Check for updates

Received: 10 November 2025 | Revised: 26 January 2026 | Accepted: 6 February 2026

DOI: 10.1002/asi.70067

RESEARCH ARTICLE

JASIST WILEY

A semantic atlas of journals: Structure, position, and dispersion

Alex J. Yang[1,2] | **Richard B. Freeman**[3,4] | **Sanhong Deng**[1,2]

[1]School of Information Management, Nanjing University, Nanjing, China

[2]International Joint Informatics Laboratory, Nanjing University, Nanjing, China

[3]Center for Labor and a Just Economy, Harvard University, Cambridge, Massachusetts, USA

[4]Department of Economics College of Economics, Harvard University, Cambridge, Massachusetts, USA

Correspondence
Sanhong Deng, School of Information Management, Nanjing University, Nanjing 210023, China.
Email: sanhong@nju.edu.cn

Richard B. Freeman, Center for Labor and a Just Economy, Harvard University, Cambridge, MA, USA; Department of Economics College of Economics, Harvard University, Cambridge, MA, USA.
Email: freeman@nber.org

Funding information
National Natural Science Foundation of China, Grant/Award Number: 724B2015

Abstract

We present semantic atlases for journals in the life sciences and social sciences by encoding over 30 million abstracts with domain-adapted transformer models, aggregating paper embeddings into journal centroids, and applying UMAP for visualization. We validate the resulting 2D manifolds through strong correlations between positional proximity and external measures of topical and authorship overlap. Semantic axes reveal interpretable gradients along biologically meaningful dimensions and a latent quality axis that recovers prestige hierarchies. The approach also positions journals with elevated retraction rates peripherally, highlighting latent signals of publication integrity. We introduce effective radius as a measure of geometric properties; larger radii are associated with higher impact and citations. These findings demonstrate that transformer embeddings provide a continuous, citation-independent framework for decoding journal roles, thematic specialization, prestige, and knowledge dispersion, offering new quantitative tools for the science of science and dynamic monitoring of scholarly ecosystems.

1 | INTRODUCTION

The rapid growth of scientific literature has created both unprecedented opportunities and significant challenges in navigating, evaluating, and understanding the structure of knowledge. With millions of papers published annually across diverse domains, traditional bibliometric tools—primarily citation networks, co-citation clusters, and keyword-based maps—have provided valuable insights into disciplinary boundaries, intellectual communities, and knowledge flows (Börner et al., 2003; Klavans & Boyack, 2009; Small, 1999; van Eck & Waltman, 2010). However, these approaches often overlook the rich semantic content embedded in the text of publications themselves, relying instead on indirect proxies such as citation linkages or term co-occurrence. Recent advances in transformer-based language models offer a powerful alternative: by encoding abstracts or full texts into dense, high-dimensional vectors that capture contextual meaning, these models enable the construction of continuous, geometry-based representations of scientific knowledge (Devlin et al., 2019; Mikolov et al., 2013).

Embedding-based methods have begun to reveal latent structures in scientific corpora, including conceptual breadth (Hao et al., 2026; Lee et al., 2025), novelty and distance metrics (Hofstra et al., 2020; Kelly et al., 2021; Scharfmann et al., 2025), semantic axes for thematic continua (Aceves & Evans, 2024b; An et al., 2018), and

 © 2026 Association for Information Science and Technology.

predictive discovery pathways (Shi & Evans, 2023; Sourati & Evans, 2023; Tshitoyan et al., 2019). In the science of science, such techniques complement citation-based maps by providing text-intrinsic views of knowledge organization, enabling fine-grained analysis of journal roles, interdisciplinary gradients, and evolving frontiers (Xiang et al., 2025; Zhang et al., 2025). Despite these advances, systematic large-scale mapping of journals—the primary units of scientific dissemination and evaluation—using modern transformer embeddings remains underexplored. Journals serve as stable, gatekeeping entities that shape research trajectories through editorial priorities, peer review, and thematic focus (Baldwin, 2019; Fersht, 2009; Garfield, 2006). Characterizing their positions, breadths, and alignments in semantic space could offer new quantitative tools for understanding prestige, specialization, integrity, and impact.

This study addresses these gaps by constructing semantic atlases of journals in two broad domains: the life sciences (using PubMed data) and the social sciences (using OpenAlex filtered to SSCI/AHCI-indexed journals). We encode abstracts with domain-adapted transformer models, aggregate paper embeddings into journal centroids, reduce dimensionality with UMAP for visualization, and derive interpretable axes for thematic and quality gradients. We further introduce the effective radius metric to quantify topical breadth and examine its associations with bibliometric outcomes. Additional analyses leverage retraction data to position potentially predatory journals and validate map integrity through correlations with external topical/author overlaps.

Our contributions are threefold. First, we provide scalable, text-driven atlases that reveal modular, hierarchical structures in journal landscapes, validated against topic and authorship networks. Second, we demonstrate the interpretive power of semantic axes to decode biologically and socially meaningful continua and recover prestige gradients independent of citations. Third, we link geometric properties to domain-specific patterns in impact, collaboration, and innovation while highlighting applications for monitoring publication integrity. These findings advance quantitative methods in the science of science, offering a complementary lens to citation-based approaches and enabling dynamic, citation-agnostic evaluation of journals and emerging knowledge structures.

2 | BACKGROUND

2.1 | Conceptual dimensions derived from embeddings

Modern embedding models map words, phrases, or documents to high-dimensional vectors in which geometric relationships capture semantic relations: closer vectors typically correspond to more semantically similar concepts (Devlin et al., 2019; Mikolov et al., 2013). In higher-dimensional neural network embeddings, semantic relations manifest as geometric patterns, enabling analogical reasoning and dimensional analysis (Kozlowski et al., 2019).

Embeddings support several key analytical constructs in conceptual analysis. First, they quantify conceptual breadth or dispersion. For instance, embeddings have been used to measure discursive breadth in texts (Lix et al., 2022), to characterize the scope of conceptual spaces across languages and domains (Aceves & Evans, 2024a), and to assess individual belief dispersion via average vectors and an effective radius metric (analogous to the radius of gyration) derived from BERT embeddings and dimensionality reduction (Lee et al., 2025). In scientific literature, similar approaches have computed paper centroids and Euclidean distances from them to quantify the extent of knowledge coverage (Hao et al., 2026).

Second, embeddings enable measures of conceptual distance and novelty. Aggregated distances have revealed biases, such as relative norm differences between group vectors and neutral terms to quantify long-term gender and ethnic stereotypes (Garg et al., 2018). In innovation contexts, embedding-based distances have assessed novelty in papers and patents (Hofstra et al., 2020; Kelly et al., 2021; Scharfmann et al., 2025).

Third, embeddings enable tracing similarity to archetypal or theoretically constructed documents, yielding measures of "X-ness" along semantic axes within cultural or knowledge continua (Aceves & Evans, 2024b; An et al., 2018). Examples include validated word sets for agency (Lawson et al., 2022), vector arithmetic for analogical reasoning (Kozlowski et al., 2019), axes to capture geographic distance of institutions (Murray et al., 2023), and aggregate differences between antonym pairs to extract directional axes for binary oppositions (Taylor & Stoltz, 2021).

Fourth, embeddings facilitate inference of latent knowledge structures from textual corpora. Much organizational and scientific knowledge is encoded symbolically in language, rendering it accessible through large-scale text data and representable within the conceptual spaces of embedding models. In materials science, such models have predicted future discoveries decades in advance (Shi & Evans, 2023; Sourati & Evans, 2023; Tshitoyan et al., 2019). Explicit geometric representation of conceptual spaces thus serves dual roles: as a "telescope" for visualizing macro-scale knowledge structures and as a "microscope" for examining the fine-grained composition of meaning that underlies larger systems (Sourati & Evans, 2023).

Within the science of science, complementary approaches have long sought to map knowledge structures. Lexical

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

diversity in article titles has been used to quantify the cognitive extent of disciplinary domains across millions of publications (Milojević, 2015). Tree-structured ontologies, such as Medical Subject Headings (MeSH), have supported semantic positioning, as in the Triangle of Biomedicine, which places publications on a trilinear graph with vertices representing animal, cellular/molecular, and human research foci; translational progress is operationalized as movement toward the human vertex (Hutchins et al., 2019; Weber, 2013). More recently, embedding-based methods have advanced dynamic inference of knowledge pathways: Xiang et al. (2025) developed a navigation model that reconstructs scholarly trajectories across millions of researchers, while Zhang et al. (2025) used document embeddings to trace citation diffusion trajectories within an embedded disciplinary landscape. These efforts underscore the growing convergence of citation-based and text-intrinsic representations for understanding the latent geometry of scientific knowledge.

2.2 | Mapping journals in science

While paper-level data and metrics have become increasingly central to modern research evaluation, journals continue to play a foundational and enduring role in the organization, dissemination, and evaluation of scientific knowledge. Since the formalization of modern science, academic journals and proceedings have served as the primary venues for scientific activity (Baldwin, 2019; Fersht, 2009; Garfield, 2006). They function as gatekeepers of quality through peer review, delineate disciplinary boundaries and subfields, and provide stable, persistent units for aggregating, archiving, and contextualizing research output (Baldwin, 2019). In bibliometrics and the science of science, journals remain critical units of analysis because they shape citation patterns, influence impact assessment frameworks, affect career trajectories, and guide the evolution of fields through editorial policies and thematic priorities (Hicks et al., 2015).

Early visions of science as a mappable and navigable enterprise, such as Bush (1945)'s proposal for linking mechanisms to address specialization and volume, and Merton (1973)'s emphasis on cumulative, norm-governed communal progress, were closely tied to the journal as an institutional anchor. Although the rise of open-access platforms, preprint servers, and paper-level analytics has diversified the landscape, journals still serve as coherent aggregators of related work, editorial signals of quality and scope, and reference points for disciplinary identity—making them a valuable lens for understanding macro-scale semantic structure and knowledge organization.

Traditional approaches to mapping the structure of science have relied predominantly on citation networks and keyword or term co-occurrence (Börner et al., 2003; Klavans & Boyack, 2009). Small (1999) pioneered the use of co-citation clusters to visualize scientific fields and identify intellectual communities. Boyack et al. (2005) constructed comprehensive maps of science using co-citation and inter-citation linkages, selecting representations based on local and structural accuracy to delineate the "backbone" of scientific domains. Rosvall and Bergstrom (2008) applied random-walk-based community detection to citation networks, decomposing them into modules that simplify and highlight regularities in journal relationships. Bollen et al. (2009) introduced a clickstream model—a first-order Markov chain derived from journal navigation patterns—to visualize interconnections among scientific domains. Tools such as VOSviewer (van Eck & Waltman, 2010) have enabled widespread construction of bibliometric maps using co-citation, bibliographic coupling, and term co-occurrence to reveal journal similarities and clusters. Boyack and Klavans (2011) decomposed journal specificity into multiple dimensions, including textual coherence, radius of influence, community leadership, and concentration, offering a multidimensional view of journal roles. Börner et al. (2012) described the methodology of the UCSD Map of Science, deriving journal similarity matrices from cited references and keywords. Boyack and Klavans (2020) compared clustering outcomes from direct citation, extended direct citation, textual relatedness, and several citation–text hybrids, evaluating accuracy using reference-based, text-based, and grant-linked measures. They found that a balanced hybrid produced the most accurate clusters overall, while also demonstrating that textual and citation logics generate substantially different cluster contents—suggesting the two are complementary rather than redundant.

More recent developments have incorporated latent representations. Citation-based embeddings have positioned journals in low-dimensional latent spaces (Peng et al., 2021). Although embedding models have powerfully represented entities at the word, document, and belief levels, relatively few studies have systematically applied them to characterize journals at scale. Recent advances in embeddings have begun to reveal diachronic changes and hidden interdisciplinary patterns in classification schemes.

This study partly builds on the pioneering work of González-Márquez et al. (2024), who constructed the first large-scale semantic atlas of biomedical papers using embeddings and demonstrated that high-dimensional representations can reveal interpretable clusters, gradients, and anomalies at the document level. The present work extends the embedding-based atlas paradigm to the journal level while incorporating additional geometric analyses and expanding coverage to the social sciences.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We propose that journals can be meaningfully characterized within high-dimensional semantic spaces derived from transformer-based abstract embeddings. Our analyses focus on five key geometric properties: (1) position—the location of a journal's centroid in multidimensional meaning space; (2) breadth—the dispersion of constituent paper embeddings around this centroid; (3) integrity—alignment of output with disciplinary norms; (4) alignment along axes—projection onto theoretically or empirically derived continua; and (5) relations to impact and innovation—associations between these properties and bibliometric outcomes.

3 | METHOD

3.1 | Paper dataset

In this study, we examined semantic structures across journals in two broad disciplinary domains: the life sciences and the social sciences (including humanities). For each domain, we drew on large-scale publication datasets that collectively cover the vast majority of journals with substantial output.

1. Life sciences dataset

We constructed the life sciences dataset from PubMed, maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health. PubMed provides access to the MEDLINE database, indexing references and abstracts in the life sciences and biomedical research. We extracted metadata for 36,555,430 publications, including PMID, title, abstract, language, journal title, ISSN, publication date, and author names. To ensure data quality and suitability for semantic embedding, we excluded non-English publications (approximately 5 million) and those lacking abstracts (approximately 11 million). This filtering yielded 23,389,083 publications.

We then integrated these records with the PubMed Knowledge Graph 2.0 (PKG 2.0) (Xu et al., 2025), which includes 36,497,767 publications, 158,953,430 disambiguated authors, MeSH, citation linkages, and journal metadata. PKG 2.0 assigns 21,382 journal IDs based on ISSN; we merged duplicate entries resulting from distinctions between print and electronic formats. To ensure robust representation, we retained only journals with at least 100 associated publications in our filtered dataset, resulting in 7279 distinct journals.

2. Social sciences dataset

We derived the social sciences (including humanities) dataset from OpenAlex (as of the end of 2024, downloaded in February 2025), a comprehensive bibliographic database containing nearly 250 million scholarly works and approximately 255,000 sources. We considered only publications assigned to journal sources. The OpenAlex dataset incorporates a well-designed author name disambiguation mechanism, which uses a machine learning model to predict the likelihood that two authors are the same on the basis of features such as their institutions, co-authors, and citations, and then applies a custom, ORCID-anchored clustering process to group their works, assigning a unique ID to each author.

We matched OpenAlex journal sources to the 2024 Journal Citation Reports (JCR) editions, restricting inclusion to journals indexed in the Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (AHCI). We extracted publications from these journals, yielding 11,777,335 records across 4882 journals. We further filtered for publications with complete titles and abstracts, retaining 6,832,446 records. To focus on journals with sufficient coverage, we excluded those with fewer than 100 publications in the filtered set, leaving 4210 distinct journals.

The combined datasets span publications from 1808 to 2023 (life sciences) and from 1834 to 2024 (social sciences and humanities). In our primary analyses, we processed all publications jointly without stratification by publication year to generate embeddings using transformer-based language models. We provide time-stratified semantic atlases as supplementary material.

3.2 | Paper embeddings

Following approaches in recent studies of scientific literature landscapes (González-Márquez et al., 2024; Hao et al., 2026), we generated dense vector representations of papers using encoding-only large language models to produce 768-dimensional embeddings.

For the life sciences dataset, we employed PubMedBERT (Gu et al., 2021) to encode abstracts. For the social sciences dataset, we used SPECTER2 (Singh et al., 2023), a transformer-based model pretrained on a broad scientific corpus and further optimized with citation-informed objectives to project abstracts into a 768-dimensional embedding space.

We processed each abstract individually. Let $A = \{a_1, a_2, ..., a_N\}$ denote the set of abstracts, where $N = 23{,}389{,}083$ for the life sciences and $N = 6{,}832{,}446$ for the social sciences. Each abstract a_i was tokenized into a sequence of tokens $T_i = \{t_{i,1}, t_{i,2}, ..., t_{i,m}\}$. The model processed T_i to produce a sequence of 768-dimensional contextual token embeddings. This yielded a single vector $\mathbf{e}_i \in \mathbb{R}^{768}$ for each abstract a_i.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The full collection of embeddings formed a dense matrix $E \in \mathbb{R}^{N\times 768}$, with each row corresponding to one abstract. In this high-dimensional space, Euclidean proximity between vectors reflects semantic similarity between the underlying abstracts, providing a continuous, data-driven map of the research literature in each domain.

3.3 | Decoding journal-level knowledge space

To characterize the knowledge space at the journal level, we aggregated the embeddings of individual papers to create journal-level representations. Given the token limit of transformer-based models, processing entire journals directly is infeasible. Instead, we computed the average embedding for each journal based on its constituent papers. For a journal j with N_j papers, the journal embedding is defined as:

$$\boldsymbol{e}_j = \sum_{i=1}^{N_j} \mathbf{b}_i^j / N_j,$$

where $\mathbf{b}_i^j$ denotes the ith article embedding of journal j, and N_j is the total number of papers associated with journal j. This allows us to measure how closely journals are positioned in the knowledge space.

To visualize and analyze the journal-level knowledge space, we applied Uniform Manifold Approximation and Projection (UMAP) (Healy & McInnes, 2024) to reduce the dimensionality of the journal embeddings from 768 dimensions to a 2D manifold. UMAP constructs a high-dimensional, fuzzy topological representation of the data using a nearest-neighbor graph, followed by optimization of a low-dimensional layout that minimizes cross-entropy between the high- and low-dimensional representations. This process preserves both local neighborhood relationships and global data structure, making it well-suited for uncovering meaningful patterns in complex datasets (Becht et al., 2019). UMAP's computational complexity is approximately $O(N^2)$, where $N = 7279$ for the life sciences and $N = 4882$ for the social sciences is the number of journals. The resulting 2D maps provide a clear view of the journal-level knowledge space. However, we interpret small differences between points with care, as UMAP involves some random variation in how it calculates the final layout.

3.4 | Semantic axes for interpreting multidimensional structures

Semantic axes (An et al., 2018; Kozlowski et al., 2019; Lee et al., 2025; Murray et al., 2023) offer a powerful framework for extracting latent dimensions from embedding spaces by defining opposing "poles" and capturing the semantic relationships along these axes through vector arithmetic. In natural language processing, these poles might represent contrasts such as "good" versus "bad," "surprising" versus "unsurprising," or "masculine" versus "feminine." Analogous to scientific polarities—such as activator-repressor dynamics in gene regulatory networks or anterior–posterior axes in embryonic development—semantic axes enable the operationalization of conceptual relationships within the knowledge space. By projecting embeddings onto these axes, we can quantify how journals or individual papers align with specific biomedical themes, revealing multidimensional structures that mirror the hierarchical and interconnected nature of scientific domains, much like phylogenetic trees or signaling pathways in science.

Formally, let $S^+ = \{\boldsymbol{e}_1^+, \boldsymbol{e}_2^+ \cdots \boldsymbol{e}_n^+\}$ and $S^- = \{\boldsymbol{e}_1^-, \boldsymbol{e}_2^- \cdots \boldsymbol{e}_n^-\}$ be the set of two pole vectors respectively. Then, the average vectors of each set can be calculated as $\boldsymbol{E}^+ = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{e}_i^+$ and $\boldsymbol{E}^- = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{e}_i^-$.

From these average vectors of each set of poles, the semantic axis is defined as $\boldsymbol{E}_{\text{axis}} = \boldsymbol{E}^+ - \boldsymbol{E}^-$. Then, a score of journal is calculated as the cosine similarity of the journal's vector with the axis

$$\text{score} = \frac{\boldsymbol{e}_j \cdot \boldsymbol{E}_{\text{axis}}}{\|\boldsymbol{e}_j\| \|\boldsymbol{E}_{\text{axis}}\|},$$

where a higher score for journal j indicates that j is more closely aligned to $\boldsymbol{V}^+$ than $\boldsymbol{V}^-$, a stronger alignment with the positive pole (S^+) relative to the negative pole (S^-). This metric allows us to position journals within conceptual continua, such as "basic research" versus "clinical application" or "molecular biology" versus "epidemiology," facilitating the decoding of their multidimensional roles in the life sciences ecosystem.

A key advantage of our approach is the flexibility in defining Semantic axes poles, which can be derived not only from aggregated journal embeddings but also from individual paper embeddings. For instance, poles could be constructed from embeddings of seminal papers representing archetypal concepts, enabling finer-grained analysis at the paper level. This duality supports both macro-scale (journal-level) and micro-scale (paper-level) explorations. Building on the analogical reasoning capabilities of embedding models—exemplified by operations like "king" − "man" + "woman" ≈ "queen"—Semantic axes can operationalize any dimension of interest. Furthermore, this method extends to anchoring low-dimensional subspaces with multiple archetypes, where journals or papers are represented

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

as convex combinations of extremal points with well-established meanings.

3.5 | Effective radius

Journals vary in the degree to which their published papers are tightly clustered or widely dispersed within the embedding space. To quantify this geometric variability, we introduce the effective radius r_g^j, a dispersion metric adapted from the radius of gyration used in physics and spatial statistics (Lee et al., 2025).

Formally, for a journal j with N_j papers, the effective radius is defined as:

$$r_g^j = \sqrt{\frac{1}{N_j}\sum_{i=1}^{N_j}\left|\mathbf{b}_i^j - \boldsymbol{e}^j\right|^2},$$

where $\mathbf{b}_i^j \in \mathbb{R}^{768}$ is the embedding of the ith paper in journal j, and $\boldsymbol{e}^j = \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{b}_i^j$ represents the centroid (average embedding) of the journal's papers. Here, $\|$ denotes the Euclidean norm. This formulation quantifies the root-mean-square distance of individual paper embeddings from the journal's central embedding, analogous to how the radius of gyration measures the spread of mass around a center of gravity in a physical system.

The effective radius thus provides a geometry-based summary of the internal scatter of a journal's content within the embedding space, independent of the absolute position of the centroid. It complements low-dimensional visualizations and semantic axis projections by offering a direct, position-invariant measure of dispersion. We computed this metric for all journals in both the life sciences and social sciences datasets using the paper and journal embeddings.

3.6 | Journal-level measures

We calculated several metrics at the journal level to evaluate the relationship between semantic dispersion and research impact, collaboration, and knowledge flow. Most measures were first calculated at the paper level and then averaged across all eligible papers in each journal.

Disruption (the CD index) quantifies whether a paper pushes a field in a new direction or consolidates existing knowledge. Following established methods (Funk & Owen-Smith, 2017; Wu et al., 2019), the CD index for a focal paper is calculated as:

$$\text{Paper CD} = p_D - p_C = \frac{n_i - n_j}{n_i + n_j + n_k},$$

where n_i represents the count of publications citing the focal paper without referencing its cited sources, n_j denotes the count of publications citing the focal paper and its references, and n_k represents the count of publications citing the focal paper's references without citing FP itself. Scores range from -1 (fully consolidating) to $+1$ (fully disruptive). We restricted calculations to papers with at least one reference and at least one citation, using a fixed 5-year citation window post-publication. Citation linkages were drawn from the PMG 2.0 (approximately 790 million citation links) for life sciences papers and OpenAlex (approximately 2.5 billion citation links) for social sciences papers. The journal-level disruption is the mean across all eligible papers:

$$\text{Avg CD} = \frac{\sum \text{Paper CD}}{N},$$

where N represents the total number of publications by the journal.

We measured interdisciplinarity using the Rao-Stirling index (Stirling, 2007), which accounts for the variety, balance, and disparity of the disciplines cited in a paper's references. References were mapped to disciplines via the MeSH hierarchy for life sciences papers and the OpenAlex topic hierarchy for social sciences papers. For a given paper with a set of references R, the fields of the references are mapped into a field distribution vector F. The paper level Rao-Stirling Index is defined as:

$$\text{Paper Rao} - \text{Stirling} = \sum_{m \neq n} d_{mnt} p_m p_n$$

where p_m denotes the fraction of references in discipline m, and d_{mnt} is the distance between field m and n in year t. Specifically, considering all papers in field m denoted by $M = \{i | \text{field}_i \in m\}$, a large-scale disciplinary co-citation matrix is constructed. This matrix forms a co-citation distribution vector V_m between field m and the other fields. The distance d_{mn} between fields m and n is computed as: $\text{CosSim}(V_m, V_n)$. We restricted calculations to papers with at least five references. We then aggregate interdisciplinarity scores at the journal level as:

$$\text{Avg Rao} - \text{Stirling} = \frac{\sum \text{Paper Rao} - \text{Stirling}}{N}.$$

We further calculated the following measures: (1) h-index: The largest integer h such that the journal has at least h papers each receiving at least h citations

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(computed over the full available citation window in the respective datasets). (2) Journal Impact Factor (JIF): The 2024 Journal Impact Factor from Clarivate Analytics JCR, for journals indexed in the relevant editions. (3) Average author count: Mean number of authors per paper across all papers in the journal. (4) Average reference count: Mean number of references per paper. (5) Average citation count: Mean citations received per paper within a five-year post-publication window. Finally, we employed ordinary least squares (OLS) regression models to examine how the effective radius relates to these metrics. All variables were z-scored prior to regression to facilitate comparison of standardized coefficients.

4 | RESULTS

4.1 | Mapping the journal knowledge landscape

Figure 1 provides an overview of our methodological pipeline and key visualizations of the semantic organization of journals. In Figure 1a, we illustrate the

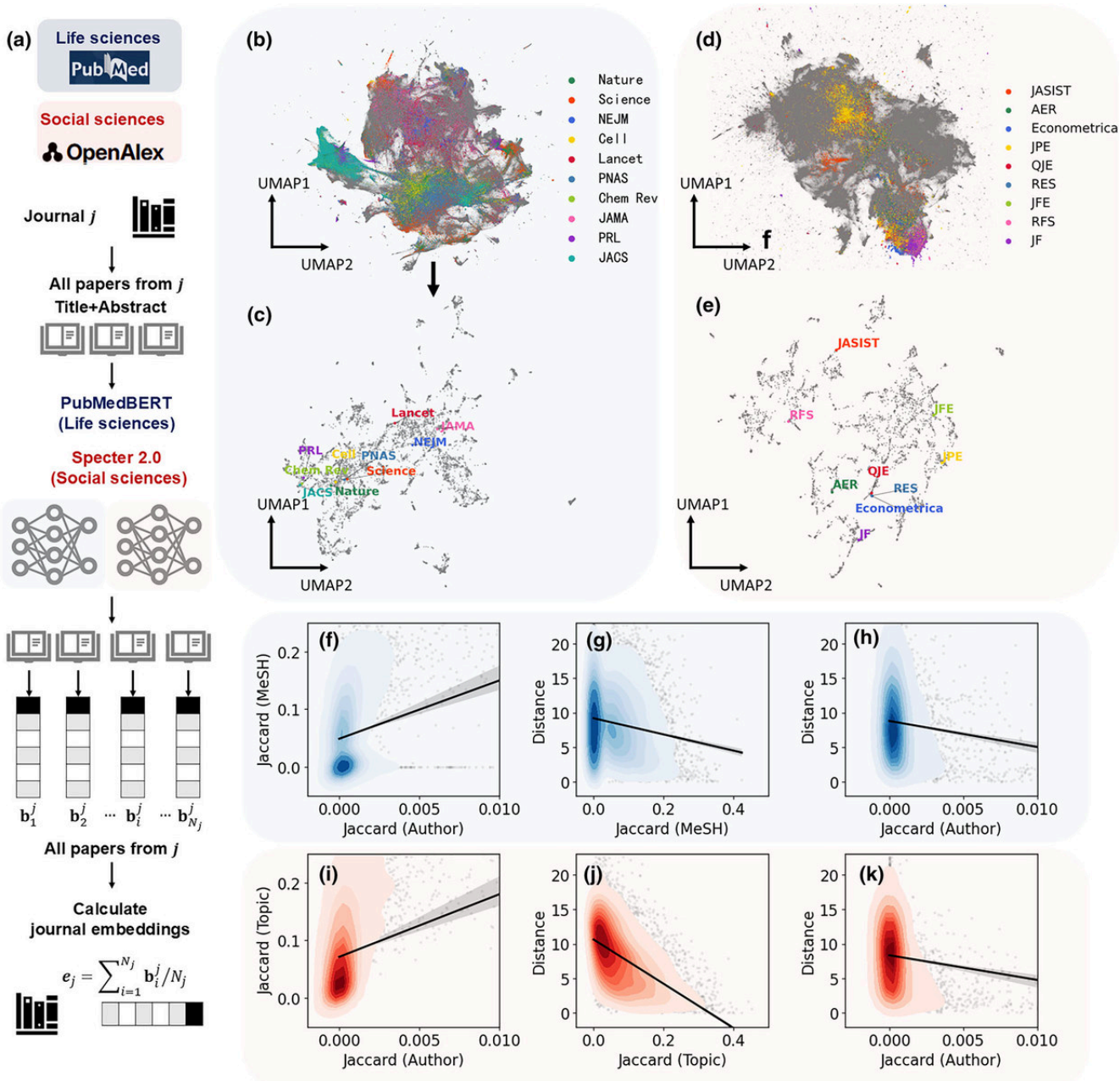


FIGURE 1 Mapping the knowledge landscape of journals.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

construction of journal embeddings. Abstracts were encoded into 768-dimensional vectors using PubMedBERT (Gu et al., 2021) for life sciences papers and SPECTER2 (Singh et al., 2023) for social sciences papers. For each journal, we computed the centroid embedding as the element-wise mean of its constituent paper embeddings, yielding a single 768-dimensional vector that summarizes the journal's position in the embedding space.

Figure 1b,d shows 2D UMAP projections of individual paper embeddings (reduced from 768 dimensions), with papers from selected high-impact journals color-coded for emphasis. In the life sciences, we highlighted papers from the top 10 journals by h-index: Nature, Science, New England Journal of Medicine, Cell, The Lancet, PNAS, Chemical Reviews, JAMA, Physical Review Letters, and Journal of the American Chemical Society. In the social sciences, we highlighted papers from the top five economics journals (American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, Review of Economic Studies), the top three finance journals (Journal of Finance, Journal of Financial Economics, Review of Financial Studies), and JASIST.

Figure 1c,e present the core journal-level UMAP projections (reduced from 768-dimensional centroids), with the same high-impact journals highlighted. These maps exhibit distinct modular clusters, where journals occupy specific intellectual niches while collectively forming a structured, cohesive knowledge landscape across each domain.

To validate the semantic meaningfulness of these journal-level maps, we examined whether positional proximity in the 2D UMAP space corresponds to external measures of topical and social overlap. In the life sciences, topical overlap was assessed using MeSH at the finest granularity (e.g., "Bone Cysts, Aneurysmal" [C04.182.089.265]); in the social sciences, we used OpenAlex topics at the finest level (e.g., topic ID 13606 for "Umberto Eco and Semiotics"). Social overlap was measured via shared authorship, using disambiguated author IDs from PubMed Knowledge Graph 2.0 and OpenAlex.

We quantified positional separation in the 2D UMAP space using Manhattan (L1) distance:

$$d_{\mathrm{L1}}(j_1, j_2) = |x_1 - x_2| + |y_1 - y_2|,$$

where (x_1, y_1) and (x_2, y_2) are the UMAP coordinates of two journals j_1 and j_2. We then assessed topical and social overlaps using Jaccard similarity. For two journals j_1 and j_2, with sets of MeSH terms M_1 and M_2 or author sets A_1 and A_2, the Jaccard similarity is computed as:

$$\mathrm{Jaccard}(M_1, M_2) = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|}, \mathrm{Jaccard}(A_1, A_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|},$$

where $|M_1 \cap M_2|$ and $|M_1 \cup M_2|$ denote the sizes of the intersection and union of topic sets, respectively, and similarly for author sets. To construct these sets, we created journal-paper-topic and journal-paper-author datasets by linking each journal to its associated papers and their corresponding topics or authors.

Given the computational complexity of evaluating all possible pairs among all journals, we randomly sampled 10,000 journal pairs to ensure a manageable yet representative analysis. Mean Jaccard similarities for topics were 0.06 ± 0.06 (life sciences) and 0.08 ± 0.07 (social sciences), indicating moderate overlap. Mean author Jaccard similarities were very low (0.0007 ± 0.002 in life sciences; 0.0005 ± 0.002 in social sciences), consistent with limited cross-journal co-authorship. We observed positive correlations between topical and author Jaccard similarities (Pearson $r = 0.33$, $p < 0.001$ in life sciences; $r = 0.38$, $p < 0.001$ in social sciences; Figure 1f,i), suggesting that journals sharing topical content are more likely to share authors.

L1 distances (mean 8.77 ± 4.85 across pairs) showed significant negative correlations with Jaccard similarities. In life sciences: L1 distance versus MeSH Jaccard ($r = -0.16$, $p < 0.001$; Figure 1g); versus author Jaccard ($r = -0.15$, $p < 0.001$; Figure 1h). In social sciences: L1 distance versus topic Jaccard ($r = -0.51$, $p < 0.001$; Figure 1j); versus author Jaccard ($r = -0.30$, $p < 0.001$). These negative associations confirm that journals closer in the UMAP-derived semantic space exhibit greater topical and author overlap, validating the interpretive power of our embeddings and projections. The maps thus capture meaningful dimensions of knowledge organization and scientific social structure in both domains.

4.2 | Decoding thematic spectra of journals across biomedical dimensions

We applied the semantic axis framework to project life sciences journal embeddings onto two complementary conceptual dimensions: a disease-oriented axis contrasting human-centric Neoplasms with Animal Diseases, and an anatomical axis contrasting Animal Structures with Plant Structures. These axes capture fundamental dichotomies within life sciences: human health versus veterinary medicine, and animal-centric versus plant-centric research.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We construct a semantic axis by defining two poles: one pole representing papers associated with Neoplasms (MeSH C04, encompassing cancers) and the other representing papers associated with Animal Diseases (MeSH C22). These poles were chosen to capture a contrast between human-focused clinical research and veterinary or comparative medicine. For the positive pole, we computed the average embedding of all papers tagged with C04 MeSH terms, denoted as $S^+ = \{e_1^+, e_2^+ \cdots e_n^+\}$, yielding $E^+ = \frac{1}{n}\sum_{i=1}^{n} e_i^+$. Similarly, for the negative pole, we averaged the embeddings of papers tagged with C22 MeSH terms, denoted as $S^- = \{e_1^-, e_2^- \cdots e_n^-\}$, yielding $E^- = \frac{1}{n}\sum_{i=1}^{n} e_i^-$. These embeddings were derived from the 768-dimensional vectors. The semantic axis was defined as $E_{\text{axis}} = E^+ - E^-$. For each journal, we calculated the score to quantify its alignment along this axis: $\text{score} = \frac{e_j \cdot E_{\text{axis}}}{\|e_j\| \|E_{\text{axis}}\|}$.

To contextualize the SemAxis scores, we classified journals based on their dominant MeSH terms within the broad Diseases category (C). A journal was assigned a specific disease category (e.g., Infections [C01], Neoplasms [C04], Cardiovascular Diseases [C14]) if its most frequent MeSH term accounted for at least 20% of its total MeSH terms. Journals not meeting this threshold were classified as multidisciplinary, reflecting their coverage of diverse topics without a singular focus. Figure 2a presents the UMAP projection of all the life sciences journal embeddings, color-coded by their assigned MeSH-based disease categories.

Figure 2b,c illustrates the distribution of journal scores along the Neoplasms–Animal Diseases semantic axis. Journals classified as Neoplasms exhibit the highest positive scores (mean = 0.008), while journals classified as Animal Diseases show the lowest negative scores (mean = −0.024). Journals associated with Infections (C01) also tend toward negative scores (mean = −0.015), suggesting closer semantic proximity to animal diseases, possibly due to shared themes in infectious disease research across species (e.g., zoonotic diseases). Other disease categories, such as Musculoskeletal Diseases (C05; mean = −0.001), Digestive System Diseases (C06; mean = 0.003), and Cardiovascular Diseases (C14; mean = −0.005), cluster near neutral scores, indicating a balanced position along this axis. Multidisciplinary journals (mean = −0.004) similarly show near-neutral scores, consistent with their broad coverage spanning multiple disease areas.

We extended our semantic axis analysis to the Anatomy dimension, contrasting animal-centric and plant-centric research. Journals were classified based on their dominant MeSH terms within the Anatomy category (A). Figure 2d presents the UMAP projection of the journal embeddings, color-coded by their assigned MeSH-based anatomical categories. Figure 2e,f illustrates the distribution of journal scores along the Animal Structures–Plant Structures SemAxis. Journals classified as Animal Structures (A13) exhibit the highest positive scores (mean = 0.022); while journals classified as Plant Structures (A18) show the lowest scores (mean = −0.004). Journals associated with Cells (A11; mean = 0.013) display a wide range of scores, suggesting their relevance to both animal and plant research, likely due to the foundational role of cellular biology across life sciences. Other anatomical categories, such as Nervous System (A08; mean = 0.019), Cardiovascular System (A07; mean = 0.021), and Embryonic Structures (A16; mean = 0.024), tend toward positive scores, indicating a stronger alignment with animal-centric research, which dominates the life sciences literature. Multidisciplinary journals (mean = 0.017) show positive but moderate scores, reflecting their balanced coverage of both animal and plant-related topics.

To examine whether these semantic axes align with the structure of the UMAP projection, we overlaid axis scores directly onto the journal-level UMAP coordinates. The resulting color gradients (Supplementary Figures S2 and S3) show that scores are heterogeneously scattered across the manifold. No coherent directional gradient emerges along the UMAP layout for either axis. These projections quantify how journals occupy positions along biologically meaningful continua, revealing hierarchical and interdisciplinary structures that complement the overall knowledge landscape.

4.3 | Ranking journals along quality dimensions

In addition to thematic axes, we explored whether semantic embeddings capture latent signals of journal prestige or "quality" by constructing a dedicated axis contrasting high- and low-prestige outlets. This dimension operationalizes perceived quality through vector differences between exemplar high-impact journals and lower-tier ones.

For life sciences journals, we defined the positive pole S^+ as the mean embedding of all papers from the top five journals by *h*-index: Nature, Science, New England Journal of Medicine (NEJM), Cell, and The Lancet. These represent archetypes of high-impact, often multidisciplinary life sciences research. The negative pole S^- was the mean embedding of papers from the bottom 100 journals ranked by *h*-index (lowest-prestige outlets in the dataset).

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Both poles used the original 768-dimensional embeddings. The axis direction was $\boldsymbol{E}_{\text{axis}} = \boldsymbol{E}^{+} - \boldsymbol{E}^{-}$, where $\boldsymbol{E}^{+}$ and $\boldsymbol{E}^{-}$ are the respective mean vectors. Journal quality scores were computed as cosine similarities to this axis.

To examine whether the quality axis aligns with the structure of the UMAP projection, we overlaid quality scores directly onto the journal-level UMAP coordinates (Figure 3a,d for life sciences and social sciences,

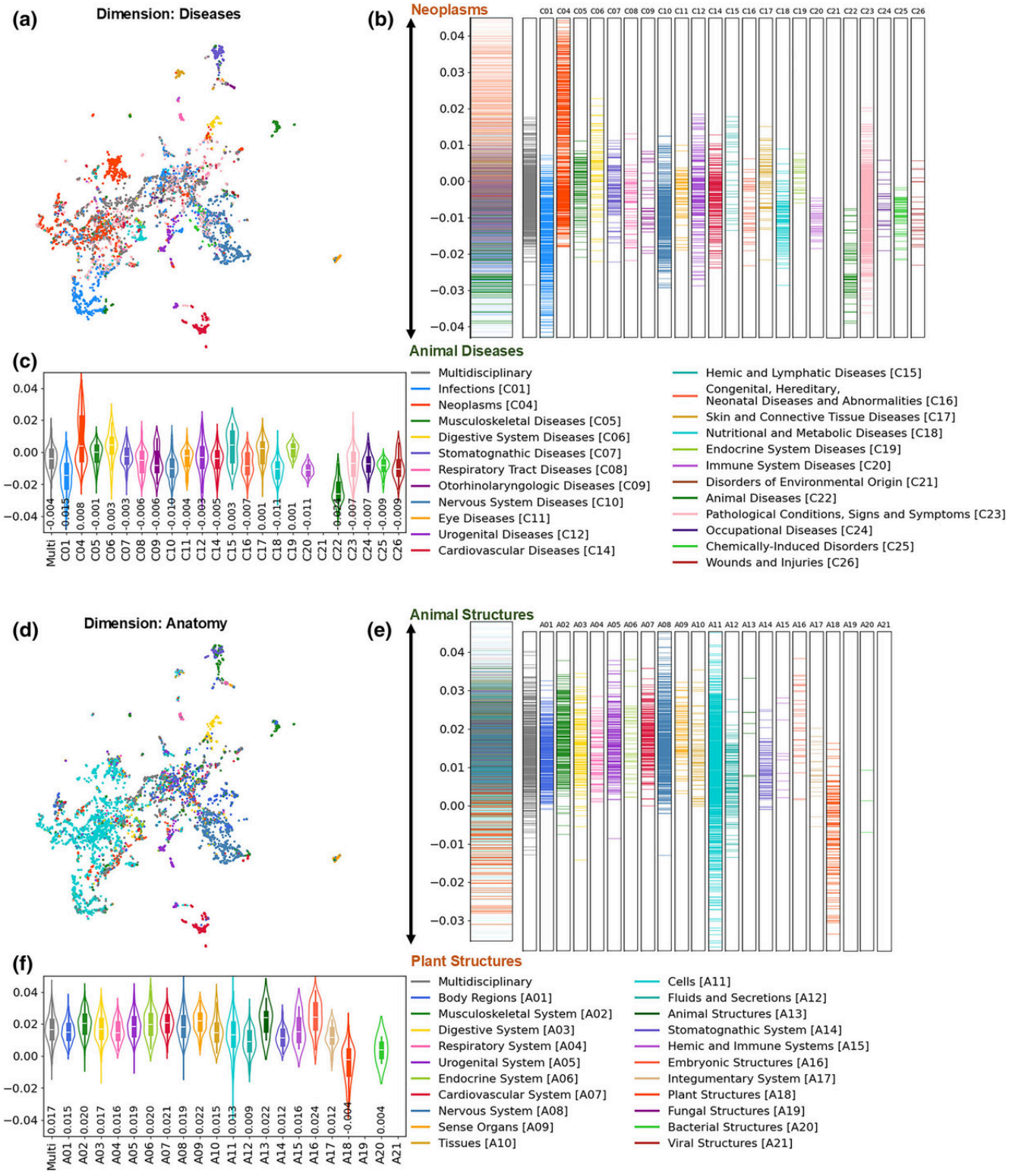


FIGURE 2 Mapping thematic spectra of journals across biomedical dimensions.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

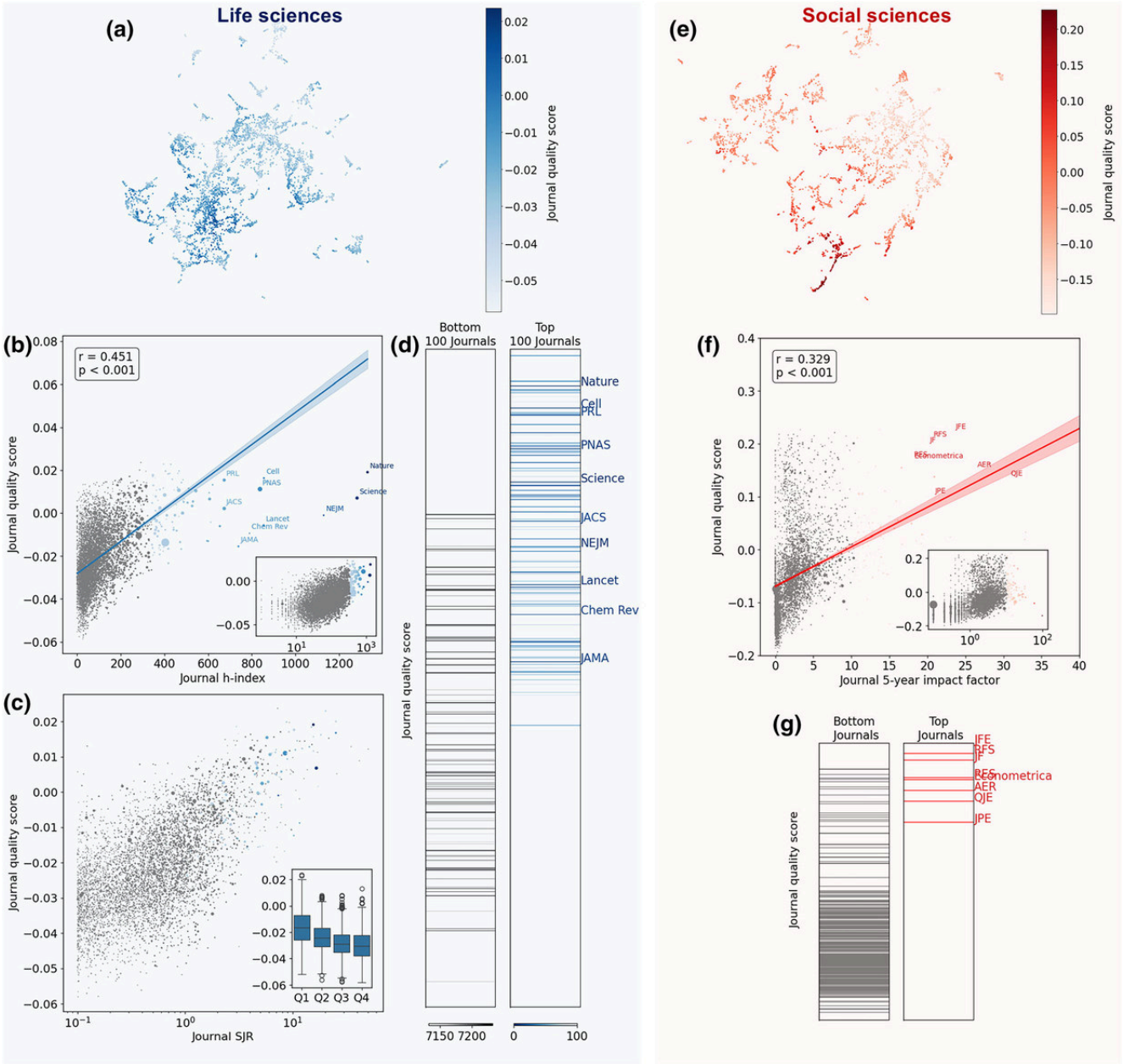


FIGURE 3 Ranking journals along quality dimensions.

respectively). The resulting gradients show that quality scores are heterogeneously scattered across the manifold, with high- and low-scoring journals intermingled within and across clusters. No coherent directional gradient emerges along the UMAP layout.

Figure 3b shows a scatter plot of quality scores against journal *h*-index for life sciences journals, with point sizes proportional to publication volume. We observed a strong positive correlation (Pearson $r = 0.451$, $p < 0.001$), indicating that higher *h*-index journals align more closely with the positive (high-prestige) pole. This suggests that embeddings implicitly capture quality-related attributes beyond explicit topical content. To further assess construct validity, we examined associations with the Scimago Journal Rank (SJR), a field-weighted citation prestige metric. Figure 3c presents quality scores versus SJR, revealing a comparable positive correlation (Pearson $r = 0.435$, $p < 0.001$). Figure 3c (inset) compares mean quality scores across SJR quartiles (Q1 highest to Q4 lowest): Q1–0.016 ± 0.013, Q2–0.024 ± 0.010, Q3 0.028 ± 0.011, Q4–0.029 ± 0.011. Pairwise *t*-tests confirmed significant differences between consecutive quartiles (all $p < 0.001$), demonstrating the axis's discriminative power across prestige tiers. Figure 3d ranks all journals along the quality score spectrum, forming a clear gradient. Top positions are occupied by the defining positive-pole

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

journals, followed by other high-impact venues (e.g., PNAS, JAMA). Lower scores correspond to specialized or lower-influence journals, mirroring conventional prestige hierarchies in a continuous, embedding-derived manner.

For social sciences journals, we applied an analogous approach in the social sciences dataset. The positive pole was the mean embedding of papers from the top 5 economics journals (American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, Review of Economic Studies) and top three finance journals (Journal of Finance, Journal of Financial Economics, Review of Financial Studies). The negative pole comprised papers from the 362 journals with 5-year JIF near zero or undefined (lowest-prestige outlets). Figure 3f also shows quality scores versus 5-year JIF (point sizes proportional to publication volume), with a positive correlation (Pearson $r = 0.329$, $p < 0.001$).

Across both domains, these quality axes demonstrate that transformer-based embeddings can recover meaningful prestige gradients from abstract text alone, independent of explicit citation information. The moderate-to-strong correlations with established bibliometric indicators (*h*-index, SJR, JIF) suggest that embeddings implicitly encode quality-related latent features—potentially including stylistic elements (e.g., clarity, structure), substantive novelty, synthesis depth, or methodological rigor—that align with conventional prestige measures. This approach offers a scalable, citation-agnostic proxy for journal evaluation, enabling text-driven analyses of scientific quality hierarchies.

4.4 | Identifying potentially predatory journals in knowledge space

Predatory journals, characterized by lax peer review, deceptive practices, and low editorial standards, pose challenges to the integrity of scientific communication. To assess whether our knowledge atlas can identify such journals, we leveraged retraction records as a proxy for questionable publishing practices. Retractions often signal issues like plagiarism, data fabrication, or ethical violations, which are disproportionately prevalent in predatory outlets. Building on prior work González-Márquez et al. (2024), we integrated retraction data into our analysis to map the distribution of potentially predatory journals within the life sciences knowledge space. We obtained the Retraction Watch Database (RWD) as of January 1, 2025, comprising 47,962 retracted records. These were matched to our dataset. We generated embeddings for these retracted papers and projected them onto the space alongside non-retracted papers.

A journal was classified as potentially predatory or questionable if at least t of its papers in our dataset were retracted, where t represents varying thresholds (0.1%, 0.5%, 1%, and 2%). This threshold-based approach allows for sensitivity analysis, capturing a spectrum from mildly suspicious to highly problematic journals. For each threshold, we overlaid the classified journals onto the atlas to visualize their positioning.

Figure 4a displays the UMAP projection of all paper embeddings, with retracted papers highlighted in red. Consistent with González-Márquez et al. (2024), retracted papers form distinct clusters, suggesting semantic anomalies such as duplicated content or superficial analyses that deviate from established scientific norms. Figure 4b illustrates the distribution of potentially predatory journals, with journals flagged at different retraction thresholds color-coded for clarity. At the lowest threshold ($t = 0.1\%$), 1168 journals (16.1% of total) are flagged, appearing diffusely across the space, including some within central clusters alongside reputable journals. This broad distribution suggests that low retraction rates may reflect isolated errors rather than systemic predatory behavior. As the threshold increases to ($t = 0.5\%$) (205 journals, 2.8%) and ($t = 1\%$) (79 journals, 1.1%), flagged journals consolidate into more distinct peripheral clusters, indicating a stronger association with questionable practices. At the stringent ($t = 2\%$) threshold (23 journals, 0.3%), three clusters emerge, predominantly occupying certain areas of the knowledge landscape, suggesting that journals with high retraction rates share semantic characteristics that set them apart from reputable outlets.

4.5 | Journal effective radius

We examined the distribution of effective radius across journals and its relationships to key bibliometric and semantic attributes in both the life sciences and social sciences datasets. Figure 5a shows the distribution of effective radii for the 7279 life sciences journals. The distribution is approximately log-normal, with a mean of approximately 1.0 (in embedding-space units). Most journals exhibit moderate average semantic distance from their centroid, with a long right tail indicating a subset of journals with substantially larger dispersion and a sharper left tail for journals with smaller dispersion. Figure 5d presents the corresponding distribution for the 4210 social sciences journals, which is closer to normal and has a markedly higher mean (approximately 5.0). This suggests that, on average, journals in social sciences exhibit greater semantic dispersion around their centroids than those in the life sciences.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

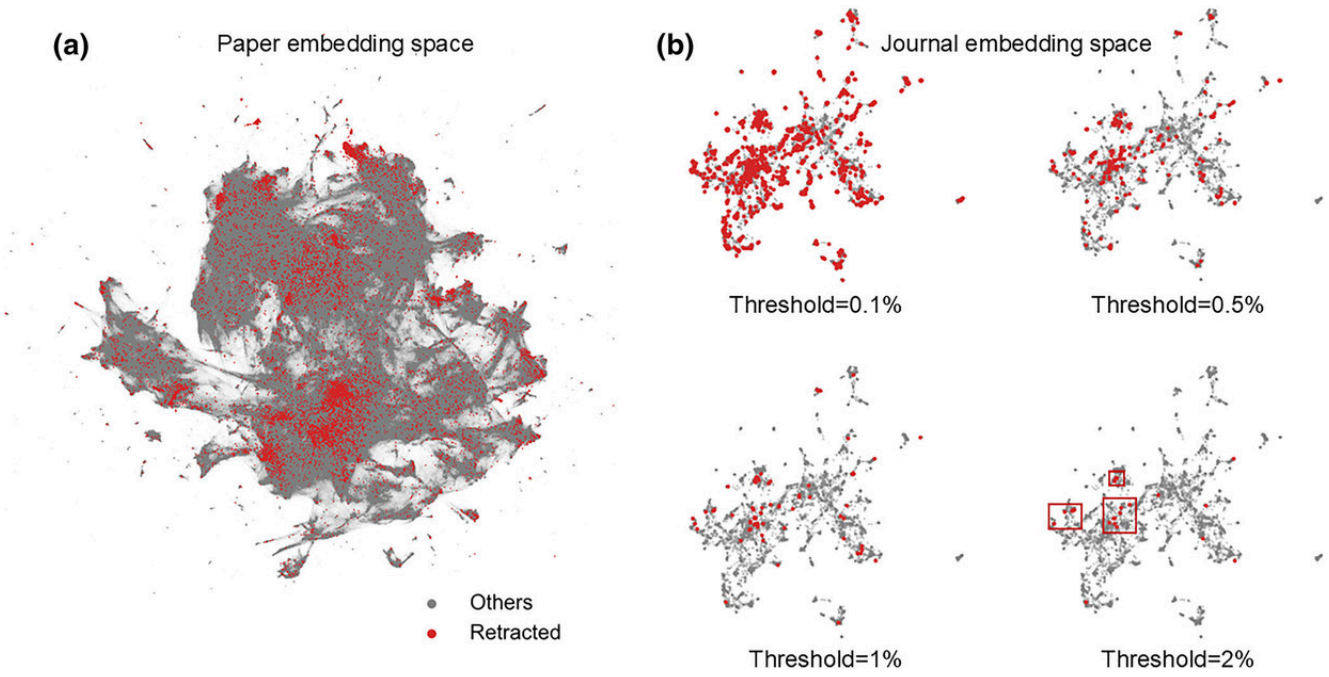


FIGURE 4 Identifying potentially predatory journals in the knowledge space.

Figure 5b,e overlay effective radius values onto the UMAP journal projections, with color intensity indicating radius magnitude (darker = larger). A clear spatial gradient emerges in both domains: journals near the center of the manifold tend to have smaller effective radii (tighter clustering of paper embeddings around the journal centroid), whereas journals toward the periphery exhibit larger radii (greater average semantic distance from the centroid). This core–periphery pattern mirrors organizational principles observed in biological systems, where central modules often display high functional specificity and peripheral elements exhibit broader exploratory variation.

To quantify associations between effective radius and journal characteristics, we computed the following paper-level metrics and aggregated them to the journal level: Journal Impact Factor, average team size, average reference count, average 5-year citation count, average Rao–Stirling interdisciplinarity index, and average disruption score (CD index). We then fitted OLS regression models with effective radius as the dependent variable and each metric as the independent variable. All variables were z-scored prior to regression to enable direct comparison of standardized coefficients (β).

Figure 5b (life sciences) and Figure 5e (social sciences) include scatter plots of effective radius against selected metrics. Figure 5c,f summarizes the standardized regression coefficients (β) and their statistical significance. Effective radius showed positive associations with impact metrics in both domains: JIF (life sciences: $\beta = 0.05$, $p < 0.001$; social sciences: $\beta = 0.13$, $p < 0.001$) and average citations (life sciences: $\beta = 0.03$, $p < 0.001$; social sciences: $\beta = 0.22$, $p < 0.001$). These results indicate that journals positioned toward the periphery of the manifold (larger effective radius, longer average semantic distance) tend to achieve higher impact and accumulate more citations than journals near the center.

The relationship between effective radius and average reference count was negative in life sciences ($\beta = -0.16$, $p < 0.001$), such that journals near the center cite more references on average; positive in social sciences ($\beta = 0.19$, $p < 0.001$), such that peripheral journals draw from a larger number of references. Similarly, the relationship between effective radius and average interdisciplinarity was negative in life sciences ($\beta = -0.10$, $p < 0.001$); positive in social sciences ($\beta = 0.19$, $p < 0.001$).

In the life sciences, effective radius showed a positive association with average disruptiveness ($\beta = 0.17$, $p < 0.001$), suggesting that peripherally positioned journals tend to publish more disruptive work. In the social sciences, the relationship between effective radius and average disruptiveness was negative ($\beta = -0.09$, $p < 0.001$), suggesting that center-located journals tend to publish more disruptive work.

5 | DISCUSSION

In this study, we constructed semantic atlases of journals in the life sciences and social sciences domains using

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

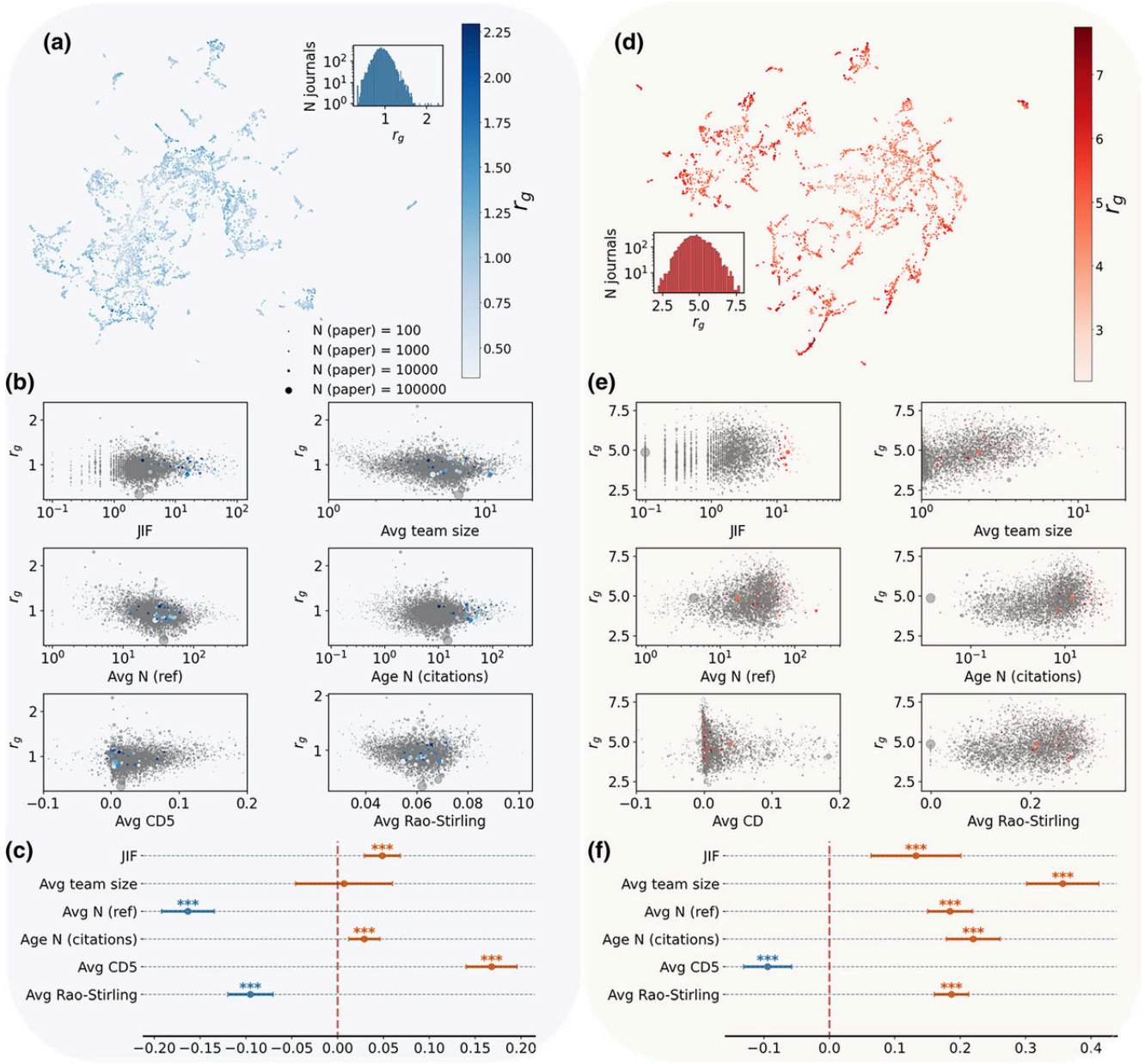


FIGURE 5 Journal effective radius in the knowledge space.

large-scale abstract embeddings, journal centroids, and dimensionality reduction techniques. Our analyses revealed structured, modular knowledge landscapes where journals cluster by thematic similarity, validated through correlations with external topical and authorship overlaps. Semantic axes illuminated gradients along disease, anatomical, and prestige dimensions, enabling quantitative positioning of journals along meaningful continua. We further demonstrated the utility of embeddings for identifying potentially predatory journals via retraction-based clustering and for characterizing topical breadth through effective radius, which exhibited domain-specific associations with impact, interdisciplinarity, and disruptiveness.

Our embedding-based semantic atlases reveal modular clusters aligned with thematic and social overlaps (Pearson correlations up to $r = 0.51$ for topical similarity). Semantic axes effectively decode gradients, such as human-centric versus veterinary disease foci (mean scores differentiating categories by up to 0.032) and animal versus plant anatomy (differences up to 0.026). The prestige axis recovers bibliometric hierarchies ($r = 0.329$–0.451 with JIF/h-index), confirming embeddings encode latent quality signals. Effective radius distributions (log-normal in life sciences, normal in social sciences) and regressions (β up to 0.22 for citations) link greater topical breadth to enhanced impact, though with domain-specific

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

patterns for interdisciplinarity and disruption. These findings interpret the scientific ecosystem as a geometrically structured space, where central journals maintain focused coherence and peripheral ones enable broader exploration and higher visibility. Additionally, the semantic axis scores show no systematic alignment with the principal directions or gradients of the UMAP projection. High- and low-scoring journals along each axis are distributed heterogeneously across the 2D manifold, indicating that the UMAP layout primarily captures broad structural and topical proximities, while the semantic axes represent orthogonal conceptual dimensions not preserved in the reduced 2D representation. This dissociation is expected given UMAP's focus on preserving local neighborhood structure rather than global linear axes, and it highlights the complementary nature of dimensionality reduction for visualization versus semantic axis projection for interpretable thematic continua.

These findings have several implications for understanding the organization of scientific knowledge. First, the embedding-based atlases provide a continuous, data-driven alternative to discrete classification systems (e.g., MeSH or OpenAlex topics), facilitating nuanced analyses of interdisciplinary boundaries and journal niches. For instance, the core–periphery gradient in effective radius suggests that peripheral journals may drive higher impact in some domains, potentially by integrating diverse ideas or exploring emerging topics. This aligns with theories of scientific innovation, where boundary-spanning work fosters breakthroughs (Uzzi et al., 2013; Wu et al., 2019). Second, the quality axis offers a text-intrinsic proxy for prestige, complementing citation-based metrics and enabling rapid evaluation of journals without long-term accrual data. This could inform funding agencies, researchers, and policymakers in assessing publication venues, particularly in fast-evolving fields. Third, the peripheral clustering of high-retraction journals highlights semantic embeddings as a tool for integrity monitoring, potentially aiding automated detection of questionable practices beyond traditional watchlists.

While our framework demonstrates applicability across datasets, we acknowledge that the life sciences currently represent the most suitable domain for constructing highly interpretable semantic axes. The hierarchical, tree-structured nature of MeSH provides well-defined, controlled vocabularies with clear oppositional categories, enabling straightforward and theoretically motivated pole selection. In contrast, many other disciplines lack equivalent comprehensive, hierarchical terminologies. In such fields, defining semantically meaningful axes may require alternative strategies, such as empirical clustering of high-impact papers, expert-curated archetypal works, or data-driven identification of antonym-like oppositions from the corpus itself. These approaches, while feasible, introduce greater subjectivity and validation challenges compared to MeSH-guided pole construction.

Our approach also opens avenues for dynamic monitoring of knowledge structures. If new publications are added, the atlas could evolve incrementally by updating embeddings and centroids, likely preserving core clusters while expanding peripheral regions to accommodate emerging subfields. However, rapid influxes might disrupt existing gradients, requiring periodic re-projection to maintain stability. To extend this framework to other disciplines, domain-specific embedding models and comprehensive datasets would be essential. Adaptations might include incorporating full-text data for deeper semantic capture or hybrid embeddings blending text with citation graphs to address domain-specific knowledge flows.

Despite these strengths, our study has limitations. First, reliance on abstracts may overlook nuances in full-text methods or discussions, potentially biasing embeddings toward high-level overviews. Second, the models (PubMed BERT and SPECTER2) are pretrained on specific corpora, which could introduce biases (e.g., underrepresentation of non-Western research). Third, UMAP projections are stochastic and preserve local structure preferentially, so fine-scale distances should be interpreted cautiously. Fourth, domain differences in effective radius correlates (e.g., opposing signs for interdisciplinarity) underscore that findings may not generalize without tailored validations. Finally, retraction-based predatory journal identification is a proxy; not all questionable outlets retract papers, and benign retractions (e.g., honest errors) could misclassify journals.

ACKNOWLEDGMENTS

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

CONFLICT OF INTEREST STATEMENT

The author declares no competing interest relevant to this article.

DATA AVAILABILITY STATEMENT

The OpenAlex dataset is available at https://docs.openalex.org/download-all-data/openalex-snapshot. The pre-trained parameters for the SPECTER 2.0 text embedding model are available at https://huggingface.co/allenai/specter2. The PubMed database is available at www.nlm.nih.gov/databases/download/pubmed_medline.html. The pre-trained www.nlm.nih.gov/databases/download/pubmed_medline.html. The pre-trained parameters for the PUBMEDBERT embedding model are available at https://huggingface.co/NeuML/pubmedbert-base-embeddings.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ORCID

Alex J. Yang https://orcid.org/0000-0002-1276-0483
Sanhong Deng https://orcid.org/0000-0002-6910-3935

REFERENCES

Aceves, P., & Evans, J. A. (2024a). Human languages with greater information density have higher communication speed but lower conversation breadth. *Nature Human Behaviour*, *8*, 644–656. https://doi.org/10.1038/s41562-024-01815-w

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

An, J., Kwak, H., & Ahn, Y.-Y. (2018). SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics*. Australia.

Baldwin, M. (2019). *Making "nature" the history of a scientific journal*. University of Chicago Press.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Bollen, J., de van Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS One*, *4*(3), e4803. https://doi.org/10.1371/journal.pone.0004803

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V., & Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS One*, *7*(7), e39464. https://doi.org/10.1371/journal.pone.0039464

Boyack, K. W., & Klavans, R. (2011). Multiple dimensions of journal specificity: Why journals can't be assigned to disciplines. In *13th conference of the international society for scientometrics and informetrics*.

Boyack, K. W., & Klavans, R. (2020). A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies*, *1*(4), 1570–1585. https://doi.org/10.1162/qss_a_00085

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.

Bush, V. (1945). *Science: The endless frontier*. US General Printing Office.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics*.

Fersht, A. (2009). The most influential journals: Impact factor and Eigenfactor. *Proceedings of the National Academy of Sciences*, *106*(17), 6883–6884. https://doi.org/10.1073/pnas.0903307106

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

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

González-Márquez, R., Schmidt, L., Schmidt, B. M., Berens, P., & Kobak, D. (2024). The landscape of biomedical research. *Patterns*, *5*(6), 100968. https://doi.org/10.1016/j.patter.2024.100968

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23.

Hao, Q., Xu, F., Li, Y., & Evans, J. (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, *649*, 1237–1243. https://doi.org/10.1038/s41586-025-09922-y

Healy, J., & McInnes, L. (2024). Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, *4*(1), 82. https://doi.org/10.1038/s43586-024-00363-x

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

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

Hutchins, B. I., Davis, M. T., Meseroll, R. A., & Santangelo, G. M. (2019). Predicting translational progress in biomedical research. *PLoS Biology*, *17*(10), e3000416. https://doi.org/10.1371/journal.pbio.3000416

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, *3*(3), 303–320. https://doi.org/10.1257/aeri.20190499

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Lawson, M. A., Martin, A. E., Huda, I., & Matz, S. C. (2022). Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language. *Proceedings of the National Academy of Sciences*, *119*(9), e2026443119. https://doi.org/10.1073/pnas.2026443119

Lee, B., Aiyappa, R., Ahn, Y.-Y., Kwak, H., & An, J. (2025). A semantic embedding space based on large language models for modelling human beliefs. *Nature Human Behaviour*, *9*, 1928–1940. https://doi.org/10.1038/s41562-025-02228-z

Lix, K., Goldberg, A., Srivastava, S. B., & Valentine, M. A. (2022). Aligning differences: Discursive diversity and team performance. *Management Science*, *68*(11), 8430–8448. https://doi.org/10.1287/mnsc.2021.4274

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Milojević, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4), 962–973. https://doi.org/10.1016/j.joi.2015.10.005

Murray, D., Yoon, J., Kojaku, S., Costas, R., Jung, W. S., Milojević, S., & Ahn, Y. Y. (2023). Unsupervised embedding of trajectories captures the latent structure of scientific migration. *Proceedings of the National Academy of Sciences*, *120*(52), e2305414120. https://doi.org/10.1073/pnas.2305414120

Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, *7*(17), eabb9004. https://doi.org/10.1126/sciadv.abb9004

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Scharfmann, E., Marx, M., & Fleming, L. (2025). Pasteur's quadrant researchers bring novelty, impact to publishing, and patenting. *Science*, *390*(6776), 891–893. https://doi.org/10.1126/science.adx3736

Shi, F., & Evans, J. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, *14*(1), 1641. https://doi.org/10.1038/s41467-023-36741-4

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2023). Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 conference on empirical methods in natural language processing*.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*(9), 799–813.

Sourati, J., & Evans, J. A. (2023). Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, *7*(10), 1682–1696. https://doi.org/10.1038/s41562-023-01648-z

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Taylor, M. A., & Stoltz, D. S. (2021). Integrating semantic directions with concept mover's distance to measure binary concept engagement. *Journal of Computational Social Science*, *4*(1), 231–242. https://doi.org/10.1007/s42001-020-00075-8

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Weber, G. M. (2013). Identifying translational science within the triangle of biomedicine. *Journal of Translational Medicine*, *11*(1), 126. https://doi.org/10.1186/1479-5876-11-126

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

Xiang, S., Jiang, X., Liu, B., Huang, Y., Tian, C., & Ma, Y. (2025). SciConNav: Knowledge navigation through contextual learning of extensive scientific research trajectories. *Journal of the Association for Information Science and Technology*, *76*(10), 1308–1339. https://doi.org/10.1002/asi.25005

Xu, J., Yu, C., Xu, J., Torvik, V. I., Kang, J., Sung, M., Song, M., Bu, Y., & Ding, Y. (2025). PubMed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *Scientific Data*, *12*(1), 1018. https://doi.org/10.1038/s41597-025-05343-8

Zhang, S., Liu, F., & Xia, H. (2025). Quantifying spatial–temporal citation diffusion of individual papers in knowledge space. *Journal of the Association for Information Science and Technology*, *76*, 1682–1700. https://doi.org/10.1002/asi.70021

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang, A. J., Freeman, R. B., & Deng, S. (2026). A semantic atlas of journals: Structure, position, and dispersion. *Journal of the Association for Information Science and Technology*, 1–17. https://doi.org/10.1002/asi.70067

23301643, 0, Downloaded from https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.70067 by Nanjing University, Wiley Online Library on [20/05/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License