nature human behaviour

Article

https://doi.org/10.1038/s41562-024-02041-0

Limited diffusion of scientific knowledge forecasts collapse

Received: 15 April 2023

Accepted: 1 October 2024

Published online: 2 December 2024

Check for updates

Donghyun Kang[1,2], Robert S. Danziger[3,4,5], Jalees Rehman[3,6,7] & James A. Evans[1,2,8,9]

Market bubbles emerge when asset prices are driven unsustainably higher than asset values, and shifts in belief burst them. We demonstrate an analogous phenomenon in the case of biomedical knowledge, when promising research receives inflated attention. We introduce a diffusion index that quantifies whether research areas have been amplified within social and scientific bubbles, or have diffused and become evaluated more broadly. We illustrate the utility of our diffusion approach in tracking the trajectories of cardiac stem cell research (a bubble that collapsed) and cancer immunotherapy (which showed sustained growth). We then trace the diffusion of 28,504 subfields in biomedicine comprising nearly 1.9 M papers and more than 80 M citations to demonstrate that limited diffusion of biomedical knowledge anticipates abrupt decreases in popularity. Our analysis emphasizes that restricted diffusion, implying a socio-epistemic bubble, leads to dramatic collapses in relevance and attention accorded to scientific knowledge.

Market bubbles emerge when widespread opinions about an asset, such as housing or securities, create self-reinforcing information that drives its price much higher than its value to society[1]. These bubbles are characterized by a swift surge in popularity, fuelled by beliefs that the value may continue to rise and persist, leading to speculation. Such bubbles burst when shifts in opinion, often catalysed by new data or events, precipitate radical discounts in pricing[2]. Science observers and researchers themselves have drawn parallels in science[3–5], which involves considerable investment in capital, attention and other resources based on highly uncertain knowledge about the outcomes of research. This exposes science to the risk of forming bubbles analogous to financial markets[4,5]. Here we operationalize the concept of scientific bubbles and their collapse, proposing a measurement framework and demonstrating that ideas and findings in science can experience abrupt booms and busts of popularity and credibility that may yield adverse consequences for science and scientists alike.

In the system of biomedical knowledge, citation counts have come to function as an operational currency[6,7], serving as a measure of the importance and impact of scientific work. This is also reflected by increasing interest in the development of indicators tracing emergent, disruptive or breakthrough science and technology[8–12], which typically incorporate citation counts as key components. The citation metric manifests some distortion, however, from the inflation of citation counts with historical growth in articles[13] and the unequal size of fields[14]. Inspired by the analogy between financial and scientific bubbles, here we forecast substantial and dramatic declines in the popularity of research ideas—the bursting epistemic bubbles—as the degree to which those ideas remain concentrated within the same collection of authors, institutions and biomedical subfields, failing to diffuse across social and scientific space despite initial popularity. We argue that this limited diffusion may indicate inflated attention to particular ideas that may not generalize or withstand broader scrutiny,

[1]Department of Sociology, University of Chicago, Chicago, IL, USA. [2]Knowledge Lab, University of Chicago, Chicago, IL, USA. [3]Division of Cardiology, Department of Medicine, University of Illinois College of Medicine, Chicago, IL, USA. [4]Department of Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [5]Department of Physiology and Biophysics, University of Illinois at Chicago, Chicago, IL, USA. [6]Department of Biochemistry and Molecular Genetics, University of Illinois, College of Medicine, Chicago, IL, USA. [7]University of Illinois Cancer Center, Chicago, IL, USA. [8]Santa Fe Institute, Santa Fe, NM, USA. [9]Paradigms of Intelligence Team, Google, Mountain View, CA, USA. e-mail: jevans@uchicago.edu

ultimately leading to disappointment and disillusionment within the scientific community.

Consider the extreme but illuminating case of cardiac regeneration in biomedicine. Dr Piero Anversa and collaborators led research in cardiac regeneration at the turn of the twenty-first century by asserting the possibility of regenerating damaged heart muscle tissue after myocardial infarction with stem cells and progenitor cells drawn from the bone marrow or within the heart[15]. During Anversa and collaborators' peak productivity, they also exercised substantial influence over the research narrative[16], sitting on editorial boards of high-profile American Heart Association journals such as 'Circulation Research' and an interlocking matrix of NIH grant review panels[15], serving on the NIH National Institute on Aging's Board of Scientific Counselors (2008–2013)[17]. Nevertheless, findings from early cardiac regeneration work not only failed to generalize, but the experiments could not be replicated by other researchers[18]. Followed by Anversa's forced departure from Harvard in December 2015[19], Harvard Medical School and Brigham and Women's Hospital announced in October 2018 the recommendation to retract more than 30 papers from leading journals due to falsified and/or fabricated data[20]. This coincided with a marked discount in citations to the subfield and diminished confidence in the prospects of cardiac regeneration with resident heart stem cells[21]. This, in turn, adversely impacted even those researchers who had been studying cardiac regeneration using more rigorous scientific approaches and had identified reproducible mechanisms underlying the phenomenon[22].

Our approach, however, aims to generalize beyond the severe research misconduct of an individual or a team of scientists. Accurate and honestly reported medical findings can still fail to generalize beyond the specific context of their initial investigation, despite optimism and hype regarding their transformative potential for medicine. More critically, as highlighted by science commentators[23] and biomedical researchers[24,25], unintended collective failures can also occur, as exemplified by the widespread use of misidentified or contaminated cell lines contributing to unjustified hype and misdirected attention and resources in the field. This phenomenon suggests the need for a more refined and multifaceted framework to better model and evaluate the trajectories of scientific attention.

In this study, we demonstrate that fragile and overhyped biomedical findings could have been anticipated by analysing their diffusion through the system of science. Utilizing PubMed Knowledge Graph (PKG)[26], a large-scale bibliographical database, we provide a framework that considers distances between publications and their citing papers within the 'scientific space' constituted by co-investigated biomedical entities and the 'social space' constituted by collaborating scientists. Specifically, we develop a diffusion index to capture whether ideas have been amplified within social and scientific bubbles[27], or diffused more widely and tested for robustness across diverse research communities[28]. This approach allows us to gain insight into the diffusion of research ideas and their impact, ultimately helping us to more rapidly assess the value and potential of scientific findings.

Our work demonstrates how a lack of diffusion measured by this framework—indicative of the existence of a scientific bubble—can anticipate a rapid decline in popularity as confidence bubbles burst. Applying the conceptual and measurement tools detailed below (Methods), we first compare two distinct trajectories from cardiac stem cell and cancer immunotherapy research papers. Figure 1a,b illustrates our approach with two contrasting papers. Figures 1a,b depicts the diffusion and citation trajectories of an early paper[29] from Dr Anversa's group on cardiac muscle regeneration using bone-marrow-derived cells and a seminal paper on cancer immunotherapy conducted by Dr Honzo[30] within scientific and social spaces, respectively. Figure 1a suggests that while cardiac stem cell research similar to this paper gained massive early attention, this was not sustained. This is contrasted in Fig. 1b with the case of cancer immunotherapy, where

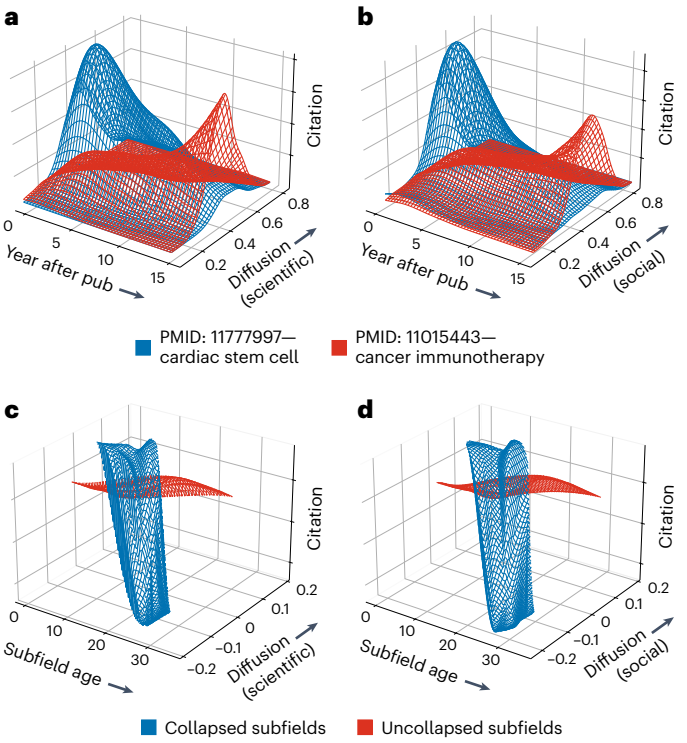


Fig. 1 | Representation of different diffusion levels and contrasting diffusion trajectories. a,**b**, 3D kernel density plots of diffusion indices and citations for PMID 11777997 (cardiac stem cell) and PMID 11015443 (cancer immunotherapy) in scientific (**a**) and social (**b**) spaces. Publication years associated with each article are aligned to zero for comparison. pub, publications. **c**,**d**, Kernel density plots based on average diffusion indices and citations, standardized within subfield ages. These plots contrast subfields that experienced collapse below the 0.5% threshold (blue) with those that did not (red), across scientific (**c**) and social (**d**) spaces respectively.

research gradually diffused to distant research groups and topics before garnering substantial attention.

Beyond papers, we trace the diffusion trajectories of 28,504 unique subfields in biomedicine through publications from elite biomedical researchers[31], encompassing nearly 1.9 million papers and more than 80 million citations. Our analysis reveals that limited diffusion of biomedical knowledge is systematically associated with an early rise and abrupt drop in popularity. Figure 1c,d displays the average trajectories of subfields by distinguishing those that experienced a sharp decline or collapse in scientific attention from those that did not by the end of 2019 (Methods). Furthermore, our post hoc analyses show that the likelihood of collapses of subfields is positively associated with the concentration of publications from elite scientists, echoing aspects of the Dr Anversa case.

In this way, our work highlights that restricted diffusion in science can effectively capture socio-epistemic bubbles. Complementing citation dynamics with diffusion patterns enriches our identification of robust insight in biomedical science, which can be readily improved by discounting bubbles and promoting convergent results sourced through social and topical diversity.

Results

Contrasting trajectories of two research papers

Applying neural embedding models to MEDLINE data enables us to project all biomedical research articles onto scientific and social manifolds[32]. As detailed in Methods and Supplementary Section 2, this allows us to locate their relative positions within collaborative

networks of scientists and biomedical entities through research. The cosine or angular distances between citing and cited research measured over social and scientific spaces aggregate into straightforward, continuous metrics of diffusion. To demonstrate the effectiveness of our approach utilizing these scientific and social spaces, we examine trajectories of two highly cited publications at the individual paper level, each drawn from cardiac stem cell and cancer immunotherapy research, respectively.

Our first case is a research article published (PMID: 11777997) in the New England Journal of Medicine in 2002[29]. Led by Dr Piero Anversa, this research supported the existence of substantial numbers of endogenous myocardial stem and progenitor cells, proposing their potential to regenerate heart muscle. This line of research initially received substantial attention, nearly 500 citations within PKG by 2007 because it suggested new possibilities for heart regeneration after severe myocardial infarctions involving massive tissue loss. This claim was later called into question by several researchers outside the Anversa network[33,34], eventually leading to the retraction of more than 30 papers by 2018 from claims of data fabrication and scientific malpractice[35].

Conversely, the second example, an article (PMID: 11015443) published in the Journal of Experimental Medicine in 2000[30] represents a study by a team of pioneering researchers in the field of cancer immunotherapy. Their work focuses on the inhibition of negative immune regulation and its implications for cancer treatment. The publication and subsequent work spurred the development of a broad spectrum of cancer immunology and immunotherapy research initiatives across many research groups and countries globally, laying the groundwork for what has become one of the most impactful innovations in cancer treatment.

In Fig. 1a,b, we visualize the contrasting temporal trajectories of these two publications in size of attention and diffusion within the scientific and social spaces, respectively, using three-dimensional (3D) kernel density estimation (see Extended Data Fig. 1 for 2D heat maps). Dr Anversa's publication experienced a meteoric rise in citations early on, accumulating 55.4% of its total citations (887 citations by the end of 2019) within the first 5 years after publication in 2002, before experiencing a sharp decline in attention. Nevertheless, our analysis of diffusion trends up to 2019 reveals no significant correlation between the time elapsed (in years) between later citing papers and our diffusion measures across either scientific ($\rho = 0.029$, $P = 0.396$) or social ($\rho = 0.005$, $P = 0.880$) space, indicating a lack of measurable diffusion.

In contrast, the article on cancer immunotherapy, which demonstrated the potential to inhibit negative immune regulation in treating cancer[36,37], gained early attention at a slower pace than the Anversa publication, accumulating 11.6% of its total citations (2,469 by 2019) in 5 years following its debut in 2000. The evolution of its diffusion metrics presents a very different picture. The ideas from the cancer immunotherapy paper diffused over time, as indicated by significant positive correlations between cosine distance and year difference in both scientific ($\rho = 0.482$, $P < 0.001$) and social spaces ($\rho = 0.470$, $P < 0.001$), suggesting that it was increasingly cited by more diverse teams and subfields. This culminated in the Drs Tasuku Honjo and James P. Allison receiving the 2018 Physiology and Medicine Nobel Prize for advancing the scientific understanding of cancer immunotherapy[38].

These contrasting cases demonstrate how our diffusion metric accounting for epistemic bubbles offers a more nuanced understanding of scientific influence than traditional citation counts, capturing the complex dynamics of diffusion through social and scientific spaces and its potential consequences.

Knowledge concentration anticipates collapse

We elevate our analysis to the level of scientific subfields to systematically test the generalizability of our approach. We apply our framework to 28,504 unique biomedical subfields curated in ref. 31. Each subfield encompasses a compactly defined set of biomedical research articles using the PubMed Related Article (PMRA) algorithm[39] applied to a given seed article. This algorithm underpins the official PubMed interface, serving as a pivotal tool for researchers to locate articles related to a focal research paper, which has been fruitfully used in various studies, such as repercussions of scientific scandal on careers[40], shifts in research focus among scientists responding to NIH funding changes[41], and the negative impact of winning prizes for recipient competitors[42]. The subfields identified by this approach enable us to analyse diffusion dynamics, epistemic bubbles and collapses of scientific attention beyond selective, high-profile papers. Specifically, if work from a focal subfield is predominantly cited by research in close social and scientific proximity, the subfield's insights may not diffuse despite its seeming popularity and could retain inflated value due to local reinforcement. In other words, we anticipate that substantial and dramatic declines in the popularity of research ideas, conceptualized as knowledge 'bubbles bursting', can be predicted by the degree to which these ideas, despite their apparent popularity, have failed to diffuse across the social and scientific space via citations.

Our primary outcome of interest is 'bubble bursting' or collapse, defined as an abrupt decline in the relevance of a given subfield of science. We time a bubble burst by comparing the standardized citation difference that a subfield garners in a given year to its performance 2 years earlier, marking if it falls below an extreme threshold. This approach allows us to distinguish subfields that experienced deflationary bursts from those that did not by using each standardized citation count difference against the values derived from 28,504 unique subfields (see Methods for details). We use the bottom 0.5% of the distribution of standardized citation differences as our threshold, which captures 4,480 out of 28,504 unique subfields as experiencing a collapse. To ensure the robustness of our results, we also apply thresholds of 0.25% and 0.1%, identifying 2,297 and 918 collapsed subfields, respectively, and report the results from parallel analyses using these thresholds throughout the following analyses and in the Supplementary Information. Across these operationalizations, the subfields that experienced a collapse also experienced a significant positive deviation from expected citation rates preceding the collapse (Supplementary Section 4.4). Fields that experience a disproportionate deflation experienced a previous inflation. In short, bubbles burst.

We compute our knowledge diffusion indices, our main predictors, for each subfield across scientific and social spaces. We identify papers published that reference at least one article within each subfield. We then calculate the average cosine distances between the referenced articles in each subfield and the citing papers with 2-year rolling windows, separately for scientific and social spaces to measure scientific and social diffusion (Methods).

Using a non-parametric Kaplan–Meier survival model to predict the probability of bubble bursting, our estimation reveals the knowledge diffusion index as a leading signal preceding a sudden collapse in attention. We employ a 1-year lag for our diffusion measures when associating them with the outcome of interest, collapse of attention. By splitting our observations into three groups with diffusion percentiles ranked by calendar year and subfield age—the bottom 10th percentile, the top 10th percentile and the middle between them—Fig. 2 visualizes that diffusion in the social space forecasts the bursting of attention bubbles captured by the 0.5% threshold. For instance, at the subfield age of 20, 90.9% of subfields in the top 10th percentile of diffusion (95% CI: 89.9%–92.0%), 86.3% of subfields in the 10th to 90th percentile range (95% CI: 85.8%–86.7%) and 77.5% of subfields in the bottom 10th percentile (95% CI: 76.0%–79.0%) avoided a burst, not experiencing a major drop in citation attention. This suggests that low diffusion rates may signal poor long-term subfield survival, while high diffusion is linked to better long-term survival, helping subfields avoid extreme deflationary events.

https://doi.org/10.1038/s41562-024-02041-0

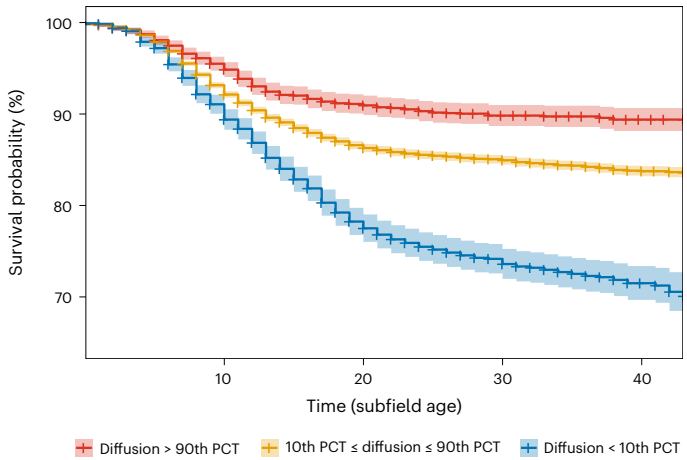


Fig. 2 | Survival probability against bubble bursting as a function of knowledge diffusion in social space. Events are defined as a sharp decline in 2-year citation counts at the subfield level with a 0.5% cut-off (see Methods). Survival refers to the converse, that is, not experiencing a subfield-level extreme deflationary event. Subfield ages are set to 0 in the year when the focal seed article spanning the subfield was published. Diffusion percentiles are ranked within calendar years and subfield ages. The crosses denote the estimated mean survival probabilities at a given subfield age, and the bands depict 95% confidence intervals.

Table 1 | Model estimates using the bottom 0.5% cut-off for citation differences in a 2-year rolling period

Dependent variable	Substantial decline of citations				
	Estimate	s.e.	*t*	*P* value	95% CI
Diffusion					
Scientific space	−0.204	0.039	−5.228	<0.001	[−0.280, −0.128]
Social space	−0.277	0.049	−5.598	<0.001	[−0.373, −0.181]
Subfield growth pattern					
Cum. subfield size (logged)	0.499	0.084	5.923	<0.001	[0.334, 0.664]
2-year subfield marginal growth	−0.217	0.010	−21.543	<0.001	[−0.237, −0.197]
Citation dynamics					
Cum. citations (logged)	−2.472	0.164	−15.003	<0.001	[−2.793, −2.151]
2-year citations (logged)	2.772	0.146	18.942	<0.001	[2.486, 3.058]
Gini coef. of cum. citations	0.012	0.006	1.843	0.065	[0.000, 0.024]
Gini coef. of 2-year citations	−0.026	0.006	−4.710	0.001	[−0.038, −0.014]
Other controls					
Retraction notice published	0.062	0.199	0.313	0.754	[−0.328, 0.452]
After death	0.152	0.129	1.186	0.236	[−0.101, 0.405]
After death × elite scientist death	−0.138	0.086	−1.601	0.109	[−0.307, 0.031]
log-likelihood	−26,289.5				
Total observations	1,313,433				

Coefficients for fixed effects of field age, calendar year and strata ID dummies are omitted. Variables under 'Knowledge diffusion', 'Subfield growth pattern' and 'Citation dynamics' are all 1-year lagged. The diffusion indices are standardized within field ages and calendar years across 28,504 subfields. Standard errors are clustered by strata ID and calendar year. The *P* values are for two-sided tests. Cum., cumulative; coef., coefficient.

We confirm this pattern, presented in Figs. 1 and 2, with discrete-time event history models that allow us to consider temporal covariates, including field size and growth rate, total cumulative citations, citation concentration across papers, paper retractions and unexpected deaths of elite scientists (see Methods). Our analysis consistently shows that the lower a paper's diffusion of influence, the greater the hazard that the subfield will experience an abrupt collapse of attention (Table 1 and Supplementary Table 1.1). For example, a reduction in diffusion in social space of one standard deviation above to one below the mean translates into a 74.02% (95% CI: 43.61%–110.85%) increase in the odds of experiencing a major reduction in scientific attention, accounting for subfield age, calendar year and other covariates. Supplementary Tables 1.2 and 1.3 in Supplementary Information show the estimations based on 0.25% and 0.1% thresholds to identify burst subfields.

Overall, we observe that limited social diffusion was more strongly linked to the likelihood of subfield collapse than scientific diffusion. We posit that this probably stems from tacit confounders in research that emerge when conducted by a concentrated, connected group of scientists. When close-knit groups perform research under uniform assumptions, methodologies and even shared resources, their findings are less likely to replicate among outsiders[28,43]. By contrast, the applicability of verified scientific findings across different biomedical domains may vary. A therapy's effectiveness for treating breast cancer is undiminished by its irrelevance for heart disease. But the failure of findings to diffuse across different groups of scientists in the same area indicates a limitation of their published scientific knowledge.

We conduct a series of subsequent analyses to gain deeper insight into characteristics of socio-epistemic 'bubbles' and consequences of their collapse. Our analysis revealed that the importance of elite scientists within a subfield, as quantified by the proportion of their publications per subfield, is positively correlated with the likelihood of collapse compared with subfields that did not burst (with the 0.5% threshold for the sudden collapse, $\beta = 1.222$; 95% CI: 1.001–1.443; $P < 0.001$; with 0.25% threshold, $\beta = 1.299$; 95% CI: 1.038–1.560; $P < 0.001$; with 0.1% threshold, $\beta = 1.320$; 95% CI: 0.961–1.680; $P < 0.001$) (Supplementary Table 4.1). Subfields dominated by the work of elite scientists[31] are more likely to have their early findings overhyped and subsequently 'burst' than subfields less influenced by those dominating elite scientists. When elite scientists' findings do not diffuse correspondingly with their seeming popularity, this indicates that others have attempted to generalize their results and failed. Correspondingly, we find a positive association between the fraction of NIH funding allocated to collaborators of elite scientists and the likelihood of attentional collapse (with the 0.5% threshold for the sudden collapse, $\beta = 0.224$; 95% CI: 0.110–0.337; $P < 0.001$; with 0.25% threshold, $\beta = 0.194$; 95% CI: 0.037–0.350; $P = 0.015$; with 0.1% threshold, $\beta = 0.231$; 95% CI: 0.004–0.458; $P = 0.046$) (Supplementary Table 4.2). This suggests that limited diffusion and subsequent collapses may be correlated with the concentration of 'scientific capital' in terms of reputation and resources[44], as exemplified by the stem cell cardiac case discussed above. We also examine the relationship between epistemic bubbles and the limits of clinical translation. These complementary analyses demonstrate how bubbles that subsequently burst are less likely translated into clinical applications (with the 0.5% threshold for the sudden collapse, $\beta = -0.509$; 95% CI: −0.672 to −0.347; $P < 0.001$; with 0.25% threshold, $\beta = -0.650$; 95% CI: −0.875 to −0.424; $P < 0.001$; with 0.1% threshold, $\beta = -0.658$; 95% CI: −1.047 to −0.268; $P < 0.001$) (Supplementary Table 4.3).

To evaluate the implications of epistemic bubbles and bursts, we compare the productivity of authors who published their articles close to the time of collapse (for example, authors who published in 2001, 2002 or 2003 when the collapse was measured in 2003) to those who published in the same subfield at an earlier time (for example, on or

https://doi.org/10.1038/s41562-024-02041-0

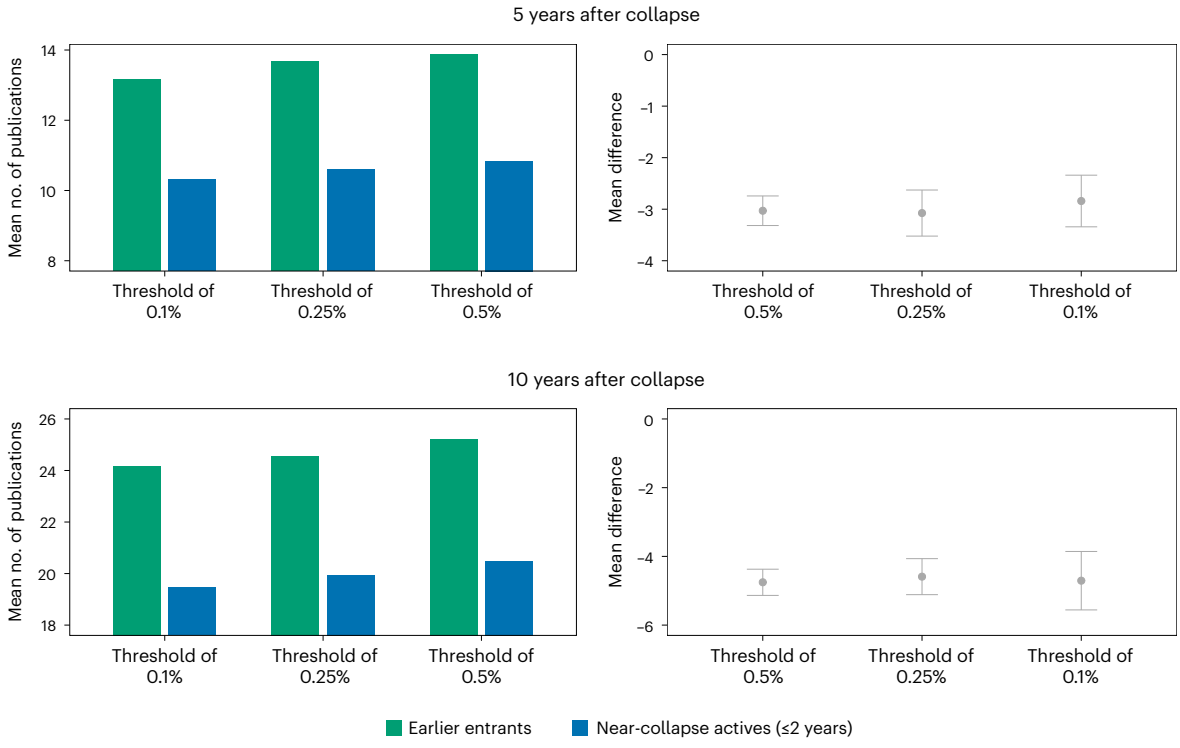


Fig. 3 | Comparison of author productivity in collapsed subfields 5 and 10 years post collapse. Top: the numbers of collapsed subfields included in the comparison 5 years after the collapse are 3,910, 1,983 and 773 for the 0.5%, 0.25% and 0.1% thresholds, respectively. Bottom: for 10 years after the collapse, the numbers are 3,605, 1,818 and 711 for the 0.5%, 0.25% and 0.1% thresholds, respectively. Error bars represent the 95% confidence intervals for the mean differences in average publication numbers. Comparisons are drawn between authors who entered the field early and those active near the collapse, based on paired *t*-tests to account for the grouped nature of the data by subfield. Extended Data Table 1 provides detailed statistical information.

before 2000). As shown in Fig. 3 and Extended Data Table 1, findings suggest that those who entered right before collapse were significantly less productive in the mean number of publications both 5 and 10 years after collapse, compared with early entrants. This suggests that subfield collapse may shape researchers' reputations and career outcomes.

We also consider the implications of bubbles for the allocation of research funding. We trace the average number of new grants acknowledged per year in papers across subfields. Our analysis shows that more than 80% of the subfields that experienced a substantial decrease in scientific attention acknowledged new grants after collapse. By the end of 2019, the median number of such grants was 6, as detailed in Extended Data Table 2. Figure 4 illustrates the trends from 15 years before to 10 years after the collapse. It shows that while peaks in the average number of new funding grants occur ~7 years before collapse, the rate at which funding decreases after the burst is markedly slower than the observed trend. The projection suggests that no funding would have been allocated 3 years after the burst. This pattern suggests a substantial lag by which money continues to support research that the broader biomedical community may perceive as less scientifically and clinically relevant.

Discussion

Current metrics of scientific attention and confidence pay scant attention to patterns of research consumption and diffusion across diverse people, institutions, disciplines, regions and beyond. This lack of consideration can lead to an incomplete understanding of a research field's true impact and potential. Our knowledge diffusion index contrasts with and complements citation counts, the conventional unit of scientific credit. Citations alone are blind to who, where and how far across the landscape of science those building on research reside, but our diffusion index provides a more comprehensive view.

Our finding suggests that constricted diffusion, captured by limited cosine distances between focal and citing papers measured across scientific and social spaces, signals an epistemic bubble. In this way, our index of limited diffusion represents an indicator of future declines in relevance and attention accorded to scientific and biomedical knowledge. Researchers can anticipate the collapse of biomedical approaches years before their occurrence by systematically tracking the diffusion of their ideas across scientists and biomedical areas. In addition, science and biomedical policy that analyses knowledge diffusion patterns can anticipate such collapses and may reduce their occurrence by incentivizing and accounting for diverse, disconnected support for robust scientific and medical claims[28].

Similar to other methods aimed at quantitatively evaluating research impact, our framework for measuring diffusion and its implementation should not replace the holistic judgement of research quality. Furthermore, while we draw on the concept of 'bubbles' in science, analogous to those in financial markets, it is worthwhile to recognize their unique aspects in the context of science. For example, small, dense research networks may be crucial for initiating high-risk projects at early stages despite a high probability of failure. In addition, scientific bubbles may not always arise from speculation but could result from authentic scientific enthusiasm or localized beliefs in a promising research direction.

Nevertheless, our finding holds implications for biomedical researchers, science-based industries and science policymakers. By accounting for diffusion and diversity, funding agencies can spot bubbles and adjust resource allocation by diversifying groups of researchers sponsored for a particular research topic. Research information platforms like PubMed, OpenAlex, the Web of Science or Google Scholar could also incorporate strong, leading signals from which analysts can anticipate the future relevance of current research. A

https://doi.org/10.1038/s41562-024-02041-0

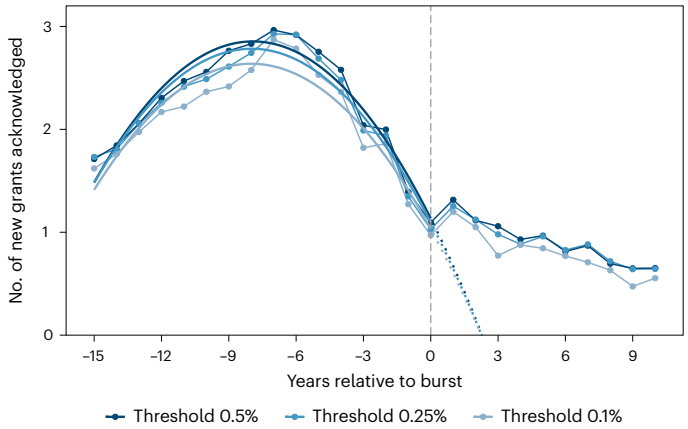


Fig. 4 | The average number of new grants acknowledged in collapsed subfields by years relative to burst. The number of collapsed subfields, using the 0.5%, 0.25% and 0.1% thresholds, is 4,480, 2,297 and 918, respectively. The quadratic fit is applied to data from years −15 to 0 relative to the burst year, with dotted lines representing extrapolations starting from year 0 onwards.

high diffusion index indicates that trending insights are more likely robust than fragile. Regular self-assessments of knowledge diffusion could enable individual researchers, teams and labs to better gauge the robustness and future impact of their work. Further, documenting associations between scientific knowledge diffusion and its applications, as in the translation of biomedical research from bench to clinic, can better inform science policy.

Our results draw on subfields identified in academic science using a particular delineation of research subfields. Nevertheless, our analysis demonstrates robust evidence for the wisdom of diverse crowds in science and technology to sustain advance. It underscores the importance of both social and scientific diversity for robust evaluation of an idea's relevance to science as a whole. Moreover, our proposed framework for measuring diffusion may extend to other domains of knowledge, such as the spread of misinformation, by allowing us to measure diversity in information consumption[45]. In social media, algorithmic metrics that account for diversity in diffusion would be far less susceptible to strategic, concentrated efforts seeking to misclassify information as a legitimate, widespread trend (for example, on Facebook's Newsfeed), just as they would decrease the intentional or unintentional illusion of scientific support.

In this way, we demonstrate the importance of idea diffusion for advancing scientific knowledge, its ability to transfer across broad science communities, and the relevance of these signals for forecasting robust ideas upon which to build novel and critical scientific and biomedical knowledge. Ultimately, our analysis underscores the relative importance of identifying the path of an idea's consumption over its point of production for predicting lasting, far-reaching impact. Accounting for this will enable the design of wise and diverse research, development and clinical crowds, leading to improved research policy, greater reproducibility and more sustained impact on future knowledge.

Methods

Manifold representations of social and scientific space

To assess the diffusion of ideas in science from biomedicine, we train two high-dimensional vector representations using neural embedding models[32] for publications catalogued in the PubMed Knowledge Graph (PKG)[26]. The PKG provides 15,530,165 disambiguated author IDs and 481,497 unique combinations of Medical Subject Headings (MeSH) from 29,339 MeSH descriptors and 76 qualifiers, each assigned to 28,329,992 and 26,666,615 MEDLINE-indexed publications, respectively, by the end of 2019. Each document in the PubMed database is assigned a unique document identifier, PMID. The database also contains the publications to the publication reference records, which integrates PubMed's citation data, NIH's open citation collection, OpenCitations and the Web of Science.

We specifically adapted the Doc2vec model[32], a variant of the Word2vec model[46], originally developed to produce dense vector representations for documents or paragraphs from the words that compose them. This approach has previously been extended to generate high-dimensional representational vectors geometrically proximate to the degree that entities frequently share neighbours, contexts[46–48], or are connected via social ties[49,50].

We considered that a biomedical research article can be characterized by a list of: (1) MeSH terms and (2) research collaborators. Consequently, we built two separate representational vector spaces to capture 'scientific space' and 'social space', respectively. For training our vector representations, we utilized the Python Gensim package[51]. We specifically used the Distributed Bag of Words (DBOW) model, analogous to the skip-gram model from the Word2vec framework, and simultaneously trained the vector position of constituting elements (MeSH terms or author IDs) along with document vectors. This resulted in two spaces trained on 100-dimensional vector representations for PMIDs and their constituent elements. Training and validation procedures are detailed in Supplementary Section 2.

Delineating biomedical subfields

Biomedical knowledge obtains influence when others recognize and build on it[44,52]. In this work, we sought to understand the dynamics of diffusion and shifting attention at the level of biomedical subfields, which we defined as a group of biomedical publications tightly related to a medically and biologically relevant research topic, identified through the PubMed similar article function powered by PubMed Related Algorithm (PMRA)[39]. This method has been previously employed in studies examining the impact of publication retraction[53], repercussions of scientific scandal on careers[40], shifts in research focus by scientists in response to NIH funding changes[41], negative impacts from prize winning on recipient competitors[42] and consequences of the premature death of elite life scientists[31] on subfields.

We specifically used the 28,504 unique seed articles curated in ref. 31, derived from publications by 'elite biomedical scientists'[6,54]. Applying the similar article function provided in PubMed enabled us to capture over 1.9 million unique articles associated with these subfields published through 2019. We then extracted ~86.8 million paper-to-paper citations identified by PKG based on them. A more comprehensive illustration of the original data source and our extension is available in Supplementary Section 3.1. To ensure robustness, we performed complementary analyses that redefined subfields on the basis of the position of papers within our scientific embedding space, resulting in the same pattern of findings. Details and results are reported in Supplementary Sections 3.2 and 3.3.

Model

Using a non-parametric Kaplan–Meier model and discrete-time event history model, we related the annual diffusion indices for each subfield calculated across social and scientific spaces with an abrupt decline in the relevance of a given subfield, or 'bubble burst,' as illustrated in Fig. 1a. Formally, the discrete-time event history analysis model can be written as:

$$\log\left(\frac{p_{ti}}{1-p_{ti}}\right) = \alpha D_{ti} + \beta x_{(t-1)i}. \quad (1)$$

p_{ti} denotes the probability of event happening at t for subfield i,
D_{ti} denotes time dummies corresponding to t with coefficients α,
x_{ti} is a vector for covariates (time varying and constant over time) with coefficients β.

The results reported in Table 1, based on equation (1), were derived from a binary outcome variable analysed through logistic regression, where a normality check required for ordinary least squares regression is not applicable. However, we accounted for potential violations of independence by clustering standard errors by strata ID and calendar year, as discussed below.

Bubble bursting as an outcome event. Our primary outcome of interest is the event of socio-epistemic bubbles bursting, characterized by an abrupt decline in popularity of a given subfield, which we measured as the decline in citation counts as illustrated in Fig. 1. Specifically, we timed bubble bursts on the basis of when the standardized citation count difference of a given year from a subfield fell below extreme cut-offs within the life cycle of each subfield. This required distinguishing subfields that experienced deflationary bursting, or collapse, from those that did not. We achieved this through the following steps.

We first computed $\Delta_{i,t} = c_i(t) - c_i(t-2)$, where $c_i(t)$ is the citations that a subfield i garnered during year t across 1970 to 2019. Unlike ref. 31 that used publications indexed both in Web of Science and MEDLINE, we used all PMID to PMID citation links identified in the PKG 2020 data to compute citation counts. (We included all MEDLINE-indexed publications, even when MeSH terms or author disambiguated IDs were not assigned to them.) Then, we standardized $\Delta_{i,t}$ within the life cycle of each subfield to make the $\Delta_{i,t}$ values comparable across 28,504 subfields. This was achieved by transforming $\Delta_{i,t}$ to $z_{i,t}$ by subtracting the mean of $\Delta_{i,t}$, $\underline{\Delta}_i = \frac{1}{N}\sum \Delta_{i,t}$ from $\Delta_{i,t}$ and dividing it by the standard deviation of $\Delta_{i,t}$ computed within a subfield. By doing so, we obtained the distribution of the standardized 2-year citation difference, $z_{i,t}$, across 28,504 subfields. The distribution of $z_{i,t}$, with the range of [−5.2, 5.52], is presented in Extended Data Fig. 2.

We operationalized bubble bursts as when the standardized citation count difference for a given year in a subfield, $z_{i,t}$, fell below extreme cut-offs, such as 0.5%, 0.25% or 0.1% of the distribution. To qualify a decline as a burst, we required that the average of $z_{i,t}$ after the drop must be negative, ensuring a continued loss of attention. In addition, the peak citation count at the subfield level should not occur in 2019, the final year of our dataset. If a subfield experienced more than one sharp decline, we considered the year with the most substantial one as the time of the burst. We note that bursts are preceded by bubbles: fields that experienced these extreme drops also manifested greater than expected citations before collapse (Supplementary Section 4.4).

Using the 0.5% cut-off (that is, $z_{i,t} < -2.64$) identified 4,480 subfields (15.7% of 28,504 subfields) that experienced a sharp decline in collective scientific attention relative to other subfields. Applying the 0.25% ($z_{i,t} < -2.91$) and 0.1% ($z_{i,t} < -3.26$) cut-offs returned 2,297 and 918 subfields with the bubble bursting events, respectively. Extended Data Fig. 3 contrasts three examples of subfields that did not experience these bubbles and bursts (top panels) with three examples that exhibited substantial declines in attention (bottom panels), according to our procedure described above.

Knowledge diffusion as a key indicator. The key leading indicator for our analysis is subfield-level knowledge diffusion. We measured knowledge diffusion by employing a 2-year rolling window approach. For each year, we identified papers published either in that year or the preceding year referencing at least one article published within a given subfield. We then separately calculated the average cosine distances (or 1-cosine similarity) between these focal articles in the subfield and the citing papers in our scientific and social spaces. This consideration led us to measure two diffusion indices: (1) diffusion across 'scientific space' and (2) diffusion across 'social space'. In our model, we incorporated a 1-year lag to assess the association between diffusion dynamics and the subsequent decline in citations.

Further characterization of subfield dynamics. To account for subfield dynamics captured by our diffusion indices, we consider the following variables in our models.

Time effect. The difference between calendar years and the year seed articles were published was captured using *subfield age* dummies, included for each subfield up to the end of 2019. This approach controls for trends related to the age of the subfield without imposing a functional form.

Subfield growth pattern. *Cumulative subfield size* is the total number of articles published in a subfield up to a given year. This measure controls for the potential impact of a subfield's size on citation dynamics. We applied a logarithmic transformation to address skewness for robust statistical comparisons between subfields of varying sizes.

Two rolling-years marginal growth is the proportion of articles published in the current year and the previous year, divided by cumulative subfield size. This metric provides a normalized indicator of how actively a subfield is growing, shrinking or remaining stagnant, adjusting for short-term fluctuations in publication activity that might affect outcomes of interest, such as citation dynamics.

Citation dynamics. *Total cumulative citations* are the aggregate citation counts that publications within a subfield have received until a specified year. We included this variable to control for the overall academic impact of a subfield, which may influence the likelihood of sudden changes in citation patterns. A natural logarithmic transformation was applied to address skewness.

Two-year rolling citation counts are the citations a subfield accumulates during the given year and the past year. We took the natural logarithm of the raw counts. This variable controls for the recent volume of citations, separate from long-term trends.

The Gini coefficient of citation counts measures the degree of centralization in citation counts within a subfield. The coefficient ranges from 0 (where every article in a subfield receives the same number of citations) to 1 (where a single article receives all citations). We computed the Gini coefficients for (1) total cumulative citations and (2) 2-year rolling citation annually to control for the potential impact of citation concentration.

Other controls. *Article retraction notification* is an indicator variable that switches from 0 to 1 once a retraction notification is observed in a subfield. It controls for the potential impact that experiencing a retraction event in the subfield level might have on overall attention the subfield receives.

After death (of elite scientists) is an indicator variable that switches from 0 to 1 with the death of elite scientists (Supplementary Section 3.1). This attempts to capture any residual temporal effects of elite scientists' death on citation dynamics[31].

After death (of elite scientists) × subfields associated with premature death of elite scientists controls for the impact of the sudden death of elite scientists on citation dynamics, reflecting how the dataset was originally constructed and the finding that elite death is positively associated with increases in subfield citation[31]. The first term is as previously described; the latter is an indicator variable that differentiates subfields associated with the premature deaths of elite scientists from those that are not.

Calendar year fixed effect is the year dummies to account for potential effects of the calendar year from 1970 to 2019. We included this to ensure that any time-specific external influences are controlled across all subfields.

Strata ID is the 3,076 'strata' IDs identified from the subfields associated with publications of prematurely deceased elite scientists[31]. These IDs were assigned to comparable 'within strata' subfields that were not experiencing a loss of elite scientists. These comparable

subfields were matched with those experiencing the loss of an elite scientist on the basis of key metrics such as (1) publication years, (2) team sizes, (3) ages of associated scientists and (4) long-run citation impact, as detailed in Supplementary Section 3.1.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

This work uses the PubMed Knowledge Graph[26] (http://er.tacc.utexas.edu/datasets/ped) and the replication data from ref. 31 (https://www.openicpsr.org/openicpsr/project/116188/version/V1/view;jsessionid=EA1E1E5A6DAB42737EE54A5F5DD4B069). Source data are provided with this paper.

Code availability

The data and code used for the figures and tables are available in GitHub[55].

References

1. Arthur, W. B. Complexity in economic and financial markets. *Complexity* **1**, 20–25 (1995).
2. Harras, G. & Sornette, D. How to grow a bubble: a model of myopic adapting agents. *J. Econ. Behav. Organ.* **80**, 137–152 (2011).
3. Goldman, A. I. & Shaked, M. An economic model of scientific activity and truth acquisition. *Philos. Stud.* **63**, 31–55 (1991).
4. Pedersen, D. B. & Hendricks, V. F. Science bubbles. *Philos. Technol.* **27**, 503–518 (2014).
5. Evans, J. P., Meslin, E. M., Marteau, T. M. & Caulfield, T. Genomics. Deflating the genomic bubble. *Science* **331**, 861–862 (2011).
6. Fortunato, S. et al. Science of science. *Science* **359**, eaao0185 (2018).
7. Partha, D. & David, P. A. Toward a new economics of science. *Res. Policy* **23**, 487–521 (1994).
8. Small, H., Boyack, K. W. & Klavans, R. Identifying emerging topics in science and technology. *Res. Policy* **43**, 1450–1467 (2014).
9. Funk, R. J. & Owen-Smith, J. A dynamic network measure of technological change. *Manage. Sci.* **63**, 791–817 (2016).
10. Klavans, R., Boyack, K. W. & Murdick, D. A. A novel approach to predicting exceptional growth in research. *PLoS ONE* **15**, e0239177 (2020).
11. Weis, J. W. & Jacobson, J. M. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nat. Biotechnol.* **39**, 1300–1307 (2021).
12. Lin, Y., Evans, J. A. & Wu, L. New directions in science emerge from disconnection and discord. *J. Informetr.* **16**, 101234 (2022).
13. Petersen, A. M., Pan, R. K., Pammolli, F. & Fortunato, S. Methods to account for citation inflation in research evaluation. *Res. Policy* **48**, 1855–1865 (2019).
14. Hutchins, B. I., Yuan, X., Anderson, J. M. & Santangelo, G. M. Relative Citation Ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS Biol.* **14**, e1002541 (2016).
15. Taylor, M. & Heath, B. Years after Brigham–Harvard scandal, U.S. pours millions into tainted stem-cell field. *Reuters* (21 June 2022).
16. Anversa, P., Kajstura, J., Leri, A. & Bolli, R. Life and death of cardiac stem cells: a paradigm shift in cardiac biology. *Circulation* **113**, 1451–1463 (2006).
17. 2009 Current Fiscal Year Report: Board of Scientific Counselors, National Institute on Aging. *The Federal Advisory Committee Act (FACA) Database* (Department of Health and Human Services, 2009); https://www.facadatabase.gov/FACA/apex/FACACommitteeLevelReportAsPDF?id=a10t0000001h2ObAAI
18. Murry, C. E. et al. Haematopoietic stem cells do not transdifferentiate into cardiac myocytes in myocardial infarcts. *Nature* **428**, 664–668 (2004).
19. Vrotsos, L. W. Harvard Medical School requests retractions for former professor's research. *The Harvard Crimson* (16 October 2018).
20. Oransky, I. & Marcus, A. Harvard and the Brigham call for more than 30 retractions of cardiac stem cell research. *STAT News* (14 October 2018).
21. Davis, D. R. Cardiac stem cells in the post-Anversa era. *Eur. Heart J.* **40**, 1039–1041 (2019).
22. Osafune, K. et al. Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat. Biotechnol.* **26**, 313–315 (2008).
23. Harris, R. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions* (Basic Books, 2017).
24. Neimark, J. Line of attack. *Science* **347**, 938–940 (2015).
25. Hughes, P., Marshall, D., Reid, Y., Parkes, H. & Gelber, C. The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? *Biotechniques* **43**, 575–586 (2007).
26. Xu, J. et al. Building a PubMed knowledge graph. *Sci. Data* **7**, 205 (2020).
27. Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K. & Evans, J. The sociology of scientific validity: how professional networks shape judgement in peer review. *Res. Policy* **47**, 1825–1841 (2018).
28. Belikov, A. V., Rzhetsky, A. & Evans, J. Prediction of robust scientific facts from literature. *Nat. Mach. Intell.* **4**, 445–454 (2022).
29. Quaini, F. et al. Chimerism of the transplanted heart. *N. Engl. J. Med.* **346**, 5–15 (2002).
30. Freeman, G. J. et al. Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192**, 1027–1034 (2000).
31. Azoulay, P., Fons-Rosen, C. & Zivin, J. S. G. Does science advance one funeral at a time? *Am. Econ. Rev.* **109**, 2889–2920 (2019).
32. Le, Q. & Mikolov, T. Distributed representations of sentences and documents. In *Proc. 31st International Conference on Machine Learning* (eds Xing, E. P. & Jebara, T.) 1188–1196 (PMLR, 2014).
33. Laflamme, M. A. & Murry, C. E. Regenerating the heart. *Nat. Biotechnol.* **23**, 845–856 (2005).
34. van Berlo, J. H. et al. C-kit+ cells minimally contribute cardiomyocytes to the heart. *Nature* **509**, 337–341 (2014).
35. Chien, K. R. et al. Regenerating the field of cardiovascular cell therapy. *Nat. Biotechnol.* **37**, 232–237 (2019).
36. Mellman, I., Coukos, G. & Dranoff, G. Cancer immunotherapy comes of age. *Nature* **480**, 480–489 (2011).
37. Finck, A., Gill, S. I. & June, C. H. Cancer immunotherapy comes of age and looks for maturity. *Nat. Commun.* **11**, 3325 (2020).
38. Smyth, M. J. & Teng, M. W. 2018 Nobel Prize in physiology or medicine. *Clin. Transl. Immunol.* **7**, e1041 (2018).
39. Lin, J. & Wilbur, W. J. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics* **8**, 423 (2007).
40. Azoulay, P., Bonatti, A. & Krieger, J. L. The career effects of scandal: evidence from scientific retractions. *Res. Policy* **46**, 1552–1569 (2017).
41. Myers, K. The elasticity of science. *Am. Econ. J. Appl. Econ.* **12**, 103–134 (2020).
42. Reschke, B. P., Azoulay, P. & Stuart, T. E. Status spillovers: the effect of status-conferring prizes on the allocation of attention. *Adm. Sci. Q.* **63**, 819–847 (2018).
43. Danchev, V., Rzhetsky, A. & Evans, J. A. Centralized scientific communities are less likely to generate replicable results. *eLife* **8**, e43094 (2019).

44. Bourdieu, P. The specificity of the scientific field and the social conditions of the progress of reason. *Soc. Sci. Inf.* **14**, 19–47 (1975).
45. Kim, J., Wang, Z., Shi, H., Ling, H.-K. & Evans, J. Individual misinformation tagging reinforces echo chambers; collective tagging does not. Preprint at https://arxiv.org/abs/2311.11282 (2023).
46. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **26**, 3111–3119 (2013).
47. Kozlowski, A. C., Taddy, M. & Evans, J. A. The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
48. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).
49. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk: online learning of social representations. In *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 701–710 (Association for Computing Machinery, 2014).
50. Grover, A. & Leskovec, J. node2vec: scalable feature learning for networks. *KDD* **2016**, 855–864 (2016).
51. Rehurek, R. & Sojka, P. Software framework for topic modelling with large corpora. In *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks* 45–50 (Univ. of Malta, 2010).
52. Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.* **80**, 875–908 (2015).
53. Azoulay, P., Furman, J. L. & Murray, F. Retractions. *Rev. Econ. Stat.* **97**, 1118–1136 (2015).
54. de Solla Price, D. J. *Little Science, Big Science—and Beyond* (Columbia Univ. Press, 1963).
55. Kang, D. Limited diffusion of scientific knowledge forecasts collapse. *GitHub* https://github.com/Donghyun-Kang-Soc/limited_diffusion (2024).

Acknowledgements

We acknowledge funding from the Fetzer Franklin Fund in association with the 2019 MetaScience Symposium (R.S.D., J.A.E., D.K.), the Air Force Office of Scientific Research (AFOSR: FA9550-19-1-0354 and FA9550-15-1-0162) (J.A.E., D.K.) and the National Science Foundation (NSF: 1829366 and 1800956) (J.A.E., D.K.). The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This work was completed in part with resources provided by the University of Chicago's Research Computing Center. We also appreciate the support from J. Xu and Y. Ding in facilitating access to PubMed Knowledge Graph.

Author contributions

D.K., R.S.D., J.R. and J.A.E. conceptualized the project. D.K. and J.A.E. developed the methodology. D.K. performed visualization. R.S.D. and J.A.E. acquired funding. D.K. and J.A.E. wrote the original draft. D.K., R.S.D., J.R. and J.A.E. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

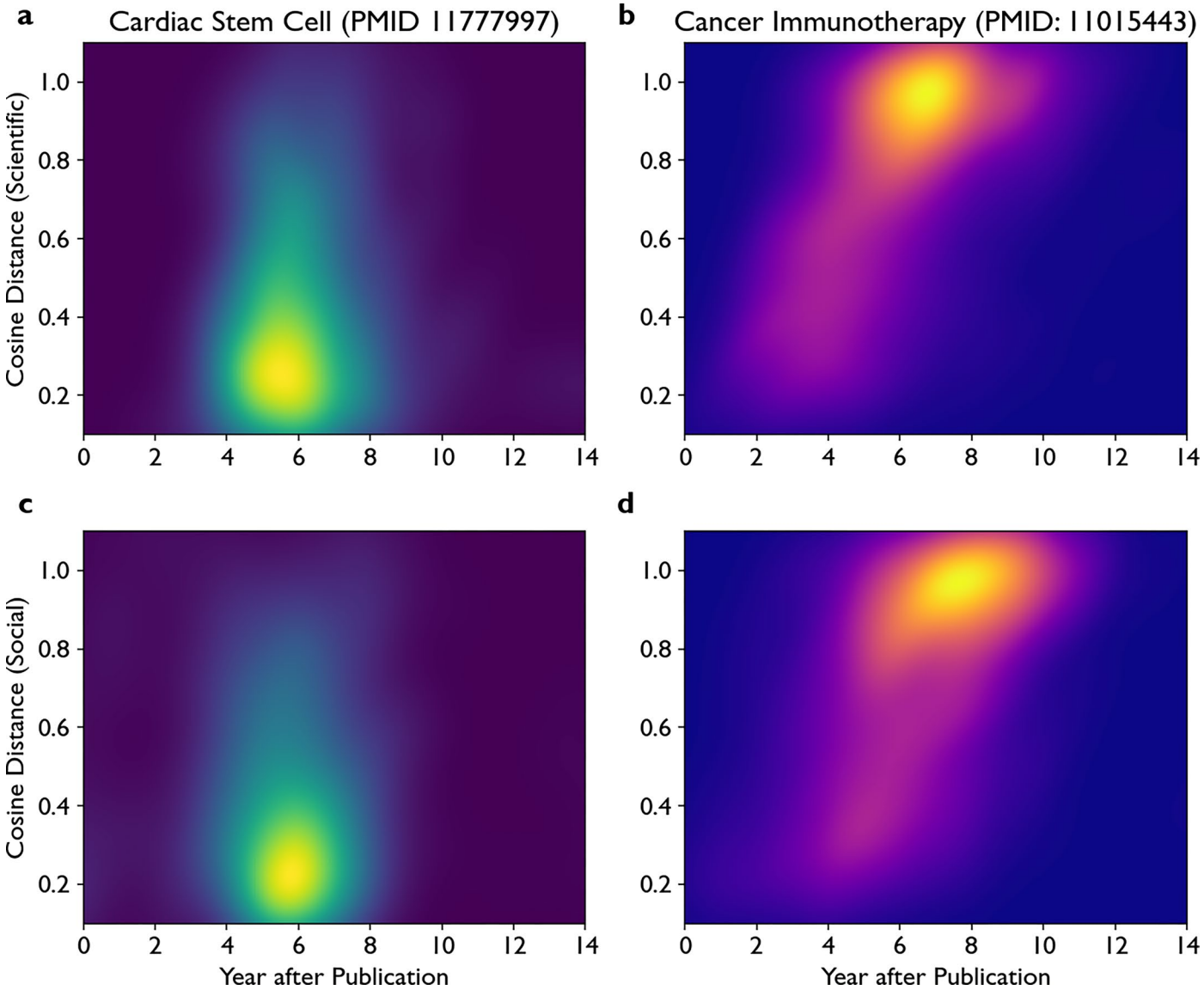
Extended data is available for this paper at https://doi.org/10.1038/s41562-024-02041-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-02041-0.

Correspondence and requests for materials should be addressed to James A. Evans.

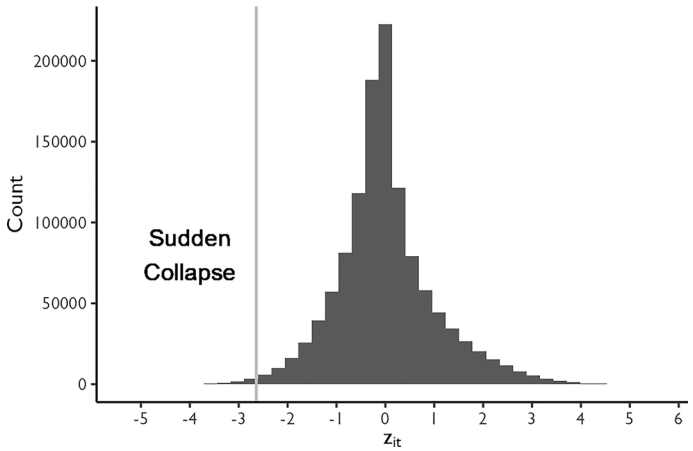
Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41562-024-02041-0



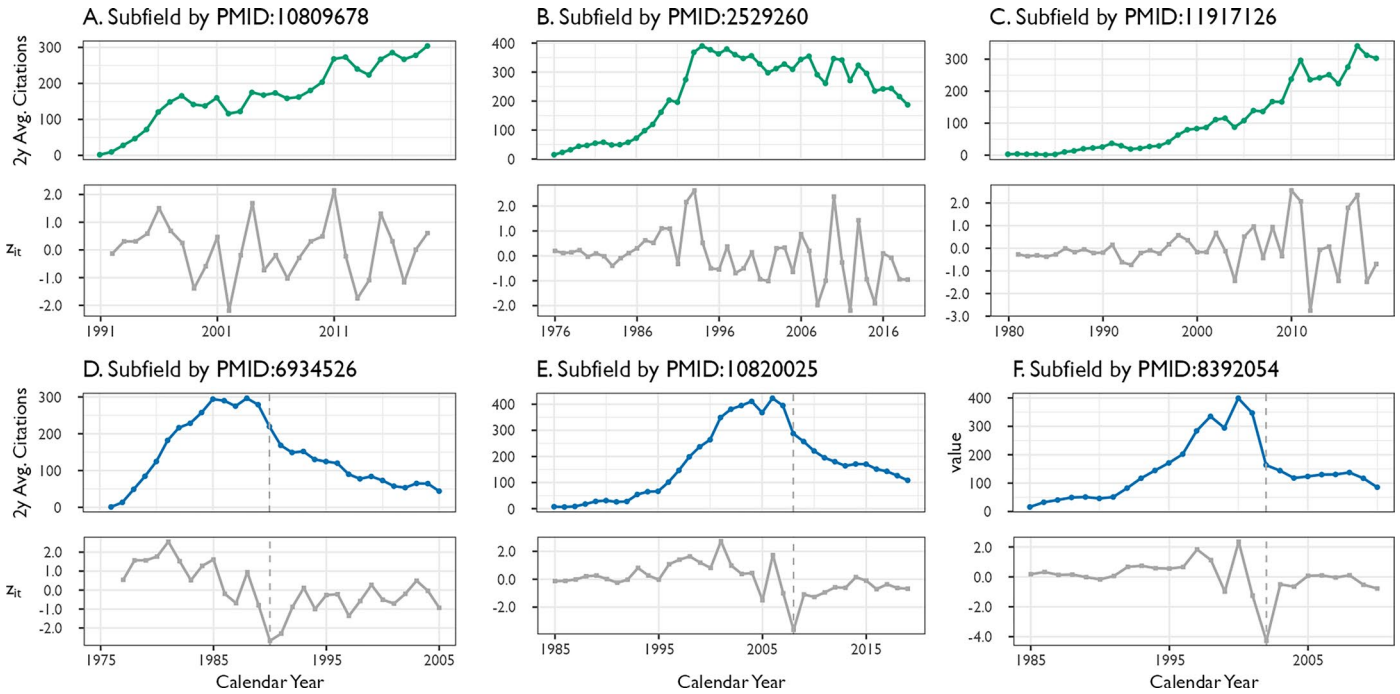
Extended Data Fig. 1 | Complementary 2D heatmaps for the upper panels of Fig. 1. Values are derived from the kernel density estimations graphed in Fig. 1 for the distribution of diffusion indices in scientific (panels **a** and **b**) and social space (panels **c** and **d**), respectively.

https://doi.org/10.1038/s41562-024-02041-0



Extended Data Fig. 2 | Distribution of $z_{i,t}$ from 28,504 subfields. We set the cutoff value for bubble bursting to −2.64, the bottom 0.5% percentile. The range of $z_{i,t}$ is [−5.2, 5.52].

https://doi.org/10.1038/s41562-024-02041-0



Extended Data Fig. 3 | Six examples of subfields. Annual citation counts aggregated at the subfield level, using forward citations to related publications. Top panels (**a**, **b**, **c**): Subfields represented by three PMIDs, illustrating cases without bubble bursting events. Bottom panels (**d**, **e**, **f**): Subfields that experienced bubble bursts, corresponding to cutoffs closest to the 0.5%, 0.25%, and 0.1% thresholds of the $z_{i,t}$ value.

https://doi.org/10.1038/s41562-024-02041-0

Extended Data Table 1 | Pairwise *t*-test comparing average productivity differences between near-collapse active scientists (≤2 years before collapse) and early entrants

	Threshold	Estimate	t	*p*-value (*d.f.*)	95% C.I.
5 Years	0.5%	-3.030	-20.67	< 0.001 (3,910)	[-3.317, -2.743]
	0.25%	-3.075	-13.48	< 0.001 (1,983)	[-3.522, -2.628]
	0.1%	-2.841	-11.13	< 0.001 (773)	[-3.342, -2.340]
10 Years	0.5%	-4.754	-24.45	< 0.001 (3,605)	[-5.136, -4.373]
	0.25%	-4.590	-17.17	< 0.001 (1,818)	[-5.114, -4.066]
	0.1%	-4.707	-10.87	< 0.001 (711)	[-5.558, -3.857]

Subfields that collapsed after 2015 were excluded from the 5-year productivity comparison. Likewise, for 10-year productivity, only subfields that collapsed on or before 2011 were included, to avoid censoring the observation window. *P*-values are for two-sided tests.

https://doi.org/10.1038/s41562-024-02041-0

Extended Data Table 2 | Proportion of subfields with newly acknowledged grants after collapse, and the Mean, 1st Quantile, Median, and 3rd Quartile of the number of new grants post-collapse

Threshold	% of Subfields with New Grants Acknowledged After Collapse	Mean	Q1	Median	Q3
0.5%	83.12%	11.8	2	6	16
0.25%	82.93%	11.3	2	6	15
0.1%	81.70%	10.1	1	6	13

Subfields that collapsed after 2015 were excluded from the 5-year productivity comparison. For the 10-year productivity analysis, only subfields that collapsed on or before 2011 were included, in consideration of the observation window size.

nature portfolio

Corresponding author(s): James A. Evans

Last updated by author(s): Aug 19, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We utilized the Biopython package (version 1.81) to collect metadata from publications indexed in PubMed (MEDLINE 2021) via the National Center for Biotechnology Information (NCBI) Entrez API (https://www.ncbi.nlm.nih.gov/home/develop/api/). The PubMed Related Algorithm (PMRA) is embedded in PubMed's 'Similar Articles' function. The PubMed Knowledge Graph (http://er.tacc.utexas.edu/datasets/ped) and replication data from Azoulay and colleagues (https://www.openicpsr.org/openicpsr/project/116188/version/V1/view;jsessionid=EA1E1E5A6DAB42737EE54A5F5DD4B069) were downloaded via the provided links.
Data analysis	We used the Gensim package (version 4.0) for Python to train our Doc2vec models and Python 3.9 for data handling and computing variables for statistical analysis with popular packages (e.g., pandas 1.4.4, scikit-learn 1.1.1). We utilized the survival package (version 3.3) in R for Kaplan-Meier estimation and the fixest package (v0.11) for fixed effect modeling.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

This work utilizes the PubMed Knowledge Graph (http://er.tacc.utexas.edu/datasets/ped) and replication data from Azoulay and colleagues (https://www.openicpsr.org/openicpsr/project/116188/version/V1/view;jsessionid=EA1E1E5A6DAB42737EE54A5F5DD4B069), both of which are publicly available. The data and codes for Figures 1a, 2, 3, 4; Table 1; Extended Data Figures 1, 2, 3; and Tables 1 and 2 are accessible at: https://github.com/Donghyun-Kang-Soc/limited_diffusion.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We did not have human research participants.
Reporting on race, ethnicity, or other socially relevant groupings	We did not have human research participants.
Population characteristics	We did not have human research participants.
Recruitment	We did not have human research participants.
Ethics oversight	We did not have human research participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study quantitatively assesses how the diffusion of scientific knowledge is associated with the likelihood of sudden collapse.
Research sample	The study utilized two datasets. The first, the PubMed Knowledge Graph, contains metadata and disambiguated author IDs for all of the 30 million papers indexed in PubMed (MEDLINE 2021) published by the end of 2020 in the biomedical and life sciences fields. This dataset is comprehensive and representative of the biomedical literature up to 2020. The second dataset, a replication dataset from Azoulay et al., includes information on 28,504 biomedical subfields based on the publications of star scientists, encompassing about 1.9 million papers. It represents the largest and most reliable dataset of biomedical subfields spanned by star scientists' publications and has undergone extensive validation, as detailed in the online Appendix (aeaweb.org/content/file?id=10303).
Sampling strategy	The Azoulay team identified star scientists based on funding levels, citation counts, patent records, membership in the National Academies, NIH MERIT awards, Howard Hughes Medical Investigator status, and early career achievements. The sampling procedure involved a stratified matching process, where 3,076 subfields impacted by the sudden deaths of 452 star biomedical scientists (out of 12,935 star scientists) were identified. These subfields were then stratified and matched by publication year, author count, citations, and the age of similar scientists without such losses, resulting in 34,218 unique pairs of scientists and subfields. Our study repurposed this dataset, covering 28,504 unique seed articles across various subfields. The samples were chosen based on the availability of comparable subfields, with the rigorous stratification and matching process controlling for potential confounders.
Data collection	We utilized the Biopython package (version 1.81) to collect metadata from publications indexed in PubMed (MEDLINE 2021) via the NCBI Entrez API (https://www.ncbi.nlm.nih.gov/home/develop/api/). The PubMed Knowledge Graph (http://er.tacc.utexas.edu/datasets/ped) and replication data from Azoulay and colleagues (https://www.openicpsr.org/openicpsr/project/116188/version/V1/view;jsessionid=EA1E1E5A6DAB42737EE54A5F5DD4B069) were downloaded via the provided links.

Timing	In April 2021, we accessed PubMed's Related Article Algorithm via the Biopython Entrez wrapper to update the Azoulay data. We used the iCite bulk repository, downloaded in October 2023, for analysis in Supplementary Information.
Data exclusions	We excluded papers from relevant variable computations when the following information was unavailable: year of publication, author IDs, or MeSH terms. Consequently, approximately 100K papers were dropped from 1.96M publications.
Non-participation	Our research did not include any direct human participants.
Randomization	Our research is observational, leveraging existing data and computational methods with measurements reflecting the data structure without controlled interventions. Thus, randomization is not involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☐ Clinical data
☐	☐ Dual use research of concern
☐	☐ Plants

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance*

Ethics oversight	*was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

(See Eklund et al. 2016)

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.