Supplementary Materials for
A crosswalk between paper and patent authors

E. Scharfmann,[1] M. Marx,[2,3*] L. Fleming[1,4]

[1]Institute for Business Innovation, Haas School of Business, U.C. Berkeley
[2]SC Johnson School of Business, Cornell University
[3]National Bureau of Economic Research
[4]College of Engineering, U.C. Berkeley
[*]Corresponding author. Email: mmarx@cornell.edu

The PDF file includes:

- Materials and Methods
- Supplementary Text
- Figs. S1 to S25
- Tables S1 to S13
- References

Other Supplementary Materials for this manuscript include the following:

- Data S1 and S2

Materials and Methods Overview of data linkage processes

In order to provide a public database that identifies individuals that both publish science and patent inventions, we linked individual identifiers in a database of disambiguated scientists (OpenAlex, hereafter "OA") with individual identifiers in a database of disambiguated inventors (PatentsView, hereafter "PV"). We refer to these individuals as "Pasteur Quadrant Researchers" or PQRs. While the OA database covers most of the science papers published worldwide since 1800, the PV database relies upon the United States Patent and Trademark Office database, which only covers inventions registered in the United States (though half are from outside the U.S., thus providing some international coverage of the typically more important inventions) since 1976. We rely upon the underlying disambiguations of OA scientists and PV inventors and while we note some of the errors in these underlying databases in passing, we do not attempt to correct these errors. Note that while we provide all linkages from 1976 through June of 2023, we usually truncate the data for analysis purposes, due to patent pendency lags (the time from patent application to patent grant, typically one to four years, but sometimes lasting longer). See the Data section below for data, code, repositories, and additional documentation.

Figure S1 provides a micro-illustration of an example illustrating all PQRs, pure scientists, and pure inventors with the relatively rare last name string of "Doudna." Considering all scientists and inventors whose last name includes the string "Doudna" highlights 18 scientists who published and seven inventors who patented. As indicated by the middle purple bar, Jennifer Doudna (of CRISPR Nobel Prize fame) and James Cate Doudna are the only individuals with a last name of "Doudna" who both published (613 papers for Doudna and 336 for Cate) and invented (216 and 22 patents, respectively).

Figure S2 provides a conceptual diagram of the processes that accomplish Figure S1. At

a high level, the process 1) "blocks" possible name matches for further examination, and for the crosswalk within each block—i.e., for every possible linked pair of OA and PV identifiers within the block, and for every pair of papers and patents within each linked pair—2) quantifies the name similarity, 3) trains and runs a classifier model that predicts whether a paper and patent are written by the same person, and 4) provides an overall confidence estimate for the proportion of correct paper-patent pairs that are within the OA-PV linked pair of identifiers.

The sections below describe in more detail the following steps:

- Section "Name cleaning and blocking": last names are first cleaned and then blocked together, based on possibly correct or *not-inconsistent* names.
- Section "Quantifying the similarity and frequency of the names": the similarity between the scientist and the inventor names is quantified based on the similarity of the name strings and frequency of the names (i.e., whether the name is common or not).
- Section "Quantifying the similarity of the documents": for all papers and patents that might be invented by the same individual (ie., for each paper-patent pair implied by the OA and PV identifier pair), we quantify their similarity. We estimate the similarity between a patent and a paper by training a random forest which relies on the documents' features (title and abstract textual similarity, geographical distances, number of similarly named co-authors, institution and assignee similarities, temporal overlaps) to return a probability that a particular paper and particular patent are both written by the same individual. Note the classifier remains independent of the name similarity.
- Section "Estimation of scientist-inventor link confidence score": lastly, we estimate the overall accuracy of each PQR (defined as an OA and PV ID pair), based on the names' and

documents' similarity. This is calculated as the proportion of paper-patent pairs within each linked pair that we assess to be correct.

Materials and Methods Description of input and output databases

For the science paper side of our linkage, we use OpenAlex (OA), a publicly-available dataset of scientific articles which provides most of the papers from 1800 to recent years (our linkage considers the papers published until June 2023). OA splits the data into multiple tables, which we uploaded and then queried using PostgreSQL. As described in the database schema Figure S3, OA provides data on scientific papers and authors, including the paper's dates, title, abstract, institutions and geographic locations. It hosts more than 275M documents and 90M disambiguated authors. OA provides the abstracts of approximately half of the documents (some are not scientific publications and lack abstracts - abstract coverage is also more likely to be missing for older works). The works are classified by OA into 19 fields of study ("concepts") based on their textual content (please see https://docs.openalex.org/api-entities/concepts). Note that OA refers to individuals as authors, while we will refer to them as scientists, in order to clearly differentiate them from inventors (who are also "authors" of patents).

Documentation for OA data can be found at https://docs.openalex.org/api-entities/entities-overview. The data used in our linkage corresponds to OA data release in August 2023 (please see OA release notes: https://github.com/ourresearch/openalex-guts/blob/main/files-for-datadumps/standard-format/RELEASE_NOTES.txt). Note that we use MAG (Microsoft Graph Academic) dates instead of OA as OA sometimes provides online publication dates (and even duplicates of the same paper, published at different times) instead of eventual issue publication dates.

For the patent invention side of our linkage, we use PatentsView (PV), a US government

database which provides all US patents from 1976 to recent years (our linkage considers the patents granted until June 2023). PV splits the data into multiple tables, which we uploaded and then queried using PostgreSQL. As described in the database schema Figure S3, PV provides data on patents and inventors, including the patent's dates, title, abstract, assignees (owners of the patents, typically firms) and geographic locations. It hosts almost 8.5M patents and 3.9M disambiguated inventors. PV provides the abstracts of approximately 91% of these (design and plant patents usually do not have an abstract and while included in the database by the transitive association of the disambiguations, are not predicted in the random forest model). The patents are classified by the hierarchical CPC categorization as defined by the World Patent Organization (please see https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html).

Documentation for PV data can be found at https://patentsview.org/download/data-download-tables. The data used in our linkage corresponds to PV data release in June 2023 (please see PV release notes: https://patentsview.org/release-notes).

The resulting database of Pasteur Quadrant Researchers provides the career cross-walks of scientist papers and inventor patents from 1976 until recent years (until June 2023 and limited to US patents). It provides the following variables (these are explained in the pertinent section and descriptive statistics provided in Tables S1 and S2).

- inventor_id: PV inventor ID number
- author_id: OA scientist ID number (OA refers to individuals as authors, we will use scientists when possible as inventors can also "author" patents)
- inventor_name: cleaned inventor name string from PV
- scientist_name: cleaned scientist name string from OA

- name_similarity: score between 0.5 and 1 reflecting likelihood that the scientist and inventor names correspond to the same person (1 is likely)
- number_of_papers: number of published papers corresponding to given author_id
- number_of_patents: number of granted patents corresponding to given inventor_id
- confidence: score between 0 and 100% indicating the proportion of paper-patent pairs that we assess to be written by the same person
- number of comparisons: denominator of confidence score based on countable paperX-patent pairs (documents must have titles and abstracts to be countable)
- number of matches: numerator of confidence score based on correct paperXpatent pairs (documents must have titles and abstracts to be assessed for correctness)

If OA or PV are updated, it is likely that their scientist and inventor IDs will change. To help scholars map the current work to those updates, Table S12 includes the current scientist ID in OA and the papers published by that scientist (paper IDs will not change). Table S13 includes the current inventor ID in PV and the patents granted to that inventor (patent numbers will not change).

Materials and Methods Name cleaning and blocking

Names from both datasets were first cleaned and then "blocked" if they might be a potential match. The purpose of cleaning and blocking is to minimize the amount of predictive modeling that is needed, i.e., you only want to spend computational cycles on scientists and inventors who might be the same person. For example, there is no value to running a complex predictive model to determine if a scientist with the last name of Oppenheimer is the same person as an inventor with the last name of Handler. Figure S4 illustrates a subset of match candidate

pairs and whether they were ultimately retained for further processing. The rest of this section describes the blocking algorithms.

The process begins by assembling a set of match candidate pairs of scientist and inventor IDs, based purely on string similarities in the name fields from OA and PV. This step essentially passes reasonable candidates on for further consideration, or in other words, all not-inconsistent name strings (e.g., L. Fleming and Lee O. Fleming would be a match candidate pair, but Ronald or R. Fleming would not be a match with either L. Fleming and Lee O. Fleming because of the different first two names). Note that this is only the first step in a lengthy process and does not assign a probability of a correct match – this first step only collects potential match candidate pairs, which are then subjected to further processing. The main steps of the process are as follows:

- The last name of an inventor is selected.
- All inventor IDs with a last name which contains the same separated last name string are collected and cleaned, e.g. "Doudna" and "Doudna Cate."
- All scientist IDs with a name containing exactly the same separated last name string are collected and cleaned, e.g. for the last name "Doudna", "Jennifer Doudna", "J. Doudna Cate. Note that the difference between the scientists and the inventors arises because OA and PV do not provide names with the same format (full name for OA, first name and last name for PV).
- For each pair of names drawn from these two different populations, the algorithm decides whether they should be considered further as a match candidate (based purely on not-inconsistent name similarities as described above).

The name string formats in OA and PV are unfortunately different; OA does not differentiate between first and last names while PV breaks names into a last name and suffix and a first and

middle name. Also note that name strings are queries with the "f_unaccent" and "ILIKE" clause in PostgreSQL which do not differentiate the upper and lower cases and the accent. This method allows us to query all the variations of a given last name (for example, if we want to select all the "Leon", this method will return all the "Leon", but also the "leon", "Léon"/"léon" and "León"/"león").

PV usually breaks names into a first and middle string and a last and prefix/suffix name string. This makes it easy to select a candidate by last name and then select all other patents with the exact same last name (we assume these fields are error free). In contrast, OA provides names in three different formats (the algorithm ignores the third format, due to the difficulty of estimating the correct string and the high likelihood of introducing errors):

1. First name (or initial or none) / middle names (or initials or none) / Last name(s), e.g., Jennifer A Doudna or Andrew H. Doudna Cate.

2. Last name(s) / first name and middle names OR first and middle names initials merged, e.g., Doudna JA, Doudna J A, Cate Doudna Jamie.

3. Others (misspelling errors, multiple authors merged as one, words that are not names etc.).

After the names are queried from the two databases, names from the first two OA formats and from PV still require additional cleaning before string matching. Capital letters were converted to small letters, a space was added between capital letters, and the following were removed:

- special characters (accent marks, non Latin letters, punctuation etc.)

- double spaces and trailing spaces (spaces at the beginning or the end of the string)

- capital letter strings: if the word is in capitals and has more than 3 letters it is assumed to be a name and simply put into lower case (e.g., "JEAN PAUL" becomes "jean paul"). If the word is less than or equal to three letters then the letters are split and a space inserted (e.g., "JP" becomes "j p").

Nobiliary particles were also merged (e.g., louis le grand becomes louis legrand). Note that names written with non Latin letters can not be cleaned properly with our algorithm and are dropped from consideration.

With these corrections in place, the algorithm starts by selecting all inventors with a last name containing a separated instance of the last name string. It then begins the scientist collection algorithm. It first selects scientists whose name strings contain a separated instance of the last name string (e.g., a separate instance of "scharfmann" is contained in "scharfmann raphael" but not "rapha alscharfmann"). It then considers first names and attempts to pass any not-inconsistent (ie., possible matches) possibilities (e.g., L. Fleming and Lee Fleming) as well as phonetically similar names (e.g., Carl and Karl). If no first name is found in the scientist name string, it is kept for further consideration (i.e. it is not-inconsistent).

The not-inconsistent (ie., possible matches) algorithm proceeds as follows. Consider, for example, Lee O Fleming and Lee R Fleming. This pair should not be kept for further consideration even though Lee is a name in common. However, Lee O Fleming and Lee Fleming are kept for further consideration. Considering only the first and middle names, the algorithm counts the number of names and the pair is kept if the number of names in common is equal to the minimum number of names, e.g., for Lee O and Lee R, the minimum number of names is 2, but the number of common names is 1, so it is not retained, because we need to have two names that match. For Lee O and Lee, the minimum number of names is 1 and hence we only

need to have one first name that is common (and we do, so the pair is kept). To summarize, the algorithm:

- Counts the minimum number of first and middle names in either the inventor or scientist name strings.
- For each scientist's first or middle name, compare it with the inventor's first or middle names.
- If the scientist's first name is an initial, and it is not-inconsistent (ie., possible matches) with an inventor's first name, keep the pair (L. And Lee are not-inconsistent, L. and Emma are inconsistent).
- If the scientist's first name is not an initial, and it matches the inventor's first name or if they sound the same (the same metaphone, e.g., Lee and Emma does not sound the same, Steven and Stephen does) or if they have a ratio of in common letters greater than 85%, keep the pair.
- Compare the number of matches and the minimum number of first and middle names between the inventor and scientist names. If it's equal, the names are considered as potentially similar. Otherwise they are dropped from further consideration.

Materials and Methods Quantifying the similarity and frequency of the names

The name of an individual is probably the most important characteristic we can use to match scientists and inventors. However, because the published names of an individual can vary greatly, be misspelled, or contain random errors, we need to determine whether similar names are actually the same person. This (intelligent) guess is more likely to be wrong if the name is common, hence, we would also like to incorporate information on the rarity of the name

into that guess. Hence the purpose of this section is three-fold: 1) to describe the algorithm that determines the lexical similarity of two names, and 2) to explain how we estimated a name's frequency from all PV names, 3) to explain how we integrated these two measures into a single name similarity metric ("similarity_score") which is used throughout the paper.

Following name cleaning and blocking and pairs of possible matches, the approach first manually assigns reasonable though subjective scores to more or less likely matches (no model is trained). Then, these scores are weighted based on the frequency of the names (a common name should intuitively increase our skepticism that two names represent the same person). Figure S5 provides a flowchart of the process and illustrates it with an example.

More specifically, for each candidate pair (OA scientist ID and PV inventor ID which may potentially represent the same individual), the process starts by defining four variables which enter into the calculation of the final "similarity_score."

- "name_score": a manually assigned subjective score to more or less likely matches (1 corresponds to a close match and 0 to a non-match). There are three cases. First, if neither database has a first name, a score of 0.4 gets assigned. Second, if both databases have entire first names, a monotonically decreasing score is assigned, ranging from 1 to 0.7, depending on the whether the strings match exactly or poorly, respectively (keep in mind from the prior step that only possible, or not-inconsistent (ie., possible matches) names, are considered in this step, in other words, obviously wrong possibilities, such as John and Mary, are not in this population). Third, if one database has just initials, a score of 0.8 is assigned if the first and middle name initials match exactly, 0.6 if only the first initials match, and 0.1 if the initial is found somewhere in the name string.

- "distribution_last_name": a score between 0 and 1 which quantifies the rarity of a last name. Low scores correspond to highly uncommon last name (for example, "Zywotko" or "scharfmann") and high scores correspond to very common last names (for example, "Kim", "Lee" or "Smith"). Last names that are not in our frequency distribution are assumed uncommon and associated with 0.

- "distribution_first_name": a score between 0 and 1 which quantifies the rarity of a first name. Low scores correspond to highly uncommon first name (for example, "Yoneno" or "Winold") and high scores correspond to very common first names (for example, "Michael", "David" or "John"). First names that are not sampled are assumed uncommon and associated with 0.

- "distribution_full_name": a score between 0 and 1 which quantifies the rarity of a full name. The metric simply multiplies the rarity of first and last names: low scores correspond to highly uncommon full names. Using the scores representing the frequency of the first and last names, we use the following formula to estimate the rarity of a full name:

$$distribution_full_name = distribution_last_name \times distribution_first_name$$

  Note that when the first name is missing or only the initial of the first name is provided, we consider that the frequency of the first name is equal to 1 (score assigned to the most frequent first names). A high score corresponds to a very common name and a low score to a rare name.

We define the a final "similarity_score" by weighting the intuitively assigned score "name_score" by the frequency of the name "distribution_full_name":

$$similarity_score = \frac{name_score}{1 + distribution_full_name}$$

If both first and last names are very rare, and multiply together to be zero, the "similarity_score" is simply the name score. In the opposite case, when the first and last names multiple together to be one, the "similarity_score" is halved. The metric is not trained on a model but it provides a rough estimate of how likely it is that the scientist (from OA) and the inventor (from PV) represent the same individual based on their names only.

Materials and Methods Quantifying the similarity of the documents

1. Creation of the training dataset

   We train and deploy a machine learning model which predicts if a patent and a paper are written by the same individual. The process begins by identifying and expanding a training set (i.e., a set of papers and patents known to be written by the same individual and a set of papers and patents known to not be written by the same individual). We then create features for these positive and negative training examples including science field and classification category similarity, similar co-authors, geographic distance, and overlapping citations to papers and patents. These data are used to train the ML model; model performance is evaluated with a standard 20% data holdout.

   The first step is to collect a sample of scientific publications and patents written by the same individual. Marx and Fuegi [2] parse and match both front page and inline citations from patents to science publications. They find 30 million instances which are reasonably certain (we only use datapoints whose confscore metric equals 10 out of a possible 10). From this set of very confident patent to paper citations, we further restrict the sample to only those that are potentially self-citations, using the Marx and Fuegi [2] metric "self-citeconf_max". This metric does not guarantee that the citations are self citations, only that they are very likely self citations. (That is, the additional presence of an overlapping

name within a patent citation to a paper is less likely to be a false positive.) To further increase the confidence of a self citation, we keep only the patent to paper citations which satisfy all of the following:

- two or more names of co-authoring scientists and co-inventors match, according to our name comparison algorithm.
- two or more matching names have a similarity score greater than 0.75 (which takes into account the similarity of the names, and whether they are common or rare, see appendix section above, "Name comparison algorithm incorporates similarity and frequency").
- at least one scientist in common has an ORCID website ID, and the ORCID website for that ID lists both the paper and the patent.

In applying this algorithm, we found 47,000 patent paper pairs in ORCID where the patent cites at least one paper corresponding to an ORCID ID. To expand the intellectual breadth of the training dataset, beyond the typically narrow correspondence between the original paper and patent pair, up to 5 papers authored by the individual with the matched ORCID ID are randomly selected. The 218,841 pairs of citing patent paper pairs corresponding to an ORCID ID are considered as being written by the same individual and constitute our positive training set.

After creating a set of papers and patents written by the same individual, we built the negative training set (papers and patents not written by the same person) by starting with the same papers and then randomly selecting patents with no name in common. We select a total of 1,000 random patents following the distribution of the number of patents

published each year. This creates a negative training set of 218,841 pairs.

2. Variable/feature definition for the random forest model

To train the machine learning model, we need features which enable the model to learn to recognize and differentiate the positive instances (papers and patents written by the same individual) from the negative instances (papers and patents which are not written by the same individual). These features need to be relevant and able to discriminate between the positive and negative linkages.

We create these features from comparisons of a number of fields in the OA and PV databases. These fields vary in their analogous similarity and contributions to the model's ability to discriminate. Figure S6 describes these fields and their relationships and shows the correlation between each feature and the data classification (written by the same individual or not). Not surprisingly, the correlation matrix shows that the greater the similarity of the titles, abstracts, institution names and citations, the more likely it is that the paper and patent are written by the same person. As expected, we have the opposite result for the distance features: the lower the distance between the scientist and the inventor or between the scientist and the assignee, the more likely it is that the paper and patent are written by the same person.

Model feature: co-scientist overlap between paper and patent

The model counts the number of co-scientist and co-inventor names which are potentially similar, using the same algorithm described above for comparing PQR name pairs. Note that only the co-scientists and the co-inventors of the specific paper and patent are con-

sidered. Consider a scientist named Lee Fleming and an inventor named Lee O Fleming. Lee Fleming and Lee O Fleming is a candidate pair (according to the process described above for comparing PQR name pairs). If the scientist has a paper written by Emma Scharfmann, Matt Marx and Lee Fleming, and the inventor has a patent written by Lee Fleming and Matt Marx, the number of co-scientist overlaps is 1 (Matt Marx). Lee Fleming is not counted as a co-scientist overlap as he is the focal/potential PQR.

Model feature: similarity from titles

The patent and paper titles are compared with the NLP (Natural language processing) tool BERT (Bidirectional Encoder Representations from Transformers), specifically, the sentence_transformers with "all-MiniLM-L6-v2" pretrained model. Each title is encoded ("text embedding") with 384 numbers between -1 and 1, which represent the meaning of the texts. A simple dot product between two encoded texts (two vectors of 384 numbers) gives us a similarity score between -1 and 1, which becomes a feature for the prediction model. Note that the papers and patents with a missing title are not taken into consideration by our algorithm, but will be included in the transitive linkage if their scientist and inventor ids are linked together.

Model feature:similarity from abstracts

OA provides inverted indexed abstracts and PV provides directly accessible abstracts. Following reconstruction of the OA abstracts, the patent and paper abstracts are compared with the NLP (Natural language processing) tool BERT (Bidirectional Encoder Representations from Transformers), specifically, the sentence_transformers with "all-MiniLM-L6-v2" pretrained model. Each abstract is encoded ("text embedding") with

384 numbers between -1 and 1, which represent the meaning of the texts. A simple dot product between two encoded texts (two vectors of 384 numbers) gives us a similarity score between -1 and 1, which is then used to predict a probability that the paper and patent have a common scientist. Note that the papers and patents with a missing abstract are not taken into consideration by the algorithm, but will be included in the transitive linkage if their scientist and inventor ids are linked together.

Model feature: geographical distance between inventor home town and scientist institution

The model uses the best available geographical data for the patent and paper and includes the distance between the inventor's hometown and the science institution or organization indicated in the paper as a feature. U.S. patents typically list each inventor's home town (though sometimes the assignee location is mistakenly used). In contrast, papers typically only list an scientist's institution. We rely on the PV and OA geographic disambiguation to get the latitudes and longitudes corresponding to the paper's institutions and the patent's inventors.

Note that PV provides the co-inventor's home town and provides the coordinates of the inventor's home town. Note also that OA provides the co-scientists of the scientific publications and quite often also provides the institution's name indicated on the scientific paper that corresponds to the scientist. OA disambiguates a large number of the available institutions' names and directly provides the geographic coordinates corresponding to these institutions. For the institutions which have not been disambiguated, only the institution name is provided.

When searching for the institutions corresponding to a scientist of a given scientific publication, three cases exist. Figure S7 illustrates these three cases and their respective conclusions.

- The institution has been disambiguated and OA provides the geographic coordinates corresponding to the institution. In this case, we rely on OA disambiguation and the coordinates of the institutions are used directly.
- The institution has not been disambiguated and OA only provides a text string corresponding to the institution. In this case, we clean the institution string and extract the city names from the string, using the NLP algorithm "spaCy". If a city name is extracted, the coordinates of the institution will be the coordinates of the city. Otherwise, we move to the next case.
- OA does not provide an institution corresponding to the scientist. In this case, we assign the institution corresponding to the scientist of the paper as the "last_known_institution" of the scientist previously provided by OA. Again, if the last institution of the scientist has not been disambiguated, we clean the institution string and extract the city names from the string using the NLP algorithm "spaCy", and use the coordinates of the extracted city as the coordinates of the institution if a city name is extracted.

Model feature: geographical distance between assignee and scientist institution

The model uses the best available geographical data for the patent and paper and includes the distance between the assignee's location and the institution indicated in the paper. U.S. patents typically list each assignee name and the assignee town. In contrast, papers typically only list a scientist's institution. We rely on the PV and OA geographic disambiguation to provide the latitudes and longitudes corresponding to the papers' institutions

and the patents' assignees.

Note that PV provides the assignees' organization names. PV disambiguates the assignees' organization names and provides the geographic coordinates corresponding to the assignee. We rely on PV's disambiguation of the assignee's location. Note also that we process missing data in the same way as described above for the recovery of OA institutional coordinates.

Model feature: comparison of scientist institution and assignee organization

The model compares the scientist institution's name if it exists (or the last known institution if not) with the assignee organization name. The following method is applied to the institution string and assignee organization string:

- the string is cleaned
- the stop words from English, Spanish, French and German are removed from the string
- each word of the string is stemmed
- the following stemmed words are removed from the string: colleg, hospit, medic, corpor, institut, center, depart, school, servic, medicin, scienc, univers, state, research, health, foundat

Once the institution and the assignee organization are cleaned and processed, the model counts the number of common words between the two strings.

Model feature: difference between mean of science and single patent publication dates

As the direct difference between a patent granted date and a paper publication date did not accurately predict whether a paper and patent were written by the same individual, we developed a feature that compared the patent grant date with the mean of the scientist's publication dates. As most scientists publish multiple papers in their career, we calculated the mean of the publication dates of the scientist's papers (in the case the scientist only published one paper, we took that date). Then, we calculated the absolute difference between the mean publication date of the scientist ID and the granted date of the focal patent. This absolute difference is used as an attribute in the prediction model. Figure S8 illustrates the distributions of the date difference for positive and negative outcomes.

Model feature: temporal correspondence with the distribution of the publication dates

As the direct difference between a patent grant date and a paper publication date does not discriminate between the papers and patents written by the same individual, we develop a feature which compares a patent's application date against the temporal distribution of a scientist's papers.

A scientist usually has multiple papers published in their career and the publication dates of the papers roughly describe the stage of the career. To model the scientist's career, we fit a distribution on the list of publication dates. This parameterization of the scientist's career allows an estimation of the probability the scientist published a patent at a given date. In more technical terms, we first fit a distribution to the list of publication dates and then calculate the likelihood of the patent grant date from the distribution. We use Maximum Likelihood Estimation (MLE) to fit the distribution and use the log likelihood as an attribute in the prediction model.

It remains to choose a distribution to fit the publication dates that represent the scientist's career. To choose the most appropriate distribution, we ran experiments on several classical distributions including Normal, Skew Normal, Chi Squared, F, and Gamma distribution, as well as shifted distributions, to model the elapsed time between the application date and the grant date. Among these, the Chi Squared appears most highly correlated with a paper and patent by the same scientist. Therefore, we conclude that Chi Square is the best distribution to model the career of a scientist, who usually has more publications in their early career and a long tail towards the end of their career.

Model feature: number of cited papers in common

The model counts the number of science papers that are cited by both the paper and the patent. The science papers cited by the paper are provided by OA. The science papers cited by the patent are provided by the Marx and Fuegi [2] dataset.

3. Random forest model

We estimated three common supervised learning models: Logistic regression, Decision tree, and Random forests. Logistic regressions are easy to interpret but do not fundamentally include interactions, Decision tree models are most easily interpreted but are prone to high variance, and Random forest models have higher accuracy but are more difficult to interpret. Results were consistent across models and we report the Random forest model because it was the most accurate. For both models, we train the model on a random sample of 80% of our training set and test it on the remaining 20%. Figure S9 shows the distribution of the predictions for the testing set: the U-shape illustrates how

well the models discriminate between papers and patents written by the same individuals.

OA and PV lack latitude and longitude data around 16% of the time. This makes it impossible to calculate the distance features. Given that machine learning models can't handle missing values, we train two distinct models:

- one model which includes the distance features, trained on the 84% of the training set for which the distance features are not missing.
- one model which does not include the distance features, trained on the entire training set.

Figure S10 (upper figure) compares results and illustrates how the Random forest classifier with the distance features returns the highest accuracy, with an overall accuracy of 96.2%, a true positive rate of 95.1%, a false positive rate of 2.8% and a AUC of 0.99. Figure S10 (lower figure) illustrates the strength of each feature.

Figure S11 (upper figure) compares results and illustrates how the Random forest classifier without the distance features returns the highest accuracy, with an overall accuracy of 93.9%, a true positive rate of 92.2%, a false positive rate of 4.6% and an AUC of 0.98. Figure S11 (lower figure) illustrates the strength of each feature.

Materials and Methods Estimation of scientist-inventor link confidence score

Recall that the initial blocking implied a crosswalk of scientists and inventors that might be the same person, based on not-inconsistent (ie., possibly correct) names. Every pair of OA and PV individual identifiers in turn implies a set of papers and patents, based on the disambiguations within OA and PV. Here we assess the likelihood that a pair of OA and PV identifiers

is correct, based on a prediction of the number of correct paper-patent pairs, divided by the number of possible paper-patent pairs. This assessment combines the similarity in names and prediction model. Using these metrics, the user can work with only the most accurate OA PV individual identifier links, all the links, or anywhere in between. Note that if a linkage pair earns no correct assessments (i.e., 0 matches divided by the number of possible paper-patent matches, for a score of 0), it is dropped from the dataset and not reported.

We first compare OA scientist names with the PV inventor names and assign the standard similarity score described above. Then, for each Random Forest link, all the OA papers and PV patents written by the PQR are compared. Two scores for each pair of paper-patents are calculated:

1. The standard "similarity_score", which gives the likelihood that the scientist name and the inventor name correspond to the same individual (described at length in an above appendix).

2. A "prediction_score", which indicates the likelihood that the paper and the patent are written by the same individual, based on other features besides the name.

The similarity metric provides much higher discrimination than the prediction metric. For example, if the "similarity_score" is high, the scientist and inventor name are very similar and rare and it is very likely that the scientist and inventor are the same individual. If low, the scientist and inventor name are not that similar and common, and the probability that the scientist and inventor are the same individual remains low.

If the "prediction_score" is high, the paper and patent are similar in content and close geographically. As many similar papers and patents are written within the same institutions, it is

unfortunately not enough to conclude that the paper and patent have been written by the same individual. If the "prediction_score" is low, the paper and patent are different in content and far apart in geographic distance. As scientists and inventors might be published in different fields and countries, again it is not enough to conclude that the paper and patent have not been written by the same individual.

Accordingly, we set the following cutoffs for the paper and patent being written by the same individual:

$$similarity_score^2 + prediction_score > 1.25$$

The squaring function incorporates the idea that the "similarity_score" discriminates more effectively. We chose the values 1.25 by considering the following limiting cases:

1. If the "similarity_score" is perfect (equal to 1), the "prediction_score" needs to be greater than 0.25.

2. If the "prediction_score" is perfect (equal to 1), the "similarity_score" needs to be greater than 0.5.

All the scientist's papers (from OA) and inventor's patents (from PV) implied by the Random Forest link are then assessed with this metric. Intuitively, if more papers and patents are correct, it should increase our confidence that the scientist and the inventor are the same individual. Hence we define a confidence score between 0 and 100 representing the percentage of patents and papers predicted as being written by the same individual by our model. In other words, the number predicted as correct, divided by the product of the number of papers and patents. Based on this confidence metric we create three categories:

- Low confidence: percentage of pairwise paper-patent combinations assessed as correct between 20% and 100%

- Middle confidence: percentage of pairwise paper-patent combinations assessed as correct between 30% and 100%
- High confidence: percentage of pairwise paper-patent combinations assessed as correct between 50% and 100%

Note that these categories are not mutually exclusive, i.e., Middle includes High, and Low includes both Middle and High. We manually evaluate the accuracy of the results obtained using these individual and aggregate thresholds in the Appendix "Accuracy assessment". Note that documents without titles and abstracts cannot be counted and are not included in this metric.

Materials and Methods Accuracy assessment

We estimate the accuracy of our linkage with Google Scholar profiles, combining manual and automated approaches. Note that all of these estimates contain unavoidable errors themselves and remain flawed, mainly due to the difficulty of establishing the exact patenting and publication record of a large and random sample of PQRs (Google Scholar and ORCID contain additional documents, duplications, and inaccurate disambiguations, and hence provide imperfect comparisons).

Some individuals present particular challenges (for example, Matt Marx patented in speech recognition in industry and is now publishing in academia in social sciences, in different organizations, locations, and time periods). Different combinations of names and first letter abbreviations exacerbate these difficulties (e.g., Matt Marx and Matthew T. Marx). Finally, common names such as Smith simply provide little purchase for accurate disambiguation and linkage.

Google Scholar is not 100% accurate and sometimes includes disambiguation errors (e.g., Stuart Russell is lumped and Alberto Sangiovanni Vincentelli is split). On the other side, OA

and PV disambiguation errors introduce errors into the down stream processes. Scholar profiles may be outdated (e.g., E. Sylvester Vizi) and sometimes do not include patents. Moreover, OA and Google Scholar are inconsistent in their listing of publications besides scientific articles (for example, they can include comments, teaching cases, discussions, revisions, datasets, books, non peer-reviewed work, etc.). They also sometimes duplicate works and list erroneous paper's titles.

Finally, we designed our accuracy estimates to require strict matching with exactly the same titles, which causes any misspelling error in the paper or patent titles to prevent a match, and thus makes our approach very conservative.

Keeping these issues in mind, we begin by randomly selecting 1,000 PQRs (we stratified sampling in order to cover a wide range of confidence scores) and different similarity scores (again randomly sampled within strata, using the name similarity measure defined above). Scientists without a Google Scholar profile were dropped from consideration. At the other extreme, some names have many more than one profile. In the case that more than five profiles correspond to a PQR name (71 Google Scholar profiles correspond to the name "John Smith"), we do not consider the name. When there are less than six Google Scholar profiles, we manually identify the correct profile. Following this we were able to find a matching Google Scholar profile for 256 PQRs which corresponded to 390 random pairs of PQR OA scientist and PV inventor IDs (the greater number of OA and PV IDs occur because of disambiguation splitting in those databases).

Once the PQRs are linked with at least one Google Scholar profile, we compare the papers and patents listed in the databases. If no patents are listed in Google Scholar, we manually

search for the names of the PQR's co-inventors (as listed in PV) in Google Scholar. If we find the patent but also find that the patent was not formatted as a patent, we code it as correct. Also, if we find co-authors of the patent in papers listed in Google Scholar, we also code the patent as correct. Each manual coding is reported and provided in the manual checking file (which is on GitHub as listed below).

The accuracy estimation process requires an exact textual match between OA and Google Scholar. Once the upper case letters, the punctuation and the trailing spaces are removed from the titles, two papers (or patents) are only considered as being the same if the titles match exactly (and for patents only, if co-inventors are listed as co-authors in Google Scholar, as described immediately above).

For each PQR, we report two metrics in the Github repository:

- the percentage of the PQR patents from PV listed under the Google Scholar profile.
- the percentage of the PQR papers from OA listed under the Google Scholar profile.

We calculated two metrics for a scientist and an inventor being the same individual, 1) if at least one of the scientist's papers from OA and at least one of the inventor's patents from PV are listed under the same Google Scholar profile, and 2) if at least 30% of the scientist's papers from OA and at least 30% of the inventor's patents from PV are listed under the same Google Scholar profile. We do not use 100% because of the errors described previously. These definitions return accuracies of 85.4% and 78.1% for our low confidence data ($conf \geq 20\%$), 86.8% and 79.2% for our mid confidence data ($conf \geq 30\%$) and 91.2% and 83.8% for our high confidence data ($conf \geq 50\%$).

We reinforced the automated comparison against Google Scholar by also conducting a manual comparison. This exercise focused on 10 known PQRs from UC Berkeley, which corresponded to 15 pairs of OA and PV IDs (two OA identifiers were found for Costas P. Grigoropoulos, three for Kristofer Pister and three for Stuart Russell), with names ranging from common to uncommon. First, we identified the Google Scholar profiles of the 10 faculty members. Subsequently, we reported the proportions of papers from OA and patents from PV found in Google Scholar, as well as the proportion of patents and papers from Google Scholar found in OA and PV (data provided in Data S1). Out of the 15 PQR pairs, 11 PQRs were confirmed as correct linkages.

All four pairs of OA and PV IDs identified as linkage errors by the manual validation had a confidence score lower than 20 (below the lower bound of our low confidence dataset). Our linkage method was also able to identify and rectify one splitting error from OpenAlex (two OA identifiers were found for Costas P. Grigoropoulos and for Kristofer Pister, which appeared correct upon manual validation against Google Scholar). These results support the choice of the confidence bounds and reinforce the concern that researchers should exercise caution in using the low confidence PQRs.

To estimate how often scientists and inventors are not being linked when they should have been (i.e., false negatives), we select a random sample of patents that cite papers and identify 12,000 pairs of scientists and inventors having similar names, using the patent-to-paper citations marked as *self*-citations in the Marx and Fuegi [2] dataset. It is unlikely that a scientist from a paper cited by a patent has the same name as one of the patent's inventors without being the same individual. For these 12,000 pairs, the low confidence dataset ($conf \geq 20\%$) includes 84.6% of the pairs, the middle confidence dataset ($conf \geq 30\%$) includes 82.7%, and the high

confidence dataset $conf \geq 50\%$) includes 77.3%. Given these estimates, researchers should consider all the data accuracy cutoffs when the cost of a false negative is high.

Supplementary Text PQR field-inventor CPC calculation for the scientist/inventor diagram

The Sankey diagram (Figure S1) illustrates the most common combinations of fields of science and classifications of technology that PQRs work in. For each of the inventors listed in the PQR dataset, the PV inventor's eventually granted patents as well as their application dates and CPC classes (the 9 highest level Cooperative Patent Classification) are queried. Similarly, for each of the scientists from the PQR dataset, the OA scientist's papers as well as their publication dates and their high level field (19 OA level 0 concepts) are queried. Note that a patent can have multiple CPC classes and a paper can have multiple fields (we select the most frequent field or class for display in the Sankey diagram).

The science field to technology class Sankey diagram in the main text considers only the papers and the patents from concurrent years (a year during which a PQR both publishes a paper and applies for a patent). For each PQR, we collect every field (level 0 concepts) of all the papers published during every active year. Similarly, we collect every CPC class of all the patents the PQR applied for during every active year. Note that in case of tie in most common field or class, we randomly select one of the most frequent fields or CPCs. In summary, each PQR is associated with a unique field and a unique CPC, which represents the most frequent field the PQR is publishing in and the most frequent CPC of the PQR's patents.

Supplementary Text Details of trends and descriptions of PQRs

To understand where PQRs publish and patent, we developed a taxonomy of four main types of possible relationships, based on the OA reliance on the Research Organization Registry (https://ror.readme.io/docs/basics) and Eurostat disambiguation [1].

- PQR publishes and patents in a company
- PQR publishes in a university and patents in a company
- PQR publishes and patents in a university
- PQR publishes in a company and patents in a university

Figure S13 illustrates the number of PQRs in each category, from 1980 to 2016. As can be seen, very few PQRs publish in a company and patent in a university. More publish in a university and patent in a company, but most common is doing both within the same type of entity, that is, publishing and patenting in either a company or a university. This makes sense, as corporations have always claimed their employees' patents, and universities have begun to do so since the Bayh-Dole Act in 1980.

Figure S14 illustrates a similar graph to S13, but illustrates how the aggregate graph masks differences by fields. For biology, the most common type is publishing and patenting within a university, while for computer science, the most common type is publishing and patenting within a firm.

We also compare the appearance over time of the number of new scientists, inventors and PQRs. For each scientist disambiguated by OA, we sum the number of first time scientists in the year they first appear in the OA database. For each inventor disambiguated by PV, we sum the number of first time inventors in the year they first appear in the PV database. For each PQR identified in our dataset, we sum the number of first time PQRs in the year they first appear in our linked database. We define the first year as the first concurrent year or the first year an allochronic PQR is observed publishing the type of document s/he hasn't yet published,

whichever is earliest. Figure S15 describes the evolution of the number of new inventors, new scientists and new PQRs from 1980 to 2016. Note the truncation bias due to patent pendency at the end of the time series.

Figure S16 illustrates the issue of patent pendency - the time between patent application and patent grant - influences the time series statistics. We chose to match patent application date with paper publication date, as this provides the best (though still approximate) temporal overlap. If we had used grant date, then most of our patents would be appearing one to two years before the science was published. To maintain consistency of results, we truncated most graphs due to patent pendency.

As can be seen, the first noticeable drop off in Figure S16 occurs in 2018-2019, and corresponds to a 13% drop. The percentage drop off in each year of the graph is as follows:

- between 2017 and 2018: 3.85%
- between 2018 and 2019: 13.05%
- between 2019 and 2020: 37.10%
- between 2020 and 2021: 57.90%
- between 2021 and 2022: 80.48%
- between 2022 and 2023: 98.18%

Supplementary Text Details of trends and descriptions of PQR gender over time

We provide the evolution over time of the PQRs by gender. We rely on PV classification of the inventors into three gender categories: female, male and unknown when the gender of the

inventors could not be assigned based on the name.

Figure S17 describes the evolution of the proportion of female PQRs and female inventors over time, as estimated by the PV gender variable. The shaded areas correspond to the uncertainty on the proportion of female PQRs and inventors due to the unknown category.

Figure S18 describes the evolution of the gender of PQRs by field of study. We use the most frequent level 0 concept provided by OA of all papers published by the scientist to classify the PQRs into field categories. Therefore, each PQR corresponds to at most one level 0 concept.

Note also that recent statistics are not reported due to the patent pendency issues described above. Also note that the PV gender identification algorithm is less accurate for non-U.S. (typically European) names, and since non-U.S. (and non-European) patents have become a greater proportion of U.S. patents over time, the number of unspecified data points has increased over time.

Supplementary Text Details of PQR geography within the U.S.

To study the locations and change in locations of the PQRs, we first associate the PQR institutions with a US county using the geographic coordinates provided by OA. Similarly, we link the PQR assignees with US counties using the coordinates provided by PV. For each PQR concurrent year (year during which the PQR applies for a patent and publishes a paper), we identify the most frequent county the PQR publishes or patents in.

Note that each PQR is associated with at most one county per year (the PQRs who only patent or publish outside of the US or the PQRs who do not publish a paper and apply for a

patent in the same year are available in the database but are not part of the analyses presented here). Also note that PQRs can be counted over multiple years if they publish and patent concurrently in multiple years.

For example, for an active PQR with one paper and one patent in 2010 in Alameda county, one paper in 2011 in Alameda county, one paper in 2010 in Santa Clara and one paper and one patent in 2012 in Santa Clara county, then the PQR is counted as part of Alameda in 2010, is not counted in 2011 and is counted as part of Santa Clara in 2012.

Figure S19 illustrates the Gini coefficient of PQR concentration over time by U.S. county. It illustrates how PQRs have been consistently more concentrated that pure inventors or pure scientists, though the concentration of pure inventors has been rising more rapidly.

Figure S20 illustrates how the concentration of PQRs in the wealthiest 1% of U.S. counties rose from 1980 to 2016 (county GDP came from https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas accessed in October 2023). Note that we calculate the top 1% counties for each given year, by each variable. The particular counties can and occasionally do change from one year to another. Note that the substantive takeaway is similar to Figure 1 in the main paper, which illustrates an unchanging set of top counties, based on the average GDP over 2001-2016.

Figure S21 illustrates the exact same variables as S20 except only for Santa Clara county. As can be seen, the proportion of inventors and PQRs in Santa Clara county increases greatly over the time period, thus accounting for a large proportion of the national trends observed in Figure S20.

Figure S22 illustrates the location of PQRs in 1990 and 2016. As can be seen, and despite the stability of the overall Gini coefficient, there is a marked change in the location of PQRs from the Northeast to the West coast. Figure S23 shows that this occurs mainly in the field of computer scientists on the west coast and in particular in Santa Clara County in Silicon Valley. Figure S24 provides an example of the primary data that gave rise to the errors observed in Figure S1.

Supplementary Text Explanation of the novelty measure

We measure the novelty of patents and papers by comparing each published paper and granted patent to the papers published, or patents granted, respectively, of the five previous years.

The patent abstracts are encoded with a fine-tuned BERT model (more specifically, the sentence transformers with "all-MiniLM-L6-v2") into a vector of size 384 of values between -1 and 1, which represents the meaning of the texts. For each patent granted from 1980 to 2022, we compare the embedded patent abstract, with the embedded patent abstracts granted in the prior five years before the patent was granted. We accomplish this by taking the dot product between the vector representing the abstract of a patent and every vector representing the patent granted in the five previous years. This returns a series of similarity scores between -1 and 1, which tell us how similar the patents are to each patent granted in the five previous years. Finally, we take the oppisite of the average of these similarity scores to quantify the novelty of the patents. Note that a low score corresponds to a less novel patent (as it corresponds to a patent more similar to the patent granted the five previous years) and a high score corresponds to a more novel patent

(the patent is less similar to the patents granted before).

We use an exactly similar method to assess of the novelty of the papers. We encoded the abstracts of the papers published from 1980 to 2022 with the same BERT model and we compare those with the abstracts of all papers published in the five previous years. We obtain a series of similarity score between -1 and 1, which tell us how similar the papers are to each of the papers published in the five previous years. Taking the opposite of the average of these similarity scores gives us a measure of the novelty of the papers. Note that a low score corresponds to a less novel papers (as it corresponds to a paper more similar to the papers published the five previous years) and a high score corresponds to a more novel paper (the paper is less similar to the papers published before).

Note that both the paper and patent novelty measures are mean standardized, in order to facilitate comparison.

Supplementary Text Summary statistics for averages by field and technology class

Figure S25 breaks out the main text Figure 2 by field of science and high level CPC class. As an be seen, medical novelty and human necessities (arguably medical science) exhibit reverse trends to the typical result, that PQRs bring greater novelty and impact to all fields and technologies. Patents in electricity and physics also do not follow the hypothesized trends, at least for the descriptive averages.

Supplementary Text Summary statistics for regressions

Table S9 describe the variables for our regressions of novelty and future citations (from papers to papers and from patents to patents). Note that we use the term "concurrent" to de-

scribe PQRs who both publish and patent in the same calendar year. "Non-concurrent" or "allochronic" PQRs publish and patent at some point over their career, but not simultaneously in a particular year of the focal patent or paper or year of the PQR models. Please note that all code to generate variables and estimations can be accessed on the Github account listed below.

Our explanatory variables are the presence of at least one concurrent or allochronic PQR on a patent or paper. We also include the log of the total number scientists or inventors on the document (results are robust to the exclusion of this control variable). For the PQR individual fixed effects models, we regress the sum of citations received in a year or the average novelty of papers or patents in that year. We also include the log of the number of the year's collaborators.

Supplementary Text Explanation of regression models

We estimate Poisson count models for the citations and linear regressions for the novelty measure, using high dimension STATA routines. The STATA ppmlhdfe and reghdfe commands support fully interacted fixed effects for the year of publishing or patenting, assignee or institutional affiliation, and science and technology field indicators (for papers and patents, respectively, at the level of 139 fields of science (using the concepts of level 1 from OpenAlex) and 128 categories of technology (using the CPC classes from PatentsView). This enables control at a very fine level of time, organizations, and technology. In addition, the career regressions include a fixed effects for each PQR. The number of absorbed fixed effects is listed at the bottom of each model and can be quite large. The results are very consistent with alternate measures, for example, the number or the ln of the number of concurrent or allochronic PQRs and likewise with the ln or number of pure scientists or pure inventors.

Supplementary Text Regression model results

Table S4 details the regression results at the publication level, i.e., for papers in models 1-2

and patents in models 3-4. In each model the magnitude of the positive effect from the indicator of a concurrent PQR upon novelty and citations is significantly greater than the positive effect of the number of allochronic PQRs.

Table S5 and S6 relax the definition of concurrent PQR to 3 and 5 year windows. Results weaken but remain substantively unchanged. Table S 7 details the Fixed Effects regression results at the PQR level, i.e., for the sum of the yearly paper citations and average paper novelty in models 13 and 14. The sum of yearly patent citations and average novelty are estimated in models 15 and 16. In both cases, concurrent PQRs earn higher citations and exhibit greater novelty, relative to non-concurrent years.

In order to validate the novelty measure, Table S8 estimates the age of backward citations for papers and patents. Consistent with the novelty measure, papers and patents with concurrent PQRs cite more recent prior art and research.

Errors Anecdotal examples of sources of error

Here we describe some of the errors we came across during our development efforts. These are far from exhaustive and are only intended to alert the user to potential problems they should be aware of in the application of the data.

ORCID errors: Two papers written by different individuals, corresponding to the same ORCID. Emmanuelle Charpentier orcid1: https://orcid.org/0000-0002-0254-0778 (Max Planck Unit for the Science of Pathogens: Berlin, DE) orcid2: https://orcid.org/0000-0001-8534-8388 (Université Toulouse Jean Jaurès: Toulouse, Occitanie, FR) https://framespa.univ-tlse2.fr/accueil/pratique/annuai emmanuelle#/ Paper: A regulatory toolkit of arabinose-inducible artificial transcription fac-

tors for Gram-negative bacteria Author and institution: 1. Institut für Biologie, Humboldt-Universität zu Berlin, Philippstrasse 13, 10115 Berlin, Germany 2. Max Planck Unit for the Science of Pathogens, Charitéplatz 1, 10117 Berlin, Germany Reported ORCID: https://orcid.org/0000-0001-8534-8388

OA lumping: author id: A5015609593 author name: Hideo UCHIDA work 1: W3005650234, DOI: https://doi.org/10.1116/1.5130487, publication date: 02/11/2020 work 2: W3005650234, DOI: https://doi.org/10.1116/1.5130487, publication date: 1927

OA date errors: Work: W416310516, DOI: https://doi.org/10.2183/pjab1912.3.567 OA publication date: 07/13/2023 True publication's year: 1984, by De Gruyter

OA institution errors: Paper: W2152048324 Author: A5062923881 Institution: Data Management (Italy) I4210163794 Raw institution: Data Management Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok

OA location errors: I4210144692 Institute for Student Achievement longitude: -7402092.0

OA institution lumping: I192545095 United States Military Academy - lumping between United States Military Academy and Merck Institute Example: Paper: W1981819970, DOI: https://doi.org/10.3181/00379727-111-27849 Author: Charles A. Winter Raw institution: Merck Inst. for Therapeutic Research, West Point, Pa. OA institution: I19254509 - United States Military Academy

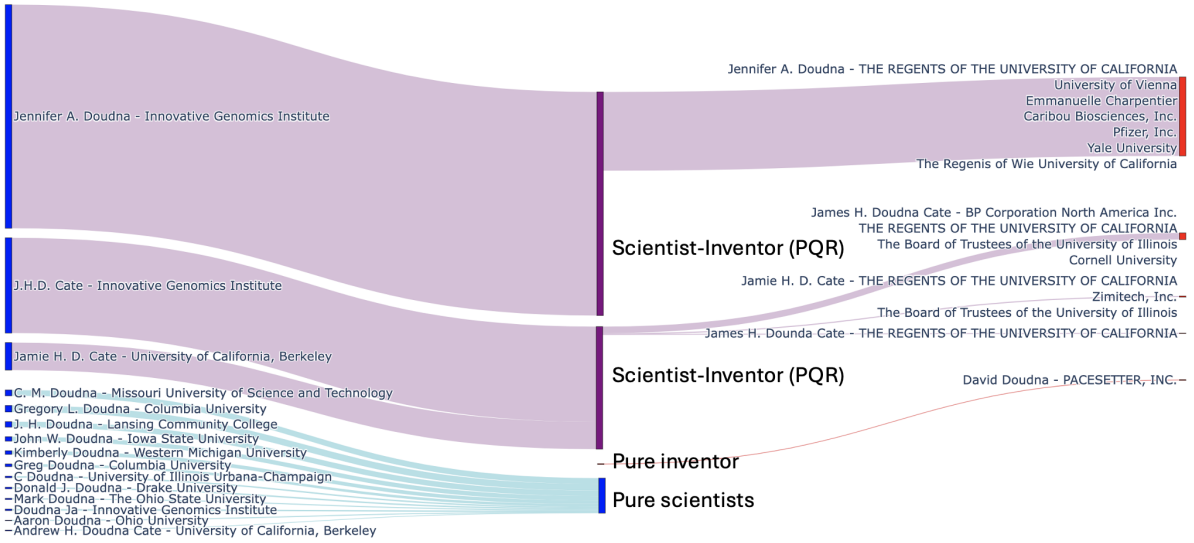


Figure S1: Sankey diagram for all scientists and inventors with the last name string "Doudna." The institution or organization of the science paper or assignee (owner) of the patent follows the name. Misspellings ("Regenis of Wie") and the patent assignee Emmanuelle Charpentier are errors introduced by our algorith but rather reflect the original data (see the pdf of US Patent 11,332,761 in Figure S24).

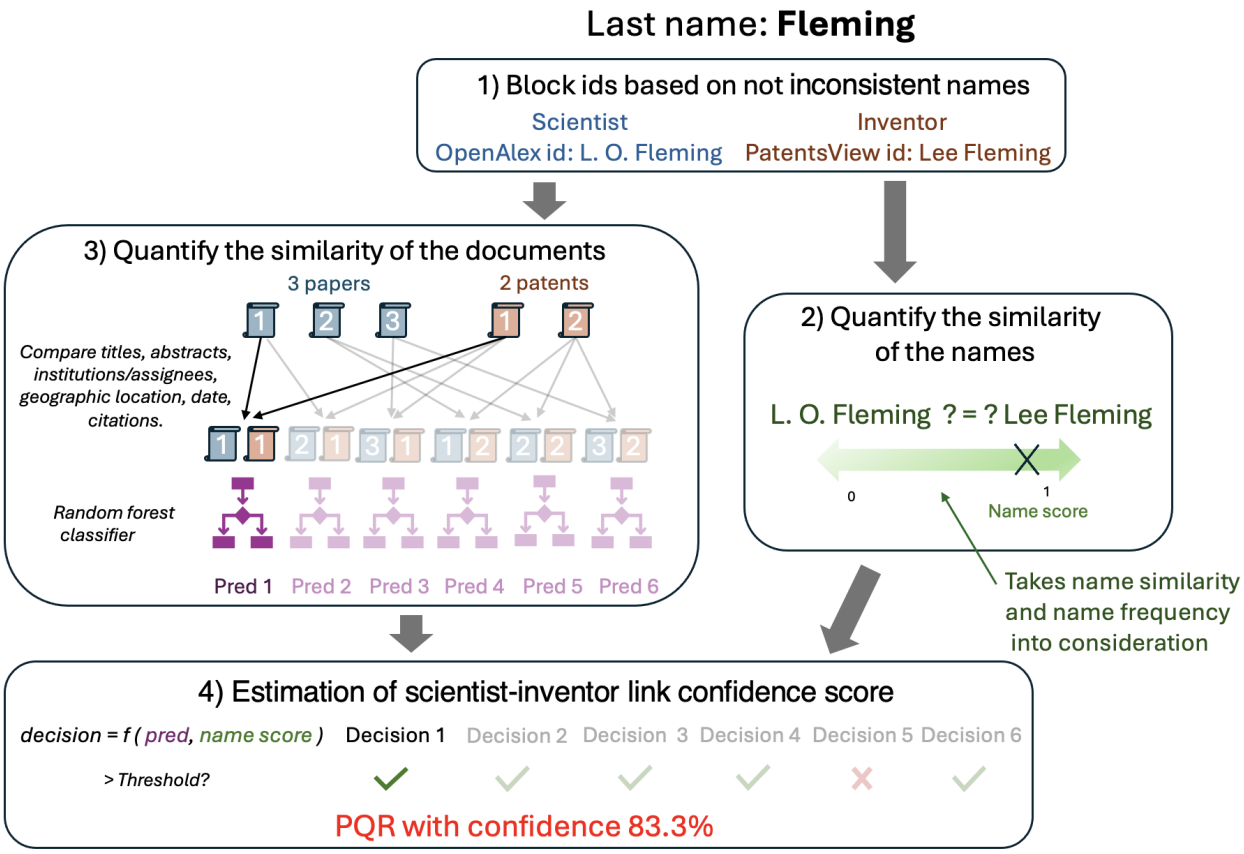


Figure S2: High level flowchart of the data processing flow.

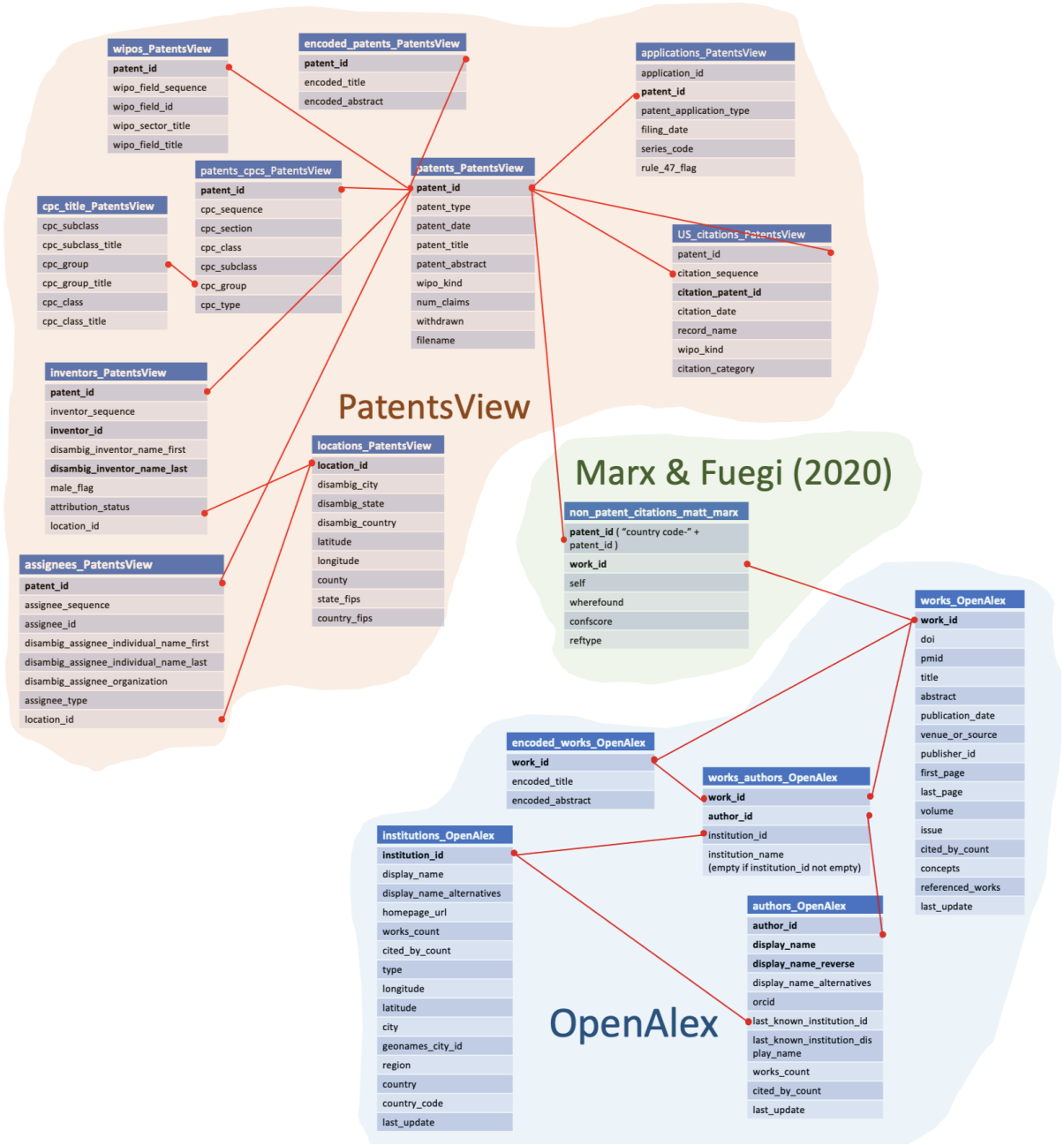
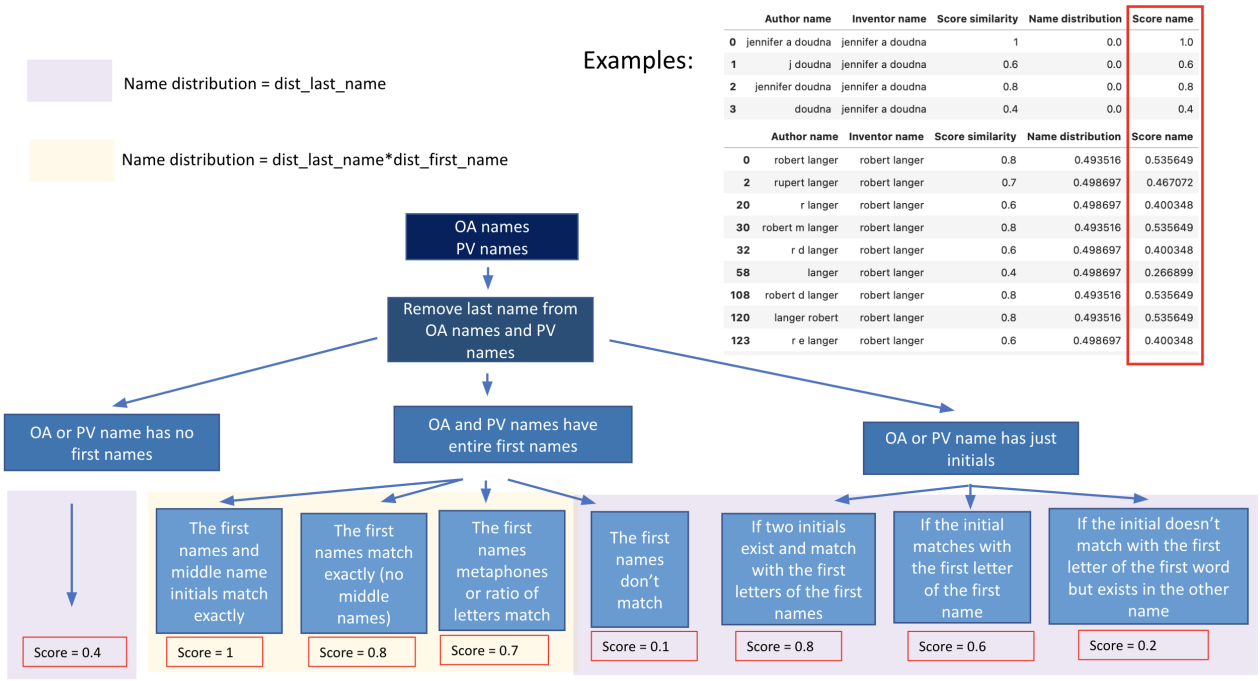


Figure S3: Schema of the input databases (OA, PV, Marx and Fuegi 2022 [2] patent to paper citations) used in the linkage process presented here.

	arnold behrman	brent r behrman	michael behrman	keith l behrman	stacey behrman	dan a behrman	andrea behrman
jere r behrman	0	0	0	0	0	0	0
andrea l behrman	0	0	0	0	0	0	1
stephen w behrman	0	0	0	0	0	0	0
harold r behrman	0	0	0	0	0	0	0
richard e behrman	0	0	0	0	0	0	0
s j behrman	0	0	0	0	1	0	0
edward j behrman	0	0	0	0	0	0	0
alison behrman	0	0	0	0	0	1	0
julia andrea behrman	0	0	0	0	0	0	0
douglas n behrman	0	0	0	0	0	0	0
rachel e behrman	0	0	0	0	0	0	0
emily l behrman	0	0	0	0	0	0	0
shannon behrman	0	0	0	0	0	0	0

Figure S4: Example of selection of match candidate pairs: OA scientist names are on the vertical axis, PV inventor names are across the horizontal, and an indicator when they are assessed to potentially be the same individual and retained for further processing.



	Author name	Inventor name	Score similarity	Name distribution	Score name
0	jennifer a doudna	jennifer a doudna	1	0.0	1.0
1	j doudna	jennifer a doudna	0.6	0.0	0.6
2	jennifer doudna	jennifer a doudna	0.8	0.0	0.8
3	doudna	jennifer a doudna	0.4	0.0	0.4

	Author name	Inventor name	Score similarity	Name distribution	Score name
0	robert langer	robert langer	0.8	0.493516	0.535649
2	rupert langer	robert langer	0.7	0.498697	0.467072
20	r langer	robert langer	0.6	0.498697	0.400348
30	robert m langer	robert langer	0.8	0.493516	0.535649
32	r d langer	robert langer	0.6	0.498697	0.400348
58	langer	robert langer	0.4	0.498697	0.266899
108	robert d langer	robert langer	0.8	0.493516	0.535649
120	langer robert	robert langer	0.8	0.493516	0.535649
123	r e langer	robert langer	0.6	0.498697	0.400348

Figure S5: Flowchart on subjective scoring of the name match probability and integration of information from the frequency of the name distribution, conditional on being a candidate match pair.

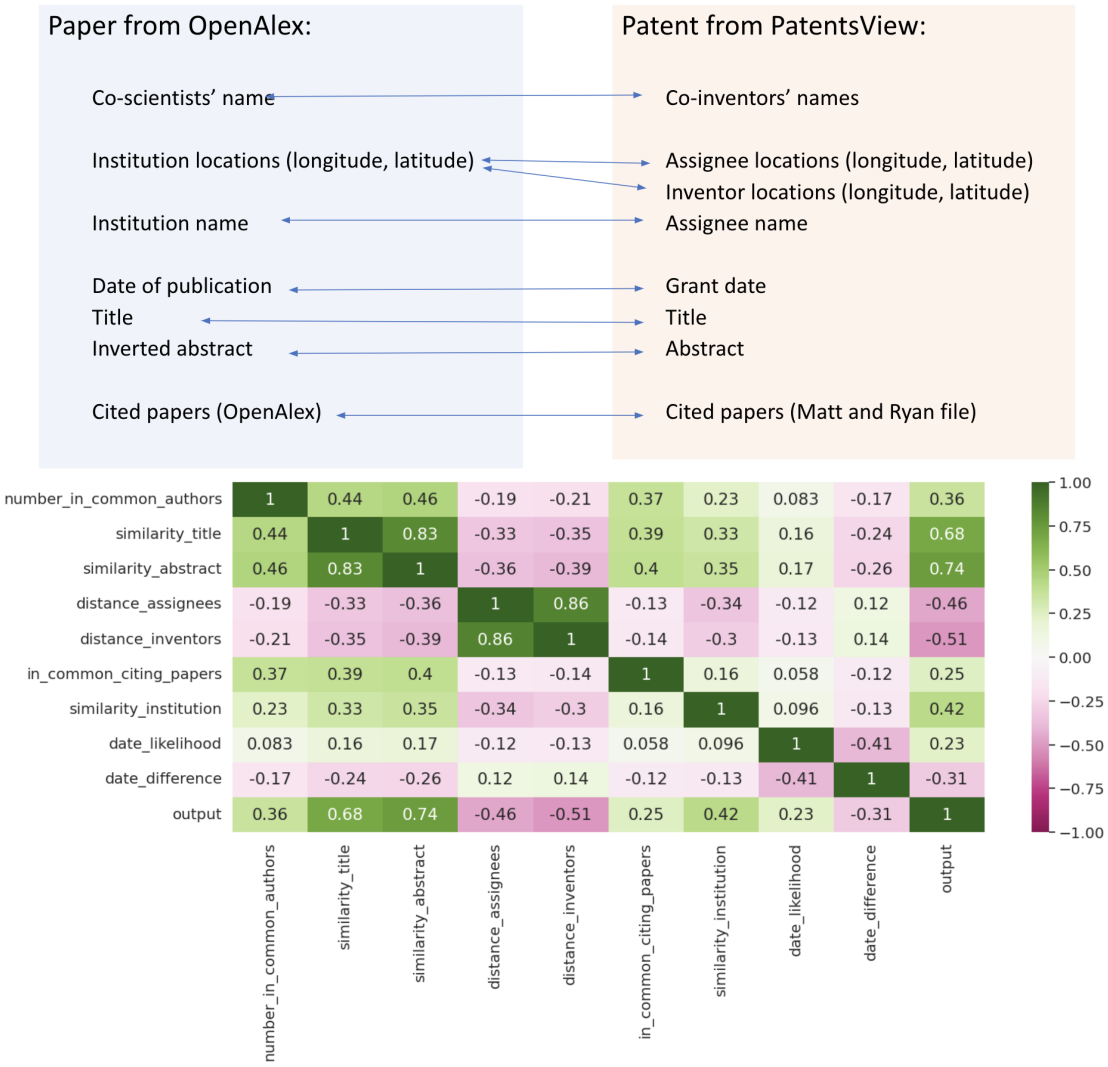


Figure S6: **Upper figure**: Available data and correspondence in fields between OA and PV. **Lower figure**: Correlation between the model features and the data classification (output=1 when the patent and the paper are written by the same individual, otherwise output=0).

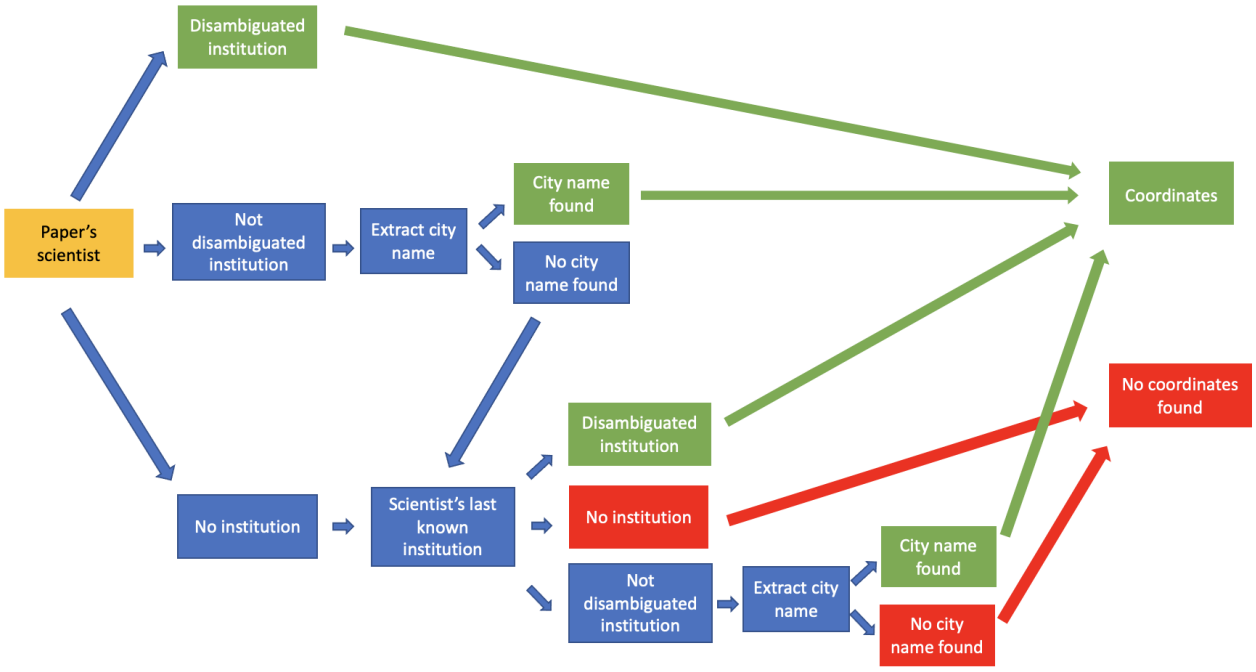


Figure S7: The process of determining a scientist's geographical location, through disambiguation of his/her institution and assignment of latitude and longitude. If OA has successfully disambiguated the institution, those coordinates are used. If the paper lacks an institution, the scientist's last institution is used. If the institution has not been disambiguated, we assign latitude and longitude through sPacey for institutions we can disambiguate. Finally, some scientist's locations cannot be assigned, and these data points are handled in a separate prediction model which does not rely on geography.

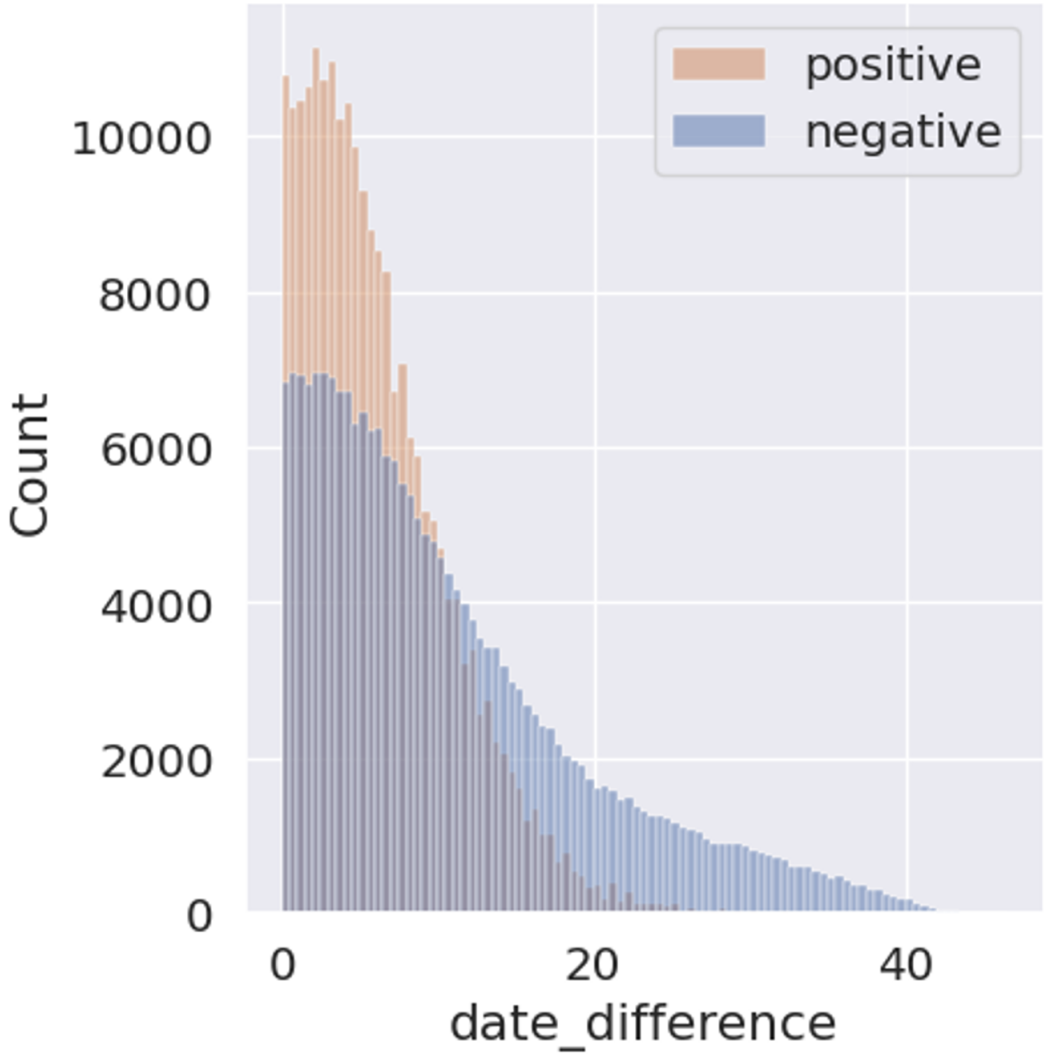


Figure S8: Plots of the positive (written by the same person) and negative (not written by the same person) distributions of the date differences (difference between the mean of the papers' publication dates and the patent grant date) of our training set. X axis date differences are in years.

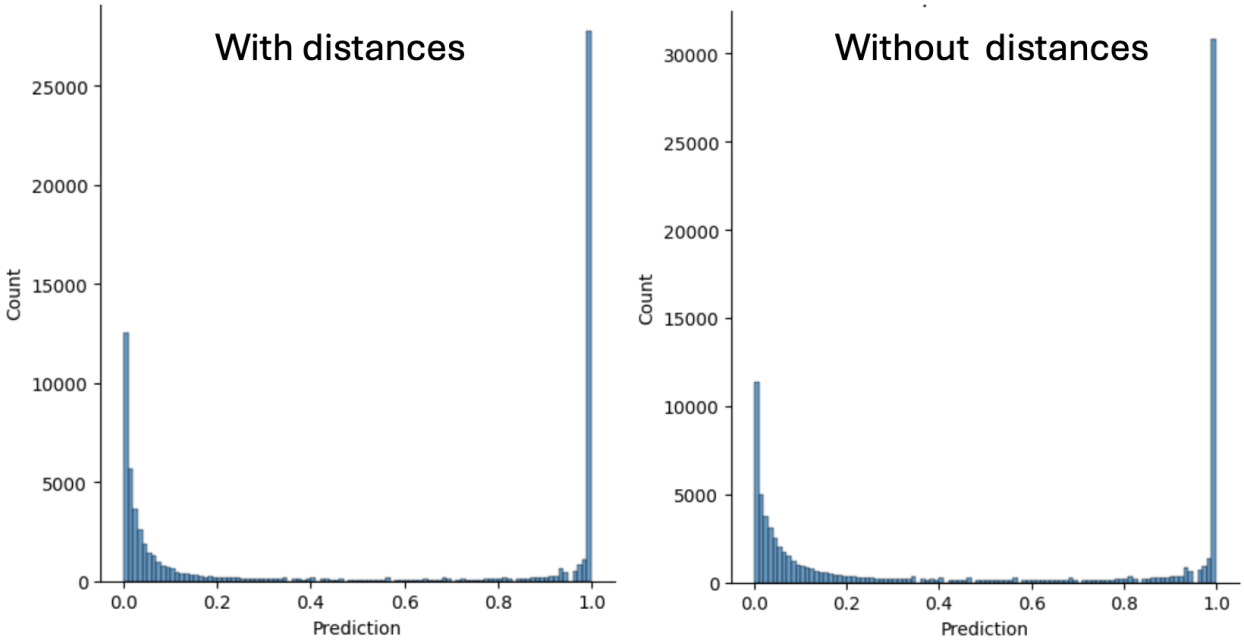
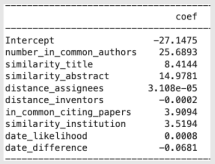
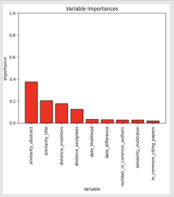


Figure S9: **Left figure**: Distribution of the predictions for the test set using the random forest model which uses geographic distances as features. **Right figure**: Distribution of the predictions for the test set using the random forest model which does not use geographic distances as features.

Model	Baseline model	Logistic model	Decision tree classifier	Random forest classifier
Interpretation	Always predicting 0 (the patent and the paper are not written by the same author)			 Threshold = 0.5
Accuracy	51.2%	94.2%	94.8%	96.2%
True positive rate (we want it equal to 100%)	0%	92.1%	93.2%	95.1%
False positive rate (we want it equal to 0%)	0%	3.9%	3.6%	2.8%

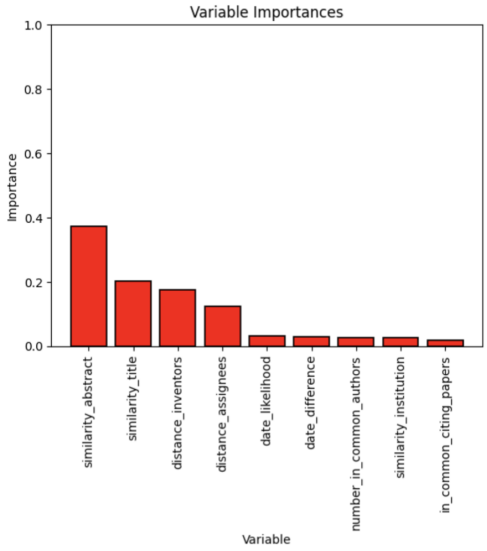
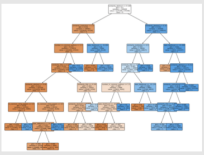
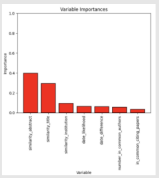


Figure S10: **Upper figure**: comparison of baseline, Logistic, decision tree classifier, gradient boosting and random forest classifier models, when geographic distance features are included. **Lower figure**: Importance of each feature for the random forest classifier model, when geographic distance features are included.

Model	Baseline model	Logistic model	Decision tree classifier	Random forest classifier
Interpretation	Always predicting 0 (the patent and the paper are not written by the same author)	coef Intercept −30.3424 number_in_common_authors 28.0278 similarity_title 8.3766 similarity_abstract 15.4946 in_common_citing_papers 3.9301 similarity_institution 3.7970 date_likelihood 0.0007 date_difference −0.0616		 Threshold = 0.5
Accuracy	49.8%	93.5%	93.1%	93.9%
True positive rate (we want it equal to 100%)	0%	91.1%	90.2%	92.3%
False positive rate (we want it equal to 0%)	0%	4.1%	4.1%	4.6%

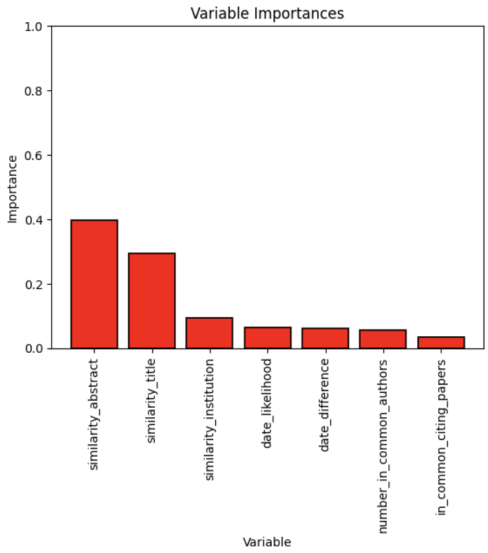


Figure S11: **Upper figure**: comparison of baseline, Logistic, decision tree classifier, gradient boosting and random forest classifier models, when geographic distance features are not available. **Lower figure**: Importance of each feature for the random forest classifier model, when geographic distance features are not available.

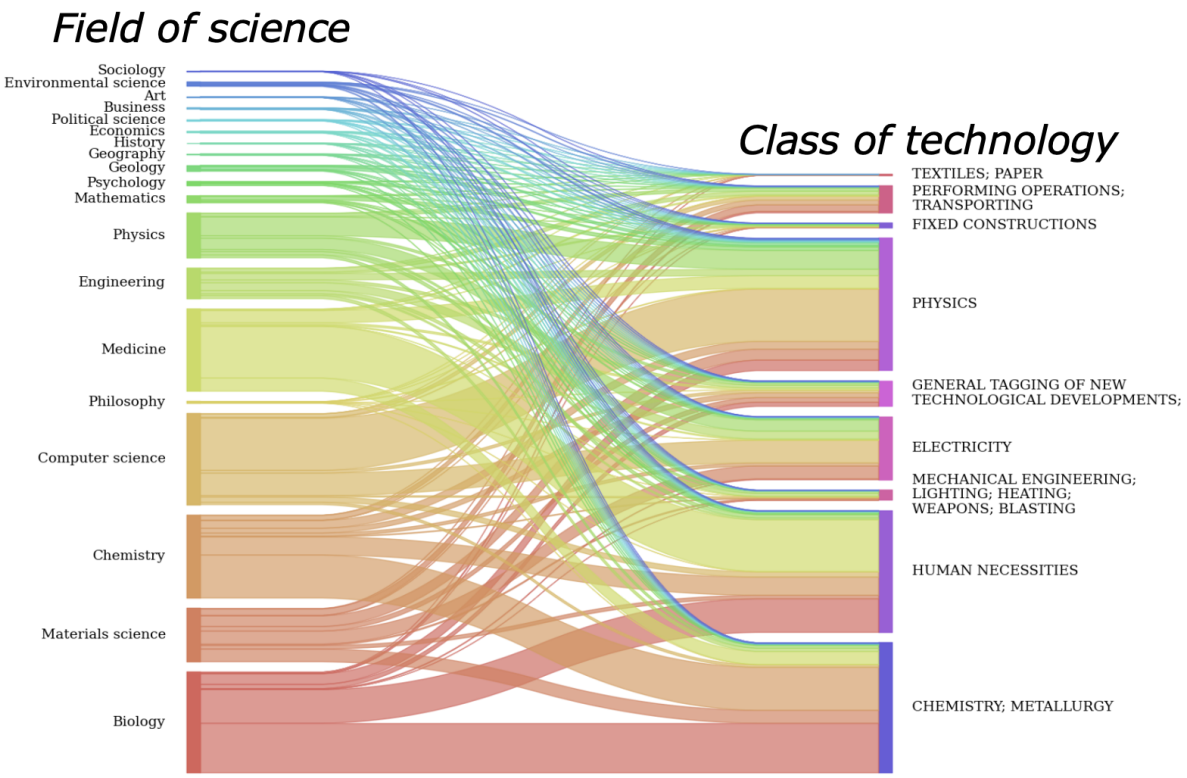


Figure S12: Left side illustrates most common field of science publication and right side illustrates most common class of technology invention, for a Pasteur Quadrant Researcher. Based on 19 OpenAlex fields of science and nine patent categories as defined by the World Patent Organization. Concurrent, U.S.-based PQRs only.

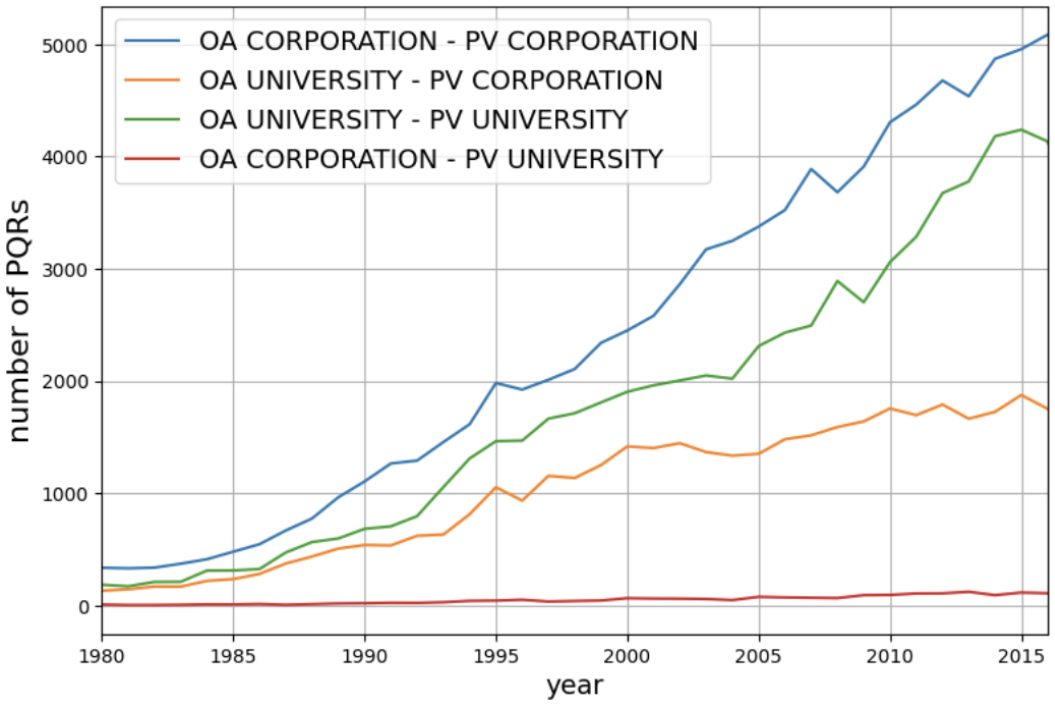


Figure S13: Corporate vs. university affiliations of PQRs. Concurrent and U.S. PQRs only.

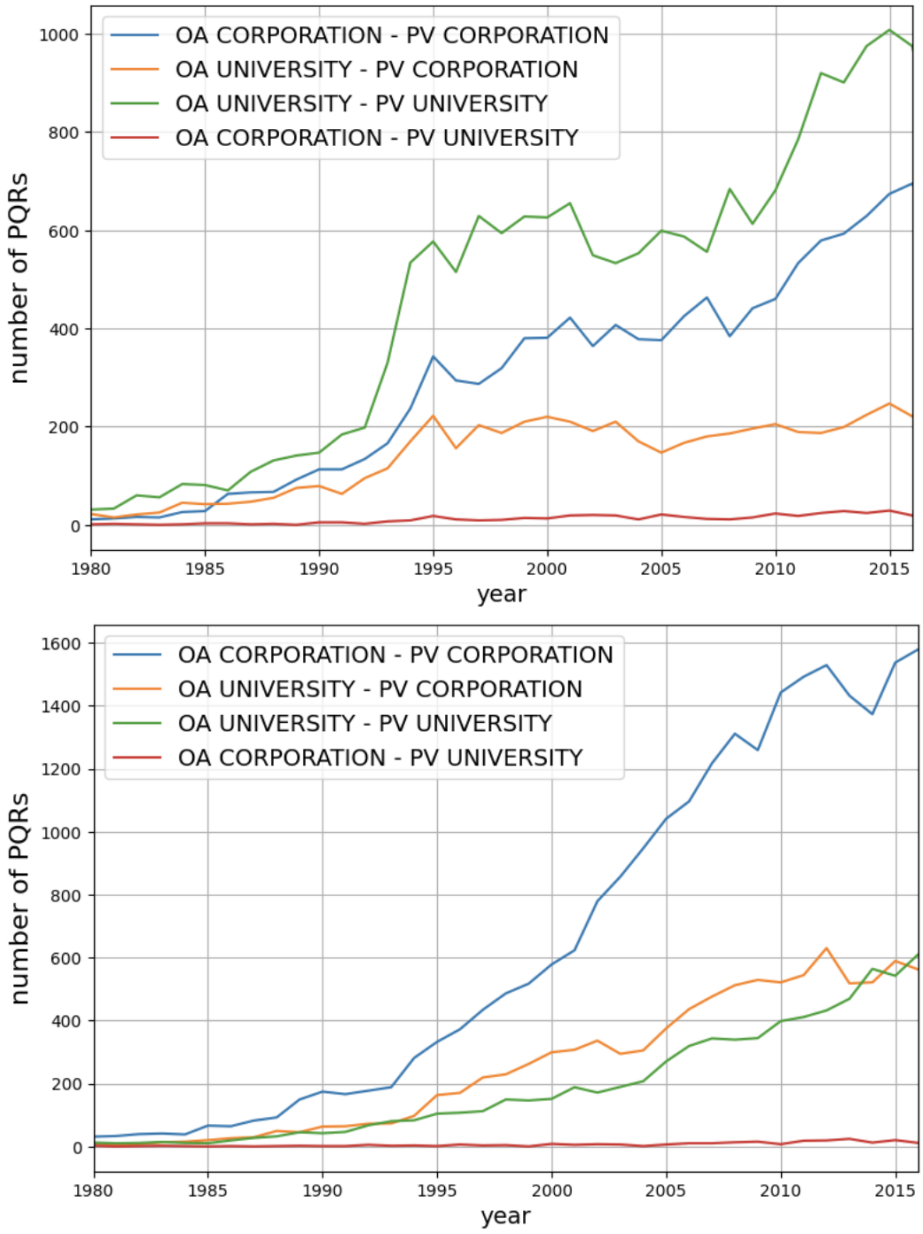


Figure S14: Corporate vs. university affiliations of PQRs. **Upper figure**: PQR's affiliations for U.S. biology PQRs. **Lower figure**: PQR's affiliations for computer science PQRs. Concurrent and U.S. PQRs only for both figures.

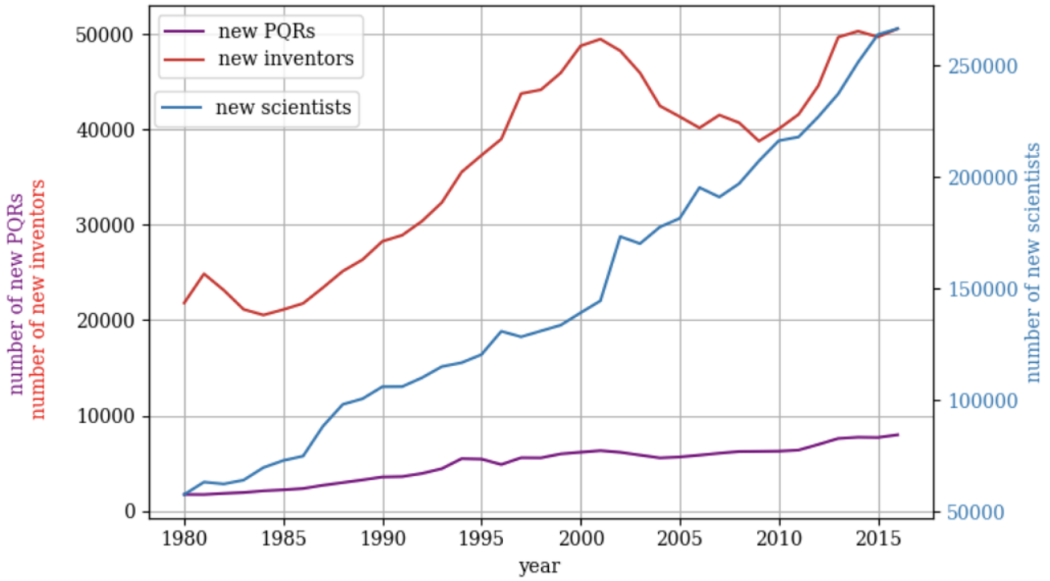


Figure S15: Evolution of the number of new PQRs each year (left axis purple line), new inventors each year (left axis red line), and new scientists each year (right axis blue line) from 1980 to 2016. First year is defined as first concurrent year in a career or first allochronic year (publication of second type of alternate document), whichever occurs first. U.S. PQRs only.

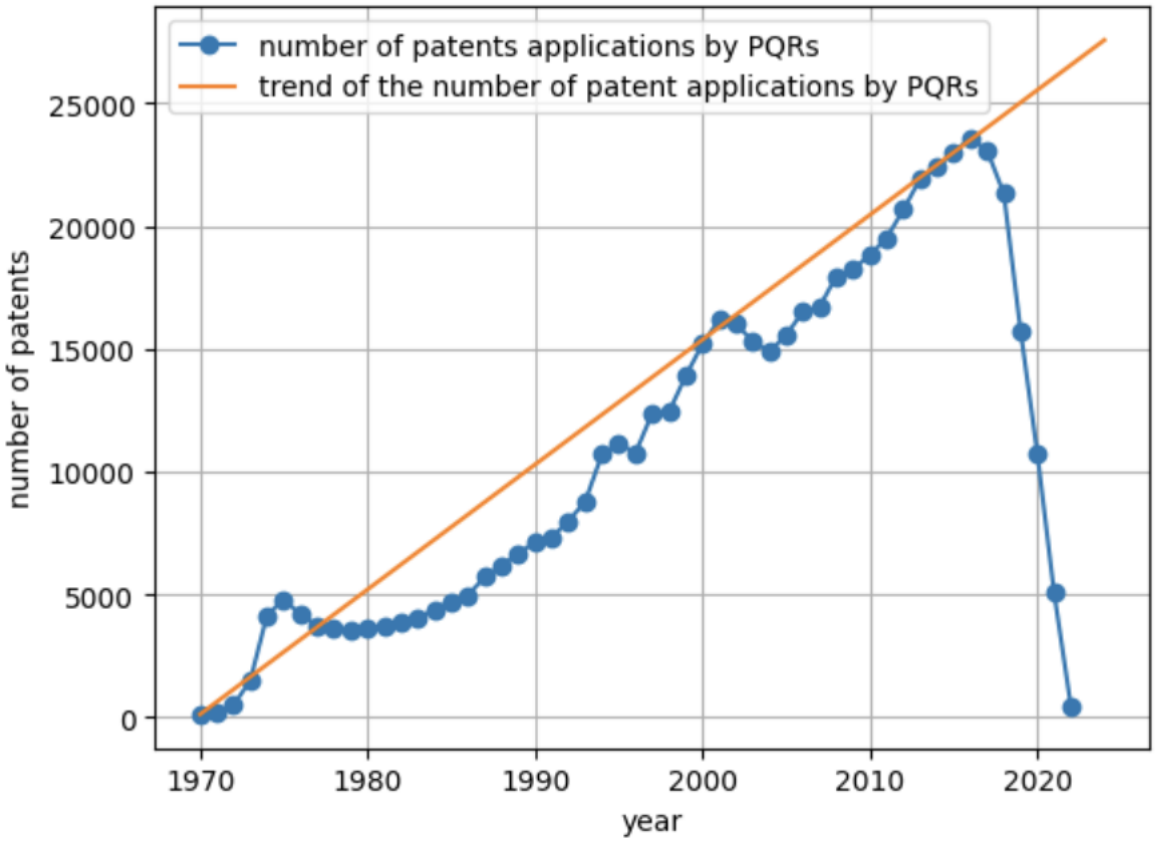


Figure S16: Illustration of patent pendency bias. All PV data.

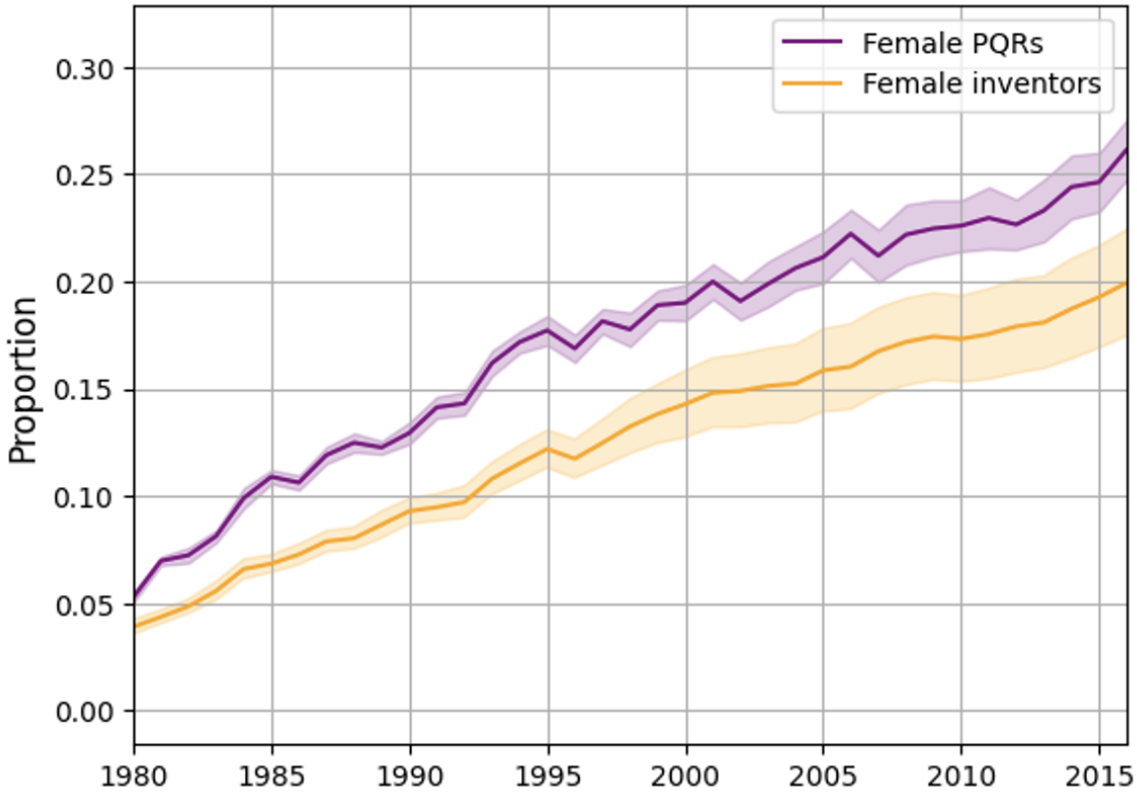


Figure S17: Evolution of the proportion of female PQRs by over time. Note that identification of gender is from PV; there is no analogous data in OA. First year of appearance and U.S. PQRs only.

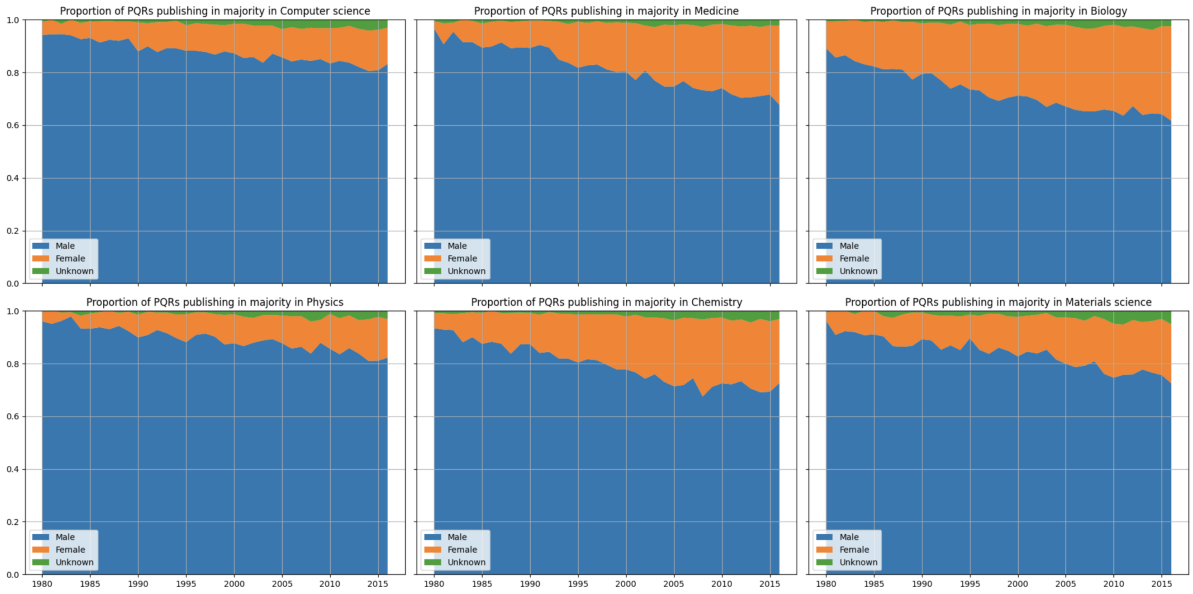


Figure S18: Evolution of the number of PQRs by field of study and gender over time. Note that identification of gender is from PV; there is no analogous data in OA.

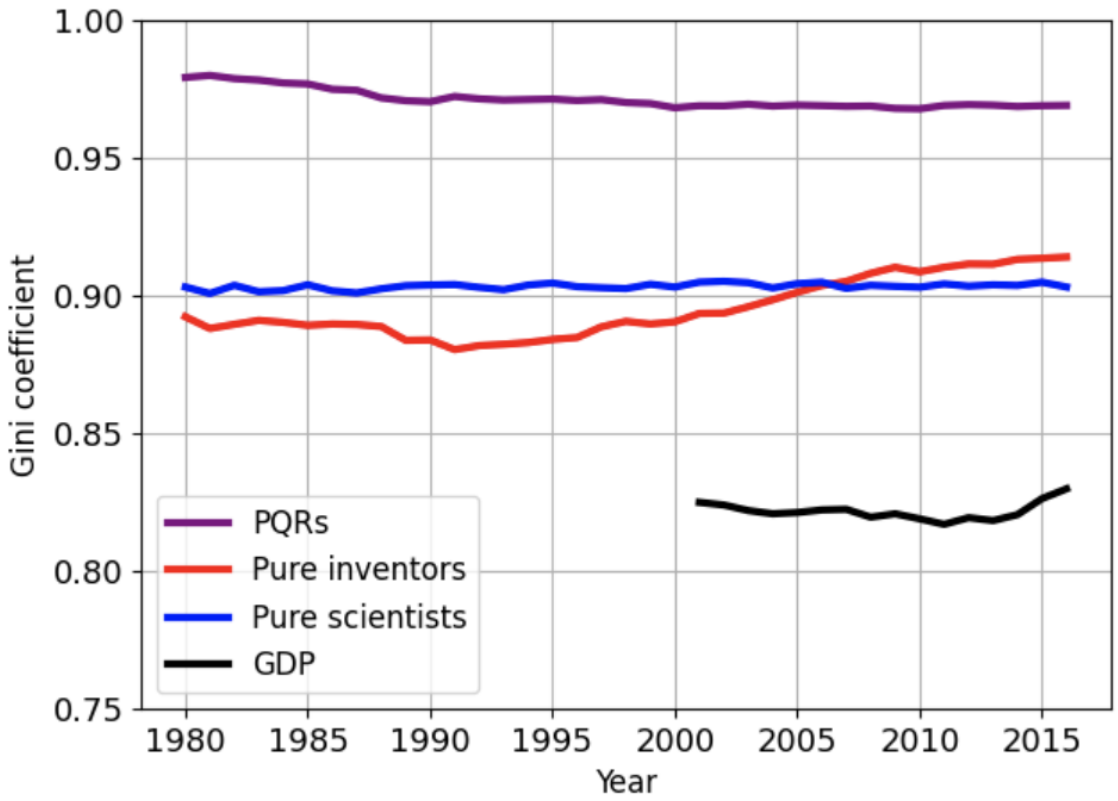


Figure S19: Gini coefficient over time by U.S. county, for PQRs, pure scientists, pure inventors, and GDP. The GDP data are taken from BEA website https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas accessed in October 2023. Concurrent and U.S. PQRs only.

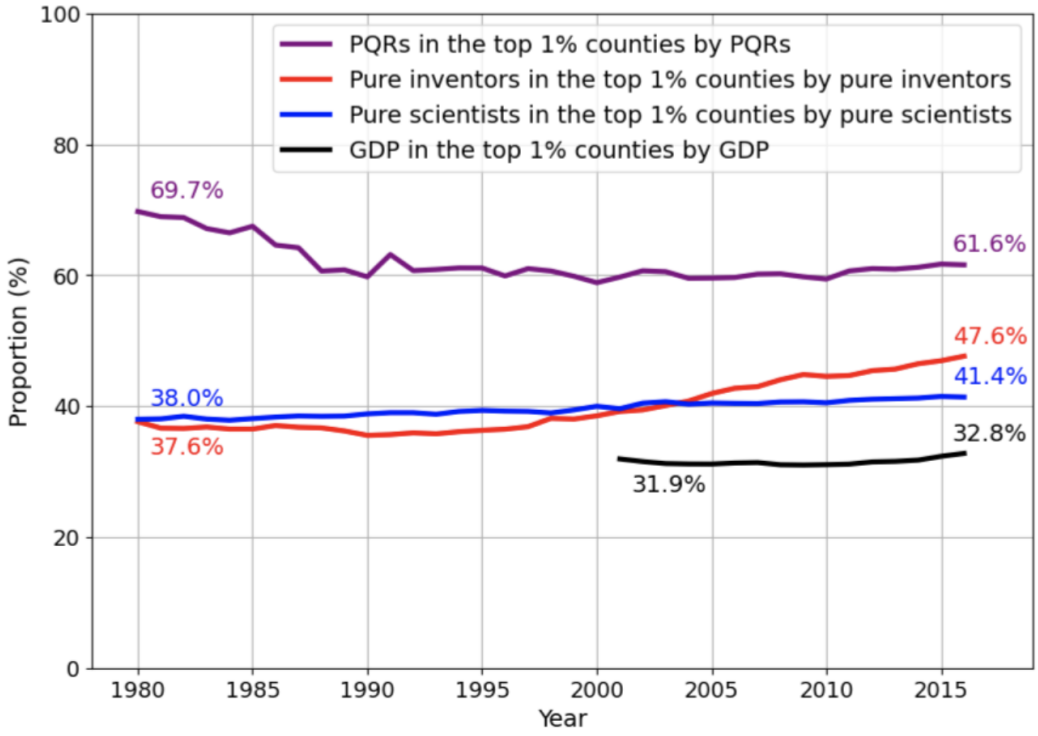


Figure S20: For the top 1% of U.S. counties, calculated each year and for each variable individually, the proportion of PQRs, pure inventors, pure scientists, and national GDP, 2001-2016. GDP data from BEA website https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas accessed in October 2023. Concurrent and U.S. PQRs only.

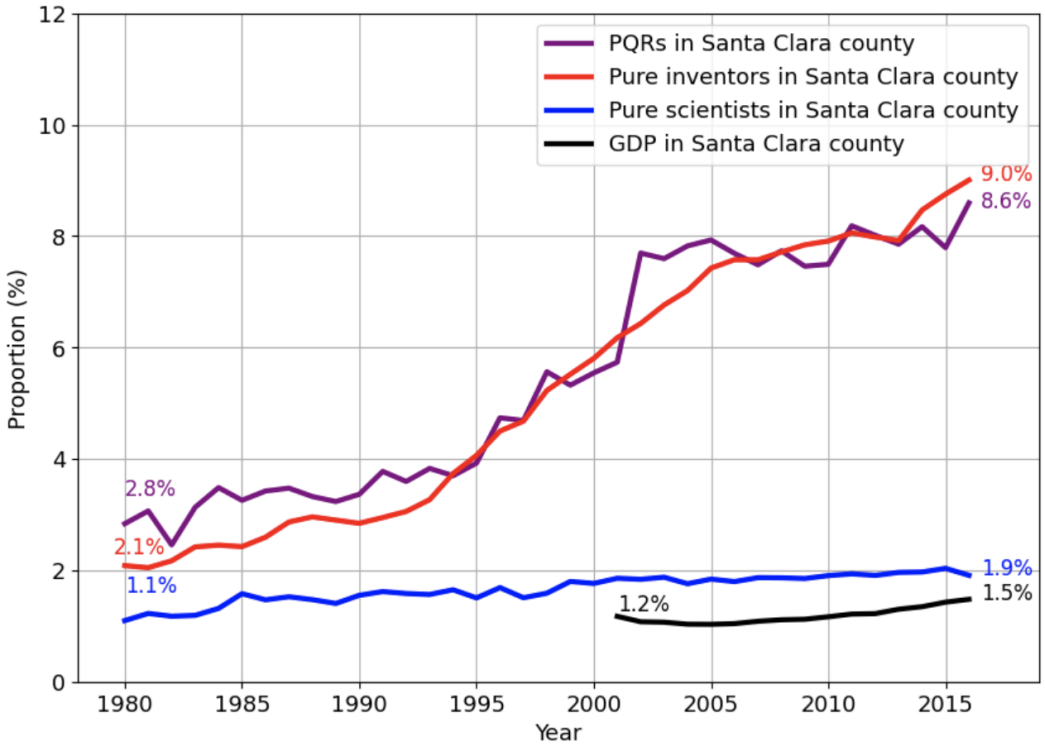


Figure S21: For Santa Clara County, relative to all other U.S. counties, the proportion of PQRs, pure inventors, pure scientists, and GDP, 1980-2016. Top 1% of counties calculated similarly to Figure 1 in the main text, i.e., by the highest average GDP of U.S. counties 2001-2016. GDP data from BEA website https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas accessed in October 2023. Concurrent and U.S. PQRs only.

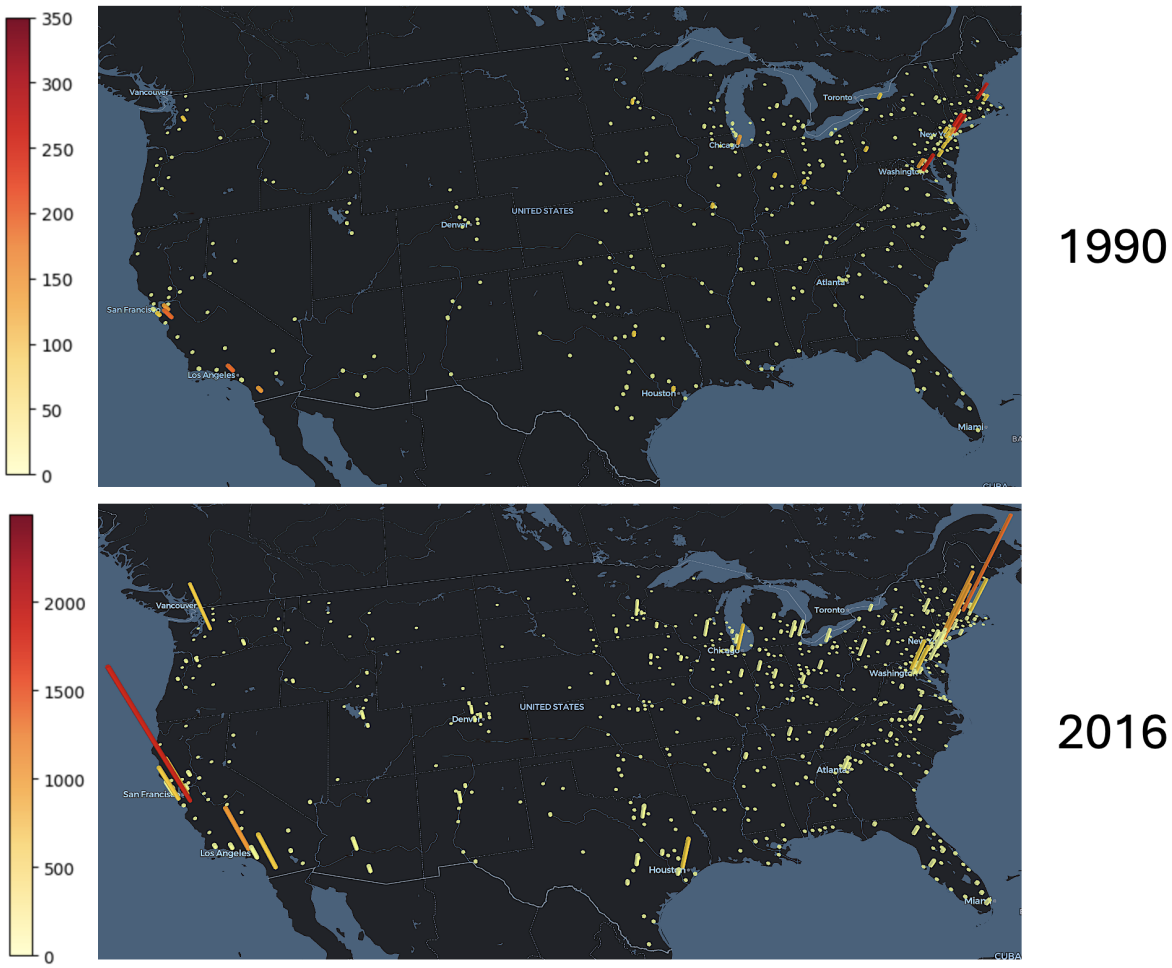


Figure S22: Choropleth of U.S. county locations of PQRs in the US in 1990 and 2016. Note heights across time periods are consistent but color intensity changes. Color intensity is uniform within time period. Concurrent and U.S. PQRs only.

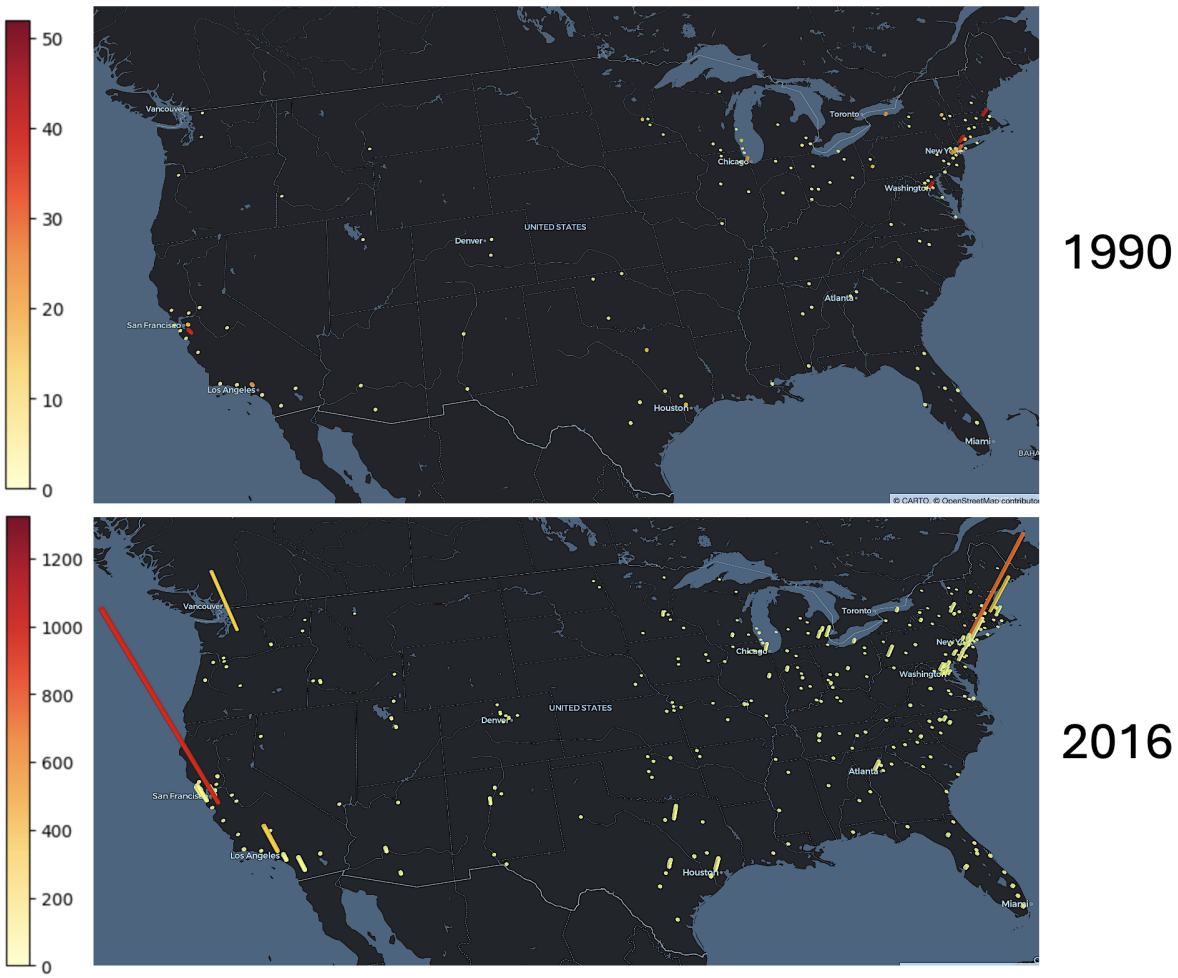
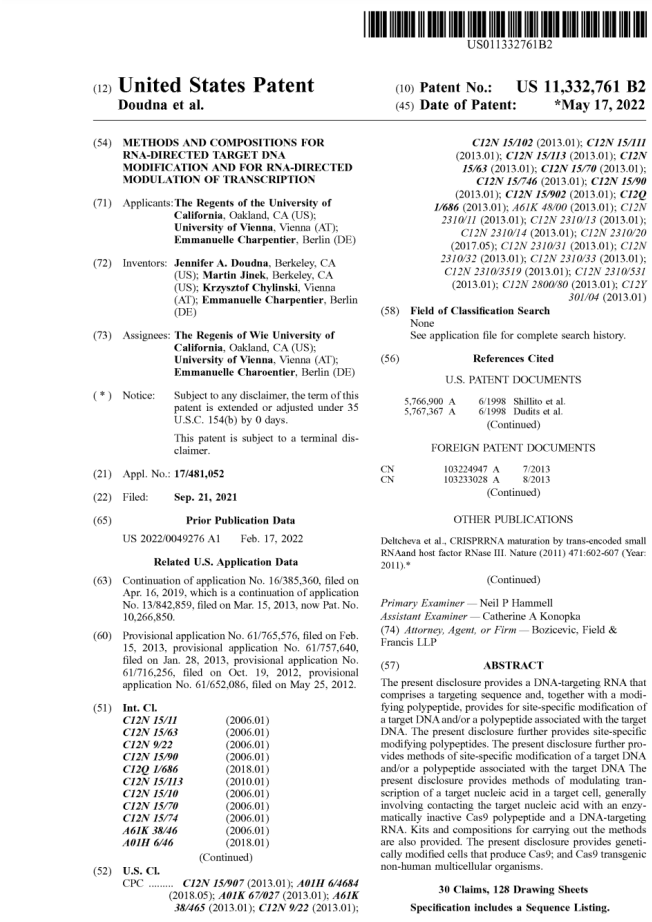


Figure S23: Choropleth of U.S. county locations of computer science PQRs in the US in 1990 and 2016. Note heights across time periods are consistent but color intensity changes. Color intensity is uniform within time period. Concurrent and U.S. PQRs only.



US011332761B2

(12) **United States Patent**
Doudna et al.

(10) **Patent No.: US 11,332,761 B2**
(45) **Date of Patent: *May 17, 2022**

(54) **METHODS AND COMPOSITIONS FOR RNA-DIRECTED TARGET DNA MODIFICATION AND FOR RNA-DIRECTED MODULATION OF TRANSCRIPTION**

(71) Applicants: **The Regents of the University of California**, Oakland, CA (US); **University of Vienna**, Vienna (AT); **Emmanuelle Charpentier**, Berlin (DE)

(72) Inventors: **Jennifer A. Doudna**, Berkeley, CA (US); **Martin Jinek**, Berkeley, CA (US); **Krzysztof Chylinski**, Vienna (AT); **Emmanuelle Charpentier**, Berlin (DE)

(73) Assignees: **The Regenis of Wie University of California**, Oakland, CA (US); **University of Vienna**, Vienna (AT); **Emmanuelle Charoentier**, Berlin (DE)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **17/481,052**

(22) Filed: **Sep. 21, 2021**

(65) **Prior Publication Data**

US 2022/0049276 A1 Feb. 17, 2022

Related U.S. Application Data

(63) Continuation of application No. 16/385,360, filed on Apr. 16, 2019, which is a continuation of application No. 13/842,859, filed on Mar. 15, 2013, now Pat. No. 10,266,850.

(60) Provisional application No. 61/765,576, filed on Feb. 15, 2013, provisional application No. 61/757,640, filed on Jan. 28, 2013, provisional application No. 61/716,256, filed on Oct. 19, 2012, provisional application No. 61/652,086, filed on May 25, 2012.

(51) **Int. Cl.**

C12N 15/11	(2006.01)
C12N 15/63	(2006.01)
C12N 9/22	(2006.01)
C12N 15/90	(2006.01)
C12Q 1/686	(2018.01)
C12N 15/113	(2010.01)
C12N 15/10	(2006.01)
C12N 15/70	(2006.01)
C12N 15/74	(2006.01)
A61K 38/46	(2006.01)
A01H 6/46	(2018.01)

(Continued)

(52) **U.S. Cl.**

CPC ***C12N 15/907*** (2013.01); ***A01H 6/4684*** (2018.05); ***A01K 67/027*** (2013.01); ***A61K 38/465*** (2013.01); ***C12N 9/22*** (2013.01); ***C12N 15/102*** (2013.01); ***C12N 15/111*** (2013.01); ***C12N 15/113*** (2013.01); ***C12N 15/63*** (2013.01); ***C12N 15/70*** (2013.01); ***C12N 15/746*** (2013.01); ***C12N 15/90*** (2013.01); ***C12N 15/902*** (2013.01); ***C12Q 1/686*** (2013.01); *A61K 48/00* (2013.01); *C12N 2310/11* (2013.01); *C12N 2310/13* (2013.01); *C12N 2310/14* (2013.01); *C12N 2310/20* (2017.05); *C12N 2310/31* (2013.01); *C12N 2310/32* (2013.01); *C12N 2310/33* (2013.01); *C12N 2310/3519* (2013.01); *C12N 2310/531* (2013.01); *C12N 2800/80* (2013.01); *C12Y 301/04* (2013.01)

(58) **Field of Classification Search**

None

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,766,900 A 6/1998 Shillito et al.
5,767,367 A 6/1998 Dudits et al.
(Continued)

FOREIGN PATENT DOCUMENTS

CN 103224947 A 7/2013
CN 103233028 A 8/2013
(Continued)

OTHER PUBLICATIONS

Deltcheva et al., CRISPRRNA maturation by trans-encoded small RNAand host factor RNase III. Nature (2011) 471:602-607 (Year: 2011).*

(Continued)

Primary Examiner — Neil P Hammell
Assistant Examiner — Catherine A Konopka
(74) *Attorney, Agent, or Firm* — Bozicevic, Field & Francis LLP

(57) **ABSTRACT**

The present disclosure provides a DNA-targeting RNA that comprises a targeting sequence and, together with a modifying polypeptide, provides for site-specific modification of a target DNA and/or a polypeptide associated with the target DNA. The present disclosure further provides site-specific modifying polypeptides. The present disclosure further provides methods of site-specific modification of a target DNA and/or a polypeptide associated with the target DNA The present disclosure provides methods of modulating transcription of a target nucleic acid in a target cell, generally involving contacting the target nucleic acid with an enzymatically inactive Cas9 polypeptide and a DNA-targeting RNA. Kits and compositions for carrying out the methods are also provided. The present disclosure provides genetically modified cells that produce Cas9; and Cas9 transgenic non-human multicellular organisms.

30 Claims, 128 Drawing Sheets

Specification includes a Sequence Listing.

Figure S24: Original front page data for US patent 11332761. Note assignee spelling error "Regenis of Wie" and assignment to Nobel Prize winner Emmanuelle Charpentier. This provides an example of an unavoidable source data error.

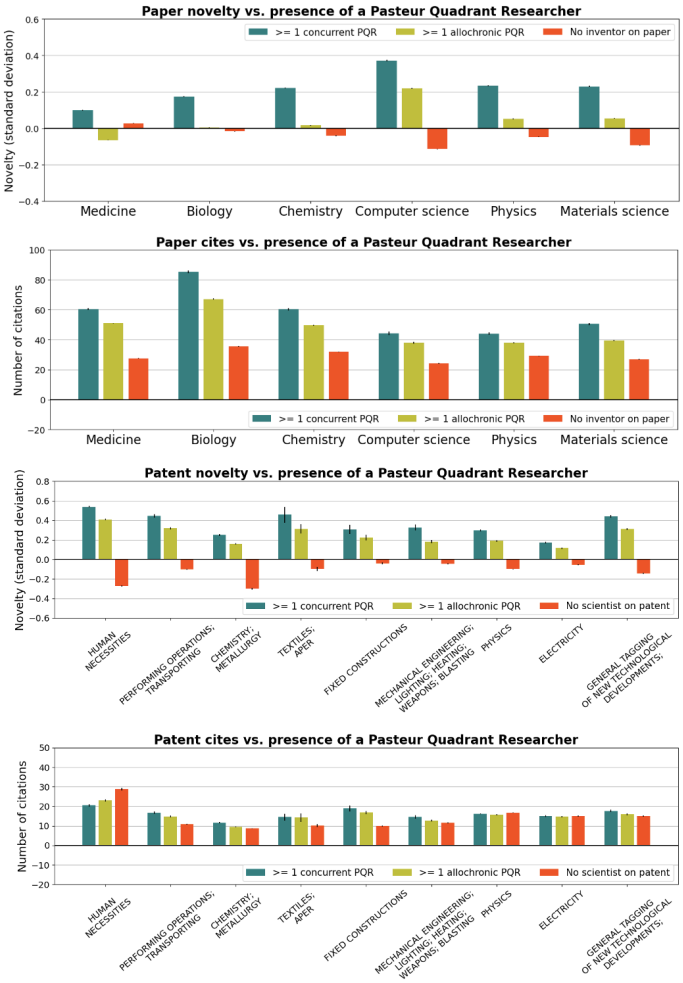


Figure S25: Descriptive statistics broken out by papers' fields and patents' CPCs for average mean standardized novelty and average citation impact, for a paper or patent publication, where the publication included a concurrent PQR, an allochronic PQR, or neither, 1980-2016, inclusive.

Table S1: Variable descriptions and statistics of the PQR dataset.

Variable name	Data provenance	Description
author_id	OpenAlex	scientist id from OA. Note that OA scientist ids may change at the next update (these scientist ids correspond to the update from August 2023). Table S12 is provided to allow user to link our dataset to new OA scientist ids.
inventor_id	PatentsView	inventor id from PV. Note that PV inventor ids may change at each PV update (these inventor ids correspond to the update from June 2023). Table S13 is provided to allow user to link our dataset to new PV inventor ids.
scientist_name	OpenAlex	(Cleaned and normalized) scientist full name from OpenAlex
inventor_name	PatentsView	(Cleaned and normalized) inventor full name from PatentsView
name_similarity	this work	A score between 0.5 and 1 indicating how likely the inventor and the scientist are the same individual based only on their names (using name similarity and frequency - see Appendix section "Quantifying the similarity and frequency of the names").
number_of_papers	OpenAlex	Number of works corresponding to the scientist id from OpenAlex
number_of_patents	PatentsView	Number of granted US patents corresponding to the inventor id from PatentsView
number_of_comparisons	this work	Only for papers and patents with a title and abstract, the number of papers multiplied by the number of patents for the identified PQRs. Note that the number of comparisons is lower than the number_of_papers multiplied by the number_of_patents as we do not compare papers and patents without titles or abstracts.
number_of_matches	this work	Number of comparisons realized by the algorithm for which the paper and the patent were predicted as being written by the same individual. Note that the number_of_matches $\leq$ number_of_comparisons.
confidence	this work	Score between 0% and 100% which estimates whether the scientist and the inventor are the same individual. This score represents the proportion of papers and patents predicted as being written by the same person by our algorithm. It incorporates the predictions of the random forest model and the similarity and frequency between the scientist and inventor names. It corresponds to the number of matches divided by the number of comparisons (for documents that have titles and abstracts).

Table S2: Descriptive summary statistics for different confidence datasets.

	PQRs dataset			OpenAlex	PatentsView
	PQRs - high conf	PQRs - mid conf	PQRs - low conf		
Number of PQRs	581,566	699,007	746,522	x	x
Scientist ids	660,984	809,823	873,191	90,274,750	x
Inventor ids	599,996	722,725	773,384	x	3,351,639
Number of works	22,584,479	25,286,393	26,866,418	275,492,415	x
Number of patents	2,412,401	2,884,842	3,138,176	x	8,436,247
Concurrent years - inventors	1,088,338	1,281,350	1,386,400	x	x
Concurrent years - scientists	1,234,463	1,496,279	1,646,281	x	x

Table S3: Detailed descriptive statistics for high confidence dataset.

	PQRs dataset - high confidence							
	Name score	# of papers	# of patents	# of comparisons	# of matches	# of inventor id per scientist id	# of scientist id per inventor id	confidence
mean	0.67	54.52	6.62	327.97	273.87	1.07	1.29	0.87
std	0.10	116.26	16.18	2,569.56	2,192.52	2.99	1.11	0.15
min	0.50	1	1	1	1	1	1	0.5
25%	0.59	3	1	5	5	1	1	0.76
50%	0.65	13	2	25	21	1	1	0.95
75%	0.73	53	6	126	106	1	1	1
max	1.0	15,395	1,645	649,491	611,643	24	225	1

Table S4: Regressions for citation and novelty by individual paper or patent - 1 year window

	Paper citations	Paper novelty	Patent citations	Patent novelty
	(1)	(2)	(3)	(4)
has_concurrent_PQR	0.2469***	0.0017***	0.1423***	0.0069***
	(0.0054)	(0.0000)	(0.0062)	(0.0001)
has_allochronic_PQR	0.1330***	0.0008***	0.0357***	0.0056***
	(0.0033)	(0.0000)	(0.0050)	(0.0001)
log_tot_scientists	0.4116***	-0.0008***		
	(0.0026)	(0.0000)		
log_tot_inventors			0.1637***	-.0008***
			(0.0035)	(0.0001)
R-squared	x	0.4799	x	0.4314
Pseudo R2	0.3315	x	0.5164	x
Number of observations	12,762,702	10,751,271	2,603,130	2,128,768
Standard deviation	167.5695	0.0182	69.2509	0.0515
Pub. year # Institution # Concept 1	y	y	n	n
App. year # Assignee # CPC class	n	n	y	y
Number of FE absorbed	1,430,849	1,346,880	300,535	249,434

Note: please see Table S9 for descriptive statistics on each dataset. Estimates from STATA PPMLHDFE (Models 1 and 3) and REGHDFE (Models 2 and 4) routines.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S5: Regressions for citation and novelty by individual paper or patent - 3 year window

	Paper citations (5)	Paper novelty (6)	Patent citations (7)	Patent novelty (8)
has_concurrent_PQRs	0.2427***	0.0017***	0.1246***	0.0073***
	(0.0040)	(0.0000)	(0.0054)	(0.0001)
has_allochronic_PQRs	0.0951***	0.0006***	0.0088	0.0047***
	(0.0032)	(0.0000)	(0.0054)	(0.0001)
log_tot_scientists	0.4093***	-0.0008***		
	(0.0025)	(0.0000)		
log_tot_inventors			0.1654***	-0.0008***
			(0.0035)	(0.0001)
R-squared	x	0.4777	x	0.4315
Pseudo R2	0.3324	x	0.5164	x
Number of observations	12,760,888	9,168,541	2,603,130	2,128,768
Standard deviation	165.893	0.0184	69.2509	0.0515
Pub. year # Institution # Concept 1	y	y	n	n
App. year # Assignee # CPC class	n	n	y	y
Number of FE absorbed	1,430,578	1,152,981	300,535	249,434

Note: estimates from STATA PPMLHDFE (Models 5 and 7) and REGHDFE (Models 6 and 8) routines.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S6: Regressions for citation and novelty by individual paper and patent - 5 year window

	Paper citations (9)	Paper novelty (10)	Patent citations (11)	Patent novelty (12)
has_concurrent_PQRs	0.2316***	0.0017***	0.1115***	0.0076***
	(0.0036)	(0.0000)	(0.0052)	(0.0001)
has_allochronic_PQRs	0.0821***	0.0004***	-0.0002	0.0040***
	(0.0030)	(0.0000)	(0.0060)	(0.0001)
log_tot_scientists	0.4087***	-0.0008***		
	(0.0024)	(0.0000)		
log_tot_inventors			0.1661***	-0.0007***
			(0.0035)	(0.0001)
R-squared	x	0.4777	x	0.4315
Pseudo R2	0.3319	x	0.5163	x
Number of observations	12,763,576	9,167,462	2,603,130	2,128,768
Standard deviation	164.4812	0.0184	69.2509	0.0515
Pub. year # Institution # Concept 1	y	y	n	n
App. year # Assignee # CPC class	n	n	y	y
Number of FE absorbed	1,431,625	1,153,673	300,535	249,434

Note: estimates from STATA PPMLHDFE (Models 9 and 11) and REGHDFE (Models 10 and 12) routines.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S7: Regressions for citation or novelty by PQR concurrent year.

	Sum papers' cites (13)	Mean papers' novelty (14)	Sum patents' cites (15)	Average patents' novelty (16)
concurrent_year	0.1177***	0.0003***	0.1537***	0.0008***
	(0.0042)	(0.0000)	(0.0063)	(0.0001)
log_nb_collaborators	0.6587***	-0.0002***	0.5586***	-0.0013***
	(0.0021)	(0.0000)	(0.0041)	(0.0001)
R-squared	x	0.7489	x	0.6327
Pseudo R2	0.7218	x	0.7465	x
Number of observations	2,411,481	2,326,286	645,689	618,556
Standard deviation	1,089.249	0.0177	160.3377	0.0541
Scientist id	y	y	n	n
Pub. year # Age count # Concept 1	y	y	n	n
Inventor id	n	n	y	y
App. year # Age count # CPC class	n	n	y	y
Number of FE absorbed	246,384 + 81,245	243,030 + 80,610	122,745 + 37,698	119,580 + 35,900

Note: please see Table S10 for descriptive statistics on each dataset. Estimates from STATA PPMLHDFE (Models 13 and 15) and REGHDFE (Models 14 and 16) routines.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S8: Regressions for the mean age of backward cites by individual paper and patent - 1 year window

	Papers age cites	Patents age cites
	(17)	(18)
has_concurrent_PQR	-0.7150***	-0.3118***
	(0.0043)	(0.0095)
has_allochronic_PQR	-0.2328***	-0.0939***
	(0.0039)	(0.0073)
log_tot_scientists	-0.4836***	
	(0.0027)	
log_tot_inventors		-0.0755***
		(0.0049)
R-squared	0.3132	0.5546
Number of observations	8,816,961	2,352,789
Standard deviation	4.7456	4.7505
Pub. year # Institution # Concept 1	y	n
App. year # Assignee # CPC class	n	y
Number of FE absorbed	1,143,590	279,880

Note: estimates from STATA REGHDFE routines. Note that results hold when dependent variable data are restricted to age of citations greater than 0 and greater than -5 years. While small negative values may be accurate, large negative values are probably an error and hence motivate the robustness check.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S9: Description of the regression data for citation and novelty by individual paper or patent.

	Number of observations	Mean	Standard deviation	Min	Max
	Regression (1): paper's cites				
Paper citations	12,762,702	35.95872	167.5695	0	144,155
has_concurrent_PQRs	12,762,702	0.0853	0.2794	0	1
has_allochronic_PQR	12,762,702	0.2868	0.4523	0	1
log_tot_scientists	12,762,702	0.9670	0.7255	0	8.0721
	Regression (2): paper's novelty				
Paper novelty	10,751,271	-0.0210	0.0182	-0.1225	.0534
has_concurrent_PQRs	10,751,271	0.0931	0.2906	0	1
has_allochronic_PQR	10,751,271	0.3080	0.4617	0	1
log_tot_scientists	10,751,271	1.0316	0.7110	0	8.0721
	Regression (3): patent's cites				
Patent citations	2,603,130	23.9912	69.2509	0	5,308
has_concurrent_PQRs	2,603,130	0.1000	0.3000	0	1
has_allochronic_PQR	2,603,130	0.1850	0.3883	0	1
log_tot_inventors	2,603,130	0.6270	0.6245	0	4.3307
	Regression (4): patent's novelty				
Patent novelty	2,128,768	-0.1769	0.0515	-.3380	0.0673
has_concurrent_PQRs	2,128,768	0.1177	0.3223	0	1
has_allochronic_PQR	2,128,768	0.2138	0.4100	0	1
log_tot_inventors	2,128,768	0.6877	0.6308	0	4.3307

Note: extreme values are not errors. For example, the paper with 144,155 citations is https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.77.3865; the high number of scientists is from the Atlas collaboration https://www.sciencedirect.com/science/article/pii/S0370269312003802; the patent with 5,308 cites is US-4683202; and the patent US-7581231 has 76 ($log(76) = 4.3307$) inventors.

Table S10: Description of the regression data for citation or novelty by PQR concurrent year.

	Number of observations	Mean	Standard deviation	Min	Max
	Regression (13): Sum PQR's papers cites				
Sum papers' cites	2,411,481	281.3552	1089.249	0	221,459
concurrent_year	2,411,481	0.1749	0.3799	0	1
log_nb_collaborators	2,411,481	2.5285	1.1442	0	9.6090
	Regression (14): Mean PQR's papers novelty				
Mean papers' novelty	2,326,286	-0.0224	0.0177	-0.1149	0.0469
concurrent_year	2,326,286	0.1761	0.3809	0	1
log_nb_collaborators	2,326,286	2.5487	1.1367	0	9.6090
	Regression (15): Sum PQR's patents cites				
Sum patents' cites	645,689	50.4702	160.3377	0	13,563
concurrent_year	645,689	0.2532	0.4349	0	1
log_nb_collaborators	645,689	1.6053	0.8706	0	5.5721
	Regression (16): Mean PQR's patents novelty				
Mean patents' novelty	618,556	-0.1510	0.0541	-0.3213	0.0673
concurrent_year	618,556	0.2578	0.4374	0	1
log_nb_collaborators	618,556	1.6305	0.8702	0	5.5721

Note: extreme values are not errors.

-

Table S11: Description of the regression data for the age of cites by individual paper or patent. Note that results hold when dependent variable data are restricted to age of citations greater than 0 and greater than -5 years. While small negative values may be accurate, large negative values are probably an error and hence motivate the robustness check.

	Number of observations	Mean	Standard deviation	Min	Max
	Regression (17): papers age cites				
Paper's age cites	8,816,961	8.0129	4.7456	-36.3233	47.0274
has_concurrent_PQRs	8,816,961	0.2034	0.4025	0	1
has_allochronic_PQR	8,816,961	0.2349	0.4240	0	1
log_tot_scientists	8,816,961	1.0635	0.6931	0	8.0721
	Regression (18): patents age cites				
Patent's age cites	2,352,789	8.8047	4.7505	-36.4767	40.7726
has_concurrent_PQRs	2,352,789	0.1005	0.3007	0	1
has_allochronic_PQR	2,352,789	0.1873	0.3901	0	1
log_tot_inventors	2,352,789	1.6481	0.6284	0	4.3307

Table S12: Scientist id - work id from OA

Variable name	Data provenance	Description
author_id	OA	scientist id from OA. Note that OA scientist ids may change at the next update (these scientist ids correspond to the update from August 2023). This table is provided to allow users to link works and individuals in our dataset to new OA scientist ids.
work_id	OA	work id from OA.

Table S13: Inventor id - Patent number from PV

Variable name	Data provenance	Description
inventor_id	PV	inventor id from PV. Note that PV inventor ids may change at each PV update (these inventor ids correspond to the update from June 2023). This table is provided to allow users to link patents and individuals in our dataset to new PV inventor ids.
patent_id	PV	patent number from PV.

References

[1] Mariëtte Du Plessis et al. “Data production methods for harmonized patent statistics: Patentee sector allocation 2009”. In: *Eurostat Working Paper-Annual Report 2009* (2010).

[2] Matt Marx and Aaron Fuegi. “Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations”. In: *Journal of Economics & Management Strategy* 31.2 (2022), pp. 369–392.

Data S1. (GitHub code repository) All code and figures can be found in the github repository:
https://github.com/EmmaScharfmann/scientists-inventors. The excel files filled out for manual and automated validation against Google Scholar can also be downloaded from this GitHub account.

The code is very extensive and documented and intended to be as transparent as possible; given its size and complexity, however, and despite our very best efforts, it may contain errors, for which we sincerely apologize, and hope that the errors will be documented and corrected by the community. We have run robustness checks and multiple analyses for all claims, such that substantive inferences should not change.

Data S2. (PQR dataset) The PQRs dataset can be downloaded from zenodo: `https://zenodo.org/records/11374125?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImJlMjFkNzk4LTRmZjYtNDhkMC1hYzM4LTU4NWRlNmM0NmY1MyIsImRhdGEiOnt9LCJyYW5kb20iOiJmYjNhODRmMzg4ZjEwYmY1MTBhZmVhYWFhZDY3YjZhYyJ9.Vy1-qRo63hmeJW9n9c0F5sEWfjWI2m7IMrYoNAh5AE9vuE7GSayFHKmEhIEXZcik6SxOZF7rvXq5Iyaz2hP2vQ`

Note that OA scientist ids or PV inventor ids are susceptible of changing when the databases are updated, therefore, OA works corresponding to each scientist id and PV patent number corresponding to each inventor id are provided to allow the user to link new scientist and inventor ids to the PQR dataset (described in appendix tables).