Knowledge-Based Systems 315 (2025) 113280

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Research hypothesis generation over scientific knowledge graphs

Agustín Borrego [a], Danilo Dessì [b], Daniel Ayala [a], Inma Hernández [a], Francesco Osborne [c,d], Diego Reforgiato Recupero [e,*], Davide Buscaldi [f], David Ruiz [a], Enrico Motta [c]

[a] University of Seville, Seville, Spain
[b] Department of Computer Science, College of Computing and Informatics, University of Sharjah, Sharjah, United Arab Emirates
[c] Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom
[d] Department of Business and Law, University of Milano Bicocca, Milan, Italy
[e] Mathematics and Computer Science Department, University of Cagliari, Cagliari, Italy
[f] Laboratoire d'Informatique de Paris Nord, Sorbonne Paris Nord University, Paris, France

ARTICLE INFO

Keywords:
Hypothesis generation
Knowledge graphs
Link prediction
Scholarly domain
Scientific facts
Artificial Intelligence

ABSTRACT

Generating research hypotheses is a crucial step in scientific investigation that involves the creation of precise, verifiable, and logically valid statements that can be empirically examined. Therefore, many efforts have been made to automate or assist this process through the use of various Artificial Intelligence solutions. However, most existing methods are tailored to very specific domains, particularly within the biomedical field. There have been recent attempts to formalize hypothesis generation as a link prediction task over knowledge graphs. This solution is potentially domain-independent and applicable across diverse disciplines. Nevertheless, current approaches for link prediction, which typically rely on embedding models or path-based methods, have shown limited success in accurately predicting new hypotheses. To address these limitations, this paper introduces ResearchLink, an innovative and domain-independent methodology for hypothesis generation over knowledge graphs. ResearchLink combines path-based features and knowledge graph embeddings with text embeddings, capturing the semantic context of entities within a given corpus, and integrates additional information from bibliometric databases to improve research collaboration predictions. To conduct a rigorous evaluation of ResearchLink, we constructed CSKG-600, a new dataset for hypothesis generation, consisting of 600 statements that were manually labeled by domain experts. ResearchLink achieved outstanding performance (78.7% P@20), significantly outperforming alternative approaches such as TransH (71.8%), TransD (71.7%), and RotatE (70.7%).

1. Introduction

Research hypothesis generation is a fundamental process in scientific inquiry that involves the formulation of specific, testable, and logically sound hypotheses that can be subjected to empirical testing. Therefore, many efforts have been made to automate or assist this process through the use of various Artificial Intelligence (AI) solutions. Indeed, the automatic generation of research hypotheses represents one of the most intriguing tasks in the field of AI as it encompasses both a creative aspect, wherein computers attempt to emulate the intuition of human researchers, and a predictive component, where they strive to anticipate future developments within a specific domain [1–3]. Furthermore, integrating AI-powered hypothesis generation into scientific practices can have a transformative impact, allowing scientists to effectively combine multidisciplinary information, accelerating the process of knowledge discovery, and leading to breakthroughs that might have otherwise remained elusive.

The research community has been actively addressing the challenge of research hypothesis generation for several years [4,5]. Current approaches generally focus on proposing novel and non-trivial associations between existing concepts (e.g., chemical compounds and diseases) that could be interpreted as testable research hypotheses. The resulting hypotheses are then evaluated by human researchers, who choose the most promising ones for further investigation. However, most existing methods are tailored to highly specific domains, such as detecting new associations between drugs and diseases, particularly within the biomedical field [6–8]. This narrow specialization

* Corresponding author.
E-mail addresses: borrego@us.es (A. Borrego), ddessi@sharjah.ac.ae (D. Dessì), dayala1@us.es (D. Ayala), inmahernandez@us.es (I. Hernández), francesco.osborne@open.ac.uk (F. Osborne), diego.reforgiato@unica.it (D. Reforgiato Recupero), davide.buscaldi@lipn.univ-paris13.fr (D. Buscaldi), druiz@us.es (D. Ruiz), enrico.motta@open.ac.uk (E. Motta).

https://doi.org/10.1016/j.knosys.2025.113280
Received 25 August 2023; Received in revised form 25 November 2024; Accepted 26 February 2025
Available online 8 March 2025

significantly limits their versatility and applicability to other research areas.

Recently, there have been some efforts to formalize hypothesis generation as a *link prediction task* over *knowledge graphs* of research concepts [3,9]. This solution is potentially domain-independent and applicable across diverse disciplines, provided they offer a machine-readable representation of the essential domain entities and their connections.

Knowledge graphs (KGs) are data structures used for organizing and representing structured information semantically [10]. By effectively capturing relationships between entities and attributes, KGs facilitate the provision of machine-actionable insights into various intelligent services [11]. The relationship between two entities is typically formalized as a triple (sometimes called statement) in the form of <*subject, predicate, object*>, e.g., <*Bill Gates, founderOf, Microsoft*>, <*Artificial_Intelligence, subtopicOf, Computer_Science*>. In recent years, there has been a notable emergence of KGs specifically dedicated to representing research knowledge. These KGs typically contain a set of entities in this space (e.g., tasks, methods, evaluation techniques, datasets, proteins, chemicals), their relations, and the relevant actors (e.g., authors, organizations) and documents (e.g., articles, books) [11–18]. This resources can be further enriched by integrating the numerous knowledge organization systems developed across scientific disciplines to describe research fields at different levels of granularity [19].

Link prediction involves the identification of additional relationships between entities in a graph [20,21]. It is commonly used for *knowledge graph completion*, whose aim is to discover missing links in the graph that may be absent due to various factors, including gaps in the process of constructing the KG [22]. In the context of hypothesis generation, link prediction solutions can utilize the contextual knowledge from a KG to identify promising but previously unknown connections between entities that may represent valid hypotheses. For example, a common pattern in Computer Science involves the reuse of algorithms developed for specific tasks on different ones. A notable instance of this concept is the recent application of BERT [23], a state-of-the-art model for Natural Language Processing, on protein prediction tasks, such as protein family classification and protein interaction prediction [24]. The intuition behind this discovery can be represented as a new relationship of type *usesMethod* between the *Protein_Prediction* task and the *BERT* method.

The standard solutions for link prediction over KGs are typically based on knowledge graph embedding (KGE) models [25,26] (e.g., $TransE$ [27], $TransH$ [28], $TransD$ [29], $RotatE$ [30]), path-based approaches [31,32], or graph convolutional networks (e.g., $RGCN$ [33]). However, while these solutions have demonstrated proficiency in KG completion tasks, their performance diminishes when tasked with predicting links that are not necessarily entailed by current knowledge, but may emerge in the future as research hypotheses. This limitation arises from two main factors. First, many statements produced by these techniques are trivial and do not necessarily qualify as hypotheses. For example, the triple `<deep_learning, uses, electricity>`, although true, does not represent a meaningful research hypothesis worth exploring. It is simply a piece of information absent from our current representation. Furthermore, predicting future links may necessitate incorporating data beyond the KG, such as the occurrences of relevant entities in bibliometric databases and their semantic meanings inferred from the scientific literature.

To address these issues, we introduce ResearchLink, an innovative and domain-independent methodology for hypothesis generation over KGs. ResearchLink combines path-based features and KGEs with text embeddings, which capture the semantic context of an entity within a given corpus, and additional information about the occurrences of relevant entities in the literature. It is adaptable to any domain that provides a KG of relevant entities.

To enable rigorous evaluation, we have developed *CSKG-600*, a new dataset for research hypothesis generation. Since ResearchLink is designed for generating hypotheses rather than identifying missing facts, it cannot be assessed using standard KG completion benchmarks such as FB15k and WN18RR. *CSKG-600* contains 600 hypotheses derived from the Computer Science Knowledge Graph (CS-KG)[1] [11] that were manually reviewed by three senior research scientists. CS-KG is an open large-scale knowledge graph comprising over 41M statements about 10M entities extracted automatically from 6.7M scientific articles. ResearchLink achieves remarkable performance without using any domain-specific features and outperforms several alternative approaches.

In summary, the main contributions of this paper are the following:

- We propose ResearchLink, a domain-independent methodology for research hypothesis generation over knowledge graphs.
- We release a new dataset for hypothesis generation: *CSKG-600*, including 600 manually annotated triples.
- We report an evaluation comparing ResearchLink against several alternative methods and analyze the impact of the novel features proposed in this paper on its overall performance.
- We make available the full source code of ResearchLink at https://github.com/dayala1/ResearchLink so that it can be easily reused in other fields.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 describes ResearchLink in detail. Section 4 reports the evaluation. Section 5 discuss the limitations of the proposed approach. Finally, Section 6 concludes the paper and presents future directions of research.

2. Related work

Automated hypothesis generation has been extensively researched in the past decades, especially in the biomedical field [34]. In the following subsection, we will summarize related work on literature-based discovery (Section 2.1), knowledge graph-based discovery (Section 2.2), and LLM-based discovery (Section 2.3). Finally, in Section 2.4 we will present an overview of the key contributions to automated hypothesis generation and discuss the research gap that we aim to address in this paper.

2.1. Literature-based discovery

Hypothesis generation can be seen as a special application of Literature-based discovery (LBD), which involves the inference of previously unseen knowledge from the automated or semi-automated analysis of large text corpora from research articles [34–36]. Most current systems focus on inferring a single type of relation between a pair of medical terms, e.g., the relation between genes and the biological functions that they regulate, or the relation between diseases and drugs that could help cure them (also known as drug discovery [37–39]). For that matter, different text mining and NLP techniques have been used in the past, such as graph diffusion [8,40], probabilistic search [7], or NLP semantic analysis [41,42]. A primary limitation of these methodologies is their predominant focus on biomedical [6–8,40] or chemical [42] data. While there are emerging attempts to adapt these techniques to other research domains, such endeavors are still in an early stage [41]. Some recent initiatives have sought to adapt LBD beyond the biomedical domain, specifically for the recommendation of scientific papers to researchers [43]. Notably, this differs from automated hypothesis generation, as its primary focus is to connect researchers with potentially relevant articles.

2.2. Knowledge graph embeddings

The automatic generation of research hypotheses can be addressed by identifying new semantic connections between concepts in the

[1] CS-KG - http://w3id.org/cskg.

Table 1
Summary of key relevant contributions for automated hypothesis generation.

Year	Reference	Domain	Approach	Technique
2014	Spangler et al. [40]	Biomedical	LBD	Graph diffusion
2015	Hristovski et al. [41]	Biomedical	LBD	NLP
2016	Workman et al. [6]	Biomedical	LBD	Human-assisted graph traversals
2017	McCusker et al. [7]	Biomedical	LBD	Probabilistic search
2018	Choi et al. [8]	Biomedical	LBD	Bag of words, Graph diffusion
2019	Tshitoyan et al. [42]	Chemical	LBD	Word2Vec embeddings
2020	Akujuobi et al. [51]	Biomedical	KG Embeddings	Node pairs embeddings
2021	Haan et al. [3]	Social Sciences	KG Embeddings	ComplEx embeddings
2022	Zhou et al. [52]	Biomedical	KG Embeddings	Temporal node pairs embeddings
2024	Park et al. [53]	Material sciences	LLM	Human-curated queries to GPT-4
2024	Elbadawi et al. [54]	Pharmaceutical	LLM	Queries to GPT-4 and DALL-E

scientific domain, such as techniques or tasks. These concepts and relationships can be represented as nodes and edges in a knowledge graph. Recent years have witnessed the rise of several knowledge graphs for capturing scholarly metadata such as authors, venues, citations, and research topics (e.g., AIDA [44], AMiner [45], Scholarly-data.org [46], OpenCitations [47], CORE [48]). An increasing number of knowledge graphs focus specifically on describing specific scientific entities and their relationships (AI-KG [14], CS-KG [11], ORKG [15], UMLS [49], TKG [17], or Nanopublications [50]).

Once scholarly information is organized into a knowledge graph, the task of inferring a new hypothesis can be framed as a link prediction problem, a subject for which numerous solutions exist in the literature. Early approaches to link prediction concentrated on analyzing multi-hop paths within graphs, serving as a means to identify potential direct relations between entities. These proposals rely on different strategies to traverse the graph, namely: random walks [55–57], deterministic graph iteration [31,32], or reinforcement learning [58–61]. Random walk-based approaches, despite being able to traverse the graph more efficiently, are prone to leave behind potentially relevant paths, which makes them less appealing in the hypothesis generation scenario. Some of the previous authors introduced embedding models in their techniques, in order to improve their performance [32,57,62,63]. Actually, recent trends in link prediction have centered on KG embedding models [25,26,64,65]. These models generate embedded representations of KG entities and relations by executing various transformations within an embedding space [66–72]. The generated embedding space is then utilized to assess the plausibility of a potential relationship between two entities. This is based on the assumption that entities intended to have a specific connection should exhibit closer proximity within the embedding space. As an example, the use of the ComplEx embedding model has recently been used for assessing research hypotheses, yielding promising results [3].

Entities and relationships within the research field exhibit distinct characteristics, including specialized technical terminology and unique syntax. This has motivated the development of different types of KG embeddings based on word embeddings (such as Word2Vec [73] or BERT [74]). Based on these representations, some proposals have been able to infer relations between concepts using embeddings that capture the semantic static association between nodes in the KG [51], or the dynamic evolution of these associations [52]. These proposals are only able to infer generic unnamed relations, whose nature has to be subsequently determined by human experts. Additionally, the topology of these graphs has some unique features, which have been addressed by a number of proposals of specific KG embeddings [21,75–79]. However, the application of these embeddings to hypothesis generation has not yet proved to provide sufficiently good performance.

2.3. Large Language Models

Large Language Models (LLMs) are having a transformative impact on a wide range of AI applications in research. These include identifying relevant papers [80–82], assisting or even automating the creation of literature reviews [83], enhancing academic writing and referencing [84,85], developing specialized conversational agents [86, 87], and much more. This growing impact has motivated researchers to explore the potential of LLMs for generating research hypotheses. However, this remains an open area of investigation, requiring further validation and refinement. Some preliminary experiments have been conducted to generate research hypotheses using the well-known GPT-4 model in the materials chemistry domain [53], and in the pharmaceutical domain [54]. The main conclusion is that, while the model can generate several new hypotheses by combination or hybridization of existing research works, a significant amount of these hypotheses is easily discardable as nonsense, being the constant corrective feedback from the human expert necessary to guide the conversation toward a reasonable and interesting hypothesis. Also, the significantly high energy consumption of this kind of systems is highlighted as another main challenge for future research.

2.4. Overview and limitations

In Table 1, we summarize the main proposals for automated hypothesis generation. The majority of previous approaches have dealt with the generation hypothesis as a LBD problem, analyzing large corpora of academical articles and applying different NLP and text mining techniques to infer the hypotheses. The main limitation of these proposals is that they are domain-centric, and therefore not suitable for their application in other scientific domains. KG embeddings-based models offer a more easily generalizable approach. However, while there exist a large number of proposals for KG-based link prediction, these approaches primarily focus on retrospective prediction. In other words, they infer missing links that should exist in the KG because they reflect accurate information within the current state of the subject area. Consequently, these models struggle to achieve comparable effectiveness when predicting future occurrences, which often require supplementary information. Finally, LLMs have shown promise in fields such as supporting literature reviews [83] and hold potential as technologies for research hypothesis generation [53,54]. However, we are still in the early stages of exploring their capabilities in this area. Current versions still suffer from hallucinations, necessitating constant human supervision, which makes them less appealing for automated hypothesis generation.

ResearchLink fills a significant research gap by framing hypothesis generation as link prediction while incorporating a set of features

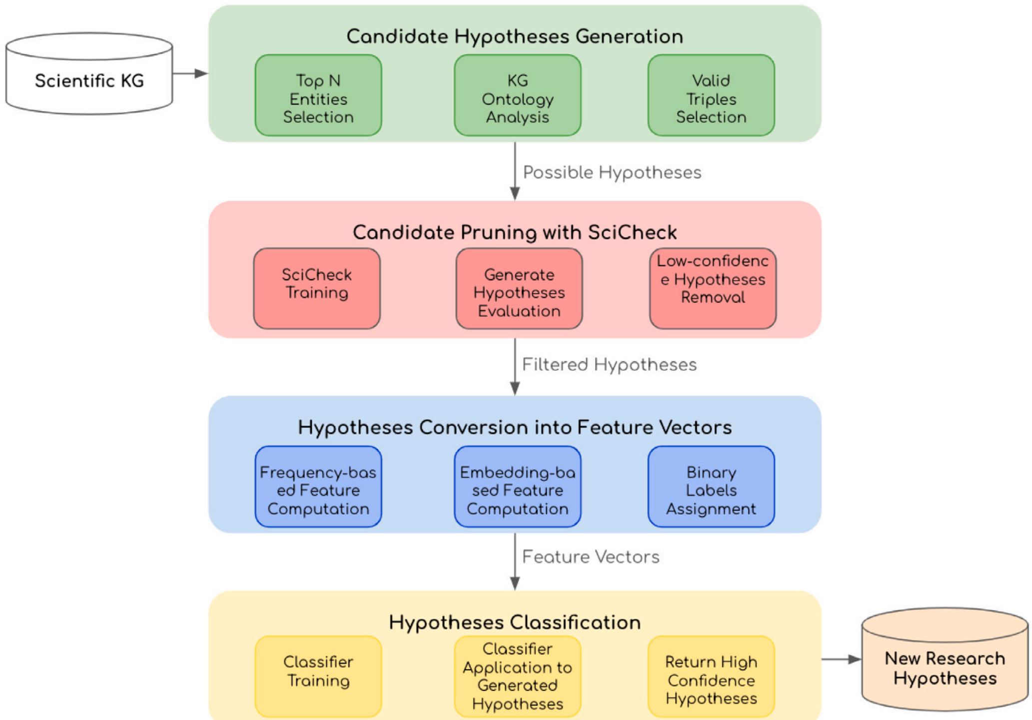

Fig. 1. The proposed approach to generate research hypotheses from a scientific KG.

that overcome the typical limitations of link prediction approaches in this domain. This is done by combining a deterministic path-based approach that leverages the structural features of the graph with state-of-the-art KG and word embeddings and additional information reflecting the frequency of relevant entities in the literature. The resulting approach is thus applicable across a variety of diverse domains and relations. Our experimental results reveal that the full version of ResearchLink, which includes this novel combination of features, outperforms traditional approaches for link prediction and KG completion.

3. Methodology

This paper introduces ResearchLink, a new domain-independent hypothesis generation methodology. ResearchLink takes in input a KG that describes the relations between research entities and returns a set of novel triples, establishing new connections between previously unrelated entities, that can be interpreted as plausible research hypotheses. Some examples of these KGs are CS-KG [11], ORKG [15], Nanopublications [18], and UMLS [16].

The input KG contains a set of statements in the form of <*subject, predicate, object*>, connecting two entities (subject and object) through a relationship [10]. They can be any type of research element, such as methods, tasks, or materials. Relationships describe the connections, e.g., a task can be addressed using a certain method, or a given method can be employed to process a specific material. KGs are structured based on a domain ontology, which formally defines entity types and relations while supporting reasoning processes [88]. The ontology also defines the domain and range of relationships, ensuring that only specific types of entities can serve as the subject and object of a triple when a given relationship is used. For instance, in an ontology designed for a medical database, the relation "has symptom" might have a domain of "Diseases" and a range of "Symptoms". This means that only entities classified as diseases can be the subject of a triple whose predicate is "has symptom".

Formally, the problem we aim to solve is: given a KG consisting of a set T of triples $\langle s, p, o\rangle$, where s and o are research entities and r a relationship between them, we want to create a model $f(T) \rightarrow H$ that, given a collection of triples belonging to T and, optionally, additional information from external sources, it generates a set of triples H where $T \cap H = \emptyset$ and $\langle s, p', o\rangle \in H$ is a plausible research hypothesis.

It is important to note that, as discussed in the introduction, p' does not correspond to a classic link prediction task. This is because research hypotheses are not merely missing facts; they possess very specific characteristics. In particular, the new link between s and o must suggest a novel and plausible piece of knowledge that warrants empirical testing by a human researcher. To evaluate this in Section 4.2, we introduce a gold standard based on manual annotations from senior researchers.

ResearchLink aims to recognize sound hypotheses by integrating a variety of features, including path-based metrics, KGEs, text embeddings, and additional information about the occurrences of relevant entities. Fig. 1 shows the architecture of ResearchLink. It includes four main stages. First, we generate a number of candidate hypotheses that connect previously unrelated entities. Next, we prune this set by ranking the candidate hypotheses using SciCheck [62], a path-based link prediction technique. We then produce for each hypothesis a richer representation that includes a diverse set of features reflecting both the occurrence of the entities in the articles and their semantic context. Finally, we apply a supervised model on this representation for ranking the remaining hypotheses.

Algorithm 1 contains a description of ResearchLink as pseudocode.

Algorithm 1 ResearchLink Algorithm

```
Input: Knowledge Graph (KG), Number of top entities (N), SciCheck threshold (0.5)
Output: Classified triples as true or false hypotheses
Step 1: Generate Candidate Triples
Identify top N entities in KG
for each entity e_i in top N entities do
    Retrieve type of e_i and possible relations r
    for each relation r do
        Find compatible entities t in top N entities
        for each entity t do
            if (e_i, r, t) ∉ KG then
                Add (e_i, r, t) to candidate triples
            end if
        end for
    end for
end for
Step 2: Initial Pruning with SciCheck
Train SciCheck on KG and compute confidence scores c for candidate triples
Prune triples with c < 0.5 and keep the rest as pruned triples
Step 3: Pre-compute Information for Features
Compute necessary pre-information (TransH embeddings, degrees, occurrences, joint occurrences)
Step 4: Compute Features
for each labeled and pruned triple (s, p, o) do
    Compute features f_1, f_2, ..., f_8
end for
Step 5: Train Model for Hypothesis Classification
Train model on labeled triples using computed features
Step 6: Apply Model for Hypothesis Evaluation
for each pruned triple (s, p, o) do
    Apply model to compute classification
    Label the triple as true or false hypothesis
end for
```

In the following, we will describe in detail each step.

3.1. Candidate Hypotheses Generation

The initial step involves generating a number of new relationships between existing entities, that constitute the set of candidate research hypotheses. This set subsequently undergoes a process of cleaning, refinement, and ranking. This candidate set should be sufficiently expansive to encompass a broad range of potentially interesting triples, yet remain computationally manageable [89].

To generate an initial set of candidate hypotheses from a KG, the following steps are undertaken. First, we identify the top N entities based on their frequency in triples. This approach guarantees that each generated triple contains at least one highly-connected entity, ensuring richer informational context in subsequent steps. We then retrieve the type of each of the selected entities from KG's ontology. For example, the entity *random forest* is a *Method*, *sentiment analysis* is a *Task*, *Wikipedia* is a *Material*, etc. This informs which relations can be employed to pair them, given that each relation has specific domain and range type constraints in the ontology. For instance, given the entity *Wikipedia* which is of type *Material*, it cannot appear as the subject of triples using relationships built on top of active verbs such as *use*, *solve*, and *acquire*.

Finally, for each entity selected in the first step, the set of all triples that respect the ontology constraints is generated. These triples have the form (e, r, t), where e is one of the top N entities, r is one of the possible relations in the KG which have the type of e in their domain and t is one of the top N entities whose type is in the range of r. Any generated triple already present in the KG is discarded. Consequently, for each entity e selected in the first step, the maximum amount of triples that are generated is:

$$\sum_{r \in r_e} t_r$$

where r_e is the set of relations that allow the type of e in their domain, and t_r is the set of entities whose types are deemed by the KG's ontology to be admissible in the range of r.

3.2. Candidate Pruning with SciCheck

Most of the previously generated candidates are not particularly promising, as they link entities without substantial affinity. Consequently, it is desirable to prune this set before engaging in more computationally intensive operations. We perform this pruning by ranking the best candidates with SciCheck [62], a path-based link prediction approach specifically tailored for scientific knowledge graphs. SciCheck evaluates the potential validity of triples, offering a confidence score based on both context-based and embedding-based features. In particular, SciCheck processes an entire KG composed of triples and generates a neural-based classifier for each relation in the KG. For a given relation r, SciCheck creates a model $f_r : (h, r, t) \rightarrow s$, which assigns a confidence score s between 0 and 1 to any given triple $\langle h, r, t \rangle$, addressing the binary classification task of determining whether the triple is correct. To input data into the model, triples are transformed into numerical vector representations using specialized features and contextual embedding representations. Detailed insights about the features used are available in [62]. It should be noted that SciCheck must be first trained on a substantial portion of the KG. For instance, when training on the full CS-KG, the KG used in the evaluation, we consider all statements extracted from at least three papers, which yields a sizable set of about 1M triples.

To apply SciCheck on the set of candidate triples, we perform the following steps. First, we train SciCheck on the entire input knowledge graph. SciCheck requires this training process to learn what constitutes plausible knowledge in the context of the KG. SciCheck also autonomously produces negative triples for the training by corrupting triples from the KG, i.e., switching the subject or object of a triple with a different entity. As the candidate hypotheses do not exist in the KG, SciCheck possesses no prior knowledge about them.

Following its training, SciCheck establishes a classification model for each KG relation. This is aimed at ensuring nuanced classification, as varied information might be optimal for different relations [31]. Candidate hypotheses are grouped by their relation and submitted to the appropriate model. Each hypothesis is then assigned a confidence value within the range [0, 1]. Hypotheses with confidence values below a predefined threshold (set at 0.5 in our current configuration) are discarded. The threshold was empirically determined based on insights from [62].

3.3. Hypotheses conversion into feature vectors

To evaluate the candidate hypotheses using a supervised classification model, we produce a detailed vector representation for each, derived from a range of diverse features.

More in detail, for an input triple $\langle s, p, o \rangle$, the specific features employed in this process include:

- **Feature f_1: SciCheck confidence score for the triple.** The outcome of applying SciCheck to the triple, as outlined in the previous section, is retained and utilized as a classification feature. This value is a decimal number within the range of $[0.5, 1.0]$. Therefore, f_1 can be defined as:

$$f_1 = \text{SciCheck}(s, p, o) \quad (1)$$

- **Feature f_2: Degree of the two entities within the KG.** This feature assesses the connectivity of the two entities in the input triple, serving as an indicator of their prominence. It is encoded as two integer values, representing the number of edges of the KG connected to the subject and object entities, respectively.
  More formally, let $deg(e)$ denote the degree of entity e. Then, f_2 can be defined as:
  $$f_2 = (\deg(s), \deg(o)) \quad (2)$$
- **Feature f_3: Number of occurrences of the two entities in scientific abstracts.** This feature quantifies the prevalence of the subject and object within the scientific literature. It is computed by performing text matching against a collection of research paper abstracts to determine the absolute frequency of each entity, which is defined as the total number of times the entity appears in the abstracts. This feature is encoded as three integer numbers: one for the subject, one for the object, and one for their harmonic mean.
  More formally, let $\text{freq}_{\text{abs}}(e)$ denote the frequency of entity e in scientific abstracts. Then, f_3 can be expressed as:
  $$f_3 = (\text{freq}_{\text{abs}}(s), \text{freq}_{\text{abs}}(o), \frac{2 \cdot \text{freq}_{\text{abs}}(s) \cdot \text{freq}_{\text{abs}}(o)}{\text{freq}_{\text{abs}}(s) + \text{freq}_{\text{abs}}(o)}) \quad (3)$$
- **Feature f_4: Number of joint occurrences of the two entities in scientific abstracts.** This feature measures how often entities have appeared together in scientific literature. It is calculated by counting how many times they appeared in the same abstract. It is encoded as an integer number.
  Let $\text{freq}_{\text{joint}}(e_1, e_2)$ denote the frequency of joint occurrences of entities e_1 and e_2 in scientific abstracts. Then, f_4 can be defined as:
  $$f_4 = \text{freq}_{\text{joint}}(s, o) \quad (4)$$
- **Feature f_5: Minimum embedding cosine distance between the triple and all the other triples in the KG.** This feature assesses the novelty level of a triple by comparing it to the existing knowledge in the input KG. This assessment holds significance as it can highlight scenarios where a closely related concept already exists, potentially indicating limited novelty. To compute this feature, all triples are converted into sentences by concatenating their subject, predicate, and object with spaces as separators. Then, the MiniLM [90] embedding model is applied to these sentences, generating a numerical vector representation for each entity based on the text's semantics. These vectors are used to identify the triple in the KG that has the highest cosine similarity (i.e., the lowest cosine distance) to the triple under evaluation. The resulting value is thus a decimal number between 0 and 1. For instance, if this feature is computed for the triple <*machine learning, applied to, plane trajectory prediction*> and the graph contains a highly similar triple <*machine learning, applied to, satellite orbit prediction*>, the cosine distance will be relatively low, indicating the presence of research closely related to the evaluated triple.
  More formally, let MiniLM(sentence) denote the MiniLM embedding of a sentence, $\text{concat}(x, y, z)$ represents the string concatenation of the three entities x, y and z resulting in the string "x y z", and let $\cos(x, y)$ denote the cosine distance between vectors x and y. Then, f_5 can be expressed as:
  $$f_5 = \min_{\forall (s',p',o') \in \text{KG}} \cos(\text{MiniLM}(\text{concat}(s, p, o)), \text{MiniLM}(\text{concat}(s', p', o'))) \quad (5)$$
- **Feature f_6: Minimum embedding Euclidean distance between the triple and all the other triples in the KG.** This feature also evaluates the novelty of a triple, but it uses Euclidean distance instead of the cosine distance. Unlike cosine similarity, which measures the alignment or angular similarity between embeddings, Euclidean distance considers the absolute differences in embedding magnitudes. This approach can capture nuances that cosine similarity might overlook, such as disparities in the embedding space caused by variations in the semantic intensity or specificity of the triples. For instance, if two triples share a similar direction in the embedding space but differ significantly in their distances from the origin, the Euclidean distance will better capture this difference. A smaller Euclidean distance indicates a closer semantic relationship, while a larger distance suggests greater novelty.
  This feature is also computed using the MiniLM model. Therefore, f_6 can be expressed as:
  $$f_6 = \min_{\forall (s',p',o') \in \text{KG}} \|\text{MiniLM}(\text{concat}(s, p, o)) - \text{MiniLM}(\text{concat}(s', p', o'))\| \quad (6)$$
- **Feature f_7: Graph embedding similarity between the two entities.** This feature quantifies the semantic similarity between two entities by calculating the cosine similarity of their graph embeddings. The embeddings are generated by using TransH [91], a well-known general-purpose KG embedding model.
  Graph embeddings provide a numerical representation of the elements within a graph, typically including nodes and, in many cases, their relationships [25]. These embeddings are generated by training a neural network to optimize an objective function that captures the structural properties of the graph, often by minimizing the distance between selected representations. This process encodes meaningful information about the graph's structure and relationships into a compact, continuous vector space. The embeddings associated with each node are expected to reflect their structural properties, capture their semantics, and support various machine learning tasks. For instance, if two nodes share connections with several common nodes in the graph, their embeddings are likely to be close in the vector space.
  Let $\text{TransH}(e)$ denote the TransH embedding of entity e. Then, f_7 can be represented as:
  $$f_7 = \cos(\text{TransH}(s), \text{TransH}(o)) \quad (7)$$
- **Feature f_8: Word embedding similarity between the two entities.** This feature measures the similarity between the text representations of the subject and object. It is calculated as the cosine similarity of the word embeddings for the two entities, generated using the MiniLM model. The resulting value is a decimal ranging from 0 to 1. Therefore, f_8 can be expressed as:
  $$f_8 = \cos(\text{MiniLM}(s), \text{MiniLM}(o)) \quad (8)$$

In conclusion, since features f_2 and f_3 produce two values each (one for the subject and another for the object), each triple is represented by a feature vector of 10 values.

3.4. Hypothesis classification

ResearchLink applies a set of random forest classification models to the representation of the candidate hypotheses in order to assess which one is valid.

Each hypothesis deemed valid through classification is also associated to the confidence values derived from the classifier. This allows ResearchLink to provide a ranked list of the generated hypotheses. This aids researchers that have then to manually inspect and interpret the hypotheses, enabling them to promptly identify the more plausible ones.

4. Evaluation

This section presents the evaluation of ResearchLink compared to alternative approaches, along with an ablation study analyzing the benefits provided by the features discussed in Section 3.3.

It is important to note that ResearchLink is designed for hypothesis generation rather than KG completion. Consequently, it cannot be evaluated using standard KG completion datasets (e.g., FB15k, WN18RR) as they primarily focus on missing facts, many of which are trivial and not suitable as hypotheses. Moreover, they lack essential features required for ResearchLink, such as collections of scientific abstracts containing relevant entities. Therefore, to conduct a rigorous evaluation of ResearchLink, we constructed *CSKG-600*, a new dataset for research hypothesis generation derived from the Computer Science Knowledge Graph.

We acknowledge that this evaluation is limited to a single scientific domain; however, this restriction is due to the substantial complexity and data requirements involved in creating a benchmark for hypothesis generation that covers all eight feature types examined in this study. Specifically, this process requires a comprehensive knowledge graph of research entities linked by relationships that could support hypothesis generation, an associated bibliographic dataset to assess entity frequencies and the manual annotation of potential research hypotheses by multiple senior researchers. CS-KG enables this approach by offering a wealth of entities, relationships, and pertinent bibliographic data. In the following, we first describe the use case based on CS-KG that we adopted for evaluating our methodology. We then describe the generation of *CSKG-600* and discuss the experimental design. Finally, we report the performance of the various techniques.

4.1. Use case: CS-KG

The Computer Science Knowledge Graph (CS-KG) [11,92] is an open, large-scale, automatically built knowledge graph describing 41M statements related to the Computer Science domain. It has been generated using the titles and abstracts of more than 6.7M research papers about Computer Science ranging from the year 2010 to the first semester of the year 2021, using the Microsoft Academic Graph[2] as a data source. CS-KG includes more than 10M research entities that can belong to 5 different classes (*cskg-ont:Material*, *cskg-ont:Task*, *cskg-ont:Method*, *cskg-ont:Metric*, *cskg-ont:OtherEntity*) described by 179 relations such as *cskg-ont:usesMethod*, *cskg-ont:includesMaterial*, *skos:broader*, etc. CS-KG statements describe the relationship that stands between two entities based on the content of scientific abstracts as a piece of information. For instance, a triple `<cskg:generative_model, usesMethod, cskg:generative_adversarial_network>` from the papers [93,94] describes the fact that there exist generative models that use generative adversarial networks. The reader can find details about CS-KG content and its ontology online.[3]

4.2. Gold standard generation

In order to evaluate our approach, we constructed *CSKG-600*,[4] a gold standard of 600 human-labeled research hypotheses. We generated these hypotheses starting from a version of CS-KG containing only data up to 2017. This enabled us to emulate the hypothesis generation process as it would have been conducted in 2017, and then validate the accuracy of the hypotheses by assessing if they were verified in the following years (2018–2023). If a triple is not present in this set, it could be due to its absence in the literature sample used to create CS-KG or because the KG generation method failed to extract it. Therefore, any additional triple can be considered novel within the context of the KG, which is a necessary (but not sufficient) condition for a valid hypothesis.

We generated an initial set of novel triples by following the methodology for producing candidate hypotheses described in Sections 3.1 and 3.2. We first materialized the Cartesian combination of all possible entities for each of the three most frequent relations in CS-KG: *usesMethod*, *usesTask*, and *usesMaterial*. We then discarded the triples that were invalid according to the ontology of CS-KG[5] or that were already in the KG. Next, we used SciCheck [62] to prune these candidate hypotheses by computing a confidence score for each of them. This was done by training SciCheck using only triples introduced in CS-KG before 2017, so as to not bias the classifier with the most recent scientific knowledge. Finally, SciCheck was used to evaluate the generated triples, and those with a confidence value below 0.5 were discarded.

To reduce the number of hypotheses manually assessed by human evaluators, we then selected a subset of 20 popular entities from CS-KG, covering a variety of different topics in Computer Science, e.g., *autonomous_vehicle*, *sentiment_analysis*, *wearable_technology*, *face_recognition*, *deep_learning*, *game_theory*. For each of them, we generated 30 hypotheses. Specifically, we generated ten hypotheses for each of the three adopted relations. This process resulted in a set of 600 candidate hypotheses, with 200 hypotheses generated for each of the three relations under analysis.

The 600 resulting hypotheses were labeled by three senior researchers as either valid or invalid. To ensure objectivity, a hypothesis was deemed valid if sufficiently precise and supported by at least a scientific article published post-2017. The scientists were asked to locate relevant evidence in the literature before assessing each hypothesis. It is crucial to highlight that both the positive and negative examples were drawn from the same pool of candidates, all of whom had high SciCheck scores. As a result, incorporating the SciCheck score as a feature in ResearchLink does not introduce any bias into the evaluation process. This systematic approach to constructing the gold standard ensures that all valid hypotheses are manually verified and supported by recent literature. In contrast, invalid hypotheses, though generated by a link prediction system and therefore realistic, are deemed invalid due to being trivial, too vague, or lacking support in current literature.

The resulting statement can be readily translated into research hypotheses, as they facilitate the combination of the three primary types of entities in CS-KG (methods, tasks, and materials). The sound hypotheses generally recommend adopting specific methods, tasks, or materials to address particular challenges.

For example, the triple `<deep_learning, usesMethod, extended_kalman_filter>` described in a paper from 2019 [95] proposes incorporating an extended Kalman filter within a deep learning framework. Similarly, the triple `<wearable_ technology, usesTask, emotion_recognition>` has been studied in the year 2023 [96] and suggests integrating emotion recognition techniques within wearable technologies. A third example we would like to mention is `<data_compression, usesMethod, gaussian_mixture _model>` from the year 2024 [97] where data quantity minimization is studied for real-time ambulatory and telemedicine systems with limited storage. As previously discussed, negative examples usually refer to trivial ideas that do not contribute to additional knowledge (e.g., `<game_theory, usesMaterial, renewable_energy>`) or are too vague (e.g., `<decision_support_system, usesMaterial, geometry>`).

[2] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[3] https://w3id.org/cskg/cskg/documentation.php

[4] https://github.com/dayala1/ResearchLink/blob/master/CSKG-600.csv

[5] https://scholkg.kmi.open.ac.uk/cskg/ontology.html

4.3. Evaluation setup

This section outlines the tested methods and the metrics we adopted for their evaluation. We compare ResearchLink to five alternative approaches generally used for link prediction and knowledge graph completion [63]: $TransE$ [27], $TransH$ [28], $TransD$ [29], $RotatE$ [30], and $RGCN$ [33]. We used the PyKEEN's implementation of these models [98].

We evaluate these methods by assessing their ability to assign a score to each hypothesis in the test set of CSKG-600, reflecting its validity. This evaluation strategy aligns with established methodologies for applying link prediction techniques to the hypothesis generation task [3]. It is important to highlight that, since all methods are tested on the same set of candidate hypotheses, ResearchLink is exclusively used to evaluate these hypotheses, as described in Sections 3.3 and 3.4 (illustrated by the blue and yellow boxes in Fig. 1).

All the methods that required training on a knowledge graph were trained on the version of CS-KG, which contained only information before 2017. Once trained, these methods are capable of scoring any triple whose subject, relation, and object appeared in the training set. Each method was then applied to the entire test set, producing a ranked list of the candidate hypotheses. The performance of the methods was measured using *Precision@N* over these ranked lists. Precision@N (P@N) is defined as $\text{P@N} = \frac{\text{TP@N}}{N}$, where TP@N represents the number of true positives among the top N predictions. It is often employed in information retrieval and ranking problems to gauge the effectiveness of the system in identifying the most relevant items within a specific subset of the total results. For example, if a method produces 7 valid hypotheses in the top 10 positions, its Precision@10 would be 70%. We computed the average P@5, P@10, P@15, and P@20 across the four cross-validation runs. In this domain, we particularly aim to achieve better results for higher values of N, such as P@15 and P@20. These measures are less sensitive to randomness and provide a realistic goal of presenting researchers with 15 to 20 hypotheses, aiming for about 70%–75% to be of good quality. Conversely, P@5 was quite unstable across different experiments, as the position of a single item can significantly affect it. Nonetheless, we still report it for the sake of completeness.

To assess the contribution of the features discussed in Section 3.3, we also conducted an ablation study by testing 15 versions of Researchlink using various combinations of these features. Specifically, we evaluated four groups of features:

- **SciCheck feature** (f_1 in Section 3.3), the confidence score for the candidate hypothesis according to SciCheck [62], a path-based link prediction technique specifically designed for scientific KGs.
- **Frequency features** (f_2, f_3, f_4), which evaluate the frequency of the entities in the KG and the literature.
- **Similarity features** (f_5, f_6), which assess the similarity of the candidate hypothesis with other triples in the KG.
- **Embedding features** (f_7, f_8), which assess the similarity of the subject and object according to their word and graph embeddings.

We then run on the gold standard 15 versions of ResearchLink using all possible combinations of these four groups. As for the other methods, all the features were computed on the version of CS-KG that only contained information available before 2017.

We evaluated all approaches using k-fold cross-validation ($k = 4$) over the gold standard defined in Section 4.2. In each of the four iterations, we chose a subset of 5 entities and the corresponding 150 hypotheses. These served as the test set, while the remaining 15 entities associated with 450 hypotheses formed the training set.

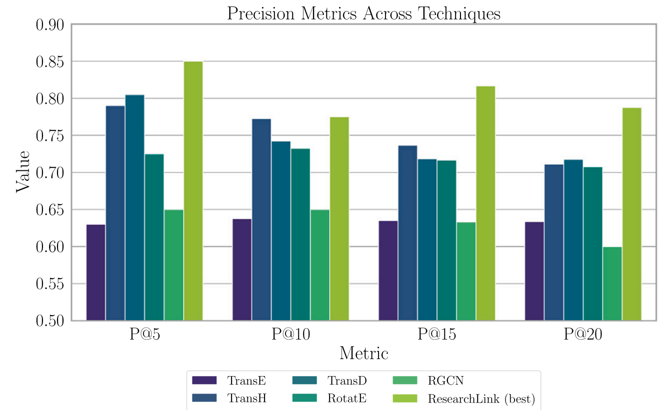

Fig. 2. Results summary of each tested method. ResearchLink (best) denotes the version of ResearchLink that obtained the best results at each metric in the ablation study.

4.4. Results

Table 2 presents the results of all experiments. The upper section details the 15 versions of ResearchLink obtained by combining the four groups of features discussed in the previous section. Each column indicates whether a specific feature set was used, with *True* denoting its inclusion and *False* indicating its exclusion. For instance, the first row corresponds to the version of ResearchLink that utilizes only the embedding features. The final row represents the full version of ResearchLink that incorporates all features. The lower section of the table reports the performance of the five alternative methods. Fig. 2 provides a summary, comparing the five baselines to the best results achieved by the different versions of ResearchLink for each P@N.

The version of ResearchLink that utilizes the complete set of features outperformed the five alternative approaches for P@10, P@15, and P@20. Notably, it was able to identify several triples representing significant hypotheses that were then materialized in recent years, such as <*face_detection, usesMethod, ensemble_learning*> [99], <*autonomous_vehicle, usesMethod, disturbance_observer*> [100], or <*structural_health_monitoring, usesTask, deep_learning*> [101]. The results for P@5 are more erratic, with TransD and TransH surpassing the full version of ResearchLink. However, the overall best method for P@5 is actually a version of ResearchLink that uses only similarity and embedding features.

It is also beneficial to examine each set of features individually to understand their distinct contributions. Focusing on the first part of Table 2, we can see that each feature group tends to enhance ResearchLink's performance. Using all features achieves the best performance for P@20 and P@10. However, other combinations are also effective and may warrant further investigation in future work. Notably, the version excluding the embedding features achieves the same performance as the full feature set for P@10, although with slightly lower performance for P@15 and P@20. The adoption of different embedding techniques may potentially further improve this approach. Similarly, the version excluding the frequency features performs best for P@15, suggesting that in some cases, integrating information about an entity's popularity is less beneficial. This may also be due to using a relatively simple proxy, namely the number of occurrences and co-occurrences of subjects and objects in scientific abstracts. We plan to explore more nuanced metrics, such as normalizing the occurrence number according to the year or the scientific field.

Finally, Fig. 2 summarizes the results, demonstrating that the general approach presented in this paper, which integrates various features relevant to the specific hypothesis generation task, is clearly superior to alternative methods designed for link prediction and knowledge graph completion. This superiority is particularly evident in the most reliable metrics, such as P@15 and P@20.

Table 2
Average Precision@N for all approaches. The best result for each metric is shown in **bold**.

Scicheck feat.	Frequency feat.	Similarity feat.	Embedding feat.	P@5	P@10	P@15	P@20
False	False	False	True	0.80000	0.75000	0.73333	0.73750
False	False	True	False	0.65000	0.70000	0.66667	0.68750
False	False	True	True	**0.85000**	0.72500	0.73333	0.75000
False	True	False	False	0.50000	0.52500	0.60000	0.65000
False	True	False	True	0.80000	0.65000	0.71667	0.72500
False	True	True	False	0.75000	0.75000	0.76667	0.72500
False	True	True	True	0.75000	0.70000	0.73333	0.73750
True	False	False	False	0.70000	0.70000	0.65000	0.66250
True	False	False	True	0.80000	0.75000	0.73333	0.72500
True	False	True	False	0.70000	0.70000	0.73333	0.71250
True	False	True	True	0.80000	**0.77500**	**0.81667**	0.77500
True	True	False	False	0.70000	0.72500	0.70000	0.70000
True	True	False	True	0.70000	0.77500	0.73333	0.73750
True	True	True	False	0.70000	**0.77500**	0.78333	0.76250
True	True	True	True	0.75000	**0.77500**	0.76667	**0.78750**
TransE				0.63000	0.63750	0.63500	0.63375
TransH				0.79000	0.77250	0.73667	0.71125
TransD				0.80500	0.74250	0.71833	0.71750
RotatE				0.72500	0.73250	0.71667	0.70750
RGCN				0.65000	0.65000	0.63333	0.60000

Table 3
Precision@N for the hypotheses generated by *ResearchLink* for a sample of entities.

Entity	P@5	P@10	P@15	P@20
action_recognition	1.000	1.000	0.800	0.800
pattern_recognition	1.000	0.900	0.867	0.850
image_analysis	1.000	0.900	0.800	0.800
tracking_system	1.000	0.900	0.733	0.750
decision_support_system	1.000	0.800	0.800	0.800
face_detection	1.000	0.800	0.600	0.550
neural_network	1.000	0.700	0.667	0.600
data_compression	0.800	0.500	0.533	0.550
deep_learning	0.800	0.800	0.867	0.850
evolutionary_algorithm	0.800	0.800	0.667	0.500
wearable_technology	0.800	0.800	0.600	0.650
autonomous_vehicle	0.800	0.700	0.733	0.600
sentiment_analysis	0.800	0.700	0.667	0.750
graphics_processing_unit	0.800	0.700	0.600	0.650
game_theory	0.400	0.700	0.667	0.550
multiagent_system	0.400	0.400	0.267	0.300

Table 3 reports the performance of ResearchLink on a sample of entities. Generally, the results are satisfactory, with valid hypotheses involving most entities ranked at the top. However, notable performance variations exist between them, implying that specific characteristics of the entities might substantially influence our ability to evaluate relevant research hypotheses. For instance, the system obtains perfect accuracy for the top 5 hypotheses on entities such as *action_recognition* and *face_detection*. Conversely, it obtains sub-par results for other ones, such as *multiagent_system*. A more in-depth investigation of this aspect will necessitate further future work.

5. Limitations

In this section, we briefly discuss the limitations of ResearchLink that need to be addressed in future work. The first limitation concerns the availability of input data. ResearchLink necessitates both a knowledge graph of research entities and an extensive corpus of papers to calculate their frequency in the literature. This information may not be readily obtainable in all fields. However, the rise of the open access movement is leading to a growing availability of scholarly data, resulting in the creation of various repositories that offer both abstracts [102] and full texts of the articles [47,48]. Concurrently, several initiatives are engaged in developing new scientific knowledge graphs to support a range of intelligent systems across multiple domains [11,15,50].

A second limitation arises from the fact that the hypothesis generation process relies solely on existing entities within the knowledge graph. Hypotheses that cannot be represented as their combination remain undiscovered when hypothesis generation is formalized as a link prediction task over a knowledge graph of research concepts. Moreover, ResearchLink may inherently prioritize highly interconnected entities, as they offer a richer representation from which further connections can be derived. This bias could lead to reduced effectiveness in generating hypotheses for less-represented or more isolated ones.

One more limitation pertains to computational resources. Specifically, some features in ResearchLink demand substantial resources, most notably f_5 and f_6. These features compute embedding similarities and distances across all hypotheses under evaluation and all triples in a KG, resulting in a complexity of $O(n^2)$. In our experiments, we used a cluster of 5 NUCs NUC11PHKi7 each equipped with 32 GB of RAM and a GPU RTX 2060. Future research aims to explore strategies to streamline this computation, for instance, by employing heuristics to reduce the computational load.

Finally, the inherent complexity of defining a good hypothesis renders the evaluation of a hypothesis generation system a challenging task. It either demands consensus among researchers on what constitutes a promising research direction, which is difficult to attain, or the capability to predict future outcomes and discern which ideas will eventually materialize, which is unfeasible. In our experiments, we set a cutoff year for the data available to our system, enabling us to formulate hypotheses using historical knowledge and assess them with current information. An alternative strategy might be to create hypotheses in the present and evaluate their validity after a specified number of years have elapsed.

6. Conclusions

In this paper, we presented ResearchLink, an innovative approach for hypothesis generation that integrates path-based features, knowledge graph embeddings, and text embeddings, while also including additional information about the occurrence of the relevant entities in the literature. This robust combination of features allows ResearchLink to outperform traditional link prediction methods in the context of hypothesis generation. ResearchLink was evaluated using a dataset consisting 600 candidate hypotheses, manually annotated by three senior researchers. It outperformed several alternative methods, especially showing marked improvement over standard link prediction techniques based on graph embeddings and path-based features.

The main contributions of this paper are as follows: (1) demonstrating that strong performance can be achieved on the challenging task of research hypothesis generation, formalized as a link prediction task over knowledge graphs of research concepts; (2) evaluating the utility of eight distinct feature types that offer valuable guidance for the design of next-generation models; (3) introducing a novel methodology that effectively leverages the rich information derived from these features; and (4) providing a benchmark and a comprehensive codebase that serve as resources to drive innovation toward new state-of-the-art solutions.

In future work, we plan to expand the size and coverage of the gold standard across multiple scientific domains. Additionally, we will include in our analysis recent link prediction solutions that incorporate large language models [103]. We also intend to investigate potential biases derived from the knowledge graph, such as reduced predictive performance for underrepresented entities. Finally, we aim to improve our approach by devising strategies to identify novel entities, such as recently introduced methodologies, which could lead to particularly impactful research hypotheses.

CRediT authorship contribution statement

Agustín Borrego: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Danilo Dessì:** Writing – review & editing, Validation, Software, Methodology, Data curation. **Daniel Ayala:** Writing – review & editing, Software, Resources. **Inma Hernández:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Francesco Osborne:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Diego Reforgiato Recupero:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Davide Buscaldi:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **David Ruiz:** Supervision, Investigation, Conceptualization. **Enrico Motta:** Supervision, Investigation, Conceptualization.

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.

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Italy, Mission 4 Component 2 Investment 1.5 - Call for tender No. 3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR), Italy.

Data availability

Data will be made available on request.

References

[1] J. Sybrandt, I. Tyagin, M. Shtutman, I. Safro, AGATHA: automatic graph mining and transformer based hypothesis generation approach, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2757–2764, http://dx.doi.org/10.1145/3340531.341268.

[2] H. Kitano, Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery, AI Mag. 37 (1) (2016) 39–49, http://dx.doi.org/10.1609/aimag.v37i1.2642.

[3] R.d. Haan, I. Tiddi, W. Beek, Discovering research hypotheses in social science using knowledge graph embeddings, in: European Semantic Web Conference, Springer, 2021, pp. 477–494, http://dx.doi.org/10.1007/978-3-030-77385-4_28.

[4] D. Cameron, R. Kavuluru, T.C. Rindflesch, A.P. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, J. Biomed. Informatics 54 (2015) 141–157, http://dx.doi.org/10.1016/j.jbi.2015.01.014.

[5] S. Henry, B.T. McInnes, Literature based discovery: models, methods, and trends, J. Biomed. Informatics 74 (2017) 20–32, http://dx.doi.org/10.1016/j.jbi.2017.08.011.

[6] T.E. Workman, M. Fiszman, M.J. Cairelli, D. Nahl, T.C. Rindflesch, Spark, an application based on serendipitous knowledge discovery, J. Biomed. Informatics 60 (2016) 23–37, http://dx.doi.org/10.1016/j.jbi.2015.12.014.

[7] J.P. McCusker, M. Dumontier, R. Yan, S. He, J.S. Dordick, D.L. McGuinness, Finding melanoma drugs through a probabilistic knowledge graph, PeerJ Comput. Sci. 3 (2017) e106, http://dx.doi.org/10.7717/peerj-cs.106.

[8] B.-K. Choi, T. Dayaram, N. Parikh, A.D. Wilkins, M. Nagarajan, I.B. Novikov, B.J. Bachman, S.Y. Jung, P.J. Haas, J.L. Labrie, et al., Literature-based automated discovery of tumor suppressor p53 phosphorylation and inhibition by NEK2, Proc. Natl. Acad. Sci. 115 (42) (2018) 10666–10671, http://dx.doi.org/10.1073/pnas.1811811115.

[9] A.A. Muñoz, E.U. Carro, L.P. Santamaría, B.O. Carrasco, E.M. Ruiz, Y.P. Gallardo, A. Rodriguez-Gonzalez, Redirection: Generating drug repurposing hypotheses using link prediction with DISNET data, in: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2022, pp. 7–12, http://dx.doi.org/10.1109/CBMS55023.2022.00010.

[10] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: opportunities and challenges, Artif. Intell. Rev. (2023) 1–32, http://dx.doi.org/10.1007/s10462-022-10397-2.

[11] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, CS-kg: A large-scale knowledge graph of research entities and claims in computer science, in: The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Springer, 2022, pp. 678–696, http://dx.doi.org/10.1007/978-3-031-19433-7_39.

[12] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Comput. Surv. 54 (4) (2021) 1–37, http://dx.doi.org/10.1145/3447772.

[13] S. Vahdati, N. Arndt, S. Auer, C. Lange, OpenResearch: collaborative management of scholarly communication metadata, in: European Knowledge Acquisition Workshop, Springer, 2016, pp. 778–793, http://dx.doi.org/10.1007/978-3-319-49004-5_50.

[14] D. Dessì, F. Osborne, D.R. Recupero, D. Buscaldi, E. Motta, H. Sack, AI-KG: an automatically generated knowledge graph of artificial intelligence, in: ISWC 2020, 12507, Springer, 2020, pp. 127–143, http://dx.doi.org/10.1007/978-3-030-62466-8_9.

[15] M.Y. Jaradeh, A. Oelen, K.E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246, http://dx.doi.org/10.1145/3360901.3364435.

[16] L. Amos, D. Anderson, S. Brody, A. Ripple, B.L. Humphreys, UMLS users and uses: a current overview, J. Am. Med. Informatics Assoc. 27 (10) (2020) 1606–1611, http://dx.doi.org/10.1093/jamia/ocaa084.

[17] A. Rossanez, J.C. dos Reis, R. da Silva Torres, Representing scientific literature evolution via temporal knowledge graphs., in: MEPDaW@ ISWC, 2020, pp. 33–42, URL https://hdl.handle.net/11250/2773742.

[18] E. Schultes, B. Magagna, T. Kuhn, M. Suchánek, L.B. da Silva Santos, B. Mons, The comparative anatomy of nanopublications and FAIR digital objects, Res. Ideas Outcomes 8 (2022) e94150, http://dx.doi.org/10.3897/rio.8.e94150.

[19] A. Salatino, T. Aggarwal, A. Mannocci, F. Osborne, E. Motta, A survey on knowledge organization systems of research fields: Resources and challenges, 2024, arXiv preprint arXiv:2409.04432.

[20] A. Kumar, S.S. Singh, K. Singh, B. Biswas, Link prediction techniques, applications, and performance: A survey, Phys. A 553 (2020) 124289, http://dx.doi.org/10.1016/j.physa.2020.124289.

[21] M. Nayyeri, G.M. Cil, S. Vahdati, F. Osborne, M. Rahman, S. Angioni, A. Salatino, D.R. Recupero, N. Vassilyeva, E. Motta, et al., Trans4E: Link prediction on scholarly knowledge graphs, Neurocomputing 461 (2021) 530–542, http://dx.doi.org/10.1016/j.neucom.2021.02.100.

[22] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semant. Web 8 (3) (2017) 489–508, http://dx.doi.org/10.3233/SW-160218.

[23] J. Devlin, M.-W. Chang, K. Lee, K.T. Google, A.I. Language, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Naacl-Hlt 2019 (Mlm), 2018, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[24] A. Nambiar, M. Heflin, S. Liu, S. Maslov, M. Hopkins, A. Ritz, Transforming the language of life: transformer neural networks for protein prediction tasks, in: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2020, pp. 1–8, http://dx.doi.org/10.1145/3388440.3412467.

[25] Y. Dai, S. Wang, N.N. Xiong, W. Guo, A survey on knowledge graph embedding: Approaches, applications and benchmarks, Electronics 9 (5) (2020) 750, http://dx.doi.org/10.3390/electronics9050750.

[26] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE TKDE 29 (12) (2017) http://dx.doi.org/10.1109/TKDE.2017.2754499.

[27] X. Han, S. Cao, X. Lv, Y. Lin, Z. Liu, M. Sun, J. Li, OpenKE: An open toolkit for knowledge embedding, in: EMNLP, 2018, pp. 139–144, http://dx.doi.org/10.18653/v1/D18-2024.

[28] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, 2014, http://dx.doi.org/10.1609/aaai.v28i1.8870.

[29] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, Knowledge graph embedding via dynamic mapping matrix, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 687–696, http://dx.doi.org/10.3115/v1/P15-1067.

[30] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, 2019, arXiv preprint arXiv:1902.10197.

[31] A. Borrego, D. Ayala, I. Hernández, C.R. Rivero, D. Ruiz, CAFE: Knowledge graph completion using neighborhood-aware features, Eng. Appl. Artif. Intell. 103 (2021) 104302, http://dx.doi.org/10.1016/j.engappai.2021.104302.

[32] Y. Shen, D. Wen, Y. Li, N. Du, H.-t. Zheng, M. Yang, Path-based attribute-aware representation learning for relation prediction, in: Proceedings of the 2019 SIAM International Conference on Data Mining, SIAM, 2019, pp. 639–647, http://dx.doi.org/10.1137/1.9781611975673.72.

[33] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, Springer, 2018, pp. 593–607, http://dx.doi.org/10.1007/978-3-319-93417-4_38.

[34] M. Thilakaratne, K. Falkner, T. Atapattu, A systematic review on literature-based discovery: general overview, methodology, & statistical analysis, ACM Comput. Surv. 52 (6) (2019) 1–34, http://dx.doi.org/10.1145/3365756.

[35] S. Henry, B.T. McInnes, Literature based discovery: Models, methods, and trends, J. Biomed. Informatics 74 (2017) 20–32, http://dx.doi.org/10.1016/j.jbi.2017.08.011, URL https://www.sciencedirect.com/science/article/pii/S1532046417301909.

[36] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al., Scientific discovery in the age of artificial intelligence, Nature 620 (7972) (2023) 47–60.

[37] D.N. Sosa, R.B. Altman, Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference, Brief. Bioinform. 23 (4) (2022) bbac268.

[38] Y. Sebastian, N.R. Smalheiser, Editorial: Emerging areas in literature-based discovery, Front. Res. Metrics Anal. 8 (2023) http://dx.doi.org/10.3389/frma.2023.1122547, URL https://www.frontiersin.org/articles/10.3389/frma.2023.1122547.

[39] J. Son, D. Kim, Applying network link prediction in drug discovery: an overview of the literature, Expert. Opin. Drug Discov. 19 (1) (2024) 43–56, http://dx.doi.org/10.1080/17460441.2023.2267020.

[40] S. Spangler, A.D. Wilkins, B.J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C.R. Pickering, A. Comer, J.N. Myers, I. Stanoi, L. Kato, A. Lelescu, J.J. Labrie, N. Parikh, A.M. Lisewski, L. Donehower, Y. Chen, O. Lichtarge, Automated hypothesis generation based on mining scientific literature, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, Association for Computing Machinery, ISBN: 9781450329569, 2014, pp. 1877–1886, http://dx.doi.org/10.1145/2623330.2623667.

[41] D. Hristovski, A. Kastrin, T.C. Rindflesch, Semantics-based cross-domain collaboration recommendation in the life sciences: Preliminary results, in: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 805–806.

[42] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K.A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (7763) (2019) 95–98.

[43] Z. Zhang, B.G. Patra, A. Yaseen, J. Zhu, R. Sabharwal, K. Roberts, T. Cao, H. Wu, Scholarly recommendation systems: a literature survey, Knowl. Inf. Syst. (2023) 1–46, http://dx.doi.org/10.1007/s10115-023-01901-x.

[44] S. Angioni, A. Salatino, F. Osborne, D.R. Recupero, E. Motta, AIDA: A knowledge graph about research dynamics in academia and industry, Quant. Sci. Stud. 2 (4) (2021) 1356–1398, http://dx.doi.org/10.1162/qss_a_00162.

[45] Y. Zhang, F. Zhang, P. Yao, J. Tang, Name disambiguation in aminer: Clustering, maintenance, and human in the loop., in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1002–1011, http://dx.doi.org/10.1145/3219819.3219859.

[46] A.G. Nuzzolese, A.L. Gentile, V. Presutti, A. Gangemi, Conference linked data: the ScholarlyData project, in: ISWC, Springer, 2016, pp. 150–158, http://dx.doi.org/10.1007/978-3-319-46547-0_16.

[47] S. Peroni, D. Shotton, F. Vitali, One year of the opencitations corpus, in: International Semantic Web Conference, Springer, 2017, pp. 184–192, http://dx.doi.org/10.1007/978-3-319-68204-4_19.

[48] P. Knoth, Z. Zdrahal, CORE: three access levels to underpin open access, D- Lib Mag. 18 (11/12) (2012) http://dx.doi.org/10.1045/november2012-knoth.

[49] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270, http://dx.doi.org/10.1093/nar/gkh061.

[50] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, Inf. Serv. Use 30 (1–2) (2010) 51–56, http://dx.doi.org/10.3233/ISU-2010-0613.

[51] U. Akujuobi, M. Spranger, S.K. Palaniappan, X. Zhang, T-pair: Temporal node-pair embedding for automatic biomedical hypothesis generation, IEEE Trans. Knowl. Data Eng. 34 (6) (2020) 2988–3001, http://dx.doi.org/10.1109/TKDE.2020.3017687.

[52] H. Zhou, H. Jiang, W. Yao, X. Du, Learning temporal difference embeddings for biomedical hypothesis generation, Bioinformatics 38 (23) (2022) 5253–5261, http://dx.doi.org/10.1093/bioinformatics/btac660.

[53] Y.J. Park, D. Kaplan, Z. Ren, C.-W. Hsu, C. Li, H. Xu, S. Li, J. Li, Can ChatGPT be used to generate scientific hypotheses? J. Mater. 10 (3) (2024) 578–584, http://dx.doi.org/10.1016/j.jmat.2023.08.007, URL https://www.sciencedirect.com/science/article/pii/S2352847823001557.

[54] M. Elbadawi, H. Li, A.W. Basit, S. Gaisford, The role of artificial intelligence in generating original scientific research, Int. J. Pharm. 652 (2024) 123741.

[55] N. Lao, W.W. Cohen, Relational retrieval using a combination of path-constrained random walks, Mach. Learn. 81 (1) (2010) 53–67, http://dx.doi.org/10.1007/s10994-010-5205-8.

[56] M. Gardner, T. Mitchell, Efficient and expressive knowledge base completion using subgraph feature extraction, in: EMNLP, The Association for Computational Linguistics, 2015, pp. 1488–1498, http://dx.doi.org/10.18653/v1/D15-1173.

[57] S. Mazumder, B. Liu, Context-aware path ranking for knowledge base completion, in: IJCAI, AAAI Press, 2017, pp. 1195–1201, http://dx.doi.org/10.24963/ijcai.2017/166.

[58] W. Xiong, T. Hoang, W.Y. Wang, Deeppath: A reinforcement learning method for knowledge graph reasoning, in: EMNLP, 2017, pp. 564—-573, http://dx.doi.org/10.18653/v1/D17-1060.

[59] R. Das, S. Dhuliawala, M. Zaheer, L. Vilnis, I. Durugkar, A. Krishnamurthy, A. Smola, A. McCallum, Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning, 2017, arXiv preprint arXiv:1711.05851.

[60] P. Tiwari, H. Zhu, H.M. Pandey, Dapath: Distance-aware knowledge graph reasoning based on deep reinforcement learning, Neural Netw. 135 (2021) 1–12, http://dx.doi.org/10.1016/j.neunet.2020.11.012.

[61] L. Chen, J. Cui, X. Tang, Y. Qian, Y. Li, Y. Zhang, Rlpath: a knowledge graph link prediction method using reinforcement learning based attentive relation path searching and representation learning, Appl. Intell. (2022) 1–12, http://dx.doi.org/10.1007/s10489-021-02672-0.

[62] A. Borrego, D. Dessì, I. Hernández, F. Osborne, D. Reforgiato Recupero, D. Ruiz, D. Buscaldi, E. Motta, Completing scientific facts in knowledge graphs of research concepts, IEEE Access 10 (2022) 125867–125880, http://dx.doi.org/10.1109/ACCESS.2022.3220241.

[63] M.U. Akhtar, J. Liu, Z. Xie, X. Liu, S. Ahmed, B. Huang, Entity alignment based on relational semantics augmentation for multilingual knowledge graphs, Knowl.-Based Syst. 252 (2022) 109494, http://dx.doi.org/10.1016/j.knosys.2022.109494.
[64] F. Sola, D. Ayala, I. Hernández, D. Ruiz, Deep embeddings and graph neural networks: using context to improve domain-independent predictions, Appl. Intell. (2023) http://dx.doi.org/10.1007/s10489-023-04685-3.
[65] P. He, G. Zhou, Y. Yao, Z. Wang, H. Yang, A type-augmented knowledge graph embedding framework for knowledge graph completion, Sci. Rep. 13 (1) (2023) 12364, http://dx.doi.org/10.1038/s41598-023-38857-5.
[66] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: NIPS, 2013, pp. 2787–2795.
[67] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, RotatE: Knowledge graph embedding by relational rotation in complex space, in: ICLR, 2019, pp. 271–280.
[68] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: ICML, ol. 48, 2016, pp. 2071–2080.
[69] S.M. Kazemi, D. Poole, Simple embedding for link prediction in knowledge graphs, in: Advances in Neural Information Processing Systems, 2018.
[70] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for knowledge graph completion., in: AAAI, Vol. 15, 2015, http://dx.doi.org/10.1609/aaai.v29i1.9491.
[71] H. Liu, Y. Wu, Y. Yang, Analogical inference for multi-relational embeddings, in: ICML-Volume 70, 2017.
[72] R. Socher, D. Chen, C.D. Manning, A. Ng, Reasoning with neural tensor networks for knowledge base completion, in: Advances in NIPS, 2013, pp. 926–934.
[73] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS '13, 2013, pp. 3111–3119, URL http://dl.acm.org/citation.cfm?id=2999792.2999959.
[74] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, CoRR abs/1907.11692, URL http://arxiv.org/abs/1907.11692.
[75] M. Nayyeri, S. Vahdati, X. Zhou, H.S. Yazdi, J. Lehmann, Embedding-based recommendations on scholarly knowledge graphs, in: European Semantic Web Conference, Springer, 2020, pp. 255–270, http://dx.doi.org/10.1007/978-3-030-49461-2_15.
[76] M. Nayyeri, G.M. Cil, S. Vahdati, F. Osborne, A. Kravchenko, S. Angioni, A. Salatino, D.R. Recupero, E. Motta, J. Lehmann, Link prediction of weighted triples for knowledge graph completion within the scholarly domain, IEEE Access 9 (2021) 116002–116014, http://dx.doi.org/10.1109/ACCESS.2021.3105183.
[77] D. Christofidellis, M. Manica, L. Georgopoulos, H. Vandierendonck, Understood in translation: Transformers for domain understanding, in: 2021 Workshop on Scientific Document Understanding, SDU 2021, 2021.
[78] M. Chen, Y. Tian, M. Yang, C. Zaniolo, Multilingual knowledge graph embeddings for cross-lingual knowledge alignment, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2017.
[79] Z. Sun, W. Hu, Q. Zhang, Y. Qu, Bootstrapping entity alignment with knowledge graph embedding., in: IJCAI, vol. 18, No. 2018, 2018.
[80] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J.F. Rousseau, X. Li, W. Xu, V.I. Torvik, et al., Building a PubMed knowledge graph, Sci. Data 7 (1) (2020) 205.
[81] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D.R. Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Eng. Appl. Artif. Intell. 133 (2024) 108166.
[82] S.S. Khanal, P. Prasad, A. Alsadoon, A. Maag, A systematic review: machine learning based recommendation systems for e-learning, Educ. Inf. Technol. 25 (4) (2020) 2635–2664.
[83] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial intelligence for literature reviews: Opportunities and challenges, Artif. Intell. Rev. 57 (259) (2024).
[84] D. Buscaldi, D. Dessí, E. Motta, M. Murgia, F. Osborne, D.R. Recupero, Citation prediction by leveraging transformers and natural language processing heuristics, Inf. Process. Manage. 61 (1) (2024) 103583.
[85] S. Brody, Scite, J. Med. Libr. Association: JMLA 109 (4) (2021) 707.
[86] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D.R. Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, IEEE Access 11 (2023) 22468–22489.
[87] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A.Y. Lau, et al., Conversational agents in healthcare: a systematic review, J. Am. Med. Informatics Assoc. 25 (9) (2018) 1248–1258.
[88] P. Hitzler, A review of the semantic web field, Commun. ACM 64 (2) (2021) 76–83, http://dx.doi.org/10.1145/3397512.
[89] A. Borrego, D. Ayala, I. Hernández, C.R. Rivero, D. Ruiz, Generating rules to filter candidate triples for their correctness checking by knowledge graph completion techniques, in: K-CAP, 2019, pp. 115–122, http://dx.doi.org/10.1145/3360901.3364418.
[90] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: NeurIPS, 2020.
[91] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph and text jointly embedding, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1591–1601, http://dx.doi.org/10.3115/v1/D14-1167.
[92] D. Dessí, F. Osborne, D.R. Recupero, D. Buscaldi, E. Motta, SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain, Knowl.-Based Syst. 258 (2022) 109945, http://dx.doi.org/10.1016/j.knosys.2022.109945.
[93] C. Wang, C. Xu, X. Yao, D. Tao, Evolutionary generative adversarial networks, IEEE Trans. Evol. Comput. 23 (6) (2019) 921–934, http://dx.doi.org/10.1109/TEVC.2019.2895748.
[94] W. Sun, S. Yu, J. Yang, B. Dong, A novel collaborative filtering framework based on variational self-attention GAN, in: GLOBECOM 2020-2020 IEEE Global Communications Conference, IEEE, 2020, pp. 1–6, http://dx.doi.org/10.1109/GLOBECOM42002.2020.9322300.
[95] Q. Li, Z.Y. Wu, A. Rahman, Evolutionary deep learning with extended Kalman filter for effective prediction modeling and efficient data assimilation, J. Comput. Civ. Eng. 33 (3) (2019) 04019014.
[96] W. Almukadi, Smart scarf: An IOT-based solution for emotion recognition, Eng. Technol. Appl. Sci. Res. 13 (3) (2023) 10870–10874.
[97] R.R. Sahoo, S. Bhowmick, D. Mandal, P.K. Kundu, A novel approach of Gaussian mixture model-based data compression of ECG and PPG signals for various cardiovascular diseases, Biomed. Signal Process. Control. 96 (2024) 106581.
[98] M. Ali, M. Berrendorf, C.T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, J. Lehmann, PyKEEN 1.0: a python library for training and evaluating knowledge graph embeddings, J. Mach. Learn. Res. 22 (82) (2021) 1–6.
[99] V.R.R. Chirra, S.R. Uyyala, V.K.K. Kolli, Virtual facial expression recognition using deep CNN with ensemble learning, J. Ambient. Intell. Humaniz. Comput. 12 (12) (2021) 10581–10599, http://dx.doi.org/10.1007/s12652-020-02866-3.
[100] X. Wang, W. Sun, Trajectory tracking of autonomous vehicle: A differential flatness approach with disturbance-observer-based control, IEEE Trans. Intell. Veh. 8 (2) (2023) 1368–1379, http://dx.doi.org/10.1109/TIV.2022.3186280.
[101] X. Ye, T. Jin, C. Yun, A review on deep learning-based structural health monitoring of civil infrastructures, Smart Struct. Syst. 24 (5) (2019) 567–585, http://dx.doi.org/10.12989/sss.2019.24.5.567.
[102] J. Priem, H. Piwowar, R. Orr, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022, arXiv preprint arXiv:2205.01833.
[103] D. Shu, T. Chen, M. Jin, Y. Zhang, M. Du, Y. Zhang, Knowledge graph large language model (KG-LLM) for link prediction, 2024, arXiv preprint arXiv:2403.07311.