Gondwana Research 145 (2025) 49–56

Contents lists available at ScienceDirect

Gondwana Research

journal homepage: www.elsevier.com/locate/gr

A century of knowledge growth in sedimentology

Yiwei Xu [a,b], Xiumian Hu [b,*], Eduardo Garzanti [c], Jiexing Qi [d], Huquan Kang [d], Hanqi Nong [d], Luwen Wu [d], Jiaxin Ding [d], Xinbing Wang [d], Chengshan Wang [e]

[a] State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, Chinese Academy of Sciences, Nanjing 210008 China
[b] State Key Laboratory of Critical Earth Material and Mineral Deposits, School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, China
[c] Department of Earth and Environmental Sciences, Università di Milano-Bicocca, Milano 20126, Italy
[d] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[e] State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Beijing 100083, China

ARTICLE INFO

Handling Editor: Damian Nance

Keywords:
Sedimentology
Knowledge growth
Natural language process
Large language model
Bibliometrics

ABSTRACT

Although the deceleration in scientific knowledge production has been well documented at the level of fields, the patterns of knowledge growth within subdisciplines remain poorly explored, primarily due to challenges in classifying papers at lower level. Our study addresses this gap by employing natural language process (NLP) tools to explore the growth of sedimentological knowledge which is a subdiscipline within geology. Utilizing SedBERT, a specialized Bidirectional Encoder Representations from Transformers (BERT) model, we accurately classify sedimentological papers, revealing an exponential growth in sedimentological publications over the past 120 years. Publications have doubled every 10.3 years between 1945 and 1980, and every 14 years from 1980 to 2021.We identify a significant paradigm shift during the 1950 s-1970 s, a period known as the 'Golden Age' of sedimentology, characterized by increased lexical diversity and myopic-referencing citation pattern. A subsequent decline in research diversity, driven by a 'follow-the-crowd' strategy, has led to a stagnation in knowledge expansion in the post-Golden Age. Our study illuminates the dynamic research landscape of sedimentology and offers a framework for analyzing the evolution of sub-disciplinary knowledge.

1. Introduction

The number of scientific publications has expanded unprecedentedly over recent decades (Milojević, 2015; Dong et al., 2017; Bornmann et al., 2021). Such information overload has made it overwhelming for researchers to identify relevant studies and stay updated with the latest developments in the field, and moreover, important findings may be scattered across multiple publications, making it harder to synthesize comprehensive insights (Cyranoski et al., 2011; Bloom et al., 2020; Chu and Evans, 2021). Consequently, the growth rate of scientific knowledge progresses linearly (Milojević, 2015), suggesting a slowdown or even stagnation in the rate of innovative idea generation (Chu and Evans, 2021; Park et al., 2023).

Despite extensive documentation of the slowdown at the level of scientific fields and large subfields, the understanding of the scientific landscape at the sub-disciplinary level remains limited. In the case of sedimentology, sedimentologists have traditionally addressed this challenge through systematic literature reviews (Miall, 1995; Friedman, 1998; Middleton, 2003). However, the exponential growth of scientific literature has made this task increasingly difficult over time.

This challenge can be addressed by applying bibliometric methods to the published literature (Donthu et al., 2021; Park et al., 2022), enabling an analysis of the temporal evolution within the sedimentology subfield. Unlike large fields such as physics, where the literature can be systematically catergorized by sets of journals (Sinatra et al., 2015), keyword-based searches are typically employed to identify relevant literature in more specialized research domain due to their narrower topic coverage (Donthu et al., 2021). There remains a longstanding challenge in precisely classifying papers at the sub-disciplinary level. Recent developments of pre-trained large language models in Natural Language Processing (NLP) provides a solution. These models can be fine-tuned for domain-specific tasks like text mining and data extraction (Tshitoyan et al., 2019; Olivetti et al., 2020). Previous studies have demonstrated the domain adaptation capabilities of Bidirectional Encoder Representations from Transformers (BERT), a language model based on the transformer architecture, in fields such as biology, chemistry, and

* Corresponding author.
E-mail address: huxm@nju.edu.cn (X. Hu).

https://doi.org/10.1016/j.gr.2025.04.017
Received 18 September 2024; Received in revised form 15 April 2025; Accepted 30 April 2025
Available online 26 May 2025

materials science (Beltagy et al., 2019; Lee et al., 2020; Gupta et al., 2022). Although studies have not yet been conducted specifically for sedimentology, BERT could address the problem of identifying sedimentological literature at sub-disciplinary level.

Scientific knowledge is constituted by concepts and relations embodied in publications. Consequently, lexical analysis of scientific texts serves as a methodological lens to decode the knowledge evolution. For example, Milojević (2015) used lexical diversity to trace the conceptual territory of science with time, while Kedrick (2024) used networks of noun phrases to reconstruct the growth of scientific knowledge. Furthermore, citation pattern also exhibits significant shifts during scientific revolutions (Sinatra et al., 2015; Funk et al., 2017; Park et al., 2023). The persistent citation of classic papers (e.g., established paradigms) becomes gradually terminated by the proliferation of newly published studies introducing novel paradigms.

In this work, we introduce SedBERT, a domain-specific BERT model trained to distinguish sedimentology articles from other scientific papers. Utilizing the resulting data set, we analyze the evolution of knowledge in the field of sedimentology over the past 120 years through lexical analysis and citation pattern.

2. Methods

We identified sedimentological publications in journals not explicitly labelled as sedimentological journals by using the SedBERT model, trained based on BERT through 4001 hand-labelled data. 3,002,923 bibliographic records of 149 geology-related journals spanning from 1902 to 2021 were obtain from the open accessed wed data. Then, publications classified as sedimentological by the SedBERT model were used for bibliometric analysis. Lexical diversity (LD) proxies, including hypergeometric distribution divergence (HD-D) and Measure of Textual Lexical Diversity (MTLD), were calculated using the Python package 'lexical-diversity.' Clusters of sedimentological publications were generated through bibliographic coupling methodology utilizing VOSviewer software (version 1.6.18).

2.1. Data collection

Given that journal articles constitute the principal type of publications in solid Earth science, sedimentological research is adequately represented by journal articles. However, the distribution of articles across journals is notably uneven; approximately 80 % of articles were published in 40 % of the journals. This disparity implies that not all journals are essential for this study. Consequently, only journals listed in the Journal Citation Reports (JCR) are considered. Initially, 444 journals categorized within the fields of Geology, Geosciences (Multidisciplinary), Geochemistry & Geophysics, and Paleontology were selected. Subsequently, 149 of these 444 journals were identified as the primary research objects, deemed to have the highest potential for containing sedimentological publications. This selection was made by experienced sedimentologists using a straightforward criterion: journals where sedimentological publications constitute more than 10 % of total output are included. This study was conducted in August 2022. Bibliometric data were collected from open web data. We only use metadata in these articles which are publicly available from the open web data. Articles on sedimentology published in these 149 selected journals are presumed to encompass the vast majority of sedimentological literature. Details of these 149 and 444 journals are illustrated in Supplementary Table S1. For analytical purposes, the data were narrowed down to cover the period from 1902 to 2021.

2.2. Construction of the SedBERT model

In the development of SedBERT, a total of 4,001 papers were selected at random from a corpus of 149 key journals. To facilitate the training process, a panel comprising experienced sedimentologists with diverse research backgrounds was convened. Their task involved annotating these papers, and determining whether each should be classified as a sedimentological study. The SedBERT model, an evolution of the BERT-base-uncased architecture, was rigorously fine-tuned using this annotated dataset, which included both titles and abstracts. This refinement has enabled SedBERT to adequately categorize articles as either pertaining to or outside the domain of sedimentology. Then, text data from articles (specifically titles and abstracts) across the 149 key journals were converted into eigenvectors and processed through SedBERT. The model employs a classification threshold of 0.4; thus, when an article yields a probability exceeding this value, SedBERT classifies it as a sedimentological publication. Interested parties can access SedBERT's pre-trained weights at https://github.com/Eden980429/sedimentology_analysis/tree/main.

2.3. Lexical diversity indices

We completed a series of preprocessing steps using spaCy, an open-source, state-of-the-art Python package for natural language processing, and textacy, a python package to clean text. To begin, we tokenized each title. From the resulting lists of tokens, we then excluded those that were tagged by spaCy as stop words, tokens consisting only of digits or punctuation. Next, we converted all remaining tokens to their lemmatized form and converted all letters to lowercase. After that we used the lexical diversity package to calculate the HD-D and MTLD.

3. Result

204,023 publications classified as sedimentological by the SedBERT model. To assess the SedBERT model's performance, three datasets were used in this study. **Dataset 1** includes 200 papers from the three journals Sedimentology (1365–3091), Sedimentary Geology (1879–0968) and Journal of Sedimentary Research (1938–3681); **Dataset 2** includes 200 papers from 149 Earth science journals; **Dataset 3** includes 1000 multidisciplinary publications from Nature (1476–4687). A control group was established through keywords-based searches, utilizing a list of 215 keywords from National Natural Science Foundation of China and Chinese Academy of Sciences (2023) to identify sedimentology literature from the Web of Science database. Evaluation metrics included Precision, Recall, and F1 score. As Table 2 illustrates, the SedBERT model outperforms keyword-based searches in classifying sedimentological papers, particularly in journals containing multidisciplinary papers.

4. Discussion

4.1. Publication growth in sedimentology

The number of sedimentology publications grew at an exponential rate following an oscillatory growth between 1902 and1920, doubling every 10.3 years since 1945 (Fig. 1A). The roughly exponentially expansion was halted during World Wars I and II, and reduced after the 1980 s when the sedimentological publications doubled every 14 years.

The rapid development of sedimentology is expected to have created new research subfields, resulting in a burst of researchers and publications, or to be related to a 'golden age' driven by the strong need for new sedimentology studies following oil price increases and worldwide growing demand. However, the growth curve observed for the sedimentology sub-field is indistinguishable from the growth of science in general, thus driven by the development of society. A similar growth is observed for many other scientific fields (Sinatra et al., 2015; Jost and Restrepo, 2022), which supports the view that the growth of science is primarily related to economic and social development (Bornmann et al., 2021).

The exponential growth of sedimentological publications results from the rapid increase in the number of researchers in the sub-field,

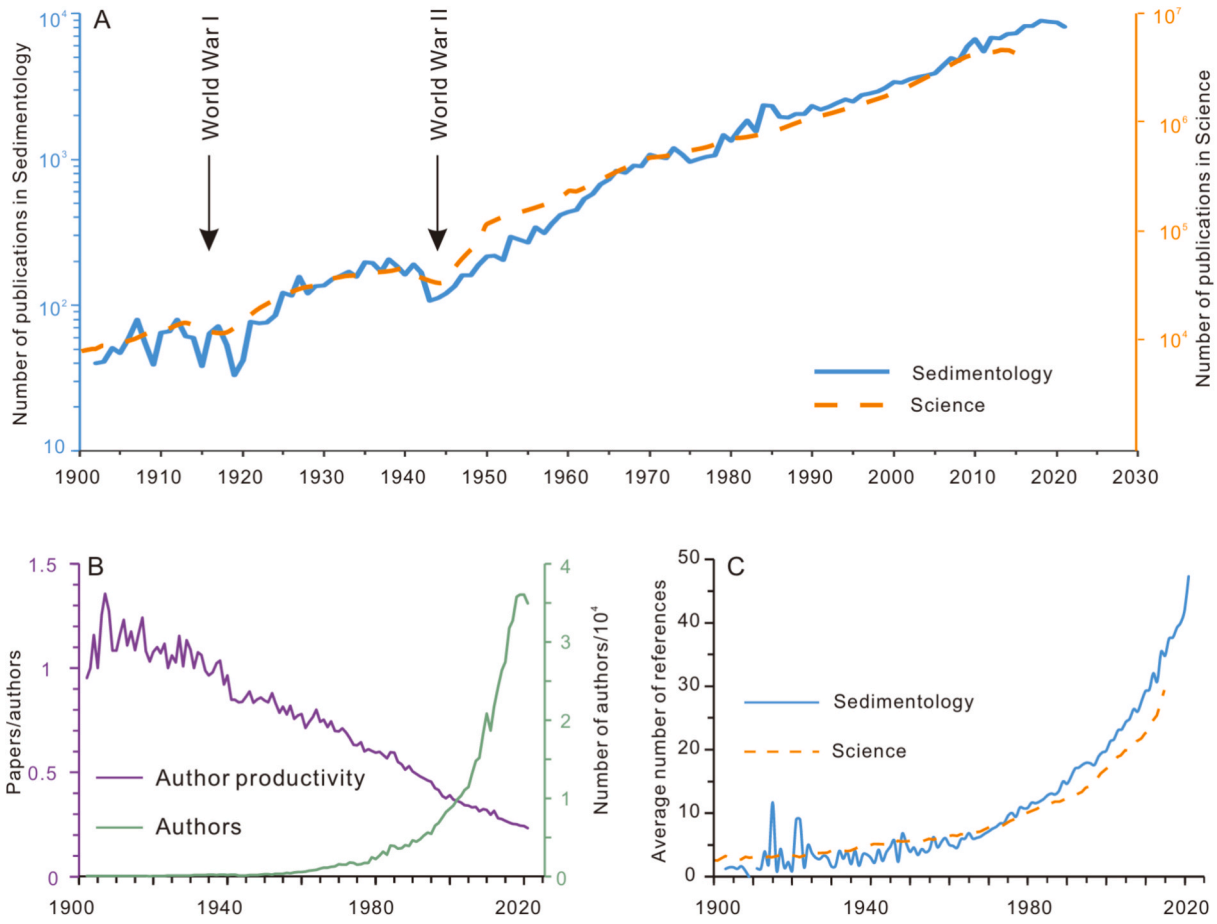


Fig. 1. The evolution of sedimentology. A) Changes in the number of papers published in sedimentology and science over the past 120 years. Growth patterns are indistinguishable, implying economic-driven increase. B) The number of authors increased at an exponential rate as the number of publications, but author productivity declined since the 1910 s. The term "author" refers to all contributors listed across the analyzed publications. C) Growth in the average number of references per paper vs. number of publications is also similar in sedimentology and science. Data for publications and references in Science after Dong et al. (2017).

coupled with the increment in the authors' yearly productivity. However, a drastic drop of 85 % is observed for the average individual productivity from the 1900 s to 2020 s, as measured by the ratio between the total number of publications and the number of active authors (Fig. 1B). This indicates the key role played by the increase of authors number rather than boosted individual productivity.

4.2. Knowledge growth in sedimentology

The number of publications increases and new data continuously accumulate, as reflected in the exponential growth of reference numbers in the fields of sedimentology and science (Fig. 1C). However, does such data expansion correspond to substantial knowledge evolution in sedimentology? If this were true, then we would expect a more rapid advance in the knowledge of sedimentology in recent decades than in the 1960–1970 'golden age,' which was previously considered as corresponding to a paradigm shift in the sub-field (Miall, 1995; Friedman, 1998).

The growth of scientific knowledge can be empirically inferred from the expansion of terms, as new words are invariably introduced with a paradigm shift (Kuhn, 1962; Park et al., 2022). The Lexical Diversity (LD) indices can measure the range and variety of unique words in the title of sedimentological papers (McCarthy and Jarvis, 2010) and are thus used to assess the growth of sedimentological knowledge. The LD indices HD-D and MTLD (see details in Table 1) exhibited significant variability between 1900 and 1950, primarily due to the limited number of papers published during this period. Both HD-D and MTLD rapidly increase in papers published between 1950–1970, suggesting a drastic knowledge expansion (Fig. 2A), followed by a steady but much slower growth of LD indices in 1970–1990 and little variation after the 1990 s, signaling the lack of major breakthroughs in sedimentology.

The citation style also reveals a similar knowledge growth pattern. The average reference age of papers (age of publication minus average age of its citing paper) reached a low point in the 1970 s (Fig. 2B), followed by a gradual increase in the age of references until the present. In general, authors tend to cite both foundational (older) works and recent developments (newer papers) to support their arguments, leading to a gradual increase in the average reference age. However, this pattern can be disrupted by paradigm shifts in scientific research. Such shifts typically result in a decline in citations of older literature, as the introduction of a new paradigm marks a departure from established frameworks (Funk and Smith, 2017). In contrast, more recent publications that adopt and promote the new paradigm tend to attract increased citation attention. Consequently, the average reference age is expected to decrease during periods of paradigm shift. This mechanism likely explains the observed decline in the average reference age within sedimentology during the 1950 s to 1970 s, coinciding with a significant revolution in the sedimentology (Miall, 1995).

Compared to citation behaviors across the sciences, the shift from deep (older) to myopic (younger) referencing styles during the 1960 s-1970 s is unique to sedimentology (Fig. 2B). Because a significant volume of new knowledge and ideas emerged in studies published in the previous seven years, sedimentologists had to focus on citing these more recent works. This drastic change coincides with a marked lexical

Table 1
Glossary of main terms as used in this paper.

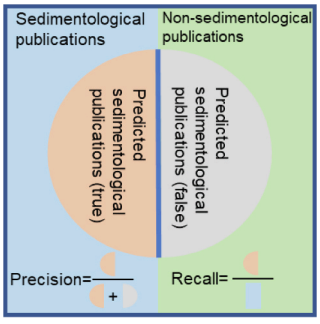
Term	Definition
BERT	BERT, a transformer-based large language model pre-trained for deep contextual language understanding, was adapted in this study to classify scientific publications as sedimentological or non-sedimentological. The methodology comprised three stages: 1. Dataset Preparation: Titles and abstracts from manually labeled sedimentological/non-sedimentological publications were compiled as input data. 2. Contextual Vector Encoding: Text (e.g., "*Carbonate are born, not made.*") was tokenized into subwords (["carbonate", "are", "born", ".", "not", "made", "."]). Each token was mapped to a 768-dimensional vector (e.g., "born" → [0.4, −0.1, 0.8, …, −0.2]) using BERT's embedding layers and transformer architecture, which dynamically adjusts vectors to reflect context. 3. Fine-Tuning: A classification layer was added to BERT's output, and the entire model was trained on labeled data to optimize weights/biases for distinguishing publication types.
Precision, Recall,F1 scores	Three metrics were employed to evaluate SedBERT's performance: 1. Precision: The proportion of correctly predicted sedimentological publications among all publications predicted as sedimentological. 2. Recall: The proportion of correctly predicted sedimentological publications among all true sedimentological publications in the dataset. 3. F1 score: is the average of Precision and Recall, prioritizing balance between the two metrics.  F1=(Precision+Recall)/2
Lexical diversity	Lexical diversity quantifies the richness and variability of distinct terms within a text. In scientific publications, where knowledge evolves through the introduction of novel concepts and their interrelations, an expansion of domain-specific terminology is expected as disciplines advance, leading to higher lexical diversity in publications over time. Several proxies are proposed the lexical diversity of text. For example, Type-Token Ratio (TTR), refers the ratio of unique words to total words. For instance, the sentence *"sedimentary rocks are formed by sedimentary processes"* contains 7 tokens (total words) but only 6 types (unique words, with "sedimentary" repeated), yielding a TTR of 6/7. But TTR decreases artificially with longer texts due to increased word repetition probabilities, thus not adopt in this study.
Measure of Textual Lexical Diversity (MTLD)	To avoid its sensitive to text length, the text is divided into sequential "segments," each extended as long as possible until its TTR (number of unique words divided by total words in the segment) falls below a predefined threshold (typically 0.72). then, the MTLD is calculated by the total words divided by the sum of segments, thus means the mean length of segments. Higher values indicate greater lexical diversity.
Hypergeometric distribution divergence (HD-D)	HD-D is another proxy of lexical diversity without being too sensitive to text length. HD-D looks at the probability of seeing each word type when drawing a small random sample (usually 42 words) from the text, based on the hypergeometric distribution. For every word type, it calculates the chance that the word would appear at least once if we randomly picked 42 words from the text. Then, all these chances are added up to get the final HD-D score. A higher HD-D score means the text has more lexical diversity (more variety of words).
Clusters	This is a proxy to deciphering research area derived from a citation network using the bibliographic coupling method. The network is composed of nodes and connections, where nodes are publications of sedimentology in this study. When two nodes share a common reference, there is a connection between them. The strength of a connection within the network increases with the number of shared cited works, leading to a closer spatial distribution and the formation of node clusters.
Cluster size	As the clusters are defined as group of publications connected by their references, the size of clusters refers the number of publications within this cluster.
Coefficient of variance (CV)	To quantify the annual publication distribution within individual clusters, we first calculated the standard deviation of publication counts per cluster within a single year. However, cross-year comparisons of publication distributions using standard deviation are not statistically valid, as this metric is inherently influenced by both the dispersion of data (variability) and the magnitude of the dataset (i.e., the total number of publications). To address this scale-dependent bias, we adopted the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean publication count. This normalization allows for unbiased comparisons of publication distribution patterns across years by eliminating the confounding effect of varying publication volumes. For example, the standard deviation of 1965, 1985 and 2020, are 3.4, 22.9 and 727.8 respectively. But their CV value is 0.34, 0.525, and 0.89.
Correlation coefficient (CC)	The correlation coefficient is a statistical measure used to assess the strength and direction of the linear relationship between two variables. In this study, we first calculated the linear relationship between cluster size (measured by the number of publications) and total citations/Top 10 % publications (representing their impact) within individual years. Subsequently, we evaluated how this relationship varied across different years by comparing the correlation coefficients obtained for each year.

Table 2
Comparison of SedBERT and Keyword Search Performance in Classifying Sedimentological Literature (F1 Score represents the average of Precision and Recall).

Datasets	SedBERT model		Keywords search	
Dataset 1	Precision	100 %	Precision	98.1 %
	Recall	87.2 %	Recall	82.4 %
	F1	93.6 %	F1	90.3 %
Dataset 2	Precision	67.2 %	Precision	52.5 %
	Recall	81.8 %	Recall	81.8 %
	F1	74.5 %	F1	67.2 %
Dataset 3	Precision	75.0 %	Precision	20.0 %
	Recall	92.3 %	Recall	100 %
	F1	83.7 %	F1	60.0 %

diversity shift during this period, indicating rapid knowledge expansion. The adoption of deep referencing citation styles in both sedimentology and broader scientific disciplines after the 1970 s has been interpreted as a consequence of the introduction of peer review and digital publication databases, which prompted authors to more meticulously credit earlier work (Sinatra et al., 2015; Dong et al., 2017). However, a slowed rate of knowledge growth may also have played a significant role.

Both lexical diversity and citation styles indicate a period (1950–1970) of rapid growth of sedimentological knowledge, potentially attributable to what is referred to as the 'golden age' of the subfield. Two revolutionary periods in sedimentology are widely

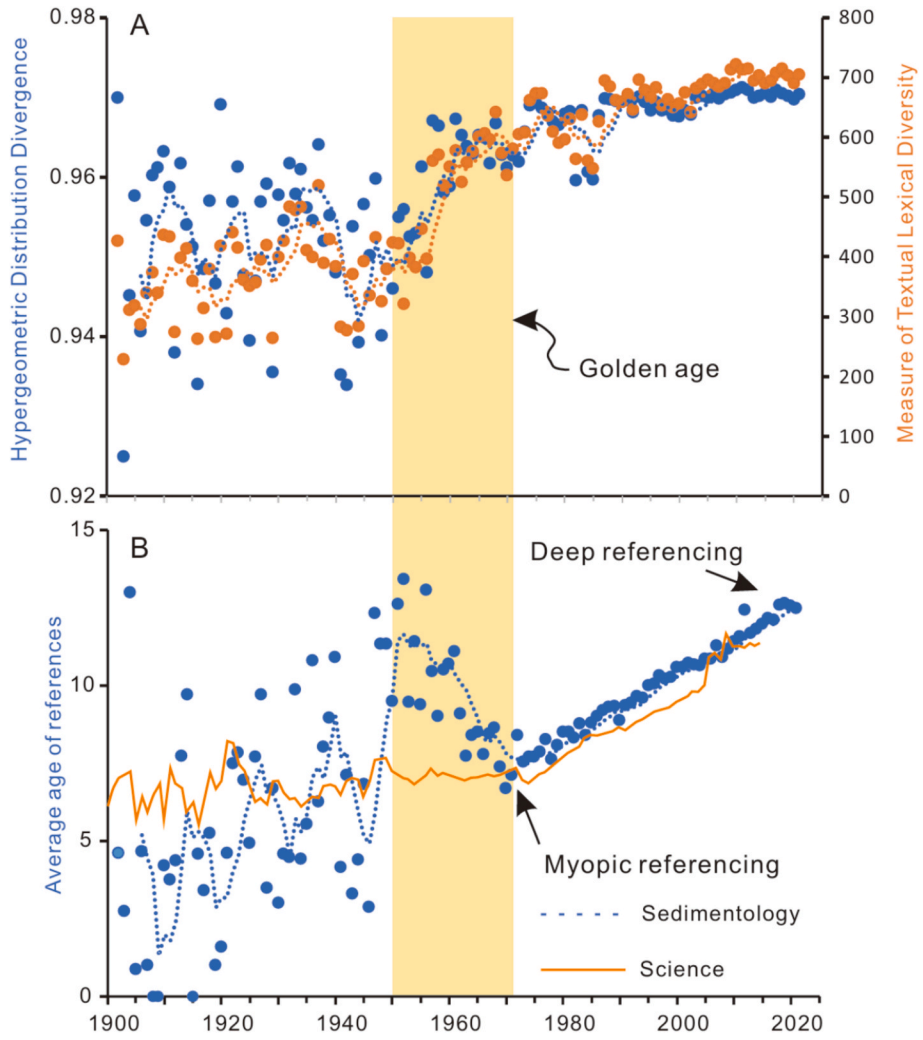


Fig. 2. The paradigm shift in sedimentology. A) Two lexical diversity proxies (HDD and MTLD) calculated from titles of sedimentology papers show a drastic increment at the 'golden age'. Because HDD and MTLD are independent of text length, the calculated lexical diversity can be used to represent knowledge growth. B) The average age of a reference is defined as the time difference between the date of publication and the average date of its citing papers. During the 1950 s-1970 s, citation style evolved from deep (referencing older papers) to myopic (referencing younger papers). Such behavior is not shown in the general field of science. The dashed lines in A-B represent a 5-point average trendline. Date for the evolution of referencing behavior in science from Dong et al. (2017).

recognized by sedimentologists (Miall, 1995; Friedman, 1998; Middleton, 2003). The first, the 'golden age', was characterized by the development of process-response sedimentary models and the application of plate-tectonic concepts. The second period (1970–1990) relates to the evolution of sequence stratigraphy. However, this study suggests that unlike the significant shift during the 'golden age', the latter period did not fundamentally impact sedimentological knowledge to the same extent.

4.3. *Evolution of research diversity after the 'golden age'*

The research area of sedimentology is continually expanding over time, and the growth of new research topics would be expected after the paradigm shift. Bibliographic coupling methods reveal that the number of publications peaked around 1985, driven by a surge in studies on sedimentary environments (Table 3), supporting rapid topic generation during the 'Golden Age' (Fig. 3A). The subsequent decline in cluster numbers since the 1990 s is associated with an exponential increase in the number of average publications per cluster (Fig. 3B). However, this growth in publication numbers is not evenly distributed across all clusters, a limited number of clusters becoming to be dominant. This trend has intensified since 2010 (Table 3), as evidenced by a sharp rise in the coefficient of variation (CV) for publication distribution from ~ 0.5 to ~ 0.9 over the past decade (Fig. 3B).

In the context of the Matthew effect (i.e., 'the rich get richer'; Price, 1976), we examined whether larger clusters, characterized by a higher number of publications, garner disproportionately greater scholarly attention, and consequently exhibit enhanced scientific impact. Our findings revealed a relatively low correlation coefficient (CC) between cluster size and citation counts until 2000. The coefficient of variation for citation distribution between clusters remained relatively low before 2010 to sharply increase thereafter (Fig. 3C). Similar results are obtained by examining the Top 10 % cited publications, with a sharp increase in the correlation coefficient value since 2000 and a significant increase in coefficient of variation values since 2010 (Fig. 3D).

Despite variations in publication volume, each cluster present in the last century has received broadly unchanged citations and scholarly attention. Each topic thus exhibited a similar level of scientific impact, reflecting high research diversity (Fig. 3E). However, such equilibrium was disrupted since 2010, when researchers primarily started to focus preferentially on the larger clusters. Research topic diversity consequently declined as larger clusters started to receive a disproportionate number of highly-cited papers, consolidating scholarly focus into fewer clusters (Fig. 3F). The observed phenomenon may have triggered a

Table 3
Comparison of research clusters of sedimentology in 1965, 1985, and 2020.

Year	Number	Cluster name	Cluster size
1965	1	Sedimentary geochemistry and diagenesis	15
	2	Sedimentary structure	14
	3	Carbon isotope	14
	4	Mineralogy of carbonate	14
	5	Marine deposits	12
	6	Sedimentary records of early life in the earth	10
	7	History of sedimentary basin	9
	8	Quaternary sea level changes	7
	9	Sea level changes since last glacial maximum	7
	10	Sedimentary structure of gravity flow	7
	11	Organic geochemistry	7
	12	Sedimentary cycles	5
1985	1	Marine geochemistry	108
	2	Sedimentary tectonics and environments	76
	3	Sedimentary processes (aeolian and lake)	64
	4	Deep water sedimentology	61
	5	Glacial sedimentology	60
	6	Fluvial sedimentology	60
	7	Diagenesis and paleoclimate	56
	8	Organic sedimentology	50
	9	Quaternary paleoceanography	50
	10	Sr isotope and clay mineral diagenesis	47
	11	Sedimentary texture (focus on grain size)	44
	12	Carbonate geochemistry	42
	13	Sandstone and tectonics	37
	14	Sedimentary records of Mass extinction	35
	15	Costal depositional system	31
	16	The last deglaciation	31
	17	Quaternary monsoon	28
	18	Quaternary paleoclimate	26
	19	Sedimentary tectonics of Pacific and Atlantic	22
	20	Paleoclimate	16
	21	Ice age	8
	22	Weathering and diagenetic process	7
2020	1	Modern sedimentary processes	2589
	2	Controls on sedimentary sequence	1754
	3	Quaternary paleoclimate	1297
	4	Paleoceanography (redox condition)	974
	5	Sedimentary tectonics (detrital zircon)	669
	6	Paleoclimate	474
	7	Paleolandscape	419

positive feedback loop, whereby an increasing number of new researchers are drawn to work in more crowded and hot topics. This attraction is fueled by the relative ease and high impact with which papers in these areas are accepted, coupled with the expectation of receiving more grants in an increasingly competitive scientific environment (Garfield, 1996; Foster et al., 2015; Chu and Evans, 2021).

This tendency for scientists to pursue hot topics is evident at both the sub-disciplinary level and higher level of fields (e.g., biomedicine, environmental science, and physics; Grandjean et al., 2011; Wei et al., 2013; Li et al., 2017). It might be expected that an increasing number of publications in subfield positively correlates with the rate of scientific knowledge growth. However, in sedimentology, the knowledge growth rate appears to have significantly slowed over the past decades, as evidenced by stable lexical diversity and deep referencing style (Fig. 2). This trend of diminishing scientific innovations and breakthroughs is not isolated but extends across broader fields (Chu and Evans, 2021; Park et al., 2023). The slowdown in knowledge growth is ascribed to the decreased diversity of research topic. At the frontier of knowledge, where rapid advancements are most expected, the focus of most researchers on hot topics implies that numerous frontiers remain underexplored, thereby decelerating the overall growth rate of knowledge (Foster et al., 2015; Rzhetsky et al., 2015). Moreover, in hot areas, researchers produce a vast number of concordant publications, which carries the risk that papers containing novel ideas are less likely to be published, read, and cited (Chu and Evans, 2021). Scholars tend to align their work with established theories to increase the likelihood of publication and garner higher citation counts (Chu and Evans, 2021). As the growth of knowledge is often spurred by disruptive new concepts that may conflict with existing paradigms, scientific knowledge tends to progress incrementally by reinforcing the classical conceptual framework of established canons (Kedrick et al., 2024). Consequently, the 'follow the crowd' strategy ensures consistent productivity but reduces research diversity, thereby hindering rapid advancements in scientific knowledge (Foster et al., 2015; Rzhetsky et al., 2015; Fortunato et al., 2018; Bhattacharya and Packalen, 2020).

Although our results support that the progression of scientific knowledge at sub-disciplinary levels mirrors the trends observed at higher levels, the timing may vary across different levels. While the rate of knowledge growth in major fields of science has been slowing since the 1950 s (Park et al., 2023), sedimentological knowledge maintained a rapid growth rate until the 1970 s (Fig. 2). This underscores the importance of studying the dynamics of knowledge evolution at the sub-disciplinary level, as such analyses are crucial for policymakers to determine the optimal timing and nature of specific strategies designed to accelerate the growth rate of scientific knowledge across different fields.

5. Conclusion and implication

The main purpose of this work is to study knowledge landscape at the sub-disciplinary level. Our survey shows how the expansion of sedimentological knowledge started during the 1950 s to 1970 s 'golden age' followed by a period of significant growth, has slowed in the past decade. A notable decline in research diversity, largely attributable to a 'follow-the-crowd' strategy, has contributed to this stagnation. Without significant changes in scientific incentives, this trend may persist into the foreseeable future.

Investigating the sub-disciplinary landscape is of fundamental importance, as it helps us to understand the growth pattern of sub-disciplines in detail. Such an analysis enables researchers to pose deeper questions concerning the patterns and reasons for knowledge progress at sub-disciplinary levels, as well as the ways in which innovations reshape the research landscape. The natural language process (NLP) tools developed in this study can address these key questions in sub-disciplines of the Earth sciences and suitably influence policy-making and research perspective. While reports in geology are limited, these methods have been effectively applied in materials science, chemistry, and physics, highlighting their broad potential (Rzhetsky et al., 2015; Tshitoyan et al., 2019; Krenn and Zeilinger, 2020).

CRediT authorship contribution statement

Yiwei Xu: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Xiumian Hu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Eduardo Garzanti:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jiexing Qi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Huquan Kang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hanqi Nong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation,

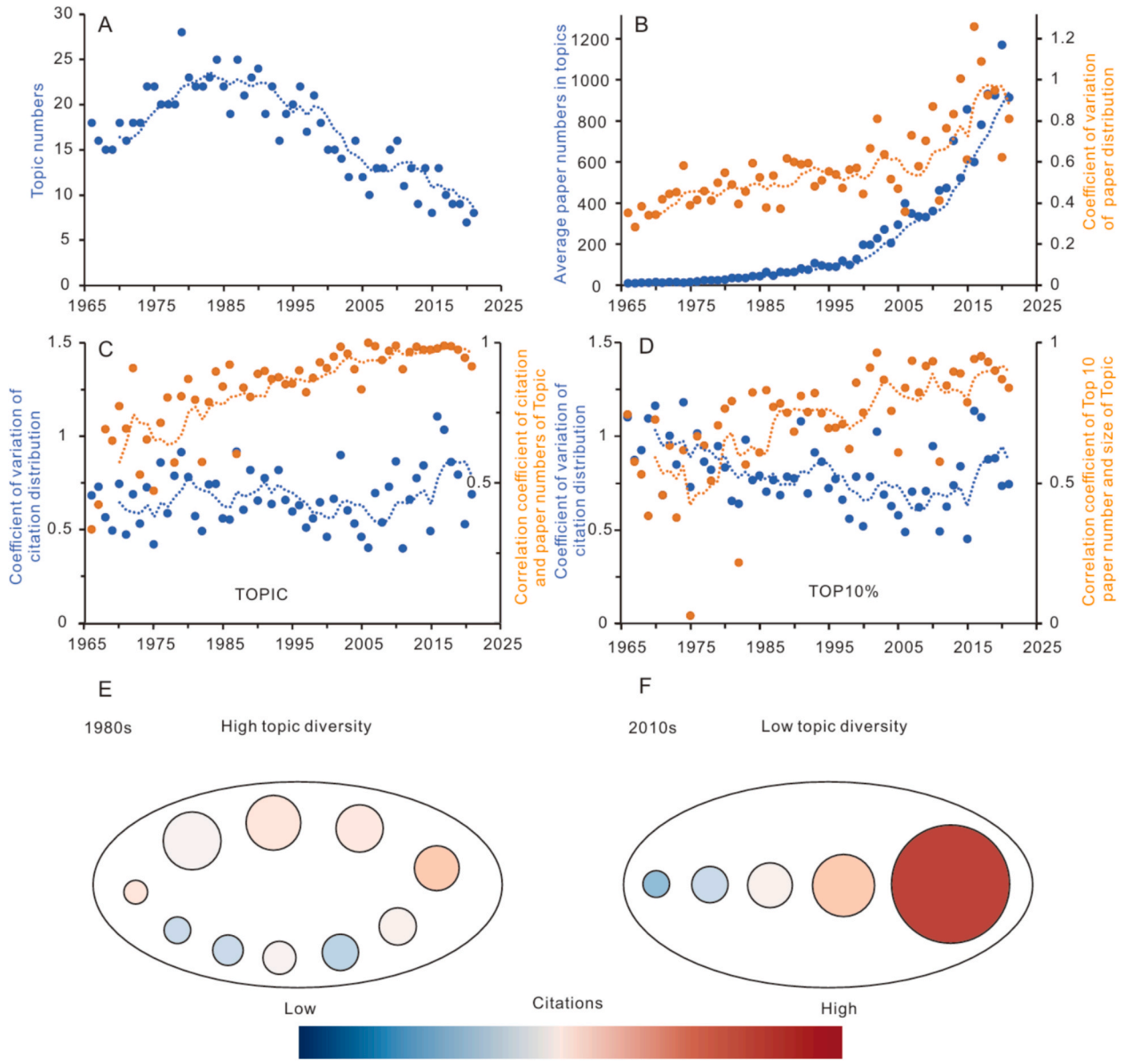


Fig. 3. The evolution of research topic diversity in sedimentology. A) The number of research topics, clustered through bibliographic coupling methods, peaked in the 1980 s, followed by a gradual decline until the present. B) Average number of publications on each topic gradually increased from 1965-1990 s, and the rate of increase became even higher afterward due to a decrease in the number of topics. The distribution pattern of publications in each topic is represented by the coefficient of variance index, calculated as the ratio of the standard deviation to the mean of publications in each topic in each year. The coefficient of variance values has increased significantly since the 2010 s, reflecting a more uneven distribution pattern. C) Coefficient of variance values indicate that also the distribution pattern among topic citations, defined as the sum of citations of all papers within the same topic, has become more uneven since the 2010 s, when a majority of citations concentrated in larger topics, as evidenced by a correlation coefficient between cluster size and citation reaching ~ 0.9. D) A similar pattern is observed in the distribution of the top 10 % cited papers each year, which tend to concentrate on larger-size topics as shown by coefficient of variance and correlation coefficient values. The dashed lines in A-D represent a 5-point average trendline. E-F). Alternative schematic models to explain the decline in research topic diversity from the 1980 s to the 2010 s, attributed to the 'follow-the-crowd' strategy. Each cycle represents a unique research topic. The diameter of a cycle indicates the volume of research within the topic, whereas its color indicates the number of citations received.

Funding acquisition, Formal analysis, Data curation, Conceptualization. **Luwen Wu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jiaxin Ding:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Xinbing Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Chengshan Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Jiangang Wang, Gaoyuan Sun, Juan Li, Shijie Zhang, Zhong Han, Anlin Ma, Wen Lai, and Wendong Liang for their help in annotation of sedimentological publication. We thank Chao Min for his valuable suggestions on this study. We also thank Peter DeCelles and one anonymous reviewer for their thoughtful feedback that helped improve this manuscript. This work was financially supported by the National Natural Science Foundation of China (42142004, 42050102), the National Natural Science Foundation of Jiangsu Province (BK20242112) and the Fundamental Research Funds for NIGPAS (NGBS202405). This paper contributes to the IUGS "Deep-time Digital Earth" Big Science Program.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gr.2025.04.017.

Data availability

Open-source code to calculate lexical diversity and sedimentological publications used in this study are available at https://doi.org/10.57760/sciencedb.12276. Analysis results of this study are included in supplementary sheets.

References

Beltagy, I., Kyle, L., and Cohan, Arman, 2019, Scibert: a pretrained language model for scientific text: Proceedings of EMNLP-IJCNLP, p. 3615–3620, Doi: 10.18653/v1/D19-1371.

Bhattacharya, J., Packalen, M., 2020. Stagnation and scientific incentives: National Bureau of. Econ. Res. https://doi.org/10.3386/w26752.

Bloom, N., Jones, C.I., Van Reenen, J., Webb, M., 2020. Are ideas getting harder to find? Am. Econ. Rev. 110 (4), 1104–1144. https://doi.org/10.1257/aer.20180338.

Bornmann, L., Haunschild, R., Mutz, R., 2021. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. Human. Soc. Sci. Commun. 8 (1), 1–15. https://doi.org/10.1057/s41599-021-00903-w.

Chu, J.S., Evans, J.A., 2021. Slowed canonical progress in large fields of science. Proc. Natl. Acad. Sci. 118 (41), e2021636118. https://doi.org/10.1073/pnas.2021636118.

Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., Yahia, M., 2011. Education: the PhD factory. Nature 472 (7343), 276–280. https://doi.org/10.1038/472276a.

Dong, Y.X., Ma, H., Shen, Z.H., Wang, K.S., 2017. A century of science: Globalization of scientific collaborations, citations, and innovations. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1437–1446. https://doi.org/10.1145/3097983.3098016.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., Lim, W.M., 2021. How to conduct a bibliometric analysis: An overview and guidelines. J. Bus. Res. 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070.

Fortunato, S., Bergstrom, C.T., Borner, K., Evans, J.A., Helbing, D., Milojevi, S., Vespignani, A., 2018. Science of science. Science v. 359 (6379), eaao0185. https://doi.org/10.1126/science.aao0185.

Foster, J. G., Rzhetsky, A., and Evans, J.A., 2015, Tradition and innovation in scientists' research strategies: American sociological review, v. 80(5), p. 875-908, Doi: 10.1177/0003122415601618.

Friedman, G.M., 1998. Sedimentology and stratigraphy in the 1950s to mid-1980s: the story of a personal perspective. Episodes 21, 172–177. https://doi.org/10.18814/epiiugs/1998/v21i3/006.

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

Garfield, E., 1996. What is the primordial reference for the phrase 'publish or perish'. The Scientist 10 (12), 11.

Grandjean, P., Eriksen, M.L., Ellegaard, O., Wallin, J.A., 2011. The Matthew effect in environmental science publication: a bibliometric analysis of chemical substances in journal articles. Environ. Health 10, 96. https://doi.org/10.1186/1476-069X-10-96.

Gupta, T., Zaki, M., Krishnan, N.A., Mausam, 2022. MatSciBERT: A materials domain language model for text mining and information extraction. npj Comput. Mater. 8 (1), 102. https://doi.org/10.1038/s41524-022-00784-w.

Jost, J., and Restrepo, G., 2022. The Evolution of chemical knowledge: a formal setting for its analysis. Springer, 122 p, Doi: 10.1007/978-3-031-10094-9.

Kedrick, K., Levitskaya, E., Funk, R.J., 2024. Conceptual structure and the growth of scientific knowledge. Nat. Hum. Behav. 8 (10), 1915–1923. https://doi.org/10.1038/s41562-024-01957-x.

Krenn, M., and Zeilinger, A., 2020. Predicting research trends with semantic and neural networks with an application in quantum physics: Proceedings of the National Academy of Sciences of the United States of America, v. 117(4), p. 1910–1916, Doi: 10.1073/pnas.1914370116.

Kuhn, T.S., 1962. The structure of scientific revolutions. University of Chicago press, Chicago.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., Chan, H.S., Kang, J., 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36 (4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

Li, M.H., Yang, L.Y., Zhang, H.N., Shen, Z., Wu, C.S., Wu, J.S., 2017. Do mathematicians, economists and biomedical scientists trace large topics more strongly than physicists? J. Informet. 11 (2), 598–607. https://doi.org/10.1016/j.joi.2017.04.004.

McCarthy, P.M., and Jarvis, S., 2010, MTLD, vocd-D, and HD-D. A validation study of sophisticated approaches to lexical diversity assessment: Behavior Research Methods, v. 42, p. 381–392, Doi: 10.3758/BRM.42.2.381.

Miall, A.D., 1995. Whither stratigraphy? Sed. Geol. 100, 5–20. https://doi.org/10.1016/0037-0738(95)00100-X.

Middleton, G.V., 2003, Sedimentology, History, in Middleton, G. V.,ed., Encyclopedia of sediments and sedimentary rocks, Berlin: Springer-Verlag, p. 628-635, Doi: 10.1007/978-1-4020-3609-5_186.

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

National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS), 2023. China's Discipline Development Strategy: Sedimentology (in Chinese). Beijing: Science Press, 346 P.

Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.J., Hiszpanski, A.M., 2020. Data-driven materials research enabled by natural language processing and information extraction. Appl. Phys. Rev. 7, 041317. https://doi.org/10.1063/5.0021106.

Park, S., Carrapa, B., Ducea, M.N., Surdeanu, M., Hayes, R., Collins, D., 2022. Answering geosciences research questions at a global scale via a hybrid machine-human learning approach: a case study of the link between climate and volcanism. GSA Today 32 (11), 4–8. https://doi.org/10.1130/GSATG528A.1.

Park, M., Leahey, E., Funk, R.J., 2023. Papers and patents are becoming less disruptive over time. Nature 613, 138–144. https://doi.org/10.1038/s41586-022-05543-x.

Price, D.D.S., 1976. A general theory of bibliometric and other cumulative advantage processes. J. Am. Soc. Inf. Sci. 27 (5), 292–306. https://doi.org/10.1002/asi.4630270505.

Rzhetsky, A., Foster, J.G., Foster, I.T., and Evans, J.A., 2015. Choosing experiments to accelerate collective discovery: Proc. Nat. Acad. Sci, v. 112(47), p. 14569-14574, Doi: 10.1073/pnas.1509757112.

Sinatra, R., Deville, P., Szell, M., Wang, D., Barabási, A.L., 2015. A century of physics. Nat. Phys. 11 (10), 791–796. https://doi.org/10.1038/nphys3494.

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

Wei, T., Li, M.H., Wu, C.S., Yan, X.Y., Fan, Y., Di, Z.G., Wu, J.S., 2013. Do scientists trace hot topics? Sci. Rep. 3, 2207. https://doi.org/10.1038/srep02207.