



Measuring knowledge complexity in the biomedical domain based on a question-method knowledge representation model

Ming Ma ^a, Jin Mao ^{b,c,*}, Zhentao Liang ^b, Zhejun Zheng ^d, Gang Li ^c

^a Research Institute for Data Management & Innovations, Nanjing University, Suzhou 215011, PR China

^b School of Information Management, Wuhan University, Wuhan 430072, PR China

^c Center for Studies of Information Resources, Wuhan University, Wuhan 430072, PR China

^d School of Information Management, Nanjing University, Nanjing 210023, PR China

ARTICLE INFO

Keywords:

knowledge complexity
question-method model
knowledge dissemination
knowledge representation

ABSTRACT

In nowadays knowledge-driven economy, knowledge complexity plays a crucial role in gaining a competitive advantage. In the biomedical field, this complexity spurs innovation and enables resource monopolization. Previous studies on knowledge complexity have primarily examined the interactions between knowledge units and nonknowledge systems from a macro perspective. These analyses often overlook how the micro-level components of knowledge influence its overall complexity. This study marks a departure from such approaches by conceptualizing biomedical knowledge in terms of questions and methods, as well as proposing a novel method to measure knowledge complexity. This approach emphasizes the exploration of connections between knowledge units by constructing a question-method bipartite network. The validity of our methodology was rigorously tested through controlled experiments involving random networks and a comprehensive review of the relevant literature. Furthermore, this study reveals the relationship between knowledge complexity and dissemination, suggesting that the more complex knowledge is, the more likely it will be cited frequently. Internal knowledge flows within the same research question exhibit greater sensitivity to knowledge complexity than external flows. This study can help demystify the sophisticated scientific knowledge system and provides detailed insights into the complexity of scientific knowledge and dissemination mechanisms.

1. Introduction

Knowledge is widely believed to be the primary driver of modern regional development and economic growth, especially in highly specialized and technology-intensive fields like biomedicine (Nissen, 2019). However, not all types of knowledge are of equal importance within a given domain. The competitive advantage of entities, i.e., nations, regions, or enterprises, often hinges on their possession of high-value, rare, and complex knowledge (Balland & Rigby, 2016). Institutions endowed with such complex knowledge typically spearhead industry innovation, utilizing their sophisticated understanding to develop new treatments, drugs, and technologies.

Taking mRNA vaccines as an example, this technology integrates diverse types of knowledge, from fundamental research in RNA molecular stability to advanced delivery systems using lipid nanoparticles, and sophisticated cold-chain preservation techniques. The

* Corresponding author.

E-mail address: danveno@163.com (J. Mao).

<https://doi.org/10.1016/j.joi.2025.101667>

Received 27 August 2024; Received in revised form 11 April 2025; Accepted 14 April 2025

Available online 24 April 2025

1751-1577/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

companies possessing such comprehensive knowledge bases, like BioNTech, were able to rapidly develop life-saving innovations during the COVID-19 pandemic. Their success not only demonstrated the value of complex knowledge but also generated substantial economic impacts by establishing new industry sectors, creating high-skilled employment opportunities, and fostering regional biotech clusters.

Given the pivotal role of complex knowledge in driving medical innovation, technological advancement, and competition for scarce resources, a critical question arises: *What is complex knowledge, and how can it be effectively measured?*

Although seemingly straightforward, the complexity of knowledge eludes precise definition (Bawden & Robinson, 2015). Regarding information quantity, complex knowledge is characterized by its capacity to perceive, store, and utilize an extensive amount of information (Cohen, 2006). If we view knowledge as an entity, that containing hard-to-access information qualifies as an example of complex knowledge (Ruelle, 1993). By modeling it as a complex system, the complexity of a knowledge system can manifest in the number of components, degree of interaction between these components, and extent of the non-linear behaviors the system exhibits (Theurer, 2014).

From a different perspective, some scholars posit that complexity measures the information required to describe the relationships between elements in a given organized system (Bawden & Robinson, 2015). Accordingly, knowledge complexity can be measured by analyzing the associations between knowledge and non-knowledge system units, e.g., geographic units in geospatial systems. Economists have viewed complexity as a manifestation of a nation's capabilities, concentrating on ownership or possession relationships between geographical areas and the knowledge represented by patents. The complexity of a nation's economy and its knowledge base can be measured by examining the lack of replicability of patents held by a country or region and the uniqueness of patent distribution over different areas (Balland & Rigby, 2016; Balland et al., 2019; Hidalgo & Hausmann, 2009; Pinter & Scherngell, 2022).

However, these studies have not touched the relationships and interactions among knowledge units within a knowledge system. Without a clear understanding of how knowledge units interact and influence each other, research institutions may misallocate resources by underestimating the complexity of unit interactions or overlooking critical connections between different types of knowledge. Moreover, insufficient analysis of these relationships can lead to fragmented knowledge integration, which may hinder scientific advances that require the synthesis of interrelated knowledge units. The lack of systematic mapping of these interactions may also impede the development of holistic research strategies that need to consider the interplay among multiple knowledge units. By modeling a knowledge system as a network, where these units serve as nodes linked by their causal or logical relationships, researchers can not only facilitate a deep understanding of the intricate structure and dynamic evolution of the knowledge system, but also gain a powerful visual framework to identify essential nodes and connection mechanisms, and uncover potential pathways for innovation.

In this study, we focus on the field of biomedicine, in which key knowledge units such as diseases and drugs, along with their interrelationships, outline hierarchical, causal, and logical connections. Therefore, we construct a knowledge network for the biomedical field to model the knowledge system. The essence of knowledge generation can be viewed as question-solving activities (Hutchins, 1977; Van Dijk, 1980), which can be abstractly and logically described as a process centered on questions and continuously seeking new methods to solve them. Constructing knowledge networks of questions and methods and analyzing the relationships between them can provide a more nuanced and detailed understanding of knowledge complexity. In the biomedical field, diseases and the chemicals used to treat them share a distinct logical correspondence. Diseases can be conceptualized as scientific questions, with the corresponding chemicals serving as methods. Accordingly, we present a bipartite knowledge network, in which diseases (questions) and chemicals (methods) form the nodes, and the relationships between them constitute the edges. It should be admitted that the biomedical knowledge modelled by our method is mainly from the question-solving perspective, which maybe the tip of the iceberg.

In such a simplified setting, we attempt to measure the complexity of biomedical knowledge (i.e., diseases and chemicals) in the question-method knowledge representation model by referring to economic complexity indicators. Specifically, this study aims to answer the following research questions:

- (1) How can the complexity of biomedical knowledge be quantified based on the heterogeneous relationships between diseases (questions) and chemicals (methods)?
- (2) How does knowledge complexity influence the dissemination of biomedical knowledge? Are there distinct patterns in how complex biomedical knowledge spread under different conditions?

This study provides a novel perspective of knowledge complexity and the corresponding approach to demystify the sophisticated scientific knowledge system. Our quantitative analysis could further the understanding of knowledge complexity and how it influences the process of knowledge dissemination.

The remainder of this paper is organized as follows. Section 2 provides a brief review of relevant research on knowledge representation models and complexity measurements. Section 3 details the research data and measurement of knowledge complexity, and introduces the steps for analyzing the impact of complex knowledge on its dissemination. Section 4 presents the results of measuring and analyzing the complexity of knowledge. Section 5 discusses the theoretical and practical implications and suggests directions for future research, followed by the conclusions in Section 6.

2. Related work

2.1. Scientific knowledge representation models

Knowledge representation models are designed to delineate the fundamental frameworks of knowledge in scientific literature,

making them measurable, computable, and comparable. These models address the intricate task of encapsulating various forms of scientific knowledge.

Term extraction and function identification are vital in constructing knowledge representation models. Researchers have developed various approaches, including a three-layer network-based model (Liang et al., 2021) and the models that employ machine learning and natural language processing techniques. These models focus on extracting paper title structures, classifying content, and categorizing concepts (Heffernan & Teufel, 2018; Tsai et al., 2013; Wang et al., 2021), achieving a progressively finer granularity of knowledge.

However, these efforts overlook the significance of research questions and methods. Research activities, especially in engineering and technology, are essentially processes of identifying and solving scientific questions (Strübing, 2007). Discourse theory in linguistics posits that the narrative structure of scientific literature centers on *the scientific questions-solving methods* (Teufel, 1999). Several studies have formalized the cognitive processes behind question-solving (Tuomaala et al., 2014) and developed classification models to distinguish phrases as either problems or methods (Heffernan & Teufel, 2018).

Questions define the direction of research, posing challenges that drive the exploration and generation of new knowledge, whereas methods provide systematic approaches to address these questions, support empirical testing, and refine theories. This philosophical view underscores scientific literature as not merely informational but also a dynamic medium that encapsulates the continuous process of problem-solving and methodological application, reflecting the core principles of scientific progression, as discussed by philosophers such as Karl Popper (1959) and Thomas Kuhn (1962).

Therefore, this study aims to construct a knowledge representation model based on questions and methods in scientific literature. By considering the logical relationships between questions and methods, this model transforms complex scientific concepts into tangible, measurable, and well-defined scientific knowledge, thereby offering a novel perspective for examining the intricacies of scientific knowledge.

2.2. Measuring knowledge complexity

In scientific research, several disciplines have interpreted and measured knowledge complexity respectively. In biology, complexity is viewed through an adaptive lens, focusing on the variety and differentiation of system components (McShea et al., 2019). In physics, complexity is closely tied to entropy and information theory, where researchers use information concepts to quantify both complexity and simplicity (Mitchell, 2006).

The exploration of complexity in social sciences can be traced back to the theory of inventive innovation (Fleming, 2001), which suggests that technologies exhibiting greater interdependence manifest higher complexity. In market economies, scholars have shifted their focus from the dependency of technological components to the complexity arising from interactions among individuals in economic activities. Hidalgo and Hausmann (2009) conceptualized trade data between countries as a country-product bipartite network and proposed the iterative reflection method to assess economic and product complexity. This method constructs symmetric metrics through successive iterations: a country's complexity is determined by the average complexity of its exported products, while a product's complexity is inversely related to the average complexity of the countries exporting it.

Economic complexity mirrors the depth and breadth of the knowledge or capabilities needed for product manufacturing (Balland et al., 2019). Consequently, some scholars hypothesize that nations or regions with higher complexity harbor knowledge that is difficult to duplicate. Researchers have quantified the knowledge complexity of various countries and regions by creating national-patent bipartite networks (Balland et al., 2019; Pintar & Scherngell, 2022). Balland and Rigby (2016) characterized knowledge complexity as an objective metric to assess the uniqueness and difficulty of replicating knowledge. Building upon this metric, Janavi et al. (2020) analyzed the structural features of university/nation-scientific publication bipartite networks to assess the disciplinary complexity of universities. Recent research has qualitatively defined the complexity of research questions from two dimensions (research levels and analytical aspects) and explored the relationship between research question complexity and academic impact (Solarino et al., 2024).

This study draws on Balland and Rigby's definition of knowledge complexity (Balland & Rigby, 2016), asserting that in the biomedical field, knowledge characterized by uniqueness and replicability challenges is more complex. Unlike previous studies that identified individual patents or scientific publications as representations of knowledge (Balland & Rigby, 2016; Janavi et al., 2020), we adopt a micro-compositional approach in which diseases and chemicals are considered fundamental knowledge units.

The binary correspondence between them forms the basis for measuring the complexity of knowledge. The complexity of the diseases is elucidated by analyzing the diversity and lack of replicability of chemicals linked to the diseases. Similarly, examining the variety and uniqueness of diseases associated with chemicals reveals the complexity of these chemicals as well. Moreover, the complexity of the knowledge generated to treat a specific disease can be assessed by evaluating the complexity of disease-chemical combinations addressing the same disease. This method of measuring knowledge complexity, based on the dependency and interactions between knowledge units, not only deepens the understanding of the complex knowledge structure in the biomedical field but also uncovers potential innovative pathways.

3. Methodology

This study focuses on measuring and analyzing the complexity of knowledge in the biomedical field, specifically examining diseases and chemicals that exhibit binary correspondence. In this study, diseases and chemicals, as well as combinations of diseases and chemicals, are considered fundamental knowledge units in the knowledge system of the biomedical field. A single disease or chemical

provides fundamental module information about knowledge, whereas disease-chemical combinations offer detailed information about the interactions between knowledge modules.

By constructing knowledge networks of diseases (questions) and chemicals (methods), we have designed and proposed a two-step experimental approach, as depicted in Fig. 1, including measuring the complexity of biomedical knowledge and analyzing the impact of complex knowledge on its dissemination.

In the first step, we construct a *question-method* knowledge representation model that maps the relationships between diseases and chemicals. This model formalizes the measurement of knowledge complexity as a task of extracting network structural information, guided by the principle of iterative reflection. In the second step, we use regression models to explore the intrinsic connections between the complexity of knowledge and its dissemination under various conditions.

3.1. Data collection

The proposed method was applied to the dataset of the Comparative Toxicogenomics Database (CTD),¹ which includes relationships between diseases and chemicals. The CTD is managed and maintained by biomedical professionals, where data managers manually review relevant literature and extract disease-chemical relationships based on molecular-level mechanisms (Davis et al., 2015). Specifically, these relationships are established through chemical-gene interactions. For instance, a chemical A may be linked to disease B because chemical A interacts with gene C, which is associated with disease B (Fig. 2 (a)). This molecular mechanism-based approach provides more reliable biological foundations and ensures the accuracy and reliability of the database.

We used the version of the CTD released in September 2023, which has 17,200 chemicals, 4100 diseases, and 11,422,697 disease-chemical combinations extracted from over 142,000 research articles (Davis et al., 2023). After excluding combinations with unknown publication years and incomplete literature identifiers, we obtained a distilled dataset with 11,195,367 disease-chemical combinations. Fig. 2 (b) illustrates the annual distribution of disease-chemicals combinations from 1946 to 2022.

The data are sparse before 1980. Similarly, only 4964 disease-chemical combinations are provided for 2022. Missing data may affect the mining of structural information regarding the relationships between diseases and chemicals. Therefore, we only included data from 1980 to 2021 in the analysis, with 11,119,593 relationships between diseases and chemicals.

3.2. Question-method based representation model for biomedical knowledge

Knowledge is the product of exploratory activities aimed at solving problems within unknown domains using specific methods as means (Hutchins, 1977; Van Dijk, 1980). For the biomedical field, we model diseases and chemicals as questions and methods with a simplified setting. Accordingly, we develop a *question-method* knowledge representation model, which could be depicted with a *disease-chemical* bipartite network, as shown in Fig. 3(a). In the bipartite network, only the interactions between different types of nodes form edges with weights, which signify the frequency of disease-chemical combinations in the dataset. This structure streamlines the correlation between questions and methods and discloses the backbone knowledge between biomedical diseases and chemicals. This question-method model does not encompass the entire spectrum of knowledge in the biomedical field but is primarily tailored to knowledge involving chemical interventions and treatments for diseases.

The bipartite network of the question-method model can facilitate the computation of knowledge complexity. Mathematically, the bipartite network can be represented by an adjacency matrix M , where the row vectors denote questions (q) and the column vectors denote methods (m). Matrix entries M_{qm} denote the frequency of their combinations, as illustrated in Fig. 3(b). Knowledge (K) can be formalized as an amalgamation of specific questions (q) and the corresponding methods, depicted through the elements and their cell values (the frequency of the question-method combination) in the question's row vector, as shown in Eq. (1):

$$K\{q\} = \{(m, M_{qm}) | m \in \text{Methods}, M_{qm} > 0\} \quad (1)$$

In this study, terms such as diseases and chemicals are sourced from the controlled vocabulary (Medical Subject Headings, MeSH), developed by the United States National Library of Medicine for the biomedical domain. We used the MeSH Tree Structure to organize these mechanism-based associations into a hierarchical structure to ensure semantic consistency and enhance understanding (Liang et al., 2023). This integration of molecular-level evidence with hierarchical semantic frameworks not only ensures the standardization of knowledge representation but also effectively addresses the matrix sparsity issues in network analysis, thereby improving the accuracy of knowledge complexity measurements. The specific operational steps are as follows:

- (1) We obtained all MeSH terms and their corresponding tree structure codes from the MeSH Browser.³
- (2) The MeSH terms were condensed to their third-level parent nodes based on the MeSH tree structure codes. Positioned centrally within the tree, these third-level nodes provide an ideal balance between semantic details and the avoidance of excessively broad conceptual descriptions.

¹ <https://ctdbase.org/>

² <http://ctdbase.org/about/dataStatus.go>

³ <https://meshb.nlm.nih.gov/>

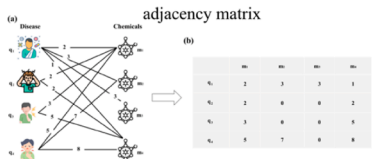
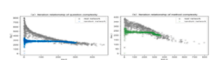
Step1	"Question-Method" knowledge representation model construction	Measurement of knowledge complexity based on iterative reflection	Validation of method effectiveness
Knowledge representation model construction and complexity measurement	<p>"disease-chemicals" bipartite network and the adjacency matrix</p> 	$QCI = k_{q,N} = \frac{1}{k_{q,0}} \sum_m M_{qm} k_{m,N-1}$ $MCI = k_{m,N} = \frac{1}{k_{m,0}} \sum_q M_{qm} k_{q,N-1}$ $KCI_q = \frac{\sum_q (QCI + MCI)}{\sum_q \text{combination number}}$	<p>Real networks vs random networks:</p> <p>The initial iteration results & the final iteration results</p>  <p>Literature validation: question, method, knowledge</p>
Step2	Regression model construction	Knowledge complexity and the external flow of knowledge	Knowledge complexity and intra-knowledge flow
Analysis of biomedical knowledge complexity and its dissemination	<p>Dependent variable: knowledge flow</p> <p>Independent variable: knowledge complexity</p> <p>Control variables: number of authors (<i>autCount</i>), research field (<i>Field</i>)</p>	<p>Under conditions of temporal accumulation and temporal slicing, various regression models were employed to examine the patterns of association between external and internal knowledge flows and knowledge complexity</p> $AvgCite_f = \beta * Complexity + \theta * autCount + \gamma * Field + \delta * Constant$	

Fig. 1. Methodological framework.

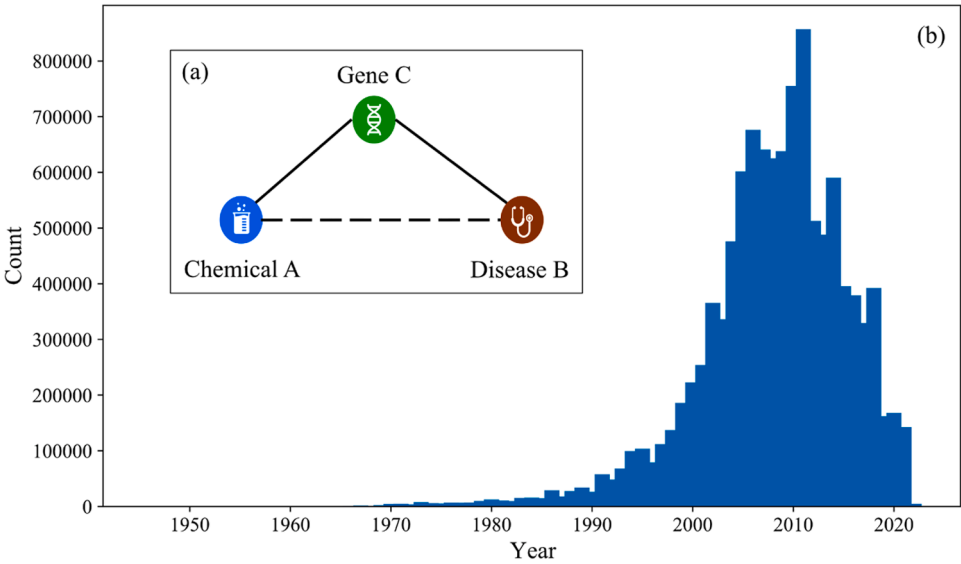


Fig. 2. (a) Disease-chemical relationships² and (b) annual distribution of disease-chemical combinations.

3.3. Measuring the complexity of biomedical knowledge

3.3.1. Indicators of knowledge complexity

Knowledge complexity can be measured based on two dimensions: diversity and uniqueness (Balland & Rigby, 2016). Diversity reflects the variety of methods available for addressing a specific question, whereas uniqueness determines whether a method is unique to that question or more broadly applicable. By examining the relationships between questions and methods, we can measure the complexity of the knowledge units they represent. In addition, the complexity of knowledge that encompasses all combinations of the same question can be obtained by summing up the complexity of these combinations. Following these conceptual foundations, we translate uniqueness and diversity into multi-level structural information extracted from bipartite networks. The detailed calculation process is as follows.

To counteract the distorting effects of data noise on our results, we set M_{qm} threshold to eliminate question-method combinations that appeared only once or twice. Following the methods in Hidalgo and Hausmann (2009) and Tacchella et al. (2012), we filtered out insignificant question-method combinations by highlighting methods that have a revealed advantage (RAM) in solving a specific question. For each question-method combination, RAM is defined as follows (Balassa, 1965):

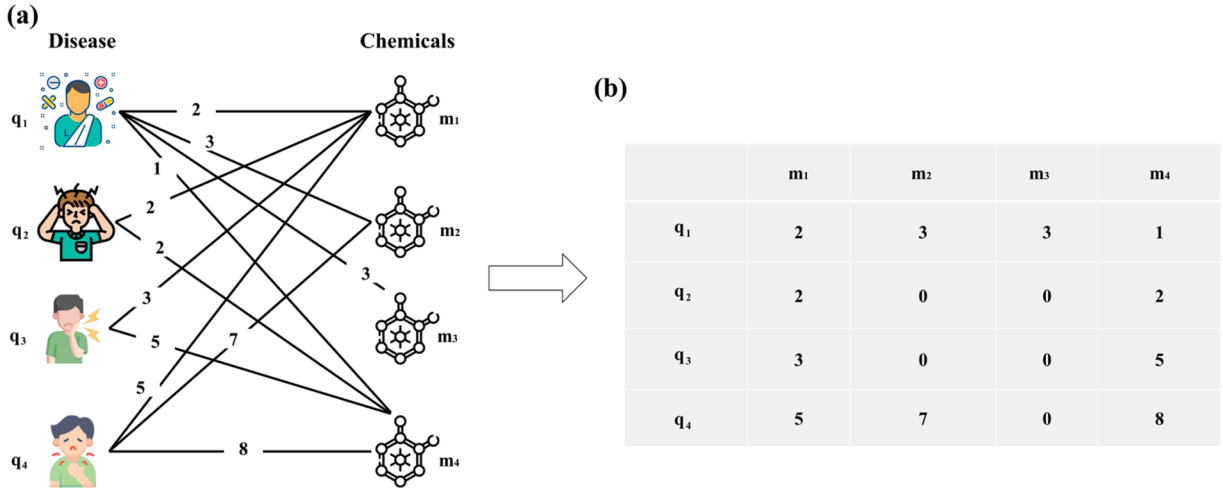


Fig. 3. (a) A question-method bipartite network model characterized by disease-chemicals and (b) the adjacency matrix of the bipartite network.

$$RAM_{qm} = \frac{freq_{q,m} / \sum_m freq_{q,m}}{\sum_q freq_{q,m} / \sum_{q,m} freq_{q,m}} \quad (2)$$

where the numerator represents the share of method m in all solutions for question q , and the denominator represents the share of method m in the solutions for all questions. $RAM_{qm} \geq 1$ indicates that method m has a relative comparative advantage in solving question q , thereby being considered a critical method for this question. Conversely, $RAM_{qm} \leq 1$ signifies that, m is not primarily used for solving q , placing m at a relative disadvantage in addressing this question. Subsequently, based on the RAM threshold, the adjacency matrix M is transformed into a binary matrix M' , as shown in Eq. (3):

$$M' = \begin{cases} 1 & \text{if } RAM_{qm} \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

In M' , if a method m is revealed advantageous for a question q , then $M'_{qm} = 1$; otherwise, $M'_{qm} = 0$. For a given question q , the number of nonzero vectors in its row indicates the number of methods with comparative advantages related to q , reflecting the diversity of the methods used to solve the question. Similarly, for method m , counting the nonzero vectors in its column provides the number of questions for which this method has a comparative advantage, indicating the uniqueness of the method. If multiple questions share the same method that has a comparative advantage, then the uniqueness of this method is lower.

We can then iterate using the question-method bipartite network structure to determine the complexity of both the questions and methods (Hidalgo & Hausmann, 2009). This iteration stops when no change occurs in the complexity of either element, as expressed in Eqs. (4) and (5):

$$QCI = k_{q,N} = \frac{1}{k_{q,0}} \sum_m M_{qm} k_{m,N-1} \quad (4)$$

$$MCI = k_{m,N} = \frac{1}{k_{m,0}} \sum_q M_{qm} k_{q,N-1} \quad (5)$$

where QCI and MCI represent the complexity of a question and method, respectively, and N denotes the number of iterations. For the initial iteration conditions of the model, $k_{q,0}$ and $k_{m,0}$, when $N \geq 1$, $k_{q,0}$ indicates the diversity of methods for solving question q , and $k_{m,0}$ represents the uniqueness of method m . The formula is as follows:

$$k_{q,0} = \sum_m M_{qm} \quad (6)$$

$$k_{m,0} = \sum_q M_{qm} \quad (7)$$

Based on the Eqs. (4) and (5), the complexity of a question-method combination can be determined. Subsequently, we can ascertain the complexity of knowledge (KCI) aimed at solving question q , as shown in Eq. (8):

$$KCI_q = \sum_q (QCI + MCI) / (l \in q) \quad (8)$$

where $(QCI + MCI)$ represents the complexity of a single question-method combination, and the entire numerator reflects the complexity of knowledge centered on question q , KCI . After dividing this by the number of question-method combinations involving q ($l \in q$), we obtained the knowledge complexity, KCI_q , which eliminates the time accumulation effect.

3.3.2. Validation of the measurement

We developed two methods to validate the proposed measurement of knowledge complexity.

(1) Real networks vs. random networks

In the absence of unified standards, comparing real-world data with random data is an effective method of validation (Hidalgo & Hausmann, 2009; Liang et al., 2023). We adopt this approach as our first method of validation. The process is as follows.

- a. Random network construction. In the bipartite network, two-degree sequences correspond to two types of nodes. To generate the corresponding random network, the edges within the real bipartite network were randomly shuffled to maintain the integrity of these degree sequences. This serves as a random simulation of RAMs for scientific questions, disregarding the actual capabilities and resources of the methods used to address the questions.
- b. Results comparison. Based on the iterative reflection principle of the knowledge complexity algorithm (Hidalgo & Hausmann, 2009), if the algorithm successfully extracts meaningful structural information that is distinct from the randomness observed in analogous random networks, the disparity between the refined structural information of higher orders and the original network structure will widen with each iteration.

Therefore, firstly, we use the same methods to mine the structural information of random networks, measuring knowledge complexity. Secondly, we compare the results from the two networks at different iterations of the algorithm to observe whether the real network exhibits differences from the random one. This difference confirms the effectiveness of the algorithm.

(2) Literature validation

The second method of validation involves correlating our findings with those reported in the literature. Specifically, by assessing whether the primary literature underpinning the crucial knowledge identified aligns with expert judgments or established knowledge frameworks, we aim to validate the accuracy and utility of our complexity algorithm. Relevant literature is reviewed to ensure the algorithm effectively captures the intrinsic differences between various types of knowledge. This approach extracts pertinent information from academic and practical sources to verify whether the acquired complexity has been discussed in previous studies or corroborated by relevant evidence. Literature validation aids in understanding the background of knowledge complexity and provides researchers with a more comprehensive understanding and preliminary assessment of its validity and credibility.

3.4. Investigating the relationship between the complexity of biomedical knowledge and its dissemination

In this study, *knowledge dissemination* represents the transfer of knowledge from the cited to the citing, including internal and external knowledge flows. We selected 1990, 2000, and 2010 as focus years because they represent three distinct periods in biomedical research development, spanning from pre-genomic era to advanced biotechnology age. These decadal intervals also provide sufficient temporal separation to observe technological evolution while ensuring complete citation data availability.

To measure knowledge dissemination, we tracked the citations of each piece of knowledge over the following five years, as this window typically captures the majority of citations. We then constructed a regression model to analyze the relationship between the complexity of biomedical knowledge and its dissemination.

3.4.1. Internal and external knowledge flow

Knowledge comprises numerous knowledge units. *Internal knowledge flow* refers to citations between knowledge units within the same research question (e.g., neuroscience papers citing other neuroscience papers), while *external knowledge flow* represents citations across different questions (e.g., neuroscience papers citing computer science papers), reflecting the depth and breadth of knowledge dissemination respectively, as shown in Fig. 4.

Citation frequency is commonly used as an objective indicator of the strength of knowledge flows (Zhai et al., 2018). We first gathered CTD citation data from the National Institutes of Health's Open Citation Collection (iCite et al., 2019) using PMIDs and organized them into three datasets (1990, 2000, and 2010) based on the publication dates of the documents. Next, we calculated the average citation frequency (internal and external citations) for each knowledge unit to measure flow strength. The citation frequency of a knowledge unit corresponds to the citation frequency of the document carrying it. The strength of the knowledge flow is represented by the average flow strength of the contained knowledge units, as shown in Eqs. (9) and (10):

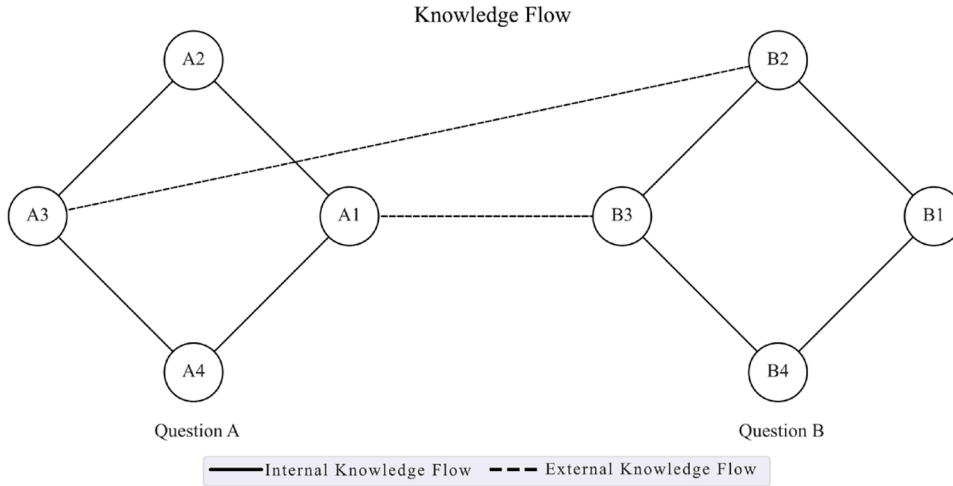


Fig. 4. Schematic representation of knowledge flows.

$$iKFS_q = \sum_{u=1}^{l(q,i)} C_u / l \in (q, i) \quad (9)$$

$$eKFS_q = \sum_{u=1}^{l(q,e)} C_u / l \in (q, e) \quad (10)$$

where $iKFS_q$ and $eKFS_q$ denote the internal and external flow strength of the knowledge centered around question q , respectively. $l \in (q, i)$ and $l \in (q, e)$ denotes the number of knowledge units involved in internal citations and external citations, respectively. C_u denote the citation frequency of the knowledge unit.

3.4.2. Regression model construction

An ordinary least squares regression model was applied to identify significant correlations between knowledge complexity and dissemination, with knowledge complexity as the independent variable and knowledge flow strength as the dependent variable.

For the control variables, we primarily considered the size of the knowledge production teams and the discipline of the knowledge. The size of the knowledge production team refers to the average number of authors of the paper in which the knowledge is located. We classify the disciplinary of the knowledge based on the field of the paper where the knowledge is found, using the classification standards provided by Science-Metric.⁴ Considering these factors, the regression model is as follows:

$$AvgCite_f = \beta * Complexity + \theta * autCount + \gamma * Field + \delta * Constant \quad (11)$$

where *Complexity* denotes the degree of complexity, *autCount* represents the average number of authors, *Field* refers to the field of the publication where the knowledge is found, and *Constant* indicates a constant term. The coefficients β , θ , γ , and δ correspond to the aforementioned factors, respectively. Since complex knowledge is more valuable and challenging to understand and master, we do not set specific expectations for the sign of β . The number of authors has been proven to affect knowledge flow positively (Klug & Bagrow, 2016; Wu et al., 2019); hence, we anticipate θ to be a positive coefficient. Owing to the significant heterogeneity among different disciplines (Ke et al., 2023), we cannot predetermine the specific sign of γ .

4. Results

4.1. Results on the complexity of biomedical knowledge

Table 1 shows detailed information on the top three ranked diseases, chemicals, and disease-chemical combinations in terms of complexity. Disease-chemical combinations were sourced from the most complex combinations within the top three most complex knowledge types.

Statistics on all questions, methods, and knowledge with complexity from 1980 to 2021, along with the distribution of the top 20 complexities, were operated to examine whether they consistently attracted attention in the previous periods. As shown in Fig. 5, the occurrence of both increased continuously over time. The growth indicates dynamic changes in the complexity of research content and

⁴ <https://science-matrix.com/classification/>

Table 1

Top 3 diseases, chemicals, and their combinations with the greatest complexity.

Type	Tree Structure Code	Name	Complexity
Diseases (Questions)	C01.920.937	Trypanosomiasis, African	2.43
	C23.550.767	Postoperative Complications	1.63
	C14.907.725	Reperfusion Injury	1.63
Chemicals (Methods)	D23.767.550	Lipofuscin	1.75
	D04.345.674	Polyketides	1.56
	D12.644.679	Peptoids	1.56
Diseases-Chemicals (Questions-Methods)	C01.920.937-D02.129.500	Trypanosomiasis, African-Melarsoprol	3.47
	C23.888.592-D23.767.550	Neurologic Manifestations-Lipofuscin	3.34
	C12.200.777-D03.132.722	Urologic Diseases-Sparteine	3.11

the academic community's concentrated interest and effort in these complex areas of study. Top 20 complexity methods, nearly absent before 1990, subsequently surged, demonstrating a shift toward more complex research methods to address the persistently complex questions.

Second, we conducted a retrospective analysis of the complexity trends of the top 20 questions, methods, and disease-chemical combinations from 1980 to 2021, examining their rankings across the previous periods, as shown in Fig. 6. The results demonstrate that complexity is a dynamic attribute. Knowledge units, including questions, methods, and question-method combinations, identified as complex in recent times were less so in the past, illustrating the cumulative and evolving nature of knowledge development over time.

4.2. Validation of the measurement

4.2.1. Real networks vs. random networks

We compared the complexity results between the real network and random networks after the first iteration of the algorithm. As shown in Fig. 7, the complexity of questions and methods in the real network demonstrates a significant negative correlation with its original network structure, while no such correlation was observed in random networks.

Furthermore, examining the overall complexity patterns of different degree sequences (questions and methods) within the bipartite network after the final iteration helps ascertain whether the complexity patterns of the real network distinctively diverge from those observed in random networks. Fig. 8 illustrates the comparative analysis results.

Fig. 8 (a) and (b) compare the complexity distributions of questions between real and random networks. The random network exhibits a right-skewed distribution (skewness: 2.98), suggesting a predominance of simple questions. In contrast, the real network approximates a normal distribution (skewness: -0.27). This pattern aligns with research realities where complex questions carry higher risks (Bateman & Hess, 2015), while overly simple questions may face publication challenges, resulting in a tendency toward moderate complexity in actual research (Liang et al., 2023).

The distributions of method complexity, shown in Fig. 8 (c), reveal distinct patterns. Random networks display a right-skewed distribution (skewness: 2.29), indicating limited complexity. Conversely, real networks show a left-skewed distribution (skewness: -1.18), reflecting the scientific community's preference for specialized methods to address specific complex questions (Ericsson & Charness, 1997). This disparity highlights how random networks fail to capture the intentional matching of methods to questions in actual scientific practice. In random networks, the introduction of hypothetical scenarios results in complex methods originally designed for specific questions and then applied to a broader range of questions. This generalized application dilutes the complexity of these methods.

Similarly, Fig. 8 (d) demonstrates contrasting knowledge complexity distributions between network types. Real networks exhibit a left-skewed distribution (skewness: -0.68), while random networks show a right-skewed pattern (skewness: 0.44). These distinct

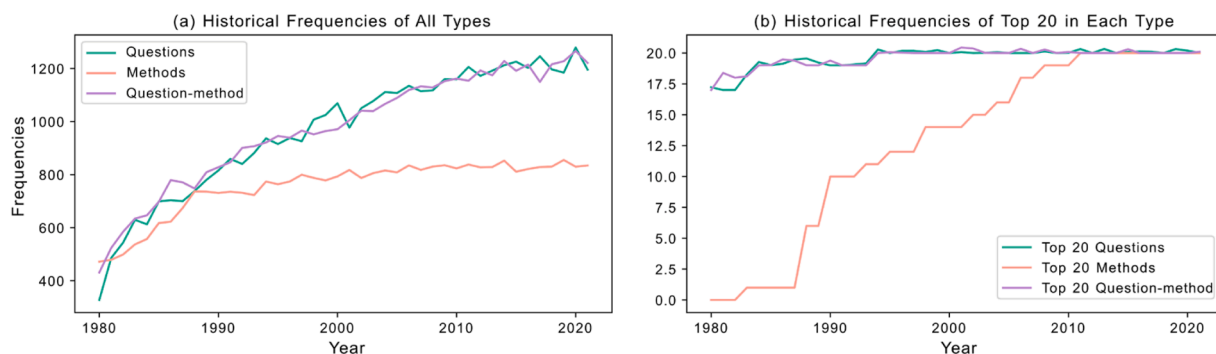


Fig. 5. The distribution over previous periods of questions, methods, and question-method combinations from 1980 to 2021.

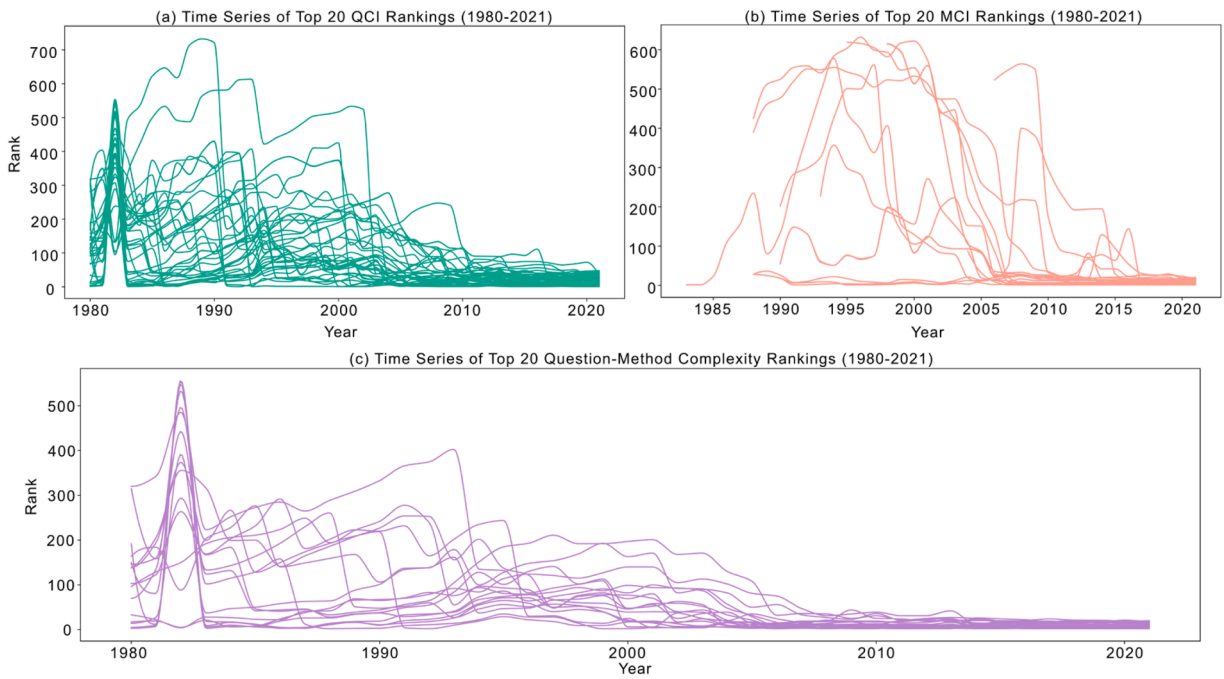


Fig. 6. The evolution of complexity rankings of questions, methods, and question-method combinations.

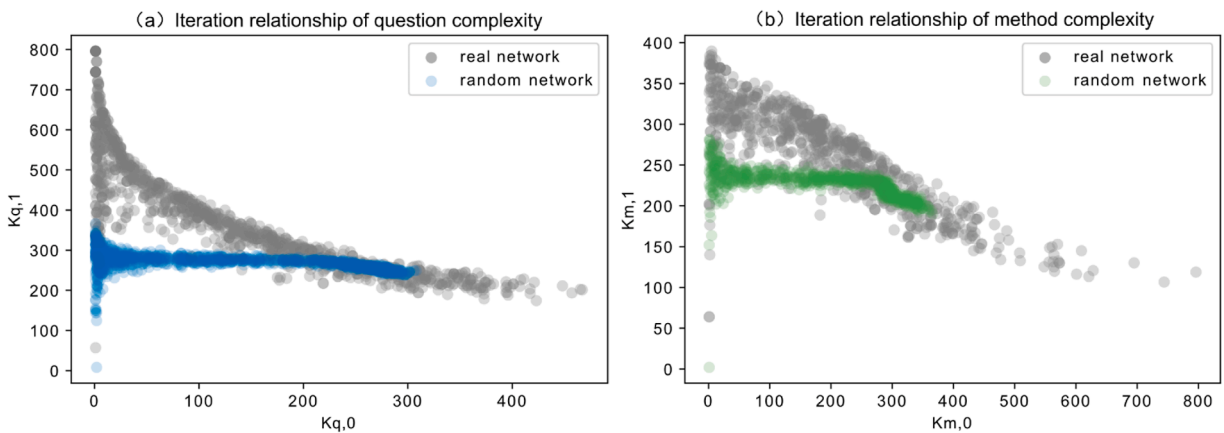


Fig. 7. Scatter plot for comparing the differences between the first iteration ($k_{q,1}$, $k_{m,1}$) and the initial values ($k_{q,0}$, $k_{m,0}$) in real-world and random networks (1980–2021).

distributions validate the effectiveness of our complexity algorithm in detecting real nonlinear patterns of question-method relationships and distinguishing the patterns from random associations. The algorithm successfully identifies structural nuances in real networks that are not mirrored in comparable random networks, thus demonstrating its operational and iterative efficacy, which underscores the method soundness of our complexity algorithm.

4.2.2. Literature validation

We conducted a detailed analysis of the related literature in response to the complex questions, methods, and combinations identified. This review substantiates the validity of the complexity rankings derived from the complexity algorithm.

- (1) Complex questions. *Trypanosomiasis, African*, also known as sleeping sickness, is a fatal disease prevalent in Africa. The World Health Organization's monitoring and control initiatives underscore the complexity of managing this disease.⁵ Surgical

⁵ [https://www.who.int/zh/news-room/fact-sheets/detail/trypanosomiasis-human-african-\(sleeping-sickness\)](https://www.who.int/zh/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness))

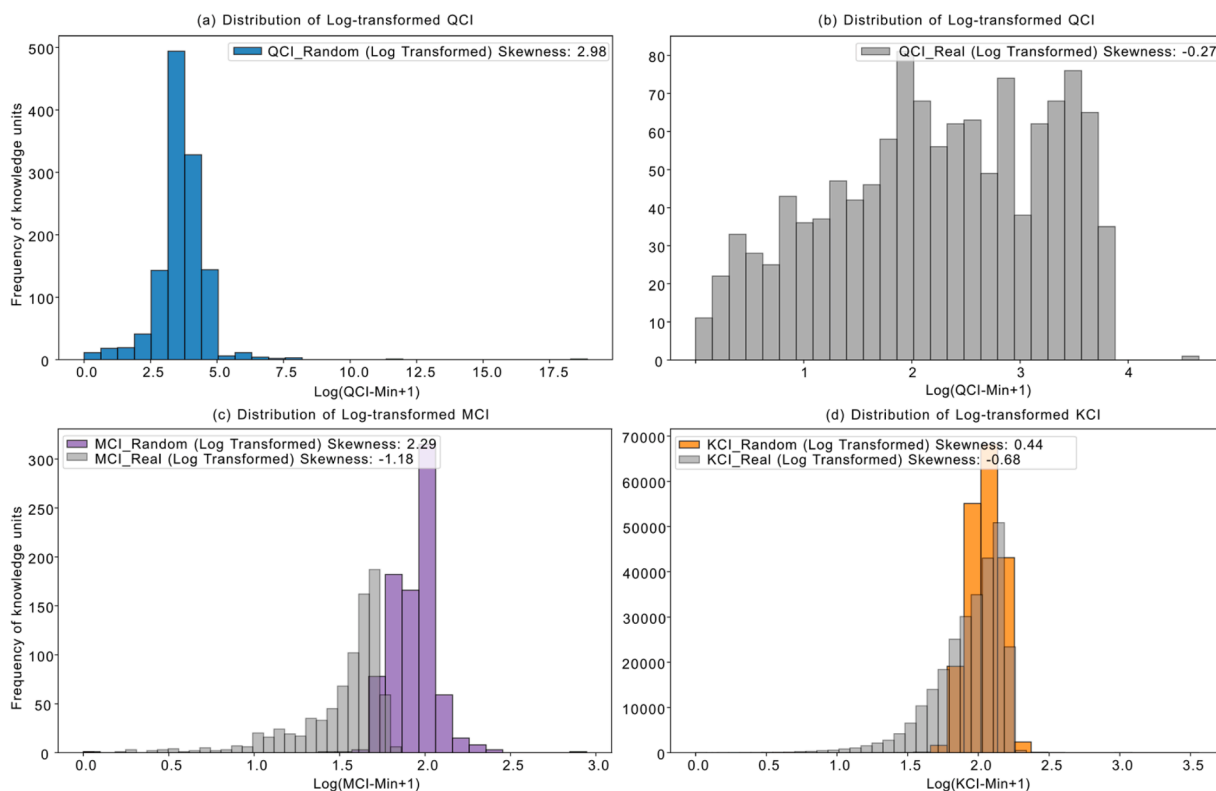


Fig. 8. Comparison of overall complexity distribution patterns of random and real networks.

interventions, while critical for life-threatening conditions, are complicated by the potential for *postoperative complications*, highlighting the difficulty in defining and grading these complications (Dindo et al., 2004) and their effective monitoring and management (Toner & Hamilton, 2013). Characterized by a high mortality rate due to the reperfusion of ischemic organs, the complex mechanisms of *reperfusion injury* render it a subject of ongoing debate (Gross & Auchampach, 2007). Advanced drug delivery methods promise more effective treatments owing to enhanced biocompatibility, controllable operations, and targeted delivery (Li et al., 2024).

- (2) Complex methods. *Lipofuscin Research* is known as the aging pigment. Investigating this "aging pigment" elucidates the intricate mechanisms of aging and related neurological diseases such as age-related macular degeneration (Di Guardo, 2015). *Polyketides* serve as a crucial source of various high-value chemicals, including antibiotics. The regulation of their biosynthesis through gene editing presents a sophisticated approach to discovering new therapeutic compounds (Weber et al., 2015). In the quest for early and accurate diagnosis of diseases such as cancer, *peptoid-based* diagnostic probes have emerged as one of the most effective methods because of their structural advantages (Giorgio et al., 2023).
- (3) Complex combinations. *Trypanosomiasis*, *African-Melarsoprol*. The treatment of *Trypanosomiasis* with *Melarsoprol* highlights the complexity of optimizing treatments for central nervous system diseases owing to severe side effects (Kennedy, 2013). In *Neurologic Manifestations-Lipofuscin*, *Neurologic Manifestations* involve various neurological diseases, including epilepsy and drug-induced akathisia (Mattoo et al., 2003), illustrating the challenges in early diagnosis and the importance of *Lipofuscin* in treatment strategies. *Urologic Diseases-Sparteine*. The side effects of arrhythmia medications in urinary system disease treatment exemplify the complexity of managing drug interactions and the need for personalized research to enable tailored treatments (Leung et al., 2009).

4.3. Impact of the complexity of knowledge on its dissemination

Knowledge is accumulative and dynamic, with its complexity varying as new questions arise and unique methods are developed. Therefore, we examined the relationship between knowledge complexity and knowledge flow within the context of temporal accumulation. The complexity of knowledge in the focus years is calculated cumulatively; for example, the complexity of knowledge in 1990 is based on the question-method network from 1981 to 1990.

In the analysis, we set the team sizes of 1–5 as the control group for the variable "autCount." This approach enabled us to assess the contributions of larger teams. For the variable "Field," given the ambiguity in the specific impact of various fields on knowledge complexity and flow, it is logical to use "Article-level Classification" provided by Science-Metric for comprehensive fields as the control

group. A comparative analysis of highly specialized and more comprehensive fields elucidates the dynamic impact of specialization in fields on knowledge complexity and flow. Table 2 presents the correlation coefficients and descriptive statistics of the variables employed in this study.

4.3.1. Knowledge complexity and internal knowledge flow

Tables 3 present the regression results of knowledge complexity and internal knowledge flow under different models. These results consistently show a significant positive correlation between the complexity of knowledge and the frequency of its internal citations across different periods, indicating that more complex knowledge tends to receive greater internal citations. Moreover, this positive relationship persists even when controlling for the number of authors and fields.

Regarding the number of authors, the data demonstrate increased internal knowledge flow strength with larger team sizes, which is particularly evident in 2010 for teams with 21–25 and 26–30 members (coefficients of 1.960 and 1.882, respectively; both with $p < 0.001$). While the impact of team size on internal knowledge flow may not be significant for teams with 6–10 members in some analyses, the influence of larger research teams becomes increasingly significant over longer periods. This trend could be due to the ability of larger teams to combine a greater pool of expertise and resources, thus enhancing the visibility and citation frequency of their knowledge outputs.

The results of Model (3) further suggests that the specific field of knowledge has little long-term effect on its internal flow, with a few exceptions in 1990, showing isolated negative effects. For instance, "Biology" and "Clinical Medicine" exhibited negative significant coefficients compared to comprehensive journals, hinting at a reduction in internal knowledge flow in these fields. However, the influence of fields on internal knowledge flow did not consistently manifest as significant negative or positive effects across periods.

4.3.2. Knowledge complexity and external knowledge flow

The results in Table 4 are generally consistent with those in Table 3, illustrating a consistent pattern in which knowledge of greater complexity tends to garner more external citations across different periods. This correlation persists even after adjusting the number of authors and research fields. A positive correlation exists between the number of knowledge authors and the strength of the external knowledge flow, whereas the impact of the research field remains minimal. Synthesizing the insights from Table 3 and Table 4 reveals the following:

- (1) Increasing complexity coefficients over time. From 1990 to 2010, the complexity coefficients for both internal and external knowledge flows have exhibited an increasing trend. This increment mirrors the natural evolution and accumulation within research fields, which, in turn, enhances the dissemination and application of knowledge.
- (2) Higher sensitivity of internal knowledge flow to complexity. The internal knowledge flow demonstrates greater sensitivity to complexity than the external flow. For instance, the complexity coefficients for internal flows consistently outpace those for external flows at comparable time intervals. For example, the coefficients of Model (3) in Table 3 are greater than those in Table 4 by approximately 40.21 % in 1990, 46.06 % in 2000, and 22.50 % in 2010. This discrepancy suggests that complex knowledge, owing to its spatial viscosity, is more readily transmitted within rather than across different knowledge spheres (Balland & Rigby, 2016). Internal flows benefit from shared contexts and prior knowledge, facilitating the decoding and utilization of complex information within organizations or close-knit professional networks. Conversely, external knowledge flows spanning diverse knowledge boundaries necessitate broader contextual interpretation and restructuring, potentially delaying the assimilation of complex knowledge.

4.3.3. Time-slicing method for analyzing knowledge complexity and flow

While measuring complexity cumulatively, we recognize the potential bias introduced by cumulative effects, which may disadvantage new knowledge. New knowledge with fewer methods to solve questions than existing knowledge results in relatively low complexity scores.

To mitigate this bias, we adopted an annual assessment strategy to independently evaluate the diversity and uniqueness of knowledge each year. This approach neutralizes the temporal cumulative disparities between new and established knowledge. We then reexamined the relationship between complexity and fluidity using the same analytical method. The results presented in Table 5 and Table 6 show the correlations between the complexity and both the internal and external knowledge flow under this time-slice approach. These findings are generally consistent with those obtained using the cumulative method.

Table 2

The correlation matrix and descriptive statistics of applied variables.

	Variable	Complexity	autCount	Mean	S.D.	Min	Max
1990 ($N = 13,925$)	Complexity	1.00	–	-1.14	1.36	-5.69	2.12
	autCount	0.15	1.00	4.09	1.48	1.00	10.00
2000 ($N = 19,510$)	Complexity	1.00	–	-0.73	1.39	-6.12	1.61
	autCount	0.35	1.00	5.99	2.83	1.00	24.00
2010 ($N = 24,445$)	Complexity	1.00	–	-0.16	1.13	-5.90	2.18
	autCount	0.08	1.00	9.43	4.97	1.00	29.00

Table 3
Knowledge complexity and internal knowledge flow (cumulative).

Model	(1)			(2)			(3)		
	1990	2000	2010	1990	2000	2010	1990	2000	2010
Complexity	0.959*** (0.032)	1.502*** (0.032)	1.470*** (0.042)	0.964*** (0.032)	1.427*** (0.034)	1.450*** (0.041)	0.945*** (0.034)	1.408*** (0.036)	1.421*** (0.042)
autCount									
6–10				-0.112 (0.114)	0.456*** (0.091)	0.658*** (0.110)	-0.125 (0.114)	0.464*** (0.091)	0.668*** (0.111)
11–15					0.723*** (0.198)	1.109*** (0.125)		0.733*** (0.199)	1.116*** (0.126)
16–20					1.184** (0.365)	2.006* (0.168)		1.188** (0.366)	2.007*** (0.168)
21–25					3.813* (1.683)	1.933*** (0.176)		3.825*** (1.685)	1.960*** (0.177)
26–30						1.874*** (0.532)			1.882*** (0.532)
Field									
Agriculture, Fisheries, & Forestry							-0.538 (0.289)	-0.090 (0.220)	-0.057 (0.195)
Biology							-1.482* (0.614)	-0.570 (0.550)	0.187 (0.298)
Biomedical Research							-0.750** (0.249)	-0.128 (0.165)	-0.209 (0.145)
Clinical Medicine							-0.663** (0.241)	-0.199 (0.159)	-0.248 (0.144)
Psychology & Cognitive Sciences							-0.934** (0.348)	0.067 (0.237)	-0.003 (0.237)
Public Health & Health Services							-0.383 (0.269)	0.001 (0.188)	0.051 (0.165)
Constant	5.472*** (0.052)	6.523*** (0.044)	6.776*** (0.042)	5.496*** (0.058)	6.228*** (0.068)	5.940*** (0.092)	6.092*** (0.227)	6.326*** (0.146)	6.010*** (0.138)
R²	0.427	0.564	0.369	0.428	0.574	0.431	0.438	0.577	0.436

Standard errors in parentheses.

*** $p < 0.001$

** $p < 0.01$

* $p < 0.05$

5. Discussion

This section presents robustness tests to validate the accuracy and reliability of our method and regression findings, followed by a discussion of potential implications, limitations, and future directions.

5.1. Robustness tests: expert validation and alternative model specifications

5.1.1. Expert validation

To further assess the accuracy of our method, we implemented expert validation through pairwise comparisons in addition to random network comparison and literature validation. We invited five experts in biological sciences to manually compare the complexity of selected pairs of knowledge units. Only disease-chemical combinations were chosen in that their complexity may be easier to perceive by experts than either diseases or chemicals. We would like to know to what degree the comparison results by our methods keep the same with the results by experts. Since the knowledge units with similar complexity may be hard to distinguish by experts, we stratified the knowledge units into 10 bins with equal interval of knowledge complexity score. From each bin, 10 knowledge unit was randomly sampled, resulting in 100 knowledge units. Then, we sampled 50 pairs of knowledge units with the constraint that the two knowledge units of a pair should come from different bins. The comparison was transformed into comparing the 50 pairs of bins. Experts were invited to determine which knowledge unit of each pair exhibits greater complexity. The comparison results by our method can be easily obtained according to the knowledge complexity score. The 50 pairs of knowledge units were assigned to one domain expert for independent assessment. For each expert, we conducted the above sampling process. Our validation design attempts to ensure that each pair of bins could expect at least 5 comparisons.

Prior to formal evaluation process, experts received standardized training: (1) Theoretical instruction on complexity metrics and computations; (2) 20 practice evaluations with consensus discussions for disputed cases. By comparing these expert assessments with our results, we could objectively evaluate the accuracy of our method. As shown in Fig. 9(a), among the five experts' evaluations, the lowest accuracy rate was 54 % (27 correct assessments), while the highest reached 78 % (39 correct assessments).

To enhance the reliability of our validation results, we incorporated quality control mechanisms into the assessment design. Specifically, we provided each expert with an extra 20 % duplicate samples and implemented cross-validation of identical sample pairs among different experts. The inter-rater agreement coefficients were calculated to assess the degree of consensus among expert

Table 4
Knowledge complexity and external knowledge flow (cumulative).

Model	(1)			(2)			(3)		
	1990	2000	2010	1990	2000	2010	1990	2000	2010
Complexity	0.709*** (0.030)	1.191*** (0.025)	1.254*** (0.028)	0.707*** (0.030)	0.981*** (0.034)	1.450*** (0.041)	0.674*** (0.031)	0.964*** (0.029)	1.160*** (0.030)
autCount									
6–10				0.538*** (0.126)	1.141*** (0.083)	0.971*** (0.092)	0.545*** (0.126)	1.142*** (0.084)	0.967*** (0.092)
11–15					2.031*** (0.179)	0.918*** (0.117)		2.041*** (0.179)	0.903*** (0.118)
16–20					0.699* (0.349)	2.219*** (0.172)		0.724* (0.349)	2.207*** (0.172)
21–25					2.947*** (0.727)	1.819*** (0.197)		2.991*** (0.725)	1.845*** (0.197)
26–30						1.947*** (0.538)			1.955*** (0.537)
Field									
Agriculture, Fisheries, & Forestry							-0.333 (0.282)	-0.330 (0.198)	-0.203 (0.165)
Biology							-0.302 (0.634)	-0.725 (0.525)	0.035 (0.276)
Biomedical Research							-0.310 (0.248)	-0.350* (0.144)	-0.310 (0.118)
Clinical Medicine							-0.500* (0.239)	-0.324* (0.139)	-0.417* (0.117)
Psychology & Cognitive Sciences							-0.556 (0.352)	0.082 (0.215)	-0.559 (0.197)
Public Health & Health Services							0.116 (0.263)	-0.061 (0.167)	-0.245 (0.137)
Constant	5.533*** (0.056)	6.596*** (0.043)	7.236*** (0.034)	5.471*** (0.058)	5.820*** (0.066)	6.339*** (0.083)	5.787*** (0.227)	6.061*** (0.135)	6.584*** (0.120)
R²	0.266	0.504	0.428	0.274	0.558	0.473	0.288	0.564	0.477

Standard errors in parentheses.

*** $p < 0.001$

** $p < 0.01$

* $p < 0.05$

judgments, resulting in an average inter-rater agreement of 64 %, which is greater than 60 % thus reflects the robustness of our measurement of knowledge complexity (McHugh, 2012).

5.1.2. Alternative model

In previous analyses, we have conducted robustness tests through alternative measurements of the dependent variable and nested experiments to verify our findings. To further validate the robustness of our regression results, we employed a negative binomial regression model as an additional method, considering the potential over-dispersion characteristics of the dependent variable. The results from this alternative model consistently supported our main findings, demonstrating the stability of our conclusions across different analytical approaches. The detailed results of negative binomial regression analyses are presented in Table A and Table B in the Supplementary Materials.

5.2. Implications

Theoretically, this study provides methodological contributions to the understanding of knowledge complexity in biomedical science. At the microscopic level, we introduce a quantitative analytical framework based on bipartite question-method networks, enabling precise characterization of disease-chemical relationships and exploration of potential structural patterns. This bibliometric-based approach transforms complex biomedical phenomena analysis into systematic observations of research outcomes. Unlike previous studies primarily focused on macro-level knowledge systems, our method enables detailed analysis of knowledge's internal structure and its impact on complexity. At the macroscopic level, our framework utilizes iterative reflection methods to reveal the complexity of biomedical knowledge systems, identifying hierarchical knowledge structures and critical nodes that conventional approaches often miss, thereby enabling more efficient research prioritization and resource allocation. Furthermore, while this study focuses on biomedicine, the analytical framework demonstrates potential applicability to other knowledge-intensive domains, enriching the theoretical foundation for knowledge complexity research.

Practically, in nowadays competitive scientific environment, where resources are limited (Ke et al., 2023), more comprehensive value assessment methods are needed for research resource allocation to address the growing biomedical challenges. Traditional metrics like journal impact factors and h-index, although widely used in today's competitive scientific environment, are inherently biased towards measuring research output quantity and citations, rather than capturing the overall quality and complexity of research.

Table 5
Knowledge complexity and internal knowledge flow (annual).

Model	(1)			(2)			(3)		
	1990	2000	2010	1990	2000	2010	1990	2000	2010
Complexity	0.431*** (0.050)	1.356*** (0.061)	1.136*** (0.114)	0.506*** (0.050)	1.230*** (0.062)	1.075*** (0.100)	0.700*** (0.038)	1.200*** (0.042)	1.117*** (0.039)
autCount									
6–10				1.111*** (0.208)	0.629*** (0.165)	0.236 (0.262)	1.004*** (0.205)	0.626*** (0.166)	0.159 (0.264)
11–15					1.989*** (0.384)	2.133*** (0.291)		1.971*** (0.402)	2.100*** (0.292)
16–20					2.436*** (0.656)	1.806*** (0.348)		2.529*** (0.663)	1.743*** (0.351)
21–25						3.567*** (0.299)			3.557*** (0.301)
26–30						2.226*** (0.456)			2.228*** (0.459)
Field									
Agriculture, Fisheries, & Forestry							-2.951** (0.864)	0.516 (1.089)	-0.135 (0.491)
Biology									0.888 (0.776)
Biomedical Research							-0.933 (0.611)	0.095 (0.352)	0.054 (0.282)
Clinical Medicine							-1.448 (0.591)	-0.141 (0.343)	-0.283 (0.280)
Psychology & Cognitive Sciences							0.299 (0.556)	0.299 (0.556)	0.563 (0.610)
Public Health & Health Services							0.140 (0.422)	0.140 (0.422)	0.358 (0.353)
Constant	4.686*** (0.104)	6.328*** (0.093)	5.968*** (0.095)	4.479*** (0.108)	5.785*** (0.136)	4.604*** (0.227)	5.540*** (0.578)	5.724*** (0.342)	4.599*** (0.316)
R²	0.152	0.483	0.127	0.206	0.521	0.371	0.291	0.532	0.386

Standard errors in parentheses.

*** $p < 0.001$

** $p < 0.01$

* $p < 0.05$

The limitation often leads to resource allocation decisions that may overlook complex biomedical research, hindering further development in related fields.

The imperative to optimize the allocation of finite research resources necessitates the implementation of a systematic, transparent, and equitable evaluation framework. In this context, our knowledge complexity measurement method offers a more effective tool for research evaluation. It enables funding agencies to make evidence-based resource allocation decisions, research institutions to identify promising research trajectories, and policymakers to develop robust knowledge dissemination mechanisms. As articulated by [Stephan \(2012\)](#), the integration of complexity metrics into research funding frameworks has the potential to catalyze substantive scientific advancement and foster more profound scholarly impact.

5.3. Limitations and future work

Firstly, our bibliometric method of studying disease-chemical relationships represents an indirect observation of research outputs rather than direct investigation of biomedical phenomena. The study only captures *published and investigated* disease-chemical relationships, not all possible biomedical knowledge connections. This literature-based paradigm may be influenced by research trends and publication preferences. Moreover, bibliometric data serves as a projection of research activities, reflecting the scientific community's cognitive understanding rather than the actual strength of relationships in biological systems.

Secondly, as a tool for identifying the primary methods for answering questions, the revealed advantage approach may overlook methods with potential but less research frequency, potentially delaying the recognition of valuable knowledge. Despite validating our approach against random networks and through a literature review, the method could benefit from refinement to consider external factors and endogenous environments. A more nuanced approach to the identification of methods with revealed advantage (RAM) can improve the timeliness and accuracy of knowledge complexity assessments ([Rousseau, 2018](#); [Rousseau & Yang, 2012](#)).

Third, the current bipartite network structure can only express direct associations between diseases and chemicals, unable to capture more complex relationships such as specific molecular mechanisms of chemical actions, synergistic effects among multiple chemicals, molecular pathways and regulatory networks of diseases, and complete action chains of drug-target-pathway-disease interactions. This structural limitation constrains our ability to fully understand the intricate relationships within biomedical knowledge systems.

Future research can enhance our understanding of knowledge complexity and dissemination through the following approaches.

Table 6
Knowledge complexity and external knowledge flow (annual).

Model	(1)			(2)			(3)		
	1990	2000	2010	1990	2000	2010	1990	2000	2010
Complexity	0.249*** (0.040)	1.025*** (0.053)	1.403*** (0.079)	0.382*** (0.036)	0.966*** (0.055)	1.410*** (0.076)	0.342*** (0.037)	0.959*** (0.057)	1.382*** (0.079)
autCount									
6–10				2.005*** (0.160)	0.521** (0.155)	0.289 (0.190)	1.838*** (0.166)	0.498** (0.156)	0.344 (0.189)
11–15					0.841** (0.333)	0.917*** (0.238)		0.891** (0.333)	0.943*** (0.235)
16–20					0.769 (0.651)	2.127*** (0.259)		0.861 (0.654)	2.168*** (0.258)
21–25						2.364*** (0.285)			2.461*** (0.284)
26–30						1.835*** (0.478)			1.803*** (0.473)
Field									
Agriculture, Fisheries, & Forestry							-1.751* (0.766)		-0.833* (0.420)
Biology									0.298 (0.692)
Biomedical Research							-1.058 (0.609)	0.223 (0.317)	-0.639** (0.232)
Clinical Medicine							-1.490* (0.598)	0.098 (0.304)	-0.701** (0.231)
Psychology & Cognitive Sciences							-0.306 (0.810)	0.544 (0.524)	-1.520 (0.573)
Public Health & Health Services							-0.550 (0.644)	0.639 (0.408)	-1.009 (0.288)
Constant	4.413*** (0.095)	5.807*** (0.093)	6.102*** (0.074)	4.153*** (0.085)	5.467*** (0.130)	5.376*** (0.155)	5.392*** (0.590)	5.278*** (0.299)	5.920*** (0.231)
R²	0.076	0.368	0.268	0.305	0.382	0.370	0.341	0.398	0.397

Standard errors in parentheses.

*** $p < 0.001$

** $p < 0.01$

* $p < 0.05$

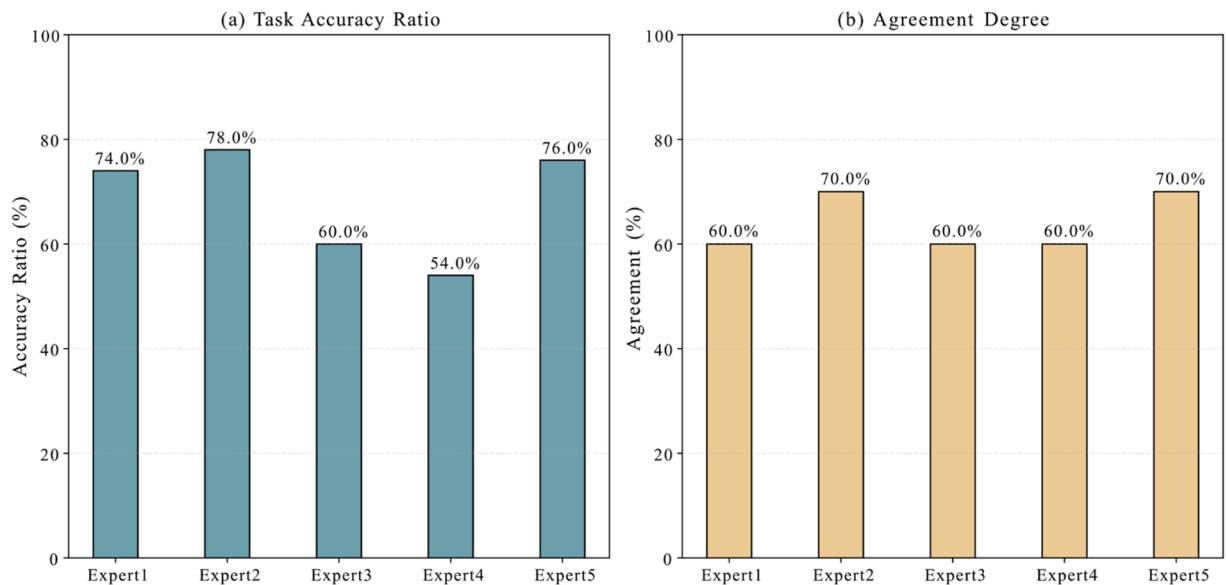


Fig. 9. The results of expert validation.

First, we will develop more sophisticated network models to capture intricate biological relationships, particularly through the use of temporal data for dynamic network analysis. It allows us to investigate the dynamic evolution of relationships between diseases and chemicals, identify time-varying characteristics of key nodes and pathways, and gain deeper insights into the dynamic features of the biomedical field. Second, we can improve RAM identification to account for a broader range of external and intrinsic factors, thereby enhancing the accuracy and timeliness of knowledge complexity assessments. Third, we may construct standardized datasets of knowledge complexity to increase the reproducibility and comparability of results. Furthermore, while our current analysis focuses primarily on author numbers and field heterogeneity, future research would benefit from examining other factors influencing knowledge dissemination, such as macro-level socioeconomic factors, technological advances, and policy changes.

6. Conclusion

This study introduces a new method for evaluating knowledge complexity by exploring the association between questions and methods within knowledge systems. Our methodology stands out for its ability to discern higher-order structural information within networks and distinguish complex knowledge from patterns that might emerge in random networks. The analysis confirms that knowledge complexity is not static; it evolves starting from simpler forms and becoming increasingly complex over time.

A key finding of our study is that complex knowledge, recognized for its intrinsic value, tends to attract a significant number of citations. Furthermore, knowledge sharing in a domain with this complex knowledge exhibits heightened sensitivity to its complexity, as evidenced by a greater volume of citations. By examining the dynamics of knowledge complexity and its dissemination characteristics, this study offers valuable insights for conducting nuanced assessments of knowledge complexity. Additionally, it contributes to the understanding of the drivers of technological innovation, enhances research performance evaluation methods, and reveals the mechanisms underlying knowledge dissemination.

CRedit authorship contribution statement

Ming Ma: Writing – review & editing, Writing – original draft, Methodology, Data curation. **Jin Mao:** Writing – review & editing, Supervision, Conceptualization. **Zhentao Liang:** Data curation, Conceptualization. **Zhejun Zheng:** Data curation. **Gang Li:** Supervision, Conceptualization.

Acknowledgment

We would like to thank the anonymous reviewers for their constructive insights which helps the improvement of our manuscript. This study is supported by the National Natural Science Foundation of China (Nos. 72174154 and 71921002). We express our profound appreciation to the five experts in biological sciences for their invaluable voluntary assistance in the study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.joi.2025.101667](https://doi.org/10.1016/j.joi.2025.101667).

References

- Balassa, B. (1965). Trade Liberalisation and “Revealed” Comparative Advantage. *The Manchester School*, 33(2), 99–123.
- Balland, P.-A., & Rigby, D. (2016). The Geography of Complex Knowledge. *Economic Geography*, 93(1), 1–23. <https://doi.org/10.1080/00130095.2016.1205947>
- Balland, P.-A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. *Regional Studies*, 53(9), 1252–1268.
- Bateman, T. S., & Hess, A. M. (2015). Different personal propensities among scientists relate to deeper vs. Broader knowledge contributions. *Proceedings of the National Academy of Sciences*, 112(12), 3653–3658.
- Bawden, D., & Robinson, L. (2015). “Waiting for Carnot”: Information and complexity. *Journal of the Association for Information Science and Technology*, 66(11), 2177–2186. <https://doi.org/10.1002/asi.23535>
- Cohen, I. R. (2006). Informational Landscapes in Art, Science, and Evolution. *Bulletin of Mathematical Biology*, 68(5), 1213–1229.
- Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., ... Mattingly, C. J. (2015). The Comparative Toxicogenomics Database's 10th year anniversary: Update 2015. *Nucleic Acids Research*, 43(D1), D914–D920. <https://doi.org/10.1093/nar/gku935>
- Davis, A. P., Wieggers, T. C., Wieggers, J., Wyatt, B., Johnson, R. J., Sciaky, D., Barkalow, F., Strong, M., Planchart, A., & Mattingly, C. J. (2023). CTD tetramers: A new online tool that computationally links curated chemicals, genes, phenotypes, and diseases to inform molecular mechanisms for environmental health. *Toxicological Sciences*, 195(2), 155–168.
- Di Guardo, G. (2015). Lipofuscin, lipofuscin-like pigments and autofluorescence. *European Journal of Histochemistry*, 59, 2485. <https://doi.org/10.4081/ejh.2015.2485>
- Dindo, D., Demartines, N., & Clavien, P. A. (2004). Classification of surgical complications—A new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Annals of Surgery*, 240(2), 205–213.
- Ericsson, K. A., & Charness, N. (1997). *Cognitive and developmental factors in expert performance* (p. 41). The MIT Press.
- Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy*, 30(7), 1019–1039. [https://doi.org/10.1016/S0048-7333\(00\)00135-9](https://doi.org/10.1016/S0048-7333(00)00135-9)
- Giorgio, A., Gatto, A. D., Pennacchio, S., Saviano, M., & Zaccaro, L. (2023). Peptoids: Smart and Emerging Candidates for the Diagnosis of Cancer, Neurological and Autoimmune Disorders. *International Journal of Molecular Sciences*, 24(22).

- Gross, G. J., & Auchampach, J. A. (2007). Reperfusion injury: Does it exist? *Journal of Molecular and Cellular Cardiology*, 42(1), 12–18. <https://doi.org/10.1016/j.jmcc.2006.09.009>
- Heffernan, K., & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116(2), 1367–1382.
- Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26), 10570–10575. <https://doi.org/10.1073/pnas.0900943106>
- Hutchins, J. (1977). On the structure of scientific texts. *UEA Papers in Linguistics*, 5(3), 18–39.
- Janavi, E., Mansourzadeh, M. J., & Samandar Ali Eshtehardi, M. (2020). A methodology for developing scientific diversification strategy of countries. *Scientometrics*, 125(3), 2229–2264. <https://doi.org/10.1007/s11192-020-03685-1>
- Ke, Q., Gates, A. J., & Barabási, A.-L. (2023). A network-based normalized impact measure reveals successful periods of scientific discovery across disciplines. *Proceedings of the National Academy of Sciences*, 120(48), Article e2309378120.
- Kennedy, P. G. (2013). Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness). *The Lancet Neurology*, 12(2), 186–194.
- Klug, M., & Bagrow, J. P. (2016). Understanding the group dynamics and success of teams. *Royal Society Open Science*, 3(4), Article 160007. <https://doi.org/10.1098/rsos.160007>
- Kuhn, T. S. (1962). The Structure of Scientific Revolutions. *Physics Today*, 16(4), 69. –69.
- Leung, N., Eirin, A., Irazabal, M. V., Maddox, D. E., Gunderson, H. D., Fervenza, F. C., & Garovic, V. D. (2009). Acute Kidney Injury in Patients with Inactive Cytochrome P450 Polymorphisms. *Renal Failure*, 31(8), 749–752. <https://doi.org/10.3109/08860220903118608>
- Li, S., Li, F., Wang, Y., Li, W., Wu, J., Hu, X., Tang, T., & Liu, X. (2024). Multiple delivery strategies of nanocarriers for myocardial ischemia-reperfusion injury: Current strategies and future perspective. *Drug Delivery*, 31(1), Article 2298514.
- Liang, Z., Liu, F., Mao, J., & Lu, K. (2021). A Knowledge Representation Model for Studying Knowledge Creation, Usage, and Evolution. In K. Toepe, H. Yan, & S. K. W. Chu (Eds.), *Diversity, divergence, dialogue* (pp. 97–111). Springer International Publishing.
- Liang, Z., Ba, Z., Mao, J., & Li, G. (2023). Research complexity increases with scientists' academic age: Evidence from library and information science. *Journal of Informetrics*, 17(1), Article 101375. <https://doi.org/10.1016/j.joi.2022.101375>
- Mattoo, S. K., Singh, G., & Vikas, A. (2003). Akathisia—Diagnostic dilemma and behavioral treatment. *Neurology India*, 51(2), 254–256.
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- McShea, D. W., Wang, S. C., & Brandon, R. N. (2019). A quantitative formulation of biology's first law. *Evolution; international journal of organic evolution*, 73(6), 1101–1115.
- Mitchell, M. (2006). Complex systems: Network thinking. *Artificial Intelligence*, 170(18), 1194–1212. <https://doi.org/10.1016/j.artint.2006.10.002>
- Nissen, M. E. (2019). Initiating a system for visualizing and measuring dynamic knowledge. *Technological Forecasting and Social Change*, 140, 169–181.
- Pintar, N., & Scherngell, T. (2022). The complex nature of regional knowledge production: Evidence on European regions. *Research Policy*, 51(8), Article 104170. <https://doi.org/10.1016/j.respol.2020.104170>
- Popper, K. R., & Weiss, G. (1959). The logic of scientific discovery. *Physics Today*, 12(11), 53–54.
- Rousseau, R., & Yang, L. (2012). Reflections on the activity index and related indicators. *Journal of Informetrics*, 6(3), 413–421.
- Rousseau, R. (2018). The F-measure for Research Priority. *Journal of Data and Information Science*, 3(1), 1–18. <https://doi.org/10.2478/jdis-2018-0001>
- Ruelle, D. (1993). *Chance and chaos*. Princeton University Press.
- iCite, Santangelo, George, & Hutchins, B. Ian (2019). *iCite database snapshots (NIH open citation collection)*.
- Solarino, A. M., Rose, E. L., & Luise, C. (2024). Going complex or going easy? The impact of research questions on citations. *Scientometrics*, 129(1), 127–146. <https://doi.org/10.1007/s11192-023-04907-y>
- Stephan, P. (2012). *How economics shapes science*. Harvard University Press.
- Strübing, J. (2007). *Research as pragmatic problem-solving: The pragmatist roots of empirically-grounded theorizing*. The sage handbook of grounded theory (pp. 580–602). SAGE Publications Ltd.. <https://doi.org/10.4135/9781848607941.n27>
- Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., & Pietronero, L. (2012). A New Metrics for Countries' Fitness and Products' Complexity. *Scientific Reports*, 2, 723.
- Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text*. University of Edinburgh Edinburgh, Scotland.
- Theurer, K. L. (2014). Complexity-based theories of emergence: criticisms and constraints. *International Studies in the Philosophy of Science*, 28(3), 277–301.
- Toner, A., & Hamilton, M. (2013). The long-term effects of postoperative complications. *Current Opinion in Critical Care*, 19(4), 364–368. <https://doi.org/10.1097/MCC.0b013e3283632f77>
- Tsai, C.-T., Kundu, G., & Roth, D. (2013). Concept-based analysis of scientific literature. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (pp. 1733–1738).
- Tuomaala, O., Järvelin, K., & Vakkari, P. (2014). Evolution of library and information science, 1965–2005: Content analysis of journal articles. *Journal of the Association for Information Science and Technology*, 65(7), 1446–1462. <https://doi.org/10.1002/asi.23034>
- Van Dijk, T. A. (1980). *Text and context explorations in the semantics and pragmatics of discourse*. Addison-Wesley Longman Ltd.
- Wang, S., Mao, J., Lu, K., Cao, Y., & Li, G. (2021). Understanding interdisciplinary knowledge integration through citance analysis: A case study on eHealth. *Journal of Informetrics*, 15(4), Article 101214.
- Weber, T., Blin, K., Duddela, S., Kim, H. U., Bruccoleri, R., ... Medema, M. H. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, 43(W1), W237–W243. <https://doi.org/10.1093/nar/gkv437>
- Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744), 378–382.
- Zhai, Y., Ding, Y., & Wang, F. (2018). Measuring the diffusion of an innovation: A citation analysis. *Journal of the Association for Information Science and Technology*, 69(3), 368–379.