*Bioinformatics*, 2026, **42**(3), btag061
https://doi.org/10.1093/bioinformatics/btag061
Published: 24 February 2026
**Original Paper | Data and text mining**

OXFORD

# Knowledge-based citation reasoning for biomedical domain

**Pengcheng Li[1,2], Kai Zhang[3], Xiaozhong Liu[3,], Xuhong Zhang[4,*,]**

[1]School of Economics and Management, Hubei University of Technology, Wuhan 430068, Hubei, China
[2]Hubei Digital Industrial Economic Development Research Center, Hubei University of Technology, Wuhan 430068, Hubei, China
[3]Computer Science and Data Science, Worcester Polytechnic Institute, Worcester, MA 01609, United States
[4]Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, United States
*Corresponding author. Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University, 700 N Woodlawn Ave, Bloomington, IN 47408, United States. E-mail: zhangxuh@iu.edu.
Associate Editor: Janet Kelso

## Abstract

***Motivation:*** **Citation is central to scholarly communication, enabling researchers to navigate rapidly expanding literature and identify relevant prior work. Yet the 'reasoning' behind why a particular paper is cited is often implicit or opaque. Although academic search engines and literature tools rank candidate papers for a query, the motivations underlying these rankings are rarely transparent, making it difficult for scholars to interpret and act on retrieved results—especially in biomedical research where domain knowledge is essential.**

***Results:*** **We propose an encoder–decoder framework that leverages curated biomedical knowledge to generate 'explanations of citation motivation' in a structured bio-triplet format. We evaluate the approach against recent families of pre-trained language models for text generation, including BERT-style (and variants) and GPT-style (and variants) models. In cancer-focused experiments using PubMed Central, we annotate over 10 000 citation relations with bio-triplets grounded in curated knowledge from multiple biomedical databases. Trained on these annotations, our model outperforms strong sequence-generation baselines, improving precision, recall, and *F*1 for citation-motivation generation.**

***Availability and implementation:*** **Code and data are available at Zenodo (archival DOI: `10.281/zenodo.14893445`) and GitHub: https://github.com/zhongxiangboy/Knowledge-based-Citation-Reasoning-for-Biomedical-Domain.**

## 1 Introduction

Citation is a crucial activity in scientific research, allowing researchers to acknowledge the contributions of others and traverse the expansive landscape of scholarly literature. Despite its importance, the underlying motivations for citation activities remain complex and not completely understood (Bronk *et al.* 2023). Numerous scientific search and recommendation systems exist to assist researchers in finding relevant literature, but the ranking algorithms employed by these systems often lack transparency (Zehlike *et al.* 2022). Consequently, it becomes difficult for researchers to discern why certain papers are recommended over others, which impedes their ability to effectively use these systems to meet their research needs.

An effective approach for citation analysis is the use of ontologies, which describe the nature of citations in terms of factual and rhetorical relationships within a specific domain (Ihsan and Muhammad 2019). Existing quantitative ontological and bibliometric citation analyses have primarily been conducted on non-biomedical datasets, with a limited number of analyzed publications (Donthu *et al.* 2021, Mejia *et al.* 2021). In the literature, citation has been explored from various perspectives, including count-based analysis (Leydesdorff and Opthof 2010, Pan and Fortunato 2014), sentiment analysis (Kochhar and Ojha 2020), co-citation behavior (Small and Garfield 1989), citation classification (Meng *et al.* 2017), and detection of citation sentiment (Athar and Teufel 2012). To gain deeper insights into citation literature, authors have identified the intent behind citations (Mercier *et al.* 2020). Intent refers to the citation's purpose, as authors cite published work for numerous reasons, such as to describe their own work or to contradict claims, which is highly relevant to our task. But unlike (Mercier *et al.* 2020), we assume that the underlying logic of citation activity is that the citing and cited papers share some common scientific knowledge. Specifically, in the biomedical field, a triplet format (e.g. <head_entity, subj, tail_entity>) is well adapted to represent domain knowledge and is organized in various databases, such as gene ontology (Ashburner *et al.* 2000) and MeSH terms (https://www.nlm.nih.gov/mesh/meshhome.html).

**Received:** 20 January 2025. **Revised:** 7 October 2025. **Accepted:** 7 January 2026

Biological triplets provide a structured, interpretable framework for representing complex biomedical relationships by explicitly denoting the subject, predicate, and object. Compared to unstructured text or pairwise relations, they offer clearer, context-rich information while facilitating efficient data retrieval and analysis. This standardization not only makes it easier to identify and understand entity roles and interactions but also supports interoperability across diverse data sources. Consequently, triplet-based representations are particularly advantageous for tasks like citation intent analysis in biomedical domains, where seamless knowledge sharing and comprehensive relationship mapping are critical. Our task can then be formalized as learning and predicting the triplet(s) that can connect the citing and cited papers.

Particularly, we use PubMed as our text source, which contains a vast volume of publications in biomedical field. The literature focused on PubMed encompasses various bibliometric studies, with some comparing PubMed to other academic search tools such as Web of Science, Google Scholar, and Scopus (Kambhampati *et al.* 2021). Other studies focus on analyzing publication trends on specific biomedical topics using PubMed and similar tools (Yoon *et al.* 2021). More recently, pre-training models have demonstrated their powerful capabilities in natural language processing (NLP). There are two main types of pre-training models: (i) BERT-like models, which are primarily designed for language understanding tasks such as sequence classification and labeling (Devlin *et al.* 2019) and (ii) GPT-like models, which excel at language generation tasks such as abstract generation (Brown *et al.* 2020). These models are initially pre-trained on large-scale corpora collected from the web using self-supervised learning tasks (e.g. masked language modeling for BERT and auto-regressive language modeling for GPT). They are then fine-tuned on specific downstream tasks for text mining and knowledge discovery in biomedical literature (Yuan *et al.* 2022).

In this study, we introduce a dual-encoder–decoder framework designed to predict the shared knowledge between pairs of citing and cited papers in terms of biological triplets, specifically tailored for text mining and inference in the biomedical domain. We compared our proposed model with multiple state-of-the-art models in NLP. Experimental results demonstrate that our proposed method can be effectively generalized across different domains to infer the citation motivation between an existing citing–cited paper pair. Our approach differs from existing recommendation ranking systems, which primarily rely on, e.g. citation counts (Pan and Fortunato 2014), key word matching (Kleminski *et al.* 2022), network-based methods (Sugishita and Asakura 2021), etc. providing a more context-specific suggestion of relevant literature and knowledge.

# 2 Related work

Relation extraction, which identifies semantic relationships between entities, is fundamental in biomedical and life science research. 5Approaches include pipeline-based methods that decompose the task into sub-tasks (often requiring additional annotations) (Wang *et al.* 2020a, 2020b), joint extraction methods (Gupta *et al.* 2019), sequence labeling approaches (where tokens are tagged for entity mentions and relationships) (Wei *et al.* 2020), and table filling techniques (which represent the task as a grid of token pairs to predict relationships) (Wang *et al.* 2020a, 2020b). Another type of related work is representation learning with co-citation analysis (Mysore *et al.* 2022), which focuses on fine-grained aspect matching to improve document similarity tasks with multi-vector representations.

Text generation methods approach the task as a sequence-to-sequence learning task, where the input sequence is the text and the target sequence is the triplet. These methods employ an encoder–decoder network to learn how to generate the triplet from the text (Giorgi *et al.* 2022). However, many joint extraction methods still require additional entity information (Wei *et al.* 2020). In this study, we approach the task as an end-to-end text generation problem. Our method takes only the text as input and generates the relational triplets directly, without the need for additional intermediate annotations (Hou *et al.* 2022).

Document classification assigns documents to one or more predefined categories, serving as a cornerstone for efficient information organization. Large pre-trained language models—such as BERT and its variants—have boosted accuracy in biomedical document classification by learning nuanced language patterns (Yasunaga *et al.* 2022). Meanwhile, generative models can produce label words directly from the text, offering flexibility beyond fixed categories and capturing more complex or context-specific content (Brown *et al.* 2020). Combining these approaches promises even greater accuracy and robustness in biomedical document classification systems (Yasunaga *et al.* 2022).

# 3 Materials and methods

## 3.1 Dataset

In our study, we utilize three primary data sources: unstructured text from PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/), structured biological triplets from multiple existing knowledge databases, and a labeled set of citing–cited paper pairs. Here, we focus on cancer research as an example, but our proposed method can be easily generalized to other biomedical domains.

### 3.1.1 PubMed Central (PMC)

PMC is a free digital archive of biomedical and life sciences journal literature, providing access to a comprehensive collection of research articles. We began our study by retrieving a significant dataset comprising 2.4 million publications from PMC. Within this dataset, we identified 53 043 works related to cancer from a selection of 72 cancer journals listed in PubMed. Each publication was then broken down into its constituent parts, including the title, abstract, and keywords, which served as input for our proposed model. To ensure data quality, we implemented several filtering processes, such as excluding publications with empty abstracts and requiring each publication to have at least one reference present in PMC. These preprocessing steps resulted in a final dataset of 11 620 publications, with the corresponding 109 596 references indexed in PMC.

### 3.1.2 Biological triplets

In our study, we utilized a biomedical triplet approach to capture the scientific reasoning behind citation relationships. A biomedical triplet consists of two biological entities and the

relationship between them, such as ⟨protein *A*, inhibits, protein *B*⟩ or ⟨gene *A*, participates, biological process *B*⟩. These triplets were sourced from public databases, including SynLethDB (Guo *et al.* 2016), CTD (Davis *et al.* 2021), COSMIC (Forbes *et al.* 2017), and UniProt (Apweiler *et al.* 2004). For instance, SynLethDB provided us with 1.4 million triplets, encompassing 24 different relationships and 11 types of medical entities. CTD contributed 1.92 million triplets, involving four types of entities (gene, chemical, disease, and phenotype) and numerous relationships. In total, our knowledge data repository comprises 11.3 million biomedical triplets. Table 1 shows a summary of top 10 biological entity categories in our dataset. Table 2 shows a summary of biological triplets in our data.

Note that we adopted the triplet mechanism rather than sentences is driven by several key considerations. First, triplets provide a compact and machine-readable representation of relationships, which is essential for downstream tasks such as paper recommendation, knowledge graph construction, and automated literature analysis. Second, triplets enable precise and interpretable relationship extraction, distilling the most relevant aspects of citation reasoning into a clear and actionable format i.e. both human-readable and machine-usable. Third, triplet generation reduces noise by focusing on the most salient relationships, avoiding the ambiguity and redundancy often present in free-text sentences. Finally, the structured nature of triplets facilitates scalability, allowing our model to efficiently process and analyze large corpora of scientific literature. By generating triplets instead of sentences, we strike a balance between capturing meaningful relationships and maintaining computational efficiency, ensuring that our approach is both practical and impactful for real-world applications. On the other hand, a meaningful work could explore the integration of triplet generation with natural language sentence generation. This hybrid approach could combine the precision and interpretability of triplets with the rich contextual information provided by sentences. Additionally, future research could investigate the use of triplets as building blocks for constructing large-scale knowledge graphs of scientific literature, enabling more advanced querying and reasoning across domains.

**Table 1** Summary of bio-entities in our study.

| | | |
|---|---|---|
| Gene | 56 428 | 2125 |
| Compound | 18 002 | 653 |
| Biological process | 12 141 | 142 |
| Phenotype | 6064 | 60 |
| Side effect | 5664 | 104 |
| Disease | 3254 | 142 |
| Molecular function | 3012 | 29 |
| Pathway | 2023 | 65 |
| Cellular component | 1619 | 56 |
| Anatomy | 390 | 59 |
| Pharmacologic class | 357 | 13 |
| Symptom | 325 | 18 |

The data sources include SynLethDB, CTD, COSMIC, and UniProt. Note that we only show the top 10 entities. The middle column is the number of each biological entity category in our dataset, and the right column is the number of entities in the annotated triplets dataset.

### 3.1.3 Data annotations

To assign triplet labels to citing–cited paper pairs for supervised model training, we first extracted the citation sentences from each citing paper. These sentences, which contain the reference to the cited paper, demonstrate the motivation behind the citing behavior. The use of citation sentences to explain the relationship between two papers has been explored in previous work (Luu *et al.* 2021), which employed an automatic citation sentence generation model. With the citation sentence, we leveraged the biological triplets to label the citation relationship between a citing–cited paper pair. Specifically, we utilized the data labeling paradigm inspired by distant supervised relationship extraction tasks (Mintz *et al.* 2009). This paradigm operates under the assumption that if a text contains two entities, e.g. $e_h, e_t$, it implies the existence of a proven relational fact represented by a triplet $\langle e_h, r, e_t \rangle$. By employing this strategy, we mitigated the challenge of limited availability of large-scale annotated corpora (Lin *et al.* 2016, Ji *et al.* 2017): We first extracted biological entities from the citation sentence. Then, we retrieved triplets containing these entities as head/tail from the integrated knowledge dataset mentioned in Section 3.1.2. Note that one citation sentence can be matched to one or multiple triplets. In total, we were able to label 13 092 citing–citation paper pairs, and divided this dataset into training, validation and testing sets according to the ratio 8:1:1.

We used citation sentences to extract triplets solely for the purpose of annotating the citing–cited paper pairs, which is essential for the supervised model training process. The learned model, once trained, not only be able to predict the citation motivation between an existing citing–cited paper pair but also infer potential citation relationships between papers that do not yet have a citation link. Consequently, our model can be

**Table 2** Summary of the top 10 bio-relationships in entire dataset (left) and also our annotated data (right).

| | | | |
|---|---|---|---|
| (Compound, decreases expression, gene) | 462 708 | (Gene, synthetic lethality, gene) | 4310 |
| (Compound, decreases expression, gene) | 425 709 | (Gene, interacts, gene) | 3885 |
| (Gene, participates, biological process) | 393 049 | (Disease, associates, gene) | 3776 |
| (Anatomy, expresses, gene) | 358 005 | (Gene, participates, biological process) | 1727 |
| (Gene, regulates, gene) | 147 639 | (Anatomy, expresses, gene) | 1626 |
| (Compound, causes, side effect) | 135 063 | (Gene, participates, pathway) | 1075 |
| (Compound, affects expression, gene) | 127 906 | (Compound, binds, gene) | 914 |
| (Gene, interacts, gene) | 87 103 | (Chemical, increases expression, gene) | 738 |
| (Gene, participates, molecular function) | 65 207 | (Compound, treats, disease) | 737 |
| (Gene, participates, cellular component) | 59 054 | (Chemical, decreases activity, gene) | 695 |

effectively utilized for scientific recommendation, identifying relevant literature, and suggesting potential references.

### 3.1.4 Annotation human validation

Given that our annotation process relies on distant supervision, we conducted a manual evaluation to ensure its quality. We randomly selected 2000 annotated samples for evaluation by two researchers with expertise in bioinformatics. Each scholar independently scored the samples based on the traditional Likert scale (Likert 1932), where a score of 0 indicates that the annotation is completely incorrect and a score of 5 indicates that the annotated triplets effectively aid in understanding the citation relationship between the two articles.

The results of the manual evaluation are present in Fig. 1. The mean scores assigned by the two evaluators were 3.74 and 3.98, respectively, indicating highly agreement between the annotation and human interpretation. This suggests that the quality of our data annotation meets the necessary standards for experimental use. Additionally, we calculated the Kappa consistency between the two evaluators' results, which was 0.81, indicating a strong level of agreement on the validation process.

## 3.2 Data augmentation

In biomedical research, word polysemy is common, and terminology evolves over time. This creates challenges when generating triplets, as older terms may surface in cited sources. Replacing outdated expressions with their more widely recognized equivalents ensures the resulting triplets remain clear and understandable. For example, "Calcimycin" is more familiar today than its older synonyms ("Carboxylic Acids" or "A-23187"). Using the well-known term helps maintain clarity and accessibility in generated triplets.

To address these challenges, we employed entity synonym replacement and masking techniques, and we also integrated the Unified Medical Language System (UMLS) (Bodenreider 2004) to aid this process. These strategies enhance our model's understanding of evolving biomedical terminology and ensure accurate and current triplet generation. Figure 2 illustrates an example of how synonym replacement was applied in our study.

- Entity synonym replacement: We initially gathered all biomedical terms and their 249 610 synonyms from MeSH, which serves as the vocabulary documentation for UMLS. Subsequently, if the entity words in the generated triplets were found in the selected MeSH term pool, we performed random replacements in the input context to assess whether the model could still generate the original entities.
- Entity masking and self-supervise: To improve the model's contextual understanding and enhance its generalization capabilities, we implemented a masking strategy inspired by the BERT model (Lewis *et al.* 2020). This strategy involved randomly replacing entities with the token "unknown". During our experiments, we progressively increased the masking ratio from 0% to 100% to assess the impact of this approach on the model's performance.

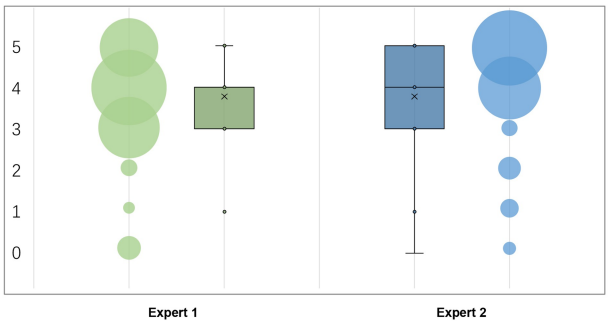


**Figure 1** Manual validation of 2000 randomly selected samples by two researchers with background in bioinformatics.

## 3.3 Method

Recent research often frames triplet extraction as an information extraction problem, where entities are tagged and relationships are classified based on text cues (Yadav and Bethard 2019, Nicholson and Greene 2020). Although these methods can identify genes, drugs, and diseases effectively, they rely heavily on local context and struggle with synonyms or alternative forms, which are widespread in biomedicine. Consequently, accurately

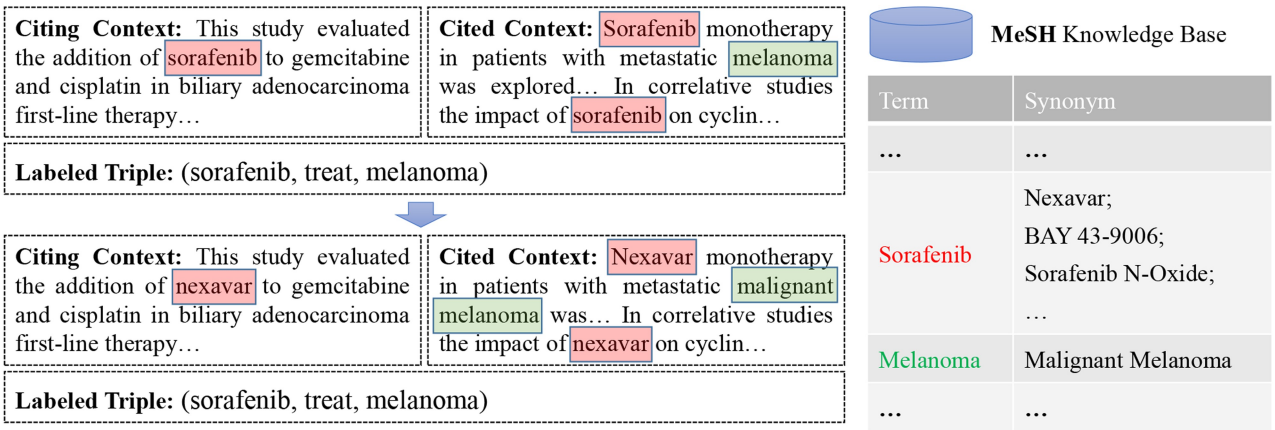


**Figure 2** An example of data augmentation with synonym replacement strategy.

recognizing and linking diverse entity names remains challenging, and information extraction techniques offer limited capacity to synthesize knowledge beyond what is immediately visible in the text.

### 3.3.1 Proposed model

Our proposed model, illustrated in Fig. 3, is an enhanced variant of the pointer-generator network (PGN) text generation model by See *et al.* (2017), enriched with an attention mechanism. Specifically, for the encoder part of our model, we adopted a single-layer bidirectional LSTM, similar to the architecture used in See *et al.* (2017). This allows the model to capture contextual information from both past and future tokens, which is crucial for understanding the relationships between entities in citation sentences. For the decoder, we used a single-layer unidirectional LSTM, which generates the output sequence step by step, ensuring efficient and accurate prediction of head entity–relation–tail entity triplets. The model is tailored to generate one or more triplets that offer biomedical insights into the citation relationship between two input papers. In contrast to prior sequence-to-sequence (seq2seq) models utilized for triplet learning (Liu *et al.* 2018), our model features a dual-encoder architecture. This innovative structure enables the model to individually process each paper as input, rather than merging them into a single sequence with constrained input capacity, e.g. input length limitation.

In the context of a citing and cited paper pair, denoted as $x'$ and $x''$ respectively, the dual-encoder model learns latent embeddings for these two papers as $h'$ and $h''$. Subsequently, at step $t$, the output $S$ from the prior time step functions as one input, and the attention score $e^t$ is computed according to Equation (1). The model's parameters, encompassing the trainable parameters associated with $h'$, $h''$, and $S$ denoted by $w$, alongside the bias term represented as $b_1$, are integrated. Additionally, a learnable weight $\nu$ assigned to the tanh activation function is included, facilitating the mapping of the hidden embedding into the scoring space. Leveraging the attention score $e^t$, the attention weights $w_{att}^t$ are computed through the softmax function, culminating in the derivation of the final context vector $C^t$. This final context vector is attained through weighted averaging, involving the weighted consideration of the hidden states $h_i'$ and $h_i''$ from the respective encoders

$$e^t = \nu^T \tanh(w_{h'} h' + w_{h''} h'' + w_s S^{t-1} + b_1) \quad (1)$$

$$w_{att}^t = \text{softmax}(e^t) \quad (2)$$

$$C^t = \sum_i w_{att}^t [h', h''] \quad (3)$$

$$P_{vocab} = \text{softmax}(v'(v([S^{t-1}, C^t]) + b_2) + b_3) \quad (4)$$

$$P_{gen} = \sigma(w_c C^t + w_s S^{t-1} + b_4) \quad (5)$$

In Equation (4), we first concatenate the context vector $C^t$ and the decoder output state $S^t$ and then pass them through two linear layers to generate the vocabulary probability distribution $P_{vocab}$. The associated learnable parameters for this process are denoted as $v$, $v'$, $b_2$, and $b_3$. We use the weight $P_{gen}$ to determine whether the generated words are taken from the vocab or from the input text. In the biomedical field, there are usually some proprietary terms, such as specific genes and drugs, that are not in a vocab consisting of commonly used words. Therefore, we hope the model could be able to copy these words from the original text when generating the triplets, if these words are irreplaceable

$$P(w) = P_{gen} P_{vocab}(w) + (1 - P_{gen}) \sum_{i:w_i=w} a_i^t \quad (6)$$

In Equation (6), we use weights $P_{gen}$ and $P_{vocab}$ as inputs to obtain the final probability. As such, $P_{vocab}(w)$ is 0 when the generated word does not exist in the vocab, and $\sum_{i:w_i=w} a_i^t$ is 0 when the word does not appear in the source input text.

### 3.3.2 Objective function

In our methodology, we address the challenge of repetitive output sequences in text generation by integrating a coverage mechanism (Tu *et al.* 2016). This mechanism acts as a safeguard against the repetition of identical sequences throughout

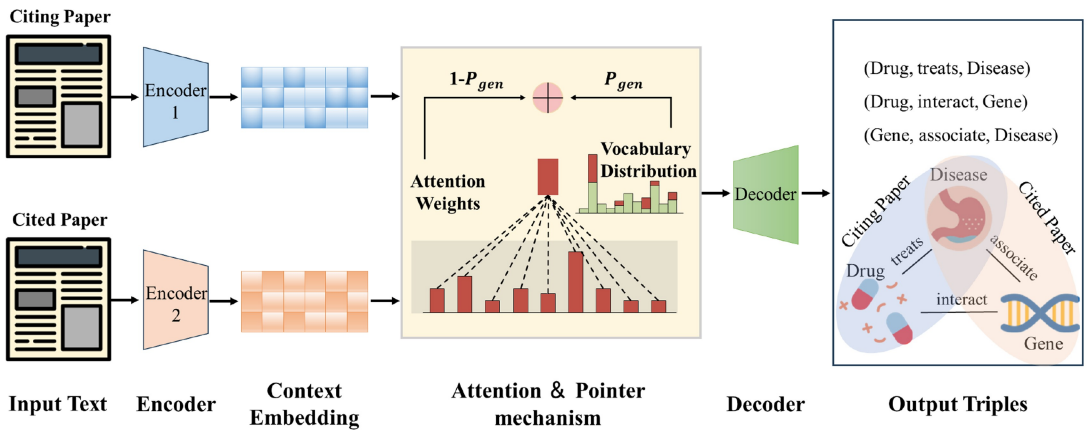


**Figure 3** Overview of the proposed model. The model is a multi-source pointer-generator network based on encoder–decoder structure: two encoders take in the citing and cited papers, respectively, to learn vector representations in the feature space. The decoder can decode one or more triples containing two medical entities and one relationship to help understand the relationship between two articles. For example, if article A studies a disease X and article B studies a drug Y that treats that disease, then using article A and article B as inputs we would expect to get (Drug Y, treatment, Disease X).

decoding. At each time step $t$, a coverage vector $Cov$ is introduced to oversee the decoding progression. The coverage vector is calculated by aggregating the attention distributions from all preceding decoder time steps. By utilizing the coverage vector, our model can monitor the attention weights allocated to distinct segments of the input sequence, ensuring that previously focused regions receive less weight in subsequent steps. This fosters diversity in the generated output. Subsequently, the resulting coverage vector is assimilated as an additional input for the final attention computation, as shown in Equation (8)

$$Cov^t = \sum_{t'=0}^{t-1} w_{att}^{t'} \tag{7}$$

$$e^t = \nu^T \tanh(w_{h'} h^{'} + w_{h''} h^{''} + w_s S^t + w_{Cov} Cov^t + b_1) \tag{8}$$

In the final objective function, we introduce a coverage loss term to penalize repetition in the generated output. This loss term encourages the model to produce diverse and non-repetitive sequences. The coverage loss is calculated by taking the element-wise minimum between the attention distribution at each time step and the coverage vector, then summing the resulting values. This loss term is weighted by a hyperparameter $\lambda$, which controls the impact of the coverage loss on the overall training objective. By tuning the value of $\lambda$, we can adjust the tradeoff between generating coherent output and avoiding repetition

$$L_t = -\log P(w^t) + \lambda \sum_i \min(a_i^t, Cov_i^t) \tag{9}$$

# 4 Results

## 4.1 Pilot evaluation: triplet explanations distinguish citation intent

To test whether triplet-based explanations capture citation intent beyond topical similarity, we conducted a pilot on 500 oncology citing papers $C_i, i = 1, .., 500$. For each $C_i$, we sampled up to 20 cited references ($R_j$; 10 000 ($C_i, R_j$) pairs total) and retrieved 20 PubMed "similar articles" ($S_i$; 10 000 ($C_i, S_j$) pairs total) via PubMed BM25 ranking. We generated biological triplets for every pair, converted them to text, embedded with BioSentVec (13 GB model; default parameters) (Zhang *et al.* 2019), and trained a logistic-regression classifier with five-fold cross-validation to predict citation status.

The classifier achieved 73.3% accuracy and 0.70 F1 under a 1 : 1 cited vs. similar setup (chance $\approx 50\%$), indicating that triplets encode signal specific to citation reasoning. Across the 500 citing papers and 10 000 top-ranked "similar" articles, only 6.42% were actually cited; for many $C$, none of their true references appeared among the top-20 similar results. Treating these as non-cited does not change the conclusion. These preliminary results suggest that triplet representations reflect meaningful citation intent and can support both explanation and discriminative citation modeling; future work will scale to larger datasets and incorporate ranking-based evaluations.

## 4.2 Evaluation metrics & implementation details

As the models produce sequences as output, the generated sequence may contain multiple triplets. To handle the output effectively, we implement a simple post-processing step where we segment the output sequence into individual triplets according to the standard triplet format $\langle e_h, r, e_t \rangle$.

Our evaluation metrics for triplet prediction are based on triplet granularity rather than token-wise matching. A predicted triplet is considered correct if it exactly matches the ground truth triplet in terms of the head entity, tail entity, and relation. Based on this criteria, our model's performance evaluation involves two main criteria: fuzzy match and exact match.

### 4.2.1 Fuzzy match

Fuzzy match allows for partial correctness, meaning some predicted triplets may be correct while others may be incorrect, missing, or additional. This criterion permits partial matches between the generated output and the ground truth. Precision, recall, and $F1$ score are computed at the triplet level. For example, suppose predictions contain two triplets A, B and ground truth contains three triplets, A, C, D, then precision = $1/2$ (only triplet A fully matches), recall = $1/3$ (only one of three required ground truth triplets is recovered). This flexibility in fuzzy match accounts for variations in prediction completeness and provides a more nuanced assessment of model performance, particularly in cases where the exact number of triplets is difficult to predict.

### 4.2.2 Exact match

For exact match, not only must every predicted triplet be correct, but the total number of predicted triplets must also match the number of ground truth triplets. Specifically, if the ground truth encompasses multiple triplets, denoted as $\langle e_{h_i}, r_i, e_{t_i} \rangle$, the generated results must precisely match each triplet to meet these criteria. This is a strict criterion for output evaluation.

Given that our input comprises two articles, while all baseline models are seq2seq models designed for a single input sequence, we tackle this incongruity by concatenating the content of the two input articles to create a unified input sequence for the baseline models. The proposed model underwent training with a batch size of 64 for 50 000 epochs. We leverage Adagrad optimization method with an initial value of 0.15 for the optimization process. All models were implemented using PyTorch libraries with an Nvidia 4090 GPU. For the baseline models discussed in the following section, we utilized pre-trained models from the Hugging Face transformers library (https://www.hugging-face.org/hugging-face-model-hub/). The models were configured with default parameters, except for adjustments made to the input length set to 1024 and the output length to 64. All baseline models compared in our experiments were fine-tuned on the same training data used for our model to ensure a fair comparison. This includes preprocessing steps, dataset related hyperparameter tuning and evaluation protocols, which were kept consistent across all models.

## 4.3 Baseline models

For comparison purpose, we adopted eight seq2seq models as baselines, which are the SOTA models for text generation.

Pointer generator network (PGN): PGN (See *et al.* 2017) is one of the most widely used sequence-generation models for tasks such as text summarization. It allows the original content from the input to be "pasted and copied" directly as output during the text generation process.

Bidirectional encoder representation from transformers (BERT): BERT (Devlin *et al.* 2019) is a pre-trained language model based on the transformer (Vaswani *et al.* 2017) encoder part. It is very robust for semantic representation learning and can be further fine-tuned for various NLP tasks. For our purpose, the decoder part of the transformer model will output a triplet interpreting the citation reasoning.

Bidirectional and auto-regressive transformers (BART): BART (Lewis *et al.* 2020) adopts a standard transformer-based encoder–decoder architecture with a strategy for noise reduction in the encoder part, including various mechanisms handling input corruptions.

BioBERT and BioBART: BioBERT (Lee *et al.* 2019) and BioBART (Yuan *et al.* 2022) are biomedical-specific versions of BERT and BART. Both models employ biomedical publications as training corpus for their pre-training tasks, and target biomedical-oriented text mining tasks.

InfLLM$_{\text{Bart}}$: InfLLM (Xiao *et al.* 2024) is an advanced method for large language models (LLMs) that effectively manages long input sequences. It stores distant contexts in additional memory units and uses an efficient mechanism to look up token-relevant units for attention computation, enabling the model to better capture long-distance dependencies. We applied this mechanism to the BART model to mitigate the effects imposed by the input text length limitation.

ChatGLM3-6B: ChatGLM3-6B (Du *et al.* 2022) is a lightweight, open-source LLM equipped with 6.2 billion parameters. Utilizing technology similar to ChatGPT, ChatGLM3-6B has been trained on a corpus containing $\sim$1 trillion tokens. The training process includes supervised fine-tuning, feedback self-help mechanisms, and human-feedback reinforcement learning. As a result, ChatGLM3-6B is capable of generating responses that are well-aligned with human interpretations.

Llama3-8B: Llama3-8B (Yang *et al.* 2023) is a next-generation, open-source large-scale language model released by Meta in April 2024, featuring 8 billion parameters. Llama3 supports long-text inputs of up to 8000 tokens and has been trained on a vast corpus exceeding 15 trillion tokens. It incorporates the grouped-query attention technique to enhance model efficiency. Compared to other models of a similar scale, Llama3 has achieved significant advancements and demonstrated superior performance across various NLP tasks.

Note that we also tested GPT-4o and o1 for our tasks with a structured prompt for triplet generation, given input paper pair. Due to their open-ended and generic characteristics, the triplets generated by GPT-4 and o1 often include prefixes, suffixes, or mutations to entities, making it challenging to align them with our labeled structured triplets extracted from existing biological knowledge databases. This misalignment renders GPT-4's outputs less suitable for precise relationship extraction and evaluation. From a practical perspective, GPT-4's massive scale and computational requirements make it infeasible for fine-tuning on our relatively small dataset, unlike the lightweight LLaMA-3-8B model, which is more resource-efficient.

## 4.4 Triplet prediction

We conducted a performance comparison of various models on triplet prediction, and the results are present in Table 3. Our proposed approach outperformed all the baseline models. The baseline models used a concatenation of the two articles' contexts into a single paragraph as input, and truncation will be applied if there is an input length limitation. Differently, our approach used a dual-encoder architecture to independently learn the contextual information from the two input articles.

Additionally, BioBERT and BERT exhibited lower overall performance compared to BioBART and BART, which can be attributed to their different pre-training strategies. BART employs a noise recovery mechanism from corrupted texts, enhancing model generalization compared to BERT.

## 4.5 Impacts of the input contents

Table 4 illustrates the impact of different paper sections on model performance. We examined four types of metadata extracted from each paper: title, abstract, keywords, and venue information. The title and keywords typically offer insights into a paper's main research topics, including the research subjects and questions addressed. Additionally, venue metadata, which encompasses journal titles and subject terms, is also pertinent

**Table 3** Model performances for the baselines and the proposed one.

| Model | *P* (%) | *R* (%) | *F*1 (%) | EM (%) |
|---|---|---|---|---|
| PGN | 16.80 | 12.07 | 14.04 | 3.46 |
| BERT | 24.41 | 21.83 | 23.05 | 17.01 |
| BioBERT | 26.65 | 24.83 | 25.70 | 20.00 |
| BART | 69.37 | 66.68 | 67.99 | 56.76 |
| BioBART | 66.96 | 64.31 | 65.61 | 54.40 |
| InfLLM$_{\text{Bart}}$ | 68.22 | 67.21 | 67.71 | 57.46 |
| ChatGLM3-6B | 63.50 | 62.70 | 63.10 | 59.26 |
| Llama3-8B | 53.43 | 55.49 | 54.44 | 45.56 |
| **Ours** | **73.65** | **73.46** | **73.55** | **66.28** |

Exact match requires the three components in a triplet, head bio-entity, relation, tail bio-entity, all match the ground truth label. Our proposed model can achieve the best score of *F*1 comparing to other five baselines. Notably, these base line models are the state-of-the-art sequence-generation models. P: Precision; R: Recall; EM: Exact Match. Best results are highlighted in bold.

**Table 4** Results based on different inputs.

| Input form | *P* (%) | *R* (%) | *F*1 (%) | EM (%) |
|---|---|---|---|---|
| A | 73.65 | 73.46 | 73.55 | 66.28 |
| AT | 77.03 | **77.00** | 77.01 | **68.91** |
| AK | 76.66 | 76.12 | 76.39 | 68.90 |
| AV | 74.35 | 74.12 | 74.24 | 67.11 |
| ATKV | **77.15** | 76.91 | **77.03** | **68.91** |

Each section of a paper may contain different, complementary or redundant information for the prediction. Our experiment demonstrates that the model with an input combination of title and abstract can produce similar result to the one with all the information. A: abstract; T: title; K: keywords; V: venue; P: Precision; R: Recall; EM: Exact Match. Best results are highlighted in bold.

to citations. For example, biomedical researchers are more likely to cite journals in fields such as biomedicine, computer science, or mathematics, and less likely to reference journals related to literature or architecture.

The proposed model demonstrates improvements in both *F*1 and Exact Match metrics by incorporating additional metadata, as opposed to using only the abstracts of the two articles as input text. Notably, the inclusion of title information results in the most significant enhancement. With the titles incorporated, the generated triplet results show a 4.7% increase in *F*1 and a 4.0% increase in Exact Match compared to using only the abstracts. In contrast, the impact of keywords and venue information on model performance is relatively modest. This may be due to the venue information providing less distinct features compared to keywords and titles.

## 4.6 Augmentation results

To augment the training data, we employed two strategies as described in data augmentation section: synonym replacement and masking. Synonym replacement involved substituting entity terms in the input context with corresponding synonyms from the MeSH knowledge bases. Masking, on the other hand, entailed replacing entity terms with meaningless strings. Specifically, we focused on replacing the head and tail entities based on the triplet labels. We applied these augmentation strategies incrementally, increasing the augmented training data by 10% at each step.

In the augmentation experiments, we used the combination of title, abstract, keywords, and venue information as the model's input. Figure 4 presents the results obtained with different augmentation ratios in the training data. The model performance, measured by *F*1 and Exact Match, improves as the augmentation ratio increases from 0% to 10%. However, beyond this range, the augmentation strategies start introducing artifacts and noise, leading to a decline in both metrics. With a 10% augmentation ratio, we can achieve optimal performance, yielding precision, recall, *F*1-score, and Exact Match scores of 79.45%, 78.96%, 79.20%, and 72.92%, respectively.

## 4.7 Comparison with natural language explanation

To further evaluate our model's interpretability, we randomly sampled 100 citing papers from the test set and collected all their reference papers, framing the task as a one-to-many citation relationship. Since each citing paper typically includes multiple references, this setup reflects the complexity of real-world citation behavior. We then used both a knowledge graph and a citation graph to represent the output of our model for each citing paper. Additionally, we randomly selected five cited references per citing paper and asked GPT4o to provide natural language explanations for the citation motivations. To assess the usefulness of the two different explanation methods, we recruited five domain researchers to evaluate how well each approach helped them understand the underlying citation intent.

We then asked five researchers to evaluate the explanation results based on the criteria listed in Table 5. These criteria are commonly adopted in the field of explainable AI and include measures such as usefulness, clarity, and relevance (Tintarev and Masthoff 2015, Zhang and Chen 2020, Kroll *et al.* 2024). Each criterion was rated on a scale from 0 to 5, and the final scores reported represent the average ratings across all five evaluators. This evaluation allows us to assess the effectiveness of the explanations in supporting human understanding of citation motivations.

Based on the human evaluation results, the usefulness score and the ease of understanding of the graph composed of ternary groups are better than that of natural language in terms of their effectiveness in helping users to understand citation behaviors. Particularly, ease of understanding is significantly higher than that of natural language results generated by GPT4o.

We also did a follow-up discussion with the evaluators and identified the key reason for the aforementioned results: the natural language explanations become excessively verbose when multiple cited documents are involved. Statistically, in our

**Table 5** Usefulness: whether the results could help users understand the underlying logic behind the citing behavior.

| Metric | Ours | GPT4o |
|---|---|---|
| Usefulness | **3.96** | 3.94 |
| Ease of understanding | **4.10** | 3.26 |
| Persuasiveness | 3.78 | **3.91** |

Ease of understanding: whether the content expression, structural organization, and presentation style are easy for users to understand. Persuasiveness: whether the interpretable results are reasonable and correct. Best results are highlighted in bold.

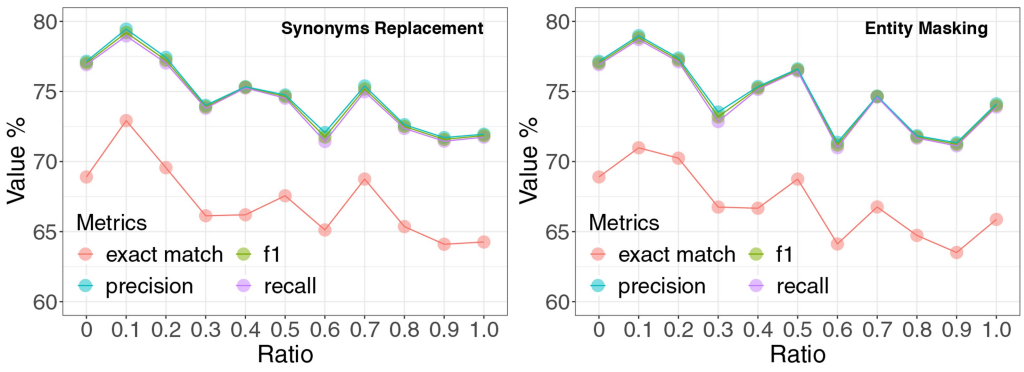


**Figure 4** Changes of the model performances as we applied synonym and masking strategies with different ratios to the input content.

experiment, the text explanations generated by GPT4o had an average length of 3595.60 characters and a word count of 486.01. These overly long texts significantly increase the user's cognitive load and hinder comprehensive understanding of the content.

# 5 Discussion

In this study, we present an encoder–decoder-based framework designed to learn and predict biomedical interpretations for existing citation relationships among published works, leveraging existing knowledge. The model is trained on data from PubMed Central and multiple biomedical relation databases that contain relationships among bio-entities. Our motivation is to enhance the transparency of the academic recommendation process, which remains opaque in publicly available search tools. This lack of transparency forces researchers to manually sift through recommendation lists to identify relevant works. Additionally, existing scholar-search tools often prioritize papers based on ranking algorithms that heavily weigh the number of citations. As a result, closely related but newly published papers struggle to appear at the top of recommendation lists.

The strategy of adopting triplets can offer a structured and compact representation that captures the core semantic relationship between citing and cited papers, making them especially useful for downstream tasks such as citation recommendation, literature summarization, and knowledge graph construction. Unlike long-form natural language explanations, triplets are easier to compare, visualize, and align across multiple citation contexts, enabling scalable and interpretable analysis. Their predefined structure also facilitates integration with existing knowledge bases and supports automated reasoning in complex multi-paper citation scenarios.

Our proposed model addresses these challenges by taking a pair of citing and cited papers as input and generating a triplet (⟨bio-entity, relation, bio-entity⟩) as output. The bio-entities represent the main subjects of the input papers, and the relation describes the known relationship between these entities. We demonstrate that our model outperforms other state-of-the-art seq2seq text generation models for this task using various evaluation criteria. Furthermore, we conduct experiments to assess the contribution of different parts of a paper to the reasoning behind citations. We also tackle the issue of terminology synonyms in the biomedical field by constructing a dictionary for entity replacement, ensuring that the original semantic meaning is preserved. Additionally, we explore data augmentation techniques, such as terminology masking, to evaluate the model's ability to accurately recover masked entities. Our task fundamentally differs from general related papers in its focus on structured relationship extraction rather than broad document similarity or co-citation analysis. While many existing approaches aim to identify related papers based on topics, keywords, or citation patterns, our work specifically targets the extraction of head entity–relation–tail entity triplets to capture the explicit rationale behind citations. By focusing on fine-grained, domain-specific relationships, our work complements broader similarity-based methods and offers a novel perspective on understanding scientific literature.

During our study, we observed that although there are multiple LLMs available for the biomedical field which have revolutionized language processing and understanding, their performance in semantic reasoning remains unsatisfactory (Wei *et al.* 2024). This variability can be problematic in biomedical contexts where precision and consistency are crucial. Moreover, these LLMs struggle with context sensitivity. They often fail to capture the comprehensive and specific relationships between different pieces of information within the same or across different documents (Giallanza and Campbell 2024, Wei *et al.* 2024). This lack of contextual understanding can lead to inaccuracies, particularly in tasks that require a deep understanding of semantic relationships, such as citation reasoning and the interpretation of scientific literature. The inability to effectively integrate context diminishes the reliability of these models in providing accurate and meaningful insights, which is a significant limitation for their application in specialized fields like biomedicine. In addition to these issues, the current LLMs often overlook the importance of domain-specific knowledge and terminologies, which are critical for accurate semantic reasoning in biomedical texts. This shortcoming underscores the need for improved models that can handle context-sensitive tasks and provide consistent, accurate predictions. Our study aims to address these gaps by developing methods that ensure better semantic reasoning, context integration, and the effective handling of domain-specific knowledge, ultimately enhancing the applicability and reliability of language models in biomedical research.

Our model's ability to generate structured ⟨bio-entity, relation, bio-entity⟩ triplets offers a complementary approach to enhancing paper recommendation systems. While traditional systems excel at identifying broadly related papers based on metrics like citation counts or keyword overlap, our approach provides fine-grained, interpretable relationships between citing and cited papers. These structured relationships can be integrated into existing recommendation frameworks to add a layer of contextual relevance. For instance, when a researcher is reading a paper on a specific method, our model can supplement traditional recommendations by suggesting related papers based on existing biological knowledge, thereby offering a more nuanced understanding of the literature. Furthermore, our model's ability to recommend newly published papers—even those with few citations—complements citation count-based systems, which often favor older, well-established papers. By incorporating the rationale behind citations, our approach not only enhances the diversity and relevance of recommendations but also provides explanations for why a paper is suggested, fostering greater trust and usability. This complementary capability makes our model a valuable addition to the toolkit for tasks like literature review, interdisciplinary research, and knowledge discovery, where understanding the why behind recommendations is as important as the recommendations themselves.

Our future work will focus on several key directions to further improve the proposed model. Firstly, we aim to enhance the learning of semantic information from input articles by exploring different strategies. This includes incorporating pre-trained language models like BioBERT to capture domain-specific knowledge, introducing self-attention mechanisms to capture internal dependencies within texts, and employing multi-headed attention to model interactions between the two articles. We can also utilize advanced models like GPT-4 as a preprocessing tool to generate candidate triplets or provide

contextual insights, which could then be refined and validated by our supervised model. This hybrid approach would combine the broad contextual understanding of GPT-4 with the precision and domain-specificity of our current method. Additionally, advancements in fine-tuning techniques and computational resources may make it feasible to adapt GPT-4 or similar models for structured triplet generation, potentially improving scalability and generalization. Secondly, we plan to refine the training process to obtain a large-scale labeled corpus. Currently, our semi-supervised approach for obtaining triplets from citation sentences introduces some noise. Implementing noise reduction mechanisms will help improve the performance and reliability of the model. Thirdly, we recognize the limitations of the knowledge bases we used, such as their focus on specific domains and limited coverage. To overcome this, we aim to leverage knowledge bases with broader coverage and higher quality, which can enhance the prediction task by providing more comprehensive and accurate information. Lastly, we plan to expand our research to cover the entire citation network. This will involve conducting citation reasoning on a broader scale, considering not only citing and cited paper pairs but also exploring the relationships within the entire network of paper references. By analyzing the entire citation network, we aim to uncover more complex patterns and connections among scientific works. This expanded scope will allow us to identify indirect relationships, discover clusters of related papers, and gain a more comprehensive understanding of the research landscape. Finally, we will expand and publish the labeled citation dataset that includes scientific reasoning associated with each reference relationship. By sharing these resources, we aim to enhance the transparency and effectiveness of academic recommendation systems.

## Author contributions

Pengcheng Li (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Methodology [equal], Resources [lead], Software [lead], Visualization [equal], Writing—original draft [Supporting], Writing—review & editing [equal]), Kai Zhang (Validation [equal], Writing—review & editing [equal]), Xiaozhong Liu (Conceptualization [equal], Writing—review & editing [equal]), and Xuhong Zhang (Conceptualization [equal], Methodology [lead], Project administration [lead], Supervision [lead], Validation [equal], Visualization [equal], Writing—original draft [lead], Writing—review & editing [equal])

## Conflicts of interest

None declared.

## Funding

Pengcheng Li is supported by Hubei Provincial Department of Education Young Talent Project: Research on Controllable Generation Methods for Scientific Literature Reviews Based on Thought Chains and Large Language Models (Q20241406).

## References

Apweiler R, Bairoch A, Wu CH *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:D115–9.

Ashburner M, Ball C, Blake J *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

Athar A, Teufel S. Detection of implicit citations for sentiment detection. In: *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse.* Jeju Island, Korea, 2012, ACL Anthology, 18–26.

Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; **32**:D267–70.

Bronk C, Reichard J, Qi L. A co-citation analysis of purpose: trends and (potential) troubles in the foundation of purpose scholarship. *J Posit Psychol* 2023;**18**:1012–26.

Brown T, Mann B, Ryder N *et al.* Language models are fewshot learners. *ACL (Volume 1: Long Papers)* 2020;**33**:1877–901.

Davis AP, Grondin CJ, Johnson RJ *et al.* Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res* 2021; **49**:D1138–43.

Devlin J, Chang M, Lee K *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In: *NAACL-HLT*. Minneapolis, US, 2019, ACL Anthology, 4171–86.

Donthu N, Kumar S, Mukherjee D *et al.* How to conduct a bibliometric analysis: an overview and guidelines. *J Bus Res* 2021; **133**:285–96.

Du Z, Qian Y, Liu X *et al.* GLM: general language model pretraining with autoregressive blank infilling. In: *ACL*. Dublin, Ireland, 2022, ACL Anthology, 320–35.

Forbes SA, Beare D, Boutselakis H *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;**45**:D777–83.

Giallanza T, Campbell D. Context-sensitive semantic reasoning in large language models. In: *ICLR 2024 Workshop on Representational Alignment.* Vienna, Austria, 2024, ICLR.

Giorgi J, Bader GD, Wang B. A sequence-to-sequence approach for document-level relation extraction. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*, Dublin, Ireland. 2022, ACL Anthology, 10–25.

Guo J, Liu H, Zheng J *et al.* SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res* 2016;**44**:D1011–7.

Gupta P, Yaseen U, Schütze H. Linguistically informed relation extraction and neural architectures for nested named entity recognition in BioNLP-OST 2019. In: *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks.* HongKong, China, 2019, ACL Anthology, 132–42.

Hou Y, Xia Y, Wu L *et al.* Discovering drug-target interaction knowledge from biomedical literature. *Bioinformatics* 2022; **38**:5100–7.

Ihsan I, Muhammad Q. CCRO: citation's context & reasons ontology. *IEEE Access* 2019;**7**:30423–36.

Ji G, Liu K, He S *et al.* Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: *AAAI 2017'*, San Francisco, US, Vol. 31. PKP Publishing Services Network, 2017.

Kambhampati SBS, Vaish A, Vaishya R *et al.* Trends of arthroscopy publications in PubMed and Scopus. *Knee Surg Relat Res* 2021;**33**:14.

Kleminski R, Kazienko P, Kajdanowicz T. Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification. *J Inf Sci* 2022;**48**:349–73.

Kochhar S, Ojha U. Index for objective measurement of a research paper based on sentiment analysis. *ICT Express* 2020; **6**:253–7.

Kroll H, Kreutz CK, Thang BM *et al.* Building an explainable graph-based biomedical paper recommendation system. In: *Proceedings of the 1st Workshop on Utilizing AI/ML to Enhance Information Extraction, Organization, and Retrieval from Large-Scale Archival Collections (AI4LAC)*, HongKong, China, 2024, CEUR-WS.

Lee J, Yoon W, Kim S *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**39**:1234–40.

Lewis M, Liu Y, Goyal N *et al.* BART: denoising sequence-to-sequence pre-training for natural language generation. In: *ACL*. online, 2020, ACL Anthology, 7871–80.

Leydesdorff L, Opthof T. Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *J Am Soc Inf Sci Technol* 2010; **61**:2365–9.

Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932;**140**: 55.

Lin Y, Shen S, Liu Z *et al.* Neural relation extraction with selective attention over instances. In: *ACL 2016'*, Berlin, Germany, Vol. 1: Long Papers. 2016, ACL Anthology, 2124–33.

Liu Y, Zhang T, Liang Z *et al.* Seq2RDF: An end-to-end application for deriving triples from natural language text. In: *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks*. Monterey, US, 2018, CEUR Workshop Proceedings, vol. 2180 (paper 37).

Luu K, Wu X, Koncel-Kedziorski R *et al.* Explaining relationships between scientific documents. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, online, 2021, ACL Anthology, 2130–44.

Mejia C, Wu M, Zhang Y *et al.* Exploring topics in bibliometric research through citation networks and semantic analysis. *Front Res Metr Anal* 2021;6:742311.

Meng R, Lu W, Chi Y *et al.* Automatic classification of citation function by new linguistic features. In: *Proceedings of iConference 2017 Proceedings.* Wuhan, China, 2017, iSchools Inc., 826–30.

Mercier D, Rizvi S, Rajashekar V *et al.* ImpactCite: an XLNet-based method for citation impact analysis. In: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021)*, 2021, SCITEPRESS-Science and Technology Publications, Lda., vol. 2, 159–68.

Mintz M, Bills S, Snow R *et al.* Distant supervision for relation extraction without labeled data. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-AFNLP 2009)*, Suntec, Singapore, 2009, ACL Anthology, 1003–11.

Mysore S, Arman C, Hope T. Multi-vector models with textual guidance for fine-grained scientific document similarity. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).* Seattle, US, 2022, ACL Anthology, 4453–70.

Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* 2020;**18**:1414–28.

Pan R, Fortunato S. Author impact factor: tracking the dynamics of individual scientific impact. *Sci Rep* 2014;**4**:4880–7.

See A, Liu P, Manning C. Get to the point: summarization with pointer-generator networks. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, ACL Anthology, 1073–83.

Small H, Garfield E. Analysis of scientific literature to assist in problem solving. *J Am Soc Inf Sci* 1989;**40**:152.

Sugishita K, Asakura Y. Vulnerability studies in the fields of transportation and complex networks: a citation network analysis. *Public Transp* 2021;**13**:1–34.

Tintarev N, Masthoff J. Explaining recommendations: design and evaluation. In: *Recommender Systems Handbook*. Boston, MA: Springer US, 2015, 353–82.

Tu Z, Lu Z, Liu Y *et al.* Modeling coverage for neural machine translation. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, 2016, ACL Anthology, Vol. 1. 76–85.

Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, US, 2017, Curran Associates, Inc., 5998–6008.

Wang D, Hu W, Cao E *et al.* Global-to-local neural networks for document-level relation extraction. In: *Proceedings of the 2020 Conference on Empirical Methods in Natual Language Processing (EMNLP)*, online, 2020, ACL Anthology, 3711–21.

Wang Y, Yu B, Zhang Y *et al.* TPLinker: single-stage joint extraction of entities and relations through token pair linking. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING),* online, 2020b, ACL Anthology, 1572–82.

Wei F, Chen X, Luo L. Rethinking generative large language model evaluation for semantic comprehension. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, 2024, ACM, no. 2153, 52525–52558.

Wei Z, Jianlin S, Wang Y *et al.* A novel cascade binary tagging framework for relational triple extraction. In: *ACL*. online, 2020, ACL Anthology, 1476–88.

Xiao C, Zhang P, Han X *et al.* InfLLM: unveiling the intrinsic capacity of LLMs for understanding extremely long sequences with training-free memory. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024, Curran Associates, Inc., no. 3801, 119638–61.

Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, US, 2018, ACL Anthology, 2145–58.

Yang R, Tan TF, Lu W *et al.* Large language models in health care: development, applications, and challenges. *Health Care Sci* 2023;**2**:255–63.

Yasunaga M, Leskovec J, Liang P. Linkbert: pretraining language models with document links. In: *ACL (Volume 1: Long Papers).* Dublin, Ireland, 2022, ACL Anthology, 8003–16.

Yoon B-H, Hwang BK, Jung E-A *et al.* Citation analysis of the journal of bone metabolism from Korean Citation Index, Web of Science, and Scopus. *J Bone Metab* 2021; **28**:193–9.

Yuan H, Yuan Z, Gan R *et al.* BioBART: pretraining and evaluation of a biomedical generative language model. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*, Dublin, Ireland. 2022, ACL Anthology, 97–109.

Zehlike M, Yang K, Stoyanovich J. Fairness in ranking, Part II: learning-to-rank and recommender systems. *ACM Comput Surv* 2022;**55**:1–41.

Zhang Y, Chen Q, Yang Z *et al.* BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* 2019;**6**:52.

Zhang Y, Chen X. Explainable recommendation: a survey and new perspectives. *Found Trends Inf Ret* 2020;**14**:1–101.


Bioinformatics, 2026, 42, 1–12
https://doi.org/10.1093/bioinformatics/btag061
Original Paper