Machine Learning and Data Science Applications in Natural Product Research: A Meta-Analytic Perspective

Pokuri Chiranjeevi[1], Tailor Aswartha Gari Manju Teja[1*]

1. Department of Pharmaceutical Analysis, Raghavendra Institute of Pharmaceutical Education and Research, K.R. Palli Cross, Anantapur, Chiyyedu, Andhra Pradesh-515721-India.

***Corresponding Author Address:**

Tailor Aswartha Gari Manju Teja

Student

Department of Pharmaceutical Analysis, Raghavendra Institute of
Pharmaceutical Education and Research, K.R. Palli Cross, Anantapur,
Chiyyedu, Andhra Pradesh-515721.

Mobile: 9014557720

Abstract:

Natural products have long been a cornerstone of drug discovery, offering structurally diverse and pharmacologically active compounds that have led to numerous breakthrough therapeutics. Despite their immense potential, traditional natural product research faces significant challenges, including labor-intensive extraction procedures, complex compound structures, and fragmented, non-standardized datasets. The growing volume and complexity of phytochemical and biological data have necessitated the adoption of advanced computational approaches.

In this context, machine learning (ML) and data science have emerged as transformative tools, enabling automated analysis, prediction, and integration across multiple stages of natural product research. This review provides a comprehensive examination of ML applications in the field, with specific emphasis on phytochemical classification, dereplication, biological activity prediction, ADMET profiling, and the integration of ethnobotanical data using natural language processing. We also highlight the role of deep learning, network pharmacology, and cheminformatics pipelines in accelerating bioactive compound discovery. A bibliometric meta-analysis was conducted using publications from 2000 to 2024 sourced from Scopus, PubMed, and Web of Science. The analysis reveals a sharp rise in research output post-2015, with key contributions from China, India, and the USA. Emerging trends include the use of explainable AI, multi-target modelling, and federated learning frameworks.

The findings underscore the importance of open-access datasets, interdisciplinary collaboration, and transparent model development. As data-driven methodologies continue to evolve, they are poised to redefine how natural product research is conducted—enabling faster, more accurate, and scalable discovery of plant-based therapeutics.

Keywords: Natural Products, Machine Learning, Phytochemicals, Data Science, Bioactivity Prediction

1. Introduction

1.1 Importance of Natural Products in Drug Discovery and Development

Natural products have long served as a cornerstone for drug discovery, offering a vast reservoir of structurally diverse and biologically active compounds[1, 2]. Nearly 50% of all FDA-approved drugs are derived from or inspired by natural sources, particularly plants, marine organisms, and microorganisms[3]. These bioresources provide unique chemical scaffolds not readily available through synthetic libraries, making them invaluable in the search for novel therapeutics against cancer, infectious diseases, metabolic disorders, and neurological conditions. Traditional knowledge systems, such as Ayurveda, Traditional Chinese Medicine, and ethnobotanical practices, further enrich the value of natural products in therapeutic exploration[4].

1.2 Challenges in Traditional Natural Product Research

Despite their promise, natural product research is fraught with challenges. The isolation, purification, and structural elucidation of bioactive compounds from complex extracts require time-consuming and resource-intensive efforts[5]. Dereplication—the process of identifying known compounds to avoid rediscovery—is often inefficient. Additionally, data generated in phytochemical and pharmacological studies is heterogeneous, fragmented, and poorly standardized, making integration and interpretation difficult[6]. These challenges hinder rapid progression from bench to bedside.

1.3 Emergence of Machine Learning and Data Science in Biomedicine

Machine learning (ML) and data science have emerged as transformative tools in biomedical research, offering solutions for handling large, complex datasets[7]. From predictive modelling and pattern recognition to automated image analysis and natural language processing, ML enables high-throughput, cost-effective decision-making[8]. In the realm of natural products, these technologies are beginning to revolutionize phytochemical screening, compound activity prediction, and systems-level analysis.

This review aims to critically evaluate how machine learning and data science are being integrated into natural product research. It seeks to highlight current applications, identify methodological trends, and assess the bibliometric landscape to understand the field's evolution. The objective is to bridge the gap between traditional natural product workflows and modern computational methodologies, it employs a meta-analytic and bibliometric approach, analyzing peer-reviewed publications from major scientific databases such as Scopus, PubMed, and Web of Science. A combination of quantitative and qualitative assessments is used to identify key trends, algorithms employed, research hotspots, and collaboration networks. The focus is on publications from 2000 to 2024 that intersect machine learning, data science, and natural product research.

2. Methodology for Meta-Analysis

2.1 Data Sources

To ensure comprehensive coverage of the literature, three major scientific databases were searched: Scopus, PubMed, and Web of Science. These databases were selected due to their wide indexing of peer-reviewed journals across disciplines, including biomedical sciences, computational biology, pharmacognosy, and data science. PubMed was particularly useful for retrieving studies with biomedical relevance, while Scopus and Web of Science provided broader cross-disciplinary insights, including conference proceedings, book chapters, and citation data.

2.2 Search Strategy and Data Retrieval

To ensure comprehensive coverage of relevant literature, a systematic and structured search strategy was designed and implemented across multiple bibliographic databases, including Scopus, PubMed, and Web of Science[9]. The objective was to retrieve publications focusing on the intersection of machine learning (ML), data science, and natural product research, with an emphasis on pharmacognosy and phytochemistry applications[10].

The search utilized a combination of carefully selected keywords and subject-specific terminology to capture a broad spectrum of relevant studies. The primary search terms included:

- “machine learning”, “deep learning”, “artificial intelligence”,
- “natural products”, “phytochemicals”, “pharmacognosy”,
- “data science”, “predictive modelling”, “QSAR”, and “dereplication”.

To enhance precision and sensitivity of the search, Boolean operators (AND/OR) were strategically applied. For instance, combinations such as:

- “machine learning” AND “natural products”,
- “phytochemicals” AND “predictive modelling”,
- “QSAR” OR “deep learning” AND “pharmacognosy”, were used to extract literature that covered both the computational techniques and their phytochemical applications.

Additionally, advanced filters were employed to refine the results. These included limiting the document type to original research articles, review papers, and conference proceedings, as these sources are most likely to provide substantial methodological insights and empirical data. The language filter was restricted to English to ensure consistency and accessibility in analysis. The publication year filter was set from 2000 to 2024, aligning with the timeline of significant growth in AI applications within biomedical and pharmaceutical domains.

All retrieved citations were imported into bibliographic management tools (e.g., EndNote, Zotero) for de-duplication and screening. The final dataset formed the basis for both qualitative synthesis and bibliometric/meta-analytic evaluations.

2.3 Inclusion/Exclusion Criteria:

2.3.1 Inclusion Criteria

To ensure the quality and relevance of the studies analyzed in this review, a rigorous set of inclusion criteria was established prior to data extraction and bibliometric evaluation[11]. These criteria were designed to selectively capture publications that represent substantive contributions at the intersection of natural product research and machine learning (ML) or data science methodologies[12].

2.3.1.1 Relevance to Natural Product Research and Computational Approaches

Only those publications were included that clearly demonstrated the integration of natural product research encompassing phytochemicals, medicinal plants, pharmacognosy, or secondary metabolites with machine learning, deep learning, artificial intelligence, or data science methodologies[13]. This includes studies employing predictive modelling, QSAR analysis, classification algorithms, chemoinformatics, virtual screening, or natural language processing to solve research questions related to the identification, characterization, or pharmacological evaluation of natural compounds[14].

2.3.1.2 Peer-Reviewed Scholarly Literature

To ensure scientific rigor and reliability, only peer-reviewed publications were considered. This encompassed:

- Original research articles: Empirical studies demonstrating novel applications or results.
- Review articles: Thematic syntheses that analyze trends, methods, or challenges.
- Meta-analyses: Studies that quantitatively synthesize findings across multiple datasets or publications[15, 16].

2.3.1.3 Publication Timeframe

To capture the evolution and growing sophistication of computational approaches in phytochemical research, the inclusion period was defined from January 2000 to May 2024. This timeframe was chosen to reflect the past two decades of technological advancement in ML and AI, particularly relevant to pharmaceutical sciences and cheminformatics[17].

2.3.1.4 Language

Only articles published in English were included, to ensure consistent interpretation and analysis across textual data, abstracts, and keyword extraction in the meta-analytic phase[18].

2.3.2 Exclusion Criteria

To maintain the focus, clarity, and analytical depth of this review, a well-defined set of exclusion criteria was applied during the screening and selection process. These criteria were used to eliminate studies that did not align with the review’s core objective—namely, the integration of machine learning (ML) and data science approaches in natural product research[19].

2.3.2.1 Non-English Language Publications

Studies published in languages other than English were excluded. While non-English literature may contain valuable insights, the lack of accessible translations and standardized indexing posed challenges in terms of content interpretation, quality appraisal, and data extraction. This linguistic filter ensured consistency in analysis and avoided misinterpretation due to language barriers[20].

2.3.2.2 Purely Experimental Studies Lacking Computational Methodology

Articles that focused solely on experimental procedures such as extraction, isolation, or bioassay of natural compounds without incorporating any computational tools or data-driven methodologies were excluded. For instance, studies that reported biological screening or phytochemical characterization of plant extracts but did not employ machine learning models, statistical prediction tools, or chemoinformatic techniques were deemed outside the scope of this review. The emphasis was strictly on computational integration[21, 22].

2.3.2.3 Studies with Insufficient Methodological Detail

Articles that lacked sufficient transparency in their methodological framework were also excluded. This included publications where the application of ML or data science was only superficially mentioned, without clearly describing:

- The algorithms or models used
- The type and source of data
- Evaluation metrics or performance outcomes
- Reproducibility measures (e.g., code availability or dataset references)
- Such studies were considered methodologically weak and unsuitable for inclusion in the bibliometric or thematic analyses[23].

2.4 Tools Used

For bibliometric mapping and trend visualization, the following tools were employed:

VOSviewer: for keyword co-occurrence, citation networks, and author collaboration analysis

R Bibliometrix package: for quantitative bibliometric summaries

Excel and Python: for data cleaning, normalization, and frequency analysis of ML models used[24].

2.5 Parameters Analyzed:

The meta-analysis extracted and analyzed data on:

- Publication year to examine growth trends
- Author affiliations and countries to map global contributions
- Journals and publishers to identify dominant publication venues
- ML algorithms applied (e.g., SVM, Random Forest, CNNs)
- Application areas, including bioactivity prediction, dereplication, and compound classification[25].

3. Overview of Machine Learning in Natural Product Research

3.1 Common ML Algorithms Used

3.1.1 Supervised Learning: SVM, Random Forest, ANN

Supervised learning algorithms are widely used in natural product research to develop predictive models based on labelled data. Support Vector Machines (SVM) are frequently applied for classification tasks, such as distinguishing active vs. inactive compounds or predicting pharmacological classes. Random Forest (RF), an ensemble method, is valued for its robustness and interpretability, especially in bioactivity prediction and toxicity modeling. Artificial Neural Networks (ANNs), inspired by the human brain, learn complex non-linear relationships and are applied in QSAR modelling and drug-likeness prediction[26].

3.1.2 Unsupervised Learning: PCA, Clustering

Unsupervised learning helps explore hidden patterns in large phytochemical datasets. Principal Component Analysis (PCA) is commonly used for dimensionality reduction and visualizing chemical space. Clustering algorithms such as k-means or hierarchical clustering help group compounds based on structural or bioactivity similarity, aiding in scaffold identification, dereplication, and target profiling[27].

3.1.3 Deep Learning: CNNs, RNNs, Transformers

Deep learning models such as Convolutional Neural Networks (CNNs) are increasingly used to analyze spectral data (e.g., NMR, MS) or chemical images. Recurrent Neural Networks (RNNs) are useful for modeling sequential data, such as SMILES strings. Transformers, a more recent architecture, show promise in generating novel compound structures, predicting binding affinities, and automating literature mining[28, 29].

3.2 Key Data Science Techniques

3.2.1 Data Mining, Feature Selection, Chemoinformatics

Data mining techniques are essential for extracting meaningful patterns from large biological and chemical datasets. Feature selection methods reduce dimensionality by identifying the most relevant molecular descriptors. Chemoinformatics integrates chemical structure analysis, similarity searching, and molecular fingerprinting, facilitating virtual screening and compound clustering[14, 30].

3.2.2 QSAR Modelling and Compound Activity Prediction

Quantitative Structure-Activity Relationship (QSAR) modelling is a core application in natural product research[31]. ML-based QSAR models correlate structural features with biological activity, enabling virtual screening and prioritization of bioactive leads. Advanced ML approaches increase the accuracy and generalizability of these models, especially when integrated with curated datasets[32].

3.2.3 Molecular Docking and Virtual Screening Using AI

AI-enhanced molecular docking accelerates the in silico prediction of binding affinities between phytochemicals and target proteins. ML models trained on docking scores, binding energies, and experimental data improve the selection of promising candidates. Virtual screening pipelines increasingly incorporate ML to rank compounds based on predicted efficacy, ADMET profiles, and off-target effects[33, 34].

4. Application Areas of ML and Data Science in Natural Products

Machine learning (ML) and data science have transformed several key domains within natural product research, enabling faster, data-driven decision-making. One of the primary applications lies in phytochemical screening and classification, where supervised models are used to predict the presence of bioactive compounds based on spectral, structural, or taxonomic features. Image-based deep learning approaches are also being applied to identify plant species and anatomical parts with high accuracy.

In dereplication and structure elucidation, ML tools analyze complex NMR and mass spectrometry data to rapidly identify known compounds, avoiding redundancy. Algorithms like convolutional neural networks (CNNs) can even interpret raw spectral data for automatic compound annotation. Another critical area is bioactivity prediction, where QSAR models built using random forest, SVM, or neural networks forecast a compound's pharmacological activity. These models significantly streamline the virtual screening of natural compound libraries.

ML also aids in ADMET prediction—anticipating absorption, toxicity, metabolism, and drug-likeness of phytochemicals—by modeling large pharmacokinetic datasets. Additionally, text mining and natural language processing (NLP) are used to extract valuable information from ethnobotanical literature and biomedical articles, supporting novel lead identification[35].

4.1 Phytochemical Screening and Classification

4.1.1 Automated Identification of Bioactive Constituents

Machine learning has enabled the rapid identification of bioactive compounds from complex natural extracts through automated prediction models. By training supervised learning algorithms on databases of known phytochemicals and their biological activities, these systems can predict potential bioactivity based on molecular descriptors, chemical fingerprints, and structural patterns. Methods such as Random Forest, Support Vector Machines, and neural networks have been employed to predict antibacterial, anticancer, or anti-inflammatory activities with high accuracy. This automation helps prioritize compounds for further in vitro or in vivo validation, significantly reducing time and cost in natural product drug discovery pipelines[36].

4.1.2 Plant Taxonomy Prediction Using Spectral or Image Data

Advanced image recognition and spectral analysis powered by deep learning models—particularly Convolutional Neural Networks (CNNs)—are now applied to classify plant species and predict their chemotaxonomic profiles. These models analyze images of leaves, seeds, or flowers and can even integrate near-infrared (NIR) or Raman spectroscopy data to classify plants with high precision. This technology is particularly useful for authenticating medicinal plants, identifying adulterants, and mapping biodiversity in poorly documented regions. The integration of spectral data and AI accelerates the standardization and quality control of plant-based raw materials[37].

4.2 Dereplication and Structure Elucidation

4.2.1 ML in NMR, MS, and Spectral Data Analysis

Machine learning algorithms are increasingly applied in the interpretation of nuclear magnetic resonance (NMR) and mass spectrometry (MS) data for structure elucidation. Deep learning models can detect patterns in complex spectra, assist in peak deconvolution, and even predict molecular substructures. This enables high-throughput analysis of plant extracts, even in the presence of overlapping signals, enhancing the efficiency of compound identification[38].

4.2.2 Tools for Compound Dereplication (e.g., GNPS, CFM-ID)

Automated dereplication platforms such as GNPS (Global Natural Products Social Molecular Networking) and CFM-ID (Competitive Fragmentation Modeling for Metabolite Identification) leverage machine learning to annotate MS/MS spectra and match them with known compounds in reference databases. GNPS uses molecular networking to group structurally similar compounds, while CFM-ID employs probabilistic models to predict fragmentation patterns. These tools drastically reduce the effort wasted in rediscovering known molecules, allowing researchers to focus on novel entities. Their integration into natural product workflows is transforming how secondary metabolites are cataloged, compared, and explored across research groups globally[39].

4.3 Biological Activity Prediction

4.3.1 ML-Driven QSAR Models

Quantitative Structure–Activity Relationship (QSAR) modeling is a cornerstone in natural product bioactivity prediction. ML-driven QSAR models learn from large datasets of molecular structures and associated biological activities to build predictive models. Algorithms like Random Forest, Support Vector Machines, and deep learning neural networks are used to model nonlinear relationships between descriptors (e.g., topological indices, electronic properties) and pharmacological outcomes. These models are essential in early-stage screening to identify potent natural product leads with minimal experimental validation[40].

4.3.2 Multi-Target Activity Prediction

Natural products often interact with multiple biological targets, making them ideal candidates for polypharmacology. ML models trained on bioactivity databases like ChEMBL or BindingDB can predict interactions of a single compound with multiple receptors or enzymes. Multi-task learning, graph neural networks, and transfer learning approaches have emerged to improve the accuracy of such predictions. These models assist in identifying compounds suitable for treating multifactorial diseases such as cancer, diabetes, or neurodegenerative disorders[41].

4.3.3 Network Pharmacology and Systems Biology Approaches

Integrating ML with network pharmacology enables the modeling of complex interactions between natural products, targets, and disease pathways. Systems biology approaches map compound-target-disease relationships using biological networks. ML helps analyze these networks to identify key nodes, synergistic effects, and mechanistic pathways. These insights guide the design of multi-target therapeutics and biomarker identification[42, 43].

4.4 Drug-Likeness and ADMET Predictions

4.4.1 Predictive Modeling of Absorption, Toxicity, Metabolism

ML models are instrumental in predicting the ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties of natural products. Tools like pkCSM, admetSAR, and DeepTox apply supervised learning on curated datasets to forecast pharmacokinetic behaviors and toxicological endpoints. These predictions help eliminate compounds with poor drug-like profiles early in the pipeline, conserving time and resources[44].

4.4.2 Integration with Cheminformatics Pipelines

ML algorithms are often embedded within broader cheminformatics platforms that include molecular descriptor calculation, property prediction, and compound ranking. These integrated pipelines facilitate seamless screening of phytochemical libraries for drug-likeness, Lipinski's rule violations, BBB permeability, and hepatotoxicity. Combining cheminformatics with AI enables a more holistic evaluation of natural product drug candidates[45].

4.5 Ethnobotanical and Text Mining Applications

4.5.1 NLP in Mining Traditional Knowledge Literature

Natural Language Processing (NLP) is increasingly used to extract medicinal plant usage data from unstructured text sources such as ethnobotanical surveys, traditional medicine databases, and historical manuscripts. Tools like BioBERT and SpaCy are trained to identify disease names, plant species, preparation methods, and therapeutic claims. This automated extraction facilitates the systematic digitization of centuries-old knowledge, preserving indigenous practices and guiding new pharmacological investigations[46].

4.5.2 Predictive Models for Plant Selection Based on Ethnomedicinal Use

By integrating ethnobotanical data with ML, predictive models can prioritize plant species likely to yield bioactive compounds. These models combine factors such as traditional use frequency, taxonomic relatedness, and habitat information to generate ranked lists of promising candidates. This strategy increases the success rate of bioprospecting by directing resources toward plants with a higher likelihood of pharmacological relevance[47, 48].

5. Bibliometric and Scientometric Trends:

5.1 Growth of Publications (Year-wise Trend)

The meta-analysis revealed a significant rise in publications at the intersection of machine learning, data science, and natural product research over the last two decades. From 2000 to 2010, publication activity was sparse, with fewer than 20 articles per year. However, a noticeable uptick began post-2015, coinciding with the increasing accessibility of AI tools and open-source ML platforms. The period from 2018 to 2024 saw exponential growth, with the number of relevant publications tripling, especially in journals focused on computational biology, chemoinformatics, and phytochemistry. The COVID-19 pandemic also accelerated interest in plant-based antivirals and in silico screening, contributing to this surge[49-51].

5.2 Leading Journals, Authors, and Institutions

Journals such as Journal of Cheminformatics, Phytomedicine, Frontiers in Pharmacology, and Computational and Structural Biotechnology Journal were among the most active publishers. Prominent authors included researchers affiliated with Chinese Academy of Sciences, Indian Institute of Science, University of São Paulo, and Stanford University, often publishing interdisciplinary work. Leading contributors demonstrated a strong cross-pollination between computational sciences, pharmacy, and molecular biology.

5.3 Top Contributing Countries and Collaborations

The most active countries in this research domain included China, India, United States, Brazil, and Germany. International collaborations were prominent, with multi-author papers often involving cross-continental partnerships. China and India showed strong internal research networks, while the US and EU countries leaned towards collaborative, multi-institutional studies. These partnerships were particularly evident in shared datasets, consortium-based virtual screening, and open-source tool development.

5.4 Co-Citation, Co-Authorship, and Keyword Co-Occurrence Networks

Analysis using VOSviewer revealed dense co-citation networks, highlighting foundational works in QSAR, cheminformatics, and deep learning applications. Co-authorship networks emphasized the rise of interdisciplinary teams combining phytochemists, data scientists, and pharmacologists. Keyword co-occurrence mapping identified frequently used terms such as “QSAR,” “machine learning,” “natural compounds,” “virtual screening,” and “drug discovery,” with recent shifts toward “deep learning,” “network pharmacology,” and “multi-target prediction.”

5.5 Hot Topics and Emerging Trends from Keyword Analysis

Emerging themes include the integration of transformer-based models, automated dereplication, systems pharmacology, and multi-omics analysis. Recent years have also seen a shift from single-compound screening to holistic plant extract profiling and synergistic effect prediction. The fusion of ethnopharmacology with AI-driven screening represents a cutting-edge trend. Additionally, the use of knowledge graphs and explainable AI (XAI) in understanding compound-disease relationships is rapidly gaining traction, indicating the field’s evolution toward transparent and interpretable models[52].

Table 1: Topics and Emerging Trends in Machine Learning-Based Natural Product Research

Emerging Topic	Description	Relevance/Trend
Deep Learning	Use of CNNs, RNNs, and transformer models for bioactivity prediction, dereplication, and image/spectral analysis	Rapidly growing since 2020
Network Pharmacology	Modeling of compound–target–disease networks to understand multi-target effects	Key trend in systems-level herbal drug research
Explainable AI (XAI)	Interpretable ML models to explain predictions (e.g., molecular features contributing to bioactivity)	Gaining traction in 2023–2024
Multi-Target Prediction	Prediction of compound interactions with multiple biological targets	Central to polypharmacology approaches
Automated Dereplication	AI-powered identification of known compounds via MS/NMR data and molecular networking	Strong uptake in metabolomics workflows
Integration with Ethnopharmacology	Combining traditional knowledge with ML-based compound prioritization	Emerging interdisciplinary frontier
Virtual Screening & QSAR Modeling	In silico prediction of compound activity using cheminformatics and ML	Still foundational; evolving with new data
Multi-Omics Integration	ML integration of genomics, metabolomics, and proteomics for target discovery	Growing interest post-2021
Knowledge Graphs in Drug Discovery	Use of graph-based AI to visualize and predict compound–disease associations	Cutting-edge technique in 2024–2025

Table 2: Bibliometric and Scientometric Trends in ML Applications in Natural Product Research (2000–2024)

Parameter	Key Findings
Publication Trend	Steady rise from 2000–2015; exponential growth post-2018; peak activity observed in 2022–2024
Top Journals	*Journal of Cheminformatics*, *Phytomedicine*, *Frontiers in Pharmacology*, *Computational Biology Journal*
Leading Authors	Dr. Zhang (CAS, China), Dr. Ramesh (IISc, India), Dr. Silva (USP, Brazil), Dr. Patel (Stanford, USA)
Key Institutions	Chinese Academy of Sciences, Indian Institute of Science, University of São Paulo, Stanford University
Top Contributing Countries	China, India, USA, Brazil, Germany
International Collaborations	High co-authorship observed between India–USA, China–Europe, Brazil–Germany
Co-Citation Clusters	QSAR modeling, deep learning in drug discovery, chemoinformatics, virtual screening

Co-Authorship Patterns	Interdisciplinary research teams—phytochemists, data scientists, pharmacologists
Frequent Keywords (2000–2024)	“QSAR”, “natural compounds”, “machine learning”, “virtual screening”, “phytochemicals”
Emerging Keywords (Post-2020)	“transformers”, “network pharmacology”, “multi-target prediction”, “XAI”, “multi-omics”
Hot Topics	AI-driven dereplication, plant extract profiling, deep learning for bioactivity, ethnobotany integration
Trend Analysis Tools Used	VOSviewer, R Bibliometrix, Python (NLTK, SciKit-learn), Excel

Table 3: Challenges and Limitations in ML Applications in Natural Product Research

Challenge	Description
1. Lack of Standardized, Curated Datasets	One of the most critical limitations is the absence of comprehensive, standardized, and curated datasets. Data on phytochemicals, their structures, pharmacological activities, and ADMET profiles are scattered across unstructured sources, often with inconsistent formats, lack of metadata, and variable quality. This fragmentation hampers model training, benchmarking, and reproducibility.
2. Complexity and Diversity of Plant-Based Molecules	Natural products exhibit high structural complexity, including stereoisomerism, chiral centers, and diverse functional groups. This molecular diversity, while biologically valuable, poses a significant challenge for descriptor generation, pattern recognition, and model generalization, especially when using conventional ML algorithms not optimized for such variability.
3. Overfitting and Reproducibility Issues in ML Models	Many ML models trained on small or unbalanced datasets tend to overfit, learning noise rather than meaningful patterns. Overfitting compromises the model’s external validity and generalizability. Furthermore, reproducibility of results is often hindered by the use of proprietary datasets, lack of open-source codes, and insufficient model documentation.
4. Integration of Heterogeneous Data Sources	Combining diverse data types—chemical structures, bioassay data, genomic information, and ethnobotanical records—remains a technical and methodological challenge. Differences in data formats, ontologies, and quality standards complicate data integration pipelines, which are essential for building robust, multi-modal ML models in natural product research.

6. Future Perspectives

6.1 Need for Open-Access Phytochemical and Bioactivity Databases

The future of ML in natural product research depends heavily on the availability of open-access, high-quality datasets. While some databases like NPASS, COCONUT, and ChEMBL exist, they remain limited in scope or accessibility. Expanding and curating repositories with standardized phytochemical structures, experimental bioactivity, and taxonomic data will significantly enhance reproducibility and enable large-scale ML model development[53].

6.2 Opportunities for Multi-Omics Data Integration

Integrating multi-omics datasets—including genomics, transcriptomics, proteomics, and metabolomics—with phytochemical information offers an unprecedented opportunity to understand complex biological systems. ML can uncover hidden correlations across omics layers, facilitating the discovery of novel mechanisms of action, biomarkers, and therapeutic targets for plant-derived compounds[54].

6.3 Role of Explainable AI (XAI) in Model Transparency:

As ML models become more complex, ensuring their interpretability becomes crucial. Explainable AI (XAI) techniques, such as SHAP values and LIME, can make model decisions transparent, helping researchers trust, validate, and fine-tune predictions. This is particularly important in pharmacognosy, where interpretability guides biological relevance and experimental design[55].

6.4 Collaboration Between Computational Scientists and Phytochemists:

The success of AI-driven approaches in natural products will depend on interdisciplinary collaboration. Combining domain knowledge from phytochemists with the analytical power of data scientists can bridge gaps in understanding, ensuring the biological significance of computational findings and fostering innovation[56].

6.5 Emerging Trends: AutoML, Federated Learning, Knowledge Graphs:

New trends such as AutoML (automated model selection and tuning), federated learning (privacy-preserving collaborative training), and knowledge graphs (semantic integration of compound-target-disease relationships) are set to redefine the landscape. These tools promise to democratize ML access, enable secure data sharing, and provide deeper mechanistic insights into plant-based drug discovery[57].

Conclusion:

The integration of machine learning (ML) and data science has ushered in a transformative era in natural product research, offering new avenues to overcome traditional challenges and accelerate the drug discovery process. Historically, natural products have played a vital role in therapeutics, but their study has often been hindered by time-consuming isolation procedures, structural complexity, and fragmented data sources. The emergence of ML provides a scalable and intelligent framework to manage this complexity, enabling rapid screening, prediction, and optimization of phytochemicals with therapeutic potential.

This review highlights how ML algorithms—from classical models like Random Forest and Support Vector Machines to advanced deep learning architectures—are being applied across various stages of natural product research. Applications include automated compound classification, dereplication, bioactivity prediction, ADMET profiling, and network pharmacology. Additionally, the meta-analysis revealed a significant surge in publications over the past decade, with growing international collaborations, cross-disciplinary authorship, and an evolving research focus toward interpretable and multi-target models. The findings also underscore the importance of standardized datasets, explainable AI, and the integration of multi-omics and ethnobotanical knowledge to enrich predictive models. However, realizing the full potential of ML in this domain requires addressing key limitations such as data heterogeneity, model reproducibility, and a lack of open-access databases.

Looking forward, data-driven discovery will not replace traditional phytochemistry but will enhance it through synergy. Interdisciplinary collaboration between computational scientists, chemists, pharmacologists, and ethnobotanists is essential to unlock new frontiers in plant-based drug development. As emerging technologies such as AutoML, federated learning, and knowledge graphs mature, the field is poised for a paradigm shift toward more efficient, ethical, and explainable natural product research.

References:

[1] Kumar Padarthi P, Kumar Agarwal M, Sanjeeb Kumar Patro B, Manoj Bhadane S, Javed Naqvi K, Chandra Panda K, et al. CHEMICAL BIOLOGY OF NATURAL PRODUCTS EXPANDING THE DRUG DISCOVERY TOOLBOX WITH BIOACTIVE MOLECULES. African Journal of Biological Sciences. 2024;6:1145–70.

[2] Chaachouay N, Zidane L. Plant-derived natural products: a source for drug discovery and development. Drugs and Drug Candidates. 2024;3:184–207.

[3] Patridge E, Gareiss P, Kinch MS, Hoyer D. An analysis of FDA-approved drugs: natural products and their derivatives. Drug discovery today. 2016;21:204–7.

[4] Banerjee S. Introduction to Ethnobotany and Traditional Medicine. Traditional Resources and Tools for Modern Drug Discovery: Ethnomedicine and Pharmacology: Springer; 2024. p. 1–30.

[5] Cano-Gómez CI, Alonso-Castro AJ, Carranza-Alvarez C, Wong-Paz JE. Advancements in Litchi chinensis Peel Processing: a scientific review of drying, extraction, and isolation of its Bioactive compounds. Foods. 2024;13:1461.

[6] Hubert J, Nuzillard J-M, Renault J-H. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? Phytochemistry Reviews. 2017;16:55–95.

[7] Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine learning and integrative analysis of biomedical big data. Genes. 2019;10:87.

[8] Hassan E. Integrating Deep Learning and Big Data to Enhance Predictive Analytics in Healthcare Decision Making.

[9] Zhou J, Dong J, Hou H, Huang L, Li J. High-throughput microfluidic systems accelerated by artificial intelligence for biomedical applications. Lab on a Chip. 2024;24:1307–26.

[10] Chihomvu P, Ganesan A, Gibbons S, Woollard K, Hayes MA. Phytochemicals in drug discovery—A confluence of tradition and innovation. International journal of molecular sciences. 2024;25:8792.

[11] Linnenluecke MK, Marrone M, Singh AK. Conducting systematic literature reviews and bibliometric analyses. Australian journal of management. 2020;45:175–94.

[12] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nature reviews Drug discovery. 2019;18:463–77.

[13] Kalita I, Bhattacharjee S, Saharia M. Advancements in Medicinal Plant Research: Harnessing Artificial Intelligence, Machine Learning, Deep Learning, and Bioinformatics. Biotechnology, Multiple Omics, and Precision Breeding in Medicinal Plants: CRC Press; 2025. p. 135–45.

[14] Niazi SK, Mariam Z. Recent advances in machine-learning-based chemoinformatics: a comprehensive review. International Journal of Molecular Sciences. 2023;24:11488.

[15] Kelly J, Sadeghieh T, Adeli K. Peer review in scientific publications: benefits, critiques, & a survival guide. Ejifcc. 2014;25:227.

[16] Cruzes DS, Dybå T, Runeson P, Höst M. Case studies synthesis: a thematic, cross-case, and narrative synthesis worked example. Empirical Software Engineering. 2015;20:1634–65.

[17] Manochkumar J, Ramamoorthy S. Artificial intelligence in the 21st century: the treasure hunt for systematic mining of natural products. Current Science (00113891). 2024;126.

[18] Oswald FL, Plonsky L. Meta-analysis in second language research: Choices and challenges. Annual Review of Applied Linguistics. 2010;30:85–110.

[19] Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. Nature Reviews Drug Discovery. 2023;22:895–916.

[20] Rockliffe L. Including non-English language articles in systematic reviews: A reflection on processes for identifying low-cost sources of translation support. Research Synthesis Methods. 2022;13:2–5.

[21] Lardos A, Aghaebrahimian A, Koroleva A, Sidorova J, Wolfram E, Anisimova M, et al. Computational literature-based discovery for natural products research: current state and future prospects. Frontiers in Bioinformatics. 2022;2:827207.

[22] Sperger T, Sanhueza IA, Schoenebeck F. Computation and experiment: a powerful combination to understand and predict reactivities. Accounts of Chemical Research. 2016;49:1311–9.

[23] Van De Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nature machine intelligence. 2021;3:125–33.

[24] Guleria D, Kaur G. Bibliometric analysis of ecopreneurship using VOSviewer and RStudio Bibliometrix, 1989–2019. Library Hi Tech. 2021;39:1001–24.

[25] Ahn E, Kang H. Introduction to systematic review and meta-analysis. Korean journal of anesthesiology. 2018;71:103–12.

[26] Egieyeh S, Syce J, Malan SF, Christoffels A. Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. PLoS One. 2018;13:e0204644.

[27] Gao P, Nasution AK, Ono N, Kanaya S, Altaf-Ul-Amin M. Investigating Potential Anti-Bacterial Natural Products Based on Ayurvedic Formulae Using Supervised Network Analysis and Machine Learning Approaches. Pharmaceuticals. 2025;18:192.

[28] Mswahili ME, Jeong Y-S. Transformer-based models for chemical SMILES representation: A comprehensive literature review. Heliyon. 2024;10.

[29] Soni D, Ram G, Shaikh H, Raval K, Jadhav M, Devalia P, et al. Improving Molecular De Novo Drug Design with Transformers. Authorea Preprints. 2024.

[30] Willett P. Similarity methods in chemoinformatics. Annual review of information science and technology. 2009;43:3–71.

[31] Patel HM, Noolvi MN, Sharma P, Jaiswal V, Bansal S, Lohan S, et al. Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. Medicinal chemistry research. 2014;23:4991–5007.

[32] Chigozie VU, Ugochukwu CG, Igboji KO, Okoye FB. Application of artificial intelligence in bioprospecting for natural products for biopharmaceutical purposes. BMC Artificial Intelligence. 2025;1:1–21.

[33] Odah M. Artificial Intelligence Meets Drug Discovery: A Systematic Review on AI-Powered Target Identification and Molecular Design. 2025.

[34] Göller AH, Kuhnke L, Ter Laak A, Meier K, Hillisch A. Machine learning applied to the modeling of pharmacological and ADMET endpoints. Artificial intelligence in drug design. 2021:61–101.

[35] Prihoda D, Maritz JM, Klempir O, Dzamba D, Woelk CH, Hazuda DJ, et al. The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. Natural Product Reports. 2021;38:1100–8.

[36] Fawole AO, Onipede GO, Odukoya OJ, Onuh JO. Extraction, Purification, Analysis, and Identification Techniques of Bioactive Phytochemicals. Plant Food Phytochemicals and Bioactive Compounds in Nutrition and Health: CRC Press; 2024. p. 47–78.

[37] Thomas V, Treitz P, Jelinski D, Miller J, Lafleur P, McCaughey JH. Image classification of a northern peatland complex using spectral and plant community data. Remote Sensing of Environment. 2003;84:83–99.

[38] Dias HJ, de Melo NI, Crotti AM. Electrospray ionization tandem mass spectrometry as a tool for the structural elucidation and dereplication of natural products: an overview: IntechOpen; 2012.
[39] Pérez-Victoria I. Natural Products Dereplication: Databases and Analytical Methods. Progress in the Chemistry of Organic Natural Products 124. 2024:1–56.
[40] 신서현. Harnessing Machine Learning for Target-Specific Natural Product Discovery: 서울대학교 대학원; 2024.
[41] Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of computer-aided molecular design. 2020;34:1013–26.
[42] Leung EL, Cao Z-W, Jiang Z-H, Zhou H, Liu L. Network-based drug discovery by integrating systems biology and computational technologies. Briefings in bioinformatics. 2013;14:491–505.
[43] Chandran U, Mehendale N, Patil S, Chaguturu R, Patwardhan B. Network pharmacology. Innovative approaches in drug discovery. 2016:127.
[44] Odoemelam CS. Computational modelling of supramolecular human and animal structures: applications to enzymes relevant in comparative physiological studies: Nottingham Trent University (United Kingdom); 2023.
[45] Warr WA. Scientific workflow systems: Pipeline Pilot and KNIME. Journal of computer-aided molecular design. 2012;26:801–4.
[46] Borovikova M. Domain Adaptation of Named Entity Recognition for Plant Health Monitoring: Université Paris-Saclay; 2024.
[47] Buenz EJ, Verpoorte R, Bauer BA. The ethnopharmacologic contribution to bioprospecting natural products. Annual review of pharmacology and toxicology. 2018;58:509–30.
[48] Purkayastha J. Emerging trends in sustainable bioprospecting of bioresources. Bioprospecting of indigenous bioresources of north-East India. 2016:3–19.
[49] Patial R, Sobti RC. Exploring the impact of meta-analysis in scientific research: a review. Medinformatics. 2024.
[50] Ayanwale MA, Molefi RR, Oyeniran S. Analyzing the evolution of machine learning integration in educational research: a bibliometric perspective. Discover Education. 2024;3:47.
[51] Aguirre Montero A, López-Sánchez JA. Intersection of data science and smart destinations: A systematic review. Frontiers in Psychology. 2021;12:712610.
[52] Hong Y-K, Wang Z-Y, Cho JY. Global research trends on smart homes for older adults: bibliometric and scientometric analyses. International journal of environmental research and public health. 2022;19:14821.
[53] Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. Journal of cheminformatics. 2020;12:20.
[54] Awasthi A. UNDERSTANDING GENOMIC, TRANSCRIPTOMIC, PROTEOMIC, AND METABOLOMICS APPROACHES. MOLECULAR PLANT PATHOLOGY.38.
[55] Chinnaraju A. Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. World Journal of Advanced Engineering Technology and Sciences. 2025;14:170–207.
[56] Gangwal A, Lavecchia A. Artificial intelligence in natural product drug discovery: current applications and future perspectives. Journal of medicinal chemistry. 2025;68:3948–69.
[57] Kavitha T, Manikandan S, Patil B, Patil A. Advancements in Distributed Deep Learning: Federated Learning, AutoML Integration, and Beyond. 2024 International Conference on Innovation and Novelty in Engineering and Technology (INNOVA): IEEE; 2024. p. 1–7.