The International Diffusion of Technology: A New Approach and Some Facts

Cyril Verluise[1] and Antonin Bergeaud[2]

***Abstract*— Building on the nascent *automated* patent landscaping literature, we introduce a new approach to study the international diffusion of technologies. We deploy and validate our approach on the *genome editing* technology. We document a significant technological lag between leading and laggard countries. This lag is even larger when we consider the country of origin of inventors and assignees instead of the granting offices.**

I. Introduction

Since the seminal contributions by [Romer, 1990] and [Aghion and Howitt, 1992], modern growth theory acknowledges the central role of technological progress in driving long-run economic growth. The creation of new inventions, but more importantly, the way they subsequently diffuse across the world have then been pointed as extremely relevant to boost productivity.[1] While patterns and determinants of technology diffusion have been largely documented at the *macro* level[2], less is known about the international diffusion of technologies at the *micro* level, that is at the level of a specific technology.[3]

This gap in the existing literature reflects the *classification* problem already identified by [Griliches, 1990]. Defining the scope of patents that falls into a technology is difficult without a prior technology classification. Early attempts to do so were based on patent offices' technology classifications. Nonetheless, the purpose of patent offices' classifications is to ease the search of prior art.[4] Resulting classifications are thus based on functional principles which are not necessarily related to economists' notion of technology[5], that is an industry or product grouping in which "the economic processes that lead to the reduction in cost of producing existing products and the development of new products and services"[6] occur.

In this context, our main contribution is to introduce a candidate approach to solve [Griliches, 1990]'s classification problem. Building on the nascent *automated* patent landscaping literature, we provide an efficient and scalable method to generate clusters of patents that belong to a well defined technology across the universe of patents and countries. This approach overcomes several important limitations that have been identified in the literature regarding the use of patent data to identify the international diffusion of technologies, in particular the sole use of citations.

The remaining of this paper is organized as follows: section II briefly summarizes the related literature, section III describes our approach, section IV deploys it for the *genome editing* technology, section V presents our preliminary results regarding the worldwide diffusion of this technology and highlights a significant lag between leading and laggard countries, section VI discusses immediate research perspectives.

II. Review of the Literature

Historically, [Trajtenberg, 1990] tackled the classification problem by manually curating patents belonging to the Computed Tomography Scanners technology. Although fruitful, this solution is too

[1]Paris School of Economics, cyril.verluise@gmail.com.

[2]Banque de France and Centre for Economic Performance, antonin.bergeaud@banque-france.fr.

[1][Eaton and Kortum, 1999] and [Keller, 2002] find that the primary sources of technological change in OECD countries are not domestic but rather from international trade and spillovers.

[2]See [Comin and Mestieri, 2014] for a review.

[3]We borrow the distinction between the *micro* and the *macro* level literature of technology diffusion from [Peri, 2005]. While the *micro* branch uses micro data such as firm-level or patent data to develop the analysis of diffusion for a well defined technology, the *macro* branch examines technological flows through large aggregated units such as countries and their productivity.

[4]See [Bergeaud et al., 2017] for a discussion.

[5]E.g. [Schmookler, 1966] reports that a subclass related to the dispensing of solids contained patents on both manure spreaders and toothpaste tubes.

[6][Griliches, 1990, p1669]

labor-intensive to be extended to a larger corpus, even more so at the international level[7].

Another approach to link technologically related patents is to rely on patent citation flows as an evidence of actual technology spillover. This approach pioneered by [Jaffe et al., 1993] had a dramatic impact on our understanding of ideas diffusion at the national level. However, the use of patent citations at the international level raises more questions. By nature, such written references miss flows that occur without explicit citations. In this context, when [Jaffe and Trajtenberg, 1999] find that "patents whose inventors reside in the same country are typically 30 to 80% more likely to cite each other, and [that] these citations come sooner", it is unclear what it really means regarding how technology spreads across borders. Is it that technology diffusion is much slower when it involves different countries? Or simply that patents have a larger propensity to cite patents from the same country?

More recently, [Webb et al., 2018] use a combination of user-defined Cooperative Patent Classification (CPC) classes and keywords to define technologies and constitute groups of US patents[8]. While the use of multiple features seems promising, this approach is also susceptible to generate a large amount of false positives and false negatives if the set of user-defined conditions is insufficiently well designed. This sends us back to the same limitation as in [Trajtenberg, 1990]. Both approaches require human input to define the set of rules that shapes a technology.

That's where the *automated* patent landscaping introduced by [Abood and Feltenberger, 2018] makes an important breakthrough. The authors develop a *semi-supervised* machine-learning framework to emulate human-made technology classification. The algorithm only requires a small set of patents as input – the *seed* – which must be representative of the technology of interest. Just as humans would do, their algorithm then *expands* to "probably related" patents. Eventually, false positives are *pruned* out using a classification algorithm trained to distinguish the seed from an *anti-seed* – a set of patent randomly drawn out of the expansion and therefore "probably unrelated" to the seed. This approach delivers a large group of patents in the technology of interest at virtually no cost. Importantly, no human intervention is needed to elaborate the set of rules determining whether a patent belongs or not to the target technology. Patterns are learned from the data.

[7]As a matter of example, there were more than 629,000 granted patents in the US in 2015.

[8]For example, they define the "Smartphones" technology through the H04 CPC-class (electric communication techniques).

III. A New Approach to Griliches' Classification Problem

In this section, we detail our approach and emphasize extensions with regard to [Abood and Feltenberger, 2018] baseline algorithm.

A. Expansion

Starting from a set of selected patents, typically between 300 and 1,000 patents representative of the technology of interest, the role of the expansion is to select all "probably related" patents. To do so, two main similarity features are available, the technological patent classes and the patent citations. At this stage, we follow the sequence proposed by [Abood and Feltenberger, 2018] (see Figure 1). Starting from the seed, we first expand to patents belonging to "important" CPC classes and then proceed to two consecutive levels of citations expansions. Citations expansions include both backward and forward patent citations.

We slightly depart from the baseline algorithm in the way we define "important" CPC classes. [Abood and Feltenberger, 2018] define the latter as CPC classes which are 1.5 more frequent in the seed than in the universe of all-time US patents. Our counterfactual distribution is drawn from the worldwide universe of patents with a publication year falling in the range of the seed publication years. This modification is motivated by two simple facts. First, we are interested in the *international* diffusion of technologies. Hence, there is no reason to restrict to US patents. Second, the technology of interest might be strictly delimited in time and we do not want the distribution of CPC classes at a very different period to influence what an important CPC class is.

B. Anti-seed

The role of the anti-seed is to define a set of examples which do not belong to the technology of interest from which the classifier model can learn seed-specific patterns at the next step. To do so,

[Abood and Feltenberger, 2018] draw a random set of patents from the complement of the expansion set (see Figure 1).

At this point, we made a substantial extension as we believe that such an anti-seed might not be sufficient to properly train the pruning classifier. Intuitively, the examples in the anti-seed are likely to be *too* far from the seed, hence making the classification task too simple. This is problematic because the classifier will later be applied to "intermediary" cases in the expansion set. In order to make sure that the classifier is trained on realistic examples, we augment the anti-seed with patents drawn from CPC classes which share a common CPC *group* with an important CPC *class* but are not important themselves. For example, if the CPC class `A01B1/002` has been found important, any patent in a CPC class from the `A01B1/` CPC *group*[9] might be selected to be part of the anti-seed provided it is *not* in the expansion set.

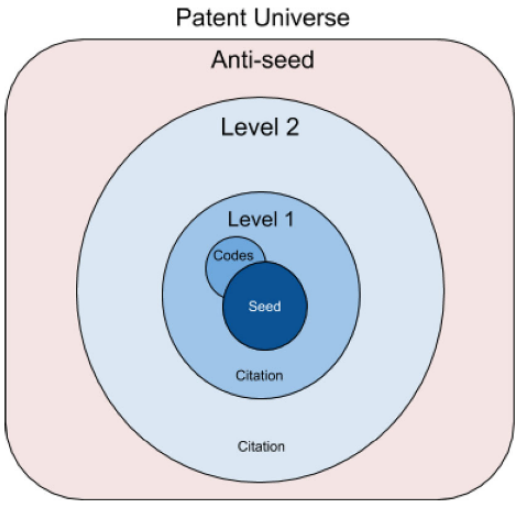

Fig. 1. Graphical representation of the expansion step (from [Abood and Feltenberger, 2018]). Starting from a seed (`Seed`), we first expand to patents belonging to "important" CPC classes (`Codes`) and then proceed to two consecutive levels of citations expansions (`Level 1` and `Level 2`). Citations expansions include both backward and forward patent citations. The anti-seed (`Anti-seed`) is randomly drawn from the complementary set of the expansion.

[9] `A01B`: Soil working in agriculture or forestry; Parts, details, or accessories of agricultural machines or implements, in general, `A01B1/00`: Hand tools.

C. Pruning

The role of the pruning stage is to remove patents which are in the expansion but do not belong to the technology of interest. The baseline idea is to train a binary classifier on the seed versus the anti-seed and then to apply the fitted model to patents in the expansion set. If the predicted class is `seed`, then the patent is kept as part of the technology of interest. Otherwise, it is discarded. [Abood and Feltenberger, 2018] consider different types of classification models based on at least 3 features: CPC classes, references and abstracts.

In our view, both the CPC classes and the references imply potential pitfalls. First, the use of CPC classes in both the expansion and the classification model can generate pathological cases. Assume that all CPC classes in the seed are found important, then the anti-seed and the seed have no CPC in common which makes the classification task trivial. Second, by construction, patents in the `Level 2` expansion have no references in common with the seed. Hence, considering references in the classification task implies a systematic and uncontrolled bias against patents in the `Level 2` expansion which is undesirable. In our view, this motivates a significant rethinking of the pruning model.

In this context, we consider only the patents abstract[10] in our classification model. The classification task is carried by a Convolutional Neural Network (CNN) on top of trainable word embeddings as we know since [Kim, 2014] that these models achieve high performances on a large set of text classification benchmarks. Intuitively, CNNs are a class of neural network primarily designed to deal with sequential inputs and are able to detects sub-sequences (e.g. n-grams) in a shift invariant way. These characteristics can account for their many successes in text classification and in Natural Language Processing (NLP) in general since text is sequential by nature. Hyper-parameters are tuned[11] following the conventional practices listed in [Zhang and Wallace, 2015]. We select the best model

[10] We use machine-translated abstracts available in the Google Patents Research Data. See Appendix A for more details on the data.

[11] See Appendix B.

according to the product of the F1-scores[12] of the two classes. Finally, we apply the selected model to the expansion and keep only patents which abstract is predicted in the seed-class.

IV. Deployment and Validation on the Genome Editing Technology

We are unaware of any previous applications of automated patent landscaping in economics. Hence, we believe that this is crucial to walk through a simple example and to cautiously validate the output in order to convince the reader. This is the purpose of the next section. Importantly, this example ran through our *standardized* pipeline (see section III), meaning that it is fully reproducible and can be thought as representative of the quality of our output in general.

A. A brief presentation of the genome editing technology

Genome editing is a rapidly growing scientific field. The three competing technologies in the field are *zinc fingers*, *TALENs* and *CRISPR-Cas9*.[13] As described by [Ledford, 2015], researchers used to rely on zinc fingers, a class of enzymes, in order to accurately edit genomes. However, such enzymes were rather expensive[14]. In 2012, CRISPR-Cas9 – CRISPR later on – was introduced. In addition to being more efficient and easy to use, it is also much cheaper than previous technologies, including TALENs, the third competing method. As an order of magnitude, CRISPR costs about 150 times less than zinc fingers. It is now widely used and a very active subject of research and invention.

B. Deployment

We start from a seed of 300 patents representative of the genome editing technology. This seed was generated from the 300 most similar patents to the `US8697359B1` patent (*CRISPR-Cas systems and methods for altering expression of gene products*) based on the Google Patents similarity feature[15]. Following the routine described in section III-A, the algorithm expand to 1,429,615 unique patents. The best model obtained from the routine described in section III-C is a shallow CNN with just one convolutional layer[16]. It achieves a F1-score of .7 on the seed class and .96 on the anti-seed class. Note that the lower score obtained on the seed class, although still reasonable, comes from a lower recall score. Intuitively, this means that this model tends to be "conservative" rather than "over-inclusive" when it comes to classifying a patent as belonging to the genome editing technology. At this point, for the sake of analysis, we restrict to patents granted by the G7 and BRICS patent offices[17]. After model pruning, we end up with 15,949 patents classified in the seed technology, that is around 0.1% of the expansion set.

C. Validation

We now discuss the relevance of our final set of genome editing technology-related patents. We show evidence that the output of the above described methodology is sensible based on qualitative information we were able to collect on this technology.

1) Abstracts: As a first exercise, we consider the abstract of the final group of patents.

To begin with, we look at the share of patent abstracts which contain a list of n-grams known to be related to the genome editing technology. We find that `crispr`, `zinc finger` and `talens` are respectively in 8.2%, 3.3% and 0.5% of the patents abstract. More general but still related n-grams such as `gene` and `dna` are respectively in 58.8% and 41% of the patents abstract.[18] There are two competing explanations at this point. Either we have a lot of false positives, or, there are patents belonging to the genome editing technology which do not use these words. In order to distinguish between these two alternatives, we randomly browse patent abstracts which do not contain *any* of the n-grams mentioned above, that is 19.7% of the patents.

Qualitative assessment provide strong support for the second alternative. Patents which abstract does not

[12]For a given class, the F1-score is the harmonic mean of the precision and the recall score. The precision score is the proportion of samples classified in the class which were actually correct while the recall score is the proportion of samples belonging to the class which were actually identified as such.

[13]See [Gaj et al., 2013] for a more technical and thorough comparison of different genome editing technologies.

[14]5,000 dollars or more in 2015

[15]This is a simple and time-efficient way to generate a seed. Future work will investigate alternatives and their impact on the final output.

[16]`blocks=1`, `filters=64` and `kernel_size=7` Other hyper-parameters are reported in Appendix B.

[17]G7: Canada, France, Germany, Italy, Japan, Great Britain and the United States of America. BRICS: Brazil, Russia, India, China and South Africa.

[18]See Appendix D.

contain any of the keyword mentioned above but appear to be in our final group of patents are however related to the genome editing technology. For example, we report the abstract of patent `JP2012500627A` below:

> "To promote targeted integration of one or more foreign sequences, discloses methods and compositions for producing single-stranded breaks in the target sequence herein."

Of course, we are aware that a single example cannot be sufficient. Hence, we provide the list of the 15,949 patents in our final group with their abstracts in our online data appendix. We encourage any reader to go though the data to make his own mind on their overall consistency.

This gives us the opportunity to discuss an important issue. Keyword boolean queries, which tend to get a larger audience as the use of text gets popular, could well provide low-quality patent groups. In the present example, this kind of clustering techniques based on domain-specific keywords (`crispr`, `cas9`, `zinc finger` and `talens`) would have led to neglect *85.6%* of the patents we find in the genome editing technology but which abstract does not contain any of these keywords. This is where the machine-learned text classifier introduces a major breakthrough. It learns complex patterns at *no* human cost.

2) Inventors: Our second exercise consists in looking at the most prolific inventors in our sample.

Rank	Name	Number of patents
1	Feng Zhang	75
2	Kurt Nurith	52
3	Willem P. C. Stemmer	50
4	Jay M. Short	48
5	David Liu	42
6	Fyodor Urnov	41
7	Jeffrey Miller	38
8	Michael C. Holmes	31
9	Wang Jianbin	30
10	George M. Church	28

TABLE I

TOP 10 INVENTORS IN THE GENOME EDITING TECHNOLOGY

We report the top 10 inventors in Table I. All these are famous specialists of the technology and can be easily tracked online. For example, Feng Zhang's Wikipedia page[19] states that he is "most well-known for his central role in the development of *optogenetics* and *CRISPR* technologies". Similarly, Kurt Nurith is "the scientific founder of NuGEN and is responsible for the creation of the Company's foundational amplification technology. (...) Nurith (...) has numerous scientific publications and issued patents in the field of *nucleic acid amplification* and *detection*."[20] A last example is Jay M. Short, founder and CEO of the antibody drug company BioAtla, LLC. According to his Wikipedia page[21]: "while at Diversa, Short invented methods of protein and pathway discovery via *metagenomics*, in addition to evolution technologies *gene site saturation mutagenesis (GSSM)* and *GeneReassembly*, and was the first to combine these discovery and evolution technologies". We also searched for relevant genome editing-related material for each of the top 10 inventors. The reader should feel free to follow the links in Table I.

3) Assignees: Finally, our third exercise replicates the same analysis but this time we consider assignees instead of inventors.

Rank	Name	Number of patents
1	Sangamo Therapeutics	311
2	Harvard College	275
3	University of California	225
4	Pioneer Hi-Bred	176
5	Regeneron Pharma	150
6	Massachusetts Institute of Technology	145
7	Life Technology Corp	120
8	Stanford University	109
9	New England BioLabs	109
10	Dow Agrosciences	102

TABLE II

TOP 10 ASSIGNEES IN THE GENOME EDITING TECHNOLOGY

We report the top 10 assignees in Table II. We begin with non-university assignees. Sangamo Therapeutics, the biggest assignee in our sample, is a California-based biotechnology company whose genome editing technology involving *zinc finger nuclease* is currently under trials[22]. Sangamo's scientists have also published

[19] See https://en.wikipedia.org/wiki/Feng_Zhang.

[20] See https://biography.omicsonline.org/india/nugen/nurith-kurn-1305187.

[21] See https://en.wikipedia.org/wiki/Jay_Short.

[22] See https://en.wikipedia.org/wiki/Sangamo_Therapeutics.

numerous related scientific publications[23]. The second largest assignee in this group is Pioneed Hi-Bred, a major producer of genetically modified organisms, including genetically modified crops with insect and herbicide resistance. More generally, firms listed in Table II are either biotech or chemical/agricultural companies[24], which is consistent with the genome editing technology.

Regarding universities in the top 10 assignees, we should not be surprised to find top-tier universities. This is typical for technologies at their early development stages[25]. In addition, the Massachussets Institite of Technology is known to be the host university of leading academics in the genome editing field such Feng Zhang, our top inventor.

All in all, these qualitative evidence suggest that our patent group is indeed representative of genome editing technological ecosystem, with consistent *contents*, well-known *inventors* and *assignees* that cover different sectors and end-of-use products.

V. Preliminary Results

The purpose of this report is mainly to introduce and validate a new approach. We do *not* pretend to establish stylized facts on the international diffusion of technologies in the following lines. That being said, we also believe that some simple results obtained on the genome editing patent landscape are worth mentioning and raise challenging questions for future work.

A. The geography of patent grants

We find that patent grants are highly concentrated in a small subset of patent offices. Within the G7 and the BRICS, 71% of the patents granted on the genome editing technology were granted by the US office, 11% by the Chinese office, 8% by the Japanese office and 6% by the Canadian office. The remaining 7 offices account for less than 4% of the total set of patents.

B. Cross country patent grant dynamics

Patent grants in the genome editing technology has known a first peak in the early 2000s in the US while it was virtually non-existing in China at that time. Since the early 2010s, Chinese patent grants have taken up, although China is still lagging behind the US in absolute terms. This lag in patent grants dynamics and the rapid increase in China is consistent with the usual technological catch-up story. In the middle of these two polar cases, Japan and Canada started granting patent in the technology as early as the US but in much smaller quantities and achieved their maximum number of grants in the late 2010s. Overall, we observe a *10-year lag* between leading and laggard countries in terms of patent grants dynamics.

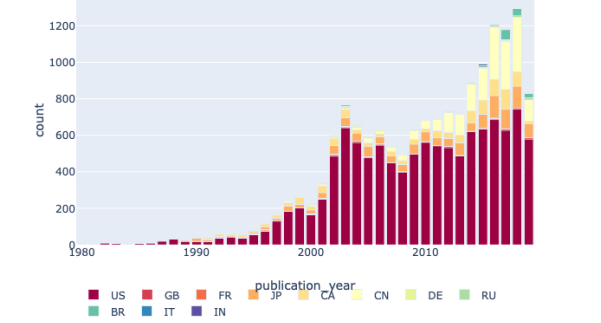

Fig. 2. Number of patents in the genome editing technology granted by year and country office.

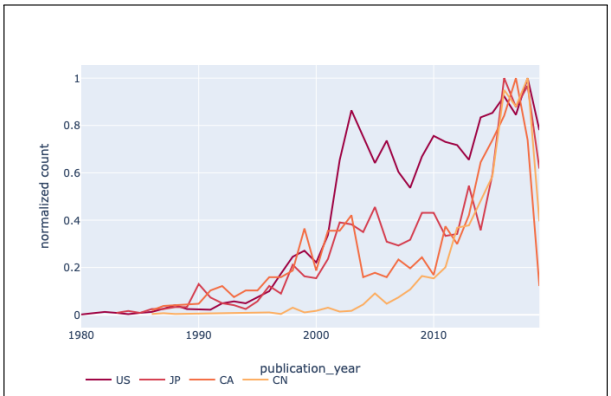

Fig. 3. Normalized count of patent grants in the genome editing technology by year in the four major granting countries. Each patent office time series is scaled by the largest number of patents it granted at some point in time.

C. The geography of patentees

As stated earlier, patent grants out of the US, China, Japan and Canada are residual. In particular, the Chinese patent office has been found to represent up to 11% of all-time G7 and BRICS patent grants in the

[23] See https://www.sangamo.com/technology/publications.

[24] Follow the links in Table II for relevant material.

[25] See [Jaffe et al., 1989] *inter allia*.

genome editing technology. This rises to 22% in recent years. In this context, it is striking to see that China is the country of origin of less than 2% of assignees and inventors. This suggests that the time lag in terms of technology adoption by local inventors and assignees might be even *larger* than the aforementioned 10 years. More generally, we observe that the geography of patentees is markedly different from the geography of granting offices. While the US patent office granted 71% of all-time patents in the technology, US assignees and inventors are actually responsible for 59% and 51% of this technology patents, respectively. On the other side, German, British and French assignees and inventors issued between 2 and 3% of all-time patents each, while their host countries have close to 0 patenting activity in this technology. This gap between the geography of patent grants and the geography of patentees suggests a clear distinction between the international *diffusion* of technologies, which can be proxied by patent grants, and the international *absorption* of technologies, which would be better reflected by domestic inventors and assignees appropriating the technology and holding patents.

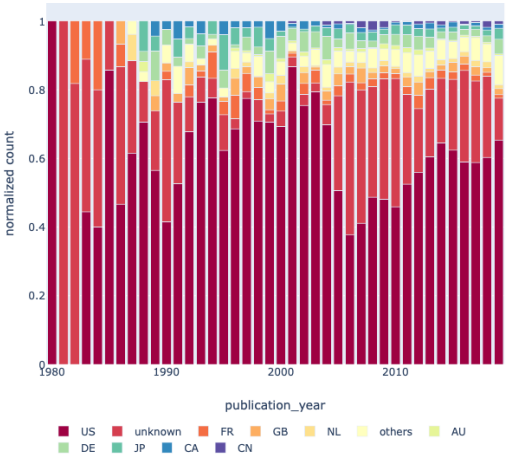

Fig. 4. Assignees' country of origin of genome editing patent granted in the G7 and BRICS patent offices, in proportion.

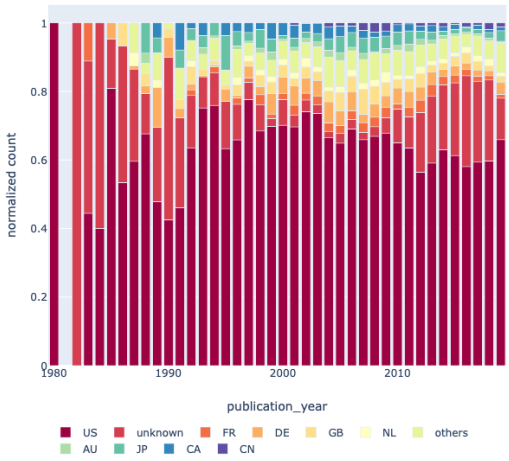

Fig. 5. Inventors' country of origin of genome editing patent granted in the G7 and BRICS patent offices, in proportion.

VI. Perspectives

At this point, we believe that we can be reasonably confident in our ability to leverage our automated patent landscaping algorithm at a larger scale. Our immediate next step is thus to implement it on a large set of technologies in order to identify the *stylized facts* of the international diffusion of technologies.

Beyond that, we plan i) to contextualize our findings in a longer historical time span and ii) investigate the determinants of the international technology diffusion. To do so, we have identified two main gaps in existing datasets: the location of patentees before the 1980s and the interactions between patents and science for non-US patents. We are currently deploying efforts to develop open source projects in these two directions[26].

[26] See [Bergeaud and Verluise, 2019] and [de Rassenfosse and Verluise, 2019].

APPENDIX

A. app:data

Variable	Source	Code
Patent office	PAT	country_code
CPC class	PAT	cpc.code
Backward citations	PAT	citation .publication_number
Forward citations	GPR	cited_by .publication_number
Abstract	GPR	abstract
Inventor name	PAT	inventor_harmonized .name
Inventor country	PAT	inventor_harmonized .country_code
Assignee name	PAT	assignee_harmonized .name
Assignee country	PAT	assignee_harmonized .country_code

Where PAT and GPR respectively refer to the patents-public-data:patents.publications and patents-public-data:google_patents_research.publications tables publicly available in the Google Patents Public Dataset.

B. Model tuning

Hyper-parameter	Value range
learning_rate	1e-3
epochs	100[27]
batch_size	64
blocks	[1, 2, 3]
filters	[64, 128]
dropout_rate	0.2
embedding_dim	100
kernel_size	[3, 5, 7]
pool_size	3

C. Number of patents by expansion level

Expansion level	Number of patents
SEED	300
CPC expansion	48,959
L1-BACK	48,287
L1-FOR	96,388
L2-BACK	504,601
L2-FOR	731,440
ANTISEED-AF	750
ANTISEED-AUG	750

L1 and L2 refer to the Level 1 and Level 2 expansions, BACK and FOR suffixes indicate backward and forward citations and AF and AUG suffixes designate the anti-seed *à la* [Abood and Feltenberger, 2018] and the *augmented* anti-seed.

[27] Early-stopping when the loss does not decrease for two consecutive epochs.

D. Patents in the genome editing technology with <N-gram> in the abstract

N-gram	Number of patents	Share of patents
gene	9,376	58.8%
dna	6,522	41.0%
edit	963	6.0%
crispr	1,313	8.2%
cas9	1,063	6.7%
zinc finger	540	3.3% %
talens	81	0.5%
any	12,793	80.3%

ACKNOWLEDGMENT

Cyril Verluise is thankful to Google for supporting this project through a grant from the Google Cloud Platform (GCP research credits programme grant).

DISCLAIMER

Preliminary draft. Do not cite, do not circulate.

REFERENCES

[Abood and Feltenberger, 2018] Abood, A. and Feltenberger, D. (2018). Automated patent landscaping. *Artificial Intelligence and Law*, 26(2):103–125.

[Aghion and Howitt, 1992] Aghion, P. and Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, 60(2):323–351.

[Bergeaud et al., 2017] Bergeaud, A., Potiron, Y., and Raimbault, J. (2017). Classifying patents based on their semantic content. *PloS one*, 12(4):e0176310.

[Bergeaud and Verluise, 2019] Bergeaud, A. and Verluise, C. (2019). Patentcity: An open source framework to extract inventors location from patent documents.

[Comin and Mestieri, 2014] Comin, D. and Mestieri, M. (2014). Technology Diffusion: Measurement, Causes, and Consequences. In *Handbook of Economic Growth*, volume 2 of *Handbook of Economic Growth*, chapter 2, pages 565–622. Elsevier.

[de Rassenfosse and Verluise, 2019] de Rassenfosse, G. and Verluise, C. (2019). Scicit: An open source worldwide patent-to-science dataset.

[Eaton and Kortum, 1999] Eaton, J. and Kortum, S. (1999). International technology diffusion: Theory and measurement. *International Economic Review*, 40(3):537–570.

[Gaj et al., 2013] Gaj, T., Gersbach, C. A., and Barbas III, C. F. (2013). Zfn, talen, and crispr/cas-based methods for genome engineering. *Trends in biotechnology*, 31(7):397–405.

[Griliches, 1990] Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4):1661–1707.

[Jaffe et al., 1989] Jaffe, A. B. et al. (1989). Real effects of academic research. *American economic review*, 79(5):957–970.

[Jaffe and Trajtenberg, 1999] Jaffe, A. B. and Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology*, 8(1-2):105–136.

[Jaffe et al., 1993] Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations*. *The Quarterly Journal of Economics*, 108(3):577–598.

[Keller, 2002] Keller, W. (2002). Geographic localization of international technology diffusion. *The American Economic Review*, 92(1):120–142.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[Ledford, 2015] Ledford, H. (2015). Crispr, the disruptor. *Nature News*, 522(7554):20.

[Peri, 2005] Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *The Review of Economics and Statistics*, 87(2):308–322.

[Romer, 1990] Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2):S71–S102.

[Schmookler, 1966] Schmookler, J. (1966). *Invention and Economic Growth*. Harvard U.P.

[Trajtenberg, 1990] Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*, 21(1):172–187.

[Webb et al., 2018] Webb, M., Short, N., Bloom, N., and Lerner, J. (2018). Some facts of high-tech patenting. Technical report, National Bureau of Economic Research.

[Zhang and Wallace, 2015] Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.