Scientometrics
https://doi.org/10.1007/s11192-022-04587-0

Do academic inventors have diverse interests?

Shuo Xu[1] · Ling Li[1] · Xin An[2]

Received: 6 July 2021 / Accepted: 7 November 2022

Abstract
Academic inventors bridge science and technology, and have attracted increasing attention. However, little is known about whether they have more diverse research interests than researchers with a single role, and whether their important position for science–technology interactions correlates with their diverse interests. For this purpose, we describe a rule-based approach for matching and identifying academic inventors, and an author interest discovery model with credit allocation schemes is utilized to measure the diversity of each researcher's interests. Finally, extensive empirical results on the DrugBank dataset provide several valuable insights. Contrary to our intuitive expectation, the research interests of academic inventors are the least diverse, while those of authors are the most. In addition, the important position of the researchers has a certain relation with the diversity of research interests. More specifically, the degree of centrality has a significant positive correlation with the diversity of interests, and the constraint presents a significant negative correlation. A significant weaker negative correlation can also be observed between the diversity of research interests of academic inventors and their closeness centrality. The normalized betweenness centrality seems be independent from interest diversity. These conclusions help understand the mechanisms of the important position of academic inventors for science–technology interactions, from the perspective of research interests.

Keywords Academic inventors · Author interest discovery · Science–technology linkage · Interest diversity

✉ Xin An
anxin@bjfu.edu.cn

Shuo Xu
xushuo@bjut.edu.cn

Ling Li
infinitell@emails.bjut.edu.cn

[1] College of Economics and Management, Beijing University of Technology, Beijing 100124, People's Republic of China

[2] School of Economics and Management, Beijing Forestry University, Beijing 100083, People's Republic of China

Published online: 16 January 2023

Introduction

Science–technology linkages have received considerable attention over recent decades, due to increasing recognition of the fundamental role of knowledge and innovation in fostering economic growth, technological performance, and international competitiveness (Arrow, 1962; Nelson, 1959; Van Looy et al., 2006). In the literature, scientific publications and patents usually act as the respective proxies of scientific research and technical development (Dubaric et al., 2011; Xu et al., 2012, 2019, 2021c). To understand the knowledge association and interaction mechanism between science and technology, the following three perspectives have been exploited: (1) mutual citations between scholarly articles and patents (Glänzel & Meyer, 2003; Huang et al., 2015; Narin & Noma, 1985); (2) lexical- and topic-based linkages between these two resources (Bassecouolard & Zitt, 2004; Shibata et al., 2011; Xu et al., 2012, 2019, 2021c); and (3) academic inventors bridging science and technology (Guan & Wang, 2010; Li et al., 2020; Meyer, 2006; Noyons et al., 1994; Zhang et al., 2019).

Historically, Narin and Noma (1985) pioneered the linkages between scientific publications and patents by analyzing nonpatent references (NPRs) on the front pages of patent documents. Meyer (2000) also did a lot of work on the basis of NPRs, most of which focused on the field of nanotechnology. Apart from citations of patents to scholarly articles, Glänzel and Meyer (2003) explored the citations of patents in scientific publications, and Huang et al. (2015) exploited two-way citations between papers and patents. However, only about 30–40% of patent documents contain NPRs (Callaert et al., 2006), and chemistry-related research dominates the citations from academic articles to patents (Glänzel & Meyer, 2003).

As for lexical- and topic-based linkages, a popular pipeline research framework (Ba & Liang, 2021; Shibata et al., 2010; Xu et al., 2012, 2019, 2020) is to extract respective thematic structures from scholarly articles and patents, to calculate the similarities between them, and then to construct topic linkages. However, the performance of such a framework is inadequate (Shibata et al., 2010; Xu et al., 2012), since noncomparable themes with different distributions are generated from scientific publications and patents (Xu et al., 2019). This makes it difficult to link the uncovered themes only according to calculated similarities. Although a joint research framework has been developed by Xu et al. (2021c) on the basis of topic models for multiple collections of documents, the lexical- and topic-based linkages often require advanced text mining and machine learning techniques.

Academic inventors are known to author scientific publications and patent inventions simultaneously. In other words, these researchers have two roles: authorship and inventorship. The relationship between their publishing and patenting activities has been investigated in the literature, and both activities are found to be rather complementary than substitutional (Azoulay et al., 2009; Stephan et al., 2007; Thursby et al., 2007). Recent studies have even observed a U-inverted shape pattern (Crespi et al., 2011; Kang et al., 2020), so that beyond a certain level of commercial engagement, patenting starts being a substitute for publishing. Compared with their noninventing/nonpublishing peers, academic inventors tend to outperform in terms of publication/patent counts, citation frequency, and h-index (Guan & Wang, 2010; Meyer, 2006; Van Looy et al., 2006).

With the development of social network theories and methods, several studies have mapped researchers to the interconnection of nodes in the network by their coauthoring and coinventing behaviors. The node position importance (Balconi et al., 2004; Zamzami et al., 2015; Zhang et al., 2019) and key role as gatekeepers (Breschi & Catalini, 2010; Li et al., 2020; Lissoni, 2010) of academic inventors in scientific and technological (S and T) networks have also been investigated. However, little is still unknown on the characteristics of academic inventors, especially their research interests, and the relation between their interests and their position in S&T networks. For this purpose, we have identified the following open questions:

- Do academic inventors have more diverse research interests than those with a single role?
- Does the position of academic inventors in S&T networks correlate with their diverse research interests?

This article is arranged as follows: after the literature review is briefly introduced, our research framework and methodology are put forward. Then, several core modules, such as the identification of academic inventors, interest discovery models, and diversity indicators, are described in more detail. Finally, extensive experiments are conducted on the DrugBank dataset to obtain several valuable insights about diversity of research interests and the relation between interest diversity and position characteristics.

Related work

Before delving into more specifies, the literature pertinent to academic inventors, interest discovery models, and concepts and measurements of diversity is discussed.

Academic inventors

The linkages between science and technology are attracting increasing attention. To exploit these interactions, one research stream of the science–technology linkages mainly focuses on researchers active in both academia and industry, and a number of different terminologies have been used, such as "inventor–author" (Boyack & Klavans, 2008; Noyons et al., 1994; Zhang et al., 2019), "author–inventor" (Wang & Guan, 2011), "patenting–publishing scientist" (Breschi & Catalini, 2010), and "academic inventor" (Balconi et al., 2004; Forti et al., 2013; Lissoni, 2010). Here, academic inventor is used to collectively refer to this type of researchers.

For identifying this kind of researchers, the following strategies have been utilized by previous studies: (1) Czarnitzki et al. (2016) observed that the title "Prof. Dr." was usually taken as a name affix in German, so they searched this title in the inventor field; (2) when a list of staff in universities and research institutes is available, each individual in this list can be linked with the resulting inventors in patent documents (Azoulay et al., 2009; Carayol & Carpentier, 2021; Ejermo & Toivanen, 2018; Hvide & Jones, 2018); (3) the authors of scientific publications can be directly linked with the inventors in patent documents (Boyack & Klavans, 2008; Breschi & Catalini, 2010; Forti et al., 2013; Lissoni, 2010; Maraut & Martínez, 2014; Noyons et al., 1994; Wang & Guan, 2011; Zhang et al., 2019). The first two strategies mainly focus on the employees with patenting activity in universities and institutes. The latter reduces this limitation, which enables it to encompass professors, researchers, and engineers with both publishing and patenting activities. Strictly speaking, this study follows the definition in the latter, but our academic inventors (see *Dataset*) all happen to affiliate with at least one academic institution.

Whichever strategy is adopted for academic inventors, the names of authors and inventors first need to be disambiguated. Many different solutions have been put forth in the literature for author name ambiguity (Caron & van Eck, 2014; Han et al., 2017; Kim, 2018; Torvik & Smalheiser, 2009; Xu et al., 2021b). Most of them follow a two-step process: feature extraction and clustering/classifying. Similarly, inventor name disambiguation is an important issue for patent data. Correspondingly, many approaches have been proposed to disambiguate inventor names (Li et al., 2014; Pezzoni et al., 2014; Raffo & Lhuillery, 2009; Yang et al., 2017). Different from disambiguating author names, most of these are three-step processes: parsing, matching, and filtering, such as the Massacrator© algorithm (Lissoni et al., 2006; Pezzoni et al., 2014).

After disambiguating the authors and inventors, an automatic method can be used to match and identify the academic inventors by text content similarity (Cassiman et al., 2007) or string matching (Boyack & Klavans, 2008). Another common way is to compare the list of disambiguated authors and inventors in a semi-automatic way (Breschi & Catalini, 2010; Li et al., 2020; Lissoni, 2010; Wang & Guan, 2011). More specifically, after given and family names of each author and inventor are normalized, desktop research is conducted, including manually checking and leveraging extra information from other databases, the internet, or questionnaires. Although this kind of conservative approach is time consuming, the reliability of linkage results is convincing.

Once the identification of the academic inventors has been made, many studies have exploited whether there is a balance between patenting and publishing activities. Several empirical investigations found increased patenting activities may undermine the performance of basic research (Agrawal & Henderson, 2002; Blumenthal et al., 1997; Fabrizio & Di Minin, 2008). Another stream of empirical investigations showed that the patenting activities were positively related to the number and quality of publications (Azoulay et al., 2007; Grimm & Jaenicke, 2015; Van Looy et al., 2006). Recent studies suggested that there was a curvilinear (inverted-U) correlation between patenting and publishing activities (Crespi et al., 2011; Lee, 2019; Kang et al., 2020). In more detail, an increase in patenting activity initially promotes the number and quality of publications up to a peak, and after this peak, it lowers the number and quality of publications.

Some studies have also focused on the special role of academic inventors, who bridge science and technology, and the network structure and position characteristics have been widely measured. In the coauthorship and coinventorship network, academic inventors have more central and better connected positions (Balconi et al., 2004; Forti et al., 2013; Zamzami & Schiffauerova, 2015). It is highly likely that these significant network characteristics should be attributed to their role as the gatekeepers between science and technology (Breschi & Catalini, 2010; Lissoni, 2010). Zhang et al. (2019) and Li et al. (2020) found that academic inventors promote the knowledge transfer between science and technology. In addition, academic inventors also play an important role in entrepreneurial firm development (Murray, 2004), breakthrough scientific research (Winnink & Tijssen, 2014), and technological innovation processes (Quatraro & Scandura, 2019).

Interest discovery models

Every researcher has their own research interests, which can be readily obtained from the curriculum vitae (CV) of the focal researcher. However, since these may not be regularly updated and many CVs are not available from the internet, several data-driven topic models for discovering interests from their research outputs are proposed in the literature.

One popular model is the Author–Topic (AT) model (Rosen-Zvi et al., 2010), which integrates author information into the standard Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). Several variants have since been proposed, such as the Author–Persona–Topic (APT) model (Mimno & McCallum, 2007), the Author–Interest–Topic (AIT) model (Kawamae, 2010), and the Author–Topic over Time (AToT) model (Shi et al., 2013; Xu et al., 2014a, b), and so on. In these models, each research output is modeled as if it is generated by a two-stage stochastic process. A researcher's interests are represented by a multinomial distribution over topics, and each topic is represented as a multinomial distribution over words. The probability distribution over topics in a multi-author paper or multi-inventor patent is a mixture of the distributions associated with their authors or inventors.

All these models are actually members of generative probabilistic topic models for uncovering main themes from a collection of documents (Blei, 2012). Hence, each model can be viewed as a generative process. For example, in the AT model (Rosen-Zvi et al., 2010), to generate each word in a document, a researcher index is uniformly drawn from its author/inventor list. Then, a topic index is drawn from their multinomial distribution over topics (viz. research interests). Finally, a word token is drawn from the multinomial distribution of that topic.

From the generative process above, it is not difficult to see that these models share the following same assumption: the author/inventor list of a document is uniformly distributed. Currently, the knowledge for addressing these issues is more diverse and specialized (Leahey, 2016), and increasing cooperation in science and technology is a general trend (Adams et al., 2005; Wuchty et al., 2007). It is obviously inappropriate to implicitly assume that each coauthor/coinventor contributes equally to a target document. Therefore, the $\mathrm{AT}^{\mathrm{credit}}$ model (Xu et al., 2021a, 2022), which powers the AT model's abilities with the credit allocation schema, is adopted in this work.

Diversity: concept and measurement

In real-world scenarios, many instances of diversity can be observed, such as diverse ecological species, diverse crystal structures, and diverse disciplines in the science and technology field. Stirling (2007) argued that diversity is a characteristic of any system whose elements could be apportioned into categories. Further, three basic properties, "variety," "balance," and "disparity" were proposed (2007), each of which is a necessary and insufficient property for diversity.

- Variety is the number of categories to which the elements in a focal system are assigned. It can be quantified as an integer (enumerating categories). When all else is equal, the greater the variety, the greater the diversity.
- Balance is a function of the pattern of assignment of elements across categories. It can be quantified as a vector of fractions summing to unity (apportioning elements). When all else is equal, the more even the balance, the greater the diversity.
- Disparity is the degree to which categories in a focal system are different from each other. It can be quantified as a matrix of distances (differentiating elements). When all else is equal, the more disparate the disparity, the greater the diversity.

To measure the diversity of an interested system, Rao (1982) and Stirling (2007) presented a general quantitative nonparametric heuristic indicator, *Rao-Stirling*. To the best

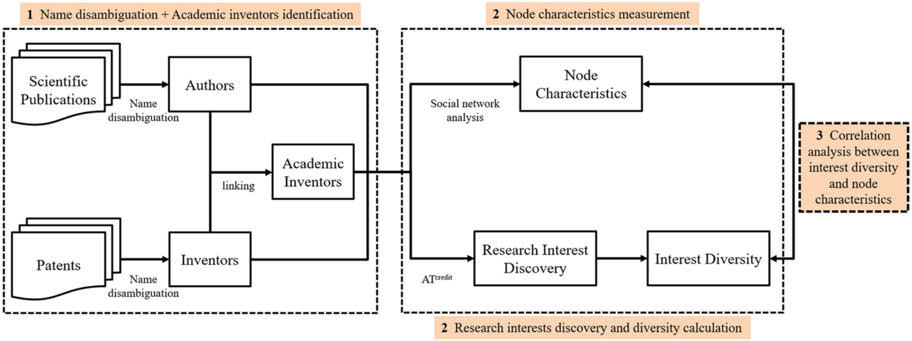

1 Name disambiguation + Academic inventors identification
Scientific Publications
Authors
Name disambiguation
linking
Academic Inventors
Patents
Inventors
Name disambiguation
2 Node characteristics measurement
Social network analysis
Node Characteristics
3 Correlation analysis between interest diversity and node characteristics
Research Interest Discovery
Interest Diversity
AT^credit
2 Research interests discovery and diversity calculation

Fig. 1 Research framework for measuring diversity of interests of the academic inventors

of our knowledge, this was the first systematic and transparent approach in the treatment of scientific and technological diversity in a broad range of fields. Since then, many alternatives have been proposed in the literature. Generally speaking, these indicators can be divided into three groups according to the basic properties above: (a) measures sensitive to balance, (b) measures sensitive to balance and disparity, and (c) measures sensitive to variety, balance, and disparity.

The measures sensitive to balance mainly focus on the distribution of different categories of elements in the system, such as Shannon entropy (Shannon, 1950) and Simpson diversity (Simpson, 1949). This type of indicator makes the following implicit assumption: the categories of system elements are completely different from each other. Obviously, this is not in line with many real-world scenarios. The measures sensitive to balance and disparity consider the balance and disparity of system elements at the same time. Two instances of this type of measures are the Rao–Stirling (Rao, 1982; Stirling, 2007) and $^{2}D^{S}$ (Zhang et al., 2016), which are closely related. The measures sensitive to variety, balance, and disparity, as their names imply, simultaneously operationalize three basic properties. The DIV (Leydesdorff et al., 2019) is one such indicator, and the superiority of the DIV indicator has been validated by Bu et al. (2020). Hence, the Rao–Stirling and DIV are both utilized here to calculate the diversity of interests of academic inventors and their peers.

Research framework and methodology

To answer the research questions in the *Introduction*, our research framework consists of three phases, as shown in Fig. 1. After disambiguating the names of authors and inventors, and linking and identifying the academic inventors in the first phase, the second phase measures the node characteristics of authors, inventors, and academic inventors with the help of a social network analysis. In this phase, the research interests of each researcher are also discovered by the AT$^{\text{credit}}$ model (Xu et al., 2021a, 2022), and then the diversity of each researcher's interests is measured by the Rao–Stirling and DIV indicators. Finally, we analyze the correlation between the node characteristics and the interest diversity in the last phase. In the following subsections, several core modules will be described in more detail.

Table 1 Examples of whether an author and an inventor should be linked

	First name	Last name	Linkage
1	Alon	Harris	✗
	Alan	Harris	
2	A S	Douglas	✗
	Alan W	Douglas	
3	Kenji	Ohmori	✓
	Kenji	Ohmori	
4	K	Seibert	✓
	Karen	Seibert	
5	Derek	Norris	✓
	Derek J	Norris	
6	K	Kimura	✓
	K	Kimura	

For each pair, the first line is the author from a paper, and the second line is the inventor from a patent

Identifying academic inventors

To identify academic inventors, the names of authors and inventors must first be disambiguated. To the best of our knowledge, the authors in several bibliographic databases (such as Web of Science and Scopus) are neigher fully unambiguously identified, nor are the inventors in the intellectual property databases [such as United States Patent and Trademark Office (USPTO) and European Patent Office (EPO)]. Hence, a revised rule-based scoring and clustering method (Xu et al., 2021b) is utilized here for disambiguating the authors. As for the inventors, we adopted a semi-automatic method. More specifically, after the first and last names of each inventor were split and checked, the inventors were disambiguated by several manually curated rules on the basis of the applicants, co-inventors, address, theme of the resulting patent, and so on.

To identify academic inventors, we first excluded the nonindividual entities in the inventor field, such as research team (laboratory/group/international organizations/institute), company (Co./Corp./LLC/PLC/AG/GmbH), university (Univ.), hospital, etc. Then, we matched the last name and initials of each pair of author and inventor. This step can group the paired researchers as follows: (a) inconsistent pairs are filtered out, such as pairs 1 and 2 in Table 1b, while the consistent pairs are linked directly to an academic inventor, such as pair 3 in Table 1c. Ambiguous pairs, such as pairs 4, 5, and 6 in Table 1, are manually checked for whether the following factors overlap, including the authors' affiliation and assignee, coauthors and coinventors, themes from the resulting publication and patent, and so on. For example, pairs 4 and 5 in Table 1 share the same research institutions and coauthor information, so we identify them as academic inventors. No evidence can be found to support pair 6 in Table 1 as the same individual, so they cannot be linked.

Interest discovery model

To discover the research interests of researchers objectively and accurately, this work adopts the AT$^{\text{credit}}$ model (Xu et al., 2021a, 2022) using the author's credit allocation. This model is a generalization of the AT model by introducing a set of hidden random variables

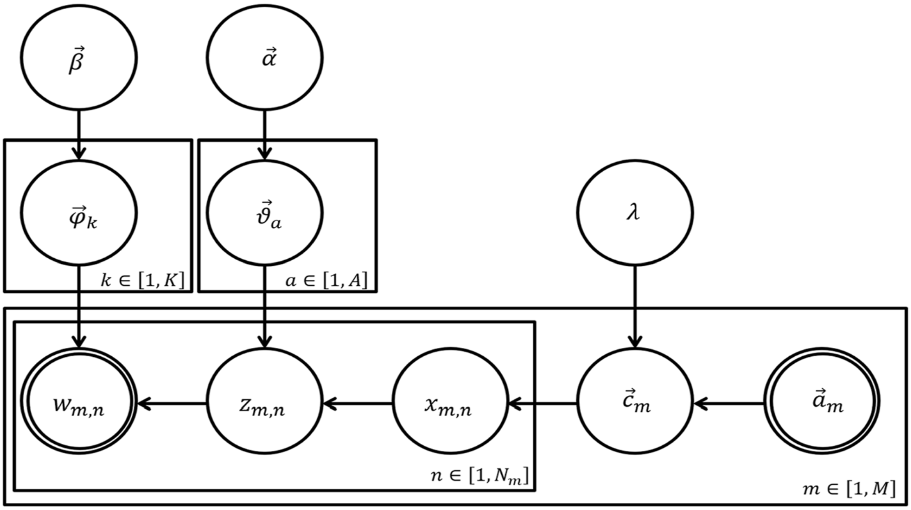


Fig. 2 Graphical representation of the AT$^{\text{credit}}$ model

$\{\vec{c}_m\}$. When the indiscriminate counting scheme is adopted, it degenerates into the AT model (Rosen-Zvi et al., 2010). Though many credit allocation schemes have been put forward in the literature, this study prefers to use the sequence-determines-credit (SDC) schema (Tscharntke et al., 2007) because (a) the SDC schema takes "hyper-authorship" (more than ten coauthors or coinventors) into consideration and (b) this scheme effectively combines the advantages of the harmonic counting scheme (Hagen, 2013) and indiscriminate counting scheme.

The graphical model representation of the AT$^{\text{credit}}$ model is illustrated in Fig. 2. Here, K, M and A represent the number of topics, documents, and unique authors/inventors, respectively. $\vec{\varphi}_k$ and $\vec{\vartheta}_a$ denote respective multinomial distribution of words specific to the topic k and of topics specific to the author/inventor a. $\vec{\alpha}$, and $\vec{\beta}$ are the Dirichlet hyperparameter. The byline information of the document m is encoded in the variable $\vec{a}_m$, and $\vec{c}_m$ assigns the authorship credit to each coauthor/coinventor in the document m according to a specified schema with the parameter λ. In addition, $z_{m,n}$ and $x_{m,n}$ are the topic and author/inventor assignment associated with the n-th word token $w_{m,n}$ in the document m.

The model can also be described from the viewpoint of generative process as follows. After $\vec{\varphi}_k$ ($k \in [1, K]$) and $\vec{\vartheta}_a$ ($a \in [1, A]$) are drawn respectively from the Dirichlet ($\vec{\beta}$) and Dirichlet ($\vec{\alpha}$), the authorship credits are calculated for each document $m \in [1, \mathrm{M}]$ by following a designated authorship credit allocation schema with a parameter λ. Finally, for each document $m \in [1, \mathrm{M}]$, and each word token $\mathrm{n} \in \left[1, N_m\right]$ in the document m, $x_{m,n}$ is drawn from $\vec{c}_m$, $z_{m,n}$ from $\vec{\vartheta}_{x_{m,n}}$, and then $w_{m,n}$ from $\vec{\varphi}_{z_{m,n}}$. As for many Bayesian models, posterior inference cannot be done exactly in this model. The collapsed Gibbs sampling algorithm was originally utilized in Xu et al. (2021a) and Xu et al. (2022) to approximate the posterior of the AT$^{\text{credit}}$ model. Please refer to Xu et al. (2021a) and Xu et al. (2022) for more detail. In this work, symmetric Dirichlet priors α and β are set at 0.5 and 0.01, respectively. The collapsed Gibbs sampling is run for 2000 iterations, including 500 for the burn-in period.

Table 2 Several formulas for operationalizing the disparity

Disparity	Formula
Symmetrized KL divergence	$symKL(\overrightarrow{\varphi}_i \parallel \overrightarrow{\varphi}_j) = \frac{1}{2}\left[KL(\overrightarrow{\varphi}_i \parallel \overrightarrow{\varphi}_j) + KL(\overrightarrow{\varphi}_j \parallel \overrightarrow{\varphi}_i)\right]$ where $KL(\overrightarrow{\varphi}_i \parallel \overrightarrow{\varphi}_j) = \sum_v \varphi_{i,v} \log \frac{\varphi_{i,v}}{\varphi_{j,v}}$
JS divergence	$JS(\overrightarrow{\varphi}_i, \overrightarrow{\varphi}_j) = \frac{1}{2}\left[KL(\overrightarrow{\varphi}_i \parallel \overrightarrow{\varphi}) + KL(\overrightarrow{\varphi}_j \parallel \overrightarrow{\varphi})\right]$ where $\overrightarrow{\varphi} = \frac{1}{2}(\overrightarrow{\varphi}_i + \overrightarrow{\varphi}_j)$
Cosine distance	$1 - cos(\overrightarrow{\varphi}_i, \overrightarrow{\varphi}_j) = 1 - \frac{\overrightarrow{\varphi}_i \bullet \overrightarrow{\varphi}_j}{\lVert\overrightarrow{\varphi}_i\rVert \times \lVert\overrightarrow{\varphi}_j\rVert}$

Diversity indicators

Since the Rao–Stirling (Rao, 1982; Stirling, 2007) and *DIV* (Leydesdorff et al., 2019) can simultaneously consider at least two basic properties in a system, these measures were adopted to measure the diversity of interests of each researcher in this work. A larger value of these two measures indicates more diverse interests. They can be defined formally as follows:

$$RS = \sum_{i,j(i \neq j)} \vartheta_{a,i} \times \vartheta_{a,j} \times d_{ij} \tag{1}$$

$$DIV = \left(\frac{n_{a,k}}{K}\right) \times (1 - Gini) \times \sum_{i,j(i \neq j)}^{n_{a,k}} \frac{d_{ij}}{n_{a,k}\left(n_{a,k} - 1\right)} \tag{2}$$

Here, $i \in [1, K]$ and $j \in [1, K]$ represent any two different topics. $\vartheta_{a,i}$ and $\vartheta_{a,j}$ are the probability of the topic i and j specific to the author/inventor a, respectively. $n_{a,k}$ denotes the number of themes that the author/inventor a prefers. $(1 - Gini)$ indicates the balance of the diversity, and d_{ij} is the degree of difference (i.e., disparity of the diversity) between the theme i and j. In our work, the disparity d_{ij} between the theme i and j is operationalized with symmetrized Kullback–Leibler (KL) divergence, Jensen–Shannon (JS) divergence, and cosine distance. See Table 2 for more detail on how to operationalize the disparity.

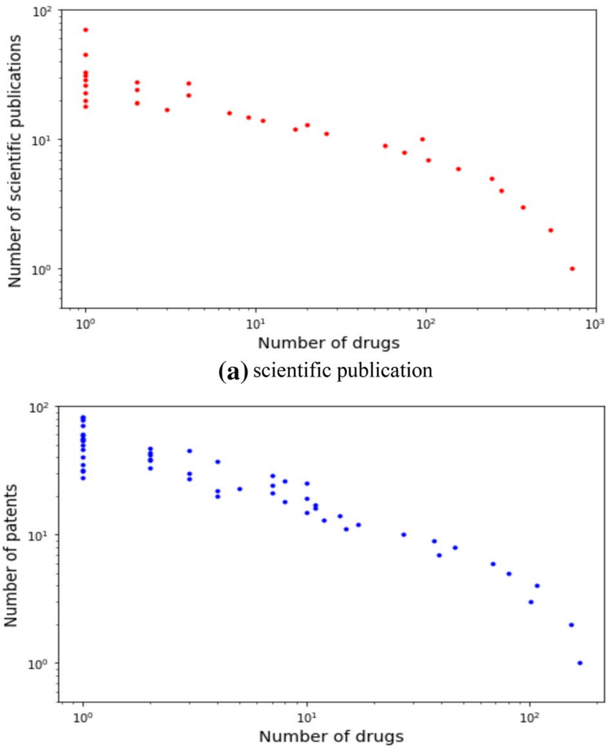
Empirical results and discussions

Dataset

It is well known that the research and development (R&D) procedure of a novel drug often involves rich scientific knowledge, intellectual property protection, and reliable clinical trials. Thus, science–technology interactions in the pharmaceutical field are prominent (Glänzel & Meyer, 2003). Hence, this work used the DrugBank[1] database (version: 1 November 2019) as our dataset, which is the world's largest online database of drug and drug-target information. The DrugBank database is a free-to-access resource for academic users. Each

[1] https://go.drugbank.com/.

Fig. 3 The number of scientific publications **a** and patents **b** (*y* axis) linked to drugs (*x* axis). Both axes are shown on a log scale. The power-law-like distribution is evident from the near linear pattern (in log space)



(a) scientific publication

(b) patents

drug in this database may be issued multiple patents and has an attached list of scholarly articles, which provides us an opportunity for further science–technology linkage research.

The DrugBank dataset was downloaded on 1 November 2019 in XML, and parsed to the MySQL database. There are 13,339 unique drugs, 10,355 unique scientific publications, and 5,932 granted patents in this dataset. Although the drugs can act as a bridge to link scientific publications and patents, not all drugs have explicitly attached scientific or technological knowledge. In this dataset, 2768 drugs have linked with 10,836 scientific publications, 1026 drugs linked with 7880 granted patents, and 804 drugs with 3713 scholarly articles and 6535 granted patents simultaneously. Note that this study does not merge the closely related patents derived from the same core technology but issued by different authorities into a patent family.

Figure 3 illustrates the linkage relations between drugs and scientific publications (a), and drugs and patents (b). More specifically, the number of unique articles and patents (*y* axis) that have linked to *k* drugs are plotted as a function of *k* [*x*-axis]. A power-law like distribution of the number of scholarly articles and patents can be noted from Fig. 4. That is to say, the vast majority of drugs are linked to few articles or patents, but several drugs are associated with a large of articles or patents. For example, Imidacloprid, a neonicotinoid insecticide, links up to 71 scientific publications, and Metformin, an oral blood glucose-lowering drug, first approved in Canada in 1972 followed by 1995 in the USA, associates with 83 patents. The number of unique academic articles and patent documents attached to drugs are 10,257 and 5930, respectively.

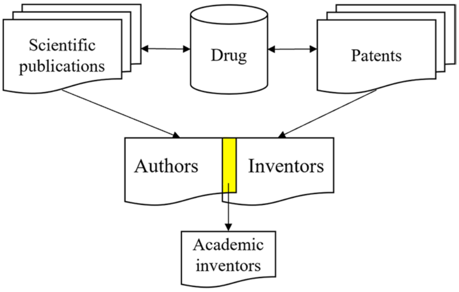


Fig. 4 Procedure of how to determine the academic inventors

Figure 4 intuitively illustrates the procedure for determining the academic inventors in this study. This just considers scientific publications and patents directly attached to each drug, and the academic inventors are limited to the intersection between the authors in scientific publications and the inventors in patents. It is worth mentioning that the authors in scientific publications could be patenting beyond the DrugBank dataset, and the inventors in patents could be publishing beyond this dataset. This study does not take these situations into consideration.[2] Hence, the number of academic inventors in this study may be underestimated, and the results from this study should be interpreted with some caution.

Given that the DrugBank dataset only records the patent numbers and PubMed Unique Identifier (PMIDs), we further collected the title, abstract, inventor and author list, and other related information for each granted patent and scientific publication from the EPO database with OPS API[3] and PubMed database with E-Fetch API.[4] To identify academic inventors, the name of each author and inventor was disambiguated with a revised rule-based scoring and clustering method (Xu et al., 2021b) and checked manually. After this operation, we obtained 43,087 unique authors and 8738 unique inventors. Then, by following the procedure in *Identifying academic inventors*, we finally obtained 805 unique academic inventors. That is, academic inventors account for 1.9% and 9.2% of authors and inventors, respectively.

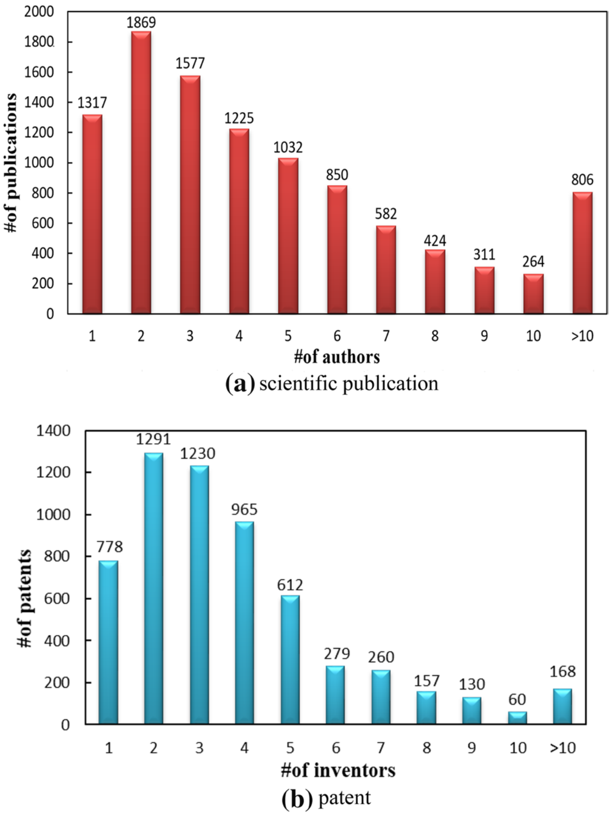
These figures are lower than those observed in previous studies, such as Breschi and Catalini (2010) and Carayol and Carpentier (2021). Since this study does not consider the scholarly articles and patents beyond the DrugBank dataset, it may result in an underestimation of academic inventors. In addition, the number of authors on a scientific publication is frequently higher than that of inventors listed in a patent (Breschi & Catalini, 2010; Lissoni & Montobbio, 2008), so the proportion of academic inventors among authors is lower than among inventors. This point can be validated from the distribution of the number of documents with the number of authors and inventors in Fig. 5. The average number of coauthors per article (4.93) is larger than that of coinventors per patent (3.87).

[2] These situations have not not considered in this study because (1) as mentioned in *Identifying academic inventors*, the authors in several bibliographic databases and the inventors in the intellectual property databases are not disambiguated at all. This makes it these situations difficult to overcome when only using the search interface provided by these databases, and (2) to estimate the number of academic inventors beyond the DrugBank dataset, this study randomly draws 1500 solely authors and 1000 solely inventors, and then manually checks the retrieved patents and publications from the EPO and PubMed databases, respectively. Only one researcher (*Anhalt, Grant J.*) was identified as an academic inventor. That is to say, the rate of academic inventors beyond the DrugBank dataset is about 0.04%. Therefore, we argue that classification of individuals into three groups (academic inventors, solely authors, solely inventors) should not affect the main conclusions regarding the topic interests of each group in this study.

[3] http://ops.epo.org/.

[4] https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch.

Fig. 5 The distribution of the number of scientific publications with the number of authors **a** and the number of patents with the number of inventors **b**



(a) scientific publication

(b) patent

Intuitively, the incentive and funding systems in different countries may affect the propensity of each individual to either publish or patent, or both. Hence, this study further collected the country information of each author and inventor. As for the authors affiliated with multiple countries, we only kept the country of the first affiliation of each author in the byline information of the resulting scientific publications. In the end, the authors in our DrugBank dataset come from 132 countries, and the inventors from 35 countries. The authors and inventors in the USA (63.31% versus 55.30%) dominate, followed by the UK (6.64% versus 4.99%), Japan (5.55% versus 13.03%), and Germany (4.54% versus 8.15%).

The pre-processing steps in this study are very similar to those in Xu et al. (2021b, 2021c). The sentences in the titles and abstracts were detected with *geniass* (Sætre et al., 2007), and then the split sentences were tokenized and lemmatized with *geniatagger* (Tsuruoka et al., 2005). To filter stopwords, the English stopword list from Natural Language Toolkit (NLTK) was used to filter stopwords and all numbers were replaced with a special word *NUMBER*. To reduce the interference of unrelated information, copyright information was removed with human-curated rules based on regular expressions.

Descriptive statistics

To make the comparison of the three types of researchers fair, we kept scientific publications and patents with at less one academic inventor for further analysis. That is to say, only

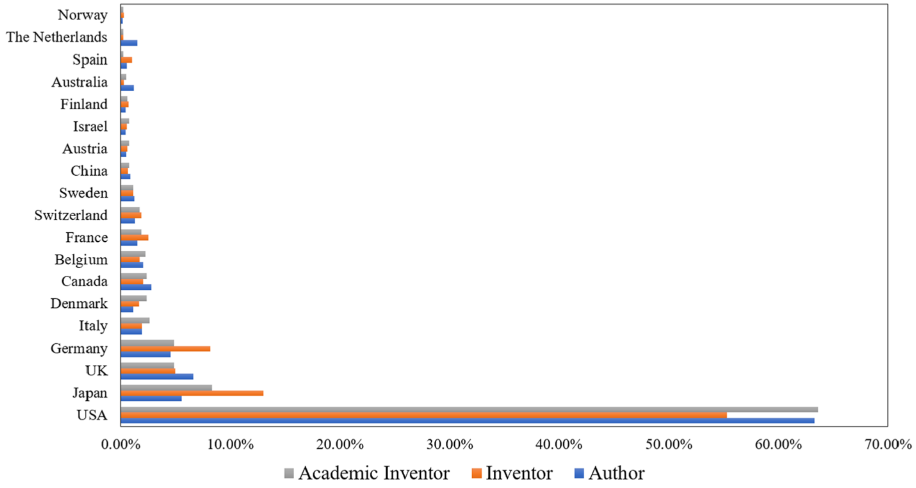


Fig. 6 The country distribution of three types of researchers

academic inventors, and their coauthors and coinventors were included in our final dataset. In the end, this dataset included 603 scientific publications and 1285 patent documents, involving 805 academic inventors, 4088 solely publishing peers and 1803 solely patenting peers. Hereinafter, for convenience, the solely publishing peers and solely patenting ones are specifically referred to as researchers with a single role.

Among the authors, 16.45% of them also applied for patents, and among the inventors, 30.87% of them also published academic articles. This is similar to observations in the fields of nanoscience and fuel cells (Guan & Wang, 2010; Klitkou et al., 2007; Meyer, 2006). Then, we compared the publishing performance of academic inventors with solely publishing peers in terms of the number of articles per author, and their patenting performance with solely patenting peers in terms of the number of patents per inventor. As a whole, most academic inventors are highly productive researchers. In more detail, the academic inventors are superior to their solely publishing peers (1.42 > 1.10) and solely patenting peers (3.11 > 2.44). This is consistent with the findings of Guan and Wang (2010), but is different from those of Meyer (2006).

In our dataset, 521 (64.72%) academic inventors participated in the basic and applied research of drugs at the same time. Surprisingly, 24 academic inventors contributed to successful delivery of Ombitasvir, an antiviral medication used as part of combination therapy to treat chronic hepatitis C. Furthermore, they are more inclined to apply for patents than to publish academic papers. The number of articles per author and that of patents per inventor are 1.43 and 3.47, respectively. For example, Soni, Paresh patented 23 inventions about the methods of treating and/or preventing cardiovascular-related disease, and published three articles on the topics of pharmacokinetics and/or clinical application of icosapent ethyl for the treatment of hypertriglyceridemia, which is an important risk factor for cardiovascular-related diseases.

Table 3 Statistics for coauthorship, coinventorship, and hybrid networks

	Coauthorship	Coinventorship	Hybrid
Number of nodes	4893	2608	6696
Number of edges	38,282	10,305	47,536
Number of components	277	337	207
Number of isolates	13	22	0
Nodes in the giant (% of all nodes)	411 (8.40%)	128 (4.91%)	1784 (26.64%)

Further, Fig. 6 illustrates the country distribution of academic inventors, solely publishing peers, and solely patenting peers. Researchers from the USA dominate, followed by Japan, the UK, and Germany. On closer examination, three interesting phenomena can be observed: (1) researchers from Japan and Germany tend to patent rather than publish; (2) researchers from USA and the UK are more inclined to publish articles; and (3) in Italy and Sweden, no significant differences between publishing and patenting were observed. In our opinion, these phenomena may be related to the incentive and funding systems of each country.

Network size and node characteristics

For an overall intuitive understanding of the three types of researchers, we constructed coauthorship, coinventorship, and hybrid networks. Several statistics are reported in Table 3. The number of components and isolated nodes in the hybrid network are less than the sum of those in other two networks. The number of nodes in the giant component of the hybrid network is much more than that in the giant component of the coauthorship or coinventorship network. This shows that the academic inventors can effectively bridge science and technology and connect more authors and inventors with each other, as shown in Fig. 7.

Here, four indicators ("degree centrality," "normalized betweenness centrality," "closeness centrality," and "constraint") in Table 4 are adopted to measure the importance and influence of the resulting researchers in the hybrid network, as presented in Table 5. A higher degree of centrality means that the resulting researcher cooperated with more researchers. If a researcher can bridge more pairs of researchers through the shortest paths, which do not have direct connectivity between them, their normalized betweenness centrality will assume a higher value. Similarly, in a network, if a researcher occupies the more central position with the shortest average distance to other researchers, they will have a higher closeness centrality value. A lower value of constraint implies that the corresponding researcher occupies a less constrained position, thereby brokering more extensively in the network. Table 5 shows that the academic inventors are more sociable, more "in between," more centrality positioned, and more likely to be structural whole than their peers, which is in line with the observation of Breschi and Catalini (2010).

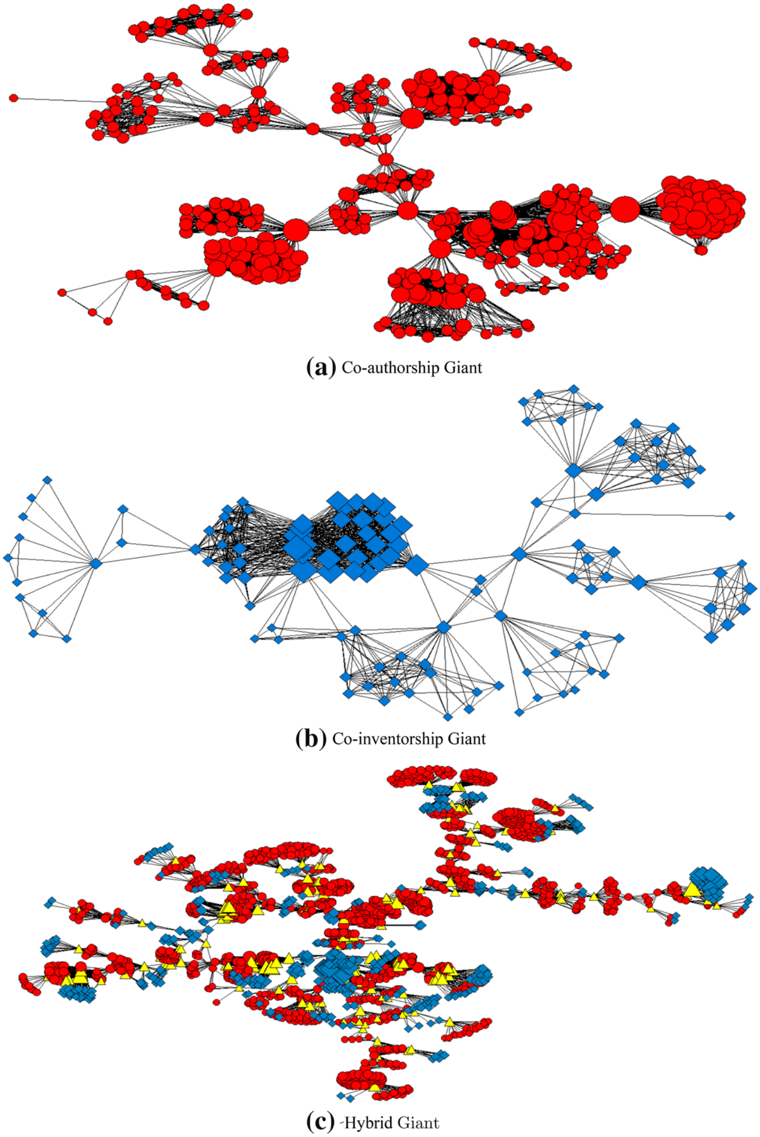


Fig. 7 The giant component of coauthorship **a**, coinventorship **b** and hybrid **c** network. The red circle nodes denote solely publishing authors, blue diamond nodes represent the solely patenting inventors, and yellow triangle nodes are academic inventors. The nodes are sized with their degrees, and the edges are thickened by cooperation strength of the resulting researchers

Research interest discovery

In this subsection, we first identify the number of interest topics, and then answer the question: do academic inventors have more diverse research interests than those with a single role?

Table 4 Four indicators for measuring node characteristics

Indicator	Description	Formula
Degree centrality	Degree centrality for a node v is the fraction of nodes it is connected to	$DC_v = \frac{n_v}{n-1}$
Normalized betweenness centrality	Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v	$NBC_v = \frac{2}{(n-1)(n-2)} \sum_{s \neq v \neq t} \frac{n_{st}^{v}}{g_{st}}$
Closeness centrality	Closeness centrality of a node v is the reciprocal of the average shortest path distance to v over all $n - 1$ reachable nodes	$CC_v = \frac{n-1}{\sum_{u=1}^{n-1} d_{vu}}$
Constraint	Constraint is a measure of the extent to which a node v is invested in those nodes that are themselves invested in the neighbors of v	$C_v = \sum_{u \in N_{(v)} \setminus \{v\}} \left(P_{vu} + \sum_{w \in N_{(u)}} P_{vw} P_{wu} \right)^2$

n is the number of nodes in G; n_v is the number of edges connected to node v; g_{st} and n_{st}^{v} represents the number of shortest paths connecting (s, t) and that pass through v; d_{vu} denotes the shortest path distance to node v; $N_{(v)} \setminus \{v\}$ is the subset of the neighbors of node v that are either predecessors or successors of node v; $N_{(u)}$ is the set of neighbors of node u; P_{vu} is the normalized mutual weight (sum of the weights of edges) of the edges joining v and u

Table 5 Node characteristics of solely publishing authors, solely patenting inventors, and academic inventors

	Authors	Inventors	Academic inventors
Degree centrality	0.2313 (0.1792)	0.1217 (0.0896)	**0.3171** (0.2539)
Normalized betweenness centrality	0.0054 (0.0000)	0.0056 (0.0000)	**0.0512** (0.0000)
Closeness centrality	35.7021 (31.6667)	29.3608 (23.4676)	**39.0719** (34.4828)
Constraint	31.1741 (27.5268)	46.2873 (43.5887)	**23.6752** (19.3102)

The table reports average and median (in parentheses) of each centrality indicator. All values have been increased 100 times

Bold indicates the best results corresponding to each indicator

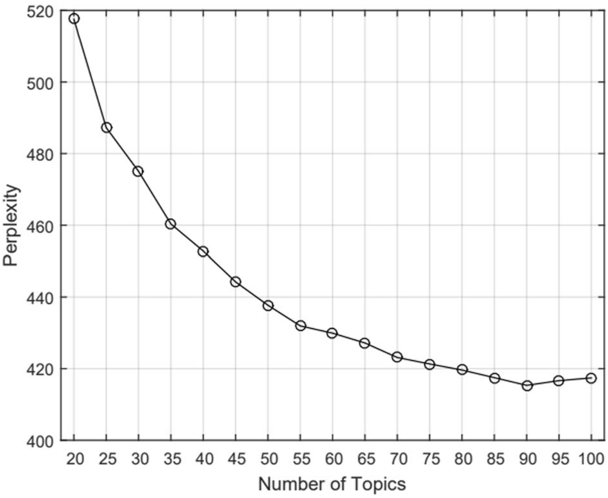


Fig. 8 The perplexity with different number of topics

Number of topics

To identify the proper number of research interest topics, the perplexity is calculated for a different number of topics K. As a standard measure for model selection, this index is defined as the exponential of the negative normalized predictive likelihood of the word observations under the trained model $\mathcal{M}$ [Equations (3 and 4)], and a lower value on the test corpus indicates a better generalization performance.

$$\Pr\left(\vec{\tilde{w}}|\vec{\tilde{a}},\mathcal{M}\right)=\exp-\frac{\sum_{m=1}^{M}\log\Pr\left(\vec{\tilde{w}}_{m,\bullet}|\vec{\tilde{a}}_m,\mathcal{M}\right)}{\sum_{m=1}^{M}N_m} \tag{3}$$

For the AT$^{\text{credit}}$ model, the likelihood of a document of the test corpus $\Pr\left(\vec{\tilde{w}}_{m,\bullet}|\vec{\tilde{a}}_m,\mathcal{M}\right)$ can be directly expressed as a function of the multinomial parameters (*Interest discovery model*) as follows:

Table 6 Diversity of research interests for solely publishing authors, solely patenting inventors, and academic inventors

	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	7.360 (± 0.261)	**7.371 (± 0.446)**	7.100 (± 0.758)
RS (JS divergence)	**0.617 (± 0.020)**	0.611 (± 0.038)	0.589 (± 0.062)
RS (cosine distance)	**0.955 (± 0.033)**	0.946 (± 0.061)	0.910 (± 0.099)
DIV_0.80 (symmetrized KL divergence)	**4.684 (± 1.277)**	4.455 (± 1.464)	2.993 (± 1.614)
DIV_0.80 (JS divergence)	**0.394 (± 0.107)**	0.374 (± 0.123)	0.252 (±0.136)
DIV_0.80 (cosine distance)	**0.613 (± 0.167)**	0.582 (± 0.192)	0.391 (± 0.211)
DIV_0.85 (symmetrized KL divergence)	**5.088 (± 1.263)**	4.862 (± 1.458)	3.389 (± 1.665)
DIV_0.85 (JS divergence)	**0.429 (± 0.106)**	0.409 (± 0.123)	0.285 (± 0.140)
DIV_0.85 (cosine distance)	**0.667 (± 0.166)**	0.636 (± 0.192)	0.443 (± 0.218)
DIV_0.90 (symmetrized KL divergence)	**5.510 (± 1.239)**	5.286 (± 1.440)	3.819 (± 1.687)
DIV_0.90 (JS divergence)	**0.464 (± 0.104)**	0.444 (± 0.121)	0.321 (± 0.142)
DIV_0.90 (cosine distance)	**0.720 (± 0.161)**	0.690 (± 0.188)	0.499 (± 0.220)

Standard deviation is shown in parentheses

Bold indicates the best results corresponding to each indicator

$$\Pr\left(\vec{\tilde{w}}_{m,\bullet} | \vec{\tilde{a}}_m, \mathcal{M}\right) = \prod_{n=1}^{N_m} \sum_{a=1}^{A_m} \sum\nolimits_{k=1}^{K} \varphi_{k,\tilde{w}_{m,n}} \vartheta_{a,k} c_{m,a} \tag{4}$$

Figure 8 depicts the perplexity with different number of topics. From Fig. 8, it is not difficult to see that the perplexity of the AT$^{\text{credit}}$ model converges when the number of topics K is 90. Hence, the number of interest topics was fixed at 90 in this work.

Interest diversity

Two diversity indicators, Rao–Stirling and *DIV*, are utilized here to calculate the diversity of research interests of solely publishing authors, solely patenting inventors, and academic inventors. To determine the number of preferred research interest topics of an author/inventor ($n_{a,k}$) in the *DIV* indicator [Equation (2)], we set three cumulative probability thresholds of interest topics (0.80, 0.85, 0.90). Whichever value the cumulative probability threshold takes, the following conclusion can be drawn: solely publishing authors have the most diverse research interests, followed by the solely patenting inventors, and the research interests of academic inventors are least diverse. Please check Table 6 for more detail.

This observation does not seem to be in line with our intuitive understanding of academic inventors. In fact, we reconducted all experiments on the whole DrugBank dataset, and a similar conclusion was drawn (Table 11 in Appendix). In our opinion, the main reasons are twofold: (1) most patents on drugs come from pharmaceutical companies with a clear R&D goal, but the authors can usually carry out free exploratory research, and even

Table 7 Diversity of research interests for solely publishing authors, solely patenting inventors, and academic inventors from the USA (a), Japan (b), the UK (c), Germany (d), and Italy (e)

(a) The USA	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	**7.371 (± 0.227)**	7.367 (± 0.471)	7.103 (± 0.798)
RS (JS divergence)	**0.618 (± 0.017)**	0.610 (± 0.041)	0.588 (± 0.066)
RS (cosine distance)	**0.957 (± 0.028)**	0.944 (± 0.065)	0.909 (± 0.104)
DIV_0.85 (symmetrized KL divergence)	**5.181 (± 1.204)**	4.794 (± 1.513)	3.424 (± 1.707)
DIV_0.85 (JS divergence)	**0.436 (± 0.101)**	0.403 (± 0.128)	0.288 (± 0.144)
DIV_0.85 (cosine distance)	**0.679 (± 0.158)**	0.627 (± 0.199)	0.447 (± 0.224)

(b) Japan	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	7.300 (± 0.364)	**7.409 (± 0.299)**	6.978 (± 0.801)
RS (JS divergence)	0.612 (± 0.027)	**0.615 (± 0.025)**	0.580 (± 0.064)
RS (cosine distance)	0.948 (± 0.042)	**0.953 (± 0.039)**	0.894 (± 0.103)
DIV_0.85 (symmetrized KL divergence)	4.755 (± 1.304)	**5.117 (± 1.220)**	2.997 (± 1.629)
DIV_0.85 (JS divergence)	0.400 (± 0.110)	**0.431 (± 0.103)**	0.252 (± 0.137)
DIV_0.85 (cosine distance)	0.623 (± 0.171)	**0.670 (± 0.161)**	0.392 (± 0.214)

(c) The UK	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	**7.362 (± 0.279)**	7.346 (± 0.397)	7.255 (± 0.506)
RS (JS divergence)	**0.617 (± 0.023)**	0.607 (± 0.034)	0.600 (± 0.041)
RS (cosine distance)	**0.955 (± 0.039)**	0.939 (± 0.053)	0.925 (± 0.076)
DIV_0.85 (symmetrized KL divergence)	**5.141 (± 1.283)**	4.342 (± 1.750)	3.679 (± 1.382)
DIV_0.85 (JS divergence)	**0.433 (± 0.108)**	0.365 (± 0.148)	0.309 (± 0.116)
DIV_0.85 (cosine distance)	**0.674 (± 0.168)**	0.568 (± 0.230)	0.480 (± 0.181)

(d) Germany	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	7.343 (± 0.235)	**7.398 (± 0.270)**	7.122 (± 0.522)
RS (JS divergence)	**0.616 (± 0.019)**	0.614 (± 0.023)	0.589 (± 0.047)
RS (cosine distance)	**0.954 (± 0.032)**	0.950 (± 0.039)	0.908 (± 0.079)
DIV_0.85 (symmetrized KL divergence)	4.801 (± 1.319)	**5.023 (± 1.268)**	3.085 (± 1.580)
DIV_0.85 (JS divergence)	0.404 (± 0.111)	**0.422 (± 0.107)**	0.259 (± 0.133)
DIV_0.85 (cosine distance)	0.629 (± 0.173)	**0.657 (± 0.167)**	0.403 (± 0.207)

(e) Italy	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	**7.369 (± 0.154)**	7.173 (± 0.920)	7.193 (± 0.514)
RS (JS divergence)	**0.618 (± 0.010)**	0.595 (± 0.077)	0.599 (± 0.043)
RS (cosine distance)	**0.957 (± 0.017)**	0.922 (± 0.120)	0.928 (± 0.068)
DIV_0.85 (symmetrized KL divergence)	**4.998 (± 1.144)**	4.496 (± 1.694)	2.997 (± 1.629)
DIV_0.85 (JS divergence)	**0.421 (± 0.096)**	0.378 (± 0.143)	0.313 (± 0.119)
DIV_0.85 (cosine distance)	**0.655 (± 0.150)**	0.589 (± 0.223)	0.486 (± 0.186)

Bold indicates the best results corresponding to each indicator

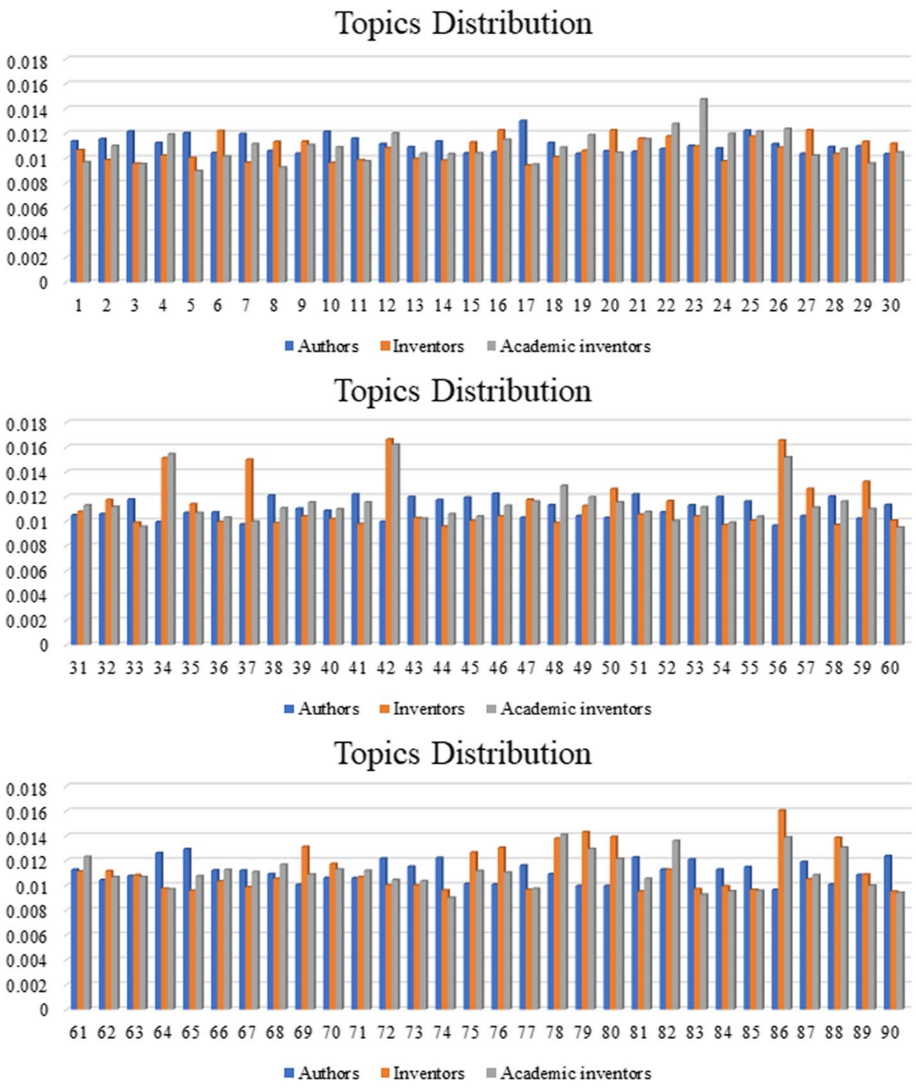


Fig. 9 The distribution of interest topics of solely publishing authors, solely patenting inventors, and academic inventors

constantly adjust their research directions according to hot themes; and (2) as science–technology gatekeepers, academic inventors span across academia and industry. Correspondingly, their interests are mainly at the interface between science and technology. In this way, they can seek a trade-off between research significance and the risk of patent uncertainty under the circumstance of market economy (Li et al., 2020). Hence, the scope of their interests may not be as diverse as the researchers with a single role.

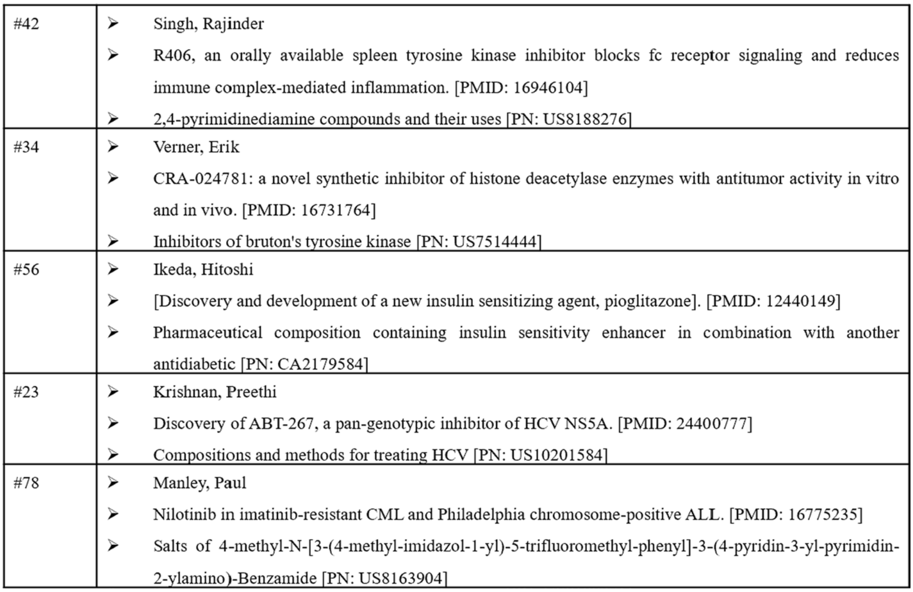
Table 8 An illustration of 12 themes from a 90-topic solution with the AT^{credit} model. Each theme is shown with top ten words conditioned on that theme

17	23	25	28	37	48
Safety	Virus	Number	Heart	Alkyl	Disorder
Study	Hcv	Aminoglycoside	Hypertension	Hydrogen	Serotonin
Placebo	Antiviral	Related	Vascular	Aryl	Antidepressant
Controlled	Infection	Spar	Cardiovascular	Substitute	Depression
Baseline	Hepatitis	Sequencing	Mortality	Cycloalkyl	Norepinephrine
Efficacy	HIV	Kd	Dysfunction	Halogen	Mood
Evaluate	Replication	Intrinsic	Angiotensin	Alkylene	Anxiety
Significantly	Genotype	Potassium	Cardiac	Alkoxy	Symptom
Difference	Nucleoside	Deprive	Prevention	Alkenyl	Depressive
Compare	Immunodeficiency	Hydroxytryptophan	Ischemia	Carboxy	Escitalopram
59	63	64	69	82	89
Atom	Pain	Metabolite	Process	Number	Disorder
carbon	Opioid	Metabolism	Preparation	Phenyl	Treat
Hydrogen	Chronic	Oral	Crystalline	Fluoro	Weight
Fluorine	Analgesic	Urine	Aripiprazole	Combine	Acid
Phenyl	Fentanyl	Excretion	Product	Fluorophenyl	Gel
Formula	Sublingual	Metabolic	Solubility	Phthalazin	Inflammatory
Acyl	Morphine	Excrete	Convert	Acetamide	Erythema
Amide	Ziconotide	Mass	Hygroscopic	Cyclopropyl	Dermatological
Nitrogen	Management	Pathway	Original	Phenylamino	Topically
Methoxy	Continuous	Metabolize	Substance	Tosylate	Cream

To check whether the interest diversity varies by country, the researchers from the following five countries were further analyzed: the USA, Japan, the UK, Germany, and Italy, presented in Table 7. Since the diversity of research interests is not sensitive to the threshold (Table 6), a threshold of 0.85 is utilized here. From Table 7, it is evident that the research interests of academic inventors are less diverse than those of the researchers with a single role across countries. This indicates that this conclusion seems be independent from national incentive and funding systems.

Distribution and preferred content of interest topics

To determine which interest topics the solely publishing authors, the solely patenting inventors, and the academic inventors prefer, Fig. 9 illustrates the distribution of interest topics. In Fig. 9, the horizontal axis denotes the topic identification, and the vertical axis represents the average probability distribution of a focal topic. Among these three types of researchers, the solely publishing authors have most even probability distribution. This is consistent with having the most diverse research interests (Table 6). Further, we can



#42	➢ Singh, Rajinder ➢ R406, an orally available spleen tyrosine kinase inhibitor blocks fc receptor signaling and reduces immune complex-mediated inflammation. [PMID: 16946104] ➢ 2,4-pyrimidinediamine compounds and their uses [PN: US8188276]
#34	➢ Verner, Erik ➢ CRA-024781: a novel synthetic inhibitor of histone deacetylase enzymes with antitumor activity in vitro and in vivo. [PMID: 16731764] ➢ Inhibitors of bruton's tyrosine kinase [PN: US7514444]
#56	➢ Ikeda, Hitoshi ➢ [Discovery and development of a new insulin sensitizing agent, pioglitazone]. [PMID: 12440149] ➢ Pharmaceutical composition containing insulin sensitivity enhancer in combination with another antidiabetic [PN: CA2179584]
#23	➢ Krishnan, Preethi ➢ Discovery of ABT-267, a pan-genotypic inhibitor of HCV NS5A. [PMID: 24400777] ➢ Compositions and methods for treating HCV [PN: US10201584]
#78	➢ Manley, Paul ➢ Nilotinib in imatinib-resistant CML and Philadelphia chromosome-positive ALL. [PMID: 16775235] ➢ Salts of 4-methyl-N-[3-(4-methyl-imidazol-1-yl)-5-trifluoromethyl-phenyl]-3-(4-pyridin-3-yl-pyrimidin-2-ylamino)-Benzamide [PN: US8163904]

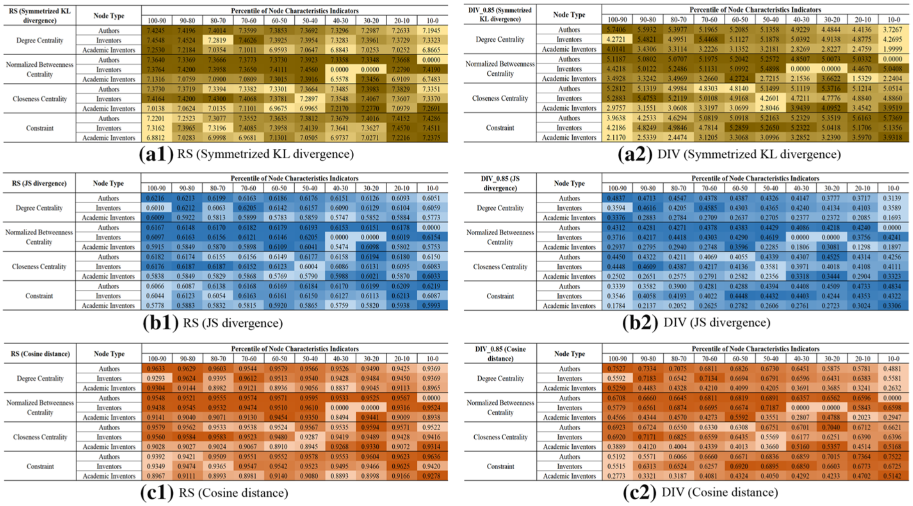
Fig. 10 An illustration of five interest topics of academic inventors. Each topic is attached with one researcher, one scientific publication, and one patent document

Table 9 Top two interest topics for two representatives from each type of researcher

	Name	Topic distribution	
Authors	Shen, Jianwei	64: 62.44%	25: 5.56%
	Ringold, Forrest G	17: 36.79%	87: 21.79%
Inventors	Bando, Takuji	69: 61.79%	6: 5.36%
	Arimilli, Murty N	37: 70.90%	34: 2.24%
Academic inventors	De Clercq, Erik	23: 43.22%	61: 7.72%
	Wang, Bing	82: 56.25%	78: 11.32%

observe the following commonality and specialty in terms of interest topics by inspecting the difference between the resulting average probability distributions. The topics 63, 25, and 28 are shared by all three types of researchers since their average probabilities are very closer to each other. Different from academic inventors, the researchers with a single role (solely publishing or patenting) are interested in topics 82, 89, and 23. In addition, the solely publishing authors prefer the topics 17, 64, and 90; solely patenting inventors prefer the topics 37, 59, and 69; and the academic inventor prefer the topics 23, 82, and 48. For ease of understanding, Table 8 illustrates examples of 12 themes.

As for academic inventors, Fig. 10 shows the top five interest topics in terms of average probability distribution, in which each topic is attached with one researcher, one scientific publication, and one patent document. Theme 42 is on drugs with blocking function,



RS (Symmetrized KL divergence)	Node Type	Percentile of Node Characteristics Indicators									
		100-90	90-80	80-70	70-60	60-50	50-40	40-30	30-20	20-10	10-0
Degree Centrality	Authors	7.4245	7.4196	7.4014	7.3599	7.3853	7.3692	7.3296	7.2987	7.2633	7.1945
	Inventors	7.4548	7.4524	7.2819	7.4626	7.3925	7.3954	7.3283	7.3961	7.3729	7.3323
	Academic Inventors	7.2530	7.2184	7.0354	7.1011	6.9593	7.0647	6.8843	7.0253	7.0252	6.8665
Normalized Betweenness Centrality	Authors	7.3640	7.3369	7.3666	7.3773	7.3730	7.3923	7.3358	7.3348	7.3668	0.0000
	Inventors	7.3764	7.4200	7.3958	7.3650	7.4111	7.4560	0.0000	0.0000	7.2790	7.4190
	Academic Inventors	7.1316	7.0759	7.0900	7.0809	7.3015	7.3916	6.5578	7.3456	6.9109	6.7483
Closeness Centrality	Authors	7.3730	7.3719	7.3394	7.3382	7.3301	7.3664	7.3485	7.3983	7.3829	7.3351
	Inventors	7.4164	7.4200	7.4300	7.4068	7.3781	7.2897	7.3548	7.4067	7.3607	7.3370
	Academic Inventors	7.0138	7.0624	7.0135	7.1101	6.9675	6.9965	7.2170	7.2770	7.0979	7.2691
Constraint	Authors	7.2201	7.2523	7.3077	7.3552	7.3635	7.3812	7.3679	7.4016	7.4152	7.4286
	Inventors	7.3162	7.3965	7.3196	7.4085	7.3958	7.4139	7.3641	7.3627	7.4570	7.4511
	Academic Inventors	6.8812	7.0283	6.9998	6.9681	7.1301	7.0505	6.9737	7.0271	7.2216	7.2375

(a1) RS (Symmetrized KL divergence)

DIV_0.85 (Symmetrized KL divergence)	Node Type	Percentile of Node Characteristics Indicators									
		100-90	90-80	80-70	70-60	60-50	50-40	40-30	30-20	20-10	10-0
Degree Centrality	Authors	5.7406	5.5932	5.3977	5.1965	5.2085	5.1358	4.9229	4.4844	4.4136	3.7267
	Inventors	4.2721	5.4821	4.9951	5.4468	5.1127	5.1878	5.0392	4.9138	4.8775	4.2695
	Academic Inventors	4.0141	3.4306	3.3114	3.2226	3.1352	3.2181	2.8269	2.8227	2.4759	1.9999
Normalized Betweenness Centrality	Authors	5.1187	5.0802	5.0707	5.1975	5.2042	5.2572	4.8507	5.0073	5.0332	0.0000
	Inventors	4.4218	5.0122	5.2486	5.1131	5.0992	5.4898	0.0000	0.0000	4.4670	5.0408
	Academic Inventors	3.4928	3.3242	3.4969	3.2660	4.2724	2.7215	2.1536	3.6622	1.5329	2.2404
Closeness Centrality	Authors	5.2812	5.1319	4.9984	4.8303	4.8140	5.1499	5.1119	5.3716	5.1214	5.0514
	Inventors	5.2883	5.4753	5.2119	5.0108	4.9168	4.2601	4.7211	4.7776	4.8840	4.8860
	Academic Inventors	2.9757	3.1551	3.0608	3.3197	3.0699	2.8046	3.9439	4.0952	3.4542	3.9519
Constraint	Authors	3.9638	4.2533	4.6294	5.0819	5.0918	5.2163	5.2329	5.3519	5.6163	5.7369
	Inventors	4.2186	4.8249	4.9846	4.7814	5.2859	5.2650	5.2322	5.0418	5.1706	5.1356
	Academic Inventors	2.1170	2.5339	2.4474	3.1205	3.3068	3.0996	3.2852	3.2390	3.5970	3.9318

(a2) DIV (Symmetrized KL divergence)

RS (JS divergence)	Node Type	Percentile of Node Characteristics Indicators									
		100-90	90-80	80-70	70-60	60-50	50-40	40-30	30-20	20-10	10-0
Degree Centrality	Authors	0.6216	0.6213	0.6199	0.6163	0.6186	0.6176	0.6151	0.6126	0.6093	0.6051
	Inventors	0.6010	0.6212	0.6063	0.6205	0.6142	0.6157	0.6090	0.6129	0.6104	0.6059
	Academic Inventors	0.6009	0.5922	0.5813	0.5899	0.5783	0.5859	0.5747	0.5852	0.5884	0.5773
Normalized Betweenness Centrality	Authors	0.6167	0.6148	0.6170	0.6182	0.6179	0.6193	0.6153	0.6151	0.6178	0.0000
	Inventors	0.6097	0.6163	0.6156	0.6121	0.6146	0.6205	0.0000	0.0000	0.6019	0.6154
	Academic Inventors	0.5915	0.5849	0.5870	0.5898	0.6109	0.6041	0.5474	0.6098	0.5802	0.5753
Closeness Centrality	Authors	0.6182	0.6174	0.6155	0.6156	0.6149	0.6177	0.6158	0.6194	0.6180	0.6150
	Inventors	0.6176	0.6187	0.6187	0.6152	0.6123	0.6004	0.6086	0.6131	0.6095	0.6083
	Academic Inventors	0.5838	0.5849	0.5829	0.5868	0.5769	0.5790	0.5988	0.6021	0.5870	0.6033
Constraint	Authors	0.6066	0.6087	0.6138	0.6168	0.6169	0.6184	0.6170	0.6199	0.6209	0.6219
	Inventors	0.6044	0.6123	0.6054	0.6163	0.6161	0.6150	0.6127	0.6113	0.6213	0.6087
	Academic Inventors	0.5778	0.5883	0.5832	0.5815	0.5920	0.5865	0.5759	0.5820	0.5938	0.5993

(b1) RS (JS divergence)

DIV_0.85 (JS divergence)	Node Type	Percentile of Node Characteristics Indicators									
		100-90	90-80	80-70	70-60	60-50	50-40	40-30	30-20	20-10	10-0
Degree Centrality	Authors	0.4837	0.4713	0.4547	0.4378	0.4387	0.4326	0.4147	0.3777	0.3717	0.3139
	Inventors	0.3594	0.4616	0.4205	0.4585	0.4303	0.4365	0.4240	0.4134	0.4103	0.3589
	Academic Inventors	0.3376	0.2883	0.2784	0.2709	0.2637	0.2705	0.2377	0.2372	0.2085	0.1693
Normalized Betweenness Centrality	Authors	0.4312	0.4281	0.4271	0.4378	0.4383	0.4429	0.4086	0.4218	0.4240	0.0000
	Inventors	0.3716	0.4217	0.4418	0.4303	0.4290	0.4619	0.0000	0.0000	0.3756	0.4241
	Academic Inventors	0.2937	0.2795	0.2940	0.2748	0.3596	0.2285	0.1806	0.3081	0.1298	0.1897
Closeness Centrality	Authors	0.4450	0.4322	0.4211	0.4069	0.4055	0.4339	0.4307	0.4525	0.4314	0.4256
	Inventors	0.4448	0.4609	0.4387	0.4217	0.4136	0.3581	0.3971	0.4018	0.4108	0.4111
	Academic Inventors	0.2502	0.2651	0.2575	0.2791	0.2582	0.2356	0.3318	0.3444	0.2904	0.3323
Constraint	Authors	0.3339	0.3582	0.3900	0.4281	0.4288	0.4394	0.4408	0.4509	0.4733	0.4834
	Inventors	0.3546	0.4058	0.4193	0.4022	0.4448	0.4432	0.4403	0.4244	0.4353	0.4322
	Academic Inventors	0.1784	0.2137	0.2052	0.2625	0.2782	0.2606	0.2761	0.2723	0.3024	0.3306

(b2) DIV (JS divergence)

RS (Cosine distance)	Node Type	Percentile of Node Characteristics Indicators									
		100-90	90-80	80-70	70-60	60-50	50-40	40-30	30-20	20-10	10-0
Degree Centrality	Authors	0.9633	0.9629	0.9603	0.9544	0.9579	0.9566	0.9526	0.9490	0.9425	0.9369
	Inventors	0.9293	0.9624	0.9395	0.9612	0.9513	0.9540	0.9428	0.9484	0.9450	0.9369
	Academic Inventors	0.9304	0.9144	0.8982	0.9121	0.8936	0.9056	0.8837	0.9045	0.9113	0.8965
Normalized Betweenness Centrality	Authors	0.9548	0.9521	0.9555	0.9574	0.9571	0.9595	0.9533	0.9525	0.9567	0.0000
	Inventors	0.9438	0.9545	0.9532	0.9474	0.9510	0.9610	0.0000	0.0000	0.9316	0.9524
	Academic Inventors	0.9141	0.9040	0.9071	0.9130	0.9454	0.9350	0.8494	0.9441	0.9009	0.8938
Closeness Centrality	Authors	0.9579	0.9562	0.9533	0.9538	0.9524	0.9567	0.9535	0.9594	0.9571	0.9522
	Inventors	0.9560	0.9584	0.9583	0.9523	0.9480	0.9287	0.9419	0.9489	0.9428	0.9416
	Academic Inventors	0.9028	0.9027	0.9024	0.9067	0.8910	0.8945	0.9268	0.9330	0.9072	0.9314
Constraint	Authors	0.9392	0.9421	0.9509	0.9551	0.9552	0.9578	0.9553	0.9604	0.9623	0.9636
	Inventors	0.9349	0.9474	0.9365	0.9547	0.9542	0.9523	0.9495	0.9466	0.9625	0.9420
	Academic Inventors	0.8967	0.9111	0.8993	0.8981	0.9140	0.9080	0.8893	0.8998	0.9166	0.9278

(c1) RS (Cosine distance)

DIV_0.85 (Cosine distance)	Node Type	Percentile of Node Characteristics Indicators									
		100-90	90-80	80-70	70-60	60-50	50-40	40-30	30-20	20-10	10-0
Degree Centrality	Authors	0.7527	0.7334	0.7075	0.6811	0.6826	0.6730	0.6451	0.5875	0.5781	0.4881
	Inventors	0.5592	0.7183	0.6542	0.7134	0.6694	0.6791	0.6596	0.6431	0.6383	0.5581
	Academic Inventors	0.5250	0.4483	0.4328	0.4210	0.4099	0.4205	0.3691	0.3685	0.3241	0.2632
Normalized Betweenness Centrality	Authors	0.6708	0.6660	0.6645	0.6811	0.6819	0.6891	0.6357	0.6562	0.6596	0.0000
	Inventors	0.5779	0.6561	0.6874	0.6695	0.6674	0.7187	0.0000	0.0000	0.5843	0.6598
	Academic Inventors	0.4566	0.4344	0.4570	0.4273	0.5592	0.3551	0.2807	0.4788	0.2023	0.2947
Closeness Centrality	Authors	0.6923	0.6724	0.6550	0.6330	0.6308	0.6751	0.6701	0.7040	0.6712	0.6621
	Inventors	0.6920	0.7171	0.6825	0.6559	0.6435	0.5569	0.6177	0.6251	0.6390	0.6396
	Academic Inventors	0.3889	0.4120	0.4004	0.4339	0.4013	0.3660	0.5160	0.5357	0.4514	0.5168
Constraint	Authors	0.5192	0.5571	0.6066	0.6660	0.6671	0.6836	0.6859	0.7015	0.7364	0.7522
	Inventors	0.5515	0.6313	0.6524	0.6257	0.6920	0.6895	0.6850	0.6603	0.6773	0.6725
	Academic Inventors	0.2773	0.3321	0.3187	0.4081	0.4324	0.4050	0.4292	0.4233	0.4702	0.5142

(c2) DIV (Cosine distance)

Fig. 11 The diversity of research interests of each type of researcher, with the percentile of node characteristic indicators

which have two major classes as inhibitors and antagonists. This kind of drugs plays a role in reducing activity and alleviating symptoms. For example, R406 is an orally available spleen tyrosine kinase (Syk) inhibitor, which blocks Fc receptor signaling and reduces immune complex-mediated inflammation. It potentially modulates Syk activity in human disease (Braselmann et al., 2006). Theme 34 discusses antitumor drugs, which inhibit tumor cell growth and induce apoptosis, providing promising therapeutic options for patients. Theme 56 discusses antidiabetic drugs and focuses on insulin, which has always been the primary pharmacological agent for treating diabetes and preventing its complications (Falcetta et al., 2022). In addition, academic inventors are also interested in antiviral drugs (Theme 23): two RNA-viruses: human immunodeficiency virus (HIV) and hepatitis C virus (HCV) received the most attention. Both have similar blood and mother-to-child transmission routes, but act on different types of cells. The former mainly infects human immune cells, and the latter infects liver cells. Theme 78 is clinical application and preparation of compound drugs. Once the clinical drug discovery is demonstrated, the researchers under this topic are more invested in the production of related drug reagents and products.

To verify whether the discovered research interests make sense, two representatives from each type of researchers were taken as examples, as reported in Table 9. Shen, Jianwei published two articles in our dataset about metabolism and disposition of inhibitor of hepatitis C. Bando, Takuji patented six inventions on the preparation of low hygroscopic aripiprazole drug. As an academic inventor, Wang, Bing worked at BioMarin Pharmaceutical Inc. His research interests include clinical application research and preparation of anticancer agents, involving one scientific publication and five patents in our dataset.

Table 10 Spearman rank correlation coefficient of interest diversity and node characteristic indicators

		RS (symmetrized KL divergence)	RS (JS divergence)	RS (cosine distance)	DIV (symmetrized KL divergence)	DIV (JS divergence)	DIV (cosine distance)
Degree centrality	Authors	**0.383****	**0.482****	**0.387****	**0.446****	**0.446****	**0.446****
	Inventors	0.056*	**0.334****	**0.289****	**0.305****	**0.306****	**0.307****
	Academic inventors	**0.283****	**0.256****	**0.232****	**0.278****	**0.277****	**0.278****
Normalized betweenness centrality	Authors	**0.047****	**0.057****	0.032*	**0.056****	**0.056****	**0.056****
	Inventors	0.017	**0.107****	**0.083****	**0.094****	**0.094****	**0.093****
	Academic inventors	0.031	0.041	0.029	0.056	0.056	0.056
Closeness centrality	Authors	**−0.054****	**−0.050****	−0.028	−0.021	−0.021	−0.021
	Inventors	0.005	0.055*	**0.088****	**0.074****	**0.074****	**0.074****
	Academic inventors	**−0.176****	**−0.192****	**−0.174****	**−0.198****	**−0.197****	**−0.198****
Constraint	Authors	**−0.374****	**−0.468****	**−0.373****	**−0.428****	**−0.427****	**−0.427****
	Inventors	**−0.065****	**−0.315****	**−0.269****	**−0.290****	**−0.291****	**−0.292****
	Academic inventors	**−0.263****	**−0.240****	**−0.215****	**−0.263****	**−0.262****	**−0.262****

*Correlation is significant at the 0.05 level (two-tailed)

**Correlation is significant at the 0.01 level (two-tailed)

Bold indicates the results of significance at the 0.01 level

Table 12 in the Appendix illustrates the titles of scientific publications and patents of these six researchers.

Interest diversity with position characteristics

In this subsection, we answer the other question: does the position of the researcher correlate with their diverse research interests? From Table 6, one can see that the diversity of research interests is not sensitive to the threshold. Therefore, a threshold 0.85 was fixed for the following analysis.

Figure 11 illustrates the distribution of the diversity of research interests of each type of researcher with the percentile of node characteristic indicators. The patterns are not consistent across the three types of researchers. The diversity of research interests of the researchers with one single role mainly concentrates on the top percentile of the degree centrality indicator and the bottom percentile of the constraint indicator. Inventors who are closer to the geometric center of the network and with medium extent of control over non-adjacent nodes have more diverse research interests.

To find out whether the interest diversity correlates with the position characteristics, we implemented Spearman rank correlation coefficient test, as presented in Table 10. The "degree centrality" and "constraint" present significant positive correlation and negative correlation with the diversity of interests of researchers, respectively. That is, whichever role one researcher has, the more widely connected, and the more structural nodes in the network, the more diverse their research interests tend to be.

In term of "closeness centrality," the results are mixed. More specifically, only academic inventors show a significant weaker negative correlation. For the researchers with one single role, we cannot conclude that the important positions at the center of a network correlate with the diversity of interests. As for "normalized betweenness centrality," nearly half of the cells in Table 10 are nonsignificant, and the other half have low values. We argue that this indicator does not correlate with the interest diversity.

In summary, the position of the researchers in a cooperative network does have a certain relation with the diversity of their research interests. For all researcher types, those with more social and more as structural nodes in the network have more diverse research interests.

Conclusions

Academic inventors play an import role in the knowledge diffusion between science and technology. Considerable efforts have been spent analyzing academic inventors in the literature. However, it is still unknown whether they have more diverse interests than researchers with one single role, and whether their position in science–technology interactions correlates with their interest diversity.

To answer these two questions, this study puts forward a rule-based identification approach of academic inventors. After research interests of each researcher were identified by an interest discovery AT^{credit} model, two diversity indicators with three disparity measurements were calculated for each type of researcher. Extensive empirical results on the DrugBank dataset indicate that academic inventors have less diverse research interests than the researchers with a single role, followed by solely patenting inventors.

The position of the researchers has a certain relation with the diversity of their research interests. The "degree centrality" has a significant positive correlation with the diversity of research interests, and the "constraint" presents a significant negative correlation. Among the three types of researchers, interest diversity of only academic inventors shows a weakly negative correlation with the "closeness centrality," and does not correlate with the "normalized betweenness centrality."

There are several limitations of this study. As mentioned in the *Dataset* subsection, this study only considers scholarly articles and patents attached to each drug in the DrugBank database, which may result in a lower proportion of academic inventors. In the near future, we will retrieve scholarly articles and patents authored or patented by each researcher in the DrugBank database from comprehensive bibliographic databases. Further, the identification of academic inventors can benefit from a good name disambiguation method, and the rules for identifying academic inventors will be further enriched in our next study. In addition, we will try to identify the factors contributing to the position and interest diversity of academic inventors for science–technology interactions.

Appendix

See Tables 11 and 12.

Table 11 Diversity of research interests for solely publishing authors, solely patenting inventors, and academic inventors in the whole DrugBank dataset

	Authors	Inventors	Academic inventors
RS (symmetrized KL divergence)	**9.251 (± 0.263)**	9.231 (± 0.522)	9.146 (± 0.555)
RS (JS divergence)	**0.643 (± 0.016)**	0.636 (± 0.036)	0.629 (± 0.037)
RS (cosine distance)	**0.965 (± 0.025)**	0.954 (± 0.055)	0.943 (± 0.059)
DIV_0.80 (symmetrized KL divergence)	**5.517 (± 1.787)**	5.245 (± 2.047)	4.212 (± 2.046)
DIV_0.80 (JS divergence)	**0.380** (±0.123)	0.361 (± 0.141)	0.290 (± 0.141)
DIV_0.80 (cosine distance)	**0.571 (± 0.185)**	0.542 (± 0.212)	0.435 (± 0.212)
DIV_0.85 (symmetrized KL divergence)	**6.035 (± 1.777)**	5.759 (± 2.055)	4.723 (± 2.081)
DIV_0.85 (JS divergence)	**0.417 (± 0.123)**	0.397 (± 0.142)	0.326 (± 0.144)
DIV_0.85 (cosine distance)	**0.625 (± 0.185)**	0.596 (± 0.214)	0.489 (± 0.216)
DIV_0.90 (symmetrized KL divergence)	**6.569 (± 1.748)**	6.294 (±2.039)	5.271 (± 2.083)
DIV_0.90 (JS divergence)	**0.454 (± 0.121)**	0.434 (± 0.141)	0.364 (± 0.144)
DIV_0.90 (cosine distance)	**0.681 (± 0.181)**	0.652 (± 0.212)	0.546 (± 0.216)

Standard deviation is shown in parentheses

Bold indicates the best results corresponding to each indicator

Table 12 The title of scientific publications and patents authored by inventor, author, and academic inventor

	Name	Title
Authors	Shen, Jianwei	Metabolism and disposition of hepatitis C polymerase inhibitor dasabuvir in humans [PMID: 27179126] Metabolism and disposition of pan-genotypic inhibitor of hepatitis C virus NS5A ombitasvir in humans. [PMID: 27179128]
	Ringold, Forrest G	Sufentanil sublingual tablet system for the management of postoperative pain following open abdominal surgery [PMID: 25318408]
Inventors	Bando, Takuji	Low hygroscopic aripiprazole drug substance and processes for the preparation thereof [PN: US8017615/PN: US8399469/PN: US8580796/PN: US8642760/PN: US8993761/PN: US9359302]
	Arimilli, Murty N	Nucleotide analogs [PN: CA2261619] Nucleotide analog compositions [PN: CA2298057/US6451340] Antiviral phosphonomethyoxy nucleotide analogs having increased oral bioavarilability [PN: US5922695/US5977089/US6043230]
Academic inventors	De Clercq, Erik	HIV resistance to reverse transcriptase inhibitors. [PMID: 7508227] Approved antiviral drugs over the past 50 years. [PMID: 27281742] Emerging anti-HIV drugs. [PMID: 15934866] Specific phosphorylation of 5-ethyl-2′-deoxyuridine by herpes simplex virus-infected cells and incorporation into viral DNA. [PMID: 2822705] N-phosphonylmethoxyalkyl pyrimidines and purines and therapeutic application thereof [PN: CA1340856] N-phosphonylmethoxyalkyl derivatives of pyrimidine and purine bases and a therapeutical composition therefrom with antiviral activity [PN:US5142051]
	Wang, Bing	Discovery and characterization of (8S,9R)-5-Fluoro-8-(4-fluorophenyl)-9-methyl-1H-1,2,4-triazol-5-yl)-2,7,8,9-tetrahydro-3H-pyrido[4,3,2-de]phthalazin-3-one (BMN 673, Talazoparib), a novel, highly potent, and orally efficacious poly(ADP-ribose) polymerase-1/2 inhibitor, as an anticancer agent. [PMID: 26652717] Crystalline (8S,9R)-5-fluoro-8-(4-fluorophenyl)-9-(1-methyl-1H-1,2,4-triazol-5-yl)-8,9-dihydro-2H-pyrido[4,3,2-de]phthalazin-3(7H)-one tosylate salt [PN:US10189837/US8735392] Dihydropyridophthalazinone inhibitors of poly(ADP-ribose)polymerase (PARP) [PN:US8012976/US8420650/US9820985]

Acknowledgements This research received the financial support from the National Natural Science Foundation of China under grant numbers 72074014 and 72004012. Our gratitude also goes to the anonymous reviewers and the editor for their valuable comments.

References

Adams, J., Black, G. C., Clemmons, J. R., & Stephan, P. E. (2005). Scientific teams and institutional collaborations: evidence from U.S. universities, 1981–1999. *Research Policy, 34*(3), 259–285.

Agrawal, A., & Henderson, R. (2002). Putting patents in context: exploring knowledge transfer from MIT. *Management Science, 48*(1), 44–60.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. *The rate and direction of inventive activity: economic and social factors* (pp. 609–626). Princeton University Press.

Azoulay, A., Ding, W., & Stuart, T. (2007). The determinants of faculty patenting behavior: demographics of opportunities? *Journal of Economic Behavior and Organization, 63*(4), 599–623.

Azoulay, P., Ding, W., & Stuart, T. (2009). The impact of academic patenting on the rate, quality and direction of (public) research output. *The Journal of Industrial Economics, 57*(4), 637–676.

Ba, Z., & Liang, Z. (2021). A novel approach to measuring science-technology linkage: from the perspective of knowledge network coupling. *Journal of Informetrics, 15*(3), 101167.

Balconi, M., Breschi, S., & Lissoni, F. (2004). Networks of inventors and the role of academia: an exploration of Italian patent data. *Research Policy, 33*, 127–145.

Bassecouolard, E., & Zitt, M. (2004). Patents and publications: the lexical connection. In H. F. Moed, W. Glänzel, & U. Schoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 665–694). Springer.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet al location. *Journal of Machine Learning Research, 3*, 993–1022.

Blumenthal, D., Campbell, E., Anderson, M., Causino, N., & Louis, K. (1997). Withholding research results in academic life science: evidence from a national survey of faculty. *Journal of the American Medical Association, 277*(15), 1224–1228.

Boyack, K., & Klavans, R. (2008). Measuring science-technology interaction using rare inventor-author names. *Journal of Informetrics, 2*, 173–182.

Braselmann, S., Taylor, V., Zhao, H., Wang, S., Sylvain, C., Baluom, M., Qu, K., Herlaar, E., Lau, A., Young, C., Wong, B., Lovell, S., Sun, T., Park, G., Argade, A., Jurcevic, S., Pine, P., Singh, R., Grossbard, E., … Masuda, E. (2006). R406, an orally available spleen tyrosine kinase inhibitor blocks fc receptor signaling and reduces immune complex-mediated inflammation. *Journal of Pharmacology and Experimental Therapeutics, 319*(3), 998–1008.

Breschi, S., & Catalini, C. (2010). Tracing the links between science and technology: an exploratory analysis of scientists' and inventors' networks. *Research Policy, 39*(1), 14–26.

Bu, Y., Li, M., Gu, W., & Huang, W. (2020). Topic diversity: A discipline scheme-free diversity measurement for journals. *Journal of the Association for Information Science and Technology, 72*, 523.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: an analysis of nonpatent references found in patent documents. *Scientometrics, 69*(1), 3–20.

Carayol, N., & Carpentier, E. (2021). The spread of academic invention: A nationwide case study on French data (1995–2012). *The Journal of Technology Transfer, 47*, 1395.

Caron, E., & van Eck, N.-J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In *Proceedings of the 19th International Conference on Science and Technology Indicators* (pp. 79–86).

Cassiman, B., Glenisson, P., & Van Looy, B. (2007). Measuring industry-science links through inventor-author relations: a profiling methodology. *Scientometrics, 70*(2), 379–391.

Crespi, G., D'Este, P., Fontana, R., & Geuna, A. (2011). The impact of academic patenting on university research and its transfer. *Research Policy, 40*(1), 55–68.

Czarnitzki, D., Doherr, T., Hussinger, K., Schliessler, P., & Toole, A. (2016). Knowledge creates markets: the influence of entrepreneurial support and patent rights on academic entrepreneurship. *European Economic Review, 86*, 131–146.

Dubaric, E., Giannoccaro, D., Bengtsson, R., & Ackermann, T. (2011). Patent data as indicators of wind power technology development. *World Patent Information, 33*(2), 144–149.

Ejermo, O., & Toivanen, H. (2018). University invention and the abolishment of the professor's privilege in Finland. *Research Policy, 47*(4), 814–825.

Fabrizio, K., & Di Minin, A. (2008). Commercializing the laboratory: faculty patenting and the open science environment. *Research Policy, 37*(5), 914–931.

Falcetta, P., Aragona, M., Bertolotto, A., Bianchi, C., Campi, F., Garofolo, M., & Del Prato, S. (2022). Insulin discovery: a pivotal point in medical history. *Metabolism, 127*, 154941.

Forti, E., Franzoni, C., & Sobrero, M. (2013). Bridges or isolates? Investigating the social networks of academic inventors. *Research Policy, 42*(8), 1378–1388.

Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: an exploratory study of 'reverse' citation relations. *Scientometrics, 58*(2), 415–428.

Grimm, H., & Jaenicke, J. (2015). Testing the causal relationship between academic patenting and scientific publishing in Germany: crowding-out or reinforcement? *Journal of Technology Transfer, 40*(3), 512–535.

Guan, J., & Wang, G. (2010). A comparative study of research performance in nanotechnology for China's inventor–authors and their non-inventing peers. *Scientometrics, 84*(2), 331–343.

Hagen, N. (2013). Harmonic coauthor credit: a parsimonious quantification of the byline hierarchy. *Journal of Informetrics, 7*(4), 784–791.

Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., & Xu, S. (2017). Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics, 111*(3), 1879–1896.

Huang, M. H., Yang, H. W., & Chen, D. Z. (2015). Increasing science and technology linkage in fuel cells: a cross citation analysis of papers and patents. *Journal of Informetrics, 9*(2), 237–249.

Hvide, H., & Jones, B. (2018). University innovation and the professor's privilege. *American Economic Review, 108*(7), 1860–1898.

Kang, B. (2020). Impact of academic patenting on scientific publication quality at the project level. *Asian Journal of Technology Innovation, 29*(2), 258–282.

Kawamae, N. (2010). Author interest topic model. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 887–888). ACM.

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: a case of DBLP. *Scientometrics, 116*(3), 1867–1886.

Klitkou, A., Nygaard, S., & Meyer, M. (2007). Tracking techno-science networks: a case study of fuel cells and related hydrogen technology R&D in Norway. *Scientometrics, 70*(2), 491–518.

Leahey, E. (2016). From sole investigator to team scientist: trends in the practice and study of research collaboration. *Annual Review of Sociology, 42*, 81–100.

Lee, S. (2019). Academic entrepreneurship: exploring the effects of academic patenting activity on publication and collaboration among heterogeneous researchers in South Korea. *Journal of Technology Transfer, 44*(6), 1993–2013.

Leydesdorff, L., Wagner, C., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics, 13*(1), 255–269.

Li, G., Lai, R., D'Amour, A., et al. (2014). Disambiguation and co-authorship networks of the US patent inventor database. *Research Policy, 43*(6), 941–955.

Li, X., Zhao, D., & Hu, X. (2020). Gatekeepers in knowledge transfer between science and technology: an exploratory study in the area of gene editing. *Scientometrics, 124*(2), 1261–1277.

Lissoni, F., Sanditov, B., & Tarasconi, G. (2006). The Keins database on academic inventors: Methodology and contents. Università commerciale Luigi Bocconi

Lissoni, F., Montobbio, F. (2008). Inventorship and authorship in Patent-Publication pairs: An enquiry into the economics of scientific credit. *Working Papers*, 224

Lissoni, F. (2010). Academic inventors as brokers. *Research Policy, 39*(7), 843–857.

Maraut, S., & Martínez, C. (2014). Identifying author-inventors from Span: methods and a first insight into results. *Scientometrics, 101*(1), 445–476.

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy, 29*(3), 409–434.

Meyer, M. (2006). Are patenting scientists the better scholars? An exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology. *Research Policy, 35*(10), 1646–1662.

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 500–509). ACM.

Murray, F. (2004). The role of academic inventors in entrepreneurial firms: sharing the laboratory life. *Research Policy, 33*(4), 643–659.

Narin, F., & Noma, E. (1985). Is technology becoming science? *Scientometrics, 7*(3–6), 369–381.

Nelson, R. (1959). The simple economics of basic scientific research. *Journal of Political Economy, 67*, 297–306.

Noyons, E., Raan, A., Grupp, H., et al. (1994). Exploring the science and technology interface: Inventor-author relations in laser medicine research. *Research Policy, 23*(4), 443–457.

Pezzoni, M., Lissoni, F., & Tarasconi, G. (2014). How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation. *Scientometrics, 101*(1), 477–504.

Quatraro, F., & Scandura, A. (2019). Academic inventors and the antecedents of green technologies. A regional analysis of Italian patent data. *Ecological Economics, 156*, 247–263.

Raffo, J., & Lhuillery, S. (2009). How to play the name game: patent retrieval comparing different heuristics. *Research Policy, 38*(10), 1617–1627.

Rao, C. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology, 21*(1), 24–43.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems, 4*(1–4), 38.

Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., & Ohta, T. (2007). AKANE system: Protein-protein interaction pairs in the BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the 2nd BioCreative challenge evaluation workshop* (pp. 209–212).

Shannon, C. (1950). The mathematical theory of communication. *Bell Labs Technical Journal, 3*(9), 31–32.

Shi, Q., Qiao, X., Xu, S., & Nong, G. (2013). Author-topic evolution model and its application in analysis of research interests evolution. *Journal of the China Society for Scientific and Technical Information, 32*(9), 912–919.

Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Extracting the commercialization gap between science and technology—case study of a solar cell. *Technological Forecasting and Social Change, 77*(7), 1147–1155.

Shibata, N., Kajikawa, Y., & Sakata, I. (2011). Detecting potential technological fronts by comparing scientific papers and patents. *Foresight, 13*, 51–60.

Simpson, A. (1949). Measurement of diversity. *Nature, 163*, 4148.

Stephan, P., Gurmu, S., Sumell, A., & Black, G. (2007). Who's patenting in the university? Evidence from the survey of doctorate recipients. *Economics of Innovation and New Technology, 16*(2), 71–99.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719.

Thursby, M., Thursby, J., & Gupta-Mukherjee, S. (2007). Are there real effects of licensing on academic research? A life cycle view. *Journal of Economic Behavior & Organization, 63*(4), 577–598.

Torvik, V., & Smalheiser, N. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 11*, 1–11.

Tscharntke, T., Hochberg, M., Rand, T., et al. (2007). Author sequence and credit for contributions in multiauthored publications. *PLoS Biology, 5*(1), e18.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., & Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In P. Bozanis & E. N. Houstis (Eds.), *Proceedings of the 10th Panhellenic Conference on Informatics* (pp. 382–392).

Van Looy, B., Callaert, J., & Debackere, K. (2006). Publication and patent behavior of academic researchers: conflicting, reinforcing or merely co-existing? *Research Policy, 35*(4), 596–608.

Wang, G., & Guan, J. (2011). Measuring science–technology interactions using patent citations and author-inventor links: an exploration analysis from Chinese nanotechnology. *Journal of Nanoparticle Research, 13*(12), 6245–6262.

Winnink, J., & Tijssen, R. (2014). R&D dynamics and scientific breakthroughs in HIV/AIDS drugs development: the case of integrase inhibitors. *Scientometrics, 101*(1), 1–16.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039.

Xu, S., Zhu, L., Qiao, X., Shi, Q., & Gui, J. (2012). Topic linkages between papers and patents. In *Proceedings of the 4th international conference on advanced science and technology* (pp. 176–183).

Xu, H., Winnink, J., Yue, Z., & Liu, Z. (2020). Topic-linked innovation paths in science and technology. *Journal of Informetrics, 14*(2), 101014.

Xu, S., Hao, L., Yang, G., Lu, K., & An, X. (2021). A topic models based framework for detecting and forecasting emerging technologies. *Technological Forecasting and Social Change, 162*, 120366.

Xu, S., Li, L., An, X., Hao, L., & Yang, G. (2021). An approach for detecting the commonality and specialty between scientific publications and patents. *Scientometrics, 126*(9), 7445–7475.

Xu, S., Li, L., Hao, L., An, X., & Yang, G. (2021). An author interest discovery model armed with authorship credit allocation scheme. In K. Toeppe, H. Yan, & S. K. W. Chu (Eds.), *Diversity, divergence, dialogue. iConference 2021. Lecture notes in computer science* (p. 12645). Springer.

Xu, S., Li, L., Wang, C., An, X., & Yang, G. (2022). An improved Author-Topic (AT) model with authorship credit allocation schemes. *Journal of Information Science, 15*, 101201.

Xu, S., Shi, Q., Qiao, X., Zhu, L., Jung, H., Lee, S., & Choi, S.-P. (2014). Author-topic over time (AToT): A dynamic users' interest model. In J. J. J. H. Park, H. Adeli, N. Park, & I. Woungang (Eds.), *Mobile, Ubiquitous, and Intelligent Computing* (pp. 239–245). Springer.

Xu, S., Shi, Q., Qiao, X., Zhu, L., Zhang, H., Jung, H., Lee, S., & Choi, S.-P. (2014). A dynamic users' interest discovery model with distributed inference algorithm. *International Journal of Distributed Sensor Networks, 2014*, 280890.

Xu, S., Zhai, D., Wang, F., An, X., Pang, H., & Sun, Y. (2019). A novel method for topic linkages between scientific publications and patents. *Journal of the Association for Information Science and Technology, 70*(9), 1026–1042.

Yang, G., Chen, L., Zhang, J., Wang, D.-R., & Zhang, H.-C. (2017). A mixture record linkage approach for US patent inventor disambiguation. In J. Park, S. C. Chen, & K. K. Raymond Choo (Eds.), *Advanced multimedia and ubiquitous engineering. Futuretech 2017, MUE 2017. Lecture notes in electrical engineering* (p. 448). Springer.

Zamzami, N., & Schiffauerova, A. (2015). Evaluating the science–technology interaction in nanotechnology: A simulation-based study. In *Proceedings of the Winter Simulation conference* (pp. 242–253).

Zhang, G., Liu, L., & Wei, F. (2019). Key nodes mining in the inventor–author knowledge diffusion network. *Scientometrics, 118*(3), 721–735.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator for interdisciplinarity of journals: taking similarity between subject fields into account. *Journal of the American Society for Information Science and Technology, 67*(5), 1257–1265.