Journal of Informetrics 20 (2026) 101840

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research Paper

Distant but bridgeable: how geographic and cognitive distance jointly shape interdisciplinary knowledge production

Alex J. Yang [a,b], Yujie Shi [a,b], Fanming Wang [a,b], Xinyu Song [a,b], Sanhong Deng [a,b,*]

[a] School of Information Management, Nanjing University, Nanjing China
[b] International Joint Informatics Laboratory, Nanjing University, Nanjing, China

ARTICLE INFO

Keywords:
Interdisciplinarity
Geographic distance
Cognitive distance
Scientific collaboration

ABSTRACT

We introduce a distance mismatch perspective and examine how geographic and cognitive distance jointly influence interdisciplinarity. Using a large-scale dataset of over 31 million team-authored publications from the OpenAlex database, we measure interdisciplinarity using the Rao–Stirling index based on the disciplinary composition of references. Team cognitive distance is constructed from time-aware author embeddings derived from SPECTER2, while geographic distance is measured using institutional affiliation coordinates. We find that both geographic and cognitive distance are positively associated with interdisciplinarity when considered independently. However, their joint effect reveals a more nuanced pattern. Geographic distance enhances interdisciplinarity only when teams are cognitively diverse. When cognitive distance is low, geographically local teams perform equally well or better. Decomposing interdisciplinarity into variety, balance, and disparity further shows that geographic distance amplifies the breadth and cognitive reach of knowledge integration, but at the cost of reduced balance. These findings demonstrate that the benefits of geographic dispersion are conditional on cognitive structure, highlighting the importance of aligning spatial and intellectual dimensions in scientific collaboration.

1. Introduction

The organization of scientific collaboration has undergone profound transformation over the past decades (Cao et al., 2024; Jones, 2021; Jones et al., 2008; Wuchty et al., 2007; Xie et al., 2023; Yadav et al., 2023; Zou et al., 2023). Advances in communication technologies, increased mobility of researchers, and the globalization of science have enabled teams to collaborate across institutional, national, and continental boundaries (Jones, 2021; Jones et al., 2008; Wuchty et al., 2007; Zou et al., 2023). At the same time, the growing complexity of scientific problems has intensified the need for integrating knowledge across disciplinary domains. Together, these trends have made geographic dispersion and cognitive diversity central features of modern scientific production.

A large body of research in the science of science and innovation studies has examined these two dimensions separately. On the one hand, geographically dispersed collaborations are often associated with access to diverse resources, perspectives, and institutional environments, which can foster novel re-combinations of knowledge (Jones et al., 2008; Lin et al., 2023; Zou et al., 2023). On the other hand, cognitive diversity—differences in expertise, disciplinary background, and knowledge domains—has been shown to play a

* Corresponding author.
E-mail address: sanhong@nju.edu.cn (S. Deng).

https://doi.org/10.1016/j.joi.2026.101840
Received 28 September 2025; Received in revised form 31 May 2026; Accepted 3 June 2026
Available online 17 June 2026
1751-1577/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

crucial role in generating interdisciplinary and innovative outcomes (Aceves & Evans, 2024; Fleming, 2001; Uzzi et al., 2013). However, both forms of distance also entail coordination costs, communication barriers, and integration challenges, suggesting that their effects on scientific outcomes are inherently ambiguous.

Despite extensive work on geographic and cognitive distance, the existing literature has largely treated them as independent drivers of scientific outcomes. This separation overlooks a fundamental feature of collaborative knowledge production: geographic and cognitive distances are often intertwined, and their effects may depend on each other. For example, geographically distant teams may be better positioned to access non-overlapping knowledge pools, but the benefits of such access depend critically on whether team members possess sufficient cognitive proximity to integrate their ideas. Conversely, cognitively diverse teams may struggle to coordinate effectively when combined with spatial separation. As a result, focusing on either dimension in isolation provides an incomplete understanding of how interdisciplinary knowledge is produced.

This paper addresses this gap by introducing a distance mismatch perspective, which emphasizes the joint role of geographic and cognitive distance in shaping interdisciplinary outcomes. We argue that geographic distance does not uniformly promote or hinder interdisciplinarity; rather, its effect depends on the cognitive structure of the team. When cognitive distance is low, geographic separation may add little new knowledge while increasing coordination costs. In contrast, when cognitive distance is high, geographic dispersion can facilitate access to heterogeneous knowledge environments, enabling more extensive recombination across distant domains. This interaction suggests that the relationship between distance and interdisciplinarity is fundamentally conditional rather than additive.

To empirically examine this mechanism, we construct a large-scale dataset of >31 million team-authored publications from OpenAlex. We measure interdisciplinarity using the Rao–Stirling diversity index based on the disciplinary composition of references, and we quantify team cognitive distance using time-aware author embeddings derived from SPECTER2 (Singh et al., 2023). Geographic distance is operationalized using institutional affiliation coordinates. This combination of fine-grained bibliometric data and embedding-based representations allows us to capture both the structural and semantic dimensions of scientific collaboration at an unprecedented scale.

Our findings reveal three main patterns. First, both geographic distance and cognitive distance are positively associated with interdisciplinarity when considered independently. Second, cognitive distance emerges as a primary driver of interdisciplinary knowledge integration, with strong and consistent effects across fields. Third, and most importantly, we uncover a robust interaction between the two dimensions: geographically distant teams exhibit higher levels of interdisciplinarity only when they are also cognitively diverse. When cognitive distance is low, geographically local teams perform equally well or better. This “distance mismatch” pattern is consistent across fields and across multiple dimensions of interdisciplinarity.

These findings contribute to the literature in several ways. First, we move beyond the dominant focus on single-dimensional measures of distance by developing a joint framework that integrates geographic and cognitive perspectives. Second, we provide large-scale empirical evidence that the benefits of distance are conditional, helping reconcile mixed findings in prior studies regarding the effects of spatial dispersion. Third, by decomposing interdisciplinarity into variety, balance, and disparity, we show that distance affects different dimensions of knowledge integration in distinct ways, highlighting trade-offs that are obscured by aggregate measures.

2. Background

2.1. Geographic distance and scientific collaboration

Geographic distance has long been recognized as a fundamental constraint on collaboration and knowledge production. Early work in economic geography and the science of science emphasizes the importance of spatial proximity for facilitating communication, trust, and the exchange of tacit knowledge (Jones et al., 2008; Zou et al., 2023). Co-located teams benefit from frequent face-to-face interaction, shared institutional environments, and lower coordination costs, all of which support efficient collaboration and incremental knowledge development (Abramo et al., 2024; Gui et al., 2019; Ma & Huang, 2024; Pialli, 2023).

However, advances in transportation and communication technologies have substantially reduced the friction of distance, enabling the rise of geographically dispersed collaborations. A growing body of research shows that such collaborations can access more diverse knowledge pools, institutional contexts, and problem-solving approaches (Fortunato et al., 2018; Jones, 2021; Wuchty et al., 2007). These advantages are often linked to higher levels of novelty, creativity, and impact (Bernal et al., 2022). At the same time, distance introduces coordination challenges, including communication delays, cultural differences, and difficulties in integrating distributed knowledge (Seo et al., 2017). As a result, empirical findings on the effect of geographic distance remain mixed, with some studies highlighting its benefits and others emphasizing its costs.

The "death of distance" hypothesis posits that technological advancements have reduced the importance of physical proximity, facilitating seamless collaboration across vast distances (Jones et al., 2008). Digital tools, including email, video conferencing, and collaborative platforms, have expanded scientific networks, fostering diverse team compositions (Hoekman et al., 2010). Yet, the impact of geographical distance on collaborative success remains debated, with studies reporting conflicting results (Abramo et al., 2024; Gui et al., 2019; Ma & Huang, 2024; Pialli, 2023). Knowledge spillovers can amplify or hinder collaboration (Bernal et al., 2022), while the effectiveness of R&D cooperation varies by institutional and sectoral conditions (Seo et al., 2017). Media richness theory (Daft & Lengel, 1986) posits that face-to-face interactions, which are more feasible at shorter distances, are optimal for conveying complex ideas due to their high information-carrying capacity. This suggests that proximity enhances tacit knowledge exchange and coordination. Conversely, media synchronicity theory (Dennis et al., 2008) argues that asynchronous tools (e.g., email)

are effective for information sharing, while synchronous tools (e.g., Zoom) facilitate consensus-building. These theories imply that remote collaboration can be productive if teams leverage appropriate communication channels for their tasks.

A key limitation of this literature is that geographic distance is typically treated as a stand-alone factor. Yet, the value of spatial dispersion likely depends on the nature of the knowledge being combined. Without considering the intellectual composition of teams, it is difficult to determine when geographic distance facilitates meaningful knowledge recombination and when it simply adds coordination burdens.

2.2. Cognitive distance and knowledge recombination

Cognitive distance—defined as differences in knowledge bases, expertise, and disciplinary backgrounds among collaborators—has emerged as a central concept in understanding innovation and scientific discovery. Rooted in theories of knowledge recombination, cognitive distance captures the extent to which actors draw on distinct but potentially complementary domains of expertise. Prior research suggests that innovation arises from combining previously unconnected knowledge elements, making cognitive diversity a key driver of novelty (Fleming, 2001; Uzzi et al., 2013).

Empirical studies provide substantial support for this view. Uzzi et al. (2013) show that high-impact scientific papers tend to balance conventional knowledge with atypical combinations, highlighting the importance of spanning distant intellectual domains. Similarly, Fleming (2001) demonstrates that recombinant search across technological domains increases the likelihood of breakthrough inventions, albeit with greater uncertainty. In the context of team science, cognitively diverse teams are more likely to generate novel insights by integrating heterogeneous perspectives and problem-solving approaches (Wuchty et al., 2007).

At the same time, the benefits of cognitive distance are not unbounded. The literature emphasizes a fundamental trade-off between novelty and coordination. As cognitive distance increases, differences in language, conceptual frameworks, and methodological approaches can hinder communication and reduce mutual understanding (Levinthal & March 1993). This can lead to difficulties in integrating knowledge, increased coordination costs, and a higher risk of project failure. As a result, several studies suggest that the relationship between cognitive distance and innovation is contingent, with the benefits of diversity depending on teams' ability to bridge cognitive gaps.

Recent advances in computational methods have enabled more precise measurement of cognitive distance using text-based embeddings and citation-informed representations of scientific knowledge (Aceves & Evans, 2024). These approaches capture fine-grained semantic relationships between research outputs and allow for continuous measurement of intellectual diversity at scale. However, despite these methodological advances, most empirical work continues to treat cognitive distance as an isolated determinant of innovation, without considering how it interacts with other structural features of collaboration, such as geographic dispersion.

2.3. Interdisciplinarity as knowledge integration

Interdisciplinary research has become a defining feature of contemporary science, particularly in addressing complex problems that transcend traditional disciplinary boundaries. At its core, interdisciplinarity reflects the integration of knowledge from multiple domains, requiring both the combination of diverse perspectives and the reconciliation of distinct conceptual frameworks. As such, it provides a natural setting for studying knowledge recombination in scientific production.

In bibliometric research, interdisciplinarity is commonly operationalized through the disciplinary diversity of a paper's references. Among the various measures proposed, the Rao–Stirling index has gained prominence as a comprehensive indicator that captures three key dimensions: variety (the number of distinct disciplines involved), balance (the distribution of attention across disciplines), and disparity (the cognitive distance between them) (Leydesdorff, 2018; Stirling, 2007). This framework allows for a nuanced characterization of interdisciplinary knowledge integration, distinguishing between breadth, evenness, and intellectual distance.

Empirical evidence suggests that interdisciplinary research plays a complex role in scientific outcomes. On the one hand, integrating distant knowledge domains can foster novelty and lead to high-impact discoveries (Uzzi et al., 2013; Wang et al., 2017). On the other hand, interdisciplinary work often faces higher uncertainty, greater coordination challenges, and barriers to evaluation within established disciplinary communities (Leahey et al., 2017). These tensions highlight that interdisciplinarity involves both opportunities for creative recombination and risks associated with integrating heterogeneous knowledge.

Importantly, prior research has largely focused on interdisciplinarity as an outcome of collaboration, without fully examining the conditions under which it emerges. In particular, while both geographic and cognitive dimensions of collaboration are likely to influence the structure of knowledge integration, their joint role in shaping interdisciplinarity remains underexplored. Understanding how these dimensions interact is therefore essential for explaining variation in interdisciplinary research across teams and contexts.

2.4. Toward a distance mismatch perspective

Although geographic and cognitive distance have each been studied extensively, relatively little attention has been paid to their joint effects. In practice, these two dimensions are closely intertwined: geographically distant collaborations often involve partners embedded in different institutional and intellectual environments, while cognitively diverse teams may span multiple locations. However, the extent to which these dimensions complement or substitute for each other remains unclear.

We argue that the effect of geographic distance on interdisciplinarity depends critically on the cognitive structure of the team. Geographic dispersion can expand access to heterogeneous knowledge environments, but the benefits of such access are contingent on whether team members possess sufficiently diverse knowledge bases to generate meaningful knowledge recombination. Conversely,

when cognitive distance is limited, geographic separation may add coordination costs without introducing substantial new knowledge.

This perspective suggests a distance mismatch mechanism, in which the alignment between spatial and cognitive distance determines the effectiveness of knowledge integration. Rather than viewing geographic distance as uniformly beneficial or detrimental, we conceptualize its impact as conditional on cognitive distance. Geographic distance may facilitate interdisciplinarity when teams bridge substantial cognitive differences and gain access to distinct knowledge environments, but it may provide limited advantages when collaborators possess highly similar knowledge backgrounds.

Building on this perspective, the present study investigates three related questions. First, how are geographic distance and cognitive distance individually associated with interdisciplinary knowledge production? Second, does the relationship between cognitive distance and interdisciplinarity vary between geographically local and geographically dispersed teams? Third, do these relationships differ across the constituent dimensions of interdisciplinarity—variety, balance, and disparity?

To address these questions, we analyze >31 million team-authored publications and examine how geographic and cognitive distance jointly shape interdisciplinary knowledge integration. By focusing on the interaction between spatial and intellectual dimensions of collaboration, we seek to provide a more comprehensive understanding of when and how distance contributes to interdisciplinary research.

3. Data and methods

3.1. Data

Our study draws on the OpenAlex dataset (Priem et al., 2022), an open and comprehensive catalog of global scholarly publications maintained by OurResearch. OpenAlex integrates data from multiple sources, including Microsoft Academic Graph, Crossref, PubMed, ORCID, and publisher websites, providing a unified and standardized repository of over 250 million scholarly works spanning the period 1666–2025. These records include a wide range of research outputs, such as journal articles, conference papers, books, and datasets, together with rich metadata on authorship, institutional affiliations, citations, concepts, venues, and funding sources. A key strength of OpenAlex lies in its large-scale entity disambiguation system, which assigns unique identifiers to works, authors, and institutions using information on names, publication histories, citation patterns, and ORCID records.

The version of OpenAlex used in this study employs a four-level hierarchical classification system consisting of domains, fields, subfields, and topics. At the highest level, domains are grouped into four broad categories: Health Sciences, Life Sciences, Social Sciences, and Physical Sciences. The second level comprises 26 fields (e.g., Medicine, Mathematics), followed by 252 subfields and 4516 fine-grained topics. Each work is assigned topics algorithmically based on its title, abstract, journal, and citation context, with the highest-scoring topic determining its corresponding subfield, field, and domain. In our analysis, subfields are used as fixed effects to control for disciplinary heterogeneity, while higher-level classifications (fields and domains) are used for aggregation and robustness analyses.

To ensure data quality, we follow the filtering strategy of Arts et al. (2025) and apply several restrictions. First, we retain only non-retracted, English-language journal articles and conference proceedings, identified by requiring the "primary_location" source type to be either "journal" or "conference," the language to be "en," and the "is_retracted" flag to be false. Second, we exclude publications from journals without a publisher or with an Impact Factor of zero. Third, we remove duplicate records based on identical titles or abstracts, keeping a single instance in each case. Finally, we drop publications with missing titles or without listed authors. After these steps, the dataset contains 75,612,702 publications. We additionally construct text-based representations of scientific content using SPECTER2 (Singh et al., 2023), a transformer-based model pretrained on a large scientific corpus and optimized with citation-informed objectives. Titles and abstracts of all papers are projected into a 768-dimensional embedding space, which forms the basis for measuring cognitive distance. Recent studies have increasingly employed SPECTER2-based embeddings to characterize the semantic structure of science and to measure relationships among scientific documents, journals, and researchers. For example, Hao et al. (2026) use SPECTER2 embeddings to examine how artificial intelligence tools reshape the focus and impact of scientific research, Furnas et al. (2025) leverage embedding-based representations of scientific knowledge to analyze partisan differences in the use of science in policymaking, and Yang et al. (2026) construct a semantic atlas of scientific journals using SPECTER2-derived document representations. These applications demonstrate the utility of SPECTER2 for capturing fine-grained semantic relationships within the scientific literature.

We further restrict the sample to team-based publications—defined at the paper level as works with more than one author—published between 2000 and 2024. This yields a final analytic sample of 31,186,933 papers. Throughout the analysis, each paper is treated as a distinct team, allowing us to examine how team-level characteristics shape scientific outcomes.

3.2. Quantifying team geographic distance

To distinguish between co-located and geographically dispersed teams, we construct a continuous measure of spatial dispersion based on the institutional affiliations of a paper's authors. OpenAlex provides latitude and longitude coordinates for each institution, typically representing geographic centroids.

For each publication in our dataset, we compute the pairwise great-circle distances between the centroids of the affiliated institutions. The team-level geographic distance is then defined as the weighted average pairwise distance:

$$G_p = \sum_{i \neq j} d_{ij} a_i a_j,$$

where a_i denotes the fraction of authors in country or affiliation i, and d_{ij} is the distance between country or affiliation i and j, computed using Geographiclib-2.0 based on latitude and longitude coordinates. This metric captures the spatial dispersion of team members, providing a continuous measure of collaborative distance for each paper.

For the main analysis, we convert this continuous measure into a binary indicator. A team is classified as geographically distant (D_p=1) if its average institutional distance exceeds 10 km, and as local (D_p=0) otherwise. The 10 km threshold is chosen to identify teams whose members are effectively co-located—typically within a single campus or compact urban area—while distinguishing collaborations that involve meaningful physical separation across cities, regions, or countries. Single-institution papers are automatically classified as local teams.

3.3. Quantifying team cognitive distance

To quantify the intellectual diversity within research teams, we construct a measure of team cognitive distance based on text embeddings derived from SPECTER2 (Singh et al., 2023). This approach captures the semantic distance between researchers' knowledge bases in a continuous and high-dimensional space.

Each paper p is represented by a 768-dimensional vector $\mathbf{e}_p \varepsilon \mathbb{R}^{768}$, obtained by applying SPECTER2 to its title and abstract. These embeddings encode semantic relationships between scientific documents, such that papers with similar research content are located closer in the embedding space.

We construct time-aware author-level embeddings to approximate each researcher's knowledge profile at the time of publication. For author i contributing to paper p published in year t, we define the author embedding as the average embedding of all papers authored by i up to (but not including) year t:

$$\mathbf{a}_{i,t} = \frac{1}{|P_{i,<t}|} \sum_{q \varepsilon P_{i,<t}}^{q} \mathbf{e}_q,$$

where $\mathbb{P}_{i,<t}$ denotes the set of papers authored by i prior to year t. This construction ensures that the cognitive representation of each author reflects their accumulated knowledge without incorporating information from the focal paper.

Given a team of authors on paper p, we compute pairwise cognitive distances between authors based on their embedding vectors. Specifically, the distance between authors i and j is defined as one minus the cosine similarity:

$$d_{ij}^{c} = 1 - \frac{\mathbf{a}_{i,t} \cdot \mathbf{a}_{j,t}}{\| \mathbf{a}_{i,t} \| \| \mathbf{a}_{j,t} \|},$$

The team-level cognitive distance is then defined as the average pairwise distance:

$$C_p = \frac{2}{n_p (n_p - 1)} \sum_{i<j}^{i} d_{ij}^{c},$$

where n_p is the number of authors on paper p. This measure ranges from 0 to 1, with higher values indicating greater cognitive diversity within the team.

To facilitate interpretation and comparison across specifications, we normalize C_p to the unit interval using its empirical distribution, resulting in the percentile-based measure *cog distance* $\in [0,1]$, which is used in the regression analysis.

3.4. Quantifying interdisciplinarity

Our primary measure of interdisciplinarity is the Rao–Stirling index, which captures the extent to which a paper integrates knowledge from multiple and cognitively distant disciplines. Interdisciplinarity is constructed based on the disciplinary composition of each paper's reference list, using the third-level classification provided by the OpenAlex taxonomy. This approach ensures interpretability and consistency with established bibliometric practices, as it relies on explicit field labels rather than embedding-based semantic similarity.

For each focal paper p, let $R_p = \{r_1, r_2, \ldots, r_k\}$ denote its set of references. Each reference r_k is mapped to a subfield label, yielding a distribution of disciplines across the reference list. Let p_m denote the proportion of references belonging to subfield mmm, such that $\sum_m p_m = 1$. To capture the knowledge distance between disciplines, we construct a field-to-field distance matrix based on global co-citation patterns. For each subfield m, we define a citation profile vector V_m, representing the distribution of citations from papers in subfield m to all other subfields. The distance between subfields m and n is then computed as the cosine distance between their citation profiles, i.e., $d_{mn} = 1 - \frac{V_m \cdot V_n}{\|V_m\| \|V_n\|}$. Using these pairwise distances, the Rao–Stirling index for paper p is defined as:

$$\text{Rao} - \text{Stirling}_p = \sum_{m \neq n} d_{mn} p_m p_n.$$

This measure increases when a paper draws on a broader set of disciplines, distributes its references more evenly across them, and

combines more cognitively distant fields. It therefore provides a comprehensive measure of interdisciplinary knowledge integration.

To complement the main measure, we also report three auxiliary indicators that capture distinct dimensions of interdisciplinarity. Variety is defined as the number of unique subfields represented in the reference list. Balance is measured as Balance $= 1 - \sum_m p_m^2$, reflecting the evenness of the distribution across disciplines. Disparity is defined as the weighted average cognitive distance between cited subfields, given by Disparity $= \sum_{m \neq n} d_{mn} p_m p_n$.

3.5. Empirical strategy

To examine how geographic distance and cognitive distance jointly shape interdisciplinarity, we estimate a series of ordinary least squares (OLS) models at the paper (team) level. The primary dependent variable is the Rao–Stirling index, which captures the extent to which a paper integrates knowledge from diverse and cognitively distant disciplines. In additional specifications, we also consider variety, balance, disparity, and citation-based impact as alternative outcomes. Our main empirical model takes the following form:

$$Y_p = \beta_1 D_p + \beta_2 C_p + \beta_3 (D_p \times C_p) + \gamma X_p + \delta_{f,t} + \varepsilon_p,$$

where Y_p denotes the interdisciplinarity (or impact) of paper p, D_p is a binary indicator equal to one if the team is geographically distant and zero otherwise, and C_p is the team-level cognitive distance. The interaction term $D_p \times C_p$ captures whether the relationship between cognitive diversity and scientific outcomes differs between local and geographically dispersed teams. The vector X_p includes control variables, $\delta_{f,t}$ denotes field-by-year fixed effects, and ε_p is an error term. This specification allows us to test whether geographic distance amplifies or attenuates the contribution of cognitive diversity to interdisciplinary knowledge production.

To isolate the relationship of interest, we include a comprehensive set of controls capturing team composition, author characteristics, and paper attributes. First, we control for the number of references, which directly influences the measurement of interdisciplinarity based on reference lists. Second, we include team size, as larger teams may have access to more diverse knowledge resources. We further account for characteristics of the first and last authors, who typically represent leading and senior contributors within the team. Specifically, we include their prior scholarly impact measured by the h-index and their academic age.

We also control for the gender of the first and last authors. Gender is inferred from authors' first names using the probabilistic model developed by Van Buskirk et al. (2023), which associates names with gender probabilities based on large-scale name–gender data. Following their approach, authors with an estimated probability greater than 0.5 of being female are coded as female (=1), while all others are coded as male (=0). Validation reported by Van Buskirk et al. (2023) indicates an accuracy rate of approximately 72–76% across different populations. In the descriptive statistics, the reported values therefore represent the proportion of authors classified as female within each group.

In addition, we control for the institutional status of the first author's affiliated institution, which serves as a proxy for differences in research environments, scientific resources, and organizational prestige. Institutional status is derived from OpenAlex institutional citation records. Specifically, we rank all institutions according to their cumulative citation counts in OpenAlex and assign them to four status tiers based on their position in the global citation distribution. Tier 1 institutions comprise the top 1% of institutions by total citations, Tier 2 institutions fall between the 1st and 5th percentiles, Tier 3 institutions fall between the 5th and 20th percentiles, and Tier 4 institutions comprise the remaining 80% of institutions. This control accounts for differences in institutional resources, visibility, and research capacity that may influence collaboration patterns and interdisciplinary outcomes.

All specifications include subfield and year fixed effects to account for systematic differences across disciplines and temporal trends in interdisciplinarity. These fixed effects absorb variation in citation practices, field-specific collaboration norms, and broader changes in the scientific system over time.

Table 1
Summary statistics for geographically distant and local teams.

Variable	Geo distant teams		Geo local teams		Diff.	P values (Welch T)
	Sample size	Mean ± S.D.	Sample size	Mean ± S.D.		
Geo distance	14,701,742	2931.860 ± 3538.235	16,485,191	0.072 ± 0.684	2931.787	$p < 0.001$
Cog distance	14,486,689	0.504 ± 0.275	15,464,754	0.496 ± 0.301	0.008	$p < 0.001$
Citation percentile	14,701,742	0.632 ± 0.284	16,485,191	0.577 ± 0.292	0.055	$p < 0.001$
Rao–Stirling	14,701,742	0.232 ± 0.107	16,485,191	0.224 ± 0.110	0.008	$p < 0.001$
Variety	14,701,742	6.583 ± 4.300	16,485,191	5.916 ± 4.096	0.666	$p < 0.001$
Balance	14,701,742	0.441 ± 0.233	16,485,191	0.455 ± 0.238	−0.014	$p < 0.001$
Disparity	13,937,660	0.885 ± 0.055	15,264,407	0.883 ± 0.058	0.002	$p < 0.001$
Team size	14,701,742	6.518 ± 25.576	16,485,191	4.423 ± 9.998	2.094	$p < 0.001$
Number of references	14,701,742	37.599 ± 31.169	16,485,191	32.642 ± 29.292	4.958	$p < 0.001$
First author h-index	14,701,742	6.466 ± 8.920	16,481,497	5.014 ± 7.640	1.452	$p < 0.001$
First author academic age	14,701,742	12.926 ± 14.477	16,481,497	11.627 ± 14.006	1.3	$p < 0.001$
First author female	12,863,457	0.368 ± 0.482	14,205,573	0.363 ± 0.481	0.005	$p < 0.001$
First author institution tier	14,701,742	2.456 ± 1.058	16,485,191	2.491 ± 1.071	−0.035	$p < 0.001$
Last author h-index	14,701,742	10.852 ± 13.381	16,481,497	8.356 ± 11.423	2.496	$p < 0.001$
Last author academic age	14,701,722	17.106 ± 16.110	16,481,438	15.195 ± 15.555	1.911	$p < 0.0017$
Last author female	12,764,239	0.321 ± 0.467	14,103,497	0.327 ± 0.469	−0.007	$p < 0.001$

In this framework, the coefficient β_2 captures the relationship between cognitive distance and interdisciplinarity for local teams, while β_3 measures how this relationship differs for geographically distant teams. A positive and significant interaction term indicates that cognitive diversity contributes more strongly to interdisciplinary outcomes when teams are geographically dispersed.

4. Results

4.1. Descriptive statistics

We begin by presenting descriptive statistics comparing key variables across team types defined by geographic and cognitive distance.

Table 1 reports summary statistics stratified by team geographic distance, distinguishing between geographically distant teams (weighted average institutional distance $>$ 10 km; (N $\approx$ 14.70 million) and local teams ($\leq$ 10 km; (N $\approx$ 16.49 million). All differences are statistically significant at the ($p < 0.001$) level based on Welch's t-tests.

Geographically distant teams exhibit slightly higher cognitive distance on average (0.504 vs. 0.496), and are associated with substantially higher scientific impact, as measured by citation percentile (0.632 vs. 0.577). In terms of interdisciplinarity, distant teams show higher Rao–Stirling diversity (0.232 vs. 0.224), greater variety in referenced disciplines (6.583 vs. 5.916), and marginally higher disparity (0.885 vs. 0.883), but lower balance (0.441 vs. 0.455), indicating a more uneven distribution of knowledge across fields.

Geographically distant teams are also characterized by systematically different team composition and resources. They tend to cite more references (37.599 vs. 32.642), and their authors have higher prior scholarly impact (first author h-index 6.466 vs. 5.014, last author h-index 10.852 vs. 8.356), greater academic age (first author age 12.926 vs. 11.627, last author h-index 17.106 vs. 15.195), and are affiliated with higher-tier institutions (2.456 vs. 2.491). Taken together, these patterns suggest that geographically dispersed teams combine broader and more heterogeneous knowledge sources, while also drawing on more experienced and better-resourced researchers.

Table 2 presents similar comparisons based on team cognitive distance, dividing the sample into above-median and below-median groups. Again, all differences are statistically significant at ($p < 0.001$).

Teams with higher cognitive distance display markedly higher interdisciplinarity. They have higher Rao–Stirling diversity (0.235 vs. 0.221), greater variety (6.370 vs. 6.197), and higher disparity (0.886 vs. 0.882), but lower balance (0.433 vs. 0.463), indicating that cognitively diverse teams draw on a wider and more distant set of fields, albeit in a less even manner.

In contrast to geographic distance, however, higher cognitive distance is associated with lower citation impact (0.582 vs. 0.635). Moreover, cognitively distant teams tend to have fewer references, lower author h-indices, lower academic age, and are affiliated with lower-tier institutions. This suggests that while cognitive diversity promotes interdisciplinary knowledge integration, it may also be associated with fewer resources or greater coordination challenges that limit short-term impact.

Overall, these descriptive patterns highlight that geographic and cognitive distance are associated with distinct and sometimes contrasting team characteristics and outcomes.

4.2. Distance mismatch and interdisciplinarity

We next examine how geographic and cognitive distance jointly shape interdisciplinarity across scientific fields. Our analysis covers all 26 fields defined in OpenAlex, including disciplines such as Medicine, Physics, Computer Science, and Social Sciences. We begin by documenting temporal patterns, and then turn to multivariate regression results.

Fig. 1 presents the temporal evolution of the Rao–Stirling index from 2000 to 2024 across all fields, separately for geographically distant teams (red lines) and local teams (blue lines). Two patterns emerge. First, interdisciplinarity has increased steadily over time in

Table 2
Comparison of team characteristics by cognitive distance.

Variable	Cognitive distance > median		Cognitive distance < median		Diff.	P values (Welch T)
	Sample size	Mean ± S.D.	Sample size	Mean ± S.D.		
Citation percentile	14,882,105	0.582 ± 0.293	15,069,338	0.635 ± 0.279	−0.053	p < 0.001
Rao–Stirling	14,882,105	0.235 ± 0.108	15,069,338	0.221 ± 0.109	0.014	p < 0.001
Variety	14,882,105	6.370 ± 4.211	15,069,338	6.197 ± 4.225	0.173	p < 0.001
Balance	14,882,105	0.433 ± 0.231	15,069,338	0.463 ± 0.239	−0.03	p < 0.001
Disparity	14,006,812	0.886 ± 0.055	14,100,868	0.882 ± 0.057	0.004	p < 0.001
Team size	14,882,105	5.538 ± 5.957	15,069,338	5.202 ± 5.556	0.335	p < 0.001
Number of references	14,882,105	33.580 ± 27.329	15,069,338	37.159 ± 33.163	−3.579	p < 0.001
First author h-index	14,882,105	5.197 ± 7.736	15,069,338	6.467 ± 8.902	−1.27	p < 0.001
First author academic age	14,882,105	12.059 ± 14.220	15,069,338	12.839 ± 14.307	−0.78	p < 0.001
First author female	12,908,541	0.375 ± 0.484	13,206,936	0.355 ± 0.479	0.019	p < 0.001
First author institution tier	14,882,105	2.499 ± 1.066	15,069,338	2.415 ± 1.055	0.084	p < 0.001
Last author h-index	14,882,105	8.573 ± 11.507	15,069,338	10.985 ± 13.410	−2.412	p < 0.001
Last author academic age	14,882,064	15.791 ± 15.836	15,069,338	17.009 ± 15.848	−1.218	p < 0.001
Last author female	12,813,543	0.332 ± 0.471	13,104,377	0.314 ± 0.464	0.018	p < 0.001

nearly all fields, indicating a broad shift toward more integrative research. The only exception is the Energy field, which exhibits a slight decline over the period. Second, geographically distant teams consistently exhibit higher levels of interdisciplinarity than local teams. This difference is statistically significant in 25 out of the 26 fields examined (Welch's *t*-test, $p < 0.001$). The sole exception is Physics and Astronomy, where geographically distant teams exhibit lower levels of interdisciplinarity than local teams. Despite this exception, the overall pattern across fields suggests a positive association between geographic distance and interdisciplinary knowledge integration.

These descriptive patterns are confirmed in the regression analysis reported in Table 3 (Models 1–2). Without controls, geographically distant teams are associated with a substantially higher normalized Rao–Stirling index (coef. = 0.0596, s.e. = 0.0004). After controlling for team characteristics, author attributes, and fixed effects, the coefficient remains positive and highly significant, albeit attenuated (coef. = 0.0143, s.e. = 0.0004). This suggests that part of the observed advantage reflects differences in team composition, but a robust positive association persists.

We next examine differences by team cognitive distance, dividing teams into above-median (cognitively distant) and below-median (cognitively proximate) groups. Fig. 2 shows that, across all 26 fields, cognitively distant teams consistently produce more interdisciplinary work, as measured by the Rao–Stirling index.

This pattern is again supported by the regression results in Table 3 (Models 3–4). Cognitive distance is strongly positively associated with interdisciplinarity, both without controls (coef. = 0.2361, s.e. = 0.0006) and with controls (coef. = 0.2862, s.e. = 0.0006). The magnitude of these coefficients indicates that cognitive distance is a primary driver of interdisciplinary knowledge integration.

We now turn to our central question: how geographic and cognitive distance jointly shape interdisciplinarity. Fig. 3 presents the relationship between cognitive distance (binned) and the Rao–Stirling index, separately for geographically distant (red) and local (blue) teams across fields.

A clear and consistent pattern emerges. When cognitive distance is low, local teams tend to exhibit higher interdisciplinarity. In contrast, when cognitive distance is high, geographically distant teams outperform local teams in producing interdisciplinary work. This crossover pattern is observed in the majority of fields.

This evidence points to a distance mismatch mechanism: geographic separation enhances interdisciplinarity only when teams are already cognitively diverse. When cognitive distance is limited, spatial dispersion does not translate into meaningful knowledge recombination and may even be associated with lower interdisciplinarity.

The regression results in Table 3 (Models 5–6) formally confirm this pattern. The interaction between geographic distance and cognitive distance is positive and highly significant, both without controls (coef. = 0.1527, s.e. = 0.0012) and with controls (coef. = 0.1497, s.e. = 0.0012). This indicates that the marginal effect of cognitive distance on interdisciplinarity is substantially stronger for geographically distant teams.

Taken together, these findings suggest that geographic distance does not uniformly promote interdisciplinarity. Rather, its effect depends critically on the cognitive composition of the team: distant collaboration is most conducive to interdisciplinary knowledge integration when it bridges substantial cognitive differences.

4.3. Distance mismatch and the components of interdisciplinarity: variety, balance, and disparity

We next unpack the distance–interdisciplinarity relationship by examining the three constituent dimensions of interdisciplinarity—variety, balance, and disparity—following the decomposition introduced in Section 3.4. This allows us to assess whether the patterns observed for the Rao–Stirling index are driven by differences in the breadth, distribution, or cognitive distance of referenced knowledge.

Figures S1–S3 present the temporal evolution of variety, balance, and disparity across all fields, separately for geographically distant and local teams. Two broad patterns emerge. First, both variety and disparity increase steadily over time in nearly all fields, mirroring the overall rise in interdisciplinarity. Second, geographically distant teams consistently exhibit higher levels of variety and disparity than local teams, but lower levels of balance. This indicates that distant teams tend to draw on a broader and more cognitively distant set of disciplines, while concentrating their references more unevenly across them. These descriptive patterns are confirmed by the regression results reported in Tables 4–6 (with full specifications including all controls provided in Tables S1–S3). After accounting for team characteristics, author attributes, and fixed effects, geographically distant teams are associated with significantly higher variety and disparity, but significantly lower balance.

We next examine differences by team cognitive distance, dividing teams into above-median (cognitively distant) and below-median (cognitively proximate) groups. As shown in Figures S4–S6, cognitively distant teams consistently exhibit higher variety and disparity, but lower balance, across all 26 fields. The regression results in Tables 4–6 corroborate these findings. Cognitive distance is strongly positively associated with variety and disparity, but negatively associated with balance. This suggests that cognitively diverse teams tend to combine a wider range of disciplines and more distant knowledge domains, but do so in a less even manner.

We now turn to the interaction between geographic and cognitive distance for each dimension of interdisciplinarity. Figures S7–S9 plot the relationship between binned cognitive distance and each outcome, separately for geographically distant and local teams. For variety and disparity, a clear interaction pattern emerges. When cognitive distance is low, the difference between distant and local teams is minimal. However, as cognitive distance increases, geographically distant teams increasingly outperform local teams, producing work that spans a broader and more cognitively distant set of disciplines. This crossover pattern is consistent across most fields. In contrast, the pattern is reversed for balance. As cognitive distance increases, geographically distant teams exhibit lower balance relative to local teams, indicating a more uneven distribution of references across disciplines. The regression results in Tables 4–6 formally confirm these patterns. For variety and disparity, the interaction between geographic distance and cognitive distance is

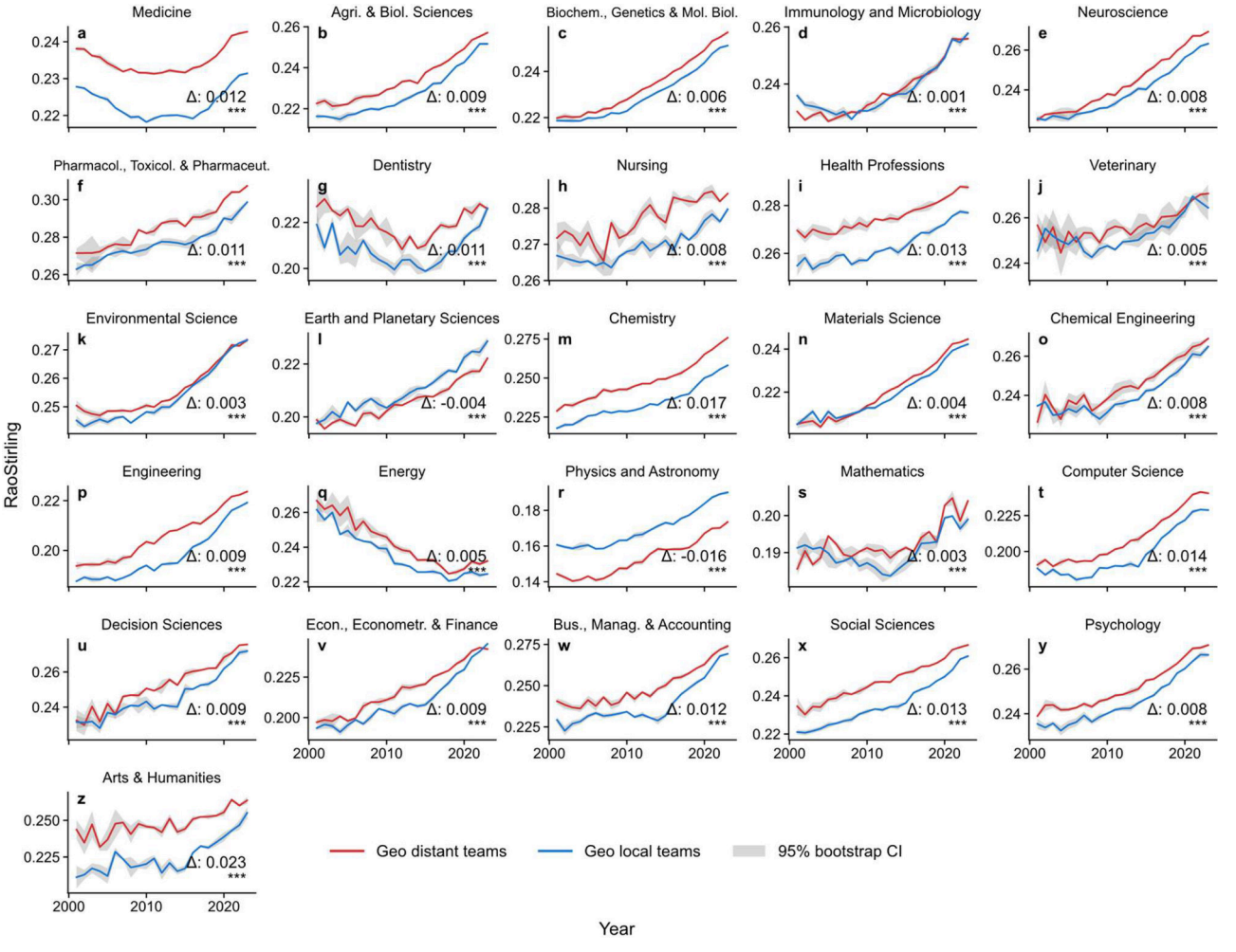


Fig. 1. Temporal trends in Rao–Stirling index for geographically distant versus local teams.
Notes: Lines show the annual mean Rao–Stirling index for geographically distant teams (red) and geographically local teams (blue) across the 26 OpenAlex fields. The reported Δ value denotes the average difference in the Rao–Stirling index between geographically distant and local teams over the study period (distant minus local). Positive values indicate higher interdisciplinarity among geographically distant teams, whereas negative values indicate higher interdisciplinarity among local teams. *** indicates Welch's *t*-test $p < 0.01$.

Table 3
. Geographic distance, cognitive distance, and their interaction in predicting Rao–Stirling index.

Models	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Var.	Rao-Stirling (z score)					
Independent Var.						
Geo distant = 1	0.0596***	0.0143***			−0.0244***	−0.0635***
	(0.0004)	(0.0004)			(0.0007)	(0.0007)
Cog distance			0.2361***	0.2862***	0.1677***	0.2209***
			(0.0006)	(0.0006)	(0.0008)	(0.0008)
Geo distant = 1 × Cog distance					0.1527***	0.1497***
					(0.0012)	(0.0012)
Control Var.						
log Number of references		0.2907***		0.2952***		0.2944***
		(0.0003)		(0.0003)		(0.0003)
log Team size		0.0363***		0.0185***		0.0151***
		(0.0004)		(0.0004)		(0.0004)
log First author h-index		−0.0197***		−0.0085***		−0.0083***
		(0.0003)		(0.0003)		(0.0003)
log First author academic age		0.0101***		0.0051***		0.0049***
		(0.0003)		(0.0003)		(0.0003)
First author female		0.0519***		0.0507***		0.0506***
		(0.0004)		(0.0004)		(0.0004)
Institution tier = 1		0.0560***		0.0564***		0.0565***
		(0.0005)		(0.0005)		(0.0005)
Institution tier = 2		0.0389***		0.0403***		0.0408***
		(0.0005)		(0.0005)		(0.0005)
Institution tier = 3		0.0277***		0.0303***		0.0303***
		(0.0005)		(0.0005)		(0.0005)
log Last author h-index		−0.0091***		0.0044***		0.0045***
		(0.0003)		(0.0003)		(0.0003)
log Last author academic age		0.0129***		0.0057***		0.0055***
		(0.0003)		(0.0003)		(0.0003)
Last author female		0.0409***		0.0416***		0.0416***
		(0.0004)		(0.0004)		(0.0004)
Year	Yes	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes	Yes
Observations	31,186,933	31,186,933	29,951,443	29,951,443	29,951,443	29,951,443
R^2	0.04864	0.08709	0.05282	0.09280	0.05397	0.09329
Within R^2	0.00092	0.04130	0.00479	0.04680	0.00600	0.04732

Note: All models are estimated using OLS regression. Standard errors are reported in parentheses. Statistical significance is indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

positive and highly significant, both with and without controls. For balance, the interaction is negative and highly significant.

Taken together, these findings suggest that the "distance mismatch" mechanism operates differently across the dimensions of interdisciplinarity. Geographic distance amplifies the expansion of disciplinary breadth and cognitive reach (variety and disparity) when teams are cognitively diverse, but at the cost of reduced balance in how knowledge sources are integrated.

5. Discussion

This study examines how geographic and cognitive distance jointly shape interdisciplinary knowledge production. Moving beyond approaches that consider these dimensions in isolation, we develop a distance mismatch perspective and show that the relationship between collaboration distance and interdisciplinarity is fundamentally conditional.

Our findings highlight three key insights. First, both geographic and cognitive distance are positively associated with interdisciplinarity when considered independently. Geographically distant teams tend to draw on a broader and more diverse set of knowledge sources, while cognitively distant teams integrate ideas from more distinct disciplinary domains. These patterns are consistent with prior work emphasizing the role of diversity and recombination in scientific discovery.

Second, cognitive distance emerges as a primary driver of interdisciplinary knowledge integration. Across fields and specifications, teams with greater cognitive diversity consistently exhibit higher levels of Rao–Stirling diversity, as well as higher variety and disparity. This suggests that differences in knowledge backgrounds are central to the production of interdisciplinary research, reinforcing the idea that intellectual heterogeneity is a key ingredient for combining distant ideas.

Third, and most importantly, we uncover a robust interaction between geographic and cognitive distance. Geographic distance enhances interdisciplinarity only when teams are already cognitively diverse. When cognitive distance is low, geographically local teams perform equally well or even better. This crossover pattern suggests that spatial separation does not inherently generate novel knowledge combinations; rather, its benefits depend on whether teams span sufficiently distinct knowledge domains. In other words, geographic distance amplifies the returns to cognitive diversity but does not substitute for it.

Decomposing interdisciplinarity into variety, balance, and disparity further clarifies the mechanisms underlying this relationship.

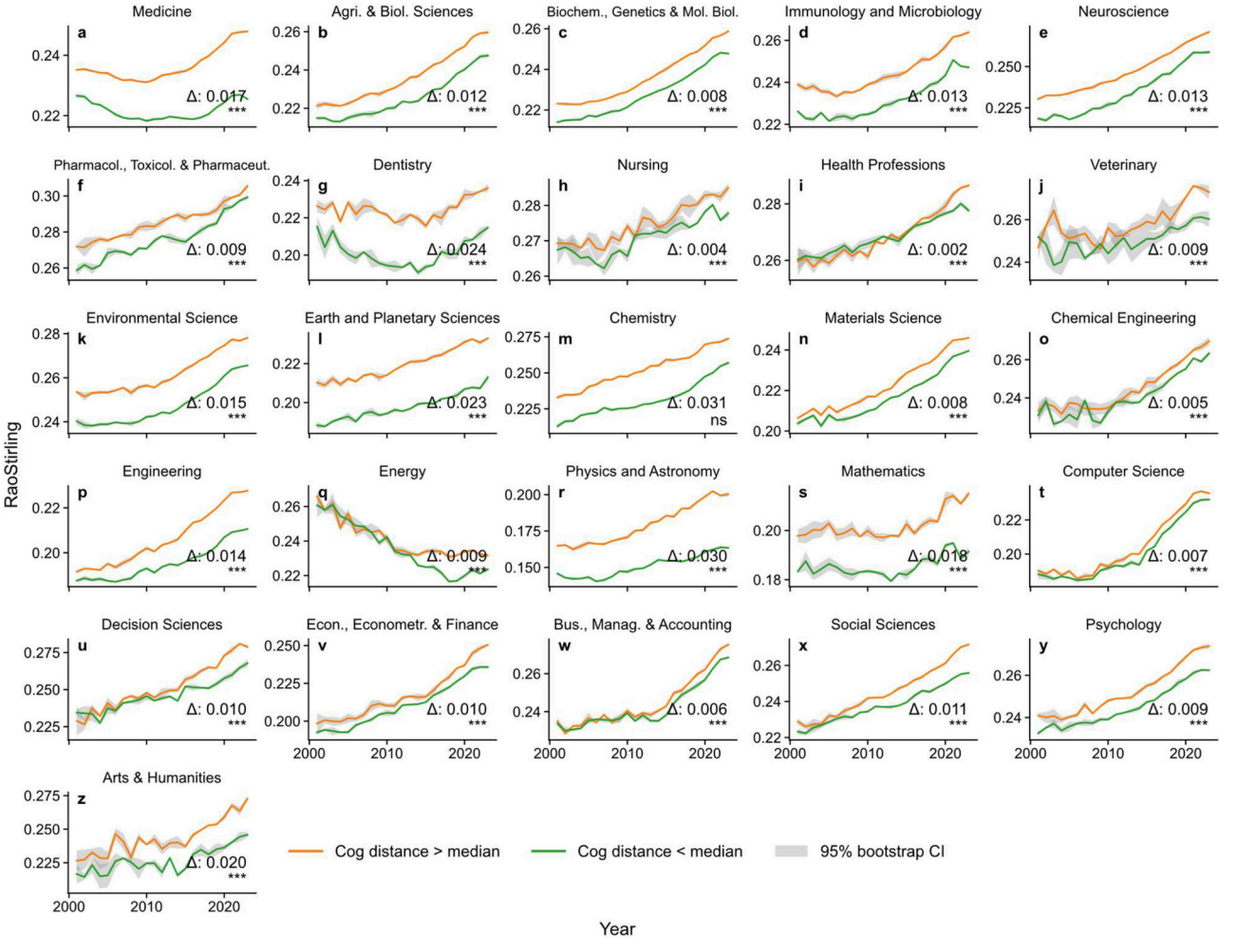


Fig. 2. Temporal trends in Rao–Stirling index for cognitively distant versus proximate teams.
Notes: Lines show the annual mean Rao–Stirling index for cognitively distant teams (red) and cognitively proximate teams (blue) across the 26 OpenAlex fields. The reported Δ value denotes the average difference in the Rao–Stirling index between cognitively distant and cognitively proximate teams over the study period (distant minus proximate). Positive values indicate higher interdisciplinarity among cognitively distant teams, whereas negative values indicate higher interdisciplinarity among cognitively proximate teams. *** indicates Welch's *t*-test $p < 0.01$.

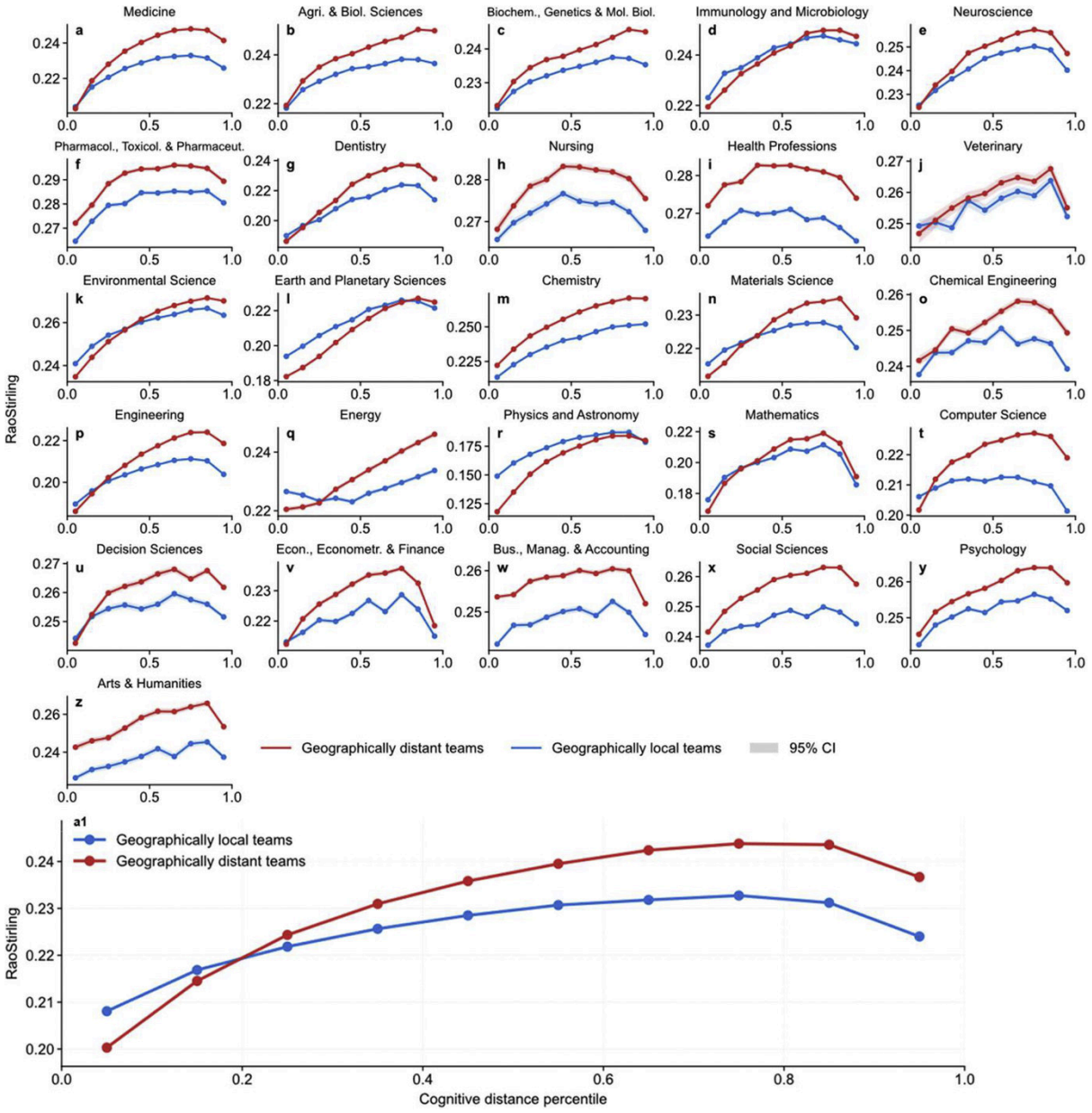


Fig. 3. Rao–Stirling index by cognitive distance, comparing geographically distant and local teams.

Geographic distance and cognitive distance both increase variety and disparity, indicating that distant teams draw on a wider range of disciplines and more cognitively distant knowledge domains. However, both dimensions are associated with lower balance, suggesting that the integration of knowledge becomes more uneven as diversity increases. This highlights an important trade-off: while distance promotes the expansion and reach of knowledge recombination, it may also complicate the even integration of different knowledge components.

These findings contribute to several strands of literature. First, they help reconcile mixed evidence on the role of geographic distance in scientific collaboration. Prior studies have reported both positive and negative effects of spatial dispersion, often without a unifying explanation (Abramo et al., 2024; Gui et al., 2019; Ma & Huang, 2024; Pialli, 2023). Our results suggest that these inconsistencies may arise from differences in the cognitive composition of teams. Second, by integrating geographic and cognitive perspectives, this study advances a more comprehensive framework for understanding collaboration and knowledge production. Third, the decomposition of interdisciplinarity provides a more nuanced view of how different dimensions of diversity are shaped by team structure.

Beyond the academic literature, our findings have implications for the organization of scientific work. As collaboration becomes increasingly global, simply expanding geographic reach may not be sufficient to foster interdisciplinary innovation. Instead, effective

Table 4
. Geographic distance, cognitive distance, and their interaction in predicting variety.

Models	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Var.	Variety					
Independent Var.						
Geo distant = 1	0.1169***	0.0012***			0.0430***	−0.0601***
	(0.0003)	(0.0003)			(0.0007)	(0.0006)
Cog distance			0.0420***	0.1977***	−0.0168***	0.1463***
			(0.0006)	(0.0005)	(0.0008)	(0.0006)
Geo distant = 1 × Cog distance					0.1271***	0.1180***
					(0.0012)	(0.0010)
Control Var.	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes	Yes
Observations	31,186,933	31,186,933	29,951,443	29,951,443	29,951,443	29,951,443
R^2	0.12073	0.44367	0.11784	0.44454	0.12096	0.44483
Within R^2	0.00381	0.36969	0.00016	0.37044	0.00369	0.37076

Note: All models are estimated using OLS regression. Standard errors are reported in parentheses. Statistical significance is indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5
. Geographic distance, cognitive distance, and their interaction in predicting balance.

Models	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Var.	Balance					
Independent Var.						
Geo distant = 1	−0.0518***	−0.0165***			0.0238***	0.0535***
	(0.0003)	(0.0004)			(0.0007)	(0.0007)
Cog distance			−0.2283***	−0.2719***	−0.1658***	−0.2130***
			(0.0006)	(0.0006)	(0.0008)	(0.0008)
Geo distant = 1 × Cog distance					−0.1395***	−0.1349***
					(0.0012)	(0.0012)
Control Var.	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes	Yes
Observations	31,186,933	31,186,933	29,951,443	29,951,443	29,951,443	29,951,443
R^2	0.06200	0.09307	0.06591	0.09885	0.06683	0.09927
Within R^2	0.00070	0.03380	0.00453	0.03964	0.00552	0.04009

Note: All models are estimated using OLS regression. Standard errors are reported in parentheses. Statistical significance is indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6
. Geographic distance, cognitive distance, and their interaction in predicting disparity.

Models Dependent Var.	(1) Disparity	(2)	(3)	(4)	(5)	(6)
Independent Var.						
Geo distant = 1	0.0233***	0.0228***			−0.0015*	0.0013. (0.0007)
	(0.0004)	(0.0004)			(0.0007)	
Cog distance			0.1436***	0.1557***	0.1234***	0.1370***
			(0.0006)	(0.0006)	(0.0008)	(0.0008)
Geo distant = 1 × Cog distance					0.0442***	0.0420***
					(0.0012)	(0.0012)
Control Var.	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes	Yes
Observations	29,202,067	29,202,067	28,107,680	28,107,680	28,107,680	28,107,680
R^2	0.11858	0.12034	0.11794	0.11987	0.11808	0.12002
Within R^2	0.00015	0.00216	0.00192	0.00411	0.00208	0.00428

Note: All models are estimated using OLS regression. Standard errors are reported in parentheses. Statistical significance is indicated as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

team design requires aligning spatial dispersion with cognitive diversity. Teams that are both geographically distributed and cognitively complementary are more likely to achieve meaningful knowledge integration, whereas mismatches between these dimensions may limit the benefits of collaboration.

At the same time, several limitations should be noted. First, our analysis is based on observational data, and although we control for

a rich set of covariates and fixed effects, causal interpretation remains limited. Second, our measure of cognitive distance relies on embedding-based representations of prior publications, which, while powerful, may not fully capture all dimensions of expertise or tacit knowledge. Third, the binary classification of geographic distance simplifies a more continuous spatial structure, potentially obscuring finer-grained patterns. Future research could explore causal identification strategies, alternative measures of cognitive diversity, and more detailed spatial metrics to further refine our understanding.

CRediT authorship contribution statement

Alex J. Yang: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Yujie Shi:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation. **Fanming Wang:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis. **Xinyu Song:** Writing – review & editing, Writing – original draft, Validation. **Sanhong Deng:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Investigation, Conceptualization.

Declaration of competing of Interest

The authors declare no competing interest relevant to this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China Young Student (Doctoral) Project (#724B2015).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.joi.2026.101840.

References

Abramo, G., Apponi, F., & D'Angelo, C. A (2024). The moderating role of the territorial research infrastructure on the geographic proximity effect in research collaborations: A regional-based view. *Scientometrics, 129*(6), 3149–3168.

Aceves, P., & Evans, J. A. (2024). Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science. *Organization Science, 35*(3), 788–814. https://doi.org/10.1287/orsc.2023.1686

Arts, S., Melluso, N., & Veugelers, R. (2025). Beyond citations: Measuring novel scientific ideas and their impact in publication text. *The Review of Economics and Statistics*, 1–33. https://doi.org/10.1162/rest_a_01561

Bernal, P., Carree, M., & Lokshin, B. (2022). Knowledge spillovers, R&D partnerships and innovation performance. *Technovation, 115*, Article 102456. https://doi.org/10.1016/j.technovation.2022.102456

Cao, Y., Sickles, R. C., Triebs, T. P., & Tumlinson, J. (2024). Linguistic distance to English impedes research performance. *Research Policy, 53*(4), Article 104971. https://doi.org/10.1016/j.respol.2024.104971

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, Media richness and structural design. *Management Science, 32*(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554

Dennis, A. R., Fuller, R. M., & Valacich, J. S. (2008). Media, tasks, and communication processes: A theory of Media synchronicity. *MIS Quarterly, 32*, 575–600.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 47*(1), 117–132. https://doi.org/10.1287/mnsc.47.1.117.10671

Fortunato, S., Bergstrom, C. T., Boerner, K., Evans, J. A., Helbing, D., Milojevic, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabasi, A.-L. (2018). Science of science. *Science (New York, N.Y.), 359*(6379). https://doi.org/10.1126/science.aao0185. Article eaao0185.

Gui, Q., Liu, C., & Du, D. (2019). Globalization of science and international scientific collaboration: A network perspective. *Geoforum; Journal of Physical, Human, and Regional Geosciences, 105*, 1–12.

Jones, B. F. (2021). The rise of research teams: Benefits and costs in economics. *Journal of Economic Perspectives, 35*(2), 191–216. https://doi.org/10.1257/jep.35.2.191

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science (New York, N.Y.), 322*(5905), 1259–1262. https://doi.org/10.1126/science.1158357

Leahey, E., Beckman, C. M., & Stanko, T. L. (2017). Prominent but less productive:The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly, 62*(1), 105–139. https://doi.org/10.1177/0001839216665364

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal, 14*(S2), 95–112.

Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics, 116*(3), 2113–2121. https://doi.org/10.1007/s11192-018-2810-y

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

Ma, X., & Huang, T. (2024). Proximity still matters in research collaboration! evidence from the introduction of new airline routes and high-speed railways in China. *Scientometrics, 129*(4), 2227–2253.

Pialli, G. (2023). The effects of limited exhaustibility of knowledge and geographical distance on the quality of R&D collaborations: The European evidence 2000–2012. *Journal of Evolutionary Economics, 33*(4), 1125–1169.

Priem, J., Piwowar, H.A., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint*.

Seo, H., Chung, Y., & Yoon, H. (2017). R&D cooperation and unintended innovation performance: Role of appropriability regimes and sectoral characteristics. *Technovation, 66-67*, 28–42. https://doi.org/10.1016/j.technovation.2017.03.002

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2023). Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

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

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science (New York, N.Y.), 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, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science (New York, N.Y.), 316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Xie, X., Liu, X., & Chen, J. (2023). A meta-analysis of the relationship between collaborative innovation and innovation performance: The role of formal and informal institutions. *Technovation, 124*, Article 102740. https://doi.org/10.1016/j.technovation.2023.102740

Yadav, A., McHale, J., & O'Neill, S. (2023). How does co-authoring with a star affect scientists' productivity? Evidence from small open economies. *Research Policy, 52*(1), Article 104660. https://doi.org/10.1016/j.respol.2022.104660

Zou, B., Wang, Y., Kwoh, C. K., & Cen, Y. (2023). Directed collaboration patterns in funded teams: A perspective of knowledge flow. *Information Processing & Management, 60*(2), Article 103237. https://doi.org/10.1016/j.ipm.2022.103237