

MACHINE LEARNING IN ENVIRONMENTAL METAGENOMIC APPROACHES: A BIBLIOMETRIC AND THEMATIC ANALYSIS

Sevtap TIRINK^{1*}

¹Iğdır University, Health Services Vocational School, Department of Medical Services and Techniques, Environmental Health Program, Iğdır, Türkiye

Abstract: Machine learning is increasingly integrated with metagenomic analysis, yet the development, conceptual organization, environmental applications, and recent thematic directions of this research area require systematic characterization. This study examined English-language article and review records indexed in the Web of Science Core Collection from 2016 to 2025 using a predefined search strategy. In this study, bibliometric and science-mapping analyses were conducted to examine annual scientific production, source productivity, keyword co-occurrence, thematic structure, trend topics, and geographic and collaboration patterns. The analysis covered 286 documents published across 139 sources, with an annual growth rate of 54.79%, and publication output reached 102 documents in 2025. Metagenomics and machine learning formed major conceptual foundations, with thematic structure linking computational methods to microbial ecology, diversity, soil, and bacterial communities. Keyword co-occurrence further distinguished interconnected groupings centered on machine learning–metagenomics–bioinformatics, soil–rhizosphere and nutrient-related research, bioremediation–biodegradation, and resistome–wastewater surveillance. Recent trend topics included artificial intelligence, prediction, and degradation. When considered together with the thematic-map position of degradation–bioremediation–artificial intelligence, these patterns support a cautious interpretation that this research direction is recently gaining visibility rather than representing a definitively emerging or dominant research front. Scientific production was geographically concentrated, particularly in China and the United States, while international collaboration was also evident. Overall, the field shows increasing methodological diversity, with reference-based approaches being complemented by predictive and explainable methods

Keywords: Environmental metagenomics, Machine learning, Science mapping, Microbial ecology, Bibliometrics

*Corresponding author: Iğdır University, Health Services Vocational School, Department of Medical Services and Techniques, Environmental Health Program, Iğdır, Türkiye

E mail: sevtaptirink@gmail.com (S. TIRINK)

Sevtap TIRINK  <https://orcid.org/0000-0003-0123-0054>

Received: April 21, 2026

Accepted: June 06, 2026

Published: June 15, 2026

Cite as: Tırink, S. (2026). Machine learning in environmental metagenomic approaches: a bibliometric and thematic analysis. *Black Sea Journal of Artificial Intelligence*, 2(1), 7-17.

1. Introduction

Environmental metagenomics provides a culture-independent framework for characterizing microbial communities directly from environmental samples through the analysis of their genetic material (Zhang et al., 2021). It has been applied across diverse matrices, including soils, surface and groundwater, aquatic sediments, air and bioaerosols, and sewage-associated environments, allowing microbial communities to be investigated across a wide range of terrestrial and aquatic ecosystems (Liu et al., 2025). Unlike cultivation-dependent approaches, metagenomic analysis does not require microorganisms to be isolated and grown under laboratory conditions before their genetic characteristics are examined. This is particularly relevant in environmental microbiology, where many microorganisms remain difficult to cultivate using conventional laboratory methods. By reducing this dependence on cultivation, metagenomic approaches extend access to microbial diversity that may be poorly represented in culture-based studies, including difficult-

to-culture and low-abundance members of microbial communities (Chettri et al., 2024).

Building on these capabilities, environmental metagenomics are widely used to characterize microbial diversity and taxonomic structure across natural ecosystems (Ghannam and Techtmann, 2021). Beyond taxonomic characterization, genome-resolved metagenomic analyses can link uncultivated microbial lineages with metabolic functions involved in major biogeochemical processes, including carbon, nitrogen, and sulfur transformations that contribute to ecosystem functioning (Mirete et al., 2025). Metagenomic information is also used in environmental monitoring, where changes in microbial community composition and source-associated microbial markers can support the assessment and tracing of contamination within environmental systems (Böke Özkoç et al., 2025; Chettri et al., 2024). In remediation contexts, metagenomic analysis can further identify microbial functions associated with pollutant transformation and degradation, thereby informing the evaluation of strategies such as



biostimulation and bioaugmentation (Böke Özkoç et al., 2025; Chettri et al., 2024). The increasing use of metagenomics across these ecological and applied contexts, however, also produces datasets that place substantial demands on data processing, interpretation, and integration. High-throughput metagenomic sequencing generates large datasets that require considerable computational resources for storage, preprocessing, and downstream analysis (Sharma et al., 2025). In shotgun metagenomics, these challenges are further intensified by heterogeneous DNA mixtures, uneven organismal abundance and sequence coverage, and complex community composition, all of which can complicate genome reconstruction and downstream analysis (Mathieu et al., 2022). Metagenomic datasets are also frequently high-dimensional, with the number of taxa, genes, or sequence-derived features often exceeding the number of available samples. Their analysis is further complicated by sparsity and compositional constraints, which can affect conventional statistical interpretation if not appropriately addressed (Sharma et al., 2025; Roy et al., 2024). Incomplete taxonomic and functional reference databases present an additional limitation, particularly for sequences derived from poorly characterized or uncultivated microorganisms (Pallavi and Uday Bhaskar, 2025). Moreover, differences in sample processing, DNA extraction, sequencing platforms, and related methodological choices can introduce technical heterogeneity and affect reproducibility across studies (Papoutsoglou et al., 2023). Addressing these challenges requires analytical approaches capable of handling high-dimensional, heterogeneous, and structurally complex metagenomic data. Machine-learning approaches have consequently been applied across several stages of metagenomic analysis. Sequence-derived features can be used by supervised learning methods to classify metagenomic reads or contigs into taxonomic groups, supporting taxonomic assignment in large and heterogeneous sequence datasets (Tonkovic et al., 2020). Machine-learning methods have also been applied to metagenomic binning and genome reconstruction, where neural-network and representation-learning approaches can assist in grouping contigs and recovering metagenome-assembled genomes from complex sequence data (Herazo-Álvarez et al., 2025). Metagenomic features can likewise be incorporated into predictive models to support the classification or inference of functional characteristics (Wassan et al., 2019). In environmental research, machine learning has been applied to microbial community data for environmental assessment and prediction, including the identification of patterns associated with ecosystem conditions and contamination status (Ghannam and Techtmann, 2021). Environmental microbial profiles may also support geographic or source classification of samples (Ghannam and Techtmann, 2021). The breadth of these applications makes it increasingly relevant to examine how research linking machine learning with metagenomics has developed and

been organized within the scientific literature.

Previous reviews and bibliometric studies have approached this intersection from several related perspectives. Tonkovic et al. (2020) provided the closest synthesis to the present topic through a scoping review of applied machine learning in metagenomic classification; however, its scope is broadly metagenomic and only partially environmental. Marcos-Zambrano et al. (2021), by contrast, reviewed statistical and machine-learning approaches primarily in human and clinical microbiome research, including disease-related applications. At a broader disciplinary level, Xu et al. (2022) mapped artificial intelligence research across biotechnology and applied microbiology rather than focusing specifically on environmental metagenomics. Abdelwahab et al. (2024) similarly examined the development of AI-microbiome research, with much of the literature centered on gut microbiota, disease prediction, and other human-health contexts. Velasco Cardona et al. (2025) analyzed AI applications in bacteriology, particularly bacterial identification, diagnostics, and antimicrobial resistance. Thus, although these studies demonstrate substantial interest in AI- and machine-learning-related microbial research, their primary scopes address domains that only partially overlap with machine learning in environmental metagenomics.

Taken together, these studies provide valuable perspectives on related areas, but they only partially capture the bibliometric and thematic development of research at the intersection of machine learning and environmental metagenomics. The present study therefore characterizes the literature at this intersection during 2016–2025 by examining publication dynamics and productivity, influential sources, geographical and collaborative patterns, citation impact, and the thematic structure and temporal development of research topics

2. Material and Methods

2.1. Data Source and Search Strategy

The scientific dataset used in this study was obtained from the Web of Science (WoS) Core Collection, one of the most widely recognized and rigorously curated multidisciplinary citation databases. WoS is commonly used in scientometric studies due to its transparent indexing criteria, comprehensive metadata, and robust citation structure (Donthu et al., 2021). The search was designed to identify publications at the intersection of metagenomics, machine learning or artificial intelligence, and environmental applications. Search terms were applied to the title (TI), abstract (AB), and author keyword (AK) fields. The search expression used in WoSCC was retained without modification:

(TI=(metagenom*) OR AB=(metagenom*) OR AK=(metagenom*)) AND (TI=("machine learning" OR "deep learning" OR "artificial intelligence" OR "random forest*" OR "support vector machine*" OR "neural network*" OR "convolutional neural network*" OR "graph neural network*" OR "gradient boosting" OR XGBoost OR

LightGBM) OR AB=("machine learning" OR "deep learning" OR "artificial intelligence" OR "random forest*" OR "support vector machine*" OR "neural network*" OR "convolutional neural network*" OR "graph neural network*" OR "gradient boosting" OR XGBoost OR LightGBM) OR AK=("machine learning" OR "deep learning" OR "artificial intelligence" OR "random forest*" OR "support vector machine*" OR "neural network*" OR "convolutional neural network*" OR "graph neural network*" OR "gradient boosting" OR XGBoost OR LightGBM)) AND (TI=("environmental metagenom*" OR "environmental microbiome*" OR soil* OR rhizosphere* OR sediment* OR freshwater* OR "surface water*" OR river* OR stream* OR lake* OR estuar* OR marine OR ocean* OR seawater OR groundwater OR wastewater OR sewage OR sludge OR "activated sludge" OR wetland* OR landfill* OR compost* OR "built environment*" OR airborne OR "air microbiome*" OR bioremediat*) OR AB=("environmental metagenom*" OR "environmental microbiome*" OR soil* OR rhizosphere* OR sediment* OR freshwater* OR "surface water*" OR river* OR stream* OR lake* OR estuar* OR marine OR ocean* OR seawater OR groundwater OR wastewater OR sewage OR sludge OR "activated sludge" OR wetland* OR landfill* OR compost* OR "built environment*" OR airborne OR "air microbiome*" OR bioremediat*) OR AK=("environmental metagenom*" OR "environmental microbiome*" OR soil* OR rhizosphere* OR sediment* OR freshwater* OR "surface water*" OR river* OR stream* OR lake* OR estuar* OR marine OR ocean* OR seawater OR groundwater OR wastewater OR sewage OR sludge OR "activated sludge" OR wetland* OR landfill* OR compost* OR "built environment*" OR airborne OR "air microbiome*" OR bioremediat*))

The search results were restricted to the 2016–2025 publication period using the recorded WoS search-history expression #2 AND (PY=(2016–2020) OR PY=(2021–2025)). Records assigned to 2026 were subsequently excluded (#3 NOT PY=(2026)).

The search results were restricted to the 2016–2025 period and to English-language Article and Review Article records. Articles and Reviews were retained because they represent the principal peer-reviewed research and synthesis outputs used for bibliometric comparison, while Early Access records were excluded to avoid duplicate or unstable year assignments and to maintain a consistent closed publication window through 2025. These database-level restrictions resulted in 286 documents. No manual relevance screening was performed beyond these restrictions.

2.2. Bibliometric and Temporal Analysis

Bibliometric and science-mapping analyses were conducted in R version 4.6.0 using the bibliometrix package and the Biblioshiny interface (Aria and Cuccurullo, 2017; Aria et al., 2026; R Core Team, 2026). Performance analyses included annual scientific production, source productivity, institutional and geographic contributions, and citation indicators,

following established bibliometric principles (Donthu et al., 2021). Geographic and collaboration patterns were examined through affiliation and corresponding-author country analyses, including single-country publications (SCP), multiple-country publications (MCP), country-level citation patterns, and international collaboration networks. Conceptual and thematic structure was evaluated using keyword-frequency analysis, keyword co-occurrence networks, thematic mapping, and Trend Topics (Aria and Cuccurullo, 2026). Source concentration was assessed using Bradford's Law and the Kolmogorov–Smirnov goodness-of-fit test.

Temporal publication dynamics were additionally evaluated using the Life Cycle of Scientific Production analysis based on a logistic growth model. Model fit was summarized using R^2 and RMSE, while model-derived indicators included the estimated saturation level, inflection point, peak annual publication output, and the interval between 10% and 90% of the estimated saturation level.

3. Results

3.1. General Characteristics and Temporal Development of Literature

The bibliometric dataset comprised 286 documents published across 139 sources between 2016 and 2025. These publications involved 1,854 authors, 17,638 cited references, and 944 distinct author keywords. Scientific production showed an annual growth rate of 54.79% over the study period. The literature was relatively recent, with a mean document age of 2.74 years, while the average citation count was 29.37 per document. Collaborative authorship was also prominent, with an average of 7.29 co-authors per document and international co-authorship accounting for 34.27% of the publications (Table 1).

Annual publication output increased substantially between 2016 and 2025 (Figure 1a). Publication output rose from two documents in 2016 to 13 in 2020 and 16 in 2021. The increase became more pronounced thereafter, reaching 31 documents in 2022, 39 in 2023, 66 in 2024, and 102 in 2025. Thus, more than one-third of all documents in the dataset were published in the final year of the study period. The Life Cycle of Scientific Production analysis provided a model-based characterization of the publication trajectory (Figure 1b). The fitted logistic model showed a close correspondence with the observed data ($R^2 = 0.995$; RMSE = 2.13). The model estimated a saturation level of 2,860 publications and an inflection point at approximately 2030.1, corresponding to a modeled maximum annual output of 339 publications. The estimated interval between 10% and 90% of the saturation level was 9.3 years. By 2025, the cumulative output of 286 documents corresponded to 10.0% of the modeled saturation level. Because the inflection point and modeled peak occur beyond the observed 2016–2025 period, these estimates represent model-based extrapolations rather than observed publication counts.

Table 1. General bibliometric characteristics of the dataset

Indicator	Value
Timespan	2016–2025
Sources	139
Documents	286
Annual growth rate	54.79%
Authors	1,854
Authors of single-authored documents	1
Co-authors per document	7.29
International co-authorship	34.27%
Author keywords	944
References	17,638
Document average age	2.74 years
Average citations per document	29.37

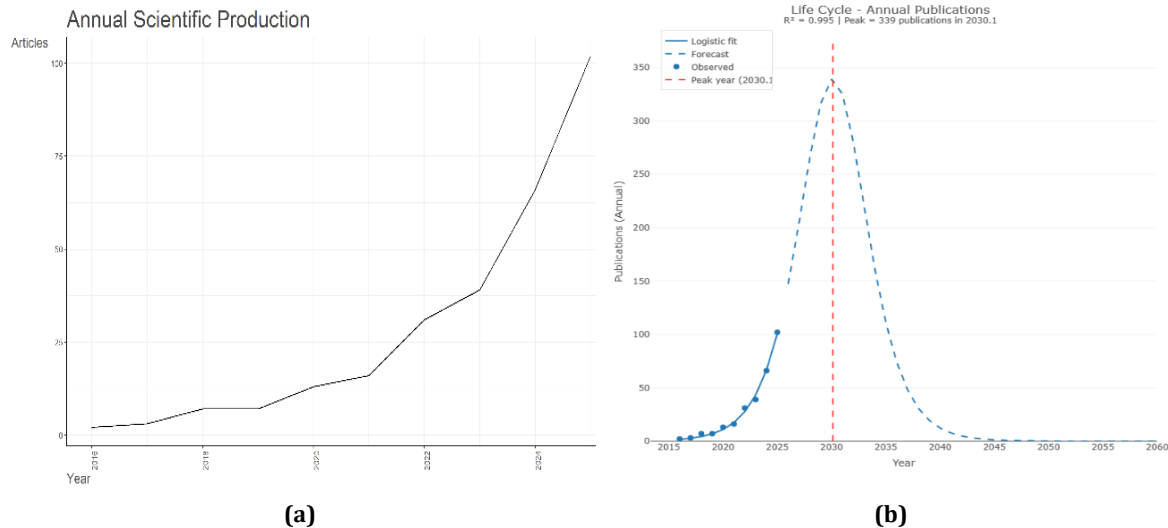


Figure 1. Temporal development of research on machine learning in environmental metagenomics: (a) annual scientific production during 2016–2025 and (b) Life Cycle of Scientific Production analysis showing the model-derived annual publication trajectory.

3.2. Scientific Sources and Citation Structure

Water Research was the most productive source with 14 documents, followed by Frontiers in Microbiology and Microbiome with 12 documents each. Water Research also had the highest local h-index ($h = 11$), whereas Frontiers in Microbiology received the highest number of local citations (599). The Bradford model showed a high coefficient of determination ($R^2 = 0.9766$) and a multiplier of 2.94. The three Bradford zones contained 11, 33, and 95 sources, respectively, with each zone contributing approximately one-third of the total publications. The Kolmogorov–Smirnov test did not provide sufficient evidence to reject the specified Bradford distribution ($D = 0.1367$, $P = 0.149$); this non-significant result should not be interpreted as proof of exact conformity.

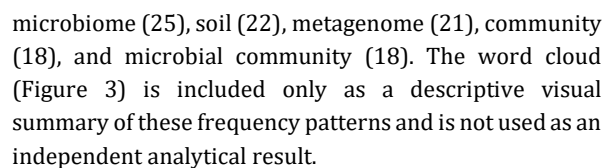
Citation analysis identified two documents with substantially higher global citation counts than the other highly cited publications in the dataset. Here, global citations refer to citations received by a document across the full Web of Science citation database, whereas local citations refer to citations received from documents within the 286-record collection analyzed in this study.

The most globally cited document was a 2022 article published in Nature Communications, with 865 citations, followed by a 2018 article published in Microbiome, with 699 citations. Other highly cited documents included articles published

Other highly cited documents included articles published in Environmental Science and Technology in 2017 (158 citations), FEMS Microbiology Ecology in 2018 (153), Nature Communications in 2022 (152), and Nature Microbiology in 2023 (151). Within the analyzed collection, the 2018 Microbiome article with 699 global citations and the 2022 Nature Communications article with 865 global citations each received 14 local citations.

3.3. Institutional, Geographic, and International Collaboration Patterns

Scientific production was unevenly distributed across institutions and countries. The Chinese Academy of Sciences was the most frequently represented affiliation, with 48 publications. It was followed by the University of California System with 20, Cornell University with 18, and the United States Department of Energy with 16. East China Normal University and Weill Cornell Medicine each



The keyword co-occurrence network identified four interconnected conceptual groupings (Figure 4). The first was centered on machine learning, metagenomics, microbiome, and related bioinformatics terms. The second comprised diversity, soil, rhizosphere, carbon, and nitrogen, reflecting a cluster of soil- and ecosystem-related concepts. A third grouping linked bacteria with bioremediation, biodegradation, and identification, whereas the fourth was characterized by resistome, antibiotic resistance genes, and wastewater. Connections between these groupings indicated that the identified conceptual domains were not completely isolated from one another.

The thematic map provided a complementary representation based on centrality and density (Figure 5). The bacteria–community–microbial community cluster occupied the motor-theme quadrant. Two groups were positioned among the basic themes: metagenomics–machine learning–microbiome and diversity–bacterial–soil. The nitrogen–carbon–rhizosphere and random forest–metabolism–denitrification groups occurred in the niche-theme quadrant. In contrast, degradation–bioremediation–artificial intelligence was located in the quadrant conventionally labeled emerging or declining themes.

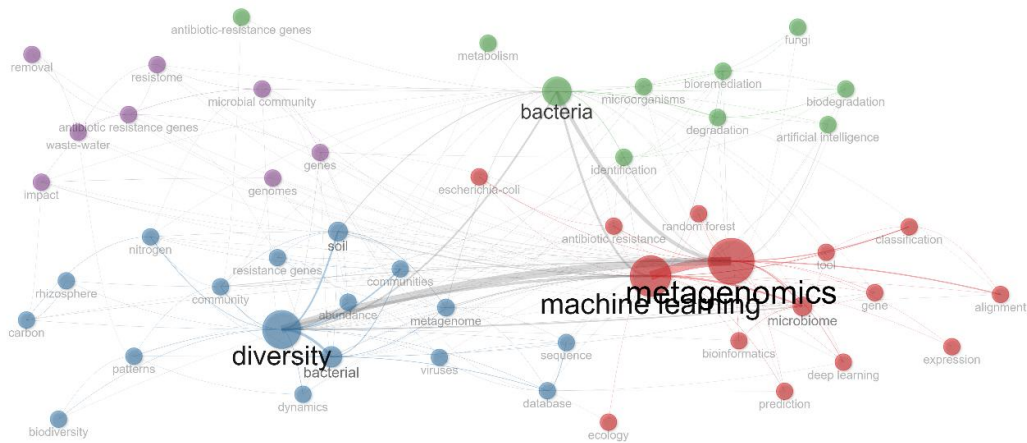


Figure 4. Author-keyword co-occurrence network showing the principal conceptual groupings in research on machine learning and environmental metagenomics.

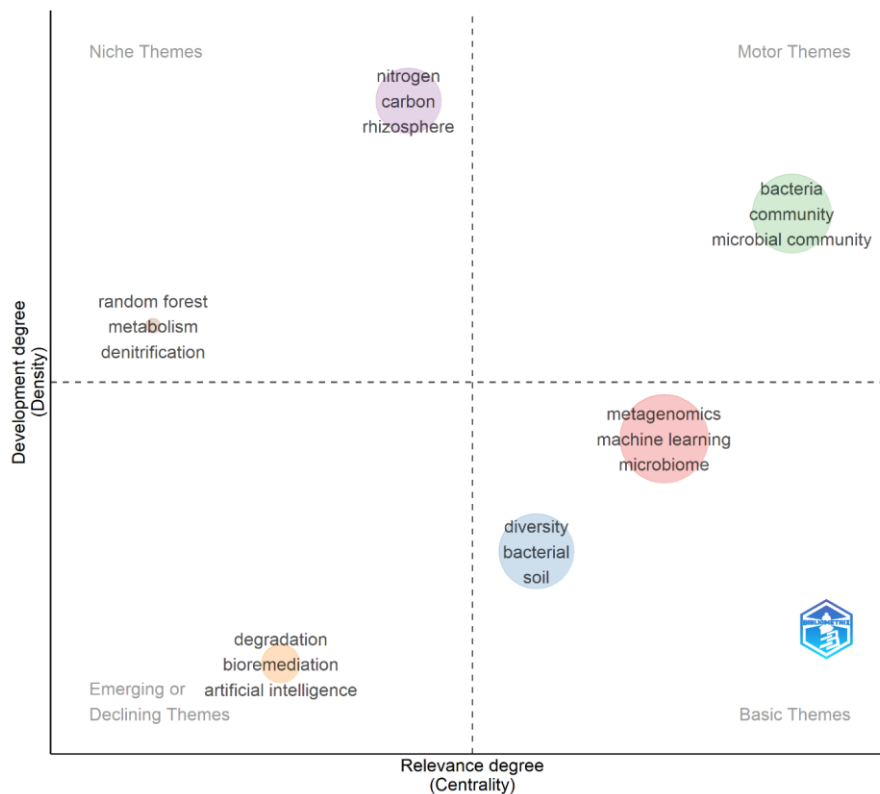


Figure 5. Thematic map of author keywords according to centrality and density.

3.5. Temporal Evolution of Research Topics

The Trend Topics analysis showed differences in the temporal positioning of frequently used terms (Figure 6). Alignment, resistome, and *Escherichia coli* were centered around 2022, whereas metagenomics, bacteria, and microbiome were positioned around 2023. Machine learning, diversity, and bacterial were centered around 2024 and were represented by comparatively larger

frequency markers than several of the more recent terms. The most recent group was positioned in 2025 and included artificial intelligence, prediction, and degradation. These terms had smaller frequency markers than the dominant terms such as machine learning and metagenomics, but their temporal placement distinguished them from earlier terminology.

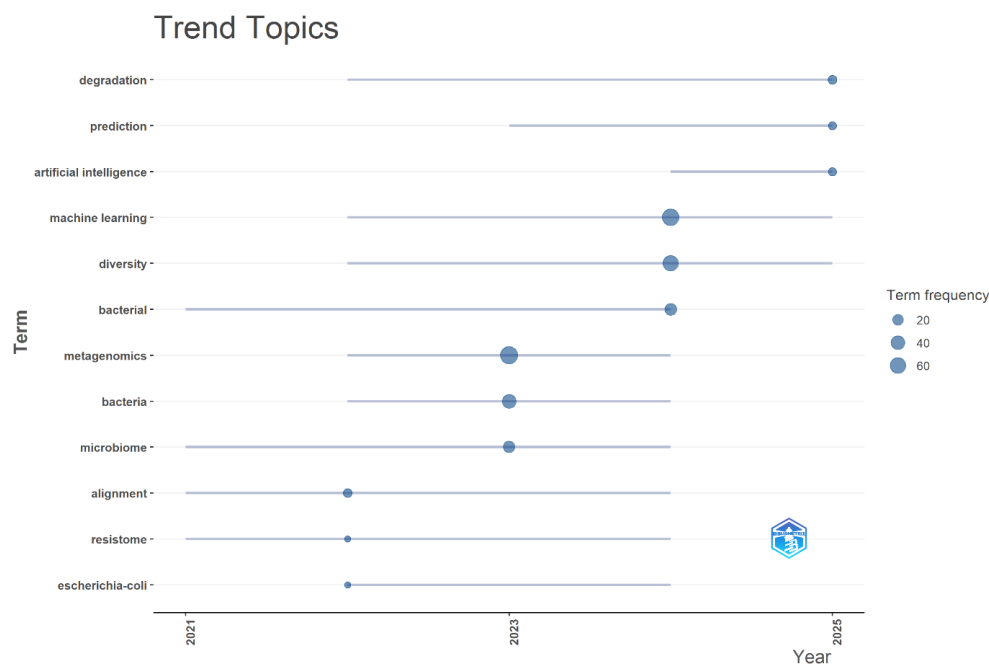


Figure 6. Trend Topics analysis showing the temporal positioning and relative frequency of selected author keywords. Bubble size represents term frequency; temporal recency indicates the location of recent use within the selected analysis settings and does not by itself indicate dominance or growth rate.

4. Discussion

4.1. Rapid Growth and Developmental Trajectory of Machine Learning in Environmental Metagenomics

The bibliometric results indicate a marked recent expansion of research at the intersection of machine learning and environmental metagenomics. Across 286 documents identified from 139 sources, the annual growth rate was 54.79%, and publication output reached 102 documents in 2025. The mean document age of 2.74 years indicates that much of the literature has accumulated relatively recently, while the average of 29.37 citations per document suggests that this literature has already generated a visible citation footprint. This publication trajectory is consistent with broader methodological developments in metagenomics. Mathieu et al. (2022) described analytical limitations associated with conventional alignment- and reference-dependent approaches in metagenomic analysis, including computational demands and incomplete representation of microbial diversity, providing an important context for complementary machine-learning and deep-learning strategies. Roy et al. (2024) likewise documented an expanding range of deep-learning applications across metagenomic tasks and increasing diversity in model

architectures and analytical purposes, although their review addressed metagenomics broadly rather than environmental applications alone.

The increasing availability of sequencing data and the analytical demands imposed by large, complex datasets may therefore help explain the growing interest in scalable computational approaches. Herazo-Álvarez et al. (2025), for example, documented increasing use of neural-network architectures for metagenomic binning, including applications involving environmental datasets and complex microbial communities. Such methodological diversification is consistent with the widening computational scope accompanying the publication growth observed here. The fitted life-cycle model ($R^2 = 0.995$) indicated that cumulative publication output in 2025 corresponded to approximately 10% of the modeled saturation level. This suggests that the field remains in an expansion phase, although the model-derived estimates should be interpreted cautiously.

4.2. Conceptual Foundations and Thematic Organization of the Field

Beyond publication growth, the conceptual structure of literature was dominated by terms linking metagenomic methodology with microbial ecology. Metagenomics (78

occurrences) and machine learning (71) were the two most frequent keywords, accompanied by diversity (57), bacteria (41), bacterial (27), microbiome (25), and terms associated with soil and microbial communities. Consistent with this pattern, metagenomics–machine learning–microbiome and diversity–bacterial–soil was positioned as basic themes, whereas bacteria–community–microbial community occupied the motor-theme quadrant. Together, these results indicate that computational analysis and community-level ecological characterization form closely connected components of the field.

The basic-theme position of metagenomics–machine learning–microbiome is consistent with the integration of machine learning into metagenomic analysis. Mathieu et al. (2022) described ML and DL applications in taxonomic classification and functional annotation, while Ghannam and Techtman (2021) showed how machine-learning approaches can relate high-dimensional microbial-community data to environmental conditions and predictive outcomes. Within the thematic-map framework, the basic-theme position therefore reflects broad connectivity to the field rather than limited relevance.

The diversity–bacterial–soil theme represents a similarly broad ecological foundation. Bahram et al. (2018) demonstrated relationships among bacterial diversity, community composition, functional gene repertoires, and environmental conditions in global topsoil microbiomes. Jansson and Hofmockel (2018) emphasized the diversity and spatial complexity of soil microbiomes and the role of metagenomics in characterizing microbial identities and functional potential. These findings provide an ecological context for the prominence of soil-associated bacterial diversity in the thematic structure identified here.

Within the same mapping framework, the motor-theme position of bacteria–community–microbial community reflects both strong connectivity with the wider network and comparatively high internal development. Grossart et al. (2020), for example, described how metagenomic information can link microbial community composition and interactions with ecological functions and biogeochemical processes in aquatic systems. By contrast, nitrogen–carbon–rhizosphere and random forest–metabolism–denitrification occupied niche-theme positions, indicating more specialized combinations of ecological processes, settings, and analytical approaches. Trend Topics further showed that artificial intelligence, prediction, and degradation were among the most recently positioned terms. Their recent temporal position, together with the placement of degradation–bioremediation–artificial intelligence in the thematic map, indicates increasing attention to this research direction, although it has not yet reached the prominence of the dominant themes. Overall, the thematic structure depicts a field organized around a broad metagenomic–computational foundation, a well-developed community–ecology axis, and several more specialized process- and

method-oriented streams.

4.3. Environmental Application Domains of Machine Learning in Metagenomic Research

This conceptual organization is further resolved by the keyword co-occurrence network, which indicates several differentiated but interconnected environmental application domains rather than a single homogeneous research stream. Four principal structures were evident: a machine learning–metagenomics–microbiome/bioinformatics domain; a diversity–soil–rhizosphere–carbon/nitrogen domain; bacteria–bioremediation–biodegradation–identification domain; and a resistome–antibiotic resistance genes–wastewater domain. These groupings suggest that computational approaches are associated with distinct environmental research contexts involving ecological prediction, contamination assessment, remediation, and environmental surveillance.

Within the soil-related domain, Gardiner et al. (2024) used shotgun metagenomic data to combine taxonomic and genetically encoded functional information in models of soil organic carbon, including graph-based approaches representing relationships between microbial taxa and their associated functions. This illustrates how environmental metagenomic information can contribute to predictive models of ecosystem properties. However, model importance or predictive association should not, by itself, be interpreted as evidence of causal microbial control over carbon dynamics.

Aquatic monitoring and source tracking provide another application. Zhu et al. (2024) applied machine learning to riverine nitrogen-source classification using environmental information that included shotgun-metagenomic taxonomic and functional profiles as well as 16S rRNA amplicon data. The metagenomic component demonstrates how functional information can complement community composition in source-attribution analyses. This is consistent with broader applications of microbial-community profiles to environmental prediction and monitoring described by Ghannam and Techtman (2021) and with the supervised-learning and de novo bioindicator framework discussed by Cordier et al. (2021). However, microbiome profiling, 16S rRNA amplicon sequencing, metabarcoding, and shotgun metagenomics represent distinct molecular and analytical approaches; examples from these categories are therefore discussed as related evidence rather than treated as methodologically equivalent.

The bioremediation–biodegradation domain similarly connects metagenomic characterization with predictive analysis. Wijaya et al. (2024) combined taxonomic and metabolic profiles from groundwater metagenomes with machine-learning and explainable-AI approaches to diagnose petroleum contamination and identify degradation-associated predictors. Moreover, Chettri et al. (2024) described the use of metagenomics to identify pollutant-degradation genes and pathways, trace contamination sources, and characterize microbial

responses relevant to bioremediation. Similarly, Böke Özkoç et al. (2025) highlighted the integration of AI and machine learning with metagenomic data in contaminated environments, including applications for predicting pollutant-degradation pathways and supporting bioremediation strategies in polluted soils, waters, and sediments. In the context of potentially toxic element (PTE)-contaminated aquatic environments, Farooq Tayeb Alsaadawi et al. (2026) further highlighted metagenomic functional profiling as a complementary approach for identifying microbial signatures related to metal resistance, detoxification, oxidative-stress response, and redox transformations, thereby extending sediment assessment beyond concentration-based measurements alone.

Finally, the resistome-antibiotic resistance genes-wastewater cluster identifies a distinct environmental-surveillance stream. Metagenomic analysis can support resistome characterization and contamination-source assessment, although the evidence considered here supports the specific integration of machine learning with wastewater shotgun metagenomics less directly than the soil, aquatic, and contaminated-groundwater applications. Detection of resistance genes also indicates genomic presence or potential rather than expression or ecological activity. Collectively, the co-occurrence structure therefore portrays machine learning as supporting multiple environmental metagenomic tasks rather than a single uniform analytical application.

5. Conclusion

The literature linking machine learning with environmental metagenomics expanded rapidly between 2016 and 2025, indicating a comparatively young research area. The dataset comprised 286 documents and showed an annual growth rate of 54.79%, with publication output reaching 102 documents in 2025. This recent acceleration indicates that the integration of computational learning approaches with environmental metagenomic research has developed into a broader and increasingly visible body of literature.

The conceptual structure of the field is organized around metagenomics and machine learning as broad conceptual foundations, with microbial diversity, bacterial communities, soil, and related ecological processes forming major thematic components. The co-occurrence network further indicated differentiated but interconnected research domains involving a machine learning-metagenomics-microbiome/bioinformatics core, soil-rhizosphere and nutrient-related research, bioremediation and biodegradation, and resistome and wastewater surveillance. Trend Topics indicated increasing visibility of machine learning, artificial intelligence, prediction, and degradation-related terminology, although the *degradation-bioremediation-artificial intelligence* direction is more appropriately interpreted as recently gaining visibility than as a definitively established emerging field. Methodologically,

predictive and explainable approaches are broadening the analytical repertoire of environmental metagenomic research.

The field is internationally connected but geographically uneven, with scientific production concentrated particularly in China and the United States. These geographical comparisons are descriptive and reflect publication and citation counts within the Web of Science-indexed dataset; they should not be interpreted as direct measures of research quality because they are influenced by database coverage, publication volume, collaboration patterns, and publication age. Future research may benefit from broader geographic participation, greater integration across diverse environmental systems, and rigorous validation of predictive AI/ML outputs against experimentally or independently measured environmental and biological processes.

Author Contributions

The percentages of the author' contributions are presented below. The author reviewed and approved the final version of the manuscript.

	S.T.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C= concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declares that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

- Abdelwahab, S. I., Taha, M. M. E., Jerah, A. A., Farasani, A., Abdullah, S. M., Aljahdali, I. A., Ibrahim, R., Oraibi, O., Oraibi, B., Alfaifi, H. A., Alzahrani, A. H., & Babiker, Y. O. H. (2024). Artificial intelligence and microbiome research: Evolution of hotspots, research trends, and thematic-based narrative review. *Cellular and Molecular Biology*, 70(10), 182-192. <https://doi.org/10.14715/cmb/2024.70.10.24>
- Aria, M. & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis, *Journal of Informetrics*, 11(4), pp 959-975. <https://doi.org/10.1016/j.joi.2017.08.007>
- Aria, M., & Cuccurullo, C. (2026). *Science Mapping Analysis - A*

- primer with Biblioshiny. McGraw-Hill, ISBN: 978-88-386-2297-7.
- Aria, M., Cuccurullo, C., D'Aniello, L., & Spano, M. (2026). Biblioshiny and the SAAS Workflow: An integrated framework for transparent and reproducible science mapping. A demonstration through the replication of a study. *Journal of Informetrics*, 20(3), 101837 <https://doi.org/10.1016/j.joi.2026.101837>
- Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Pölme, S., Sunagawa, S., Ryberg, M., ... Bork, P. (2018). *Structure and function of the global topsoil microbiome*. *Nature*, 560(7717), 233–237. <https://doi.org/10.1038/s41586-018-0386-6>
- Böke Özkoç, H., Tırink, S., & Farooq Tayeb Alsaadawi, S. (2025). Metagenomic approaches to monitoring and management of potential toxic element pollution. *UBAK Publishing House*. <https://doi.org/10.5281/zenodo.17939033>
- Chettri, D., Verma, A. K., Chirania, M., & Verma, A. K. (2024). Metagenomic approaches in bioremediation of environmental pollutants. *Environmental Pollution*, 363, 125297. <https://doi.org/10.1016/j.envpol.2024.125297>
- Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A. A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, 30(13), 2937–2958. <https://doi.org/10.1111/mec.15472>
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- Farooq Tayeb Alsaadawi, S., Tırink, S., & Böke Özkoç, H. (2026). Integrating metagenomic functional profiling to evaluate public health risks derived from PTE-contaminated river sediments. In *International Congress on Natural Sciences and Applied Mathematics* (pp. 259–275). <https://doi.org/10.5281/zenodo.19459760>
- Gardiner, L., Marshall, M., Reusch, K., Dearden, C., Birmingham, M., Carrieri, A. P., Pyzer-Knapp, E. O., Krishna, R., Neal, Andy (2024) DGCNN approach links metagenome-derived taxon and functional information providing insight into global soil organic carbon. *NPJ Biofilms and Microbiomes*, 10. p. 113. [10.1038/s41522-024-00583-9](https://doi.org/10.1038/s41522-024-00583-9)
- Ghannam, R. B., & Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, 19, 1092–1107. <https://doi.org/10.1016/j.csbj.2021.01.028>
- Grossart, H. P., Massana, R., McMahon, K. D., & Walsh, D. A. (2020). Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography*, 65(S1), S2–S20. <https://doi.org/10.1002/lno.11382>
- Herazo-Álvarez, J., Mora, M., Cuadros-Orellana, S., Vilches-Ponce, K., & Hernández-García, R. (2025). A review of neural networks for metagenomic binning. *Briefings in Bioinformatics*, 26(2), bbaf065. <https://doi.org/10.1093/bib/bbaf065>
- Jansson, J. K., & Hofmockel, K. S. (2018). The soil microbiome—from metagenomics to metaphenomics. *Current Opinion in Microbiology*, 43, 162–168. <https://doi.org/10.1016/j.mib.2018.01.013>
- Liu, S., Rodriguez, J. S., Munteanu, V., Ronkowski, C., Sharma, N. K., Alser, M., ... & Mangul, S. (2025). Analysis of metagenomic data. *Nature Reviews Methods Primers*, 5, 5. <https://doi.org/10.1038/s43586-024-00376-6>
- Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovic, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., ... Truu, J. (2021). Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology*, 12, 634511. <https://doi.org/10.3389/fmicb.2021.634511>
- Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., & Droit, A. (2022). Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. *Frontiers in Microbiology*, 13, 811495. <https://doi.org/10.3389/fmicb.2022.811495>
- Mirete, S., Sánchez-Costa, M., González-de-Figueras, C., Díaz-Rullo, J., Martínez-Rodríguez, P., & González-Pastor, J. E. (2025). Metagenome-assembled genomes (MAGs): Advances, challenges, and ecological insights. *Microorganisms*, 13(5), 985. <https://doi.org/10.3390/microorganisms13050985>
- Pallavi, H., & Uday Bhaskar, N. (2025). Computational and machine learning frameworks for microbial data analysis: A systematic review. *International Journal of Advanced Research in Computer Science*, 16(6), 60–69. <https://doi.org/10.26483/ijarcs.v16i6.7377>
- Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahim, E., Eckenberger, J., Novelli, P., Tonda, A., Simeon, A., Shigdel, R., Béreux, S., Vitali, G., Tangaro, S., Lahti, L., Temko, A., Claesson, M. J., & Berland, M. (2023). Machine learning approaches in microbiome research: challenges and best practices. *Frontiers in Microbiology*, 14, 1261889. <https://doi.org/10.3389/fmicb.2023.1261889>
- R Core Team (2026). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL: <https://www.R-project.org/>.
- Roy, G., Prifti, E., Belda, E., & Zucker, J.-D. (2024). Deep learning methods in metagenomics: A review. *Microbial Genomics*, 10(4), 001231. <https://doi.org/10.1099/mgen.0.001231>
- Sharma, S., Narahari, H. P., & Raman, K. (2025). Harnessing machine learning for metagenomic data analysis: Trends and applications. *mSystems*, 10(11), e01642-24. <https://doi.org/10.1128/msystems.01642-24>
- Tonkovic, P., Kalajdziski, S., Zdravevski, E., Lameski, P., Corizzo, R., Pires, I. M., Garcia, N. M., Loncar-Turukalo, T., & Trajkovic, V. (2020). Literature on applied machine learning in metagenomic classification: A scoping review. *Biology*, 9(12), 453. <https://doi.org/10.3390/biology9120453>
- Velasco Cardona, D. C., Cardona-Acevedo, S., Valencia-Arias, A., Pérez-Delgado, O., & Pinella Vega, M. (2025). Bibliometric analysis of the application of artificial intelligence techniques in bacteriology: A PRISMA-guided research agenda. *Frontiers in Microbiology*, 16, 1641967. <https://doi.org/10.3389/fmicb.2025.1641967>
- Wassan, J. T., Wang, H., Browne, F., & Zheng, H. (2019). A comprehensive study on predicting functional role of metagenomes using machine learning methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3), 751–763. <https://doi.org/10.1109/TCBB.2018.2858808>
- Wijaya, J., Park, J., Yang, Y., Siddiqui, S. I., & Oh, S. (2024). A metagenome-derived artificial intelligence modeling framework advances the predictive diagnosis and

- interpretation of petroleum-polluted groundwater. *Journal of Hazardous Materials*, 472, 134513. <https://doi.org/10.1016/j.jhazmat.2024.134513>
- Xu, D., Liu, B., Wang, J., & Zhang, Z. (2022). Bibliometric analysis of artificial intelligence for biotechnology and applied microbiology: Exploring research hotspots and frontiers. *Frontiers in Bioengineering and Biotechnology*, 10, 998298. <https://doi.org/10.3389/fbioe.2022.998298>
- Zhang, L., Chen, F., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y., Hao, H., Yi, W., Li, M., & Xie, Y. (2021). Advances in metagenomics and its application in environmental microorganisms. *Frontiers in Microbiology*, 12, 766364. <https://doi.org/10.3389/fmicb.2021.766364>
- Zhu, Z., Ding, J., Du, R., Zhang, Z., Guo, J., Li, X., Jiang, L., Chen, G., Bu, Q., Tang, N., Lu, L., Gao, X., Li, W., Li, S., Zeng, G., & Liang, J. (2024). Systematic tracking of nitrogen sources in complex river catchments: Machine learning approach based on microbial metagenomics. *Water Research*, 253, 121255. <https://doi.org/10.1016/j.watres.2024.121255>