The Murmuration of Scientists

Munjung Kim[1] and Yong-Yeol Ahn[1,+]

[1]School of Data Science, University of Virginia, Charlottesville, USA
[+]Corresponding authors: yyahn@iu.edu

September 1, 2025

Abstract

Scientists shift their research interests—or "move" the space of knowledge—sparking new ideas and novel methodologies. While prior studies have explored how individual characteristics or direct collaborations shape scientists' research topic shifts [1–3], they have largely overlooked the impact of the broader, dynamically evolving behavior of the peer community. Here, we create a continuous mapping of authors' positions in a high-dimensional topic space to examine how scientists "move" through the knowledge space in relations with their peers. Drawing inspiration from the boids model of collective animal behaviors, we characterize individual research movements using simple rules: alignment (matching the movement of peer scientists), cohesion (clustering with peers on similar topics), and separation (avoiding overlap when peers are too close). Our empirical analysis demonstrates that authors exhibit these propensities, with alignment and cohesion being modulated with the academic prominence. Furthermore, we develop a generative model grounded in these simple rules, which effectively predict future topic shifts based on observed peer dynamics. These findings reveal that the complex evolution of research interests may be driven by simple collective dynamics principles that resemble animal behaviors in physical space.

1 Introduction

Researchers navigate a knowledge landscape to discover new knowledge. The progress of scientific studies often arises either from continued immersion in a single domain—where methodical inquiry

and specialization gradually push disciplinary boundaries—or from venturing into unfamiliar territory and synthesizing insights across multiple fields [4, 5]. By focusing on one specialized area, scientists can accumulate substantial expertise and cultivate a high level of technical mastery over time [6]. Conversely, radical domain shifts or cross-disciplinary exploration can yield unexpected breakthroughs; such boundary-crossing efforts frequently introduce new theoretical constructs or methodologies that spark entirely novel lines of inquiry [5, 7].

Given its significant implications for scientific progress, the dynamics by which researchers choose their next research topics and by which researcher alter their research interests have garnered increasing attention in the literature. With the advance of data analysis techniques and large-scale datasets of scholarly information, the transitions of research interests of scientists are captured by the community detection based on co-citation network [2, 8], topic classification code [1, 3], and network of entities, such as a molecule, in science [4, 9]. Through the various ways of detecting the transition, scholars have examined how various attributes of individual scientists—such as collaboration [2, 3], prior training [10], and propensity for exploration or exploitation [1, 9]—influence patterns of field entry and exit.

Despite recognizing the importance of such contextual forces, most existing work has paid limited attention to how peer behavior shapes researchers' decisions to venture into new domains. Studies of topic change of researchers in relationship with other researchers frequently restrict their focus to only collaborators and often treat these collaborators' research interests as fixed. Yet in reality, collaborators' own scientific foci also evolve, and researchers might choose to align with or diverge from their colleagues that they have not collaborated with before for myriad strategic reasons—ranging from leveraging shared expertise to avoiding competition in the same topic. Consequently, whether scientists closely follow peers' dynamic shifts or deliberately separate from them remains poorly understood. Clarifying these peer-driven dynamics could reveal a fundamental mechanism through which entire disciplines reorganize in response to broader scientific currents.

In this paper, we address these questions by modeling researchers' movements in the "knowledge space," drawing inspiration from the boids model. We posit that authors' subsequent moves are shaped by three peer-related mechanisms: (i) alignment, in which authors mirror the directional shifts of their peers, (ii) cohesion, in which authors gravitate toward their peers' positions, and (iii) separation, in which authors maintain distance to sidestep overcrowding within a given area. We empirically assess the propensity of these forces using large-scale publication data. By representing knowledge and individual research trajectories in a high-dimensional embedding space—constructed

from text neural embeddings of publication abstracts—we quantify scientists' directional transitions of research interests. Then we compute how the transitions of peer authors, who are defined as the closest neighbors in the embedding space, are related to the transition of the focal author in the next step.

2 Results

2.1 Scientist Swarming Model in the Topic Space

The core premise of our framework is that scientists choose their subsequent movements in a knowledge space based on both their own state and that of their peers (Figure 1a). Each researcher's state here is defined by two components: position and velocity of the movement in the knowledge space.

We begin by conceptualizing a "topic" not as a discrete label (e.g., a subject category) but as a point in a high-dimensional semantic space derived from textual representations of scientific publications. In this view, the "knowledge space" is a continuous embedding space where proximity reflects topical similarity between research outputs. This allows us to model gradual shifts and overlaps between topics, rather than enforcing rigid categorical boundaries.

To empirically instantiate this space, we use the OpenAlex dataset [11], which includes author-disambiguated data along with their publication titles and abstracts. We generate vector representations of each paper's research topics using the SPECTER embedding model [12]. For each author, we calculate their position by averaging the topic vectors of all papers they published within a five-year window. To ensure reliability, we include only authors who have published at least five papers in each window, and we focus on those who meet this threshold across three consecutive windows to capture sustained research trajectories. As shown in Figure 1b, an author's velocity is computed as the change in their topic-space position between consecutive time windows.

In addition to tracking the movement of focal authors—those who published at least five papers in each of three consecutive five-year periods—we also identify other researchers who may have influenced them. These individuals, whom we refer to as potential peers, are defined as authors who published at least three papers in each of two consecutive five-year windows. Notably, focal authors themselves meet this criterion and are therefore included in the pool of potential peers.

Our main observation period runs from 1999 to 2018, divided into four five-year windows

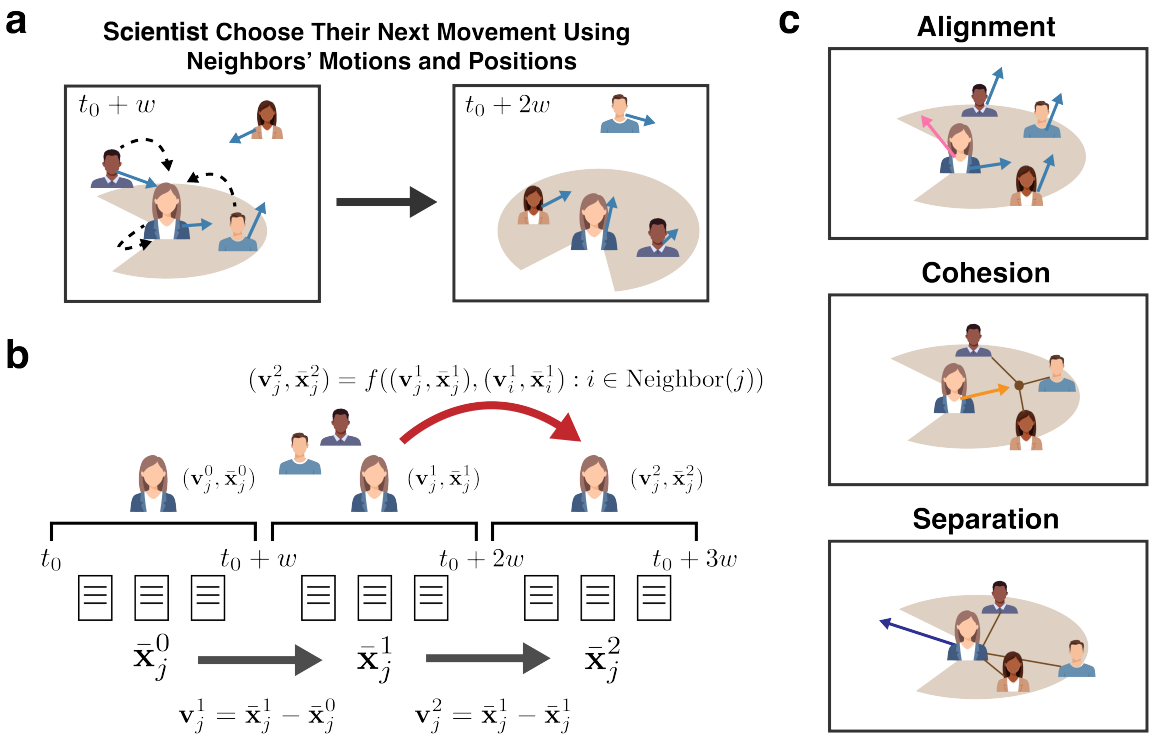

Figure 1: **Schematics of the scientist agent's research topic transition framework.** **(a)** A scientist agent navigates the knowledge landscape to pursue novel ideas and research topics. The decision on where to move next is influenced by the agent's current movement (velocity) and position as well as by the movements and positions of its peers. The set of peers influencing the agent can change over time, depending on who is closest in the research topic space at each step. **(b)** The author's research interests state is captured within a time window of length w. For each window, we compute the embedding vectors of all papers published by the author and take their average to represent the author's current research position, denoted as $\bar{\mathbf{x}}_j^t$. The author's movement, or velocity $\mathbf{v}_j^t$, is then defined as the difference between their positions in two consecutive time windows. To model how an author evolves in topic space, we define their next position and velocity, $\left(\mathbf{v}_j^{t+1}, \bar{\mathbf{x}}_j^{t+1}\right)$, using a function f that takes into account two factors: (1) the author's own prior state $\left(\mathbf{v}_j^t, \bar{\mathbf{x}}_j^t\right)$ and (2) the states of their peer authors at the same time, $(\mathbf{v}_i^t, \bar{\mathbf{x}}_i^t) : i \in \text{Neighbor}(j)$. This setup captures how both personal research trajectories and the surrounding scholarly environment influence an author's future direction. **(c)** The function f is modeled using three core dynamics inspired by the boids model of collective behavior. First, alignment encourages authors to adjust their velocity in the direction of their peers' average movement. Second, cohesion draws authors toward the central position of their neighboring peers. Third, separation pushes authors away from others who are too close in topic space, reducing overcrowding and potential competition. Together, these dynamics simulate how individual researchers respond to the collective motion of the scholarly community.

(1999–2003, 2004–2008, 2009–2013, and 2014–2018). To estimate the underlying mechanisms of the authors' movement, we restrict our analysis to the 1999–2013 period, which comprises three consecutive five-year windows, and subsequently use the 2014–2018 period to assess the predictions derived from these mechanisms. This approach yields a final sample of 1,024,539 focal authors (each publishing at least five papers per window across all three intervals) and 1,909,038 potential peer scientists (each publishing at least three papers per window across the first two intervals) for the 1999–2013 period. Then we link focal authors with their neighbor peers by identifying each focal author's 14 nearest neighbors based on cosine distance.

We then compute each author's velocity—that is, the change in their position between consecutive time windows—by taking the difference between positions in adjacent intervals. For an author j, the velocity in the time window t is calculated as $\mathbf{v}^t = \bar{\mathbf{x}}_j^t - \bar{\mathbf{x}}_j^{t-1}$, where $\bar{\mathbf{x}}_j^t$ is the average embedding vectors of the papers published by author j in time window t and $\bar{\mathbf{x}}_j^{t-1}$ is the average embedding during the preceding window. Because positions prior to the first window are not observable, velocities are only computed starting from the second window. Figure 1b illustrates how each author's position and velocity are derived. In our framework, an author's subsequent movement is determined by their own prior state as well as by the states of their peers. This relationship is formalized as

$$(\mathbf{v}_j^t, \bar{\mathbf{x}}_j^t) = f((\mathbf{v}_j^{t-1}, \bar{\mathbf{x}}_j^{t-1}), (\mathbf{v}_i^{t-1}, \bar{\mathbf{x}}_i^{t-1}) : i \in \text{Neighbor}(j)), \tag{1}$$

indicating that author j's state in window t is determined by their own prior state in window $t-1$ and by the states of their peer authors in the same period.

To construct the function f, we take inspiration from the boids model [13]. The boids model is a classic example of emergent behavior in agent-based simulations, originally proposed to capture how birds flock. Each "boid" follows three simple rules: separation (avoid getting too close to neighbors), alignment (match heading with neighbors), and cohesion (move toward the center of the group). Despite the simplicity of these rules, complex collective patterns emerge from local interactions alone.

We adapt this concept to model how scientists move through a high-dimensional topic space. Here, each author acts as an agent whose position and movement are shaped by analogous social and intellectual forces. *Alignment* captures the tendency to move in the direction that peers are also heading, reflecting a belief that convergence toward certain areas may signal emerging

opportunities or intellectually fertile ground. *Cohesion* reflects the drive to remain within the intellectual bounds of a research community. Scientists often stay close to the center of disciplinary conversations to maintain shared language, contribute to dominant paradigms, and avoid becoming isolated—resonating with Thomas Kuhn's ideas about normal science and paradigm-driven inquiry. *Separation* represents the desire to differentiate one's work from others. This includes avoiding overcrowded topics, but also reflects a deeper aspiration: to carve out a niche where one's expertise is distinctive and difficult to replace. In this sense, separation is not just about physical distance in topic space, but about establishing a unique and recognizable identity within the scientific landscape.

2.2 Alignment, Cohesion, and Separation Propensity

Building on the boids-inspired framework, we assess how each of the three behavioral dynamics manifests among scientists in the topic space. Specifically, for *alignment*, we investigate whether changes in a focal author's velocity, defined as $\Delta \mathbf{v}_j^t = \mathbf{v}_j^t - \mathbf{v}_j^{t-1}$ for an author j, are influenced by the difference between the average peer velocity and their own velocity:

$$\mathbf{A}_j^t = \left(\frac{1}{N} \sum_{i \in \text{Neighbor(j)}} \mathbf{v}_i^{t-1} \right) - \mathbf{v}_j^{t-1}. \tag{2}$$

To quantify the strength of this relationship, we computed the cosine similarity between $\Delta \mathbf{v}_j^t$ and $\mathbf{A}_j^t$. As a benchmark, we compare this value to a random baseline in which the author's prior velocity $\mathbf{v}_j^{t-1}$ is replaced with that of a randomly selected author $\mathbf{v}_m^{t-1}$. As shown in Figure 2a, the mean cosine similarity in the focal author scenario is 0.753, far higher than a random baseline of -0.0445 (Cohen's $d = 5.8$; $p < 0.001$). Thus, authors appear to adjust their velocity in alignment with the difference between their peers' average velocity and their own.

We next assessed *cohesion* by determining whether an author tends to shift their research position toward the center of their peer group. Concretely, we measured the cosine similarity between the change of the position of author j, $\Delta \mathbf{x}_j^t = \mathbf{x}_j^t - \mathbf{x}_j^{t-1}$, and the cohesion vector, which is defined as the difference between the average position of the author's peers and their own prior position:

$$\mathbf{C}_j^t = \left(\frac{1}{N} \sum_{i \in \text{Neighbor}(j)}^{k} \bar{\mathbf{x}}_i^{t-1} \right) - \bar{\mathbf{x}}_j^{t-1}. \tag{3}$$

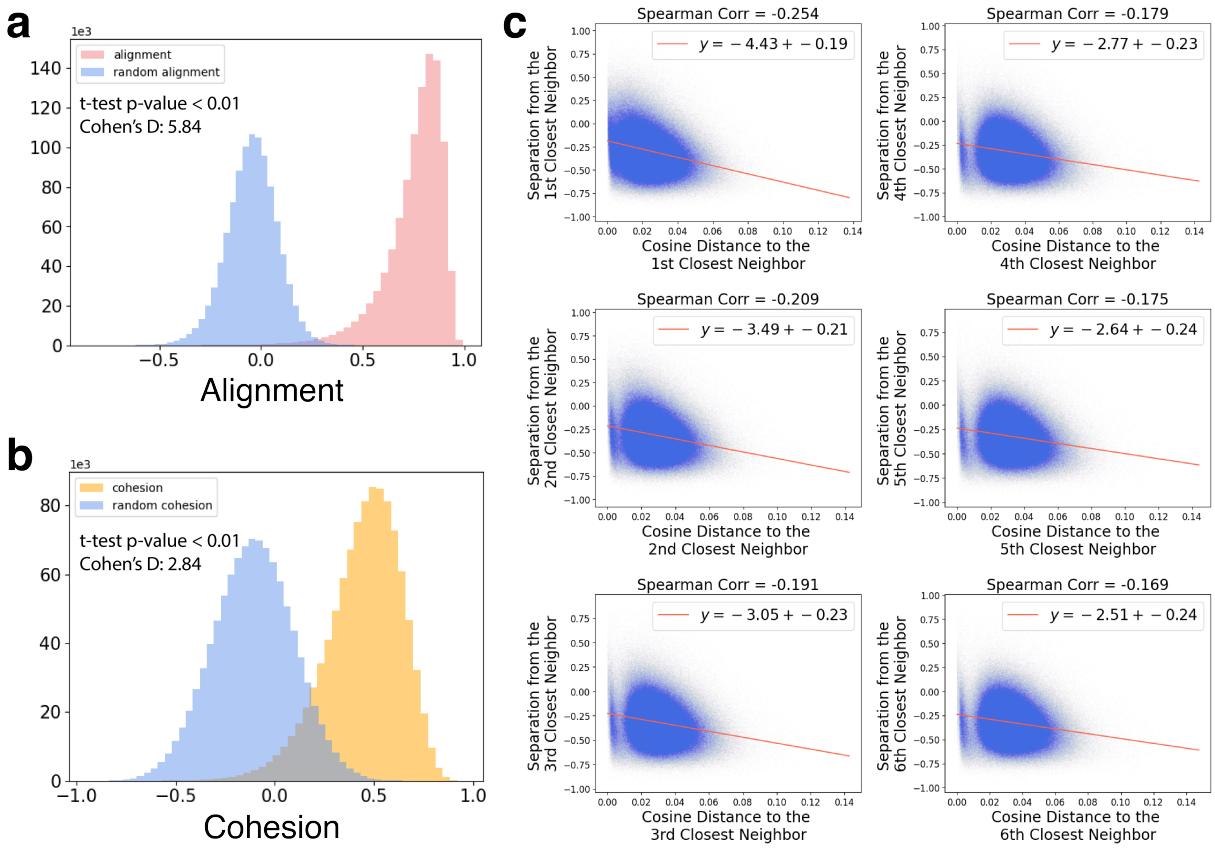

Figure 2: **Three patterns of topic shifts among scientists, resembling the boids model.** **(a)** The cosine similarity between the change in an author's velocity and the alignment vector is significantly higher than the similarity between the velocity change and an alignment vector calculated from a randomly selected author's movement, highlighting the influence of peer movement alignment. **(b)** The cohesion effect is captured by the cosine similarity between the author's positional change and the vector pointing from their current position to the centroid of their peers' positions. **(c)** The cosine distance between the focal author and their peers is negatively correlated with movement direction, suggesting that authors tend to move away from peers who are closer in topic space. This negative correlation between distance and separation impact diminishes with increasing neighbor rank—that is, the repulsive effect is strongest for the nearest peers and gradually weakens for more distant ones.

Analogous to the alignment case, we establish a random baseline by replacing $\bar{\mathbf{x}}_j^{t-1}$ with $\bar{\mathbf{x}}_m^{t-1}$ in the calculation of $\mathbf{C}_j^t$. Figure 2b shows that the resulting cohesion distribution differs markedly from the random baseline, with an average similarity of 0.453, compared to a random baseline of -0.109 (Cohen's $d = 2.8$; $p < 0.001$), suggesting authors often shift their position toward the "center" of their peer group.

Finally, we examined *separation*, the idea that authors may be inclined to move away from peers who are very close in topic space—potentially to reduce redundancy and establish distinct research identities. Detecting this behavior is inherently challenging, as it can be masked by opposing cohesive forces. Therefore, instead of calculating the separation propensity, we tested the hypothesis that the repulsive effect of separation should be strongest when authors are in close proximity.

To test this, for a focal author j and their closest neighbor i_1 in window t, we calculate the distance $\bar{\mathbf{x}}_j^t - \bar{\mathbf{x}}_{i_1}^t$, and compared its direction with the author's subsequent positional shift, $\bar{\mathbf{x}}_j^t - \bar{\mathbf{x}}_j^{t-1}$. This process is repeated for peers at increasing distances $(i_2, i_3, \dots)$.

As shown in Figure 2c, we find a clear negative correlation between peer proximity and the separation effect. This repulsive tendency is strongest for the nearest peers and gradually weakens as peers become more distant in topic space. These results suggest that authors may intentionally distance themselves from those with highly similar research trajectories—potentially to avoid direct competition and to establish a distinct, recognizable niche within the broader scientific landscape. However, it is important to note that this observed pattern may also be influenced by cohesion dynamics. As peers become more distant, the cohesive pull toward the broader center of the peer group may grow stronger, potentially offsetting or even dominating the separation effect. Therefore, further analysis is needed to disentangle the relative contributions of these competing forces—a task we leave for future research.

2.3 Peer's Prominence and Alignment, Cohesion impact

An author's tendency to follow the movements of their peers may depend not only on topical proximity but also on peer attributes such as academic prominence. In particular, when a highly influential peer shifts toward a new research direction, a focal author may be more likely to follow that path. Conversely, more prominent authors may be less susceptible to peer influence, instead charting more independent research trajectories.

To test the first hypothesis, that prominent peers exert greater influence, we compare the in-

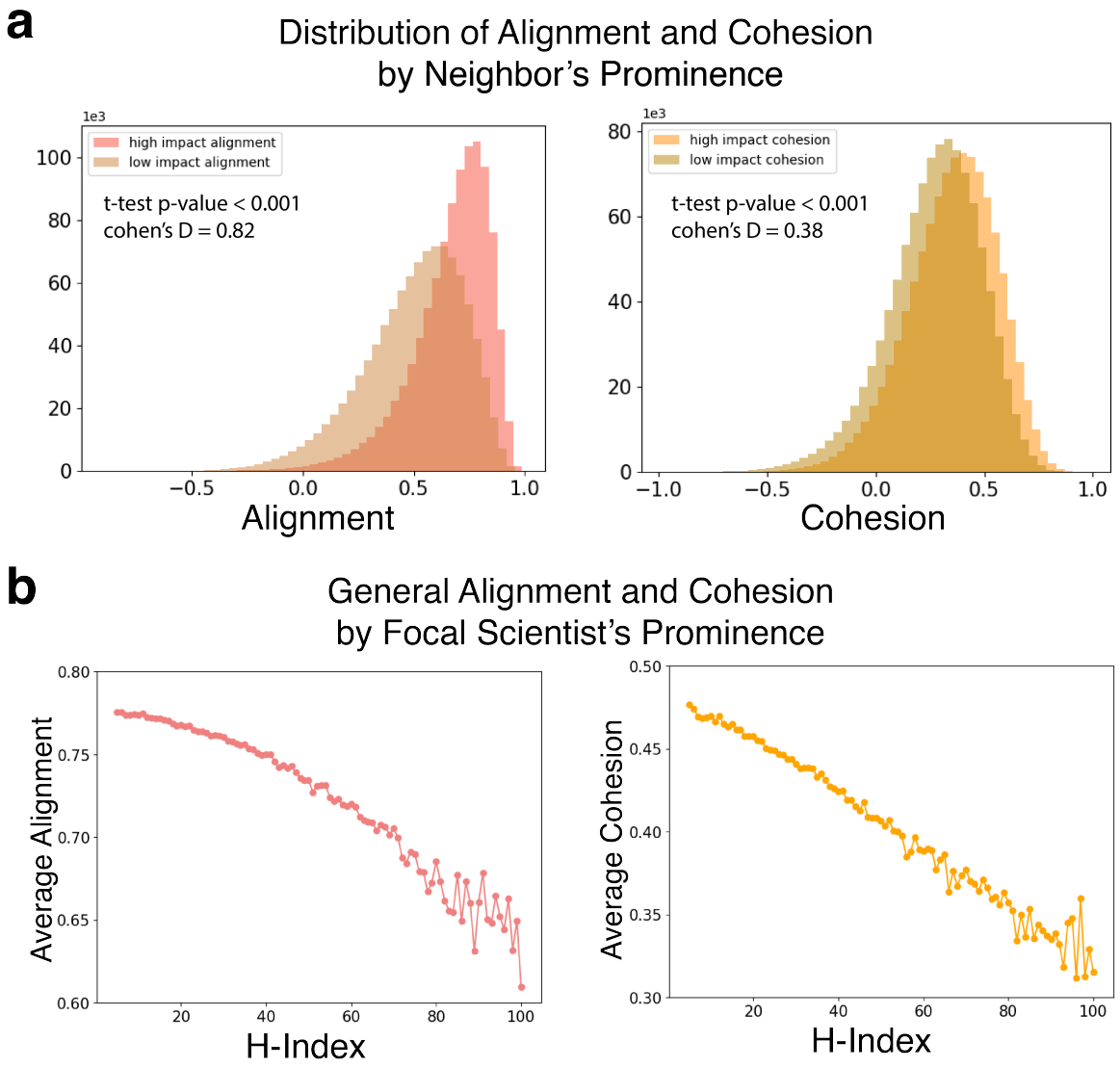

Figure 3: **Prominent Scientists Are More Likely to Be Followed by Their Peers, Yet Less Likely to Follow Them.** **(a)** The propensity for alignment the neighbor with the highest h-index is significantly greater than that toward the neighbor with the lowest h-index. Similarly, the cohesion propensity toward the highest h-index neighbor is significantly bigger than that observed for the lowest h-index. **(b)** Average alignment and cohesion scores as a function of the focal author's h-index. Both measures decrease with increasing author prominence, suggesting that more influential researchers are less likely to follow the movements of their peers.

fluence of a focal author's most prominent peer with the highest h-index against that of the least prominent peer with the lowest h-index within their local neighborhood.

For the alignment analysis, we modified Equation 2 by replacing the average peer velocity with the velocity of either the highest or lowest h-index peer. Similarly, for cohesion, we substituted the average peer position with the position of the highest or lowest h-index peer. We then computed the cosine similarity between the focal author's movement and these adjusted vectors.

As shown in Figure 3a, the cosine similarity between the alignment vector $\mathbf{A}^t$ and the change of the velocity of a focal author $\Delta\mathbf{x^t}$ is significantly higher when using the highest h-index peer than when using the lowest h-index peer (Cohen's $d = 0.82$; $p < 0.001$). This pattern also holds for cohesion (Cohen's $d = 0.38$; $p < 0.001$). These results suggest that academically prominent peers exert greater influence on the research trajectories of others.

To evaluate the second hypothesis, that a focal author's prominence reduces their susceptibility to peer influence, we examine how alignment and cohesion vary with the author's h-index. As shown in Figure 3b, both average alignment and cohesion scores decrease as the focal author's h-index increases. This pattern indicates that more prominent researchers are less influenced by their peers and tend to follow more independent research trajectories.

2.4 Prediction

After quantifying the alignment, cohesion, and separation propensities of scientists, we evaluated the model's predictive capability on an out-of-sample period by using data from the 2014-2018 window, which is a period excluded from the initial analysis.

Based on the inverse correlation between neighbor distance and separation impact, we examine two functional forms of this relationship: an inverse linear form and an inverse quadratic form. Specifically, the prediction is formulated as follows:

$$\begin{aligned}\mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \alpha^{\text{alignment}}\mathbf{A}_j^t + \alpha^{\text{cohesion}}\mathbf{C}_j^t \\ + \alpha^{\text{separation}}\mathbf{S}_j^t.\end{aligned} \tag{4}$$

Here, $\mathbf{A}_j^t$ and $\mathbf{C}_j^t$ are defined in Eq. 2 and 3, respectively. The term $\mathbf{S}_j^t$ captures the separation impact, which we previously showed to be negatively associated with the distance between the focal

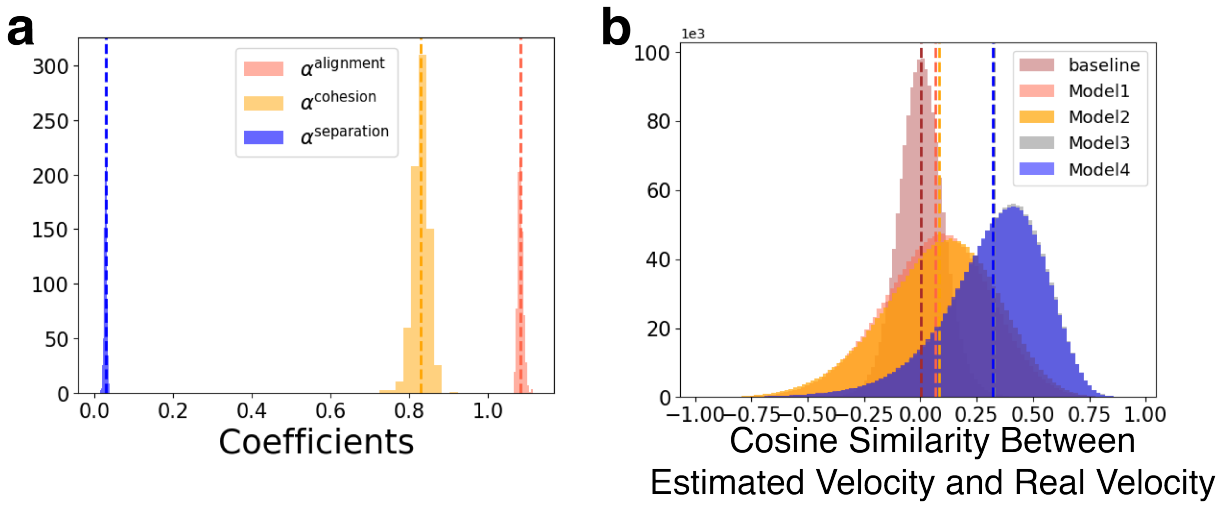

Figure 4: **Impact of Peer's Academic Prominence and Prediction of the Boids Model.** **(a)** The estimated coefficients across different dimensions, calculated from the linear regression of model 4, are highly consistent with small standard deviation ($\mu_{\alpha^{\text{alignment}}} = 1.08$, $\sigma_{\alpha^{\text{alignment}}} = 0.007$, $\mu_{\alpha^{\text{cohesion}}} = 0.830$, $\sigma_{\alpha^{\text{cohesion}}} = 0.020$, and $\mu_{\alpha^{\text{separation}}} = 0.030$, $\sigma_{\alpha^{\text{separation}}} = 0.004$). **(b)** Cosine similarity between the estimated velocity the generative models and the actual velocity indicates that Models 3 and 4 have the highest similarity, followed by Models 2 and 1, with all models performing significantly better than the baseline. These findings demonstrate that the combination of alignment and cohesion propensities provides a substantial level of explainability for how authors navigate the topic space.

author and their neighbors. While several functional forms can model this inverse relationship, we adopt a simple approach: an inverse-linear relationship between cosine distance and separation effect. Specifically, the separation term is defined as $\alpha^{\text{separation}} \left(\sum_{i \in \text{Neighbor}(j)} \frac{1}{(1+\text{cosdis}(x_j^t, x_i^t))} (\mathbf{x}_j^t - \mathbf{x}_i^t) \right)$, where $\text{cosdis}(x_j^t, x_j^t)$ indicates the cosine distance between x_j^t and x_i^t.

To estimate the coefficients $\alpha^{\text{alignment}}$, α^{cohesion}, and $\alpha^{\text{separation}}$, we employ linear regression. Since both the inputs and outputs are vectors, we perform the regression separately for each dimension. Note that our goal here is not to get the exact value of the estimated coefficient but rather to get the approximate magnitude and relative scale of the coefficients across the different behavioral forces.

Notably, the estimated coefficients were almost consistent across dimensions with low standard deviation. For example, the $\alpha^{\text{alignment}}$ values were centered around 1.08 , as were the α^{cohesion} values 0.83. while the $\alpha^{\text{separation}}$ values were approximately 0.0296, each exhibiting small standard deviations (Figure 4b).

To evaluate model performance, we compared four models:

$$
\begin{aligned}
\textbf{Model baseline:} \quad & \mathbf{v}_j^{t+1} = \mathbf{v}_m^t, \\
\textbf{Model 1:} \quad & \mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \alpha^{\text{alignment}} \, \mathbf{A}_j^t, \\
\textbf{Model 2:} \quad & \mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \alpha^{\text{cohesion}} \, \mathbf{C}_j^t, \\
\textbf{Model 3:} \quad & \mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \alpha^{\text{alignment}} \, \mathbf{A}_j^t + \alpha^{\text{cohesion}} \, \mathbf{C}_j^t, \\
\textbf{Model 4:} \quad & \mathbf{v}_j^{t+1} = \mathbf{v}_j^t + \alpha^{\text{alignment}} \, \mathbf{A}_j^t + \alpha^{\text{cohesion}} \, \mathbf{C}_j^t \\
& \qquad + \alpha^{\text{separation}} \left(\sum_{i \in \text{Neighbor}(j)} \frac{1}{1 + \text{cosdis}(\mathbf{x}_j^t, \mathbf{x}_i^t)} \frac{\mathbf{x}_j^t - \mathbf{x}_i^t}{\|\mathbf{x}_j^t - \mathbf{x}_i^t\|} \right).
\end{aligned}
$$

The $\mathbf{v}_m^t$ is the velocity of a random focal author. The predicted movements were then evaluated by computing the cosine similarity between the predicted and actual movements in the 2014–2018 window.

The results of this evaluation are presented in Figure 4c. Models 3 and 4 produced the best results, with average cosine similarities of 0.321 and 0.326, respectively. Models 1 and 2 also significantly outperformed the baseline, achieving average scores of 0.0811 and 0.0656. These findings highlight the strong predictive value of alignment and cohesion dynamics in modeling scientific movement.

3 Discussion

In this study, we present a novel framework for understanding how scientists shift their research focus through the dynamic influence of their peers. By applying neural text embeddings to represent research topics in a continuous space and drawing inspiration from the boids model of collective behavior, we identify three simple yet powerful dynamics—alignment, cohesion, and separation—that drive the evolution of individual research trajectories. Our generative model not only accounts for past movements but also predicts future shifts, underscoring the crucial role of peer influence in shaping the intellectual landscape of science.

This work makes several key contributions. First, we extend the analysis of peer influence beyond direct collaborators by incorporating all scientists engaged in related fields. Second, we capture the dynamic nature of peer interactions by tracking the evolving research interests of peers, moving beyond the static representations of collaborators' interests used in previous studies [3]. Third, it introduces the use of continuous neural text embeddings to represent research topics, enabling a more flexible and nuanced analysis of how individual research interests shift across time.

Despite these advances, our study has several limitations. First, our boids-like model focuses on its simplicity only based on peer interactions and largely omits exogenous drivers such as funding availability, institutional priorities, or broader societal events. In reality, the trajectories of scientists are shaped by a range of external influences, including grant programs [14], scientific prizes [15], industry demand [16], or global crises [17, 18]. Combining our approach with analyses of these external factors can increase the explainability. Second, the exact formulation of the separation component in our model remains an open question. Although earlier analysis revealed a negative relationship between peer proximity and separation behavior, the addition of the separation term in Model 4 yielded no predictive improvement. This suggests that our current formulation—an inverse-linear function of peer distance—may not adequately capture the complexity of separation dynamics. It is possible that the influence of separation is more subtle or non-linear, or that it interacts with other forces in ways not captured by our current model. Moreover, the neural text embedding model we used to represent each author's position inherently restricts multiple authors from occupying the same position in topic space unless their documents are identical, potentially suppressing meaningful separation behavior. In this sense, the current separation term might function more as noise than as a reliable signal. Finally, our predictive analysis relies on embedding vectors derived from data before 2014 to forecast outcomes for 2014–2018. However, these vectors may already incorporate

some information from the 2014–2018 period due to their inclusion in SPECTER's training process. Future research should use papers published in periods excluded from the SPECTER training dataset to enhance prediction accuracy. Overall, these limitations provide promising avenues for refining our understanding of the interplay between social influence and scientific innovation.

4 Methods

4.1 Dataset

4.1.1 Papers

We collected the article types papers published between 1999 and 2023 from OpenAlex. There are 184,234,031 number of papers in total. Among these, we further selected the articles with title where the length of the titles are bigger than 10 characters, resulting in xxx papers. Then we

for the embedding, which result in 52,692,352 papers.

4.1.2 Authors

We used an author-disambiguated dataset from OpenAlex. For robustness, we defined focal authors as those who published at least five papers in three consecutive time windows, and potential peer authors as those who published at least three papers in the first two consecutive windows. This filtering produced 1,024,539 focal authors for the 1999–2013 period and 1,909,038 potential peer authors for the 1999–2008 period. For the prediction task, we applied the same criteria to the 2004–2018 data and selected focal authors appearing in both the 1999–2013 and 2004–2018 periods, yielding 862,320 authors and 2,756,783 potential peer authors for the 2004–2013 window.

4.2 Identifying Peers

To identify the peers of each focal author, we first computed embedding vectors for each publication using the SPECTER model. We then determined the average position of each focal and potential peer author's publications within a five-year window by averaging their respective publication embeddings. For each time window, we identified the 15 nearest potential peer authors—those who had published at least three papers during the period—whose average positions in the embedding space were closest to that of the focal author, who published at least five papers in the time window.

Here, we assume that individuals tend to interact with a limited number of nearest neighbors rather than with all others within a fixed metric radius, which also aligns with the studies of the empirical observation of flocking birds [19]. Given the computational intensity of finding nearest neighbors in a 768-dimensional embedding space, we employed Facebook's FAISS library [20], which is optimized for efficient similarity search and clustering of dense vectors .

4.3 Code Availability Statement

The code to reproduce the results is available at https://github.com/MunjungKim/Boids_Scientists.

References

[1] Tao Jia, Dashun Wang, and Boleslaw K. Szymanski. Quantifying patterns of research-interest evolution. *Nature Human Behaviour*, 1(4):0078, March 2017.

[2] An Zeng, Ying Fan, Zengru Di, Yougui Wang, and Shlomo Havlin. Impactful scientists have higher tendency to involve collaborators in new topics. *Proceedings of the National Academy of Sciences*, 119(33):e2207436119, 2022.

[3] Sara Venturini, Satyaki Sikdar, Francesco Rinaldi, Francesco Tudisco, and Santo Fortunato. Collaboration and topic switches in science. *Scientific Reports*, 14(1):1258, January 2024.

[4] Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015.

[5] Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso. Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, 42(3):527–554, 2011.

[6] Erin Leahey. Not by productivity alone: How visibility and specialization contribute to academic earnings. *American Sociological Review*, 72:533 – 561, 2007.

[7] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[8] An Zeng, Zhesi Shen, Jianlin Zhou, Ying Fan, Zengru Di, Yougui Wang, H. Eugene Stanley, and Shlomo Havlin. Increasing trend of scientists to switch between topics. *Nature Communications*, 10(1):3439, July 2019.

[9] Andrey Rzhetsky, Jacob G. Foster, Ian T. Foster, and James A. Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574, 2015.

[10] Enrico Berkes, Monica Marion, Staša Milojević, and Bruce A. Weinberg. Slow convergence: Career impediments to interdisciplinary biomedical research. *Proceedings of the National Academy of Sciences*, 121(32):e2402646121, 2024.

[11] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022.

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online, July 2020. Association for Computational Linguistics.

[13] Craig W. Reynolds. Flocks, herds, and schools: a distributed behavioral model. *Seminal graphics: pioneering efforts that shaped the field*, 1987.

[14] Kyle Myers. The elasticity of science. *American Economic Journal: Applied Economics*, 12(4):103–34, October 2020.

[15] Ching Jin, Yifang Ma, and Brian Uzzi. Scientific prizes and the extraordinary growth of scientific topics. *Nature Communications*, 12(1):5619, October 2021.

[16] David Mowery and Nathan Rosenberg. The influence of market demand upon innovation: a critical review of some recent empirical studies. *Research Policy*, 8(2):102–153, 1979.

[17] Shir Aviv-Reuven and Ariel Rosenfeld. Publication patterns' changes due to the COVID-19 pandemic: a longitudinal and short-term scientometric analysis. *Scientometrics*, 126(8):6761–6784, August 2021.

[18] Holly Else. How a torrent of covid science changed research publishing - in seven charts. *Nature*, 588:553, 2020. News feature, published 16 December 2020, clarification 17 December 2020.

[19] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1232–1237, 2008.

[20] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025.

Acknowledgements

We thank Dr. James Evans, Dr. Filippo Menczer, Dr. Haewoon Kwak, Dr. Nadav Kunievsky, and Dr. Rachith Aiyappa for their helpful feedback.

Author Contributions

Competing Interests

The authors declare no competing interests.

Supplementary Information

Flocking Scientists in Research Trajectories

Munjung Kim, Yong-Yeol Ahn

A Excluding Collaborators

Since collaborators share the same publications, having many collaborators could influence both alignment and cohesion effects. To test whether our findings are not driven solely by collaboration ties, we repeated the analysis while excluding collaborators from the peer set. The results remained consistent even when collaborators were removed.

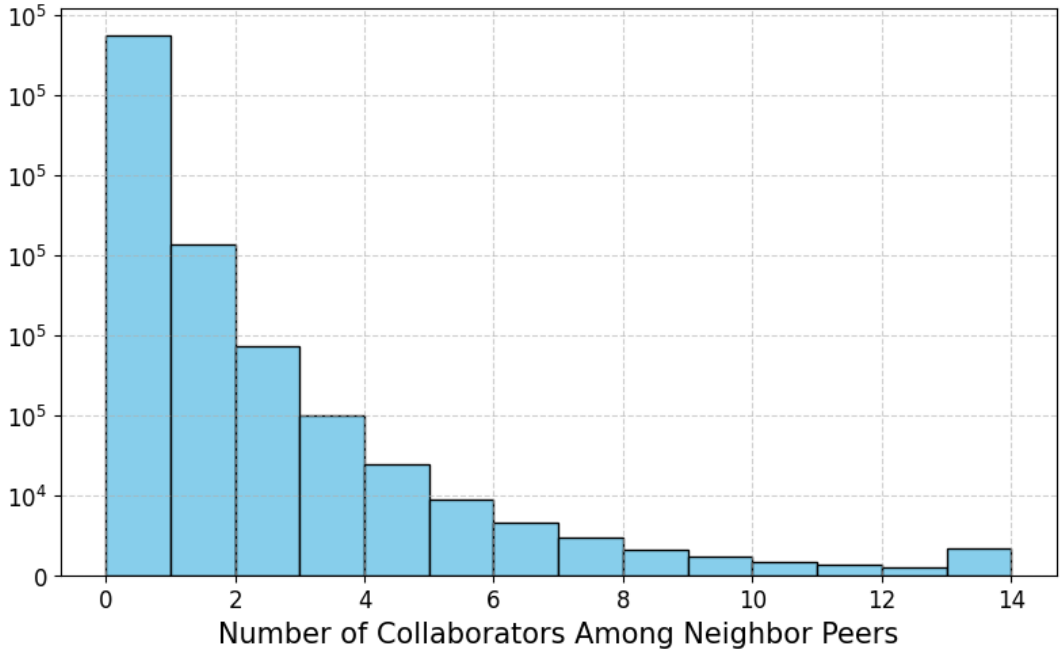

Figure 5: The number of collaborators found within each focal author's 14 nearest neighbors

Figure 5 shows the distribution of the number of focal authors based on different number of collaborators present in their 14 nearest peers. Notably, about one-third of the focal authors had no collaborators among their 14 closest peers.

After excluding the collaborators in the nearest peer set, we re-calculated the alignment, cohesion, and separation propensities using this revised peer set. Hence, the authors whose entire peer group consisted of collaborators are excluded in the analysis. Figure 6 presents the results. Consistent with the findings reported in the main text (where collaborators were included), both

the alignment and cohesion scores—measured via cosine similarity—remained significantly higher than the baseline, reinforcing the robustness of our results.

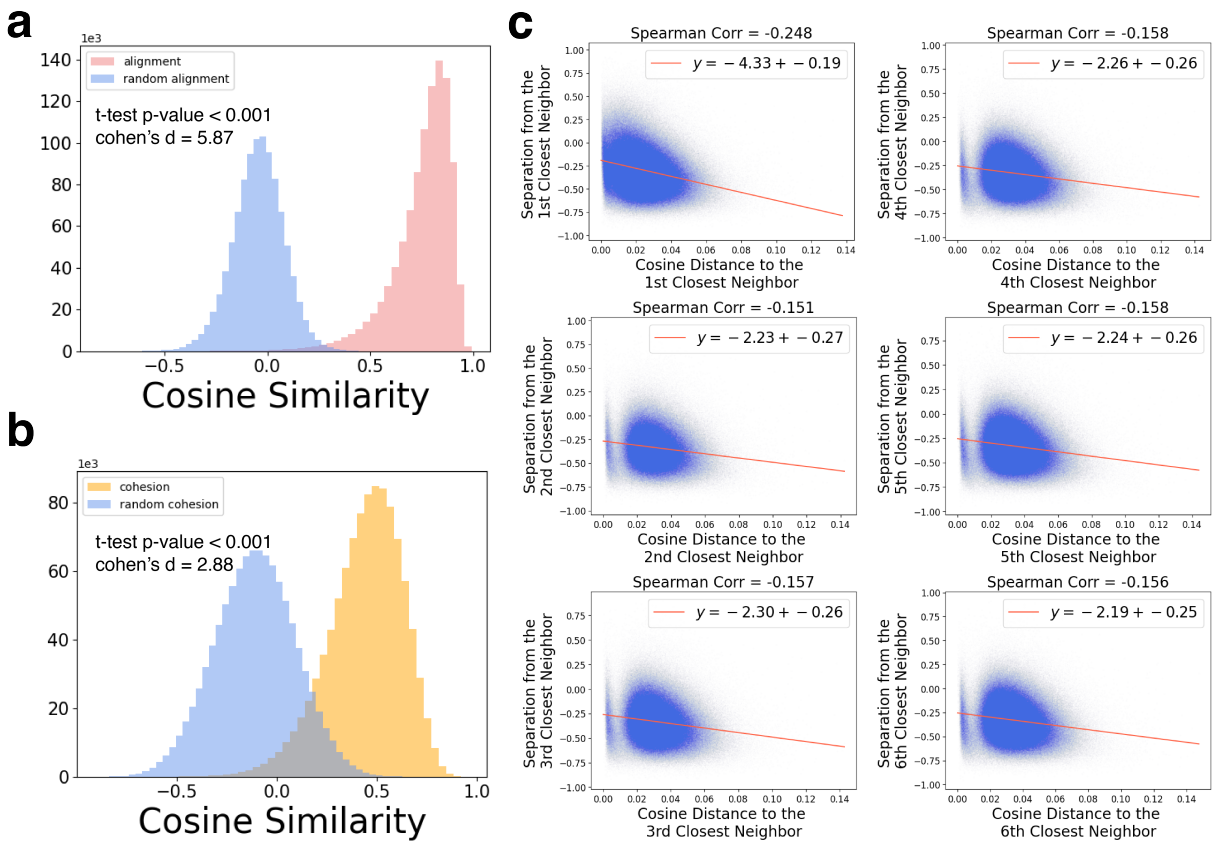

Figure 6: The three patterns of topic shifts among scientists persist even when collaborators are excluded from the set of peers.

B Different Number of Neighbors

We test the robustness of our findings by varying the number of neighbors considered as peers. Specifically, we increase the number of nearest neighbors in the embedding space from 15 (as used in the main analysis) to 30. The results remain consistent, confirming the stability of our findings. This robustness check is presented in Figure 7.

C Field Impact

A potential concern is that the observed alignment and cohesion effects merely reflect broader field-level trends—namely, that researchers within the same discipline tend to move in similar

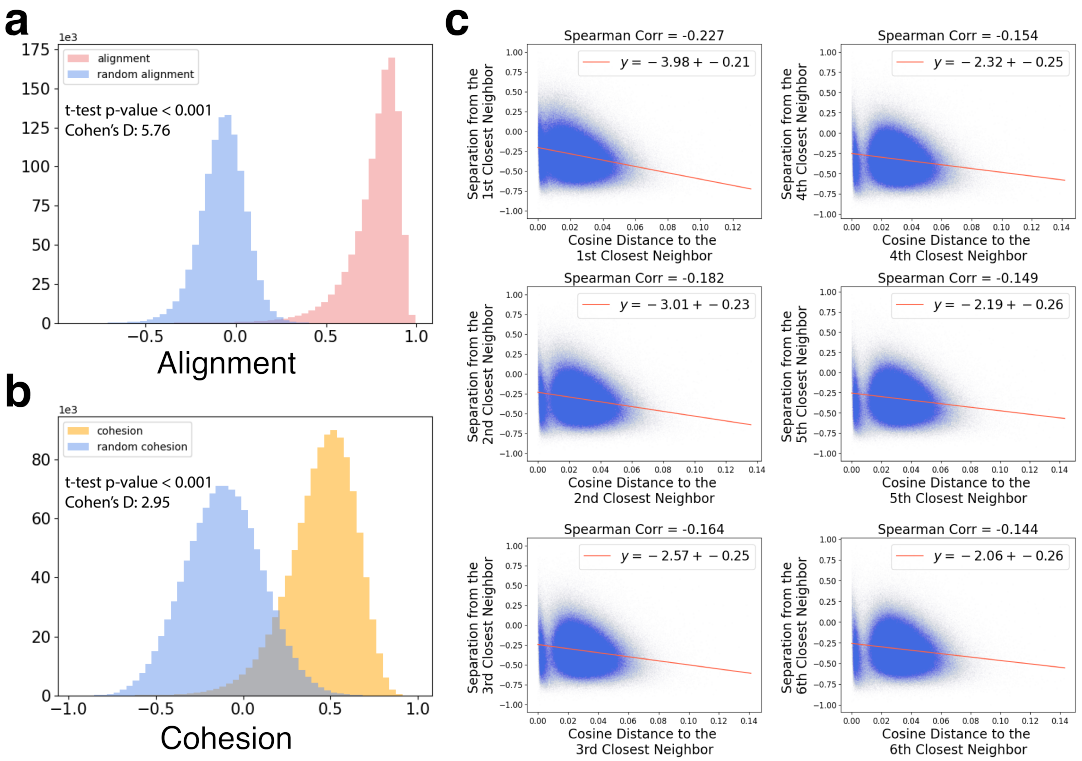

Figure 7: The three patterns of topic shifts among scientists persist even when the number of neighbors is increased to 30.

topical directions over time. However, our operationalization of alignment specifically addresses this possibility. Rather than measuring absolute similarity in research direction, alignment is defined as the cosine similarity between the focal author's velocity change at time $t+1$ and the difference in velocity between the focal author and their nearest neighbors at time t. This formulation captures the extent to which the author *adjusts* their trajectory in response to the relative motion of peers, rather than following a shared disciplinary drift.

To empirically distinguish localized peer influence from general field-level convergence, we introduce a field-matched random baseline. In this baseline, we compute alignment as the cosine similarity between the alignment vector of a randomly selected author from the same field and the actual velocity change of the focal author. Similarly, the cohesion baseline is calculated as the cosine similarity between the cohesion vector of a random author from the same field and the focal author's observed position change. By comparing these baseline scores with those derived from actual neighbors, we assess whether alignment and cohesion impact arise from specific peer dynamics rather than generalized movement within a field.

The results show that alignment and cohesion scores are significantly higher in the original model

than in the field-level random baseline. This finding suggests that the observed effects cannot be explained solely by broader topical trends within a field, but rather reflect localized, peer-specific influences on the evolution of individual research trajectories.

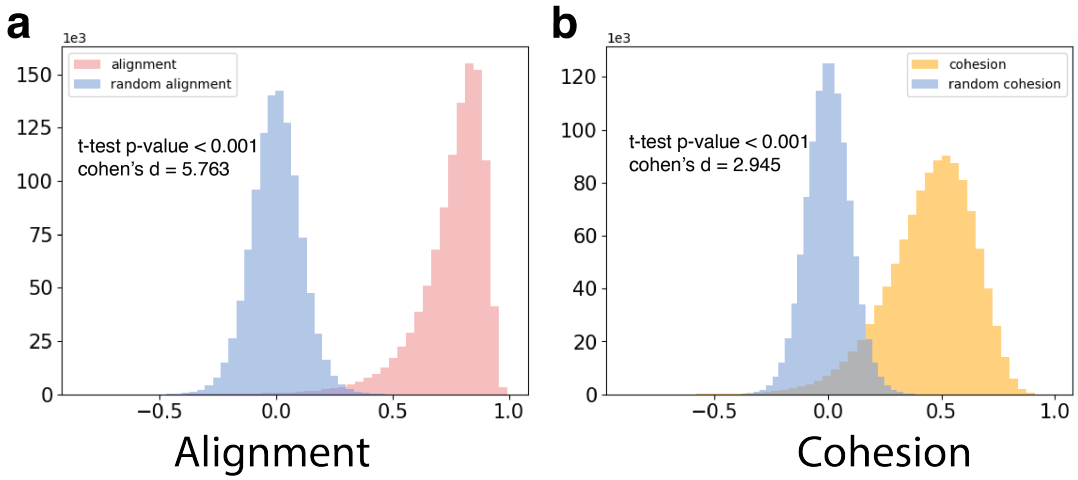

Figure 8: Alignment and cohesion scores remain significantly higher than those in the random baseline, where authors are randomly selected from the same field. This suggests that alignment and cohesion effects are not solely driven by field-level topical similarity, but also reflect meaningful peer influence.

D Author Map

We visually assess whether the average embedding vectors reflect the authors' primary research interests. Specifically, we plot the UMAP projection of the average embedding vectors for the 2004–2008 period, coloring each point by the author's major field—defined as the most frequent field in which they published during that time. The resulting visualization is shown in Figure 9. Authors associated with the same field tend to appear near one another in the projection, suggesting that the average embedding vectors meaningfully capture patterns aligned with disciplinary focus.

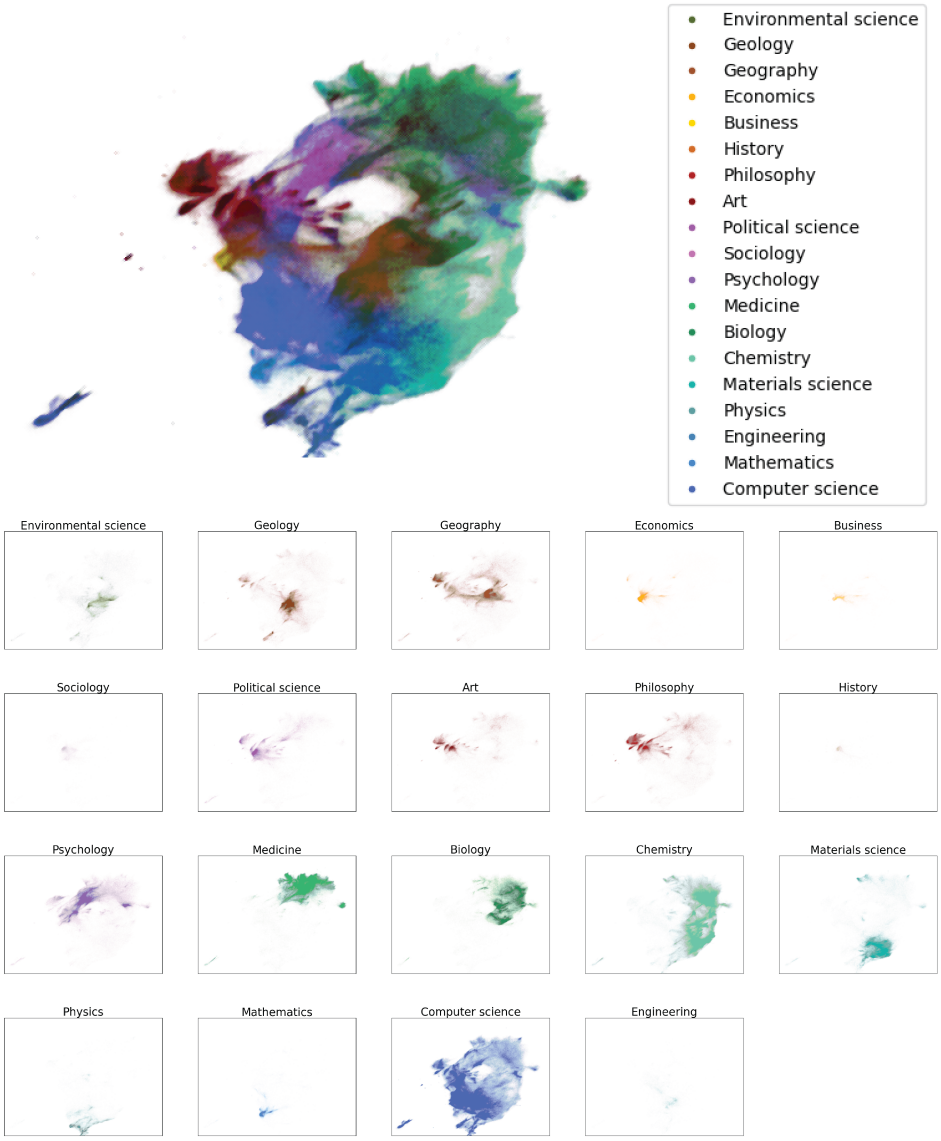

Figure 9: UMAP projection of average embedding vectors for authors during the 2004–2008 period. Each point represents an author and is colored by their major field of study, defined as the most frequent field in which they published during this time. The lower panels show the position of the authors from different fields in the same region of the embedding space.