A.; Dancik, V.; Dumontier, M.; Fecho, K.; et al. Biolink Model: A Universal Schema for Knowledge Graphs in Clinical, Biomedical, and Translational Science. *Clin. Transl. Sci.* **2022**, *15*, 1848–1855. [CrossRef] [PubMed]

37. Caufield, J.H.; Putman, T.; Schaper, K.; Unni, D.R.; Hegde, H.; Callahan, T.J.; Cappelletti, L.; Moxon, S.A.; Ravanmehr, V.; Carbon, S.; et al. KG-Hub—Building and Exchanging Biological Knowledge Graphs 2023. *Bioinformatics* **2023**, *39*, btad418. [CrossRef] [PubMed]

38. Hamilton, W.L.; Ying, R.; Leskovec, J. Representation Learning on Graphs: Methods and Applications. *arXiv* **2017**, arXiv:1709.05584.

39. Chami, I.; Abu-El-Haija, S.; Perozzi, B.; Ré, C.; Murphy, K. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *J. Mach. Learn. Res.* **2022**, *23*, 3840–3903.

40. Cappelletti, L.; Fontana, T.; Casiraghi, E.; Ravanmehr, V.; Callahan, T.J.; Joachimiak, M.P.; Mungall, C.J.; Robinson, P.N.; Reese, J.; Valentini, G. GRAPE: Fast and Scalable Graph Processing and Embedding 2021. *Nat. Comput. Sci.* **2023**, *3*, 552–568. [CrossRef]

41. Ilievski, F.; Garijo, D.; Chalupsky, H.; Divvala, N.T.; Yao, Y.; Rogers, C.; Li, R.; Liu, J.; Singh, A.; Schwabe, D. KGTK: A Toolkit for Large Knowledge Graph Manipulation and Analysis. In Proceedings of the The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, 2–6 November 2020; pp. 278–293.

42. Nelson, C.A.; Bove, R.; Butte, A.J.; Baranzini, S.E. Embedding Electronic Health Records onto a Knowledge Network Recognizes Prodromal Features of Multiple Sclerosis and Predicts Diagnosis. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 424–434. [CrossRef] [PubMed]

43. Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2724–2743. [CrossRef]

44. Li, X.; Chen, W.; Chen, Y.; Zhang, X.; Gu, J.; Zhang, M.Q. Network Embedding-Based Representation Learning for Single Cell RNA-Seq Data. *Nucleic Acids Res.* **2017**, *45*, e166. [CrossRef]

45. Liu, X.; Yang, Z.; Sang, S.; Lin, H.; Wang, J.; Xu, B. Detection of Protein Complexes from Multiple Protein Interaction Networks Using Graph Embedding. *Artif. Intell. Med.* **2019**, *96*, 107–115. [CrossRef]

46. Wang, X.; Gong, Y.; Yi, J.; Zhang, W. Predicting Gene-Disease Associations from the Heterogeneous Network Using Graph Embedding. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine—BIBM 2019, San Diego, CA, USA, 1 November 2019; pp. 504–511.

47. Xu, B.; Liu, Y.; Yu, S.; Wang, L.; Dong, J.; Lin, H.; Yang, Z.; Wang, J.; Xia, F. A Network Embedding Model for Pathogenic Genes Prediction by Multi-Path Random Walking on Heterogeneous Network. *BMC Med. Genom.* **2019**, *12*, 188. [CrossRef]

48. Malec, S.A.; Taneja, S.B.; Albert, S.M.; Elizabeth Shaaban, C.; Karim, H.T.; Levine, A.S.; Munro, P.; Callahan, T.J.; Boyce, R.D. Causal Feature Selection Using a Knowledge Graph Combining Structured Knowledge from the Biomedical Literature and Ontologies: A Use Case Studying Depression as a Risk Factor for Alzheimer's Disease. *J. Biomed. Inform.* **2023**, *142*, 104368. [CrossRef]

49. Nicholson, D.N.; Greene, C.S. Constructing Knowledge Graphs and Their Biomedical Applications. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1414–1428. [CrossRef] [PubMed]

50. Arp, R.; Smith, B. Function, role and disposition in basic formal ontology. *Nat. Preced.* **2008**. [CrossRef]

51. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for The Unification of Biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]

52. The Gene Ontology Consortium; Aleksander, S.A.; Balhoff, J.; Carbon, S.; Cherry, J.M.; Drabkin, H.J.; Ebert, D.; Feuermann, M.; Gaudet, P.; Harris, N.L.; et al. The Gene Ontology Knowledgebase in 2023. *Genetics* **2023**, *224*, iyad031. [CrossRef]

53. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res.* **2016**, *44*, D1214–D1219. [CrossRef] [PubMed]

54. Natale, D.A.; Arighi, C.N.; Barker, W.C.; Blake, J.A.; Bult, C.J.; Caudy, M.; Drabkin, H.J.; D'Eustachio, P.; Evsikov, A.V.; Huang, H. The Protein Ontology: A Structured Representation of Protein Forms and Complexes. *Nucleic Acids Res.* **2010**, *39*, D539–D545. [CrossRef]

55. Vasilevsky, N.A.; Matentzoglu, N.A.; Toro, S.; Flack, J.E.; Hegde, H.; Unni, D.R.; Alyea, G.F.; Amberger, J.S.; Babb, L.; Balhoff, J.P.; et al. Mondo: Unifying Diseases for the World, by the World. *medRxiv* **2022**. [CrossRef]

56. Köhler, S.; Doelken, S.C.; Mungall, C.J.; Bauer, S.; Firth, H.V.; Bailleul-Forestier, I.; Black, G.C.M.; Brown, D.L.; Brudno, M.; Campbell, J.; et al. The Human Phenotype Ontology Project: Linking Molecular Biology and Disease through Phenotype Data. *Nucleic Acids Res.* **2014**, *42*, D966–D974. [CrossRef] [PubMed]

57. Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **2021**, *49*, D1207–D1217. [CrossRef]

58. Gkoutos, G.V.; Schofield, P.N.; Hoehndorf, R. The Anatomy of Phenotype Ontologies: Principles, Properties and Applications. *Brief. Bioinform.* **2018**, *19*, 1008–1021. [CrossRef] [PubMed]

59. Mungall, C.J.; Torniai, C.; Gkoutos, G.V.; Lewis, S.E.; Haendel, M.A. Uberon, an Integrative Multi-Species Anatomy Ontology. *Genome Biol.* **2012**, *13*, R5–R20. [CrossRef] [PubMed]

60. Haendel, M.A.; Balhoff, J.P.; Bastian, F.B.; Blackburn, D.C.; Blake, J.A.; Bradford, Y.; Comte, A.; Dahdul, W.M.; Dececchi, T.A.; Druzinsky, R.E. Unification of Multi-Species Vertebrate Anatomy Ontologies for Comparative Biology in Uberon. *J. Biomed. Semant.* **2014**, *5*, 21. [CrossRef]

61. Rosse, C.; Mejino, J.L.V. A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. *J. Biomed. Inform.* **2003**, *36*, 478–500. [CrossRef]

62. Cox, A.P.; Jensen, M.; Ruttenberg, A.; Szigeti, K.; Diehl, A.D. Measuring Cognitive Functions: Hurdles in the Development of the NeuroPsychological Testing Ontology. In Proceedings of the ICBO, Montreal, QC, Canada, 7–12 July 2013; pp. 78–83.

63. Gomez-Valades, A.; Martinez-Tomas, R.; Rincon, M. Integrative Base Ontology for the Research Analysis of Alzheimer's Disease-Related Mild Cognitive Impairment. *Front. Neuroinformatics* **2021**, *15*, 561691. [CrossRef]

64. Peters, B.; OBI Consortium, T. Ontology for Biomedical Investigations. *Nat. Preced.* **2009**, *1*. [CrossRef]

65. Bandrowski, A.; Brinkman, R.; Brochhausen, M.; Brush, M.H.; Bug, B.; Chibucos, M.C.; Clancy, K.; Courtot, M.; Derom, D.; Dumontier, M. The Ontology for Biomedical Investigations. *PLoS ONE* **2016**, *11*, e0154556. [CrossRef]

66. Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A.L.; Fang, T.; Doncheva, N.T.; Pyysalo, S.; et al. The STRING Database in 2023: Protein–Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest. *Nucleic Acids Res.* **2023**, *51*, D638–D646. [CrossRef]

67. Gillespie, M.; Jassal, B.; Stephan, R.; Milacic, M.; Rothfels, K.; Senff-Ribeiro, A.; Griss, J.; Sevilla, C.; Matthews, L.; Gong, C.; et al. The Reactome Pathway Knowledgebase 2022. *Nucleic Acids Res.* **2022**, *50*, D687–D692. [CrossRef] [PubMed]

68. Diehl, A.D.; Meehan, T.F.; Bradford, Y.M.; Brush, M.H.; Dahdul, W.M.; Dougall, D.S.; He, Y.; Osumi-Sutherland, D.; Ruttenberg, A.; Sarntivijai, S. The Cell Ontology 2016: Enhanced Content, Modularization, and Ontology Interoperability. *J. Biomed. Semant.* **2016**, *7*, 44. [CrossRef] [PubMed]

69. Nadendla, S.; Jackson, R.; Munro, J.; Quaglia, F.; Mészáros, B.; Olley, D.; Hobbs, E.T.; Goralski, S.M.; Chibucos, M.; Mungall, C.J.; et al. ECO: The Evidence and Conclusion Ontology, an Update for 2022. *Nucleic Acids Res.* **2022**, *50*, D1515–D1521. [CrossRef]

70. Malone, J.; Holloway, E.; Adamusiak, T.; Kapushesky, M.; Zheng, J.; Kolesnikov, N.; Zhukova, A.; Brazma, A.; Parkinson, H. Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics* **2010**, *26*, 1112–1118. [CrossRef] [PubMed]

71. Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A.R.; Binz, P.-A.; Deutsch, E.W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; et al. The HUPO Proteomics Standards Initiative- Mass Spectrometry Controlled Vocabulary. *Database* **2013**, *2013*, bat009. [CrossRef]

72. Stefancsik, R.; Balhoff, J.P.; Balk, M.A.; Ball, R.L.; Bello, S.M.; Caron, A.R.; Chesler, E.J.; de Souza, V.; Gehrke, S.; Haendel, M.; et al. The Ontology of Biological Attributes (OBA)—Computational Traits for the Life Sciences. *Mamm. Genome* **2023**, *34*, 364–378. [CrossRef]

73. Scheuermann, R.H.; Ceusters, W.; Smith, B. Toward an Ontological Treatment of Disease and Diagnosis. *Summit Transl. Bioinforma.* **2009**, *2009*, 116.

74. Hicks, A.; Hanna, J.; Welch, D.; Brochhausen, M.; Hogan, W.R. The Ontology of Medically Related Social Entities: Recent Developments. *J. Biomed. Semant.* **2016**, *7*, 47. [CrossRef] [PubMed]

75. Kurlowicz, L.; Wallace, M. The Mini-Mental State Examination (MMSE). *J. Gerontol. Nurs.* **1999**, *25*, 8–9. [CrossRef]

76. Fillenbaum, G.G.; Mohs, R. CERAD (Consortium to Establish a Registry for Alzheimer's Disease) Neuropsychology Assessment Battery: 35 Years and Counting. *J. Alzheimers Dis.* **2023**, *93*, 1–27. [CrossRef] [PubMed]

77. Quental, N.B.M.; Brucki, S.M.D.; Bueno, O.F.A. Visuospatial Function in Early Alzheimer's Disease—The Use of the Visual Object and Space Perception (VOSP) Battery. *PLoS ONE* **2013**, *8*, e68398. [CrossRef]

78. Bowie, C.R.; Harvey, P.D. Administration and Interpretation of the Trail Making Test. *Nat. Protoc.* **2006**, *1*, 2277–2281. [CrossRef]

79. Mainland, B.J.; Shulman, K.I. Clock Drawing Test. In *Cognitive Screening Instruments: A Practical Approach*; Springer: London, UK, 2017; pp. 67–108. [CrossRef]

80. Benton, A.L.; de Hamsher, S.; Sivan, A.B. Controlled Oral Word Association Test. *Arch. Clin. Neuropsychol.* **1994**. [CrossRef]

81. Jack, C.R., Jr.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Dunn, B.; Haeberlein, S.B.; Holtzman, D.M.; Jagust, W.; Jessen, F.; Karlawish, J.; et al. NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease. *Alzheimers Dement.* **2018**, *14*, 535–562. [CrossRef] [PubMed]

82. Fecho, K.; Thessen, A.E.; Baranzini, S.E.; Bizon, C.; Hadlock, J.J.; Huang, S.; Roper, R.T.; Southall, N.; Ta, C.; Watkins, P.B.; et al. Progress toward a Universal Biomedical Data Translator. *Clin. Transl. Sci.* **2022**, *15*, 1838–1847. [CrossRef]

83. Matentzoglu, N.; Goutte-Gattat, D.; Tan, S.Z.K.; Balhoff, J.P.; Carbon, S.; Caron, A.R.; Duncan, W.D.; Flack, J.E.; Haendel, M.; Harris, N.L.; et al. Ontology Development Kit: A Toolkit for Building, Maintaining, and Standardising Biomedical Ontologies. *Database* **2022**, *2022*, baac087. [CrossRef]

84. Osumi-Sutherland, D.; Courtot, M.; Balhoff, J.P.; Mungall, C. Dead Simple OWL Design Patterns. *J. Biomed. Semant.* **2017**, *8*, 18. [CrossRef] [PubMed]

85. Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P.F.; Rudolph, S. OWL 2 Web Ontology Language. Available online: https://www.w3.org/TR/owl2-primer/ (accessed on 30 August 2023).

86. Lawrence Berkeley National Laboratory. (BBOP), Lawrence Berkeley National Knowledge Graph Hub. Available online: https://kghub.org/ (accessed on 30 August 2023).

87. KGX Format—Kgx 1.5.1 Documentation. Available online: https://kgx.readthedocs.io/en/latest/kgx_format.html (accessed on 30 August 2023).

88. KG-OBO. Available online: https://github.com/Knowledge-Graph-Hub/kg-obo (accessed on 30 August 2023).

89. Relation-Graph. Available online: https://github.com/INCATools/relation-graph (accessed on 30 August 2023).

90. Balhoff, J.P.; Bayindir, U.; Caron, A.R.; Matentzoglu, N.; Osumi-Sutherland, D.; Mungall, C.J. Ubergraph: Integrating OBO Ontologies into a Unified Semantic Graph. In *CEUR Workshop Proceedings*; CEUR-WS: Aachen, Germany, 2022; Volume 1613, p. 73.

91. Kostovska, A.; Tolovski, I.; Maikore, F.; Initiative, A.D.N.; Soldatova, L.; Panov, P. Neurodegenerative Disease Data Ontology. In Proceedings of the Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, 28–30 October 2019; pp. 235–245.

92. Vita, R.; Zheng, J.; Jackson, R.; Dooley, D.; Overton, J.A.; Miller, M.A.; Berrios, D.C.; Scheuermann, R.H.; He, Y.; McGinty, H.K.; et al. Standardization of Assay Representation in the Ontology for Biomedical Investigations. *Database* **2021**, *2021*, baab040. [CrossRef]

93. Fischl, B. FreeSurfer. *Neuroimage* **2012**, *62*, 774–781. [CrossRef] [PubMed]

94. Yushkevich, P.A.; Pluta, J.B.; Wang, H.; Xie, L.; Ding, S.-L.; Gertje, E.C.; Mancuso, L.; Kliot, D.; Das, S.R.; Wolk, D.A. Automated Volumetry and Regional Thickness Analysis of Hippocampal Subfields and Medial Temporal Cortical Structures in Mild Cognitive Impairment. *Hum. Brain Mapp.* **2015**, *36*, 258–287. [CrossRef] [PubMed]

95. Basser, P.J.; Mattiello, J.; LeBihan, D. MR Diffusion Tensor Spectroscopy and Imaging. *Biophys. J.* **1994**, *66*, 259–267. [CrossRef] [PubMed]

96. Low, A.; Mak, E.; Stefaniak, J.D.; Malpetti, M.; Nicastro, N.; Savulich, G.; Chouliaras, L.; Markus, H.S.; Rowe, J.B.; O'Brien, J.T. Peak Width of Skeletonized Mean Diffusivity as a Marker of Diffuse Cerebrovascular Damage. *Front. Neurosci.* **2020**, *14*, 238. [CrossRef] [PubMed]

97. Fladby, T.; Pålhaugen, L.; Selnes, P.; Waterloo, K.; Bråthen, G.; Hessen, E.; Almdahl, I.S.; Arntzen, K.-A.; Auning, E.; Eliassen, C.F.; et al. Detecting At-Risk Alzheimer's Disease Cases. *J. Alzheimers Dis.* **2017**, *60*, 97–105. [CrossRef]

98. Marcus, D.S.; Olsen, T.R.; Ramaratnam, M.; Buckner, R.L. The Extensible Neuroimaging Archive Toolkit: An Informatics Platform for Managing, Exploring, and Sharing Neuroimaging Data. *Neuroinformatics* **2007**, *5*, 11–34. [CrossRef]

99. Fillenbaum, G.G.; van Belle, G.; Morris, J.C.; Mohs, R.C.; Mirra, S.S.; Davis, P.C.; Tariot, P.N.; Silverman, J.M.; Clark, C.M.; Welsh-Bohmer, K.A.; et al. Consortium to Establish a Registry for Alzheimer's Disease (CERAD): The First Twenty Years. *Alzheimers Dement.* **2008**, *4*, 96–109. [CrossRef] [PubMed]

100. Kirsebom, B.E.; Espenes, R.; Hessen, E.; Waterloo, K.; Johnsen, S.H.; Gundersen, E.; Botne Sando, S.; Rolfseng Grøntvedt, G.; Timón, S.; Fladby, T. Demographically Adjusted CERAD Wordlist Test Norms in a Norwegian Sample from 40 to 80 Years. *Clin. Neuropsychol.* **2019**, *33*, 27–39. [CrossRef]

101. Espenes, J.; Hessen, E.; Eliassen, I.V.; Waterloo, K.; Eckerström, M.; Sando, S.B.; Timón, S.; Wallin, A.; Fladby, T.; Kirsebom, B.-E. Demographically Adjusted Trail Making Test Norms in a Scandinavian Sample from 41 to 84 Years. *Clin. Neuropsychol.* **2020**, *34*, 110–126. [CrossRef]

102. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery, San Francisco, CA, USA, 13–17 August 2016; pp. 855–864.

103. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.

104. Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.