Beyond Citations: Integrating Finding-Based Relations for Improved Biomedical Article Representations

Yuan Liang
Queen Mary University
London, UK
yuan.liang@qmul.ac.uk

Roonak Rezvani
Recursion Pharmaceuticals, Inc
Oxford, UK
roonak.rezvani@recursion.com

Massimo Poesio
Queen Mary University
University of Utrecht
London, UK
Utrecht, Netherland
m.poesio@qmul.ac.uk

Abstract

High-quality scientific article embeddings are essential for tasks like document retrieval, citation recommendation, and classification. Traditional citation-based approaches assume citations reflect semantic similarity—an assumption that introduces bias and noise. Recent models like SciNCL and SPECTER2 have attempted to refine citation-based representations but still struggle with noisy citation edges and fail to fully leverage textual information. To address these limitations, we propose a hybrid approach that combines Finding-Citation Graphs (FCG) with contrastive learning. Our method improves triplet selection by filtering out less important citations and incorporating finding similarity relations, leading to better semantic relationship capture. Evaluated on the SciRepEval benchmark, our approach consistently outperforms citation-only baselines, showing the value of text-based semantic structures. While we do not surpass state-of-the-art models in most tasks, our results reveal the limitations of purely citation-based embeddings and suggest paths for improvement through enhanced semantic integration and domain-specific adaptations.

1 Introduction

High-quality scientific article embeddings are essential for various downstream tasks, including citation recommendation, article retrieval, and classification (Cunningham and Greene, 2023). These effective representations accelerate research progress by enhancing knowledge discovery. However, generating high-quality embeddings remains challenging, largely due to the limitations of existing methods that rely primarily on citation networks.

Traditional approaches use Large Language Models (LLMs) to generate article embeddings directly, but research shows this method often underperforms compared to basic baseline models like GloVe (Reimers and Gurevych, 2019). To enhance embedding quality, researchers have turned to contrastive learning for refining document representations (Cohan et al., 2020). This method uses a triplet-based training framework, where each triplet includes a query article, a similar article (positive sample), and a dissimilar article (negative sample). These triplets are typically drawn from citation networks, based on the assumption that citation relationships indicate semantic similarity.

Over the years, researchers have made various improvements to optimize triplet selection. SPECTER (Cohan et al., 2020) introduced a unidirectional citation-based approach, using cited papers as positive samples and non-cited papers as negative samples. However, this method created inconsistencies in triplet generation, as the same paper could be both a positive and negative sample in different contexts. To address this issue, SciNCL (Ostendorff et al., 2022) eliminated citation directionality and implemented graph embeddings and k-nearest neighbors (KNN) sampling to identify positive and negative samples. This change significantly improved embedding quality by reducing triplet formation inconsistencies.

Recent advances have further refined this pipeline. SPECTER2 (Singh et al., 2023) developed task-specific embeddings by generating a general representation and then fine-tuning it for different downstream tasks. Other approaches explore multi-faceted embeddings, generating multiple representations of a paper to capture various aspects of its content (Zhang et al., 2023).

Despite these advances, current methods rely solely on citation networks for triplet construction, overlooking the many semantically similar articles that lack direct citation links. This limitation creates biases in representation learning and constrains the quality of scientific embeddings. To address these challenges, we propose a hybrid approach that enhances contrastive learning by combining Finding-Citation Graphs (FCG) with text-based se-

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 297–306

mantic relationships. Our contributions include:

- Filtering less important citations using an LLM-based classification mechanism to remove noisy edges.
- Incorporating finding similarity relations to establish meaningful connections between semantically related papers.

We evaluate our approach on SciRepEval (Singh et al., 2023), a benchmark for assessing scientific embeddings across multiple tasks. Our method outperforms citation-only baselines, demonstrating the effectiveness of integrating text-based semantic structures into contrastive learning. While it does not surpass state-of-the-art models in all tasks, our results highlight the importance of moving beyond purely citation-based embeddings toward richer, more semantically aware representations.

2 Related Work

Researchers have developed various models and methodologies to improve scientific text representation, ranging from traditional keyword-based methods and vector space models to modern deep-learning approaches. Beyond general-purpose techniques, specialized approaches exist specifically for scientific articles.

General-Purpose Methods Early scientific article representations relied primarily on word-level features. The Bag-of-Words (BoW) model represented documents as vectors of word frequencies—a simple but limited approach that suffered from sparsity and lost semantic relationships (Salton et al., 1975). Latent Semantic Analysis (LSA) addressed these limitations by introducing dimensionality reduction and capturing latent word relationships (Deerwester et al., 1990). The field then progressed to probabilistic topic modeling with Latent Dirichlet Allocation (LDA), which effectively modeled texts as mixtures of latent topics (Blei et al., 2003). LDA became particularly valuable in scientific literature analysis by enabling researchers to extract thematic structures from large document collections.

A major breakthrough came with word embedding techniques like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which transformed scientific text representation through dense vector spaces. These models excel at capturing semantic relationships, leading to improved information retrieval and document clustering. However, they face limitations in handling polysemy and contextual variations.

The field has recently advanced further with transformer-based models, notably BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019)—models specifically trained on scientific corpora. These architectures have dramatically improved contextual representation and now power various tasks including citation prediction, summarization, and scientific question answering. SciBERT stands out by outperforming generic language models in domain-specific applications, demonstrating the value of domain-adapted pretraining.

Scientific-Specific Methods Most methods for associating embeddings to scientific papers rely on citation networks, which represent articles as nodes connected by citation links to analyze influence patterns and research trends (Page et al., 1999). Several approaches have developed universal embeddings for articles, such as SPECTER (Cohan et al., 2020) and SciNCL (Ostendorff et al., 2022), as discussed in Section 1. Other approaches generate multiple embeddings for scientific articles, each serving a distinct purpose. For instance, SPECTER2 (Singh et al., 2023) creates task-specific embeddings, producing four different representations per article for tasks like classification, regression, ad-hoc search, and proximity. Similarly, ASPIRE (Mysore et al., 2022) generates aspect-specific embeddings for each article, such as method embeddings and finding embeddings. Despite the noise in citation networks, these models outperform traditional embeddings by leveraging citation relationships, resulting in improved downstream performance in retrieval, classification, and clustering tasks.

3 Methodology

Our goal is to learn citation-informed and text-informed representations for scientific documents. Given a document's textual content d, we aim to generate a dense vector representation e that effectively encodes both the document's information and the citation's information for downstream tasks. Following previous work (Cohan et al., 2020; Ostendorff et al., 2022; Singh et al., 2023), we developed an information-enriched network combining citation networks with finding similarity relations. Using this network, we sample triplets to learn document embeddings through contrastive learning. In the following subsections, we describe the creation of the information-enriched network, the triplet

sampling approach, and the contrastive learning approach.

3.1 Information-Enriched Network Construction

To enhance the semantic similarity of the citation network, we combined citation networks with finding similarity relations to create an information-enriched network. We improved semantic accuracy by filtering out less important citations—those that contribute minimally to the new study. We established new relations between articles with similar findings based on the Finding-Citation Graph (FCG) (Liang et al., 2024). The resulting network contains links—both citations and finding similarity relations—that better represent semantic relationships beyond simple citations.

3.1.1 Citation Filtering

Although citation networks form the foundation of many scientific embedding models, they can introduce noise since not all citations reflect meaningful content similarity (Ostendorff et al., 2022). To address this issue, we implemented a large language model (LLM)-based filtering mechanism that evaluates citations by assessing their contribution to the citing study, thereby determining their relevance.

Due to the lack of open-source datasets for this task, we used Mistral-7B-Instruct (Jiang et al., 2023) with few-shot in-context learning to classify citations into three categories: Highly Important, Moderately Important, and Less Important. Our analysis of citation importance considered three key elements: the citation sentence, the abstract of the citing paper, and the title of the cited paper. The prompt can be seen in Appendix A. Less relevant citations were removed from the network to reduce noise and improve the quality of triplet selection.

3.1.2 Finding Similarity Relations

Beyond citations, scientific findings provide a more precise measure of content similarity between papers. To incorporate additional semantic relationships, we utilized the Finding-Citation Graphs (FCG). We used Contriever (Lei et al., 2023), a dense retrieval model, to convert scientific findings into embeddings. We then calculated pairwise cosine similarity between findings and added new finding similarity edges to the network when pairs exceeded a similarity threshold.

Through these two enhancements—removing noisy citations and introducing new semantic edges—we created an information-enriched citation network that better reflects the true relationships between papers.

3.2 Triplet Sampling

Contrastive learning relies on high-quality triplets—sets of (query, positive, negative) samples to train models to differentiate between similar and dissimilar documents. To enhance our model's performance, we optimized triplet selection by combining citation-based and finding-based similarity measures. Following Ostendorff et al. (2022), we trained node embeddings on the combined network using PyTorch BigGraph (Lerer et al., 2019). For each article d^Q, we used the k nearest neighbors (KNN) method to identify similar (positive) and dissimilar (negative) articles.

For positive article sampling, following Wang and Isola (2022) and Ostendorff et al. (2022), we selected positive articles from locations distant from the query. Specifically, we sampled c^+ positive articles from a close neighborhood around the query article—those within the range $(k^+ - c^+, k^+]$, where k^+ represents the k parameter in the KNN method.

For negative article sampling, we considered two types of negative articles: easy negatives c^-_{easy} and hard negatives c^-_{hard}. Easy negatives can be obtained through simple random sampling. Hard negatives are crucial for contrastive learning—the more challenging the negative samples, the better the model training becomes. We used a sampling method similar to positive article sampling, selecting articles within the range $(k^-_{hard} - c^-_{hard}, k^-_{hard}]$, where k^-_{hard} represents the k parameter in the KNN method.

3.3 Contrastive Learning

Once triplets are constructed, we train our embedding model using contrastive learning with a triplet margin loss function (Schroff et al., 2015). The method's core principle is to minimize the distance between similar (positive) samples in the latent space while maximizing the distance between dissimilar (negative) samples. To implement contrastive learning, we fine-tuned SciBERT (Beltagy et al., 2019), a domain-specific transformer model for scientific text, to generate embeddings for each article.

$$\mathcal{L} = max\{\left\|d^Q - d^+\right\|_2 - \left\|d^Q - d^-\right\|_2 + \xi, 0\} \quad (1)$$

4 Experiment Setup

This section describes our experimental setup, detailing the datasets, model training configurations, and baseline comparisons.

4.1 Dataset

4.1.1 Training Dataset: Finding-Citation Graph (FCG)

For training, we utilized the Finding-Citation Graph (FCG) derived from the Europe PMC dataset (Liang et al., 2024). This biological FCG encompasses 16 million nodes—consisting of 6 million papers and 10 million findings—and 27 million edges, comprising 17 million citations and 10 million paper-finding generation relations. After preprocessing the dataset to filter out noisy citations and incorporate finding similarity relations, as described in the methodology section, this enriched network forms the foundation for our triplet sampling strategy.

4.1.2 Evaluation Dataset: SciRepEval

For evaluation, we used SciRepEval (Singh et al., 2023), the first large-scale benchmark for evaluating scientific document embeddings. SciRepEval encompasses 24 tasks across four evaluation formats—Ad-Hoc Search, Proximity, Classification, and Regression—spanning multiple scientific domains. We primarily used the "Out-of-Train" datasets in SciRepEval. Table 1 provides an overview of the dataset statistics and evaluation metrics.

4.2 Model Training and Implementation

4.2.1 Input Network Variations

To assess performance, we generated different variations of the citation network through distinct preprocessing methods.

- **Citation** – The original unprocessed citation network.
- **Citation (Filtered)** $\mathbb{F}$ – Noisy citations removed.
- **Citation (Finding Similarity)** $\mathbb{T}$ – New finding-based relations added.
- **Citation (Combine)** $\mathbb{FT}$ – Both filtering and finding similarity applied.

4.2.2 Training Configuration

For filtering less important citations, we utilized Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) with two-shot learning on a single NVIDIA GeForce A100 GPU, processing each sample in approximately 0.8 seconds. To identify finding similarities, we used Contriever (Lei et al., 2023) to generate embeddings for all 10 million findings and performed similarity searches, with each search taking about 0.87 seconds.

For triplet generation and contrastive learning, we closely replicated SciNCL's training setup. We implemented the KNN strategy using FAISS (Johnson et al., 2019) with a flat index and maintained the same KNN parameters: $k^+ = 25$ and $k^- = 4000$. For contrastive learning, we used Huggingface Transformers (Wolf et al., 2020) and initialized the model with SciBERT's weights (Beltagy et al., 2019), training it with triplet loss. The training process used the Adam optimizer (Kingma and Ba, 2017) with weight decay and a learning rate of $\lambda = 2^{-5}$. The model was trained for 2 epochs on a single NVIDIA GeForce RTX A100 (40G) GPU with a batch size of 14, completing in approximately 8 hours.

4.3 Baselines

We compared our method against two existing contrastive learning-based scientific embedding models: SciNCL (Ostendorff et al., 2022), and SPECTER2 (Singh et al., 2023). Since these baselines were trained on multi-domain datasets, their results serve as a reference point rather than direct competitors. Our primary goal is to assess whether removing noisy citations and incorporating text-based similarity relations improves embedding quality. Therefore, our true baseline is the unprocessed citation network, which we used to generate embeddings without any filtering or augmentation.

5 Overall Results

We evaluated our approach by building multiple input networks using different preprocessing strategies and comparing them to baselines. Our main goal was to determine if filtering less important citations and incorporating finding similarity relations would enhance the quality of biomedical article embeddings.

Table 2 presents the statistics of each network variant. Due to time constraints, we analyzed citation importance and generated finding similarity

Task Format	Name	Test	Eval Metric	Source
		Out-of-Train		
CLF	Biomimicry	10,991	Binary F1	Shyam et al. (2019)
	DRSM	7,520 S; 955 G	Macro F1	Burns (2022)
	SciDocs MAG	23,540	Macro F1	Cohan et al. (2020)
	SciDocs MeSH Diseases	25,003	Macro F1	Cohan et al. (2020)
RGN	Peer Review Score	10,210	Kendall's $\mathcal{T}$	Singh et al. (2023)
	h-Index of Authors	8,438	Kendall's $\mathcal{T}$	Singh et al. (2023)
	Tweet Mentions	25,655	Kendall's $\mathcal{T}$	Jain and Singh (2021)
PRX	S2AND	X: 68,968 Y: 10,942	B^3 F1	Subramanian et al. (2024)
	Paper-Reviewer Matching	Q:107 P: 1,729	P@5, P@10	Mimno and McCallum (2007)
	RELISH	Q: 3190 P: 191,245	nDCG	Zhao et al. (2022)
	SciDocs Co-view	Q: 1,000 P: 29,978	MAP, nDCG	Cohan et al. (2020)
	SciDocs Co-read	Q: 1,000 P: 29,977	MAP, nDCG	Cohan et al. (2020)
	SciDocs Cite	Q: 1,000 P: 29,928	MAP, nDCG	Cohan et al. (2020)
	SciDocs Co-cite	Q: 1,000 P: 29,949	MAP, nDCG	Cohan et al. (2020)
SRCH	NFCorpus	Q: 323 P: 44,634	nDCG	Boteva et al. (2016)
	TREC-CoVID	Q: 50 P: 69,318	nDCG	Voorhees et al. (2021)

Table 1: Dataset statistics and evaluation metrics for different tasks in SciRepEval benchmark.

relations for only a subset of nodes—detailed information is available in Appendix B. Table 3 summarizes the performance in different evaluation tasks in SciRepEval.

Table 3 shows that both removing less important citations ($\mathbb{F}$) and adding finding-based relations ($\mathbb{T}$) improved performance compared to the raw citation network, with the combined approach ($\mathbb{FT}$) achieving the best results. Significantly, adding the finding similarity relations proved more effective than citation filtering alone, indicating that citation-based embeddings do not fully capture the semantic structure of scientific literature.

	Node_Num	Edge_Num
Citation	6,013,398	17,795,8624
Citation $\mathbb{F}$	6,013,398	17,769,1665
Citation $\mathbb{T}$	6,013,398	38,650,3618
Citation $\mathbb{FT}$	6,013,398	38,624,0425

Table 2: Input network with different process methods

We also evaluated our model against two state-of-the-art scientific embedding models—SciNCL (Ostendorff et al., 2022) and SPECTER2 (Singh et al., 2023)—both trained on multi-domain scientific datasets. While our approach did not achieve state-of-the-art performance in most tasks from Table 3, it demonstrated competitive results in regression and search tasks, where semantic relationships are particularly important.

6 Discussion

Our experimental results demonstrate that integrating finding similarity relations into citation networks improves the quality of scientific article embeddings, particularly in search and regression tasks. This section explores the implications of these findings, addresses the limitations of purely citation-based approaches, and discusses potential avenues for further improvements.

6.1 The Limitations of Citation Networks for Embedding Learning

Citation networks have traditionally been used to model relationships between scientific articles, operating on the assumption that citations indicate semantic similarity. However, this assumption has several fundamental flaws due to the diverse motivations behind citations:

- Papers are often cited to provide background context or build a research narrative, rather than signifying true conceptual similarity.
- Many papers with strong semantic similarities lack citation connections to each other.
- Citations are subject to various biases, including popularity effects, disciplinary silos, and self-citation patterns.

Our results demonstrate that simple citation-based triplet selection produces suboptimal contrastive learning outcomes. The enhanced performance we observed with finding similarity relations indicates that citation-based methods alone inadequately capture content-based relationships, highlighting the necessity for alternative similarity measures in scientific document embeddings.

6.2 Effect of Citation Filtering and Finding Similarity Relations

A key contribution of our work is demonstrating how citations vary in their importance for learning

Task	Metric	SciNCL	SPECTER2	citation	citation $\mathbb{F}$	citation $\mathbb{T}$	citation $\mathbb{FT}$
			Out-of-Train				
Classification							
Biomimicry	Wt. F1	50.22	**53.20**	48.50	48.51	49.13	49.29
DRSM	Wt. F1	65.10	**68.9475**	62.32	62.78	66.23	66.01
SciDocs MAG	F1	81.11	**82.55**	81.16	80.96	82.24	82.22
SciDocs MeSH	F1	89	**89.72**	88.88	89.09	89.56	88.65
Proximity							
Relish	nDCG	90.67	**91.65**	91.22	91.22	91.05	91.18
S2AND	B^3 F1	93.98	92.8	95.6	95.4	95.3	**95.67**
Peer Reviewer Matching	Avg	45.40	**45.44**	43.83	44.58	44.86	44.67
SciDocs Co-View	MAP	**85.28**	84.68	82.15	82.18	83.25	83.71
	nDCG	**92.23**	92.04	90.71	90.79	91.34	91.47
SciDocs Co-Read	MAP	**87.69**	86.29	83.99	84.6	84.69	84.85
	nDCG	**94**	93.36	92.14	92.57	92.51	92.6
SciDocs Cite	MAP	93.55	**94.08**	84.07	83.89	85.93	87.14
	nDCG	97.35	**97.59**	92.91	92.77	93.83	94.42
SciDocs Co-Cite	MAP	**91.66**	90.58	88	88.13	88.23	88.79
	nDCG	**96.44**	95.99	94.75	94.89	94.93	95.21
Regression							
Review Score	Avg	18.87	**21.79**	18.59	19.71	20.42	19.37
Max h-Index	K Tau	11.3	12.83	12.26	13.13	**14.14**	12.63
Tweet Mentions	K Tau	**25.78**	24.56	23.04	22.89	23.75	25.57
Search							
NFCorpus	nDCG	70.85	70.18	69.7	70.24	**71.47**	70.89
TREC CoVID	nDCG	87.67	**90.87**	89.34	89.39	88.03	88.37
Average Exp. SciDocs	-	56	**57.23**	55.4	55.8	56.4	56.4
Overall Average	-	73.4	**73.95**	71.7	71.9	72.5	72.6

Table 3: Performance metrics across different methods and tasks. The columns labeled citation, citation $\mathbb{F}$, citation $\mathbb{T}$, and citation $\mathbb{FT}$ show our experimental results. The SciNCL and SPECTER2 columns present experimental results from (Ostendorff et al., 2022) and (Singh et al., 2023).

high-quality embeddings. By filtering out less important citations, we reduced noise and achieved modest improvements. However, our most significant gains came from incorporating finding similarity relations, which create direct links between papers based on their research findings rather than citations alone.

6.3 How Does Our Method Compare to Existing Models?

While our approach outperforms the baseline citation network, it does not surpass state-of-the-art models like SPECTER2 in most tasks. This is expected, as SPECTER2 and similar models are trained on larger, more diverse datasets and benefit from task-specific fine-tuning. However, our findings suggest that incorporating additional semantic relations—like findings, methodologies, or co-authorship networks—could help close this performance gap.

Notably, our method achieved competitive performance in regression and search tasks, demonstrating that text-based semantic relations complement citation-based embeddings effectively. This strengthens our argument that citation networks alone cannot fully capture the contextual and conceptual relationships between scientific articles.

6.4 Limitations

Despite its benefits, our approach has some limitations. First, due to computational constraints, we applied citation filtering and finding similarity generation to only a subset of the dataset. A more comprehensive application across a larger scientific corpus may yield even stronger improvements.

Additionally, we limited our exploration of text similarity relations to research findings, excluding other important aspects like methodology. While we believe findings are the most crucial part of scientific papers, examining other aspects could yield valuable insights.

Furthermore, our approach of generating a single universal embedding per article may result in the loss of important information.

These limitations point to clear opportunities for future improvements.

7 Conclusion

In this study, we introduced an enhanced approach to biomedical article embedding by integrating Finding-Citation Graphs (FCG) with contrastive learning. Our method overcomes the limitations of traditional citation-based embeddings by filtering out less important citations and incorporating text-

based semantic relationships into triplet selection. This refined network improves the representation quality of scientific documents, particularly in the biomedical domain.

Our experiments show that removing noisy citations and leveraging finding similarity relations enhance contrastive learning performances. Though our approach did not exceed state-of-the-art methods like SciNCL and SPECTER2, it consistently performed better than the original citation network, demonstrating the value of context-aware triplet formation.

In conclusion, our work establishes a foundation for enhancing scientific document representations through a balanced approach that combines citation analysis with semantic similarity. By improving the construction of scientific embeddings, we deliver more accurate, domain-specific, and semantically meaningful representations—enabling better information retrieval and knowledge discovery in biomedical research.

Ethics Statement

This research focuses on improving scientific article embeddings through Finding-Citation Graphs (FCG) and contrastive learning. Our approach enhances document representations for biomedical scientific articles to improve downstream tasks like retrieval, classification, and citation recommendation. We conducted our research using only open-source datasets.

Acknowledgments

This work was supported by the UKRI Biotechnology and Biology Sciences Research Council [BB/X511833/1], Digital Environment and Research Institute (DERI), the Queen Mary University of London, and Recursion Pharmaceuticals Inc.

We thank Arkaitz Zubiaga, Dan Crowther, and Anniek Myatt for their valuable feedback and suggestions on the project. We are grateful to the Semantic Scholar team for assisting with data access. Additionally, we thank Apocrita (King et al., 2017) and its ITS team for providing and maintaining the HPC resources.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Gully Burns. 2022. Drsm-corpus v1.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Eoghan Cunningham and Derek Greene. 2023. Graph embedding for mapping interdisciplinary research networks. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 784–789, New York, NY, USA. Association for Computing Machinery.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Naman Jain and Mayank Singh. 2021. Tweetpap: A dataset to study the social media discourse of scientific papers. *Preprint*, arXiv:2106.07213.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Thomas King, Simon Butcher, and Lukasz Zalewski. 2017. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. Pytorch-biggraph: A large-scale graph embedding system. *Preprint*, arXiv:1903.12287.

Yuan Liang, Massimo Poesio, and Roonak Rezvani. 2024. A fine-grained citation graph for biomedical academic papers: the finding-citation graph. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 416–426, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 500–509, New York, NY, USA. Association for Computing Machinery.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. 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*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Vikram Shyam, Lauren Friend, Brian Whiteaker, Nicholas Bense, Jonathan Dowdall, Bishoy Boktor, Manju Johny, Isaias Reyes, Angeera Naser, Nikhitha Sakhamuri, Victoria Kravets, Alexandra Calvin, Kaylee Gabus, Delonte Goodman, Herbert Schilling, Calvin Robinson, Robert Omar Reid II, and Colleen Unsworth. 2019. Petal (periodic table of life) and physiomimetics. *Designs*, 3(3).

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Shivashankar Subramanian, Daniel King, Doug Downey, and Sergey Feldman. 2024. S2and: A benchmark and evaluation system for author name disambiguation. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries*, JCDL '21, page 170–179. IEEE Press.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Tongzhou Wang and Phillip Isola. 2022. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *Preprint*, arXiv:2005.10242.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023. Pre-training multi-task contrastive learning models for scientific literature understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12259–12275, Singapore. Association for Computational Linguistics.

Yue Zhao, Ajay Anand, and Gaurav Sharma. 2022. Reviewer recommendations using document vector embeddings and a publisher database: Implementation and evaluation. *IEEE Access*, 10:21798–21811.

A LLM Prompt

The prompt to analyze the importance of each citation can be seen here.

You are an AI language model tasked with analyzing the importance of specific citations within a research paper. Each citation is provided with three pieces of information:
- Citation Sentence: The sentence shows why and what the citation occurs.
- Abstract of the Citing Paper: A summary of the research of the citing paper.
- Title of the Cited Paper: The title of the cited paper.
Based on this information, your task is to analyze and determine the importance of the citation to the citing paper.
Your thinking logic chain should follow the following diagram:
- Abstract Analysis: Identify key goals, methods, and findings.
- Citation Sentence Analysis: Determine citation context and purpose.
- Title Analysis: Check for alignment of scope and key themes.
- Cross-Referencing: Is the cited work foundational to methods, key concepts, or outcomes? Does it appear crucial for the execution of the citing study?
- Explanation: Provide a concise explanation for the classification based on analysis.
- Importance Classification:
- Highly Important: Core foundation (methods, key framework).
- Moderately Important: Background, context, secondary relevance.
- Less Important: General information, historical context.

Here are some examples:
{Examples}

Just output the importance classification result and explanation.

B Preprocessing Citation Network

For the citation filtering, we examined approximately 1.46 million citations, classifying 28.4% as highly important, 44.8% as moderately important, and 26.8% as less important. Since papers can cite others multiple times using different citation sen-

tences, the same citation pair sometimes receives different importance classifications. In such cases, we retained citations marked as less important if they also appeared in the highly important category. Ultimately, we removed only about 260,000 citations from the total of $\tilde{1}7$ million citations, as we only have those citation analysis results.

For the finding similarity relation, we searched for similar findings for 392,505 (Total 10 million) findings. When the two papers shared similar findings, we created a new relation between them. Through this process, we generated approximately 200 million relations between papers.