Artificial Intelligence in the Life Sciences 4 (2023) 100078

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research Article

A natural language processing system for the efficient updating of highly curated pathophysiology mechanism knowledge graphs

Negin Sadat Babaiha[a,b], Hassan Elsayed[a], Bide Zhang[a], Abish Kaladharan[c], Priya Sethumadhavan[c], Bruce Schultz[a], Jürgen Klein[a], Bruno Freudensprung[d], Vanessa Lage-Rupprecht[a], Alpha Tom Kodamullil[a], Marc Jacobs[a], Stefan Geissler[d], Sumit Madan[a,e], Martin Hofmann-Apitius[a,b,*]

[a] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[b] Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany
[c] Causality Biomodels, Kinfra Hi-Tech Park, Kalamassery, Cochin 683503, Kerala, India
[d] Kairntech SAS, 29 Chemin du Vieux Chêne 38240 Meylan, France
[e] Institute of Computer Science, University of Bonn, 53115 Bonn, Germany

ARTICLE INFO

Keywords:
Knowledge graphs
Relation extraction
Natural language processing
Biomedical text mining
Biological expression language (BEL)
Human brain pharmacome (HBP)

ABSTRACT

Background: Biomedical knowledge graphs (KG) have become crucial for describing biological findings in a structured manner. To keep up with the constantly changing flow of knowledge, their embedded information must be regularly updated with the latest findings. Natural language processing (NLP) has created new possibilities for automating this upkeep by facilitating information extraction from free text. However, due to annotated and labeled biomedical data limitations, the development of completely autonomous information extraction systems remains a substantial scientific and technological hurdle. This study aims to explore methodologies best suited to support the automatic extraction of causal relationships from biomedical literature with the aim of regular and rapid updating of disease-specific pathophysiology mechanism KGs.

Methods: Our proposed approach first searches and retrieves PubMed abstracts using the desired terms and keywords. The extension corpora are then passed through the NLP pipeline for automatic information extraction. We then identify triples representing cause-and-effect relationships and encode this content using the Biological Expression Language (BEL). Finally, domain experts perform an analysis of the completeness, relevance, accuracy, and novelty of the extracted triples.

Results: In our test scenario, which is focused on the KG regarding the phosphorylation of the Tau protein, our pipeline successfully contributed novel data, which was then subsequently used to update the KG leading to the identification of six additional upstream regulators of Tau phosphorylation.

Conclusion: Here, it is demonstrated that the NLP-based workflow we created is capable of rapidly updating pathophysiology mechanism graphs. As a result, production-scale, semi-automated updating of pre-existing, curated mechanism graphs is enabled.

1. Introduction

Biomedical literature is a valuable source of information; typically expressing high-level knowledge in the form of scientific narratives often in an unstructured format. Presently, publications are growing at an unprecedented rate, therefore, staying up to date on the latest findings by manually reading the literature is no longer feasible for a given field of study. Thus, automated extraction of relevant information by applying intelligent natural language processing (NLP) algorithms has become a crucial and non-trivial task [1]. One of the most difficult challenges in biomedical information extraction and literature mining is the development of effective methods and strategies for transforming scientific text into structured knowledge representations such as ontologies and knowledge graphs (KGs) [2]. Biomedical KGs are a powerful means of representing biomedical concepts and relations in the form of nodes and edges within networks. They have been utilized in a range of applications including identifying protein functions [3], prioritizing genes associated with disease [4], as well as drug discovery [5]. Drug discovery is the process of finding new potential medications, and conventionally it is an extremely time-consuming and costly procedure. Biomedical KGs

* Corresponding author.
E-mail address: martin.hofmann-apitius@scai.fraunhofer.de (M. Hofmann-Apitius).

https://doi.org/10.1016/j.ailsci.2023.100078
Received 1 June 2023; Accepted 12 June 2023
Available online 13 June 2023

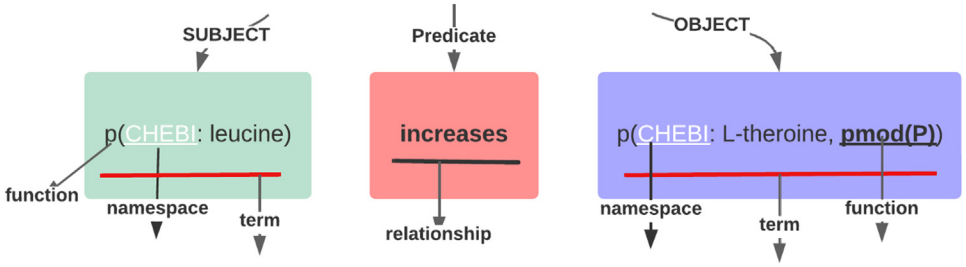

Fig. 1. Example of a BEL statement encoding a biological causal relationship from the following sentence: "Leu markedly increased phosphorylation of Thr residues 36, 47, and 70 on 4E-binding protein (BP)1 in muscle from rats not receiving EtOH" [26]. Different parts of a BEL statement such as function, namespace, or relationship type, are illustrated in different colors.

have been essential for the acceleration of data-driven drug discovery and drug repurposing as they can incorporate a broad range of biological scales and concepts [6]. For example, Domingo-Fernández et al. developed an algorithm that prioritizes drugs for a given disease by reasoning over causal paths in a KG [6]. In the context of drug discovery, Knowledge Graph Embedding (KGE) models have also been explored and utilized frequently [5]. KGE models learn vector representation of graph nodes and edges in low dimensional space and embed entities and relations of knowledge graphs while keeping the original structural information (such as the relations between entities) [7]. In their study, Bonner et al. investigated a variety of KGE models to assess how different factors such as the training setup or their hyperparameters can result in better prediction in drug discovery tasks [5]. In this regard, N. Sosa et al. [8] implemented a literature-based KGE method with drug repurposing application in rare diseases.

Inherently, knowledge evolves rapidly in many research domains of current interest, and thus an efficient update procedure is in high demand for KG maintenance [9]. Generally, biomedical KGs can be generated in multiple ways, ranging from manually aggregating curated databases [2,10,11] to techniques such as NLP [12–14]. To cope with the increased number of scientific papers published every day, NLP techniques are quickly becoming the primary method by which rapid overviews of thousands of publications can be generated in seconds. NLP is often compromised of a combination of other technologies and techniques such as named entity recognition (NER) [15], relation extraction (RE) [16], and document or sequence classification [17]. These core methods have also been widely utilized in domains such as cancer research and clinical decision support [18]. Pre-trained language models such as BERT/BioBERT are currently among the most widely employed language representation models that have proven to be exceptionally powerful in biomedical text mining [19,20]. As an example, in a recent work, a BERT-based model was applied for generation of the high-quality molecular representations in drug discovery issues [21]. BERT models can employ the encoder part of the transformer architecture so that the attention mechanism can then be utilized for learning context and relationships in sequential data [20]. BERT is trained on text corpora of general domains, whereas BioBERT, which is an extension of BERT, is further re-trained on domain-specific biomedical corpora including PubMed database [19].

To represent scientific findings in a computable format, modeling of a data exchange language should be considered. Extensible Markup Language (XML) [22], Biological Pathways Exchange (BioPAX) [23], Systems Biology Markup Language (SBML) [24], and Biological Expression Language (BEL) [25] are currently among the most widely utilized language models in systems biology. BEL was developed to assemble causal relationships among biological entities with the underlying meta data such as support evidence [23,24]. Furthermore, BEL makes use of the concept of namespaces (e.g., Human Gene Nomenclature (HGNC)) to flexibly normalize entities. Fig. 1 depicts a more detailed example of a BEL statement. More information about BEL language specifics is provided in the Supplementary file.

Further research was recently conducted to enable the automatic extraction and translation of biological relationships directly from scientific papers into BEL. The results of these studies can be subsequently utilized to improve the development and updating of causal KGs in biomedical domains [27,28]. In the context of drug discovery, KGs representing causality are becoming essential analytical and inference tools that can help in areas like target identification [29]. One recent example of a cause-and-effect KG is the Human Brain Pharmacome (HBP), a drug-target mechanism-oriented data model that combines knowledge from multiple sources [30,31]. Pharmacomes link drug-target information to specific pathophysiology mechanisms and are ideal for rational approaches toward the modulation of pathophysiology mechanisms [32].

Although the need for constant maintenance of KGs may appear trivial, the regular updating of "pathways" and other biological associations remains a challenge for biocurators and biomedical knowledge workers. Though in most cases, changes brought on by evolution and the completion of knowledge occur gradually and smoothly [33], a fully automated KG updating procedure imposes a substantial challenge to the biomedical research community, given the rapid pace of publishing in the field. Hence, our work had a singular, pragmatic focus: How can a tool be developed for semi-automatic updating of the mechanism and disease-specific physiology KGs starting with limited data? In this work, we begin with the HBP and, due to its significance in neurodegenerative diseases like Alzheimer's disease (AD), we focus on the phosphorylated Tau (pTau) as a starting point. Next, we introduce and employ an NLP-based workflow to extract the BEL relationships from relevant literature contained in the PubMed database and verify the correctness and relevance of the extracted triples by engaging experts before adding the newly encoded information to the KG. Finally, we investigate the NLP workflow's shortcomings by analyzing what is missing during information extraction and attempt to improve it.

2. Related work

Retrieval of essential information from text, and therefore mapping, creation, and extension of KGs are gaining more attention in biomedical research. Santos et al. [34], as an example, have presented a clinical KG that supports the harmonization of proteomics with other omics data while integrating biomedical databases and scientific publications. In another work, Thomas et al. [35] have recently developed a framework for modeling causal activities called Gene Ontology Causal Activity Modeling (GO-CAM) to represent more complex statements about biological functions in a structured and scalable manner. Gene-product interactions are modeled qualitatively and causally using GO-CAM, which can in turn help answer complex questions about how gene product activities interact to carry out biological functions [35]. Their research also reveals a growing trend toward modeling cause-and-effect relationships in the context of gene ontology. Therefore, there are several substantial efforts being made towards the construction of ontologies and causal KGs, most of which can be broadly classified as manual or automatic/semi-automatic and are summarized below [33,36].

2.1. Construction of KGs with manual curation

Biomedical KGs can be constructed from manually curated databases which contain pre-existing data that can be merged into a single KG

[10,11,37–39]. Manual construction of KGs requires that experts continuously read publications to annotate and identify relevant relationships or concepts of interest, which does ultimately result in a high degree of accurate information, but at the expense of time. One of the characteristics of KGs built in this manner is that they contain precise data, although the recall is low [2,40–43].

In a recent investigation, Lage-Rupprecht et al. [31], developed the Human Brain Pharmacome (HBP), a manually and highly-curated drug-mechanism-oriented model that encompasses a wide range of biological processes relevant to the pathophysiology of AD. The HBP includes a highly granular representation of mechanisms describing tauopathies, a group of neurodegenerative diseases distinguished by the deposition of the abnormal microtubule-associated Tau protein in the brain in the form of neurofibrillary tangles [31]. In their work, they established a comprehensive framework in silico for identifying druggable mechanisms in the context of Tau phosphorylation (pTau) and created a workflow for evaluating likely therapeutics that target AD-specific pathophysiology processes. Briefly, they first created an initial version of AD-specific KG using the eBEL software package [44] which generated a network using a combination of manually curated, AD-focused BEL relations as well as established biological pathways derived from public repositories including KEGG [45] and Reactome [46]. Using this KG, they selected pTau as an entry point and identified upstream regulators (secondary interactors) in the network which modulate that protein [31]. The secondary interactors were then ranked according to defined criteria (including draggability), and the top hits were selected for further evaluation. In the case of the Tau drug repurposing example, these biochemical assays led to the discovery of novel target/compound combinations modulating posttranslational modification of the Tau protein [31]. Since the introduction of HBP, several studies have worked on drug discovery for AD [47–52], which can in turn highlight the necessity for keeping HBP up-to-dated with recent findings.

Though manual curation is a labor-intensive task and makes the generation of KGs an expensive undertaking, it is still required for the creation of the highest quality KGs. However, to accelerate KG construction, automatic/semi-automatic approaches are greatly desirable.

2.2. Automatic/semi-automatic approaches for information extraction

The automated and large-scale generation of biomedical KGs can greatly benefit several fields of research, including drug repurposing and drug-disease identification. In this regard, Xu et al. [53] created a pattern matcher system to recognize words in PubMed abstracts that indicate drug-disease therapies. Most automatic information retrieval systems and KG creation approaches contain NER and relation extraction (RE) subtasks in their pipelines. Regarding this, in an investigation, Kocaman et al. [15] presented a deep learning framework that performs noticeably well on biomedical benchmarks for NER task. Deep neural network topologies have also been proposed regularly for RE to improve the prediction performance on several standard benchmarks [16,34,54–56]. An article by Tudor et al. [57] explored the use of text-mining algorithms to detect protein phosphorylation events in relevant research publications. Likewise, a knowledge discovery framework was similarly designed to extract protein-protein interaction, gene-disease and drug-disease interaction from biomedical scientific publications via text mining [58,59]. Specifically, Wang et al. [60] examined the text mining approaches, modeling resources, systems, and shared tasks introduced following the emergence of Coronavirus pandemic (COVID-19). In their review, the qualitative description and evaluation of each system's performance can reveal how these models can assist and provide information, such as in the detection of novel features, or the potential to link scientific articles and clinical trials [60].

In our work presented here, we describe our solution for a rapid and efficient extension of an existing mechanism KG using NLP technologies. In the following section, we provide a detailed description of the primary data resources and methodologies that are adopted.

Table 1
The total number of BEL edges, BEL nodes, distinct PMIDs, and distinct BEL triples in Tau KGs and training dataset.

	Tau KG	Training dataset
BEL edges	5,704	45,992
BEL nodes	4,219	17,400
Distinct PMIDs	256	2,750
Distinct BEL triples	5,173	39,099

Table 2
Statistics of reach keyword search in PubMed in the timeframe of 2020-2022.

Keyword Num	Abstracts
Tau	11,596
Tau phosphorylation	1,180
Tau phosphorylation AND post-translational modification	65

3. Materials and methods

3.1. Datasets

3.1.1. Raw training dataset

To train the RE module of the NLP pipeline, we utilized a training dataset generated by combining several published KGs. These KGs were built by expert manual curation on various topics such as AD [28,61–68], Parkinson's Disease [68], epilepsy [69], amyotrophic lateral sclerosis (unpublished), type 2 diabetes [70,71], post-traumatic stress disorder (unpublished), traumatic brain injury (unpublished), schizophrenia, bipolar disorder (unpublished), COVID-19 [62,72,73], and pTau modulation [31]. The BEL statements underlying these KGs are extracted from PubMed abstracts as well as full-text documents. Table 1 summarizes the information regarding the total number of BEL nodes, edges, distinct PMIDs, and unique BEL statements in this training dataset.

3.1.2. Tau KG

One subgraph of HBP, which focuses on the context of pTau, serves as the primary disease-specific BEL causal KG that we sought update using our workflow [31]. Table 1 summarizes the key information present in this Tau KG: it contains a total of 4,219 BEL nodes, 5,704 BEL edges, and 5,173 unique BEL triples all of which were derived from 256 different publications identified using their PubMed Identifiers (PMID).

3.1.3. Extension corpus

The extension corpus was created for updating and expanding the Tau KG. To create this collection, we used the E-utilities API [74] to search and retrieve abstracts from PubMed using "Tau" as the primary keyword, and extracted all abstracts published between May 2020 and May 2022. The distribution of abstracts belonging to the context of Tau phosphorylation and its post-translational modification are shown in Table 2.

3.2. Knowledge graph expansion workflow

A pictorial overview of the KG expansion workflow is provided in Fig. 2 and consists of the following steps:

- Starting with the initial mechanism KG, we query PubMed for relevant medical subject headings (MeSH) terms and keywords consistent with the context of the KG. Here, we extracted PubMed abstracts dealing with Tau phosphorylation and its post-translational modification. The abstracts extracted from this step constitute the extension corpus.
- Next, an NLP-based BEL extractor pipeline that is pre-trained to extract BEL relationships from biomedical text is applied to the extension corpus.

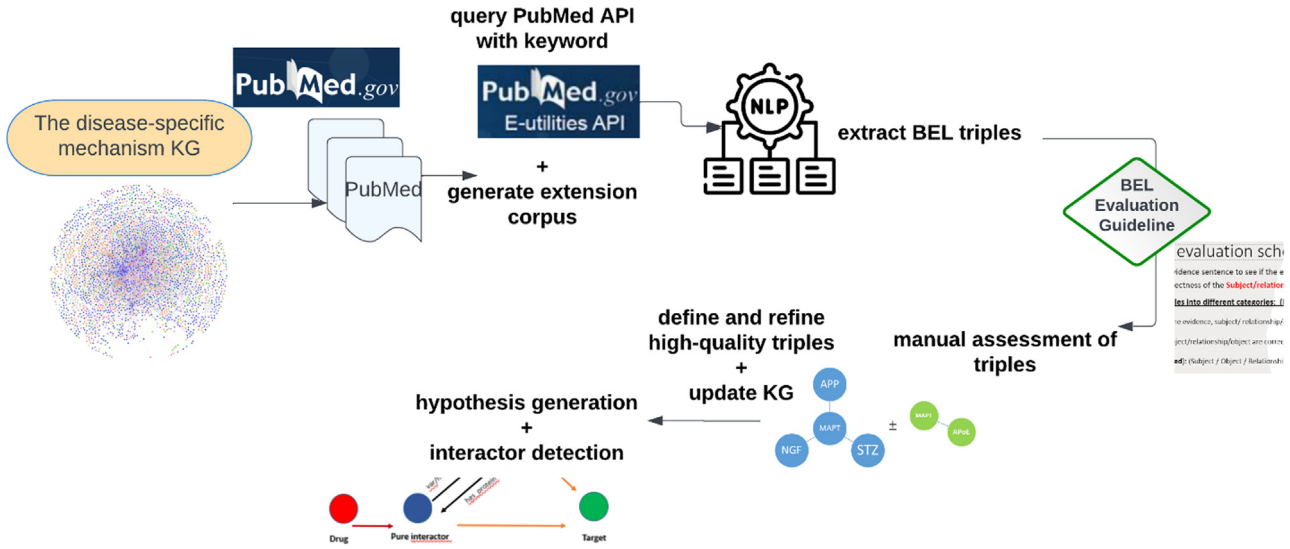

Fig. 2. Overview of NLP-based disease-specific mechanism KG extension.

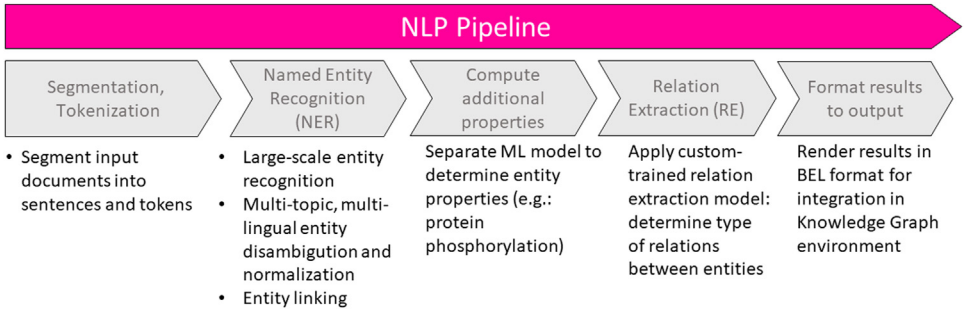

Fig. 3. A high-level overview of the subtasks within the NLP pipeline.

- The extracted BEL statements are then manually checked for correctness and completeness by domain experts and only the high-quality triples are added to the KG.
- Finally, by applying graph-based mining and analysis, we search for novel regulators of pTau in the updated KG.

3.3. Training and development of NLP workflow

To extract BEL statements from unstructured text and test the problem-solving capacity in translational biomedicine application examples, we partnered with the Kairntech research and product development team [75]. We employed the Kairntech Sherpa workflow, which is a web-based collaborative and completely open source platform that provides essential NLP requirements for the extraction of cause-and-effect statements [76]. In this pipeline, the NLP workflow is subdivided into multiple subtasks as represented in Fig. 3. First, after tokenizing and segmenting the input text, NER is applied and biomedical entities are identified in the text. The extracted entities are then linked to their specific database identifiers through a named entity disambiguation (NED) approach. In Sherpa, NER and NED are integrated into the same component using a machine learning tool named Entity Fishing [77]. Then, to determine specific properties of the recognized entities (e.g., protein modifications), a dedicated model is trained and inserted into the NLP pipeline. In the following step, a neural network-based RE module known as OpenNRE [78] is used to determine the type of relations between entities. OpenNRE is an open-source tool that allows one to perform RE in different settings such as at sentence-level or at bag-level. In this work, we used sentence-level RE and utilized the default BERT-base-uncased tokenizer [20,78] as the basis for our entity encoder. The RE model was then fine-tuned using the recognized entities and their corresponding sentences from the BEL training corpus which was formerly described.

Finally, the pipeline encodes results in the BEL format for subsequent integration into the KG. Before the training began, a randomly selected subset of the training corpus was defined and set aside to monitor the training procedure.

3.4. Evaluation scheme

Automated extraction of BEL triples from text is a difficult task due to the high level of expressiveness in the biomedical domain. Besides, there are some limitations of the BEL syntax, terms, and functions [79], moreover, not all information encoded in a BEL document can be directly found in the evidence text. There may also be several different perspectives on how or which specific information should be coded in BEL [79]. Therefore, our goal was to create flexible evaluation guidelines and a schema to assess the performance of the BEL extraction module so that a BEL statement can be accurately labelled based on how "correct" it is. As depicted in Fig. 4, this evaluation rubric assigns one of six categories to each triple according to how accurately the BEL statement describes the associated support evidence. This schema also helps to determine the level of correctness of each extracted statement and to better understand where the NLP pipeline can be improved when extracting BEL relationships. Based on these guidelines, we selected triples that were of high-quality to be further evaluated. Domain experts then analyzed and filtered the high-quality triples, and this final subset was subsequently added to the Tau KG. Table 3 showcases a few examples of extracted BEL triples and their associated correctness label. These correctness labels are defined as the following:

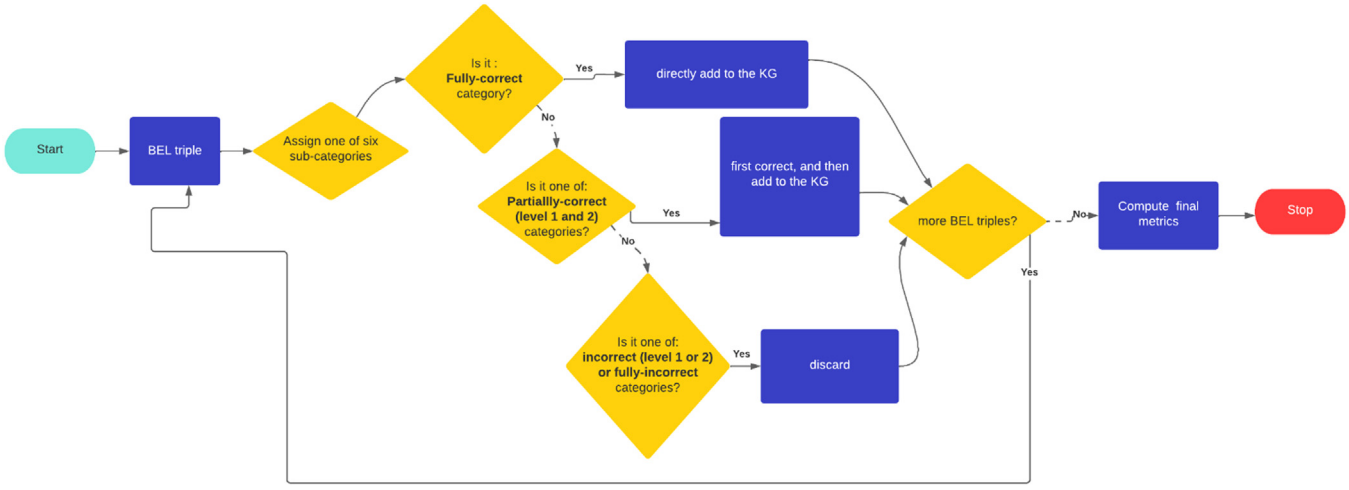

Fig. 4. The evaluation scheme for checking the quality of extracted BEL triples to be added to the KG.

Table 3
Examples of extracted triples and their assigned label.

Extracted BEL triple	Support evidence	Assigned label	Comment
a(CHEBI:"Magnolol") **increases** *p(HGNC:"CHRM1")*	Magnolol contributed to the activation of the cAMP/PKA/CREB pathway through enhancing CHRM1 level [80].	**Fully correct**	All parts are correct.
path(MESHD:"preeclampsia" **positive_correlation** *a(CHEBI:"beta-amyloid"*	Beta-amyloid(1-40)/(1-42) ratio was significantly higher in HELLP syndrome than in severe preeclampsia (8.49 + 2.73 vs. 4.71 + 1.65; $p = 0.007$) [81].	**Partially correct level-1**	Modification needed: beta-amyloid should be mentioned as: amyloid-beta.
path(MESHD:"major depression") **positive_correlation** *path(MESHD:"Alzheimer disease"*	Late-life major depression (LLMD) is a risk factor for the development of mild cognitive impairment and dementia, including Alzheimer's disease (AD) and vascular dementia [82].	**Partially correct level-2**	Relationship can be "association" instead of "positive correlation". According to MESH, "Major depression" should be encoded as "Depressive Disorders, Major".
p(HGNC:"C3") ***positive_correlation*** *p(HGNC:"MAPT")*	We did not find any significant association of C3 with the AD biomarkers Aβ42 reflecting plaque pathology, P-tau related to tau pathology or the neurodegeneration biomarker T-tau [82].	**Partially incorrect level-1**	Relationship is wrong.
a(CHEBI:"insulin") ***increases*** *a(CHEBI:"advanced glycation end products")*	In the central nervous system, insulin has crucial regulatory roles, while chronic hyperglycemia leads to formation and accumulation of advanced glycation end products (AGEs) [83].	**Partially incorrect level-2**	Subject is wrong. It should be "Hyperglycemia".
a(CHEBI:"sevoflurane") ***association*** *p(HGNC:"APOE")*	This research aimed to reveal the role of ApoE in the pathogenesis of cognitive deficiency caused by sevoflurane anesthesia and the protective mechanism of CoQ10 in a multiple sevoflurane treatment model of young mice [84].	**Fully incorrect**	No cause-and-effect fact is given in the support evidence.

- **Fully correct:** The subject, relationship, object, function, and syntax are all correct.
- **Partially correct level-1**: The subject, relationship, object, function, and syntax are correct, but some minor modification (e.g. in the BEL function) is needed.
- **Partially correct level-2**: The subject, object, function, and syntax are correct, but some knowledge is missing such as in the relationship type (for instance, if the system misinterprets the use of positive correlation instead of association).
- **Partially incorrect level-1**: The relationship is wrong.
- **Partially incorrect level-2:** Either the subject, object or both are wrong.
- **Fully incorrect**: There are no causal relations included in the support evidence.

Table 4
Performance of the NLP model on the test dataset.

Relationship type	Precision	Recall	F-score
Increase	74%	62%	67%
decrease	59%	66%	62%
directly increase	60%	33%	42%
directly decrease	60%	42%	49%
regulate	66%	61%	63%
association	65%	54%	58%
positive correlation	64%	61%	62%
negative correlation	43%	53%	47%
biomarker for	66%	40%	50%
has component	100%	33%	50%
is a	61%	30%	40%
cause no change	34%	37%	35%

4. Results

4.1. Performance of NLP pipeline on the test dataset

A 10% of the training dataset was randomly selected to create the test dataset to assess the performance of the NLP pipeline on the unseen data. The precision, recall, and F-score [85] for each relationship type are represented in Table 4. Furthermore, a manual analysis of the results by domain experts revealed that the performance of the model on the BEL triple extraction is comparable to other relationship mining approaches (Alpha Tom Kodamullil et al., unpublished).

4.2. Performance of NLP pipeline on the extension corpus

The performance of the NLP pipeline on the extension corpus is summarized in Table 5. More detailed information, such as the distribution of each relationship type or most common namespaces occurrences, is described in the Supplementary file (Figs. S1- S8). To assess the correct-

Table 5
Details of the BEL statements extracted by NLP pipeline from the extension corpus.

	Count
BEL edges	3,161
BEL nodes	4,219
Distinct PMIDs	1,499
Distinct BEL triples	2,051

Table 6
The result of the first manual evaluation team.

Triple Category	Count
Fully correct	379
Partially correct level 1	460
Partially correct level2	69
Partially incorrect level 1	53
Partially incorrect level 2	205
Fully incorrect	272
Total annotated BEL triples	1,438

Table 7
The result of the second manual annotation team and the rate of inter-annotator agreement.

Triple Category	Count
Fully correct	97
Partially correct level 1	312
Partially correct level2	67
Partially incorrect level 1	25
Partially incorrect level 2	270
Fully incorrect	229
Total annotated BEL triples	1,000
Inter-annotator agreement	37%

ness of the extracted triples, a BEL curation team consisting of four annotators with both biology and data science backgrounds was formed. We then randomly selected more than 50% of the extracted triples and the team manually evaluated the quality of this dataset based on the above-defined evaluation criteria. The result of this evaluation is represented in Table 6. Additionally, the inter-annotator agreement rate for the entire dataset was computed by three BEL-expert annotators with several years of experience in BEL curation. For this calculation, 1,000 annotated triples were randomly selected and given to this second BEL expert team who then computed the percentage of annotations that match across all annotators. In this study, the inter-annotator agreement rate, meaning the percentage of BEL triples that were annotated with the same sense by all annotators, stands as 371 out of 1000 triples. Finally, we chose a selection of high-quality triples as decided by subject-matter experts for inclusion in the KG. Table 7 illustrates the result of the second round of manual evaluation as well as the inter-annotator agreement rate.

4.3. Enrichment analysis

The visualization of the updated BEL KG centered around pTau is represented in Supplementary File (Fig. S9). Moreover, this KG can also be visualized with the Biomedical Knowledge Miner (BiKMi) web tool, which is accessible at https://bikmi.pharmacome.scaiview.com/. After the high-quality BEL triples were added to the KG, we used BiKMi and identified six additional interactors that are related to the modulation of Tau phosphorylation. Briefly, we first chose the target, which is the Tau protein, and only selected the entities that have phosphorylation as the protein modification. Then, we identified upstream regulators of the modified target by considering the direct causal relations. Next, we selected interactors that are druggable, meaning those for which there exist drugs from DrugBank [86] that target them. These upstream interactors are apolipoprotein E (ApoE), Legumain (LGMN), phosphodiesterase 11A (PDE11A), protein kinase cGMP-dependent 1v (PRKG1), Alpha-synuclein (SNCA) and TANK-binding kinase 1 (TBK1). Among them, only ApoE, PDE11A, and TBK1 were shown to be druggable. Finally, we searched several sources and databases, including protein-protein interaction networks, gene-disease databases, and pathway databases to validate and enrich the information about the identified potential targets against neurodegenerative diseases, including AD. We have summarized this information in the following sections.

4.3.1. Interactor TBK1

TBK1 is a protein-coding gene that has been associated with diseases such as encephalopathy, acute frontotemporal dementia, and amyotrophic lateral sclerosis 4 [87]. TBK1 is also shown to be involved in autophagy and inflammatory responses, too [88]. More specifically, loss-of-function (LoF) mutations in the TBK1 gene, including frameshift mutations and inflame amino acid deletions have been associated to a variety of neurodegenerative diseases [88]. However, the role of TBK1 in the modification of tau phosphorylation linked to AD or other tauopathies, is not yet well-understood [89]. In a recent study, H.Abreha et al. employed immunoaffinity enrichment coupled with mass spectrometry (MS) and identified TBK1 in human cortical brain tissue as a regulator of tau hyperphosphorylation [89]. In a related study, Xiang et al. summarized the knowledge on inhibitors of TBK1 that may be candidates for therapeutic approaches in many diseases, such as inflammation or metabolic diseases or pancreatic cancer [90]. However, no TBK1 kinase inhibitor is currently authorized by the US Food and Drug Administration (FDA).

4.3.2. Interactor Apolipoprotein E (ApoE)

ApoE gene is known to be a major genetic risk factor of late-onset AD, with the ApoE $^{*}\varepsilon 4$ allele being shown to be a high-risk element, while the ApoE $^{*}\varepsilon 2$ allele serving as a potential neuroprotective variant [91]. Following this, Yu et al. reported in a recent study that ApoE may be protective in neuronal activity, but its toxic fragments are linked to neurodegeneration and neurocognitive impairment among those with AD [92]. Their study also highlights that the differences in ApoE fragments expressed in the hippocampus of mice of different ages lead to variability in Tau phosphorylation and neurocognitive functions and mechanisms [92]. Various animal studies have demonstrated that pathological (polymerized) Tau drives cerebral atrophy, however, the mechanisms that exacerbate Tau pathology or Tau hyper-phosphorylation in ApoE4-carriers are not known [93]. R. Saroja et al., claim that the astrocyte-secreted protein glypican-4 is a key driver of ApoE4-mediated Tau abnormal hyperphosphorylation [93]. A variety of therapeutic approaches on targeting ApoE4 for preventing or treating Alzheimer's disease have been highlighted in a recent review by Williams et al. [94]. However, because the mechanisms underlying ApoE isoform-specific differences in brain homeostasis are complex, it remains unclear which ApoE-related therapeutic approaches are applicable in vivo [95].

4.3.3. Interactor PDE11A

PDE11A is a protein-coding gene that is related to primary pigmented nodular adrenocortical disease and is associated with the GPCR downstream signaling pathway [87]. In a recent study, exome sequencing performed by Qin et al., revealed that the PDE11A variant led to significantly higher Tau phosphorylation levels and therefore PDE11A could be a candidate gene for early-onset AD [96]. Previously, two rare mutations in the phosphodiesterase PDE11A gene were identified in people with early-onset AD [97]. Their study confirmed significantly decreased protein levels of PDE11A in brain samples of AD patients [97]. Furthermore, the expression of PDE11A variants was proven to be associated with the increased Tau hyperphosphorylation in context of AD [97]. In another research, Fawcett et al. used molecular cloning to characterize PDE11A and reported that it is responsive to zaprinast, dipyridamole, and the nonselective PDE inhibitor 3-isobutyl-1-

methylxanthine (IBMX) [98]. It has also been observed that each of the four PDE11A splice variations (PDE11A1-4) appears to have a different tissue expression profile as well as a separate N-terminal regulatory area, implying that each isoform could be targeted independently with a small molecule or biologic [99,100]. Though PDE11A inhibitors and activators have been identified so far, more examinations in patient tissue must yet be performed for the establishment of a therapeutic indication [99].

5. Discussion

5.1. A practical and efficient system for the extension of mechanism KGs

By utilizing AI-based relationship extraction workflows, we demonstrate that regular updating of highly curated KGs is feasible. The relationship-extraction workflow Sherpa is an open-source and freely accessible platform that combines different machine learning and NLP techniques to enhance biomedical text analysis procedures. Initially, the corpus for KG extension was extracted by searching PubMed for research abstracts involving Tau phosphorylation. Then, the NLP-based module performed NER, NED and RE on this extension corpus to identify all BEL relationships embedded within. As a final step, manual assessment was performed based on defined BEL evaluation guidelines to assess and refine the extracted BEL triples for further integration in the initial KG. In our example scenario that is highly relevant for drug repurposing in neurodegenerative diseases, the subsequent analysis of the updated KG revealed previously unreported interactors related to the modulation of Tau phosphorylation. While the results of our overall approach are promising, we acknowledge that substantial work is needed to improve the efficacy of the underlying NLP-based module. Though the development of complete autonomous systems for extraction of cause-and-effect triples is a complex task for which there is no near-optimal prediction model yet.

5.2. Principal findings achieved by the BERT-based NLP pipeline

At the core of Sherpa NLP workflow stands the Entity-Fishing tool, a machine learning service that aims to automate entity recognition and disambiguation generically while minimizing domain-specific restrictions. To perform entity disambiguation, a supervised machine learning technique based on Random Forest and Gradient Tree Boosting exploiting multiple features was utilized. The Sherpa pipeline includes an open-source and extensible tool named OpeNRE, which provides a framework for the implementation of neural models for RE based on TensorFlow and PyTorch, and is trained with a BERT-based encoder. The performance and error analysis revealed the success of the BERT-based BEL triple extraction module in the rapid expansion of the Tau KG. By mining the updated KG, our investigations and enrichment analysis yielded six potential target candidates relevant to the regulation of Tau phosphorylation concerning neurodegenerative diseases. However, additional biological screening systems are undoubtedly required to learn more about the role of these interactors in modulation of Tau phosphorylation in the context of AD.

5.3. Hypothesis generation and validation of identified regulators

For identifying the novel upstream regulators of pTau, we applied a graph-based algorithm using BiKMi, an in-house software package which was designed to explore pathways and interactions within a BEL network. Currently, there is an ongoing demand for ML-based approaches to improve not only the hypothesis generation itself, but also the speed at which it is performed. For instance, the path ranking algorithm (PRA) is an alternative that can help with knowledge reasoning and knowledge recommendation tasks [101]. PRA computes the feature matrix on a pair of graph nodes and edges and assigns a score to each triple. As the KG representations are used to predict links that may indicate potential relations between entities in the KG, hypothesis generation approaches can be influenced by a variety of factors, such as how well the KG is constructed or how well the KGE algorithm works. In this regard, Zeng et al. [102] provided a review of hypothesis generation approaches for drug discovery using KGs and explain that designing and validating automated workflows which can generate hypotheses, is still a massive challenge in the biomedical domain. Our work aims to meet this challenge and present a viable method to overcome the hurdles of hypothesis generation.

To validate and enrich our understanding of the found interactor candidates, we searched a variety of gene, protein, and disease databases for additional information. By comparing proteins, protein families, and protein-protein interaction networks, one can generate a wealth of information about the relationship between proteins within a genome or across different species and related to different diseases. Additionally, human disease databases are also an essential component of biological research that aid in understanding the importance and context of these proteins. Among the identified interactors described in the results section, a major allelic variant of ApoE gene, ApoE $\varepsilon 4$, is likely the most well-known risk factor for late onset AD [95]. However, it is unclear whether ApoE $\varepsilon 4$ is linked to increased tau phosphorylation and neurofibrillary tangle development, both of which are markers of AD and cause structural changes in the neuronal cytoskeleton [103]. Interestingly, Zhou et al. investigated genotype-specific effect of ApoE on tau phosphorylation in mice and found that the ApoE $\varepsilon 4$ genotype increases this type of phosphorylation via the calpain-CDK5 signaling pathway [103]. Following the fact that modulations of ApoE are related to and promising in AD drug development, Ymazaki et al. [95] summarized the associated therapies in three main categories: 1) regulation of ApoE quantity; 2) modification of ApoE properties; and 3) indirect therapeutic approaches. Despite extensive research conducted regarding the role of ApoE in neurodegeneration, no therapeutic interventions has been successfully found that targets the ApoE gene directly [95]. This could be due to the complicacy of mechanisms underlying ApoE isoform-specific differences on brain homeostasis [95]. However, though ApoE is not directly targeted, the ApoE genotype is expected to affect the responses in several potential AD therapies [95].

5.4. Corpora generation and possible improvement strategies

Despite the efficiency of the RE module, more thorough evaluations and adaptations of our workflow are still required to further improve performance. One area that can be revised in our proposed methodology is the strategy for searching for relevant abstracts and publications. As stated previously, we used specific keywords and looked through PubMed database via their API, however, even with a sophisticated search phrase, querying PubMed may not always return pertinent articles in their entirety [104]. To improve the search strategy, Subramanian et al. [104] introduced a shallow filtering approach to retrieve relevant literature in the context of drug repurposing. Previously, this matter was also investigated, and a systematic methodology was introduced to develop better strategies and enhance biomedical literature searching [105]. Moreover, there are now alternative platforms such as A2A [106] which support searching through biomedical publications. Such tools will also be considered in our future investigations toward more efficient context-specific corpora generation [106].

5.5. Model optimization strategies for an enhanced BEL relation extraction

Training deep learning models for BEL triple extraction is difficult due to the small amount of BEL training data available. One way to alleviate this and allow for a more easily generated annotated dataset, would be the use of distant supervision techniques [107]. Distant supervision, also known as weak supervision, can help with generating

labeled datasets, though these labels may also contain noise [108]. Interestingly, active learning approaches have recently been shown to reduce the labeling effort required for learning a RE model. This is accomplished by incorporating a human annotator into the learning loop and carefully selecting a small number of samples to be labeled [109]. Active learning methodologies are especially useful when there are plenty of unlabeled, available, and easy-to-access samples, but labeled data is rare and costly to provide [110]. As instance, in their work, Rosales et al. [111] used active learning to extract and classify clinical concepts. Ideally, we would in the future design and utilize an active learning framework combined with weak supervision to maximize the BEL triple extraction performance by minimizing the number of samples that require domain-expert manual BEL annotation.

Another possible solution for improving RE performance would be to combine the co-occurrence-based strategies in our pipeline. In a recent study, a scoring method was previously proposed that relies on entity co-occurrences to extract relationships from biomedical literature [112]. In their work, they considered how often two entities appear together and aggregated these counts over the whole corpus [112]. Co-occurrence based RE approaches are very advantageous since they do require annotated training yet maintain a relatively high recall rate. In our future work, we will implement these strategies to enhance our platform and overcome its limitations.

6. Conclusion

This work demonstrated the feasibility of an NLP-based pipeline for the semi-automated updating of a highly curated KG. Our strategy is based on a pragmatic approach to rapidly generate biomedical corpora enriched with recent publications which are then fed into an NLP workflow for BEL statement detection. Despite the efficiency of our pipeline, clearly there is a demand for accelerated manual BEL annotation as well as an expansion of training datasets for NLP workflow improvement.

We obviously did not solve the problem of full automation of the KG updating process. However, we foresee that soon complete triple indices will be available for entire PubMed (including related full-text literature collections). This will make finding “related” and “recent” triples merely a graph query task. We are currently working on such “triple index” for entire indication areas and it is only a matter of time before we will see the first PubMed-wide cause-and-effect indices.

Moreover, the breathtaking dynamics we currently observe in the field of large language models (LLMs) open new perspectives for the interplay between LLMs and knowledge graphs and we are looking forward to the first BART or ChatGPT plugins that support the updating of knowledge graphs. The work here sheds a spotlight on the challenge of updating knowledge graphs; however, it needs to be seen as part of a much broader development in the field of language, knowledge and their utility.

Data and code availability

The data and materials such as the initial Tau KG, new high-quality triples and the scripts for analyzing the Tau KG from BiKMi, are publicly available from our GitHub repository https://github.com/SCAI-BIO/tau-kg-extension-nlp.

Funding

This work was supported by the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI) and the Bonn-Aachen International Center for Information Technology (b-it) foundation, Bonn, Germany.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link for the used code and data in the main manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100078.

References

[1] Zhao S, Su C, Lu Z, Wang F. Recent advances in biomedical literature mining. Brief Bioinform 2021;22(3). doi:10.1093/bib/bbaa057.
[2] Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 2020;18:1414–28. doi:10.1016/j.csbj.2020.05.017.
[3] Crichton G, Guo Y, Pyysalo S, Korhonen A. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. BMC Bioinform 2018;19(1):176. doi:10.1186/s12859-018-2163-9.
[4] Hu J, Lepore R, Dobson RJB, Al-Chalabi A, Bean DM, Iacoangeli A. DGLinker: flexible knowledge-graph prediction of disease–gene associations. Nucleic Acids Res 2021;49(W1) W153–W161. doi:10.1093/nar/gkab449.
[5] Bonner S, et al. Understanding the performance of knowledge graph embeddings in drug discovery. Artif Intell Life Sci 2022;2:100036. doi:10.1016/j.ailsci.2022.100036.
[6] Domingo-Fernández D, et al. Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. PLOS Comput Biol 2022;18(2):e1009909. doi:10.1371/journal.pcbi.1009909.
[7] Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models. Brief Bioinform 2021;22(2):1679–93. doi:10.1093/bib/bbaa012.
[8] Sosa DN, Derry A, Guo M, Wei E, Brinton C, Altman RB. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. Pac Symp Biocomput Pac Symp Biocomput 2020;25:463–74.
[9] Wewer C, Lemmerich F, Cochez M. Updating embeddings for dynamic knowledge graphs; 2021. [Online]. Available http://arxiv.org/abs/2109.10896 .
[10] Liu H, Song Y, Guan J, Luo L, Zhuang Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. BMC Bioinform 2016;17(Suppl 17):539. doi:10.1186/s12859-016-1336-7.
[11] “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information - PubMed.” https://pubmed.ncbi.nlm.nih.gov/28924171/ (Accessed 4 October, 2022).
[12] Xu D, et al. DTMiner: identification of potential disease targets through biomedical literature mining. Bioinformatics 2016;32(23):3619–26. doi:10.1093/bioinformatics/btw503.
[13] Exploiting graph kernels for high performance biomedical relation extraction. J Biomed Semant 2022. Full Text https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0168-3 Accessed 13 October.
[14] Warikoo N, Chang Y-C, Hsu W-L. LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task. Database 2018;2018 bay108. doi:10.1093/database/bay108.
[15] V. Kocaman and D. Talby, “Biomedical Named Entity Recognition at Scale.” arXiv, Nov. 12, 2020. doi: 10.48550/arXiv.2011.06315.
[16] Zhou D, Zhong D, He Y. Biomedical relation extraction: from binary to complex. Comput Math Methods Med 2014;2014.
[17] Harish BS, Udayasri B. Document classification: an approach using feature clustering. In: Thampi SM, Abraham A, Pal SK, Rodriguez JMC, editors. Recent advances in intelligent informatics. Advances in intelligent systems and computing. Cham: Springer International Publishing; 2014. p. 163–73. doi:10.1007/978-3-319-01778-5_17.
[18] Zhu F, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform 2013;46(2):200–11. doi:10.1016/j.jbi.2012.10.007.
[19] “BioBERT: a pre-trained biomedical language representation model for biomedical text mining | Bioinformatics | Oxford Academic.” https://academic.oup.com/bioinformatics/article/36/4/1234/5566506?login=false (Accessed 29 May 2023).
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi:10.48550/arXiv.1810.04805.
[21] B. Fabian et al., “Molecular representation learning with language models and domain-relevant auxiliary tasks.” arXiv, Nov. 26, 2020. doi:10.48550/arXiv.2011.13230.
[22] Achard F, Vaysseix G, Barillot E. XML, bioinformatics and data integration. Bioinformatics 2001;17(2):115–25. doi:10.1093/bioinformatics/17.2.115.
[23] Demir E, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol 2010;28(9):9. doi:10.1038/nbt.1666.
[24] Hucka M, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003;19(4):524–31. doi:10.1093/bioinformatics/btg015.
[25] “PyBEL: a computational framework for Biological Expression Language | Bioinformatics | Oxford Academic.” https://academic.oup.com/bioinformatics/article/34/4/703/4557184 Accessed 4 October 2022.

[26] Lang CH, et al. Alcohol impairs leucine-mediated phosphorylation of 4E-BP1, S6K1, eIF4G, and mTOR in skeletal muscle. Am J Physiol-Endocrinol Metab 2003;285(6) E1205–E1215.. doi:10.1152/ajpendo.00177.2003.

[27] Madan S, et al. The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track. Database 2019;2019:baz084. doi:10.1093/database/baz084.

[28] Hoyt CT, et al. Re-curation and rational enrichment of knowledge graphs in Biological Expression Language. Database 2019;2019:baz068. doi:10.1093/database/baz068.

[29] MacLean F. Knowledge graphs and their applications in drug discovery. Expert Opin Drug Discov 2021;16(9):1057–69. doi:10.1080/17460441.2021.1910673.

[30] "Human Brain PHARMACOME," Fraunhofer Institute for Algorithms and Scientific Computing SCAI. https://www.scai.fraunhofer.de/en/projects/Human-Brain-Pharmacome.html (Accessed 13 October 2022).

[31] Lage-Rupprecht V, et al. A hybrid approach unveils drug repurposing candidates targeting an Alzheimer pathophysiology mechanism. Patterns 2022;3(3). doi:10.1016/j.patter.2021.100433.

[32] Hofmann-Apitius M. PHARMACOMES: cause-and-effect models linking drugs, targets and disease mechanisms; 2022. doi:1014293/S2199-rexpo22006v1.

[33] N. Abbas et al., "Knowledge Graphs Evolution and Preservation – A Technical Report from ISWS 2019," ArXiv201211936 Cs, Dec. 2020, Accessed: Mar. 25, 2022. [Online]. Available: http://arxiv.org/abs/2012.11936

[34] Santos A, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol 2022;40(5):5. doi:10.1038/s41587-021-01145-6.

[35] Thomas PD, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet 2019;51(10):1429–33. doi:10.1038/s41588-019-0500-1.

[36] Tamašauskaitė G, Groth P. Defining a knowledge graph development process through a systematic review. ACM Trans Softw Eng Methodol 2022. doi:10.1145/3522586.

[37] Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. PLOS Comput Biol 2012;8(7):e1002574. doi:10.1371/journal.pcbi.1002574.

[38] Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform 2012;45(2):363–71. doi:10.1016/j.jbi.2011.11.017.

[39] Himmelstein DS, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife 2017;6:e26726. doi:10.7554/eLife.26726.

[40] Salwinski L, et al. Recurated protein interaction datasets. Nat Methods 2009;6(12):860–1. doi:10.1038/nmeth1209-860.

[41] Cusick ME, et al. Literature-curated protein interaction datasets. Nat Methods 2009;6(1):39–46. doi:10.1038/nmeth.1284.

[42] Keseler IM, et al. Curation accuracy of model organism databases. Database J Biol Databases Curation 2014;2014 bau058. doi:10.1093/database/bau058.

[43] Müller H-M, Van Auken KM, Li Y, Sternberg PW. Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. BMC Bioinform 2018;19(1):94. doi:10.1186/s12859-018-2103-8.

[44] "Christian Ebeling Schultz" Bruce, "ebel: e(BE:L) - validation and extension of BEL networks." Accessed: Oct. 17, 2022. [OS Independent]. Available: https://github.com/e-bel/ebel

[45] Kanehisa M. The KEGG database. In: 'In silico' simulation of biological processes. John Wiley & Sons, Ltd; 2002. p. 91–103. doi:10.1002/0470857897.ch8.

[46] "Reactome Pathway Knowledgebase | Nucleic Acids Research | Oxford Academic." https://academic.oup.com/nar/article/46/D1/D649/4626770 (Accessed 14 April 2023).

[47] "Hyperphosphorylated tau (p-tau) and drug discovery in the context of Alzheimer's disease and related tauopathies - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S135964462300003X?via%3Dihub (Accessed 8 May 2023).

[48] Khachaturian AS, Dengel A, Dočkal V, Hroboň P, Tolar M. Accelerating innovations for enhanced brain health. Can artificial intelligence advance new pathways for drug discovery for Alzheimer's and other neurodegenerative disorders? J Prev Alzheimers Dis 2023;10(1):1–4. doi:10.14283/jpad.2023.1.

[49] Spooner A, Mohammadi G, Sachdev PS, Brodaty H. A. Sowmya, and for the Sydney Memory and Ageing Study and the Alzheimer's Disease Neuroimaging Initiative, "Ensemble feature selection with data-driven thresholding for Alzheimer's disease biomarker discovery,". BMC Bioinformatics Jan 2023;24(1):9. doi:10.1186/s12859-022-05132-9.

[50] Kostidis S, Sánchez-López E, Giera M. Lipidomics analysis in drug discovery and development. Curr Opin Chem Biol 2023;72:102256. doi:10.1016/j.cbpa.2022.102256.

[51] K. Savva, M. Zachariou, M. Bourdakou, N. Dietis, and G. M. Spyrou, "DReAmocracy: a method to capitalize on prior drug discovery efforts to highlight candidate drugs for repurposing." bioRxiv, p. 2023.01.12.523717, Jan. 16, 2023. doi:10.1101/2023.01.12.523717.

[52] Jabeen F, et al. Deep learning-based prediction of inhibitors interaction with Butyrylcholinesterase for the treatment of Alzheimer's disease. Comput Electr Eng Jan 2023;105:108475. doi:10.1016/j.compeleceng.2022.108475.

[53] Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics Jun 2013;14(1):181. doi:10.1186/1471-2105-14-181.

[54] Peng Y, Lu Z. Deep learning for extracting protein-protein interactions from biomedical literature. In: BioNLP 2017. Vancouver, Canada: Association for Computational Linguistics; 2017. p. 29–38. doi:10.18653/v1/W17-2304.

[55] Kavuluru R, Rios A, Tran T. Extracting drug-drug interactions with word and character-level recurrent neural networks. In: 2017 IEEE international conference on healthcare informatics (ICHI); 2017. p. 5–12. doi:10.1109/ICHI.2017.15.

[56] Björne J, Salakoski T. biomedical event extraction using convolutional neural networks and dependency parsing. In: Proceedings of the BioNLP 2018 workshop. Melbourne, Australia: Association for Computational Linguistics; 2018. p. 98–108. doi:10.18653/v1/W18-2311.

[57] Tudor CO, Ross KE, Li G, Vijay-Shanker K, Wu CH, Arighi CN. Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. Database J Biol Databases Curation 2015;2015 bav020. doi:10.1093/database/bav020.

[58] Song M, Kim WC, Lee D, Heo GE, Kang KY. PKDE4J: Entity and relation extraction for public knowledge discovery. J Biomed Inform Oct 2015;57:320–32. doi:10.1016/J.JBI.2015.08.008.

[59] Yuan J, et al. Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowl Inf Syst 2020;62(1):317–36. doi:10.1007/s10115-019-01351-4.

[60] Wang LL, Lo K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. Brief Bioinform Mar 2021;22(2):781–99. doi:10.1093/bib/bbaa296.

[61] Khanam Irin A, Kodamullil ATom, Gündel M, Hofmann-Apitius M. Computational modelling approaches on epigenetic factors in neurodegenerative and autoimmune diseases and their mechanistic analysis. J Immunol Res 2015;2015:737168. doi:10.1155/2015/737168.

[62] Kodamullil AT, Younesi E, Naz M, Bagewadi S, Hofmann-Apitius M. Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. Alzheimers Dement J Alzheimers Assoc Nov 2015;11(11):1329–39. doi:10.1016/j.jalz.2015.02.006.

[63] Naz M, Kodamullil AT, Hofmann-Apitius M. Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases. Brief Bioinform 2016;17(3):505–16. doi:10.1093/bib/bbv063.

[64] Domingo-Fernández D, et al. Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. Bioinforma Oxf Engl 2017;33(22):3679–81. doi:10.1093/bioinformatics/btx399.

[65] Kodamullil AT, Iyappan A, Karki R, Madan S, Younesi E, Hofmann-Apitius M. Of mice and men: comparative analysis of neuro-inflammatory mechanisms in human and mouse using cause-and-effect models. J Alzheimers Dis JAD 2017;59(3):1045–55. doi:10.3233/JAD-170255.

[66] Emon MAEK, Kodamullil AT, Karki R, Younesi E, Hofmann-Apitius M. Using drugs as molecular probes: a computational chemical biology approach in neurodegenerative diseases. J Alzheimers Dis JAD 2017;56(2):677–86. doi:10.3233/JAD-160222.

[67] "A Systems Biology Approach for Hypothesizing the Effect of Genetic Variants on Neuroimaging Features in Alzheimer's Disease - PubMed." https://pubmed.ncbi.nlm.nih.gov/33554913/ (Accessed 4 October 2022).

[68] Younesi E, et al. PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. Theor Biol Med Model 2015;12:20. doi:10.1186/s12976-015-0017-y.

[69] Hoyt CT, Domingo-Fernández D, Balzer N, Güldenpfennig A, Hofmann-Apitius M. A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities. Database J Biol Databases Curation 2018;2018. doi:10.1093/database/bay050.

[70] R. Karki, S. Madan, Y. Gadiya, D. Domingo-Fernández, A. T. Kodamullil, and M. Hofmann-Apitius, "Data-driven modeling of knowledge assemblies in understanding comorbidity between type 2 diabetes Mellitus and Alzheimer's disease," J Alzheimers Dis, vol. 78, no. 1, pp. 87–95, doi:10.3233/JAD-200752.

[71] Karki R, Kodamullil AT, Hofmann-Apitius M. Comorbidity analysis between Alzheimer's disease and type 2 diabetes Mellitus (T2DM) based on shared pathways and the role of T2DM Drugs. J Alzheimers Dis JAD 2017;60(2):721–31. doi:10.3233/JAD-170440.

[72] Schultz B, et al. A method for the rational selection of drug repurposing candidates from multimodal knowledge harmonization. Sci Rep 2021;11(1):11049. doi:10.1038/s41598-021-90296-2.

[73] Hopp M-T, et al. Linking COVID-19 and heme-driven pathophysiologies: a combined computational–experimental approach. Biomolecules 2021;11(5):644. doi:10.3390/biom11050644.

[74] Sayers E. A General Introduction to the E-utilities. 2009 May 26 [Updated 2022 Nov 17]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK25497/

[75] Geißler S. Natürliche Sprachverarbeitung und Künstliche Intelligenz – ein wachsender Markt mit vielen Chancen: Das Beispiel Kairntech. Inf - Wiss Prax 2020;71(2–3):115–18. doi:10.1515/iwp-2020-2079.

[76] Geißler S. The Kairntech Sherpa – An ML Platform and API for the Enrichment of (not only) Scientific Content. In: Proceedings of the 1st international workshop on language technology platforms. Marseille, France: European Language Resources Association; 2020. p. 54–8. Accessed: Dec. 12, 2022. [Online]. Available. https://aclanthology.org/2020.iwltp-1.9 .

[77] P. Lopez, "entity-fishing." Oct. 08, 2022. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/kermitt2/entity-fishing

[78] "OpenNRE." THUNLP, Oct. 18, 2022. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/thunlp/OpenNRE

[79] Rinaldi F, et al. BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language. Database J Biol Databases Curation 2016;2016. doi:10.1093/database/baw067.

[80] Zhu G, Fang Y, Cui X, Jia R, Kang X, Zhao R. Magnolol upregulates CHRM1 to attenuate Amyloid-β-triggered neuronal injury through regulating the cAMP/PKA/CREB pathway. J Nat Med 2022;76(1):188–99. doi:10.1007/s11418-021-01574-2.

[81] Lederer W, et al. Cerebrospinal beta-amyloid peptides(1-40) and (1-42) in severe preeclampsia and HELLP syndrome - a pilot study. Sci Rep 2020;10(1):5783. doi:10.1038/s41598-020-62805-2.
[82] Pillai A, et al. Complement component 3 levels in the cerebrospinal fluid of cognitively intact elderly individuals with major depressive disorder. Biomark Neuropsychiatry 2019;1:100007. doi:10.1016/j.bionps.2019.100007.
[83] Shieh JC-C, Huang P-T, Lin Y-F. Alzheimer's disease and diabetes: insulin signaling as the bridge linking two pathologies. Mol Neurobiol 2020;57(4):1966–77. doi:10.1007/s12035-019-01858-5.
[84] Yang M, Tan H, Zhang K, Lian N, Yu Y, Yu Y. Protective effects of Coenzyme Q10 against sevoflurane-induced cognitive impairment through regulating apolipoprotein E and phosphorylated Tau expression in young mice. Int J Dev Neurosci Off J Int Soc Dev Neurosci 2020. doi:10.1002/jdn.10041.
[85] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Losada DE, Fernández-Luna JM, editors. Advances in information retrieval. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2005. p. 345–59. doi:10.1007/978-3-540-31865-1_25.
[86] Wishart DS, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res Jan 2008;36 no. Database issueD901-906. doi:10.1093/nar/gkm958.
[87] "GeneCards - Human Genes | Gene Database | Gene Search." https://www.genecards.org/ Accessed 1 March 2023).
[88] Verheijen J, et al. Common and rare TBK1 variants in early-onset Alzheimer disease in a European cohort. Neurobiol Aging 2018;62:245 e1-245.e7. doi:10.1016/j.neurobiolaging.2017.10.012.
[89] Abreha MH, et al. TBK1 interacts with tau and enhances neurodegeneration in tauopathy. J Biol Chem 2021;296:100760. doi:10.1016/j.jbc.2021.100760.
[90] Xiang S, et al. TANK-binding kinase 1 (TBK1): An emerging therapeutic target for drug discovery. Drug Discov Today 2021;26(10):2445–55. doi:10.1016/j.drudis.2021.05.016.
[91] Liao F, Yoon H, Kim J. Apolipoprotein E metabolism and functions in brain and its role in Alzheimer's disease. Curr Opin Lipidol Feb 2017;28(1):60–7. doi:10.1097/MOL.0000000000000383.
[92] Yu Y, Yang M, Zhuang X, Pan J, Zhao Y, Yu Y. Effects of toxic apolipoprotein E fragments on Tau phosphorylation and cognitive impairment in neonatal mice under sevoflurane anesthesia. Brain Behav 2022;12(8):e2702. doi:10.1002/brb3.2702.
[93] Saroja SR, Gorbachev K, TCW J, Goate AM, Pereira AC. Astrocyte-secreted glypican-4 drives APOE4-dependent tau hyperphosphorylation. Proc Natl Acad Sci 2022;119(34):e2108870119. doi:10.1073/pnas.2108870119.
[94] Williams T, Borchelt DR, Chakrabarty P. Therapeutic approaches targeting Apolipoprotein E function in Alzheimer's disease. Mol Neurodegener Jan 2020;15(8). doi:10.1186/s13024-020-0358-9.
[95] Yamazaki Y, Painter MM, Bu G, Kanekiyo T. Apolipoprotein E as a therapeutic target in alzheimer's disease: a review of basic research and clinical evidence. CNS Drugs 2016;30(9):773–89. doi:10.1007/s40263-016-0361-4.
[96] Qin W, et al. Exome sequencing revealed PDE11A as a novel candidate gene for early-onset Alzheimer's disease. Hum Mol Genet 2021;30(9):811–22. doi:10.1093/hmg/ddab090.
[97] wei qin et al., "Two Novel PDE11A Genetic Variants Increase Tau Phosphorylations in Early-onset Alzheimer's Disease," In Review, preprint, Aug. 2020. doi:10.21203/rs.3.rs-60929/v1.
[98] L. Fawcett et al., "Molecular cloning and characterization of a distinct human phosphodiesterase gene family: PDE11A".
[99] Kelly PM. Does Phosphodiesterase 11A (PDE11A) hold promise as a future therapeutic target? Curr Pharm Des 2015;21(3):389–416. doi:10.2174/1381612820666140826114941.
[100] Kelly MP. A role for phosphodiesterase 11A (PDE11A) in the formation of social memories and the stabilization of mood. Adv Neurobiol 2017;17:201–30. doi:10.1007/978-3-319-58811-7_8.
[101] Mazumder S, Liu B. Context-aware path ranking for knowledge base completion. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence; 2017. p. 1195–201. doi:10.24963/ijcai.2017/166.
[102] Zeng X, Tu X, Liu Y, Fu X, Su Y. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol Feb 2022;72:114–26. doi:10.1016/j.sbi.2021.09.003.
[103] "APOE4 Induces Site-Specific Tau Phosphorylation Through Calpain-CDK5 Signaling Pathway in EFAD-Tg Mice."
[104] S. Subramanian et al., "A Natural Language Processing System for Extracting Evidence of Drug Repurposing from Scientific Publications," p. 6.
[105] Bramer WM, De Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc 2018;106(4). doi:10.5195/jmla.2018.283.
[106] Rybinski M, Karimi S, Nguyen V, Paris C. A2A: a platform for research in biomedical literature search. BMC Bioinformatics 2020;21(19):572. doi:10.1186/s12859-020-03894-8.
[107] W. Hogan, "An Overview of Distant Supervision for Relation Extraction with a Focus on Denoising and Pre-training Methods." arXiv 2022. doi:10.48550/arXiv.2207.08286.
[108] Ravikumar K, Liu H, Cohn JD, Wall ME, Verspoor K. Literature mining of protein-residue associations with graph rules learned through distant supervision. J Biomed Semant 2012;3(3) S2. doi:10.1186/2041-1480-3-S3-S2.
[109] A. Primpeli, "Reducing the labeling effort for entity resolution using distant supervision and active learning," p. 242.
[110] Naseem U, Khushi M, Khan SK, Shaukat K, Moni MA. A comparative analysis of active learning for biomedical text mining. Appl Syst Innov 2021;4(1):1. doi:10.3390/asi4010023.
[111] Rosales R, Krishnamurthy P, Rao RB. Semi-supervised active learning for modeling medical concepts from free text. In: Sixth international conference on machine learning and applications (ICMLA 2007). Cincinnati, OH, USA: IEEE; 2007. p. 530–6. doi:10.1109/ICMLA.2007.103.
[112] Junge A, Jensen LJ. CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision. Bioinformatics 2020;36(1):264–71. doi:10.1093/bioinformatics/btz490.