BETA
Bureau
d'économie
théorique
et appliquée

Documents
de travail

«A Novel Matching Algorithm for Academic Patent Paper Pairs: An Exploratory Study of Japan's national research universities and laboratories»

Auteurs

Van-Thien Nguyen, René Carraz

Document de Travail n° 2023 – 29

Septembre 2023

**Bureau d'Économie
Théorique et Appliquée
BETA**

www.beta-economics.fr

@beta_economics

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Université de Strasbourg   cnrs   UNIVERSITÉ DE LORRAINE   INRAE   AgroParisTech

"A Novel Matching Algorithm for Academic Patent Paper Pairs: An Exploratory Study of Japan's national research universities and laboratories"

Van-Thien Nguyen
Toyo University, Department of Global Innovation Studies

René Carraz
Toyo University, Department of Global Innovation Studies
BETA, CNRS, 67000, Strasbourg
carraz@toyo.jp
Corresponding author

ABSTRACT

This paper proposes a new method for matching patents with academic publications to create patent-paper pairs (PPP). These pairs can identify instances where a research result is both applied in a patent and published in a paper. The study focuses on a sample of top research-intensive universities and laboratories in Japan, utilizing a new dataset that contains patent-to-article citations and a machine learning model as part of the matching process. Expert consultations were conducted to enhance the robustness of the methodology. Focusing on a set of 14 Japanese universities and 3 national research laboratories, using patent (USPTO) and publication data (OpenAlex) between 1998 and 2018, we built a dataset of 3,177 PPPs out of 7,766 granted patents and 91,213 publications. The results demonstrate that this phenomenon is widespread in academia and our data show the diversity of the academic disciplines and technical field involved, highlighting the intricate connections between scientific and technical concepts and communities. On the methodological side, we documented in-depth complementary validation techniques to enhance the precision and reliability of our matching algorithm. Using open-source data, our methodology is adaptable to diverse national contexts and can be readily adopted by other research teams investigating similar topics.

Keywords: #Patent Paper Pair #Methodology #Matching algorithm #Academic patent #Japan

JEL Codes: 031, 034, 05

1. INTRODUCTION

The Bayh-Dole Act, which was introduced in the United States in 1980, allowed for US universities to patent and exclusively license their inventions, even if they originated from publicly funded research projects. Similar legislations were subsequently enacted in most European countries (Mowery and Sempat, 2005), Japan (Kneller, 2007) and Asia (Wong, 2011); resulting in an increase in academic patenting, even though other factors influenced this upward trend (Mowery et al. 2001; Sampat, 2006).

While the impact of an increasing reliance on intellectual property rights (IPR) by academic scientists has initiated intense academic and policy discussions (Compagnucci and Spigarelli 2020), we are mostly interested in this paper in exploring instances where knowledge has both scientific value and commercial potential. Stokes (1997) proposed a classification where both objectives can be met through "use-inspired basic research," conducted in "Pasteur's quadrant." Fiona Murray developed a practical approach to implement this concept, utilizing patent-paper pair (PPP) as a means of identifying instances where a discovery/idea is both applied in a patent and published in a paper (Murray, 2002; Murray and Stern, 2007).

Several pairing strategies have been proposed to find PPPs. These approaches have primarily focused on the biotechnology domain and include manual pairing methods used by Ducor (2000), Murray (2002), Murray and Stern (2007), and Martelli and Remo (2019). In biotechnology as well, Magerman et al. (2015) proposed text-mining algorithms to perform the pairing process, while Lissoni et al. (2013) initiated a data-mining approach, which focuses on a set of Italian academic inventors.

In this study, we present a novel method for matching patents with academic publications. Our approach involves three key modifications to existing methods. First, rather than limiting our investigation to a particular field, we have chosen to focus on institutions with shared patenting and publication strategies. Specifically, we selected a sample of top research-intensive universities and laboratories in Japan. Secondly, we utilized a new dataset (Marx and Fuegi, 2020, 2022) that contains patent-to-article citations, which enabled us to identify common citation patterns between patents and papers. This approach is particularly relevant for our study, given that academic scientists rely heavily on citations as a means of allocating credit, and it is deeply ingrained in their social practices (Merton, 1973). Finally, we propose the use of innovative semantic analysis and machine techniques as part of our matching process.

Our model was tested in Japan, a country with a university system that has undergone significant changes in the past two and a half decades, placing a strong emphasis on commercialization, patenting, and technology transfer activities. These changes were initiated by the *Science and Technology Basic Law* in 1995 and the *Incorporation* of National Universities in 2004 (Carraz and Harayama, 2008). Furthermore, Japan’s national government has invested heavily in upgrading its national innovation capabilities with a strong emphasis on upgrading research facilities at national research laboratories and universities (Carraz and Harayama, 2018). Expert consultations were conducted to discuss the patenting and publication processes at national universities and laboratories, thereby enhancing the robustness of our methodology.

In this paper, we present a novel method for building PPPs. As an application we constructed a sample of 3,177 PPPs between 1998 and 2018, focusing on a set of 14 Japanese universities and 3 national research laboratories. The scientific domains of interest include Computer Science, Environmental Science, Engineering, Materials Science, Chemistry, Biology and Physics (classified by the OpenAlex Class 0 list[1]). To create our sample, we first obtained 7,766 USPTO granted patents for all the institutions and matched the inventors of these patents to 91,213 papers using OpenAlex database, resulting in 17,286 potential PPPs. Next, we developed our score system and employed machine learning to develop an indicator model that assesses the likelihood of a match.

The paper proceeds as follows. Section 2 provides a review of the existing literature and proposes several paths to improve the matching process. Section 3 describes the Japanese context. While section 4 outlines the data collection process, Section 5 and Section 6 introduce our scoring system, validation strategies and the PPP identification model used in this study. Section 7 presents the results and provides some descriptive statistics. Finally, Section 8 offers concluding remarks and possibilities for further research.

2. LITERATURE REVIEW

2.1. Patent-Paper Pair (PPP)

Investigating the emergence of scientific and technological innovations has led to a growing interest in the interaction between science and technology. Murray (2002) proposed an original approach for documenting the connections between the scientific and technical community. She introduced the concept of patent-paper pair (PPP). PPPs are formed when a scientific discovery or invention described in a published research paper is at the same time granted as patent, thus indicating the convergence of scientific and technical concepts.

As Murray (2002, 1392) puts it, these two documents can be used for: “a natural experiment because they transcribe the same idea and yet the texts are distinct—a paper describes experimental results, while a patent defines utility and makes claims on inventiveness.” The emergence of this concept is rooted in the recognition that the communication of scientific research findings through publications and the protection of inventions through patents are two separate, but closely related activities that serve different purposes. Nobel Laureate William

[1] The concepts and their position on the list can be found here: https://docs.openalex.org/api-entities/concepts | Accessed 25.02.2023

Schockley's invention of the transistor is a classic example of a PPP. He started to do research on the transistor after the end of the War in 1945 which led to a major discovery in January 1948, he filed for a US patent in June of the same year and published his theory in 1949 (Huang and Murray, 2009; Gertner, 2012).

PPPs provide a valuable resource for examining situations when science and technology have common roots. They demonstrate how scientific and technical concepts and communities are interconnected. By analyzing the overlap between scientific research and innovation, researchers could better understand the factors that drive technological progress and the role of scientific knowledge in that process. PPPs can also provide insights in a variety of subjects addressed in the literature, such as the use of IPR's strategies by academic researchers (Powell and Owen-Smith, 2008), academic contribution to innovation (Kang and Motohashi, 2020), the diffusion and localization of scientific knowledge (Zucker et al. 1998; Bonaccorsi and Daraio, 2005), the impact of patenting on scientific publication (Azoulay et al., 2009), and the potential barriers to the dissemination of scientific knowledge (Walsh et al., 2007; Heller and Eisenberg, 1998). In addition, a PPP presents an opportunity for analyzing two distinct, yet potentially overlapping, citation networks: a technical and a scientific (Jaffe and Rassenfosse, 2019).

2.2. Patent-Paper Pair Generation

One of the challenges in conducting empirical studies on the interaction between scientific research and patenting is to create an appropriate matching of patent applications and scientific publications. Manual matching method has been adopted by many researchers as it is a relatively simple way to link patents to papers (Ducor, 2000; Murray, 2002; Murray & Stern, 2007; Martinelli & Romito, 2019). For example, Murray and Stern (2007) built a dataset of 169 PPPs to test the anti-common hypothesis. They first looked at 340 novel research articles in the journal *Nature Biotechnology* and then found compatible patents in USPTO data based on authors' information, affiliation, and the article-patent content. Martinelli and Romito (2019) followed a three-step process to match 1,652 patents in the field of cancer detection with publications in the *Web of Science* (WoS) database and identified 373 valid results. The matching process involved searching for patent inventors who were also authors, comparing patent and paper publication dates within a two-year limit, and ensuring correspondence of topics by matching patent and paper abstracts and contents using keyword searches. While these methods have the advantage of being comprehensive and reliable, it has some limitations as well. For example, it is possible that an inventor did not contribute to the paper or that the paper was written by someone else who did not file the patent. Additionally, some inventors or authors may use different names in different contexts, which can make linking patents to papers more difficult.

Another approach is to use semi-automated and automated techniques that rely on search engine and text mining algorithms to identify the associations between patents and scientific publications. Those techniques have become increasingly popular due to the large volume of data involved and the need for accuracy and efficiency. For instance, out of a population of 4,270 human gene patents (covering almost 20% of 23,688 known human genes), Huang and Murray (2009) could identify 1,279 human gene PPPs. These pairs were selected by the shared disclosure of a gene sequence in the “gene paper” and in the claims of the “gene patent” (Murray and Stern, 2007; Huang and Murray, 2008). The matching process was therefore limited by the availability of gene sequences as discriminatory elements for the matching process. On the other end of the spectrum, Magerman et al. (2015) proposed a matching process based on the content similarity of titles and abstracts of patents and publications, as derived by text mining algorithms. From a starting database of 88,248 biotech patents and 948,432 biotech publications, they found 645 PPPs. Even if their method seems to be robust and can potentially identify pairs that were missed by the inventor/author-based approach, they are most likely excluding many valid pairs (false negative), as the content of patent and publication abstracts are believed to be different (Myers, 1995).

Until now, the construction of patent-paper pairs has been impeded by common challenges, including the significant amount of time required to perform the matching process, as well as being narrowly tailored to specific fields such as biotechnology. These obstacles have hindered the applicability to more diverse areas of research, the generalizability of their findings, and made it difficult for researchers to establish a comprehensive understanding of the interactions between concomitant scientific research and patenting.

Considering these shared limitations, there is a pressing need for a new and hybrid approach to overcome these challenges. We aim to address these limitations by proposing a novel method that incorporates a scoring system and machine learning techniques to create PPPs accurately and efficiently. Therefore, our study offers a potentially useful tool for discovering the complex interplay between public and private knowledge.

3. JAPANESE CONTEXT

In Japan, university-industry collaboration increased significantly after the introduction of the *Science & Technology Basic Law* in 1995, the *Act on the Promotion of Technology Transfer from Universities to Private Business Operators* (TLO Act) in 1998 which facilitated transparent and contractual transfers of university discoveries, and the Japanese Bayh-Dole Act, part of the *Industrial Revitalization Special Law,* in 1999 (Takenaka, 2005).

In 1999, the *General Law on Administrative Agency* was adopted, as a result most national research institutes were incorporated and became *Independent administrative agencies* (*Dokuritsu Gyôsei Hôjin*) by 2001 and were separated from their regulating ministries. The aim of this reform was to give agencies considerable autonomy in their operations, in how to use their budgets, and taking advantage of the flexibility given to the structures to achieve higher performance (Shiozawa and Ichikawa, 2005).

In April 2004, national universities were also incorporated as independent administrative agencies (Carraz and Harayama, 2008). As a result, national universities gained greater autonomy in managing their intellectual properties (IP). For instance, they can now manage the ownership of their invention, which was seldom the case before the Incorporation, and directly manage their relations with outside partners (Takahashi and Carraz, 2011). As a result, the number of university-owned patents greatly increased since 2004, before patent applications were mostly applied by a university partner with university researchers being usually listed as inventors (Kanama and Okuwada, 2008).

We initiated data collection from 1998 due to the heightened commercialization of university inventions, sparked by the implementation of the TLO Act and the subsequent year's Japanese Bayh-Dole Act. This led to a notable surge in universities' patent ownership. Furthermore, the management of intellectual property has had time to become more professionalized since 1998. In order to have an up-to-date understanding of the current IP practices among national universities and laboratories, we conducted two interviews in November 2022 with IP experts from Tohoku University, a national public university, and RIKEN (Institute of Physical and Chemical Research), a national research laboratory. At Tohoku University, we talked with an IP specialist and a lawyer at the New Industry Creation Hatchery Center, the innovation center of the university that promotes technology transfer and supports academic entrepreneurs. While at RIKEN, we discussed with a former executive director in charge of technology transfer. Each interview lasted for around 60 minutes.

During the interviews, we discussed 3 dimensions: the patent application procedure; the policy of the institutions regarding authorship of a paper and inventorship in a patent; and whether or not they were aware if ‘same’ research being used in both a patent and a paper by researcher within their institution (patent-paper pair).

Regarding the procedure, for national universities and research laboratories, the processes are quite similar. They are described in Figure 1, first an invention disclosure is made by a scientist, second it is evaluated by the IP division of the university and internal committees, then it is sent to a patent attorney and finally it is submitted to the Japan Patent Office. Usually, all selected patents are first applied at the Japanese Office, then depending on cost considerations and marketability they may be applied abroad.

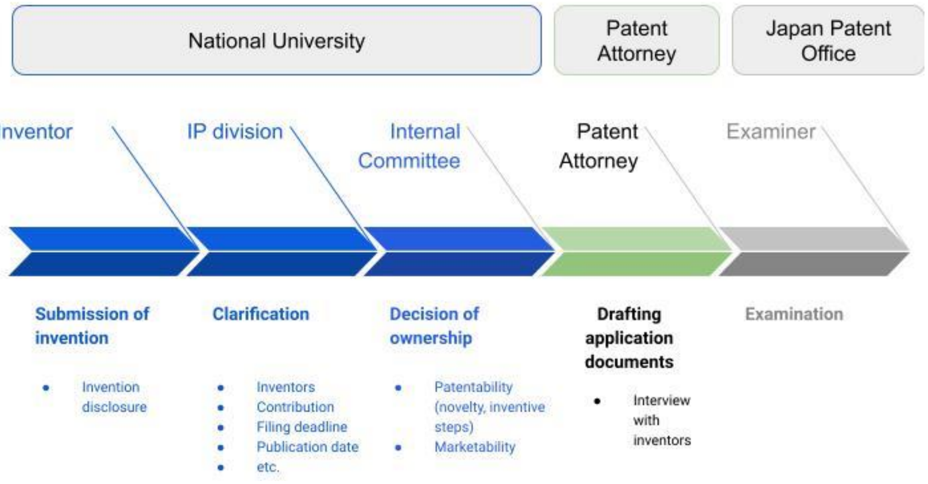
In terms of authorship, Tohoku University follows an internal guideline for the *Appropriate Publication of Research Results*[2], where the scope of authorship is “based on substantial contributions to research” and as a general rule they follow guidelines from the *Japanese Society for the Promotion of Science*, a leading research funding agency in Japan, which stipulate that authorship involves “Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work”[3]. While for inventorship they follow the Japan Patent Office guidelines for inventorship where the inventor as to be an actual contributor for the creation of the invention[4].

[2] http://bureau.tohoku.ac.jp/kenkyo/fb/files/rules/7.pdf | Accessed 21.08.2023

[3] https://www.jsps.go.jp/file/storage/general/j-kousei/data/rinri_e.pdf | Accessed 21.08.2023

[4] https://www.jpo.go.jp/resources/shingikai/sangyo-kouzou/shousai/tokkyo_shoi/document/seisakubukai-06-shiryou/paper07_1.pdf | Accessed 21.08.2023

Figure 1: Patenting process at National Universities



Although we do not claim legal expertise and acknowledge the complexity of the topic. Based on our conversations and analysis of relevant documents, we offer a hypothesis that the inventorship process appears to exhibit a relatively uniform approach across national universities and laboratories, whereas the criteria for inventorship appears to be comparatively more stringent. This observation is consistent with the findings of Lissoni et al.'s (2013) research on Italian inventors.

Regarding the association between patents and publications, the three experts we interviewed suggested that while they were unaware of the extent of the phenomenon, they believed that such an association does exist. As a rule, universities generally advised researchers to prioritize patenting before publication on a discovery, although this practice cannot be strictly enforced. With regard to the contents of abstracts in research papers and patents, the experts noted that differences are likely to arise due to the involvement of patent attorneys in drafting patent abstracts. Additionally, they also highlighted that the level of researcher input during this process may vary depending on factors such as individual circumstances, interest, and experience, creating possible heterogeneity in the process.

As for the selection of our unit of analysis for this on-going research, we decided to select the top 14 universities in Japan in terms of their active foreign patent portfolio; they represent 55.72% of all the applications, with 10,636 active patents (see Annex Table A1). On top of that we choose the top 3 Japanese national research laboratories, in terms of research personnel: RIKEN (Institute of Physical and Chemical Research), AIST (National Institute of Advanced Industrial Science and Technology) and MIMS (National Institute for Materials Science)[5]. 17 research-intensive institutions are included in the sample.

[5] https://www.mext.go.jp/en/about/relatedsites/title01/detail01/1373967.htm & https://www.meti.go.jp/policy/economy/gijutsu_kakushin/4kokuken/index.html | Accessed 25.02.2022

4. DATA COLLECTION

4.1. Potential Patent-Paper Pair Identification

4.1.1. Patent Selection

We acknowledge that academic institutions play a crucial role in fostering technical innovation through patenting and scientific knowledge creation through publishing. They provide the necessary resources and infrastructure to support research and development activities that lead to patentable inventions and scientific publications (Klevorick et al. 1995, Etzkowitz et al., 2000; Shane, 2004; Mowery and Sampat, 2005). Therefore, to guarantee that our analysis accurately captures the trends and patterns in patenting behavior of Japanese institutions, we focus on exploring patents granted to the top 14 universities in Japan and the top 3 research institutions (see Annex Table A1). These institutions were selected because they best represent the diverse range of research conducted in the country, they have different institutional settings (public and private; medical and nonmedical; university and national laboratory) and cover a broad range of scientific fields.

We obtained the patent data from the United States Patent and Trademark Office (USPTO) through *PatentsView*[6], a patent data visualization and analysis platform that offers publicly accessible patent research datasets with detailed documentation. We first extracted patent data by *Assignee* where the associated assignee belongs to one of our target institutions. All possible name variances of these institutions were also considered to ensure that we gathered all relevant patents. Next, we filtered the data by date, covering the period from 1998 to 2018. For each patent, we considered the earliest filing date as the priority date. In cases where the patent was not first filed in the US, we compared the US filing date with the foreign priority date and selected the earlier one. This resulted in a set of 7,766 USPTO patents.

Based on the unique US patent ID, we then retrieved patent information such as patent title, patent abstract, publication date and inventor names. Additionally, we accessed a public dataset provided by Marx and Fuegi (2021, 2022) to obtain information about publications cited in each patent. This dataset contains citations from both U.S. patents (1947-2018) and non-US patents (1782-2018) to scientific articles (1800-2018), comprising approximately 22 million patent citations to scientific literature. The citations were provided in the form of Digital Object Identifier (DOI) numbers, which we use to identify and access the corresponding scientific articles.

[6] For more information: https://patentsview.org/

4.1.2. Publication Selection

The process of selecting relevant publications for our study involved the use of OpenAlex[7] as a primary source. OpenAlex is an online open database that provides access to a comprehensive up-to-date collection of scholarly papers. By leveraging this resource, we aim to identify all publications where one of the inventors of our retrieved patents was a first or last author.

To guarantee that we are only securing publications authored by our list of inventors, we used a combination of author name and authorship institution information to filter our works. We started with a pool of 31,193 unique inventors belonging to our selected patents. We found 14,730 name matches in the OpenAlex database, with a total of 91,222 unique author IDs associated. We denominated these as candidates. This is due to the fact that several distinct authors can have the same name. Then, we queried OpenAlex again to obtain 295,387 works where at least one of the candidates is either first or last author. Next, we filtered out those works where the authorship information does not include any of our whitelisted institutions. This ensures that for each paper, one of the candidates has participated while being associated with one of our 17 institutions of interest. In total, this yielded 91,239 publications that meet our stringent selection criteria.

However, it is important to consider the nature of certain publications in this count. Experts and researchers we consulted noted that review papers pose challenges in our investigation. These papers frequently lack direct relevance to specific research projects and tend to cover a wide range of topics, complicating the direct correlation between a patent and a paper. To circumvent this issue and maintain the integrity of our selection criteria, we implemented an additional step. We excluded 26 papers from our dataset that had the word "review" in their titles. This strategic exclusion helped us to further refine our dataset, ensuring that our analysis is based on direct patent-paper pairs rather than on broader review papers. Consequently, our final dataset was narrowed down to a total of 91,213 publications, providing a more accurate and relevant basis for our analysis.

At this stage, we also identify the related work associated with each of the publications through a function offered by OpenAlex. The system uses an algorithm to find recently published works that share the most common concepts with the original paper[8]. OpenAlex's concept is the modified version of the Microsoft Academic Graph (MGA) classification system, identified by an automated classifier trained on MAG data. To find the related work of a paper, OpenAlex seeks publications that have similar concepts in its pools of more than 65,000 unique concepts. This information is critical in our study as it enables us to later compare it with that of publication in potential matches. Finally, we randomly selected a related work among those provided by OpenAlex. The process helps to avoid bias towards any specific publication and to have a representative sample of related work as our control sample to compare our findings against.

[7] For more information: https://openalex.org/about

[8] For more information: https://docs.openalex.org/api-entities/concepts

4.1.3. Matching Process

Our objective is to pinpoint accurate patent-publication pairs within a defined time frame. We aim to establish the necessary time constraint to facilitate the identification of matches from a pool of 7,766 USPTO patents and 91,213 publications. To verify the eligibility of the patent-publication pairs, we employed a time condition which mandates that the publication date falls between the patent priority date and patent grant date. The time frame considered is such that it satisfies the following requirement:

*(1) - 365 days < Time difference between paper publication's date and patent priority date < 365*2 days.*

This time range has been set to ensure that the identified patent-publication pairs have a significant temporal overlap, and any similarities or commonalities between them can be effectively analyzed. Specifically, it allows for a time lag of up to three years between the patent priority filing date and paper publication date. This is founded on the belief that three years provides sufficient time for the inventors to prepare manuscripts, submit them for publication and get their paper published.

Utilizing the established time range, we were able to identify a total of 17,286 potential matches. These potential matches are based on a set of specific conditions that we have set forth to ensure the validity and accuracy of the matches. These conditions include (1) a match where the paper's author is one of the inventors named in the patent, (2) at least one author's affiliation in the paper is one of our 17 Japanese institutions, and (3) the publication date falls within the proposed time frame. Moreover, we also generate a control sample in the form of PPPs that share similarities but are deemed as non-matches. To achieve this, we employed OpenAlex's *related_works* function, which utilizes an algorithm to identify recently published works that share the most common concepts with the original paper. The rationale behind creating this control sample is to establish a baseline for comparison with potential matches. By doing so, we aim to evaluate the reliability of our scoring system and to test our hypothesis that the scores assigned to the control sample will be comparatively lower than those assigned to potential matches. Additionally, this enables us to scrutinize the behavior of our scoring system and to identify any potential anomalies.

4.2. Labeled Data Acquisition

For the purpose of training and evaluating a machine learning model which is able to identify a match, we create a labeled dataset consisting of 726 PPPs. This dataset consists of two distinct segments: one verified by the researchers themselves and the other manually validated by our team.

The first portion of the dataset was validated by the researchers directly involved in the creation of PPPs. We initiated contact with researchers who were listed as inventors in patents and authors in research papers simultaneously. We achieved this through email by having 90 researchers assess the accuracy of our suggested PPPs. Selected from our potential PPP pool, participants included the 455 researchers with over 3 potential PPPs, with the highest count being 178 for a single researcher. Subsequently, we randomly selected 3 to 20 PPPs for each researcher and inquired via email about their evaluation of them as potential matches. We identified 365 email contacts and sent our request between January and March 2023, including one reminder. Table 1 illustrates the achieved response rate of 24.65%. From this process, we obtained 195 instances labeled as True Matches and 317 instances labeled as False Matches. Instances with missing values were excluded, leaving us with a dataset of 147 True Matches and 259 False Matches.

Table 1: Results for Patent-Paper Pairs Emails and Matching

Researchers with more than 3 potential PPPs	455
Email contacts & sent	365
Responses	90
Response rate	24.65%
Number of PPPs evaluated	512
True matches	195
False matches	317

The second part of our dataset was validated through a meticulous visual validation process which was executed in a series of steps to ensure the accuracy and reliability of the dataset. To ensure the robustness of our findings, we randomly selected 600 pairs from the pool of 17,286 potential PPPs, and did evaluate the adequacy of the match. Firstly, we checked the information related to the individuals involved in both the patent and publication. This included a review of the authors and inventors associated with each document, as well as their respective affiliations and contributions. Next, the title and abstract of both the patent and publication were analyzed to determine if they shared any common themes or topics. This involved a detailed examination of the content in both documents using keywords search, as well as an assessment of the main ideas presented in each. The final and most critical step involved searching for similarities in the images, charts, and diagrams in both documents. If identical graphs or images were found in both documents, we classified them as a PPP with high confidence, despite any potential imperfections. This approach originated from dialogues with some of the 90 surveyed researchers, serving as a pairing strategy. However, the absence of identical graphs in both documents did not automatically disqualify a pair as a PPP. To avoid the risk of mislabeling or overlooking potential PPPs, pairs without identical graphs were not included in the labeled dataset, rather than being labeled as *False Matches*. Through this visual validation process, we identified 216 PPPs, which were subsequently labeled as *True Matches*.

In order to maintain the label balance of our training dataset, it is crucial to incorporate an equal representation of both *True* and *False* instances. This balanced approach is essential to prevent the model from developing a bias towards the majority class, which could lead to inaccurate predictions and undermine the model's performance. Consequently, we supplemented our training dataset with 104 negative matches derived from our control sample of related works. These instances were subsequently labeled as *False Matches*. The inclusion of these negative matches serves to enhance the diversity of the training data, thereby improving the model's ability to generalize from the training data to unseen instances.

The evaluation and labeling process resulted in a dataset comprising 363 *True Matches* and 363 *False Matches*, providing a robust foundation for subsequent analyses. Table 2 provides the descriptive statistics of our labeled data.

Table 2: Summary statistics of a labeled dataset of True matches and False matches

	True Matches	**False Matches**
Mean number of inventors	3.94	4.73
Mean number of authors	4.80	5.26
Patent time range	1998 - 2017	2000 - 2017
Paper time range	1999 - 2018	2001 - 2020
Mean time difference (days)	297.58	830.24
Count	363	363

5. SCORING SYSTEM

To assess the similarity between patents and academic papers, we develop a new matching algorithm that integrates multiple factors such as *inventor_score, semantic_similarity_score*, *word_overlap_score*, and *citation_overlap_score*. Each of these scores provides a unique perspective on the relationship between a given patent and scholarly paper and helps to determine whether they are a pair or not.

The inventor score is an important component of the system as it captures the level of representation of those who are inventor-author in the inventive process. Previous studies have highlighted the importance of considering the involvement of individuals who contribute to creating invention and scientific knowledge when evaluating PPPs. While Huang and Murray (2009) as well as Martinelli and Romito (2019) used a restrictive definition that all inventors had to be authors, Lissoni et al. (2013) adopted a looser condition which only required one of the inventors belong to the author group regardless of the position of the author. Our approach falls somewhere in the middle, striking a balance between being sufficiently flexible to accommodate a range of factors, yet not so permissive as to compromise the integrity of our analysis. As a part of our method, we have set the requirement that the first author or last author must belong to the inventor group. This is because the first author of a research paper is often the person who contributed the most to the research and who is primarily responsible for the content of the paper. Meanwhile, the last author often represents the senior author or principal investigator who supervised the research. In many scientific disciplines, the last author position denotes a leadership role, reflecting significant intellectual contribution and overall responsibility for the research study (Baerlocher et al, 2007; Müller, 2014). This dual focus helps to establish a clear connection between the research and the patent, and ensures that the key contributors to the research and the invention process are properly credited. Based on our proposed requirement, we calculate the inventor score by dividing the number of overlapping inventors and authors by the total number of inventors.

(2) *Inventor_score = (Number of researchers who appear in both patent and paper)/(Total number of inventors in the patent)*

Another component of the scoring system is the *semantic_similarity_score*. It measures the similarity between the title and abstract of patents and papers. The title and abstract of a patent and a paper provide important information about their content and can be used to identify potential matches (Landauer et al., 2007; Magerman et al., 2015). To calculate the score, we leveraged *S-BERT*, a language model trained to obtain meaningful sentence or paragraph embeddings, with the goal of leveraging them in tasks such as semantic similarity. The model determines the semantic similarity between two texts by computing the cosine similarity among the embedding vectors of the titles and abstracts in each PPP.

We chose this approach due to its stronger results than *TF-IDF* (Term Frequency - Inverse Document Frequency) and *GloVe (Global Vectors for Word Representation)* embeddings in semantic similarity benchmark as reported by Reimers and Gurevych (2019), as well as the ease of application thanks to the *Python* framework *SentenceTransformers* provided by these authors.[9]

One powerful property of this method is that the embedding is computed in a context-aware manner: the model considers the meaning of each word within the context of the surrounding words, thus being able to capture nuances in word meaning and how they contribute towards the meaning of the whole text. Since the direction of each word's representation vector symbolizes its semantics, and the angle formed among a pair of vectors can show their closeness, the cosine of the angle between the vectors can symbolize how semantically similar two texts are. Hence, the cosine similarity metric is able to capture semantic similarity among words.

(3) *Semantic_similarity_score = (Semantic similarity score of patent and paper's title + Semantic similarity score of patent and paper's abstract)/2*

The *word_overlap_score* is another measure of the similarity between a patent and a paper based on the number of overlapping words in their respective titles and abstracts. It is worth noting that the *word_overlap_score* may not be a strong signal due to the difference in writing style in patent and publication documents (Myers, 1995). However, we believe that this score can still be useful

[9] For more information: https://www.sbert.net/index.html

in indicating PPP. To calculate the score, we first lemmatized and removed stopwords from both texts. We then calculated the number of overlapping words between the patent and paper and divided it by the total number of words in the patent title and abstract. This score ranges from 0 to 1, with higher scores indicating a higher degree of overlap between the patent and paper.

(4) Word_overlap_score = (Number of overlapping terms in patent and paper's title and abstract)/(Total number of terms in the patent title and abstract)

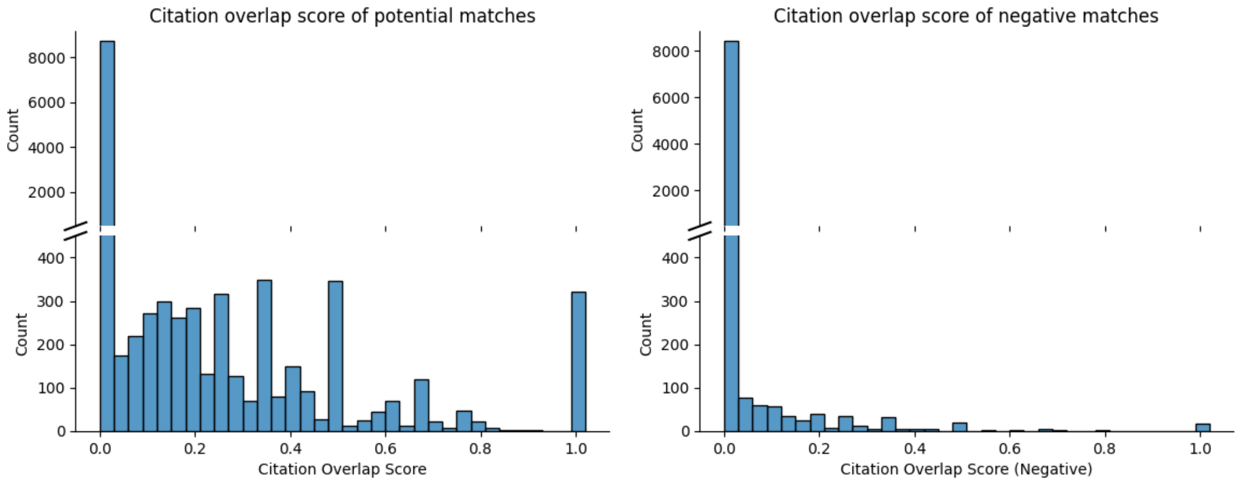
Finally, we consider the overlap of publication references (*citation overlap score*) between the patent and paper as an indicator that evaluates the extent to which a patent and paper references the same previously disclosed scientific knowledge. This is important as it enables us to trace scientific development of each PPP and investigate their connections. Both patents and publications need to establish that the idea they are describing is new and specially for patents, non-obvious. As a result, they often cite prior art; existing publications that describe similar or related inventions as a way of presenting the context of their work or demonstrating the relevance of their findings. Since both patent and publication are trying to establish the novelty of the invention they are describing, PPPs may end up citing some of the same prior art. To the best of our knowledge, no prior study has utilized the citation overlap score for patent-paper matching evaluation. Our approach thus provides a unique perspective on the topic and has the potential to offer valuable insights into the relationship between patents and academic research. The *citation_overlap_score* is then calculated by dividing the number of publication references appearing in both documents by the total number of publication references in the patent.

(5) Citation_overlap_score = (Number of overlapping publication references in patent and paper)/(Total number of publication references in patent)

We aimed to test the validity of this relatively new metric by comparing with a control sample of negative potential matches, which exhibit similarities in the knowledge mobilized but are ultimately considered non-matches. We limited our negative matches to those papers that met our time limit, Equation (1) criteria, and had no common authors with our matching candidate. Subsequently, we calculated scores using Equation (4) for all our matches and this control sample.

The results of our analysis are presented in Figure 2, which compares the citation overlap scores for our matching sample with those of our control sample of related works. Notably, the scores of our potential match sample are significantly higher than those of the control sample in all the distributions. Thus, we hypothesize that the overlap of publication references between a patent and paper within a pair is a strong indicator of a match. Our findings support the utility of the new dataset gathered by Marx and Fuegi (2020, 2022) and contribute to the development of effective matching instruments.

Figure 2: Citation_overlap_score between potential matches and negatives matches (control sample)



6. PPPs IDENTIFICATION MODEL

6.1 Logistic Regression Model

In the next step, we used four features which are the *inventor_score*, *semantic_similarity_score*, *word_overlap_score*, and *citation_overlap_score* as predictors in our model. We believe that these features capture different aspects of the problem and are complementary to each other.

In this research, we aimed to develop a binary logistic regression model for predicting the binary outcome of PPPs. To achieve this, we utilized *scikit-learn*, a popular and efficiently implemented machine learning toolkit in *Python*, with default parameters for binary logistic regression[10] (Hao et al., 2019). Logistic regression is a type of statistical model that predicts the probability of a binary outcome based on one or more predictor variables. We chose a logistic regression model for its simplicity and interpretability. Our input data consisted of PPPs, which were relatively simple and easily separable. Therefore, we did not need to use a more complex or advanced approach.

The logistic function, also known as a type of sigmoid function, is a mathematical function that maps any real-valued number to a value between 0 and 1. This makes it particularly useful for predicting probabilities. In our binary logistic regression model, the logistic function is used to transform a linear combination of predictor variables into a probability that the outcome variable

[10] For more information: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

belongs to the positive class. The formula for binary logistic regression used in our study is as follows:

(6) $P(Y=1|X) = 1 / (1 + e\hat{}-(\beta_0 + \beta_1 * inventor\_score + \beta_2 * citation\_overlap\_score + \beta_3 * semantic\_similarity\_score + \beta_4 * word\_overlap\_score))$

where *P(Y=1|X)* represents the probability of the positive class given the values of the predictor variables which are *inventor_score*, *citation_overlap_score, semantic similarity_score* and w*ord overlap_score*; β_0 is the intercept; and β_1 to β_4 are the coefficients that represent the effect of each predictor variable on the outcome.

6.2 Comparative models

In our research, we utilize our labeled data to establish an equitable testbed for the comparison of various methodologies, including our own linear model, and those proposed by Lissoni et al. (2013), and Magerman et al (2015).

We replicated the method proposed by Lissoni et al. (2013), which involves creating a binary bag of words fitted on the training data, consisting of patent and paper documents (title and abstract) from the training subset of the data. We applied stopword removal as described in their methodology but refrained from stemming or lemmatization. Subsequently, we computed the cosine similarity among the bag-of-word vectors of the training data and determined the similarity score at a predetermined percentile (the 90th percentile or top 10%, as used by Lissoni et al. (2013)). To predict whether a new pair is a match, we vectorized the documents using the same bag-of-words transformation, computed the cosine similarity. The model returns true if it exceeds the threshold value.

We also re-implement the methodology of Magerman et al. (2015), which does not necessitate fitting on any training data, as it employs predefined threshold values. Initially, we removed stop words and applied the *Porter stemmer*. Subsequently, we computed two word overlap metrics:

- *(Number of overlapping words)/(Minimum number of words among patent document and paper document)*
- *(Number of overlapping words)/(Maximum number of words among patent document and paper document)*

For a pair to be considered a match, it must satisfy three conditions: the first score must exceed 0.6, the second score must exceed 0.3, and at least one inventor must be an author. The latter condition is consistently satisfied in our labeled data.

6.3 Training and evaluation methodology

To evaluate the performance of the model on the labeled data, we used stratified 5-fold cross-validation. This means that the data is split into five buckets, out of which 4 are taken for training and the 5th is used for testing model performance. This is done 5 times, each time the testing will be done on a different bucket. Stratification means that the buckets are created with an equal proportion of true and false labels.

Additionally, while creating a test scenario for fair comparison with the models by Lissoni et al. (2013) and Magerman et al. (2015), we used one 80-20 split, and calculated the performance of all three models on the same test split. Since machine learning models perform better when the training data is normalized, we normalized the training and testing splits before each training phase. We used a standard scaler that shifts the training data to a *mean* of 0 and a *standard deviation* of 1. Once fitted on the training data, the scaler normalizes the test data using the same parameters.

We used accuracy, precision, recall and F1 scores to measure model performance. These metrics are derived from the values in the confusion matrix, as illustrated in Figure 3, providing a comprehensive understanding of the model's performance.

Figure 3: Confusion matrix

	Actual	
Prediction	Negative	Positive
Negative	TRUE NEGATIVE (TN)	FALSE NEGATIVE (FN)
Positive	FALSE POSITIVE (FP)	TRUE POSITIVE (TP)

Accuracy measures the proportion of correctly classified instances out of all instances in the dataset, while precision measures the proportion of true positive predictions out of all instances predicted as positive. Recall measures the proportion of true positive predictions out of all instances that are actually positive. F1 score is a metric that combines precision and recall, which provides a balance between the two metrics. Those four parameters are calculated as below:

(7) Accuracy = (Number of TN + TP)/(Number of TN + TP + FN + FP)

(8) Precision = (Number of TP)/(Number of TP + FP)

(9) Recall = (Number of TP)/(Number of TP + FN)

*(10) F1 Score = 2 * (Precision * Recall)/(Precision + Recall)*

7. RESULTS

7.1 Model Results

In this section, we present and evaluate the results obtained after training the model. At the standard prediction threshold of 0.5 our logistic regression model achieved very good performance, Table 3 shows the following indicators: *accuracy* of 0.82, *precision* of 0.8462, *recall* of 0.7872, and *F1 score* of 0.8144. These scores indicate that the data was easily separable, and the model was highly effective at predicting PPPs.

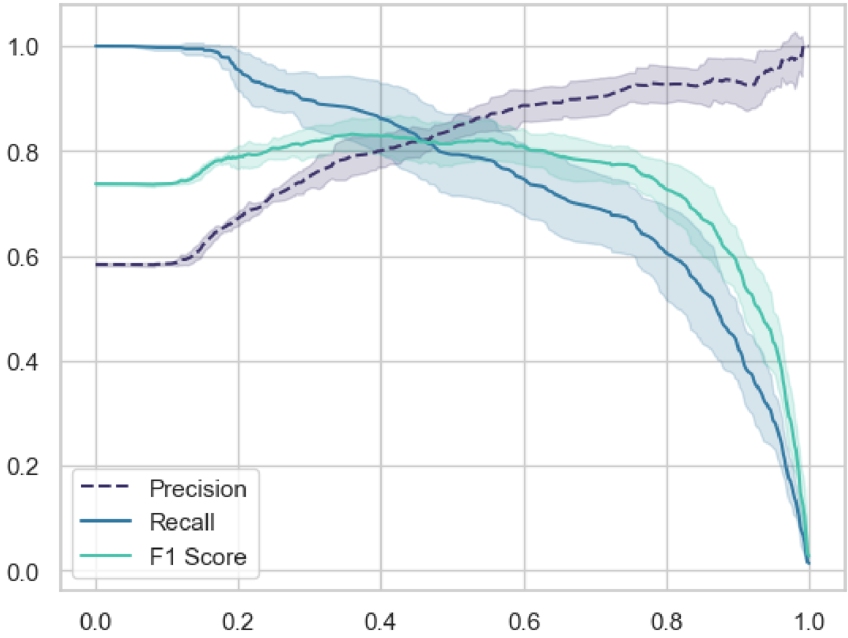
Table 3: Model performance (5-fold evaluation)

Metric	***Value***
Accuracy	0.82 (± 0.02)
Precision	0.8462 (±0.0410)
Recall	0.7872 (±0.0357)
F1 Score	0.8144 (±0.0230)

We then further investigated the precision and recall curves, as illustrated in Figure 4, to comprehend their relationship with the prediction threshold. Our analysis revealed a noteworthy balance between precision and recall over a threshold range from 0 to 1. The F1 score, a harmonic mean of precision and recall, exhibits an optimal value within the threshold interval of 0.2 to 0.6, peaking around 0.4. This precision-recall plot proves instrumental in determining the appropriate threshold for new data point predictions, thereby enhancing the predictive capacity of our model.

In order to further validate our proposed model, we conducted a comparative analysis with the models proposed by Lissoni et al. (2013) and Magerman et al. (2015). The comparative analysis was performed by applying their original parameters to our data set. However, both models yielded almost no positive matches. We hypothesize that this discrepancy may be attributed to several factors including the complexity of our dataset, which make the original authors' thresholds too strict, and the differences in methodology that we could not account for. In order to achieve a comparable positive prediction rate with the aforementioned models, we adjusted the thresholds. Specifically, the percentile in the Lissoni et al’s model was set to 40%, and the minimum and maximum thresholds in the Magerman et al’s model were set to 0.11 and 0.04 respectively, resulting in a positive prediction rate of approximately 47% in both cases.

Figure 4: Precision - Recall Curve



Under these improved conditions, our model still outperformed the models proposed by Lissoni et al. (2013) and Magerman et al. (2015) in key performance metrics, as shown in Table 4. With a precision of 0.7857, our model is more accurate in its positive predictions compared to the other models (0.6811 and 0.7000 respectively). Our model also demonstrates superior recall (0.7971), indicating its effectiveness in identifying positive instances, in contrast to the 0.6811 and 0.7101 recall rates of the Lissoni et al. (2013) and Magerman et al. (2015) models. Lastly, our model yields a higher F1 score (0.7913), signifying a better balance between precision and recall, and thus a more robust overall performance. The superior performance of our model can be attributed to the incorporation of a larger number of features and the model's inherent ability to determine the importance of each feature.

Table 4: Comparative results

Methodology / Metric	**Prediction Rate**	**Precision**	**Recall**	**F1 Score**
Lissoni et al. (2013)	0.4726	0.6811	0.6811	0.6811
Magerman et al. (2015)	0.4794	0.7000	0.7101	0.7050
Van-Thien & Carraz (2023)	0.4794	0.7857	0.7971	0.7913

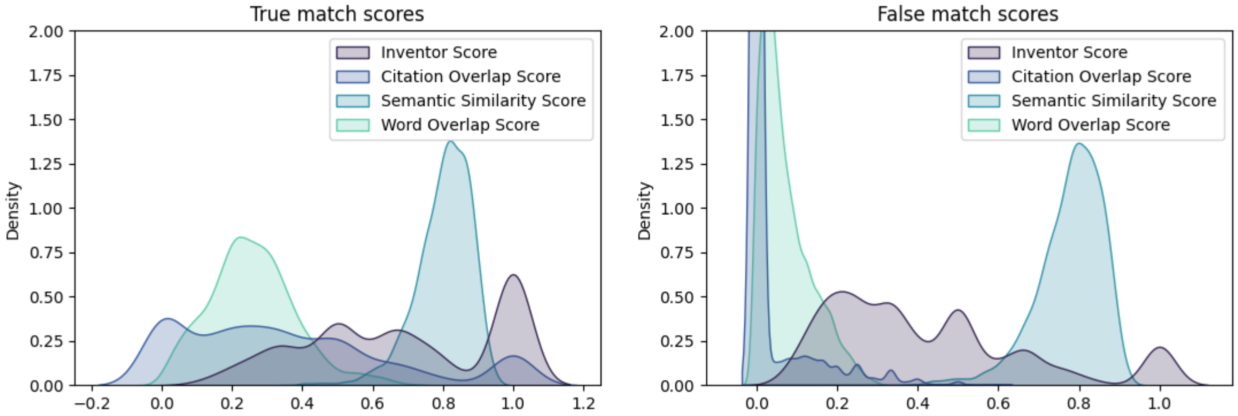
We also examined the model coefficients of one of the models in the cross-validation to identify which features were most important for predicting patent-paper pairs, results are presented in Table 3. The *intercept* for the model was 0.18. Our analysis revealed that *citation_overlap_score* and *word_overlap_score* were the most important features for prediction, as they had high and significant coefficients. The coefficients for *citation_overlap_score* (β_2 = 1.13***) and *word_overlap_score* (β_4 = 1.05***) are both statistically significant at the $p < 0.01$ level, which suggests that they are important predictors of our matching process. The other coefficients are not significant for our current model. While we hypothesize that the reason might be due to our data collection approach, where inventor overlap and semantic similarity are strongly present in both positive and negative matches.

Table 5: Model's coefficient and p-value

	Coefficient	***p-value***
Intercept	0.18	0.240916
Inventor score	0.39**	0.035047
Citation overlap score	1.13***	0.000000
Semantic similarity score	0.10	0.510181
Word overlap score	1.05***	0.000000
Note: Significance levels: *p<0.1, **p<0.05, ***p<0.01		

Because of absent data points, we were only able to analyze 12,627 matches out of 17,286 potential matches. Based on the predictions generated by the model at the standard prediction threshold of 0.5, we identified 3,177 *True matches* and 9,450 *False matches* from 12,627 matches considered. The observed rate of PPP generation was found to be relatively high, reflecting the effectiveness of the approach used. Figure 4 provides a comparison between our true and false matches. For instance, the distribution of *citation_overlap_score* and *word_overlap_score* are noticeably different between the two sets of data, false matches are mostly aggregated around 0, while true matches are more widely spread.

Figure 4: Score comparison between True PPPs and False PPPs



7.2. Descriptive Statistics of the true PPPs

To provide further insights into the characteristics of the identified PPPs, we conducted a descriptive analysis of the statistics associated with the pairs. The results of the analysis are presented in Table 5, which includes various measures such as mean, standard deviation, and range of the score we used for the matching process. Table A2 presents information on the quantity of patents and papers per year between 2002 and 2016.

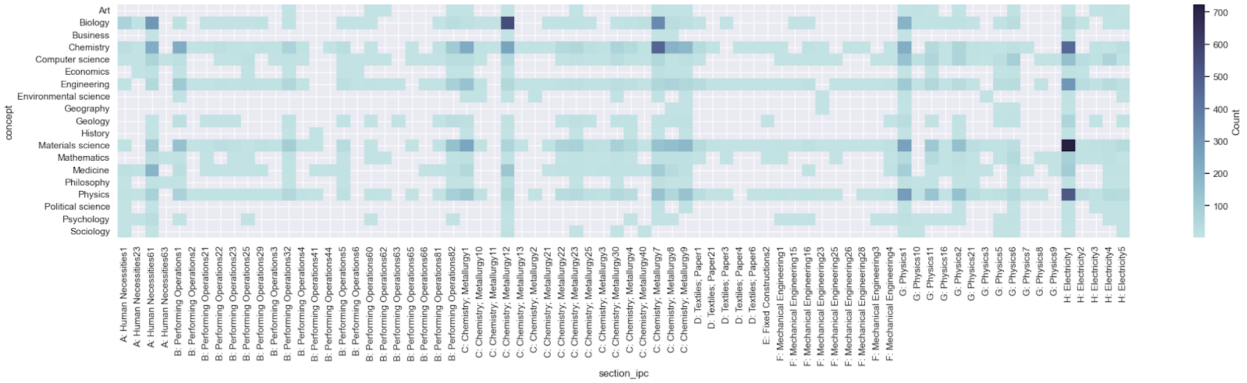
Table 5: Summary statistics of the Patent-Paper Pairs

	Inventor score	***Citation score***	***Semantic Similarity score***	***Word overlap score***	***Time difference (days)***
Count	3177	3177	3177	3177	3177
Mean	0.68	0.35	0.80	0.26	309.26
Median	0.67	0.31	0.81	0.25	280.00
Std. dev.	0.27	0.30	0.08	0.12	199.61
Min	0.08	0.00	0.35	0.00	0.00
Max	1.00	1.00	0.96	0.73	729.00

In terms of affiliations, the proportion of applicants is depicted in Figure A1: AIST constitutes 18.5% of the pairs, followed by Tokyo universities at 11.1%, Tohoku University at 10.3%, and RIKEN at 9.8%. These results show a resemblance in the ranking as compared to the Top 14 Universities (refer to Table A1), with the addition of three national research laboratories.

Previous studies have largely focused on biotechnology, but our research differs by concentrating on a selection of research-intensive institutions in Japan. Figure 5 visually illustrates how combinations of *concepts* and International Patent Classification (IPC) occur. We looked at the *concept* associated with each paper as listed by OpenAlex, each paper had multiple concepts (min= 1 and max=32) and for each patent we looked at the IPC related to the works (min=1 and max=8), it enabled us to create a list of matching *concepts* and IPC 3-digit. The results show that certain combinations are more common than others, such as "Materials science" linked to "H01: Basic Electric Elements" 723 times, "Biology" linked to "C12: Biochemistry; Beer; Spirits; Wine; Vinegar; Microbiology; Enzymology; Mutation Or Genetic Engineering" 551 times, and "Physics" linked to "H01: Basic Electric Elements" 513 times. The top three most commonly used concepts are Chemistry, Materials Science, Physics (3,705 times), and IPC 3-digit codes are H01: Basic Electric Elements, C07: Organic Chemistry, C12: Biochemistry; Beer; Spirits; Wine; Vinegar; Microbiology; Enzymology; Mutation Or Genetic Engineering (2,074 times). This implies that pairings are not only present in biotechnology, but span a broader range of fields warranting further investigation.

Figure 5: Representation of IPC (patents) and Concept (papers) Combinations



8. CONCLUSION

In this study, we have developed a novel approach for the identification of patent-paper pairs (PPPs), specifically within the context of leading research-intensive universities and laboratories in Japan. A prediction model was constructed using a logistic regression algorithm, a choice that, while simple, allowed for a high degree of flexibility and interpretability. The strength of our model lies in its ability to effectively discern and prioritize important features, thus enabling a more accurate matching of patents with corresponding academic papers. By utilizing the model, we constructed a sample of 3,177 (PPPs) spanning from 1998 to 2018 in several scientific domains such as Materials Science, Physics, Chemistry using a logistic regression model. To achieve this, we utilized a new dataset that includes patent-to-article citations (Marx and Fuegi, 2020, 2022), as well as the OpenAlex database, which was created in 2022.

While our methodology represents a significant improvement over existing approaches, several avenues for improving our methodology and the relevance of our results remain. First, we plan to collect more data in order to be able to fully rely on data points validated by experts. This will allow us in the short term to increase the validity of our results and in the long term to train a fully automated matching process based on using patent ID and publication DOI. Second, in the mid-term, we aim to expand the scope of our analysis to include all 1040 higher education institutions and 27 national research institutions in Japan. Third, we focused on the first author and last author in papers for the matching process to limit computational complexity, but we need to explore alternative solutions. Fourth, our results suggest that semantic analysis of abstracts may not provide significant information for our matching process, and we need to investigate this further. Fifth, we believe that spending more time analyzing the academic and technical discipline of the patent-paper pairs would provide a richer understanding of the landscape of the pairs. Finally, the same methodology could be applied to examine other national contexts, opening new avenues for international comparisons.

Moving forward, there are several opportunities for research that could build on the findings presented in this methodological paper. For example, future studies could compare the citation patterns of paired patent and publication to those of non-paired patent and publication from Japanese research institutions. Another potential area for research could be to examine the effect of patents on follow-on research developments, as the change in the citation rate of a paper after a patent is granted would indicate an impact on the diffusion of public knowledge, in other words examining the presence or not of an anti-commons phenomenon. Besides, as most of the studies have so far focused on life science, it would be worthwhile to explore other fields falling within "Pasteur's quadrant." Finally, we could evaluate novelty and disruptiveness measures for our PPPs using the *Novelpy* package (Pelletier and Wirtz, 2022), which incorporates novelty measures from Uzzi et al. (2013) and Shibayama et al. (2021), as well as disruptiveness measures from Wu et al. (2019), and using the methodology developed by Park et al. (2023). In conclusion, the study's proposed robust methodology demonstrates broad applicability across various topics, with PPPs potentially serving as valuable tools for exploring scientific production and innovation.

REFERENCES

Azoulay, P., Ding, W., & Stuart, T. (2009). The impact of academic patenting on the rate, quality and direction of (public) research output. *The Journal of Industrial Economics*, 57(4), 637-676. https://doi.org/10.1111/j.1467-6451.2009.00395.x

Baerlocher, M. O., Newton, M., Gautam, T., Tomlinson, G., & Detsky, A. S. (2007). The meaning of author order in medical research. *Journal of Investigative Medicine*, 55(4), 174-180. https://doi.org/10.2310/6650.2007.06044

c C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics* 63, 87–120 https://doi.org/10.1007/s11192-005-0205-3

Carraz, R. & Harayama, Y. (2008). Japanese university reform seen through bureaucratic reform and changes in patterns of scientific collaboration. In Weber, Luc, E. and Duderstadt, James, J., editors, *The Globalization of Higher Education,* chapter 8, pages 93–107. Economica.

Carraz, R., & Harayama, Y. (2018). Japan's innovation systems at the crossroads: Society 5.0. *Digital Asia*, *13*(12), 33-45.

Compagnucci, L., & Spigarelli, F. (2020). The Third Mission of the university: A systematic literature review on potentials and constraints. *Technological Forecasting and Social Change*, 161, 120284. https://doi.org/10.1016/j.techfore.2020.120284

Ducor, P. (2000). Coauthorship and co-inventorship. *Science*, 289(5481), 873-875. https://doi.org/10.1126/science.289.5481.873

Etzkowitz, H., Webster, A., Gebhardt, C., & Terra, B. R. C. (2000). *The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm*. Research policy, 29(2), 313-330. https://doi.org/10.1016/S0048-7333(99)00069-4

Gertner, J. (2012). *The idea factory: Bell Labs and the great age of American innovation*. Penguin.

Geuna, A., & Muscio, A. (2009). The governance of university knowledge transfer: A critical review of the literature. Minerva, 47, 93-114. https://doi.org/10.1007/s11024-009-9118-2

Geuna, A., & Nesta, L. (2006). University patenting and its effects on academic research: The emerging European evidence. *Research policy*, 35(6), 790-807. https://doi.org/10.1016/j.respol.2006.04.005

Hao, J., & Ho, T. K. (2019). Machine learning made easy: a review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, 44(3), 348-361. https://doi.org/10.3102/1076998619832248

Heller, M. A., & Eisenberg, R. S. (1998). Can patents deter innovation? The anticommons in biomedical research. *Science*, 280(5364), 698-701. https://doi.org/10.1126/science.280.5364.698

Huang, K. G., & Murray, F. E. (2009). Does patent strategy shape the long-run supply of public knowledge? Evidence from human genetics. *Academy of management Journal*, *52*(6), 1193-1221. https://doi.org/10.5465/amj.2009.47084665

Jaffe, A. B., & De Rassenfosse, G. (2019). Patent citation data in social science research: Overview and best practices. *Research handbook on the economics of intellectual property law*, 20-46. https://doi.org/10.1002/asi.23731

Kanama, D. & Okuwada, K. (2008). *A study on university patent portfolios: The impact of intellectual property related policies and the change into corporations of national universities* (In Japanese). Technical Report 154, National Institute of Science and Technology Policy, Tokyo, Japan.

Klevorick, A. K., Levin, R. C., Nelson, R. R., & Winter, S. G. (1995). *On the sources and significance of interindustry differences in technological opportunities.* Research policy, 24(2), 185-205. https://doi.org/10.1016/0048-7333(93)00762-I

Kneller, R. (2007). The beginning of university entrepreneurship in Japan: TLOs and bioventures lead the way. *The Journal of Technology Transfer*, *32*, 435-456. https://doi.org/10.1007/s10961-006-9024-9

Kang, B., & Motohashi, K. (2020). Academic contribution to industrial innovation by funding type. *Scientometrics*, *124*, 169-193. https://doi.org/10.1007/s11192-020-03420-w

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Lawrence Erlbaum Associates Publishers.

Lissoni, F., Montobbio, F., & Zirulia, L. (2013). Inventorship and authorship as attribution rights: An enquiry into the economics of scientific credit. *Journal of Economic Behavior & Organization*, *95*, 49-69. https://doi.org/10.1016/j.jebo.2013.08.016

Magerman, T., Van Looy, B., & Debackere, K. (2015). Does involvement in patenting jeopardize one’s academic footprint? An analysis of patent-paper pairs in biotechnology. *Research Policy*, *44*(9), 1702-1713. https://doi.org/10.1016/j.respol.2015.06.005

Martinelli, A., & Romito, E. (2019). When authors become inventors: an empirical analysis on patent-paper pairs in medical research (No. 2019/32). *LEM Working Paper Series,* Pisa, Italy.

Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, *41*(9), 1572-1594. https://doi.org/10.1002/smj.3145

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. https://doi.org/10.1111/jems.12455

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. University of Chicago press.

Myers, G. (1995). From discovery to invention: The writing and rewriting of two patents. *Social Studies of Science*, 25(1), 57-105. https://doi.org/10.1177/030631295025001004

Mowery, D, Nelson, R, Sampat, B, & Ziedonis, A. (2001). The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh–Dole act of 1980. *Research policy*, *30*(1), 99-119. https://doi.org/10.1016/S0048-7333(99)00100-6

Mowery, D., & Sampat, B. (2005). The Bayh-Dole Act of 1980 and university-industry technology transfer: a model for other OECD governments?. *Essays in Honor of Edwin Mansfield: The Economics of R&D, Innovation, and Technological Change*, 233-245. https://doi.org/10.1007/s10961-004-4361-z

Murray, F. (2002). Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. *Research policy*, *31*(8-9), 1389-1403. https://doi.org/10.1016/S0048-7333(02)00070-7

Murray, F., & Stern, S. (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization*, *63*(4), 648-687. https://doi.org/10.1016/j.jebo.2006.05.017

Müller, R. (2014). Postdoctoral life scientists and supervision work in the contemporary university: A case study of changes in the cultural norms of science. *Minerva*, 52(3), 329-349. https://doi.org/10.1007/s11024-014-9257-y

Myers, G. (1995). From discovery to invention: The writing and rewriting of two patents. *Social Studies of Science*, *25*(1), 57-105. https://doi.org/10.1177/030631295025001004

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*(7942), 138-144. https://doi.org/10.1038/s41586-022-05543-x

Powell, W. W., & Owen-Smith, J. (1998). Universities and the market for intellectual property in the life sciences. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *17*(2), 253-277. https://doi.org/10.1002/(SICI)1520-6688(199821)17:2%3C253::AID-PAM8%3E3.0.CO;2-G

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1908.10084

Sampat, B. (2006). Patenting and US academic research in the 20th century: The world before and after Bayh-Dole. *Research Policy*, *35*(6), 772-789. https://doi.org/10.1016/j.respol.2006.04.009

Shane, S. A. (2004). *Academic entrepreneurship: University spinoffs and wealth creation*. Edward Elgar Publishing.

Shiozawa, B., & Ichikawa, T. (2005). Japan's industrial technology and innovation policies and the effects of "Agencification". *Governance of Innovation Systems*, *2*, 139-76.

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. PloS one, 16(7), e0254034. https://doi.org/10.1371/journal.pone.0254034

Stokes, D. (2011). Pasteur's Quadrant: Basic Science and Technological Innovation. United States: Brookings Institution Press.

Takahashi, M., & Carraz, R. (2011). Academic patenting in Japan: illustration from a leading Japanese university. *Academic Entrepreneurship in Asia. Edward Elgar Publishing, Cheltenham, UK*, 86-107.

Takenaka, T. (2005). Technology licensing and university research in Japan. *International Journal of Intellectual Property-Law, Economy and Management*, *1*(1), 27-36. https://doi.org/10.2321/ijip.1.27

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468-472. DOI: 10.1126/science.1240474

Walsh, J. P., Cohen, W. M., & Cho, C. (2007). Where excludability matters: Material versus intellectual property in academic biomedical research. *Research Policy*, 36(8), 1184-1203. https://doi.org/10.1016/j.respol.2007.04.006

Wong, P. K. (Ed.). (2011). *Academic entrepreneurship in Asia: The role and impact of universities in national innovation systems*. Edward Elgar Publishing

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566(7744), 378-382. https://doi.org/10.1038/s41586-019-0941-9

Zucker, L. G., Darby, M. R., & Armstrong, J. (1998). Geographically localized knowledge: spillovers or markets?. *Economic inquiry*, *36*(1), 65-86. https://doi.org/10.1111/j.1465-7295.1998.tb01696.x

ANNEX

Table A1: Foreign Active Patents: Top 14 Universities in Japan

University	Type	Rank	Foreign Active Patents	Percentage of total
University of Tokyo	National University	1	2,349	12.31%
Kyoto University	National University	2	1,651	8.65%
Tohoku University	National University	3	1,545	8.09%
Osaka University	National University	4	1,404	7.36%
Tokyo Institute of Technology	National University	5	680	3.56%
Kyushu University	National University	6	629	3.30%
Hokkaido University	National University	7	476	2.49%
Nagoya University	National University	8	419	2.20%
University of Tsukuba	National University	9	371	1.94%
Keio University	Private University	10	331	1.73%
Waseda University	Private University	12	273	1.37%
Tokyo Medical and Dental University	National Medical University	17	218	1.06%
Sapporo Medical University	National Medical University	23	182	0.91%
Doshisha University	Private University	40	108	0.54%
Top 14 universities			10,636	55.72%
All Universities			19,088	100%

Source: https://www.mext.go.jp/a_menu/shinkou/sangaku/1413730_00013.htm | Access 23.02.2022

Table A2: Number of Patents and Number of Papers in PPPs by year

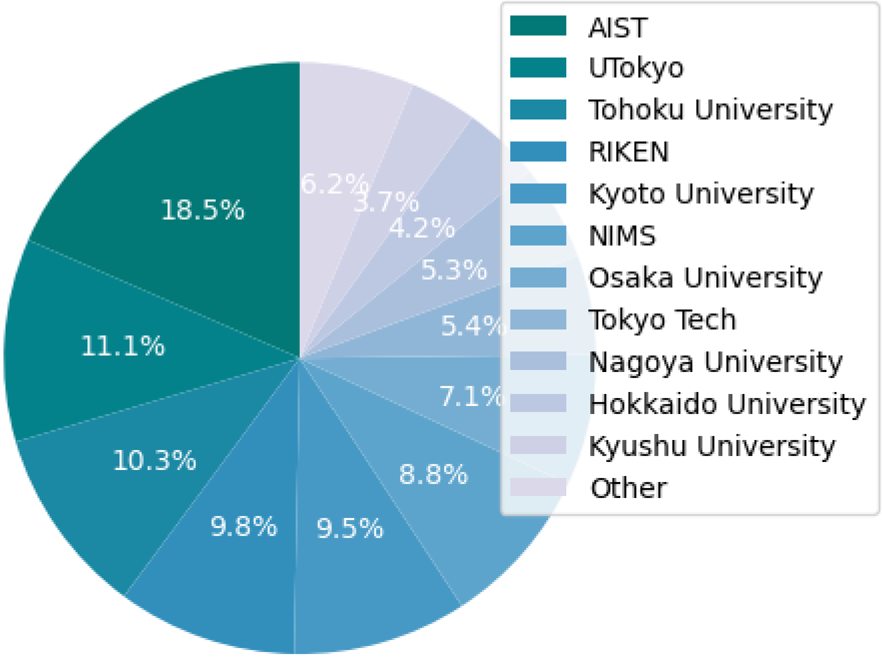
Year	Num. Patents	Num. Papers	Year	Num. Patents	Num. Papers
1997	0	1	2009	215	218
1998	18	6	2010	256	227
1999	25	17	2011	213	245
2000	32	35	2012	225	2a35
2001	74	41	2013	253	225
2002	105	83	2014	202	223
2003	113	105	2015	224	214

2004	152	125	2016	192	213
2005	131	149	2017	83	145
2006	229	190	2018	17	55
2007	216	195	2019	0	18
2008	202	212			

Table A3: Top Concept - IPC Section combination for True PPPs

Concept	**IPC Section**	**Count**
Materials science	H01: Basic Electric Elements	723
Biology	C12: Biochemistry; Beer; Spirits; Wine; Vinegar; Microbiology; Enzymology; Mutation Or Genetic Engineering	551
Physics	H01: Basic Electric Elements	513
Chemistry	C07: Organic Chemistry	472
Chemistry	H01: Basic Electric Elements	456
Biology	C07: Organic Chemistry	344
Engineering	H01: Basic Electric Elements	299
Biology	A61: Medical Or Veterinary Science; Hygiene	298
Physics	G01: Measuring; Testing	279
Materials science	G01: Measuring; Testing	265

Figure A1: Institutions Representation in PPPs


AIST
UTokyo
Tohoku University
RIKEN
Kyoto University
NIMS
Osaka University
Tokyo Tech
Nagoya University
Hokkaido University
Kyushu University
Other
18.5%
11.1%
10.3%
9.8%
9.5%
8.8%
7.1%
5.4%
5.3%
4.2%
3.7%
6.2%