Received: 30 September 2024 | Revised: 15 February 2025 | Accepted: 11 March 2025

DOI: 10.1002/asi.25005

RESEARCH ARTICLE

JASIST WILEY

SciConNav: Knowledge navigation through contextual learning of extensive scientific research trajectories

Shibing Xiang[1,2] | **Xin Jiang**[3,4] | **Bing Liu**[5,6] | **Yurui Huang**[1] | **Chaolin Tian**[1] | **Yifang Ma**[1]

[1]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong, China

[2]Pengcheng Laboratory, Shenzhen, Guangdong, China

[3]Zhongguancun Laboratory, Beijing, China

[4]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, School of Artificial Intelligence, Beihang University, Beijing, China

[5]LMIB, School of Mathematical Sciences, Beihang University, Beijing, China

[6]Zhengzhou Aerotropolis Institute of Artificial Intelligence, Zhengzhou, Henan, China

Correspondence
Yifang Ma, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong, China.
Email: mayf@sustech.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: NSFC62006109, NSFC12031005; The Major Key Project of PCL, Grant/Award Number: PCL2023A09; The Stable Support Plan Program of Shenzhen Natural Science Fund, Grant/Award Number: 20220814165010001

Abstract

New knowledge builds upon existing foundations, which means an interdependent relationship exists between knowledge, manifested in the historical records of the scientific system for hundreds of years. By leveraging natural language processing techniques, this study introduces the Scientific Concept Navigator, an embedding-based navigation model to infer the "knowledge pathway" from the research trajectories of millions of scholars. We validate that the learned representations effectively delineate disciplinary boundaries and capture the intricate relationships between diverse concepts. Utility of the navigation space is showcased through multiple applications. Firstly, we demonstrate the multi-step analogy inferences between concepts from various disciplines. Secondly, we formulate the cross-domain conceptual dimensions of knowledge, observing the distributional shifts of 19 disciplines along these conceptual dimensions, including "Theoretical" to "Applied," and "Societal" to "Economic," highlighting the evolution of functional attributes across diverse domains. Lastly, by analyzing the knowledge network structure, we find that knowledge connects with shorter global pathways, and interdisciplinary concepts play a critical role in enhancing accessibility. Our framework offers a novel approach to mining knowledge inheritance pathways from extensive scientific literature, which is of great significance for understanding scientific progression patterns, tailoring scientific learning trajectories, and accelerating scientific progress.

1 | INTRODUCTION

In the current science and technology landscape, we are experiencing an unprecedented expansion of knowledge and information overload. Knowledge manifests in various tangible forms, including reports, papers, books, mathematical formulas, and diagrams (Abdullah et al., 2002). This growth is evident in the increasing number of publications, research topics, and patents, all documented in academic databases and extensive scientific knowledge graphs like Wikipedia, Web of Science (Xu & Dan, 2023), Scopus, PubMed (White, 2020),

Dimensions (Hook et al., 2018), OpenAlex (Priem et al., 2022), and Aminer (Tang, 2016).

Researchers constantly face the challenge of selecting the next topic for exploration. The overwhelming influx of research and information makes it difficult for scientists to stay abreast of the latest developments, adopt interdisciplinary approaches, and identify the most rational research paths to scientific discoveries and advanced technologies (Shirah & Sidney, 2023). This challenge is particularly daunting for early-career scientists who may not yet have a broad knowledge base in their respective fields. Knowledge navigation (Hammond, Burke, & Lytinen, 1995; Hammond, Burke, Martin, & Lytinen, 1995; Li et al., 2014) has emerged as a viable solution, serving as intelligent assistance for navigation (Amant et al., 1998), providing access to valuable and reliable information (Jadad & Gagliardi, 1998), assisting learners and researchers in navigating through the expanding information spaces (Benyon, 2001; Burke et al., 1996) and the complexities of modern scientific research. Understanding complex knowledge structures and their dependency relationships is crucial for effectively navigating the intellectually structured vast knowledge landscape and providing access to valuable and reliable information (Jadad & Gagliardi, 1998).

Previous researches on knowledge navigation emphasize computer-based systems or platforms for information retrieval and well-organized knowledge repositories with complex interactions (Patel & Kushniruk, 1998) to facilitate access and intuitive navigation. For instance, the McSyBi (Yamamoto & Takagi, 2007) navigation system assists in acquiring knowledge from biomedical literature, offering researchers a comprehensive overview of topics and their interrelationships. The TaxoFolk (Kiu & Tsui, 2010) introduces the hybrid taxonomy-folksonomy classifications to enhance knowledge retrieval. Similarly, CoNavigator (Hao & Park, 2021) employs formal concept analysis for domain-specific knowledge acquisition, particularly in the context of Coronavirus Disease 2019 (COVID-19). The Knowledge Navigator Model (KNM) (Hsieh et al., 2009; Hsieh et al., 2020) guides structured knowledge management within organizations, representing high-level behavioral or implementation navigation. These models help users acquire meta-knowledge; however, these methods rarely integrate the latest representation learning techniques to capture contextual semantics in natural language processing (NLP), fail to leverage knowledge interdependencies for accurate retrieval, and often overlook the identification of the most relevant global knowledge pathways.

In the field of the Science of Science (Fortunato et al., 2018; Zeng et al., 2017), the knowledge mapping (Vail, 1999) formulates the graphical knowledge map to analyze the latent structure of knowledge (Chiu & Pan, 2014; Miao et al., 2022), reveals learning dependency (Liu et al., 2012) and extensive information from academic entities. Mapping refers to the process of creating a visual representation (Wexler, 2001) of graphical maps to display knowledge entities and their interrelationships, such as the co-occurrence or collaboration networks (He, Wang, Shang, et al., 2022). Leading software like CiteSpace (Chen, 2006) and VOSViewer (Van Eck & Waltman, 2010) are continuously evolving to enhance knowledge visualization. Recent studies aim to uncover the latent knowledge landscape (González-Márquez et al., 2024) using embedding techniques in NLP. Notable efforts involve creating embedding atlases for journals (Peng et al., 2021), papers (Ganguly & Pudi, 2017), and concepts (Chen et al., 2020; Choi et al., 2016), as well as tracing the trajectories of research affiliations (Murray et al., 2023).

Given the abundance of scientific information available, navigating an efficient and logical knowledge pathway for learning purposes and acquiring knowledge is essential. Earlier navigation efforts have primarily focused on explicit knowledge management, effective sharing, or facilitating knowledge queries. However, these graphical knowledge maps are static and fail to incorporate embedding techniques. Current embeddings of scientific entities are primarily used for macro-level visualization and are seldom applied to knowledge navigation, particularly at the level of hierarchical scientific concepts. Moreover, all these methods often neglect the intrinsic knowledge dependencies embedded in the historical records of scientific publications, such as prerequisite relations (Manrique et al., 2019; Scheines et al., 2014) or precedence relations (Manrique et al., 2018; Xiao et al., 2021).

To address the limitations of the aforementioned research, we utilize concept data in the OpenAlex (Priem et al., 2022) dataset as meta-knowledge to construct millions of research trajectories (RT), viewing scientists' careers as comprehensive mentors enriched with prior knowledge. Building on this foundation, we develop the Scientific Concept Navigator (SciConNav), which leverages these extensive RT as a training corpus to learn the concept representations (Li et al., 2013), enabling fine-grained knowledge navigation and providing dynamic, personalized pathways. Our study addresses the following questions: (1) topic selection: How can researchers select a topic aligned with their expertise when transitioning to a new field? (2) interdependent learning pathways: How can researchers identify interdependent learning pathways when exploring unfamiliar topics across domains or investigating interdisciplinary subjects?

Due to the lack of directly comparable baselines for quantitative evaluation, Table A1 provides a qualitative comparison showcasing the advantages of SciConNav over existing approaches. By integrating concept embeddings with shortest-path algorithms, SciConNav identifies efficient knowledge pathways within large-scale scientific concept networks and offers personalized navigation recommendations. It reveals how knowledge connects across domains through bridging concepts, uncovering related topics and forming step-by-step learning pathways that reflect precedence and temporal dependencies across disciplines. These pathways enable efficient knowledge acquisition and reveal dependencies between related entities. SciConNav supports a variety of applications, including effective research or curriculum planning, adaptive learning guidance (Chiou et al., 2010; Díaz & Nussbaum, 2024), precise knowledge exploration, and the facilitation of scientific discovery (Yan et al., 2024). Ultimately, our model represents an initial exploration into the interconnected landscape of scientific concepts, empowering researchers to navigate complex knowledge spaces and strategically refine their research directions and career paths. Our key contributions are:

1. We learn the embedding space for knowledge navigation based on the RT filled with concepts of millions of scholars.
2. We demonstrate the utility of the embedding space for multi-step analogy inference of new concepts to enhance decision making and creative thinking.
3. We introduce *SciConNav* as a novel approach to navigating the knowledge inheritance pathways learned from extensive scientific RT.
4. We identify key concepts using centrality measures, highlighting the critical bridging role of interdisciplinary concepts in the global knowledge network.

2 | METHODS AND CONCEPTUAL VALIDATION

Our work focuses on three types of entities from the OpenAlex dataset: Authors, Works, and Concepts. Specifically, we utilize a total of 1,332,254 authors whose works exceed 50. Together, these authors comprise a total of 70,600,299 works, which collectively span 64,976 concepts. For each researcher, we construct concept sequences and train embeddings using curated sequences from millions of authors. We validate the effectiveness and similarity properties with the OpenAlex concept tree. Additionally, we analyze attribute roles and variance based on a specifically designed conceptual dimension to understand the semantic functionality of knowledge across disciplines.

2.1 | Knowledge entity and discipline concept classification

In this study, we utilize concepts[1] in the publicly accessible OpenAlex[2] dataset as knowledge entities, which are organized using a hierarchical tree structure. There are 19 root-level concepts (Level 0), each representing a distinct discipline and indexed with a unique number (1–19) on the left side of Figure 2. These root concepts further branch out into five levels of descendant concepts (Levels 1–5).[3] We initially classify each concept based on the number of root-level concepts it is associated in the concept hierarchy (see Figure B1) into three categories: Zero-root, Single-root, and Multi-root concepts. Zero-root concepts refer to those that have no connection to Level 0 concept in the hierarchy. Single-root concepts are associated with a single Level 0 concept as their root, while Multi-root concepts are connected with multiple Level 0 concepts. Leveraging the concept tree structure for discipline classification (as shown in Appendix B.2.2) faces challenges, especially when dealing with Zero-root concepts and cases where multiple roots share the same maximum path count. Thus we leverage five latest large language models (LLMs) for concept discipline annotation (shown in Appendix B.2.1), which can better capture the semantics and associations between concepts and disciplines.

Details of the annotation framework is shown in Figure B2, each concept is assigned a primary discipline concept (DC) from the 19 root-level concepts based on majority vote, thus the rest 18 root-level concepts are the non-discipline concepts (NDCs) of this concept. Concepts with full agreement (five votes) among five LLMs are further classified as "Mono-disciplinary" (Chen & Luetz, 2020) (or simply "Disciplinary"), as they align clearly with a single discipline. Concepts with less than full agreement, including near-full (four votes), moderate (three votes, see Figure B2), low (two votes), or no consensus, are classified as "Multi-disciplinary" (Chen & Luetz, 2020), reflecting ambiguity or overlap across multiple disciplines. We refer to "Mono-disciplinary" and "Multi-disciplinary" concepts as "Mono" and "Multi" concepts for simplicity, and denote them as C_s and C_m, respectively.

By adopting the LLM annotation method, we achieved 100% classification coverage, compared to 75.83% with the path quantity-based classification. This improvement ensures that all concepts are included in the analysis, avoiding biases caused by omissions and

providing more accurate discipline labels. The findings are more comprehensive and representative of the entire dataset, which could not be fully captured using the pathway quantity-based method. Additionally, the LLM annotation method introduces a confidence metric through majority voting, enabling better interpretation of classification reliability. Furthermore, the LLM annotation method enhances the identification of "Multi-disciplinary" concepts, which is essential for analyzing cross-disciplinary pathways.

2.2 | Extraction and learning of research trajectory

A research trajectory refers to the chronological sequence of scholarly outputs, primarily represented by a scholar's published papers, with each work tagged with multiple concept entities to categorize its key themes or topics. The methods for tagging these concepts are detailed in.[4] In academic research advancement, scientists often explore new research topics based on their previous background, revealing an underlying interdependent prerequisite or precedence relationship (Lu et al., 2019) between past and new topics. We highlight the temporal order and progression patterns among knowledge entities in the research trajectory. The prerequisite relationships (Tang et al., 2023) among concepts are crucial for clarifying knowledge causality and form the foundational basis for various educational applications, including personalized learning (Feldman-Maggor et al., 2024) and adaptive teaching strategies (Chen et al., 2018). They also guide the development of instructional rules for course design (Gasparetti et al., 2015) and curriculum planning (Sun et al., 2022), enhancing knowledge recommendation systems (He, Wang, Pan, et al., 2022). By leveraging these relationships, educators and researchers can create more effective and personalized learning experiences, ultimately improving educational outcomes.

We extract the RTs of millions of scholars using their historical publication records from the OpenAlex database. To ensure sufficient trajectory length and effective embedding, we focus on scholars with over 50 publications, resulting in a dataset of 1,332,254 authors. For each scholar, we construct an ordered concept list that capture the sequence of topics they explored over time. The RTs, enriched with semantic information and prerequisite relations (Scheines et al., 2014), provide essential context for understanding scholars' research interests and transitions across disciplines. By offering a solid foundation of prior knowledge or expertise (Backfisch et al., 2020), these trajectories enable the SciConNav model to recommend logical research pathways and facilitate effective knowledge navigation. Figure 1 illustrates this approach for author a_i, where all works are sorted by publication year. For each work $\mathbb{W}_t$, where $t \in [1,k]$, the labeled concepts are represented as $\mathbb{W}_t = \{C_i^t | i = 1,2,...,n_t\}$. These concepts are chronologically ordered into the research trajectory $\mathbb{A}_i = \{\mathbb{W}_1, \mathbb{W}_2 \cdots, \mathbb{W}_k\}$, effectively encoding the interdependent prerequisite relationships between concept entities. To facilitate knowledge navigation, we apply the Word2Vec model (Mikolov et al., 2013; Peng et al., 2021) to learn concept representations. This model trains vectors for each concept using the RT of millions of selected authors, represented as $\mathbb{RT} = \{\mathbb{A}_1, \mathbb{A}_2, \cdots, \mathbb{A}_N\}$ in Figure 1. While researchers often utilize higher dimensions (e.g., 100, 128, 256, or more), according to the evaluation in Gu et al. (2021), a dimension of 22 is small enough to be efficient, hence we use 22 as the lower bound. To determine the optimal dimensionality, we select six dimensions ranging from 24 to 128 (related to powers of 2) and evaluate their performance in terms of in-domain propensity (IDP) and cross-domain propensity (CDP), details can be seen in Appendix B.3. The experimental results presented in Figure B3a show that the dimension of 24 achieves the highest in-domain similarity. Additionally, we show that the sub-concepts of "Mathematics"

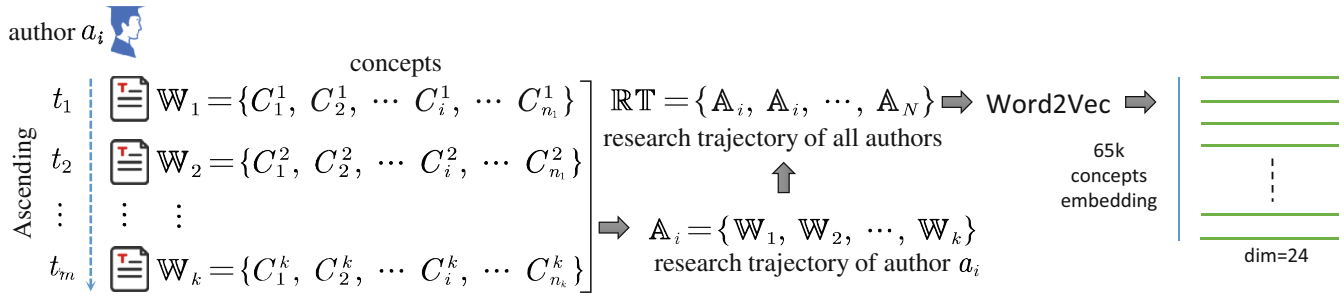


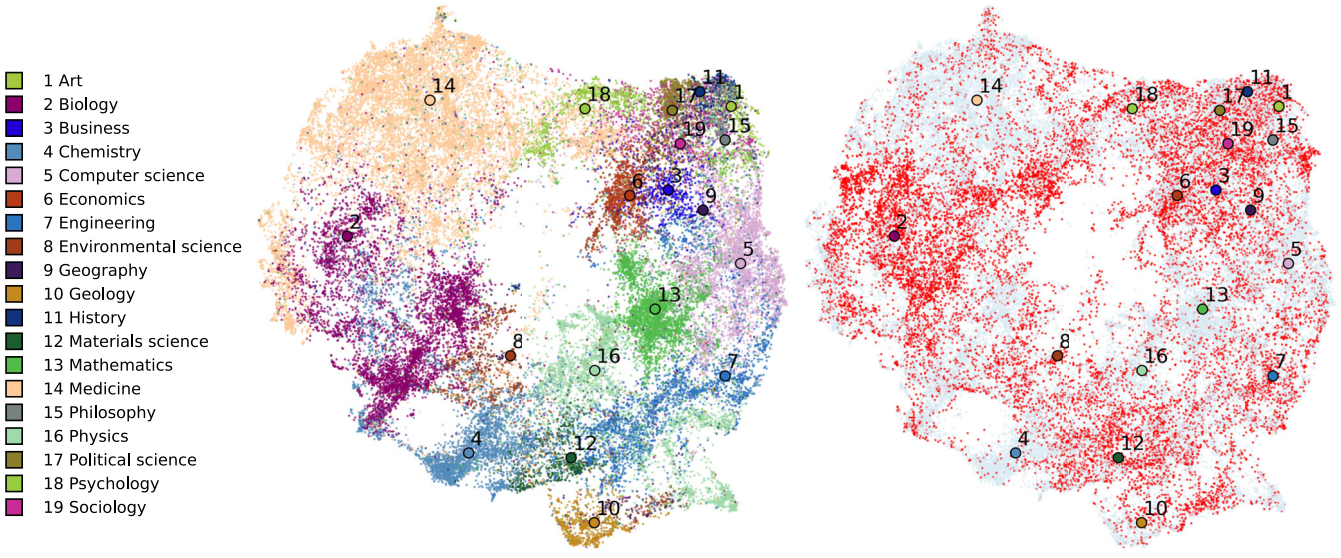
FIGURE 1 Concept representation learning from extensive research trajectories. For each author, the research concepts of each of his/her papers are sorted by the publication years. Then, we concatenate the concept list of all selected authors and obtain embedding vectors of all concepts by training concatenated concept lists with the Word2Vec model.

exhibit greater "non-relatedness" to less relevant disciplines as the dimensionality decreases, as illustrated in Figure B3b. The sub-concepts of "Mathematics" show higher similarity to "Mathematics" itself, while demonstrating lower similarity to disciplines such as "Medicine" and "Biology" as the dimensionality decreases. This suggests that lower-dimensional representations effectively distinguish between different disciplines and enhance similarity contrast. Our knowledge navigation application benefits from lower dimensionality, as it captures broad IDP while enhancing contrast across domains. Therefore, we select the dimension of 24 in experiments to prevent sparsity in high-dimensional spaces while preserving high semantics, balancing dimensionality with effectiveness.

To better visualize the structure of the learned embedding space, we use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) for dimensionality reduction. UMAP is a dimensionality reduction technique that preserves both local and global data structures by assuming the data lies on a low-dimensional manifold. It constructs a weighted graph to capture local relationships in high-dimensional space and optimizes its low-dimensional representation. Compared to t-distributed Stochastic Neighbor Embedding (t-SNE), UMAP is faster, scales better to large datasets, and preserves more global structure.

The 2D visualization of concept embedding with UMAP dimension reduction is shown in Figure 2. Specifically, the Figure 2a shows the "Mono" concepts (C_s), where each discipline is indexed with a number (1–19) and marked with a corresponding color, and the highest density point of each color is labeled with a number to indicate the discipline. The "Mono" concepts show an apparent spatial concentration within each discipline, while concepts from different disciplines occupying separate regions and exhibit distinct spatial modules. To highlight the spatial distribution differences, we color all "Multi" concepts in red and "Mono" concepts in cyan, as shown in Figure 2b. The "Multi" concepts in Figure 2b span across the entire embedding map and form a continuous spread with varying densities, unlike the distinct spatial concentration of "Mono" concepts in Figure 2a. To quantify this difference, we divide the embedding map into a 10×10 grid (see Figure B4a) for each discipline, and analyze the Bar Chart of Grid Counts (BCGC) as shown in Figure B4b. The BCGC analysis reveals distinct distribution patterns: "Mono" concepts from 19 disciplines show concentrated peaks and steep gradients, confirming their spatial clustering, while "Multi" concepts display gradual density variations, indicating a wider spatial dispersion.

We see that there are overlaps among related disciplines as shown in Figure 2a, particularly in the



(a) Concept of 19 distinct disciplines (b) Comparison with multi-disciplinary

FIGURE 2 Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction of concept representation. For each color-coded set of dots, the point with the highest density is circled in black, and the adjacent number indicates the corresponding discipline shown on the left side of the diagram. (a) Embedding map of "Mono-disciplinary" concepts from all disciplines with UMAP 2D dimension reduction, concepts with the same color belong to the same discipline, while each discipline is represented and sorted by a number from 1 to 19. (b) Overview of "Multi-disciplinary" concepts (red) and comparison with other "Disciplinary" concepts (cyan).

Humanities (1. "Art," 11. "History," and 15. "Philosophy") and Social Sciences (17. "Political Science" and 19. "Sociology"), which are closely distributed and intermixed in specific regions. These relations are verified in the mutual information matrix shown in Figure B4c; we show that the "Multi" cocnepts have low mutual information with the "Mono" concepts of 19 disciplines, and the "Multi" concepts are less interacted with "Mono" concepts. We also verify that the "Multi" concepts are generally more distant from their parent discipline compared with "Mono" concepts as shown in Figure B5b, which further supports the conclusion. We further verify that the "Mono" concepts in Figure 2a exhibit higher modularity compared with "Multi" concepts in Figure 2b. We construct "Mono" and "Multi" networks by connecting concept pairs whose similarities exceed specified percentile thresholds of the pairwise similarity distribution. The Modularity values are then computed for both networks across different thresholds. The results in Figure B5a show that the Modularity curve for "Mono" concepts is consistently and significantly higher than that of "Multi" concepts. Overall, the analysis of spatial distribution patterns reveals a fundamental difference in disciplinary embedding: "Mono" concepts exhibit strong discipline-specific clustering with high modularity, while "Multi" concepts show broader spatial dispersion and weaker disciplinary boundaries, suggesting their role as potential bridges across different academic domains.

2.3 | Discipline propensity validation of concepts

To verify the degree of propensity between a given concept c and its classified discipline d, denoting $\vec{i}$ as the embedding vector of concept i, we calculate the cosine similarity between concepts i and j as follows

$$\mathrm{sim}\left(\vec{i},\vec{j}\right)=\frac{\vec{i}\times\vec{j}}{|\vec{i}|\times|\vec{j}|}. \quad (1)$$

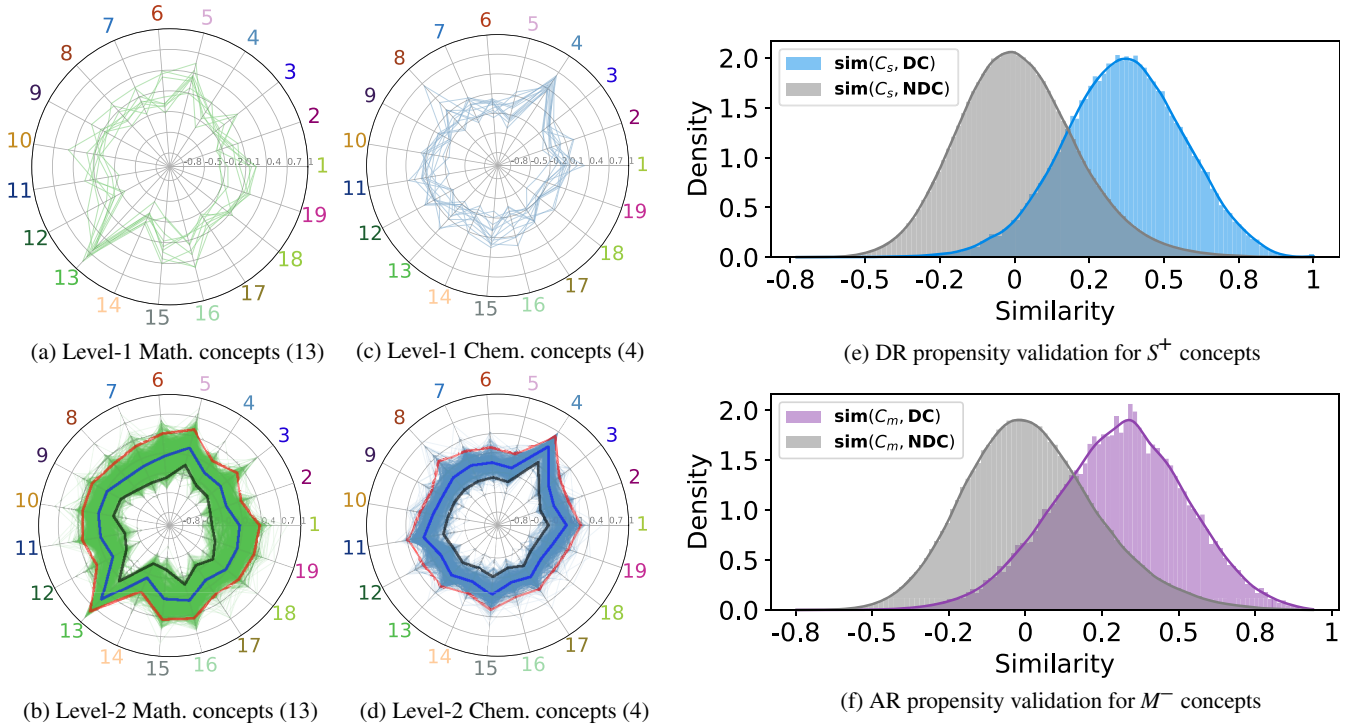
This metric will be denoted as s_{ij} in subsequent text for simplicity. The results of discipline propensity validation were quantitatively represented on a radar map, each one of the 19 polar axes is marked by a number symbolizing a discipline shown on the left side of Figure 2a. To simplify our analysis, we consider concepts in C_s (Mono), concepts in C_s (Multi) respectively. To validate the disciplinary propensity of all sub-concepts toward their DCs, we further divide the 19 root disciplines into two categories: DC versus NDC. Specifically, for each concept c, the set $\mathrm{DC}(c)$ consists of the Level 0 concept annotated as the DC of c, while the complementary set $\mathrm{NDC}(c)$ includes the remaining 18 disciplines that are NDCs of c.

For concepts in C_s, we verify the DC propensity of concepts in discipline "Mathematics." Its affiliated Levels 1 and 2 descendant concepts ($c \in C_s$) are expected to demonstrate significantly higher cosine similarity with "Mathematics" compared to the other 18 disciplines (NDC). We project the sub-concepts (Levels 1 and 2 concepts) of Mathematics onto each direction shown in Figure 3a,b respectively. The result coincides with the expectation, which is clearly illustrated by the most outward-pointing angle (MOPA) toward the discipline "Mathematics" (13) on the radar map. Both Level 1 concepts in Figure 3a and Level 2 concepts in Figure 3b exhibit an obvious MOPA toward its discipline "Mathematics," indicating a strong discipline propensity of descendant concepts to its labeled discipline "Mathematics" (DC) compared with the rest NDCs. We show another example of DC propensity validation for subconcepts (Levels 1 and 2) of "Chemistry" respectively shown in Figure 3c,d, both radar maps exhibit an obvious MOPA toward discipline "Chemistry" (DC) compared with NDCs, also supporting the conclusion. More examples of radar mappings for propensity validation of Levels 1 and 2 concepts across various disciplines are shown respectively in Figures C6 and C7 in Appendix C. The radar mapping for "Art" is excluded due to the limited number of sub-concepts.

To further verify the overall DC propensity for concepts in C_s and C_m across all disciplines collectively, rather than examining each discipline individually (Figure 3a–d). We calculate the set of similarities between concepts in C_s and their corresponding DC as

$$\mathrm{sim}(C_s,DC)=\left\{\mathrm{sim}\left(\vec{c},\vec{d}\right)|c\in C_s,d\in \mathrm{DC}(c)\right\}. \quad (2)$$

For simplicity, we refer to $\mathrm{sim}(C_s,\mathrm{DC})$ as the corresponding similarity distribution between concepts in C_s and their corresponding DC. Similarly, expressions such as $\mathrm{sim}(C_s,\mathrm{NDC})$, $\mathrm{sim}(C_m,\mathrm{DC})$ or $\mathrm{sim}(C_m,\mathrm{NDC})$ represent distributions of the corresponding sets of similarity values. For "Mono" concepts in C_s, we compare the distributions $\mathrm{sim}(C_s,\mathrm{DC})$ and $\mathrm{sim}(C_s,\mathrm{NDC})$ and observe a noticeable rightward shift of $\mathrm{sim}(C_s,\mathrm{DC})$ compared to $\mathrm{sim}(C_s,\mathrm{NDC})$. This shift indicates that C_s concepts have significantly higher similarity with their DC. We further validate the discipline propensity of "Multi" concepts in C_m, as shown in Figure 3f. The distribution $\mathrm{sim}(C_m,\mathrm{DC})$ is markedly shifted to the right compared to $\mathrm{sim}(C_m,\mathrm{NDC})$. This rightward shift pattern is also observed both in "Mono" concepts in C_s and "Multi"



(a) Level-1 Math. concepts (13)
(c) Level-1 Chem. concepts (4)
(e) DR propensity validation for S^+ concepts
(b) Level-2 Math. concepts (13)
(d) Level-2 Chem. concepts (4)
(f) AR propensity validation for M^- concepts

FIGURE 3 Propensity validation of concepts with discipline classification. (a) Level 1 Math. concepts (13). The vector similarity between Level 1 sub-concepts ($c \in C_s$) of "Mathematics" and the 19 root concepts. The disciplines are encoded by numbers listed in Figure 2a, thus values on each axis are the cosine similarities between sub-concepts and corresponding discipline. (b) Level 2 Math. concepts (13). The vector similarity between Level 2 sub-concepts ($c \in C_s$) of Mathematics and the 19 root concepts. The plot verifies that vectors of all math sub-concepts are semantically closer to the discipline "Mathematics." The 5-th, 50-th, and 95-th quantiles are marked with black, blue and red lines respectively from center to outside circle. (c) Level 1 Chem. concepts (4) and (d) Level 2 Chem. concepts (4) are another example of discipline propensity verification for Level 1 and Level 2 sub-concepts of discipline "Chemistry". (e) DC propensity validation for "Mono" concepts (C_s). Right shift of sim(C_s,DC) compared with **sim**(C_s,**NDC**) (f) DC propensity validation for "Multi" concepts (C_m). Right shift of sim(C_m,DC) compared with sim(C_m,NDC). Both plots in (e) and (f) exhibit a significantly higher similarity with their DC compared with their NDCs.

concepts in C_m, reinforcing the conclusion that concepts exhibit a strong propensity toward DC.

2.4 | Functionality projections of knowledge

We explore functional variance among 19 disciplines by calculating the cosine similarity between concept embeddings and specially designed conceptual axes. Analyzing the positioning of concepts along these axes reveals distributional shifts, providing insights into their diverse functional properties. These distributions highlight each discipline's unique characteristics and functional roles, showing how semantic functionality converges within and diverges across disciplines.

We define the functional groups based on key differences in disciplinary characteristics. For example, Theoretical disciplines (e.g., "Mathematics" and "Physics") emphasize abstract reasoning and fundamental principles, while Applied disciplines (e.g., "Computer Science" and "Engineering") prioritize innovative approaches to real-world challenges. Details of the eight discipline groups are provided in Table 1. For the theoretical concept group, we selected rigorously defined Level 1 sub-concepts from "Mathematics" and "Physics," along with their descendants in the concept tree, while explicitly excluding less theoretical concepts and their descendants. Table B5 provides detailed information on the selected and excluded Level 1 concepts. The axes are derived by subtracting the average embedding vectors of two discipline groups, highlighting differences in functionality and semantic roles. Let G denote a functional group of concepts, and let $V_G = (1/|G|)\sum_{c \in G} \vec{c}$ denote its average embedding. The functional axis (FA) from group G_1 to G_2 is then given by $\mathrm{FA}_{G_1 \to G_2} = V_{G_2} - V_{G_1}$.

The results for the two selected axes are presented in Figure 4. Specifically, the axis $\mathrm{FA}_{\mathrm{Theoretical} \to \mathrm{Applied}}$ is constructed to differentiate disciplines based on their tendency toward theoretical exploration or practical

TABLE 1 Discipline group partition of 19 disciplines.

Functional group	Disciplines	Functional group	Disciplines
Theoretical	Mathematics, Physics	Societal	Sociology, Political science, Psychology
Applied	Computer science, Engineering	Economic	Economics, Business
Chemical	Chemistry, Materials science	Humanities	Philosophy, History, Art
Biomedical	Biology, Medicine	Geographical	Geography, Geology, Environmental science

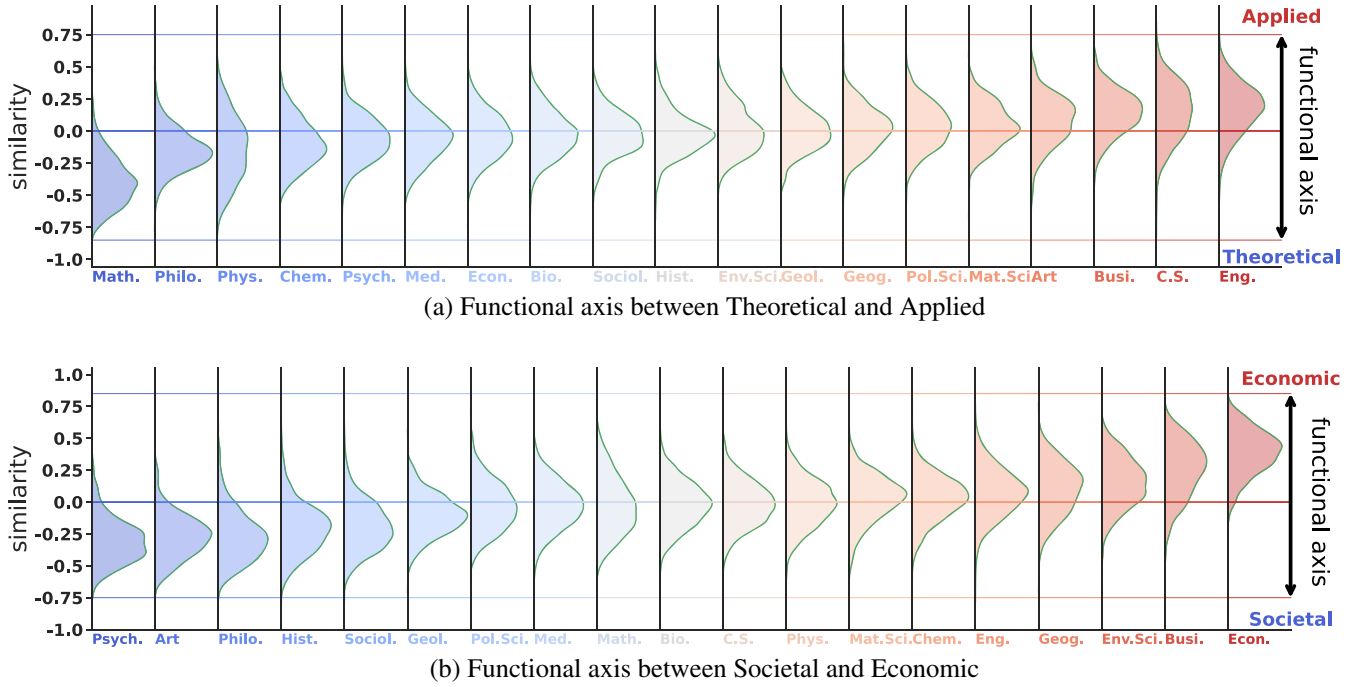


FIGURE 4 Functionality projections of knowledge from 19 disciplines on predefined axes. To avoid overlap caused by lengthy terms, we use abbreviations for discipline categories listed in Table C6 in Appendix C. (a) Functional axis between Theoretical and Applied. Projection of all disciplines on the functional axis from Theoretical (Mathematics and Physics) to Applied (Engineering and Computer science). (b) Functional axis between Societal and Economic. Projection of all disciplines on the functional axis from Societal to Economic.

application. As shown in Figure 4a, the probability density functions (PDFs) of all disciplines are projected along this axis. Concepts from "Mathematics" are located at the theoretical end, while those from "Computer science" positioned at the applied end. Concepts from other disciplines occupy intermediate positions, reflecting varying levels of theoretical or applied characteristics. The projection results on axis $FA_{Societal \rightarrow Economic}$ are shown in Figure 4b, which highlights how disciplines such as "Psychology," "Art," "Philosophy," and "History" align more with Societal sciences, while fields like "Environmental science," "Economics," and "Business" are closer to Economic sciences. Two additional examples of functional projections along the axes $FA_{Chemical \rightarrow Biomedical}$ and $FA_{Humanities \rightarrow Geographical}$ are presented in Figure C8. Collectively, these axes offer a comprehensive perspective on the functional trends and relationships among concepts across 19 academic disciplines, while also validating their effectiveness in distinguishing between disciplines.

Overall, we examine and verify the embedding representations of knowledge from various disciplines and analyze the semantic functionality along the crafted conceptual dimensions. This provides an intuitive understanding of the similarities and differences in domain knowledge across various fields. In the next section, we utilize the semantics of knowledge to demonstrate meaningful applications, such as analogy inference and navigation.

3 | RESULTS

In this section, we propose the SciConNav, and demonstrate its impactful applications in analogy inference and knowledge navigation. SciConNav addresses challenges

in selecting suitable research topics and inferring interdependent learning pathways. Firstly, we employ multi-step analogy inference to explore analogical concepts along predefined axes, resulting in meaningful inference graphs. Secondly, we conduct global knowledge navigation and examine the accessibility between knowledge domains. Lastly, we highlight the critical role of interdisciplinary concepts in the global knowledge network.

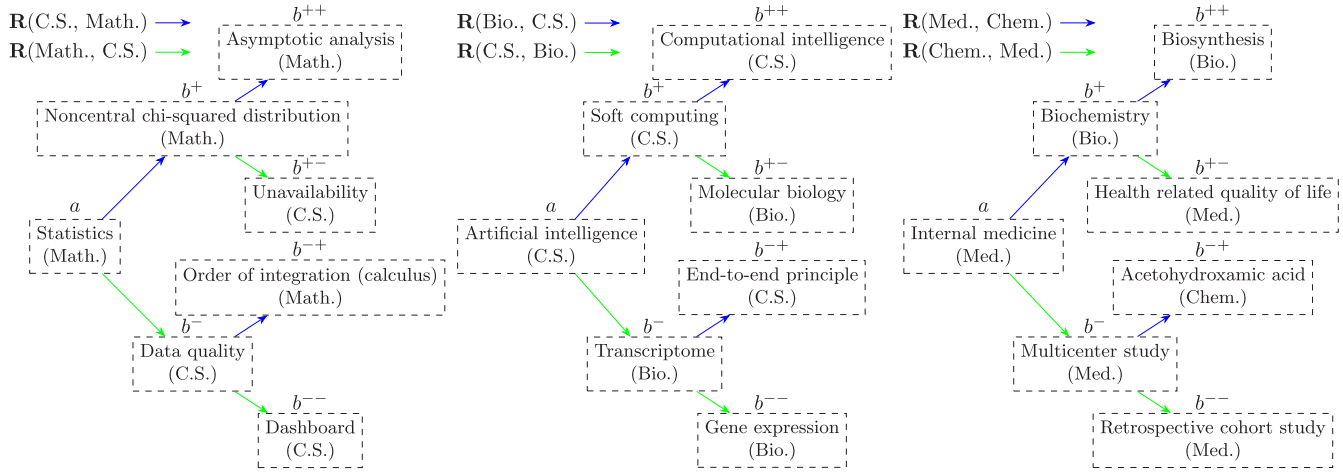
3.1 | Analogies of interconnected knowledge

Concept analogy inference derives related concepts from seed concepts with functional or causal relations. The SciConNav model strengthens analogical reasoning between scientific concepts, enabling researchers to uncover interconnected knowledge and identify highly relevant topics aligned with their research background. The analogy is expressed in its general form as "a is to b as c is to d" (Drozd et al., 2016), a well-known illustration is the analogy "king is to? as man is to woman" (Ethayarajh et al., 2018). The word analogies fundamentally capture the shared relational patterns between pairs of words. Let $R(a,b)$ represents the semantic relationship between two words a and b. Hence, the relational analogy between two pairs of words can be expressed as $R(\text{man},\text{woman}) \sim R(\text{king},\text{queen})$, which captures the same pattern of Gender Difference: man and woman represent this relationship in the context of general human identity, while king and queen reflect it within the realm of royal hierarchy.

Let $\vec{c}$ denote the embedding vector for concept c, thus the linear relation between two words c and d can be represented as a relation vector in the embedding space, defined by $R(c,d) = \vec{d} - \vec{c}$, which captures the semantic difference between the two concepts. The analogical reasoning for new concepts can be solved through vector arithmetic, exploring both positive and negative analogical relations that

$$b^{\pm} = \arg\max_{b} \left\{ \text{sim}\left(\vec{b}, \vec{a} \pm R(c,d)\right) \right\}, \tag{3}$$

where b^+ and b^- are obtained through analogy to the positive and negative relations respectively. We illustrate the two-step concept analogy inference in Figure 5, where new concepts are derived step by step, selecting newly inferred concepts as seed concepts at each step. The analogy relations of Figure 5a are positively $R(\text{C.S.},\text{Math.})$ and negatively $R(\text{Math.},\text{C.S.})$. From the seed concept "Statistics," with the positive relation, "Noncentral chi-squared distribution" is inferred. Starting from "Noncentral chi-squared distribution," "Asymptotic analysis" and "Unavailability" are derived in positive and negative relations. The "Data quality" is obtained from seed concept "Statistics" with the negative relation, and then starting from "Data quality," "Order of integration (calculus)" is inferred in the positive relation, while "Dashboard" is derived in the negative relation. In Figure 5b, the analogy relations are positively



(a) Analogy inference 1 (b) Analogy inference 2 (c) Analogy inference 3

FIGURE 5 Examples of two-step analogy inference. (a) Analogy inference 1. The analogy relations exist between "C.S." and "Math.," while the seed concept is "Statistics." (b) Analogy inference 2. The analogy relations lie between "Bio." and "C.S.," with the seed concept being "Artificial intelligence." (c) Analogy inference 3. The analogy relations reside between "Med." and "Chem.," the seed concept is "Internal medicine."

TABLE 2 Interpretations of the analogy examples.

Analogy example	Interpretation
$R(\text{Statistics}, \text{Data quality}) \sim R(\text{Math.}, \text{C.S.})$	"Data quality" is an application area of "Statistics."
$R(\text{Data quality}, \text{Dashboard}) \sim R(\text{Math.}, \text{C.S.})$	"Dashboard" is a specific implementation of "Data quality."
$R(\text{Artificial intelligence}, \text{Transcriptome}) \sim R(\text{C.S.}, \text{Bio.})$	AI (a tool from "C.S.") supports and is applied to the study of Transcriptomes (an application in "Bio.").
$R(\text{Soft computing}, \text{Molecular biology}) \sim R(\text{C.S.}, \text{Bio.})$	Soft computing is a computational tool applied to Molecular biology akin to how "C.S." is applied to "Bio."
$R(\text{Internal medicine}, \text{Biochemistry}) \sim R(\text{Med.}, \text{Chem.})$	Internal medicine heavily relies on Biochemistry to understand diseases at a molecular level, just as "Med." relies on "Chem." for drug design and physiological understanding.
$R(\text{Biochemistry}, \text{Biosynthesis}) \sim R(\text{Med.}, \text{Chem.})$	Biosynthesis (e.g., synthesis of biomolecules) is a direct application of biochemical principles, much like Medicine applies Chemistry.
$R(\text{Internal medicine}, \text{Multicenter study}) \sim R(\text{Chem.}, \text{Med.})$	Internal medicine often identifies clinical questions requiring large-scale studies, similar to how medicine shapes chemistry's focus.

$R(\text{Bio.}, \text{C.S.})$ and negatively $R(\text{C.S.}, \text{Bio.})$, with the seed concept being "Artificial intelligence." Similarly, in Figure 5c, the analogy relations are positively $R(\text{Med.}, \text{Chem.})$ and negatively $R(\text{Chem.}, \text{Med.})$, the seed concept is "Internal medicine."

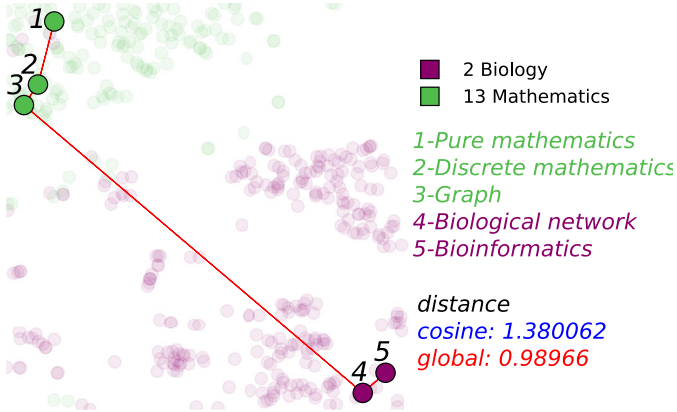
We provide interpretations of the analogy examples, as shown in Table 2. While the relationships of concept analogies may not present easily interpretable as $R(\text{man}, \text{woman}) \sim R(\text{king}, \text{queen})$, they still exhibit clear transitions between concepts, reflecting meaningful relations in scientific connections. In our example in Figure 5a, we can model the "C.S." as the application of "Math.," and "Math." as the foundational tool of "C.S." We see that $R(\text{Statistics}, \text{Data quality}) \sim R(\text{Math.}, \text{C.S.})$ because "Data quality" is an application area of "Statistics," and $R(\text{Data quality}, \text{Dashboard}) \sim R(\text{Math.}, \text{C.S.})$ for "Dashboard" is a specific implementation of "Data quality." Still in Figure 5b, $R(\text{Artificial intelligence}, \text{Transcriptome}) \sim R(\text{C.S.}, \text{Bio.})$ because AI (a tool from "C.S.") supports and is applied to the study of Transcriptomes (an application in "Bio."). And in Figure 5c, $R(\text{Biochemistry}, \text{Biosynthesis}) \sim R(\text{Med.}, \text{Chem.})$ because Biochemistry is fundamental to understanding and driving Biosynthesis processes, just as "Chem." underpins "Med."

The examples above demonstrate that conceptual analogical reasoning improves knowledge understanding by revealing semantic similarities, taxonomic hierarchies, and causal mechanisms among knowledge concepts. Furthermore, the inferred concepts can provide valuable insights for knowledge expansion, inspire creative thinking in the learning process, and cross-disciplinary research. Overall, this process drives innovation in applications such as information retrieval and question-answering systems by leveraging analogical reasoning to uncover deeper conceptual analogical connections. Additional examples of two-step analogy inference are provided in Table C7.

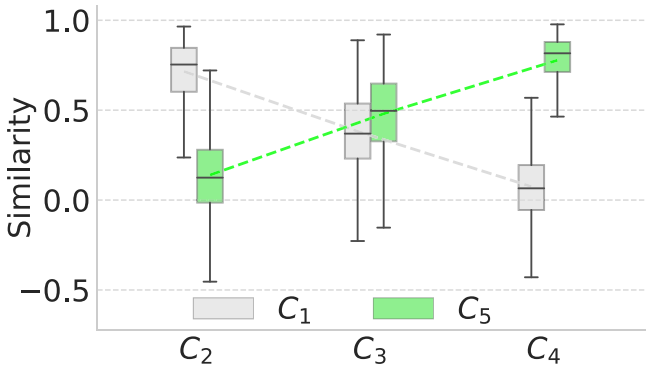
3.2 | Global knowledge navigation and network accessibility

The concepts of analogy inference and functionality inspire us to investigate global connections across all disciplines through the shortest path, extending beyond the typical focus on local similarity through analogy inference. This approach addresses the challenge of identifying interdependent learning pathways between selected topics. SciConNav constructs a fully connected network using the top n (=20,000) representative concepts with the highest number of related papers, ensuring that each selected concept has at least 3410 associated works. The cosine distance $w_{ij} = 1 - s_{ij}$ between concept i and j serves as the edge weight. We calculate the pairwise weighted shortest path (WSP) for each pair of concepts, record the corresponding shortest distance (SD) d_{ij} and path length of the WSP between two concepts i and j, and further analyze the SD distribution from "Mathematics" to all disciplines. The results of global knowledge navigation are shown in Figure 6.

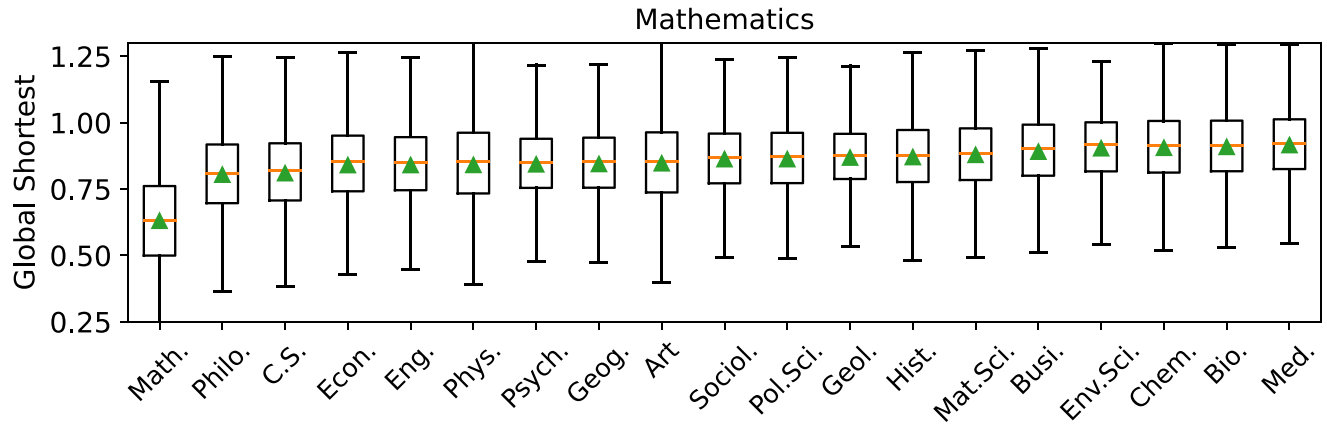
To demonstrate the practical utility of SciConNav, we analyzed the trajectory of a researcher transitioning from "Pure mathematics" to "Bioinformatics" in Figure 6a. The model identifies "Discrete mathematics," "Graph" and "Biological network" as bridging concepts, providing a logical pathway for interdisciplinary exploration, highlighting its ability to guide efficient cross-disciplinary transitions. The global SD along this path is 0.99, calculated as the sum of cosine distances between adjacent



(a) Global navigation between Pure Math. and Bioinformatics



(b) Validation of pathway effectiveness



(c) Knowledge accessibility of concepts form Mathematics to 19 disciplines

FIGURE 6 Global knowledge navigation and network accessibility. (a) Global navigation between Pure Math. and Bioinformatics. Illustrates the shortest path from "Pure mathematics" to "Bioinformatics," highlighting the cross-disciplinary connections. (b) Validation of pathway effectiveness. Distribution of similarities between middle concepts and endpoint concepts. The box plots show the similarity distribution between bridging concepts (C_2, C_3 and C_4) and two endpoints (C_1 and C_5), while the dashed lines connect the mean values to illustrate the overall trend. (c) Knowledge accessibility of concepts form Mathematics to 19 disciplines. Distributions of shortest path distances from the sub-concepts of "Mathematics" to the sub-concepts of 19 disciplines.

concepts along the path, and is shorter than the direct cosine distance of 1.38 between the two endpoint concepts. This reduction is achieved by linking multiple globally relevant concepts along the path. The WSP not only shortens the distance between two concepts but also highlights the importance of knowledge connections in academic exploration and discovery across diverse disciplines. Additionally, it indicates prerequisite correlations, suggesting that concepts along these paths reflect historical relationships in knowledge advancement.

To validate the overall effectiveness and semantic coherence of the obtained knowledge paths, we analyze all five-concept paths $(C_1, \cdots, C_5)$ by examining the similarities between the bridging concepts (C_2, C_3, and C_4) and two endpoints (C_1 and C_5). As shown in Figure 6b, the similarity distributions reveal a clear pattern: with respect to the first concept C_1, the average similarities decrease along the path ($C_2 > C_3 > C_4$), while for the last concept C_5, the average similarities increase ($C_2 < C_3 < C_4$). This systematic pattern demonstrates a gradual semantic transition from the first concept to the last concept. Notably, the same trend is observed when the path is reversed, further confirming the robustness and reliability of the knowledge paths.

We calculate the path length distribution (PLD) of paths from sub-concepts of each discipline to all concepts, the details provided in Appendix C.4 and the results are shown in Figure C9. The first column, labeled "All" shows the PLD for paths between all concept pairs, while the remaining columns represent distributions for paths starting from sub-concepts of each discipline and reaching all concepts. Notably, 99% of WSPs across

disciplines are shorter than 6, with most path lengths being 3 or 4, as shown in Figure C9, indicating that only one or two intermediate steps are typically needed to connect two concepts.

We analyze the shortest path distances from the concepts of one discipline to others. Taking "Mathematics" as an example in Figure 6c, we examine the distribution of shortest path distances from "Mathematics" concepts to those in 19 disciplines. The SDs are observed within "Mathematics" itself, followed by "Philosophy" and "Computer Science." This indicates that "Mathematics" concepts are semantically closer to "Philosophy" and "Computer Science" in scientific research. While the distributions of shortest path distances to other disciplines are similar, their mean distances show subtle but discernible differences, with "Biology" and "Medicine" having the greatest mean distance from "Mathematics."

By mapping these conceptual pathways, we offer a structured approach to navigating the complex web of scientific knowledge. Our framework provides a logical, step-by-step learning trajectory across diverse fields, focusing on global shortest paths between concepts. Knowledge navigation explores how researchers and learners can efficiently engage with new topics incrementally through interdependent pathways, offering vital insights into the interconnected nature of knowledge progression. The identified pathways guide strategic research planning, facilitate interdisciplinary collaboration, and promote innovation. Our approach reveals efficient paths for knowledge acquisition and highlights potential fields for cross-disciplinary breakthroughs. This enables researchers to expand their expertise, discover novel research directions, and advance science beyond traditional boundaries. To deepen our understanding of the structural foundations of such transitions, we analyze the concepts that serve as crucial bridges between knowledge pathways in the following section.

3.3 | Analysis of bridging concepts in knowledge pathways

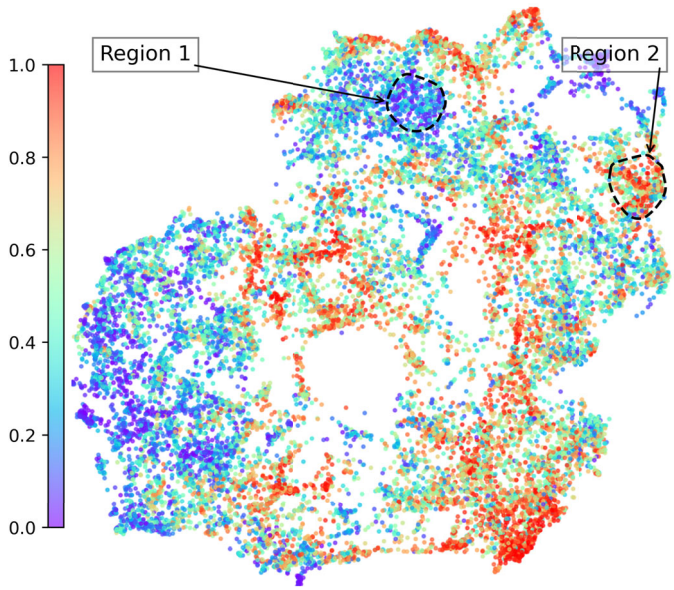
3.3.1 | Connectivity patterns with contrasting centralities

We define centrality as the accessibility of a node, which indicates the significance of a concept. In our global knowledge network among selected 20,000 concepts, we focus on two types of centralities: closeness and betweenness. Closeness centrality is calculated as the inverse of the average SD from the focal to other concepts, which measures the proximity of a concept to all others, indicating its efficiency in interacting with a wide range of concepts across different fields. Betweenness centrality is calculated as the fraction of the shortest paths that pass through the focal concept, which measures a concept's role in information transfer, acting as the bridge between different clusters or disciplines. We denote H as the set of selected concepts with high centrality and L as the set of selected low centrality concepts. We further visually highlight concepts with high closeness centrality in red on the embedding maps in Figure 7. Our analysis reveals a distinct pattern that concepts with high closeness centrality tend to cluster within specific fields shown in Figure 7a, in contrast to the concepts with high betweenness centrality, which are more widely distributed across various disciplines (as illustrated in Figure C13 in Appendix C).

To elucidate the varying centrality patterns, we select two representative regions with contrasting centralities, showcasing the concentration of main concepts through word clouds in Figure 7b. Our findings indicate that L concepts in Region 1 with low accessibility, primarily from natural sciences and technological domains, are specialized within their fields. These include fundamental physics ("Spin," "Quantum," and "Magnetic"), materials science ("Silicon"), energy technology ("Solar"), and biological fundamentals ("Cell"). In contrast, H concepts in Region 2, primarily fundamental mathematical and theoretical constructs, serve a bridging role, aligning with the foundational role of basic sciences like mathematics in real-world research. This includes spatial concepts, functional analysis, algebraic structures ("Polynomials" and "Groups"), matrix theory, and advanced methods like differential equations. These cornerstone H concepts exhibit high centrality due to their widespread applications and connections across various scientific disciplines, crucial for modeling complex systems, underpinning numerous scientific theories, and addressing practical challenges. As illustrated in Figure C10, the figure primarily presents the top 1%–20% of H concepts ranked by closeness centrality, while the left panel specifically shows the overall proportions for each discipline as a baseline. These high-centrality concepts span 19 disciplines, with proportions stabilizing as the ratio increases. Figure C11 illustrates the word clouds for the top 1%, 2%, 3%, 4%, 6%, and 10% of high closeness centrality concepts, while the right panel shows the corresponding concept level distribution.

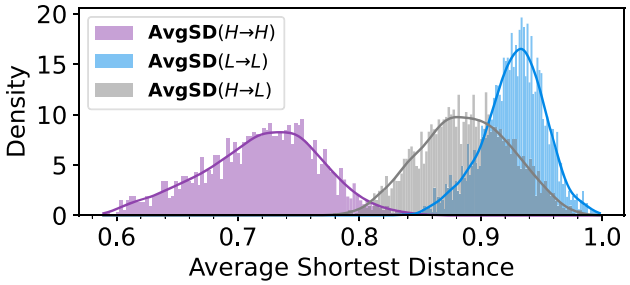
We selected top 10% high closeness centrality concepts (2000 H concepts) for SD analysis, and calculate the set of SDs for all pairs of concepts within H as

$$\mathrm{SD}(H,H) = \left\{ d_{ij} | i \neq j, i \in H, j \in H \right\}. \quad (4)$$



(a) Heatmap of concept closeness centrality

(b) Concept word clouds of two selected small regions



(c) AvgSD between 2k H and 2k L concepts by closeness

FIGURE 7 Centrality measure the importance of interdisciplinary concepts. (a) Heatmap of concept closeness centrality. Embedding map of 20,000 key concepts, color-coded by closeness centrality. (b) Concept word clouds of two selected small regions, Region 1 mainly contains concepts with low closeness centrality, and Region 2 mainly contains concepts with high closeness centrality. (c) Average shortest distance (AvgSD) between 2k H and 2k L concepts by closeness.

For simplicity, we use the notation $\mathrm{SD}(H,H)$ to represent the distribution of SDs for all pairs of concepts within H, and $\mathrm{SD}(L,L)$ represents the same for L. The results in Figure C12 show that concepts within H have shorter SDs than those in L, as demonstrated for closeness in Figure C12a and for betweenness in Figure C12b. To analyze how tightly connected these concepts are within or across H and L concepts, we measure the average distance within and cross groups: $L \to L, H \to H$, and $H \to L$. We define the average shortest distance (AvgSD) of a concept i as:

$$\mathrm{AvgSD}(i) = \frac{1}{|R(i)|} \sum_{j \in R(i)} d_{ij}, \tag{5}$$

where $R(i)$ is the set of concepts reachable from i.

We denote the set of AvgSDs from H concepts to L concepts as

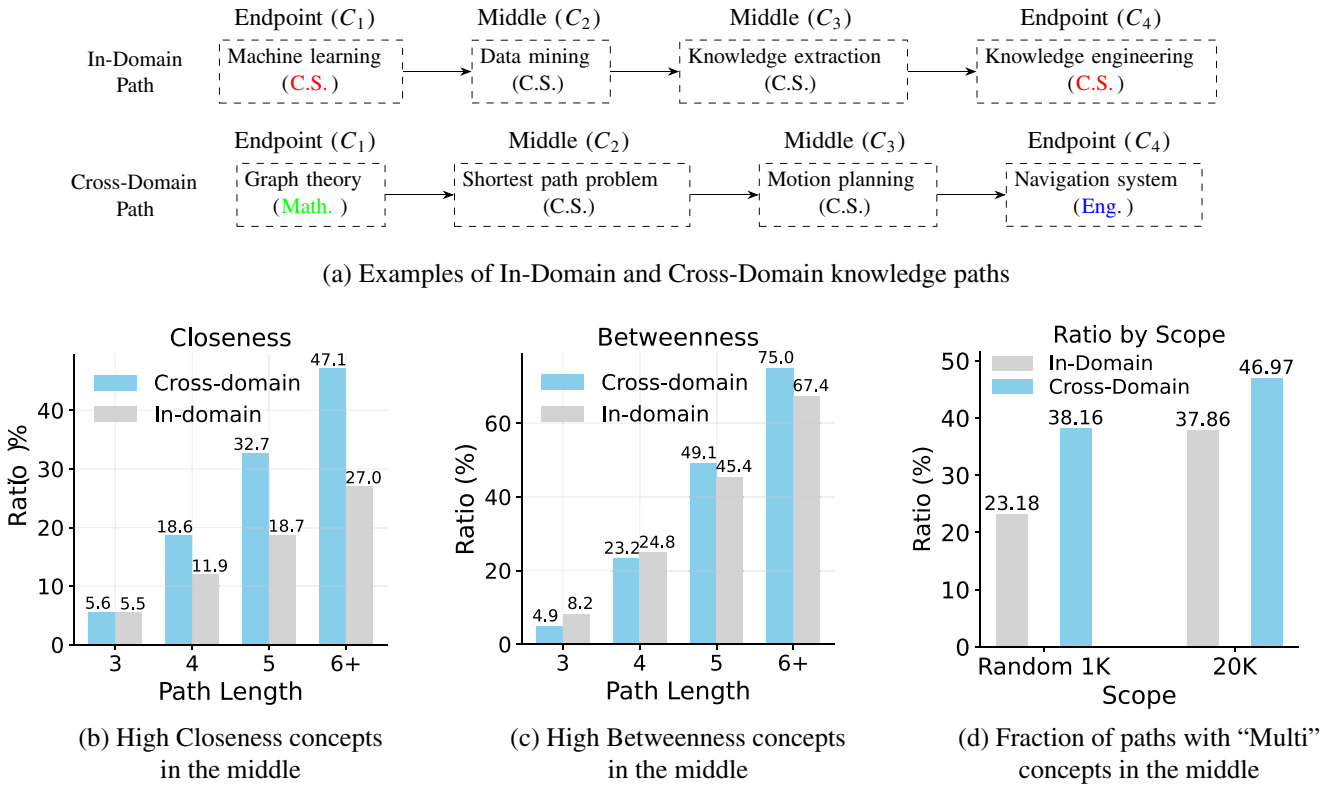
$$\mathrm{AvgSD}(H \to L) = \{\mathrm{AvgSD}(i) | i \in H, R(i) = L\}. \tag{6}$$

For simplicity, we use the set notation $\mathrm{AvgSD}(H \to L)$ to denote the corresponding distribution. Similarly, $\mathrm{AvgSD}(H \to H)$ represents the distribution of concepts in H reaching all nodes in H, and $\mathrm{AvgSD}(L \to L)$ for concepts in L.

The results of the AvgSD analysis of within and cross H and L concepts are shown in Figure 7c, aligning with results from two selected smaller regions shown through word clouds in Figure 7b. We observe the leftward position of $\mathrm{AvgSD}(H \to H)$, which indicates shorter average semantic distances among high-centrality concepts, suggesting dense interconnectivity in fundamental mathematical and theoretical constructs. On the other hand, $\mathrm{AvgSD}(L \to L)$ at the right end indicates larger average semantic distances among low-centrality concepts, suggesting less interconnectivity in specialized domains like natural sciences and technologies. Finally, the central position of $\mathrm{AvgSD}(H \to L)$ between $\mathrm{AvgSD}(H \to H)$ and $\mathrm{AvgSD}(L \to L)$ confirms that high-centrality concepts act as bridges, facilitating interdisciplinary connections and knowledge integration.

3.3.2 | Bridging role validation of high-centrality concepts

To validate that H concepts act as bridges and enhance cross-domain connections, we analyze pairwise paths generated from 20,000 selected concepts, and we calculate the proportions of In-Domain and Cross-Domain paths where H concepts serve as bridges. As illustrated in Figure 8a, In-Domain paths connect two endpoint concepts in the same discipline, while Cross-Domain paths link endpoints from different disciplines. Bridging nodes are the intermediate concepts along the path, excluding



(a) Examples of In-Domain and Cross-Domain knowledge paths

(b) High Closeness concepts in the middle

(c) High Betweenness concepts in the middle

(d) Fraction of paths with "Multi" concepts in the middle

FIGURE 8 Validation of the bridging role of H concepts and "Multi" concepts, with an increased proportion in Cross-Domain knowledge paths. (a) Examples of In-Domain and Cross-Domain knowledge paths. (b) High Closeness concepts in the middle. (c) High Betweeness concepts in the middle. (d) Fraction of paths with "Multi" concepts in the middle.

the two endpoints, while H concepts are defined as those ranking among the top 5% in centrality. Figure 8b shows that high closeness concepts (H concepts) more frequently act as bridges in Cross-Domain paths than In-Domain paths. The proportion of paths where H concepts serve as bridging nodes increases as path length extends from 3 to 6+. Compared to In-Domain paths, Cross-Domain paths show a steeper rise in the proportion where H concepts serve as bridges as path length grows, confirming that H concepts enhance cross-domain connections. Similarly, Figure 8c reveals that high betweenness concepts also demonstrate enhanced bridging roles in Cross-Domain paths as path length increases.

We also verify that "Multi" concepts (C_m) serve as bridging roles, as shown in Figure 8d. Among the pairwise paths of 1000 randomly selected concepts with a discipline distribution matching the overall ratio, the results indicate that "Multi" concepts serve as bridges in 38.16% of Cross-Domain paths, significantly higher than the 23.18% in In-Domain paths. This trend is consistent when analyzing all 20,000 concepts, where "Multi" concepts serve as bridges in 46.97% of Cross-Domain paths, compared to 37.86% in In-Domain paths. This suggests that "Multi" concepts also facilitate connections across diverse disciplines.

The results above shown in Section 3 demonstrate how our framework achieves two core research objectives. For the first objective of enabling analogical reasoning, Figure 5 highlight meaningful conceptual analogies across disciplines, showcasing how researchers can systematically uncover related concept through step-by-step analogy. Additionally, we provided detailed explanations and interpretations to examine the validity of these analogies. For the second objective of facilitating knowledge pathway navigation, we provide an example of effective knowledge navigation in Figure 6a and subsequently conduct a series of analyses to quantify the efficiency of the knowledge path, evaluate the underlying knowledge structure, and analyze its bridging role in the knowledge path. The PLD shown in Figure C9 indicates that majority of path length are less than 6, reflecting a high degree of interconnectedness within the knowledge network. Figure 6c quantifies the SD of knowledge pathways from "Mathematics" to other disciplines, revealing the relative semantic proximity within the knowledge structure. We further validates the effectiveness of the global knowledge navigation approach by verifying the five-concept paths. Figure 7 highlights concepts with contrasting centralities, enabling the identification of bridging concepts and their distinct connectivity strength in facilitating

knowledge connections. The results in Figure 8 validate the bridging role of high-centrality concepts, with a notably higher proportion of H concepts in Cross-Domain paths. We also verify that "Multi" concepts appear more frequently in Cross-Domain paths (38.16%) compared to In-Domain paths (23.18%).

Overall, our research aims to propose the SciConNav for identifying knowledge analogies and pathways. The results validate: (1) the effectiveness of concept analogies and knowledge pathways in connecting diverse domains, (2) the high degree of interconnectivity within the knowledge structure, and (3) the crucial role of high-centrality concepts in facilitating cross-domain connections. Collectively, these findings confirm the ability of SciConNav to enable cross-disciplinary analogical reasoning and facilitate efficient knowledge navigation.

4 | DISCUSSION

The SciConNav model utilizes Word2Vec to learn concept representations from the RT of millions of scholars, integrating concept embeddings with network-based methods to infer concept analogies and knowledge pathways. By modeling author trajectories enriched with topic dependencies and potential prerequisite relationships, the SciConNav model uncovers common patterns in academic careers and serves as a practical tool for both researchers and learners. This model maps dependency connections across 19 disciplines, tracking the evolution of scientific knowledge and emerging technologies. It empowers users to identify promising research directions, explore new topics, and plan efficient transition paths into emerging fields, while also offering insights into potential interdisciplinary collaborations. By capturing typical career transitions, SciConNav recommends logical pathways to support informed decision-making. The resulting dependency network accelerates scientific discovery, enhances knowledge exploration, and fosters interdisciplinary collaboration, offering guidance and valuable resources for early-career scientists.

Our work analyze several density figures to uncover patterns and variations. First, we calculate $\mathrm{sim}(C_s, \mathrm{DC})$ and observe a noticeable rightward shift compared to $\mathrm{sim}(C_s, \mathrm{NDC})$, indicating higher similarity between C_s concepts and their DC. A similar pattern is observed for "Multi" concepts in C_m, where $\mathrm{sim}(C_m, \mathrm{DC})$ also shifts markedly to the right, highlighting the strong discipline propensity of concepts. Additionally, we analyze the semantic distances to reveal the connectivity patterns. The leftward shift of $\mathrm{AvgSD}(H \to H)$ indicates dense interconnectivity among high-centrality concepts, the rightward shift of $\mathrm{AvgSD}(L \to L)$ reflects weaker connectivity among low-centrality concepts, and the intermediate position of $\mathrm{AvgSD}(H \to L)$ suggests moderate connectivity. These findings highlight the intricate relationships between concepts, their structural organization, and functional alignment, providing valuable insights into the underlying structure, dependencies, and semantics of academic disciplines.

The SciConNav model with conceptual knowledge space holds significant values in several aspects. First, it provides scientists with a comprehensive overview of scientific knowledge. By projecting concepts onto the predefined functional axes, such as $\mathrm{FA}_{\mathrm{Theoretical}} \to \mathrm{FA}_{\mathrm{Applied}}$, we provide insights into the theoretical and applied aspects of concepts, demonstrating the effectiveness of these axes in distinguishing disciplines and highlighting their distinct positional alignments and functional differences. Additionally, the concept analogy inference, undertaken through a step-by-step local approach, allows for the exploration of cross-disciplinary knowledge and the drawing of analogical relations from diverse fields. This enhances logical reasoning, supports decision-making (Schulz, 2023), and fosters creative thinking (Wegerif et al., 2010; Xiong et al., 2022). Moreover, the global connections through the shortest cosine distance paths navigate the research progression, offering a customized learning pathway from the current knowledge base to desired knowledge, particularly for researchers less familiar with certain areas. By investigating the global accessibility of concepts using centrality measures, we can apprehend the connectivity across disciplines and identify key concepts that highlight the bridging role of interdisciplinary concepts.

Despite these contributions, our study has limitations. The dataset comprises 64,976 concepts, which are validated by Wiki and could be expanded for greater diversity and depth. Advanced language models could offer richer semantic encoding, and more sophisticated network methods could enhance cross-domain knowledge inference. Additionally, the high dimensionality of the embedding space presents challenges in interpreting the knowledge structure, suggesting a need for more nuanced analytical tools to unpack the black box.

In conclusion, while there is room for improvement and further exploration, our conceptual SciConNav model represents a significant advancement in knowledge navigation and scientific discovery, which informs the decision-making processes, promoting more effective collaboration and investment strategies in the evolving landscape of scientific research and

education. Future work may incorporate advanced language models and interpretable frameworks, enabling more refined tasks such as causal knowledge reasoning, structured knowledge retrieval, and personalized academic exploration.

ENDNOTES

[1] https://docs.openalex.org/api-entities/concepts.

[2] https://docs.openalex.org/download-all-data/download-to-your-machine.

[3] https://docs.google.com/spreadsheets/d/1LBFHjPt4rj_9r0t0TTAlT68NwOtNH8Z21lBMsJDMoZg.

[4] https://docs.google.com/document/d/1OgXSLriHO3Ekz0OYoaoP_h0sPcuvV4EqX7VgLLblKe4.

REFERENCES

Abdullah, M. S., Benest, I., Evans, A., & Kimble, C. (2002). Knowledge modelling techniques for developing knowledge management systems. In *Third European Conference on Knowledge Management: Trinity College Dublin, Ireland*.

Amant, R. S., Long, T., & Dulberg, M. S. (1998). Experimental evaluation of intelligent assistance for navigation. *Knowledge-Based Systems*, *11*(1), 61–70.

Backfisch, I., Lachner, A., Hische, C., Loose, F., & Scheiter, K. (2020). Professional knowledge or motivation? Investigating the role of teachers' expertise on the quality of technology-enhanced lesson plans. *Learning and Instruction*, *66*, 101300.

Benyon, D. (2001). The new HCI? Navigation of information space. *Knowledge-Based Systems*, *14*(8), 425–430.

Burke, R. D., Hammond, K. J., & Young, B. C. (1996). Knowledge-based navigation of complex information spaces. In *Proceedings of the National Conference on Artificial Intelligence* (p. 462, 468). IEEE.

Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.

Chen, J.-M., & Luetz, J. M. (2020). Mono−/inter−/multi−/trans−/anti-disciplinarity in research. In: Leal Filho, W., Azul, A., Brandli, L., Özuyar, P., Wall, T. (eds) *Quality Education* (pp. 562–577). Encyclopedia of the UN Sustainable Development Goals. Springer, Cham. https://doi.org/10.1007/978-3-319-69902-8_33-1.

Chen, P., Lu, Y., Zheng, V. W., & Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 39–48). IEEE.

Chen, Q., Lee, K., Yan, S., Kim, S., Wei, C.-H., & Lu, Z. (2020). Bioconceptvec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Computational Biology*, *16*(4), e1007617. https://doi.org/10.1371/journal.pcbi.1007617

Chiou, C.-K., Tseng, J. C., Hwang, G.-J., & Heller, S. (2010). An adaptive navigation support system for conducting context-aware ubiquitous learning in museums. *Computers & Education*, *55*(2), 834–845. https://doi.org/10.1016/j.compedu.2010.03.015

Chiu, D.-Y., & Pan, Y.-C. (2014). Topic knowledge map and knowledge structure constructions with genetic algorithm, information retrieval, and multi-dimension scaling method. *Knowledge-Based Systems*, *67*, 412–428.

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., & Sun, J. (2016). Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1495–1504). ACM.

Díaz, B., & Nussbaum, M. (2024). Artificial intelligence for teaching and learning in schools: The need for pedagogical intelligence. *Computers & Education*, *217*, 105071. https://doi.org/10.1016/j.compedu.2024.105071

Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king-man +woman= queen. In *Proceedings of coling 2016, the 26th International Conference on Computational Linguistics. Technical papers* (pp. 3519–3530). The COLING 2016 Organizing Committee.

Ethayarajh, K., Duvenaud, D., & Hirst, G. 2018. Towards understanding linear word analogies. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3253–3262). Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1315.

Feldman-Maggor, Y., Tuvi-Arad, I., & Blonder, R. (2024). Navigating the online learning journey by self-regulation: Teachers as learners. *Computers & Education*, *219*, 105074. https://doi.org/10.1016/j.compedu.2024.105074

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., Barabási, A.-L., & others. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Ganguly, S., & Pudi, V. (2017). Paper2vec: Combining graph and text information for scientific paper representation. In *Advances in information retrieval: 39th European conference on IR Research, ECIR 2017, Aberdeen, UK, April 8–13, 2017 Proceedings* (Vol. 39, pp. 383–395). Springer.

Gasparetti, F., Limongelli, C., & Sciarrone, F. (2015). Exploiting wikipedia for discovering prerequisite relationships among learning objects. In *2015 International Conference on Information Technology Based Higher Education and training (ITHET)* (pp. 1–6). IEEE.

González-Márquez R. Schmidt L. Schmidt B. M. Berens P. Kobak D. (2024). The landscape of biomedical research. *Patterns*, *5*(6).

Gu, W., Tandon, A., Ahn, Y.-Y., & Radicchi, F. (2021). Principled approach to the selection of the embedding dimension of networks. *Nature Communications*, *12*(1), 3772.

Hammond, K., Burke, R., Martin, C., & Lytinen, S. (1995). Faq finder: A case-based approach to knowledge navigation. In *Proceedings the 11th Conference on Artificial Intelligence for Applications* (pp. 80–86). IEEE.

Hammond, K. J., Burke, R. D., & Lytinen, S. L. (1995). A case-based approach to knowledge navigation. In *Proceedings of the*

AAAI'94 Workshop on Knowledge Discovery in Databases (pp. 383–393).

Hao, F., & Park, D.-S. (2021). *Conavigator: A framework of fca-based novel coronavirus covid-19 domain knowledge navigation*. Human-centric Computing and Information.

He, P., Wang, T.-Y., Shang, Q., Zhang, J., & Xu, H. (2022). Knowledge mapping of e-commerce supply chain management: A bibliometric analysis. *Electronic Commerce Research*, *24*, 1889–1925.

He, Y., Wang, H., Pan, Y., Zhou, Y., & Sun, G. (2022). Exercise recommendation method based on knowledge tracing and concept prerequisite relations. *CCF Transactions on Pervasive Computing and Interaction*, *4*(4), 452–464.

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, *3*, 23. https://doi.org/10.3389/frma.2018.00023

Hsieh, P. J., Lin, B., & Lin, C. (2009). The construction and application of knowledge navigator model (KNM™): An evaluation of knowledge management maturity. *Expert Systems with Applications*, *36*(2), 4087–4100.

Hsieh, P. J., Lin, C., & Chang, S. (2020). The evolution of knowledge navigator model: The construction and application of KNM 2.0. *Expert Systems with Applications*, *148*, 113209.

Jadad, A. R., & Gagliardi, A. (1998). Rating health information on the internet: Navigating to knowledge or to babel? *Jama*, *279*(8), 611–614.

Kiu, C.-C., & Tsui, E. (2010). Taxofolk: A hybrid taxonomy–folksonomy classification for enhanced knowledge navigation. *Knowledge Management Research & Practice*, *8*(1), 24–32. https://doi.org/10.1057/kmrp.2009.33

Li, L.-Y., Chen, G.-D., & Yang, S.-J. (2013). Construction of cognitive maps to improve e-book reading and navigation. *Computers & Education*, *60*(1), 32–39.

Li, Y.-M., Lin, L.-F., & Lin, Y.-H. (2014). A recommender mechanism for social knowledge navigation in an online encyclopedia. *Information Processing & Management*, *50*(5), 634–652.

Liu, J., Wang, J., Zheng, Q., Zhang, W., & Jiang, L. (2012). Topological analysis of knowledge maps. *Knowledge-Based Systems*, *36*, 260–267.

Lu, W., Zhou, Y., Yu, J., & Jia, C. (2019). Concept extraction and prerequisite relation learning from educational data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 9678–9685.

Manrique, R., Pereira, B., Marino, O., Cardozo, N., & Wolfgand, S. (2019). Towards the identification of concept prerequisites via knowledge graphs. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (Vol. 2161, pp. 332–336). IEEE.

Manrique, R., Sosa, J., Marino, O., Nunes, B. P., & Cardozo, N. (2018). Investigating learning resources precedence relations via concept prerequisite learning. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 198–205). IEEE.

McInnes, L., Healy, J., & Melville, J. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://arxiv.org/abs/1802.03426.

Miao, L., Murray, D., Jung, W.-S., Larivière, V., Sugimoto, C. R., & Ahn, Y.-Y. (2022). The latent structure of global scientific development. *Nature Human Behaviour*, *6*(9), 1206–1217.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv Preprint arXiv:1301.3781

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

Patel, V. L., & Kushniruk, A. W. (1998). Understanding, navigating and communicating knowledge: Issues and challenges. *Methods of Information in Medicine*, *37*, 460–470.

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

Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint at arXiv. https://arxiv.org/abs/2205.01833.

Scheines, R., Silver, E., & Goldin, I. M. (2014). Discovering prerequisite relationships among knowledge components. In Proceedings of the 7th International Conference on Educational Data Mining (EDM) (pp. 355–356).

Schulz, M. (2023). Knowledge and inquiry—The missing key for a knowledge-based decision theory. *Asian Journal of Philosophy*, *2*(2), 54.

Shirah, J. F., & Sidney, P. G. (2023). Computer-based feedback matters when relevant prior knowledge is not activated. *Learning and Instruction*, *87*, 101796.

Sun, H., Li, Y., & Zhang, Y. (2022). Conlearn: Contextual-knowledge-aware concept prerequisite relation learning with graph neural network. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)* (pp. 118–126). Society for Industrial and Applied Mathematics.

Tang, J. (2016). Aminer: Toward understanding big scholar data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (p. 467). ACM.

Tang, X., Liu, K., Xu, H., Xiao, W., & Tan, Z. (2023). Continual pretraining of language models for concept prerequisite learning with graph neural networks. *Mathematics*, *11*(12), 2780.

Vail, E. F. (1999). Knowledge mapping: Getting started with knowledge management. *Information Systems Management*, *16*(4), 1–8.

Van Eck, N., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Wegerif, R., McLaren, B. M., Chamrada, M., Scheuer, O., Mansour, N., Mikšátko, J., & Williams, M. (2010). Exploring creative thinking in graphically mediated synchronous dialogues. *Computers & Education*, *54*(3), 613–621.

Wexler, M. N. (2001). The who, what and why of knowledge mapping. *Journal of Knowledge Management*, *5*(3), 249–264.

White, J. (2020). Pubmed 2.0. *Medical Reference Services Quarterly*, *39*(4), 382–387.

Xiao, K., Bai, Y., & Wang, S. (2021). Mining precedence relations among lecture videos in MOOCS via concept prerequisite learning. *Mathematical Problems in Engineering*, *202*, 1–10.

Xiong, Z., Liu, Q., & Huang, X. (2022). The influence of digital educational games on preschool children's creative thinking. *Computers & Education*, *189*, 104578.

Xu, X., & Dan, Z. (2023). Exploring the evolution of energy research in hospitality: Mapping knowledge trends, insights, and frontiers. *Energy Reports*, *10*, 864–880.

Yamamoto, Y., & Takagi, T. (2007). Biomedical knowledge navigation by literature clustering. *Journal of Biomedical Informatics*, *40*(2), 114–130.

Yan, L., Na, C., & Kang, J. (2024). The impact of team synchrony on argument construction and science knowledge acquisition: Insights from a science learning game. *Journal of Science Education and Technology*, *33*, 633–646.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714*, 1–73.

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: 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

APPENDIX A

A.1 | A COMPARISON WITH EXISTING APPROACHES

TABLE A1 Comparison with existing approaches.

	SciConNav	Existing approaches
Corpus	Scientific research trajectories	Knowledge or information in a certain field
Domain	Across 19 disciplines	Domain-Specific
Techniques	Concept embeddings, word analogy, shortest-path	Static knowledge graph, co-occurrence or citation-based networks
Efficiency	Computationally efficient for navigating large-scale scientific knowledge networks	Manual curation or static graph structures, which can be time-consuming and less scalable
Accuracy	Semantic dependencies between concepts	Manual information retrieval through matching
Applicability	Identifies bridging concepts and facilitates cross-domain exploration and knowledge navigation	Lack the flexibility to accommodate diverse information source

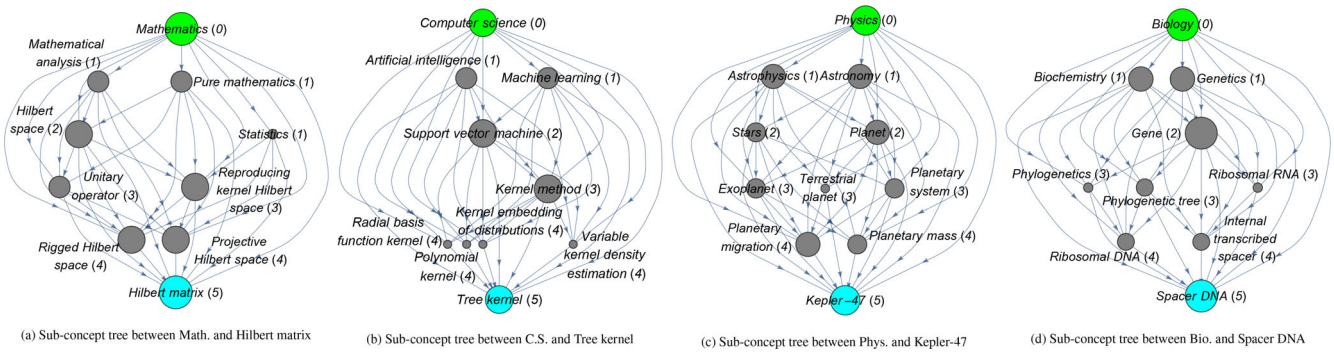
Abbreviation: SciConNav, Scientific Concept Navigator.

APPENDIX B: MATERIALS AND METHODS

B.1 | Dataset description

In our study, we utilize the OpenAlex dataset, which organizes concepts hierarchically across six levels (from Levels 0 to 5). Each publication within this dataset is linked to multiple concepts, spanning Levels 0 through 5. The dataset encompasses approximately 65,000 concepts in total. For a focused and comprehensive analysis, we select the top 20,000 concepts based on the highest number of associated works. At the foundation of the hierarchy are 19 Level 0 concepts, each representing a distinct academic discipline. Illustrative examples of these sub-concept trees are provided in Figure B1, each originating from a distinct discipline concept and extending to Level 5 concepts. Specifically, the trees start from the disciplines of Mathematics, Computer Science, Physics, and Biology, connecting to their respective Level 5 concepts: Hilbert matrix, Tree kernel, Kepler 47, and Spacer DNA.

Taking Figure B1a as an illustration, we observe multiple directed graph paths bridging the discipline of Mathematics with the concept of Hilbert matrix. One such path unfolds as follows: (Mathematics → Mathematical analysis → Hilbert space → Unitary operator → Hilbert matrix). This and similar sub-concept trees delineate the ancestral structure, shedding light on the underlying structure of knowledge within each discipline.



(a) Sub-concept tree between Math. and Hilbert matrix

(b) Sub-concept tree between C.S. and Tree kernel

(c) Sub-concept tree between Phys. and Kepler-47

(d) Sub-concept tree between Bio. and Spacer DNA

FIGURE B1 Four examples sub-concept tree in OpenAlex dataset. The size of the nodes corresponds to the degree of the nodes. Each directed edge in the graph represents the ancestral relationship between the ancestor and descendant concept. The number within the parentheses following the concept name signifies the concept's level. The hierarchy of the concept tree starts from Level 0 at the top and gradually descends to Level 5 at the bottom. (a) Sub-concept tree between Math. and Hilbert matrix. (b) Sub-concept tree between C.S. and Tree kemel. (c) Sub-concept tree between Phys. and Kepler-47. (d) Sub-concept tree between Bio. and Spacer DNA.

B.2 | Concept discipline classification

B.2.1. | LLM discipline annotation

The LLM discipline annotation framework is illustrated in Figure B2. Each concept and the 19 root-level concepts are used to construct a prompt, queried through five LLMs to yield responses in a specified format. The five LLMs from top to bottom are respectively: *GPT-4o*, *Claude-3.5-Sonnet*, *Grok-2*, *Gemini-2.0-Flash-Exp*, *Llama-3.1-Nemotron-70B-Instruct*. We annotated the disciplines of 65,068 concepts from the OpenAlex dataset, the final discipline determined by majority vote.

Specifically, if the five LLMs are unable to reach a consensus, such as in cases where the votes are distributed (e.g., [2, 2, 1] or [1, 1, 1, 1, 1]), we will rely on the top three models: GPT-4o, Claude-3.5-Sonnet, and Grok-2 to make the determination. If these three models still fail to reach a consensus (e.g., [1, 1, 1]), the response from GPT-4o will be used as the final decision, counting for only 1 vote. As our selected author trajectories encompass 64,976 concepts, thus essentially we actually use 64,976 concepts. Table B2 presents the annotation statistics, showing the distribution of 64,976 concepts across 19 disciplines based on different voting patterns.

Notably, we identified 44,413 concepts with full agreement, accounting for 68.35%, and 10,902 concepts with near-full agreement (four votes), accounting for 16.78%. Meanwhile, 8062 and 1448 concepts received three votes and two votes of agreement, respectively. Additionally, 151 concepts, where all five LLMs provided different discipline annotations (i.e., no agreement), highlight significant uncertainty, reflecting the nuanced nature of disciplinary boundaries. We initially categorize concepts as either "Zero-root" "Single-root" or "Multi-root" concepts based on concept tree structure. Table B3 shows the number of "Mono" and "Multi" concepts grouped by concept tree structure.

B.2.2. | Pathway quantity-based classification

In this section, we provide a detailed description of the classification process for each concept's discipline label. We denote $\mathbb{C}$ as the set of all concepts, and $\mathbb{C}^0$ as the set of 19 Level 0 concepts, which represent the 19 potential ancestor roots (AR), and $\overline{\mathbb{C}^0}$ as the set of descendent concepts whose level ranges from Levels 1 to 5 (as shown in Figure B1a–d). The 19 disciplines are shown in Table C6 from No.1 to No.19. To refine our analysis further, we introduce three additional categories: "Disciplinary," "Interdisciplinary," and "Multi-interdisciplinary." These categories help streamline the classification process and the specific scope of concepts under each category will be elucidated in subsequent sections. For each concept $c \in \overline{\mathbb{C}^0}$, there is no direct information to tell which discipline $d \in \mathbb{C}^0$ this concept c mainly belongs to, which we define as its discipline root (DR). As defined in the main text, $\mathrm{AR}(c)$ is the set of root-level ancestors (ARs) of concept $c \in \overline{\mathbb{C}^0}$, and $\mathrm{DR}(c) \subset \mathrm{AR}(c)$ is the set containing the single classified discipline label of c, it is natural that we select an optimum discipline $d \in \mathrm{AR}(c)$ based on the maximum number of graph paths to c as its main discipline.

Let $\mathrm{NPaths}(d, c)$ be the number of distinct paths from discipline d to concept c in the concept tree. Generally, multiple paths exist from a root concept

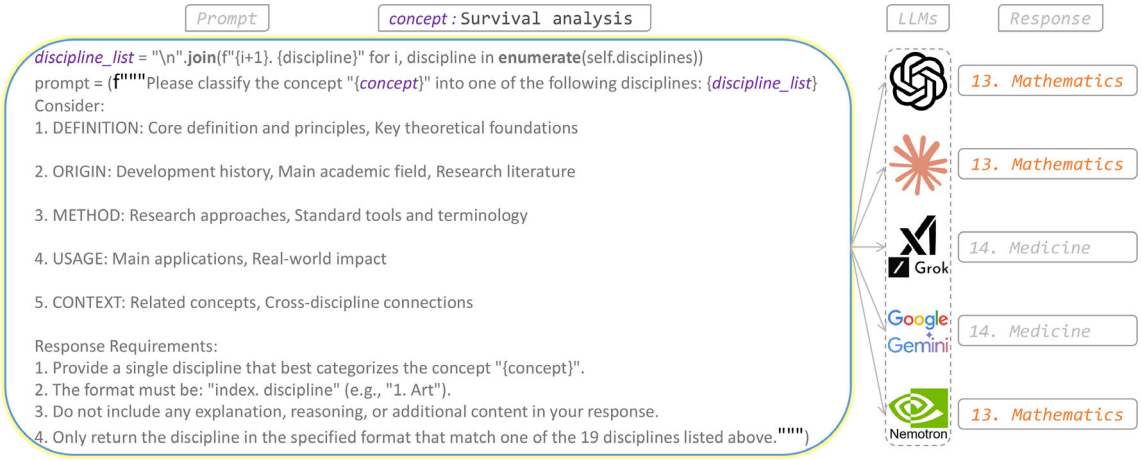


FIGURE B2 Annotation framework of five LLMs with Majority vote. The five LLMs from top to bottom are respectively: GPT-4o, Claude-3.5-Sonnet, Grok-2, Gemini-2.0-Flash-Exp, Llama-3.1-Nemotron-70B-Instruct.

TABLE B2 Number of concepts of 19 disciplines under different vote agreement with large language models annotation (total 64,976 concepts).

Vote	Discipline index 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Sum
1	9	11	28	13	13	7	3	9	7	3	6	6	6	11	1	0	4	11	3	151 (2.3)
2	77	171	139	86	86	30	95	77	81	18	63	81	46	98	62	41	20	63	114	1448 (2.23%)
3	135	1420	489	492	406	244	859	282	181	103	196	232	269	1092	179	559	215	302	407	8062 (12.41%)
4	138	3034	416	669	502	388	904	288	182	124	180	330	353	1435	162	644	208	422	523	10,902 (16.78%)
5	304	5433	729	3641	3984	1168	3078	861	408	921	452	707	3677	11,513	683	3697	962	1318	877	44,413 (68.35%)
Sum	663	10,069	1801	4901	4991	1837	4939	1517	859	1169	897	1356	4351	14,149	1087	4941	1409	2116	1924	64,976
%	1.02	15.5	2.77	7.54	7.68	2.83	7.6	2.33	1.32	1.8	1.38	2.09	6.7	21.78	1.67	7.6	2.17	3.26	2.96	1

(e.g., "Mathematics") at the top to a specific sub-concept (e.g., "Hilbert matrix") at the bottom of the concept tree, as illustrated in Figure B1a. We neglect the concepts with no ancestor root ($|\mathrm{AR}(c)|=0$), and subsequently partition the remaining concepts into two groups S and M.

1. $S=\left\{c\middle|c\in\overline{\mathbb{C}^0},|AR(c)|=1\right\}$, a total of 17,508 concepts with a single AR. This single AR of each concept c is determined as the discipline label of c, without causing any ambiguity.
2. $M=\left\{c\middle|c\in\overline{\mathbb{C}^0},|\mathrm{AR}(c)|>1\right\}$, a total of 47,002 concepts with multiple ARs. We determine the $\mathrm{DR}(c)$ for each concept c by analyzing the graph paths between its multiple ARs and concept c.

For each $c\in S$, this sole AR is classified as the DR that $\mathrm{DR}(c)=\mathrm{AR}(c)$. For each $c\in M$, we determine the main discipline root d, or the classified discipline label of concept c as the ancestor root that maximizes $\mathrm{NPaths}(d,c)$. For a concept $c\in M$, the maximum-paths (MaxPaths) from its ARs to c is calculated as

$$\mathrm{MaxPaths}(c)=\max\{\mathrm{NPaths}(d,c)|d\in\mathrm{AR}(c)\},\tag{B1}$$

thus the set of maximum-paths roots (MPR) is denoted as

$$\mathrm{MPR}(c)=\{d\in\mathrm{AR}(c)|\mathrm{NPaths}(d,c)=\mathrm{MaxPaths}(c)\},\tag{B2}$$

hence if there is a unique AR exhibiting the MaxPaths that $|\mathrm{MPR}(c)|=1$, which means that this concept c is classifiable, hence $\mathrm{DR}(c)=\mathrm{MPR}(c)$. If $|\mathrm{MPR}(c)|>1$, we assign such unclassifiable concept a new class: "Multi-interdisciplinary."

Overall, this method allows us to systematically assign a discipline category to all the concepts under consideration, ensuring each concept is accurately categorized based on its most prominent connections to discipline root within the knowledge network. The corresponding algorithm is shown in Algorithm 1, and Table B4 provides illustrative examples of the concept discipline classification process. Specifically, the concept PPADS refers to "Pyridoxal-phosphate-6-azophenyl-2',4'-disulfonic acid". We define the extended classifiable concepts S^+ as the union of concepts in S and classifiable concepts in M, comprising 49,275 concepts, which we term "Disciplinary" concepts. The remaining unclassifiable concepts in M, denoted as M^- and termed "Multi-interdisciplinary" concepts, account for 15,235 concepts (Table B4).

TABLE B3 Contingency table of concepts grouped by large language models annotation and concept tree structure.

	Mono (C_s)	**Multi** (C_m)	**Sum**
Zero-root (Z)	355	135	490 (7.5)
Single-root (S)	14,105	3383	17,488 (26.92%)
Multi-root (M)	29,953	17,045	46,998 (72.33%)
Sum	44,413 (68.35%)	20,563 (31.65%)	64,976

ALGORITHM 1 **Concept discipline root classification**

Input: Set of concepts $\mathbb{C}$, concept tree $\mathbb{G}$
Output: Sets S^+, M^-, $\mathbf{DR}(c)$ and $\mathbf{NDR}(c)$ for each concept $c \in S^+$
Initialize sets: $S^+ = S$, $M^- = \emptyset$, $\mathbb{C}^0 = \{1, 2, ..., 19\}$
$\overline{\mathbb{C}^0} \leftarrow \mathbb{C} \setminus \mathbb{C}^0$ be the concepts with level ranging from 1 to 5
$S = \left\{ c | c \in \overline{\mathbb{C}^0}, |\mathbf{AR}(c)| = 1 \right\}$
$M = \left\{ c | c \in \overline{\mathbb{C}^0}, |\mathbf{AR}(c)| \geqslant 2 \right\}$
for each concept $c \in S$ **do**
  $\mathbf{DR}(c) \leftarrow \mathbf{AR}(c)$
  $\mathbf{NDR}(c) \leftarrow \mathbb{C}^0 \mathbf{DR}(c)$
end for
for each concept $c \in M$ **do**
  $\mathbf{MaxPaths}(c) \leftarrow \max\{\mathbf{NPaths}(d, c) | d \in \mathbf{AR}(c)\}$
  $\mathbf{MPR}(c) \leftarrow \{d | d \in \mathbf{AR}(c), \mathbf{NPaths}(d, c) = \mathbf{MaxPaths}(c)\}$
  if $|\mathbf{MPR}(c)| > 1$ **then**
    $M^- \leftarrow M^- \cup \{c\}$
  else
    $\mathbf{DR}(c) \leftarrow \mathbf{MPR}(c)$
    $\mathbf{NDR}(c) \leftarrow \mathbb{C}^0 \mathbf{DR}(c)$
    $S^+ \leftarrow S^+ \cup \{c\}$
  end if
end for return sets S^+, M^-, and $\mathbf{DR}(c), \mathbf{NDR}(c)$ for each $c \in S^+$

TABLE B4 Illustration of concept discipline root classification.

Subset	***S* (single AR)**		***M* (multiple ARs)**					
c	Simplified manifold		PPADS		Glycine cleavage system		Cutter location	
AR(c)	AR	# paths	AR	# paths	AR	# paths	AR	# paths
	Math.	16	*Bio.*	54	*Bio.*	29	*C.S.*	12
			Chem.	32	Chem.	22	*Eng.*	12
			Med.	37	Phys.	8	*Mat. Sci.*	12
							Math.	4
DR	Mathematics		Biology		Biology		Multi-interdisciplinary	
Update	S^+ (classifiable)						M^- (unclassifiable)	

B.3 | Dimension analysis

For each discipline d, let Sub(d) denote the set of its sub-concepts. We investigate how dimensional reduction affects the similarities across disciplines. The cross-domain propensity (CDP) of the sub-concepts Sub(d_1) within discipline d_1 toward discipline d_2 is defined as the distribution of similarities between the concepts in Sub(d_1) and discipline d_2 that

$$\mathrm{CDP}(d_1, d_2) = \left\{ \mathrm{sim}\left(\vec{c}, \vec{d}_2\right) \middle| c \in \mathrm{Sub}(d_1) \right\}. \quad \text{(B3)}$$

We further generalize this to the in-domain propensity (IDP) and define the IDP of the sub-concepts $\mathrm{Sub}(d)$ toward their own discipline d as $\mathrm{IDP}(d) = \mathrm{CDP}(d, d)$. This metric measures how closely the sub-concepts $\mathrm{Sub}(d)$ are related to their respective discipline d.

We first analyze the IDP of 19 disciplines across six embedding dimensions shown in Figure B3a, which demonstrates a systematic decline as the embedding dimensions increase from 24 to 128. This trend is uniformly observed across all 19 disciplines, as evidenced by the progressive lowering of boxplot medians and their corresponding quartile ranges. The results demonstrate that 24 dimensions achieve optimal IDP and higher semantic relationships within their discipline, while higher dimensions (32, 48–128) lead to increased sparsity and diminishing similarity concentration.

We further analyzed the CDP patterns of the sub-concepts within "Mathematics" toward 19 other disciplines, as shown in Figure B3b. In this analysis, d_1 represents "Mathematics" while d_2 corresponds to each of the 19 disciplines. We show that "Mathematics" sub-concepts exhibit the highest similarity with "Mathematics" itself, the similarities with other 18 disciplines were notably lower. When examining the propensities toward quite unrelated "Biology" and "Medicine," we find that lower

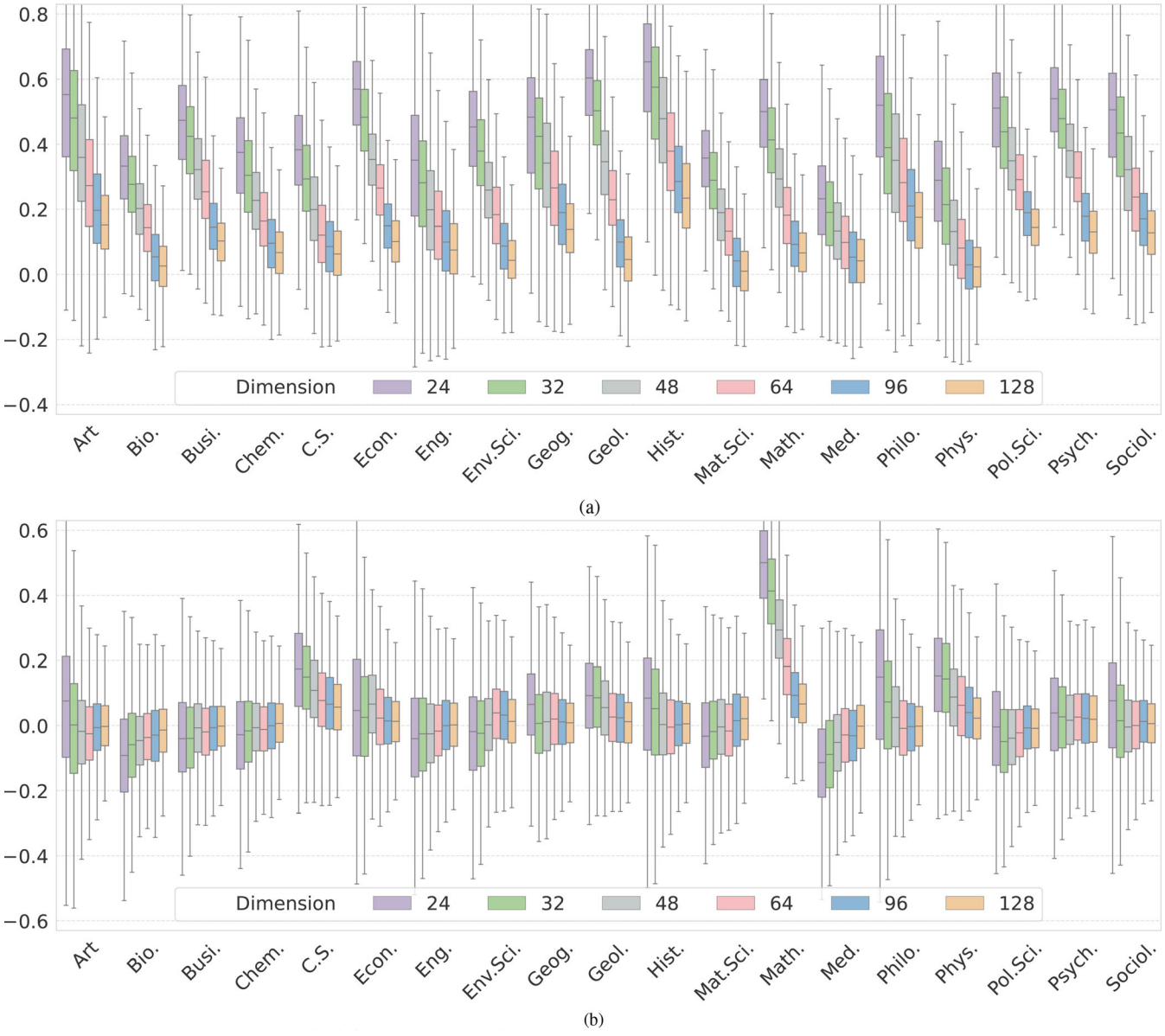


FIGURE B3 Discipline propensities analysis across different embedding dimensions: (a) Comparison of in-domain across 19 disciplines measured in six embedding dimensions (24, 32, 48, 64, 96, and 128); (b) Discipline propensities of Mathematics concepts to 19 disciplines measured in six embedding dimensions (24, 32, 48, 64, 96, and 128).

dimensions led to reduced similarity propensity toward "Biology" and "Medicine." This phenomenon validates that 24-dimensional embeddings achieve superior domain discrimination, maximizing the distinction between unrelated domains while maintaining high similarity within related domains, demonstrating its advantage over higher dimensions in semantic representation learning.

Overall, we select 24 dimensions in our experiments, as it maintains power-of-two alignment, ensuring computational efficiency, and achieves better domain discrimination by simultaneously maximizing in-domain similarities while minimizing cross-domain similarities with unrelated disciplines.

B.4 | Analysis of concepts concentration

First, We divide the embedding map into 10×10 grids for each discipline as shown in Figure B4a, and the count the

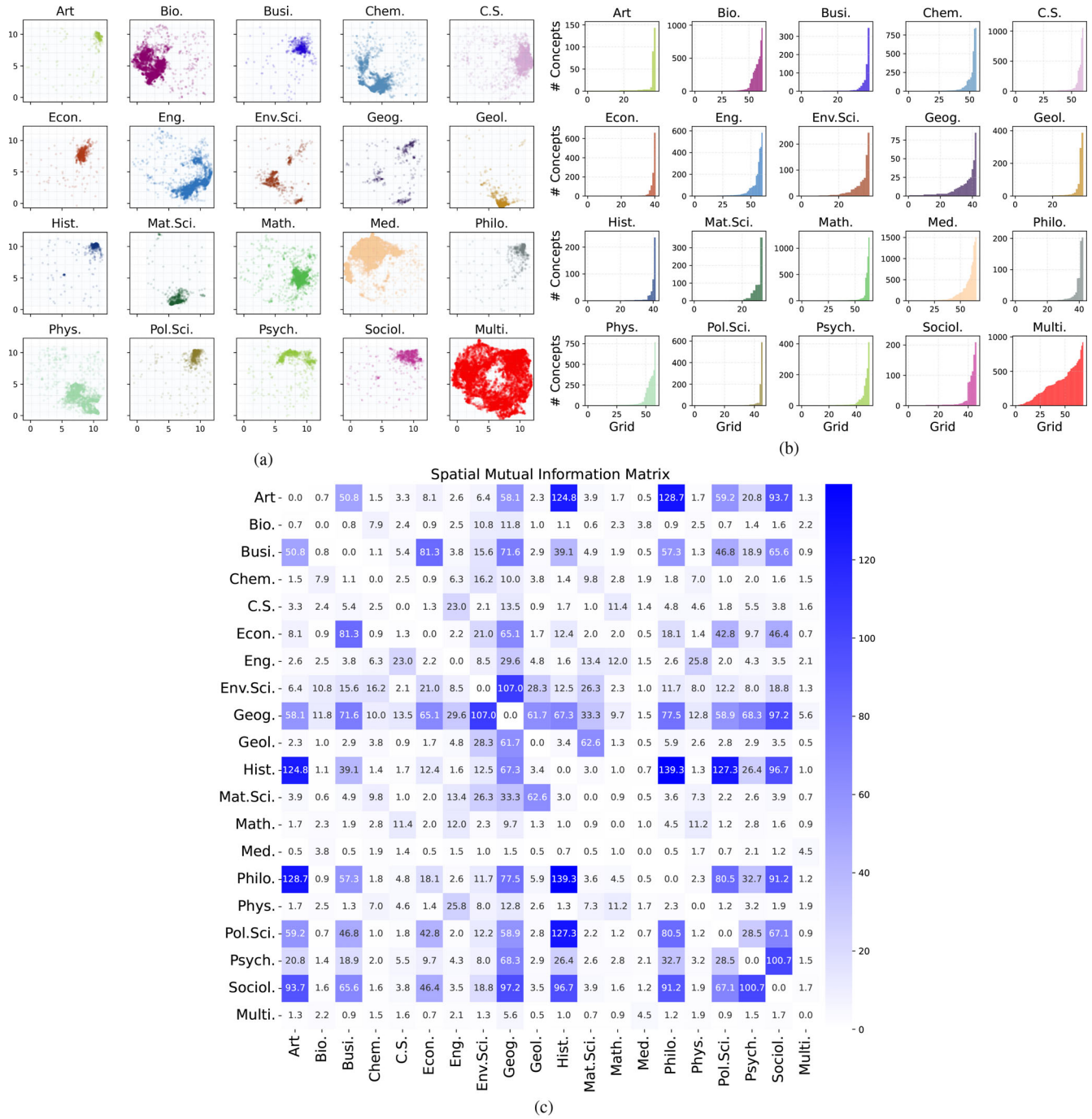


FIGURE B4 Distribution analysis of disciplinary concepts: (a) Density plot of embedding map of each discipline. Spatial density distribution in the embedding space; (b) Bar charts of grid counts for each discipline. (c) Spatial mutual information matrix.

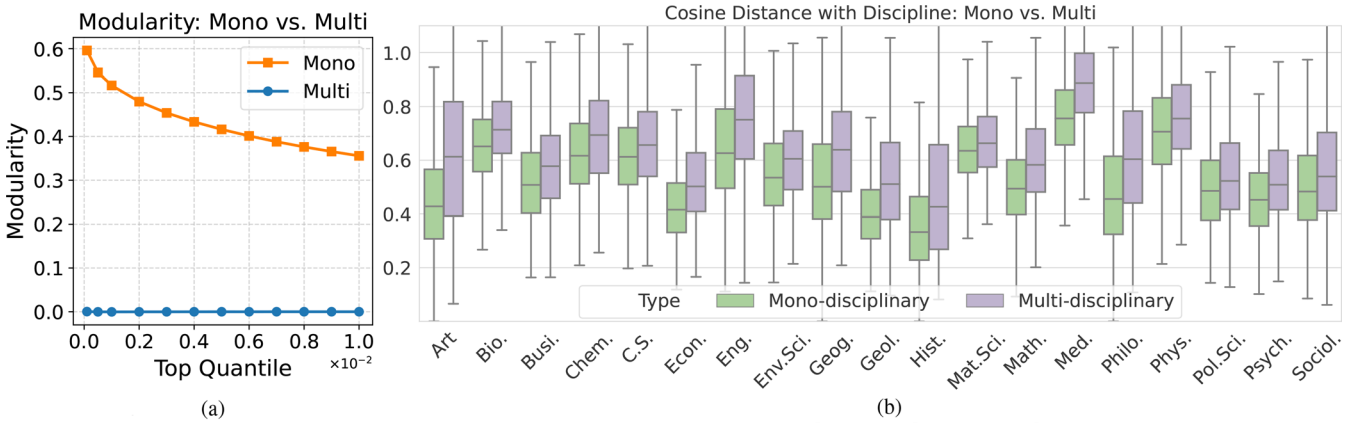


FIGURE B5 (a) The modularity of Mono and Multi concepts under different top quantiles. (b) Cosine distance with Discipline: Mono-disciplinary versus Multi-disciplinary.

number of concepts per grid for each discipline, the grid counts are ranked in ascending order and visualized as a ranked bar chart, where the bar heights represent the number of concepts in the corresponding grids. We formally refer to this visualization as the BCGC. The BCGC for "Mono" concepts of 19 disciplines and "Multi" concepts are shown in Figure B4b, a steeper rise in bar heights indicates a stronger concentration of concepts in fewer regions, while a more gradual rise suggests a broader distribution across regions. We then compute the mutual information between two disciplines d_i and d_j, we first estimate kernel density functions f_i and f_j based on their respective concept coordinates. Using the coordinates corresponding to the union of sub-concept sets $\mathrm{Sub}(d_i) \cup \mathrm{Sub}(d_j)$, which is $V = \{v_c = (x_c, y_c), c \in \mathrm{Sub}(d_i) \cup \mathrm{Sub}(d_j)\}$, we evaluate both density functions at these shared coordinates to obtain the set of density values $D_i = \{f_i(v), v \in V\}$ and $D_j = \{f_j(v), v \in V\}$. Then we calculate the mutual information using D_i and D_j to quantify the spatial correlation between the two distributions. The results in Figure B4c shows that "Multi" concepts exhibit low spatial correlations with all disciplines.

Second, we analyze the modularity of two type of networks: the "Mono" network (44,413 concepts in Figure 2a) and the "Multi" network (20,533 concepts in Figure 2b). Network edges are selected as pairs of concepts with similarities ranked among the top ratios, ranging from 10^{-4} to 10^{-2} at varying thresholds $(10^{-4}, 5 \times 10^{-4}, 10^{-3}, ..., 10^{-2})$. The modularity of both networks is calculated at each threshold, as shown in Figure B5a. The results indicate that the modularity of "Mono" concepts decreases gradually as the top ratio increases, while the modularity of "Multi" concepts remains consistently low, close to 0. This finding highlights that "Mono" concepts exhibit significantly higher modularity compared to "Multi" concepts.

TABLE B5 Selected theoretical Level 1 concepts in "Mathematics" and "Physics."

	Selected theoretical	Excluded non-theoretical
Mathematics	"Statistics," "Combinatorics," "Discrete mathematics," "Mathematical optimization," "Mathematical analysis," "Pure mathematics," "Geometry," "Arithmetic"	"Mathematics education," "Actuarial science," "Applied mathematics"
Physics	"Particle physics," "Statistical physics," "Quantum electrodynamics," "Theoretical physics," "Mathematical physics," "Mechanics," "Quantum mechanics," "Classical mechanics," "Thermodynamics," "Astrophysics"	"Astronomy," "Chemical physics," "Atomic physics," "Nuclear physics," "Medical physics," "Acoustics," "Condensed matter physics," "Computational physics," "Molecular physics," "Nuclear magnetic resonance," "Crystallography," "Geophysics"

Third, we analyze and compare the cosine distance distributions of "Mono" and "Multi" sub-concepts to their respective disciplines across each discipline. For a discipline d, let $\mathrm{Sub}_s(d) = \mathrm{Sub}(d) \cap C_s$ be the set of "Mono" sub-concepts of d, and $\mathrm{Sub}_m(d) = \mathrm{Sub}(d) \cap C_m$ be the "Multi" sub-concepts of d. We calculate the set of

cosine distances between the sub-concepts and discipline d as

$$\mathrm{CosDis}(d,\alpha)=\left\{1-\mathrm{sim}\left(\vec{i},\vec{d}\right)\middle|\, i\in \mathrm{Sub}_{\alpha}(d)\right\},\alpha\in\{s,m\},\quad \text{(B4)}$$

where $\alpha=s$ corresponds to the "Mono" sub-concepts, and $\alpha=m$ corresponds to the "Multi" sub-concepts. Then we compare the cosine distances between disciplines and their respective "Multi" and "Mono" sub-concepts. As shown in Figure B5b, the distribution of $\mathrm{CosDis}(d,m)$ consistently exceeds $\mathrm{CosDis}(d,s)$ across 19 disciplines, indicating that "Multi" concepts are generally more distant from their parent disciplines than "Mono" concepts.

APPENDIX C: EXTENDED RESULTS

C.1 | Discipline propensity of concepts from different disciplines

In addition to the examples in the main text, we show more examples for other disciplines to elucidate the discipline propensity of concepts. We employ radar maps to visualize the inclination of Levels and 2 concepts toward their respective labeled disciplines. These visual representations are provided in Figures C6 and C7, offering insights into the domain tendencies of disciplinary concepts across various fields.

Figure C6 presents a series of radar maps for 19 disciplines listed in Table C6; each radar map represents the discipline propensity of Level 1 concepts from a distinct discipline to all 19 disciplines. In these radar maps, the most outward-pointing angle (MOPA) toward the labeled discipline indicates that Level 1 concepts exhibit a notably higher cosine similarity with their labeled discipline in comparison to other disciplines. Similarly, Figure C7 displays a set of radar maps of discipline propensity of Level 2 concepts of each discipline against 19 disciplines, respectively. Consistent with Level 1 concepts, the MOPA in the radar map of each discipline toward this discipline underscores that Level 2 concepts maintain a higher cosine similarity with their labeled discipline as opposed to other disciplines.

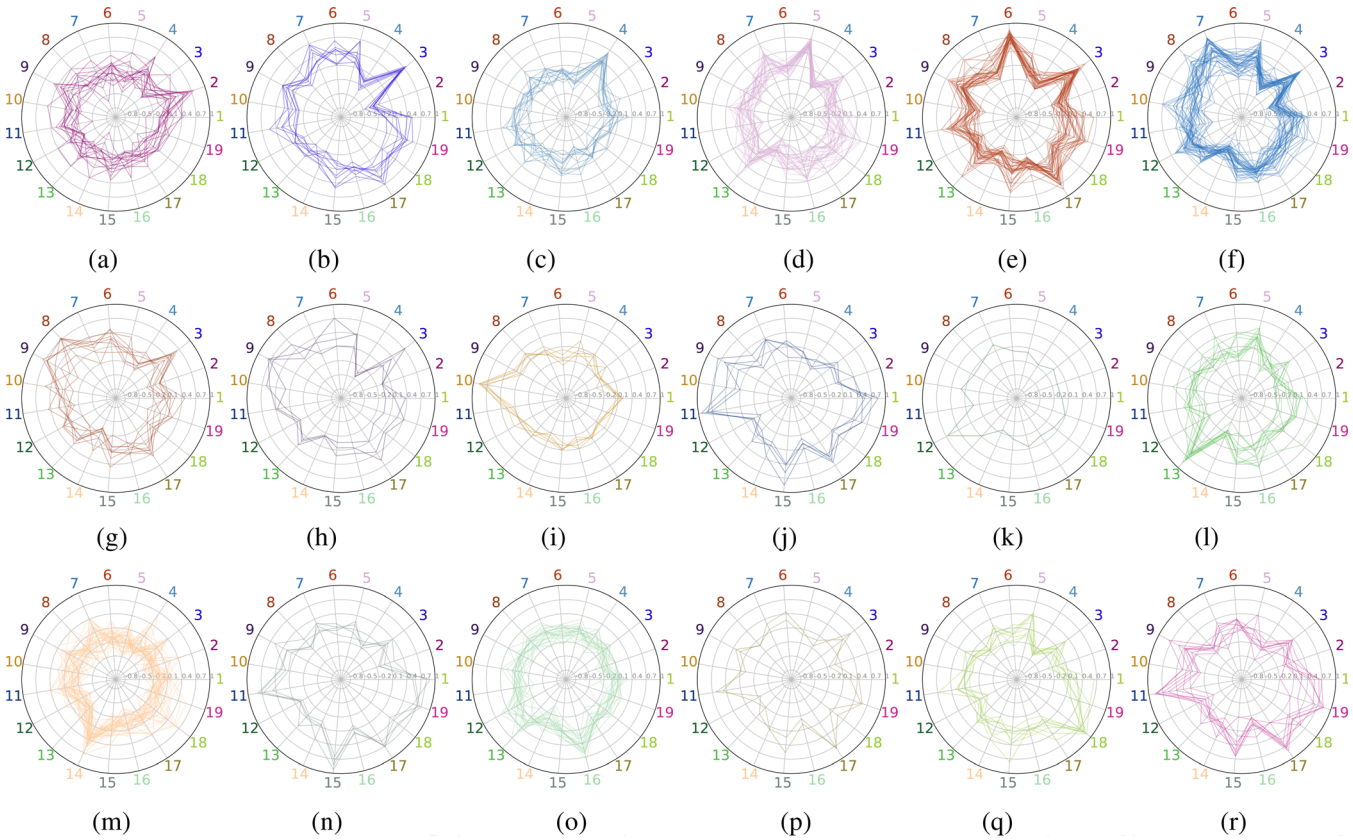


FIGURE C6 Discipline similarity radar rapping: Level 1 sub-concepts. Each radar map, numbered from (1) to (19) excluding the Art (1), illustrates the propensity of Level 1 concepts from a specific discipline toward the 19 disciplines. (a) Biology (2). (b) Business (3). (c) Chemistry (4). (d) C.S. (5). (e) Economics (6). (f) Engineering (7). (g) Env. Sci. (8). (h) Geography (9). (i) Geology (10). (j) History (11). (k) Mat. Sci (12). (l) Mathematics (13). (m) Medicine (14). (n) Philosophy (15). (o) Physics (16). (p) Pol. Sci (17). (q) Psychology (18). (r) Sociology (19).

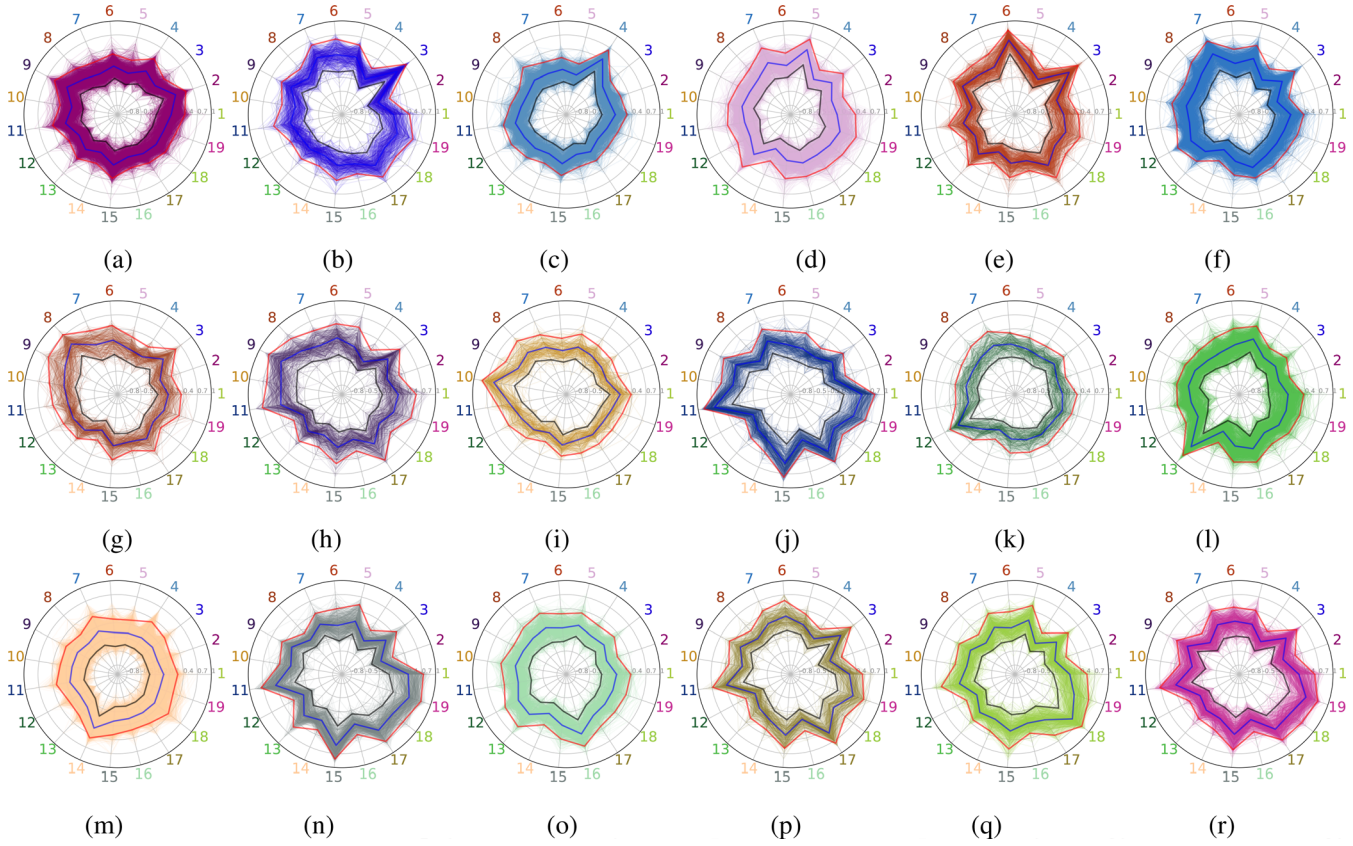


FIGURE C7 Discipline similarity radar rapping: Level 2 sub-concepts. Each radar map, numbered from (1) to (19) excluding the Art (1), illustrates the propensity of Level 2 concepts from a specific discipline toward the 19 disciplines. (a) Biology (2). (b) Business (3). (c) Chemistry (4). (d) C.S. (5). (e) Economics (6). (f) Engineering (7). (g) Env. Sci. (8). (h) Geography (9). (i) Geology (10). (j) History (11). (k) Mat. Sci (12). (l) Mathematics (13). (m) Medicine (14). (n) Philosophy (15). (o) Physics (16). (p) Pol. Sci (17). (q) Psychology (18). (r) Sociology (19).

TABLE C6 Abbreviations of 19 disciplines.

ID	Discipline	Abbr.	ID	Discipline	Abbr.	ID	Discipline	Abbr.
1	Art	Art	8	Environmental science	Env. Sci.	14	Medicine	Med.
2	Biology	Bio.	9	Geography	Geog.	15	Philosophy	Philo.
3	Business	Busi.	10	Geology	Geol.	16	Physics	Phys.
4	Chemistry	Chem.	11	History	Hist.	17	Political science	Pol. Sci.
5	Computer science	C.S.	12	Materials science	Mat. Sci.	18	Psychology	Psych.
6	Economics	Econ.	13	Mathematics	Math.	19	Sociology	Sociol.
7	Engineering	Eng.						

Figures C6 and C7 present radar charts that depict the discipline propensity of Levels 1 and 2 sub-concepts across 18 disciplines. Discipline Env. Sci (8) is excluded due to the limited number of concepts, as specified in Table C6. Each radar map displays the measured propensity of sub-concepts from a specific discipline toward all 19 disciplines. In both figures, each radar map contains 19 axes corresponding to the disciplines, with the outward-pointing degree of each axis representing the calculated propensity of the sub-concepts toward that particular discipline. For instance, in the Mathematics radar map, the MOPA aligns with the Mathematics axis, indicating that Mathematics sub-concepts have the highest calculated propensity toward their own discipline compared to the other 18 disciplines. This pattern is consistent across all disciplines and both hierarchical levels, with each exhibiting the strongest propensity toward its labeled discipline. The trend observed in Figure C6 for Level 1 sub-concepts is mirrored in Figure C7 for Level 2 sub-concepts. These visualizations effectively illustrate

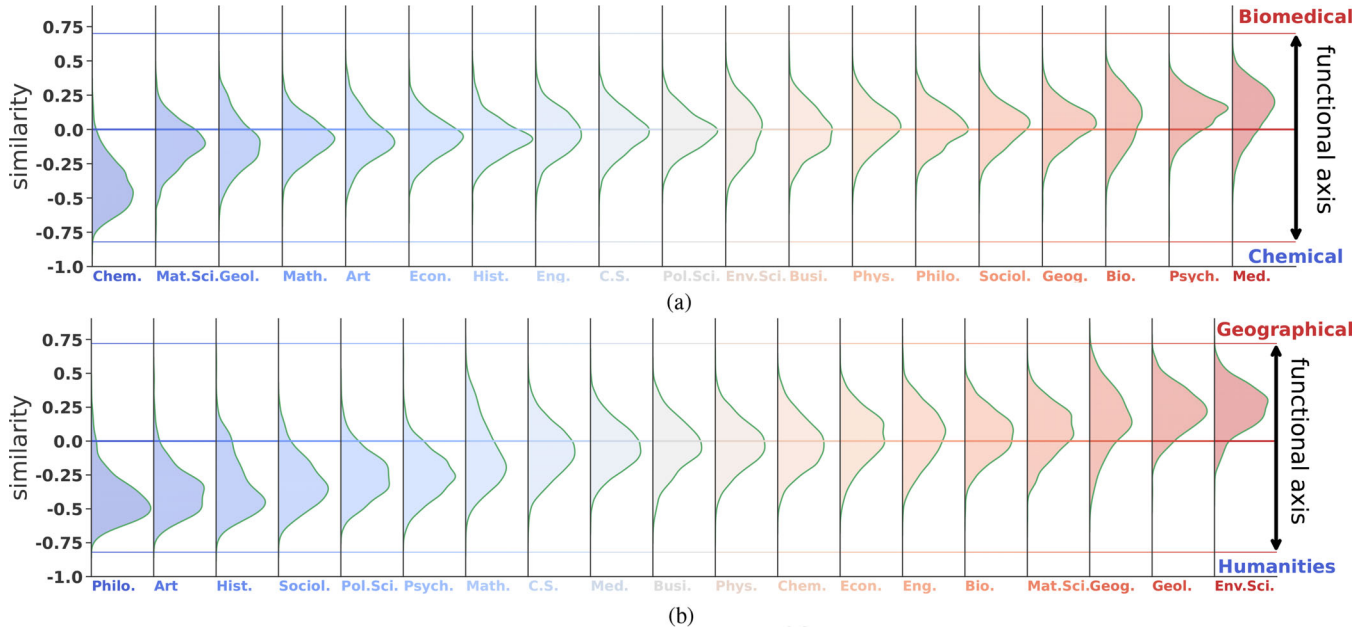


FIGURE C8 Functionality projections of knowledge from 19 disciplines on predefined axes. (a) Functional axis between Chemical and Biomedical. Projection of all disciplines on the functional axis between Chemical and Biomedical. (b) Functional axis between Humanities and Geographical. Projection of all disciplines on the functional axis between Humanities and Geographical.

the strong association between sub-concepts and their respective labeled disciplines across both hierarchical levels, as determined by our calculations, highlighting the distinctiveness and cohesiveness of disciplinary knowledge structures.

C.2 | Examples of analogy inference networks

In Table C7, we present a series of concept analogy inference examples. For each example, starting with a seed

TABLE C7 Two-step analogy inference cases in scientific concepts.

Case	No. 1	No. 2
a	Distributed computing (C.S.)	Combinatorics (Math.)
c	Computer graphics (C.S.)	Information retrieval (C.S.)
d	Statistics (Math.)	Statistics (Math.)
b^{+}	Erasure (C.S., Math.)	Noncentral chi-squared distribution (Bio., Chem., Math., Med.)
b^{++}	Binary symmetric channel (C.S., Eng., Math.)	Logistic distribution (C.S., Math., Med.)
b^{+-}	Emerging technologies (C.S., Mat. Sci.)	Negation (C.S., Philo.)
b^{-}	Emerging technologies (C.S., Mat. Sci.)	Ramsey's theorem (C. S., Math.)
b^{-+}	Computational topology (Chem., Math.,Phys.)	Conjunctive query (C.S.)
b^{--}	Desk (C.S., Eng., Pol. Sci.)	Noncentral chi-squared distribution (Bio., Chem.,Math., Med.)
Case	**No. 3**	**No. 4**
a	Database (C.S.)	Data science (C.S.)
c	Computer network(C.S.)	Computer network (C.S.)
d	Mathematical analysis (Math.)	Mathematical analysis (Math.)
b^{+}	Heat equation (Math., Phys.)	Heat equation (Math., Phys.)
b^{++}	Boundary value problem (Math., Phys.)	Boundary value problem (Math., Phys.)
b^{+-}	Live migration (C.S., Pol. Sci.)	Live migration (C.S., Pol. Sci.)
b^{-}	Virtualization (C.S., Pol. Sci.)	Cloud testing (C.S., Pol. Sci.)
b^{-+}	Boundary value problem (Math., Phys.)	Elementary theory (C.S., Math.)
b^{--}	Server (C.S.)	Virtualization (C.S., Pol. Sci.)

(Continues)

TABLE C7 (Continued)

Case	No. 5	No. 6
a	Combinatorics (Math.)	Applied mathematics (Math.)
c	Multimedia (C.S.)	Computer graphics (C.S.)
d	Applied mathematics (Math.)	Statistics (Math.)
b^{+}	Skew-symmetric matrix (Math., Phys.)	Bivariate analysis (C.S, Math.)
b^{++}	Matrix differential equation (Math., Phys.)	Logistic distribution (C.S., Math., Med.)
b^{+-}	Interactive video (C.S.)	Ray casting (C.S., Eng.)
b^{-}	Two-way communication (C.S., Eng.)	Critical system (Busi., C.S., Econ., Eng., Math., Med.)
b^{-+}	Homoclinic bifurcation (Eng., Math., Phys.)	Odds (C.S., Math., Med.)
b^{--}	Critical system (Busi., C.S., Econ., Eng., Math., Med.)	Computer Science and Engineering (C.S., Eng.)

concept a and two analogous concepts c and d, we perform a two-step analogy inference process. In the first step, we derive the initial inference results b^{+} and b^{-} from the seed concept a. Using these initial results as new seed concepts, we then proceed to the second step of inference. The outcomes of the second step of inference b^{++} and b^{+-} derived from b^{+}, and b^{-+} and b^{--} derived from b^{-} are systematically documented in each row of Table C7, respectively.

C.3 | Examples of concept pathways

We add additional examples of concept pathways in Table C8, highlighting the interconnectedness within the knowledge network. For instance, consider the pathway between the concepts "Statistics" and "Artificial Intelligence" connected by an intermediary concept "Weighting," which serves as a crucial bridge between traditional statistical methods and AI approaches. The concept of weighting is originally developed for statistical sampling and analysis; the principles have evolved to become fundamental in AI applications, where they are crucial in neural network training (e.g., weight optimization), ensemble learning (e.g., model weighting), and feature importance assessment (e.g., feature weighting). This transformation and adaptation of the weighting concept illustrate how interdisciplinary concepts facilitate knowledge transfer and drive innovation across fields.

The provided examples illustrate the shortest semantic paths between pairs of concepts using Word2Vec vector representations, with training data sorted by the publication time of scientific papers. These paths reveal the semantic and logical relationships between concepts, offering insights into their interconnectedness and practical applications. For instance, the path from "Statistics" to "Sociology" includes "Categorical variable" and "Relation (database)," indicating the importance of statistical methods in analyzing sociological data. Similarly, the path from "Artificial Intelligence" to "Bioinformatics" via "Interpretability" emphasizes the need for transparency in AI applications within bioinformatics. Each path demonstrates domain-specific associations that are logical and coherent, such as the path from "Statistics" to "Plant growth," which includes "Best linear unbiased prediction" and "Diallel cross," reflecting the critical role of statistical methods in agricultural research and plant breeding.

TABLE C8 Examples of concept pathways between interested concept pairs.

No.	Shortest concept pathway between two concepts
1	(Statistics, Categorical variable, Relation [database], Meaning [existential], Sociology)
2	(Statistics, Weighting, Artificial intelligence)
3	(Statistics, Probability distribution, Random walk, Complex system)
4	(Statistics, Best linear unbiased prediction, Diallel cross, Cultivar, Phaseolus, Germination, Plant growth)
5	(Artificial intelligence, Representation [politics], Meaning [existential], Sociology)
6	(Artificial intelligence, Representation [politics], Semiotics, Dialectic, Pragmatism, Social science)
7	(Artificial intelligence, Information processing, Judgment, CLARITY, Need to know, Nursing)
8	(Artificial intelligence, Segmentation, Lung biopsy, Bronchoscopy, Cardiothoracic surgery, Vascular surgery)
9	(Artificial intelligence, Task [project management], Driving simulator)
10	(Artificial intelligence, Interpretability, Bioinformatics)
11	(Pure mathematics, Discrete mathematics, Graph, Biological network, Bioinformatics)
12	(Applied mathematics, Convergence [economics], Optimal control, Control [management], Driving simulator)
13	(Combinatorics, Upper and lower bounds, Channel capacity, Fading, Multipath propagation, Radar)
14	(Green building, ASHRAE 90.1, Natural ventilation, Airflow, Venturi effect, Blood viscosity)
15	(Green building, Energy engineering, Energy resources, Gene cluster, Genetics, Human genetics)

These paths underscore the interdisciplinary nature of modern research and can guide researchers in identifying key intermediate concepts and methodologies that bridge different fields. For example, the connection between "Artificial Intelligence" and "Sociology" via "Representation (politics)" and "Meaning (existential)" suggests that AI's impact on societal structures and existential questions is a critical area for future research. Additionally, the path from "Pure mathematics" to "Driving simulator" via "Optimal control" and "Control (management)" indicates the application of mathematical optimization and control theory in developing advanced driving simulations. By revealing logical connections between seemingly disparate fields, these paths can foster cross-disciplinary collaboration. For instance, the path from "Green building" to "Human genetics" through "Energy engineering" and "Gene cluster" suggests potential interdisciplinary research opportunities in sustainable building practices and genetic studies. In summary, these shortest semantic paths provide a structured and meaningful way to understand the relationships between concepts across different domains, offering valuable insights for guiding research directions and enhancing cross-disciplinary collaboration.

C.4 | Path length distribution

For the overall evaluation of PLD, denote the set of all concepts as $\mathbb{C}$ (All). We calculate the path PLD from each discipline to all concepts (All) as follows:

$$\mathrm{PLD}(d) = \left\{ \mathrm{PathLen}\left(c_i, c_j\right) \middle| c_i \in \mathrm{Sub}(d), c_j \in \mathbb{C} \right\}. \quad \text{(C5)}$$

where $\mathrm{PathLen}\left(c_i, c_j\right)$ is the path length between concepts c_i and c_j, and $\mathrm{Sub}(d)$ denotes the set of sub-concepts in discipline d. The results of PLD form 19 disciplines to all concepts are shown in Figure C9. Specifically, the first column ("All") represents the PLD of pairwise paths among all concepts in $\mathbb{C}$.

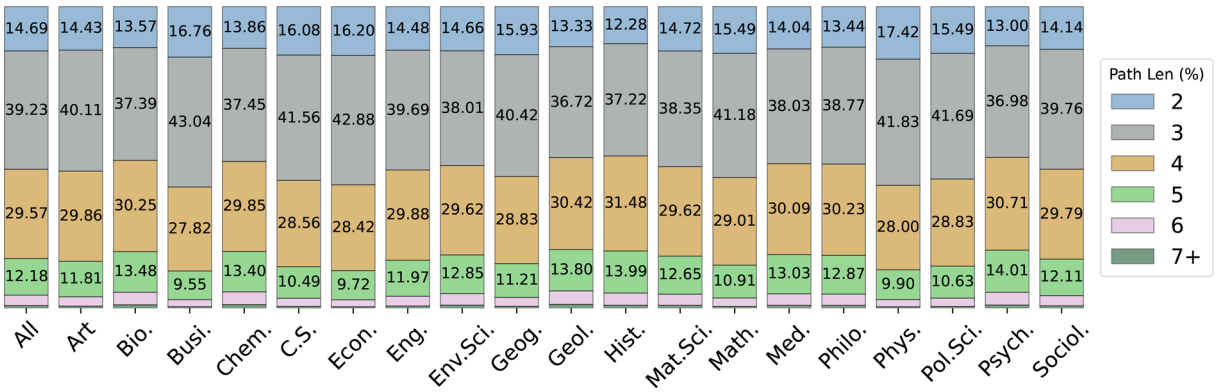


FIGURE C9 Path length distributions of pairwise paths between concepts within each discipline and all concepts.

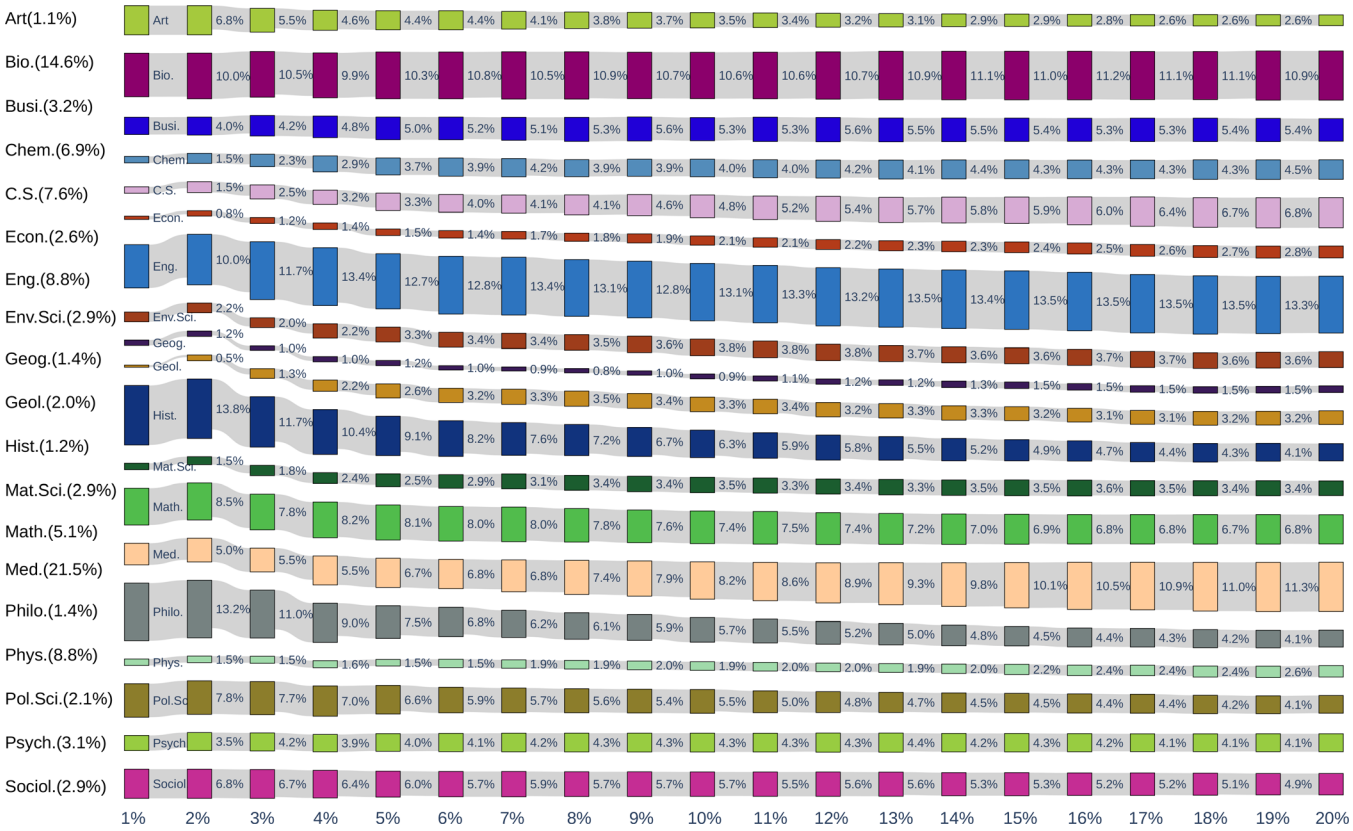


FIGURE C10 Proportion of disciplines of High Closeness centrality concepts from top 1% to top 20%.

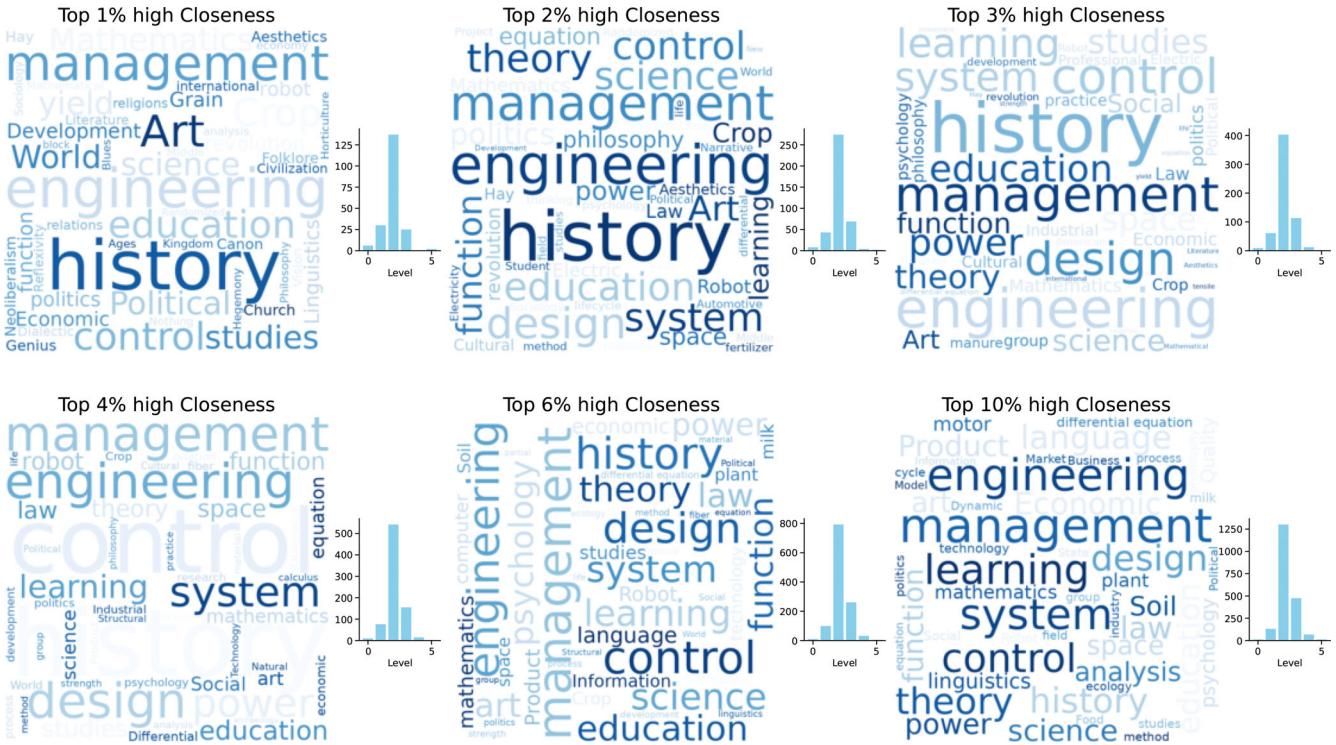


FIGURE C11 High Closeness centrality concepts word clouds (Top 1%, 2%, 3%, 4%, 6%, and 10%).

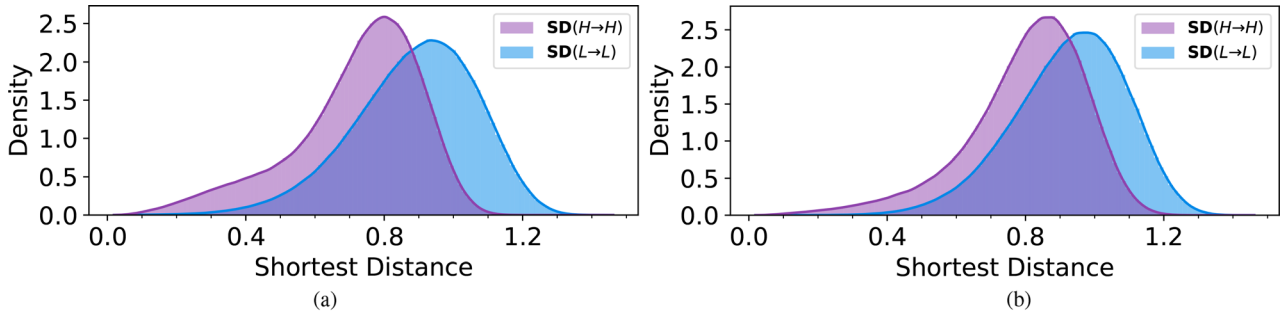


FIGURE C12 Shortest distance (SD) between concepts with different centrality. (a) SD within *H* versus *L* concepts by closeness. Comparison of SD distribution between high closeness centrality concepts versus low closeness centrality concepts. (b) SD within *H* versus *L* concepts by betweenness. Comparison of SD distribution between high betweenness centrality concepts versus low betweenness centrality concepts.

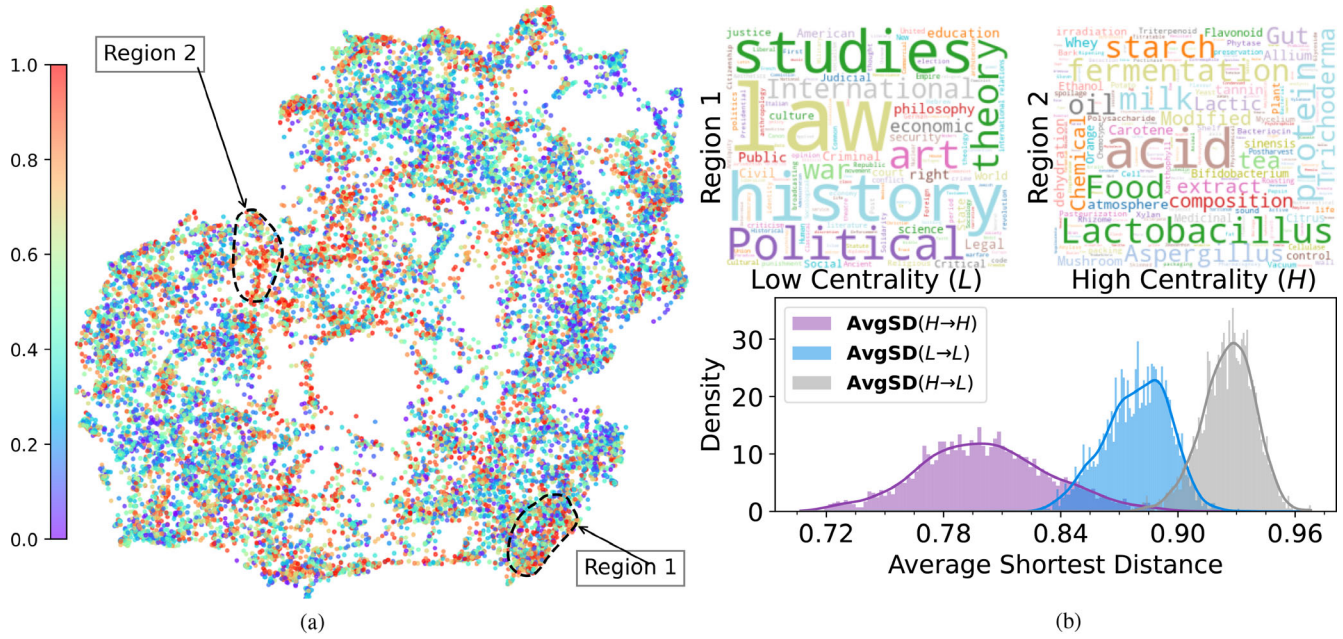


FIGURE C13 Centrality measure the importance of interdisciplinary concepts. (a) Heatmap of concept betweenness centrality. Embedding map of 20,000 key concepts, color-coded by betweenness centrality. (b) Average shortest distance (AvgSD) between 2k *H* and 2k *L* concepts by betweenness. Concept word clouds of two selected regions: Region 1 mainly contains concepts with low betweenness centrality, and Region 2 mainly contains concepts with high betweenness centrality.