ISSN: 2458-8989

Natural and Engineering Sciences

NESciences, 2026, 11 (2): 264-287 doi: 10.28978/nesciences.262024

Exploring the Links Between Artificial Intelligence and its Applications in the Fields of Biosynthesis

Mohammed.Ahb.AL-krdoshi[1], Hamza. A. Saadallah[2], Ali Q Saeed[3], Shadan. S. Ismail[4], Osama. Y. Mohammed[5], Thakir. A. Abboosh[6], Afaf. N. Yousif[7], Eesha. I. Mohammed[8]

[1]Department Natural Products Research, Technical Center for Research, Northern Technical University, Mosul, Iraq. E-mail: mohammed.ahb.student@ntu.edu.iq, ORCID: https://orcid.org/0009-0005-4884-9032

[2]Department Natural Products Research, Technical Center for Research, Northern Technical University, Mosul, Iraq. E-mail: hamza.atheer.saadallah@ntu.edu.iq, ORCID: https://orcid.org/0009-0005-6258-9361

[3]Artificial Intelligence Techniques Engineering Department, Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. E-mail: ali.qasim@ntu.edu.iq, ORCID: https://orcid.org/ 0000-0002-2276-3776

[4]Artificial Intelligence Techniques Engineering Department, Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. E-mail: Shadn.sabah@ntu.edu.iq, ORCID: https://orcid.org/0009-0002-4720-4097

[5]Artificial Intelligence Techniques Engineering Department, Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. E-mail: osama.yassin@ntu.edu.iq, ORCID: https://orcid.org/0000-0002-5278-2468

[6]Artificial Intelligence Techniques Engineering Department, Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. E-mail: thaker.a.a1975@ntu.edu.iq, ORCID: https://orcid.org/0009-0004-7052-4069

[7]Cyber Security Techniques Engineering Department, Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. E-mail: afaf.nasser@ntu.edu.iq, ORCID: https://orcid.org/0009-0009-9241-756X

[8]Computer Techniques Engineering Department, Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. E-mail: aysha.ibrahim@ntu.edu.iq, ORCID: https://orcid.org/0009-0001-9964-7060

Abstract

Objective: To study the key areas and emerging trends in AI research in biosynthesis.
Methods The Web of Science Core Collection's Science Citation Index Expanded was the source of the research literature on AI in biosynthesis. To learn more about publishing years, countries/regions, journals, institutions, citations, and keywords, this data was studied. The online analytic platform's

Corresponding Author: Mohammed.Ahb.AL-krdoshi, E-mail: mohammed.ahb.student@ntu.edu.iq

CiteSpace tools and VOSviewer library were used to create co-occurrence network graphs.
Results 44 pertinent research publications from various nations were chosen in total. Every year since 1998, the quantity of pertinent articles has grown dramatically. With 96.76% of the total, the three nations with the most publications were the USA (n = 25.7%), the UK (n = 8.58%), and China (n = 11.43%). The two journals with the most articles were IEEE Access (n = 97, 18.44%) and Multimedia Tools and Applications (n = 61, 11.59%). The three universities with the most publications were Capital Medical University (n = 7.21%), Nankai University (n
= 11.26%), and Chinese Academy of Sciences (n = 7.66%). Through keyword analysis, we discovered that the primary subjects of the current study are biosynthesis and biomanufacturing.
Conclusions The study found important areas of attention and new developments in AI for biomanufacturing diagnostics and its uses, indicating that AI will become more and more important in this sector going forward.

Keywords:
Artificial intelligence, biosynthesis, bibliometric analysis, Citespace, VOSviewer.

Available online: 02/06/2026

1. Introduction

The complex biological processes by which living things produce natural substances, such as necessary medications and bioactive chemicals, are referred to as biosynthesis. The use of machine learning to this field has become a game-changing development, greatly improving the accuracy and effectiveness of biosynthetic route design and prediction. This multidisciplinary collaboration optimizes current biosynthetic processes while hastening the discovery of new natural compounds. Important facets of this connection are explored in the next sections, which emphasize the significance of sophisticated algorithms in template-free route design, which improves catalytic efficiency and enzyme function prediction (Xie et al., 2024). Bio Navi-NP, a widely utilized tool, employs deep learning to identify biosynthetic pathways for 90.2% of chemicals (Zheng et al., 2022). Machine learning was employed to evaluate omics data, revealing plant biosynthesis pathways (Durand et al., 2024).
Biosynthesis has greatly benefited from machine learning; nonetheless, challenges persist. The inherent complexity of biological systems and the necessity for high-quality, complete information can restrict model generalization and predictive accuracy. The efficiency of biosynthesis, molecular diversity, and target specificity objectives influence the synthetic processes for bioactive chemicals. These objectives dictate the efficacy and relevance of established pathways, influencing metabolic engineering methodologies and instruments (Lv and Wang, 2024). A deeper comprehension of intricate natural product biosynthetic mechanisms facilitates the identification of essential enzymes that may be reused for synthetic applications (Jamieson et al., 2021). The incorporation of specific substrates into specialized pathways facilitates the manufacture of both natural and synthetic compounds, enhancing the versatility and applicability of synthetic biosynthetic systems (Kufs et al., 2020).
Biosynthesis encounters numerous problems, such as ineffective retrosynthesis prediction, intricate metabolic pathway design, and challenging management of organic electrosynthesis selectivity. AI provides distinctive approaches that enhance biosynthesis accuracy and efficiency to address these challenges.
Conventional methods necessitate expert knowledge, which can be expensive and result in suboptimal synthesis (Kufs et al., 2020). AI-driven retrosynthesis employs extensive datasets to enhance the prediction of reaction pathways, diminishing reliance on human expertise. Numerous metabolic pathways are inadequately comprehended, complicating bioproduct synthesis. AI methodologies can generate innovative biosynthetic pathways from extensive information, enhancing the efficacy and optimization of metabolic pathways (Jiang et al., 2023).
Combining linguistics, statistics, and mathematics, bibliometrics is an interdisciplinary topic that makes it easier to quantitatively assess and analyze research patterns. VOSviewer and CiteSpace (Yang et al., 2023) are two frequently used tools for graphical representation and data visualization in bibliometric investigations (Zhong et al., 2023). Tao et al. Recognized for its simplicity, dependability, and effectiveness, bibliometric analysis has been widely used in the field of biosynthesis (Baumgart, Beck & Ghezel-Ahmadi, 2024; Donnelly, 2023). To investigate current developments and new study

fields in biosynthetic imaging, for example, Zhang et al. (2024) have employed bibliometric approaches. The specific areas of study and potential paths for bibliometrics' application to biosynthesis are still little understood, though. An innovative attempt to use bibliometric analysis to pinpoint important areas of emphasis and new trends is this research.

2. Materials and Method

2.1. Data source and research process

On October 25, 2024, information was retrieved from the Web of Science Core Collection's Science Citation Index Expanded Citation Index. (((((All= (Deep Learning)) and All=(Biosynthesis)) or All= (Deep Learning)) or All=(Biofabrication)) or All=(CNN)) or All=(GAN)) or All= (Generative Adversarial Networks)) and All=(Biofabrication))is the search formula. English was chosen as the search language, and computer science research articles released during the previous five years were chosen. All studies that satisfied the aforementioned requirements underwent separate screening by two researchers. Retracted articles, data papers, books, and early access were not included. Carefully review each literature's abstract and title to weed out the remaining pieces. The following are the manual exclusion criteria: 1)
There is no association between artificial intelligence and biosynthesis; 2) Artificial intelligence and biomanufacturing were the main research objectives. In case of conflict, the decision was discussed in a group meeting. The research and analysis process are depicted in Figure 1.

2.2. Bibliometric analysis

To conduct a thorough bibliometric analysis of the literature covering many facets of the research issue, this study makes use of three analytical techniques. Initially, cooperative network linkages between nations and areas were shown using the Library Online Analysis Platform (https://bibliometric.com/). The second tool used was VOSviewer 1.6.19.0 (Leiden University, Leiden, Netherlands), a program for document-based data analysis with essential features including visualization and "co-occurrence clustering." In this study, the co-occurrence network of terms across all papers was constructed and analyzed using VOSviewer. With keywords shown as circles whose widths matched their frequency and co-occurrence strength, the resulting visualizations employed various hues to indicate discrete clusters. Lastly, citation data was assessed using the Publish or Perish tool, which quantified the influence of publications within the field. Finding the top 20 most popular terms and assessing their corresponding temporal patterns and intensity were the specific goals of the investigation.

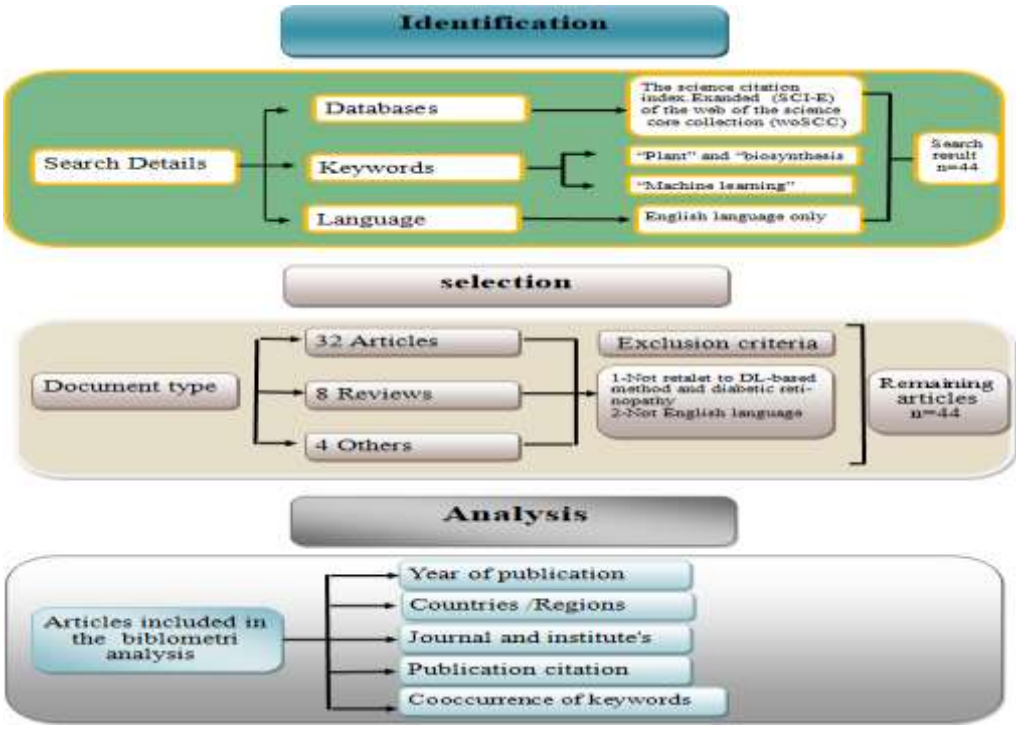

Figure (1): Flowchart of document collection and selection process with Research Framework

3. Results and Discussion

3.1. Publications per year

The growth strategy for 2011–2024 is shown in the bar chart. Growth was modest and steady from 2011 and 2019, with little upticks in 2018 and 2019. 2020 marks the start of a notable increase, which picks up speed in 2021 and peaks in 2023. 2024 is still greater than prior years, despite a little drop from the 2023 peak.

The plan proposes a strategy centered on gradual advancement, followed by rapid growth beginning in 2020. The 2023 peak marks a high point in the project, and this increasing tendency is a result of greater investment, innovation, or market demand. After fast development, the minor decline in 2024 can be a sign of stabilization or recalibration.

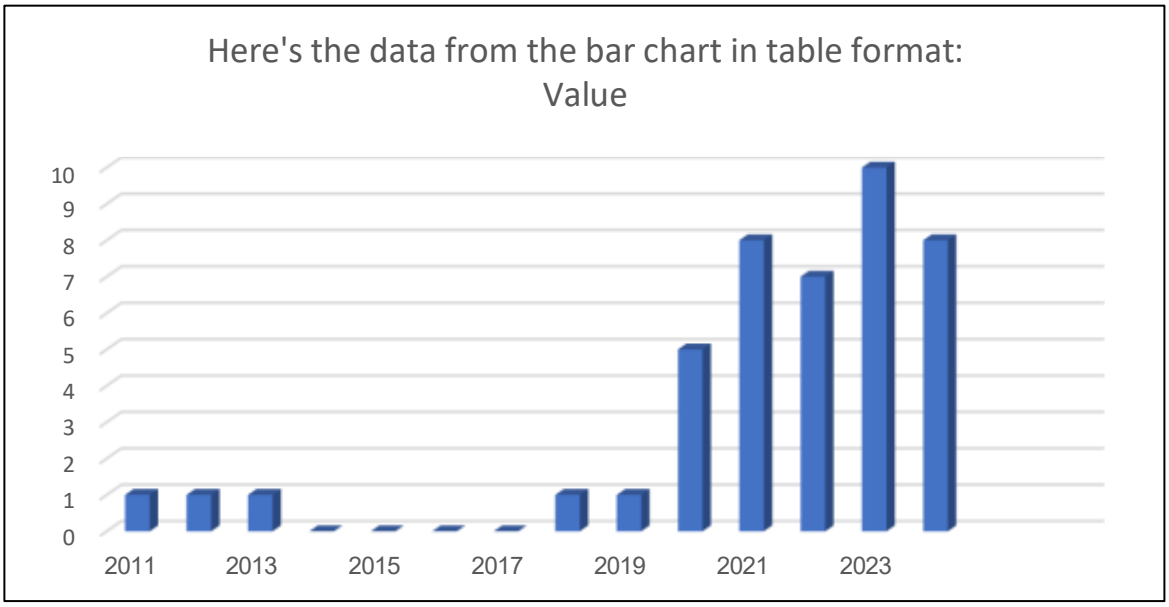

Figure (2) shows Literature Publications

3.2. Countries/Regions Publications and Collaborations

A bar graph showing the frequency of different keywords from 2011 to 2024 is displayed in the supplied graphic. A particular term is represented by each color on the graph, and the height of the bars indicates how frequently that keyword occurs in a given year. An overview of term usage trends throughout time is provided by this graphic depiction.

The graph's general trend shows a consistent rise in these terms' usage over the course of the analysis. This increasing trend is especially noticeable starting in 2020, when the bars become noticeably longer, indicating a large increase in keyword frequency.

Different patterns emerge when specific terms are examined more closely. terms such as genome. Consistent use of keywords such as Genome, Annotation, and Escherichia-Coli throughout time suggests that these topics continue to be of interest to researchers. Genes, pathways, and synthetic biology continue to be significant, indicating ongoing study in biological and genetic engineering. Nonetheless, the frequency of Stress and Insights climbed in 2020, indicating an intensified emphasis on stress responses and novel research.

The prominence of Arabidopsis thaliana and gene expression is increasing, indicating a growing significance in plant biology and gene expression research. Ongoing research in metabolism and protein frequently utilizes the terms "Metabolism" and "Protein." Beginning in 2020, Gene Expression and Secondary Metabolites see substantial growth, indicating heightened interest. Ultimately, "prediction," Evolution, Identification, and Biosynthesis exhibit moderate expansion, signifying an increasing interest in predictive modeling, evolutionary research, and innovative chemical identification and biosynthesis. The increased frequency of keywords may suggest an expanding scope within a scientific area such as biology or bioinformatics. The increasing emphasis on stress responses, gene expression, and secondary metabolites indicates a shift in research goals; nonetheless, the persistent utilization of genomic and genetic terminology underscores their significance. Nonetheless, the graph possesses limitations. Due to the lack of specified contexts for the terminology employed,

further investigation is required to comprehend the trends and their implications. Examinations of these concepts and external factors including as funding, publication frequency, and patents may provide further insights. Future study may concentrate on stress responses, secondary metabolites, and gene expression to address the increasing interest in the topic.
This bar graph illustrates the evolution of research concentration within a scientific discipline. Recognizing these trends enables researchers to identify emerging subjects, allocate resources effectively, and adjust their research in accordance with the evolving direction of the discipline.

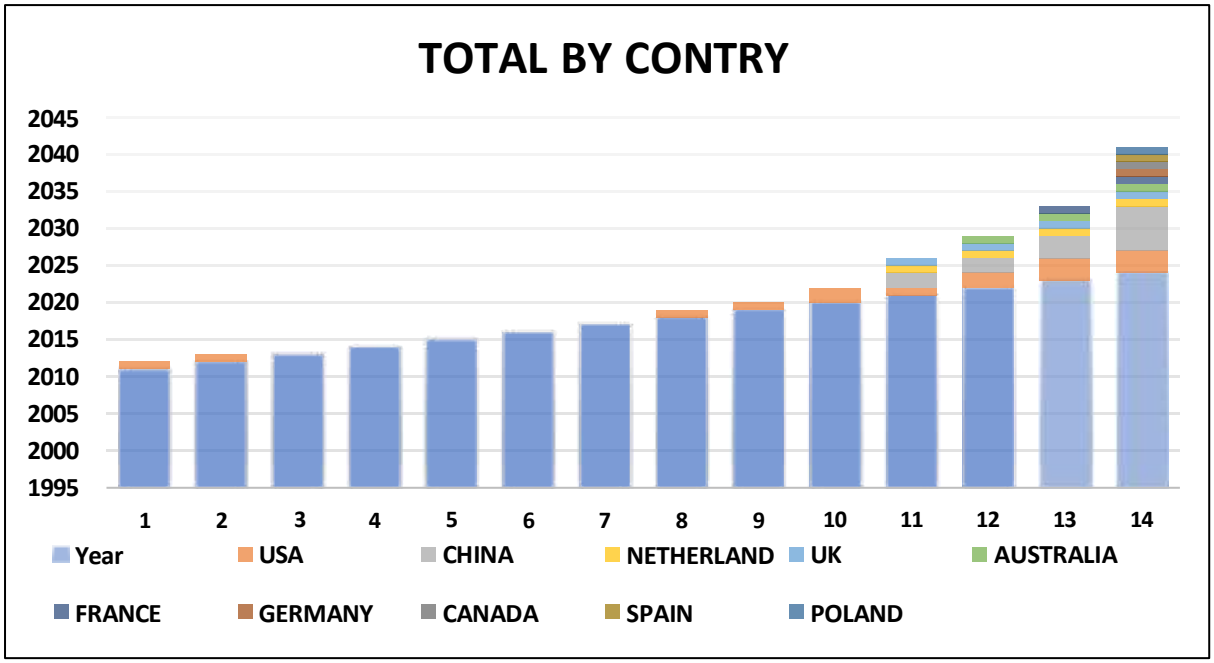

Figure (3) shows publication per county

A chord diagram, a visualization method for showing relationships between items, is shown in the accompanying image. In this instance, the chord diagram seems to show the relationships between several nations.

Interpretation:
The chord diagram uses linking chords and arcs to graphically depict the links between nations. Along the border of the circle, each nation is represented by a colored arc, the thickness of which may indicate the significance of the nation or any other numerical indicator. Chords linking the arcs represent the frequency or intensity of international relations, potentially influencing chord width. This approach offers a rapid overview of global dynamics and elucidates intricate international connections. The importance of the connections in the figure is contingent upon the data utilized; hence, further context is required for comprehension. In the absence of facts, the image possesses various interpretations.
The significance of the chord diagram is contingent upon the relationships it illustrates. Trade contacts utilize chords whose width signifies worldwide trade. Thicker cables underscore economic linkages and reinforce commercial connections. Wider chords signify larger migration flows between nations, therefore potentially reflecting migration patterns. It may also denote technical collaboration, with larger cables symbolizing more profound research connections. Cultural exchanges, such as travel or student programs, may be represented with stronger chords to signify their significance. The adaptability of the chord diagram allows for its application in various global interactions, albeit its interpretation is contingent upon the dataset. Despite its utility, the chord diagram possesses limitations. Its appearance is captivating; nonetheless, it is devoid of quantitative data, hindering precise examination. While it illustrates overarching tendencies, the image fails to elucidate the reasons for the linkages. Additional context and statistics are required for more profound understanding. Modifying layouts, labels, and color palettes to enhance readability may further augment the visualization. Commentaries or stories may elucidate the relationships. Future designs may mitigate these issues to enhance the diagram's informational utility and public accessibility.

The chord diagram elucidates intricate global connections. It underscores trends and discrepancies in international relations and their associated patterns. It achieves its maximum potential alone through quantitative and contextual data. When utilized in isolation, it offers a broad overview; nevertheless, additional details can enhance its quality. The chord diagram could evolve into a more effective instrument for comprehending the complex interconnections of nations across many industries through enhanced development and sophistication.

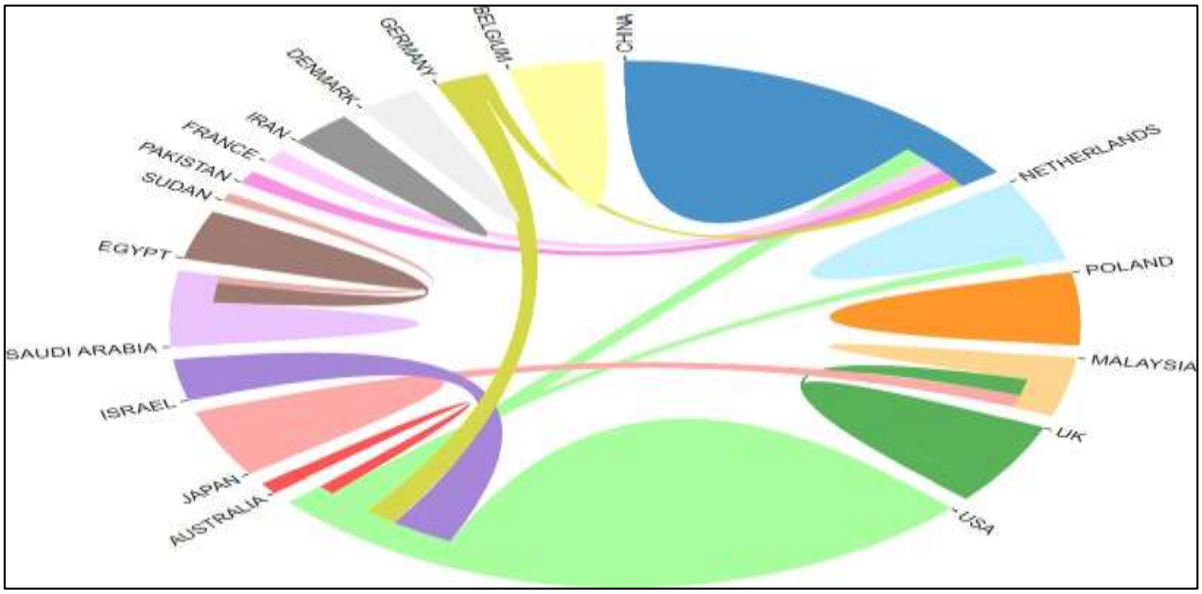

Figure (4) shows Relation between countries

3.3. Publications per Journals, Publishers and Institutions

The table contrasts scientific journals based on publication frequency, quartile classification, impact factor, and geographical origin. The text examines article dissemination and the impact of journals on academics and researchers. The table reveals that journals in the "Q1" quartile, signifying academic excellence and prestige, are prevalent. Many of the journals included, such as Plant Methods, Plant Cell Physiology, and Natural Product Reports, possess significant scientific impact. Only Metabolites and BMC Genomics occupy the "Q2" quartile, indicating a lower ranking yet a significant impact.

The impact factor of each article, a crucial measure of journal influence, varies. The greatest impact factor was 10.2 for Natural Product Reports, followed by Plant Cell at 10.0 and a journal from the United States at 9.4. These ideals indicate their commitment to study and exemplary scholarship. Nonetheless, journals with lesser impact factors—3.3 and 3.4, respectively—such as Phytobiomes Journal and Metabolites retain their significance.

The periodicals are geographically distributed throughout several nations, with the United States and England leading the list. Notable publications such as Natural Product Reports, Plant Methods, and BMC Genomics are representative of England. Conversely, notable American publications such as Plant Cell and PLOS Computational Biology are featured. Switzerland and Japan are included among the nations represented, underscoring the global reach of scientific publication and research.

. A few of journals have several entries for publication frequency; BMC Genomics, Plant Methods, and Plant Cell Physiology each contributed three articles. This demonstrates their proactive involvement in sharing scientific discoveries. On the other hand, journals with a single publication, such Phytobiomes Journal, Antioxidants, and Metabolites, suggest a more focused or niche emphasis. All things considered, the chart shows a wide range of high-impact journals that make substantial contributions to the advancement of science. Additionally, it emphasizes the value of regional variety and the range of effect these publications have in their respective professions.

Table (1) showing Journals with Maximum Number of Publications

Journal Name	Publication	Quartile category	Impact factor	Country
BMC GENOMICS	3	Q2	3.5	ENGLAND
PLANT METHODS	3	Q1	4.7	ENGLAND

PLANT AND CELL PHYSIOLOGY	3	Q1	4.7	JAPAN
NATURAL PRODUCT REPORTS	2	Q1	10.2	ENGLAND
PLANT CELL	2	Q1	10	UNITED STATES
UNITED STATES	1	Q1	9.4	UNITED STATES
PLOS COMPUTATIONAL BIOLOGY	1	Q1	3.8	UNITED STATES
ANTIOXIDANTS	1	Q1	6	SWITZERLAND
PHYTOBIOMES JOURNAL	1	Q1	3.3	UNITED STATES
METABOLITES	1	Q2	3.4	SWITZERLAND

Wageningen University and Research and the University of Wisconsin lead among the organizations listed in terms of total publications, each contributing five publications, one of which is a first-author publication. Both institutions are located in nations with significant research investments—Russia for Wageningen University and New Zealand for the University of Wisconsin—highlighting the global nature of their collaborative academic endeavors. The table illustrates the quantity of first-author publications produced by various organizations across different countries. This table serves to evaluate research production and academic publishing.

Subsequently, the University of Florida, Chinese Academy of Agricultural Sciences, and University of California, Davis each submitted four papers. The University of California, Davis is distinguished by its two first-author publications, the highest number among the listed institutions, and its contributions to its academic disciplines.

The data indicates that universities are situated in North America (UC Davis; UF; MN), Asia (Chinese Academy of Agricultural Sciences in China and University of Science and Technology Malaysia), and Europe. A multitude of universities contributes to this dataset, ensuring robust representation of the US. The substantial number of U.S. institutions demonstrates the nation's robust academic and research framework. Three publications and one first-author article from institutions such as the University of Science and Technology Malaysia illustrate Malaysia's expanding academic influence in global research. The Netherlands fosters significant intellectual advancements, as evidenced by Wageningen University, despite its lesser scale compared to the US and China.

The data indicates that academic publication has proliferated worldwide, with contributions from both established and nascent research domains. The prevalence of first-author publications, particularly from institutions such as UC Davis, highlights individual contributions to research fields. The geographical diversity of contributing institutions illustrates the diverse and collaborative nature of contemporary academic research, which addresses global issues beyond borders.

This table analysis provides a lens through which to view academic contributions in a globally interconnected research ecosystem by illuminating the interactions among regional diversity, institutional reputation, and research authorship dynamics.

Table (2) Top10 institutes with maximum number of publications

n	Organization Name	Publication	Total number of first-author Publication	Country
1	Wageningen Univ & Res	5	1	Russia
2	Univ Wisconsin	5	1	New Zealand
3	Chinese Acad Agr Sci	4	1	China
4	Univ Calif Davis	4	2	UNITED STATES
5	University of Florida	4	1	UNITED STATES
6	Univ Nottingham	4	1	UNITED STATES
7	University of Oklahoma	4	1	UNITED STATES
8	Wageningen Univ	3	1	Netherlands

9	University of Science and Technology Malaysia	3	1	Malaysia
10	University of Minnesota	3	1	UNITED STATES

The table presented provides a comprehensive overview of various academic publications, summarizing key metrics related to citation performance, authorship, and publication details. It encompasses ten distinct publications, each identified by a unique index number. The table is structured to convey detailed information across multiple columns, including the authors, title of the paper, year of publication, source (journal or conference), publisher, citation performance (expressed as citations per paper, per year, and per author), the number of authors contributing to the paper, and the age of the publication in years.

A critical aspect highlighted in the table is the citation performance, which serves as an indicator of the scholarly impact of each publication. The citation data is broken down into three categories: total citations per paper, average citations per year, and citations attributed to each author. For example, the highest-performing paper in terms of total citations (76) was publishe
in 2011 under the journal *Plant Cell* by the American Society of Plant Biologists, reflecting its long-standing influence over a span of 13 years. Conversely, more recent publications, such as the 2023 paper in *Biotechnology Advances*, show a relatively lower citation count of 49, which is expected given the shorter period since publication.

The table emphasizes collaboration by enumerating the authors of each paper. Collaborative papers, such as the 2019 Communications Biology article, feature 12 authors, signifying a multidisciplinary approach. In contrast, the 2020 Journal of Biological Chemistry article features fewer authors (7), indicating a more compact research team made a greater contribution.The publication age of any document signifies its longevity and utility. Earlier publications, such as the 2013 Plant and Cell Physiology article, exhibit sustained citation activity, whereas more recent works indicate emerging research themes. The temporal dimension facilitates the study of both immediate and long-term academic impacts.

The table provides a detailed analysis of academic publications, facilitating the evaluation of scholarly contributions, collaborative dynamics, and research impact in plant biology and associated disciplines.

Table (3) Top 10 most cited articles in the field of deep learning and retinopathy

Authors	**Year**	**Source**	**Publisher**	**Citation within WOS only Per paper per year per author**	**Author count in the paper**	**Age**
(Durand et al., 2024)	2024	PLANT CELL	AMER SOC PLANT BIOLOGISTS	76 15 38	5	13
(Jiang et al., 2023)	2023	BIOTECHNOLOGY ADVANCES	PERGAMON-ELSEVIER SCIENCE LTD	49 12. 5	4	1
(Jamieson et al., 2021)	2021	NATURAL PRODUCT REPORTS	ROYAL SOC CHEMISTRY	48 12 4	4	3
(Kufs et al., 2020)	2020	COMMUNICATIONS BIOLOGY	NATURE PORTFOLIO	44 4 12	12	5
(Lv and Wang, 2024)	2024	ANTIOXIDANTS	MDPI	43 11 4	4	4
(Xie et al., 2024)	2024	BIOMOLECULES	MDPI	31 8 4	4	3
(Yang et al., 2023)	2023	JOURNAL OF BIOLOGIC AL CHEMISTRY	ELSEVIER	25 4 7	7	4

(Zheng et al., 2022)	2022	PLANT METHODS	BIOMED CENTRAL LTD	25 6 4	4	6
(Zhong et al., 2023)	2023	PLANT AND CELL PHYSIOLO GY	OXFORD UNIV PRESS	23 2 13	13	11
(Bai et al., 2024)	2012	PLANT CELL	AMER SOC PLANT BIOLOGISTS	21 4 5	5	12

3.4. Co-occurrence of keywords

The illustration depicts keyword frequency from 2011 to 2024 in the form of a bar graph. Each hue in the graph signifies a keyword, while the elevation of the bars indicates its yearly frequency. This image illustrates the temporal trends in term usage.

The graph illustrates a rise in the employment of these terms throughout the analysis. Commencing in 2020, the bars ascend, signifying a substantial augmentation in term frequency. A meticulous examination of the phrases uncovers patterns. The frequent use of terms such as "genome," "annotation," and "Escherichia coli" indicates ongoing scholarly interest in these topics. Synthetic Biology, Genes, and Pathways predominate, indicating a focus on biological and genetic engineering research. Since 2020, there has been a substantial growth in Stress and Insights, indicating an emphasis on stress responses and emerging research.

The prominence of Arabidopsis thaliana and gene expression is increasing, indicating a growing significance in plant biology and gene expression research. Keywords such as "Metabolism" and "Protein" signify active research in the fields of metabolism and protein studies. Subsequent to 2020, there was a considerable increase in secondary metabolites and gene expression. highlighting these subjects. The terms "prediction," "evolution," "identification," and "biosynthesis," reflecting an increasing interest in evolutionary research, predictive modeling, and compound identification and biosynthesis, have also risen.

The increased frequency of keywords may suggest the expanding scope of a scientific field such as biology or bioinformatics. The increasing emphasis on stress responses, gene expression, and secondary metabolites indicates a shift in research goals; nonetheless, the persistent utilization of genomic and genetic terminology underscores their significance.

Nevertheless, the graph possesses limitations. Further research is required to elucidate the underlying patterns and their ramifications, as it does not clarify the usage of the terminology. Examining these terms in relation to financing, publication numbers, and patents may yield further insights. These patterns indicate the necessity of investigating stress responses, gene expression,and secondary metabolites to address the increasing interests within the discipline. This bar graph illustrates the evolution of the study focus within a scientific discipline. Recognizing these trends enables researchers to identify new subjects, allocate resources effectively, and adjust their research in accordance with the evolving direction of the discipline.

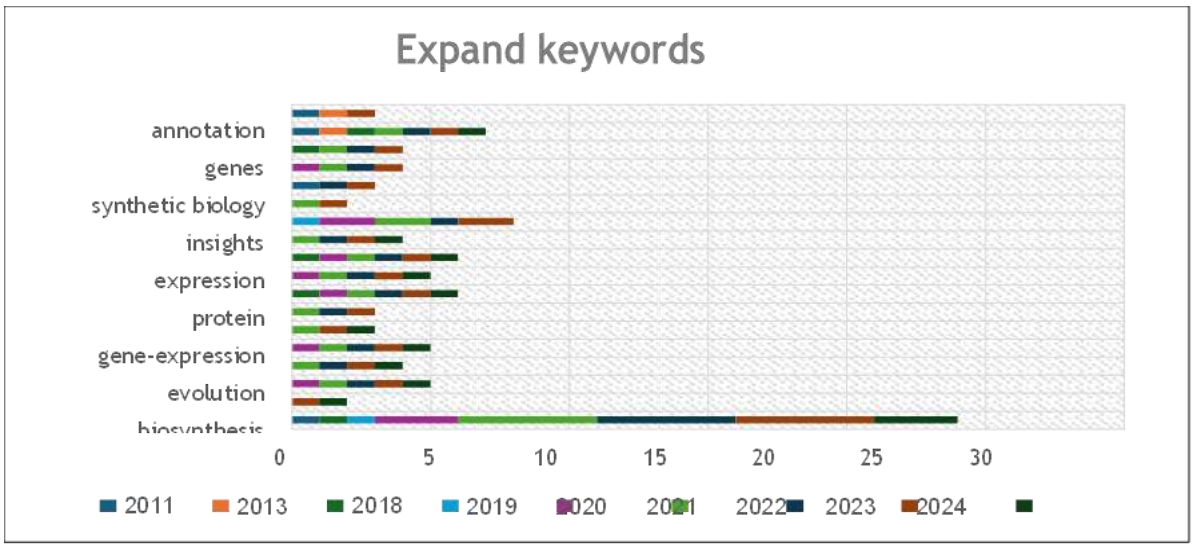

Figure (4) shows keywords

The picture above is a network visualization depicting the interconnections among several concepts related to artificial intelligence (AI) and its applications in areas such as machine learning, deep learning, and fungal research.
The node with the name "artificial intelligence," which is at the heart of the network, signifies that it is the main object of the visualization. Because AI is interdisciplinary and relevant to many scientific fields, this primary node is connected to a number of other important concepts. The terminology associated with machine learning is emphasized in the left green cluster of the diagram. The significance of "classification" and "prediction" underscores their importance in machine learning. These nodes likely signify traditional machine learning model tasks.
emphasizing their significance in AI-facilitated decision-making. In the lower section of the image, "deep learning" is depicted as a distinct node connected to AI, illustrating its unique yet essential function within the realm of artificial intelligence. The identification and diagnosis within the red cluster illustrate the application of deep learning in fields necessitating sophisticated data processing and pattern recognition.The blue cluster on the right, centered on "fungi," is associated with "secondary metabolites" and "molecular docking." This network segment indicates an increasing interest in the convergence of biological sciences and artificial intelligence, encompassing fungal species and biochemistry. The relationships between AI and phrases connected to fungus might indicate how useful AI is for improving ecological research, medication development, and bioinformatics. An intuitive comprehension of the connections between AI and its subfields is made possible by the diagram's color-coded clusters, which represent thematic groups. The different node sizes probably reflect the relative frequency or significance of these phrases in scholarly literature, emphasizing regions of considerable study interest. All things considered, this illustration well capt
ures the interdependence and wide range of applications of artificial intelligence in several fields, highlighting its crucial significance in current scientific research.

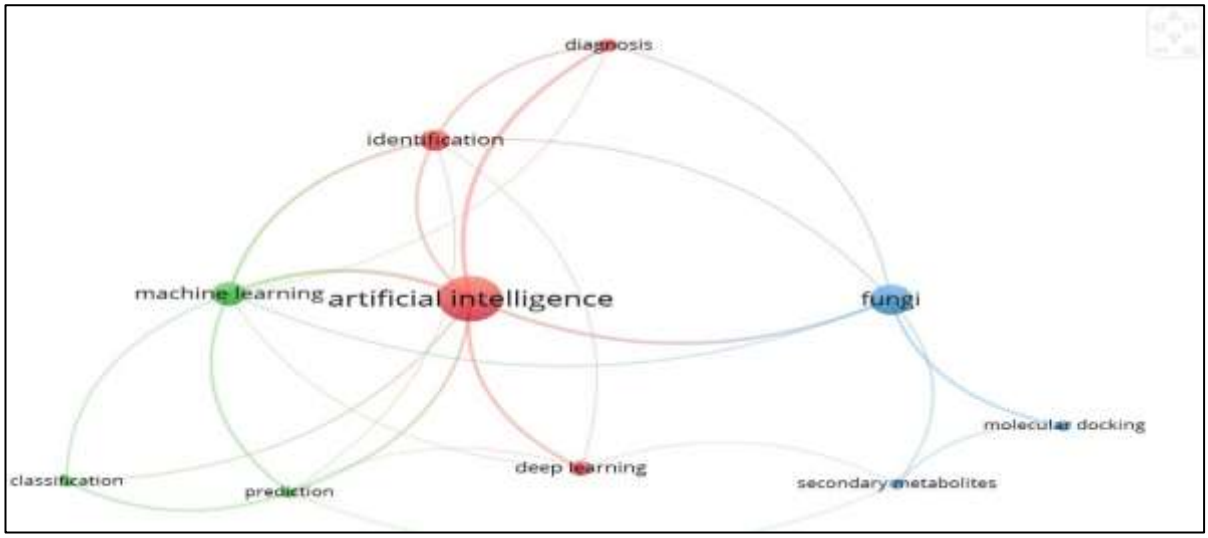

Figure (shows 5) Co-occurrence of keywords

Keywords:
Selected keywords are included in the table along with their frequency and overall link strength. Its essential importance is indicated by the fact that "artificial intelligence" appears the most frequently (21 times) and has the strongest relationship (28). "Fungi" and "machine learning" both have close relationships. Although they are less common, concepts like "prediction," "diagnosis," and "deep learning" are nonetheless important. Key study issues and their relationships are highlighted by the data.

Table (4) shows the keywords

Selected	Keyword	Occurrences	Total link strength
✓	artificial intelligence	21	28
✓	machine learning	11	19
✓	fungi	14	17

✓	identification	10	14
✓	prediction	5	12
✓	diagnosis	6	11
✓	deep learning	7	8
✓	classification	5	7
✓	molecular docking	5	6
✓	secondary metabolites	5	6

4. Summary of Current Studies

Recent research at the intersection of artificial intelligence (AI) and biosynthesis highlights the growing role of AI in decoding complex biological systems, improving the production of valuable compounds, and advancing our understanding of how living organism's function. The existing literature spans several thematic areas, reflecting both the complexity and opportunity in biosynthetic research.

A major area of exploration involves gene discovery and regulation in biosynthetic pathways. For example, machine learning was applied to identify long non-coding RNAs (lncRNAs) involved in secondary cell wall formation in moso bamboo (Phyllostachys edulis), offering insights into lignin and flavonoid biosynthesis (Abdullah-Zawawi et al., 2022). Similarly, in Brassica napus, AI techniques were used to investigate transcriptional variability and reveal gene networks that influence seed oil production, uncovering key regulatory elements such as eQTLs and transcription factors like NAC13 and SCL31 (García-Pérez et al., 2020). Another study using Arabidopsis integrated genomics and proteomics through automated machine learning to predict genes responsible for specialized metabolite biosynthesis (Bassel et al., 2011). Advances have also been made in predicting transcription factor-binding sites in plants using supervised machine learning, with chromatin structure emerging as a significant feature (Rodrigues and Diosdado, 2023). Even in cases where traditional biosynthetic genes are absent, AI has helped uncover functional genes—such as nitrogen fixation-related genes in maize-associated Lactococcus—by combining population genomics with machine learning techniques (Jiang et al., 2023).

Another key area involves metabolomics and chemical profiling. AI and chemoinformatics have been combined to classify angiosperms based on sulfur-containing metabolites, revealing evolutionary relationships through metabolic diversity (Bai et al., 2024). Broader plant metabolomics research is benefiting from data-driven tools like machine learning, network analysis, and statistical models, which are being used to explore plant growth patterns and environmental responses (Dean et al., 2022). Notably, spatial metabolomics paired with random forest algorithms has mapped how influenza virus infection impacts the lung tissue metabolome, identifying key metabolic disruptions (Rodrigues and Deusdado, 2023). Machine learning has also played a pivotal role in profiling volatile organic compounds (VOCs) in Streptomyces, helping to annotate new compounds with diverse biological functions (Rivière et al., 2022).

Table (5) Summary of Current Studies

Article ID	AI Technique Used	Biological Application	Main Outcome/Contribution
(Abdullah-Zawawi et al., 2022)	Machine Learning	lncRNA identification in secondary cell wall biosynthesis (moso bamboo)	Identified lncRNA-coding gene networks and SCW-related lncRNAs.
(Bai et al., 2024)	Machine Learning (Tanimoto coefficient, clustering)	Angiosperm classification based on sulfur-containing compounds	Classified Angiosperms and showed association between metabolite structure and plant phylogeny.
(Bassel et al., 2011)	Automated Machine Learning (AutoGluon-Tabular)	Prediction of plant specialized metabolite biosynthesis genes	Identified genomics and proteomics as crucial features for predicting PSM biosynthesis genes.

(Cabanás and Mercado-Blanco, 2025)	Artificial Neural Networks	Optimizing phenolic compound production in in vitro plant culture	Optimized phenolic extraction and identified factors influencing biosynthesis.
(Cangi and Yagcı, 2017)	Machine Learning (Review)	Streamlining natural product biomanufacturing in microbes	Reviewed integration of omics and ML for optimizing biosynthetic pathways.
(Dean et al., 2022)	Machine Learning, Network Analysis, Statistical Modeling (Review)	Plant metabolomics and understanding plant metabolism	Highlighted potential of data science to revolutionize plant metabolomics.
(Elsayed et al., 2021)	XGBoost, Basenji (MachineLearning)	Gene networks regulating seed oil content in Brassica napus	Identified eQTLs and transcription factors (NAC13, SCL31) regulating seed oil content.
(García-Pérez et al., 2020)	Definitive Screening Design, Artificial Neural Networks	Bioconversion of date palm fronds into citric acid	Optimized one-step bioprocessing for citric acid production.
(Gou et al., 2024)	Machine Learning	Discovery of biocontrol bacteriawith fungicidal activity	Accelerated discovery rate and identified novel fungicidal taxonomic groups.
(Higdon et al., 2020)	Machine Learning (Proposed Framework)	'Yeastizing' plant enzymes forimproved performance in yeast hosts	Proposed a computational pipeline to redesign plant enzymes for better adaptation.
(Jamieson et al., 2021)	Machine Learning	Prediction of adenylate-forming enzyme function and substrate specificity	Predicted and experimentally validated a novel β-lactone biosynthesis pathway in Nocardia.
(Jiang et al., 2023)	Population Genomics, Genome-Wide Association Study, Random Forests (MachineLearning)	Identification of nitrogen fixation genes in Lactococcus	Identified novel genes for BNF in Lactococcus lacking canonical nif genes.
(Kisiel et al., 2023)	Network Analysis, MachineLearning	Prediction of metabolic pathways from tomato metabolomics data	Detected novel metabolic pathways and validated one in vivo.
(Knaack et al., 2022)	Machine Learning, Deep Learning (Review)	Biotic and abiotic stress management in plants	Summarized advanced applications of ML/DL for stress detection and mitigation.
(Kufs et al., 2020)	Logistic Regression, RandomForest, XGBoost (Machine Learning)	Identification of rye wax genes	Mapped gene for epicuticular wax formation and identified putative candidate gene (ABCG11).
(Lim et al., 2023)	Rule-Based Machine Learning	Functional network construction in Arabidopsis	Identified novel regulators of seed germination and functional gene associations.
(Liu et al., 2022)	Supervised Machine Learning	Prediction of transcription factor- binding sites in plants	Improved TFBS prediction by integrating genomic features, especially chromatin state.

(Lv and Wang, 2024)	Meta-learning (SVM, KLR, meta-SVM, meta-KLR)	Bacteria classification and identification of informative genes in plant-microbe interaction	Identified informative genes in Bacillus megaterium-tomato root interaction.
(Priya et al., 2018)	Machine Learning (Hidden Markov Models)	Identification and analysis of the plant terpenome	Developed Terzyme tool for predicting terpene synthase and prenyl transferase enzymes.
(Ramzi et al., 2020)	Knowledge-Based Fuzzy Adaptive Resonance Theory (KB-FuzzyART)	Gene networks in leaf development in Arabidopsis	Revealed critical role of ETT pathway in adaxial–abaxial patterning and cell division.
(Rivière et al., 2022)	Machine Learning (MSHub/GNPS workflow)	Analysis of Streptomyces volatilomes	Identified and annotated VOCs with diverse biological activities.
(Robinson et al., 2020)	Supervised Machine Learning	Effect of plant-derived microbial soil legacy in grafting system	Identified key microbial genera influencing plant growth and optimized microbiota benefits.
(Rodrigues and Deusdado, 2023)	Random Forest (Machine Learning)	Spatial metabolomics of influenza virus infection	Identified localized metabolic perturbations in lung tissue.
(Sajid et al., 2021)	Deep Transfer Learning (Graph Transformer, CNN)	Prediction of plant secondary metabolic pathways	Achieved high accuracy in classifying KEGG and plant secondary metabolic pathways.
(Takahashi et al., 2013)	Ensemble Learning (NRTPredictor)	Identifying rice root cell state in single-cell RNA-seq	Predicted cell states and identified marker genes involved in phenylpropanoid biosynthesis.
(Tan et al., 2022)	Machine Learning	Deciphering genetic basis of metabolism in polished rice	Combined hyperspectral imaging with ML to assess crop metabolites and identify novel genes.

In the area of bioproduction and metabolic engineering, AI is being used to improve the efficiency of compound synthesis. For example, integrating plant tissue culture with machine learning has enabled researchers to optimize conditions for extracting phenolic compounds from medicinal plants (Cabanás and Mercado-Blanco, 2025). In microbial biotechnology, combining omics data with AI is helping enhance microbial systems for producing natural compounds like flavonoids and terpenes (Cangi and Yagcı, 2017). One innovative study used artificial neural networks and screening designs to optimize the conversion of date palm waste into citric acid in a one-step bioprocess (García-Pérez et al., 2020). Moreover, some researchers propose using machine learning to redesign plant enzymes so they can function more effectively in yeast systems, facilitating scalable biomanufacturing (Higdon et al., 2020).

AI is also making strides in agriculture and plant health. Genomics combined with machine learning has accelerated the discovery of biocontrol bacteria with antifungal properties, revealing new microbial groups with agricultural potential (Gou et al., 2024). AI models, including deep learning, are increasingly used for early detection of plant diseases and stress factors, supporting precision farming efforts (Knaack et al., 2022). Supervised machine learning has been applied to study how grafted plants benefit from microbial legacies in the soil, identifying key microbes that promote plant growth and resilience (Robinson et al., 2020). At the cellular level, ensemble learning models like NRTPredictor have been created to analyze single-cell RNA sequencing data in rice roots, aiding in the identification of stress-related cell types and biomarkers (Takahashi et al., 2013). Other studies use deep transfer learning to accurately predict secondary metabolic pathways and classify plant-derived compounds (Sajid et al., 2021), while tools like Ter zyme utilize machine learning to automatically identify

enzymes involved in terpene synthesis (Priya et al., 2018).
Taken together, these findings demonstrate the powerful role of AI in driving innovation across the biosynthesis landscape—from fundamental gene discovery to real-world applications in agriculture and bioengineering. A consistent theme across the studies is the integration of complex biological data with advanced computational models, offering deeper insights and more streamlined solutions for biosynthetic research.

5. Methodologies Used in Reviewed Studies: Strengths and Limitations

The methodologies applied in AI-biosynthesis research are as varied as the biological questions they aim to address. A shared pattern across these studies is the reliance on computational models—especially machine learning (ML) and deep learning (DL)—to make sense of complex, high-throughput biological data. This section outlines the primary methodological approaches employed in the reviewed papers, highlighting both their strengths and the limitations encountered.

Data-Driven Discovery and Functional Prediction
Several studies center around the use of large-scale biological datasets to identify new genes, predict regulatory elements, or assign biological functions. For example, researchers studying Phyllostachys edulis (moso bamboo) (Abdullah-Zawawi et al., 2022). utilized machine learning to analyze a comprehensive set of transcriptomic data, including 231 RNA-Seq datasets, one Iso-Seq, and one full-length cDNA dataset. This rich data context allowed for meaningful insights into lncRNA regulatory networks; however, the absence of information about the specific machine learning models or parameters used makes the study hard to reproduce or critically assess.
A study on Xanthomonas strains (Bao et al., 2023) used a combination of CART, Lasso, and Random Forest algorithms to analyze protein domains from 118 strains. The comparative analysis of different ML models provided robust predictive capabilities and helped unify fragmented data from various sources. Still, the limited sample size and potential classification bias derived from literature curation present challenges to generalizability and accuracy.
In Arabidopsis, a study integrated genomics, transcriptomics, proteomics, and epigenomics with AutoGluon-Tabular (an AutoML framework) to predict specialized metabolite biosynthesis genes (Bassel et al., 2011). The automation of feature selection and external validation across multiple plant species adds substantial value. However, the lack of specificity regarding which features within genomics or proteomics carried the most weight limits biological interpretation.
For Lactococcus isolates lacking known nitrogen fixation genes (Jiang et al., 2023), researchers combined pangenomic analysis, GWAS, and Random Forests. This integrative strategy was effective in discovering candidate genes via patterns rather than known signatures. Nonetheless, the study's computational nature necessitates further laboratory validation, and again, details about the ML model design were sparse.
Predicting transcription factor binding sites (TFBSs) in plants (Liu et al., 2022) used supervised ML enriched by incorporating chromatin accessibility and conservation data. The key advantage here was the clear improvement in TFBS prediction accuracy and the development of a user-friendly tool, Wimtrap. However, the reliance on available ChIP-chip/ChIP-seq data restricts the method's broader applicability, especially for underrepresented transcription factors.

Omics Integration and Analysis
Integrating multi-omics datasets is a recurring strategy. A study on seed oil content in Brassica napus (Elsayed et al., 2021) employed XG Boost and Basenji models across hundreds of RNA-seq and ATAC-seq datasets. The robustness of the model was supported by experimental verification of key transcription factors. Still, the underlying molecular mechanisms for observed e QTL effects remain unclear, and the model's performance in varying conditions remains to be tested.
A review on plant metabolomics (Dean et al., 2022) discussed the fusion of ML, network analysis, and statistical tools in interpreting mass spectrometry data. While informative, the discussion remained high-level, with limited exploration of challenges like inconsistent data quality or the reproducibility of specific methods.
In spatial metabolomics (Rodrigues and Diosdado, 2023), LC-MS/MS combined with 3D tissue models and Random Forest algorithms to explore metabolic responses to influenza infection in mice. This technique allowed for precise localization of biochemical changes. Nevertheless, the reliance on

animal models and the complex nature of spatial data analysis poses obstacles for broader application and clinical translation.

Optimization of Biosynthetic Pathways

AI also plays a key role in enhancing bioproduction and optimizing metabolic pathways. In a study on Bryophyllum (Cabanás and Mercado-Blanco, 2025), Artificial Neural Networks (ANNs) were trained on experimental data to improve phenolic compound yield from in vitro cultures. This highlights ANNs' power in modeling non-linear, multifactorial interactions. However, the narrow focus on a single plant genus and lack of transparency about ANN architecture limit broader impact.

A comprehensive review (Cangi and Yagcı, 2017) described how omics datasets are being utilized in conjunction with ML to streamline microbial engineering for natural product synthesis. Although the overview is valuable, it lacks specific methodological details and experimental validation.

Date palm frond bioconversion into citric acid was explored using a blend of Definitive Screening Design (DSD) and ANN (García-Pérez et al., 2020). The one-step approach is innovative, and the ANN's predictive accuracy outperformed DSD. Yet, the experiment's scope was limited to a single microbial strain and substrate, and more details about the ANN's parameters were needed.

A conceptual paper (Higdon et al., 2020) introduced the idea of "yeastizing" plant enzymes using a machine learning strategy to adapt them for expression in yeast. The theoretical model is promising for improving biosynthesis in heterologous hosts, though it currently lacks experimental proof.

Agricultural and Plant Health Applications

AI is increasingly applied to crop health and agricultural productivity. A study (Gou et al., 2024) used a novel ML workflow to identify fungicidal traits in biocontrol bacteria, accelerating discovery rates and revealing new microbial candidates. However, the methodology is limited to only two plant pathogens, and algorithm details remain vague.

The application of AI to stress detection in plants was explored in a review (Knaack et al., 2022), covering DL and ML strategies for high-throughput phenotyping. While the breadth of the review is commendable, it lacks deep dives into implementation and scalability of the technologies discussed.

Research on Streptomyces volatilomes (Rivière et al., 2022) used the MS Hub/GNPS pipeline alongside traditional methods to classify VOCs. This integration improved annotation rates and uncovered compounds with promising bioactivities. Still, many VOCs remained unidentified, and cross-validation across tools highlighted discrepancies.

Deep transfer learning using a hybrid Graph Transformer-CNN (GTC) model was proposed for classifying secondary metabolic pathways (Sajid et al., 2021). The approach delivered impressive accuracy and rich molecular feature extraction. Yet, as with many DL models, interpretability is limited, and performance depends heavily on the quality of training databases.

NRTPredictor (Takahashi et al., 2013), an ensemble model developed for single-cell RNA-seq analysis in rice roots, demonstrated strong accuracy in identifying cellular states and biomarkers. Its integration with bulk RNA-seq provided context to the results. However, as with other ensemble methods, the specific algorithms used were not detailed, and biological validation remains pending.

A final example (Tan et al., 2022) combined hyperspectral imaging with ML to study polished rice metabolomics. This cost-effective and non-invasive technique successfully pinpointed candidate genes and was validated through gene editing. Still, the setup costs may be prohibitive for some labs, and the findings are specific to one crop variety.

Table (6) Methodology Used in Each Study

Article ID	Methodology (AI/ML Techniques)	Strengths	Limitations
(Abdullah-Zawawi et al., 2022)	Machine Learning on RNA-Seq, Iso-Seq, cDNA datasets	Comprehensive data, machine learning for lncRNA identification.	Specific ML algorithms/parameters not detailed; experimental validation needed.
(Bai et al., 2024)	Chemoinformatics, Tanimoto coefficient, Machine Learning (clustering)	Large database (KNApSAck); novel chemical-based classification; phylogenetic alignment.	Focus on SCCs only; specific ML algorithm not detailed; biological implications need more study.

(Bao et al., 2023)	CART, Lasso, Random Forest on Pfam protein domains	Comparison of multiple ML algorithms; unification of fragmented literature data.	Potential for bias from literature mining; relatively small genomic dataset; specific functions of domains not detailed.
(Bassel et al., 2011)	AutoGluon-Tabular (Automated ML) on multi-omics data	Streamlined prediction; identification of crucial features (genomics, proteomics); external validation.	Specific features within omics not detailed; generalizability limited by training species diversity; biological interpretation limited.
(Cabanás and Mercado-Blanco, 2025)	Artificial Neural Networks (ANNs) for data modeling	ANNs model complex relationships; practical insights for optimization.	Focus on specific genus; ANN architecture/training details not extensive; further experimental validation needed.
(Cangi and Yagcı, 2017)	Review of Omics and Machine Learning integration	Comprehensive overview; highlights synergistic potential; addresses bottlenecks.	No new data; limited detail on specific ML algorithms/integration strategies; potential selection bias.
(Dean et al., 2022)	Review of Machine Learning, Network Analysis, Statistical Modeling	Comprehensive overview of data science in metabolomics; potential for crop improvement.	No new data/case studies; general discussion on data quality/complexity; no concrete solutions.
(Elsayed et al., 2021)	XGBoost, Basenji (ML) on RNA-seq, ATAC-seq	Robust data foundation; integration of two ML models; experimental validation of TFs.	Precise molecular mechanisms of eQTLs not fully elucidated; generalizability to other conditions needs study.
(García-Pérez et al., 2020)	Definitive Screening Design (DSD), Artificial Neural Networks (ANNs)	Novel one-step bioprocessing; superior ANN predictive capability.	Focus on single strain/substrate; ANN architecture/training details not extensive; economic feasibility not discussed.
(Gou et al., 2024)	Novel Machine Learning workflow for genomic features	Threefold improvement in discovery rate; identified new fungicidal taxonomic groups.	Specific ML algorithm details lacking; focus on two diseases; exact metrics for improvement not fully elaborated.
(Higdon et al., 2020)	Proposed Data-Driven Machine Learning Framework	Addresses critical bottleneck (enzyme adaptation); broadly applicable computational solution.	Conceptual framework, no experimental validation; specific ML algorithms not detailed; data acquisition challenges.
(Jamieson et al., 2021)	Machine Learning for protein sequences; phylogenetic reconstruction	Significant advancement in AFE function prediction; broad genomic analysis; experimental validation.	ML model accuracy depends on training data; focus on AFEs; detailed ML parameters not fully described.
(Jiang et al., 2023)	Pangenome analysis, GWAS, Random Forests (ML)	Multi-pronged approach for novel gene identification; addresses non-canonical genes.	Experimental validation of gene roles needed; specific ML details not extensive; transferability to other systems needs study.

(Kisiel et al., 2023)	Correlation-based Network Analysis, Machine Learning	Powerful for de novo pathway discovery; identifies pathways even without specific metabolites.	Not all pathway components fully resolved; generalizability to other species needs study; specific ML algorithm not named.
(Knaack et al., 2022)	Review of Machine Learning, Deep Learning with HTP phenotyping	Comprehensive overview of AI in stress management; highlights HTP potential.	No new data; general algorithm discussion; practical deployment challenges not extensively discussed.
(Kufs et al., 2020)	Logistic Regression, Random Forest, XGBoost (ML)	Innovative combination of traditional mapping with ML; high-density markers; novel gene mapping.	Experimental validation of gene function needed; specific ML parameters not detailed; focus on single trait.
(Lim et al., 2023)	Rule-Based Machine Learning (Coprediction)	Beyond coexpression analysis; predicts functional associations irrespective of expression similarity; public web tool.	Focus on Arabidopsis seed; specific ML algorithm details not extensive; biological interpretation of rules complex.
(Liu et al., 2022)	Supervised Machine Learning integrating genomic features	Significant improvement in TFBS prediction (chromatin features); user-friendly tool (Wimtrap).	Reliance on ChIP-chip/seq data; specific ML algorithm details not extensive; perfect accuracy remains a challenge.
(Lv and Wang, 2024)	Meta-learning (SVM, KLR, meta-SVM, meta-KLR)	Sophisticated approach for classification; identifies informative genes; functional enrichment.	Specific implementation details not extensive; focus on single PGPR/host; experimental validation needed.
(Priya et al., 2018)	Machine Learning (Hidden Markov Models)	Dedicated automated tool for terpenome analysis; integrates sequence, taxonomy, ML; webserver available.	Does not directly predict specific compounds; accuracy depends on training data; HMMs may not capture all nuances.
(Ramzi et al., 2020)	Knowledge-Based Fuzzy Adaptive Resonance Theory (KB-FuzzyART)	Novel clustering algorithm for gene expression; integrates genetic/molecular analysis.	Focus on Arabidopsis; specific algorithm details not extensive; precise biochemical interactions not fully elucidated.
(Rivière et al., 2022)	MSHub/GNPS (Machine Learning workflow)	Enhanced VOC annotation; identifies diverse bioactivities; molecular networking for unknown VOCs.	High number of unknown VOCs; small overlap between methods; focus on specific geographic region.
(Robinson et al., 2020)	Supervised Machine Learning	Well-designed experimental setup; comprehensive omics integration; identifies key microbial genera.	Focus on specific plant/grafting system; specific ML algorithms not detailed; further experimental validation needed.
(Rodrigues and Deusdado, 2023)	Random Forest (ML) with LC-MS/MS and 3D models	Powerful for localized metabolic changes; comprehensive data analysis; insights into pathogenesis.	Mouse model (translation to human needs validation); specific biological pathways not detailed; data complexity.

(Sajid et al., 2021)	Deep Transfer Learning (Graph Transformer, CNN)	Innovative hybrid DL architecture; high accuracy; comprehensive molecular representation; user-friendly tool.	Interpretability challenges; reliance on existing databases; experimental validation needed.
(Takahashi et al., 2013)	Ensemble Learning (NRTPredictor)	Robust cell state prediction; high accuracy/recall; model interpretability; webserver available.	Specific ensemble algorithm details not extensive; generalizability to other tissues/species needs study; experimental validation needed.
(Tan et al., 2022)	Machine Learning with Hyperspectral Imaging	Inexpensive and powerful approach for metabolite assessment; integrates phenomics, metabolomics, genomics; gene editing validation.	Initial setup cost; focus on polished rice; specific ML algorithms not explicitly detailed.

6. Challenges and Recommendations

The integration of artificial intelligence into biosynthesis research has opened up new possibilities, but it also presents a number of challenges. Across the literature reviewed, recurring themes emerge—ranging from limitations in data availability to the complexity of biological systems and the need for greater interdisciplinary collaboration. Below, we outline these challenges along with practical recommendations drawn from the studies.

1. Data-Related Challenges and Recommendations Challenges:

Limited Data Availability and Quality: A recurring issue in many studies is the scarcity of comprehensive, high-quality datasets. Particularly in research involving non-model organisms or less-studied biosynthetic processes, the lack of sufficient experimental data limits the training and effectiveness of AI models (Abdullah-Zawawi et al., 2022؛Jamieson et al., 2021 ؛Liu et al., 2022). Additionally, fragmented and inconsistent data across studies hampers efforts at large-scale integration (Bao et al., 2023). In metabolomics, the vast diversity of plant metabolites further complicates data collection and standardization (Dean et al., 2022).

Complexity of Biological Data: Even when data is available, its structure can be highly complex—often high-dimensional and sparse (as seen in sc RNA-seq studies (Takahashi et al., 2013). Deep learning models, though powerful, often lack transparency, making it difficult to interpret predictions or extract biological meaning (Sajid et al., 2021).

Recommendations:

Standardization and Open Sharing: To address data inconsistencies, researchers should work towards standardized protocols for data collection and sharing. Establishing public repositories with well-annotated omics datasets (genomic, transcriptomic, proteomic, metabolomic) would facilitate more robust model training and validation (Dean et al., 2022؛Góralska et al., 2020).

Advanced Preprocessing and Feature Engineering: Improving how biological data is encoded and prepared for AI models is critical. Developing new preprocessing techniques and biologically-informed feature engineering methods can help extract meaningful insights and improve predictive performance (Goralski et al., 2020؛Sajid et al., 2021).

Improved Model Interpretability: A shift toward more explainable AI (XAI) is essential. Researchers should aim to make complex models more transparent, so that predictions can be linked back to biological mechanisms, not just computational patterns (Sajid et al., 2021).

2. Biological System Complexity and Validation Challenges Challenges:

Complex Biosynthetic Networks: Gene regulatory networks and biosynthetic pathways often involve intricate, multi-layered interactions. In some cases—such as identifying non-canonical nitrogen fixation genes in Lactococcus (Jiang et al., 2023)—the complexity of these systems makes it difficult to infer accurate biological functions from computational models alone. Chromatin accessibility, gene expression regulation, and epigenetic factors further complicate these analyses (Biggs et al., 2021).

Validation Bottlenecks: While AI can generate numerous hypotheses, experimentally verifying these predictions remains a major challenge. Studies often highlight the gap between computational discovery and biological confirmation, which can be labor-intensive and time-consuming (Bassel et al., 2011). Findings in model organisms may also not always transfer reliably to other species (Ramzi et al., 2020). Challenges in Translation to Practice: Bridging the gap from theoretical models to real-world applications—such as crop improvement or industrial-scale bioproduction—raises issues of scalability, cost, and environmental variability (García-Pérez et al., 2020).

Recommendations:

Adopt Integrated Multi-Omics Approaches: Combining genomic, transcriptomic, proteomic, metabolomic, and phenotypic data provides a more comprehensive view of biosynthetic systems (Tan et al., 2022). Multi-omics integration helps uncover interactions that single data types might miss.

Develop High-Throughput Validation Systems: Accelerating the pace of validation requires tools such as CRISPR-based gene editing and high-throughput functional assays to experimentally verify AI-driven predictions (Tan et al., 2022).

Investigate Mechanistic Pathways More Deeply: Moving beyond high-level associations to study the detailed biochemical mechanisms of predicted genes or enzymes is vital for both scientific understanding and applied innovation (Robinson et al., 2020).

3. Methodological and Interdisciplinary Challenges Challenges:

Algorithm Selection and Reproducibility: Choosing the most appropriate AI model for a given biological problem—and properly optimizing it—can be difficult, especially given the wide range of available methods (Cabanás and Mercado-Blanco, 2025). Several studies lack sufficient detail about the models used, which makes it harder to reproduce or build upon their work (Gou et al., 2024).

Interdisciplinary Knowledge Gaps: Effective AI-driven biosynthesis research demands collaboration across biology, chemistry, computer science, and engineering. However, bridging disciplinary boundaries and fostering mutual understanding remains a major hurdle (Biggs et al., 2021).

Modeling Dynamic Systems: Biological systems are inherently dynamic. Capturing temporal shifts—such as plant development stages or stress responses—requires more advanced AI methods capable of modeling complex, time-varying interactions (Biggs et al., 2021).

Recommendations:

Promote Methodological Transparency: Authors should clearly document their model architectures, training parameters, and evaluation metrics. Open-source tools and publicly available codebases help ensure reproducibility and foster collaboration (Abdullah-Zawawi et al., 2022؛Cabanás and Mercado-Blanco.(Priya et al., 2018 Takahashi et al., 2013).

Encourage Interdisciplinary Training and Teams: Integrating biological knowledge with computational expertise should be a core priority. Training programs, joint research centers, and cross-disciplinary initiatives can promote the synergy needed for impactful research (Dean et al., 2022).

Develop Domain-Specific AI Tools: There is growing need for customized AI frameworks tailored to biosynthesis challenges—such as enzyme redesign, pathway optimization, or metabolite classification (Higdon et al., 2020؛ Sajid et al., 2021). These tools should reflect both computational advances and biological constraints.

Embed AI in Iterative Design Cycles: Integrating AI models into iterative design–build–test–learn (DBTL) cycles can streamline optimization and accelerate discovery, particularly in synthetic biology and metabolic engineering contexts (Higdon et al., 2020).

Table (7): Challenges and Recommendations

Article ID	Challenges Identified	Recommendations Proposed
(Abdullah-Zawawi et al., 2022)	Lack of detailed ML information; experimental validation needed.	Further experimental validation of lncRNA functions.
(Bai et al., 2024)	Limited to sulfur-containing compounds;specific ML not detailed.	Expansion to other compound classes; further integration with phylogenetic data.
(Bao et al., 2023)	Bias from literature mining; relativelysmall dataset; fragmented literature.	Standardized data collection; larger, more diverse datasets; detailed characterization of protein domains.

(Bassel et al., 2011)	Generalizability limited by training species diversity; biological interpretation.	Further validation across diverse species; deeper biological interpretation of features.
(Biggs et al., 2021)	Understanding chromatin accessibility;dynamic gene expression.	Multi-omics integration; advanced AI for dynamic modeling;experimental validation of regulatory mechanisms.
(Cabanás and Mercado-Blanco, 2025)	Focus on specific genus; insufficient ANN details; generalizability.	Broader application to other plants; detailed reporting of ML parameters; economic feasibility studies.
(Cangi and Yagcı, 2017)	Lack of new experimental data; limited detail on specific ML algorithms.	More experimental validation; detailed reporting of ML algorithms and integration strategies.
(Dean et al., 2022)	Data quality and availability; complexity of plant metabolome.	Standardized data collection; improved data preprocessing; interdisciplinary collaboration.
(Elsayed et al., 2021)	Precise molecular mechanisms of eQTLs; generalizability.	Further investigation into molecular mechanisms; broader application to different conditions.
(García-Pérez et al., 2020)	Focus on single strain/substrate; insufficient ANN details; economic feasibility.	Broader application to other strains/substrates; detailed reporting of ML parameters; economic feasibility studies.
(Góralska et al., 2020)	Data preprocessing; feature engineering; model interpretability.	Improved data preprocessing; novel feature engineering; enhanced model interpretability.
(Gou et al., 2024)	Lack of specific ML algorithm details; focus on two diseases.	Detailed reporting of ML algorithms; broader application to more diseases.
(Higdon et al., 2020)	Conceptual framework, no experimental validation; data acquisition.	Experimental validation; detailed reporting of ML algorithms; addressing data acquisition challenges.
(Jamieson et al., 2021)	ML model accuracy depends on training data; focus on AFEs.	Expansion of training data; broader application to other enzyme families.
(Jiang et al., 2023)	Experimental validation of gene roles; specific ML details.	Experimental validation of gene functions; detailed reporting of ML parameters.
(Kisiel et al., 2023)	Not all pathway components fully resolved; generalizability.	Further resolution of pathway components; broader application to other species.
(Knaack et al., 2022)	General algorithm discussion; practical deployment challenges.	More specific algorithm details; addressing practical deployment challenges.
(Kufs et al., 2020)	Experimental validation of gene function; specific ML parameters.	Experimental validation of gene function; detailed reporting of ML parameters.
(Lim et al., 2023)	Focus on Arabidopsis seed; specific ML algorithm details; biological interpretation.	Broader application to other systems; detailed reporting of ML parameters; improved biological interpretability.
(Liu et al., 2022)	Reliance on ChIP-chip/seq data; specific ML algorithm details.	Expansion of ChIP-chip/seq data; detailed reporting of ML parameters; development of methods for data-scarce TFs.
(Lv and Wang, 2024)	Specific implementation details; focus on single PGPR/host; experimental validation.	Detailed reporting of ML parameters; broader application to other PGPR/hosts; experimental validation.
(Priya et al., 2018)	Does not directly predict specific compounds; accuracy depends on training data.	Integration with specific compound prediction; continuous updating of training data.

(Ramzi et al., 2020)	Focus on Arabidopsis; specific algorithm details; precise biochemical interactions.	Broader application to other plants; detailed reporting of ML parameters; deeper investigation into biochemical interactions.
(Rivière et al., 2022)	High number of unknown VOCs; small overlap between methods.	Further refinement of ML workflows; experimental validation of bioactivities; broader sampling.
(Robinson et al., 2020)	Focus on specific plant/grafting system; specific ML algorithms.	Broader application to other plants/grafting systems; detailed reporting of ML parameters; further experimental validation.
(Rodrigues and Deusdado, 2023)	Mouse model (translation to human needs validation); data complexity.	Validation in human samples; deeper investigation into metabolic pathways; improved data analysis tools.
(Sajid et al., 2021)	Interpretability challenges; reliance on existing databases.	Improved model interpretability; expansion of training data; experimental validation.
(Takahashi et al., 2013)	Specific ensemble algorithm details; generalizability; experimental validation.	Detailed reporting of ML parameters; broader application to other tissues/species; experimental validation.
(Tan et al., 2022)	Initial setup cost; focus on polished rice.	Cost reduction strategies; broader application to other rice tissues/developmental stages.

7. Limitations

While this bibliometric study was designed to methodically and quantitatively analyze the selected body of literature, several limitations should be acknowledged. First, the analysis was confined to the SCI-E index of the Web of Science Core Collection (WoSCC), a widely recognized source for academic citations. However, this choice inevitably narrowed the scope of the research, as potentially relevant studies outside this database were excluded. This may have impacted the overall depth of the findings.

Second, the inclusion criteria were limited to English-language publications and peer-reviewed articles. As a result, significant research published in other languages or in non-traditional formats such as conference proceedings or reports may have been overlooked. Additionally, although the screening process was conducted carefully by subject experts, the manual nature of this task carries an inherent risk of bias or omission due to individual knowledge limitations. Third, the study relied on keyword-based retrieval using terms such as “deep learning” and “biosynthesis.” While this approach was helpful for targeting relevant research, it may have unintentionally omitted studies that used alternative terminology or that explored the topic from different but related perspectives. Moreover, frequently recurring general terms—such as “biosynthesis,” “machine learning,” and “artificial intelligence”—tended to dominate the keyword analysis without necessarily offering specific insights tied to the study’s core objectives.

Finally, access to the full texts of all identified papers posed a challenge. Out of 44 initially selected articles from the WoS database, only 31 were available in full. The remaining 13 papers were inaccessible, which limited the completeness of the content analysis and may have excluded valuable findings that could have enriched the study’s conclusions.

Future work will seek to broaden the range of data sources, improve keyword strategies, and leverage interdisciplinary collaboration to overcome these limitations and deliver a more comprehensive analysis.

8. Conclusion

In relation to biosynthesis, this article reveals research directions and critical points for artificial intelligence. All aspects of biosynthetic diagnosis can benefit greatly from the validated models, but their effectiveness must be demonstrated using a large amount of data over an extended period of time.

This study will help the relevant government agencies and researchers better understand the changing trends in bio fabrication and combine them with artificial intelligence in the future to produce results that benefit world kind.

Conflict of Interest:
The authors declare that there is no conflict of interest for this paper.

Acknowledgments
The authors sincerely thank the Technical Research Center at Northern Technical University for their invaluable support, resources, and guidance, which were essential in completing this research. Their contributions significantly enhanced the quality and progress of this work.

References

1. Durand, M., S. Besseau, N. Papon, and V. Courdavault. 2024. Unlocking plant bioactive pathways: omics data harnessing and machine learning assisting. Curr. Opin. Biotechnol. 87:103135.
2. Jamieson, C.S., J. Misa, Y. Tang, and J.M. Billingsley. 2021. Biosynthesis and synthetic biology of psychoactive natural products. Chem. Soc. Rev. 50:6650–7008.
3. Jiang, Y., Y. Yu, M. Kong, Y. Mei, L. Yuan, Z. Huang, K. Kuang, Z. Wang, H. Yao, and J. Zou. 2023.
4. Artificial intelligence for retrosynthesis prediction. Engineering 25:32–50.
5. Kufs, J.E., S. Hoefgen, J. Rautschek, A.U. Bissell, C. Graf, J. Fiedler, D. Braga, L. Regestein, M.A. Rosenbaum, and J. Thiele. 2020. Rational design of flavonoid production routes using combinatorial and precursor-directed biosynthesis. ACS Synth. Biol. 6:1823–1832.
6. Lv, Y., and W. Wang. 2024. Metabolic design–build–test–learn cycle used for the biosynthesis of plant-derived bioactive compounds. Elsevier.
7. Xie, X., L. Gui, B. Qiao, G. Wang, S. Huang, Y. Zhao, and S. Sun. 2024. Deep learning in template-free de novo biosynthetic pathway design of natural products. Brief. Bioinform. 25:bbae465.
8. Yang, X.X., S.J. Zhang, and J. Luo. 2023. Plant phylogenetics: a bibliometric analysis based on vosviewer and citespace.. Appl. Ecol. Environ. Res. 21.
9. Zheng, S., T. Zeng, C. Li, B. Chen, C.W. Coley, Y. Yang, and R. Wu. 2022. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nat. Commun. 13:3342.
10.Zhong, Z., J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, T. Hou, and M. Song. 2023. Recent advances in artificial intelligence for retrosynthesis. arXiv Prepr. arXiv2301.05864.
11.Wang, J., Hou, Y., Wang, Y., & Zhao, H. 2021. Integrative lncRNA landscape reveals lncRNA-coding gene networks in the secondary cell wall biosynthesis pathway of moso bamboo (Phyllostachys edulis). BMC genomics. 22:638.
12.Abdullah-Zawawi, M. R., Govender, N., Karim, M. B., Altaf-Ul-Amin, M., Kanaya, S., & Mohamed-Hussein, Z. A. 2022. Chemoinformatics-driven classification of angio-sperms using sulfur-containing compounds and machine learning algorithm. Plant Methods. 18:118.
13.Te Molder, D., Poncheewin, W., Schaap, P. J., & Koehorst, J. J. 2021. Machine learning ap-proaches to predict the plant-associated phenotype of Xanthomonas strains. BMC ge-nomics. 22:848.
14.Bai, W., Li, C., Li, W., Wang, H., Han, X., Wang, P., & Wang, L. 2024. Machine learning assists prediction of genes responsible for plant specialized metabolite biosynthesis by integrating multi-omics data. BMC genomics. 25:418.
15.Knaack, S. A., Conde, D., Chakraborty, S., Balmant, K. M., Irving, T. B., Maia, L. G. S., ... & Roy, S. 2022. Temporal change in chromatin accessibility predicts regulators of nodu-lation in Medicago truncatula. BMC biology. 20:252.
16.García-Pérez, P., Lozano-Milo, E., Landín, M., & Gallego, P. P. 2020. Combining medicinal plant in vitro culture with machine learning technologies for maximizing the production of phenolic compounds. Antioxidants. 9:210.
17.Ramzi, A. B., Baharum, S. N., Bunawan, H., & Scrutton, N. S. 2020. Streamlining natural products biomanufacturing with omics and machine learning driven microbial engi-neering. Frontiers in Bioengineering and Biotechnology. 8:608918.
18.Kisiel, A., Krzemińska, A., Cembrowska-Lech, D., & Miller, T. 2023. Data science and plant metabolomics. Metabolites. 13:454.
19.Cabanás, C. G. L., & Mercado-Blanco, J. 2025. Groundbreaking technologies and the biocon-trol

of fungal vascular plant pathogens. Journal of Fungi. 11:77.
20.Tan, Z., Peng, Y., Xiong, Y., Xiong, F., Zhang, Y., Guo, N., ... & Zhao, H. 2022. Comprehen-sive transcriptional variability analysis reveals gene networks regulating seed oil con-tent of Brassica napus. Genome biology. 23:233.
21.Sajid, M., Channakesavula, C. N., Stone, S. R., & Kaur, P. 2021. Synthetic biology towards improved flavonoid pharmacokinetics. Biomolecules. 11:754.
22.Elsayed, M. S., Eldadamony, N. M., Alrdahe, S. S., & Saber, W. I. 2021. Definitive screening design and artificial neural network for modeling a rapid biodegradation of date palm fronds by a new Trichoderma sp. PWN6 into citric acid. Molecules. 26:5048.
23.Lim, P. K., Julca, I., & Mutwil, M. 2023. Redesigning plant specialized metabolism with su-pervised machine learning using publicly available reactome data. Computational and Structural Biotechnology Journal. 21:1639-1650.
24.Biggs, M. B., Craig, K., Gachango, E., Ingham, D., & Twizeyimana, M. 2021. Genomics-and machine learning-accelerated discovery of biocontrol bacteria. Phytobiomes Journal. 5:452-463.
25.Van Gelder, K., Lindner, S. N., Hanson, A. D., & Zhou, J. 2024. Strangers in a foreign land: 'Yeastizing' plant enzymes. Microbial Biotechnology. 17:e14525.
26.Robinson, S. L., Terlouw, B. R., Smith, M. D., Pidot, S. J., Stinear, T. P., Medema, M. H., & Wackett, L. P. 2020. Global analysis of adenylate-forming enzymes reveals β-lactone biosynthesis pathway in pathogenic Nocardia. Journal of Biological Chemistry. 295:14826-14839.
27.Higdon, S. M., Huang, B. C., Bennett, A. B., & Weimer, B. C. 2020. Identification of nitrogen fixation genes in Lactococcus isolated from maize using population genomics and ma-chine learning. Microorganisms. 8:2043.
28.Toubiana, D., Puzis, R., Wen, L., Sikron, N., Kurmanbayeva, A., Soltabayeva, A., ... & Elovici,
29.Y. 2019. Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. Communications biology. 2:214.
30.Gou, C., Zafar, S., Fatima, N., Hasnain, Z., Aslam, N., Iqbal, N., ... & Abbas, M. 2024. Machine and deep learning: artificial intelligence application in biotic and abiotic stress manage-ment in plants. Frontiers in Bioscience-Landmark. 29.
31.Góralska, M., Bińkowski, J., Lenarczyk, N., Bienias, A., Grądzielewska, A., Czyczyło-Mysza, I., ... & Myśków, B. 2020. How machine learning methods helped find putative rye wax genes among GBS data. International Journal of Molecular Sciences. 21:7501.
32.Bassel, G. W., Glaab, E., Marquez, J., Holdsworth, M. J., & Bacardit, J. 2011. Functional net-work construction in Arabidopsis using rule-based machine learning on large-scale data sets. The Plant Cell. 23:3101-3116.
33.Rivière, Q., Corso, M., Ciortan, M., Noël, G., Verbruggen, N., & Defrance, M. 2022. Exploit-ing genomic features to improve the prediction of transcription factor-binding sites in plants. Plant and Cell Physiology. 63:1457-1473.
34.Rodrigues, V., & Deusdado, S. 2023. Meta-learning approach for bacteria classification and identification of informative genes of the Bacillus megaterium: tomato roots tissue in-teraction. 3 Biotech. 13:271.
35.Priya, P., Yadav, A., Chand, J., & Yadav, G. 2018. Terzyme: a tool for identification and anal-ysis of the plant terpenome. Plant Methods. 14:4.
36.Takahashi, H., Iwakawa, H., Ishibashi, N., Kojima, S., Matsumura, Y., Prananingrum, P., ... & Machida, C. 2013. Meta-analyses of microarrays of Arabidopsis asymmetric leaves1 (as1), as2 and their modifying mutants reveal a critical role for the ETT pathway in stabilization of adaxial–abaxial patterning and cell division during leaf development. Plant and cell physiology. 54:418-431.
37.Liu, J., Clarke, J. A., McCann, S., Hillier, N. K., & Tahlan, K. 2022. Analysis of Streptomyces volatilomes using global molecular networking reveals the presence of metabolites with diverse biological activities. Microbiology Spectrum. 10:e00552-22.
38.Wang, T., Ruan, Y., Xu, Q., Shen, Q., Ling, N., & Vandenkoornhuyse, P. 2024. Effect of plant-derived microbial soil legacy in a grafting system—a turn for the better. Microbiome. 12:234.
39.Dean, D. A., Klechka, L., Hossain, E., Parab, A. R., Eaton, K., Hinsdale, M., & McCall, L. I. 2022. Spatial metabolomics reveals localized impact of influenza virus infection on the lung tissue metabolome. Msystems. 7:e00353-22.
40.Bao, H., Zhao, J., Zhao, X., Zhao, C., Lu, X., & Xu, G. 2023. Prediction of plant secondary metabolic pathways using deep transfer learning. BMC bioinformatics. 24:348.
41.Wang, H., Lin, Y. N., Yan, S., Hong, J. P., Tan, J. R., Chen, Y. Q., ... & Fang, W. 2023.

42.NRTPredictor: identifying rice root cell state in single-cell RNA-seq via ensemble learning. Plant Methods. 19:119.
43.Feng, H., Li, Y., Dai, G., Yang, Z., Song, J., Lu, B., ... & Yang, W. 2025. Integrative phenom-ics, metabolomics and genomics analysis provide new insights for deciphering the ge-netic basis of metabolism in polished rice. Genome Biology. 26:55.