EPO ARP grant 2020

Linking patents to scientific publications through in-text reference mining

Final report

December 2023

PI: Jian Wang
Co-PI: Suzan Verberne
Team members: Zahra Abbasiantaeb, Jia Zhang

Leiden University

Project summary

Patent in-text references (i.e., references in the full text of the patent) to the scientific literature provide a valuable paper trail of knowledge flow from science to technological innovation and encode very different kinds of information than patent front-page references that are commonly used. However, patent in-text references are unstructured and difficult to extract. This project aims to (1) develop a high-performing machine learning method to extract patent in-text references and then match them to the Web of Science (WoS) database of scientific publications, (2) implement this method to EPO and USPTO patents to create a large-scale dataset linking patents to publications, and (3) uncover what kinds of science lead to more valuable patents and how in-text references are different from front-page references.

To this end, we developed a three-stage pipeline for extracting and matching patent in-text references. The first stage extracts reference strings in patent texts, the second stage extracts fields (e.g., author name, journal name, title) from reference strings, and the third stage matches references to WoS publications based on extracted fields. To train our model, we randomly sampled 2,000 EPO patents and 4,000 USPTO patents, screened out a subset of patents that are unlikely to have in-text references, and manually annotated 725 EPO patents and 650 USPTO patents by labelling the scientific in-text references. Out of these, 392 EPO patents made 3900 references and 2088 of them can be matched to a WoS publication, and 319 USPTO patents made 3901 references and 2247 of them can be matched to WoS publications. To evaluate the performance of our pipeline, we reserved 20% patents for testing. The first stage reference extraction model achieved a precision of 98.9% and a recall of 97.7%, at the reference level, and the whole pipeline achieved a precision of 96.8% and a recall of 91.9%, at the unique patent-paper-pair level.

Subsequently, we implemented this pipeline to the corpus of EPO and USPTO patent full texts granted between 1990 and 2022. We identified 5,438,836 references from 492,469 EPO patents and matched 2,763,779 (51%) of these references to WoS publications. We identified 20,432,189 references from 1,449,398 of USPTO patents and matched 11,069,995 (54%) of these references to WoS publications.

Using a subset of biotech USPTO patents and in-text reference therein, we studied what kinds of scientific publications lead to more valuable patents, as measured by patent forward citations and the stock market response to the issuing of the patent. We found (1) a positive

effect of the number of referenced scientific papers, (2) an inverted U-shaped effect of basicness, (3) an insignificant effect of interdisciplinarity, (4) a discontinuous and nonlinear effect of novelty, and (5) a positive effect of scientific citations for patent market value but an insignificant effect on patent citations. When comparing patent in-text and front-page references, we found a remarkable low overlap between them; the overlapping references only account for 20% of all in-text references and 27% of all front-page references. In-text references are more basic and have more scientific citations than front-page references. The difference in interdisciplinarity and novelty is small when comparing at the reference level and insignificant when comparing at the patent level. In addition, in-text referenced papers have a higher chance of being listed on the front-page of the same patent when they are moderately basic, less interdisciplinary, less novel, and more highly cited. Accordingly, using front-page reference yields substantially different results than using in-text references to study what kinds of science lead to more valuable patents. These results suggest that findings regarding science-technology-linkages might be sensitive to which type of patent references are analyzed.

Table of Contents

1. Motivation and project objectives

1.1. Motivation

How science feeds into technology is a long-standing question for economists, sociologists, management and policy scholars. One obstacle for this line of inquiry is the lack of a large-scale empirical identification strategy for tracing knowledge flows from science to technology. To tackle this challenge, this project aims to develop a high-performing text mining method to extract and match patent in-text references to the scientific literature. This method was then applied to the corpus of EPO and USPTO patents at a large scale, resulting in a dataset linking patents to publications. This dataset will be an invaluable resource for studies of science and innovation. Building on this dataset, we explored how different types of science contribute to technological innovation differently.

References in patents to science provide a paper trail of the knowledge flow from science to patented inventions. Since the pioneer work of Nunn and Oppenheim (1980), Narin and Noma (1985), and Tamada et al. (2006), they have been widely used for science and innovation studies, science policy, and business intelligence, for example, for studying industrial dependence on public science (Narin et al., 1997), examining concordances and distances between scientific disciplines and technological fields (Ahmadpoor & Jones, 2017; Callaert et al., 2014), uncovering mechanisms through which firm innovation benefits from science (Cassiman et al., 2008; Fleming & Sorenson, 2004), quantifying economic returns of science (Li et al., 2017; Watzinger & Schnitzer, 2019), and identifying characteristics of scientific contributions that are useful for patented inventions (Poege et al., 2019; Veugelers & Wang, 2019).

However, the current practice relies mostly on patent front-page references but neglects the more difficult patent in-text references. Front-page references are the references listed on the front page of the patent document, which are deemed as relevant prior art for assessing patentability by examiners, as well as inventors and patent attorneys. In-text references are references embedded in patent text, serving a very similar role as references in scientific publications. Because of their different generation processes, front-page and in-text references embody different information. Prior studies have documented very low overlap between patent in-text and front-page references (Bryan et al., 2020; Marx & Fuegi, 2022; Verberne et al., 2019). For example, Verberne et al. (2019) reported that the majority (88%)

of the extracted in-text references from 33,338 biotech patents are not listed on the front page. Several recent studies suggest that in-text references are a better indication of knowledge flow than front-page references (Bryan & Ozcan, 2020; Bryan et al., 2020; Nagaoka & Yamauchi, 2015).

While patent front-page references are readily retrievable from the metadata in patent databases, in-text references are part of the unstructured, running text. Therefore, identifying the start and end of a reference is a challenge. Furthermore, patent in-text references are shorter and contain less information than front-page references (e.g., the title of the publication is typically not included), adding to the difficulty of matching in-text references to scientific publications.

For example, USPTO patent "CRISPR-Cas systems and methods for altering expression of gene products," in its first paragraph containing in-text references, cites four publications, in different formats and none of them appears on the front page.

- Goeddel, GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990).
- Boshart et al. Cell, 41:521-530 (1985)
- Mol. Cell. Biol. Vol. 8(1), p. 466-472, 1988
- Proc. Natl. Acad. Sci. USA., Vol. 78(3), p.1527-31, 1981

Therefore, this project aims to solve the challenging problem of extracting patent in-text references, by developing advanced text mining methods.

1.2. Prior studies

There are three general approaches to the problem: The first approach (i.e., Bryan et al., 2020) skips the challenge of reference extraction and accordingly avoids extraction errors. However, a disadvantage is that starting from scientific publications instead of patent references is computationally inefficient, considering that WoS has more than 70 million publications and more than 21 thousand journals, but only a very small share of them are cited by patents. Bryan et al. (2020) only attempted to match 248 journals covering less than 5% WoS publications. The second approach uses regular expressions (i.e., Marx & Fuegi, 2022; Tamada et al., 2006), which has lower computational burden and can achieve a high level of precision, which is however usually at the expense of recall. The third approach (i.e.,

Rassenfosse & Verluise, 2020; Verberne et al., 2019; Voskuil & Verberne, 2021) relies on machine learning methods, and recently developed language models such as BERT-based models (i.e., Voskuil & Verberne, 2021) have shown great potential for achieving high recall and precision, while the used training and testing datasets are relatively small and homogeneous.

To the best of our knowledge, Tamada et al. (2006) conducted the first systematic analysis of patent in-text references. They used regular expressions to extract in-text references from granted Japanese patents. They extracted 9379 non-patent references from a sample of 1500 patents across five technology fields, with recall of 98.2% and precision of 98.1%.

Bryan et al. (2020) skipped the step of reference extraction but started from a set of scientific publications and then searched for coarse matches between the metadata of these publications and patent full texts. They covered 3,389,853 articles published between 1984 and 2016 in 248 prominent academic journals, which are cited collectively 2,786,041 times in 342,667 USPTO patents granted since 1984, with 1,573,143 front-page references and 1,212,898 in-text references. Recall and precision were not reported.

Our own prior work, Verberne et al. (2019) approached the reference extraction problem as a sequence labeling task and used Conditional Random Fields and Flair. We trained models on 22 patents with 1,952 manually labeled references and found that CRF obtained better results on citation extraction than Flair, with a precision of 83% and a recall of 81.3% at the complete reference level. However, Flair extracted many more references from the large collection than CRF, and more of those can be matched to WoS publications. Voskuil and Verberne (2021) further improved the training data and used BERT-based models (e.g., BERT, bioBERT, sciBERT). They achieved higher performance: testing recall and precision were 94.7% and 95.4% respectively for words at the beginning of citations, and 98.6% and 97.6% for words inside citations, while such metrics were not reported at the complete reference level.

Rassenfosse and Verluise (2020) applied the GROBID model to the full texts of USPTO patents but did not report recall or precision statistics.

Marx and Fuegi (2022) combined the rule-based method and the GROBID model for reference extraction and evaluated model performance using a set of 5,939 references.

Depending on the chosen level of confidence score, prevision ranged from 93.53% to 100%, and recall from 82.05% to 57.70%.

1.3. Project objectives

This project aims to advance this line of literature with three main objectives:

Objective 1: Developing a high-performing text mining method for extracting and matching patent in-text references to scientific publications.

Building on our prior work (i.e., Verberne et al., 2019; Voskuil & Verberne, 2021), we aim to develop a method to extract and match patent in-text references with high precision and recall, by building a larger and more diverse training dataset and testing multiple more recently developed language models. This research contributes to computer science, in particular the blooming fields of natural language processing and machine learning. The named entity recognition (NER) methods we build upon were initially designed for extracting short strings (e.g., names) but are not tailored for relatively long strings such as references. Our project explores a new domain of applications for these methods and provides insights into the difficulties and possible solutions. Also, the patent domain is notorious for its long and complex sentences, which complicates automated text analysis (Verberne et al., 2010). This research makes direct contributions to research in the intersection between text mining and patent data. In addition, the training dataset that we build will be a valuable resource for the text mining community for developing new methods.

Objective 2: Building a large-scale dataset linking individual patents and publications that are cited in patent text.

This project applies our final method on the corpus of USPTO patents and EPO patents. After extracting in-text references from these patents with our text mining methods, we automatically match them to the Web of Science (WoS) database of scientific publications, which has high data quality and is widely used for science studies and policy. This large-scale dataset of publication-patent-links will be an invaluable resource for studying science and innovation. It will benefit researchers in various disciplines.

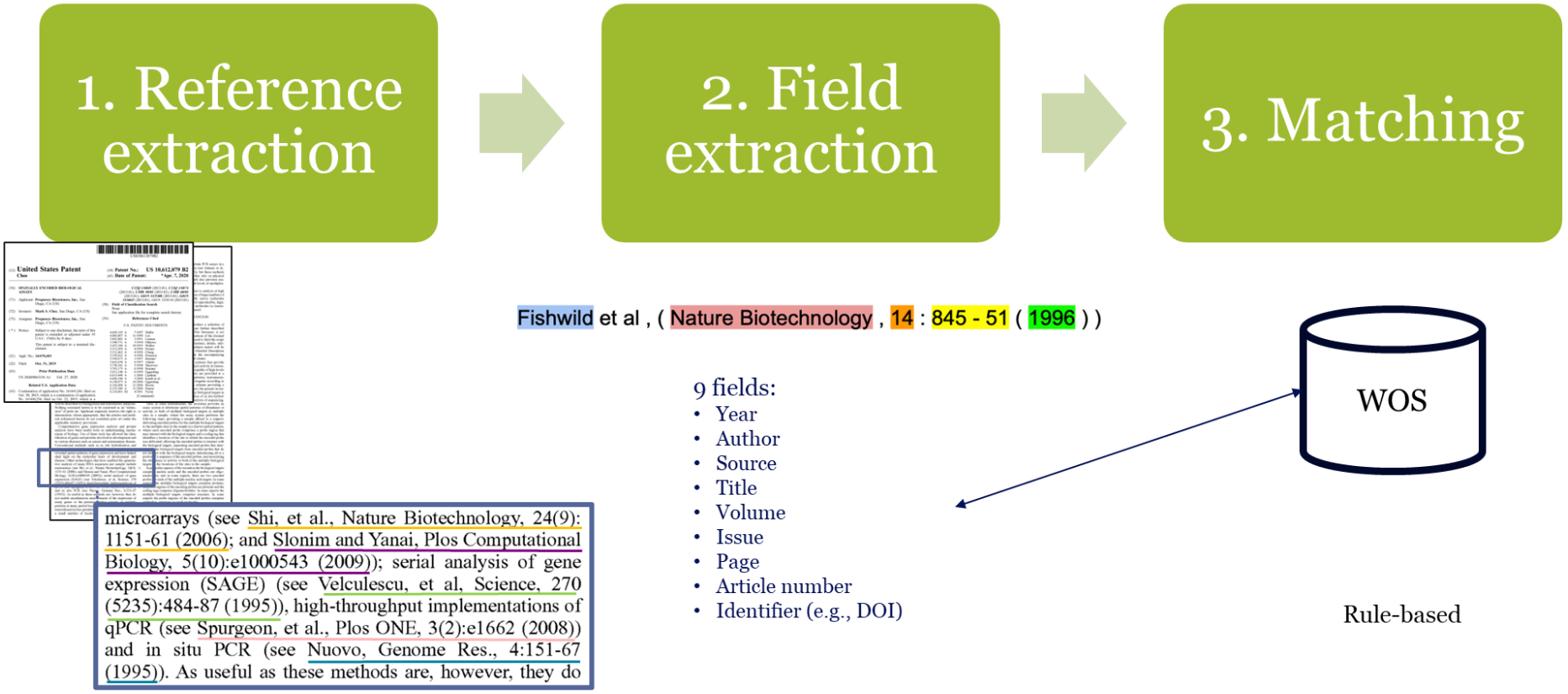
Objective 3: Uncovering characteristics of scientific publications that are particularly useful for patented inventions and whether and how the cited science leaves an imprint on the citing patent.

Previous studies have provided many insights into the mechanisms through which companies may benefit from science and factors that may facilitate the knowledge transfer from the academia to the industry. However, we know little about what types of scientific research are more useful for technological development. This project fills this gap in the literature, by exploring which types of scientific references lead to more valuable patents. Furthermore, as our previous studies suggest that patent in-text and front-page references embody different kinds of information, we compare in-text and front-page references. This research advances our understanding on the relationship between science and technology, as well as informs science and innovation policy and management. Specifically, research findings are relevant for policymakers and university administrators to understand what types of research should be encouraged, if the goal is to make science more useful for the industry. Research findings are also relevant for R&D managers for understanding what types of science are of high value, depending on their innovation strategies.

2. Method: A three-stage pipeline

We developed a three-stage pipeline for extracting the in-text scientific references from patents and matching them to WoS publications, as illustrated in Figure 1. The first stage is reference extraction. In this stage, given the text of the patents, the in-text references are extracted using a sequence labeling model. In the second stage, field extraction, the fields of the references (author name, year, journal name, etc.) are extracted from the reference texts. In the third stage, matching, the extracted fields are matched with entries in the Web of Science (WoS) publication database to find the corresponding scientific publication. We use pre-trained sequence labeling models for the reference extraction and field extraction stages that we fine-tune for our task. For training these models and evaluating our proposed pipeline, we have collected a manually annotated benchmark dataset. We provide a detailed explanation of each of the stages in our pipeline in the following sections.

Figure 1. Three-stage pipeline for extracting and matching patent in-text references.



2.1. Reference extraction

We approach the problem of reference extraction as a sequence labeling task. Given the text of the patent, the model labels each token with BIO labels, where "B" means the beginning token of a reference, "I" means a token inside a reference, and "O" means a token outside of the reference. We fine-tune different pre-trained BERT-based language models for context modeling with a linear layer on top for classification. We experiment with patent-specific language models, i.e., PatentBert (Lee & Hsiang, 2019) and Bert for Patents (Srebrovic & Yonamine, 2020), as well as other language models including BERT (Devlin et al., 2018) and SciBERT (Beltagy et al., 2019).

To form the input sequences of the model, the text of the patent is segmented based on the following rules: We need to make sure the length of the sequence does not exceed the maximum sequence length allowed by the model. Therefore, we first tokenize the patent text using the tokenizer of the same language model, then segment the text in sequences of at most 512 tokens.

After labeling texts with the fine-tuned sequence labeling model, the references are identified based on the predicted BIO labels. Each "B" label indicates the start of a new reference and the end token of a reference is the last "I" label after the "B" label.

2.2. Field extraction

The field extraction stage is also a sequence labeling model based on BERT models. Our method includes a pre-trained BERT-based language model and a classification layer on top of each output. The input of this model is the extracted reference from the previous stage. We do not give the context of the reference to the field extraction model and only rely on the reference text itself. We define 14 labels:Year, Author-B, Author-I, Source-B, Source-I, Title-B, Title-I, Page-B, Page-I, Volume, Number, Issue, Identifier-B, Identifier-I. The Identifier can be the DOI, ISSN, or BSSN number, Title is the name of the publication, Source can be the name of the journal, conference, or name of the book when the reference is to a book chapter. Using the above-mentioned labels, we extract (at most) 9 fields from references and pass them to the next stage.

2.3. Matching

In this stage, the fields extracted from the reference text are used to identify the referred scientific paper. We only process the references that have a year field and for which the value of the year is later than 1980 because our WoS database only includes publications starting from 1980. Some of the fields, like Author, can have multiple values. For the fields with multiple values, we only consider the first author. We create a vector representation of the reference and match the vector with all of the publications in that year. For the Author field, we consider a match when the author's name of the record is a sub-string of the author's name of the reference or vice versa. For the Journal field, we consider it a match when the extracted journal name is a sub-string of the full name, or standard abbreviations provided by the WoS database, or vice versa. For the rest of the fields, we consider a match when we have an exact match between the fields of the reference and the fields of that record. The count of the matched fields for each record is calculated and stored as the *match_score s*, indicating the number of fields out of the 9 fields (excluding the year, which was a prerequisite for the matching) of the reference that are matched with this record. The matched entity is the ID of the record that is matched to the reference. In addition, we define a Boolean variable *exact_match* that indicates whether the reference is matched to a record in the WoS database or not. The *exact_match* variable is True if the (1) Title or Identifier (DOI) of the reference is matched with exactly one record in the WOS database or (2) the maximum number of *match_score s* is at least 3 and only there is one record with this *match_score*, or (3) four fields of reference including volume, issue, page, and source are matched with the record.

3. Dataset for training the pipeline

3.1. Sample

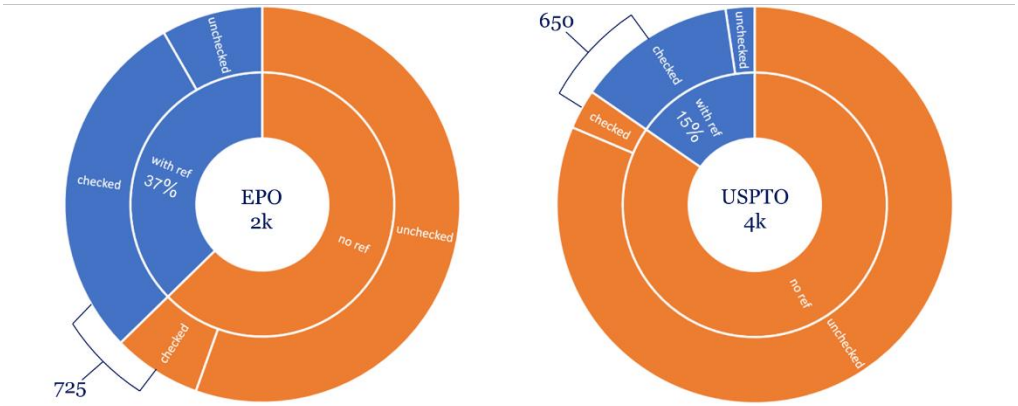
We sampled a set of EPO patents and a set of USPTO patents, separately, for student assistants to manually annotate. We gradually sampled 1000 random patents and released them for annotation, until we exhausted all our person-hours. Furthermore, only patents that have in-text references are useful for training the model, while the majority of patents do not have any in-text preferences. Therefore, for the sampled 1000 patents, we kept all patents that

were predicted to have at least one in-text reference for annotation using the model developed by Verberne et al. (2019). For every four such patents, we added one randomly selected patent that was predicted not to have any in-text references. We discarded other patents that were predicted not to have any in-text references from annotation work. Another point of consideration was to have a dataset with a balanced number of annotated EPO and USPTO patents. In the following, we provide further details regarding the sample (see also Figure 2).

For EPO patents, we downloaded and processed "EP full-text data for text analytics" from: https://www.epo.org/searching-for-patents/data/bulk-data-sets.html. The version we downloaded covers patents up to Week 30 of 2021, including file up to "EP3800000.txt." We kept patents that meet the following criteria: (1) published between 1990-2022, (2) are in English, (3) have title and description fields, and (4) are granted utility patents (type B1, B2, B3, and B9), when a patent has multiple published versions, we kept the most recent version. Then we randomly sampled 2000 EPO patents in two batches. In order to pre-filter the patents without any in-text reference, we used the reference extraction model developed by Verberne et al. (2019) to identify the references of the patents in this sample. According to the prediction of this model, 746 of them had at least one reference. For annotation, we kept all these 746 patents and additionally sampled 187 patents from the rest of the patents which were predicted not to have any references. We gradually released them for annotation. In the end, we managed to annotate 725 patents (consisting of 580 predicted to have references and 145 predicted not to have any).

For USPTO patents, we downloaded "Patent Grant Full Text Data (No Images) (JAN 1976 - PRESENT)" from: https://bulkdata.uspto.gov/. The version we downloaded covers patents up to 12 March 2022, including files up to "ipg220412.xml." We kept only B1 or B2 versions of utility patents published between 1990 and 2022. Then we randomly sampled 4000 patents in four batches. We used the same model for predicting the references of the sample of patents. According to the prediction of the model, 615 of them had at least one reference. For annotation, we kept all these 615 patents and randomly sampled 154 from the rest of the patents that were predicted to have no references. We gradually released them for annotation. In the end, we annotated 650 USPTO patents of which 520 were predicted to have references and 130 not.

Figure 2. Sample composition for training the pipeline.



3.2. Annotation

In the pre-processing step, we added whitespace before and after each punctuation mark in the text of patents, because it helps the annotators to select the exact span easier and, in this way, we can ensure that they are correctly tokenized by the tokenizer of BERT models.

We had two rounds of annotation, (1) the first for the first stage reference extraction model and (2) the second for the second stage field extraction model and the third stage of matching. In the first round, we hired 8 master students at Leiden University for 40 hours each in one month, for annotating references in patent texts. We designed a guideline for annotators and trained them in one session. We had one pilot annotation step in which all annotators annotated the same 10 patents. We evaluated their performance in the pilot step and gave them feedback on their performance, and additional instructions for the actual annotation.

The annotators were given a total number of 4 batches of data to annotate (one batch per week). We included some overlap patents in each batch between different persons to measure their inter-rater agreement.

The annotated references were then checked by two senior members of the project team and were modified based on the agreement between them. In addition, some human errors were spotted during the second round of annotation (see the next paragraph) and corrected accordingly. Furthermore, in the evaluation step, we manually checked the prediction of the model for each fold of the dataset and modified the dataset based on the predicted references

if needed. For the overlap patents, we included all of the references annotated by both annotators.

In the second round of annotation, for stages 2 and 3 of the pipeline, we hired four master students at Leiden University for 40 hours each in one month. We asked them to (1) annotate fields of the references, (2) find the corresponding WoS publication of the reference, and (3) determine the type of reference (which can be Journal, book, manual, and other).

The statistics of the annotated dataset is shown in Table 1.

Table 1. Statistics of the annotated dataset.

	# Extracted References	# Patents	# Matched References
EPO	3900	392	2088 (53.5%)
USPTO	3901	319	2247 (57.6%)
All	7801	711	4335 (55.6%)

4. Performance evaluation

We first evaluate the performance of each stage model and then the whole pipeline, end to end.

4.1. Reference extraction

We compared multiple pre-trained BERT-based models including patent specific models, i.e., BERT for patent and PatentBERT, and other models like SciBERT and BERT-base. In this stage, a model with higher recall is preferred because we want to extract as many as the references from the patent text, and precision can be improved in the latter stages, more specifically, non-scientific or wrong references would be discarded in the matching stage when the matching model cannot match them with any record in the WoS dataset. We used five-fold cross-validation for evaluating the models. The results of evaluating the BERT models are shown in Table 2. SciBERT achieved the highest recall. We found the best parameter setting using five-fold cross-validation. We fine-tuned the SciBERT model using the best parameter setting on the entire labeled dataset for the main pipeline.

Table 2. Performance of different language models for reference extraction (Stage 1).

Model	Recall B	Precision B	Recall I	Precision I
SciBERT	**0.953**	0.760	**0.985**	0.780
BERT for patents (large)	0.952	**0.767**	0.983	**0.781**
BERT	0.903	0.725	0.976	0.760
PatentBERT	0.943	0.752	0.981	0.775

4.2. Field extraction

We compared three BERT-based language models including NER-BERT (Liu et al., 2021), SciBERT (Beltagy et al., 2019), and BERT-base (Devlin et al., 2018). We evaluated the models using five-fold cross-validation. We had nine labels and selected the best performance of the models based on average Precision and Recall for these nine labels. We have selected the best parameter setting based on five-fold cross-validation and used it for training the model for the main pipeline on the entire dataset. For extracting the fields, both precision and recall are important because not only we need to extract as many fields as possible, but also the incorrect fields can lead to incorrect matching in the next stage. The Results of different models are reported in Table 3. The SciBERT model achieved the best performance for field extraction.

Table 3. Performance of different language models for field extraction (Stage 2).

Model	Precision	Recall	F1	Accuracy
SciBERT	**0.945**	**0.958**	**0.951**	**0.966**
NER-BERT	0.942	0.956	0.949	0.965
BERT	0.941	0.956	0.948	0.965

4.3. Matching

The matching model is not a supervised model but a rule-based (as specified in the method subsection). The result of evaluating the matching model using the entire dataset is reported in Table 4. We removed duplicate reference strings from the set of extracted references and

used the unique references (in each patent) for evaluating the matching model. The collected dataset has 6956 unique references.

Table 4. Performance of matching (Stage 3).

TP	TN	FN	FP	Precision	Recall	Accuracy
59%	37.2%	2.95%	0.83%	98.6%	95.2%	96.2%

4.4. End-to-End

For the end-to-end evaluation of the whole pipeline consisting of three models, we used 80% patents to train the models and the remaining 20% for testing. We used the best parameter setting found using five-fold cross-validation for training stage 1 and 2 models. The predicted references by the stage 1 reference extraction model are passed to the stage 2 field extraction model and extracted fields from the stage 2 model are then passed to the stage 3 matching model.

We first evaluate the stage 1 reference extraction model at the reference level (as subsection 4.1 only reported performance at levels of B and I tokens). Table 5 shows different types of outcomes. In total, the reference extraction model extracted 863 reference strings that match exactly with human-annotated reference strings. 38 references are extracted with minimal differences (e.g., missing the punctuation marks at the end), and 28 with small differences (e.g., some author names were not included) which however does not affect further matching. Interestingly, the model extracted 27 references that were not annotated by human annotators, but our further investigation concluded that the model prediction was correct while human annotators missed them. We classify all these four types of cases as true positives. Our model failed to extract 15 references and extracted 8 references with substantial differences (e.g., only include author name) such that these references cannot be matched. We classify these two types of cases as false negatives. It is interesting to note that these cases coincide with difficult cases reported by our human annotators during the annotation process; our human annotators were not confident about whether they should be annotated or not. Our model extracted 11 references that are not actually references, which we label as false positives. Overall, the reference extraction model achieved the precision of 98.9% and recall of 97.7%, which are very high.

Table 5. Reference extraction: performance at the reference level.

Label	N	Description
True positive	863	exact extraction
	38	extracted with minimal differences, e.g., punctuation
	28	extracted with small differences
	27	extracted but not in gold data
False negative	15	extraction failed
	8	extracted with substantial differences
False positive	11	false extraction

Prec	Recall	F1
98.9%	97.7%	98.3%

Finally, we evaluate the end-to-end performance by comparing the set of unique patent-paper-pairs identified by our model and those identified by human annotators. The end-to-end precision is 96.8% and recall is 91.9%, which are satisfactory.

Table 6. Whole pipeline end-to-end performance.

Prec	Recall	F1
96.8%	91.9%	94.3%

5. Implementation

The whole pipeline was trained using the fully annotated dataset and then implemented for extracting and matching the references from the corpus of EPO and USPTO patents.

5.1. EPO

The pipeline is executed on all utility patents of EPO that are published between 1990-2022. More specifically, we kept patents with types of B1, B2, B3, or B9, written in English, and then processed the most recent version with both description and title information. From 492,469 of these patents, our pipeline extracted at least one reference, and these patents collectively made 5,438,836 references. The average number of references for the patents

that have at least one reference is 11. About 51% (2,763,779 references) of the extracted references were matched to scientific publications in WoS.

The distribution of the *match_score s* for the extracted references is shown in Table 7. The value of s = -2 means that the reference does not include publication year information and s = -1 means that the reference has a publication year before 1980. We do not have the record of the publications before the year 1980 in WoS. References that do not have publication year information or were published before 1980 account for around 42% (1,123,513) of the non-matched references and 21% of the total references.

Table 7. The Distribution of the match_score for extracted references.

match_score s	EPO	USPTO
-2	638,393	2,445,168
-1	485,120	1,351,659
0	226,341	728,867
1	1,125,257	4,034,114
2	299,178	1,159,912
3	615,509	2,209,559
4	1,650,860	6,508,584
5	348,116	1,766,368
6	49,765	225,776
7	297	2,182

5.2. USPTO

The pipeline is executed on all utility patents of USPTO that are published between 1990-2022. Our pipeline has extracted 20,432,189 references from patents of 1990-2022. These references are extracted only from 1,449,398 unique patents. The average number of references for each USPTO patent (that has at least one reference) is 14. Among 20,432,189 extracted references, 11,069,995 of them (54%) are matched and 9,362,194 of them (46%) are not matched with a publication.

The distribution of the matching score for the extracted references is shown in Table 7. According to this, 34% of the non-matched references do not have a year or are published after 1980. These references account for 19% of the total references.

6. Investigating science-technology-linages

We investigated how patent value is affected by referencing different types of science, in terms of basicness, interdisciplinarity, novelty, and scientific impact, using a subset of 33,337 biotech utility patents granted by USPTO from 2006 to 2010, and their 860,879 in-text references matched to the Web of Science (WoS) database. We focus on one particular field to control for field heterogeneities, and we choose biotech because it is a sector relying heavily on science.

6.1. Measures

Patent measures (dependent variables)

For each patent, we constructed two measures to capture its value: (1) Patent citations, which is the number of times a patent is cited by future patents, using a five-year citations time window, following the common practice. (2) Market value, which is based on the stock market response to the issuing of the patent, in million US dollars, developed by Kogan et al. (2017). Information about the market value is available for a subset of 7,336 patents in our sample.

Science measures (independent variables)

We first constructed measures for individual scientific papers. For Basicness, we adopted the measure proposed by Weber (2013) for biomedical research, which classifies a paper as highly basic if it only has cell/animal-related MeSH terms but no human-related MeSH terms, moderately basic if it has both cell/animal- and human-related MeSH terms, and not basic (i.e., clinical) if it only has human-related but not cell/animal-related MeSH terms. This measure is an ordinal measure, and its attributes 1, 2, and 3 correspond to not basic, moderately basic, and highly basic, respectively.

For Interdisciplinarity, we adopted the Rao-Stirling measure (Stirling, 2007), which captures all the three diversity dimensions (i.e., variety, balance, and disparity) of the involved disciplines underlying a study. More specifically, it equals $\sum_{i \neq j} p_i p_j d_{ij}$, where i and j are indices of a paper's referenced disciplines (i.e., WoS subject categories), p_i is the proportion

of references to discipline i, and d_{ij} is the distance between discipline i and j, measured as 1 – cosine similarity between discipline i and j based on their co-citation matrix. This measure is a continuous measure ranging from 0 to 1.

For Novelty, we adopted the measure developed by Wang et al. (2017), which follows the combinatorial novelty perspective and identifies novel paper as the ones that makes unprecedented combinations of pre-existing knowledge components, where knowledge components are proxied by referenced journals. This measure is a binary variable: 1 if novel and 0 if not novel.

For Scientific citations we count the number of forward citations a scientific paper receives from future papers in the Web of Science (WoS) database, using a five-year citation time window, following the common practice.

At the patent level, for quantifying a patent's profile of referenced science, in terms of basicness, interdisciplinarity, novelty, and scientific citations, we take the average of these four measures across its referenced scientific papers: Avg(Basicness), Avg(Interdisciplinarity), Avg(Novelty), and Avg(Scientific citations). In addition, our focal explanatory variables also include I(sNPR), which indicates whether a patent has any scientific references, and sNPRs, which is the number of unique WoS papers referenced by a patent. Descriptive statistics are reported in Table 8.

Table 8. Descriptive statistics (Unit of analysis: patent)

	N	Mean	Std. Dev.	Min	Max
Patent citations	33,337	3.766	8.238	0	549
Market value (m$)	7,336	46.757	87.453	0.001	993.601
I(sNPR)	33,337	0.806	0.395	0	1
sNPRs	26,872	32.036	43.005	1	711
Avg(Basicness)	26,012	2.445	0.394	1	3
Avg(Interdisciplinarity)	26,709	0.254	0.032	0.037	0.429
Avg(Novelty)	26,872	0.155	0.157	0	1
Avg(Scientific citations)	26,872	124.208	191.441	0	5871

6.2. Results

We estimate how the characteristics of referenced science affect patent value, as measured by patent forward citations, where referenced science is based on in-text references. The dependent variable is an over-dispersed count variable, so we fit a series of Negative Binomial (NB) models. Regression results are reported in Table 9. Column 1 reports the NB model that uses whether having scientific references as the focal independent variable and incorporates the complete set of patent's issuing year and IPC class dummies. The result suggests that patents having in-text scientific references receive 29.1% more patent citations than patents not having in-text scientific references, issued in the same year and IPC class. Within the set of patents that have in-text scientific references, we further examine the intensity of reliance on science, that is, the number of referenced scientific papers. This independent variable is also a count variable and has a skewed distribution, so we take its natural logarithm for regression analysis. Column 2 shows that as a patent's number of referenced papers increases by 1%, its patent citations increase by 0.122%.

Then we move on to explore the characteristics of referenced science. Column 3-6 each uses average basicness, interdisciplinarity, novelty, and scientific citations of referenced papers as the focal independent variable. In all these models, the ln(number) of scientific references is controlled for, in addition to patent issuing year and IPC class. Avg(Scientific citations) is skewed so it takes natural logarithm transformation for regression analysis. Column 3 shows that, as the average basicness of referenced papers increases by 1, patent citations decrease by 7.0%, holding all other variables fixed. Column 4 suggests no significant effects of interdisciplinarity. Column 5 shows that, as the average novelty of referenced papers increases by 1, patent citations increase by 15.6%, holding all other variables fixed. Column 6 suggests no significant effects of scientific citations. Column 7 further fits a model with all these four variables together and yields consistent results as running separate models for each independent variable (i.e., Column 3-6). In summary, patents building on less basic but more novel science are more impactful in the technological domain.

Table 9. In-text scientific references and patent citations

	Patent citations NB						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I(sNPR)	0.291***						
	(0.029)						
ln(sNPRs)		0.122***	0.133***	0.122***	0.122***	0.124***	0.136***
		(0.011)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)
Avg(Basicness)			-0.070*				-0.073*
			(0.034)				(0.035)
Avg(Interdisciplinarity)				0.622			0.219
				(0.446)			(0.506)
Avg(Novelty)					0.156+		0.175*
					(0.082)		(0.086)
ln(Avg(Scientific citations) +1)						-0.009	-0.009
						(0.013)	(0.014)
Issue year	Y	Y	Y	Y	Y	Y	Y
IPC class	Y	Y	Y	Y	Y	Y	Y
N	33337	26872	26012	26709	26872	26872	25930
BIC	152066	123120	118841	122524	123115	123130	118555

Robust standard errors in parentheses. *** p < .001, **p < .01, *p < .05, +p < .10.

We then use the market value (in million US dollars) of the patent based on stock market reaction to the event of patent being issued as the dependent variable. This variable is also skewed but not a count variable, so we take natural logarithm transformation and then fit Ordinary Least Squares (OLS) models. Results are reported in Table 10. Results show that patents with in-text scientific references worth 86.8% more than patent without in-text scientific references (Column 1). Within the set of patents having in-text scientific references, patent market value increases by 0.190% as the number of referenced scientific papers increases by 1% (Column 2). As the average basicness of referenced science increases by 1, patent market value decreases by 37.8% (Column 3). Interdisciplinarity and novelty has no significant effects on patent market value (Column 4 and 5). As the average scientific citations of referenced science increase by 1%, patent market value increases by 0.094% (Column 6). These results are robust when fitting a model with all four variables together (Column 7). In summary, patents building on papers that are less basic but more highly cited in science generate higher private market value.

Table 10. In-text scientific references and patent market value

	ln(Market value) OLS						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
I(sNPR)	0.868***						
	(0.079)						
ln(sNPRs)		0.190***	0.191***	0.182***	0.190***	0.164***	0.163***
		(0.019)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)
Avg(Basicness)			-				-
			0.378***				0.380***
			(0.081)				(0.082)
Avg(Interdisciplinarity)				-0.523			0.122
				(0.929)			(1.095)
Avg(Novelty)					0.003		-0.073
					(0.189)		(0.223)
ln(Avg(Scientific citations)+1)						0.094**	0.116***
						(0.029)	(0.032)
Issue year	Y	Y	Y	Y	Y	Y	Y
IPC class	Y	Y	Y	Y	Y	Y	Y
N	7336	6181	5984	6143	6181	6181	5969
R2	0.091	0.074	0.077	0.073	0.074	0.076	0.079
BIC	30944	25341	24427	25197	25341	25337	24374

Robust standard errors in parentheses. *** p < .001, **p < .01, *p < .05, +p < .10.

Our regression models assume a linear equation, where the left-hand side is the natural log of a dependent variable (patent citations or market value), and the right-hand side consists of a series of independent variables. This setup is flexible for fitting positive (or negative) effects at an increasing or decreasing rate. However, it does not allow nonmonotonic effects (e.g., inverted U-shaped) or discontinuous effects. Therefore, we reexamine the effects using a more non-parametric approach without assuming a linear equation. Specifically, we categorize our independent variables into 10 ordered and evenly sized groups and then estimate the expected patent citations and market value for each group. Because of ties, not all groups are evenly sized. Taking the number of in-text scientific references as an example, 6,465 patents with 0 references are classified into Group 1, the next 1,891 patents with only 1 reference are classified into Group 2, …, the last 3,330 patents with 71 to 711 references are classified into Group 10. Then we use group numbers as a factor/categorical variable for regression analysis, and results are reported in Table 11 and 12. Based on the regression results, we can estimate the expected value of the dependent variables for each group for an average patent (i.e., issuing year is 2010, IPC class is C12N, and other control variables (if any) at the mean). Figure 3 plots these estimates.

Table 11. In-text references and patent citations: A nonparametric approach

	Patent citations NB				
	(1)	(2)	(3)	(4)	(5)
Group:	sNPRs	Basicness	Interdisciplinarity	Novelty	Scientific citations
2	0.018	0.170**	-0.063	/	0.009
	(0.048)	(0.061)	(0.058)		(0.053)
3	0.153*	0.027	-0.069	-0.176**	0.105
	(0.074)	(0.051)	(0.060)	(0.056)	(0.077)
4	0.040	0.066	-0.056	-0.094	-0.018
	(0.043)	(0.054)	(0.062)	(0.058)	(0.054)
5	0.171***	-0.062	-0.066	-0.039	0.099+
	(0.041)	(0.054)	(0.060)	(0.054)	(0.057)
6	0.401***	-0.018	-0.138*	0.026	-0.012
	(0.046)	(0.049)	(0.059)	(0.056)	(0.058)
7	0.352***	-0.048	-0.009	0.047	0.012
	(0.045)	(0.051)	(0.059)	(0.054)	(0.060)
8	0.275***	-0.067	0.003	0.040	-0.226***
	(0.043)	(0.051)	(0.057)	(0.051)	(0.062)
9	0.432***	-0.027	-0.022	0.161**	-0.030
	(0.042)	(0.045)	(0.057)	(0.050)	(0.061)
10	0.616***	/	0.090	0.021	0.113+
	(0.044)		(0.077)	(0.070)	(0.061)
ln(sNPRs)	/	0.129***	0.135***	0.127***	0.130***
		(0.013)	(0.010)	(0.013)	(0.012)
Issue year	Y	Y	Y	Y	Y
IPC class	Y	Y	Y	Y	Y
N	33337	26012	26709	26872	26872
BIC	151854	118890	122573	123135	123129
Chi2	314***	22**	16+	44***	53***
LR Chi2	54***	22**	22**	51***	73***

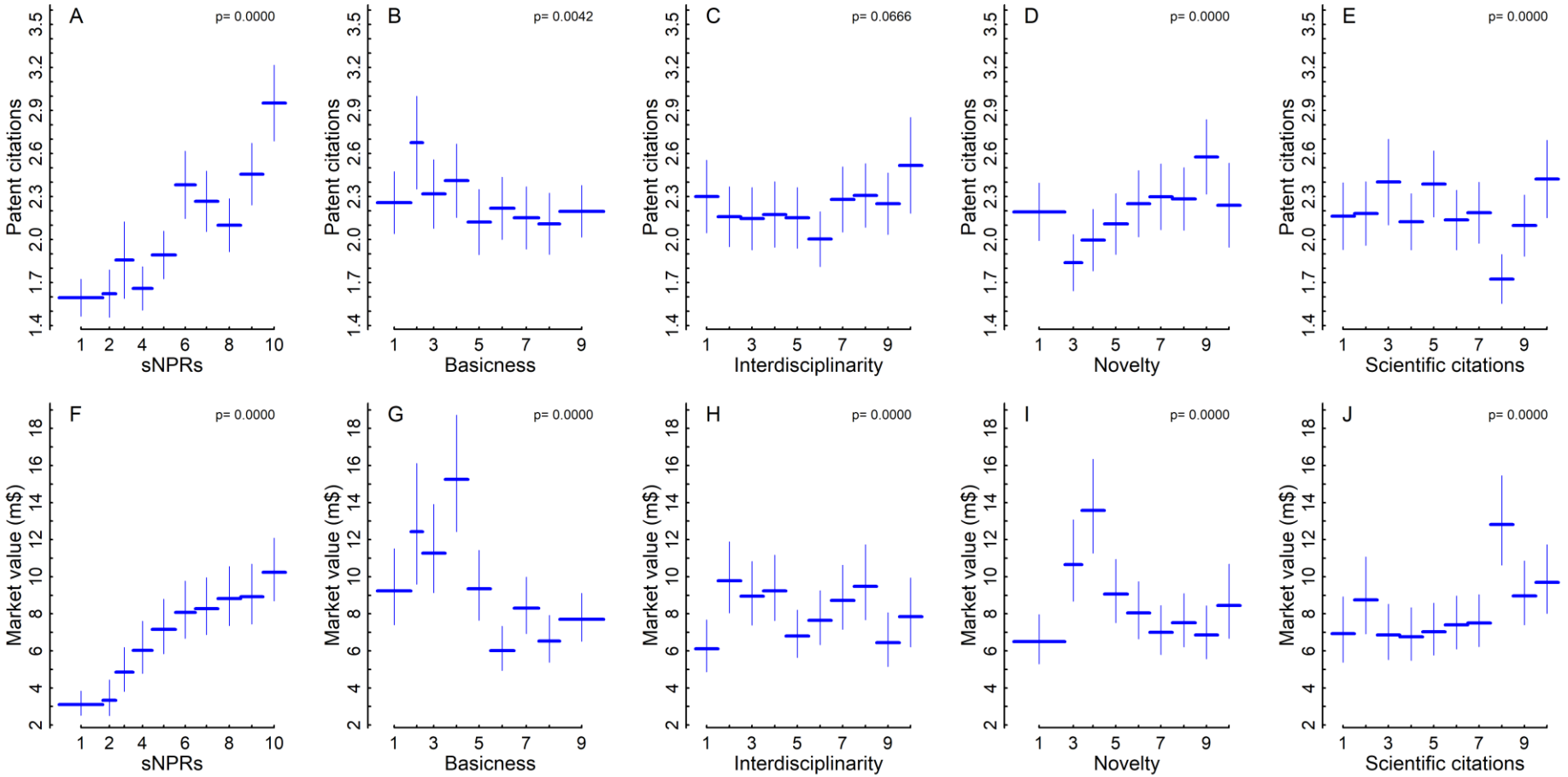
This table repeats the analysis reported in Table 9 but uses the categorized science measures as independent variables instead. Take the variable *sNPRs Group* as an example, we code it as 1 if a patent's number of referenced papers is among the lowest 10%, 2 if among the next 10%, … 10 if among the top 10%. The complete categorization scheme is reported in Appendix Table A1. Due to ties, not all groups are evenly sized, and sometimes two groups are be merged. For example, for Avg(Basicness), group 9 and 10 are merged and labelled as group 9; for Avg(Novelty), group 1 and 2 are merged and labelled as group 1. Then we use these categorical variables as the independent variable for regression. Group 1 is the reference group. Coefficients indicates the difference between a focal group and the reference group (i.e., Group 1). Take Column 1 as an example, the coefficient of 0.018 for Group 2 means that patents in Group 2 have 1.8% more patent citations than Group 1. Chi2 tests the joint significance of all the levels of the group variable (H0: all coefficients of Group 2, 3, …,10 equal to 0). LR Chi2 reports the likelihood ratio test between the focal regression model and the model with the raw uncategorized independent variable. More specifically, it tests the model in Table 4 column 2 against the model in Table 2 column 3, Table 4 Column 3 vs. Table 2 Column 4, Table 4 Column 4 vs. Table 2 Column 5, and Table 4 Column 5 vs. Table 2 Column 6. For Table 4 Column 1, we fit another model with *ln*(sNPRs +1) as the independent variable, because Table 2 Column 2 only includes patents with at least one reference. If we test Table 4 Column 1 against Table 2 Column 1, then the result is: 295***. Being significant here means the model in this table fits the data better than its corresponding model in Table 2. Robust standard errors in parentheses. *** p < .001, **p < .01, *p < .05, +p < .10.

Table 12. In-text references and patent market value: A nonparametric approach

	ln(Market value) OLS				
	(1)	(2)	(3)	(4)	(5)
Group	sNPRs	Basicness	Interdisciplinarity	Novelty	Scientific citations
2	0.070	0.298*	0.470***	/	0.234+
	(0.146)	(0.139)	(0.118)		(0.139)
3	0.447***	0.199+	0.381**	0.495***	-0.010
	(0.122)	(0.116)	(0.115)	(0.112)	(0.135)
4	0.662***	0.502***	0.412***	0.737***	-0.025
	(0.120)	(0.115)	(0.115)	(0.107)	(0.137)
5	0.835***	0.013	0.107	0.333**	0.015
	(0.107)	(0.115)	(0.116)	(0.112)	(0.135)
6	0.955***	-0.428***	0.225+	0.213+	0.066
	(0.101)	(0.116)	(0.117)	(0.112)	(0.134)
7	0.979***	-0.105	0.356**	0.075	0.079
	(0.100)	(0.108)	(0.118)	(0.111)	(0.134)
8	1.042***	-0.346**	0.439***	0.146	0.615***
	(0.099)	(0.116)	(0.125)	(0.108)	(0.133)
9	1.055***	-0.181+	0.052	0.054	0.257+
	(0.100)	(0.107)	(0.127)	(0.108)	(0.133)
10	1.193***	/	0.251+	0.262*	0.337*
	(0.092)		(0.132)	(0.114)	(0.131)
ln(sNPRs)	/	0.168***	0.166***	0.126***	0.163***
		(0.022)	(0.022)	(0.026)	(0.022)
Issue year	Y	Y	Y	Y	Y
IPC class	Y	Y	Y	Y	Y
N	7336	5984	6143	6181	6181
R2	0.107	0.093	0.079	0.087	0.086
BIC	30865	24374	25206	25314	25334
F	26***	16***	5***	12***	9***
LR Chi2	32***	106***	43***	88***	64***

This table repeats the analysis reported in Table 10 but uses the categorized science measures as independent variables instead. F tests the joint significance of all the levels of the group variable (H0: all coefficients of Group 2, 3, …,10 equal to 0). LR Chi2 reports the likelihood ratio test between the focal regression model and the model with the raw uncategorized independent variable. More specifically, it tests the model in Table 5 column 2 against the model in Table 3 column 3, Table 5 Column 3 vs. Table 3 Column 4, Table 5 Column 4 vs. Table 3 Column 5, and Table 5 Column 5 vs. Table 3 Column 6. For Table 5 Column 1, we fit another model with *ln*(sNPRs +1) as the independent variable, because Table 3 Column 2 only includes patents with at least one reference. If we test Table 5 Column 1 against Table 3 Column 1, then the result is: 133***. Being significant here means the model in this table fits the data better than its corresponding model in Table 2. Robust standard errors in parentheses. *** p < .001, **p < .01, *p < .05, +p < .10.

Figure 3. In-text scientific references and patent value.



This figure plots the estimated value of patent value for an average patent in different science measure groups. Plot A-E and F-J are based on regression models reported in Table 11 Column 1-5 and Table 12 Column 1-5, respectively. An average patent means its issuing year is 2010, IPC is C12N, and for Plot B-E and G-J, the natural log number of references takes the mean value. The length of the blue horizontal lines is proportional to the group size. The blue vertical lines mark the 95% confidence interval. p-value is for the joint significance of all the levels of the group variable, corresponding to Chi2 in Table 11 and F in Table 12.

Regarding the number of in-text references, consistent with the result reported in the preceding section, Fig 3A&F display a roughly continual increase in patent citations and market value, as the number of in-text references move from 0 (Group 1) to 1 (Group 2), and then further increases (from Group 2 to Group 10). There is consistent evidence that citing science and citing more scientific papers have a positive effect on patent value.

For average basicness, the previously reported result suggests that it has negative effects on patent citations and market value. However, results in Fig 3B&G suggest inverted U-shaped effects. Both patent citations and market value first increase and then decrease as the average basicness increases. Patent citations reach the peak point at Group 2, and if we dismiss Group 2 due to its small group size, then the peak point is reached at Group 4. Patent market value reaches its peak at Group 4. These results suggest that a moderate level of basicness is optimal for patent value, while too applied or too basic lead to lower patent value.

In terms of interdisciplinarity, our previous result suggests insignificant effects of interdisciplinarity on patent citations and market value. Fig 3C seems to suggest a U-shaped effect, but the p-value for the joint significance of interdisciplinarity groups is larger than 0.05. Fig 3H suggests no clear association between average interdisciplinarity and patent market value. Taken together, we conclude no significant effect of average interdisciplinarity on patent value.

Our previous result suggests that average novelty has a positive effect on patent citations but an insignificant effect on patent market value. Fig 3D&I reveal more complex patterns. According to Fig 3D, as a patent moves from having no novel references to having novel references, there is a sudden drop in patent citations. As the average novelty further increases, patent citations rise and slowly reach a plateau (or even go down). According to Fig 3I, there is a disruptive rise in patent market value when moving from having no novel references to having novel references. However, as the average novelty further increases, patent market value decreases and slowly flattens (or even bounces up). Taken together, these results suggest a structural change between patents building on novel science and those not. As a patent builds on novel science, its technological impact drops, potentially due to uncertainties introduced by sourcing novel science. Its technological impact then recovers and reaches a higher point than patents not building on novel science, indicating that sourcing more novel science leads to broader and more unexpected applications. However, this

increasing trend does not continue unlimitedly, the benefit from sourcing novel science stops at certain level of average novelty. Regarding patent market value, sourcing novel science brings a jump in the stock market reaction to the patented technology, reflecting market's appreciation of novelty. However, further increase in the novelty of sourced science reduces market value as the patent might become too remote from marketable applications.

Consistent with our previous result that average scientific citations have an insignificant effect on patent citations but a positive effect on patent market value, Fig 3E exhibits no clear associations between scientific citations and patent citations, but only fluctuations around a flat line, and Fig 3J displays an increasing trend with some fluctuations. Therefore, we conclude an insignificant effect of scientific citations on patent citations but a positive effect on patent market value. This suggests that the criteria of usefulness might not be perfectly aligned between science and technology, scientific outputs that are (perceived) useful for others to do follow-on scientific research (i.e., receive more scientific citations) do not necessarily leads to technologies that are (perceived) useful for others to develop follow-on technologies (i.e., receive more patent citations). On the other hand, there is neither a constriction between them as no significantly negative relation is observed. Regarding patent market value, scientific outputs that are highly recognized by other scientists are positively associated with technologies that are highly appreciated by the stock market, reflecting a certain level of alignment between scientists' interest and the market's interest.

7. Comparing patent in-text and front-page references

Patent in-text and front-page references are generated through different processed and accordingly record different kinds of information (Bryan et al., 2020; Marx & Fuegi, 2022; Verberne et al., 2019). Therefore, it is important to examine their differences and test whether findings regarding science-technologies are sensitive to which type of references are being analyzed. We use the same subset of USPTO biotech patents, more specifically, the 33,337 USPTO biotech patents and their 860,879 in-text references and 637,570 front-page references to WoS publications.

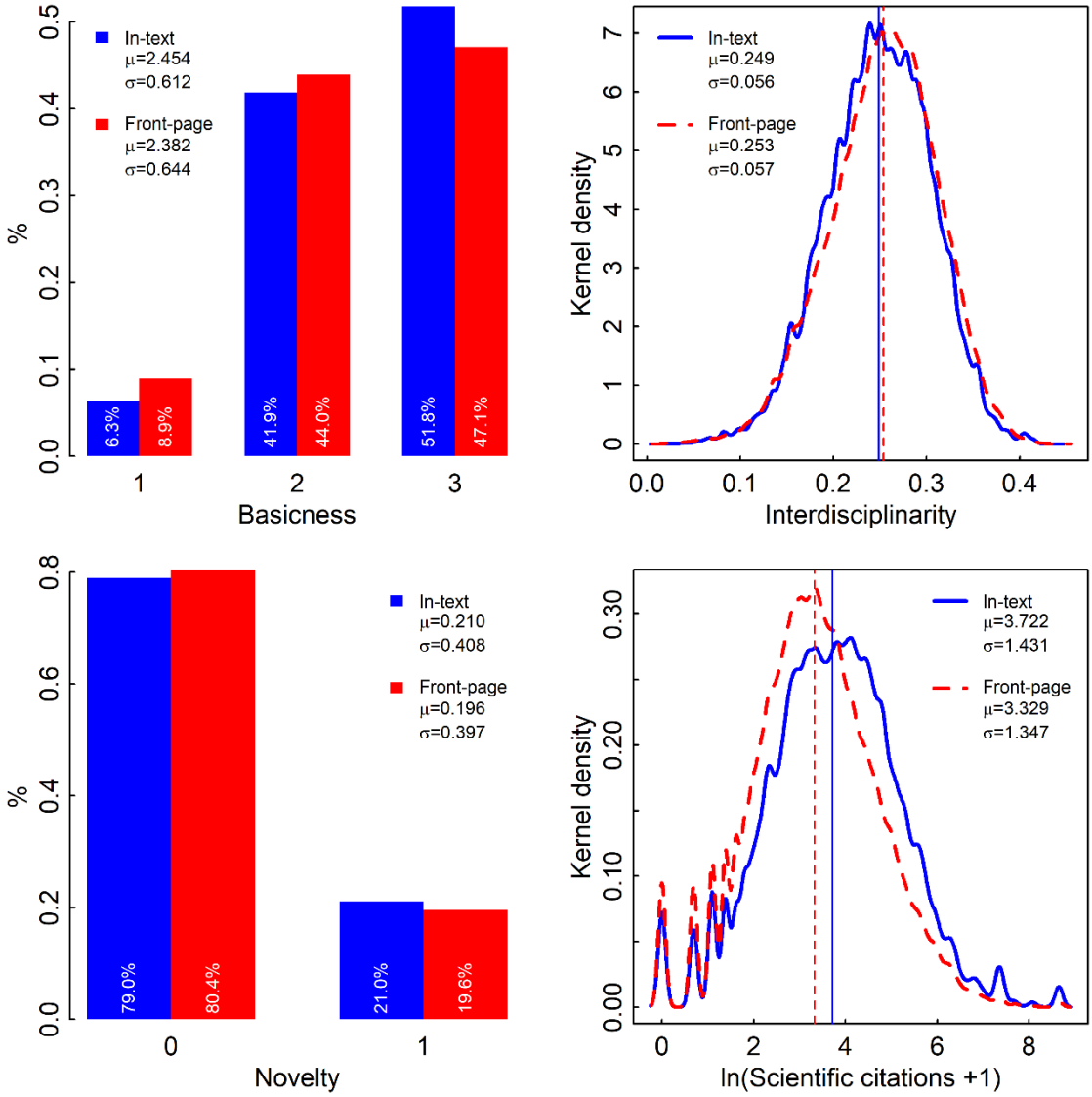
7.1. Reference comparison

Figure 4 reports the overlap between in-text and front-page references. In total, 173,281 references appear both in the text and on the front page of the same patent, which accounts for only 20% of all in-text references and 27% of all front-page references. This observed low overlap is consistent with prior observations (Bryan et al., 2020; Marx & Fuegi, 2022; Verberne et al., 2019).

Figure 4. Overlap between in-text and front-page references.

A scientific paper can be cited by multiple patents. The 1,325,168 total references are linked to 336,522 unique papers, the 860,879 in-text references are linked to 195,988 unique papers, and the 637,570 front-page references are linked to 245,852 unique papers. Although in-text references have a larger volume (i.e., more paper-patent-links), they are linked to fewer unique papers, compared with front-page references. In other words, in-text references are concentrated in a smaller set of papers than front-page ones. In-text referenced papers are cited more often than front-page referenced papers. On average, in-text referenced papers are cited by 5.4 patents in our sample in text or on front page, 4.4 patents in text, and 1.9 patents on front-page, and the corresponding numbers are 4.6, 2.7 and 2.6 for front-page referenced papers, respectively.

Figure 5. Distribution of basicness, interdisciplinarity, novelty, and scientific citations, by in-text and front-page references.



Plots for basicness and novelty are simple proportions by category. Plots for interdisciplinarity and *ln*(Scientific citations +1) are kernel densities where the vertical lines mark the mean values. Scientific citations are logarithm transformed because it is highly skewed.

We further assess the difference between in-text and front-page references in terms of their basicness, interdisciplinarity, novelty, and scientific citations. Figure 5 plots the distributions of these four measures for in-text and front-page references separately. Because the sample size is large, all the mean differences are highly significant (i.e., $p<0.001$) according to Welch

two sample t-tests and Wilcoxon rank sum tests, although the difference in interdisciplinarity and novelty seem very small in size. Taken together, results show that in-text references are more basic and have more scientific citations than front-page references. In-text references are less interdisciplinary but more novel than front-page references, but the differences are small. This finding suggests that studies of which kinds of science is more cited by patents might be sensitive to whether the data come from patent in-text or front-page references.

7.2. Patent level comparison

Descriptive statistics for patent-level variables are reported in Table 13. 80.6% of our sampled patents have in-text scientific references, while 87.3% have front-page references. Among those with in-text references, they cite on average 32.0 scientific papers in-text. Among those with front-page references, they cite on average 21.9 papers on front-page. These differences are significant according to Wilcoxon matched-pairs signed-rank test at significance level of 0.05. In the previous section, we have shown that in-text references have a larger volume but are concentrated in a smaller set of scientific papers. It appears that in-text references are also concentrated in a smaller set of patents.

Table 13. Descriptive statistics (Unit of analysis: patent)

	N	Mean	Std. Dev.	Min	Max
In-text					
I(sNPR)	33,337	0.806	0.395	0	1
sNPRs	26,872	32.036	43.005	1	711
Avg(Basicness)	26,012	2.445	0.394	1	3
Avg(Interdisciplinarity)	26,709	0.254	0.032	0.037	0.429
Avg(Novelty)	26,872	0.155	0.157	0	1
Avg(Scientific citations)	26,872	124.208	191.441	0	5871
Front-page					
I(sNPR)	33,337	0.873	0.333	0	1
sNPRs	29,110	21.902	31.368	1	1064
Avg(Basicness)	27,999	2.401	0.435	1	3
Avg(Interdisciplinarity)	28,982	0.256	0.037	0.023	0.419
Avg(Novelty)	29,110	0.162	0.178	0	1
Avg(Scientific citations)	29,110	62.580	88.399	0	5871

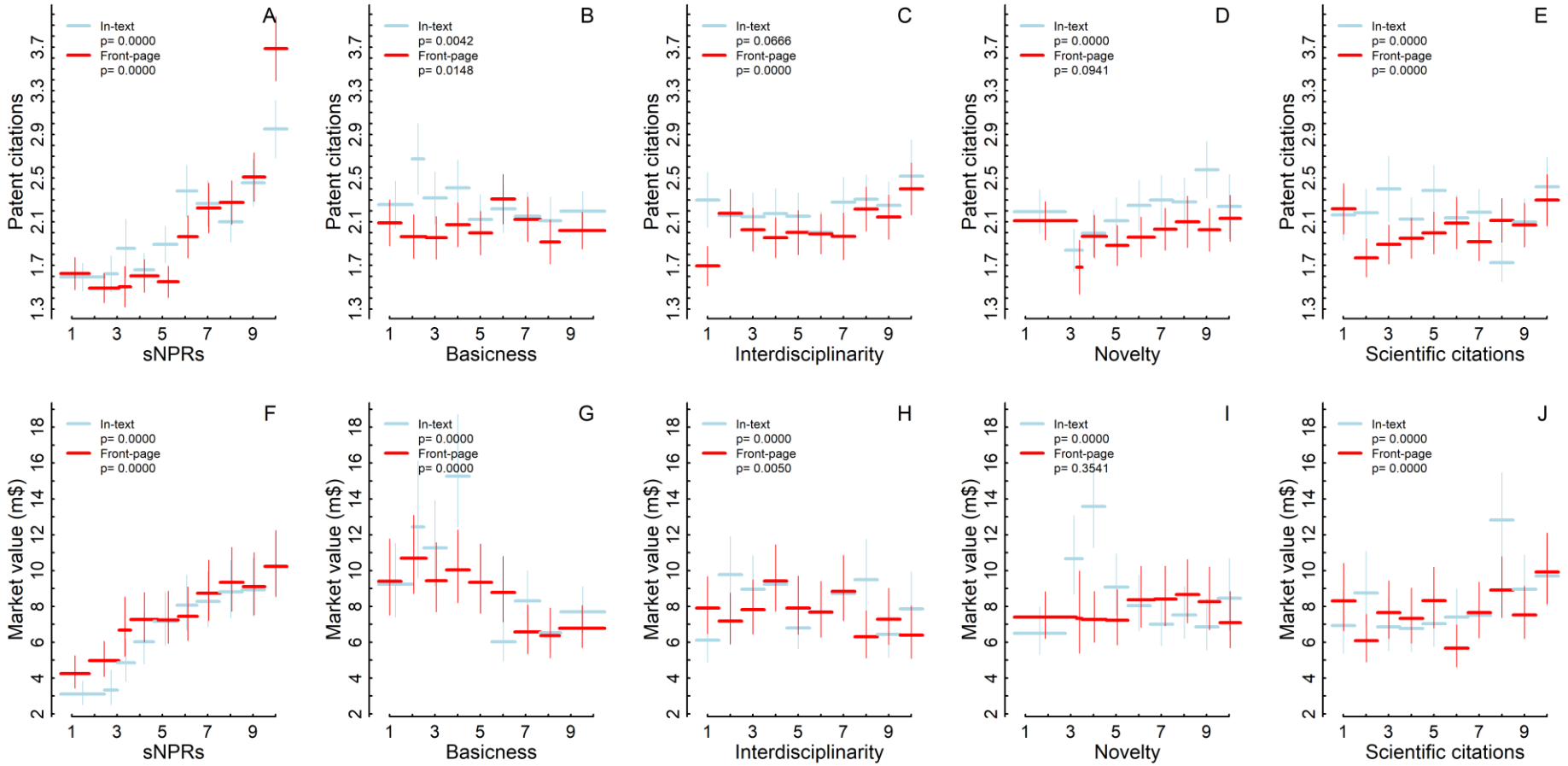
Wilcoxon matched-pairs signed-rank tests also suggest that the average basicness and scientific citations of papers in a patent's in-text references are significantly higher than that

of front-page references in the same patent, while there are no significant differences in average interdisciplinarity or novelty.

Correlations between the variables based on in-text and front-page references are moderate. The Spearman correlation is 0.466 between whether having in-text references and whether having front-page references. The correlations between two versions of variables (i.e., in-text and front-page) are 0.324, 0.602, 0.533, 0.261, and 0.430, for the number of referenced papers, average basicness, average interdisciplinarity, average novelty, and average scientific citations, respectively. These moderate correlations suggest that, if we rank patents by their number of scientific references or the average basicness, interdisciplinarity, novelty, and scientific citations of their referenced scientific papers, using in-text and front-page references will produce rankings that are substantially different. Furthermore, if we study the association between the characteristics of patents and the characteristics of their referenced scientific papers, we might come to different conclusions depending on whether in-text or front-page references are used.

Figure 6 replicates results reported in Figure 3 based on patent in-text reference but used front-page references instead. Analytical results are quite different.

Figure 6. Scientific references and patent value: Front-page vs. in-text.



The light blue lines in this plot are identical to the blue lines in Figure 3. They represent results based on in-text references. We overlay them with the red lines (results based on front-page references) for an easier comparison. Specifically, we repeat the same procedure for producing Figure 3 but use science measures based on patent front-page references instead.

7.3. Why do front-page and in-text references yield different results?

The inconsistencies between the results based on front-page and in-text references are not surprising, considering their low overlap and the moderate correlations between science measures based on front-page and in-text references. We then attempt to explore why such inconsistences emerge, by looking into the processes through which in-text and front-page references are generated. As discussed before, in-text references document various sources of knowledge that are instrumental to the patented technology, while front-page references are listed for disclosing prior arts that are relevant for assessing patentability. Sampat (2010) argued that, for patents that are expected to be more valuable, patent applicants may perform a more comprehensive prior art search, to prevent the chance that the patent application is rejected due to failure of disclosure. This more comprehensive search may result in a longer list of front-page references. Therefore, we view in-text references as an unbiased (but noisy) representation of the scientific outputs underlying a focal patented technology. In comparison, front-page references also reflect this unbiased representation but are subject to additional biases introduced by patent applicants' strategic behavior. It is possible that certain types of scientific papers are valuable for inspiring the patented technology but not so relevant for assessing its patentability and therefore are not listed on the front page. To explore this, we analyze individual in-text references and examine which types of in-text referenced papers are more likely to be listed on the front page of the same patent.

Using in-text references (i.e., paper-patent-links) as the unit of analysis, we fit conditional fixed-effects logistic models, with patent fixed effects to account for patent heterogeneities. Regression results are reported in Table 14. Column 1 shows that, for the same patent, among its in-text referenced papers, moderately basic papers have the highest chance of being listed on its front page, followed by highly basic, and lastly not basic papers. Column 2 and 3 show that more interdisciplinary and novel papers among the in-text referenced papers of the same patent are less likely to be listed on the front page of that patent. In contrast, papers with more scientific citations have a higher chance of being listed on the front page of the same patent (Column 4).

Table 14. What types of in-text referenced papers are more likely to be listed on the patent front page?

	I(Front-page) Conditional fixed-effects logit			
	(1)	(2)	(3)	(4)
Basicness=1	-0.225***			
	(0.019)			
Basicness=2	0.106***			
	(0.009)			
Interdisciplinarity		-0.638***		
		(0.077)		
Novelty			-0.025*	
			(0.011)	
ln(Scientific citations +1)				0.065***
				(0.003)
Paper: Publication year	Y	Y	Y	Y
Paper: Scientific field	Y	Y	Y	Y
N obs	570408	634225	613065	655311
N patents	15374	16217	16093	16394

Unit of analysis: in-text references, i.e., paper-patent-links through in-text referencing. All models incorporate patent fixed effects, so that estimates are about within-patent differences. The dependent variable I(Front-page) is a binary variable: 1 if an in-text referenced paper is listed on the front page of the same patent, and 0 otherwise. Each column reports one conditional fixed-effects logistic regression model. For Column 1 we treat basicness as a categorical variable with three levels: 1, 2, and 3. Level 3 is used as the reference group, so the coefficient reports the difference between a focal basicness group and the reference group. For example, the coefficient of -0.225 means the log-odds of being listed on the front-page (vs. not being listed) is 0.225 smaller for patents that are not basic (Basicness =1) than for patents that are highly basic (Basicness = 3, the reference group). Robust standard errors in parentheses. *** p < .001, **p < .01, *p < .05, +p < .10.

The inverted U-shaped relationship between basicness and the likelihood of being listed on patent front page is in line with the observed inverted U-shaped relationship between average basicness and patent value. This means that a moderate level of basicness is not only positively associated with higher patent value but also higher degree of relevance for assessing patentability. Both interdisciplinarity and novelty deviates from the existing paradigm, and their contribution to the patented technology might be rather unexpected. Therefore, their intellectual link to the patent is relatively distant and tenuous, and their relevance for assessing patentability is relatively low. On the other hand, highly cited papers have generated more follow-on research and therefore is also possible to have more direct relevance for assessing patentability. In addition, highly cited papers are more visible in both domains of science and technology, such that missing them would bring a higher risk of being rejected due to failure of disclosure. Since front-page references systematically under-

represent interdisciplinary and novel papers but over-represent moderately basic and highly cited papers. We can expect that using front-page references will yield substantially different results than using in-text references when analyzing these science measures.

8. Conclusion and deliverables

This project aimed to (1) develop a high-performing machine learning method to extract patent in-text references and then match them to the Web of Science database of scientific publications, (2) implement this method to EPO and USPTO patents to create a large-scale dataset linking patents to publications, and (3) uncover what kinds of science lead to more valuable patents and how in-text references are different from front-page references.

We developed a three-stage pipeline for extracting and matching patent in-text references and accomplished high performance (precision = 96.8% and recall of 91.9%). We implemented the pipeline to full texts of EPO and USPTO patents granted between 1990 and 2022. From 492,469 EPO patents we identified 5,438,836 references and matched 2,763,779 (51%) references to WoS publications. From 1,449,398 USPTO patents we identified 20,432,189 references and matched 11,069,995 (54%) references to WoS publications.

We uncovered the relationship between patent value and science basicness, interdisciplinarity, novelty, and scientific citations, as well as differences between patent in-text and front-page references.

This project produced the following deliverables:

- A large-scale dataset linking EPO and USPTO patents to WoS publications.
- A training dataset for future text-mining tasks.
- A methodology paper about the pipeline.
- A scientific paper about science-technology linkages.

Acknowledgments

We are grateful for the financial support and valuable feedback provided by EPO ARP. We thank Linda Andersson and Juan Carlos Gomez Carranza for their constructive discussions at two EPO ARP workshops, and engaging discussions from other participates at these workshops, as well as a series of other conferences and workshops where we presented results from this project. We thank Ellen Kageler, Emilio Sanchez Olivares, Famke Nouwens, Federica Spiga, Floris Tomassen, Luuk van den Nouweland, Marika Bazelmans, Vasiliki Kogia for their diligent and meticulous annotation work.

References

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, *357*(6351), 583-587. https://doi.org/10.1126/science.aam9527

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),

Bryan, K. A., & Ozcan, Y. (2020). The Impact of Open Access Mandates on Invention. *The review of Economics and statistics*, *0*(ja), 1-45. https://doi.org/10.1162/rest_a_00926

Bryan, K. A., Ozcan, Y., & Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy*, *49*(4), 103946. https://doi.org/https://doi.org/10.1016/j.respol.2020.103946

Callaert, J., Vervenne, J., Van Looy, B., Magermans, T., Song, X., & Jeuris, W. (2014). *Patterns of Science-Technology Linkage*.

Cassiman, B., Veugelers, R., & Zuniga, P. (2008). In search of performance effects of (in)direct industry science links. *Industrial and Corporate Change*, *17*(4), 611-646. https://doi.org/10.1093/icc/dtn023

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*(8-9), 909-928. https://doi.org/10.1002/smj.384

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics*, *132*(2), 665-712. https://doi.org/10.1093/qje/qjw040

Lee, J.-S., & Hsiang, J. (2019). Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*.

Li, D., Azoulay, P., & Sampat, B. N. (2017). The applied value of public investments in biomedical research. *Science*, *356*(6333), 78-81. https://doi.org/10.1126/science.aal0010

Liu, Z., Jiang, F., Hu, Y., Shi, C., & Fung, P. (2021). NER-BERT: a pre-trained model for low-resource entity tagging. *arXiv preprint arXiv:2112.00405*.

Marx, M., & Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, *31*(2), 369-392.

Nagaoka, S., & Yamauchi, I. (2015). *The Use of Science for Inventions and its Identification: Patent level evidence matched with survey*.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, *26*(3), 317-330. https://doi.org/10.1016/s0048-7333(97)00013-9

Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics*, *7*(3), 369-381. https://doi.org/10.1007/BF02017155

Nunn, H., & Oppenheim, C. (1980). *A patent journal citation network on prostaglandins*. Elsevier.

Poege, F., Harhoff, D., Gaessler, F., & Baruffaldi, S. (2019). Science quality and the value of inventions. *Science Advances*, *5*(12), eaay7323. https://doi.org/10.1126/sciadv.aay7323

Rassenfosse, G. d., & Verluise, C. (2020). PatCit: A Comprehensive Dataset of Patent Citations.". In.

Sampat, B. (2010). When Do Applicants Search for Prior Art? *The Journal of Law and Economics*, *53*(2), 399-416. https://doi.org/10.1086/651959

Srebrovic, R., & Yonamine, J. (2020). Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery. *White paper*.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society [Journal Article]. *Journal of the Royal Society Interface*, *4*(15), 707-719. https://doi.org/10.1098/rsif.2007.0213

Tamada, S., Naito, Y., Kodama, F., Gemba, K., & Suzuki, J. (2006). Significant difference of dependence upon scientific knowledge among different technologies. *Scientometrics*, *68*, 289-302.

Verberne, S., Chios, I., & Wang, J. (2019). Extracting and matching patent in-text references to scientific publications. Proceedings of the 4th Joint Workshop on Bibliometric-enhanced

Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019), Paris, France.
Verberne, S., D'hondt, E., Oostdijk, N., & Koster, C. (2010). Quantifying the Challenges in Parsing Patent Claims. Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval at ECIR 2010,
Veugelers, R., & Wang, J. (2019). Scientific novelty and technological impact. *Research Policy*, *48*(6), 1362-1372. https://doi.org/https://doi.org/10.1016/j.respol.2019.01.019
Voskuil, K., & Verberne, S. (2021). Improving reference mining in patents with BERT. *arXiv preprint arXiv:2101.01039*.
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416-1436. https://doi.org/https://doi.org/10.1016/j.respol.2017.06.006
Watzinger, M., & Schnitzer, M. (2019). Standing on the Shoulders of Science.
Weber, G. M. (2013). Identifying translational science within the triangle of biomedicine. *Journal of Translational Medicine*, *11*(1), 126. https://doi.org/10.1186/1479-5876-11-126