Cognitive Distance in Scientific Teams and Innovative Research Outcomes

Abstract: Scientific innovation increasingly emerges from collaborative research teams, yet the relationship between collaborators' prior knowledge differences and innovative outcomes remains insufficiently understood. We introduce a large-scale measure of team cognitive distance using transformer-based scientific embeddings derived from more than 37 million OpenAlex publications between 1970 and 2024. Dynamic author-level knowledge representations are constructed from researchers' prior publication histories, and team cognitive distance is defined as the average semantic distance among coauthors with usable pre-publication knowledge embeddings in the embedding space. We examine how this measure is associated with two complementary outcomes: scientific disruptiveness and combinatorial novelty. Greater team cognitive distance is positively associated with both outcomes, and the relationship remains positive across scientific fields, publication cohorts, and team sizes. The pattern is also observed when disruptiveness is measured over a common five-year citation window, when nonlinear terms are included, and in a biomedical replication based on PubMed records and PubMedBERT embeddings. Interaction analyses show that the strength of the association varies across organizational and compositional contexts.

Keywords: collaboration; cognitive distance; disruptiveness; novelty; scientific embeddings

Introduction

Scientific research has become increasingly collaborative across disciplines, reflecting the growing scale, specialization, and technical complexity of knowledge production [1–3]. As teams become the dominant organizational form of science, an important question is how the knowledge backgrounds of collaborators are related to the novelty and disruptiveness of the work they publish. This question

concerns not only whether teams are diverse in observable characteristics, but also whether their members bring substantively different prior knowledge to the collaboration.

Prior research links team diversity to scientific outcomes through two competing processes. Heterogeneous teams can access a wider range of concepts, methods, and problem-solving approaches, which may facilitate recombinant innovation[4,5]. At the same time, differences in expertise, language, and research practices can increase communication and coordination costs[6]. The net relationship is therefore theoretically ambiguous: cognitive differences may expand the opportunity set for knowledge recombination while making integration more difficult.

Existing studies have examined interdisciplinarity, international and institutional collaboration, gender composition, career-stage diversity, hierarchy, and shared leadership[7–14]. These dimensions are important, but they are imperfect proxies for collaborators' underlying knowledge. Researchers classified in the same field can follow very different intellectual trajectories, whereas researchers assigned to different fields can work on closely related problems. Consequently, categorical indicators may miss substantial within-field cognitive variation and may overstate differences at disciplinary boundaries.

Recent advances in scientific representation learning make it possible to measure these differences more directly. Transformer-based models represent publications as dense vectors based on their textual content and, in the case of SPECTER2, citation-informed training objectives[15,16]. Aggregating embeddings from each researcher's prior publications yields a dynamic representation of accumulated knowledge, which can then be compared across collaborators. This continuous approach complements traditional measures by capturing fine-grained semantic differences both within and across fields.

We define *team cognitive distance* as the average pairwise semantic distance among coauthors with usable pre-publication knowledge embeddings. The concept refers to separation in accumulated scientific knowledge before the focal collaboration, rather than demographic difference, organizational separation, or the topical distance of the focal paper itself. We apply this measure to large-scale OpenAlex publication and citation data and examine its association with two distinct forms of innovation: disruptiveness, which captures whether subsequent work cites a focal paper without also citing its predecessors, and combinatorial novelty, which captures atypical combinations of referenced journals[17].

The analysis addresses three research questions. First, how does team cognitive distance vary across scientific fields, countries, and historical periods? Establishing these descriptive patterns is foundational because it identifies where cognitively heterogeneous collaboration is concentrated and whether its

prevalence is changing as science becomes more specialized. Second, how is team cognitive distance associated with disruptive and novel research? Third, how does this association vary across institutional, international, career-stage, gender, and disciplinary forms of team diversity?

The analytic sample contains 37,230,124 team-authored publications from 1970–2024. Team cognitive distance varies substantially across fields and countries and declines over time across most of the scientific system. In regression models with publication-year and field fixed effects, paper- and author-level controls, and heteroskedasticity-robust standard errors, cognitive distance is positively associated with both disruptiveness and novelty. Separate models estimated within each field and each publication year yield consistently positive associations. The relationship is also examined using a common five-year citation window for papers published through 2019, raw outcome measures, a quadratic specification, and a PubMed/PubMedBERT replication.

These results contribute a fine-grained measurement strategy for studying cognitive heterogeneity in science and document broad, robust associations between collaborators' prior knowledge distance and innovative outcomes. The observational design does not identify a causal effect: team formation is endogenous, and unobserved researcher, institutional, funding, or topic characteristics may influence both cognitive distance and subsequent outcomes. We therefore interpret the findings as associational evidence and discuss causal identification as an important direction for future research.

Background

Cognitive distance and knowledge integration

Cognitive distance describes the separation between actors' knowledge bases, interpretive frames, or technological capabilities. In collaborative research, it captures how different team members' accumulated expertise is before they work together. The concept is closely related to theories of organizational learning and recombinant innovation: greater distance can provide nonredundant knowledge and expand opportunities for recombination, whereas very large differences can make knowledge transfer, mutual understanding, and integration more difficult[4–6,18]. Empirical work on interorganizational innovation has consequently emphasized an "optimal cognitive distance," while research on scientific evaluation shows that intellectual distance can also shape how novel ideas are recognized and assessed[18,19]. These arguments motivate examination of both the average association and possible nonlinearity in scientific teams.

The scientific-team setting is especially relevant because collaboration increasingly joins specialized researchers whose expertise is distributed across

people and organizations. A team may benefit from access to distinct intellectual resources, yet the resulting coordination burden depends on whether members can translate and combine those resources. Cognitive distance is therefore conceptually different from diversity in demographic attributes or affiliations: it concerns the content and history of what collaborators know.

Team composition and scientific innovation

A large literature examines how scientific team composition is associated with impact and innovation. International and multi-institutional collaborations can broaden access to resources and audiences [8]; interdisciplinary teams can recombine atypical knowledge [7,20]; and fresh collaborative ties can increase originality [21]. Team hierarchy and career-stage composition also matter. Flatter teams are associated with more disruptive work [10], career-age differences reveal power dynamics within scientific teams [11], and heterogeneous shared leadership is associated with higher team impact than homogeneous leadership [12].

Demographic diversity constitutes another related but distinct dimension. Prior work documents persistent differences in participation and recognition across demographic groups and reports associations between gender-diverse collaboration and novelty or impact [9,22]. Recent studies further show that the relationship can vary by field, team composition, and internal team roles, and may be nonlinear [13,14]. These findings reinforce the importance of distinguishing multiple dimensions of heterogeneity rather than treating diversity as a single construct.

Measuring team cognitive distance

Traditional measures of cognitive difference typically rely on disciplinary labels, keyword overlap, topic distributions, patent or journal classifications, or citation patterns. These approaches are interpretable and often scalable, but each imposes a relatively coarse representation. Classification-based measures are sensitive to taxonomy boundaries; keyword measures depend on vocabulary choices; topic models require a prespecified dimensional structure; and citation-based measures may partly reflect strategic referencing or field-specific citation practices.

Scientific document embeddings offer a complementary approach. Scientific-domain language models such as SciBERT learn representations from scientific text, while SPECTER and SPECTER2 further optimize document representations for scientific retrieval and related tasks [15,16,23]. Such continuous representations preserve within-field variation, allow comparisons across formal disciplinary boundaries, and can be aggregated over an author's prior publications to create time-varying knowledge profiles. The approach therefore aligns the

empirical measure more closely with the theoretical definition of cognitive distance as separation between accumulated knowledge bases.

SPECTER2 uses citation-informed representation learning, which raises the possibility that an embedding-based predictor and citation-based outcomes share measurement structure. We address this concern in two ways. First, novelty is defined from atypical reference combinations rather than future citation links and yields similar results. Second, Supplementary Note 5 reports a biomedical replication using PubMed records and PubMedBERT, a text-pretrained model that does not use citation links in its pretraining objective[24].

Research gap and expectations

Despite extensive research on team diversity, large-scale evidence on latent semantic distance between collaborators' prior knowledge bases remains limited. The present study addresses this gap by constructing dynamic author profiles from publication histories and examining cognitive distance across more than 37 million team-authored papers. We expect greater cognitive distance to be associated with more atypical recombination and greater departure from established citation trajectories, while the balance between knowledge variety and coordination costs leaves open the possibility of a nonlinear relationship at high levels of distance.

Data and Methods

Paper Dataset

This study uses publication data from OpenAlex[25], an open and comprehensive database of global scholarly research maintained by OurResearch. OpenAlex integrates bibliographic information from multiple sources, including Microsoft Academic Graph, Crossref, PubMed, ORCID, and publisher websites, providing standardized metadata for more than 250 million scholarly works published between 1666 and 2025. The database covers a wide range of research outputs, including journal articles, conference proceedings, books, and datasets, and includes detailed information on authors, institutional affiliations, citations, publication venues, concepts, and funding sources. An important advantage of OpenAlex is its large-scale entity disambiguation system, which assigns persistent identifiers to authors, institutions, and publications using information on names, publication histories, citation networks, and ORCID records.

OpenAlex organizes scientific knowledge in a hierarchical taxonomy of domains, fields, subfields, and topics. At the highest level, publications are grouped into four broad domains: Health Sciences, Life Sciences, Social Sciences, and Physical Sciences. These domains are further divided into 26 fields, 252 subfields, and 4,516 fine-grained topics. Topic assignments are

generated algorithmically from publication titles, abstracts, venues, and citation relationships. We use the 26 field categories to characterize disciplinary variation and include field fixed effects in the regression models to account for broad differences across areas of science.

The analysis uses the January 2025 OpenAlex snapshot. The original OpenAlex data contain 249,803,279 publication records and 100,418,971 disambiguated authors. We retain non-retracted English-language journal articles and conference proceedings, remove duplicate or incomplete publication records, and require each retained publication to contain a title or abstract for embedding, following the general procedure of Arts et al.[26]. After these preprocessing steps, the dataset contains 75,547,123 publications associated with 35,258,587 authors. We then require authors to have at least two publications with a title or abstract so that a prior knowledge embedding can be constructed. This publication-history requirement leaves 14,036,737 authors linked to 44,876,812 publications. Supplementary Note 1 describes these filtering steps in greater detail.

To capture the semantic content of scientific publications, we use the `allenai/specter2_base` checkpoint of SPECTER2[16], a transformer-based representation model trained on large-scale scientific corpora using citation-informed objectives. Titles and abstracts are concatenated and projected into a 768-dimensional embedding space, allowing semantic similarity and cognitive distance between publications to be measured quantitatively.

The analysis then restricts the publication set to team-produced research published from 1970 through 2024, with more than one author, a title or abstract, at least one observed citation, at least one reference, and complete author-affiliation metadata for the variables used in the analysis. The resulting analytic universe contains 37,230,124 publications. The disruption regressions contain 37,213,395 papers for which the disruption outcome and all covariates are available. The novelty regressions contain 33,740,972 papers because novelty additionally requires at least two references to form a journal pair. For the five-year CD analysis, we include only papers published through 2019 so that every focal paper has at least a five-year citation window, yielding 32,928,637 observations. Throughout the analysis, each publication is treated as a research-team observation.

Constructing Scientific Embeddings

To measure cognitive distance between researchers, we first construct a continuous representation of scientific knowledge based on the semantic content of publications. We operationalize scientific knowledge using SPECTER2 embeddings[16], a transformer-based language model trained on large-scale scientific corpora and optimized using citation-informed contrastive learning

objectives. Compared with conventional keyword- or citation-based approaches, embedding models provide a dense and continuous representation of scientific content that enables fine-grained measurement of conceptual similarity across publications.

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ denote the set of publications in the sample. Each publication p_i contains textual information, including a title and abstract, which is converted into a sequence of tokens and processed by the pretrained SPECTER2 encoder. Formally, the embedding function is defined as

$$f_\theta : X_i \rightarrow e_i, \tag{1}$$

where X_i denotes the tokenized text of publication i, $f_\theta(\cdot)$ represents the pretrained SPECTER2 model parameterized by θ, and $e_i \in \mathbb{R}^{768}$ is the resulting 768-dimensional embedding vector.

To facilitate similarity comparisons, all embeddings are normalized to unit length. Semantic similarity between two publications p_i and p_j is then measured using cosine similarity:

$$\text{sim}(p_i, p_j) = \tilde{e}_i \cdot \tilde{e}_j, \tag{2}$$

where $\tilde{e}_i$ and $\tilde{e}_j$ denote normalized embedding vectors. Higher cosine similarity indicates greater semantic overlap between publications, while lower similarity reflects larger cognitive distance.

Collecting all publication embeddings yields the matrix

$$E = \begin{bmatrix} \tilde{e}_1 \\ \tilde{e}_2 \\ \vdots \\ \tilde{e}_N \end{bmatrix} \in \mathbb{R}^{N \times 768}, \tag{3}$$

which defines a continuous semantic representation of the scientific literature. In this embedding space, publications addressing related research questions, methods, or conceptual domains are located closer together, whereas intellectually distant works occupy more separated regions.

Compared with discrete field classifications, embedding-based representations preserve continuous variation in scientific knowledge and allow comparisons both within and across disciplines. This feature is particularly important for studying interdisciplinary collaboration, where cognitive differences may arise even among researchers formally classified within the same field.

Measuring Team Cognitive Distance

We conceptualize team cognitive distance as the extent to which collaborators differ in their accumulated knowledge bases prior to collaboration. To operationalize this concept, we construct dynamic author-level knowledge embeddings using researchers' publication histories.

Let $\mathcal{P}_{i,<t}$ denote the set of publications authored by researcher i before year t. The knowledge representation of author i at time t is defined as the centroid of embeddings from all prior publications:

$$a_{i,t} = \frac{1}{|\mathcal{P}_{i,<t}|} \sum_{p_k \in \mathcal{P}_{i,<t}} \tilde{e}_k, \tag{4}$$

where $\tilde{e}_k$ is the normalized embedding vector of publication p_k, and $a_{i,t} \in \mathbb{R}^{768}$ represents the latent cognitive position of researcher i immediately prior to year t.

This approach assumes that a researcher's cognitive orientation is reflected in the semantic content of prior publications. Averaging across historical outputs provides a stable representation of accumulated expertise while reducing noise associated with individual papers. To preserve temporal ordering and avoid mechanical endogeneity, the focal publication is excluded from the construction of $a_{i,t}$. In addition, we restrict the analysis to authors with at least two prior publications before year t to ensure reliable estimation of knowledge profiles. Coauthors who do not meet this history requirement are not assigned a knowledge embedding and are excluded from the pairwise-distance calculation for the focal paper. A publication enters the cognitive-distance analysis only when at least two coauthors have usable pre-publication histories.

Using these author-level embeddings, we compute cognitive distance between collaborators based on cosine distance. For any pair of authors i and j with usable pre-publication knowledge embeddings, pairwise cognitive distance is defined as

$$d_{ij}^{(c)} = 1 - \frac{a_{i,t} \cdot a_{j,t}}{\|a_{i,t}\| \|a_{j,t}\|}. \tag{5}$$

Lower values indicate cognitively similar researchers with overlapping expertise, whereas higher values reflect greater intellectual separation between collaborators. Although cosine distance theoretically ranges from 0 to 2, empirical values are concentrated within $[0, 1]$ because scientific embeddings exhibit strong positive clustering. We use cosine distance rather than Euclidean distance because cosine-based measures capture angular differences independent of vector magnitude. This property is important in high-dimensional embedding spaces,

where variation in publication volume may affect embedding norms without necessarily reflecting substantive differences in knowledge content.

For each publication p, let $\mathcal{A}_p$ denote the set of coauthors with usable pre-publication knowledge embeddings and let $n_p = |\mathcal{A}_p|$. Team cognitive distance is defined as the average pairwise cognitive distance among coauthors in $\mathcal{A}_p$:

$$C_p = \frac{2}{n_p(n_p - 1)} \sum_{i<j} d_{ij}^{(c)}. \tag{6}$$

This measure aggregates all unique pairs of coauthors in $\mathcal{A}_p$ and divides by $n_p(n_p - 1)/2$, the number of pairs formed by these coauthors. Higher values indicate teams composed of researchers with more distinct knowledge bases, whereas lower values correspond to cognitively homogeneous teams.

To facilitate interpretation and comparability across specifications, we normalize the raw cognitive distance measure using its empirical percentile distribution:

$$\widetilde{C}_p = \frac{\text{rank}(C_p)}{N}, \tag{7}$$

where $\text{rank}(C_p)$ denotes the percentile rank of publication p in the pooled global analytic sample and N is the total number of observations. Percentiles are therefore not recomputed within year, field, or field-year cells. The normalized measure ranges from 0 to 1; year and field differences are handled separately through descriptive stratification and fixed effects.

Conceptually, this framework treats scientific collaboration as the integration of distinct knowledge trajectories within a shared semantic space. Team cognitive distance therefore captures the extent of intellectual recombination embedded in collaborative research production.

Figure 1 summarizes the workflow used to construct the measure of team cognitive distance. Figure 1b illustrates the semantic embedding space generated from publication embeddings, where geometric proximity reflects conceptual similarity between scientific works. Publications from related research areas form local clusters, whereas intellectually distant topics occupy more separated regions of the space.

To characterize researchers' knowledge bases, we construct dynamic author-level embeddings from publication histories prior to collaboration. As illustrated in Figure 1c, the embedding for author i at time t is computed as the centroid of all publications authored before year t. This procedure generates a time-varying representation of accumulated expertise while preserving temporal ordering by excluding the focal publication.

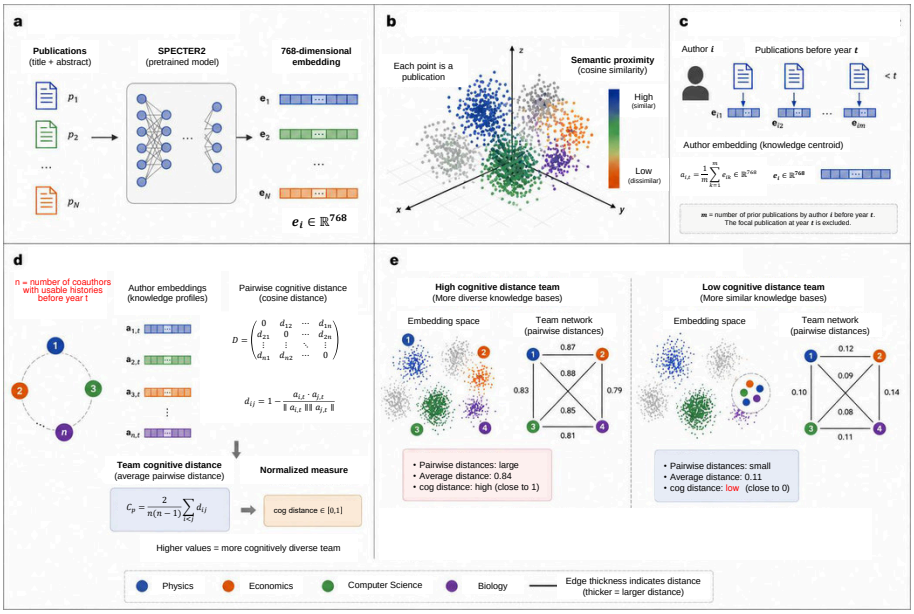


Figure 1. Framework for measuring team cognitive distance using scientific embeddings. (a) Scientific publications are embedded using SPECTER2. (b) Papers are mapped into a semantic embedding space. (c) Author knowledge profiles are constructed from at least two publications before the focal publication year. (d) Team cognitive distance is the average pairwise cosine distance among coauthors with usable pre-publication knowledge embeddings. Here, $n = n_p = |\mathcal{A}_p|$ counts only these coauthors, rather than all listed authors, and $n \geq 2$. (e) Illustration of high- and low-distance research teams, showing coauthors with usable pre-publication knowledge embeddings.

Figure 1d illustrates the construction of team cognitive distance. For each publication, pairwise cosine distances are computed only between coauthors with usable pre-publication knowledge embeddings, and the team-level measure is the average over all unique pairs within this subset. In Figure 1d, n denotes the number of these coauthors (i.e., $n = n_p = |\mathcal{A}_p|$), each with at least two prior publications before year t; the measure requires $n \geq 2$. Figure 1e contrasts teams with high and low cognitive distance. The illustrated authors are coauthors with usable pre-publication knowledge embeddings. High-distance teams consist of collaborators located in distant regions of the embedding space, reflecting the integration of heterogeneous expertise from multiple scientific domains. In contrast, low-distance teams are composed of researchers positioned in nearby regions of the embedding space, indicating substantial overlap in prior knowledge and intellectual background.

Quantifying Disruptiveness and Novelty

To characterize the innovative properties of scientific publications, we examine two complementary dimensions of research output: scientific disruptiveness and combinatorial novelty. Disruptiveness captures the extent to which a publication redirects subsequent scientific attention away from prior work, whereas novelty reflects the degree to which a paper recombines previously disconnected bodies of knowledge.

We measure scientific disruptiveness using the Consolidation–Disruption (CD) index[27]. The CD index evaluates whether a publication disrupts existing scientific trajectories or instead consolidates prior knowledge by examining the local citation network surrounding the focal paper.

The intuition underlying the measure is that disruptive papers reduce the future relevance of the studies on which they build. If subsequent publications cite the focal paper without also citing its predecessors, the focal paper effectively substitutes for earlier work and is therefore considered disruptive. By contrast, if future studies continue to cite both the focal paper and its references, the focal paper is interpreted as consolidating existing knowledge rather than redirecting scientific attention.

Formally, the CD index for focal paper p is defined as

$$CD = \frac{1}{n} \sum_{i=1}^{n} \frac{-2 f_i b_i + f_i}{w_i}, \tag{8}$$

where $f_i = 1$ if citing paper i cites the focal paper and 0 otherwise, and $b_i = 1$ if citing paper i also cites any of the focal paper's references and 0 otherwise. Here, n denotes the total number of papers connected to either the focal paper or its predecessors, and w_i represents citation weights.

Under this formulation, citations that reference the focal paper but not its predecessors increase the disruptive character of the focal paper, whereas citations that reference both the focal paper and its predecessors increase its consolidating character. The CD index ranges from -1 to $+1$. Values close to -1 indicate highly consolidating research that reinforces existing knowledge structures, whereas values close to $+1$ indicate highly disruptive work with greater potential to redirect scientific trajectories. Large-scale studies using this family of measures document substantial temporal variation in disruptiveness, underscoring the importance of accounting for publication cohorts[28].

In the main analysis, we calculate the CD index using the full observed citation network. Because citation opportunity differs across publication cohorts, we treat this measure as a long-window outcome rather than as a directly comparable exposure period. We therefore also calculate CD using citations

received during the first five years after publication. This fixed-window analysis includes only papers published through 2019, ensuring at least five complete years of follow-up for every focal paper. Table S2 reports percentile outcomes, and Table S3 reports raw CD and CD5 outcomes.

Because the distribution of CD scores is highly skewed, we follow Wu et al.[29] and convert raw CD values into percentile-based measures of disruptiveness. When multiple papers share identical CD scores, we assign the corresponding median percentile rank. This transformation improves comparability across specifications while preserving the ordinal structure of disruptiveness.

To measure the novelty of scientific research, we adopt the atypical knowledge recombination framework proposed by Uzzi et al.[7]. This approach evaluates the extent to which a publication combines journals that are rarely cited together relative to expectations derived from randomized citation networks.

Following prior work, we generate ten Monte Carlo simulations of reshuffled citation networks by randomly reassigning citation links while preserving the temporal and degree-distribution characteristics of the observed network. Using both the empirical and randomized networks, we calculate the frequency with which journal pairs appear jointly within the reference lists of focal papers.

For each journal pair (m, n), we compute a standardized z-score:

$$Z_{mn} = \frac{\text{obs}(pair_{mn}) - \exp(pair_{mn})}{\sigma(pair_{mn})}, \tag{9}$$

where $\text{obs}(pair_{mn})$ denotes the observed frequency of journal pair (m, n) in the empirical citation network, $\exp(pair_{mn})$ represents the expected frequency across randomized networks, and $\sigma(pair_{mn})$ is the corresponding standard deviation estimated from the Monte Carlo simulations.

For each focal paper, the collection of journal-pair z-scores characterizes the atypicality of its referenced knowledge combinations. A paper must contain at least two references to form at least one reference pair and enter the novelty analysis. In the main analysis, we summarize the distribution using the median z-score across all journal pairs cited by the paper. Lower median z-scores indicate more atypical combinations of prior knowledge sources and therefore greater novelty.

Following prior literature, we transform the measure to reduce skewness and facilitate interpretation:

$$Novelty = \begin{cases} -\ln(Z_{\text{med}} + 1), & \text{if } Z_{\text{med}} > 0, \\ \ln(-Z_{\text{med}} + 1), & \text{if } Z_{\text{med}} \leq 0, \end{cases} \tag{10}$$

where Z_{med} denotes the median journal-pair z-score for the focal paper. Higher values indicate more atypical combinations of knowledge and therefore greater novelty in the recombination process underlying scientific discovery.

As an additional robustness check, the Supplementary Materials report results using the 10th percentile of the journal-pair z-score distribution, following the original specification proposed by Uzzi et al.[7]. The substantive results remain unchanged across alternative operationalizations.

Empirical Specification

To account for alternative explanations of scientific innovation, we include a broad set of paper- and author-level controls commonly used in the science of science literature. At the paper level, we control for the number of references and team size. Reference-list length captures variation in the breadth of prior knowledge incorporated into a publication and may mechanically influence both disruptiveness and novelty measures. Team size accounts for differences in collaborative scale that may affect coordination, specialization, and knowledge recombination.

At the author level, we include several pre-publication characteristics of the first author because the first author commonly has a central role in developing and executing the focal study, while these variables also provide parsimonious controls for prior visibility, experience, and output. Prior citations measure accumulated scientific recognition before the focal paper; career age is the number of years since the first observed publication; and productivity is the number of prior publications. We also include an indicator for whether the first author is inferred as female. Gender is inferred from authors' first names using the open-source statistical model developed by Van Buskirk et al.[30], which estimates probabilistic name–gender associations by combining multiple sources of empirical name data. Gender is included as a compositional control rather than a primary object of analysis. Supplementary Note 1 provides further details on the procedure and its limitations. All models include publication-year and field fixed effects.

In addition to cognitive distance, prior research has emphasized the role of organizational, demographic, and disciplinary diversity in shaping scientific innovation. To examine whether these dimensions mediate or condition the relationship between team cognitive distance and innovation outcomes, we construct several additional measures of team diversity.

We define a *multi-institutional team* as a collaboration involving authors affiliated with more than one institution, and an *international team* as a collaboration spanning multiple countries. To capture heterogeneity in professional experience, we compute the Gini coefficient of coauthor career

ages within each team, where higher values indicate greater inequality in career stage and accumulated experience among collaborators.

We further measure gender diversity using team gender entropy:

$$H_p = -\sum_{g \in G} s_{pg} \ln(s_{pg}), \tag{11}$$

where s_{pg} denotes the share of team members assigned to inferred gender category g by the first-name classification model. Higher entropy values indicate a more balanced inferred gender composition within the team.

Finally, we define an *interdisciplinary team* as a collaboration involving authors associated with different scientific fields based on their prior publication histories. This measure captures disciplinary heterogeneity within collaborative teams.

To examine the relationship between team cognitive distance and scientific innovation, we estimate ordinary least squares (OLS) regression models of the following form:

$$Y_p = \beta_0 + \beta_1 C_p + \mathbf{X}_p' \boldsymbol{\gamma} + \delta_f + \delta_t + \varepsilon_p, \tag{12}$$

where Y_p denotes the innovation outcome of publication p, including disruptiveness or novelty, C_p represents team cognitive distance, and $\mathbf{X}_p$ denotes the vector of control variables. The terms δ_f and δ_t represent scientific field and publication-year fixed effects, respectively.

To examine whether the relationship between cognitive distance and innovation varies across different forms of team diversity, we additionally estimate interaction models of the following form:

$$Y_p = \beta_0 + \beta_1 C_p + \beta_2 D_p + \beta_3 (C_p \times D_p) + \mathbf{X}_p' \boldsymbol{\gamma} + \delta_f + \delta_t + \varepsilon_p, \tag{13}$$

where D_p denotes a specific diversity measure, including multi-institutional collaboration, international collaboration, career-age inequality, gender diversity, or interdisciplinarity. The interaction coefficient β_3 captures whether the association between team cognitive distance and scientific innovation differs systematically across alternative forms of team heterogeneity. The main regression tables use heteroskedasticity-robust standard errors. Supplementary Note 6 reports venue-clustered, year-clustered, and two-way year-and-field-clustered standard errors for the principal specifications in Table 1; two-way year-and-field-clustered inference for the interaction and fixed-window models; and principal-model estimates after excluding teams with more than 10 authors. Given the very large sample, we emphasize coefficient magnitudes, consistency across specifications, and confidence intervals rather than statistical significance alone.

Separate field-specific models include year fixed effects, while separate year-specific models include field fixed effects.

Skewed count variables—reference count, team size, first-author prior citations, career age, and prior productivity—are transformed as $\ln(1 + x)$. Team cognitive distance, disruptiveness, novelty, Gini coefficients, entropy measures, and binary indicators remain on the scales defined above.

Results

Patterns and trends in team cognitive distance

Before examining the relationship between team cognitive distance and scientific innovation, we first describe how cognitive distance varies across disciplines, over time, and across countries. Understanding these descriptive patterns is important because the organization of scientific collaboration differs substantially across research domains and institutional environments. Figure 2 presents three broad findings. First, team cognitive distance varies considerably across scientific fields. Second, cognitive distance declines systematically over time across nearly all disciplines. Third, substantial geographic heterogeneity exists across national research systems.

Figure 2a reports the average level of team cognitive distance across scientific fields. We observe pronounced disciplinary variation in the cognitive composition of research teams. Computer Science exhibits the highest average cognitive distance, followed by Neuroscience, Veterinary Science, Nursing, and Engineering. By contrast, Mathematics and Physics & Astronomy display the lowest levels of cognitive distance, with Chemistry and Business, Management & Accounting also located near the lower end of the distribution. These patterns suggest that applied and interdisciplinary domains rely more heavily on the integration of heterogeneous knowledge bases, whereas theoretically consolidated disciplines tend to involve collaborations among researchers with more overlapping expertise.

Despite these cross-sectional differences, Figure 2b reveals a strikingly consistent temporal trend across fields: team cognitive distance declines steadily over time throughout the scientific system. To improve readability, the figure highlights the twelve largest fields by publication volume using colored lines, while all remaining fields are shown in grey. Although fields differ in their initial levels and rates of decline, the downward trajectory is nearly universal. Fields such as Medicine, Environmental Science, Agricultural & Biological Sciences, and Materials Science begin the period with relatively high levels of cognitive distance in the 1970s but experience substantial declines over subsequent decades. Similar patterns are also observed in Engineering, Neuroscience, Social Sciences, and Earth & Planetary Sciences.

The widespread decline in cognitive distance suggests that scientific collaboration increasingly occurs among researchers with more similar prior knowledge bases. One possible explanation is that growing specialization and the increasing complexity of scientific research have encouraged collaboration within narrower intellectual domains. As the stock of scientific knowledge expands, researchers may depend more heavily on collaborators with closely related expertise, thereby reducing the cognitive breadth of research teams. At the same time, persistent differences across disciplines indicate that field-specific knowledge structures continue to shape the extent of intellectual heterogeneity in scientific collaboration.

Figure 2c further shows substantial geographic variation in team cognitive distance across countries. Higher average levels are observed in several countries in East Asia, parts of Africa, and parts of Latin America, whereas lower values are more common in Europe. These cross-national differences likely reflect variation in the organization of scientific systems, including differences in disciplinary specialization, international collaboration patterns, and institutional research capacity. Countries with smaller or less specialized research communities may rely more heavily on cognitively diverse collaborations, whereas mature scientific systems may exhibit denser within-field collaboration networks.

Team cognitive distance and scientific innovation

We next examine whether cognitively distant collaborations are associated with more innovative scientific outcomes. The analytic universe contains 37,230,124 publications; outcome-specific samples are 37,213,395 papers for disruptiveness and 33,740,972 papers for novelty, the latter requiring at least two references. Figure 3 presents descriptive relationships across publication cohorts, scientific fields, and team sizes.

To visualize these relationships, we divide the distribution of team cognitive distance into ten equally spaced bins and compute the mean outcome within each bin. Shaded regions represent 95% confidence intervals. Across all panels, we observe a consistently positive descriptive relationship between team cognitive distance and both measures of innovation. Publications authored by cognitively distant teams have, on average, higher disruptiveness and novelty than publications authored by cognitively similar teams. These binned patterns are descriptive and do not identify a causal effect of team composition.

Figure 3a shows the relationship between team cognitive distance and disruptiveness. Average disruptiveness increases monotonically across the full distribution of cognitive distance. Publications produced by teams in the lowest cognitive-distance bin exhibit an average disruptiveness percentile of approximately 0.46, whereas publications in the highest cognitive-distance

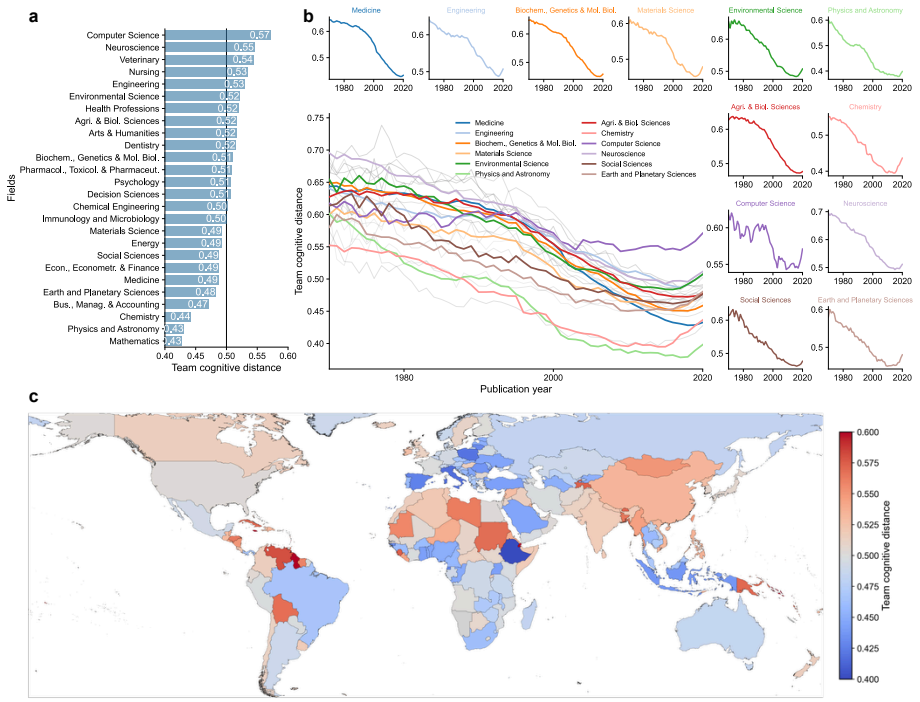


Figure 2. Distributions and temporal trends in team cognitive distance across scientific fields and countries. (a) Average team cognitive distance across scientific fields. (b) Temporal evolution of team cognitive distance by scientific field, 1970–2020. Colored lines represent the twelve largest fields by publication volume, while all remaining fields are shown in grey. (c) Geographic distribution of average team cognitive distance across countries.

bin reach approximately 0.56. The magnitude of this difference indicates a substantively meaningful association between cognitive heterogeneity and research that redirects subsequent scientific attention away from established knowledge trajectories.

A similarly strong pattern is observed for novelty. As shown in Figure 3b, publications generated by cognitively distant teams exhibit systematically higher levels of combinatorial novelty. Average novelty scores increase from approximately -5.0 among teams with very low cognitive distance to approximately -4.3 among teams with the highest levels of cognitive distance. Because higher values indicate more atypical knowledge recombination, this pattern is consistent with cognitively heterogeneous teams drawing on more unusual combinations of prior knowledge.

Importantly, the positive association between team cognitive distance and innovation outcomes remains highly stable across multiple dimensions of scientific production. Figures 3c and 3d report the relationship separately across publication cohorts from the 1970s to the 2020s. In every decade, both disruptiveness and novelty increase monotonically with cognitive distance. Although baseline levels vary somewhat across cohorts, the upward-sloping relationship remains remarkably persistent over time.

Figures 3e and 3f further demonstrate that these patterns generalize across scientific fields. Nearly all disciplines exhibit positive relationships between team cognitive distance and both disruptiveness and novelty. Although the magnitude of the relationship differs somewhat across fields, the qualitative pattern remains highly consistent. This pattern indicates that the positive association is not confined to a small set of interdisciplinary domains, but appears broadly across scientific fields.

Finally, Figures 3g and 3h examine heterogeneity across team sizes. The positive association between cognitive distance and innovation outcomes remains evident for teams of all sizes, ranging from small collaborations to large research groups. Although larger teams tend to exhibit somewhat lower levels of disruptiveness and novelty overall, the monotonic relationship between cognitive distance and innovation outcomes remains stable across all team-size categories.

Taken together, Figure 3 provides descriptive evidence that cognitive distance is systematically associated with disruptiveness and novelty across time periods, fields, and team sizes. The recurring pattern is consistent with knowledge-recombination arguments, but it should not be interpreted as evidence that cognitive distance itself causes innovative outcomes.

Regression of team cognitive distance and innovation

We next examine whether the positive association between team cognitive distance and scientific innovation persists after accounting for observable paper and author characteristics. Table 1 reports regression estimates relating team cognitive distance to two innovation outcomes: disruptiveness and novelty.

Across all specifications, team cognitive distance is positively associated with both disruptiveness and novelty. Columns (1)–(3) report results for disruptiveness, whereas Columns (4)–(6) report results for novelty. All models include publication-year and field fixed effects and report heteroskedasticity-robust standard errors. Because the sample is very large, we focus on effect sizes and stability across specifications rather than statistical significance alone. Supplementary Note 6 and Tables S6–S11 report cluster-robust inference and large-team sensitivity analyses. Confidence intervals are wider when standard errors are clustered by publication venue, publication year, or two-way by year and field, but remain

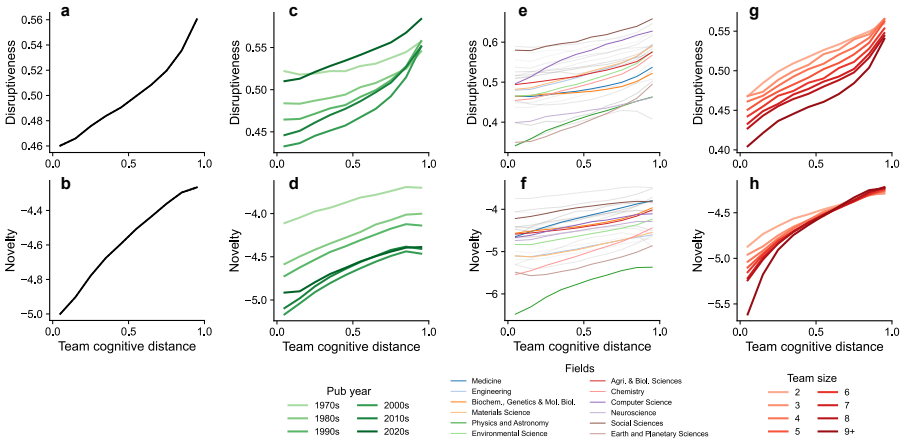


Figure 3. Relationship between team cognitive distance, disruptiveness, and novelty. Panels show mean disruptiveness and novelty across ten bins of team cognitive distance. Shaded regions represent 95% confidence intervals. (a–b) Relationship between team cognitive distance and average disruptiveness/novelty. (c–d) Relationship across publication cohorts. (e–f) Relationship across scientific fields. (g–h) Relationship across team sizes.

well above zero in the principal, interaction, and fixed-window models. Excluding teams with more than 10 authors also produces estimates close to the full-sample results.

For disruptiveness, the baseline specification reported in Column (1) yields a coefficient of 8.552 ($p < 0.01$). Because disruptiveness is measured using percentile ranks multiplied by 100, this estimate implies that moving from the minimum to the maximum level of normalized team cognitive distance is associated with an 8.55 percentile increase in disruptiveness. The coefficient declines modestly after introducing controls for reference count and team size in Column (2), and remains large and statistically significant in the fully specified model reported in Column (3), where the estimated effect equals 7.291. Thus, moving from the bottom to the top of the global cognitive-distance percentile distribution is associated with a 7.29-point higher disruptiveness percentile in the fully controlled specification.

A similarly robust relationship emerges for novelty. In the baseline model shown in Column (4), the coefficient on team cognitive distance equals 0.6319 ($p < 0.01$). The estimate decreases slightly after introducing paper- and author-level controls, but remains highly stable across specifications, reaching 0.5794 in the fully controlled model reported in Column (6). The estimates therefore indicate a positive conditional association between cognitive distance and atypical combinations of prior knowledge.

The coefficient remains positive and of similar magnitude after the inclusion of observed paper and first-author controls. This stability shows that the association is not explained by these measured covariates, but it does not rule out selection on unobserved team, topic, institution, funding, or network characteristics.

Several control variables also exhibit systematic relationships with innovation outcomes. Papers with longer reference lists and larger teams tend to be less disruptive and less novel conditional on cognitive distance. By contrast, first authors with longer career histories and higher prior productivity are associated with higher levels of disruptiveness and novelty.

Overall, the regressions document positive and substantively nontrivial associations between team cognitive distance, disruptiveness, and combinatorial novelty. The estimates reinforce the descriptive patterns in Figure 3, while remaining associational because the models do not exogenously determine team formation.

Table 1. Team cognitive distance, disruptiveness, and novelty

Dependent Variables:	Disruptiveness percentile			Novelty (median atypicality)		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Team cognitive distance	8.552***	7.710***	7.291***	0.6319***	0.5953***	0.5794***
	(0.0167)	(0.0165)	(0.0167)	(0.0010)	(0.0010)	(0.0010)
N References		-6.705***	-6.442***		-0.2778***	-0.2581***
		(0.0059)	(0.0060)		(0.0004)	(0.0004)
Team size		-2.533***	-2.300***		-0.0296***	-0.0128***
		(0.0108)	(0.0108)		(0.0006)	(0.0006)
First author citations			-1.347***			-0.1056***
			(0.0048)			(0.0003)
First author career age			0.8207***			0.0395***
			(0.0074)			(0.0004)
First author productivity			1.025***			0.1102***
			(0.0087)			(0.0005)
First-author name classified female			0.8599***			0.0991***
			(0.0099)			(0.0006)
Fixed-effects						
Year	Yes	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	37,213,395	37,213,395	37,213,395	33,740,972	33,740,972	33,740,972
R^2	0.04565	0.08723	0.09012	0.12258	0.14013	0.14511
Within R^2	0.00736	0.05061	0.05361	0.01342	0.03316	0.03876

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: This table reports OLS estimates of the relationship between team cognitive distance and scientific innovation outcomes. Columns (1)–(3) use disruptiveness percentile rank as the dependent variable, while Columns (4)–(6) use the novelty measure based on journal-pair atypicality. Disruptiveness values are multiplied by 100 for interpretability. All models include publication year and field fixed effects. Heteroskedasticity-robust standard errors are reported in parentheses. Coefficients are associational and should not be interpreted causally. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We next estimate the fully controlled specification separately within every scientific field and every publication year. Field-specific models include year fixed effects; year-specific models include field fixed effects; all use heteroskedasticity-robust standard errors. Figure 4 presents the estimated cognitive-distance coefficients and 95% confidence intervals.

Figures 4a and 4b report heterogeneity across scientific fields. Across all disciplines, team cognitive distance is positively and significantly associated with both disruptiveness and novelty. Although the magnitude of the coefficients varies across fields, the direction of the relationship remains highly consistent.

For disruptiveness, the largest coefficients are observed in Computer Science ($\beta = 10.96$), Earth & Planetary Sciences ($\beta = 10.85$), Environmental Science ($\beta = 9.36$), Engineering ($\beta = 8.47$), and Physics & Astronomy ($\beta = 8.45$). Positive and statistically significant effects are also observed in more consolidated disciplines, including Mathematics ($\beta = 6.00$), Medicine ($\beta = 5.95$), and Chemistry ($\beta = 7.92$). Even the smallest coefficients, such as those for Health Professions ($\beta = 4.00$) and Nursing ($\beta = 4.03$), remain substantively positive.

A similarly consistent pattern emerges for novelty. The estimated effects of team cognitive distance are positive across all fields, with particularly large coefficients in Medicine ($\beta = 0.82$), Physics & Astronomy ($\beta = 0.81$), Chemistry ($\beta = 0.81$), Immunology & Microbiology ($\beta = 0.65$), and Pharmaceutics ($\beta = 0.59$). Smaller but still positive effects are observed in Economics and Finance ($\beta = 0.07$), Mathematics ($\beta = 0.18$), and Business and Accounting ($\beta = 0.19$). Overall, the positive association with atypical knowledge combinations is observed across all disciplinary contexts examined.

Figures 4c and 4d further examine temporal heterogeneity in the relationship between team cognitive distance and innovation outcomes. The estimated coefficients remain positive throughout the entire observation period from 1970 to 2024. For disruptiveness, the magnitude of the coefficient increases steadily from the 1970s through the 2000s before declining modestly in recent years, although remaining strongly positive throughout. For novelty, the estimated coefficient exhibits a general upward trend over time, indicating that the novelty advantages associated with cognitive diversity have become increasingly pronounced in recent decades.

Taken together, Figure 4 shows that the positive association is robust across disciplinary and temporal subsamples. Although coefficient magnitudes differ, the estimated association remains positive and statistically distinguishable from zero in each field and year-specific analysis shown.

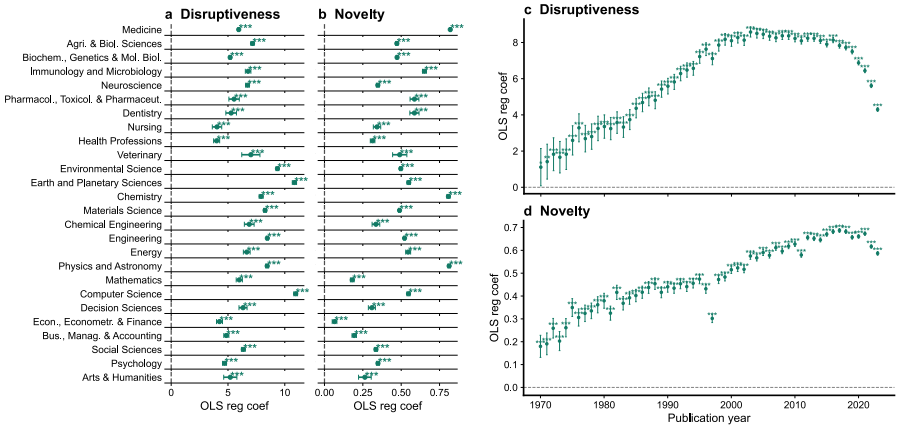


Figure 4. Heterogeneity in the relationship between team cognitive distance and innovation outcomes across fields and publication years. Panels report team-cognitive-distance coefficients from fully controlled regressions with paper- and author-level controls and heteroskedasticity-robust standard errors. Field-specific models include year fixed effects; year-specific models include field fixed effects. Error bars represent 95% confidence intervals. (a–b) Estimated coefficients across scientific fields for disruptiveness and novelty. (c–d) Estimated coefficients across publication years for disruptiveness and novelty.

Team diversity and the innovation effects of cognitive distance

Prior research has shown that multiple dimensions of team diversity shape scientific innovation and research impact. Existing studies emphasize the importance of institutional diversity, international collaboration, interdisciplinarity, demographic heterogeneity, and team hierarchy in influencing the production of novel and disruptive research. For example, gender-diverse teams have been shown to generate more novel and higher-impact scientific ideas[9], while interdisciplinary and academically diverse teams are associated with greater originality and knowledge recombination[21,31]. Research has also demonstrated that team structure and power dynamics influence collaborative outcomes, with flatter teams and teams characterized by balanced career-stage composition tending to produce more innovative research[10,11]. At the same time, geographically distributed and remote collaborations may constrain breakthrough idea generation because of coordination and communication barriers[32]. Building on this literature, we examine whether these forms of team diversity condition the relationship between team cognitive distance and scientific disruptiveness.

Table 2 reports OLS regression estimates in which team cognitive distance is interacted with several dimensions of team composition, including multi-institutional collaboration, international collaboration, career-age inequality, gender diversity, and interdisciplinarity. All models include the full set of paper- and author-level controls together with publication-year and field fixed effects.

The interaction estimates show that the association between team cognitive distance and disruptiveness varies across these dimensions. The positive interaction with multi-institutional collaboration indicates a stronger association among teams spanning institutions, and the positive interaction with international collaboration indicates a stronger association among teams spanning countries. These interaction coefficients describe conditional associations and should not be interpreted as causal benefits of organizational diversity.

A comparable pattern emerges for career-stage diversity. Although teams with greater career-age inequality exhibit lower baseline disruptiveness, the interaction between the career-age Gini coefficient and cognitive distance is strongly positive. This result indicates that the positive association with cognitive distance is larger in teams combining researchers from different career stages.

By contrast, the interactions involving gender diversity and interdisciplinarity are negative. While gender-diverse and interdisciplinary teams are themselves associated with higher levels of disruptiveness, the marginal effect of cognitive distance becomes weaker within these teams. One possible interpretation is that these forms of diversity partially overlap with the mechanisms captured by team cognitive distance, thereby reducing the incremental contribution of additional cognitive heterogeneity.

Importantly, across all specifications the coefficient on team cognitive distance remains positive and highly statistically significant. Even after accounting for multiple dimensions of team diversity and their interactions, cognitively distant teams continue to exhibit substantially higher levels of disruptiveness. The estimated coefficients remain large in magnitude, ranging from approximately 5.0 to 7.9 across models. These estimates indicate that the cognitive-distance association is not absorbed by the measured dimensions of team diversity included in these specifications.

Overall, team cognitive distance captures a dimension of collaborative heterogeneity that is empirically distinct from the organizational, demographic, career-stage, and disciplinary measures examined here. Other forms of diversity condition the magnitude of the association, but cognitive distance remains positively associated with disruptiveness in every interaction specification.

Table 2. Team diversity, cognitive distance, and disruptiveness

Dependent Variable:	Disruptiveness percentile				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Team cognitive distance	6.685***	6.661***	4.985***	7.898***	6.246***
	(0.0233)	(0.0182)	(0.0383)	(0.0237)	(0.0250)
Multi-institutional team	-1.724***				
	(0.0185)				
Team cognitive distance × Multi-institutional team	1.106***				
	(0.0313)				
International team		-3.170***			
		(0.0214)			
Team cognitive distance × International team		2.464***			
		(0.0380)			
Team age gini			-1.548***		
			(0.0612)		
Team cognitive distance × Team age gini			6.160***		
			(0.0973)		
Team gender entropy				1.724***	
				(0.0309)	
Team cognitive distance × Team gender entropy				-1.889***	
				(0.0510)	
Interdisciplinary team					3.493***
					(0.0191)
Team cognitive distance × Interdisciplinary team					-0.9838***
					(0.0332)
Controls	Yes	Yes	Yes	Yes	Yes
Fixed-effects					
Year	Yes	Yes	Yes	Yes	Yes
Field	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	37,213,395	37,213,395	37,164,190	35,113,457	37,164,190
R^2	0.09049	0.09102	0.09024	0.09083	0.09216
Within R^2	0.05400	0.05455	0.05376	0.05367	0.05576

Notes: This table reports OLS estimates examining the relationship between team cognitive distance and scientific disruptiveness conditional on multiple dimensions of team diversity. The dependent variable is the percentile rank of the disruption index (*CD_pct*), multiplied by 100 for interpretability. Each specification includes interaction terms between team cognitive distance and a specific diversity measure, including multi-institutional collaboration, international collaboration, team age inequality, gender diversity, and interdisciplinarity. All models include the full set of paper- and author-level controls reported in Table 1, together with publication year and field fixed effects. The gender-diversity model is restricted to observations for which the model-based gender classifications required to calculate team gender entropy are available. Heteroskedasticity-robust standard errors are reported in parentheses. Interaction terms describe conditional associations rather than causal moderation effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Discussion

This study develops a dynamic embedding-based measure of team cognitive distance and examines its association with scientific disruptiveness and combinatorial novelty. Across the pooled sample, field-specific models, year-specific models, alternative outcomes, a quadratic specification, and a

PubMed/PubMedBERT replication, greater cognitive distance is consistently associated with more innovative publication outcomes. The results should be read as robust observational associations rather than causal estimates.

The findings extend research on recombinant innovation by measuring heterogeneity in collaborators' accumulated knowledge directly rather than inferring it only from field labels, affiliations, or demographics. The positive association with novelty is consistent with the idea that teams whose members occupy more distant semantic positions have access to less redundant knowledge components. The positive association with disruptiveness is consistent with such teams being linked to work that departs from established citation trajectories. These interpretations describe patterns compatible with theory; they do not establish the underlying mechanism.

The analysis also distinguishes cognitive distance from other dimensions of team composition. Cognitive distance remains positively associated with disruptiveness after accounting for institutional, international, career-stage, gender, and disciplinary diversity, while the interaction results show that its association varies across these contexts. This suggests that observable diversity indicators and semantic knowledge distance overlap only partially and should not be treated as interchangeable.

The nonlinear specification adds an important qualification. The positive quadratic terms indicate an upward-curving association across the observed range: estimated differences in innovation become larger at higher levels of cognitive distance. Because both the linear and squared terms are positive, the estimated turning point lies outside the observed 0–1 range; the evidence therefore supports convexity rather than an interior U-shaped reversal. This pattern does not eliminate the possibility of coordination costs, but it indicates that such costs do not dominate within the observed range of the measure.

For research organizations and funders, the results suggest that team-formation systems should not rely solely on disciplinary categories when assessing the breadth of proposed collaborations. Semantic profiles derived from prior work may reveal complementary expertise that field labels miss. At the same time, the observational evidence does not justify mechanically maximizing cognitive distance. Teams also require communication structures, shared goals, and integration capacity; the positive average associations documented here do not guarantee that every cognitively distant collaboration will succeed.

Several limitations qualify the findings. First, team formation is endogenous. Researchers pursuing unconventional questions, working in better-resourced institutions, receiving particular forms of funding, or occupying exploratory network positions may both form cognitively distant teams and publish more innovative work. Year and field fixed effects and observed controls do not

eliminate this selection, so causal language is inappropriate. Second, SPECTER2 uses citation-informed training, which could create shared structure with citation-based outcomes. The PubMedBERT replication reduces this concern by using a text-pretrained biomedical model, but no embedding fully captures tacit knowledge, experimental skill, or unpublished expertise. Third, author disambiguation, missing text, the requirement of at least two prior publications, and the requirement of complete author-affiliation metadata generate selection into the analytic sample. These exclusions may vary across fields, countries, and periods.

Fourth, the first-name gender classification based on Van Buskirk et al.[30] is probabilistic and culturally uneven. Reported classification accuracy is approximately 72–76%, and performance may be poorer for some cultural and linguistic groups, including Asian names. The method infers binary name–gender associations rather than self-identified gender and cannot represent nonbinary identities. Because gender is included only as a compositional control in this study, gender-related coefficients are interpreted cautiously. Fifth, the full-window CD outcome provides unequal citation opportunity across cohorts. The five-year analysis restricted to papers published through 2019 imposes a common follow-up period but necessarily excludes the most recent papers. Sixth, novelty requires at least two references and is sensitive to coverage of journals and citation links. Finally, the study observes published outputs rather than failed collaborations, abandoned projects, or within-team processes.

Conclusion

Team cognitive distance, measured from collaborators' prior publication embeddings, varies substantially across fields, countries, and historical periods and has declined over time across much of science. Across multiple observational specifications, more cognitively distant teams are associated with higher disruptiveness and novelty. The contribution of the study is therefore twofold: it provides a scalable measure of latent knowledge heterogeneity and documents a broad empirical association between that heterogeneity and innovative scientific outcomes.

Future work should move from association toward identification. Promising designs include matched comparisons, author or institution panels, venue and topic-by-year fixed effects, collaboration shocks, funding or mobility discontinuities, and sensitivity analyses for unobserved confounding. Research should also examine the organizational processes that allow distant knowledge to be integrated, including communication routines, leadership, division of labor,

prior ties, and institutional support. Additional validation with alternative text-only models, multilingual corpora, richer author histories, and direct measures of collaboration processes would clarify the generalizability and mechanisms of the observed relationships.

References

1. Wuchty S, Jones BF and Uzzi B. The increasing dominance of teams in production of knowledge. *Science* 2007; 316(5827): 1036–1039. DOI:10.1126/science.1136099.
2. Jones BF, Wuchty S and Uzzi B. Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* 2008; 322(5905): 1259–1262. DOI:10.1126/science.1158357.
3. Hall KL, Vogel AL, Huang GC et al. The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist* 2018; 73(4): 532–548. DOI:10.1037/amp0000319.
4. Weitzman ML. Recombinant growth. *The Quarterly Journal of Economics* 1998; 113(2): 331–360. DOI:10.1162/003355398555595.
5. Fleming L. Recombinant uncertainty in technological search. *Management Science* 2001; 47(1): 117–132. DOI:10.1287/mnsc.47.1.117.10671.
6. Nooteboom B. *Learning and Innovation in Organizations and Economies*. Oxford: Oxford University Press, 2000. DOI:10.1093/acprof:oso/9780199241002.001.0001.
7. Uzzi B, Mukherjee S, Stringer M et al. Atypical combinations and scientific impact. *Science* 2013; 342(6157): 468–472. DOI:10.1126/science.1240474.
8. Adams J. The fourth age of research. *Nature* 2013; 497: 557–560. DOI: 10.1038/497557a.
9. Yang Y, Tian TY, Woodruff TK et al. Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences* 2022; 119(36): e2200841119. DOI:10.1073/pnas.2200841119.
10. Xu F, Wu L and Evans JA. Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences* 2022; 119(23): e2200927119. DOI:10.1073/pnas.2200927119.
11. Xu H, Bu Y, Liu M et al. Team power dynamics and team impact: New perspectives on scientific collaboration using career age as a proxy for team power. *Journal of the Association for Information Science and Technology* 2022; 73(10): 1489–1505. DOI:10.1002/asi.24653.
12. Xu H, Liu M, Bu Y et al. The impact of heterogeneous shared leadership in scientific teams. *Information Processing & Management* 2024; 61(1): 103542. DOI:10.1016/j.ipm.2023.103542.
13. Zhang C, Zeng J and Zhao Y. Is higher team gender diversity correlated with better scientific impact? *Journal of Informetrics* 2025; 19(2): 101662. DOI: 10.1016/j.joi.2025.101662.
14. Zhao Y, Zhu Y, Kim D et al. The effect of gender diversity on scientific team impact: A team roles perspective. *Journal of Informetrics* 2026; 20(1): 101766. DOI:10.1016/j.joi.2025.101766.

15. Cohan A, Feldman S, Beltagy I et al. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2270–2282. DOI:10.18653/v1/2020.acl-main.207.
16. Singh A, D'Arcy M, Cohan A et al. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 5548–5566. DOI:10.18653/v1/2023.emnlp-main.338.
17. Leahey E, Lee J and Funk RJ. What types of novelty are most disruptive? *American Sociological Review* 2023; 88(3): 562–597. DOI:10.1177/00031224231168074.
18. Nooteboom B, Van Haverbeke W, Duysters G et al. Optimal cognitive distance and absorptive capacity. *Research Policy* 2007; 36(7): 1016–1034. DOI:10.1016/j.respol.2007.04.003.
19. Boudreau KJ, Guinan EC, Lakhani KR et al. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science* 2016; 62(10): 2765–2783. DOI:10.1287/mnsc.2015.2285.
20. Leahey E, Beckman CM and Stanko TL. Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly* 2017; 62(1): 105–139. DOI:10.1177/0001839216665364.
21. Zeng A, Fan Y, Di Z et al. Fresh teams are associated with original and multidisciplinary research. *Nature Human Behaviour* 2021; 5(10): 1314–1322. DOI:10.1038/s41562-021-01084-x.
22. Hofstra B, Kulkarni VV, Munoz-Najar Galvez S et al. The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences* 2020; 117(17): 9284–9291. DOI:10.1073/pnas.1915378117.
23. Beltagy I, Lo K and Cohan A. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620. DOI:10.18653/v1/D19-1371.
24. Gu Y, Tinn R, Cheng H et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* 2021; 3(1): 1–23. DOI:10.1145/3458754.
25. Priem J, Piwowar H and Orr R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:220501833* 2022; DOI:10.48550/arXiv.2205.01833.
26. Arts S, Melluso N and Veugelers R. Beyond citations: Measuring novel scientific ideas and their impact in publication text. *The Review of Economics and Statistics* 2025; : 1–33DOI:10.1162/rest_a_01561. URL `https://doi.org/10.1162/rest_a_01561`.

27. Funk RJ and Owen-Smith J. A dynamic network measure of technological change. *Management Science* 2017; 63(3): 791–817. DOI:10.1287/mnsc.2015.2366.
28. Park M, Leahey E and Funk RJ. Papers and patents are becoming less disruptive over time. *Nature* 2023; 613(7942): 138–144. DOI:10.1038/s41586-022-05543-x.
29. Wu L, Wang D and Evans JA. Large teams develop and small teams disrupt science and technology. *Nature* 2019; 566(7744): 378–382. DOI:10.1038/s41586-019-0941-9.
30. Van Buskirk I, Clauset A and Larremore DB. An open-source cultural consensus approach to name-based gender classification. *Proceedings of the International AAAI Conference on Web and Social Media* 2023; 17(1): 866–877. DOI: 10.1609/icwsm.v17i1.22195.
31. Yang AJ, Zhao SX, Deng S et al. Quantifying the dynamics of research teams' academic diversity. *Journal of the Association for Information Science and Technology* 2025; 76(12): 1719–1735. DOI:10.1002/asi.70023.
32. Lin Y, Frey CB and Wu L. Remote collaboration fuses fewer breakthrough ideas. *Nature* 2023; 623(7989): 987–991. DOI:10.1038/s41586-023-06767-1.