Discussion Paper Series – CRC TR 224

Discussion Paper No. 656
Project B 02

Looking for Innovation Beyond the Patent System: Evidence from Research Disclosures

Bernhard Ganglmair[1]
Alexander Kann[2]

February 2025

[1]University of Mannheim, ZEW Mannheim, b.ganglmair@gmail.com
[2]University of Mannheim, ZEW Mannheim, alex.kann14@googlemail.com

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Looking for Innovation Beyond the Patent System: Evidence from Research Disclosures*

Bernhard Ganglmair† Alexander Kann‡

February 18, 2025

Abstract

We study the content, novelty, and value of defensive publications relative to patents. We use a large language model (LLM) to apply the cooperative patent classification (CPC) system to a set of defensive publications (from 1962 to 2022) from the journal Research Disclosure, thus mapping such research disclosures and patents into a common space and allowing for a direct evaluation of textual similarities between these two types of R&D outputs. We find that while in some technologies, patents and research disclosures follow similar aggregate trends, some exhibit diverging developments over time. We also document shifts in the position of research disclosures in the patenting space that are indicative of changes in the technological landscape not captured in patents. We further show that substantial numbers of research disclosures are published before their closest patents are filed, and many contain terminology before it is first used in patents. Last, we find that in several technology areas, research disclosures have evolved from being an outlet for niche results to a vehicle to publicize technological developments of high practical relevance and value. Our results imply that when we draw conclusions about the nature of technological progress or the direction of innovation based solely on patent data, we obtain an incomplete picture.

Keywords: defensive publications, disclosure, open innovation, patents, R&D, text-as-data

JEL classification: C81, O32, O34, O36

*We thank Bruno Cassiman, Laurie Ciaramella, Pete Klenow, Tim Martens, and the participants at the NBER I3 Working Group meeting in Boston, Research on Innovation, Science and Entrepreneurship Workshop (RISE), and ZEW/MaCCI Conference on the Economics of Innovation and Patenting for comments and suggestions. We also thank the staff at the British Library in London, UK, and the Technical Information Library in Hannover, Germany, for their hospitality and patience during our visits. We thank Simran Bhurat for her excellent research assistance. Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project B02) and the NBER Innovation Information Initiative (A. Kann) is gratefully acknowledged. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

†University of Mannheim and ZEW Mannheim. E-mail: b.ganglmair@gmail.com (corresponding author)

‡University of Mannheim and ZEW Mannheim. E-mail: alex.kann14@googlemail.com.

1 Introduction

When inventors develop new products or processes, they will inevitably face a decision of whether and how to reveal their inventions to the public and maybe even protect them from unlicensed use. Secrecy is one possibility. If that is not an option (maybe because things that are very visible are difficult to keep secret), then disclosure is the answer. An overly simplistic view is that an inventor then has the choice between patenting their invention or publishing it in the form of a defensive publication (without any claims to formal intellectual property).[1] An inventor will forego patent protection and simply publish when the invention is not patentable or novel (and therefore not eligible for patenting) or when it is too insignificant to warrant the costs of a patent application and subsequent patent prosecution.

The simplicity of this view is nauseating and yet powerful. Even without strategic considerations that often lie behind patenting (Hall and Ziedonis, 2001; Noel and Schankerman, 2013), the inventor's decision problem (and its solution) highlights the fact that patents may not be representative of the overall innovation landscape but rather cover a very specific region of the technology space. When we draw conclusions about the efficacy of policy in fostering technological progress or influencing the direction of innovation based solely on patent data, we may miss something. What do we miss, though? And how representative or non-representative are patents (as the preferred output of innovation for many in the literature) relative to other output domains?

In this paper, we focus on a specific form of defensive publications: *Research Disclosure* has served as an outlet for individuals and firms to publish new inventions since the early 1960s. These disclosures have long played an important role in the patent system: patent offices use the publication for the prior-art search (e.g., in 2001, it was added to the PCT Minimum Documentation standard). This means that when publishing in Research Disclosure, the invention enters the prior art (and patent offices ought to know of it), preventing others (potentially rivals) from patenting the same invention or raising the bar for patenting their own (Parchomovsky, 2000; Baker and Mezzetti, 2005; Bar, 2006). On top of their defensive role, these publications also facilitate the sharing of research findings and foster innovation within the wider community (potentially inducing spillovers). Research Disclosure is one among several outlets that serve inventors as a vehicle for effective defensive publications. For instance, the *IBM Technical Disclosure Bulletin* or the *Xerox Disclosure Journal*, primarily for their own employees' use, used to serve a similar purpose, and both outlets were made available to patent offices.

Following our simplistic argument from earlier, inventions disclosed in the outlet *Research Disclosure* (we refer to them as *lowercase* research disclosures) are different

[1] Academic publications are yet another outlet for new technological developments. They are outside the scope of this paper. There exists an active literature (e.g., Gans et al., 2017) that studies the disclosure of scientific knowledge via publications, patents, or both.

from patents (containing inventions that are not patentable subject matter). Moreover, research disclosures are laggards because inventions that are not novel (because of existing prior art) cannot be patented and are more likely published as research disclosures. Instead, novel or groundbreaking developments are patented. And last, we can expect research disclosures to be of lower significance and value. If an invention is valuable, its developer will find it profitable to seek patent protection to commercialize it and prevent unlicensed use

We use the texts of research disclosures to address the validity of these predictions. We use machine learning and text-as-data techniques to classify research disclosures using the Cooperative Patent Classification (CPC) system for patent classification. Our approach is meant to assign to a research disclosure the patent classification it would have received had it been filed as a patent application. We thus map research disclosures and patents into a common space, allowing us to evaluate textual similarities between these two types of R&D publications.

Assuming that research disclosures and patents that are textually similar also share key characteristics such as content, novelty, and value, we can "transfer" characteristics from patents (which we can measure) to their associated research disclosures (for which some of these characteristics are not available). This approach allows us to see if research disclosures are inherently different from patents and if they follow different trends in volume and content. The approach provides us with the means to identify research disclosures with novel ideas and concepts (before they appear in patents) and of high value (which we would otherwise expect in patents).

We document three main findings. First, while in some technologies, patents and research disclosures follow similar trends, some have seen diverging developments over time, specifically CPC sections B, F, and E). In all three sections, patenting has lost prominence, whereas relatively more research disclosures have been made in these technological areas. This result highlights both the dynamic landscape of innovation and how different measures of R&D output capture different aspects of this dynamism. We also document that research disclosures in the broader physics area (Section G) have been diverging from patents and started to occupy different (and novel) regions in the technology space. Shifts like these are indicative of changes in the technological landscape and an evolving nature of innovation within the affected areas that are not captured in patents.

Second, we find a significant level of technological leadership in research disclosures. Substantial numbers of research disclosures are published before their (textually) closest patents are filed, and many contain terminology before it is first used in patents (Arts et al., 2021). This is not to say that the affected patents should not have been granted. What our results rather reveal is that broader ideas and concepts are not necessarily novel when they enter the patenting space.

Last, we show evidence for high-value research disclosures across different technologies, highlighting the importance of research disclosures as a means to freely publish even high-value technological developments. Research disclosures are not just ideas and inventions that are too insignificant to warrant the costs of a patent application and subsequent prosecution. Instead, firms and inventors often decide to forego patent protection and disclose their developments for everyone to see and use. In the broader areas of physics and electricity, for instance, we find that research disclosures have evolved from being an outlet for niche results to a vehicle to publicize technological developments of practical relevance and value.

Our paper contributes to the literature on defensive publications as an IP management strategy. The strategy is widely used: Adams and Henson-Apollonio (2002) provide guidance for practitioners. Henkel and Lernbecher (2008) find (using 56 in-depth interviews with German industrial firms) that seven out of ten companies defensive publications for up to one-third of their inventions. Johnson (2014) finds that defensive publishing has become more common, particularly in response to concerns about low-quality patents in software and business methods. He argues that it is a useful strategy even for firms with patentable innovations, especially those that are less technically challenging and easier to innovate around. Using antitrust litigation against IBM and Xerox, Bhaskarabhatla and Pennings (2014) find that firms switch to more defensive publications when the costs of uncertain antitrust enforcement increase. Building theory models, Bar (2006) and Baker and Mezzetti (2005) study defensive publications in the context of R&D races. We contribute to this literature by providing a detailed empirical account of the content, novelty, and value of firms' defensive publications.

Another strand of the literature studies the incentives of rivals to exchange information without formal protection. This behavior is observed in numerous settings: von Hippel (1987) and Schrader (1991) report empirical evidence of know-how sharing of competing firms in the steel minimill industry, Bouty (2000), Häussler (2011), and Häussler et al. (2014) present results for knowledge sharing in academic research, Gächter et al. (2010) (modeling knowledge sharing as a coordination game with multiple equilibria) present experimental results for a setting of private-collective innovation (see von Hippel and von Krogh (2006)) in which private investors fund public goods innovation, and Ganglmair et al. (2020) offer experimental evidence of information sharing using a financial investor framing. Our results on the factors that induce inventors to publish research disclosures anonymously add to this literature by highlighting the importance of the positioning of the information-to-disclosed relative to the existing stock of knowledge.

Last, our text-as-data approach further contributes to the literature on patent classification. For instance, DeepPatent by Li et al. (2018) is a deep learning algorithm for patent classification that uses neural networks and word vector embeddings; PatentBERT (Lee and Hsiang, 2019a) utilizes a pre-trained BERT model (Devlin et al., 2018a) and

applies fine-tuning techniques for patent classification; And "Bert for Patents" (Srebrovic and Yonamine, 2020a) is a state-of-the-art model trained exclusively on the texts 100M+ patents. We add to this patent-centered literature by building on build on adversarial methods popularized by Goodfellow et al. (2020). Domain adaptation techniques (e.g., Ganin et al., 2016; Rozantsev et al., 2018; Kang et al., 2019) allow us to use patent texts as a labeled training sample (i.e., source domain) and research disclosures as our target domain. Our methodology to find technology classes for research disclosures has many applications; essentially, we can predict technology classes for every form of text that comes with an abstract or for which we can generate an abstract (e.g., academic publications).

The remainder of this paper is structured as follows: In Section 2, we provide a brief history of the Research Disclosure, the publication. In Section 3, we describe our methodology of applying the patent classification system to research disclosures using the Cooperative Patent Classification (CPC) system. In Section 4, we compare trends in research disclosures and patenting, document novelty in research disclosures, and construct a measure for value. In Section 5, we examine strategic motives for anonymous disclosures. We conclude in Section 6.

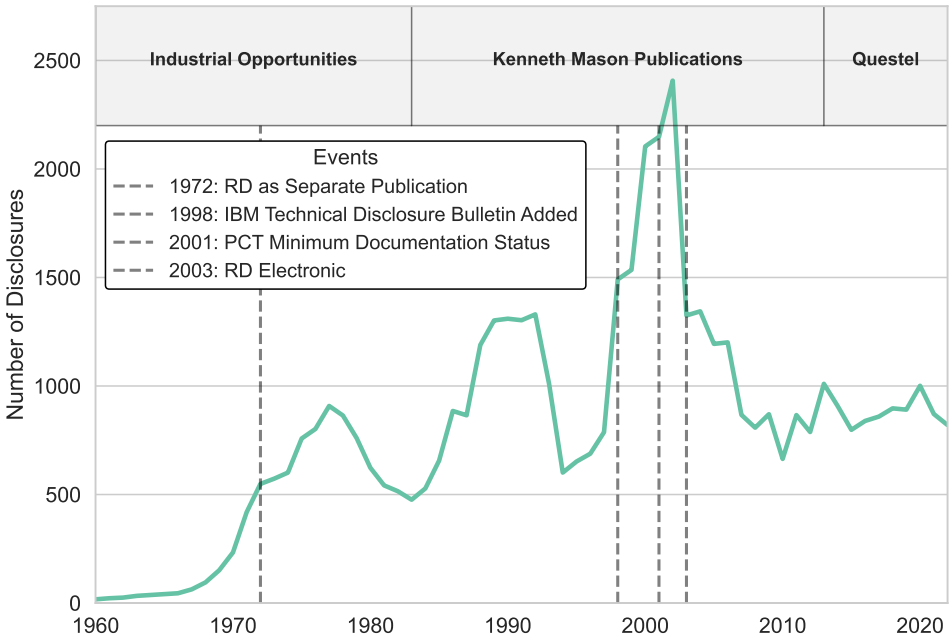
2 Research Disclosures

Research Disclosure (RD) is a monthly publication that has been publishing defensive disclosures (research disclosures) since 1960 and is currently published by Questel[2]. Firms and individuals pay to disclose their innovation or research in RD and the publisher sends each issue to patent offices all over the world (97 patent offices in 1965 and 130 in 2000). While the published innovation is not protected (i.e. can be used by anyone), it is promised that the research disclosure is seen by patent offices as prior art, which implies that no one else should be able to patent the same innovation. A previous publisher claimed that "Companies that publish with us know that they can rely on their disclosures being found by all the patent examining authorities" and "90% of the world's leading companies have published disclosures in RD" in August 2010. Today RD comes in the form of a physical copy and a database to which patent offices have free access. Disclosures can be simply made by uploading a .docx or .pdf file at https://www.researchdisclosure.com/Step/FileUpload and paying the disclosure fee. In April 1968 the first anonymous research disclosure was published and it has been possible to disclose anonymously—at least—since then.

Research Disclosure (RD) started as part of the Product Licensing Index before 1972, becoming its own publication afterward. In 1998, RD began including the IBM Technical

[2]A company that provides intellectual property (IP) solutions, including software and services for patent, trademark, design, and domain name management.

Figure 1: Number of Research Disclosures



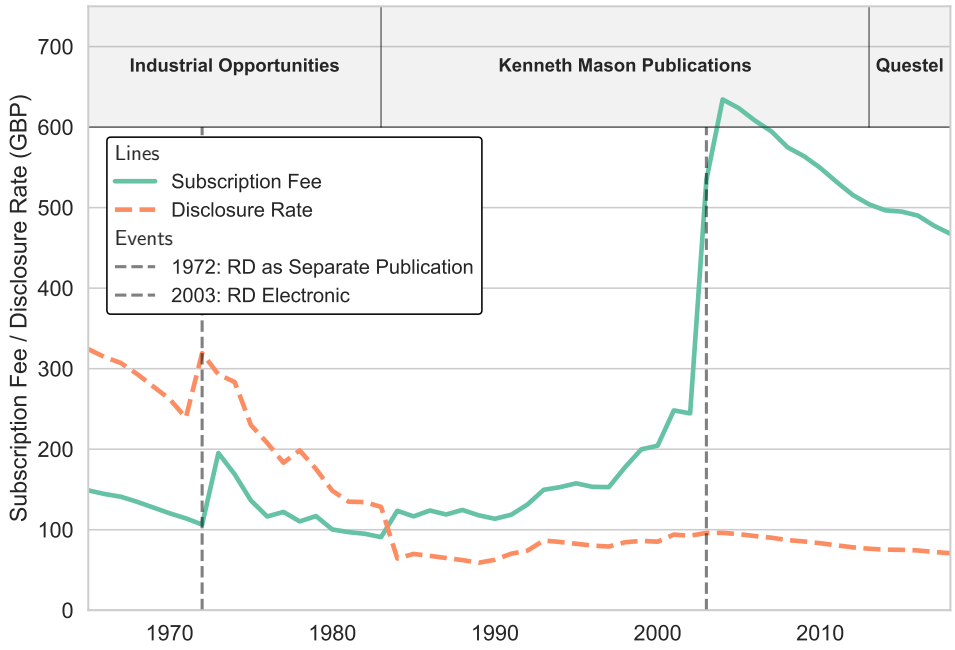
Notes: The figure depicts the annual number of research disclosures for the full sample period 1962–2022. The vertical lines mark major events in the publication's history: 1972 (Research Disclosure became its own separate publication), 1998 (the IBM Technical Bulletin was merged into Research Disclosure), 2001 (Research Disclosure achieved PCT minimum standard status), and 2003 (the electronic version of Research Disclosure was launched). The three publishing companies are listed at the top. *Source:* Research Disclosure Database (Questel).

Bulletin, which was a key reason it was added to the PCT Minimum Documentation Standard in 2001. Inclusion in the standard means that patent offices must check RD when they search for prior art. Starting in 2003, research disclosures were made available online, including historical ones, by creating a digital database. Research Disclosure changed its publisher three times: from Industrial Opportunities Inc. in 1960 to Kenneth Mason Publications Inc. in 1983 and then to Questel Ireland Inc. in 2013.

Figure 1 shows the changes in the number of research disclosures published in RD since 1960. There was an increase in disclosures around 1970, just before RD became a separate publication. The highest number of research disclosures were published between 1997 and 2004, after RD included the IBM Technical Bulletin and was added to the PCT standard in 2001. After reaching this high point, the number of research disclosures returned to the levels seen before 1998.

Research Disclosure (RD) relies on two main revenue streams: a subscription fee, which is the cost for receiving new issues of RD, and a disclosure fee, charged to inventors for publishing their work in RD. Notably, patent offices are not required to pay the subscription fee. The total cost for a disclosure depends on its length, with additional charges applied for including graphs and figures.. Figure 2 illustrates the trend in subscription fees over time.

Figure 2: Publication and Subscription Fees



Notes: The figure depicts the publication fee (orange; normalized to 600 words without figures) and the annual subscription fee (green) for Research Disclosure. Fees are in real British pounds (UK, July 2005=100). *Source:* archival collection for issues 015 (1965) to 656 (1998) (at British Library, London, UK, and Technical Information Library, Hannover, Germany).

After RD became an independent publication, the subscription fee saw its first increase, gradually rising further once Kenneth Mason Publications took over publishing responsibilities. The most significant increase in price occurred in 2003, coinciding with the launch of RD Electronic, which led to the subscription fee for the physical copy more than doubling. The initial cost for accessing RD Electronic was set at 1200 GBP (1578 GBP when adjusting for inflation).

Regarding disclosure fees, normalized for a 600-word submission without figures, there was a marked decrease from 1965 to 1995, except for a brief uptick in 1975 following RD's transition to a separate publication. Post-1985, the disclosure fee experienced a slight increase from 75 to 100 GBP in real terms. However, the correlation between the number of disclosures, as shown in Figure 1, and the disclosure fee post-1985 appears minimal.

3 Patent Classification for Research Disclosures

To empirically analyze research disclosures, it is crucial to find out what they are about and, in particular, what kind of technologies or innovations are disclosed in them. The problem we face is that research disclosures do not come with a technology classification. However, technology classes assigned by patent offices are available for patents.

Since patents and research disclosures are similar in the sense that they both disclose innovations, we develop a model using Natural Language Processing (NLP) to infer the technology classes of research disclosures based on patents..

With our methodology, we aim to answer the following question: If a research disclosure would have been a patent, what would be the technology class of that hypothetical patent? To address this, we process patent and research disclosure texts with a model that has two objectives. The first objective is to accurately predict technology classes from patent texts; a *text-classification* task. The second objective is to ensure that the technology classification learned from patent texts is useful in learning the technology classes of research disclosures; a *domain adaptation* task. Combining these two objectives results in our model, which is capable of predicting the technology classes of research disclosures.

In this section, we first describe the technology classification task and the language model we are using. Next, we describe the domain adaptation and the challenges our model faces. We proceed by illustrating out pipeline by following an example through it. Lastly, we evaluate the classification performance.

3.1 Technology Classification with BERT

As a basis for our model, we use the large language model BERT (Bidirectional Encoder Representations from Transformers) developed by Devlin et al. (2018b). Using a large language model ensures that we efficiently process and represent the textual data, in particular each word and document is represented by an numerical embedding. BERT is built on the attention algorithm popularized by Vaswani et al. (2017), which allows the model to take context into account. For example, the "bus" is processed differently if it appears in the the context of transportation, as opposed to computer architecture. Another crucial feature of BERT is that it is pre-trained on a vast amount of text before it is fine-tuned to perform a specific task. The pre-training ensures that the model already "understands" text before being adapted to perform a specific task. For example, it knows that "bus" and "train" are modes of transportation before being trained on a classification task.

BBERT has been widely used to classify texts; for example, Araci (2019) use the model to predict sentiment in financial texts. In the classification we are interested in, i.e. predicting the technology classes (CPC classes) of patents, BERT has also proven to be useful. The current state-of-the-art models are PatentBERT (Lee and Hsiang, 2019b) and PatentSBERTa (Bekamiri et al., 2021). Our approach to classifying patents is similar to Lee and Hsiang (2019b) with the main difference being that we use "BERT for patents" (Srebrovic and Yonamine, 2020b), a version of BERT specifically pre-trained to understand patent texts, as the basis.

3.2 Domain Adaptation

In our application, we are ultimately not interested in predicting the technology classes of patents. Instead, we want to predict the technology classes of research disclosures by transferring knowledge about the technology classification from patents to research disclosures. This process, known as *domain adaptation*—a sub-field of transfer learning—attempts to optimize the transfer of knowledge from a source domain (i.e., patents) to a target domain (i.e., research disclosures).[3] The source domain and target domain are different but share some crucial features. Domain adaptation ensures that the task of interest (i.e., CPC classification) is performed as well as possible with only the features shared by both domains. In our application, this means that the CPC classification should use the textual features that are present in both research disclosures and patents.

To achieve this, we add domain adaptation techniques to the prediction of patent technology classes with BERT. The underlying idea of these techniques is to train the model to excel at classifying patents while being poor at differentiating between source and target domains. In particular, the model is punished for identifying differences between domains. In our model, we implement the methodology described by Ganin et al. (2016) in their Domain Adversarial Neural Network (DANN). The idea is to add a second classification task when training the model. In the first classification task, the model tries to predict the CPC classification of patents, while the second classification task is to predict the domain, i.e. whether text belongs to a patent or an research disclosure. During training, the model aims to enhance the performance of the technology classifier while simultaneously reducing the performance of the domain classifier.". Ultimately, the model is capable of predicting technology classes but incapable of distinguishing patents and research disclosures.

In addition to providing technology classes—in the form of CPC (sub-)classes—for research disclosure our pipeline also maps research disclosures and patents into a common technology space, which enables us to evaluate technological similarities between research disclosures and patents. It also provides us with the patent-disclosure siblings.

3.2.1 Source and Target Domain

Source Domain: USPTO Utility Patent: We obtain the titles and abstracts of USPTO utility patents, granted between 1976 and 2022, from PatentsView (at https://patentsview.org/). We use the filing date of the respective patent to determine the date of disclosure. The total number of patents in our source domain sample is 6,470,704.[4]

[3] Farahani et al. (2021) provide a brief overview of domain adaptation.

[4] We lose some patents in data preparation. Around 800,000 patents are lost when merging different datasets from PatentView, including 2,256 with application date and 815,666 with CPC classification data. We also remove duplicates based on patent title, CPC group, and abstract (714,355), abstracts

Target Domain: Research Disclosures: We obtain the research disclosures from Research Disclosure Database (Questel), which includes the full text of research disclosures published between 1960 and 2022. The total number of disclosures in our target domain sample is 50,647.[5]

3.2.2 Challenges

In our application, we face several challenges due to the inherent differences between both domains. Patents and research disclosures differ in content, structure, and purpose. These differences pose unique obstacles to effective knowledge transfer.[6]

A major challenge is the heterogeneity within our text corpus of research disclosures. They vary greatly in length; some are very short, consisting of only a few sentences, while others extend over multiple pages.[7] This variation complicates the task of domain adaptation because our approach to the transfer of knowledge about CPC classes must account for disclosures of differing lengths. Another significant hurdle is identifying the patents that closely resemble research disclosures. Patent abstracts follow formatting rules (where the patent examination process ensures compliance), whereas research disclosures have no pre-defined format or structure. For our classification to work, the textual features of research disclosures (such as length and writing style) should be as uniform and similar to those of patents as possible. Lastly, long texts create problems because the BERT model employed in later stages of our pipeline is limited to processing texts of no more than 300 words.[8]

3.3 Our Pipeline

Our pipeline consists of three steps. First, we homogenize the text corpus of research disclosures. Then, we build a suitable dataset for training, validating, and testing. Last, we train a domain-adaptation model. To illustrate our pipeline, we use a practical example and trace the journey of a single research disclosure through our pipeline:

Example Disclosure (Title and Text)**.** *"Barrier layer in the metallisation of semiconductor diode lasers"*

that repeat legal requirements (126), abstracts longer than 400 words (1,296), and those shorter than 15 words (22,079).

[5]We have 50,778 research disclosures in the original dataset. For 1,848 non-English research disclosures (1,008 German, 351 French, among others), we use EasyNMT (https://github.com/UKPLab/EasyNMT) for translation. For 131 research disclosures, the translation does not work.

[6]This discussion focuses on textual and structural features of our corpus. Patents and research disclosures also differ in their potential content, as patents are subject to patentability constraints, while research disclosures can contain anything the disclosing party chooses to include in the document.

[7]The mean length of a research disclosure is 1,728 words, the median is 401 words, and the 25th and 75th percentiles are 239 and 774 words.

[8]While recent alternatives capable of handling longer texts exist (e.g., Beltagy et al., 2020), their computational demands are significantly (if not prohibitively) higher.

"Semiconductor diode lasers of e.g. III-V materials and emitting in the visible wavelength region, very attractive light sources for optoelectronic applications [...] Upside-down mounting of such a laser on a carrier improves its cooling, and thus its high-power performance. This is shown in Fig. 1 [...] the wetting of the laser 10 by the solder layer 7 is very often poor. This results in locally poor cooling of the laser 10, which is particularly detrimental if the poor wetting is near the mirror surface 6 [...] very good and homogeneous cooling [...]."

Our example research disclosure was made anonymously and published in April 1994 and is cited by U.S. patents as non-patent literature.

3.3.1 Write Abstracts for Disclosures

For the first step of our domain adaptation pipeline, we use the Large Language Model (LLM) Llama2 70B (Touvron et al., 2023) to generate abstracts for research disclosures. The goal is to convert research disclosures into a text corpus that matches the style of patent abstracts. We design a prompt for Llama2 to yield clear and concise abstracts that adhere to specific constraints, such as a 120-word limit and the exclusion of figures, numbers, and grammatical errors.[9] This step homogenizes the research disclosures by reducing the variance in text representing them. It produces abstracts that are directly comparable to patent abstracts by ensuring that both share similar lengths and writing styles. The generated abstracts are also short enough to be processed by the BERT model, which we employ in the later stages of our pipeline.

The text of our example research disclosure highlights the need for generating abstracts. The original text is too long (411 words), introduces noise by referring to figures that our model cannot process, and differs in structure and style from patent abstracts. The abstract we generate is a condensed representation of the research disclosure and eliminates computational issues related to length and figure references. It closely resembles a typical patent abstract.

[9] We use the following prompt:

"You are an expert on {title} and similar inventions. You are asked to write an abstract of the following invention. – The abstract mimics the style of a patent abstract.

- The abstract is clear and concise.
- The abstract cannot exceed 120 words.
- The abstract cannot mention any figures or images.
- The abstract cannot include numbers.
- The abstract cannot refer to itself.
- The abstract cannot include grammatical errors.
- The output only includes the abstract

Title: {title}; Invention: {text}; Abstract:"

Example Disclosure (Abstract). *"A barrier layer in the metallization of semiconductor diode lasers enhances high-power performance by improving wetting behavior. Comprising molybdenum or tungsten, this barrier layer sits between a platinum and a gold layer, preventing a solid-state reaction between indium and gold. Consequently, laser cooling is improved, boosting high-power performance."*

3.3.2 Build the Domain Adaptation Dataset

The data used to train, validate, and test any domain adaptation model are crucial. Our data construction is driven by two goals: accurately predicting CPC (sub-)classes for disclosures and mapping disclosures and patents into the same domain-invariant embedding space.

Given that our primary goal is to predict the CPC classification of research disclosures as accurately as possible, we want to train our model with patents that are similar to disclosures in terms of the technologies they describe. For example, our model should not care about correctly predicting the CPC classification of pharmaceutical drug patents if there is no disclosure about pharmaceutical drugs. We implement this notion by matching disclosures with ten patents based on lexical similarity. [10] This selection process oversamples patents that are similar to research disclosures in terms of the words used in them.

A disadvantage of this selection process is that it hampers the secondary goal, that is, mapping research disclosures and patents into a common embedding space. Without having seen patents about certain technologies, our model will not appropriately map all patents and research disclosures into a common embedding space. This is why we also include ten patents randomly drawn from each CPC subclass in each year. These patents are matched with the lexically closest research disclosure.

By constructing the data for domain adaptation in this way, we make things easier for our model by preventing it from getting stuck on technological differences between research disclosures and patents. Additionally, we reduce the computational burden by limiting the data to a subset of all patents, resulting in around 850,000 patent/research disclosure pairs. As is common practice in machine learning, we split the data into training, validation, and test sets.

[10] For the matching, we add disclosures and patents to an Elasticsearch (https://www.elastic.co/) database and match each disclosure with 10 patents that are similar in terms of words used in the title and the abstract. The matching is based on Elasticsearch implementation of the BM25 algorithm (Jones et al., 2000) (https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables). To avoid patent duplicates, we select disclosure-patent pairs based on the BM25 score. If two research disclosures are matched with the same patent, we keep the disclosure-patent pair with the higher score, and the remaining disclosure is paired with the next best matching patent, again based on the BM25 algorithm.

Our example research disclosure is paired with closely related patents, such as "System for Soldering a Semiconductor Laser to a Metal Base," filed in 1977 and cited 23 times.

3.3.3 Train a Domain Adaptation Model

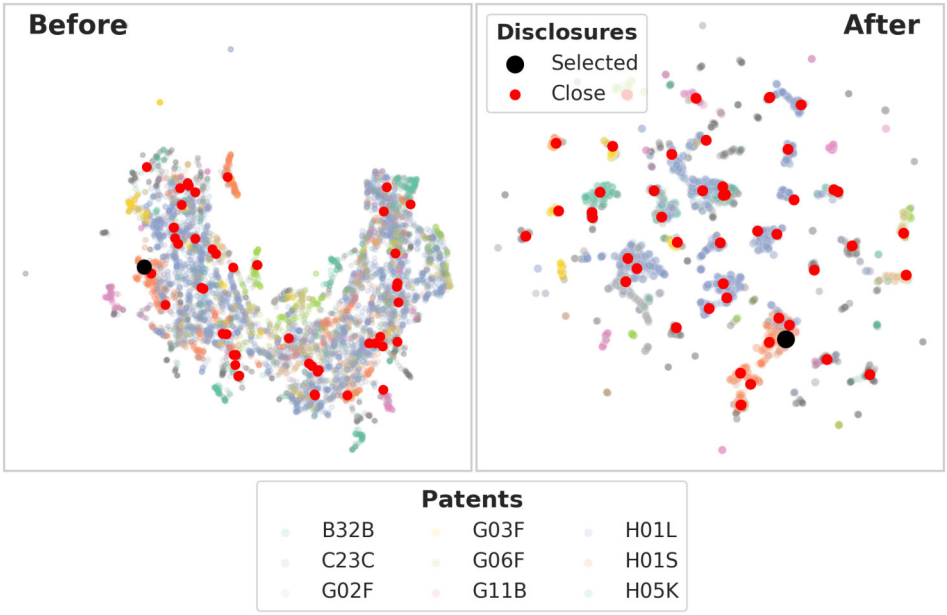
During training, each of the research disclosure/patent pairs is processed by BERT and then enters two classification tasks: the technology classification and the domain classification. Only the patent "part" of each pair enters the technology classification and the model tries to predict the its CPC (sub-)class. It then updates the parameters of the model such that it improves the performance in the technology classification task. Simultaneously, the pair enters the domain classification and the model calculates the parameter update that would improve the domain classification. Instead of doing the update, however, the model does an update of the parameters in the opposite direction.[11]In other words, the model finds out what it should do to better distinguish domains and then does the opposite. With the trained model, we extract the domain-invariant technology class prediction and a domain-invariant embedding for research disclosures and patents, mapping them into a common technology space.

We illustrate what is happening in the domain adaptation with our example disclosure in Figure 3. The figure shows embeddings of disclosures (the example research disclosure and lexically similar ones) and patents (lexically similar to the example and other research disclosures), mapped into 2D space. Initially, before training the model, both research disclosures and patents are scattered without clear grouping. This indicates that the untrained model is not capable of detecting differences between the technology classes. Additionally, there are clusters of research disclosures that are likely not driven by technological similarity but rather domain-specific attributes of disclosures. After the training, we observe a significant change. The embeddings are clustered according to technology classes, which means that the model has learned the technology classification. At the same time, there are no clear differences between domains. The dominant features that define the embedding space are technological differences, not differences between research disclosures and patents. This result highlights the model's ability to bridge the gap between disclosures and patents, ensuring that they are grouped by their technological similarities.

The example disclosure is accurately classified under the H01S CPC subclass for semiconductor laser patents, with the second most likely subclass being H01L, which covers semiconductor manufacturing processes. This illustrates the effectiveness of the pipeline in identifying technological similarities and distinctions. After the domain adap-

[11]This process is called gradient reversal and was popularized by **?**.

Figure 3: Illustration of Adversarial Learning



Notes: We selected the 49 disclosures that are closest to our example disclosure based on the BM25 algorithm. We then matched the 50 disclosures—one example disclosure and 49 close ones—with their respective 100 closest patents according to the BM25 algorithm and the 100 closest patents according to the domain invariant vectors. The figure was generated using McInnes et al. (2017).

tation, the closest patent in the domain invariant technology space is titled "Semiconductor Laser Device."

The implementation demonstrates our pipeline's capability to accurately predict CPC subclasses for research disclosures and achieve a unified representation of research disclosures and patents within the technological space defined by CPC classifications.

3.4 Classification Performance

Evaluating the classification performance of our model presents unique challenges, particularly when assessing its effectiveness on disclosures. Unlike patents, where performance metrics are straightforward due to the abundance of labeled data, disclosures pose a more complex scenario. Interestingly, our archival research unearthed CPC classifications for 4,074 disclosures between 1972 and 1977 within Research Disclosure issues, providing a rare dataset for evaluation.[12]

Table 1 contrasts the model performance across two distinct datasets: (i) patents in the test dataset from our data construction, i.e., patents that our model has not seen during training, and (ii) the historical dataset of research disclosures' CPC classifications

[12]Caution is warranted in interpreting these classifications as definitive ground truth, given the ambiguity surrounding their origin and the potential impact of technological evolution and CPC scheme updates over time. This period's data may not fully represent subsequent years, further complicating the evaluation.

Table 1: Classification Performance

	Accuracy			Recall	Precision	F1-score
	Top-1	Top-2	Top-5			
Sample of U.S. Patents						
Subclass	0.70	0.84	0.94	0.63	0.65	0.64
Class	0.77	0.89	0.96	0.72	0.73	0.72
Section	0.85	0.94	0.98	0.85	0.85	0.85
Research Disclosures (1972–1977)						
Subclass	0.56	0.69	0.80	0.27	0.27	0.25
Class	0.69	0.81	0.90	0.48	0.44	0.42
Section	0.80	0.91	0.97	0.75	0.70	0.72

Notes: The table provides performance metrics for our CPC classification of patents and research disclosures. Benchmark test samples are 147,532 patents of the dataset generated for the domain adaptation and 4,074 research disclosures (1972–1977) for which CPC labels are available. *Accuracy* is the percentage of correctly predicted patents and research disclosures; *Recall* is the percentage of successfully identified patents/research disclosures belonging to a specific CPC subclass/class/section; *Precision* is the percentage of patents/research disclosures assigned to a specific CPC subclass/class/section that are correctly specified; the *F1-score* combines Recall and Precision into a single metric. Recall, Precision, and F1-score are calculated on a macro level meaning that they are calculated for each CPC subclass/class/section separately and the reported number is the simple mean across CPC subclass/class/section. *Source:* archival collection of CPC labels (at British Library, London, UK, and Technical Information Library, Hannover, Germany); own calculations.

from 1972 to 1977. The model demonstrates commendable classification accuracy on the patent dataset, achieving around 70% accuracy at the subclass level (across 634 CPC subclasses), 76% at the class level (124 CPC classes), and 85% at the section level. Since patents often have additional CPC (sub)classes, we also evaluate whether the primary CPC (sub)class is among the 2 and 5 CPC (sub)classes with the highest probability weight assigned to them by our model. When doing so, the accuracy jumps to 84% and 95% on the subclass level, highlighting that even if the model fails to identify the primary CPC class, it is nonetheless close. Metrics sensitive to class imbalance, such as the F1-score, exhibit lower performance at the class level, although they remain comparable at the section level.

Performance on the historical disclosure dataset declines, with a drop of approximately 10 points in accuracy and a more pronounced decrease in other metrics. This discrepancy is partly attributable to the inherent challenges of transfer learning, where target domain performance typically does not match that of the source domain. Additional factors potentially influencing this outcome include the uncertain origin of the historical classifications—likely self-reported rather than expert-verified[13]—the fact that classification scheme used by disclosures is the International Patent Classification (IPC), and the changes in classification schemes over time.

[13] An issue of Research Disclosure that includes the 1972 classification scheme states: "[N]o effort has been made to classify every element of the disclosure."

4 What Really is in Research Disclosures?

In this section, we take a simplistic view of patenting and research disclosures (ignoring the obvious option of keeping one's invention secret). Inventors have the choice between patenting their invention or publishing it in the form of a research disclosure (without any claims to formal intellectual property). We should expect—somewhat naïvely—that inventors publish their technological developments in the form of research disclosures because (1) they are not patentable subject matter (and therefore not eligible for patenting), (2) they are not novel enough (and therefore not eligible for patenting), or (3) they are too insignificant to warrant the costs of a patent application and subsequent patent prosecution.

These three motivations to publish as a research disclosure, if applicable, imply that research disclosures are different from patents, and we should not expect much parallelism in how they develop over time. Moreover, research disclosures are laggards—we should not expect new ideas and concepts to appear in research disclosures because the novelty is reserved for patents. And last, we can expect research disclosures to be of lower significance and value.

We use our patent classifications for research disclosures to explore these hypotheses. Assuming that research disclosures and patents that are textually similar also share key characteristics such as content, novelty, and value, we transfer characteristics from patents (which we can measure) to their associated research disclosures (for which some of these characteristics are not available).

4.1 Comparing Research Disclosures and Patents

Assigning CPC classes to research disclosures allows us to compare long-term trends of patenting and defensive publications (in the form of research disclosures). In Table 2 we list the (eight) CPC sections, their respective short descriptions, and the number of research disclosures assigned to each section. In the aggregate, research disclosures are most relevant for inventions in physics (Section G), electricity (Section H), and performing operations and transporting (Section B). We observe the smallest number of research disclosures in textiles and paper (Section D) and fixed constructions (Section E).

The comparisons of trends provide insights into the relative importance of the technology areas in their respective domains, and how that importance changes. In this section, we first show the results of aggregate trends (and differences thereof) in the volume of research disclosures and patents. Differences in these trends suggest that conclusions we often draw from changes in aggregate patenting trends do not necessarily extend to technological developments (and innovation more generally) outside the patent space. In a second step, we document the similarity of research disclosures and patents *within* a given CPC (and changes thereof over time) to examine the position of research

Table 2: CPC Sections

Section	Description	Disclosures
A	Human necessities	2,860
B	Performing operations; transporting	8,219
C	Chemistry; metallurgy	4,231
D	Textiles; paper	909
E	Fixed constructions	684
F	Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	3,750
G	Physics	18,600
H	Electricity	8,495

Notes: This table provides the short descriptions of CPC sections (https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/classification/cpc) and the number of research disclosures in each CPC section.

disclosure within the patenting space. Low similarities imply that the content of research disclosures is different from patents and that research disclosures occupy different regions in the technology space.

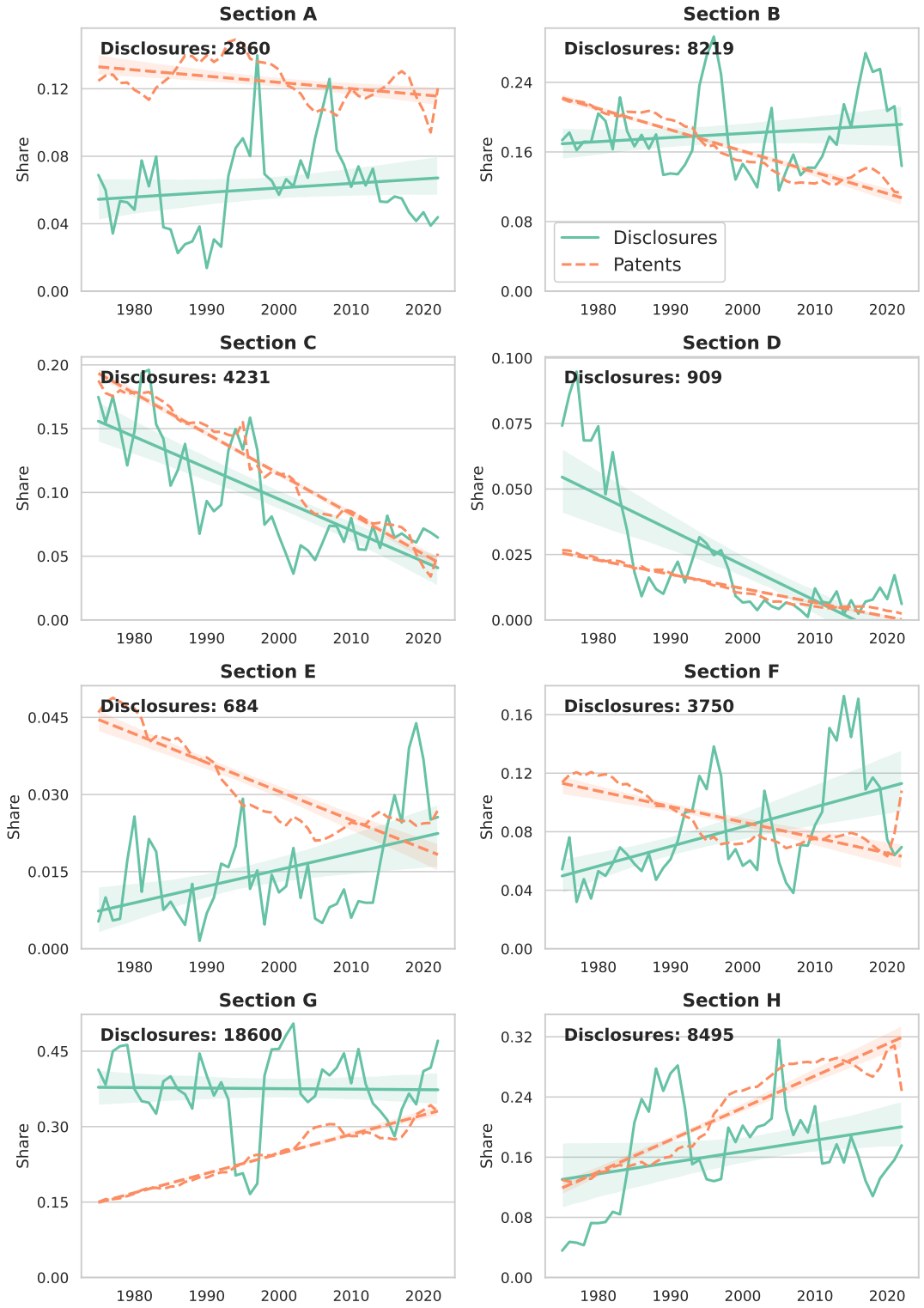
4.1.1 Aggregate Trends

In Figure 4, we show aggregate trends of research disclosures and patenting across CPC sections. We plot the share of research disclosures (in green) and patents (in orange) in a given CPC section over time (publication years for research disclosures and filing years for patents). We also include linear trend lines with 95% confidence intervals (shaded).

CPC sections C (chemistry and metallurgy) and D (textiles and paper) exhibit similar patterns. For both research disclosures and patents, these sections have similar shares and follow similar trends. In section C, we see a decline in shares for both patents and research disclosures, dropping from approximately 0.2 in 1975 to just 0.05 by 2020. This suggests a shift of focus away from this sector in both domains. A similar trend is observed in section D. In the early years of our sample period, research disclosures and patenting in section H (electricity) played similar roles. But while the share of research disclosure in that section remained fairly constant, patenting took off with the share of patents in section H approximately doubling.

In other sections, however, research disclosures and patents have different priorities or have experienced even diverging trends over time. In sections A (human necessities) and E (fixed construction), patenting plays a much more prominent role (especially in the earlier years of our sample period) than research disclosures. In section G (physics), the reverse is true. This section stands out as, on average, about 40% of all research disclosures have been in this section G. The relevance of patenting, on the other hand, has been significantly lower and has only recently caught up. This means that patenting statistics as a measure of R&D output in this CPC section ignore the technological

Figure 4: Shares of Research Disclosures and Patents



Notes: The figure depicts shares of patents (orange; by application year) and research disclosures (green; by disclosure year) for each CPC section, for 1975–2022. Linear trend lines with 95% confidence interval included. The numbers in the upper-left corner of each panel indicate the number of research disclosures assigned to a specific CPC section. *Source:* own calculations.

developments that are not patented but published as research disclosures. Research disclosures may have been a leading indicator of technological focus in this area.

Sections B (performing operations and transporting), F (mechanical engineering), and—albeit in small numbers—Section E (fixed construction) exhibit contrasting trends. In all three sections, patenting has lost prominence, whereas relatively more research disclosures have been made in these technological areas. These patterns highlight both the dynamic landscape of innovation and how different measures of R&D output capture different aspects of this dynamism. These documented differences in aggregate trends should serve as a word of caution. When we draw conclusions about changes in the direction of innovation and technology based on aggregate changes in patenting, we are ignoring the fact that the same trends do not necessarily extend over to other measures of R&D output.

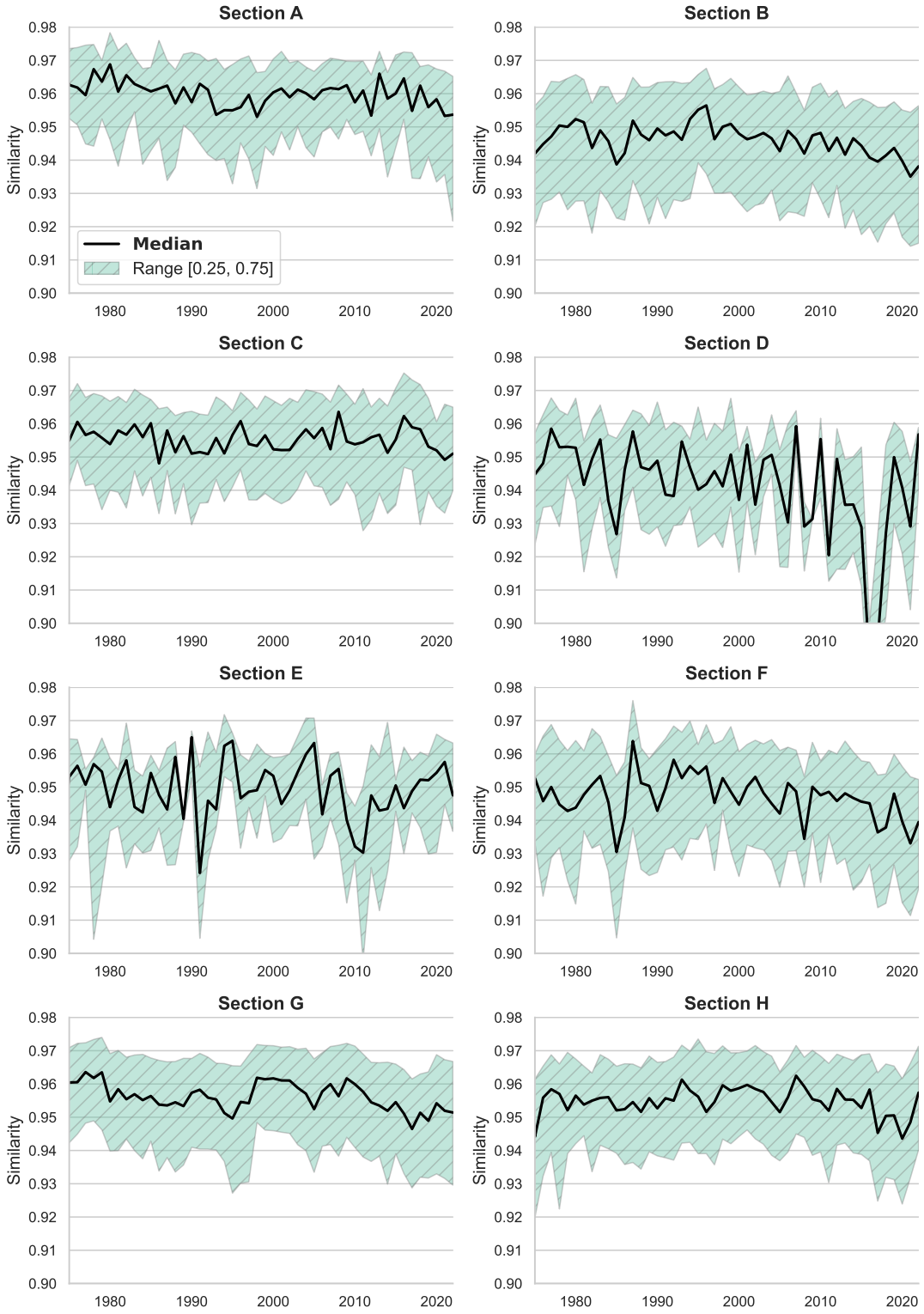
4.1.2 Technological Similarity

In the next step, we examine patents and research disclosures within a given CPC section. Examining their similarities tells us if the two domains populate the same technology space or if research disclosures are inherently different because, for instance, they contain inventions outside the scope of patentability. By considering the within-CPC similarities over time, we are also able to observe changes in these patterns. A decrease in the similarities means that the contents of patents and research disclosures diverge over time because they evolve in different directions or the scope of one of the domains changes. For instance, a change in patentability may result in a shift of inventions from patents to research disclosures, affecting the average similarities within a CPC.

In Figure 5, we plot the median similarity (and the interquartile range) of research disclosures to their ten most similar patents in a two-year window around the research disclosures publication date (using the patents' filing dates). We observe stable compositions for most CPC sections where neither medians nor the interquartile ranges follow noticeable trends. Sections A (human necessities) and C (chemistry; metallurgy) exhibit slightly stronger similarities, whereas sections B (performing operations; transporting) and F (mechanical engineering; ...) show a broader diversity of research disclosures relative to their most similar patents.

Section G (physics), which has historically been dominant in disclosures, has seen a notable drift away from patents since 2010. This indicates a potential divergence in the focus of recent disclosures from existing patents, with research disclosures thus occupying different (and novel) regions in the technology space. A similar trend is observable in disclosures in Section B (performing operations; transporting). These shifts are indicative of changes in the technological landscape and an evolving nature of innovation within these areas that are not captured in patents.

Figure 5: Similarity Between Research Disclosures and Patents



Notes: The figure depicts the median cosine similarity (and the interquartile range) of research disclosures to patents in a specific CPC section. Each disclosure is matched with the ten most similar patents in two-year window around the research disclosure's publication date. *Source:* own calculations.

4.2 Novelty in Research Disclosures

We take two approaches to identify research disclosures that contain novel content (i.e., content not yet part of a patent application). First, we examine the timing of the publication of a research disclosure relative to the filing dates of their most similar patents. A publication date of the research disclosure that precedes the filing date of the most similar patent(s) is an indicator that the research disclosure was first in disclosing the patented technology. For our second approach, we track and compare the first occurrence of technical terms in patents and research disclosures, using the list of keywords compiled by Arts et al. (2021).

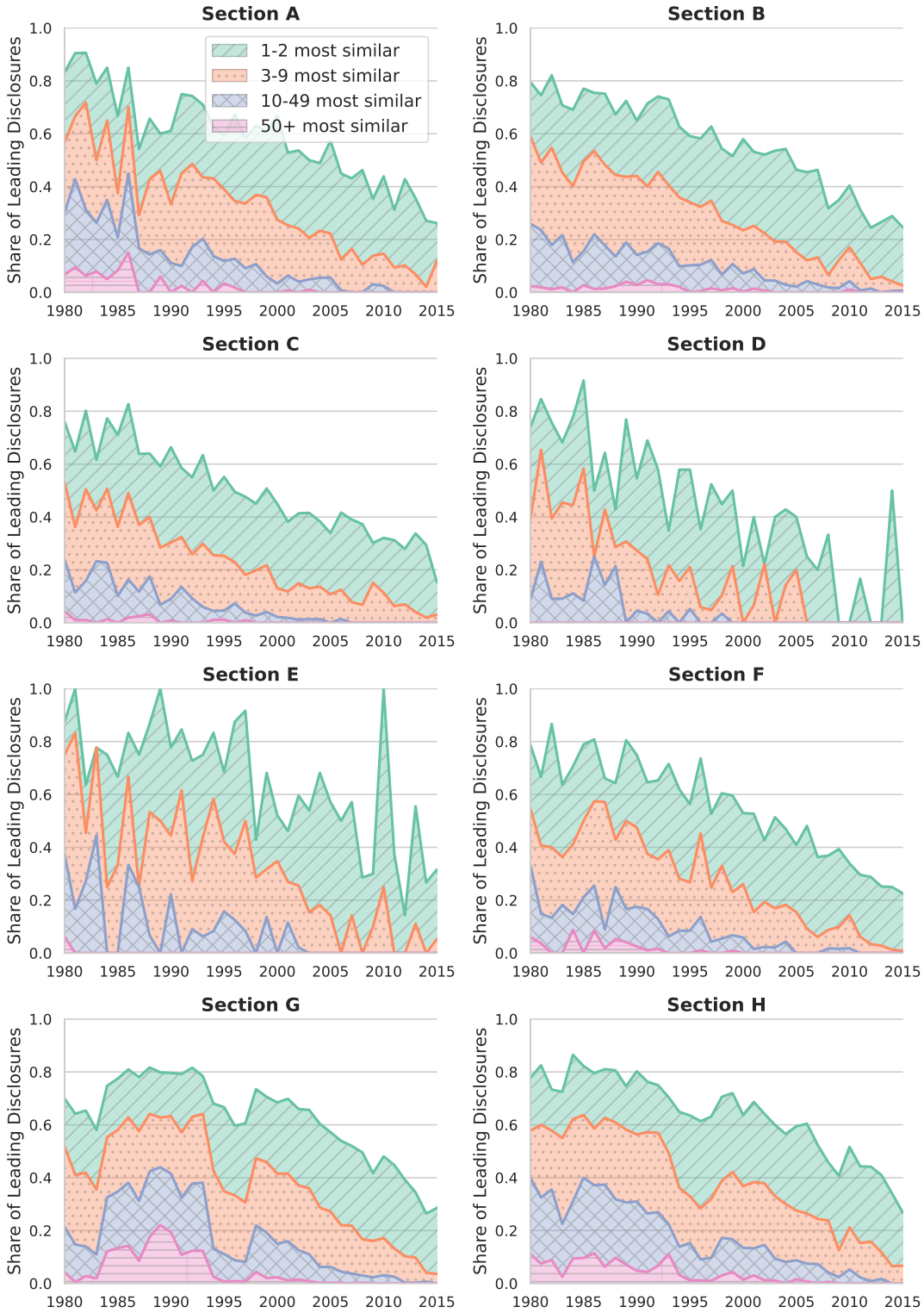
In Figure 6 we present the results for our first approach. We plot the share of disclosures (by publication date) that have a publication date before the filing date of all of its most similar patents. The green line captures the share for the most similar patent, the orange line for the 3 most similar patents, the blue line for the 10 most similar patents, and the magenta line for the 50 most similar patents. As a given research disclosure has to have been published before an increasing number of most similar patents, our approach becomes restrictive as we increase that number. We observe this in lower shares the higher the number of most similar patents increases.

The results for intermediate years are strongly suggestive of the novelty of research disclosures in their respective CPC section. For instance, in Section G (physics), in the early 1990s, more than 60% of research disclosures were published before 3 most patents were filed. This number is still at close to 40% for the 10 most similar patents. Even for our most restrictive case, around 10% of research disclosures preceded all of its 50 most similar patents. These patterns prevail (albeit weaker). In the early 2000s, every tenth research disclosure is more novel (by publication date) than its ten most similar patents, and two out of five are more novel than their three most similar patents. We see similar patterns in Section H (electricity).

The patterns we observe at both ends of the timeline (and the negative trend in between) are somewhat mechanical. Research disclosures published in 1980 are more likely to lead because we have fewer patents to compare them to, increasing the likelihood that none of the most similar patents were filed before 1980. Similarly, Research disclosures published in 2015 are less likely to lead because relatively fewer patents were filed after 2015, and many more were filed before. Additionally, the surge in patent filings post-2000 has crowded the technology space, increasing the likelihood of finding a closely related patent within this period.

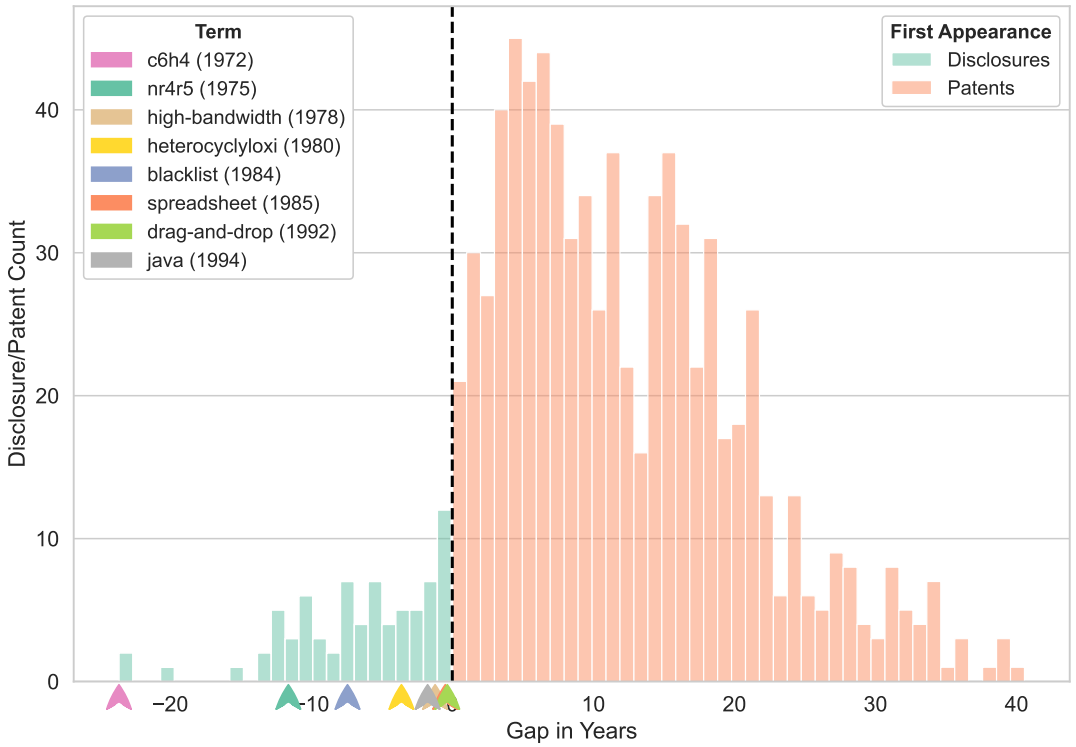
In Figure 7, we present the results for our second approach. We track the first use of specific keywords that are later adopted and frequently used by subsequent patents. This approach hinges on the premise that the introduction of new terminology within patents

Figure 6: Research Disclosures Leading Their Most Similar Patents



Notes: The figure depicts the share of research disclosures with publication dates prior to the filing dates of all of the 1 (green), 3 (orange), 10 (blue), or 50 (magenta) most similar patents, for research disclosures published between 1980 and 2015. *Source:* own calculations.

Figure 7: Research Disclosures with Novel Terms Prior to Patents



Notes: The figure depicts the count of research disclosures that use a term before (green) and after (orange) its very first use in a patent. Each bar represents the count of research disclosure by years of the respective gap. The legend provides eight examples of earlier use in research disclosures. *Source*: Arts et al. (2021) (for first-time use terms in patents) and own calculations.

can serve as a marker of innovation. We employ the list of 1000 keywords curated by Arts et al. (2021) and extend the analysis to research disclosures.

Our analysis reveals that 873 out of the 1000 keywords designated as "novel" by Arts et al. (2021) are also present in disclosures. While the majority of these keywords debut in patents (orange bars in Figure 7), a notable proportion first emerges within disclosures (green bars). This observation underscores the potential of disclosures to serve as a conduit for novel ideas and terminology, predating their formal recognition and adoption in patent documents.

Both our approaches to identifying research disclosures with novel content reveal a significant level of technological leadership in research disclosures. Substantial numbers of research disclosures are published before their (textually) closest patents are filed, and many contain terminology before it is first used in patents. This is not to say that the affected patents should not have been granted. What our results rather reveal is that broader ideas and concepts are not necessarily novel when they enter the patenting space.

4.3 Implied Value of Research Disclosures

To assess the significance of research disclosures in relation to patents, we rely on forward citations, a measure used to proxy the value of patents. We argue that a research disclosure that is similar to (and shares key features) with highly-cited patents, is of relevance and value itself. We refer to this value as the *implied value* or a research disclosure. For this implied value, we determine the proportion of a research disclosure's 100 closest patents that rank within the top 10% of most-cited patents for their respective CPC class and filing year.

Under the assumption of a random distribution, the baseline expectation for this share is 1/10 (that means, 10% of 100 random patents are in the top 10% of the most cited patents). A share below 10% indicates that a research disclosure is associated with less-cited patents, suggesting a lower implied value. Likewise, a share above 10% means a higher implied value.

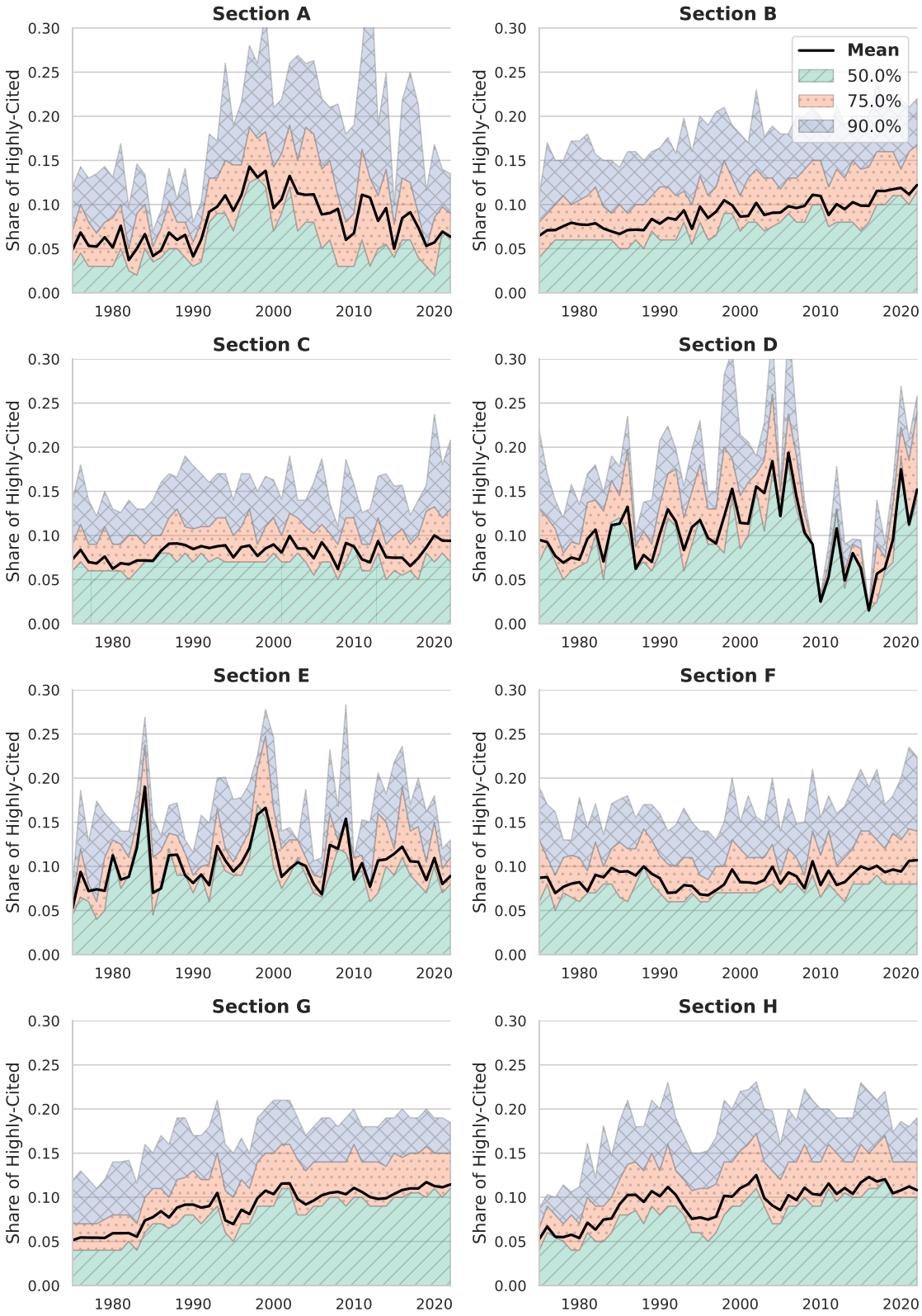
Figure 8 depicts the results of this exercise by CPC section. We plot the mean share (by publication year of the research disclosure) of highly-cited patents associated with a given research disclosure; we also provide the median (green), 75th percentile (orange), and 90th percentile (purple) of the distribution of shares.

Sections A (human necessities), D (textiles, paper), and E (fixed construction) exhibit the highest implied values. Because of their small numbers, D and E show significant levels of heterogeneity over time. Section A saw a strong increase in implied value in the 1990s, from a mean level of 5% to almost 15%. The implied level of research disclosures as decreased since.

We observe the lowest implied values in Sections C (chemistry and metallurgy) and F (mechanical engineering,...), where the mean shares fall well below 10% and are fairly constant over time. In both sections, however, we find a significant share of research disclosures with relatively high implied values (the 90th percentiles reach shares of 20% and more). Research disclosures in Sections B (performing operations, transporting), G (physics), and H (electricity) have low implied values in earlier years but exhibit an upward trend and reach levels above the baseline in the early 2000s.

Our results highlight the importance of research disclosures as a means to freely publish even high-value technological developments. Research disclosures are, therefore, not just ideas and inventions that are too insignificant to warrant the costs of a patent application and subsequent prosecution (especially in CPC Sections A, D, and E). Instead, firms and inventors often decide to forego patent protection and instead disclose their developments for everyone to see and use. For Sections B, G, and H we further document how research disclosures have evolved from being an outlet for niche results to a vehicle to publicize technological developments of practical relevance and value.

Figure 8: Implied Value (by Proximity to Highly-Cited Patents)



Notes: The figure depicts the share of highly-cited patents (ranking within the top 10% of most-cited patents of their respective CPC class and filing year) that are among the 100 closest patents to a given research disclosure. Increasing shares imply an increasing implied value of research disclosures. *Source:* PatentsView (for citations) and own calculations.

5 Anonymous Disclosure as a Strategic Choice

5.1 Sometimes Firms Disclose Anonymously

One unique aspect of disclosures is the option for anonymity. When submitting a research disclosure, the disclosing party can choose whether to reveal or conceal their identity. This feature introduces a strategic dimension to the disclosure process. For instance, a party may wish for their invention to become part of the prior art, thereby influencing patentability criteria for future inventions, without signaling to competitors their active engagement in a specific technological field.

This capability to publish anonymously fundamentally differentiates disclosures from patents. In the patent process, the disclosure of the inventor's identity is a requirement, ensuring transparency about the source of innovation. Conversely, anonymous disclosures provide a means to contribute to the collective knowledge base while maintaining a strategic silence about the contributor's identity and areas of interest. This distinction highlights the nuanced roles that research disclosures play in the broader innovation ecosystem, offering a pathway for influencing technological development and the patent landscapes without direct attribution.

In Figure 9, we show the proportion of anonymous disclosures across different CPC sections from 1975 to 2022. We observe a considerable amount of heterogeneity within and across CPC sections. These patterns reflect ever-changing incentives to disclose one's identity, likely driven by strategic considerations. In the remainder of this section, we explore the disclosure-level factors that determine whether a disclosing party chooses to do so anonymously.

5.2 Strategically Keeping Others in the Dark

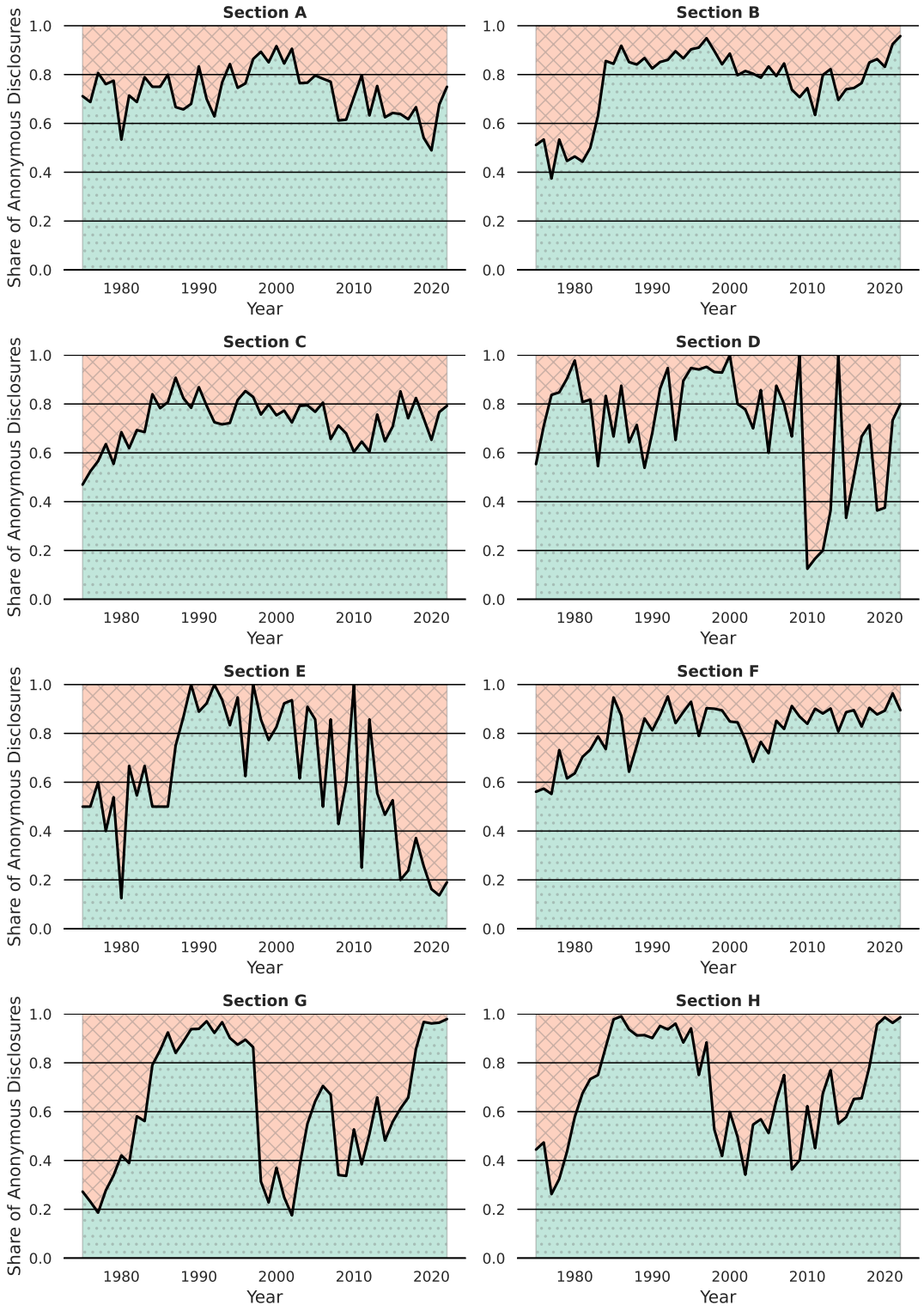
Our analysis is based on the presumption that an inventor's decision to disclose anonymously is intricately linked to the disclosure's technological positioning.[14] The decision is thus strategically influenced by how the disclosure aligns with the broader technology space. This alignment may reflect considerations around competitive positioning, intellectual property strategy, and the potential impact of the disclosure on the firm's or inventor's technological domain.

In order to explore the role of positioning in the technology space, we construct three variables.

1. $SimClose_i$ captures the research disclosure i's similarity to the closest (i.e., most similar) patent (using cosine similarity) that was filed before the research disclosure's publication.

[14] Obviously, the decision is also driven by inventor characteristics. Those we do not observe for anonymously disclosed RDs and, therefore, cannot use for our analysis.

Figure 9: Anonymous Research Disclosures



Notes: This figure depicts the annual share of research disclosures (green) that are made anonymously, by CPC section for 1975–2022. *Source:* Research Disclosure Database and own calculations.

2. $SimCPC_i$ captures the average similarity of research disclosure i and the patents within i's CPC class filed in the research disclosure's publication year.

3. $HighCit_i$ captures the proximity to highly-cited patents. It is equal to one if the research disclosure's closest preceding patent (by filing date) is among the top 10% most-cited patents for its publication year and CPC class.

Our outcome variable is y_i for a given research disclosure i is the anonymity status, with $y_i = 1$ if the inventor disclosed anonymously and zero otherwise. We employ a logistic regression model to study the relationship between the anonymity decision and the research disclosure's positioning in the patenting space. The model is specified as follows:

$$\ln\left(\frac{p(y_i = 1|\cdot)}{1 - p(y_i = 1|\cdot)}\right) = \beta_0 + \beta_1 SimClose_i + \beta_2 SimCPC_i + \beta_3 HighCit_i + \mathbf{X}_i\delta + u_i, \quad (1)$$

where β_1, β_2, and β_3 are the coefficients corresponding to our variables of interest. $\mathbf{X}_i$ is a vector of control variables (publication year FE and CPC section FE).[15] We restrict our sample of research disclosures to those published between 1980 and 2015. This restriction assures a sufficient number of patents to enter our patenting-space proximity measure.[16]

5.3 Regression Results for Anonymous Disclosure

Table 3 presents our regression results. Each column contains results from a separate regression. For each of the variables of interest, we run two separate regressions: one that uses only that variable of interest and one where the variable is interacted with CPC section indicators. The last two columns contain joint estimations of all three variables of interest. Since we are using a logistic regression model, we only interpret the sign of coefficients and relative size across section—e.g. a higher coefficient of $HighCit_i$ when interacted with Section A than when interacted with Section B.

Negative coefficients in columns (1), (3), and (7) suggest that when research disclosure i is in closer proximity to an existing patent (column (1)) or exhibits higher similarity with the patents in the respective CPC class (column (3)), the inventor is more likely to reveal their identity (anonymous disclosure becomes less likely). This association is particularly strong for research disclosures in Sections G (phyics) and H (electricity), with negative interaction terms, and weaker for Sections A (human necessities) and B (performing operations, transporting), with positive interaction terms.

[15] Note that, for statistical inference, we do not account for the variance introduced by the estimation of CPC sections and the technological positioning of disclosures and patents. This omission may result in the underestimation of the variance of our coefficients, we do not anticipate it introducing a systematic bias into our estimation results.

[16] We use PatentsView as our patent-data source. It provides patent-level information (including patent texts) for all patents granted in 1976 and onwards.

Firms with research disclosures that are not closely related to patents (in the respective patenting space) are more likely to make them anonymously. A motivation for this could be the desire to conceal the inventor's identity when their innovation strategy diverges from existing patents. This could mean that inventors seek to stay under the radar when they disclose novel (or non-patentable) technologies.

Being close to a highly-cited patent does not have a significant average effect on an inventor's decision to disclose anonymously. However, we observe some heterogeneity across sections. In Section A (human necessities), closer proximity to a highly-cited patent (yielding a higher implied value for the research disclosure) is associated with more anonymous disclosures, whereas in Section H (electricity) such a higher implied value of the research disclosure is associated with fewer anonymous disclosures.

6 Concluding Remarks

In this paper, we study the content, novelty, and value of defensive publications relative to patents. Inventors may choose defensive publications—such as those in the journal *Research Disclosure*—when their inventions are patentable or not novel (and thus not patent-worthy) or when the inventions are of insufficient value to warrant the costs of a patent application and subsequent patent prosecution. As a consequence, the set of technologies disclosed in research disclosures and patents ought to differ in content, novelty, and value.

We use a large language model (LLM) to apply the cooperative patent classification (CPC) system to a set of defensive publications (from 1962 to 2022) from the journal *Research Disclosure*, assigning disclosures to the CPC classification they would have received had they been filed as patents. With this approach, we map research disclosures and patents into a common space, allowing for a direct evaluation of textual similarities between these two types of R&D outputs.

We find that while in some technologies, patents and research disclosures follow similar aggregate trends, some exhibit diverging developments over time. We also document shifts in the position of research disclosures in the patenting space that are indicative of changes in the technological landscape not captured in patents. We further show that substantial numbers of research disclosures are published before their closest patents are filed, and many contain terminology before it is first used in patents. Last, we find that in several technology areas, research disclosures have evolved from being an outlet for niche results to a vehicle to publicize technological developments of high practical relevance and value.

Our results imply that when we draw conclusions about the efficacy of policy in fostering technological progress or influencing the direction of innovation based solely on patent data, we obtain an incomplete picture because patents are not representative of

Table 3: Strategic Choice of Anonymous Disclosures

	$SimClos_i$		$SimCPC_i$		$HighCit_i$		**All**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$SimClos_i$	-5.98***						-4.75***	
	(0.49)						(0.62)	
$SimClos_i$:[A]		0.95						-1.46
		(2.07)						(2.55)
$SimClos_i$:[B]		2.60**						3.36**
		(1.20)						(1.52)
$SimClos_i$:[C]		2.83						3.88
		(1.80)						(2.39)
$SimClos_i$:[D]		-1.95						2.88
		(4.88)						(6.59)
$SimClos_i$:[E]		-3.56						-5.21
		(4.60)						(5.63)
$SimClos_i$:[F]		-2.09						-2.40
		(2.03)						(2.72)
$SimClos_i$:[G]		-11.51***						-10.09***
		(0.72)						(0.91)
$SimClos_i$:[H]		-7.44***						-5.74***
		(1.10)						(1.36)
$SimCPC_i$			-2.37***				-0.94***	
			(0.23)				(0.29)	
$SimCPC_i$:[A]				1.72				2.20*
				(1.10)				(1.34)
$SimCPC_i$:[B]				0.48				-0.67
				(0.68)				(0.86)
$SimCPC_i$:[C]				0.43				-0.92
				(1.03)				(1.35)
$SimCPC_i$:[D]				-3.18				-4.28
				(2.68)				(3.62)
$SimCPC_i$:[E]				-0.05				1.52
				(2.70)				(3.27)
$SimCPC_i$:[F]				-0.75				0.25
				(1.14)				(1.52)
$SimCPC_i$:[G]				-3.96***				-1.03**
				(0.32)				(0.40)
$SimCPC_i$:[H]				-2.71***				-1.14*
$HighCit_i$					-0.05		-0.04	
					(0.03)		(0.03)	
$HighCit_i$:[A]						0.53***		0.53***
						(0.17)		(0.17)
$HighCit_i$:[B]						0.09		0.08
						(0.10)		(0.10)
$HighCit_i$:[C]						0.14		0.14
						(0.15)		(0.15)
$HighCit_i$:[D]						0.50		0.50
						(0.39)		(0.39)
$HighCit_i$:[E]						-0.20		-0.20
						(0.34)		(0.33)
$HighCit_i$:[F]						-0.00		0.01
						(0.17)		(0.17)
$HighCit_i$:[G]						-0.06		-0.04
						(0.05)		(0.05)
$HighCit_i$:[H]						-0.32***		-0.30***
						(0.07)		(0.07)
Observations	37488	37488	37488	37488	37488	37488	37488	37488
Log-Likelihood	-19933.75	-19861.03	-19890.05	-19777.04	-19957.98	-19921.47	-19889.84	-19726.60

Notes: Regression table from logistic regressions (equation (1)). The outcome is $y_i = 1$ if a research disclosure is published anonymously between 1980 and 2015. All regression models include publication year FE and CPC section FE. Robust standard errors in parentheses. * p<.1, ** p<.05, ***p<.01

the broader innovation landscape. Last, our methodology to find technology classes for research disclosures has many applications and can be applied to other text domains that capture R&D output. We can essentially predict technology classes for every form of text that comes with an abstract or for which we can generate an abstract (e.g., academic publications, new-product descriptions) and compare their content, novelty, and value with those of patents.

References

Adams, Stephen and Victoria Henson-Apollonio (2002) "Defensive Publishing: A Strategy for Maintaining Intellectual Property as Public Goods," ISNAR Briefing Paper 53, International Service for National Agricultural Research.

Araci, Dogu (2019) "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez (2021) "Natural Language Processing to Identify the Creation and Impact of New Technologies in Patent Text: Code, Data, and New Measures," *Research Policy*, 50 (2), 104144.

Baker, Scott and Claudio Mezzetti (2005) "Disclosure as a Strategy in the Patent Race," *Journal of Law and Economics*, 48 (1), 174–194–254.

Bar, Talia (2006) "Defensive Publications in an R&D Race," *Journal of Economics & Management Strategy*, 15 (1), 229–254.

Bekamiri, Hamid, Daniel S. Hain, and Roman Jurowetzki (2021) "PatentSBERTa: A Deep NLP based Hybrid Model for Patent Distance and Classification using Augmented SBERT," Unpublished manuscript, arXiv preprint at doi.org/10.48550/arXiv.2103.11933.

Beltagy, Iz, Matthew E Peters, and Arman Cohan (2020) "Longformer: The Long-Document Transformer," Unpublished manuscript, arXiv preprint at doi.org/10.48550/arXiv.2004.05150.

Bhaskarabhatla, Ajay and Enrico Pennings (2014) "Defensive Disclosure of Patentable Inventions under Antitrust Enforcement," *Industry and Innovation*, 21 (7-8), 533–552.

Bouty, Isabelle (2000) "Interpersonal and Interaction Influences on Informal Resource Exchanges between R&D Researchers across Organizational Boundaries," *Academy of Management Journal*, 43 (1), 50–65.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018a) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Unpublished manuscript, arXiv preprint at doi.org/10.48550/arXiv.1810.04805.

——— (2018b) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, https://arxiv.org/abs/1810.04805.

Farahani, Abolfazl, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia (2021) "A Brief Review of Domain Adaptation," in *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, 877–984: Springer.

Gächter, Simon, Georg von Krogh, and Stefan Haeflinger (2010) "Initiating Private-Collective Innovation: The Fragility of Knowledge Sharing," *Research Policy*, 39 (7),

893–906.

Ganglmair, Bernhard, Alex Holcomb, and Noah Myung (2020) "Expectations of Reciprocity when Competitors Share Information: Experimental Evidence," *Journal of Economic Behavior and Organization*, 170, 244–267.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky (2016) "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, 17 (59), 1–35.

Gans, Joshua S., Fiona E. Murray, and Scott Stern (2017) "Contracting over the Disclosure of Scientific Knowledge: Intellectual Property and Academic Publication," *Research Policy*, 46 (4), 820–835.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2020) "Generative Adversarial Networks," *Communications of the ACM*, 63 (11), 139–144.

Hall, Bronwyn H. and Rosemarie Ham Ziedonis (2001) "The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979–1995," *RAND Journal of Economics*, 32 (1), 101–128.

Häussler, Carolin (2011) "Information-Sharing in Academia and the Industry: A Comparative Study," *Research Policy*, 40 (1), 105–122.

Häussler, Carolin, Lin Jiang, Jerry Thursby, and Marie C. Thursby (2014) "Specific and General Information Sharing Among Academic Scientists," *Research Policy*, 43 (3), 465–475.

Henkel, Joachim and Stefanie M. Lernbecher (2008) "Defensive Publishing – An Empirical Study," unpublished manuscript, available at https://ssrn.com/abstract=981444.

Johnson, Justin P. (2014) "Defensive publishing by a leading firm," *Information Economics and Policy*, 28, 15–27, 10.1016/j.infoecopol.2014.05.001.

Jones, K. Sparck, Steve Walker, and Stephen E. Robertson (2000) "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments," *Information Processing & Management*, 36 (6), 779–840.

Kang, Guoliang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann (2019) "Contrastive Adaptation Network for Unsupervised Domain Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4893–4902.

Lee, Jieh-Sheng and Jieh Hsiang (2019a) "Patentbert: Patent classification with fine-tuning a pre-trained bert model," *arXiv preprint arXiv:1906.02124*, https://arxiv.org/abs/1906.02124.

——— (2019b) "PatentBERT: Patent Classification with Fine-Tuning a Pre-Trained BERT Model," Unpublished manuscript, arXiv preprint at doi.org/10.48550/arXiv.1906.02124.

Li, Shaobo, Jie Hu, Yuxin Cui, and Jianjun Hu (2018) "DeepPatent: patent classification with convolutional neural networks and word embedding," *Scientometrics*, 117 (2), 721–744, 10.1007/s11192-018-2905-5.

McInnes, Leland, John Healy, and Steve Astels (2017) "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, 2 (11), 205.

Noel, Michael and Mark Schankerman (2013) "Strategic Patenting and Software Innovation," *Journal of Industrial Economics*, 61 (3), 481–520.

Parchomovsky, Gideon (2000) "Publish or Perish," *Michigan Law Review*, 98 (4), 926–952.

Rozantsev, Artem, Mathieu Salzmann, and Pascal Fua (2018) "Beyond Sharing Weights for Deep Domain Adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (4), 801–814.

Schrader, Stephan (1991) "Informal Technology Transfer Between Firms: Cooperation Through Information Trading," *Research Policy*, 20, 153–170.

Srebrovic, Rob and Jay Yonamine (2020a) "Leveraging the BERT algorithm for Patents with TensorFlow and BigQuery," *White paper*, https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf.

——— (2020b) "Leveraging the BERT Algorithm for Patents with TensorFlow and BigQuery," White Paper, Global Patents at Google. Available at: https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf (Last accessed: February 20, 2024).

Touvron, Hugo, Louis Martin, Kevin Stone et al. (2023) "Llama 2: Open Foundation and Fine-Tuned Chat Models," Unpublished manuscript, arXiv preprint at doi.org/10.48550/arXiv.2307.09288.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017) "Attention Is All You Need," December, 10.48550/arXiv.1706.03762, arXiv:1706.03762 [cs].

von Hippel, Eric (1987) "Cooperation Between Rivals: Informal Know-How Trading," *Research Policy*, 16 (6), 291–302.

von Hippel, Eric and Georg von Krogh (2006) "Free Revealing and the Private-Collective Model for Innovation Incentives," *R&D Management*, 36 (3), 295–306.