Supplementary Material for:

Female scientists produce more novel and disruptive ideas yet receive less citation impact

Contents

Supplementary Note 1. Descriptive analysis

Table S1. Description of the variables

Type	Variables	Description/Definition	Sources
Dependent variable	**Female authorship as the first author**	If the first author is a female scientist, the dummy variable is assigned a value of 1; otherwise, it is assigned a value of 0.	(Van Buskirk et al., 2023)
	Female authorship as the last author	If the last author is a female scientist, the dummy variable is assigned a value of 1; otherwise, it is assigned a value of 0.	(Van Buskirk et al., 2023)
Independent variable	**5-Year citation count**	The cumulative number of citations a paper receives five years post-publication as a measure of its scientific impact.	(Wang et al., 2013)
	Novel paper	Scientific novelty is operationalized by scrutinizing the observed distribution of journal pairs within a paper's reference list and comparing it to the random distribution of journal pairs generated by a null model. Novel paper is a dummy variable.	Uzzi et al. (2013)
	Disruptive paper	A paper is considered to be disruptive when its 5-year CD index is above zero (The 5-year CD index is defined as the proportion of disruptive citations in the total 5-year citations received by the paper).	(Funk & Owen-Smith, 2017; Park et al., 2023; Wu et al., 2019)
Team level controls	**Team size**	The number of authors in a paper.	(Wuchty et al., 2007)
	International team	Indicates whether the paper's authorship includes individuals from different countries.	(Lee et al., 2019)
	Interdisciplinary team	Indicates whether the paper's authorship spans multiple fields of study.	(Liu et al., 2024)
Author level controls	**Career age**	The number of years elapsed from their initial publication in the MAG dataset until the publication year of the focal paper.	(Yang, Xu, et al., 2024)
	#Past publications	The total number of publications by the author up to the year the focal paper was published.	(Jones, 2009)
	#Past hit publications	the number of hit papers the author has produced up to that point. A hit paper is defined as one of the top 10% most highly cited papers in its publication year and subfields	(Mukherjee et al., 2017)
	Funding	Indicates whether the researcher was funded by the National Institutes of Health (NIH) or the National Science Foundation (NSF).	(Yang, Gong, et al., 2024)
	Focal field	We categorize the primary field of each scientist using first-level MAG field-of-study labels. If the focal paper aligns with the scientist's most frequently studied fields before its publication date, we assign it a value of 1; otherwise, we assign it a value of 0.	(Zeng et al., 2019)

Table S2. Statistical analysis of the variables.

Panel a. all	COUNT	MEAN	STD	MIN	MEDIAN	MAX
Female	11385225	0.21	0.41	0	0	1
5-year citation count	11385225	23.14	60.37	0	12	37405
Team size	11385225	4.90	25.18	1	4	5104
Interdisciplinary team	11385225	0.36	0.48	0	0	1
International team	11385225	0.21	0.41	0	0	1
Career age	11385225	16.11	11.61	0	14	60
#Past publications	11385225	76.35	112.75	0	37	1000
#Past hit publications	11385225	10.16	21.86	0	3	400
Funding	11385225	0.13	0.34	0	0	1
Focal field	11385225	0.35	0.48	0	0	1

Panel b. first author	COUNT	MEAN	STD	MIN	MEDIAN	MAX
Female	6199531	0.24	0.43	0	0	1
5-year citation count	6199531	22.43	60.18	0	11	37405
Team size	6199531	4.58	24.15	1	4	5104
Interdisciplinary team	6199531	0.33	0.47	0	0	1
International team	6199531	0.20	0.40	0	0	1
Career age	6199531	12.55	10.78	0	9	60
#Past publications	6199531	46.31	79.35	0	21	1000
#Past hit publications	6199531	5.75	14.48	0	1	400
Funding	6199531	0.13	0.33	0	0	1
Focal field	6199531	0.36	0.48	0	0	1

Panel c. last author	COUNT	MEAN	STD	MIN	MEDIAN	MAX
Female	5185694	0.17	0.38	0	0	1
5-year citation count	5185694	23.99	60.58	0	12	24550
Team size	5185694	5.28	26.35	2	4	5104
Interdisciplinary team	5185694	0.39	0.49	0	0	1
International team	5185694	0.23	0.42	0	0	1
Career age	5185694	20.37	11.13	0	19	60
#Past publications	5185694	112.27	134.22	0	68	1000
#Past hit publications	5185694	15.43	27.34	0	6	400
Funding	5185694	0.14	0.34	0	0	1
Focal field	5185694	0.34	0.47	0	0	1

Supplementary Note 2. Regression results

Table S3. Poisson regression: effect of the female scientist as the first or last author on the 5-year citation count.

Models	Poisson regression					
	(1)	(2)	(3)	(1)	(2)	(3)
Dependent variable	5-year Citation count					
	First author			Last author		
Female	-0.0559*** (0.0022)	-0.0794*** (0.0022)	-0.0437*** (0.0021)	-0.0680*** (0.0026)	-0.0541*** (0.0026)	-0.0172*** (0.0026)
ln(Team size)		0.3363*** (0.0025)	0.3264*** (0.0023)		0.3668*** (0.0032)	0.3271*** (0.0030)
Interdisciplinary team		-0.0951*** (0.0024)	-0.0982*** (0.0023)		-0.0888*** (0.0024)	-0.0972*** (0.0023)
International team		0.1840*** (0.0029)	0.1322*** (0.0028)		0.1811*** (0.0029)	0.1198*** (0.0029)
ln(Career age)			0.0187*** (0.0023)			0.0241*** (0.0026)
ln(#Past publications)			-0.3301*** (0.0025)			-0.3904*** (0.0023)
ln(#Past hit publications)			0.5714*** (0.0018)			0.5305*** (0.0016)
Funding			0.2975*** (0.0030)			0.1948*** (0.0032)
Focal field			0.0467*** (0.0024)			0.0317*** (0.0024)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,199,531	6,199,531	6,199,531	5,185,694	5,185,694	5,185,694
Squared Cor.	0.0188	0.03136	0.07323	0.0192	0.03144	0.07385
Pseudo R2	0.07125	0.10563	0.22023	0.06653	0.09582	0.2088

Note: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S4. Logistic regression: effect of the female scientist as the first or last author on the scientific novelty.

Models	Logistic regression					
	(1)	(2)	(3)	(1)	(2)	(3)
Dependent variable	P (Novel paper)					
	First author			Last author		
Female	0.0931*** (0.0020)	0.0776*** (0.0020)	0.0730*** (0.0021)	0.0681*** (0.0025)	0.0748*** (0.0025)	0.0819*** (0.0026)
ln(Team size)		0.0355*** (0.0014)	0.0286*** (0.0014)		0.0747*** (0.0019)	0.0623*** (0.0020)
Interdisciplinary team		0.4987*** (0.0021)	0.4775*** (0.0021)		0.5067*** (0.0021)	0.4832*** (0.0021)
International team		0.0121*** (0.0023)	0.0143*** (0.0023)		0.0118*** (0.0023)	0.0103*** (0.0023)
ln(Career age)			0.0065*** (0.0017)			0.0170*** (0.0021)
ln(#Past publications)			-0.0242*** (0.0015)			-0.0084*** (0.0017)
ln(#Past hit publications)			0.0268*** (0.0012)			0.0241*** (0.0012)
Funding			0.2683*** (0.0028)			0.2581*** (0.0029)
Focal field			-0.2633*** (0.0019)			-0.3100*** (0.0021)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes

Observations	6,199,523	6,199,523	6,199,523	5,185,692	5,185,692	5,185,692
Squared Cor.	0.0806	0.09184	0.09649	0.0731	0.08587	0.09182
Pseudo R2	0.06066	0.06957	0.07317	0.05497	0.06507	0.06967

Note: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S5. Logistic regression: effect of the female scientist as the first or last author on the probability of disruptive papers.

Models	**Logistic regression**					
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
Dependent variable	**P (Disruptive paper)**					
	First author			**Last author**		
Female	0.0755*** (0.0021)	0.0755*** (0.0021)	0.0684*** (0.0021)	0.0578*** (0.0025)	0.0557*** (0.0025)	0.0391*** (0.0026)
ln(Team size)		-0.0248*** (0.0015)	-0.0102*** (0.0015)		-0.0185*** (0.0020)	-0.0075*** (0.0020)
Interdisciplinary team		0.1022*** (0.0021)	0.0982*** (0.0021)		0.1133*** (0.0021)	0.1137*** (0.0021)
International team		-0.1225*** (0.0023)	-0.1126*** (0.0023)		-0.1161*** (0.0023)	-0.0980*** (0.0023)
ln(Career age)			0.1851*** (0.0017)			0.1019*** (0.0022)
ln(#Past publications)			-0.0483*** (0.0015)			0.0077*** (0.0017)
ln(#Past hit publications)			-0.1061*** (0.0012)			-0.1103*** (0.0012)
Funding			-0.1571*** (0.0026)			-0.1265*** (0.0028)
Focal field			-0.0685*** (0.0019)			-0.0728*** (0.0021)
Year FE	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Field FE	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Observations	5,852,007	5,852,007	5,852,007	4,962,797	4,962,797	4,962,797
Squared Cor.	0.05	0.05092	0.05575	0.0518	0.05289	0.05707
Pseudo R2	0.03733	0.03802	0.04171	0.03871	0.03952	0.04268

Note: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Supplementary Note 3. Novel paper

Novelty involves the introduction of new ideas or the recombination of existing fragments of knowledge (March, 1991; Nelson, 1985; Schumpeter, 1939). Scientific and technological advancements do not arise spontaneously but are derived from the existing corpus of knowledge (Arthur, 2009). Scientific breakthroughs are often likened to a journey through a vast expanse of combinatorial possibilities, leading to fresh insights and technological progress. Research studies have emphasized the crucial role of combinations in connecting innovation with scientific and technological impact (Hofstra et al., 2020; Trapido, 2015). Most efforts to model the creative process perceive it as an accumulative and interactive recombination of existing fragments of knowledge, merged in novel ways (Azoulay et al., 2011). Novel research endeavors, while potentially entailing higher risks, often strive to address challenging issues and projects sought after by policymakers (Stephan et al., 2017; Wang et al., 2018). Novel studies push the boundaries of existing knowledge and venture into unexplored domains of

scientific inquiry (Xu et al., 2022). In this study, we follow the methodology of Uzzi et al. (2013), using the yearly journal distribution in the reference list of papers to calculate scientific novelty. As our dataset includes papers with at least 10 references, this method effectively measures atypical combinations of knowledge.

To assess the novelty of papers, we examine the presence of atypical combinations of knowledge. Scientific novelty is operationalized by scrutinizing the observed distribution of journal pairs within a paper's reference list and comparing it to the random distribution of journal pairs generated by a null model. As shown in Fig. S1, we run Monte Carlo simulations by randomly switching the reference links between random paper pairs while controlling for time. Novelty as the atypical combinations of knowledge is quantified through these Monte Carlo simulations. The simulations involve creating reshuffled networks with random edge reassignment while preserving the temporal and distributional attributes of the original citation network. Each journal pairing is transformed into z-scores, representing standardized values. The computation of atypical combinations for each journal pair ($pair_{mn}$) is encapsulated by the following formula:

$$Z\ score_{m,n} = \frac{obs(pair_{mn}) - exp(pair_{mn})}{\sigma(pair_{ij})}$$

where $obs(pair_{mn})$ signifies the observed frequency of the journal pair in the actual dataset, $exp(pair_{mn})$ represents the mean, and $\sigma(pair_{mn})$ denotes the standard deviation of journal pairs obtained from 10 randomized simulations of the reshuffled network. To encapsulate the information within the distribution, denoted as a set $\{Z\ score_{m,n} \mid m,n \in J\}$, wherein J encompasses all journals within the reference list, we leverage the 10th percentile value of the novelty set as a succinct summary statistic. We denote papers as novel if their 10th percentile value of the novelty set is lower than 0; otherwise, they are not considered novel papers.

$$Novel\ paper = \begin{cases} 1, if\ 10pct\{Z\ score_{m,n} \mid m,n \in J\} < 0 \\ 0, if\ 10pct\{Z\ score_{m,n} \mid m,n \in J\} \geq 0 \end{cases}$$

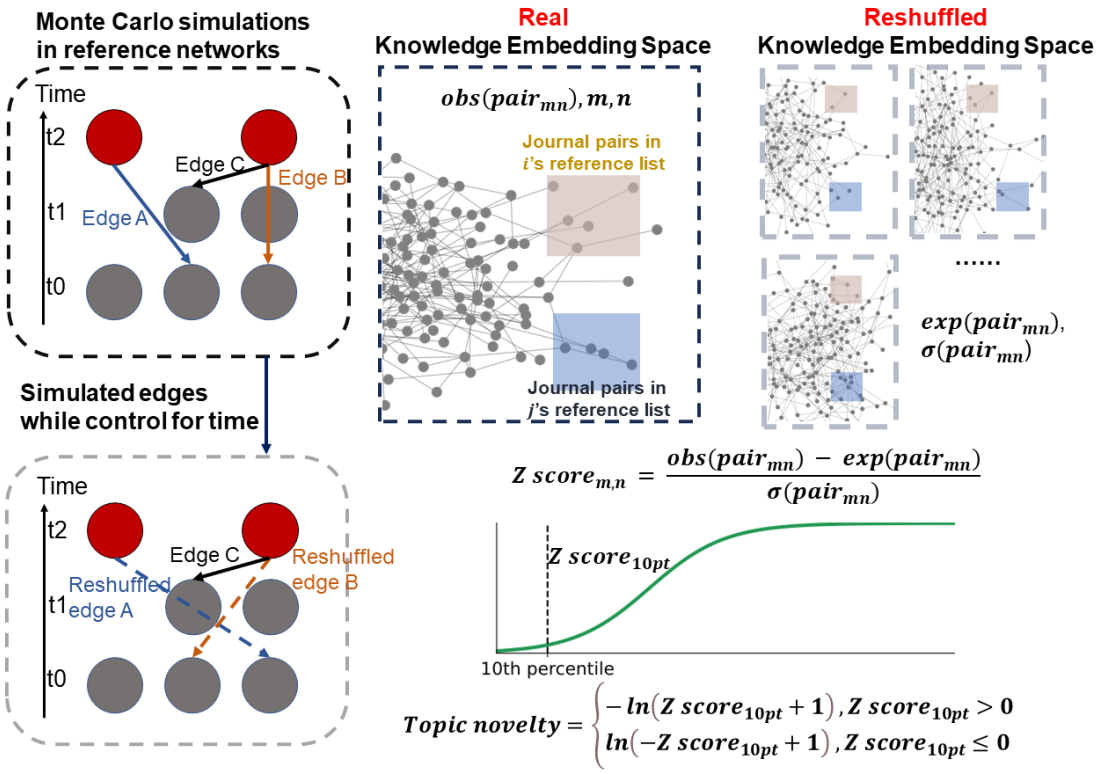

Fig. S1 Quantifying novelty in scientific papers

Supplementary Note 4. Disruptive paper

The CD index is defined as:

$$CD\ index = p_D - p_C = \frac{n_i - n_j}{n_i + n_j + n_k} \quad (1)$$

where n_i represents the number of future papers citing the focal paper (FP) that do not reference its cited sources, n_j denotes the number of future papers citing FP and its references, and n_k represents the number of papers citing FP's references without citing FP itself. A lower CD index indicates alignment with established knowledge, while a higher CD index signifies papers with the potential to induce paradigm shifts.

Leibel and Bornmann (2023) provide a comprehensive appraisal of the disruption index, deliberating on its conceptual underpinnings, capabilities, extensions, and constraints. Although it has some flaws and limitations, the CD index has been widely used in bibliometrics, the science of science, and many other fields of study. Notably, at least three papers published in the esteemed journal Nature in the last 3 years utilized the CD index as the main variable (Lin et al., 2023; Park et al., 2023; Wu et al., 2019).

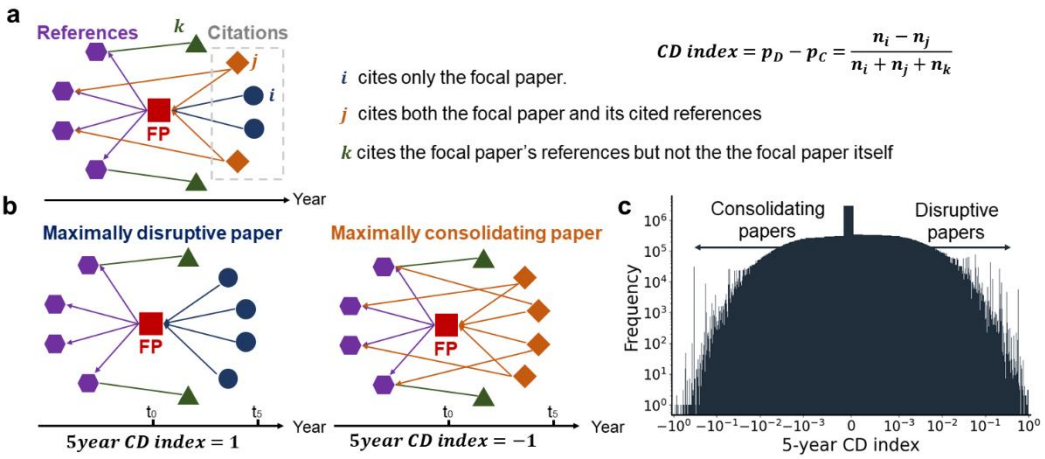

Fig. S2 Quantifying the 5-year CD index

Supplementary Note 4. Robustness check

We run OLS regression with alternative novelty scores to check the robustness of our findings:

$$Novelty\ score = \begin{cases} -ln(Z\ score_{10pt} + 1), if\ Z\ score_{10pt} > 0 \\ ln(-Z\ score_{10pt} + 1), if\ Z\ score_{10pt} \leq 0 \end{cases}$$

Table S4. OLS regression: effect of the female scientist as the first or last author on the novelty scores.

Models	Poisson regression					
	(1)	(2)	(3)	(5)	(6)	(7)
Dependent variable	Novelty score					
	First author			Last author		
Female	0.1097*** (0.0023)	0.0873*** (0.0022)	0.0796*** (0.0023)	0.0780*** (0.0027)	0.0855*** (0.0027)	0.0949*** (0.0028)
ln(Team size)		0.0507*** (0.0017)	0.0406*** (0.0017)		0.1037*** (0.0022)	0.0872*** (0.0022)
Interdisciplinary team		0.6537*** (0.0023)	0.6236*** (0.0023)		0.6652*** (0.0023)	0.6314*** (0.0023)
International team		0.0140*** (0.0026)	0.0175*** (0.0026)		0.0127*** (0.0026)	0.0096*** (0.0026)
ln(Career age)			0.0101*** (0.0019)			0.0250*** (0.0024)
ln(#Past publications)			-0.0365*** (0.0017)			-0.0197*** (0.0019)
ln(#Past hit publications)			0.0346*** (0.0013)			0.0405*** (0.0013)
Funding			0.3500*** (0.0030)			0.3318*** (0.0031)
Focal field			-0.3319*** (0.0021)			-0.3924*** (0.0024)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,199,531	6,199,531	6,199,531	5,185,694	5,185,694	5,185,694
R2	0.10669	0.12103	0.12655	0.09938	0.1158	0.12299

Note: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

References

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. Simon and Schuster.

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527-554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791-817. https://doi.org/10.1287/mnsc.2015.2366

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The Diversity-Innovation Paradox in Science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284-9291. https://doi.org/10.1073/pnas.1915378117

Jones, B. F. (2009). The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? *The Review of Economic Studies*, *76*(1), 283-317. https://doi.org/10.1111/j.1467-937X.2008.00531.x

Lee, C., Kogler, D. F., & Lee, D. (2019). Capturing information on technology convergence, international collaboration, and knowledge flow from patent documents: A case of information and communication technology. *Information Processing & Management*, *56*(4), 1576-1591. https://doi.org/10.1016/j.ipm.2018.09.007

Leibel, C., & Bornmann, L. (2023). What do we know about the disruption index in scientometrics? An overview of the literature. *Scientometrics*. https://doi.org/10.1007/s11192-023-04873-5

Lin, Y., Frey, C. B., & Wu, L. (2023). Remote collaboration fuses fewer breakthrough ideas. *Nature*, *623*(7989), 987-991. https://doi.org/10.1038/s41586-023-06767-1

Liu, X., Bu, Y., Li, M., & Li, J. (2024). Monodisciplinary collaboration disrupts science more than multidisciplinary collaboration. *Journal of the Association for Information Science and Technology*, *75*(1), 59-78. https://doi.org/10.1002/asi.24840

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, *2*(1), 71-87. https://doi.org/10.1287/orsc.2.1.71

Mukherjee, S., Romero, D. M., Jones, B., & Uzzi, B. (2017). The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science Advances*, *3*(4). https://doi.org/10.1126/sciadv.1601315

Nelson, R. R. (1985). *An evolutionary theory of economic change*. harvard university press.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*(7942), 138-144. https://doi.org/10.1038/s41586-022-05543-x

Schumpeter, J. A. (1939). *Business Cycles: A Theoretical, Historical, and Statistical Analysis of the Capitalist Process* (Vol. 1). Mcgraw-hill New York.

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature*, *544*(7651), 411-412. https://doi.org/10.1038/544411a

Trapido, D. (2015). How novelty in knowledge earns recognition: The role of consistent identities. *Research Policy*, *44*(8), 1488-1500. https://doi.org/10.1016/j.respol.2015.05.007

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468-472. https://doi.org/10.1126/science.1240474

Van Buskirk, I., Clauset, A., & Larremore, D. B. (2023). An Open-Source Cultural Consensus Approach to Name-Based Gender Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, *17*, 866-877. https://doi.org/10.1609/icwsm.v17i1.22195

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*(6154), 127-132. https://doi.org/doi:10.1126/science.1237825

Wang, J., Lee, Y.-N., & Walsh, J. P. (2018). Funding model and creativity in science: Competitive versus block funding and status contingency effects. *Research Policy*, *47*(6), 1070-1083. https://doi.org/10.1016/j.respol.2018.03.014

Wu, L. F., Wang, D. S., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378-+. https://doi.org/10.1038/s41586-019-0941-9

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036-1039. https://doi.org/10.1126/science.1136099

Xu, F., Wu, L., & Evans, J. (2022). Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences*, *119*(23), e2200927119. https://doi.org/10.1073/pnas.2200927119

Yang, A. J., Gong, H., Wang, Y., Zhang, C., & Deng, S. (2024). Rescaling the disruption index reveals the universality of disruption distributions in science. *Scientometrics*, *129*(1), 561-580. https://doi.org/10.1007/s11192-023-04889-x

Yang, A. J., Xu, H., Ding, Y., & Liu, M. (2024). Unveiling the dynamics of team age structure and its impact on scientific innovation. *Scientometrics*. https://doi.org/10.1007/s11192-024-04987-4

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1), 3439. https://doi.org/10.1038/s41467-019-11401-8