Mapping the Links Between Science and Technology With Citations and Semantics

September 2024

Jianying Liu (OST-Hcéres, Paris), Mounir Amdaoud (OST-Hcéres, Paris), Wilfriedo Mescheba (OST-Hcéres, Paris), Justin Quemener (OST-Hcéres, Paris), David Sapinho (OST-Hcéres, Paris), Jean-Marc Deltorn (CEIPI, Strasbourg University), Dominique Guellec (OST-Hcéres, Paris)

Grant 2021-8402 from the European Patent Office under their Academic Research Program (EPO-ARP).

Summary

The linkages between science and technology are important for delivering the innovation needed to increase productivity and to address societal and environmental challenges. This report aims at better mapping and explaining these linkages, focusing on selected frontier technological domains (quantum cryptography, CRISPR, CAR-T cells and mRNA). It is made of two separate parts: one that addresses the measurement issues (citations and semantics), leading to recommendation regarding statistics of knowledge transfer; the other one conducts a behavioural assessment of citations, leading to recommendations regarding the forecasting patents based on scientific publications.

In a first part, we focus on the measurement aspects: How best to map science-technology linkages? Scientific publications and patents are major sources of information on science and technology respectively, and their mutual relations can be used as signals of linkages. Citation of a scientific publication by a patent (NPL: Non patent literature) and similarity between the text of publication and a patent (NLP: Natural language processing) are two signals of possible knowledge transfer between science and technology. We compare these two signals for USPTO patents (as frontier technical fields are more present in this particular patent office and requirements for citing the prior art are broader) and find only partial convergence between them when semantic proximity is calculated with the title and abstract of documents. We then calculate semantic proximity by using the full text of scientific publications and patent claims, which significantly improves the correspondence with citations. On that basis, further analysis provides support to two explanations for this discrepancy: 1) citations reflect commonality between specific ideas of documents, that do not necessarily show up in title and abstract, as shown by an analysis of full text of documents; 2) many citations have weak relevance, they reflect informational limitations, legal constraints or strategic choices of the patent holder rather than knowledge transfer. The study brings two operational conclusions regarding the use of NLP for mapping science-technology linkages: 1) NLP on scientific publications and patents should use the full text of documents instead of the abstract; 2) we propose and test a new indicator of knowledge transfer, which combines NLP (on full text) and NPL and succeeds in exploiting their respective strengths.

In a second part, we conduct econometric analysis of the citation patterns, with the view to map issues relating to the forecasting of patent filings based on scientific publications. A first analysis examines factors affecting whether a scientific publication will be cited or not by a patent. The factors that affect the likelihood of a scientific publication to be cited by a patent are similar in CRISPR and CAR-T cells. Main drivers are: the age of the publication (quadratic), its semantic proximity to patents, whether it is open access or not, whether it was funded or not, its number of authors, whether one of them at least is from industry and whether they reside in different countries (with a negative impact). In addition, we detected a "foundation effect" in CRISPR, by which earlier publications (foundational discoveries) are more cited over time than others.

A second analysis examines the lag between cited publications and citing patents. It finds that the impact of science on technology is differentiated: scientific discoveries have a quicker impact on inventions pertaining to the same domain, a slower impact on inventions from other domains. In the case of GPTs, it means that applications of the core domain develop later than core inventions. It also finds that the impact of discoveries on inventions is spread over time: notably foundational discoveries of a domain have a long term impact. Hence a forecasting model must include some dynamics and time variability.

1. Introduction

The linkages between science and technology are the subject of a vast literature, theoretical, historical or quantitative. The core idea, accepted even with variations by most scholars, is that science has a direct and significant influence at least on certain fields of technology, science informs the direction and content of new technology (Nelson 2004). On the other hand, technology is an enabler of scientific advances, notably by providing tools for observation, experimentation and analysis. In this study we investigate the influence of science on technology, using both traditional and newly available signals of the connections between the two domains, and focusing on a few, advanced fields.

A preliminary question when addressing this matter is the difference between science and technology: before analyzing their connections, one must understand their distinction. Different perspectives can be taken: epistemological (what sorts of knowledge the two domains generate, what are the validation criteria etc.), institutional (how their missions are implemented in society and the economy), and empirical (what types of data reflect best their respective workings). From an epistemological perspective, it is well recognized that science is about seeking new knowledge about nature, while technology is about designing useful artefacts (e.g. Nelson 2004). Although quite clear in many cases, e.g. Big Bang theory vs. mouse traps design, in certain advanced fields of science and technology, the very distinction looks blurred: e.g. in artificial intelligence, AI (are "transformers" part of science or part of technology?), in genetics (CRISPR resulted in two Nobel Prizes and many patents). In such domains new applications result from the most recent knowledge, and knowledge advances are often performed with immediate applications in mind.

Regarding the linkages between the two domains, the traditional vision, as heralded notably by Schumpeter, is the "linear model", which views technology as a downstream application of science: science generates understanding of natural phenomena, and technology frames them so as to serve human purposes. Conversely, certain scholars have argued that many important inventions were made without any scientific basis (e.g., the steam engine); the "chain-linked model" (Kline and Rosenberg 1987) presents the two domains as being connected by multiple knowledge flows in both directions.

There are good empirical reasons to see the connections between science and technology as being diverse in intensity, ranging from total separation in the most theoretical scientific disciplines or the most applied technological domains, to fusion in some of the frontier domains: disconnection at one end, identity at the other, and various densities of connections between these two extreme situations. It is also recognized that the dominant, although far from exclusive, direction of influence runs from science to technology.

The relations between science and technology are not only an issue of analytical interest, but also of policy interest. The institutional foundations, incentive systems and funding mechanisms differ significantly between the two domains. There is a fierce debate in science policy about the mounting requirements imposed by funders for basic research to demonstrate its usefulness for society, for closer linkages between basic and applied research. Government is pressing academia to connect more closely its research with applications, with technology, in the context of 1) societal challenges like the environment, social inequality or national security; 2) growing involvement of government with society and the economy, that makes it increasingly responsible for areas that used to be dominated by market forces, including technology.

Our study is based on this conceptual framework and aims at testing some of its implications in a few frontier domains, where the connection between science and technology is presumably the strongest: we will attempt to identify and map various signals of connections between scientific discoveries and technological inventions, measure their strength, investigate their determinants and their speed.

These questions are not new, but they can be revisited with new data and new methods, allowing to go beyond case studies or broad statistical characterizations: thanks to new data and methods, it is now possible to identify signals that reflect various types of linkages between science and technology, and to do that on large populations of discoveries and inventions. One purpose of this report is to apply this novel approach to a few specific fields, in order to explore its capacity to draw a sound picture of linkages between science and technology, and to identify methodological limitations and potential developments. We will focus on citations of scientific papers by patents, on the one hand, and on semantic similarity, on the other hand. The research questions we address are: What are the advantages and drawbacks of citations and semantics in mapping science-technology linkages? How do these two signals compare? How can they be combined in order to improve measurement in this domain?

A second purpose of this report is to analyse specific aspects of the relationship between the two fields: what makes certain science more likely to influence inventions than others, and what are the factors that affect the timing of this influence. For these exercises we will make use of citation data. Factors that are considered include: whether the discovery/invention comes early or later in the development of the domain, whether authors are from academia or industry, whether the research benefitted from a research grant or not, whether the publication was open access or not etc. Overall, these models prove quite effective in certain disciplines, less so in others.

The report is based on specific scientific and technological fields: Quantum cryptography, CRISPR and CAR-T Cells. They have in common to be frontier technology domains that rely heavily on science. They differ by their content, as quantum cryptography is software based whereas the others pertain to biotechnology; and by their degree of application: quantum cryptography and CAR-T cells have direct industrial applications (in telecommunication and cancer cure respectively), whereas CRISPR is more of a research tool (modifying the genome by removing certain genes and inserting others). In addition we conducted investigations on mRNA, another biotechnology fields, which are reported in annex 5 (mRNA differs from others by its age, it is older than others, by its size, it is much bigger, and by its internal diversity).

2. Review of the economic literature on the connections between science and technology

This review does not cover the full scope of science-technology linkages; it focuses on the economic literature that addresses two major issues relevant to our research: what sorts of influence does science have on technology? And how can they be detected?

Knowledge transfers

Science is about understanding natural phenomena, while technology is about artefacts and processes that exploit natural phenomena to serve in a controlled manner for purposes chosen by humans. In the words of Nelson (2004, p. 457): “In recent times, virtually all powerful technologies have strong connections with particular fields of science. There is a widespread belief that modern fields of technology are, in effect, applied science, in the sense that practice

is directly drawn from scientific understanding, and that advancing technology essentially is a task of applying scientific knowledge to achieve better products and processes." There are two distinct (although not exclusive from each other) ways that science can serve technology.

1) Understanding of the laws governing a particular field provides a mapping: what the various mechanisms operating in the field are, how they work, on what higher level principles they are based, how they are articulated with each other, how they relate to other, neighboring fields and phenomena. Examples include Pasteur' germs theory of disease (bacteriology, resulting in many drugs and cures), nuclear physics (resulting in nuclear energy), quantum mechanics (resulting in quantum computing), thermodynamics (elaborated after the invention of the steam engine, but which stimulated strongly its development). Each of these theories provided principles explaining a broad range of observed natural phenomena. Based on this understanding, humans could "hijack", manipulate and reframe these natural phenomena for their own purposes. This understanding can be used to interpret and reinforce certain technical effects; to anticipate possible effects and beneficiary developments, i.e., technological opportunities.

2) Certain of the mechanisms that are discovered by scientific inquiry have a direct, specific application; they might need fine-tuning (which is not necessarily simple and might even not be achievable) but they point precisely the direction to pursue. In such cases, where the discovery is nearly actionable for technical purposes it can be both published in a scientific journal and patented. A case in point is CRISPR, a gene editing tool, that was discovered in 2012 and led to both a Nobel Prize for its discoverers and a series of patents (that were litigated, but this is not our subject here). Interestingly, a patent application had been filed already in 2008 for CRISPR, but it was rejected by the USPTO for lack of enablement: the actual mechanism of CRISPR had not been fully deciphered yet (Lander 2016). This example shows that in certain important cases scientific distinctions and patent grants can reflect convergent achievements – a property that we will exploit in our empirical work. Pasteur's quadrant (Stokes 1997) characterizes this sort of research which is dual in its findings, sometimes also in its goals, with both a discovery and an invention.

Of course, science can also help technology through other channels than publication, notably human capital (scientific training helps engineers to master their field and ameliorate existing techniques), technical transfer (highly sophisticated techniques developed for scientific purposes can then be applied to more mainstream matters), technological forecasting (science can point to possible limitations, impossible directions or possible opportunities). But for this study we will focus on the transfer of new knowledge generated by science to new technological inventions, which is well characterised by the two polar cases above.

It is also true that there is a reverse influence of technology on science. Thermodynamics is a case in point, as it was developed by Carnot, Clausius and others as "the science of the steam engine", its explicit aim was to understand why the steam engine was so effective – a phenomenon that existing physics had no explanation for. Although we will not address this issue in this study, it can be noticed that the same tools as we will be using could be adapted to detect this symmetric influence.

In certain respects, the difference between the polar ways presented above is more of degree or breadth than of nature: the first one, mapping, is broader, it is about theories, often abstract, with a wide range of relevance; while the second one, direct application, is about specific, concrete mechanisms and phenomena. Hence the science behind the first, polar way, is more distant from applications, it requires in general more developments, intermediate models, before it can translate into new technology. When asked about the role of science in their

inventions, inventors frequently mention "inspiration": scientific discoveries would have "inspired" technological inventions (Callaert et al. 2014). Inspiration can refer to either type of influence, it can be broad, abstract, a general intuition, or it can take the form of a close analogy between a natural phenomenon and a desired technical effect.

This analysis of the types of influence has impact on the choice and interpretation of indicators in the empirical inquiry (see below).

Quantitative assessments

The framework above has been illustrated by historical cases studies (see Nelson 2004 for a review). There has also been more recently quantitative analysis of the relations between science and technology. Scholars have investigated possible regularities in the relationships between data or indicators reflecting developments in science and in technology respectively. In most studies, science is represented by scientific publications (articles in scientific journals, communications in conferences etc.), and technology is represented by patents. The linkages between the two are represented by citations of non patent literature in patents. There are well recognized limitations in these data, but 1) they are the best available for now; 2) these limitations can be accounted for by appropriate statistical controls. For the specific purpose of analyzing the linkages between science and technology, it can be pointed that papers and patent do not reflect perfectly these two domains respectively: Technological inventions can be disclosed in papers rather than in patents, either because the invention would not be patentable, or for strategic purposes (OECD 2009); and certain scientific discoveries are subject to patent applications, although those applications should be rejected by the patent offices. But the boundaries of the domain where patenting is permitted are somewhat flexible and subject to debate, as the discussions around "software patents" and "gene patents" demonstrated. Hence, identifying articles and communications with science, and patents with technology has to be viewed as an approximation.

In addition, there are also limits in the ability of citation data to capture the linkages between the two domains, as we pointed above.

Ahmadpour and Jones (2017) use citations to calculate a "distance" between patents and scientific publications that depends on the number of chained citations needed to connect two documents. They set the distance D= 1 when a patent cites a publication; the concerned pair of documents defines the "frontier" between science and technology, where patents and publications are directly linked. Then D=2 for a patent citing a patent whose distance to the frontier is 1, or for a publication which is cited by a publication whose distance to the frontier is 1 etc. They calculate distances for 4.8 million USPTO patents and 32 million publications from the WoS. They find that overall 60% of patents can be ultimately connected to a scientific publication, while 80% of publications link to a patent; 21% of the connected patents and 10% of the connected publications are at the frontier. The modal connected science and engineering paper is 3 degrees from the nearest patent. The modal connected patent was 2 degrees from the nearest paper. The distance between publications and patents varies a lot across scientific disciplines and across technical fields. In addition, when taking into account in-text citations (cf. infra), the proportion of documents at the frontier in the same dataset rises to 29% (Marx and Fuegi 2021). Hence, the picture generated by citations, direct and indirect, is that science and technology are highly connected with each other.

Marx and Fuegi (2021) work out and exploit a new dataset (Reliance on Science) with "in-text citations", citations mainly of NLP found in the body text of USPTO and EPO patents. They find that 32.3% of patents have a scientific citation, with 26% having a citation appearing on

the front page and 13.4% in the body text, while 8.8% have a citation appearing both on the front page and in the body text.

Fleming et al. (2019) analyze the reliance of US technology on federal support, as detected by government ownership of patents, and acknowledgement of government support in patents or in scientific publications cited by patents. They find that the share of patents citing a federally supported patent fell from 15.8% in 2003 to 9.9% in 2017, while the share of those citing a federally supported scientific paper "reached a historical high of 10.7% in 2017, up from 6.9% in 2004." It seems that scientific publications are an increasingly important channel for transferring knowledge between government supported research and business innovation.

The overall message that comes out of this literature is that there is a dense web of knowledge transfers from publications to patents, as evidenced by citations (which is not exclusive of symmetric transfer, from patents to publications). However, as we will see, citations are one signal of transfer only, and they might not capture accurately all cases of transfer.

Patterns of science-technology linkage

Not all patents rely on publications to the same extent, and not all publications influence patents to the same extent. Using this variability in the intensity of the link, scholars have inferred the determinants and impact of such knowledge transfers. In particular, they correlated the characteristics of the connected publications and patents with the existence, or not, of a link between them. Although the specific questions addressed and the findings can vary between studies, there appears to be some robust results from this literature:

- Science-inspired patents have larger and more diversified impact than others as measured by citations by other patents (Fleming and Sorensen 2004)

- Science-inspired patents are on average more novel than others (Verhoeven et al. 2016).

- Scientific publications that are cited by patents are also more cited by other publications (Ahmadpoor & Jones 2017);

- Scientific publications which are more novel than others are more likely to be cited by patents (Veugelers and Wang 2019).

- Scientific publications which are more novel are also more likely to be cited by other publications which are themselves cited by patents – an indicator of indirect impact (Veugelers and Wang 2019).

- The link is highly dependent on the scientific field (Du et al. 2019)

- In medical fields, scientific publication that are cited are the most basic and most novel ones (Ke 2020, McMillan, Narin and Deed 2000)

- More novel publications are cited with lower lag than others (Ke 2020)

- Scientific publications can be a good predictor of patents (Ba and Liang 2021).

When science coincides with technology: Publication-patent pairs

Another approach, which has developed more recently, has investigated PPPs. The method consists in identifying publications and patents reporting the same research result, which can

then be considered as both discovery AND invention. Existing studies identified several hundreds of such pairs, which seem quite common in certain fields like biotechnology, although no representative survey has been conducted which would reveal the share of PPPs across scientific and technological fields. Anecdotal evidence would also point to material sciences and computer sciences as other fields where such cases are common. Past studies were aimed at analyzing the impact of patenting on diffusion (Murray and Stern 2007) and the influence of individual researchers' negotiation power in teams (Lissoni et al. 2013), not the connection between science and technology per se. Main conclusions from this literature are that 1) publications which are part of a PPP tend to be slightly more cited than other, similar ones; 2) there are more authors than inventors, on average almost half of article's authors drop out at the patent filing, while there is no evidence that this would be reflective of a difference between the scientific and the technological aspects of the research, which are essentially not separable. A large-scale study has been conducted by Marx and Sharfmann (2024), encompassing all technological fields. They conclude that patenting has beneficial effects on the dissemination of scientific discoveries, especially in the case of lower tier institutions.

PART I: NLP

3. NLP and NPL

We consider three types of signals of science-technology knowledge transfer: citations (of scientific publications by patents: so-called non-patent literature, NPL); semantic proximity (similarity in the text of scientific publications and patents, captured by Natural Language Processing: NLP); and publication-patent pairs (PPPs: publications and patents reporting the same discovery/ invention). Citations have been used for a long time in scientometrics, sociology or economics; what is new is the availability of in-text citations, references that are mentioned in the body-text of patents (Marx and Fuego, 2020). Semantic proximity is a more recent measure, that relies on natural language processing (NLP, a branch of both linguistics and AI), and which has made fast progress recently with the use of so-called "transformers", which are sophisticated neural networks architectures that vectorize texts. Publication-patent pairs have been identified first in the 2000s (Murray and Stern, 2007) and have been the object of a few studies, mainly based on manual identification, which makes it accurate but limited in numbers: This barrier has been crossed recently by using notably NLP, and a large dataset of PPPs is now available (Marx and Scharfmann 2024). These three types of signals reflect mutual information between science and technology: However they are generated in different ways and consequently they reflect different aspects or components of this mutual information, and are affected by different interfering phenomena.

Most existing studies exploit only citations. Although citations convey information, they do not convey all of it, and they also convey noise and biases. This was underlined by Callaert et al. (2014). Based on 33 interviews, they find that citations are incomplete and sometimes little relevant: "30% of patents that were inspired by scientific knowledge do not contain any scientific references. Moreover, if scientific references are present, half of them are evaluated as unimportant or background information by the inventor."

Nagaoka and Yamauchi (2015) asked Japanese inventors whether and which scientific sources were essential and important for the conception or implementation of their invention. Their results indicated that half the patents relied on scientific references in their research processes, among which only 17% revealed such important literature in their patent documents. This study also indicated that in the inventions that cite scientific literature in their patent documents, these

references were not essential for the conception or implementation of 82% of the inventions. Therefore, scientific references were essential for the conception and implementation of R&D in only 18% of inventions.

NPL are scientific publications cited by patents. The main advantages of citations are well known, which is why they have been used in so many studies (see above):

- They benefit from human expertise (applicant, examiner); experts identified a connection between the cited and citing document.

- Data are broadly accessible (from patent offices; now the Reliance on Science database, RoS, presented below).

Their limits are well known as well (see e.g. Jaffe, Trajtenberg and Fogarty, 2000):

- The purpose of citations is not scientific (except for some in the body of the patent), but legal (mapping the boundary of the exclusion right of the patent holder): hence the cited document which serves this legal goal is not necessarily the actual source of the idea in the patent, it could be the one where the idea is most clearly reported, or the one that the examiner knows best; moreover, if a document kills some idea (claim) in the patent, the search can stop, falling short of identifying possibly more relevant sources for this idea.

- Strategic choices of the applicant (e.g. flooding or deceiving the examiner with weakly relevant literature).

- Biases of the examiner (habits, training, "pet references" etc.)

- Priority in citations is given to patents vs. other sources, as patents have a stronger legal status. Hence certain publications, e.g. those which are part of a PPP, might be represented by a patent, then being ignored as such.

- Delay in availability (partial publication after search in EPO; after successful examination in the US), which can take 3 to 5 years after filing at the USPTO.

Some of these weaknesses can be systematic (across technical fields, time periods etc.) due to: competitive behaviours that synchronise, training of examiners etc. Hence certain fields that are judged quite similar regarding their connection with science (e.g. CRISPR and CAR-T cells) can display highly different NPL numbers (as we will see, 80% of patents in quantum cryptography have NPL, against 97% and 98% respectively in CRISPR and CAR-T cells; the average number of NPL per patent having NPL is 6, 73 and 48 respectively – for USPTO up to 2023). The extent to which these differences reflect the actual strength of the influence of science remains to be assessed.

It could therefore be extremely valuable to complement citation data by other types of signals, notably semantic proximity, with the view to broaden and clarify the empirical picture of science-technology linkages.

Semantic proximity can be defined in our context as the similarity between the texts of two documents. Its measurement makes use of recent techniques (NLP: Natural Language Processing) that translate textual information into vectors, called "embeddings". Once texts are put in such a format they become eligible to methods and principles of linear algebra, like calculation of distances. Embeddings are supposed to capture the meaning of text, and distances

between the corresponding vectors would therefore reflect the similarity between ideas embodied in the texts.

By giving directly access to the content of inventions and discoveries, semantic similarity escapes some of the limitations of citations. But it has limitations of its own:

- The techniques used to compile embeddings might not always reflect accurately the meaning of the text.

- Text could sometimes be misleading as the applicant aims to obscure some aspects of the invention or as some vocabulary is idiosyncratic.

- Importance of non-textual information (mathematical or chemical formula, images, graphs), that are more difficult to process.

- Granularity: Citation refers to a specific idea being present in the two documents; while semantic similarity refers to overall proximity between the two texts being compared, that can be entire documents or parts of them. Hence specific commonalities could be averaged out.

- Cost: the computing cost of compiling embeddings can be high; this might entail restrictions on the text to be processed, e.g. to title and abstract instead of the full text. This might aggravate the weaknesses mentioned above.

For measuring semantic proximity, we use the recently developed transformer method. Transformer is a special architecture of neural networks, which allows to capture connections between words separate in a text and to identify words carrying most meaning ("attention mechanism"). The model that we use, SPECTER, was trained to calculate an embedding (vector representation) for each document so that this document is closer, in a vector space, to documents that cite it or are cited by it (Cohan et al., 2020). We apply the language model to the title and abstract of documents (we also apply it to the full text for some tests), and we impose some time structure in order to capture possible influence of science on technology: the paper should be anterior to the patent.

The issue of the correspondence between NPL and NLP is best illustrated by some examples (see annex 1). There are cases where a patent and a publication are both semantically similar and linked by a citation (example 1, which constitutes a "patent-paper pair", as the authors are almost the same and dates of the two documents are close to each other). There are cases where semantic similarity is low but a paper is cited (examples 2 and 3). There are cases where semantic similarity is high but the paper is not cited (example 4). Hence the two indicators can sometimes converge or they can diverge: How often and why is that? The rest of this paper will map this sort of issue and test possible explanations, with the aim of articulating the two indicators so that they could be used jointly, e.g. by combining them in a composite indicator of "influence" or "knowledge sourcing".

The academic literature on NLP applied to linkages between science and technology is still tiny as the technique itself is a very recent. Masclans-Armengol et al. (2024) fine tune a language model on the title and abstract of scientific publications, that predicts whether a scientific publication will be cited or not by a patent (NPL): But they do not track the paper-patent specific correspondence and do not need to calculate embeddings of patents. Ghosh et al. (2024) train a language model on the title and abstract of documents, which predicts the probability that a specific paper would be cited by a specific patent. In the present study we use a pre-trained

language model, which is probably less precise than a model trained directly on the relevant data; but we also attempt to analyse the limitations of such a model, and show that using only the title and abstract of documents has advantages but also important drawbacks.

The core of our investigation is based on NPL and NLP. The comparison of the two sorts of signals is aimed at identifying conditions for and ways of using them jointly. For instance, semantic similarity could be used to correct and supplement citations, which offer both strengths and weaknesses.

What could semantic similarity do to complement citation data?

- Help identifying important sources among the many citations of certain fields.
- Help identifying actual scientific sources in fields where citations are scarce.
- Anticipate future possible citations for newly published patent applications.
- Allow to contextualise and extract more information from citations: e.g. by identifying what specific ideas are being present in the two documents.

4. Data

We constitute thematic corpuses of patents and scientific publications in frontier domains: quantum cryptography, CRISPR, CAR-T Cells (and mRNA, see annex 5).

Patents were extracted from Patstat. We used USPTO patents, as the USPTO has much more patents than other offices in the selected domains (5 to 10 times as many as the EPO, as the US is the largest market for technology and the number of inventors and businesses in these domains is significantly higher in the US than in other jurisdiction, leading to more direct patenting. We used only granted patents, as the citation data is not public for pending applications.

Publications were extracted from OpenAlex (quantum computing) and the Web of Science (CAR-T Cells, CRISPR). OpenAlex offers a better coverage of conference proceedings, that matter more in IT fields like cryptography, while the WoS offers more curated metadata.

The corpuses are built by using keywords and CPC categories for patents; using keywords and "concepts" (for OpenAlex) for scientific publications (see Annex 2). For instance, in the case of CRISPR publications and patents, the key words are: CRISPR, Cas9; and the CPC category is: C12N2310/20. These "core corpuses" correspond to the technology itself and the science that underlies it.

In addition, we also constituted "non core" corpuses with documents generated by citations (patents citing publications from the core, publications cited by patents from the core). That allows to capture cross-domain knowledge transfers: to trace the sources of a particular technology that are outside the corresponding scientific field (non core publications), and the inventions that make use f the technology while not being in its core, notably applications to specific uses (non core patents). Citations are identified with the Reliance on Science database (Marx & Fuegi 2020).

For reasons of data availability, we had to make restrictions on the corpuses for all or some of the statistical exercises (e.g. only granted patents have NPL; some of the RoS cited publications

could not be retrieved from the publications databases; some publications, while referenced in the databases, have no text available so that they are not eligible to semantic treatment).

Table 4.1: Number of patents, publications NPL per corpus			
	Quantum Cryptography	CRISPR	CAR-T Cells
Patent families (core)	1182	3249	1844
	(872 granted)	(931 gr.)	(561 gr.)
Patent families (citing)/non core	1674	3123	1657
Publications (core)	27607	41244	15775
Publications (cited)/non core	4166	26447	10556

As reported in table 4.1, the core corpuses of patents vary between 1200 and 3200 patents, due to the scope of the concerned fields and to their age. Publications are far more numerous than patents in all domains.

The time distribution of corpuses (see Figures 4.1 to 4.3) shows these fields are made of recent documents: they emerged in the 2000s or 2010s. As a result, the core corpuses are truncated, which affects notably the citation data (coming patents will cite many past publications already included in these statistics, some scientific publications not currently cited will be cited in the future).

Figure 4.1: Distribution of documents over time, quantum cryptography.

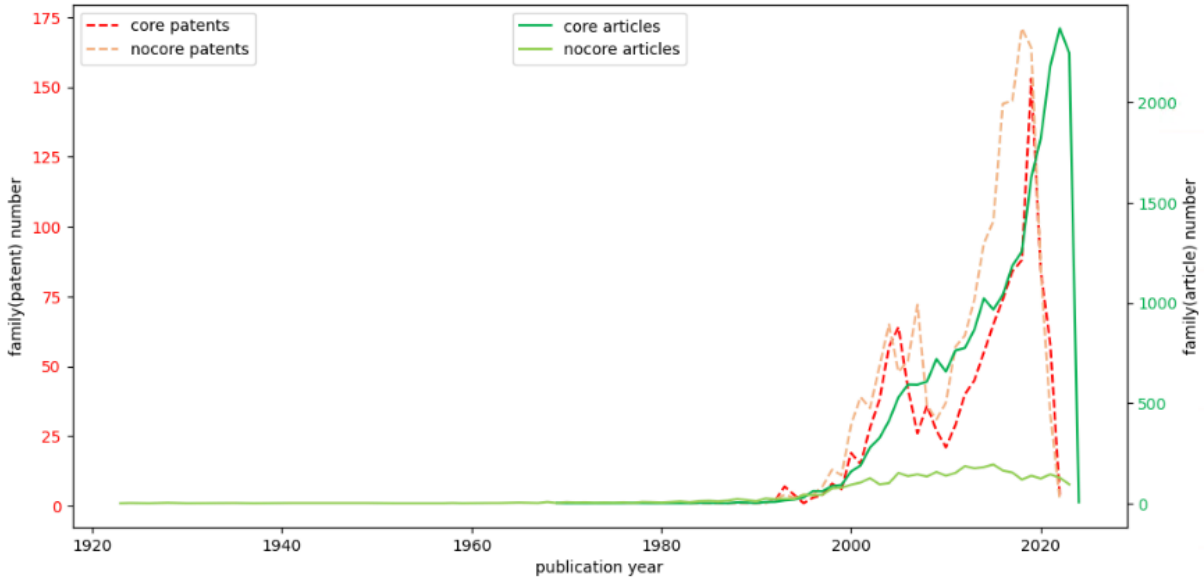


Figure 4.2: Distribution of documents over time, CRISPR.

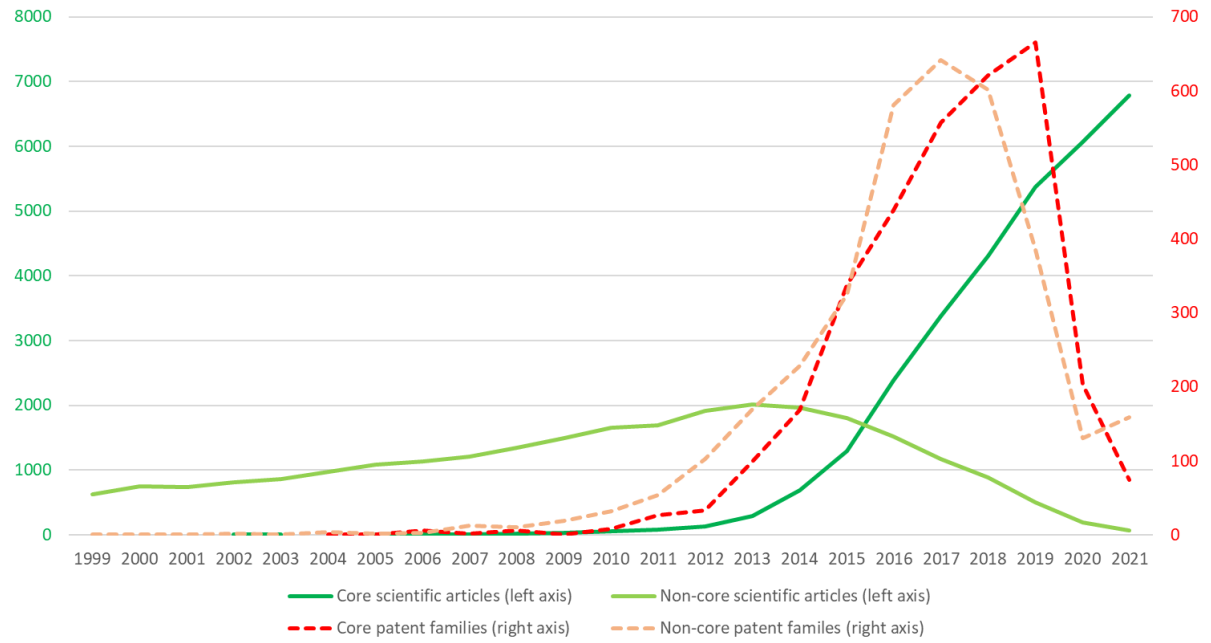
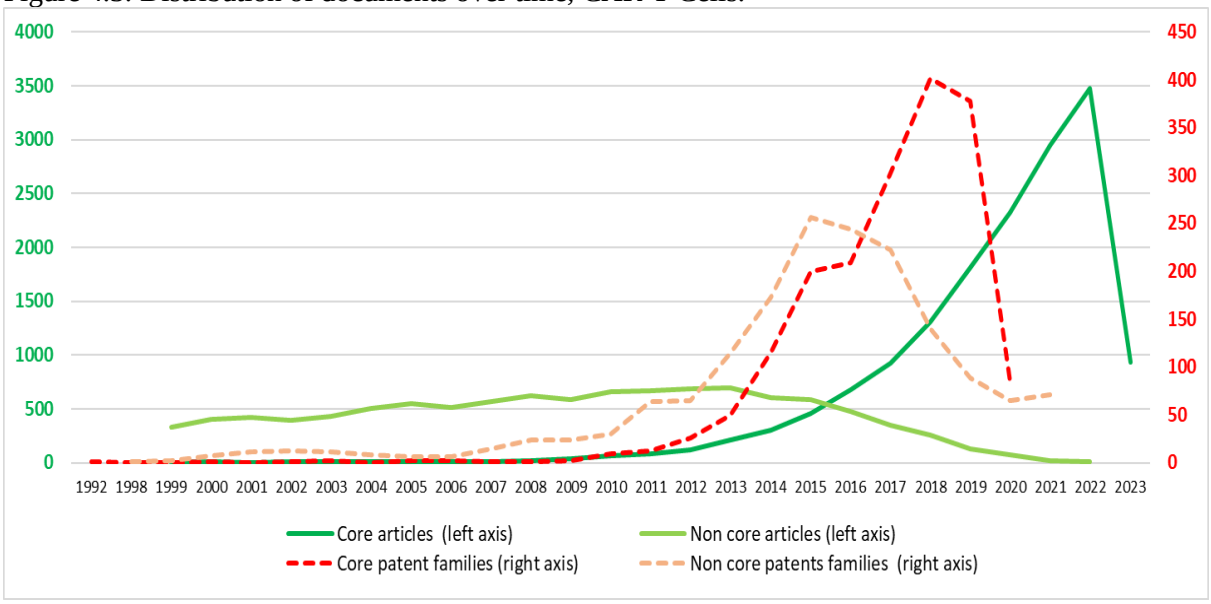


Figure 4.3: Distribution of documents over time, CAR-T Cells.



5. NPL

As compared with other fields, as reported in the literature, a high proportion of patents in our domains have NPLs (82 to 98%), and they have numbers of them (up to 73 per patent; see table 5.1). NPL are particularly numerous for the biotech fields, with a maximum in CRISPR. Numbers in quantum cryptography are much lower, pointing to a different type of liaison between science and technology in this field.

Table 5.1: NPL of core granted patents			
	Crypto	CRISPR	CAR-T
Number of granted patents with NPL (and % among granted patents)	720 families (82.6%)	902 families (97%)	551 families (98%)
Average number of NPL per patent having NPL	9.68	73.10	48.09
Median number of NPL per patents having NPL	5	32	27
Maximum number of NPL for one patent	109	1126	373

These differences in averages are due mainly to the fact that a significant proportion of patents in the biotech related fields have a high or very high (several hundreds) number of NPL (Figures 5.1 to 5.3). The distribution of NPL per patent is quite skewed, as shown by the fact that the average is much higher than the median. The maximum number of NPL per patent is above 1000 for CRISPR. This reflects certainly the deep intermeshing of technology and science in these fields. It is however dubious that an invention might rely *directly* on hundreds of scientific publications and such high numbers point to a likely mix of significant and less significant references, which blur the actual landscape of knowledge transfers.

Figure 5.1: Distribution of patents by the number of NPL – Quantum cryptography

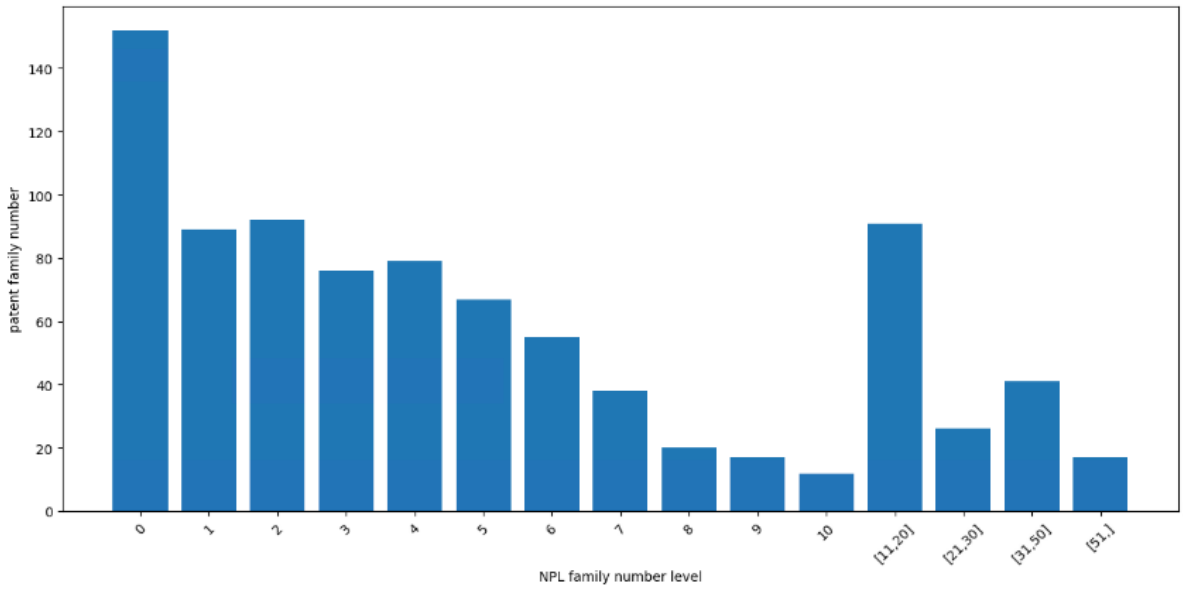


Figure 5.2: Distribution of patents by the number of NPL – CRISPR

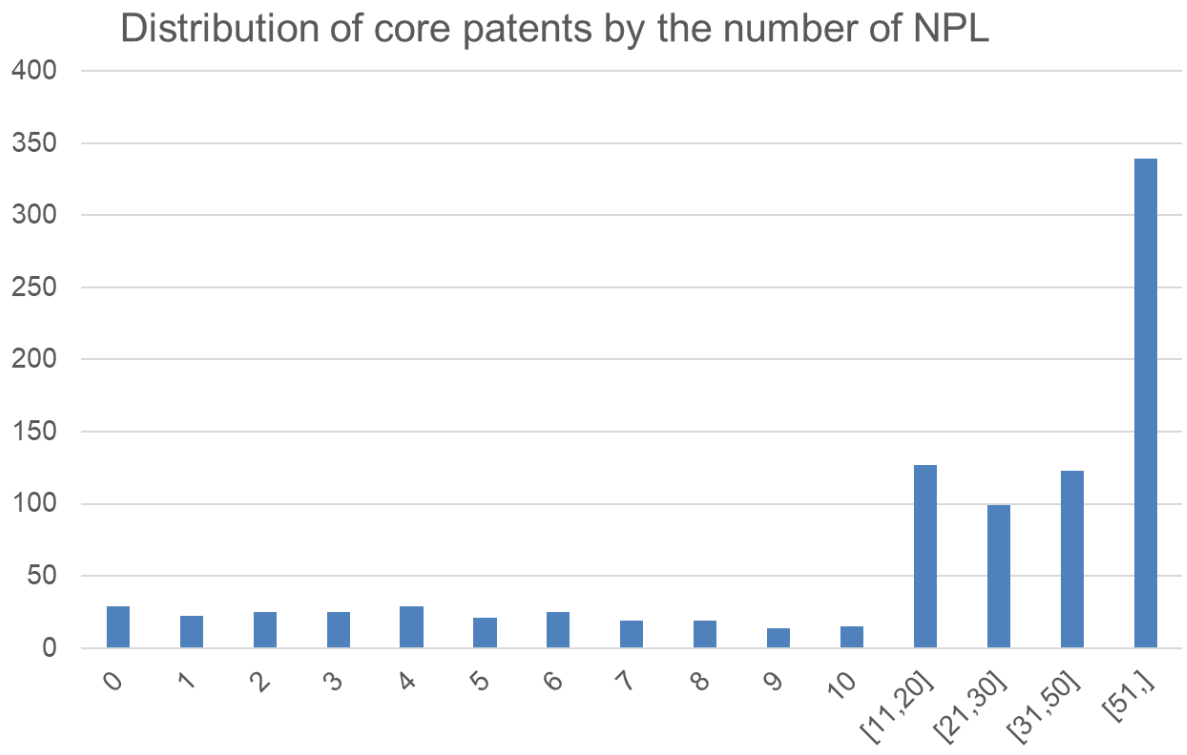
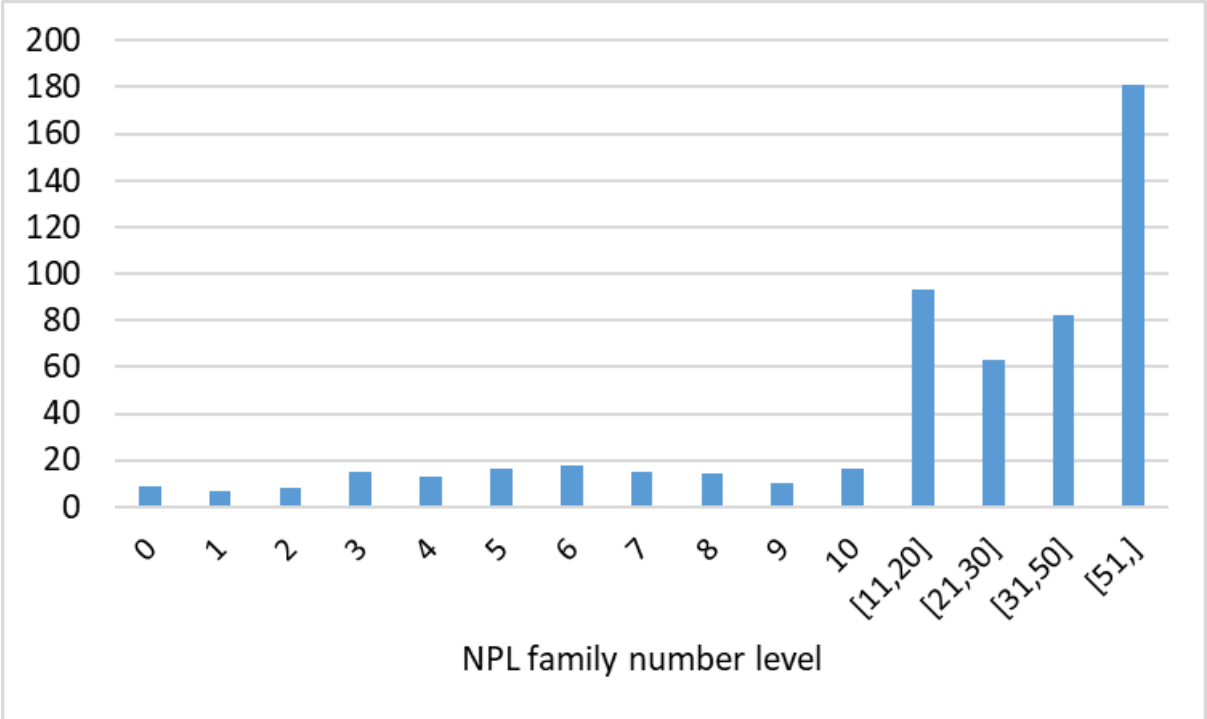


Figure 5.3: Distribution of patents by the number of NPL – CAR-T Cells



6. NLP

The semantic analysis of the corpuses consists in investigating the proximity between publications and patents. For that purpose we use “embeddings”, which are representations of

the text of documents in a vector space. In these spaces proximity is interpreted as a reflection of a similarity in the content and meaning of documents. Embeddings are calculated using "language models", most often neural networks pre-trained to predict the missing words in a text. The type of language model we used is a BERT based model, which takes into account the deep structure of texts (as reflected in relations between remote words). We chose SPECTER2-CLS, a transformer trained on scientific articles with a loss function aimed at contrasting the embeddings of cited and non cited documents.

SPECTER (Cohan et al., 2020) is a model finetuned with citation relations among articles, based on SciBERT (Beltagy et al. 2019), a Bert model pre-trained exclusively on scientific articles' full-text. This finetuning process allows the model to better represent the semantic similarity among scientific text by integrating the citation information during its training, with the assumption that an article directly cited by the focal article should be semantically more similar than articles not cited or only being cited indirectly. After training, the model needs only the text of documents for inferring embeddings, it does not need the citation information. SPECTER2 (Singh et al., 2022) is the further elaboration of SPECTER, which adds adapters allowing the model to generate multiple embeddings for different tasks in the new benchmark SciRepEval and to improve the model's performance on each task.

We chose to use SPECTER2, with the adapter trained for proximity retrieval tasks, trained on more recent articles. This adapter corresponds to our task: testing the similarities among articles or patents. As a typical choice, we use the output vector of the special token CLS as the embedding representation of the whole input text.

We have also compared SPECTER2 with another model trained on patents with the same strategy: PaECTER (Ghosh et al., 2024) on our PPP (patent-paper pairs) test for all the available pairs in our data source (RoS, PATSTAT, OPENALEX and S2OCR), which is 9860 PPP pairs for 8317 patents. Since the article and patent in a PPP pair should be based on the same invention, this article should be the nearest citation of this patent among all its NPL. Results are reported in table 6.1. The results do not show much difference between these two models, only a slightly better detection of similarity by SPECTER2. Therefore, we conclude that our analysis is not affected by the chosen model.

Table 6.1: Comparison of PPPs with semantic proximity - for title and abstract

	Number of PPP patent documents with complete data	Average number of NPL per PPP patent	Average rank of the PPP scientific publication among cited publications of the PPP patent	Median rank of the PPP scientific publication among cited publications of the PPP patent	Number of patents for which the PPP publication is the closest NPL
Specter2 (bert base)	8317	34.7	2.13	1	6454
Paecter (bert large)	8317	34.7	2.40	1	6029
Patspecter (bert base)	8317	34.7	2.26	1	6185

We measure semantic proximity with the following steps:

1. We calculate embeddings of the title+abstract of all documents, using Specter2-CLS. The choice of title and abstract is based on the fact that these parts of the text of a document are supposed to encapsulate its overall meaning. The alternative would be to use the full text, something we will do later. However, the full text is available only for a subset of documents (many are behind paywalls), and its processing requires far more computing power.

2. We calculate the cosine distance between all pairs (patent, publication) of the same corpus *in which the patent is posterior to the publication* (a logical condition of influence and to make NLP comparable to NPL).

3. We identify the "nearest neighbours" (NN), i.e. the closest publications of each patent: NN1, NN5, NN20, NN100 are the cosine similarity between a patent and the 1st, 5th, 20th and 100th publication respectively ranked by declining cosine similarity with itself.

Influence is detected through semantic proximity between documents (scientific publications and patents) and anteriority.
Comments: 1) Frontpage citations are semantically closer to the citing patent than in-text; 2) Core patents are closer to publications than non core patents.

Figure 6.1: Distribution of semantic proximity (NN1, 5, 20 and 100, here labelled "ppv") between patents and scientific publications, quantum cryptography.

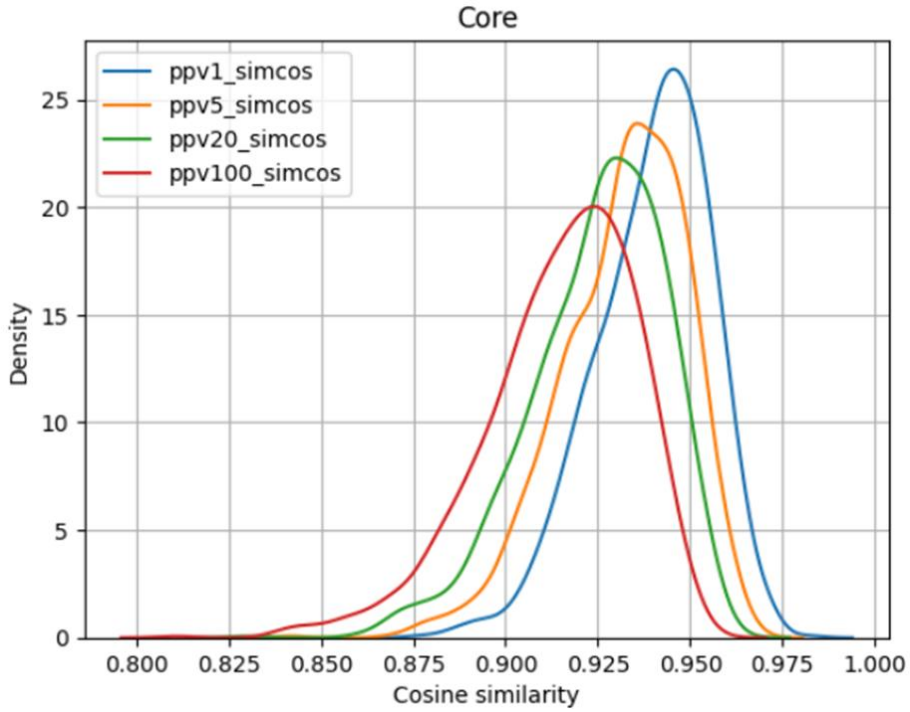


Figure 6.2: Distribution of semantic proximity (NN1, 5, 20 and 100) between patents and scientific publications, CRISPR.

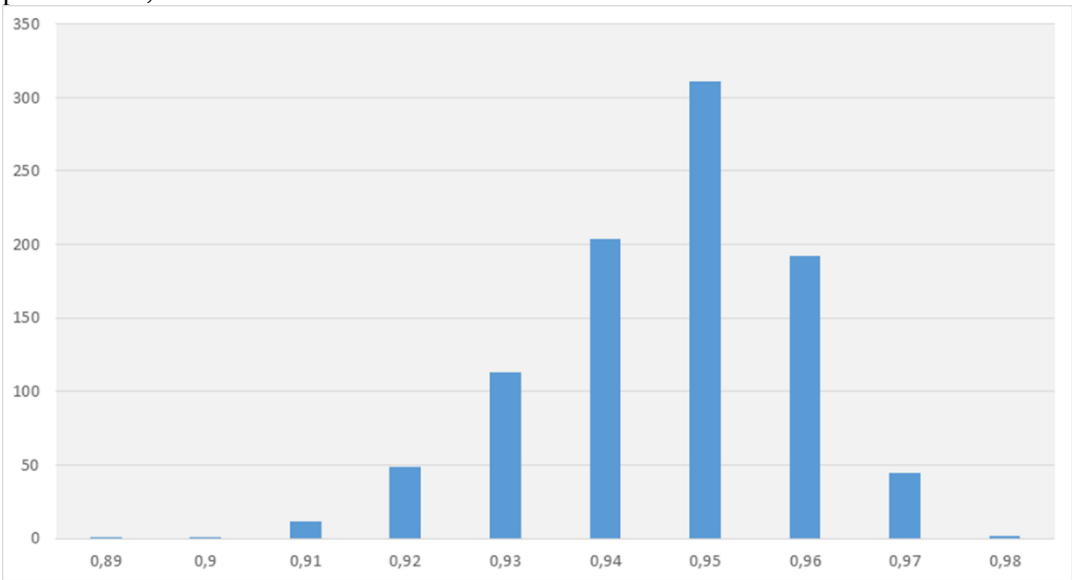
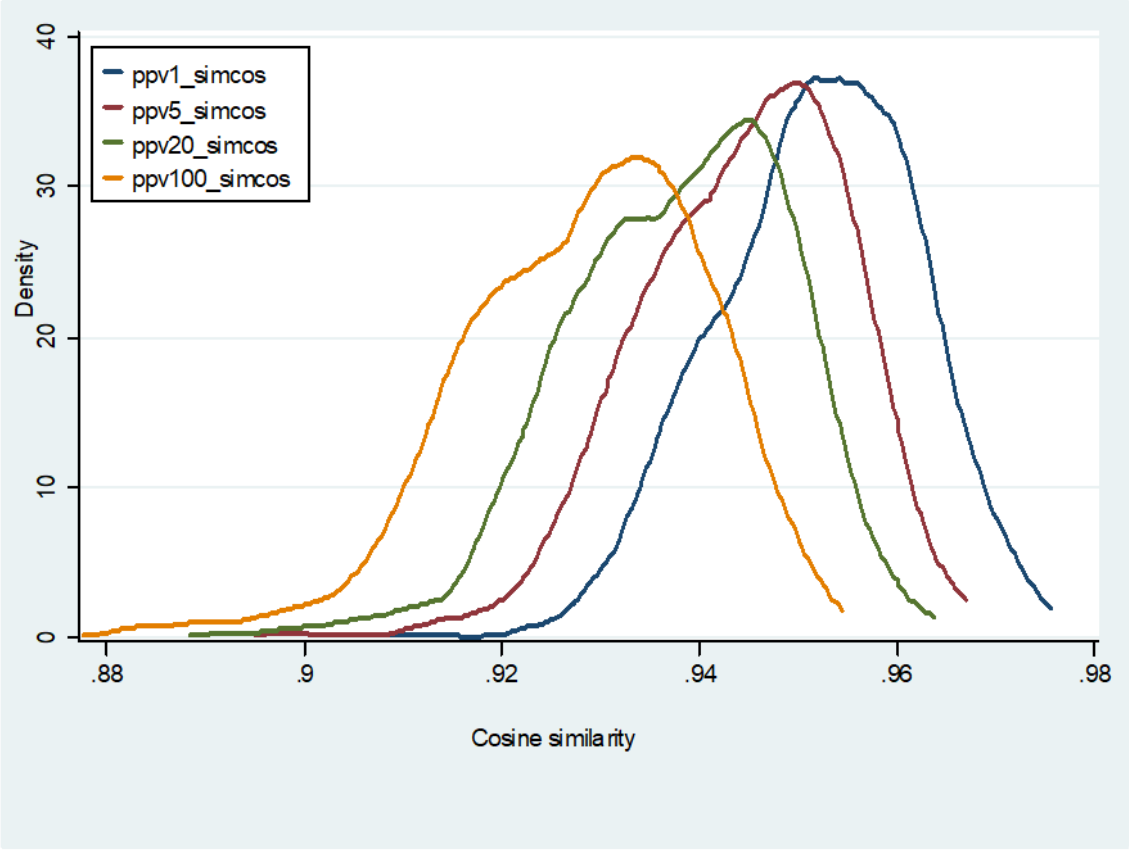


Figure 6.3: Distribution of semantic proximity (NN1) between patents and scientific publications, CAR-T Cells.



Figures 6.1 to 6.3 show the distribution of semantic similarities in the three corpuses, with the proximity between patents and their 1st, 5th, 20th and 100th nearest neighbours (NN1, NN5, NN20 and NN100 respectively). Proximities are highly concentrated: for the four corpuses the maximum of NN1 is at 0.94-0.95, and 95% of patents have their nearest neighbour between 0.92 and 0.96 (the distribution is not symmetric around its maximum). This concentration might come from the fact that the language model, SPECTER, has been trained on documents from

all scientific domains, which are widely diverse, so that within-domain distances are minimized (but not necessarily distorted!): and we compare here documents pertaining to a same domain. Then in the following even small differences in proximities could be interpreted as possibly reflecting significant differences in the actual semantics of documents.

7. Comparing NPL with NLP: Title and abstract

To what extent do NPL and NLP coincide? As the types of data generated by citations and by semantic distance differ deeply (citations are qualitative, dichotomic data whereas semantic distances are real numbers in the [0, 1] interval), this question is statistically complex and can be addressed in multiple ways.

7.1 Comparison at aggregate level

So we reformulated it in a more specific way: do citations and semantic proximity provide similar signals regarding science-technology linkages? Or, equivalently: are patents deemed closer to science according to NPL also deemed closer according to semantic distance?

The indicators often used by economists to measure the proximity of a patent to science is the number of NPL of the patent. Hence, we correlated the number of NPL of patents with their semantic proximity to science (NN1, 5, 20 and 100). We distinguished between core patents (pertaining to the technical field) and non-core patents (those which cite publications from the thematic corpus) on the one hand, core publications and all publications (including those which are not in the core but are cited by patents from the corpus) on the other hand. Tables 3 report the Pearson correlation for CRISPR and CAR-T Cells.

Results are reported in tables 7.1 and 7.2. Most correlations are significant but not very high, around 0.2. Non significant correlations are those between non core patents and core publications (for CRISPR) and between core patents and all publications (CAR-T Cells). The highest ones are between core patents and core publications, which is consistent with the idea that both NPL and NLP are able to reflect in the same way knowledge transfers *within domains*. For cross-domain transfers, NLP based on the title and abstract might have more difficulties as the cited and citing document might address different topics, or use different vocabularies that tend to hide the specific idea that is being transferred.

Table 7.1.a: Pearson Correlation between the number of NPL and semantic proximity of patents – CRISPR

	Core patents		**Non Core patents**	
(Core)	**Pearson**	**p-value**	**Pearson**	**p-value**
NN1	0.192	3.3E-09	0.223	7,4E-39
NN5	0.208	1.5E-10	0.267	1,4E-55
NN20	0.207	1.9E-10	0.257	1,2E-51
NN100	0.189	6.2E-09	0.239	9,8E-45

Table 7.1.b: Pearson Correlation between the number of core NPL and semantic proximity of patents – CRISPR				
	Core patents		**Non Core patents**	
(Core)	**Pearson**	**p-value**	**Pearson**	**p-value**
NN1	0.295	7.2E-20	0.007	0.668
NN5	0.296	5.2E-20	-0.022	0.208
NN20	0.284	1.8E-18	-0.037	0.035
NN100	0.247	3.0E-14	-0.053	0.002

Table 7.2.a: Pearson Correlation between the number of NPL and semantic proximity of patents – CAR-T Cells				
	Core patents		**Non Core patents**	
	Pearson	**p-value**	**Pearson**	**p-value**
NN1	0,070	0,098	0,210	2,20731E-15
NN5	0,069	0,100	0,210	1,71777E-15
NN20	0,056	0,187	0,202	2,08864E-14
N100	0,039	0,354	0,181	7,75254E-12

Table 7.2.b: Pearson Correlation between the number of core NPL and semantic proximity of patents – CAR-T Cells				
	Core patents		**Non Core patents**	
(Core)	**Pearson**	**p-value**	**Pearson**	**p-value**
NN1	0,225	9,77E-08	0,143	2,31E-07
NN5	0,216	2,91E-07	0,151	4,57E-08
NN20	0,203	1,50E-06	0,147	1,09E-07
NN100	0,176	3,20E-05	0,131	2,14E-06

7.2 Comparison at individual level

Another indicator that compares citations and semantic proximity is the percentage of NPL which are among the nearest neighbours of the citing patent. It can happen that a patent would be semantically close to publications and would have many NPL, but these NPL would not be publications to which the patent is close. It is therefore necessary to conduct a comparison of NPL and semantic similarity at the level of individual publications if one wants to qualify how the two measures of proximity relate to each other.

Table 7.3: Percentage of NPL citations which are among the nearest neighbours of the citing patent - CRISPR - Core patents (granted) (%)

Nb. of NPL	Nb. of families	N=1	N=5	N=20	N=100
1	3	0	0	0	0
2	5	10	20	40	40
3	5	7	13	13	27
4	7	4	14	21	43
5	16	3	5	6	15
6	7	5	12	14	36
7	5	6	11	14	20
8	5	0	0	3	5
9	9	1	5	10	20
10	6	3	13	23	25
[11-20]	70	1	4	10	19
[21-30]	62	1	2	5	12
[31-50]	102	1	1	4	10
Sup 50	254	0	1	2	7
All	556	1	3	5	12

Note: Patents are classified by the number of NPL they have. The table reads, for instance: 13% of the NPL of patents having 3 NPL are among the 5 closest documents to the patent that cites them.

	Table 7.4: Percentage of NPL citations which are among the nearest neighbours of the citing patent - CAR-T Cells- Core patents (granted)				
Nb. of NPL	Nb. of families	N=1	N=5	N=20	N=100
1	7	0	0	14	14
2	8	0	0	0	12
3	15	0	0	13	20
4	13	0	0	4	8
5	16	2	7	15	27
6	18	0	1	5	8
7	15	3	4	8	10
8	14	1	4	4	8
9	10	1	3	10	26
10	16	0	2	6	20
[11-20]	93	1	3	5	12
[21-30]	63	1	2	3	9
[31-50]	82	0	1	4	9
Sup 50	181	0	1	2	6
All	551	1	2	4	10

Overall, 12% and 10% of the NPL of patents having 50 or more NPL are among the top 100 nearest neighbours, in CRISPR and CAR-T cells respectively. Hence the coincidence between NPL and NLP does exist but it is rather weak.

7.3 Conclusion

Overall, these results point to a partial coincidence of citations and semantic proximity: cited NPL are most often but not always among the semantically closest publications to a citing patent, although rarely the single closest one; and patents that are semantically closer to scientific publications tend to have more NPL than other patents. This weak correspondence can be explained by two alternative (but non-exclusive) hypotheses:

> H1: NPL point to *specific aspects*/ideas of the two documents, while semantic proximity between abstracts reflects *overall similarity* in the topics (subject matter) addressed by the two documents. When applicants or examiners chose a NPL it often refers to a commonality between one claim of the patent and one idea or discovery in a publication, not to a global similarity between the documents. It can also happen that the documents are quite different overall, in the case for instance of the transfer of a specific idea or method across different domains (e.g. a specific CRISPR technique being implemented in modifying genetically a specific living organism): in such a case, the patent will include a lot of material that is not connected to the scientific publication, and vice versa.

H2: NPL are drawn from a broader pool of publications, all of which are NOT necessarily among the closest to the citing patent: it could be the most accessible to the applicant of the examiner, the most often cited ones ("Matthew effect"), the most recent ones, literature surveys which encapsulate many variegated ideas etc. A specific case would be that certain NPL are strategic weapons used by the applicant, with only a slight connection with the citing patent: either the applicant simply wants to cover her back regarding the "duty of candour" rule of patent law, preferring to add irrelevant references to the risk of overlooking a relevant one; or the applicant aims at flooding competitors and the examiner with many weakly relevant references in order to deceive them about the actual sources.

In the following we will test separately these two hypotheses.

8. Comparing NPL with NLP at granular level

In order to test hypothesis H1, we will 1) calculate a distance between documents which best encapsulate the notion that a NPL should reflect a commonality in at least one idea-component of the two concerned documents, and 2) check whether this distance (that we label "granular") is a better driver of NPL than the similarity between titles and abstracts that we calculated before (that we labeled "aggregate distance").

How to estimate the proximity between the closest component-ideas of a publication and a patent respectively? We took the following approach. First, we hypothesise that one idea corresponds to one claim in a patent (which is legally correct), and top one paragraph in a paper. Second, we calculate the embeddings for each claim of a patent and each paragraph of a paper. Third we calculate the semantic similarity (cosine distance) between all claims of the patent on the one hand, and all paragraphs of the paper on the other hand. Fourth we take the minimum of these distances for the concerned pair, which reflects the similarity between the closest paragraph and claim of the two documents. Once we have this bilateral distance between the patent and the paper, we compare its predictive power of a NPL relation with the predictive power of the aggregate distance, using a "twins" method.

The data source for publications full text is the S2ORC dataset: https://allenai.org/data/s2orc and the data source for patent claims is Patstat. Unfortunately, not all publications identified in OpenAlex or the WoS could be found in S2ORC, we therefore experience attrition of the data set for this test, but there is no indication that this generates a selection bias.

Let's call "minimal distance" (MinD), the lowest granular distance (highest granular similarity) for each pair: it is the distance between the most similar pair (paragraph, claim) of the paper and patent pair. We interpret MinD as the distance between the closest ideas of each publication.

For each pair (P, A) (patent, publication) with a NPL relation we select a twin pair (P, A') where A' is a publication that has the same aggregate distance from P as A, but is NOT NPL (i.e. not cited by P). The question is then: are the closest ideas between the patent and the NPL closer than the closest ideas between the patent and the NO NPL publication, which could justify the citation (although we could not check whether the citation is specifically based on this particular idea).

We apply the Wilcoxonn signed-rank that reports the statistical significance: MinD (P, A) < MinD (P, A').

Results are reported in table 8.1.

Table 8.1: Tests of relative granular proximity of cited and non cited documents to patents			
	Crypto	**CRISPR**	**CAR-T Cells**
Number of pairs	5582	58160	13980
Pairs with MinD NPL < MinD No NPL	3561 (63.8%)	31865 (53.4%)	7518 (53.8%)
Wilcoxonn signed-rank test	<0.0001	<0.0001	<0.0001

Table 8.1 shows that, as predicted from H1, granular proximity is significantly correlated with citation once we control for aggregate proximity (granular proximity of NPL is higher than granular proximity of no-NPL having a similar aggregate proximity). For two papers having a similar aggregate distance with a patent, one paper being cited and the other not being cited, the cited paper has higher granular similarity to the patent than the non-cited one. Hence hypothesis H1 is confirmed, especially in the case of quantum cryptography, although it is far from explaining entirely the discrepancy between NPL and NLP (there are still many pairs for which semantic granular semantic similarity is lower in the case of NPL than for their no NPL twin). This shows that the proximity between specific component-ideas of two documents is a better predictor of citation than proximity between titles and abstracts, which rather point to similarity in the overall topics of the two documents.

9. Semantic proximity and the number of citations

We will now test hypothesis H2, the "strategic citation" case. If there are strategic behaviours behind the inflation of citations of certain patents, one would expect the relevance of citations to a particular patent to decline with the number of citations in this patent. Increasing the number of cited documents, be it for reasons of cautiousness or for flooding examiners, implies to include publications that are less and less relevant to the protected invention. We measure relevance by semantic proximity, and test whether the average semantic proximity between a patent and its NLP declines with the number of NLPs cited in the patent.

Figure 9.1: Average semantic similarity (title+abstract) of patents to their NPL, ranked by the number of NPL of a patent - Quantum Cryptography (e.g. the average similarity of patents having 1-10 NPL to their NPL is 0.879)

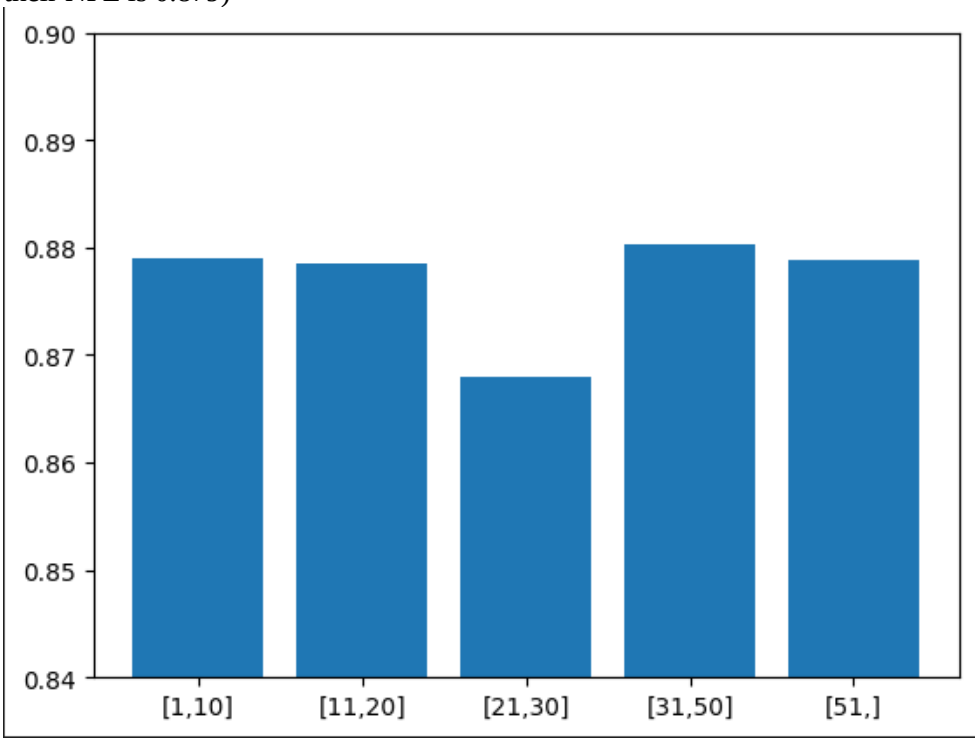


Figure 9.2: Average semantic similarity (title+abstract) of patents to their NPL, ranked by the number of NPL of a patent - CRISPR (e.g. the average similarity of patents having 1 to 10 NPL to their NPL is 0.900)

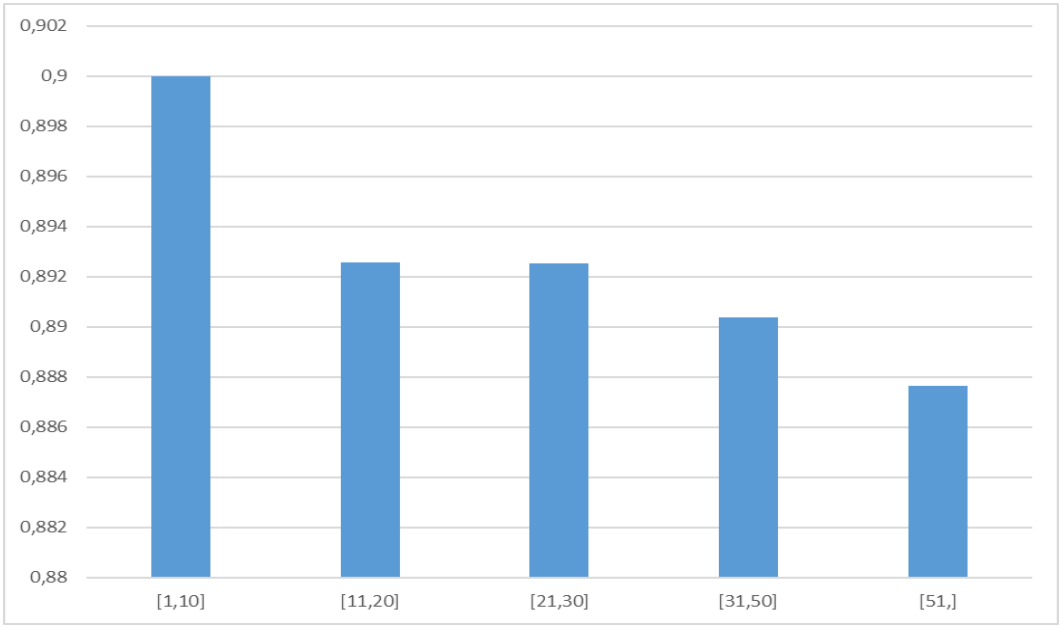
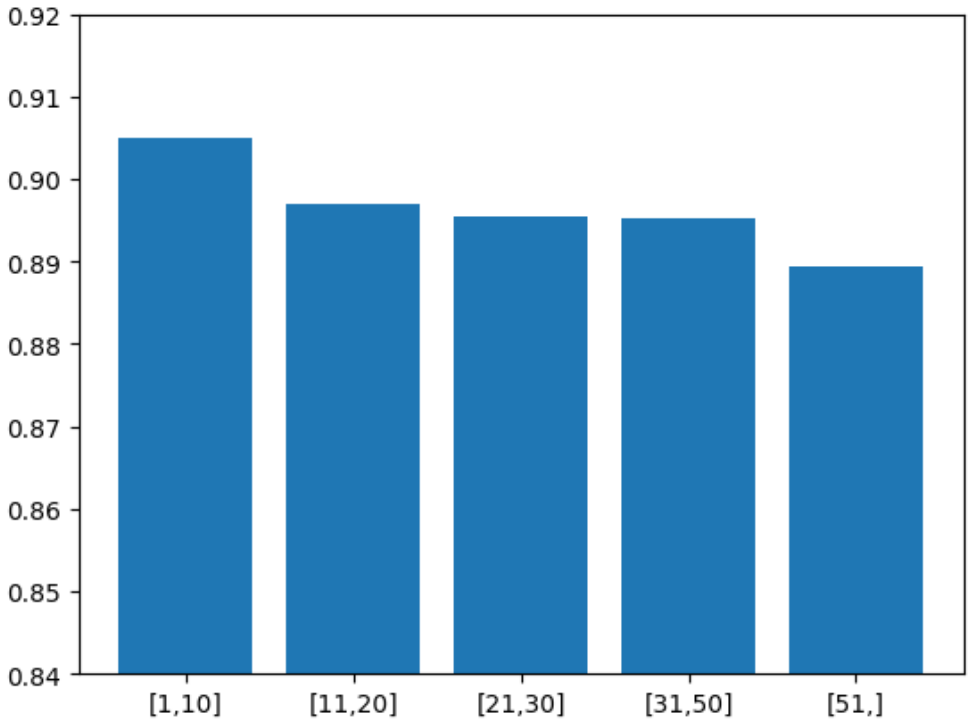


Figure 9.3: Average semantic similarity (title+abstract) of patents to their NPL, ranked by the number of NPL of a patent – CAR-T Cells (e.g. the average similarity of patents having 1 to 10 NPL to their NPL is 0.905)



For CRISPR and CAR-T cells there is clearly a decline of the average semantic proximity between a patent and its NPLs when the number of NPLs increases (figures 9.2 and 9.3), which goes in the direction of providing support to hypothesis H2. This is not the case for quantum cryptography (figure 9.1). One important difference between these fields is that the number of NPL is significantly lower in quantum cryptography than in the two other fields, pointing to a more selective behaviour in the former than in the latter. These numbers suggest that in fields where the number of NPL is particularly high, these NPL correspond partly at least to strategic choices. Hence hypothesis H2 is likely to be valid at least in certain fields, contributing top weakening the connection between NPL and NLP.

10. Combining NPL and NLP

The partial coincidence of citations and semantic proximity and its interpretation point to a complementarity between the two types of signals, and suggest their possible combination to form indicators better reflecting knowledge transfers between science and technology. One such combination is the closest scientific publications among the cited literature of a patent. Restricting the set of sources to cited documents allows to exploit human expertise, while selecting the semantically closest ones allows to sort out those who come from strategic choices. We have tested such an option by experimenting on two special types of patent-publication couples: so-called “patent-publication pairs”, PPPs, on the one hand, and publications both cited by the USPTO and EPO on the other hand. These two sets have in common to be composed of publications which are likely to be the closest to the citing patent.

10.1 PPPs

PPPs are pairs made of a patent and a publication covering the same discovery-invention. They have been extensively studied by the economic literature as they represent a unique bridge

between science and technology, a domain where the two fields are overlapping. It is a reasonable assumption that the patent and publication which are part of a PPP must be extremely close semantically. Possible limitations to this include: different modes of drafting; one patent might correspond to several publications or vice versa, hence resulting in only partial overlap. However the robustness of this assumption has been verified in a number of past studies (e.g. Murray and Stern, 2007).

For this exercise, we used a set of PPPs recently included in the RoS database, identified using metadata (authors names, close dates) and semantic techniques (Marx and Sharfmann 2024). RoS includes two sets of PPPs: one set is based on citation information, the other one ignores this information. This two pronged approach is necessary as publications are not necessarily cited in the patent that uses their discovery, notably for legal reasons (a valid patent must have priority in publication, except for "provisional patents" in the US, but not all patentees use this legal road). The semantic similarity used in the identification of PPPs concerns the titles and abstracts of the documents, what we have called here "aggregate similarity". As a consequence, this particular distance cannot be used in any test applied to this data set.

In order to determine the Power of semantic similarity to select the most significant citations, we calculated the ranking of all citations of a patent in a PPP and determined the ranking of its paired publications. It is expected that the PPP publication would most often be ranked first. This test has very high requirements in terms of data and the number of observations on which it can be implemented is restricted accordingly. The following conditions have to apply (see annex for more detail):

- The patent and publication must be in a PPP in RoS
- The title and abstract of the publication and of the patent must be available
- The patent must have other citations beyond the PPP publication (otherwise the ranking would be meaningless).
- The claims of the patent and the full text of the publications (PPP and other citations) must be available.

A further issue is that certain patents have several PPP publications (this happens when the patented discovery was published in more than one article).

We did that testing on the three thematic corpuses and on the entire USPTO dataset.

Table 10.1: Testing the semantic proximity ranking of PPPs among NPLs

Corpus	Number of patents satisfying conditions	Average number of NPL per patents	Average rank of the PPP publication among NPLs	Number of PPP publications ranked first (%)
Crypto	57	8.69	3.01	34 (60%)
CRISPR	173	48.14	26.50	26 (15%)
CART-T	53	5.82	1.60	37 (70%)
ALL	8199	10.24	3.21	4956 (60%)

Results are reported in table 10.1. Except for CRISPR, results are highly consistent across the technical fields and at the aggregate level: in around 60% of PPPs, the cited publication of the PPP is the closest to the patent. The weaker result for CRISPR is related to the fact that many CRISPR patents are part of more than one PPP, hence at least one of the concerned publications cannot be ranked first.

10.2 USPTO-EPO families

The citation practice differs significantly across jurisdictions. The "duty of candour" of the USPTO leads applicants to cite extensively all publications related, even weakly, to the patent application, as the legal sanction in case of missing a relevant reference is high. Conversely, for instance, the EPO encourages applicants to keep to citations with a significant, material connection with the application. Hence, it is to be expected that 1) the number of references in EPO applications will be lower than in USPTO applications; 2) the average relevance of references to the citing document will be higher. We focus on applications pertaining to a same patent family having both a USPTO and an EPO member, and we compare the NPL profile of the two members. The first assumption is verified for NPL in our three corpuses (the number of references in EPO documents is less than half the number in USPTO: see tables 10.1 and 10.2). Based on the second assumption, a test of the ability of semantic proximity to capture the most relevant citations is to measure the relative proximity of citations by USPTO-only and USPTO-EPO citations with the citting patent. Results of the test are reported in table 11.2. It turns out that in the three corpuses NPL of the USPTO which are also NPL of the EPO family member are often the closest scientific publication to the citing patent.

Table 10.2: Testing the semantic ranking of EPO cited publications among NPLs

Corpus	Number of patents satisfying conditions	Average number of NPL per patents	Average rank of the EPO cited publication among NPLs	Number of EPO cited publications ranked first (%)
Crypto	102	3.26	1.49	53 (52%)
CRISPR	931	27.69	4.22	241 (26%)
CART-T	570	7.75	2.82	268 (47%)

A comparison of tables 10.1 and 10.2 shows that the EPO test captures more patents than the PPP test (i.e. more UPSTO patents are part of a family including an EPO member than part of a PPP). It also suggests that being part of a PPP is more significant than being part of a family with an EPO member in the case of quantum cryptography and of Car-T cells, but less so in the case of CRISPR.

11. Conclusion on semantics

The first part of the report investigates the correspondence between citations (NPL) and semantic proximity (NLP) of scientific publications and patents (title and abstract), which are two different signals of science-technology linkage. We find that NPL and NLP correlate significantly but weakly: semantic proximity coincides with citation, but only so far. The report then investigates two possible explanations to this discrepancy, related to limitations of NLP and NPL respectively.

First, we calculate semantic proximity at a more granular level (full text and claims), and find that at this level the correspondence with citations is stronger than at an aggregate level (title and abstract). This is interpreted the following way. Aggregate semantic proximity comes from similarity in the topics addressed in the two documents, that are presented in broad terms in the abstract; whereas granular similarity comes from a close proximity between at least one specific idea present in the two documents, a characteristic that is better reflected in citations. A direct implication of this finding is that studies or language models that are trained to analyse or predict citations would better be trained on the full text of documents, rather than the title and abstract as it is often, the case.

As for interpreting the finding in analytical terms, the two types of signals (granular similarity and citations on the one hand, aggregate similarity on the other hand) correspond to the notions of mapping and of sourcing ideas respectively: science is recognized 1) as a source of both maps of the laws of nature, that can inform downstream inventive activity, and as 2) a source of specific ideas, that can be implemented in inventions after some modifications (Nelson 2004). If confirmed, this view would point to complementarity between the two indicators, granular semantics and citations, as they would reflect different types of influence of science on technology.

Second, we correlate the proximity of NPL to their citing patent with the number of NPL of the patent, and find a negative relation: Patents with more NPL have NPL semantically further from themselves. We interpret this result as evidence of a tendency of applicants to cite an excessive number of weakly relevant references. This could be due to a cautious implementation of the "duty of candour", or to strategic orientations (e.g. flooding examiners and competitors with weakly relevant references so as to hide the important ones). This result suggests that using NPL as indicators of scientific sourcing should be done in a cautious way.

The report then suggests and tests a new indicator of knowledge flows from science to technology, combining semantics and citations. This indicator consists in the selecting, among the cited publications, the ones that are semantically closer to the citing patent (full text). We validate this indicator on two sub-set of patents, PPPs and EPO-USPTO families.

Further testing would be necessary to address two pending issues:

- For an indicator that keeps to the closest documents among cited ones, how many of those documents to select? One possible solution would be to select all documents with proximity above a certain threshold, to be determined.
- What to do with patents with NO NPL? In certain domains, patents tend to cite little or no NPL. This might reflect a weak connection with science, but also in certain cases a specific approach to citation. One possible solution would be to identify documents that are semantically closest to the patents and select those whose proximity exceeds a certain threshold.

In addition to the conclusions that come out of the present study, solving those two problems would open the way to more accurate indicators of the reliance of technology on science.

PART II: Econometric Investigations

12. What are the factors for a scientific publication to be cited by a patent?

In this section, we aim to identify features of publications that make them more or less likely to be cited by a patent, by using econometric estimates. Not all science is equally likely to be cited. Certain publications are cited while others are not. The likelihood for a scientific publication to be cited by a patent may depend on several factors. First are the relevance of its scientific content to technological inventions and its accessibility to potential users. Publications might also be cited thanks to the tendency of patent applicants to fulfil their "duty of candour", which motivates an inflation of little relevant documents in order to be legally "on the safe side" (see above).

Although certain characteristics themselves are not directly observable, they might be reflected in or correlated with measurable variables. The exercise is performed on two domains, CRISPR and CAR-T cells. As citations take time and we are working of recently developed domains, this exercise is subject to data truncation. The patent data are patent grants by the USPTO. The USPTO publishes most citations at the time of grant, and consequently we miss the most recent years (it takes usually 2 to 4 years for a patent application to be processed).

The model and variables

We estimate determinants on the likelihood to be cited in a patent by using logistic regression.

This model allows to explain a binary variable (the citation) with multiple factors that can be quantitative (such as age) or qualitative (e.g. discipline). For each dependent variable entered in the model, we estimate an odds ratio that quantifies the relative strength of the association with the variable to explain.

The variables are reported in table 12.1 (CAR-T cells) and table 12.2 (CRISPR).

- Citation: the explained variable is a dummy, equal to "1" for a publication that has been cited in a patent of the core less than two years after it issued, "0" otherwise. Testings show that the time horizon and the restriction to the core do not affect significantly the estimates (see Tables 3 and 5). This variable is independent of the number of citations above 1, then highly cited papers are treated equally as less cited ones. Overall, 19% of publications are cited by core patents in CAR-T cells (13% in the next two years) and 13% in CRISPR (9% in the next two years). Citations by non core patents are 25% and 17% respectively.

- NN100: it is equal to "1" if the scientific publication is among the 100 nearest neighbours of any patent of the core, it is "0" otherwise". It reflects the similarity in the text of the scientific publication and technology. The more similar the texts, the more relevant the science could be to invention. Overall, 80% of CAR-T cells and 71% of CRISPR publications are among the NN100.

- Age_publi: it is the age of the publication at the time of the estimation (2024), and its impact can be complex. More recent publications could be more relevant, but they are less well-known (then less accessible); older publications might be less relevant but their reputation is broader, due to diffusion lags. The impact of age could also depend on the domain, notably its maturity: emerging technologies rely on more recent science, but they might also be subject to a "foundation effect" (whereby initial scientific discoveries that launched the domain years ago

are still cited). Summary statistics for CAR-T cells and CRISPR are reported in the annex. As we are interested in citation, and citation takes time we run into a problem of data truncation. To mitigate that, 1) we restricted to citations in the two years following publication (it does not affect the estimates, see below); and 2) we restricted to publications published until 2019, as patents could cite older publications and more recent publications might suffer more from the truncation in patent data. In addition, age is among the variables of the model, so that we control for this phenomenon. The reference year being 2024, the average age of publications (pre-2019) is 7.25 years for CAR-T and 6.75 for CRISPR.

- Discipline: the technological relevance of science might differ across sub-disciplines, requiring that we breakdown CRISPR and CAR-T into more granular domains: 4 domains for CAR-T and 10 domains for CRISPR.

- Open_access: Some publications are stored behind paywalls, while others are accessible freely. Previous studies have shown that open access publications are more cited than closed access ones by other publications: we test here whether this result holds for citation by patents. Overall, 49% of CAR-T and 66% of CRISPR publications are open access.

- Funded: This variable reflects the fact that the research reported in the publications has benefitted from specific funding, of any nature (competitive funding from a national agency, grant by a firm or foundation etc.). Previous work has found that it correlates with "quality": funded research tends to me more cited than other research. The proportion of funded publications is 44% for CAR-T and 71% for CRISPR.

- Private: It is equal to "1" when one or more of the authors belong to a firm. It is the case for 21% of CAR-T and 9% of CRISPR publications. The impact of this variable could reflect the fact that businesses tend to invest in research that is more relevant to industry, or that businesses are more aware of research results coming from their own labs or other industry labs.

- Pub_type: scientific publications can be of various types, each corresponding to a special way of communicating knowledge. The types we used (from the classification of the WoS) include: journal articles (31% of CAR-T cells, 66% of CRISPR), review articles (22% and 12%), meeting abstracts (38% of CAR-T, not accounted for CRISPR), editorial material (5% for both fields) and "other" (4% and 18%).

- Nb_auth: number of authors of the publication: in the literature, the number of authors of a publication has been found to be positively correlated with the number of citations it receives: that could be due to a "network effect" (more authors means more promoters), and/or to a quality effect (more labour involved in the research would increase the quality-weighted output).

Inter_collab: It is equal to 1 when at least two authors are located in different countries. It is the case for 20% of CAR-T cells and 28% of CRISPR publications.

Regression results

Regression results are reported in tables 12.3 and 12.4**.** Interestingly, estimates are quite similar across regressions, between CRISPR and CAR-T cells, and between the various subcategories of publications and citations; subcategories with some dissimilarity are publications cited in the body of the patents, and the 2010-2014 set of publications, whose small numbers affect the significance of the results. This tends to show robustness in the relationships identified, at least

in the domains covered. Our analysis addresses the odd ratios (OR), which are more directly comparable between variables than regression coefficients. Main results are as follows.

- The impact of age is quadratic: it increases but at a declining rate. This shape is traditional in the literature and is interpreted as resulting from the composition of the opposite effects of depreciation and reputation mentioned above.
- Semantic proximity increases the likelihood of citation, as it would be expected (OR are around 1.6 and 2.0 for CAR-T cells and CRISPR respectively).
- For the type of publication: journal articles have the highest probability of citation for CAR-T, whereas for CRISPR it is literature reviews. Editorial material is less often cited, as it usually has less original scientific content than articles.
- Open access publications are more often cited, by far, with an OR around 2 in both CAR-T and CRISPR. This confirms previous results in the literature.
- Funded publications also are more often cited, with an OR close to the one of Open access. That could reflect either scientific quality of the publication (the underlying research was selected ex ante), or its relevance to industry (some public funds and most industry funds take into account the prospective relevance of research to industry) - these two factors are not exclusive of either other
- Publications with more authors are more often cited, which can be due to their quality (the number of authors is correlated with the cost of a research project, which is correlated with the quality of its results) or to a network effect (that boosts diffusion with the number of "diffusers"). It can be noticed that the citation differential is the highest for the body citations.
- The participation of one author from industry increases the likelihood of citation, and the OR of CAR-T and CRISPR are close to each other (around 1.7).
- Finally, and more puzzling, publications resulting from an international cooperation are less often cited than purely national ones (the effect is not significant for CAR-T cells). This tends to run against the literature, which tends to associate international cooperation with all attributes of quality and diffusion, including citation. How to interpret this result? In this regression, certain factors of quality often related to international cooperation are already accounted for (funding, number of authors), and what remains is the pure impact of the international nature of the authorship. One possible interpretation would be that, once other drivers are factored in, citation of a publication by a patent is associated with the applied character of the publication (closeness to technology), and applied research is less internationalised than basic research. This hypothesis would deserve further testing, that we do not conduct here.

Results are also very coherent across the various categories of publications and citations, with just a few exceptions. Hence, the determinants of citation are similar for core patents and non core patents; they are similar for a two years horizon and an open time horizon (which suggests that the data truncation issue does not affect significantly our estimates), for body and frontpage citations, and to a lesser extent for the 2010-2014 and the 2015-2019 periods. The most notable exceptions are as follows: Body citations for CAR-T cells are not boosted by semantic similarity or private participation; the same happens for CAR-T cells publications of the 2010-2014 period, whereas for CRISPR older publications, semantic similarity is neutral. It is also noticeable that the estimate on the 2010-2014 period gives an extremely high value for the variable "age". This corresponds to the emergence of the field, when a number of foundational articles were published. These articles, including Nobel prize nominated ones, have been and are still hugely cited: this is a "foundation effect", that we will also identify in the estimates of the "citation lag" model (section 13).

Forecasting

We then tested the use of our models for forecasting the likelihood of a paper to be cited by a patent. For that matter we used the model as reported in annex 4.

We just eliminate the NN100 variable, as it is revealed only in the future (it cannot be known at the time of publication): hence, all the variables that are left are known at the time of the prediction.

Our method is to estimate a model up to year Y, and to use it to predict the likelihood of publications of year Y+1 to be cited in the next two years. In order to obtain a meaningful prediction, we need 1) enough data for the model to be significant; 2) a sufficient time lag so that most if not all citations have been generated. While constraint 1 favours working on a recent year, constraint 2 pleads for going far in the past. We solved this trade-off by choosing Y= 2015, then predicting citation of publications of year 2016.

In addition to the econometric model, we also tested a neural network, trained with the same data, over the same period, and adjusted to predict citations in year 2016.

Confusion matrices are reported in tables 12.5 (CAR-T) and 12.6 (CRISPR).

Accuracy is around 0.75 (CAR-T) and 0.88 (CRISPR), which is not bad. Unsurprisingly, the prediction is more accurate for the largest categories, the negatives (no citation), where it is true in more than 80% of all cases. However, the prediction of citations (positives) is of lesser quality: precision (the percentage of true positive predictions among all positive predictions) is around 50% for CAR-T, and around 30% for CRISPR; and sensitivity, or recall (the percentage of true positive predictions among all real positives) is 45 to 60% for CAR-T cells, but only 2 to 7% for CRISPR. This is to be related to the lower proportion of positives (citations) among CRISPR publications, which makes them more difficult to predict, ceteris paribus. The neural network seems slightly better than the logistic model for predicting citations of CRISPR publications, but it is still weak.

Discussion and conclusion

The factors that affect the likelihood of a scientific publication to be cited by a patent are similar in CRISPR and CAR-T cells. Main drivers are: the age of the publication (quadratic), its semantic proximity to patents, whether it is open access or not, whether it was funded or not, its number of authors, whether one of them at least is from industry and whether they reside in different countries (with a negative impact). In addition, we detected a “foundation effect” in CRISPR, by which earlier publications (foundations) are more cited over time than others.

This model cannot however be used as such for prediction, due variables that are not known at the time of publication. Forecasting of citations is made difficult by the fact that citation is a relatively rare event (19% of CAR-T cells publication, 13% of CRISPR) but the model performs quite well for CAR-T cells and less so for CRISPR.

Table 12.1: Summary statistics for CAR-T cells

Variable	Obs	Mean	Std. Dev.	Min	Max
Pub_Cit_Cor	6099	0,19	0,39	0	1
Pub_Cit_Core_NnCor	6099	0,25	0,43	0	1
Pub_Cit_2Y	6099	0,13	0,34	0	1

Pub_Front_Cit_2Y	6099	0,13	0,33	0	1
Pub_Body_Cit_2Y	6099	0,08	0,27	0	1
Pub_Both_Cit_2Y	6099	0,07	0,26	0	1
Pub_Cit_2Y_10-14	784	0,38	0,48	0	1
Pub_Cit_2Y_15-19	5178	0,10	0,30	0	1
Age_pub	6099	7,25	2,69	5	25
Age_sq_pub	6099	15,04	34,82	0,30	422,30
NN100	6099	0,80	0,40	0	1
Onco (ref.)	6099	0,29	0,46	0	1
Hemato	6099	0,31	0,46	0	1
Immuno	6099	0,25	0,43	0	1
Other_discip	6099	0,15	0,35	0	1
Art (ref.)	6099	0,31	0,46	0	1
Edit_mat	6099	0,05	0,22	0	1
Meet_abst	6099	0,38	0,48	0	1
Review	6099	0,22	0,42	0	1
Other_type	6099	0,04	0,19	0	1
Open_access	6099	0,49	0,50	0	1
Funded	6099	0,44	0,50	0	1
1_3_auth	5672	0,32	0,47	0	1
4_7_auth	5672	0,28	0,45	0	1
8_ 11_auth	5672	0,20	0,40	0	1
More_11_auth	5672	0,20	0,40	0	1
Private	5672	0,21	0,40	0	1
Inter_collab	5672	0,20	0,40	0	1

Table 12.2: Summary statistics for CRISPR corpora

Variable	Obs	Mean	Std. Dev.	Min	Max
Pub_Cit_Cor	18080	0,13	0,33	0	1
Pub_Cit_Core_NnCor	18080	0,17	0,37	0	1
Pub_Cit_2Y	18080	0,09	0,28	0	1
Pub_Front_Cit_2Y	18080	0,08	0,28	0	1
Pub_Body_Cit_2Y	18080	0,05	0,21	0	1
Pub_Both_Cit_2Y	18080	0,04	0,20	0	1
Pub_Cit_2Y_10-14	1242	0,38	0,48	0	1
Pub_Cit_2Y_15-19	16752	0,07	0,25	0	1
Age_pub	18080	6,75	1,82	5	24
Age_sq_pub	18080	8,47	15,34	0,28	381,23
NN100	18080	0,71	0,45	0	1
Onco (ref.)	18080	0,07	0,26	0	1
Biochim_biol_molec	18080	0,35	0,48	0	1
Bio_cel	18080	0,05	0,21	0	1
Biotech_microbio	18080	0,14	0,35	0	1
Botan_bio_veg	18080	0,04	0,20	0	1
Genet_hered	18080	0,03	0,17	0	1
Hema_imuno	18080	0,05	0,21	0	1
Microbio	18080	0,08	0,27	0	1
Neurosc	18080	0,03	0,17	0	1
Other_discip	18080	0,16	0,36	0	1
Art (ref.)	18080	0,66	0,47	0	1
Edit_mat	18080	0,05	0,21	0	1
Other_type	18080	0,18	0,38	0	1
Review	18080	0,12	0,32	0	1
Open_access	18080	0,66	0,47	0	1
Funded	18080	0,74	0,44	0	1
1_4_auth (ref.)	17315	0,36	0,48	0	1
5_6_auth	17315	0,19	0,39	0	1
7_10_auth	17315	0,26	0,44	0	1
More_10_auth	17315	0,18	0,39	0	1
Private	17315	0,09	0,29	0	1
Inter_collab	17315	0,28	0,45	0	1

Table 12.3: Odds ratios of factors affecting patent-paper citation in CAR-T-cells

	Pub_Cit_Cor	Pub_Cit_Core_NnCor	Pub_Cit_2Y	Pub_Front_Cit_2Y	Pub_Body_Cit_2Y	Pub_Both_Cit_2Y	Pub_Cit_2Y_10-14	Pub_Cit_2Y_15-19
Age_pub	2.044***	1.881***	3.188***	3.346***	2.841***	3.147***	1.743	3.847***
	(0.074)	(0.061)	(0.201)	(0.220)	(0.208)	(0.253)	(1.561)	(0.764)
Age_sq_pub	0.971***	0.976***	0.924***	0.922***	0.939***	0.933***	0.956	0.904***
	(0.002)	(0.002)	(0.005)	(0.005)	(0.006)	(0.006)	(0.058)	(0.032)
NN100	1.317 **	1.612***	1.487**	1.625***	1.377	1.612**	0.419	1.491**
	(0.165)	(0.185)	(0.236)	(0.272)	(0.278)	(0.363)	(0.335)	(0.248)
Discipline *(ref. onco)*								
Hemato	1.266**	1.123	1.481***	1.544***	1.287	1.379*	0.986	1.599***
	(0.141)	(0.112)	(0.197)	(0.214)	(0.226)	(0.262)	(0.321)	(0.240)
Immuno	1.362***	1.304***	1.698***	1.725***	1.733***	1.813***	1.676**	1.675***
	(0.147)	(0.129)	(0.212)	(0.222)	(0.266)	(0.296)	(0.421)	(0.251)
Other_discip	0.478***	0.564***	0.687*	0.750	0.823	1.030	0.415*	0.834
	(0.091)	(0.093)	(0.153)	(0.172)	(0.254)	(0.340)	(0.191)	(0.213)
Pub_type *(ref. art)*								
Edit_mat	0.649**	0.765	0.599**	0.619*	0.270***	0.259***	0.530	0.658
	(0.142)	(0.149)	(0.150)	(0.159)	(0.114)	(0.118)	(0.235)	(0.197)
Meet_abst	0.231***	0.166***	0.272***	0.250***	0.098***	0.034***	0.275***	0.281***
	(0.036)	(0.024)	(0.050)	(0.049)	(0.033)	(0.018)	(0.117)	(0.058)
Review	0.813*	1.087	0.939	0.964	0.982	1.037	1.256	0.862
	(0.090)	(0.110)	(0.117)	(0.123)	(0.146)	(0.162)	(0.317)	(0.128)
Other_type	0.774	0.743	0.547	0.603	0.170*	0.202	4.789	0.381*
	(0.249)	(0.213)	(0.249)	(0.276)	(0.174)	(0.207)	(5.547)	(0.207)
Open_access	2.357***	2.221***	2.567***	2.698***	2.017***	2.175***	2.444***	2.691***
	(0.232)	(0.196)	(0.291)	(0.316)	(0.289)	(0.334)	(0.514)	(0.375)
Funded	2.153***	2.157***	2.122***	2.211***	2.484***	2.795***	2.946***	1.922***
	(0.246)	(0.217)	(0.285)	(0.308)	(0.471)	(0.584)	(0.851)	(0.296)
Nb_auth *(ref. 1_3_auth)*								
4_7_auth	1.178	1.367***	1.132	1.164	1.530***	1.649***	1.678**	0.996
	(0.136)	(0.142)	(0.148)	(0.156)	(0.248)	(0.281)	(0.422)	(0.158)
8_ 11_auth	1.557***	1.703***	1.405**	1.437**	1.836***	1.969***	1.985**	1.149
	(0.212)	(0.215)	(0.214)	(0.224)	(0.344)	(0.389)	(0.603)	(0.212)
More_11_auth	2.173***	2.186***	1.993***	2.002***	3.559***	3.898***	3.699***	1.606**
	(0.309)	(0.291)	(0.318)	(0.329)	(0.694)	(0.805)	1.270)	(0.302)
Private	1.699***	1.880***	1.679***	1.760***	1.041	1.079	0.853	1.849***
	(0.189)	(0.195)	(0.219)	(0.237)	(0.183)	(0.204)	(0.334)	(0.262)
Inter_collab	0.976	0.987	0.901	0.926	0.813	0.841	0.791	0.970
	(0.098)	(0.092)	(0.105)	(0.110)	(0.116)	(0.126)	(0.212)	(0.130)
Pseudo R^2	31.54%	31.62%	32.37%	33.88%	35.62%	39.71%	29.16%	27.87%
Log Likelihood	-1,948.40	-2,226.25	-1,556.99	-1,478.43	-1,063.36	-943.31	-353.82	-1,138.87
Correct classification	85.45%	82.95%	88.58%	88.93%	92.86%	93.35%	76.60%	90.42%
Observations	5672	5672	5672	5672	5672	5672	748	4789

Notes: Standard errors are shown in parentheses. ***, **, * = significance at 1%, 5% and 10%, respectively

Table 12.4: Odds ratios of factors affecting patent-paper citation in CRISPR

	Publ_Cited_Core	Publ_Cited_Core_NonCore	Publ_Cited_2Years	Publ_Front_Cited_2Years	Publ_Body_Cited_2Years	Publ_Both_Cited_2Years	Pub_Cited_2Years_2010-14	Pub_Cited_2Years_2015-19
Age_pub	1.987***	2.001***	2.317***	2.540***	2.338***	2.868***	14.555***	1.432***
	(0.062)	(0.058)	(0.091)	(0.107)	(0.115)	(0.165)	(11.381)	(0.159)
Age_sq_pub	0.973***	0.974***	0.959***	0.953***	0.964***	0.952***	0.827***	1.042**
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.046)	(0.021)
NN100	2.237***	2.038***	2.003***	2.160***	1.800***	2.123***	1.544	2.170***
	(0.185)	(0.140)	(0.198)	(0.230)	(0.247)	(0.343)	(0.443)	(0.232)
Discipline (ref. onco)								
Biochim_biol_molec	5.134***	2.082***	6.264***	5.979***	6.652***	6.072***	0.931	7.021***
	(0.910)	(0.234)	(1.497)	(1.470)	(2.282)	(2.214)	(0.734)	(1.821)
Bio_cel	0.896	0.617***	0.933	0.894	1.008	0.9336	0.504	0.848
	(0.242)	(0.112)	(0.336)	(0.333)	(0.515)	(0.511)	(0.499)	(0.347)
Biotech_microbio	6.630***	2.754***	8.288***	7.919***	7.016***	6.339***	2.336	9.175***
	(1.206)	(0.327)	(2.016)	(1.981)	(2.450)	(2.353)	(1.919)	(2.418)
Botan_bio_veg	1.923***	0.920	2.382***	2.310**	2.824**	2.741**	1.177	2.384**
	(0.471)	(0.164)	(0.761)	(0.765)	(1.268)	(1.327)	(1.235)	(0.828)
Genet_hered	1.384	0.654**	1.713	1.767	0.773	0.816	0.771	1.515
	(0.371)	(0.131)	(0.583)	(0.611)	(0.470)	(0.507)	(0.749)	(0.593)
Hema_imuno	2.121***	1.348*	2.847***	2.633***	3.290***	2.907**	0.269	3.361***
	(0.487)	(0.211)	(0.834)	(0.802)	(1.345)	(1.284)	(0.301)	(1.046)
Microbio	1.941***	0.774*	2.306***	2.035***	2.392**	1.863	0.320	2.889***
	(0.381)	(0.106)	(0.594)	(0.540)	(0.870)	(0.721)	(0.257)	(0.854)
Neurosc	1.185	0.803	0.824	0.753	0.440	0.245	1.359	0.637
	(0.334)	(0.154)	(0.354)	(0.342)	(0.346)	(0.261)	(1.662)	(0.330)
Other_discip	1.433*	0.775*	1.783**	1.707*	1.053	0.848	0.175*	2.207***
	(0.298)	(0.110)	(0.485)	(0.478)	(0.442)	(0.389)	(0.156)	(0.646)
Pub_type (ref. art)								
Edit_mat	0.628***	0.563***	0.639**	0.646**	0.411***	0.402***	0.664	0.705
	(0.104)	(0.086)	(0.122)	(0.126)	(0.129)	(0.132)	(0.271)	(0.157)
Review	1.198**	1.380***	1.194*	1.184*	1.107	1.079	1.561**	1.193*
	(0.095)	(0.099)	(0.109)	(0.111)	(0.136)	(0.141)	(0.331)	(0.126)
Other_type	0.478***	0.324***	0.521***	0.491***	0.531***	0.468***	0.799	0.454***
	(0.069)	(0.043)	(0.086)	(0.084)	(0.123)	(0.117)	(0.272)	(0.092)
Open_access	1.898***	1.839***	2.030***	1.994***	1.738***	1.664***	2.657***	1.938***
	(0.130)	(0.113)	(0.165)	(0.166)	(0.187)	(0.188)	(0.485)	(0.181)
Funded	1.826***	1.743***	1.514***	1.496***	1.732***	1.705***	2.287***	1.492***
	(0.178)	(0.151)	(0.168)	(0.170)	(0.267)	(0.277)	(0.599)	(0.194)
Nb_auth *(ref. 1_4_auth)*								
5_6_auth	1.215**	1.124*	1.317***	1.364***	1.291**	1.374***	1.642***	1.303***
	(0.092)	(0.078)	(0.113)	(0.121)	(0.148)	(0.166)	(0.302)	(0.133)
7_10_auth	1.307***	1.308***	1.380***	1.428***	1.446***	1.552***	0.992	1.541***
	(0.095)	(0.087)	(0.115)	(0.123)	(0.159)	(0.180)	(0.180)	(0.150)
More_10_auth	1.483***	1.816***	1.619***	1.599***	2.067***	2.120***	1.593*	1.733***
	(0.126)	(0.137)	(0.157)	(0.161)	(0.257)	(0.281)	(0.391)	(0.191)
Private	1.608***	1.850***	1.594***	1.662***	1.940***	2.161***	1.695*	1.614***
	(0.138)	(0.140)	(0.155)	(0.166)	(0.233)	(0.273)	(0.471)	(0.171)
Inter_collab	0.880**	0.881**	0.880*	0.880*	0.762***	0.744***	1.003	0.864**
	(0.053)	(0.047)	(0.061)	(0.062)	(0.070)	(0.072)	(0.168)	(0.066)
Pseudo R2	23.56%	22.79%	23.67%	25.32%	24.56%	28.43%	15.35%	19.11%
Log Likelihood	-5108.05	-6115.12	-4050.76	-3831.35	-2560.40	-2265.12	-679.07	-3250.54
Correct classification	88.04%	85.01%	91.05%	91.55%	94.98%	95.34%	69.43%	93.14%
Observations	17315	17315	17315	17315	17315	17315	1207	16025

Notes: Standard errors are shown in parentheses. ***, **, * = significance at 1%, 5% and 10%, respectively

Table 12.5: Confusion matrix for CAR-T cells

From logistic regression	**Actual class: Yes**	**Actual class: No**	
Predicted class: Yes	TP: 72	FP: 70	*Sensitivity 45.6%*
Predicted class: No	FN: 86	TN: 392	*Specificity 84.8%*
	Precision 50.7%	*Negative predictive value 82.0%*	*Accuracy 74.8%*

From neural network	**Actual class: Yes**	**Actual class: No**	
Predicted class: Yes	TP: 96	FP: 104	*Sensitivity 60.8%*
Predicted class: No	FN: 62	TN: 358	*Specificity 77.5%*
	Precision 48.0%	*Negative predictive value 85.2%*	*Accuracy 73.2%*

Table 12.6: Confusion matrix for CRISPR corpora

From logistic regression	**Actual class: Yes**	**Actual class: No**	
Predicted class: Yes	TP: 6	FP: 11	*Sensitivity 2.2%*
Predicted class: No	FN: 261	TN: 1983	*Specificity 99.4%*
	Precision 35.3 %	*Negative predictive value 88.4%*	*Accuracy 88.0 %*

From neural network	**Actual class: Yes**	**Actual class: No**	
Predicted class: Yes	TP: 20	FP: 51	*Sensitivity 7.5%*
Predicted class: No	FN: 247	TN: 1943	*Specificity 97.4%*
	Precision 28.2%	*Negative predictive value 88.7%*	*Accuracy 86.8%*

13. What are the factors that affect the citation lag?

Introduction

This section addresses the determinants of the time lag between a patent and the science that it refers to, focusing on four frontier domains, CRISPR, CAR-T cells, mRNA and Quantum Cryptography. In the analysis of the relations between science and technology, time matters, as illustrated by the label "accelerator" given, to public organisations in charge of transferring public science to business. Delays imply that new scientific ideas are slow to be translated into innovations, resulting in reductions in productivity or welfare gains. In a competitive context, longer lags can mean loss in competitiveness for businesses or countries that are the slowest.

Multiple factors can influence the length of the lag from science to technology, including:

- The positioning of an innovation in the scientific and technological cycle (is it closer to the scientific discovery or to customer driven market demand?).
- The proximity of the patent to the core of the domain: a patent can either contribute to developing the core technology, or to apply it to specific uses (some patents do both of course). Our assumption is that the closest a patent is to the core of the domain, the most recent the science it refers to is, as the complementary knowledge combined for applications is usually older than the core knowledge: an important characteristic of general purpose technologies is that their development re-dynamises technologies that had been developed some time before, by combining with them.
- The weight of the science that is core to the domain among the sources of the patent: this mechanism operates according to the same logic as the previous one, that it reinforces on the science side. In addition, a lower weight might mean a stronger interdisciplinary character of the patent, and interdisciplinarity might involve further delays: Within-domain diffusion channels are more efficient (then quicker) than cross-domain channels, as scientists have set up networks and institutions mainly on a disciplinary basis.
- The institutional sector of the assignee (public or business) is a particular focus of this study. Inventions issued by the public sector, notably universities, are expected to rely on more recent science as most science is realised in public labs, involving shorter diffusion channels for the concerned knowledge.

Our model explains the lag between a patent and its cited scientific publications, in order to estimate whether this invention relies on recent or older science. The analysis is performed from the patent perspective, then restricted to discoveries that actually diffused to technology – it does not explain the fact that a publication is cited or not. The study is conducted on emerging technology fields (CRISPR and CART-Cell appeared in 2012).

Description of the variables

Patent publication lag: The dependant variable is the lag between the priority date of the patent family and the date of cited scientific articles (NPL). One patent can have several NPL citations, up to several hundred; therefore, we computed the median date of citations (to lower the impact of extreme values).

Core/non_core patents: A patent is either from the core of the technology field (see definitions in appendix 1) or it is non_core, an application of the core to another technology field (i.e. patents that cite the publications from the field but are not from the core). As pointed above, this we expect that core patents will have a higher proximity to the science of the field, and then smaller delay of citations.

Institutional sector and size of patentee: The institutional sector of the patent assignee (public[1] or business) may play a role in the technology transfer, as mentioned above. Most scientific publications originate from the public sector, and within sector diffusion is usually quicker than cross sector diffusion, notably for institutional and legal reasons. In addition, patents taken by public sector entities are usually closer to science, which corresponds better to the skills and missions of these entities.

We combine this institutional dimension with the size of businesses (measured by the number of patents in its name). We consider that businesses with more than 150 patent families in the database as large patentees, and businesses with fewer patents as small patentee businesses. In the case of co-patenting, the institutional categories of the partners (public and business, business only or public only).

Age of the patent: The connection of inventions with science can be affected by their position in the technology trajectory, whether they are from an early or later stage. It is usually expected that earlier discoveries rely on more recent science, not least because the technology field that they are creating directly comes out of recent discoveries. As a technology trajectory develops it passes from an exploration to an exploitation regime, where basic, scientific knowledge plays a lesser role. The reference year used for the computation of the age is 2022, the year of the most recent patents in pour dataset.

NN1: this indicator measures the semantic similarity between a patent and it nearest neighbour among scientific articles of the same field (for instance CRISPR patents are compared with articles in the CRISPR field). It is calculated as the highest cosine of the embedding of the patent and the embeddings of scientific publications of the field.

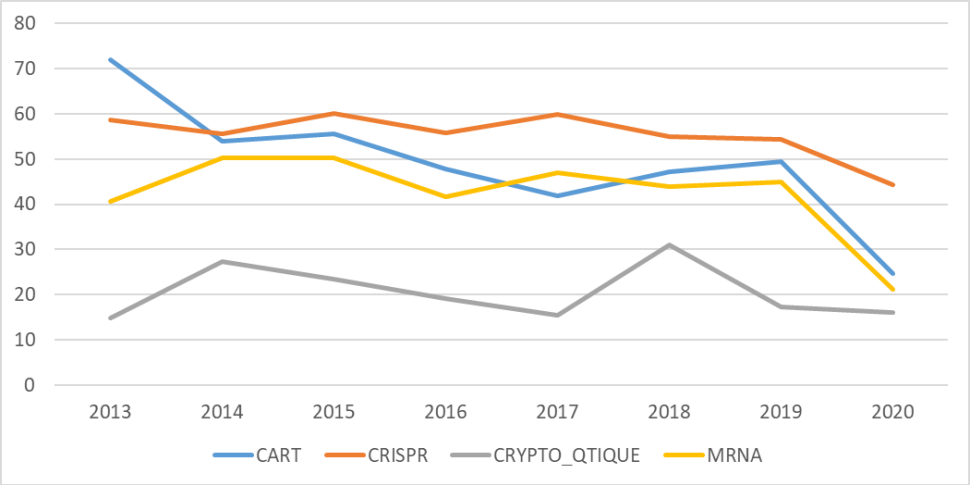
PPP: this dummy indicates whether the patent is part of a patent/publication pair (PPP). PPPs are made of documents that cover a same idea (discovery, invention). We use the ROS database (which takes citations as a reference).

Core_NPL_share: The ratio between the number of NPL from the core and the total number of NPL. This indicator shows how much an invention relies on the field, relatively to its other influences, it reflects the degree of multidisciplinary in the science used. **Inventor_non_USA**: This variable indicates whether there is an inventor from another country than the USA for this patent. It is expected that geographical barriers (between countries) might slowdown the circulation of knowledge, hence patents with both US and non US inventors would cite older science.

[1] Public assignees include Higher Education institutions, Non-profit research Institutes, Government institutions and Hospitals.

Descriptive charts

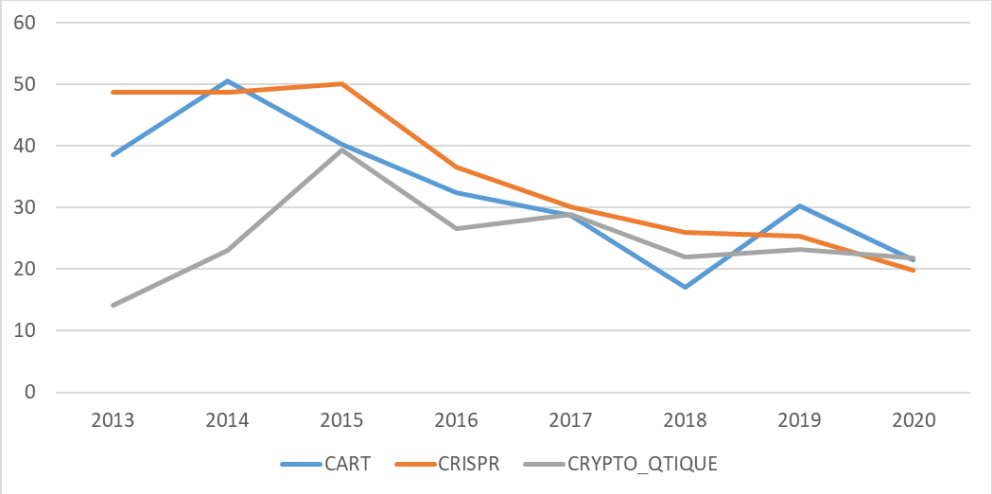
Figure 13.1: The share of the public institutional sector (in %) in four technology fields at the USPTO – core technology (year 2020 is incomplete)



Source: Patstat 2023

Figure 13.1 shows the role of the public sector in our four fields (relating to the assignee type). It focuses on core patents from these four fields, on the 2013-2020 period which is the common period observed across fields (as CRISPR and CART fields emerged recently). The role of the public sector appears to be very significant in the three fields related to biotechnology (CART, CRISPR and mRNA), accounting for more than 40% of the patents which is significantly more than the average at USPTO for all fields (around 5%). The role of the public sector seems to be fairly constant over time, if we leave aside the last year 2020, which is incomplete. The public sector in the quantum cryptography field accounts for around 20%, below these three fields, but still far above the average at USPTO.

Figure 13.2: The share of the public sector in "non core patents" of four technology fields at the USPTO



Source: Patstat 2023

Figure 13.2 is based on patents from the "non-core", i.e applications of the field to other technologies; the role of the public sector appears to be significant for CART and CRISPR in the first years (2013-2015) but it decreases over time, probably because businesses are progressively investing more in the field to develop new technologies. This is consistent with the view of the technology trajectory that recognises an increasing role over time for commercial entities, as knowledge that originated in academia diffuses to businesses. In the case of quantum cryptography, the number of observations for "non-core" patents is low preventing to interpret the figures.

Figure 13.3: Patent publication lag in years, CAR-T and CRISPR, by assignee institutional sector, core inventions

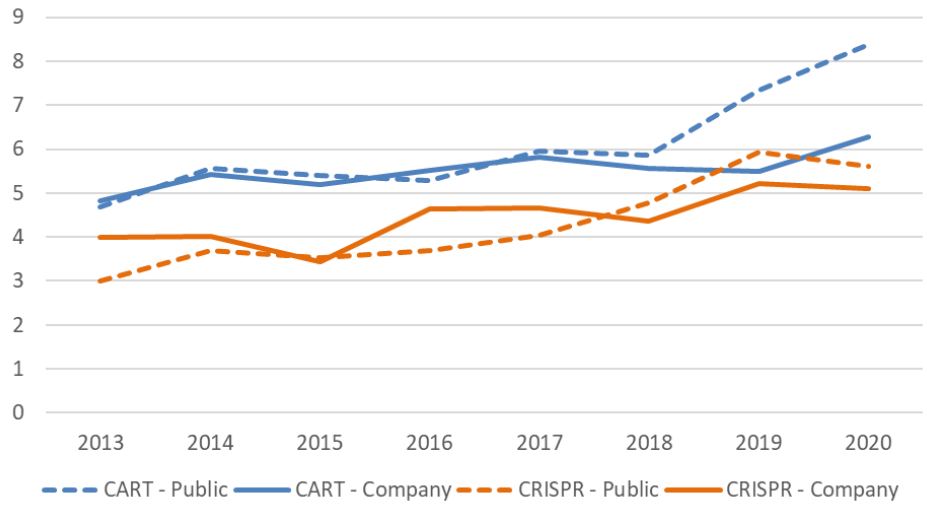


Figure 13.3 shows the average citation lag of publication in patents over time for two fields – focusing on their core technologies-, CRISPR and CAR-T cell, and differentiating public institutions and companies. The patent-publication lag appears to increase over time for these two fields. The lag appears to be rather similar for public institutions and businesses, which might contradict the hypothesis mentioned above: but no control variable is taken into account in this figure.

Figure 13.4: Patent publication lag in years, CAR-T and CRISPR, by assignee sector, non core inventions

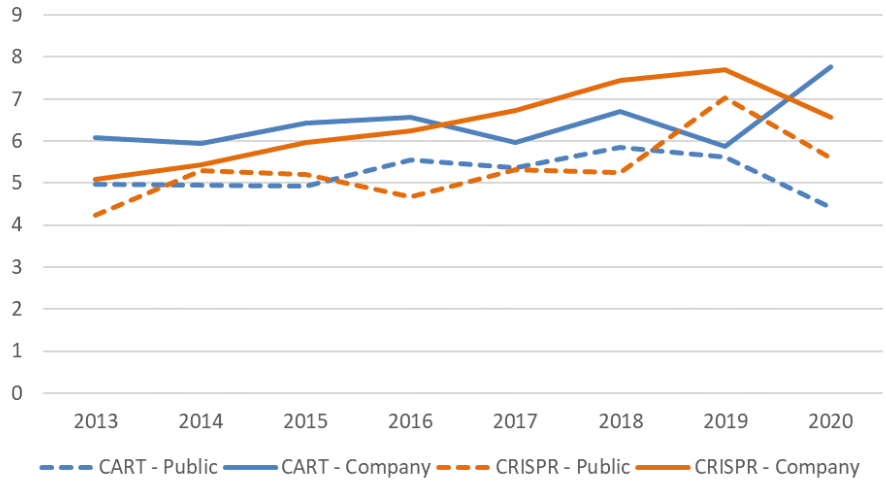


Figure13.4 shows that when we focus on non core patents, the public sector in these two fields have on average a lower patent publication lag.

Table 13.1: Mean values of the variables used in the regression – All (core and non-core) patents

	CRISPR (2013-2020)	**CAR-T (2013-2020)**	**mRNA (1983-2020)**	**QUANTUM CRYPTOGRAPHY (1983-2020)**
Patent_science_lag	5.45	5.24	6.19	5.79
Non_core patent	0.78	0.69	0	0.83
Age	5.46	6.21	18.28	10.18
Institutional sector				
Large_pat_business_only	0.39	0.11	0.21	0.50
Small_pat_business_only	0.28	0.49	0.27	0.21
Public_only	0.29	0.31	0.36	0.24
Public & business	0.03	0.08	0.05	0.02
PPP	0.07	0.08	0.15	0.05
Inventor_non_USA	0.29	0.35	0.49	0.47
NN1	0.93	0.94	0.93	0.93
Core_NPL_Share	0.29	0.29	0.36	0.63

Table 13.1 reports the average value for the variables used in the following regressions (accounting for both the core and non-core patents). The institutional sector variable is qualitative; therefore, we computed the share of granted patents for each individual sector. Few comments:

1) The mean age of patents is significantly different across fields. It is on average 5-6 years for CRISPR and CAR-T (CAR-T and CRISPR emerged in 2012-2013), compared to 10 years for quantum cryptography and 18 years for mRNA. For this analysis we restricted the CRISPR and CAR-T samples to 2013-2020, but this should not impact the results as these two fields are very recent.
2) The average patent-publication lag is comprised between 5.24 for CAR-T and 6.19 for mRNA. This value is lower for the most recent fields (CAR-T and CRISPR) and higher for the two older fields.
3) The weight of the public sector is significant in the four fields (usually around 30% of the filing are made by public organizations only). The co-filing of public organizations with businesses is much less frequent (between 2% and 8% of patents), However, in these domains the frontier between basic research and commercial development is not always clearcut, and universities give birth to many spinoff involving academic researchers, or even founded by them, whose objective is to commercialise the underlying, academic discovery. The intellectual property framework for such operations often includes licensing agreement, exclusive or not, which is not reflected in our data. It is well known for instance that the discoverers of CRISPR, Emmanuelle Charpentier and Jennifer Doudna are engaged in several such startup companies. In such cases it is to be expected that science diffuses rapidly to technology, as the two activities are conducted by the same people. Hence one would like to differentiate businesses between academic spinoff and others: although we don't have such information, the "size of the business" variable can partly capture it, as spinoffs are usually of small size.

Table 13.2: Mean values of the variables used in the regression –Core patents

	CRISPR (2013-2020)	**CAR-T (2013-2020)**	**mRNA (1983-2020)**	**QUANTUM CRYPTOGRAPHY (1983-2020)**
Patent_science_lag	3.62	4.90	6.19	5.55
Age	5.98	6.25	18.28	10.41
Institutional sector				
Large_pat_business_only	0.14	0.05	0.21	0.51
Small_pat_business_only	0.32	0.46	0.27	0.21
Public_only	0.48	0.35	0.36	0.22
Public & business	0.04	0.14	0.05	0.02
PPP	0.19	0.16	0.15	0.09
Inventor_non_USA	0.29	0.41	0.49	0.57
NN1	0.95	0.95	0.93	0.94
Core_NPL_Share	0.36	0.33	0.36	0.37

Descriptive statistics from the core patents (table 2) show that the patent publication lag is lower than for non core patents (very significantly for CRIPSR), while the patent age is sensibly the same.

The main differences come from the sectoral breakdown:

For CRISPR and CAR-T cell the share of large patentee businesses is significantly lower in core than non core patents, while the share of the public sector is significantly larger (in particular for CRISPR), as is the share of small businesses and of the collaboration between the public and the business sector for CAR-T. For CAR-T, it might show the importance of spin-offs in this domain, while for CRISPR we notice a more direct role of the public sector.

Econometric results – citation lag between science and technology

The dependent variable is the lag between patent family priority date and the median date of its scientific citations

Table 13.3: Regression results – Explained variable: median lag of publications cited by a patent. All patents

	CRISPR	**CAR-T**	**mRNA**	**QUANTUM CRYPTOGRAPHY**
Intercept	29.02***	-0.47	15.2**	1.73
Core patent	-0.79***	-0.30(.)		-1.13***
Patent age	-0.50***	-0.50***	-0.23***	-0.12***
Sector_Public_only	-1.15***	-0.56*	-1.29***	0.13
Sector_Public & business	-1.24**	0.14	-0.24	-0.06
Sector_Small_pat_business_only	-0.82***	-0.42(.)	-0.41(.)	0.49(.)
PPP	-0.19	-0.22	-0.49*	-2.06***
Inventor_non_USA	0.20*	0.40**	-0.79***	-0.29
NN1	-19.71***	11.54*	-3.22	8.16
Core_NPL_Share	-5.73***	-5.93***	-1.78***	-3.22***
Adjusted R-squared	**0.497**	**0.283**	**0.228**	**0.072**

Signif. codes: *** significant at 0.001, '**' 0.01 '*' 0.05 '(.)' 0.1

Main conclusions from these regressions are as follows:

1) The quality of regressions is not bad for CRISPR, CAR-T and mRNA, with R2 of 0.50, 0.28 and 0.23 respectively. It is lower for Quantum cryptography, a domain where there are far less NPL than in the two others. Our comments mainly focus on the' three significant regressions.
2) The coefficients are generally of the same sign across all technology fields, and sometimes very close to each other, except one (NN1) between CRISPR and CART-cell, and one between mRNA and the two other fields (Inventor_non_USA).
3) The sign is negative and very significant for CRISPR for "core" patents (the reference being "non-core" patents). It means that on average core technologies cite more recent science than the applications of the field to other technologies (e.g CRISPR applied in agriculture), which usually combines citations from different scientific fields beyond the core field. This result holds for quantum cryptography. CAR-T is a technique used to fight cancer, a narrower application than CRISPR, thus both core and non-core CAR-T cell patents samples have an applied aspect, which may explain that the selected core and non-core samples give similar results in this field. For mRNA, a very broad and older field, we only focused on the core of the field in the sample, thus this variable is irrelevant and was not included in the regression. For CAR-T, this coefficient is not very significant.
4) The patent age has a negative sign, meaning that newer patents cite relatively older science as the field develops. This is consistent across the four fields. This relates at least partly to a sort of "foundation effect", as foundational papers (e.g. Charpentier and Doudna for CRISPR, published in 2012) obviously get older with time while they are still cited.
5) The (institutional) sector variables' coefficients have to be interpreted relatively to the reference "Large patentee business". The public sector variable has a negative sign for CRISPR, CAR-T and mRNA meaning that public institutions generally rely on more recent science than large patentee businesses. The most basic research is also at the frontier of knowledge and relies on most recent advances. Interestingly, collaboration between the public sector and businesses also has a negative and significant sign for CRISPR, as well as for small patentee businesses that seem to rely on more recent science than large businesses, at least in the CRISPR field. In this field, many spin-offs from the public sector emerged, and are accounted as "small patentee businesses" according to our criteria which might explain this effect.
6) The effect of the share of core science of the field among cited references is negative and strong in both fields. References from the core are younger than other cited references. This is consistent with the fact that patents from the core cite more recent literature than non-core patents: core science and technology define the frontier and must rely on frontier references.
7) The effect of semantic proximity has an opposite sign in CRISPR and CAR-T fields and is not significant for mRNA. CRISPR patents that are semantically closer to science tend to cite more recent publications, while the opposite holds for CART. More inquiry into the specificities of these two fields would be needed in order to bring interpretation of this result.
8) The "Inventor_non_USA" variable has a positive and significant sign for CRISPR and CAR-T showing that USA inventors cite in average more recent science than inventors from other countries. It can be explained by the fact that the bulk of global research on CRISPR is performed in the USA. For mRNA, a field that emerged in the 80s and 90s,

this variable has a positive sign, which might underline that that the bulk of global research became less concentrated in favour of the USA in this field over time.

9) Results for quantum cryptography are consistent with other fields for the age, non-core patents and the share of core science variables but differ for the institutional sector - where no specific sector has a significant effect – and for the PPP variables, where it appears that patents with a PPP cite newer scientific articles.

Discussion and conclusion

This study addresses the determinants of the time lag between a patent and the science that it refers to, focusing on four frontier domains, CRISPR, CAR-T cells, mRNA and Quantum Cryptography.

The results show that more recent publications are cited by: patents from the core of science-based domains rather than patents applying the core knowledge to other domains; by patents that cite a larger share of publications from the core of the domain; by patents which are closer in time to the emergence of the domains (meaning that later patents keep citing older publications); by patents assigned to public sector entities or small companies, as opposed to large companies.

Hence, time lag can be considered as an indicator of closeness to the core knowledge of scientific domains. Closeness of a technology field to the corresponding "core science" obeys both scientific and social factors: it is driven by similarity in knowledge, by incentives, institutional constraints and by the density of communication channels. If one takes these factors as references, then the results above make perfect sense. Knowledge related closeness shows in the effect of being in the core of the technology or citing more core publications; it also shows through the higher closeness of technology to science in earlier stages of the technology cycle. Social factors show through the higher closeness of public entities and small businesses (among them are startups) with science, but also through the share of core publications, as knowledge channels are denser within domains than across domains.

Although this study cannot give rise directly to a forecasting exercise, it can inform such an exercise. A model that would attempt to predict patent filings from scientific publications should take into account the two following factors:

- The impact of science on technology is differentiated: scientific discoveries have a quicker impact on inventions pertaining to the same domain, a slower impact on inventions from other domains. In the case of GPTs, one would expect that applications develop later than core inventions.
- The impact of discoveries on inventions is spread over time: notably foundational discoveries of a domain have a long term impact. Hence a forecasting model must include some dynamics and time variability.

References

Ahmadpoor, M., Jones, B.F., 2017. The dual frontier: patented inventions and prior scientific advance. Science 357, 583–587. https://doi.org/10.1126/science.aam9527

Ba Z., Liang Z. (2021) A novel approach to measuring science-technology linkage: From,the perspective of knowledge network coupling. Journal of Infometrics. https://www.sciencedirect.com/science/article/abs/pii/S1751157721000389?via%3Dihub

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Boyack, K. W., & Klavans, R. (2020). A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. Quantitative Science Studies, 1(4), 1570–1585. https://direct.mit.edu/qss/article/1/4/1570/96116/A-comparison-of-large-scale-science-models-based

Callaert, J., Pellens, M., Looy, B.V., 2014. Sources of inspiration? making sense of scientific references in patents. Scientometrics 98, 1617–1629. https://doi.org/10.1007/s11192-013-1073-x.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. arXiv preprint arXiv:2004.07180.

Du, J., Li, P., Guo, Q. and Tang, X., 2019. Measuring the knowledge translation and convergence in pharmaceutical innovation by funding-science-technology-innovation linkages analysis. Journal of informetrics, 13(1), pp.132-148.

Fani, H., Jiang, E., Bagheri, E., Al-Obeidat, F., Du, W. and Kargar, M., 2020. User community detection via embedding of social network structure and temporal content. Information Processing & Management, *57*(2), p. 102056.

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strat. Manage. J. 25, 909–928. https://doi.org/10.1002/smj.384

Fleming L., Greene, H., G Li, Marx, M., Yao, D. (2019), Government-funded research increasingly fuels innovation. Science 364(1139-1141).

Fortunato S. et al. (2018) Science of Science. Science 359(6379). DOI: 10.1126/science.aao0185

Gilliland, C., Zuk, D., Kocis, P. et al., 2016. Putting translational science on to a global stage. Nat Rev Drug Discov 15, p. 217–218.

Gosh Mainak, Sebastian Erhardt, Michael E. Rose, Erik Buunk, Dietmar Harhof (2024), "PaECTER: Patent-level Representation Learning using Citation-informed Transformers". https://arxiv.org/pdf/2402.19411

Griliches Zvi (1990), "Patent Statistics as Economic Indicators: A survey". JEL Vol. XXVIII 1661-1707. https://www.jstor.org/stable/2727442

Jaffe Adam B., Manual Trajtenberg, Michael S. Fogarty (2000), "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors". American Economic Review May 2000. https://www.aeaweb.org/articles?id=10.1257/aer.90.2.215

Ke Q. (2020), Technological impact of biomedical research: The role of basicness and novelty. Research Policy, 49(7). https://doi.org/10.1016/j.respol.2020.104071

Kim, J., Yoon, J., Park, E., & Choi, S. (2020). Patent document clustering with deep embeddings. Scientometrics, p. 1-15.

Kline S. J. and N. Rosenberg (1987), An Overview of Innovation, in Studies on the Science and Innovation Process: Selected Works by Nathan Rosenberg, 2009.

Lissoni F., F. Montobbio, L. Zirulia (2013) Inventorship and authorship as attribution rights: An enquiry into the economics of scientific credit - Journal of Economic Behavior & Organization. 95(49-69).

Marx M., A. Fuegi (2020) Reliance on science: Worldwide front-page patent citations to scientific articles. Strategic Management. Volume41, Issue9, September 2020, Pages 1572-1594 https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.3145

Marx Matt, Emma Scharfmann (2024), "Does Patenting Promote the Progress of Science?"

Masclans-Armengol Roger, Sharique Hasan, Wesley M. Cohen (2024), "Measuring the Commercial Potential of Science". NBER Working Paper 32262. https://www.nber.org/papers/w32262

McMillan, G.S., Narin, F., Deeds, D.L., 2000. An analysis of the critical role of public science in innovation: the case of biotechnology. Res. Policy 29, 1–8. https://doi.org/10.1016/S0048-7333(99)00030-X.

Murray F. and S. Stern (2007), Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. Journal of Economic Behaviour and Organization. 63(648-687)

Nagaoka S. and I. Yamauchi (2015), An economic analysis of deferred examination system: Evidence from a policy reform in Japan. International Journal of Industrial Organisation 39, 19-28.

Nelson R. R. (2004), The Market Economy and the Scientific Commons, Research Policy 33(455-471) https://www.econstor.eu/bitstream/10419/89454/1/391320661.pdf

OECD (2009), Patents Statistics Manual.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. arXiv preprint arXiv:2211.13308.

Verhoeven, D., Bakker, J., Veugelers, R., 2016. Measuring technological novelty with patent-based indicators. Res Policy 45, 707–723. https://doi.org/10.1016/j.respol. 2015.11.010.

Annexes

Annex 1: Examples of patents and papers

NOTA: Commonalities between the texts of the patent and the article are covered with a same colour.

Example 1: Citation and high semantic similarity (0.979)

Patent: USPTO publication number: 9409994

Title: High-affinity monoclonal antibodies to glypican-3 and use thereof

Abstract: Described herein is the identification of a panel of high affinity monoclonal antibodies that bind GPC3. The disclosed antibodies recognize native GPC3 on the surface of cancer cells, as well as soluble GPC3. The highest affinity antibody (YP7) was further characterized and shown to be highly sensitive in that it was capable of detecting cancer cells with low expression of GPC3. YP7 also exhibited significant HCC tumor growth inhibition in vivo. Immunotoxins comprising the antibodies disclosed herein fused to PE38 exhibited very high binding affinity for GPC3-expressing cells and significantly inhibited GPC3-expressing cancer cell growth. Thus, the high-affinity monoclonal antibodies disclosed herein can be used for the diagnosis and treatment of GPC3-expressing cancers.

Inventors: Mitchell H., Yen T. Phung, Wei Gao, Yifan Zhang

Publication:

Title: High-affinity monoclonal antibodies to cell surface tumor antigen glypican-3 generated through a combination of peptide immunization and flow cytometry screening

Abstract: Isolating high-affinity antibodies against native tumor antigens on the cell surface is not straightforward using standard hybridoma procedures. Here, we describe a combination method of synthetic peptide immunization and high-throughput flow cytometry screening to efficiently isolate hybridomas for cell binding. Using this method, we identified high-affinity monoclonal antibodies specific for the native form of glypcian-3 (GPC3), a target heterogeneously expressed in hepatocellular carcinoma (HCC) and other cancers. We isolated a panel of monoclonal antibodies (YP6, YP7, YP8, YP9 and YP9.1) for cell surface binding. The antibodies were used to characterize GPC3 protein expression in human liver cancer cell lines and tissues by flow cytometry, immunoblotting and immunohistochemistry. The best antibody (YP7) bound cell surface-associated GPC3 with equilibrium dissociation constant, K-D = 0.3 nmol/L and was highly specific for HCC, not normal tissues or other forms of primary liver cancers (such as cholangiocarcinoma). Interestingly, the new antibody was highly sensitive in that it detected GPC3 in low expression ovarian clear cell carcinoma and melanoma cells. The YP7 antibody exhibited significant HCC xenograft tumor growth inhibition in nude mice. These results describe an improved method for producing high-affinity monoclonal antibodies to cell surface tumor antigens and represent a general approach to isolate therapeutic antibodies against cancer. The new high-affinity antibodies described here have significant potential for GPC3-expressing cancer diagnostics and therapy.

Authors: Yen Phung, Wei Gao, Yan-Gao Man, Satoshi Nagata, Mitchell H.

Example 2: Citation but low semantic similarity (0.876)

Patent: USPTO 7994467 (filed, 2008; granted: 2011).

Title: Optical cavity emitter arrangements with spectral alignment and methods therefor

Abstract: Aspects of the disclosure are directed to optical microcavities and emitters that are spectrally aligned in an arrangement having an array of such microcavity-emitter combinations. The spectral alignment can be selective, in that a portion of the array of microcavity-emitter combinations, or a single microcavity-emitter combination, can be individually spectrally aligned. In specific examples, light is coupled within a semiconductor device having wavelength-dependent structures and optical cavities optically couple to the wavelength-dependent structures. One of the optical cavities and a wavelength-dependent structure are spectrally aligned, independent of another of the optical cavities.

Paper: in Applied Physics, 2007

Title: Local quantum dot tuning on photonic crystal chips

Abstract: Quantum networks based on InGaAs quantum dots embedded in photonic crystal devices rely on QDs being in resonance with each other and with the cavities they are embedded in. We developed a new technique based on temperature tuning to spectrally align different quantum dots located on the same chip. The technique allows for up to 1.8nm reversible on-chip quantum dot tuning.

Authors: Andrei Faraon, Dirk Englund, Ilya Fushman, Nick Stoltz, Pierre Petroff, Jelena Vuckovic

Example 3: Citation and low similarity (0.900)

Patent: applied 2016, granted 2021

Title: Method of identifying human compatible T cell receptors specific for an antigenic target

Abstract: Methods are provided for identifying T cell receptors that specifically bind a particular antigenic target and can be used as therapeutics against disease.

Inventors: Harlan S. Robins, Aude Georgiana Chapuis, Thomas M. Schmitt,Philip Greenberg, Anna Sherwood.

Paper: published in Genome Medicine, 2013

Title: Sequence analysis of T-cell repertoires in health and disease

Abstract: T-cell antigen receptor (TCR) variability enables the cellular immune system to discriminate between self and non-self. High-throughput TCR sequencing (TCR-seq) involves the use of next generation sequencing platforms to generate large numbers of short DNA sequences covering key regions of the TCR coding sequence, which enables quantification of T-cell diversity at unprecedented resolution. TCR-seq studies have provided new insights into the healthy human T-cell repertoire, such as revised estimates of repertoire size and the

understanding that TCR specificities are shared among individuals more frequently than previously anticipated. In the context of disease, TCR-seq has been instrumental in characterizing the recovery of the immune repertoire after hematopoietic stem cell transplantation, and the method has been used to develop biomarkers and diagnostics for various infectious and neoplastic diseases. However, T-cell repertoire sequencing is still in its infancy. It is expected that maturation of the field will involve the introduction of improved, standardized tools for data handling, deposition and statistical analysis, as well as the emergence of new and equivalently large-scale technologies for T-cell functional analysis and antigen discovery. In this review, we introduce this nascent field and TCR-seq methodology, we discuss recent insights into healthy and diseased TCR repertoires, and we examine the applications and challenges for TCR-seq in the clinic.

Authors: Daniel J Woodsworth, Mauro Castellarin, and Robert A Holt

Example 4: High similarity (0.966) but NO citation

Patent: US 10,472,427; applied 2013; granted 2019.

Title: Heterodimeric proteins

Abstract: The invention provides heterodimeric antibodies comprising a first heavy chain comprising a first Fc domain and a single chain Fv region that binds a first antigen. The heterodimeric antibodies also comprise a second heavy chain comprising a second Fc domain, a first variable heavy chain and a first variable light chain, wherein the first and second Fc domains are different.

Inventors: Desjarlais J.; Moore G.; Rashid R. ; Bernett M.

Paper: published in mAbs, 2013

Title: Progress in overcoming the chain association issue in bispecific heterodimeric IgG antibodies

Abstract: The development of bispecific antibodies has attracted substantial interest, and many different formats have been described. Those specifically containing an Fc part are mostly tetravalent, such as stabilized IgG-scFv fusions or dual-variable domain (DVD) IgGs. However, although they exhibit IgG-like properties and technical developability, these formats differ in size and geometry from classical IgG antibodies. Thus, considerable efforts focus on bispecific heterodimeric IgG antibodies that more closely mimic natural IgG molecules. The inherent chain association problem encountered when producing bispecific heterodimeric IgG antibodies can be overcome by several methods. While technologies like knobsinto-holes (KiH) combined with a common light chain or the CrossMab technology enforce the correct chain association, other approaches, e.g., the dual-acting Fab (DAF) IgGs, do not rely on a heterodimeric Fc part. This review discusses the state of the art in bispecific heterodimeric IgG antibodies, with an emphasis on recent progress.

Authors: Christian Klein, Claudio Sustmann, Markus Thomas, Kay Stubenrauch, Rebecca Croasdale, Jürgen Schanzer, Ulrich Brinkmann, Hubert Kettenberger, Jörg T. Regula and Wolfgang Schaefer.

Annex 2: Method for classifying documents in the three corpuses

1. Quantum cryptography

Patents:

CPC: 'H04L9/0852', 'H04L9/0855', 'H04L9/0858'

Key Words: bb84; quantum cryptography protocol%; QKD system%; QKD network%; quantum key distribution%; qubit% AND cryptography; key distillation%; quantum network%; quantum coin flipping; position\-based quantum cryptography; device\-independent quantum cryptography; quantum computing AND cryptography; quantum computing AND cybersecurity; quantum communication%; noisy intermediate\-scale quantum comput% AND cryptography; NISQ comput% AND cryptography; quantum repeater% AND quantum communication%; quantum teleportation% AND cryptography; entanglement% AND cryptography; entangled photon% AND cryptography; post quantum cryptography; quantum cryptography; lattice based cryptography

Publications:

OpenAlex concepts: "BB84","Lattice-based cryptography", "Quantum cryptography", "Post-quantum cryptography"; plus a keywords' (same as the above list) search in titles and abstracts of publications

2. CRISPR

Patents: %CRISPR%; CAS9

CPC: C12N2310/20

Publications: %CRISPR%; CAS9

3. CART-T Cells

Patents: two queries proposed by EPO patent examiners (EPO, 2019): https://link.epo.org/web/patent_insight_report-chimeric_antigen_receptor_t-cell_immunotherapy_en.pdf).

A first query identifies patents that were classified in the following Cooperative Patent Classification (CPC) :

C07K2319/03 in combination with either C07K2317/622 or C07K2317/55

A second query combines two classes of International Patent Classification (IPC) (C07K14/705 or A61K35/17) and search of terms ("CAR T" or "chimeric antigen receptor" or "chimeric T cell receptor") in the title or the abstract of patent documents.

Patent classification symbol	Type of classification	Description
C07K2319/03	CPC	Fusion polypeptide – containing a transmembrane segment
C07K2317/622	CPC	Immunoglobulins specific features – characterized by non-natural combinations of immunoglobulin fragments – comprising only variable region components – Single chain antibody (scFv)
C07K2317/55	CPC	Immunoglobulins specific features – characterized by immunoglobulin fragments – Fab or Fab'
C07K14/705	IPC	Receptors; Cell surface antigens; Cell surface determinants
A61K35/17	IPC	Blood; Artificial blood- Lymphocytes; B-cells; T-cells; Natural killer cells; Interferon-activated or cytokine-activated lymphocytes

Publications: we adopted a topic strategy search (including title and abstract of the academic papers) in the Web of Science database (WoS), version of June 05, 2023. The series of searching words were "CAR T", "CAR T cell", "CAR T therapy", "chimeric T cell receptor", and "chimeric antigen receptor".

Annex 3: Front-page and in-text (body) citations

The position of a citation in the patent (front-page vs. body) affects its legal status, and is sometimes interpreted as reflecting different levels of knowledge transfer: citations in the text would come more often from inventors and would reflect the actual use of knowledge, while citations in the front-page would rather come from attorneys and would reflect legal aspects, including strategies. Alternatively, it has been argued that front-page citations are more carefully curated than in-text ones, because they have legal implications.

Average proximities by categories of patents are reported in tables A3.1 to A3.3. Front-page NPL are clearly closer to the citing patent than in-text NPL, for the 3 corpuses. This holds both for title+abstract similarity and for full text similarity. These differences are statistically significant (Kruskal-Wallis test, <0.001). On this basis, it remains to be investigated what specific information body citations add to front-page ones.

Table A3.1: Semantic similarity with front-page and body NPL – quantum cryptography		
	Aggregate similarity	Granular similarity
Body	0.874	0.910
Front-page (incl. when body as well)	0.882	0.913

Table A3.2: Semantic similarity with front-page and body NPL – CRISPR		
	Aggregate similarity	Granular similarity
Body	0.869	0.876
Front-page (incl. when body as well)	0.886	0.882

Table A3.3: Semantic similarity with front-page and body NPL – CAR-T Cell		
	Aggregate similarity	Granular similarity
Body	0.883	0.910
Front-page (incl. when body as well)	0.894	0.916

Annex 4: Predictive models of citation

Table A4.1: Estimation of the predictive model for CAR-T cells (1999-2015)

	Publ_Cited_1999_15
Age_pub	0.373**
	(0.183)
Age_sq_pub	-0.029**
	(0.011)
Discipline *(ref. onco)*	
Hemato	-0.009
	(0.235)
Immuno	0.575***
	(0.192)
Other_discip	-0.361
	(0.331)
Pub_type *(ref. art)*	
Edit_mat	-0.595*
	(0.349)
Meet_abst	-1.516***
	(0.306)
Review	-0.123
	(0.188)
Other_type	1.140
	(1.080)
Open_access	0.778***
	(0.159)
Funded	0.830***
	(0.212)
Nb_auth *(ref. 1_3_auth)*	
4_7_auth	0.216
	(0.194)
8_ 11_auth	0.556**
	(0.226)
More_11_auth	1.074***
	(0.248)
Private	-0.118
	(0.274)
Inter_collab	-0.284
	(0.195)
Constant	-4.509***
	(1.505)
Pseudo R2	26.82%
Log Likelihood	-615.61
Correct classification	77.98%
Observations	1326

Table A4.2: Estimation of the predictive model for CRISPR corpora (2000-2015)

	Publ_Cited_2000_15
Age_pub	0.755***
	(0.161)
Age_sq_pub	-0.039***
	(0.010)
Discipline *(ref. onco)*	
Biochim_biol_molec	1.225***
	(0.453)
Bio_cel	-0.170
	(0.604)
Biotech_microbio	1.611***
	(0.465)
Botan_bio_veg	1.160*
	(0.647)
Genet_hered	0.238
	(0.585)
Hema_imuno	0.361
	(0.613)
Microbio	0.140
	(0.468)
Neurosc	0.044
	(0.775)
Other_discip	0.110
	(0.502)
Pub_type *(ref. art)*	
Edit_mat	-0.598**
	(0.274)
Review	0.194
	(0.147)
Other_type	-0.586**
	(0.247)
Open_access	0.753***
	(0.126)
Funded	0.448**
	(0.175)
Nb_auth *(ref. 1_4_auth)*	
5_6_auth	0.314**
	(0.131)
7_10_auth	0.150
	(0.132)
More_10_auth	0.288*
	(0.164)
Private	0.486***
	(0.179)
Inter_collab	-0.011
	(0.114)
Constant	-9.025***
	(1.323)
Pseudo R2	11.65%
Log Likelihood	-1,377.47
Correct classification	72.19%
Observations	2524

Annex 5: mRNA

Messenger RNA (mRNA) has become central to recent scientific advances, generating interest across various research fields. Once viewed as simply a genetic information carrier between DNA and proteins, mRNA now shows versatile applications and technological potential. Huang et al. (2022) provide a detailed overview of the rise of mRNA-based nanomedicine, highlighting its revolutionary role in science and medicine, as seen in the success of mRNA COVID-19 vaccines. However, significant challenges remain for this technology to reach its full potential.

mRNA regulates gene expression and contributes to protein synthesis essential for cell function. Its study helps researchers understand complex processes like development, cell differentiation, and responses to environmental stimuli, with implications for genetic diseases, cancer, and aging. Several studies (Crick, 1970; MSc, 2023; Schwanhäusser et al., 2011; Zhong, 2009) emphasize the importance of mRNA in molecular and cellular mechanisms, crucial for addressing future health challenges.

Despite progress, mRNA technology faces challenges in design, safety, efficacy, and delivery (Uddin & Roni, 2021; Verbeke et al., 2019; Zhang, 2023). Effective delivery while maintaining mRNA stability, especially in developing countries, remains a key obstacle.

Nevertheless, mRNA's role in advancing genetic engineering and biotechnology is evident. By improving the understanding of genetic mechanisms like gene expression and cell differentiation, mRNA supports the development of solutions for genetic diseases and cancer. Its success in COVID-19 vaccines shows the transformative potential of mRNA in medicine and biotechnology, driving innovation and accelerating new therapeutic, vaccine, and diagnostic technologies (Huang et al., 2022).

Methodology for Corpus Construction

The publications in the mRNA corpus were obtained by querying the PubMed API with the keyword 'mRNA'. To build the corpus of scientific publications on mRNA, a targeted query was submitted to the PubMed API, using the main keyword 'mRNA'. This query allows the identification of all relevant publications related to mRNA by leveraging associated terms through the MeSH (Medical Subject Headings) thesaurus. Table 1 presents the list of MeSH terms corresponding to 'mRNA' (unique MeSH ID: D012333), as well as the synonyms used in indexed publications.

This list of keywords was then used to query the Patstat database to gather a list of patents mentioning mRNA in the title, abstract, or claims.

Table A5.1: MeSH terms for mRNA (unique ID: D012333) and synonyms

MAIN TERME	SYNONYMS
MRNA	Poly(A) Tail ; RNA, Messenger ; Polyadenylated
MESSENGER RNA	mRNA ; Non-Polyadenylated ; Messenger RNA ; Polyadenylated
RNA, POLYADENYLATED	Polyadenylated Messenger RNA ; Non-Polyadenylated mRNA
POLY(A) RNA	mRNA, Polyadenylated, Poly(A)+ mRNA

To refine the corpus, we created a 'core' subset, defined as a research domain specifically aimed at advancing mRNA-related science and technologies. For scientific publications, only those assigned the MeSH term 'mRNA' by NLM experts were included in this 'core' subset. As for patents, we selected those mentioning a relevant keyword in the text and whose CPC code primarily falls under the biotechnological categories C12N and C12Q. This approach targets the technological and scientific developments most directly related to mRNA.

The core corpus was then enriched with documents—both patents and publications—having a direct citation link to one of the core elements. Thus, the 'non-core publications' subset includes all scientific publications cited by patents without belonging to the core of mRNA research. On the other hand, the 'non-core patents' subset consists of patents other than those in the 'mRNA core' but that cite 'mRNA core' publications.

Table A5.2: Number of patents, publications NPL per corpus

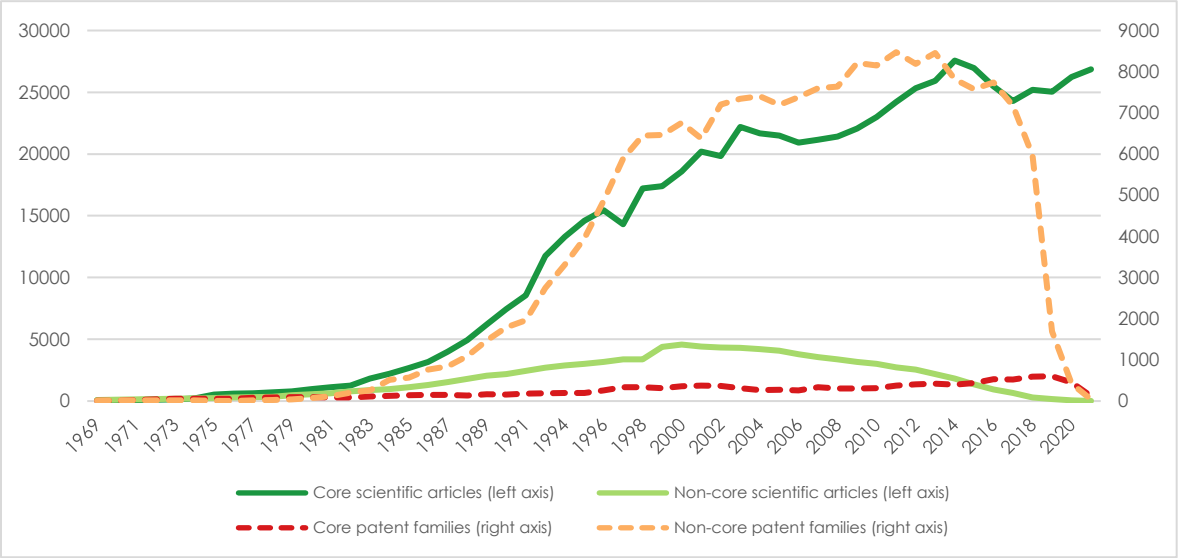
Corpus	**Core**	**Non core**
Publications	387 362	42 085
Patent families	5 037 (3347 granted)	142 099 (125 812 granted.)

Patents that only mention mRNA in the claims and are not cited by core publications form a separate subset that will not be detailed here. The two corpora are not limited in time, as the interest here is also to examine trends over time. The figure below provides an overview of temporal trends in these numbers.

According to Figure 1, we observe that the rise of mRNA research really begins in the 1990s. The curve for 'core' scientific publications shows a strong growth from this period, illustrating an explosion in mRNA research activities. However, on the patent side, 'core' patent families remain relatively low in number, with much more moderate growth. This confirms the presence of significant obstacles in the technological development of mRNA, despite considerable scientific advances.

Additionally, we note a rise in related technologies, represented by 'non-core' patents and publications, which build upon advancements in mRNA scientific research. These technologies, while not directly part of mRNA developments, leverage the knowledge generated in this field to innovate and thrive.

Figure A5.1: mRNA corpuses – time trends



Citation links between patents and publications

Among mRNA granted patent families, 98%, or 3,279, cite at least one scientific publication. Each family cites an average of 48.17 publications, with one citing up to 1,429. These figures highlight the strong dependence of mRNA technologies on academic work.

45% of NPL references in mRNA patents come from the core of the field, while non-core patents draw 61% of their references from the core of mRNA publications. This illustrates the significant influence of core mRNA knowledge, even for technologies that are not directly related to this field.

Semantic links

Another exercises addresses semantic proximities between scientific publications and patents with a direct citation link. The objective is to determine whether this citation link reflects a real semantic proximity. In our set of core patent families, the average similarity between patents and their cited publications is 0.823. However, in the case of mRNA technologies, this measure does not significantly distinguish between the core and non-core sets.

Lastly, for the core patent set, we compared the number of NPL references (non-patent literature) with the level of semantic proximity (nearest neighbour: NN1, NN5, NN20 and NN100), as defined in the main text. The correlation levels is positive and significant, but rather weak (around 0.2 for core patents, 0.15 for non core patents).

Table A5.3: Pearson Correlation between the number of core NPL and semantic proximity of patents – mRNA				
	Core patents		**Non Core patents**	
(Core)	Pearson	p-value	Pearson	p-value
NN1	0,225	9,77E-08	0,143	2,31E-07
NN5	0,216	2,91E-07	0,151	4,57E-08
NN20	0,203	1,50E-06	0,147	1,09E-07
NN100	0,176	3,20E-05	0,131	2,14E-06

Table A5.4 presents the percentage of non-patent literature (NPL) citations that are among the nearest neighbors of the citing patent, grouped by different ranges of the number of NPLs cited per family. The columns correspond to different levels of nearest neighbors (N=1, N=5, N=20, N=100). The percentages reflect how often the cited NPLs appear in the top semantic matches for each patent.

Tableau A5.4: Percentage of NPL citations which are among the nearest neighbours of the citing patent - mRNA patents (granted)

NB NPL	NB of Family id	N=1	N=5	N=20	N=100
1	154	0,00%	0,00%	0,00%	0,00%
2	160	0,00%	0,00%	0,00%	0,00%
3	181	0,00%	0,00%	0,00%	0,00%
4	181	0,00%	0,00%	0,00%	0,00%
5	174	0,00%	0,00%	0,00%	0,00%
6	140	0,00%	0,00%	0,00%	0,00%
7	125	0,00%	0,00%	0,00%	0,00%
8	101	0,00%	0,00%	0,00%	0,00%
9	83	0,00%	0,00%	0,00%	0,00%
10	87	0,00%	0,00%	0,00%	0,00%
[11,20]	773	0,00%	0,00%	0,13%	0,52%
[21,30]	474	0,00%	0,00%	0,00%	1,48%
[31,50]	670	0,00%	0,00%	0,15%	0,75%
[51,]	999	0,00%	0,40%	0,80%	3,10%

References to annex 5:

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

Huang, X., Kong, N., Zhang, X., Cao, Y., Langer, R., & Tao, W. (2022). The landscape of mRNA nanomedicine. *Nature Medicine*, *28*(11), 2273–2287. https://doi.org/10.1038/s41591-022-02061-1

MSc, A. D. B. (2023, January 6). *mRNA Medicine: What's next after the COVID-19 vaccine?* News-Medical. https://www.news-medical.net/health/mRNA-Medicine-whate28099s-next-after-the-COVID-19-vaccine.aspx

Schwanhäusser, B., Busse, D., Na Li, Dittmar, G., Schuchhardt, J., Wolf, J., Wei Chen, & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature (London)*, *473*(7347), 337–342.

Uddin, M. N., & Roni, M. A. (2021). Challenges of Storage and Stability of mRNA-Based COVID-19 Vaccines. *Vaccines*, *9*(9), Article 9. https://doi.org/10.3390/vaccines9091033

Verbeke, R., Lentacker, I., De Smedt, S. C., & Dewitte, H. (2019). Three decades of messenger RNA vaccine development. *Nano Today*, *28*, 100766. https://doi.org/10.1016/j.nantod.2019.100766

Zhang, C. Y. W. (2023). Revolutionizing vaccinology: The rise of mRNA vaccine. *Theoretical and Natural Science*, *21*(1), 103–108. https://doi.org/10.54254/2753-8818/21/20230839

Zhong Wang, Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics (Print)*, *10*(1), 57–63.