

Adaptive Evolution of Data-Driven Life Science Research: A Complex Adaptive Systems Analysis of Bioeconomy Themes

Hanmei Jin^{ID}, Jianhui Li

School of Philosophy, Beijing Normal University, Beijing, China

Email: 1072590665@qq.com

How to cite this paper: Jin, H. M., & Li, J. H. (2026). Adaptive Evolution of Data-Driven Life Science Research: A Complex Adaptive Systems Analysis of Bioeconomy Themes. *Open Journal of Social Sciences*, 14, 527-551.
<https://doi.org/10.4236/jss.2026.1410025>

Received: September 28, 2026

Accepted: October 25, 2026

Published: October 28, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution-NonCommercial International License (CC BY-NC 4.0).
<http://creativecommons.org/licenses/by-nc/4.0/>



Open Access

Abstract

Large-scale data and algorithmic tools are changing how life science research problems are organized and redirected. Using bioeconomy research as an empirical window, this study examines the adaptive evolution of research attention under data-driven conditions. It analyzes 5429 archived model-input rows from the China National Knowledge Infrastructure and the Web of Science Core Collection between 2014 and October 2024, and reports revised statistics for 4937 retained text-year records after an exact-duplicate check. BERTopic is used to identify thematic structures and trace temporal changes in Chinese- and English-language corpora, and the observed patterns are interpreted through complex adaptive systems theory. The results show distinct but internally concentrated configurations. Chinese research centers on industrial innovation and application-oriented issues, while English-language research centers on the circular bioeconomy. Growth during 2020-2022 is concentrated mainly in these existing cores. Smaller topics display limited, corpus-specific variation. Three descriptive comparison windows show thematic persistence together with localized changes in research attention. The observed patterns are interpreted through operability boundaries, positive-feedback stabilization, and window-period acceleration. The study connects topic-level evidence with changes in the organization of data-driven life science research and identifies institutional priorities for standards, collaboration, and evaluation.

Keywords

Data-Driven Research, Life Sciences, Bioeconomy, Complex Adaptive Systems, BERTopic, Thematic Evolution

1. Introduction

The growing use of large-scale data and algorithmic tools is reshaping how life science research is organized. As data collection, integration, and computation become more central to scientific work, research problems are increasingly selected and advanced under conditions defined by data availability, infrastructural access, and computational feasibility rather than by experimental design alone (Hey et al., 2009; Leonelli, 2014; Leonelli, 2016). In the accumulated record of publications, this transformation becomes visible through shifts in research attention. Certain topics gain temporary centrality, stabilize around shared problem settings, and are later reorganized as new methods, materials, and external demands enter the research process. Tracing these changes in thematic prominence therefore offers an empirical way to observe how data-driven conditions alter the organization of scientific inquiry (Small, 1973).

Bioeconomy-related research provides a useful empirical window for addressing this problem. As an interdisciplinary domain built on the development and use of biological resources and biotechnological innovation, the bioeconomy spans healthcare, agriculture, forestry, energy, environmental governance, and new materials (Zheng et al., 2020). Its knowledge base remains closely connected to life science research, while its expansion across heterogeneous application contexts makes it especially suitable for observing how data-driven conditions shape research attention beyond any single discipline or technical trajectory. Because many bioeconomy-related studies rely on shared data infrastructures, computational standards, and increasingly reusable analytical tools, their thematic structure and temporal evolution make it possible to examine how research priorities are organized, stabilized, and redirected under changing conditions.

Existing studies have provided important insights into data-driven life science research, bioeconomy development, and the changing conditions of knowledge production. However, less attention has been paid to how research attention itself becomes concentrated at particular moments, how its center of gravity shifts over time, and how these changes can be identified from the evolving structure of the literature. A cross-database thematic analysis is useful in this regard because it makes it possible to observe not only what topics dominate at specific moments, but also how the structure of research attention is reorganized over time.

Drawing on bioeconomy-related publications indexed in CNKI and the Web of Science Core Collection from 2014 to 2024, this study uses BERTopic to identify the thematic structure of Chinese- and English-language literature and to trace changes in topic prominence over time (Grootendorst, 2022). On this basis, it compares the structural configurations and period-specific changes of the two corpora and further interprets the observed shifts through the lens of complex adaptive systems (Mitchell, 2009; Holland, 1995). The analysis aims to show how data-driven conditions are reflected in the organization and reorganization of research attention, and to provide a structured account of the evolving dynamics of life science research across heterogeneous application contexts.

2. Literature Review

2.1. The Evolution of Bioeconomy Research

Bioeconomy research has developed through several overlapping conceptual stages. Early discussions were closely tied to concerns about resource limits, ecological boundaries, and the biophysical conditions of economic activity. In this line of thought, the bioeconomy was not primarily framed as a technological or industrial strategy, but as a way of rethinking the relationship between economic systems and biological limits (Georgescu-Roegen, 1975). As life sciences and biotechnology advanced, however, the meaning of the bioeconomy gradually expanded. The emergence of the knowledge-based bioeconomy reframed the field around scientific knowledge, technological innovation, and the productive use of biological resources, shifting attention from ecological constraint alone to the role of research and innovation in economic transformation (Kircher et al., 2022). More recently, circular bioeconomy has become one of the dominant formulations in the field, linking biological resource use to circularity, sustainability transitions, and the reorganization of value chains across sectors (D'Amato et al., 2017).

This conceptual evolution has been accompanied by a broadening of research concerns. Recent studies have examined policy narratives, techno-political choices, sustainability promises, transition pathways, and governance arrangements in the bioeconomy (Hausknost et al., 2017; Gould et al., 2023; Eversberg et al., 2023; Khanna et al., 2024; Gottinger et al., 2020). These studies show that the bioeconomy is not a single, internally coherent research field. Instead, it is characterized by the coexistence of multiple application contexts, strategic narratives, and problem framings. In the Chinese literature, discussion has often focused on development strategies, sectoral trends, financing, regional measurement, and industrial upgrading, reflecting a stronger emphasis on development pathways and implementation conditions (Deng & Yan, 2001; Deng & Zhang, 2010; Wang et al., 2018; Wu et al., 2021; Meng & Liu, 2022; Ding et al., 2022). Related work examines policy responses to global bioeconomy development (Jiang, 2022; Pu et al., 2022) and pathways to high-quality development in China (Li, 2022). Taken together, these studies suggest that bioeconomy research has become a heterogeneous and expanding domain in which life science knowledge is continuously reorganized across diverse application settings. This makes it particularly suitable for observing how research attention is distributed and reconfigured under data-driven conditions.

At the same time, existing bioeconomy studies remain limited in one important respect. Much of the literature has focused on concepts, strategies, transition barriers, or sector-specific development patterns, while paying less attention to how the structure of research attention itself changes over time. Even when recent work uses bibliometric or review-based approaches, the main concern is often to summarize themes, trends, or policy agendas rather than to explain how certain thematic centers emerge, persist, and give way to new configurations (Gould et

al., 2023). As a result, the literature provides substantial knowledge about what the bioeconomy is argued to be, but less about how its evolving thematic structure can be used to observe broader changes in data-driven life science research.

2.2. Explaining Data-Driven Changes in Life Science Research

The transformation of life science research under data-driven conditions has been examined from several related perspectives. One line of research addresses paradigm-level change. It argues that the growing centrality of data, computation, and machine learning is altering the epistemic basis of life science research, reshaping how problems are framed, how evidence is generated, and how knowledge is organized (Leonelli, 2014; Leonelli, 2016; Li & Li, 2023). Another line of research focuses more directly on infrastructural and technical conditions, including large-scale biological data, computational modeling, intelligent algorithms, digital platforms, and the integration of wet-lab and dry-lab practices (Li & Yu, 2024; Jiang et al., 2024). A further line, especially in recent philosophical and methodological discussions, emphasizes that data-driven research does not simply add new tools to existing scientific practice, but reorganizes the pathways through which knowledge is discovered, validated, and circulated (Kitchin, 2014).

These studies have clarified many important aspects of data-driven life science research. They have shown, for example, how expanding data resources and algorithmic capabilities support new forms of prediction, screening, modeling, and coordination; how biological databases and platform infrastructures facilitate the reuse of materials and methods; and how the increasing reliance on computational procedures changes the relationship between experimentation, inference, and discovery (Callahan et al., 2020; Rigden & Fernández, 2023; Pardi et al., 2018; Wang et al., 2023). However, much of this literature remains organized around specific technologies, application domains, or infrastructures. It is therefore more effective at explaining how particular tools or systems work than at accounting for how research attention becomes concentrated, stabilized, and redirected at the level of the literature as a whole.

This limitation creates a space for a more structural approach. If data-driven change is reflected not only in new techniques and platforms but also in the organization of research attention, then thematic evolution in the literature becomes an important empirical object in its own right. Topic prominence, clustering, and redirection do not directly identify causal processes, but they do reveal how research priorities become organized across time and across application settings. In this respect, thematic analysis makes it possible to move from isolated accounts of technological change to a broader account of how problem spaces are formed, maintained, and reorganized under changing conditions. This is the point at which a complex adaptive systems perspective becomes useful.

Complex adaptive systems theory has its roots in general systems theory and later work on complexity, feedback, and adaptive change (Von Bertalanffy, 1975; Mitchell, 2009; Holland, 1992). Its relevance here lies in its capacity to relate structural

persistence and structural change to the interaction between internal accumulation and external disturbance. Rather than treating thematic evolution as a simple descriptive trend, this perspective draws attention to how research directions become viable through the formation of workable material and methodological conditions, how existing thematic centers are reinforced through repeated use and evaluative convergence, and how external pressures can accelerate reorganization when sufficient internal capacities are already in place. Recent studies have used related perspectives to discuss bioeconomy governance and transformation pathways, but they have rarely applied them to the evolution of research attention in data-driven life science research itself (Holmgren et al., 2022; Ramcilovic-Suominen, 2023).

Taken together, the existing literature provides strong foundations for understanding data-driven life science research, bioeconomy development, and adaptive system dynamics. Yet two gaps remain. First, current studies do not adequately capture how research attention evolves across heterogeneous application contexts at the level of thematic structure. Second, they offer only limited explanation of why research priorities become concentrated at specific moments, why some thematic configurations remain stable, and why key shifts occur in particular periods. To address these gaps, this study combines cross-database topic analysis with an interpretive perspective from complex adaptive systems. In doing so, it seeks to explain not only what thematic changes can be observed in bioeconomy-related literature, but also how those changes reflect the evolving organization of data-driven life science research.

3. Research Design

3.1. Data Sources

This study draws on bioeconomy-related publications indexed in the China National Knowledge Infrastructure and the Web of Science Core Collection between 2014 and October 2024. CNKI uses the Subject field with 生物经济 (bioeconomy); the journal branch selects PKU Core, CSCD, CSSCI, and AMI source categories, and dissertations form a separate branch. The English corpus uses the topic terms “bioeconomy,” “bio-based economy,” and “bio-economy,” with document types limited to articles and reviews. The archived model inputs contain 181 CNKI and 5248 Web of Science rows. Removing identical processed-text and publication-year repeats leaves 171 and 4766 text-year records for the revised summaries. The saved topic assignments are retained. Appendix A.1 gives the retrieval specification and screening counts.

As the two databases differ in coverage, disciplinary emphasis, and publication conventions, the comparison in this study is not intended to establish one-to-one equivalence between individual topics across the two corpora. Instead, the comparison is used to identify differences in thematic structure and relative changes in topic prominence over time. This makes it possible to examine how research attention is organized and adjusted across two heterogeneous but comparable literature systems. The unequal samples should be interpreted in light of different

database coverage, language scope, and source eligibility; CNKI applies narrower journal-source restrictions. Comparisons therefore use within-corpus topic shares, and the small Chinese sample is not treated as a measure of the overall scale of Chinese bioeconomy research.

3.2. Data Processing and Topic Modeling

The analysis uses the processed abstract-based texts and publication years archived with the original model inputs. Chinese and English texts were processed separately to account for linguistic differences. The revision checks identical processed text and year, retaining the first occurrence in the archived input order. It recalculates summaries from the saved assignments without refitting the topic model; Appendix A.3 specifies the calculation procedure.

For the Chinese corpus, word segmentation was conducted using jieba and a custom dictionary entry 可持续 (sustainable). A stop-word list was then applied to remove high-frequency function words and general noise terms. For the English corpus, whitespace was normalized and the vectorizer used the scikit-learn English stop-word option. This separate preprocessing strategy was necessary because the two corpora differ in tokenization logic, word form variation, and the distribution of common functional terms.

Topic identification was conducted using BERTopic, a topic modeling approach that combines semantic embeddings, dimensionality reduction, density-based clustering, and class-based TF-IDF to generate interpretable topic representations (Grootendorst, 2022). Text embeddings were generated using the multilingual SentenceTransformer model distiluse-base-multilingual-cased-v1 to support semantic representation across Chinese and English texts (Reimers & Gurevych, 2019). In the vectorization stage, CountVectorizer was used to construct the term matrix, and the supplied custom vectorizers used the effective unigram range (1, 1). Dimensionality reduction was performed with UMAP (McInnes et al., 2018), and topic clusters were identified using HDBSCAN (McInnes et al., 2017). To enhance reproducibility, the UMAP random seed was fixed at 42. The embedding, UMAP, HDBSCAN, vectorizer, and topic-reduction settings from the original scripts are listed in Appendix A.2.

Topic labels were checked against the saved top-ten c-TF-IDF terms and 47 selected non-duplicate processed texts, including both dominant topics and all retained CNKI topic 4 and CNKI topic 5 records. CNKI topic 0 and Web of Science topic 0 are broad umbrella clusters. The selection procedure and label-scope decisions are given in Appendix A.4.

3.3. Analytical Strategy

The analysis proceeds in three steps. First, BERTopic was used to identify topic clusters in the Chinese and English corpora separately and to generate representative keywords for each topic. This step provides the basis for describing the thematic composition of the two literature systems. Second, topic space distributions

were examined to identify the relative positions, clustering patterns, and degrees of separation among topics. Topic space maps are used here to reveal whether research attention is organized around dominant cores, relatively independent subclusters, or more continuous semantic groupings. Third, changes in topic prominence over time were traced through BERTopic's temporal analysis function. The archived frequency figures are complemented by calendar-year counts and normalized topic shares. Annual topic share is calculated by dividing the retained count assigned to a topic by the total retained count for that corpus and year, including outliers, and multiplying by 100. Counts describe record volume; shares describe relative attention within each corpus.

The temporal comparison uses three non-overlapping descriptive windows: 2014-2019, 2020-2022, and 2023-October 2024. The 2020 boundary follows the pandemic-era context, and 2022 is the annual record-count peak in both corpora. These windows organize the observed series rather than estimate common developmental stages or change points. The 2024 sample covers a partial year.

The purpose of combining these three steps is not only to describe what themes appear in the literature, but also to capture how thematic structures are reorganized across time. On this basis, the analysis moves from the empirical description of thematic clustering and topic evolution to an interpretation of the conditions under which particular research directions emerge, stabilize, or gain momentum. This is the point at which the complex adaptive systems perspective is introduced, allowing the observed changes in the literature to be interpreted in relation to internal accumulation, feedback processes, and external disturbance (Mitchell, 2009; Holland, 1995).

4. Thematic Structure of Bioeconomy Research

4.1. Topic Composition

The BERTopic analysis identified fifteen bioeconomy-related topics across the Chinese and English corpora. Their retained counts, proportions, and representative keywords are presented in **Table 1**. In the Chinese corpus, Topic 0, industrial innovation and bioeconomy strategy, accounts for 69.17% of the 120 assigned CNKI records and occupies a clearly dominant position. Topic 1, green biomanufacturing and forestry utilization, accounts for 13.33%; Topic 3, marine resources and fisheries management, accounts for 6.67%. Topic 2, digital integration and biological data, is a smaller cluster at 4.17%. This distribution suggests that Chinese research is strongly organized around industrial development, technological application, and related forms of sectoral expansion. The prominence of these themes is consistent with a research agenda shaped by development-oriented priorities and the extension of life science knowledge into strategic application domains.

In the English-language corpus, Topic 0, circular bioeconomy, and Topic 1, forestry bioeconomy, account for 74.70% and 16.47% of the 3115 assigned Web of Science records, respectively. Together, these two topics form the central structure of the international literature. Compared with the Chinese corpus, the English-

language literature is more strongly concentrated around circularity, biomass use, and sustainability-related concerns. At the same time, the persistence of forestry as a relatively large and distinct topic suggests the continued importance of specialized research traditions within the broader bioeconomy agenda. This indicates that the international literature is structured around a dominant sustainability-oriented core while also retaining established application domains with their own cumulative research trajectories.

Table 1. Topic composition and main topic words of retained records.

Source	Topic ID	Count	Assigned share (%)	Topic label	Main topic words
CNKI	0	83	69.17%	Industrial innovation and bioeconomy strategy	technology, innovation, industry, strategy
	1	16	13.33%	Green biomanufacturing and forestry utilization	green, manufacturing, forest, forestry
	2	5	4.17%	Digital integration and biological data	digital, diversity, integration, data
	3	8	6.67%	Marine resources and fisheries management	policy, squid, fishing, resources
	4	5	4.17%	Biopharmaceutical R&D facilities	experiment, animal, antibody, R&D
	5	3	2.50%	Bioeconomic population and harvesting models	capture, refuge, population, system
WoS	0	2327	74.70%	Circular bioeconomy	biomass, sustainable, circular, environmental
	1	513	16.47%	Forestry bioeconomy	forest, wood, policy, forestry
	2	96	3.08%	Bio-based materials and agricultural residues	olive, oil, film, palm
	3	64	2.05%	Biomass conversion and biocatalysis	lignin, fiber, chemistry, biocatalysis
	4	38	1.22%	Coffee-residue valorization	coffee, scg, spend, ground
	5	24	0.77%	Digital governance and biosecurity	digital, cyber, risk, vulnerability
	6	22	0.71%	Socio-bioeconomic imaginaries and biopolitics	argentina, imaginary, bioeconomic, reproductive
	7	17	0.55%	Grape and winery resources	grape, wine, winery, pomace
	8	14	0.45%	Reproductive labor and egg donation	egg, donation, reproductive, donor

Note. WoS denotes Web of Science. Shares use assigned-record denominators of 120 for CNKI and 3115 for Web of Science. Outliers account for 51 of 171 and 1651 of 4766 retained records, respectively. Annual shares in **Table 2** and Appendix B use all retained records, including outliers.

Taken together, the topic proportions reveal both convergence and divergence. In both corpora, research attention is concentrated in a small number of high-weight topics rather than evenly distributed across many themes. However, the Chinese corpus is more strongly centered on industrial innovation and application-oriented development, whereas the English-language corpus is more strongly

centered on circularity and resource transformation. This difference suggests that data-driven life science research is not organized around a single global thematic structure. Instead, its thematic emphasis is shaped by different problem settings and application priorities across literature systems.

4.2. Topic Space Distributions

Topic space maps provide a visual representation of the relative positions of themes in semantic space. In these maps, proximity indicates stronger semantic similarity, while more distant positions suggest clearer thematic boundaries. The Chinese topic space distribution, shown in **Figure 1**, presents a pattern characterized by a dominant central topic together with several relatively separated thematic clusters. Topic 0, industrial innovation and bioeconomy strategy, occupies the core position and covers a broad semantic area, indicating strong associations with multiple other research directions. Topic 2, digital integration and biological data, is located relatively close to this central cluster, and the partial continuity between them suggests that digital technology-related discussions are semantically linked to industrial innovation narratives. By contrast, topics such as green biomanufacturing, marine resources and fisheries, biopharmaceutical R&D facilities, and population and harvesting models form more distinct clusters located at some distance from the central core. This pattern indicates that although these themes remain connected to the broader field, they retain clearer boundaries in terms of research objects, problem framings, or methodological language.

The English-language topic space distribution, shown in **Figure 2**, is closer to a structure composed of a large main cluster and a smaller number of relatively independent subclusters. Topic 0, circular bioeconomy, forms a broad and continuous semantic region, within which several more specific themes are embedded. This suggests that circularity-related narratives possess stronger integrative capacity in the international literature, enabling a wider range of themes to be organized around a common semantic core. Topic 1, forestry bioeconomy, appears as a relatively distinct subcluster, but it is not fully detached from the main cluster. This indicates that forestry remains a specialized domain with its own research continuity while still maintaining semantic links to the dominant circular bioeconomy framework.

The comparison between **Figure 1** and **Figure 2** points to a notable structural difference. In the Chinese corpus, the thematic center is more sharply defined and surrounding themes are more clearly segmented. In the English-language corpus, the dominant thematic core is more continuous and capable of incorporating adjacent directions into a larger semantic grouping. The first pattern can be described as a more centralized and segmented organization of research attention, while the second is closer to a dominant core with semantically connected extensions. These differences matter because they show that topic prominence alone is not sufficient to describe the structure of research attention. Thematic organization also depends on how topics are clustered, how close or distant they are from one another, and whether a dominant theme functions as a narrow center or as a wider integrating framework.

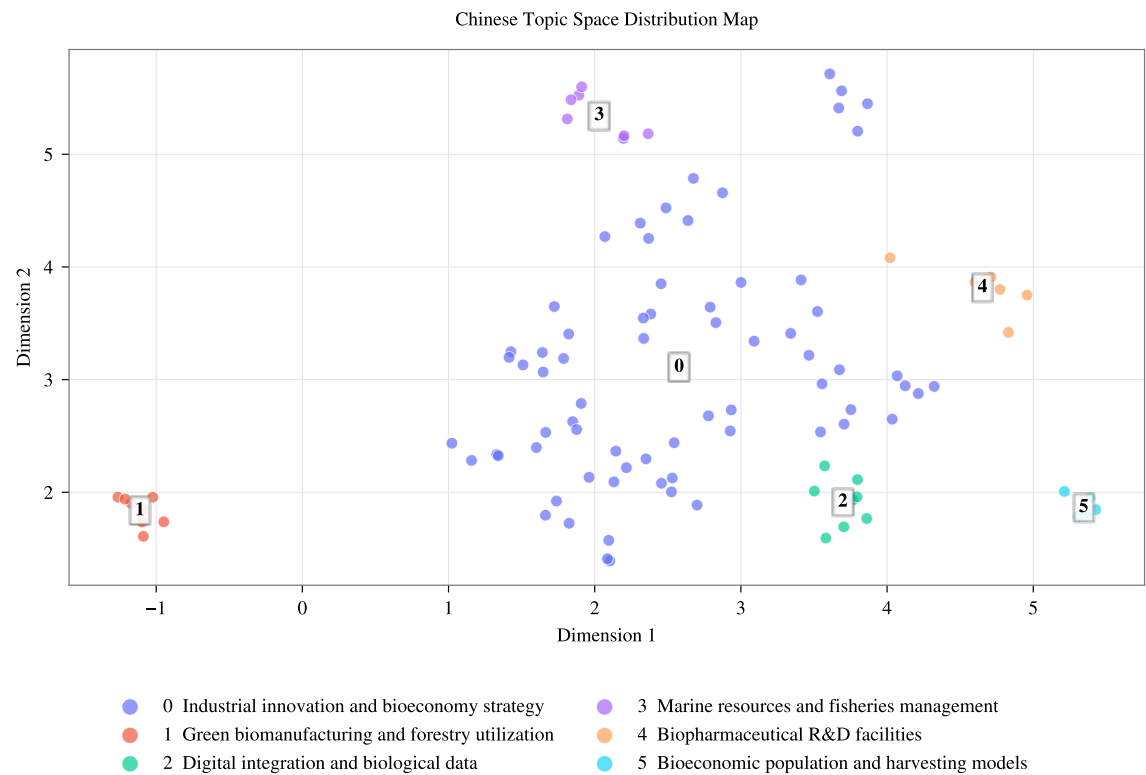


Figure 1. Topic space distribution of chinese bioeconomy research. Topic IDs follow [Table 1](#); the original point coordinates are retained and plotting labels are corrected.

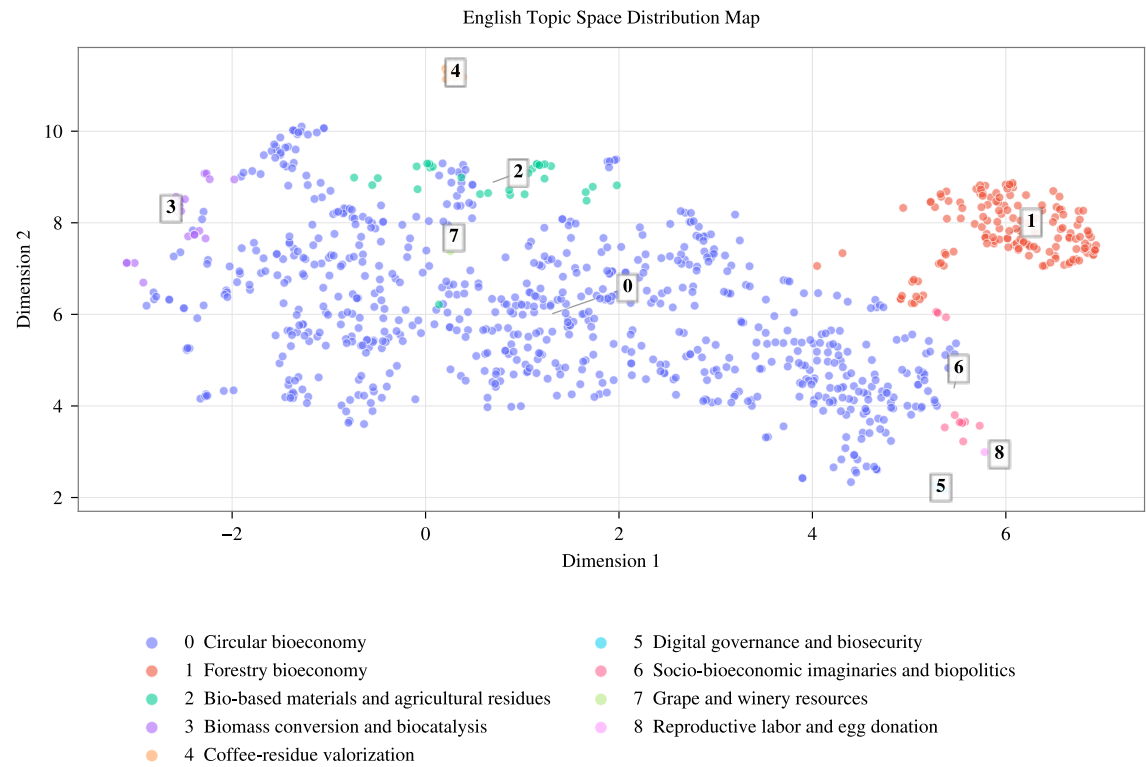


Figure 2. Topic space distribution of English-language bioeconomy research. Topic IDs follow [Table 1](#); the original point coordinates are retained and plotting labels are clarified.

4.3. Temporal Evolution of Research Attention

Figure 3, Figure 4 retain the archived model-input frequency values. Their year labels identify the publication-year groups represented by the saved time bins. **Table 2** reports calendar-year totals and selected topic counts and shares after the exact-duplicate check; Appendix B provides annual shares for all named topics. These normalized values provide the basis for interpreting relative attention.

During 2014-2019, English output grows from 36 to 321 retained records, while circular bioeconomy remains the largest named topic. Its annual share changes from 38.89% to 45.79%, with fluctuations between those years. Chinese annual totals remain small, ranging from 6 to 14 records. Research attention is organized around persistent broad cores with variable annual shares.

During 2020-2022, English output increases from 533 to 936 records. Web of Science topic 0 rises from 255 to 479 records, while its share changes from 47.84% to 51.18%. Chinese CNKI topic 0 contains 12, 7, and 24 records, with shares of 75.00%, 35.00%, and 52.17%. Growth is mainly associated with existing industrial-strategy and circular-bioeconomy themes. Web of Science topic 5, digital governance and biosecurity, contains 4, 4, and 3 records, with shares of 0.75%, 0.49%, and 0.32%. Chinese CNKI topic 4 concerns R&D facilities and contains 1, 0, and 0 records. These series do not establish a shared concentration around emergency public health or biosecurity.

Table 2. Annual totals and selected topic counts and shares.

Year	CNKI total	Industrial innovation n (%)	Digital integration n (%)	WoS total	Circular bioeconomy n (%)	Digital governance and biosecurity n (%)
2014	14	3 (21.43%)	0 (0.00%)	36	14 (38.89%)	0 (0.00%)
2015	8	3 (37.50%)	0 (0.00%)	63	40 (63.49%)	0 (0.00%)
2016	6	4 (66.67%)	0 (0.00%)	82	34 (41.46%)	0 (0.00%)
2017	6	3 (50.00%)	0 (0.00%)	161	82 (50.93%)	0 (0.00%)
2018	11	4 (36.36%)	0 (0.00%)	254	118 (46.46%)	2 (0.79%)
2019	12	4 (33.33%)	0 (0.00%)	321	147 (45.79%)	5 (1.56%)
2020	16	12 (75.00%)	0 (0.00%)	533	255 (47.84%)	4 (0.75%)
2021	20	7 (35.00%)	0 (0.00%)	809	408 (50.43%)	4 (0.49%)
2022	46	24 (52.17%)	1 (2.17%)	936	479 (51.18%)	3 (0.32%)
2023	19	15 (78.95%)	0 (0.00%)	902	427 (47.34%)	2 (0.22%)
2024*	13	4 (30.77%)	4 (30.77%)	669	323 (48.28%)	4 (0.60%)

Note. N(y) includes all retained records in the corpus and year, including outliers. Topic IDs and full names are given in **Table 1**. *2024 is incomplete; the original manuscript records coverage through October 2024. **Figure 3, Figure 4** retain archived model-input frequencies; this table reports retained text-year records after the exact-duplicate check. Appendix B gives annual shares for all named topics.

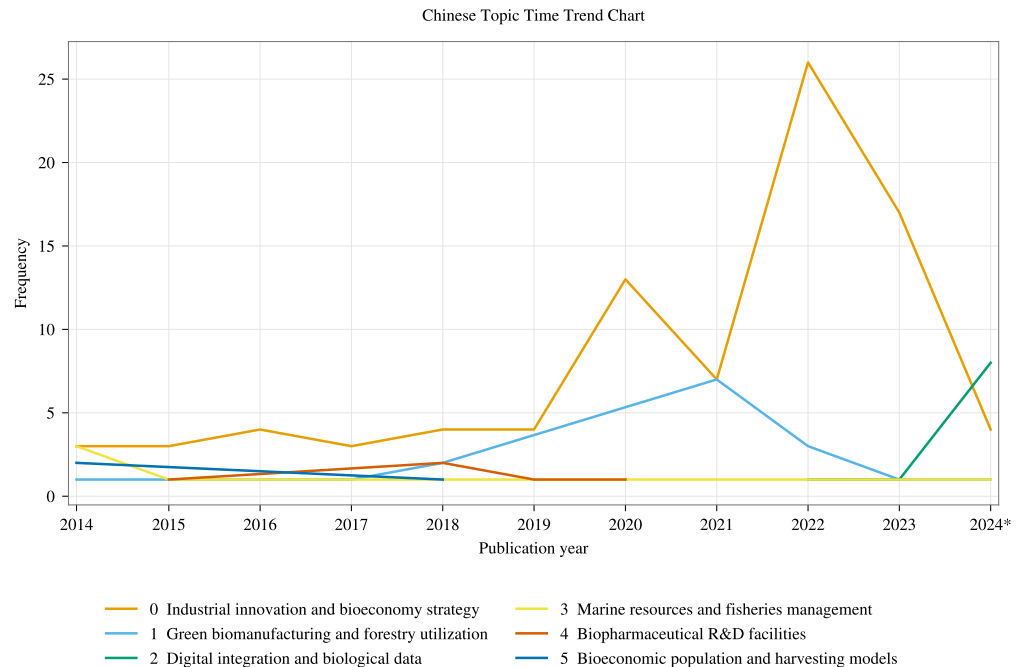


Figure 3. Temporal trends of chinese research topics. Archived frequency values and time-bin coordinates are retained; year labels identify their corresponding publication-year groups. The incomplete 2024 group is marked with an asterisk. Calendar-year counts and within-corpus shares after the exact-duplicate check are reported in **Table 2** and Appendix B; relative attention is interpreted from those normalized values.

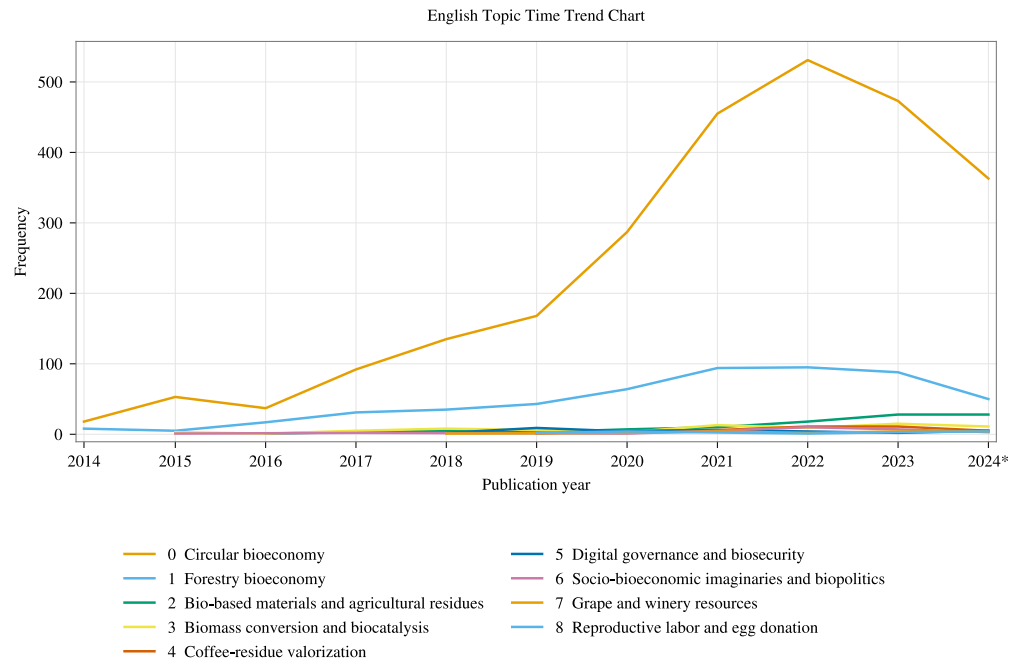


Figure 4. Temporal trends of English-language research topics. Archived frequency values and time-bin coordinates are retained; year labels identify their corresponding publication-year groups. The incomplete 2024 group is marked with an asterisk. Calendar-year counts and within-corpus shares after the exact-duplicate check are reported in **Table 2** and Appendix B; relative attention is interpreted from those normalized values.

English Web of Science topic 0 remains prominent in 2023 and the partial 2024 sample, accounting for 47.34% and 48.28%, respectively. The Chinese digital-integration topic CNKI topic 2 contains one record in 2022, none in 2023, and four of thirteen retained records in the partial 2024 sample. This is a localized digital theme whose small sample requires a limited interpretation.

The temporal analysis shows persistent dominant topics alongside selected changes in smaller topics. Annual shares distinguish growth in output from changes in relative attention. The three comparison windows organize these patterns descriptively, providing a bounded empirical basis for the complex adaptive systems interpretation.

5. Adaptive Mechanisms of Thematic Evolution

The empirical results presented above show that thematic evolution in bioeconomy-related literature does not follow a smooth and linear path. Research attention becomes concentrated around a small number of themes, some of these thematic centers remain stable for extended periods, while smaller themes show localized variation. Topic space distributions reveal the structural side of this process, while temporal trends reveal persistence and selected changes. The remaining question is how these observed patterns can be interpreted in relation to the changing organization of data-driven life science research. Drawing on a complex adaptive systems perspective, this study interprets thematic evolution through three mechanisms: operability boundaries, positive-feedback stabilization, and window-period acceleration. Together, these mechanisms provide an interpretive account of how research directions become viable, how they are reinforced over time, and how external disturbances can trigger rapid reorganization when sufficient internal capacities are already in place (Mitchell, 2009; Holland, 1995).

5.1. Operability Boundaries

The first mechanism concerns the conditions under which a research direction becomes workable enough to support sustained accumulation. Not every emerging topic can become a durable center of attention. For this to happen, a field must possess a sufficiently stable combination of data sources, analytical tools, validation procedures, and material inputs that allows researchers to organize work around a shared problem space. This is what is meant here by operability boundaries. They define the conditions under which a topic becomes not merely imaginable, but practically researchable in a repeatable and expandable way. Once such conditions are in place, dispersed problem-solving activities can begin to converge around common materials and methods, making thematic concentration more likely.

Developments in protein structure research illustrate this mechanism clearly. The increasing use of predictive tools such as AlphaFold changed the way structural information entered the research process. Once structure prediction reached a level of performance that made it usable as a routine research input, structure

no longer had to depend exclusively on the specialized experimental capabilities of a relatively small number of laboratories. The establishment and expansion of the AlphaFold Protein Structure Database, which provides open access to structure predictions for more than 214 million protein sequences, further transformed structural information into a widely accessible and reusable input for downstream research (Jumper et al., 2021; Varadi et al., 2024). AlphaFold2-based methods have also been applied to predicting protein-protein interactions, extending the use of a common predictive framework to protein complexes (Bryant et al., 2022). This did not simply accelerate existing work. It altered the threshold of entry into the research process and enabled more researchers to work on related problems through a shared computational interface. As predicted structures became more searchable, reusable, and easier to integrate into larger inference sequences, related lines of inquiry were more likely to converge around them. In thematic terms, this is the kind of condition under which a high-weight topic can form and begin to occupy a central position.

From this perspective, topic concentration is not adequately explained by scientific or technological importance alone. Themes rise when the conditions for organizing continuous work around them become available. In data-driven life science research, those conditions increasingly depend on whether data can be standardized and reused, whether model outputs can be integrated into broader research workflows, and whether validation procedures are stable enough to sustain further accumulation. Operability boundaries therefore explain why certain research directions become thematic centers while others remain fragmented or peripheral.

5.2. Positive-Feedback Stabilization

The second mechanism concerns the persistence of thematic centers once they have formed. The results in the previous section show that some topics maintain high prominence over extended periods rather than disappearing after short bursts of attention. This suggests that once a workable thematic center emerges, it can be reinforced through repeated use, cumulative referencing, coordinated division of labor, and converging standards of evaluation. This reinforcing process is described here as positive-feedback stabilization. Under such conditions, past accumulation lowers the cost of subsequent work, making it easier for researchers to continue moving along existing paths than to construct entirely new ones.

The case of forestry bioeconomy and related governance arrangements offers a useful example. Research on the Swedish forest bioeconomy has shown that industrial organizations, forest-owner associations, regional institutions, and research actors can become linked through relatively stable governance and resource channels, with policy interfaces and shared narratives reinforcing the field's continuity (Holmgren et al., 2022). Stability in such a system does not simply mean that actors remain connected. It also means that the categories through which work is evaluated, coordinated, and legitimated become more aligned over time. In a

related way, biofoundry research demonstrates how repeated standardization of workflows, data recording, and iterative model use can strengthen reproducibility and cross-team transferability, thereby increasing the durability of certain research trajectories (Kim et al., 2025). Once data formats, procedures, and evaluative expectations become sufficiently consistent, thematic centers are not reproduced by attention alone. They are reproduced through the infrastructural and organizational conditions that make continued work within them comparatively efficient and credible.

Positive-feedback stabilization helps explain why certain high-weight topics remain central for relatively long periods. It is not only that they are important. It is also that they become easier to continue. Reusable data, standardized workflows, familiar problem framings, and relatively settled evaluative criteria all reinforce the existing structure of research attention. In this sense, thematic stability reflects not mere inertia, but an actively reproduced alignment between methods, infrastructures, and organizational expectations.

5.3. Window-Period Acceleration

The third mechanism explains why major shifts in research attention occur abruptly in particular periods rather than through gradual redistribution alone. External pressures do not create new thematic centers from nothing. However, they can sharply accelerate reorganization when research systems already possess latent capacities that can be activated, scaled, and redirected under time pressure. This process is described here as window-period acceleration. It refers to the concentration of resources, urgency, and expected returns within a limited period, such that a topic can rise rapidly and restructure the thematic field around it.

The rapid development of mRNA vaccines during the COVID-19 period illustrates this mechanism. Their acceleration did not result from emergency conditions alone. It depended on prior accumulations in nucleotide modification, lipid nanoparticle delivery, prototype pathogen preparedness, and related sequence-based design strategies that had already been developed over many years (Pardi et al., 2018; Corbett et al., 2020). When SARS-CoV-2 sequence information became available, Moderna was able to move from sequence release to phase I clinical testing of mRNA-1273 within 66 days (Corbett et al., 2020; National Institutes of Health, 2020), while large-scale organizational initiatives such as Pfizer's "Project Lightspeed" compressed production timelines even further. The subsequent publication of strong clinical evidence for the safety and efficacy of mRNA vaccines helped transform short-term emergency mobilization into a more stable expectation of continued investment and research continuity (Baden et al., 2021; Polack et al., 2020). In thematic terms, this is the kind of process through which an external disturbance can rapidly reorder research attention, not because it generates a wholly new direction, but because it activates an already prepared combination of data, methods, infrastructures, and organizational capabilities.

The mRNA vaccine case illustrates how external pressure can accelerate work

by activating capacities accumulated in advance. Its role here is to clarify the proposed mechanism of window-period acceleration. The topic series show growth around existing broad cores during 2020-2022, with limited variation in smaller topics. The illustrative case supports the mechanism discussion; the corpus analysis establishes thematic persistence and local variation rather than a common crisis-driven sequence.

Taken together, these three mechanisms provide an interpretation of thematic evolution that is more specific than a generic claim about complexity. Operability boundaries explain how research directions become viable centers of work. Positive-feedback stabilization explains how such centers persist through cumulative reinforcement. Window-period acceleration explains how external pressure can trigger rapid reorganization when internal capacities are already in place. Read together, they show that the thematic evolution of bioeconomy literature reflects broader changes in the organization of data-driven life science research. The literature does not merely record shifting interests. It records the changing conditions under which certain forms of inquiry become possible, sustainable, and expandable.

6. Discussion and Conclusion

The analysis shows that bioeconomy-related literature provides a useful window onto the evolving organization of data-driven life science research. Across the Chinese and English corpora, research attention is concentrated in a limited number of high-weight topics rather than evenly dispersed across many directions. At the same time, the two corpora display different thematic centers. Chinese-language research is more strongly organized around industrial innovation and application-oriented agendas, whereas English-language research is more strongly centered on circular bioeconomy themes and related sustainability concerns. These differences indicate that data-driven life science research does not develop through a single globally uniform thematic structure. Instead, it is organized through distinct but internally clustered configurations shaped by different application contexts and problem settings.

The temporal analysis identifies persistent broad cores together with localized changes in smaller topics. Publication-record growth during 2020-2022 is mainly associated with industrial-strategy and circular-bioeconomy themes. Annual shares distinguish this output growth from changes in relative attention. The incomplete 2024 observations are reported descriptively and do not establish a new developmental stage. The three descriptive comparison windows organize these observations, while complex adaptive systems theory offers an interpretation of thematic persistence and possible routes of adaptation.

The complex adaptive systems perspective helps interpret these changes in a more structured way. The three proposed mechanisms specify conditions under which thematic centers may emerge and change; they are interpretive propositions rather than causal effects estimated from the topic series. Operability bound-

aries explain how certain research directions become viable through the availability of reusable data, analytical tools, and workable validation paths. Positive-feedback stabilization explains how thematic centers persist through repeated use, accumulated coordination, and increasingly aligned evaluative expectations. Window-period acceleration explains how external disturbances can rapidly amplify themes that already rest on sufficiently mature internal capacities. Taken together, these mechanisms suggest that thematic evolution in the literature reflects a dynamic relationship between internal accumulation and external pressure rather than a simple sequence of topic replacement.

The study also has implications for how data-driven life science research may be understood more broadly. Existing discussions often focus either on specific technological systems or on general claims about paradigm change. The present analysis suggests that an intermediate level of observation is equally important: the organization of research attention in the literature. By tracing thematic concentration, stability, and redirection across two databases, this study shows how literature-based evidence can help connect paradigm-level claims to observable changes in research organization. In this respect, the contribution of the study lies not only in mapping bioeconomy research, but also in showing how cross-database thematic analysis can be used to examine the evolving structure of data-driven scientific inquiry.

These findings also suggest three practical priorities for research policy. First, data definitions and processing standards should be institutionally coordinated so that distributed resources can produce comparable and cumulative outputs across teams. Second, executable rules for benefit sharing and the allocation of responsibility should convert cross-organizational connections into durable divisions of labor. Third, data contributions, method development, and critical intermediate outputs should be incorporated into formal evaluation systems. These measures would improve the research community's capacity to mobilize when external opportunities arise and help translate temporary concentrations of resources into longer-term research capabilities.

Several limitations should be noted. First, the analysis is based on titles and abstracts rather than full texts, which means that some thematic distinctions are necessarily coarser than those that might emerge from full-document analysis. Second, although CNKI and the Web of Science Core Collection provide broad coverage, differences in indexing practices, disciplinary scope, and publication conventions shape the resulting corpora and constrain direct equivalence between them. The Chinese corpus is small, especially at the annual and minor-topic levels, and 2024 is only partially observed. The retained text-year records are not a census of all publications, so absolute corpus sizes are not directly compared. The original partitions remain fixed after the exact-duplicate check and may retain effects of repeated inputs on clustering. Third, thematic evolution does not by itself establish causal explanation. The interpretive framework proposed here provides plausible mechanism-based readings of observed patterns, but further work is needed

to connect topic shifts more systematically to specific events, infrastructures, organizational networks, and methodological changes. Future research could therefore combine topic modeling with citation networks, collaboration structures, funding sources, or event-sequence analysis to strengthen the explanatory linkage between thematic change and the reorganization of data-driven research.

Acknowledgements

This research was supported by the National Social Science Foundation of China through the project “Research on the Paradigm Transformation of Life Sciences Driven by Big Data” (Grant No. 22&ZD045).

Author Contributions

Hanmei Jin: Writing - review and editing, Writing - original draft, Software, Methodology, Funding acquisition, Data curation, Conceptualization. Jianhui Li: Supervision, Funding acquisition.

Conflicts of Interest

The authors report there are no competing interests to declare.

References

- Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R. et al. (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*, 384, 403-416. <https://doi.org/10.1056/nejmoa2035389>
- Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved Prediction of Protein-Protein Interactions Using AlphaFold2. *Nature Communications*, 13, Article No. 1265. <https://doi.org/10.1038/s41467-022-28865-w>
- Callahan, T. J., Tripodi, I. J., Pielke-Lombardo, H., & Hunter, L. E. (2020). Knowledge-based Biomedical Data Science. *Annual Review of Biomedical Data Science*, 3, 23-41. <https://doi.org/10.1146/annurev-biodatasci-010820-091627>
- Corbett, K. S., Edwards, D. K., Leist, S. R., Abiona, O. M., Boyoglu-Barnum, S., Gillespie, R. A. et al. (2020). SARS-CoV-2 mRNA Vaccine Design Enabled by Prototype Pathogen Preparedness. *Nature*, 586, 567-571. <https://doi.org/10.1038/s41586-020-2622-0>
- D’Amato, D., Droste, N., Allen, B., Kettunen, M., Lahinen, K., Korhonen, J. et al. (2017). Green, Circular, Bio Economy: A Comparative Analysis of Sustainability Avenues. *Journal of Cleaner Production*, 168, 716-734. <https://doi.org/10.1016/j.jclepro.2017.09.053>
- Deng, X. A., & Yan, J. S. (2001). Who Will Replace the Information Economy: Knowledge Economy or Bioeconomy? *Global Science, Technology and Economy Outlook*, No. 1, 7-8. (In Chinese).
- Deng, X. A., & Zhang, Y. L. (2010). On the Evolution of Economy Eras and the Laws of Bioeconomy. *Journal of Zhejiang University (Humanities and Social Sciences)*, 40, 144-151. (In Chinese).
- Ding, C. J., Chen, F., Han, Q. et al. (2022). Methodology for Measuring Bioeconomy in Some Countries/Regions and Its Enlightenment to China. *China Biotechnology*, 42, 154-162. (In Chinese). <https://doi.org/10.13523/j.cb.2205110>

- Eversberg, D., Koch, P., Lehmann, R., Saltelli, A., Ramcilovic-Suominen, S., & Kovacic, Z. (2023). The More Things Change, the More They Stay the Same: Promises of Bioeconomy and the Economy of Promises. *Sustainability Science*, 18, 557-568. <https://doi.org/10.1007/s11625-023-01321-4>
- Georgescu-Roegen, N. (1975). Energy and Economic Myths. *Southern Economic Journal*, 41, 347-381. <https://doi.org/10.2307/1056148>
- Gottinger, A., Ladu, L., & Quitzow, R. (2020). Studying the Transition towards a Circular Bioeconomy—A Systematic Literature Review on Transition Studies and Existing Barriers. *Sustainability*, 12, Article 8990. <https://doi.org/10.3390/su12218990>
- Gould, H., Kelleher, L., & O'Neill, E. (2023). Trends and Policy in Bioeconomy Literature: A Bibliometric Review. *EFB Bioeconomy Journal*, 3, Article ID: 100047. <https://doi.org/10.1016/j.bioeco.2023.100047>
- Grootendorst, M. (2022). *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*. arXiv: 2203.05794. <https://doi.org/10.48550/arXiv.2203.05794>
- Hausknost, D., Schriebl, E., Lauk, C., & Kalt, G. (2017). A Transition to Which Bioeconomy? an Exploration of Diverging Techno-Political Choices. *Sustainability*, 9, Article 669. <https://doi.org/10.3390/su9040669>
- Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.
- Holland, J. H. (1992). Complex Adaptive Systems. *Daedalus*, 121, 17-30. <https://www.jstor.org/stable/20025416>
- Holland, J. H. (1995). *Hidden Order: How Adaptation Builds Complexity*. Addison-Wesley.
- Holmgren, S., Giurca, A., Johansson, J., SöderlundKanarp, C., Stenius, T., & Fischer, K. (2022). Whose Transformation Is This? Unpacking the ‘Apparatus of Capture’ in Sweden’s Bioeconomy. *Environmental Innovation and Societal Transitions*, 42, 44-57. <https://doi.org/10.1016/j.eist.2021.11.005>
- Jiang, H. P., Gao, C. C., Liu, W. H. et al. (2024). Advances in Data-Driven Life Sciences Research. *Bulletin of Chinese Academy of Sciences*, 39, 862-871. (In Chinese). <https://doi.org/10.16418/j.issn.1000-3045.20240225003>
- Jiang, J. (2022). Evolution Patterns and Development Layout of the Global Bioeconomy. *Renmin Luntan (People's Tribune)*, No. 17, 12-16. (In Chinese).
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596, 583-589. <https://doi.org/10.1038/s41586-021-03819-2>
- Khanna, M., Zilberman, D., Hochman, G., & Basso, B. (2024). An Economic Perspective of the Circular Bioeconomy in the Food and Agricultural Sector. *Communications Earth & Environment*, 5, Article No. 507. <https://doi.org/10.1038/s43247-024-01663-6>
- Kim, H., Hillson, N. J., Cho, B.-H., Sung, B. H., Lee, D.-H., Kim, D.-M. et al. (2025). Abstraction Hierarchy to Define Biofoundry Workflows and Operations for Interoperable Synthetic Biology Research and Applications. *Nature Communications*, 16, Article No. 6056. <https://doi.org/10.1038/s41467-025-61263-6>
- Kircher, M., Maurer, K.-H., & Herzberg, D. (2022). KBBE: The Knowledge-Based Bioeconomy: Concept, Status and Future Prospects. *EFB Bioeconomy Journal*, 2, Article ID: 100034. <https://doi.org/10.1016/j.bioeco.2022.100034>
- Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, 1. <https://doi.org/10.1177/2053951714528481>
- Leonelli, S. (2014). What Difference Does Quantity Make? On the Epistemology of Big Data

- in Biology. *Big Data & Society*, 1. <https://doi.org/10.1177/2053951714534395>
- Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. University of Chicago Press. <https://doi.org/10.7208/chicago/9780226416502.001.0001>
- Li, J. H., & Li, C. H. (2023). Paradigm Transformation of Life Science Research Driven by Big Data. *Philosophical Trends*, No. 10, 97-105. (In Chinese)
- Li, X. J. (2022). High Quality Development Approach of Bioeconomy in China. *Renmin Luntan-Xueshu Qianyan (Academic Frontiers)*, No. 14, 61-67. (In Chinese) <https://doi.org/10.16619/j.cnki.rmltxsqy.2022.14.006>
- Li, X., & Yu, H. C. (2024). A New Paradigm of Life Science Research Driven by Artificial Intelligence. *Bulletin of Chinese Academy of Sciences*, 39, 50-58. (In Chinese) <https://doi.org/10.16418/j.issn.1000-3045.20231211001>
- McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical Density Based Clustering. *The Journal of Open Source Software*, 2, Article 205. <https://doi.org/10.21105/joss.00205>
- McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3, Article 861. <https://doi.org/10.21105/joss.00861>
- Meng, Q. Z., & Liu, J. (2022). Construction of Evaluation Index System of Bioeconomic Development Level in China: An Empirical Study Based on Provincial Data. *Science and Technology Management Research*, 42, 39-48. (In Chinese)
- Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.
- National Institutes of Health (2020). *NIH Clinical Trial of Investigational Vaccine for COVID-19 Begins*. <https://www.nih.gov/news-events/news-releases/nih-clinical-trial-investigational-vaccine-covid-19-begins>
- Pardi, N., Hogan, M. J., Porter, F. W., & Weissman, D. (2018). mRNA Vaccines—A New Era in Vaccinology. *Nature Reviews Drug Discovery*, 17, 261-279. <https://doi.org/10.1038/nrd.2017.243>
- Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S. et al. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, 383, 2603-2615. <https://doi.org/10.1056/nejmoa2034577>
- Pu, A. Q., Liu, Q. Q., Guo, S. H. et al. (2022). The Evolution Law of the Global Bioeconomy and the Suggestions for the Development of Chinese Bioeconomy. *China Biotechnology*, 42, 146-153. (In Chinese) <https://doi.org/10.13523/j.cb.2205109>
- Ramcilovic-Suominen, S. (2023). Envisioning Just Transformations in and beyond the EU Bioeconomy: Inspirations from Decolonial Environmental Justice and Degrowth. *Sustainability Science*, 18, 707-722. <https://doi.org/10.1007/s11625-022-01091-5>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/d19-1410>
- Rigden, D. J., & Fernández, X. M. (2023). The 2023 *Nucleic Acids Research* Database Issue and the Online Molecular Biology Database Collection. *Nucleic Acids Research*, 51, D1-D8. <https://doi.org/10.1093/nar/gkac1186>
- Small, H. (1973). Co-Citation in the Scientific Literature: A New Measure of the Relationship between Two Documents. *Journal of the American Society for Information Science*, 24, 265-269. <https://doi.org/10.1002/asi.4630240406>

- Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M. et al. (2024). AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences. *Nucleic Acids Research*, 52, D368-D375. <https://doi.org/10.1093/nar/gkad1011>
- Von Bertalanffy, L. (1975). *Perspectives on General System Theory: Scientific-Philosophical Studies*. George Braziller.
- Wang, H. G., Zhu, S., Yin, Z. X. et al. (2018). Trends and Features in the Development of International Bio-Economy. *Forum on Science and Technology in China*, No. 5, 158-164. (In Chinese) <https://doi.org/10.13580/j.cnki.fstc.2018.05.019>
- Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z. et al. (2023). Scientific Discovery in the Age of Artificial Intelligence. *Nature*, 620, 47-60. <https://doi.org/10.1038/s41586-023-06221-2>
- Wu, X. Y., Chen, F., Ding, C. J. et al. (2021). Analysis of Global Bioeconomy Status, Trends and Financing Prospects. *China Biotechnology*, 41, 116-126. (In Chinese) <https://doi.org/10.13523/j.cb.2107015>
- Zheng, S. Q., Han, Q., Chen, Y. P. et al. (2020). A Review of Recent Foreign Bioeconomy Strategies and Their Implications for China. *China Biotechnology*, 40, 108-113. (In Chinese) <https://doi.org/10.13523/j.cb.1909069>

Appendix A. Retrieval and Reproducibility Details

Retrieval Specification and Record Screening

The retrieval specification below reconstructs the scope documented for the original collection. It defines the search fields, concepts, eligibility filters, and observation period. The original manuscript records a cutoff of October 2024. Screening counts start from the archived processed inputs used by the saved model (**Table A1**).

CNKI: Subject = 生物经济; publication years 2014-2024 through the cutoff. The journal branch selects PKU Core, CSCD, CSSCI, and AMI source categories; dissertations are a separate branch. Web of Science Core Collection: TS = (bioeconomy OR “bio-based economy” OR “bio-economy”) AND PY = (2014-2024), with Article and Review selected using the document-type filter. The archived English inputs comprise a base bioeconomy workbook and four technology-tagged workbooks. The core expression specifies the documented scope rather than a preserved export of every executed query.

Table A1. Archived input screening.

Screening item	CNKI	Web of science
Archived model-input rows	181	5248
Identical processed-text and year repeats removed	10	482
Retained text-year records	171	4766
Assigned to named topics	120	3115
Outliers (topic –1)	51	1651

An exact duplicate is a row with the same processed text and publication year as an earlier archived row. The first occurrence is retained together with its saved topic assignment. This rule identifies repeated inputs at the text-year level; the resulting counts are not identifier-verified unique-publication totals. One CNKI and 38 Web of Science duplicate groups contain different saved assignments. The first-occurrence rule makes their treatment deterministic, while retaining the original partition means that duplication may have influenced the fitted clustering.

Original Modeling Settings

Both scripts use SentenceTransformer distiluse-base-multilingual-cased-v1. The Chinese tokenizer uses jieba, removes the supplied stop words, retains trimmed tokens longer than one character, and adds the custom dictionary entry 可持续 (sustainable). The English vectorizer uses the scikit-learn English stop-word option. The supplied custom vectorizers have the effective unigram range (1, 1); the BERTopic configuration value (1, 2) does not override that vectorizer. The settings are summarized in **Table A2**.

Table A2. Settings recorded in the original modeling scripts.

Component	Chinese corpus	English corpus
UMAP	n_neighbors=15; n_components=2; min_dist=0.1; metric=cosine; random_state=42	Same settings
HDBSCAN	min_cluster_size=3; min_samples=2; metric=euclidean; cluster_selection_method=leaf; prediction_data=True; gen_min_span_tree=True	min_cluster_size=10; min_samples=5; metric=euclidean; cluster_selection_method=eom; prediction_data=True
CountVectorizer	max_features=10000; min_df=1; max_df=0.95; jieba tokenizer; supplied stop words; ngram_range=(1,1)	max_features=10000; min_df=2; max_df=0.95; stop_words=english; ngram_range=(1,1)
BERTopic	nr_topics=auto; min_topic_size=3; top_n_words=10; calculate_probabilities=True	nr_topics=auto; min_topic_size=10; top_n_words=10; calculate_probabilities=True
Topic reduction	If topic-info rows exceed 11, reduce to nr_topics=7	If topic-info rows exceed 15, reduce to nr_topics=10

Reproduction of the Revised Statistics

The revised summaries use the archived aligned records, publication years, saved topic assignments, and their original row order. Apply the exact-text/year rule in A.1, group retained assignments by corpus, topic and calendar year, and divide each topic count by the total retained corpus count for that year. Annual denominators include topic – 1. For **Table 1**, divide topic counts by the number assigned to named topics. These operations reproduce the revised summaries without embedding, dimensionality reduction, clustering, or topic reduction being fitted again.

The calculation package contains data/cnki_aligned_records.json, data/wos_aligned_records.json, saved topic assignments, the matched original scripts, chinese_custom_dictionary.txt, chinese_stopwords.json, recalculate_results.py, recalculated/results.json, and recalculated/label_audit.json. Its recorded calculation environment is Python 3.12.14, numpy 2.3.5, pandas 2.2.3, matplotlib 3.10.8, and scikit-learn 1.8.0. These versions identify the revision calculations; the historical model-training package versions are not recorded. The archived assignments provide the fixed reference for reproducing the reported tables.

Figure 1, **Figure 2** retain their original document-point coordinates. **Figure 3**, **Figure 4** retain the saved frequency values and time-bin coordinates, with tick labels identifying the corresponding publication years. Topic names are corrected

by saved topic ID. The figures are re-exported from archived plotting data with readable labels and vector artwork. Calendar-year statistics for retained records are reported in **Table 2** and Appendix B.

Topic Label Check

For each topic, the saved top-ten c-TF-IDF terms were compared with the processed text having the strongest top-word-weight match and distinct texts from the earliest and latest available years. All five retained CNKI topic 4 texts and all three CNKI topic 5 texts were checked. The resulting 47 selected texts are recorded with source file, source row, year, and saved topic ID in label_audit.json. This is a targeted semantic check of label coverage rather than an independent estimate of cluster purity or inter-coder agreement.

CNKI topic 0 covers industrial innovation and bioeconomy strategy, while Web of Science topic 0 covers a broad circular-bioeconomy agenda. The smaller-topic labels are constrained to their textual coverage: CNKI topic 4 describes biopharmaceutical R&D facilities; CNKI topic 5 describes bioeconomic population and harvesting models; Web of Science topic 5 combines digital governance and biosecurity; Web of Science topic 7 concerns grape and winery resources; and Web of Science topic 8 concerns reproductive labor and egg donation. **Table 1** and the figure labels use these interpretations consistently.

Appendix B. Annual Topic Shares

The following tables report percentages of all retained records in each corpus and year, including outliers in the denominator. Annual totals $N(y)$ appear in **Table 2**. A count can be recovered by multiplying the annual total by its reported percentage, dividing by 100, and rounding to the nearest record. Topic IDs refer to **Table 1**, and –1 denotes outliers. *2024 is incomplete; the cutoff recorded in the original manuscript is October 2024. Annual shares appear in **Table B1**, **Table B2**.

Table B1. CNKI annual shares (%).

Year	Outliers	0	1	2	3	4	5
2014	35.71	21.43	7.14	0.00	21.43	0.00	14.29
2015	37.50	37.50	0.00	0.00	12.50	12.50	0.00
2016	16.67	66.67	0.00	0.00	16.67	0.00	0.00
2017	33.33	50.00	16.67	0.00	0.00	0.00	0.00
2018	18.18	36.36	18.18	0.00	0.00	18.18	9.09
2019	50.00	33.33	0.00	0.00	8.33	8.33	0.00
2020	12.50	75.00	0.00	0.00	6.25	6.25	0.00
2021	30.00	35.00	35.00	0.00	0.00	0.00	0.00
2022	39.13	52.17	6.52	2.17	0.00	0.00	0.00
2023	15.79	78.95	5.26	0.00	0.00	0.00	0.00
2024*	23.08	30.77	7.69	30.77	7.69	0.00	0.00

Table B2. Web of science annual shares (%).

Year	Outliers	0	1	2	3	4	5	6	7	8
2014	38.89	38.89	22.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2015	25.40	63.49	6.35	0.00	3.17	0.00	0.00	1.59	0.00	0.00
2016	35.37	41.46	20.73	1.22	1.22	0.00	0.00	0.00	0.00	0.00
2017	24.84	50.93	18.63	1.86	2.48	0.00	0.00	1.24	0.00	0.00
2018	34.65	46.46	13.78	1.97	1.97	0.00	0.79	0.00	0.39	0.00
2019	37.07	45.79	12.15	0.93	1.56	0.31	1.56	0.00	0.00	0.62
2020	35.46	47.84	11.82	1.31	0.56	0.94	0.75	0.19	0.38	0.75
2021	33.99	50.43	10.88	1.24	1.36	0.62	0.49	0.49	0.49	0.00
2022	33.65	51.18	10.04	1.60	0.96	1.18	0.32	0.75	0.21	0.11
2023	36.03	47.34	9.42	2.88	1.55	1.22	0.22	0.55	0.44	0.33
2024*	36.02	48.28	7.47	3.89	1.49	0.75	0.60	0.30	0.60	0.60