



Evolution of research portfolios of universities and heterogeneity in the returns to research specialization

Adrià Plazas^{1,2*}, Solange Chavel², Roger Guimerà^{1,3*} and Marta Sales-Pardo^{1*}

Handling Editor: Johannes Wachs

*Correspondence:

adria.plazas@sirisacademic.com;
roger.guimera@urv.cat;
marta.sales@urv.cat

¹ Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain

Full list of author information is available at the end of the article

Abstract

Science is a dynamic process involving actors at different scales: from papers to countries going through authors and institutions. Nonetheless, the role of institutions - which is critical for assessing impact of some policies - has been largely understudied. To bridge this gap, we study the scientific portfolio of 26 institutions across the globe in the period 2000-2020. By analyzing 2.4 Mi papers, we find a correlation between portfolio dynamism and diversification, with more dynamic institutions tending to produce research across a broader range of non-related topics. Additionally, we find that not all research topics are equal in terms of impact: High-yielding topics show a significantly positive correlation between relative specialization and scientific impact per paper and are concentrated around a few research areas. Our results reveal robust patterns in university portfolio evolution that open the door to the development of mechanistic theories that explain the effect of university policies in the evolution of their portfolios.

Keywords: Science of Science; Complex Systems; Social Computational Sciences

1 Introduction

Science is a dynamic social phenomenon that continuously changes its theoretical frameworks by incorporating novel concepts arising from new evidence. In particular, the very subjects of scientific research change following a rise and decay pattern: research topics disappear in virtue of the emergence of new ones [1–4]. In this ever changing scientific landscape, researchers might decide to work on a specific topic with a certain team not only because of academic interest but also in response to different pressures including topic relevance [5], their current research environment [6] or the perceived benefits of strategic collaborations [7, 8]. Nonetheless, a general picture that explains not only the emergence of new ideas, but also how the specific context facilitates the adoption of new research topics and what are the consequences of these choices in terms of scientific and societal impact is still lacking.

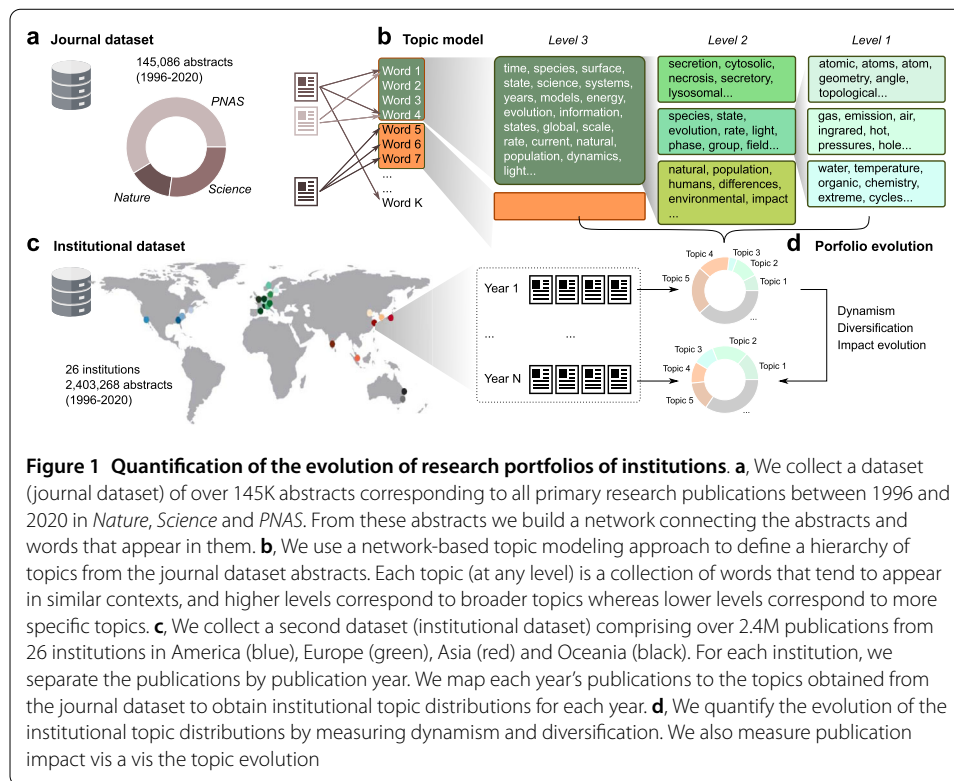
Partly, this is due to the fact that the scientific process involves multiple agents at different scales that play different roles in the generation and adoption of research topics.

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

For instance, at the level of individual papers, studies have found that disruptive, non-conventional science that combines concepts from diverse areas can lead to higher impact [9, 10] even though it might take longer to get noticed [4]. Instead, for individual researchers, scientific diversification does not necessarily pay-off in terms of the overall impact of their research [11]. At the country-level, research portfolio diversification is an indicator of scientific relevance, as nations that publish the most and receive the most citations are also those publishing in more diverse areas [12]. In fact, at the level of countries economic development and research diversification are tightly linked, so that developing countries follow a well-set diversification path by adding research areas going from the life sciences to the physical and engineering sciences to the social sciences [13].

In this context, a scale that has been largely understudied is that of research institutions. And yet, through their internal policies these agents play a key role in the way research topics evolve. Indeed, institutions have the power to decide what institutional equipment and what strategic research lines to invest on, decisions that can have a major impact in the evolution of their own research portfolio and their future innovation capabilities. Some studies have explored what is the role of universities and their prominence in shaping collaboration among scientists and how this is related to scientific impact [7, 14]. However, we do not know how institutions change their research portfolios and if these changes are related to the overall impact of their research output, a key aspect for institutional managers.

To cover this gap, we study the evolution of the scientific portfolio of 26 technical universities in America (5), Europe (13), Asia (6), and Oceania (2) in the period 2000-2020 and analyze how this evolution is related to scientific impact. In particular, we take advantage of the availability of publication data for universities around the world and quantify shifts in the language used in more than two million scientific articles [5, 15, 16]. We find that, as expected, universities change their portfolios over time and that, similarly to countries, institutions tend to diversify their portfolios. We also find that dynamism and diversification are correlated, and that universities change their portfolios by incorporating non-related research topics. However, we find that overall portfolio characterization is not sufficient to anticipate the impact of specific institutional strategies due to the diversity of effects of specialization/diversification strategies in different research areas. By analyzing individual research topics, we find that, in terms of investment returns (measured as the relative scientific impact per paper), not all research topics are equal: a small (but larger than expected) fraction of topics are high-yielding, that is, research topics for which the more specialized the institution is, the larger the relative impact per paper in that topic. Additionally these high-yielding topics are concentrated around a few research areas, thus indicating that there are specific scientific areas rather than specific topics where higher specialization is linked to higher returns. Overall, our findings thus reveal that there are some macroscopic non-random patterns in the way university portfolios evolve that link changes in portfolio to portfolio diversification, and at the same time that specialization in specific high-yielding areas has had increased impact returns. Our findings open the door to developing sound theories that explain the mechanisms behind these patterns and help us understand innovation processes within universities and research institutions at large.



2 Results

2.1 Dynamical institutions have a tendency to increase the diversity of their portfolios by exploring unrelated research topics

To investigate the evolution of institutional research portfolios, we consider a worldwide set of 26 diverse universities focused in the natural sciences and technology (Fig. 1 and Supplementary Table S1), and collect all the abstracts of the research papers published in English in which at least one of the authors is affiliated in one of these institutions during the period 1996-2020 (2,403,268 papers in total; see Methods for details).¹ For each article, we also collect the citation counts received until 2020. We select technically and natural science oriented institutions to ensure that published articles are representative of the majority existing research activities, which is not necessarily the case for institutions with a large focus in the social sciences and humanities in which research results are not necessarily disseminated as articles. Note that for our analysis of abstract content to be sound, we need universities that publish enough papers per year; in our analysis, we have fix that number to be at least 500 articles for each year within the study period. Among technically oriented institutions, only universities meet this condition, so we could not include other research oriented institutions in our analysis.

To quantify the evolution of the research portfolios of these institutions over time, we follow an approach that relies on defining a comprehensive space of topics of scientific knowledge onto which we can then project publications of institutions to define institutional portfolios and quantify temporal changes of those portfolios [6, 18]. To define a research topic space for the natural sciences that is comprehensive and not biased to-

¹ We do not collect papers before 1996 as there are digitization issues before that date and the coverage would be biased [17].

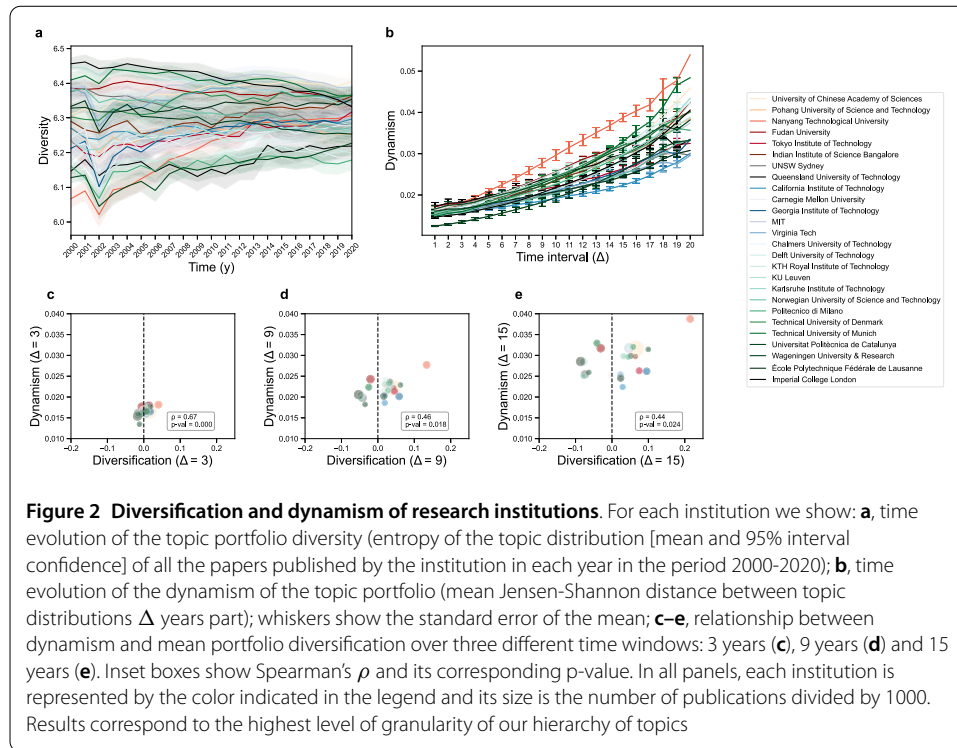
ward more recent years given the large growth rate in document production (11.6% annual growth rate in the corpus of 2.4M papers), we consider instead 145,086 abstracts published in English in *Nature*, *Science* and *PNAS* from 1996 until 2020. These are multidisciplinary journals that capture the most prominent advances in the natural sciences and technology. Additionally, the annual growth rate of article output in this corpus is moderate (1.6%), and therefore is a more adequate article corpus to identify research topics in the period 1996-2020 that avoids substantial bias toward recently published research.

We then use a network inference approach to topic modeling [19] to obtain a nested hierarchy of topics that defines a bottom-up map of scientific production (Fig. 1; see Methods, Supplementary Table S2). Note that to obtain these topics, we filter words that are commonplace and that cannot be unambiguously assigned to a specific topic, so that each topic is a list of informative words that typically appear in the same abstracts (Methods). Our topic map comprehensively represents the 2.4M papers in the institutional dataset, as less than 5% of abstracts have fewer than 10 words classified within the topics we identify (Supplementary Fig. S1). Our approach allows us to avoid relying on preexisting taxonomies that do not take into account the specificity of the dataset and are not useful for quantifying context-specific semantic shifts [20].

Given this space of topics, we can then characterize the research portfolio of institution k in year y by counting the number $w_i^k(y)$ of words in abstracts published by the institution in that year that belong to topic i . Note that because we have a hierarchy of topics, we can define institutional portfolios at any level in the hierarchy (Supplementary Table S2 and Supplementary File S1). In what follows, we focus in the most detailed topic level unless otherwise stated.

To characterize research portfolios, we first analyze the diversity of research topics. We measure portfolio diversity of a university k with the entropy of the topic distribution for each year, $S^k(y)$ (Fig. 2a; Methods), so that the more diverse the portfolio, the larger the entropy value. Our results show that the universities have diverse research portfolios and that for most of them portfolio diversity grows over the years, independently of the level of granularity of the topic distribution; for very few institutions entropy remains stable or has a small decrease (Supplementary Figs. S2, S3 a). Additionally, we observe a convergence to more similar values of portfolio diversity over time ($p=0.035$ - permutation test on the variance across universities of the entropy in year 2000 with respect 2020). The tendency to diversify scientific output over time is similar to what has been observed at country level and has been linked to economic development [13]. Indeed, this tendency for diversification has also been observed in the exports of countries [21] and the products of co-located industries in cities [22], which has often been linked to economic development and technological innovation [21].

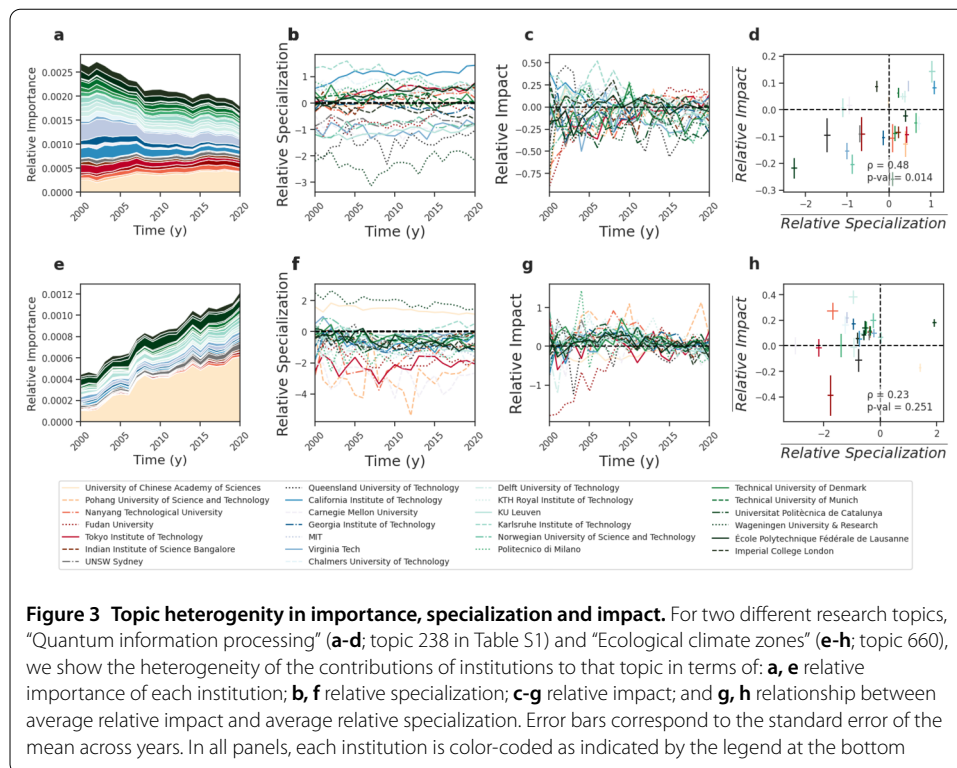
In this context, an issue that has received much attention is how the addition of novel exports/products occurs, and specifically whether it occurs by adding related or unrelated items to the existing portfolio [22, 23]. In our context, while it is clear that we cannot formalize the concept of product as in an economic context, the hierarchy of topics on which we define institutional portfolios establishes a clear measure of semantic relatedness in portfolio evolution which we investigate in what follows. Specifically, we quantify the dynamism of the research portfolios of institutions by measuring the average Jensen-Shannon distance between topic distributions Δ years apart. The more dynamic the institution, the larger the difference between portfolios over time windows Δ , and the larger



the Jensen-Shannon distance (Methods; Fig. 2b). Our results show that, over time and independently of the topic granularity and time window, all institutions change their portfolios by monotonically increasing their dynamism although some institutions change at faster rates than others (Supplementary. Figures S2 and S3). Note that this sustained increase discards the existence of short-term cycles in portfolios in which institutions could revert the changes after Δ years, as we would observe dips or a systematic flattening of the curves.

The dynamism we observe can be the result of different processes. For example, dynamism can arise due to loss of research topics leading to lower diversity or to gain of topics leading to higher diversity; or it can arise due to changes in topics without significant changes in diversity. And, in any scenario, changes in topics may be among different but closely related topics or among totally unrelated topics. To clarify how university portfolios evolve, we look at the relationship between the dynamism of an institution and its diversification, that is, the average change in diversity in windows of Δ years (Fig. 2c–d). We find that dynamism and diversification are significantly and positively correlated—the larger the diversification, the higher the dynamism of an institution. This correlation between dynamism and diversification is robust to changes in both the time window and topic granularity (Supplementary Figs. S2, S3).

These findings shed light on different questions related to portfolio evolution. First, our results show that institutions that loose diversity (negative diversification), tend to do it slowly (low dynamism). Second, we find that institutions with constant diversity (diversification close to zero) are nevertheless dynamic, indicating that there is change in the research topic portfolio; however, those changes are smaller than in portfolios of institutions with large diversification, which tend to be the most dynamic ones.



Finally, the fact that correlation between dynamism and diversification is significant at all levels in the hierarchy of topics indicates that the change and increase in diversity involves the exploration of new unrelated topics [24] instead of the exploitation of topics in related areas to those already present in the institution [23]. (Indeed, if dynamism was due to moves between closely related topics, the correlation between dynamism and diversification would disappear at higher levels of the topic hierarchy, that is, when considering fewer, more general topics encompassing related lower-level topics.)

Importantly, we find that there is no significant correlation between overall diversification and variables such as university size in terms of Full-Time Equivalent Researchers or age (Supplementary Material), therefore implying that there underlying mechanisms that explain these observations are not trivial.

2.2 Specialization in certain research areas correlates with higher scientific impact

Beyond the overall relationship between dynamism and diversification discussed so far, we now investigate the effect of research portfolio trajectories on research impact, as this is a metric that could ultimately be a quantifiable effect of institutional research policies related to diversification and specialization. At the country level, research portfolio diversification has been linked to greater scientific production and impact [12]. However, as in the Simpson paradox, global assessments (be it at the country level or at the institution level) can obscure trends associated to individual research areas, which are rather heterogeneous in terms of productivity and impact [25, 26]. This is particularly relevant because the relative contribution of each institution to the yearly production in a given research topic is highly variable over time and across topics (Fig. 3 a, e). Therefore, if we

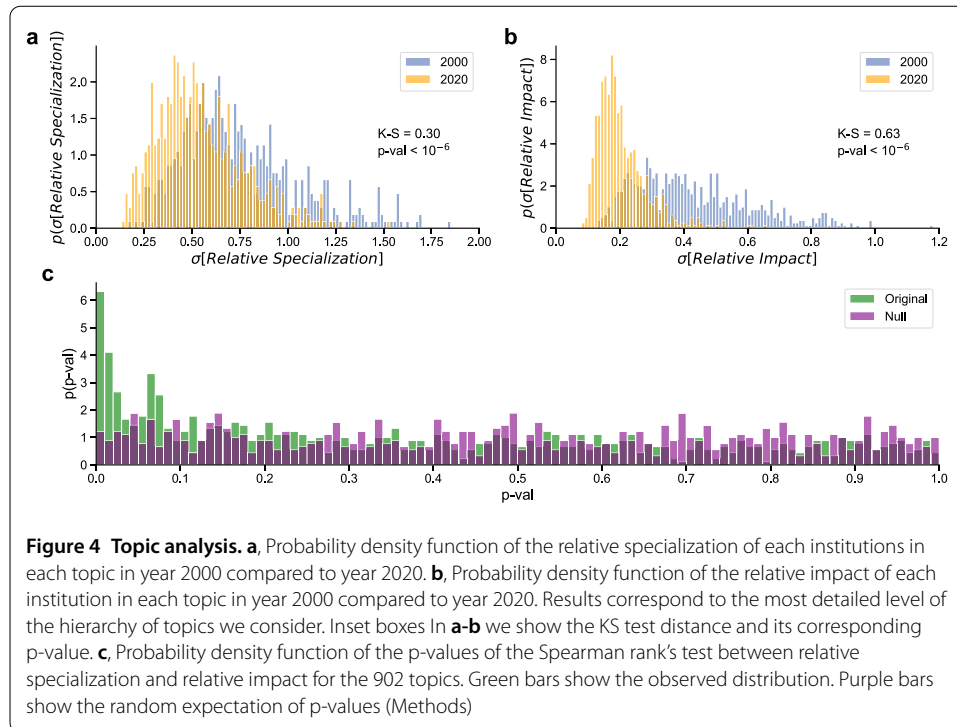
want to understand how institutions navigate through the space of research topics and the effects of specific research policies it is critical that we analyze the research contribution of institutions and its impact for each topic separately.

First, to quantify the level of specialization of an institution k in a research topic i in year y , we define its relative specialization $SP_i^k(y)$ as the log-ratio between the within-institution relative weight of a topic i (fraction of words belonging to that topic) and the overall relative weight of that topic (Methods; Fig. 3b, f and Supplementary Fig. S4). Positive log-ratios indicate that the relative weight of the topic within the institution exceeds that of the whole pool of institutions and, therefore, it is an indicator of the relative specialization of the institution in the research topic; negative log-ratios indicate under-representation of the topic in the institution. Our results in Fig. 3b, f show that there are typically a few specialist institutions for each topic, and that, while there are fluctuations, the majority of institutions maintain their specialist/non-specialist status over time.

Second, to measure scientific impact, we look at the number of citations received until 2020 by each article in our corpus. Then, for each institution and year, we measure the log-impact per paper on each topic (Methods). For year y , the relative impact per paper $RC_i^k(y)$ of institution k in topic i is the log-ratio between the log-impact per paper of that institution in that topic, and the log-impact per paper of all the institutions in that topic (Methods; Fig. 3c, g, and Supplementary Fig. S4). Positive $RC_i^k(y)$ indicates above-average log-impact per paper in that topic, and vice versa. We find that, unlike relative specialization, there is no clear tendency for one or a few institutions to systematically have more impact than the others (Fig. 3c, g).

Third, to understand whether relative specialization of an institution in a topic is related to higher impact in that topic, we analyze the correlation between the average specialization and the average relative impact over time (Fig. 3d, h). We find that for some topics like the topic labeled “Ecological climate zones” relative impact and relative specialization are not significantly correlated (Fig. 3h). However, in others like “Quantum information processing” specialization and relative impact are positively correlated (Fig. 3d)—we call these high-yielding topics. For these high-yielding topics, only specialized institutions tend to have positive relative impacts; if research specialization occurs as a response to institutional policies, such policies may then result in positive returns.

Given that individual topics have different behaviors, we finish by investigating all topics to establish what patterns related to specialization, impact and their correlation are globally outstanding. A common feature that we observe both in relative specialization and in relative impact (Fig. 3b, c, f, g) is that the variability across institutions decreases over time. Indeed, we find that there has been a tendency for the majority of topics to see institutions converge to more similar relative specializations and relative impacts (log-ratio closer to zero) (Fig. 4a–b). In particular, relative specialization has converged for 89% of the 902 topics (compared to 50% null expectation; $p\text{-value} < 10^{-6}$) and relative impact for 96% of them (compared to 50% null expectation; $p\text{-value} < 10^{-6}$). For 68% of the topics, both specialization and relative impact have converged (compared to 25% null expectation, $p\text{-value} < 10^{-6}$). A possible explanation of the observed convergence could be the irruption of Asian institutions as scientific leaders in the observed period (Supplementary Fig. S5). However, even after removing these institutions we still observe convergence in specialization and impact (Supplementary Fig. S6). Therefore, convergence seems to result from deeper and more general mechanisms.



Unlike this tendency to convergence, which affects the majority of topics, the correlation between relative impact and relative specialization observed in topics like “Quantum information processing” is specific to a small fraction of the topics which we call high-yielding topics: 6.3% (16.2%) of topics show a significant correlation between relative impact and relative specialization (Spearman ρ , $p < 0.01$ ($p < 0.05$)). Despite the fact that the existence of correlations is thus relatively rare, it is still much more frequent that one would expect by chance (Fig. 4 c). Furthermore, if we take into account the nested hierarchy of topics, high-yielding topics are not homogeneously distributed. The analysis of coarser research areas (level 2 in the hierarchy of topics; 39 topics) reveals that high-yielding topics tend to cluster around a few areas (Supplementary Table S4). Indeed, we find two research areas whose number of high-yielding topics is higher than expected by chance ($p < 0.05$, permutation test, Methods; Supplementary Table S4): Structural Biology & Molecular Chemistry (20/60 high-yielding topics) and Atomic, Electronic & Materials Physics (15/43 high-yielding topics).

3 Discussion

Our analysis reveals several robust patterns that characterize portfolio evolution of research intensive technical universities across the world. As we have already mentioned, some of these patterns bear resemblance with some observations of research portfolios at the level of countries that have been linked with economic development [12, 13] and from which we can establish some parallels with studies of the evolutions of product portfolios of companies and industries.

Nonetheless, as institutional portfolios are the sum of the articles published by individual researchers, our findings can also be understood in part as the result of decisions made by individual researchers and decisions made by the institutions themselves and in part as the outcome of specific institutional policies (that enabled hiring in the first place).

Following this line of thought, a striking pattern we observe is the convergence we observe across universities not only in portfolio diversity but also in relative specialization and relative impact of specific topics. Convergence in diversity could be compatible with studies showing that co-located researchers tend to collaborate more and over time become more similar research-wise [27]. However, this fact alone would not explain the convergence we observe on relative specialization and impact across institutions and topics.

Indeed, researchers often face a trade-off when choosing topics: working on well-established areas can increase visibility but also competition, while exploring new topics involves higher risk but potentially higher rewards [28–30]. Previous work suggests that researchers balance these forces by combining stable research lines with gradual exploration of new directions, although frequent shifts may reduce short-term impact due to a “pivot penalty” [31]. Further issues that might affect topic selection are the interdisciplinarity of the topic [9, 32] or perceived importance of the topics [33]. From this perspective, the coexistence of diversification as a general trend while maintaining specialization in some areas at the institutional level could reflect up to some extent the aggregation of these individual strategies and not only a single coordinated institutional plan.

Several mechanisms may underlie these dynamics that could have direct implications for policies. Topic choice is strongly shaped by funding systems and other external factors such as the construction of large infrastructures [34], which can bias attention toward specific areas and influence diversification or concentration strategies [35]. At the same time, institutions evolve through hiring, mobility, and collaboration networks: inflows of researchers introduce new expertise and reshape local research directions [27, 36]. Conversely, persistent specialization—particularly in high-yielding topics—likely on top of being a signature of maintaining past successful lines of research could also reflect long-term investments in infrastructure and capital-intensive research environments, which attract talent and reinforce cumulative advantage [34]. Importantly, our data do not allow us to distinguish whether portfolio changes arise from recruitment of new researchers or topic shifts of existing researchers, a distinction with fundamentally different policy implications that can be further assessed.

Our results should also be interpreted within their empirical scope: we focus on large, research-intensive institutions due to technical constraints on data, excluding smaller organizations with lower publication volumes. One of the strengths of our approach is that we rely on a topic space derived from natural science journals that allows us to avoid the effect of the exponential increase in the number of publications. However, this choice also limits our ability to generalize to non-technical universities.

4 Conclusions

We have studied research portfolios of 26 different natural science and technology oriented institutions over time, by analyzing their content of close to two million abstracts to quantify research dynamism, and explored its correlation with portfolio diversification and its implications in terms of scientific impact.

The analysis of institutional research portfolios reveals that there is a positive significant correlation between dynamism and portfolio diversification that is independent of the granularity of scientific topics we consider. This finding is interesting because of several reasons: first, because it shows that change in research portfolios comes from typically adding topics to the research portfolio and not just changing them; second, in this

diversification process research portfolios add new, semantically unrelated topics to those already present in the portfolio as the observed tendency is not level dependent. Therefore, our results show that, at an institutional level, dynamism happens via the exploration of unrelated scientific topics, instead of the exploitation of topics similar to existing ones.

These observations draw interesting parallels to what has been observed for country level scientific diversification [12, 13] and also to studies on the evolution of exports and products by countries or co-located industries [21–23, 37], which tend to diversify over time. Indeed, studies of international trade [37] show that while developing economies exploit the space of related exports to diversify, more mature economies also explore unrelated products. In this context, our results including the fact that we also observe a convergence of portfolio diversity as well as convergence of relative specialization and relative impact of individual topics across institutions might also be an indicator of the universities we study reaching a similar degree of scientific stability that translates into the way they update their research portfolios.

However, overall portfolio diversification and dynamism are not useful to quantify the impact of institutional policies. We show that to properly assess in detail the effects of specialization in research areas, we must look at the impact received by each research topic individually. In particular, we show that there are more than expected high-yielding topics for which knowledge specialization correlates with a greater relative scientific impact that concentrate around a few research areas. With our data we cannot assess the specific mechanisms that explain this pattern, but it opens the window for further exploration in future studies specially to determine whether institutional policies have an effect or whether this is a response to external factors such as funding, or the perception of importance of the topic.

Our results should be interpreted in the context of this study and the specific data used. This research examines English-language publications on emerging topics from top journals—PNAS, Science, and Nature—across 26 universities. These universities were chosen for their geographic diversity, size, and overall productivity. Further research should include private, government, and non-profit organizations as their motivations for conducting specific research programs might differ from those of universities. Our work also opens the way for a broader analysis of science, including fields beyond natural science and technology which will necessitate building a broad, unbiased corpus of documents from which it is possible to build a comprehensive space of topics.

Appendix: Data and methods

A.1 Data

From Open Alex,² a free and open catalog of the world's scholarly research system, we obtained two datasets:

- All the publications with abstracts in English published in *Nature*, *Science* and *PNAS* from 1996 to 2020 (both included),³ resulting in a set of 145,086 unique publications.
- Abstracts of indexed works produced by 26 different technical institutions (Supplementary Table S1). This dataset contains 2,403,268 unique publications with

²<https://openalex.org/about>.

³In order to overcome frontier effects we have focused our analysis in the period 2000 to 2020 (both included).

English abstracts from 1996 to 2020. For these publications, we also obtained all the citations received by each publication from the publication date until 2020.

A.2 Map of research topics

Topic modeling To define the map of research topics that characterize research portfolios, we use a complex networks approach to topic modeling [19]. This approach represents the content of a corpus of documents as a bipartite network of words and documents (abstracts in our case) and finds the hierarchical stochastic block model that best compresses the network [19] (implemented using the Graph-tools python library⁴). The result is a nested hierarchy of groups of unique words that constitute our map of research topics. Note that the choice of this algorithm is not arbitrary, this algorithm has been shown to outperform algorithms like LDA and other heuristic approaches at retrieving research topics from data [19, 38]. Moreover, the emerging topics (groups of unique words) are maximally compressive (it is impossible to design an algorithm based on the assumptions mentioned above that compresses the word content of articles more) and maximally predictive (we cannot design any algorithm that predicts better the word content of an article from knowledge of the group to which the article belongs). Finally, [19] naturally provides a hierarchy of topics (also automatically setting the number of topics for each level due to its ability to maximally compress), with the most detailed level of the hierarchy (the level we use) being the most compressive/predictive, and higher levels providing convenient ways to summarize the topics at the lowest level.

Topic Labels have been generated prompting in ChatGPT: “I have a set of documents describing a single thematic topic. If I provide you a list of the most representative keywords for each topic, sorted by importance in the topic, can you give me a short human-readable label for each topic? No need to justify the answer.”

Supplementary Table S2 and Supplementary File S1 show the number of topics per level in the hierarchy, the most ocurent words in each topic and the automatic generated label.

Text pre-processing In order to get an informative set of word topics from documents, it is important that first we clean-up de the text and identify what words or combinations of words we want to keep in a context dependent fashion (see [6, 20] for similar pipelines). To that end we carry out the following steps:

1. Identification of n -grams. We grouped and substituted statistically significant n -grams, that is we have identified groups of $n = 2, 3$ words that appear consecutively in the abstracts we consider more than expected by chance.
2. Removal of non-informative words. A typical approach is to remove a set of pre-defined stop words that are in principle devoid of semantic information. However, in a specific scientific context there are words that are ubiquitous and therefore devoid of meaning in this context. To identify words that carry information in a context dependent manner we use an information-theoretic approach [39], and select only those words that are most informative.

After the whole process, we end up with the 34,000 most informative words of the journal dataset, corresponding to a mean size of 40 words per abstract.

⁴<https://graph-tool.skewed.de/>.

A.3 Metrics

Portfolio of an institution We consider all abstracts published in institution k in year y , $a^k(y)$. Given a set of topics $\{i\}$ each abstract $a \in a^k(y)$ has w_{ai}^k words that belong to topic i (note that the explicit dependency on y is carried by a). Then we define the portfolio of an institution k in year y by and counting the number of words in abstracts $a \in a^k(y)$ that belong to topic i , $w_i^k(y) = \sum_a w_{ai}^k$.

In order to compute metrics derived from information theory, such as entropy or Jensen-Shannon distance, we have to take into account that we do not observe a distribution over topics, so we have to count the words. In a frequentist or maximum likelihood approach, one could estimate the distribution over topics of institution k in year y as $\rho^k(y) = \{\rho_i^k(y)\}$ with $\rho_i^k(y) = w_i^k(y) / \sum_i w_i^k(y)$. However, this estimation can lead to very biased estimations of information-theoretic quantities, especially if the number of observed words in a topic is low [40]. To avoid these issues, we follow the Bayesian approach in [40], and estimate $\rho_i^k(y)$ as

$$\rho_i^k(k) = \frac{w_i^k(y) + \beta}{w^k(y) + K} \quad (1)$$

where $w_y^k = \sum_i w_i^k(y)$, K is the number of topics in the level in the hierarchy we consider, and β is the parameter of a Dirichlet prior over ρ estimated using an empirical Bayes approach (see Ref. [40] for details and code).

Sampling factor All computations of the entropy or the Jensen-Shannon (JS) divergence, accounted the influence of non negligible biases in discrete distributions, particularly when the sample size deviates from the limit $N \gg K$, where N represents the number of words in the research portfolio and K denotes the number of word topics. The bias depends on the so-called sampling factor N/K [40]. Therefore, by controlling the sampling factor we can compare different entropy or JS measures. For each experiment, we have set a common sampling-factor by considering subsets of 500 publications per institution and year (about 2000 words per year). The final entropy or JS we report for each year and institution is the result of 1000 sub-samples of 500 publications.

Diversity To measure portfolio diversity of institution k in year y , we use entropy, S , an information-theoretic metric that quantifies distribution diversity,

$$S^k(y) = - \sum_i \rho_i^k(y) \log(\rho_i^k(y)) \quad (2)$$

with $\rho_i^k(y)$ defined as in Eq. (1).

Dynamism To compute the dynamism of an institution, we compute the difference between distributions $\rho^k(y)$ Δ years apart using the Jensen-Shannon distance,

$$JS^k(y, \Delta) = \frac{1}{2} (S^k(y) + S^k(y + \Delta)) - \tilde{S}^k(y, \Delta), \quad (3)$$

where $\tilde{S}^k(y, \Delta)$ is the entropy of the average distribution $\tilde{\rho}^k(y, \Delta) = \frac{1}{2} (\rho^k(y) + \rho^k(y + \Delta))$.

Then we define the dynamism D as,

$$D^k(\Delta) = \frac{1}{|Y|} \sum_{y \in Y} JS^k(y, \Delta), \quad (4)$$

where Y is the set of years in our dataset for which there exists data for y and $y + \Delta$.

Relative importance We define the relative importance of the contribution to topic i of institution k in year y as

$$RI_i^k(y) = \frac{w_i^k(y)}{\sum_{k',j} w_j^{k'}(y)}, \quad (5)$$

where $\sum_{k,j} w_j^k(y)$ is the contribution of all institutions in a given year y .

Relative specialization For a given topic i and year y , we define relative specialization $SP_i^k(y)$ as the log-ratio between the within-institution relative contribution to topic i in year y compared to the relative contribution of the whole corpus to topic i in year y .

$$SP_i^k(y) = \log \frac{\frac{w_i^k(y)}{\sum_j w_j^k(y)}}{\frac{\sum_{k'} w_i^{k'}(y)}{\sum_{k',j} w_j^{k'}(y)}} \quad (6)$$

Note that positive log ratios indicate that in institution k and year y , the relative contribution to that topic within the institution is larger than that in the overall output of the institutions we consider.

Note that in both this metric and relative impact (below), we choose to make the comparisons using only articles published by the institutions we consider rather than using the background corpus to ensure that the metrics capture cross-institutional variation rather than differences between the background corpus and the institutional corpus.

Relative impact Our goal is to compare the impact per paper of institution k on topic i (in a given year y) to the overall impact per paper on that topic (in the same year). We start by considering all articles $a^k(y)$ published by the institution in year y , and the logarithm of the citations received by those articles, $\{lc_a^k; a \in a^k(y)\}$. (Note that we use the logarithm of the number of citations because the number of citations can span several orders of magnitude; using logarithms avoids impact metrics being dominated by one or a few highly-cited papers [41].) We then estimate the impact $lc_i^k(y)$ of topic i by assigning to the topic a fraction of the citations of each publication $a \in a^k(y)$, proportionally to the weight $\tilde{w}_{ai}^k$ of that topic in the abstract of the publication

$$lc_i^k(y) = \sum_{a \in a^k(y)} \tilde{w}_{ai}^k lc_a^k \quad \text{with} \quad \tilde{w}_{ai}^k = \frac{w_{ai}^k}{\sum_j w_{aj}^k}. \quad (7)$$

In this way, we account for the extent to which each article contributes to multiple topics.

Because total impact is related to total output and we are interested in impact per paper, we divide the log-impact by the total contribution of the institution to the topic, $p_i^k(y) = \sum_{a \in a^k(y)} \tilde{w}_{ai}^k$, to obtain an effective log-impact per paper

$$c_i^k(y) = \frac{lc_i^k(y)}{p_i^k(y)}. \quad (8)$$

Note that, by taking into account all institutions we can define the same overall impact per paper of topic i as

$$c_i(y) = \frac{\sum_k lc_i^k(y)}{\sum_k p_i^k(y)}. \quad (9)$$

Finally, we define the relative impact of the scientific production in topic i of institution k in year y as the log-ratio between the log-impact per paper of the institution and the overall log-impact per paper

$$RC_i^k(y) = \log \left(\frac{c_i^k(y)}{c_i(y)} \right). \quad (10)$$

A.4 Hypothesis testing

Testing correlation among pairs of variables To investigate general monotonic (in general non-linear) correlation patterns between variables such as relative specialization and relative impact, we use Spearman's rank correlation coefficient ρ . We computed the coefficient and its associated p-value using the SciPy Python package.

Null model for Spearman's rank correlation significances To assess whether the p-values we observed from Spearman's rank correlation tests were expected because of the underlying distributions of citation counts, we constructed a null expectation by randomly shuffling the citation counts of each paper. For each randomization, we recomputed the relative impact and obtained p-values for each Spearman's rank test. We repeated this process 10 times to obtain an expected distribution of p-values (Fig. 4c).

Null model to detect concentration of high-yielding topics in specific research areas To assess whether high-yielding topics are concentrated in specific research areas (i.e. parent topics), we estimate the distribution of expected high-yielding topics per research area. To that end, we randomly assign topics to research areas and compute the number of high-yielding topics for each research area. We repeat this procedure 10,000 times. We then compare the observation with the distribution of the random expectation and obtain a p-value for each research area. Note that p-values differ from those obtained assuming a binomial distribution of high-yielding topics with parameters $p = 0.2$ because the number of high-yielding topics is fixed.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-026-00669-8>.

Additional file 1. (PDF 1.8 MB)

Additional file 2. (XLSX 1.1 MB)

Acknowledgements

We thank Open Alex for making research data available. This research was funded by projects PID2022-142600NB-I00 and DIN2020-010974 (A.P.) from MCIN/AEI/10.13039/501100011033 FEDER, UE, and by the Government of Catalonia (2021SGR-633).

Author contributions

A.P. collected data and wrote code. All authors designed research. A.P., M. S.-P. and R.G. analyzed results. All authors wrote the manuscript.

Funding information

This research was funded by projects PID2022-142600NB-I00 and DIN2020-010974 (A.P.) from MCIN/AEI/10.13039/501100011033 FEDER, UE, and by the Government of Catalonia (2021SGR-633).

Data availability

Data availability All data are available as described in the Methods section. Code Availability The code for the topic modeling is available through the repository associated to [Gerlach, M., Peixoto, T. P. & Altmann, E. G. A network approach to topic models. *Sci. Adv.* (2018).]. The code for data analysis will be available at the time of publication

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain. ²SIRIS Lab, Research Division of SIRIS Academic, 08003 Barcelona, Catalonia, Spain. ³ICREA, 08007 Barcelona, Catalonia, Spain.

Received: 20 October 2025 Accepted: 17 May 2026 Published online: 20 May 2026

References

1. He Q, et al (2009) Detecting topic evolution in scientific literature: how can citations help? In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09. Association for Computing Machinery, New York, pp 957–966. <https://doi.org/10.1145/1645953.1646076>
2. McMahan P, McFarland DA (2021) Creative destruction: the structural consequences of scientific curation. *Am Sociol Rev* 86:341–376. <https://doi.org/10.1177/0003122421996323>
3. Wu L, Wang D, Evans J (2019) Large teams develop and small teams disrupt science and technology. *Nature*
4. Lin Y, Evans JA, Wu L (2022) New directions in science emerge from disconnection and discord. *J Informetr* 16:101234. <https://doi.org/10.1016/j.joi.2021.101234>
5. Lou W, Meng J (2023) The diversity of canonical and ubiquitous progress in computer vision: a dynamic topic modeling approach. *Inf Process Manag* 60:103238. <https://doi.org/10.1016/j.jipm.2022.103238>
6. Danus L, Davis RH, Guimera R, Sales-Pardo M (2024) Gender and the influence of research environment in topic selection of early-career faculty in stem. *arXiv:2407.19607*
7. Jones BF, Wuchty S, Uzzi B (2008) Multi-university research teams: shifting impact, geography, and stratification in science. *Science* 322:1259–1262. <https://doi.org/10.1126/science.1158357>
8. Sekara V, et al (2018) The chaperone effect in scientific publishing. *Proc Natl Acad Sci USA* 115:12603–12607. <https://doi.org/10.1073/pnas.1800471115>
9. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342:468–472. <https://doi.org/10.1126/science.1240474>
10. PLoS ONE 10:e0135095. <https://doi.org/10.1371/journal.pone.0135095>
11. Abramo G, D'Angelo CA, Di Costa F (2019) Diversification versus specialization in scientific research: which strategy pays off? *Technovation* 82–83:51–57
12. Cimini G, Gabrielli A, Sylos Labini F (2014) The scientific competitiveness of nations. *PLoS ONE* 9:e113470. <https://doi.org/10.1371/journal.pone.0113470>
13. Miao L, et al (2021) The latent structure of global scientific development. *Nat Hum Behav* 6:1206–1217
14. Burghardt K, He Z, Percus A, Lerman K (2021) The emergence of heterogeneous scaling in research institutions. *Commun Phys*
15. Milojević S, Sugimoto CR, Yan E, Ding Yg (2011) The cognitive structure of library and information science: analysis of article title words. *J Am Soc Inf Sci Technol* 62:1933–1953. <https://doi.org/10.1002/asi.21602>
16. Kuhn T, Perc Mcv, Helbing D (2014) Inheritance patterns in citation networks reveal scientific memes. *Phys Rev X* 4:041036. <https://doi.org/10.1103/PhysRevX.4.041036>
17. Task Force on Archiving of Digital Information (1996) Preserving digital information: report of the Task Force on Archiving of Digital Information. Tech. Rep., Commission on Preservation and Access and the Research Libraries Group Washington, DC. Also available through ERIC, ED395602
18. Barron A, Huang J, Spang R, DeDeo S (2018) Individuals, institutions, and innovation in the debates of the French revolution. *Proc Natl Acad Sci USA*

19. Gerlach M, Peixoto TP, Altmann EG (2018) A network approach to topic models. *Sci Adv*
20. Font-Pomarol L, et al (2023) Socially disruptive periods and topics from information-theoretical analysis of judicial decisions. *EPJ Data Sci*
21. Sarin V, Mahapatra SK, Sood N (2022) Export diversification and economic growth: a review and future research agenda. *J Public Aff* 22:e252. <https://doi.org/10.1002/pa.2524>
22. Boschma R, Balland P-A, Kogler DF (2014) Relatedness and technological change in cities: the rise and fall of technological knowledge in us metropolitan areas from 1981 to 2010. *Ind Corp Change* 24:223–250. <https://doi.org/10.1093/icc/dtu012>
23. Hidalgo C, et al (2018) The principle of relatedness. *Springer proceedings in complexity*. Springer, pp 451–457
24. Foster J, Rzhetsky A, Evans J (2015) Tradition and innovation in scientists' research strategies. *Am Sociol Rev* 80(5):875–908
25. Stringer MJ, Sales-Pardo M, Amaral LAN (2008) Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE* 3:1–8. <https://doi.org/10.1371/journal.pone.0001683>
26. Duch J, et al (2012) The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS ONE* 7
27. Catalini C (2018) Microgeography and the direction of inventive activity. *Manag Sci* 64:4348–4364. <https://doi.org/10.1287/mnsc.2017.2798>
28. Leahey E (2007) Not by productivity alone: how visibility and specialization contribute to academic earnings. *Am Sociol Rev* 72:533–561. <https://doi.org/10.1177/000312240707200403>
29. Stremersch S, Verniers I, Verhoef PC (2007) The quest for citations: drivers of article impact. *J Mark* 71:171–193. <https://doi.org/10.1509/jmkg.71.3.171>
30. Aleta A, Meloni S, Perra N, Moreno Y (2019) Explore with caution: mapping the evolution of scientific interest in physics. *EPJ Data Sci* 8:27. <https://doi.org/10.1140/epjds/s13688-019-0205-9>
31. Hill R, Yin Y, Stein C, et al (2025) The pivot penalty in research. *Nature* 642:999–1006. <https://doi.org/10.1038/s41586-025-09048-1>
32. Xiang S, Romero DM, Teplitskiy M (2025) Evaluating interdisciplinary research: disparate outcomes for topic and knowledge base. *Proc Natl Acad Sci USA* 122:e2409752122. <https://doi.org/10.1073/pnas.2409752122>
33. Xiang S, Romero D, Teplitskiy M (2025) Research topic choice: motivations, strategies, and consequences. *Quant Sci Stud* 6:623–651. https://doi.org/10.1162/qss_a_00365
34. Helmers C, Overman HGM (2017) My precious! The location and diffusion of scientific research: evidence from the synchrotron diamond light source. *Econ J* 127:2006–2040. <https://doi.org/10.1111/eoj.12387>
35. Hoppe TA, Litovitz A, Willis KA, et al (2019) Topic choice contributes to the lower rate of nih awards to African-American/black scientists. *Sci Adv* 5:eaaw7238. <https://doi.org/10.1126/sciadv.aaw7238>
36. Duede E, Teplitskiy M, Lakhani K, Evans J (2024) Being together in place as a catalyst for scientific advance. *Res Policy* 53:104911. <https://doi.org/10.1016/j.respol.2023.104911>
37. Nomaler O, Verspagen B (2024) Related or unrelated diversification: what is smart specialization? *Struct Chang Econ Dyn* 70:503–515. <https://doi.org/10.1016/j.strueco.2024.05.019>
38. Grootendorst M (2022) Bertopic: neural topic modeling with a class-based tf-idf procedure. *arXiv preprint*. [arXiv:2203.05794](https://arxiv.org/abs/2203.05794)
39. Gerlach M, Shi H, Amaral L (2019) A universal information theoretic approach to the identification of stopwords. *Nat Mach Intell*
40. Piga A, Font-Pomarol L, Sales-Pardo M, Guimerà R (2024) Bayesian estimation of information-theoretic metrics for sparsely sampled distributions. *Chaos Solitons Fractals* 180:114564. <https://doi.org/10.1016/j.chaos.2024.114564>
41. Stringer MJ, Sales-Pardo M, Amaral LAN (2010) Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *J Am Soc Inf Sci Technol* 61:1377–1385. <https://doi.org/10.1002/asi.21335>

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.