Tracing the Flow of Knowledge From Science to Technology Using Deep Learning

Michael E. Rose Erik Buunk Sebastian Erhardt
Cheng Li Mainak Ghosh Dietmar Harhoff

Max Planck Institute for Innovation and Competition
Munich, Germany

Email: {michael.rose, erik.buunk, sebastian.erhardt,
cheng.li, mainak.ghosh, dietmar.harhoff }@ip.mpg.de

Abstract

We identify a similarity model suitable for working with patents and scientific publications at the same time. As transformer-based language and similarity models are trained on a single corpus (such as patents), they often perform poorly on other corpora (such as publications). Especially in the patent-publication nexus, researchers are interested in linking semantically similar documents. In a horse race-style evaluation involving eight language (similarity) models, we find that the SPECTER model fine-tuned on patents–the Pat-SPECTER–performs best. In three real-world scenarios (separating patent-paper-pairs, suggesting patent-paper-pairs, prior art search) we demonstrate the capabilities of the Pat-SPECTER.

1 Introduction

Tracing the flow of knowledge from science to technology is pivotal for understanding and fostering innovation processes. The dominant approach for tracing knowledge flows is the analysis of explicit citations in scientific publications and patents. Such citations acknowledge intellectual priority and delineate the foundations upon which new inventions are built. This method, however, has significant limitations, such as the sparsity of citation graphs and the strategic or opportunistic selection of citations.

To overcome this problem, we present a transformer-based similarity model that can be used for patent-publication comparisons. We identify four suitable transformer models and test them rigorously in a horse race. These models have SPECTER, developed by Devlin et al. (2019) for work with publications, and PaECTER, developed by Ghosh et al. (2024) for work with patents, as well as derivations from these two models. We benchmark the models using standard test metrics on small datasets derived from Patent Paper citations by Marx and Fuegi (2022) and Marx (2023).

The winning model, Pat-SPECTER, is a derivative of the SPECTER model fine-tuned on patents, more specifically on the training dataset for the PaECTER. To show its strength, we apply the model to two real-world scenarios, a prior art search akin to patent examination and the prediction of patent paper pairs, where the goal is to find relevant publications among thousands or even millions candidates. We utilize an ElasticSearch database in the Logic Mill search system (Erhardt et al., 2024) to identify (approximate) nearest neighbors among all of PATSTAT and OpenAlex.

In the patent domain, citations to scientific literature are used to link patented inventions to their scientific underpinnings. As many patents do not have direct references to the non-patent literature, Ahmadpoor and Jones (2017) propose the notion of "distance to the science frontier". Patents citing scientific publications directly are taken to be at the science frontier.

Poege et al. (2019) show that highly cited scientific publications often serve as the foundation for particularly impactful patented inventions, highlighting the critical role of science in driving technological innovation. They linked around 950,000 patent families to over 2.2 million scientific articles, demonstrating that foundational scientific research significantly contributes to breakthrough inventions. Related studies such as Marx and Fuegi (2020) also use large-scale citation graphs to derive conclusions for innovation policy.

However, the use of explicit citations has serious disadvantages, even if indirect links are taken into account. Citations are rare, and they citations may be selected strategically or opportunistically (Jaffe & Rassenfosse, 2017).

As an alternative to citations, scholars have investigated the possibility of using text as input data and comparing two documents to gauge their similarity. When vectors represent individual documents, vector operations (such as Cosine distance, Manhattan distance, Euclidean distance or others) yield similarity estimates.

In his dissertation, Natterer (2014) uses a term frequency–inverse document frequency (TF-IDF) model of textual similarity for technical texts. A more recent example in this growing literature is Kelly et al. (2021), who identify breakthrough patents as patents with low textual similarity to the existing patent text corpus.

They then establish that sectors with many breakthrough patents experience higher growth. This relationship could not have been demonstrated convincingly using patent citations, as the authors point out.

Yet, turning massive amounts of text into data is challenging, and so far there exists no scalable solution that at the same time spans diverse text corpora. All applications intended to find similarities between patents or scientific publications suffer from one or several of the following shortcomings: they are specific to only one text corpus (i.e., only patents, such as patent maps); they do not account for semantic structure across different text corpora; they do not scale well.

A common yet simple approach to transform text into a vector is the bag-of-word approach. Typically researchers used a weighted incarnation of this approach, the so-called term frequency–inverse document frequency (TF-IDF).

TF-IDF vectorization, simple as it is, has two important limitations. First, it ignores the relative positioning of terms (to each other and within the document) and scales badly. For example, including new documents in the corpus may require the re-computation of the entire matrix. Thus, it becomes computationally expensive with the growing number of documents. Secondly, the TF-IDF matrix is sparse and high-dimensional, which leads to higher memory consumption. The loss of information of the location of a term within a sentence, within a paragraph and within a document is presumably the most severe limitation.

Recent advances in Natural Language Processing, especially the Word2vec (Mikolov et al., 2013), and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), have made it possible to non-computer scientists to work with textual data while retaining full syntactial information.

The BERT model leverages the attention layer (Vaswani et al., 2017), which efficiently estimates which tokens of a sentence to what extent are valuable in understanding the sentence. In this family of models, words in a sentence do not have equal weight, as is the case for TF-IDF. BERT is pre-trained on newspaper articles and Wikipedia and focuses on the 40k most important tokens.[1]

While to-date multiple thousand specialized transformers exist, two notable models are SciBERT (Beltagy, Lo, & Cohan, 2019) and BERT for patents (Srebrovic & Yonamine, 2020). SciBERT is a BERT model specific to scientific publications. It has learned the 40k most important tokens used in scientific publications. In the patent domain, Google's BERT for patents was trained on 3 million granted USPTO patents.

The computational requirements to train a BERT from scratch are immense, however. Beltagy, Lo, and Cohan (2019) state on their SciBERT paper that 16 TPUv3 (Tensor Processing Units of the third generation) chips were used for 4 days to train BERT's largest model - 8 GPUs (Graphics Processor Unit) are expected to take 40-70 days for the same task. Training a base model can become very expensive and time-consuming. Thus, in our approach, we fine-tune existing models, as this is more efficient and allows us to leverage the knowledge already incorporated in the more fundamental models.

Since BERT's token vocabulary is likely not representative of every domain, multiple domain-specific BERT derivatives have been trained. For example, SciBERT, whose 30k most important tokens overlap with the original general purpose BERT's vocabulary at the rate of 42%.

[1] A token usually represents a word or a part of a word.

In the patent domain, there are three models trained on patent text, namely PatentBERT by Lee and Hsiang (2020), Google's BERT for patents (Srebrovic & Yonamine, 2020) and the SEARCHFORMER (not publicly available) by Vowinckel and Hähnke (2023).

Because BERT models were trained to predict masked tokens and the next sentence, they do not perform effectively at identifying similar documents. This is a necessary prerequisite to identify potential knowledge flows between documents. The SPECTER model (Cohan et al., 2020) and its successor SPECTER2 (Singh et al., 2023) address these limitations through citation-informed learning. The difference between SPECTER and SPECTER2 are simply the underlying training data set: At the time of training of SPECTER, Semantic Scholar consisted only of biomedical publications and those from computer science. SPECTER2 uses a more recent version of Semantic Scholar, encompassing all scientific disciplines. More specifically, the authors fine-tuned SciBERT via contrastive learning between two publications, an actually cited publication and an uncited random publication.

However, SPECTER was only trained on scientific publications and the language specific to patents likely differs from that relevant to scientific publications. Recently, Ghosh et al. (2024) develop the PaECTER model. It was trained in a similar way as SPECTER/SPECTER2 as it leverages credible citation information between patents. Credible citation information refers to citations added by examiners at the European Patent Office (EPO).

2 Data

For training and the evaluations, we use the *Reliance on Science* dataset provided by Marx and Fuegi (2020, 2022). In particular, we look at Patent-Paper Pairs (PPP) and Patent-Paper Citations (PPC). Both of these data sets are rigorously curated, lending a high degree of reliability and applicability. All datasets contain the patent publication number and the OpenAlex IDs of the publications.

The PPP data set contains around 548k pairs with 310k unique USPTO patents and 336k unique publications. A patent-paper pair (PPP) is a unique combination where a scientific publication (the paper) is the foundation for the corresponding patented invention which is described in the patent document (the patent).

The PPC data set is a patent data set with all the cited scientific publications for each patent. The set has around 47 million records, with 7 million unique patents. It includes many authorities such as USPTO (approx. 34 million), EPO(5.7 million), WIPO (4.1 million), and many others. In our case, we use USPTO or EPO, depending on the use case. The PPC dataset provides confidence scores based on how certain their algorithm and human judgment are that the reference and the scientific publication belong together. To be on the conservative side, we only selected observations from the PPC involving the highest confidence scores.

In some cases, the PPC and PPP overlap. Yet in most of the cases, there are no direct citations between the patent and the paper or vice versa. In the most obvious case, this would not be possible. If a patent cites a publication on the same invention, it would destroy the novelty. However, there are many publications with an earlier publication date than the patent. The reason is the one-year grace period. In these cases, the publication already exists but is not cited by the patent.

3 Fine-tuning SPECTER2 and PaECTER

Hitherto, in innovation studies, there exist only language models specific to one corpus: either publications or patents. Thus we develop two cross-corpus models, Pat-SPECTER and Pub-PaECTER.

Pat-SPECTER and Pub-PaECTER are derivatives of SPECTER and PaECTER respectively. Pat-SPECTER is the SPECTER2 model fine-tuned on the training data set for the PaECTER, and the Pub-PaECTER is the original PaECTER fined-tuned on the training data set of the SPECTER.

The training dataset of SPECTER and SPECTER2 are constructed in the same way: For each focal document, they use five triples, where the second element is a cited document (called "positive") and the third element is a non-cited document (called "negative"). Negatives are furthermore separated into easy negatives and hard negatives. Easy negatives are randomly selected uncited documents, while hard negatives are indirectly cited documents (publications cited by cited publications but not by the focal document).

The training dataset of PaECTER (Ghosh et al., 2024) is akin to the SPECTER/SPECTER2 approach: For each of the 300k patent documents, there are also five triplets with positives and negatives. Positives are patents cited with citation category X, Y, I and A. Easy negatives are patents sharing at least some CPC classes, while hard negatives are indirectly cited patents. Crucial for the analysis, only EPO patents are considered, because at the EPO references to other patents are solely added by examiners, and with specific citation categories.

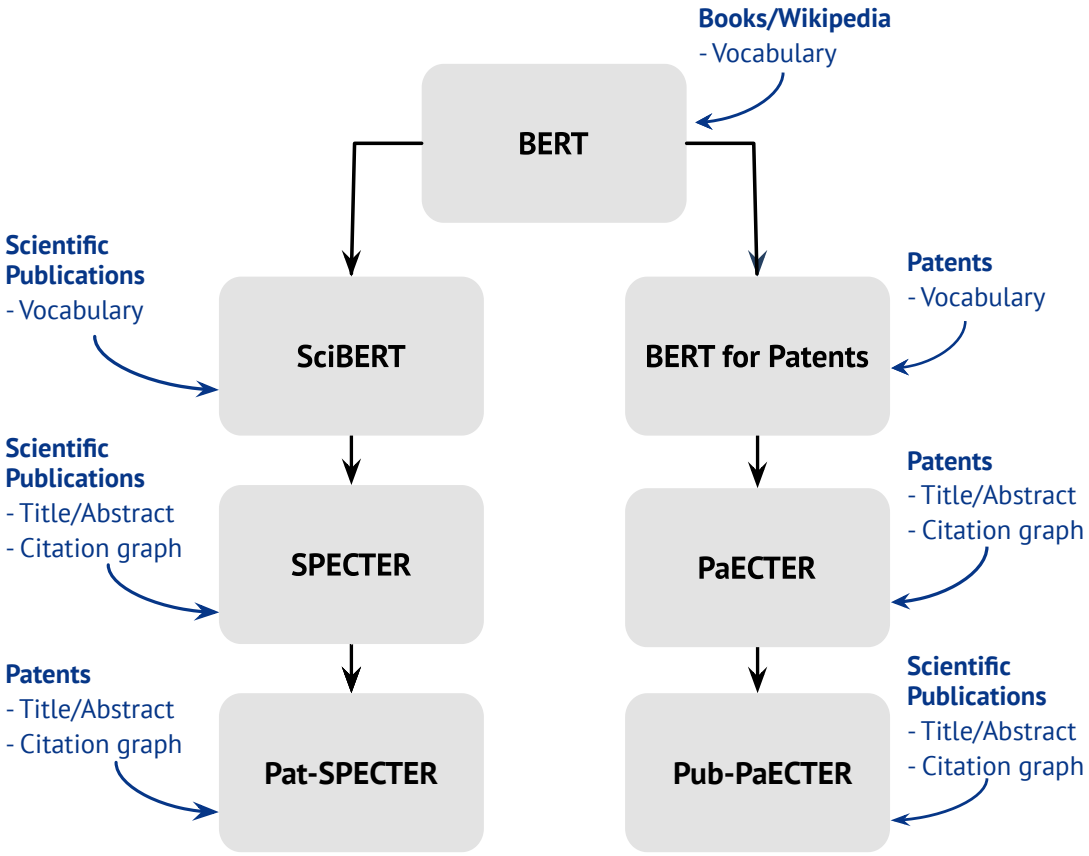
Equipped with these models and datasets, we fine-tune them in a cross-corpus manner: First we fine-tune SPECTER2 (base) on the training dataset of PaECTER. The resulting model is called *Pat-SPECTER*. Second we fine-tune PaECTER on the training dataset of SPECTER2. The resulting model is called *Pub-PaECTER*.

4 Finding a cross-corpus language model

4.1 Experimental setup

Our goal is to assess which similarity models trace knowledge flows from science to technology best. We compare three the following models against each other: BERT, SciBERT, BERT for Patents, SPECTER, SPECTER2, PaECTER, Pat-SPECTER, Pub-PaECTER. Figure 1 illustrate how the models relate to each other.

Figure 1: Overview of the models used in the comparisons



Notes: BERT (Devlin et al., 2019) is a general languagel model for the English language trained on a variety of public corpora. SciBERT (Beltagy, Lo, & Cohan, 2019) is a BERT with a vocabulary specific to scientific papers (biomedical and computer science). BERT for Patents (Srebrovic & Yonamine, 2020) is a BERT with a vocabulary specific to patents. SPECTER (Cohan et al., 2020) augments the SciBERT model using paper citations. PaECTER (Ghosh et al., 2024) augments the BERT for Patents model using patent citations. Pat-SPECTER and Pub-PaECTER are developments reported in this paper.

For our comparisons we need textual data, notably title and abstracts. For publications, we use the OpenAlex (Priem, Piwowar, & Orr, 2022) database. OpenAlex is a comprehensive, community-curated database of scholarly works, authors, institutions, and more. For patents, we use PATSTAT 2023 Spring version. PATSTAT (Patent Statistical Database) is a comprehensive database produced by the European Patent Office (EPO).

Our dataset consists of 1,000 randomly selected triplets: Every record contains a focal EPO patent (A1, A2 or B1), five cited publications (positives), and 25 publications (negatives). We sampled the negatives randomly from OpenAlex, provided their abstract is in English.

To compare the models, we use ranking metrics. For each test:

- Embeddings are generated for all 31 documents in a triplet.
- The cosine distance metric is calculated between the focal patent and the other

30 documents using the embeddings.

- The documents are then ranked from most similar to least similar based on the cosine distance.
- This process is repeated for each of the 1,000 triplets.

Ideally, the positive documents should be ranked in the top five positions. The metrics used (RFR, MAP, and MRR@10) provide different perspectives on the ranking results.

Table 1: Rank-Aware Evaluation of Different Models on the Cross-Corpus Dataset

	Avg. RFR		MAP		MRR@10	
Model	CLS	Mean	CLS	Mean	CLS	Mean
BERT	2.52	1.29	46.76	79.05	69.86	91.52
SciBERT	2.48	1.37	49.23	71.14	71.75	90.55
BERT for Patents	1.20	1.10	78.51	85.99	93.14	96.97
SPECTER	1.08	1.13	**91.64**	86.57	97.72	96.23
SPECTER2	1.11	1.35	88.29	76.51	95.94	91.62
Pat-SPECTER	**1.05**	1.12	91.38	87.24	**98.04**	96.06
PaECTER	1.13	1.07	86.45	89.72	96.12	97.55
Pub-PaECTER	1.32	1.25	76.49	79.57	92.51	94.15

Notes: "Avg. RFR" is the average rank of the first relevant (i.e., actually cited) publication. "MAP" is the mean-average precision, which takes into account the precision and the recall at every position where a relevant item appears. "MRR@10" is the mean reciprocal rank of the first relevant publication within the 10 closest publications. "CLS" and "Mean" refer to different ways we compute embeddings for a text of multiple sentences: "CLS" concatenates all sentences into one sentence while "Mean" takes the average of the sentences' embeddings.

The results in Table 1 reveal that Pat-SPECTER performs well in terms of *avg. RFR* and *MRR@10* compared to other models. However, it comes out only second to SPECTER in terms of *MAP* metric, though the difference is not statistically significant (see Table 3 in Appendix).

Since *avg. RFR* and *MRR@10* evaluate the rank of the most similar publication for a given patent, Pat-SPECTER's well performance in these metrics suggests that it excels at early detection of the most relevant publications. In general, its slightly lower performance in *MAP*, which considers the ranks of all relevant publications, indicates that it may need to retrieve more publications to identify all relevant ones.

4.2 Conclusion

The evaluation suggests that Pat-SPECTER is the most suitable model for cross-corpus comparison involving scientific publications and patents. The runner-up model is SPECTER, but Pat-SPECTER statistically dominates SPECTER. The other evaluations suggest it to be a suitable model, though not the best.

5 Applications

To make the model more palpable, we test it on three applications. Working with such large datasets is a more realistic scenario than working with small document samples.

The goal of the first application is to separate "Patent Paper Pairs" (PPPs) from other pairs of publications and patents.

The second evaluation makes use of the PPPs as well, but it aims to predict the paper in the pair. We aims to identify the related paper among a the 1000 most similar publications for a given patent.

The final evaluation is analogous to the prior art search by patent offices. Prior art search is a crucial task in the patent examination process, as it aims to find existing patents, publications, and other relevant articles that might challenge the novelty of the patent application under prosecution.

5.1 Separating Patent Paper Pairs from other Patent Paper Citations

We use our trained transformer models to create embeddings and then identify textually similar candidates for a PPP. We aim to determine a similarity cut-off value or separation model, that we can use as a threshold to separate pairings with high similarity from actual PPPs.

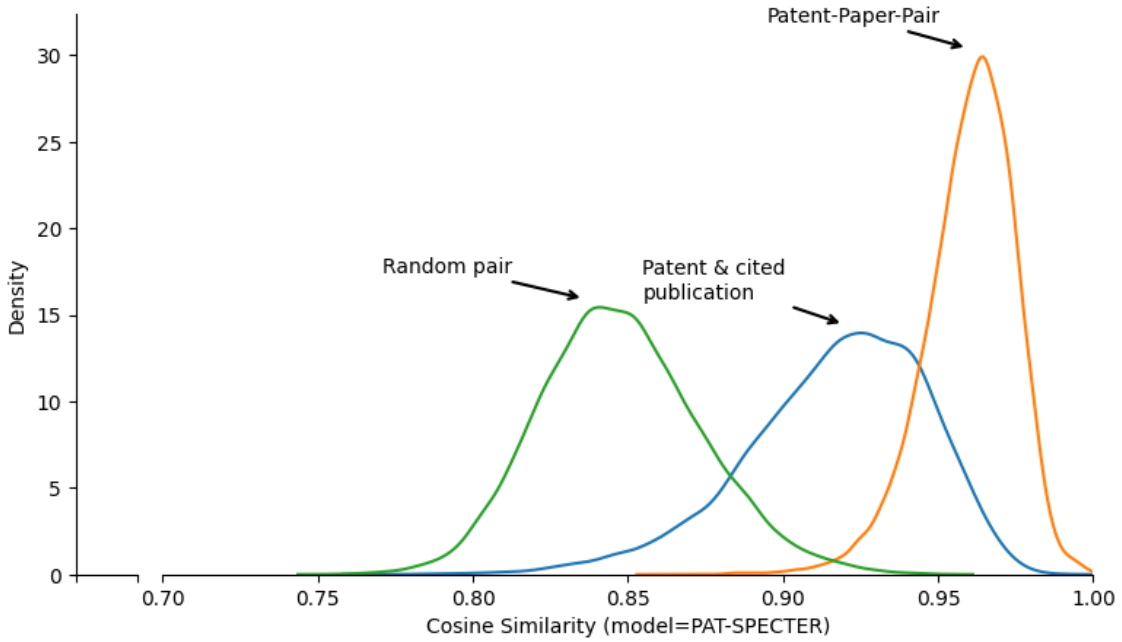
In our dataset, we have three different classifications. The PPPs, the Non-patent literature (NPL), and patents paired with random publications. We try to identify the PPPs in the total set.

We employ the cosine similarity metric in combination with various separation methods to predict PPPs. This approach is essential as it reduces the decision variable into a single scalar metric. As proposed here, our method does not utilize supplementary information such as dates, confidence scores, author names, CPC classes, or concepts, which could potentially enhance the predictive power. The results should, therefore, be taken as a very conservative approach. This approach differs from previous methods, such as those used by Marx and Scharfmann (2024), which incorporate additional metadata like author names, similarities, and institutional distance into their prediction model. Our decision is based on efficiency and scalability. We aim to achieve high-quality outcomes with minimal input data.

In practical applications, this additional metadata may serve as pre- or post-filtering criteria. For instance, author names could be used to pre-filter documents, refining the search space for identifying PPPs. Within this refined subset, a nearest-neighbor search would be conducted. Documents exceeding the similarity threshold would then be treated as PPPs. We choose not to apply such filtering in order to demonstrate the unconstrained performance of our algorithms.

For our analysis, we utilize a random selection of 20,000 Patents from the PPP dataset. We then include all pairs with an English abstract in the OpenAlex database. We add a Non-Patent Literature (NPL) set, which are patents from the United States Patent and Trademark Office (USPTO). This approach ensures a consistent citation strategy, as citation rules may vary across different jurisdictions. For the random dataset, we use the PPP and NPL patent selection and add the random publications. This leads to a sample that is almost twice as large.

Figure 2: KDE of cosine distribution PAT-SPECTER



The next step is the cleaning of the Open Alex abstracts. Some abstracts contain structure elements (such as headings) or contain copyright statements. These elements are typical for certain journals. Since the abstracts appear to be more similar, it may introduce an unwanted bias. We remove these elements with our abstract cleaning tool.

Given the embeddings obtained with the Pat-SPECTER, we calculate the cosine similarity between for each patent paper pair. We use an 80/20 split for training and testing for optimizing the separation method.

Additionally, we need to adjust the fraction of PPPs compared to the other pair types. Patent-paper-pairs are not so common in the real world; hence, we adjust the fraction of PPPs. If the fraction is too high, the results will be unrealistic compared to the real world. If the fraction is too low, the modeling results will be inaccurate since we have too few positives. In the latter case, whether a single item is found will make a big difference in the results. In our setup, we have chosen for 10% PPPs and 90% random and NPL. An iterative approach is used to make sure the resulting data set sizes are still around the 80/20 split. We ensure that the training and test sets do not have any overlapping papers or patents. We end up with a split of 24,035 pairs for training and 5,661 for testing.

The results in Figure 2 demonstrate a clear separation between the PPP, the NPL pairs, and the random pairs. However, the distributions overlap partially.

Thus we strive to compute an optimal cut-off value of cosine similarity to separate PPPs from other pairs. One of the simplest method is what we call 'Cut-off by F1'. The F1 score balances the rate of false positives to false negatives.[2] We aim to determine the cosine similarity value where the F1 score is maximized, indicating the best balance between precision and specificity for identifying PPPs.[3]

[2]The F1 score is the harmonic mean of precision and recall. Precision is the share of true positives over all positives. Recall is the share of all positives that was detected.

[3]We have experimented with different methods. Variations of logistic regression models led to

We find the maximum F1 score at 0.76 for a cosine similarity of 0.949, which is thus the optimal threshold value. Document pairs with a cosine similarity equal or higher than this value should be categorized as PPP.

5.2 Searching Patent Paper Pairs in the patent-publication universe

A similar task is to identify papers that are very similar or identical to the patent, using the patent-paper-pair (PPP) dataset from the Reliance on Science project as a ground truth (Marx & Scharfmann, 2024). The goal is to evaluate the performance of our model to predict PPPs.

Based on the model's predictions, each pair is assigned to one of four categories reflecting the confidence score: very high for category 4 (prediction greater than 0.99), high for category 3 (prediction between 0.90 and 0.99), medium for category 2 (prediction between 0.80 and 0.90), and low for category 1 (prediction between 0.70 and 0.80). For each unique US patent in the PPP dataset, we retrieve its embedding from our DocDB index and then search for the corresponding 1,000 approximated nearest neighbors in our OpenAlex index. Furthermore, we calculate the rank of the paired papers for each focal patent in the PPP dataset when they appear within the top 1000 results.

In the PPP dataset, we were able to match 342,252 pairs ($\approx$ 62%) within the first 1,000 nearest neighbors, where the missing instances are mainly due to the imperfect datasets as discussed in section 5.3. For the matched PPPs, about 65% have a rank between 1 and 100, and nearly 90% have a rank between 1 and 500. The cumulative distribution of the rank is skewed towards lower values and gradually levels off as the rank increases (Figure 3).

Further, we investigated the distribution of ranks from the matched PPPs across different categories, highlighting similarities and differences in the distribution patterns within each category by violin and box plots in Figure 4. Across all categories, the PPP ranks are predominantly low, as shown by the high density of data points at lower ranks. Categories 1 and 2 show more variability in ranks, with Category 1 having the most dispersed distribution. In contrast, Categories 3 and 4 have a higher concentration of lower-ranked PPPs and less variability, with Category 4 being the most concentrated. Hence, our predictions are more confident in categories 3 and 4 than in categories 1 and 2. This is consistent with the definition of the confidence level in Marx and Scharfmann (2024).

the same results when the hyper-parameters were optimized and with or without normalized cosine values. The order of magnitude for all separation method remain the same, depending on the data sample, with F1 scores roughly between 0.73 and 0.76)

Figure 3: Empirical Cumulative Distribution of the rank for the matched PPPs

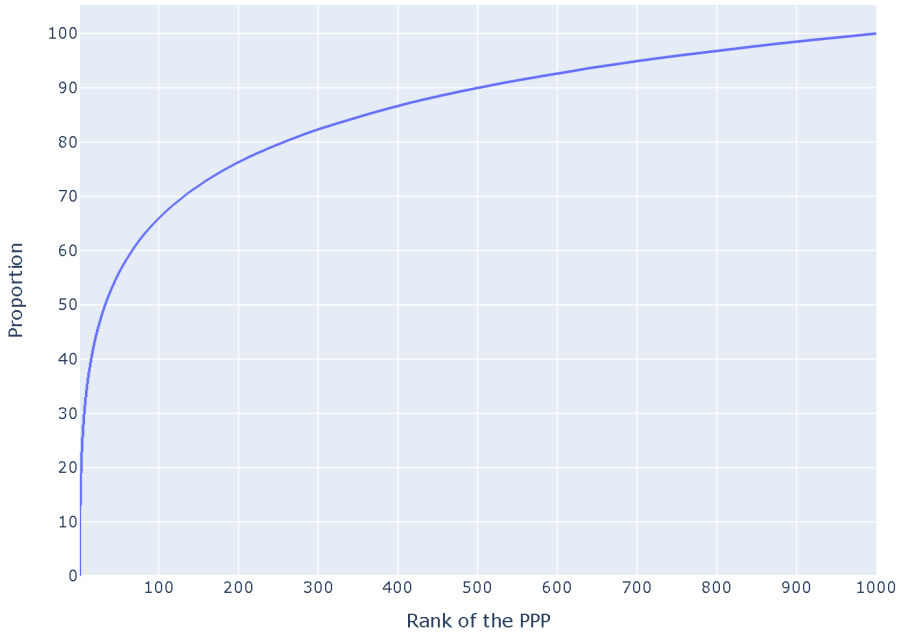
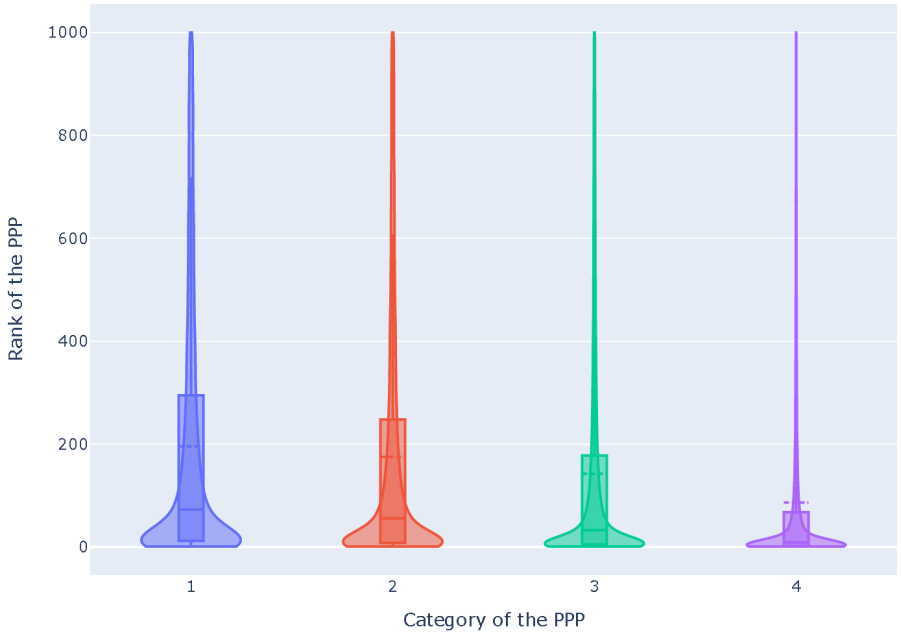


Figure 4: Distribution of the rank for the matched PPPs in different categories



Notes: Categories relate to algorithmic confidence of the Reliance of Science dataset, with category 4 being the highest confidence.

5.3 Prior Art Search

When performing a prior art search, patent examiners seek to find documents that help assess whether the patent application is novel or not. These are usually patents, but also scientific publications may be relevant. There we conduct two exercises, a patent to patent prior art search, and a patent to publication prior art search.

To perform these two tasks, we rely on the Logic Mill system, developed by the authors (Erhardt et al., 2024). Logic Mill is a scalable and openly accessible software system with the goal to identify semantically similar documents within either one domain-specific corpus or multi-domain corpora. Logic Mill provides embeddings for scientific publications (all of OpenAlex) and patent documents (all patent documents from the three jurisdictions EPO, USPTO and WIPO) as computed by Pat-SPECTER, provided they have an English abstract. As it uses ElasticSearch, we can leverage the approximate nearest-neighbor search algorithm to find the most similar documents with a very high probability for any query document within milliseconds.

To establish a ground truth for these citations, we leverage the Reliance on Science dataset, which links USPTO patent IDs to their corresponding citations within the OpenAlex database.

Each patent-paper citation in this dataset is assigned a confidence score ranging from 1 to 10. For our analysis, we exclusively consider citations with the highest confidence score of 10.

Of initially 10,000 randomly selected samples, we found at least one cited publication among the top 100 results in 3,740 cases ($\approx 37\%$). As seen in Figure 5, the distribution is not as steep as in the patent-to-patent case. Table 2 presents the MRR and MAP for various $k \in \{5, 10, 20, 50, 100\}$.

The results presented in Table 2 indicate that, in $3,740/10,000 \approx 37\%$ of the cases, on average, users find a relevant item at position $1/0.16 \approx 6$ given a universe of 100+ million publication documents.

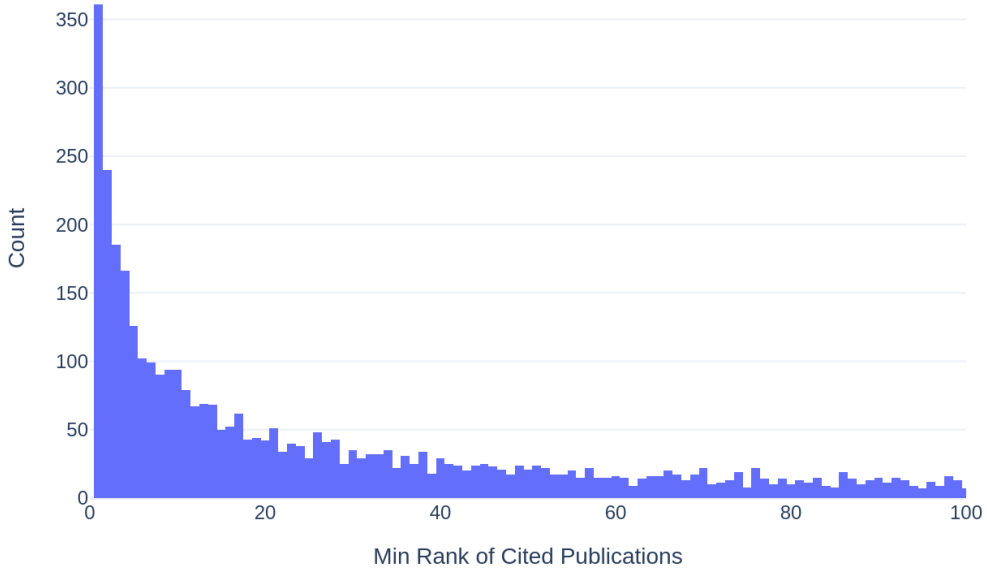
To understand the omission of a significant share of citations, we manually examined anecdotal cases where matches were not found within our approximate nearest neighbor (ANN) results. Our analysis reveals that many of these citations were applicant-added rather than examiner-cited. These applicant-added citations often appear in the patent description, a context not considered during model training

Table 2: MRR and MAP for patent-to-publication citations

k	MRR	MAP
5	16.2932	5.2162
10	17.9573	6.6367
20	19.0259	8.0288
50	19.7886	9.3623
100	20.0598	10.0609

Notes: Table shows the Mean Reciprocal Rank (MRR) and Mean Average Rank (MAP) at different rank cutoffs k.

Figure 5: Distribution of the first ranked publication cited by the reference patent (N=3,740 of 10,000)



nor encoded in our system.

Additionally, a lack of semantic similarity between the cited publication and the focal patent's abstract makes associating the two challenging even for humans. Another critical aspect is the incompleteness of the underlying OpenAlex database.

6 Conclusion

Fine-tuned from the SPECTER2 model, Pat-SPECTER demonstrated the best performance in predicting citations from patents to scientific publications.

In three real-world scenarios, the model was subjected to prior art tests to evaluate its practical utility in identifying relevant prior art documents within a large dataset. This involved using Pat-SPECTER to search extensive databases for patents and scientific publications that closely matched specific patent documents. The model's performance was assessed based on its ability to rank the most relevant prior art at the top of the search results. The tests demonstrated that Pat-SPECTER could effectively identify pertinent prior art.

The models are capable of picking up differences between the text corpora. The results show that the different models lead to similar results, and the separation methods do not outperform each other.

The F1 cutoff value can be directly used as a guide as to which documents are to be selected. In this case, the (Approximate) Nearest Neighbor search operates on one similarity metric, thus allowing for considerable speed, even while searching through millions of documents. After the initial selection, one could build a more informed machine learning model that uses more metadata of the documents (such as authors, dates, CPC-classes, and concepts). This model can be more precise and only needs to be applied to a very limited set of documents.

The most severe limitation stems from lack and incompleteness of our data sources. OpenAlex misses a substantial amount of abstracts. In the May 2024 version, we found that as many as 45.5% of all works have no indexed abstract.[4] While many of them are not in English, a great deal of them belong to Springer journals which does not share abstracts with OpenAlex.

References

Ahmadpoor, Mohammad and Benjamin F. Jones (2017). "The dual frontier: Patented inventions and prior scientific advance". In: *Science* 357.6351, pp. 583–587. DOI: 10.1126/science.aam9527.

Beltagy, Iz, 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3613–3618. DOI: 10.18653/v1/D19-1371.

Cohan, Arman et al. (2020). "SPECTER: Document-level Representation Learning using Citation-informed Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2270–2282. DOI: 10.18653/v1/2020.acl-main.207.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Erhardt, Sebastian et al. (2024). "Logic Mill – A Knowledge Navigation System". In: *Proceedings of the 5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2024), Washington D.C., USA, July 28th, 2024*. Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech. arXiv:2301.00200 [cs]. Washington D.C., USA: arXiv. DOI: 10.48550/arXiv.2301.00200.

Ghosh, Mainak et al. (2024). *PaECTER: Patent-level Representation Learning using Citation-informed Transformers*. Version Number: 1. DOI: 10.48550/ARXIV.2402.19411.

Jaffe, Adam B. and Gaétan de Rassenfosse (2017). "Patent citation data in social science research: Overview and best practices". In: *Journal of the Association for Information Science and Technology* 68.6, pp. 1360–1374. DOI: 10.1002/asi.23731.

Kelly, Bryan et al. (2021). "Measuring Technological Innovation over the Long Run". In: *American Economic Review: Insights* 3.3, pp. 303–320. DOI: 10.1257/aeri.20190499.

Lee, Jieh-Sheng and Jieh Hsiang (2020). "Patent classification by fine-tuning BERT language model". In: *World Patent Information* 61, p. 101965. DOI: 10.1016/j.wpi.2020.101965.

[4] It should be noted that "Works" included documents published by journals which are not commonly referred to as scientific publications. front matters, back matters, table of contents, advertisements, etc.

Marx, Matt (2023). *From Patent-Paper Citations to Patent-Paper Pairs*. Cambridge, MA, USA.

Marx, Matt and Aaron Fuegi (2020). "Reliance on science: Worldwide front-page patent citations to scientific articles". In: *Strategic Management Journal* 41.9, pp. 1572–1594. DOI: 10.1002/smj.3145.

— (2022). "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations". In: *Journal of Economics & Management Strategy* 31.2, pp. 369–392. DOI: 10.1111/jems.12455.

Marx, Matt and Emma Scharfmann (2024). "Does Patenting Promote the Progress of Science?" In: *Unpublished manuscript*.

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs]. DOI: 10.48550/arXiv.1301.3781.

Natterer, Michael (2014). "Entwicklung, empirische Validierung und ökonomische Anwendung eines texbasierten Ähnlihckeitsmaßes". Inauguraldissertation. Munich: Ludwig Maximilians University.

Poege, Felix et al. (2019). "Science quality and the value of inventions". In: *Science Advances* 5.12, eaay7323. DOI: 10.1126/sciadv.aay7323.

Priem, Jason, Heather Piwowar, and Richard Orr (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. Version Number: 2. DOI: 10.48550/ARXIV.2205.01833.

Singh, Amanpreet et al. (2023). "SciRepEval: A Multi-Format Benchmark for Scientific Document Representations". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 5548–5566. DOI: 10.18653/v1/2023.emnlp-main.338.

Srebrovic, Rob and Jay Yonamine (2020). *Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery*. Tech. rep. Google.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. DOI: 10.5555/3295222.3295349.

Vowinckel, Konrad and Volker D. Hähnke (2023). "SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search". In: *World Patent Information* 73, p. 102192. DOI: 10.1016/j.wpi.2023.102192.

7 Appendix: Statistical Significance Test Pat-Specter

Table 3: Statistical Significance Testing of Rank-Aware Evaluation of Pat-SPECTER versus Different Models on a Cross-Corpus Dataset

	Avg. RFR	MAP	MRR@10
BERT	0.239***	-0.123***	-0.065***
	(0.035)	(0.008)	(0.007)
SciBERT	0.322***	-0.202***	-0.075***
	(0.035)	(0.008)	(0.007)
BERT for Patents	0.050	-0.054***	-0.011
	(0.035)	(0.008)	(0.007)
SPECTER	0.025	0.003	-0.003
	(0.035)	(0.008)	(0.007)
SPECTER2	0.057	-0.031***	-0.021**
	(0.035)	(0.008)	(0.007)
PaECTER	0.015	-0.017*	-0.005
	(0.035)	(0.008)	(0.007)
Pub-PaECTER	0.196***	-0.118***	-0.039***
	(0.035)	(0.008)	(0.007)
Constant	1.051***	0.914***	0.980***
	(0.025)	(0.006)	(0.005)
Adjusted R2	0.020	0.130	0.026
Observations	8,000	8,000	8,000

Notes: Standard errors in parentheses. Base level model is Pat-SPECTER. All models use their best pooling method, as derived from Table 1.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$