Natural Products in AI-Driven Drug Discovery: A Bibliometric and Thematic Review of Computational Methods, Disease Areas, and Emerging Frontiers

João Silva[1]*, Pedro Costa[1], Ana Beatriz[2], Rafael Mendes[3]

[1]Department of AI for Phytopharmacology and Amazonian Biodiversity, Faculty of Pharmacy, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
[2]Department of Computational Pharmacology and ADME Prediction, Faculty of Pharmaceutical Sciences, University of Campinas, Campinas, Brazil.
[3]Department of Natural Product Drug Discovery, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil.

ABSTRACT

Natural products remain a major source of chemically diverse scaffolds for drug discovery, while artificial intelligence has created new opportunities to mine, represent, prioritize, and interpret complex natural product chemical space. This original bibliometric and thematic review examines how AI-driven natural product drug discovery is organized as a research field, with attention to computational methods, disease-area priorities, data infrastructures, emerging frontiers, and translational barriers. The review applies bibliometric logic by separating search strategy, dataset construction, eligibility criteria, metadata cleaning, keyword normalization, research-cluster interpretation, and thematic synthesis. Because verified database-export counts were not available for direct computation in the present execution, numerical bibliometric indicators, ranking metrics, keyword frequencies, and cluster-size claims are not reported. Instead, the analysis maps the approved evidence base through structured charting and cautious thematic interpretation. The review focuses on natural products, phytochemicals, medicinal plants, microbial and marine natural products where supported, natural product databases, machine learning, deep learning, molecular representation learning, network pharmacology, virtual screening, target prediction, ADMET and toxicity prediction, and natural product-inspired molecular design. Disease-area mapping is interpreted cautiously across cancer, infectious disease, inflammatory, metabolic, neurodegenerative, antimicrobial, and antiviral discovery themes where the mapped evidence supports discussion. The synthesis identifies a field shaped by database quality, molecular-standardization needs, representation learning, target and mechanism inference, screening prioritization, explainability, generative design, and validation constraints. The review contributes a research agenda for more reproducible, interpretable, experimentally validated, and translationally responsible AI-enabled natural product discovery.

Key Words: *Natural products, Artificial intelligence, Drug discovery, Bibliometric review, Thematic review, Machine learning*

eIJPPR 2025; 14(2): 12-22

HOW TO CITE THIS ARTICLE: Silva J, Costa P, Beatriz A, Mendes R. Natural Products in AI-Driven Drug Discovery: A Bibliometric and Thematic Review of Computational Methods, Disease Areas, and Emerging Frontiers. Int J Pharm Phytopharmacol Res. 2025;15(2): **12-22**. https://doi.org/10.51847/H3b2B00eye

INTRODUCTION

Natural products have long provided pharmacologically productive chemical scaffolds for drug discovery, and contemporary AI-enabled approaches have renewed attention to how these molecules can be mined, represented, prioritized, and interpreted within computational discovery pipelines [1–3]. Their chemical

Corresponding author: João Silva
Address: Department of AI for Phytopharmacology and Amazonian Biodiversity, Faculty of Pharmacy, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
E-mail: ✉ joao.silva@ufrj.br
Received: 24 November 2024; **Revised:** 28 February 2025; **Accepted:** 07 March 2025

diversity, stereochemical richness, biosynthetic origins, and biological relevance make them attractive but technically challenging substrates for computational modeling. AI-driven natural product discovery therefore sits at the intersection of medicinal chemistry, natural product chemistry, cheminformatics, bioinformatics, pharmacology, and data science.

Machine learning has become increasingly visible across drug discovery and development, including molecular-property prediction, target identification, virtual screening, toxicity estimation, and prioritization of candidate compounds [4]. In the natural product context, these tasks are complicated by heterogeneous structural data, inconsistent metadata, difficult dereplication, limited negative examples, assay variability, and uneven biological annotation. As a result, AI should be understood as a prioritization and hypothesis-generation framework rather than a substitute for pharmacological experimentation.

A bibliometric and thematic review is especially appropriate for this field because AI-driven natural product discovery is not a single method, disease area, or model family. It includes natural product databases, molecular fingerprints, SMILES strings, molecular graphs, deep learning classifiers, target prediction tools, drug–target interaction models, network pharmacology, virtual screening, ADMET prediction, toxicity modeling, and emerging generative design approaches. Mapping the field therefore requires attention to knowledge structures as well as thematic interpretation.

The aim of this review is to examine how AI-driven natural product drug discovery has evolved as a research field, which computational methods and research clusters are most visible in the mapped literature, which disease areas receive thematic attention, and which emerging frontiers and limitations shape future research. The review asks how natural product data sources, molecular representation methods, target and mechanism prediction, virtual screening, network pharmacology, ADMET modeling, and validation constraints interact in the current research landscape.

Bibliometric review rationale

Bibliometric review methods are useful when a research field is heterogeneous, rapidly expanding, and distributed across multiple disciplinary communities. In such contexts, performance analysis, science mapping, keyword normalization, co-occurrence analysis, and thematic mapping can help clarify how publications, methods, and concepts are organized, provided that bibliometric visibility is not equated with scientific quality or translational success [5–7]. This distinction is crucial for AI-driven natural product discovery because highly visible computational terms may reflect methodological popularity rather than validated pharmacological value.

A bibliometric approach can identify publication patterns, source distribution, recurring keywords, method families, and conceptual clusters. However, bibliometric mapping alone cannot determine whether a predicted target is experimentally validated, whether a virtual-screening hit has biological activity, or whether an ADMET estimate predicts in vivo behavior. For this reason, the present review treats bibliometric mapping as a field-structure tool and thematic synthesis as the interpretive layer that connects methods, data sources, tasks, disease priorities, and translational barriers.

Thematic synthesis complements bibliometric mapping by examining what the mapped literature is actually about. In AI-driven natural product discovery, this means distinguishing database infrastructure from model development, molecular representation from biological interpretation, docking-linked prioritization from target validation, and computational ADMET screening from experimental safety assessment. It also means examining how disease areas are framed: as areas of computational interest, not as domains where AI-derived natural product candidates have necessarily achieved clinical validation.

The rationale for combining bibliometric and thematic review logic is therefore methodological as well as substantive. Bibliometric mapping can show how the field is structured, while thematic synthesis can explain why certain clusters matter, where evidence remains weak, and what standards are needed for future progress. This dual approach is well suited to a field in which computational enthusiasm must be balanced with chemical curation, biological plausibility, reproducibility, interpretability, and experimental confirmation.

Search strategy and dataset construction

The review protocol separates the bibliometric dataset from the fixed manuscript reference set. A full bibliometric dataset would require direct export from scholarly databases, including complete metadata for titles, abstracts, authors, journals, keywords, affiliations, references, and DOI records. Because verified database-export files were not available for computation in the present execution, this manuscript does not report annual publication counts, duplicate counts, keyword frequencies, citation metrics, author rankings, country rankings, journal rankings, co-occurrence values, or cluster sizes.

The planned search strategy uses multidisciplinary citation databases, biomedical indexing, and publisher platforms to capture peer-reviewed journal literature on natural products in AI-driven drug discovery. The search logic combines natural product terms with artificial intelligence, machine learning, deep learning, graph models, generative

models, cheminformatics, bioinformatics, network pharmacology, virtual screening, target prediction, ADMET, toxicity, drug repurposing, and drug design terms. Bibliometric software selection and reporting should be explicit, because different tools support different metadata-cleaning, mapping, visualization, and export functions [8].

Eligibility criteria prioritize peer-reviewed journal articles focused on natural products, phytochemicals, medicinal plants, marine or microbial natural products, or natural product-inspired molecules with clear AI, machine learning, computational drug discovery, cheminformatics, bioinformatics, or network-pharmacology relevance. Bibliometric precedents in AI drug discovery show the value of separating search strategy, metadata handling, screening decisions, and future-agenda interpretation [9]. Exclusion criteria remove conference papers, preprints, book chapters, editorials, patents, websites, generic AI articles without drug-discovery relevance, wet-lab-only pharmacology studies without computational relevance, and records lacking sufficient metadata.

The dataset-construction process includes DOI-based deduplication, title-level normalization, abstract screening, article-type checking, metadata cleaning, author-keyword and index-keyword normalization, and thematic coding. Bibliometric studies of AI drug discovery also illustrate why indicator selection must be reported clearly and interpreted cautiously, particularly when publication trends and research clusters are used to guide field-level conclusions [10]. In the present review, all numerical bibliometric outputs that cannot be verified from actual exported records are intentionally omitted.

Table 1 summarises the bibliometric search, dataset construction, screening, keyword normalization, and thematic mapping protocol used to examine AI-driven natural product drug discovery.

Table 1. Bibliometric Search, Dataset Construction, and Thematic Mapping Protocol for AI-Driven Natural Product Drug Discovery: Databases, Search Terms, Eligibility Criteria, Deduplication, Metadata Fields, Keyword Normalization, Cluster Interpretation, and Reporting Safeguards

Protocol component	Operational definition	Implementation rule	Data field required	Quality-control action	Risk of bias or limitation	Reporting requirement
Database selection	Identification of scholarly sources for peer-reviewed journal literature	Use multidisciplinary, biomedical, and publisher databases; use supplementary checking only for verification	Database name; record source; DOI	Compare metadata across indexing and publisher records	Database coverage varies by discipline, publisher, and indexing policy	State all databases searched and distinguish primary search sources from supplementary checking
Search-string construction	Combination of natural product, AI, computational, and drug-discovery terms	Combine natural product terms with AI, machine learning, cheminformatics, network pharmacology, virtual screening, target prediction, ADMET, toxicity, and drug-design terms	Search string; date searched; field searched	Test broad and narrow strings for topic relevance	Overly broad strings may retrieve generic AI or wet-lab-only records	Report exact strings or search logic without claiming unverified counts
Eligibility screening	Inclusion and exclusion of records by article type, topic, and metadata sufficiency	Include peer-reviewed journal articles with natural product and computational drug-discovery relevance	Title; abstract; article type; DOI; journal	Screen title, abstract, and metadata; inspect full text where needed	Abstracts may not fully describe computational methods or natural product scope	Report inclusion and exclusion rules transparently
Deduplication	Removal of repeated records across databases	Deduplicate by DOI first, then normalized title and metadata inspection	DOI; title; authors; year; journal	Standardize DOI strings, punctuation, capitalization, and spacing	Different databases may record titles or author names inconsistently	Describe deduplication rules and avoid invented duplicate counts
Metadata cleaning	Preparation of bibliographic fields for mapping	Standardize journal names, author names, years, keywords, and DOI fields	Authors; title; journal; year; abstract; keywords; references	Reconcile conflicting metadata using publisher or indexing records	Incomplete metadata can reduce mapping reliability	State which fields were required for inclusion

ISSN (Online) 2249-6084 (Print) 2250-1029

Keyword normalization	Cleaning author and index keywords for thematic mapping	Merge synonyms cautiously; preserve method-specific terms such as graph neural network, target prediction, ADMET, and network pharmacology	Author keywords; index keywords; abstract terms	Remove generic terms and harmonize spelling variants	Excessive merging can erase meaningful conceptual distinctions	Report normalization rules and examples of retained distinctions
Computational-method coding	Classification of AI and computational approaches	Code machine learning, deep learning, molecular representation, graph models, network pharmacology, docking-linked screening, ADMET, toxicity, and generative models where present	Abstract; methods; keywords	Code only methods explicitly reported or clearly described	Reviews may discuss methods broadly without study-level implementation	Avoid claiming method prevalence without verified counts
Disease-area coding	Identification of therapeutic domains discussed in the literature	Code cancer, infectious disease, inflammation, metabolic disease, neurodegenerative disease, antimicrobial, and antiviral themes only where supported	Abstract; keywords; disease terms	Distinguish computational attention from validated efficacy	Disease mentions may not indicate experimental or clinical validation	Present disease areas as thematic priorities, not proven therapeutic impact
Research-cluster interpretation	Interpretation of conceptual groupings in mapped literature	Label clusters using keywords, titles, abstracts, and charted evidence	Keywords; abstracts; thematic codes	Validate labels against record content	Cluster labels may overgeneralize heterogeneous papers	State that clusters indicate knowledge structure, not clinical success
Bibliometric indicators	Planned quantitative descriptors of the field	Use publication patterns, journal distribution, keyword co-occurrence, and mapping outputs only if export data are available	Bibliographic export; cleaned metadata	Verify outputs against raw records	Citation and visibility indicators do not equal quality	Do not report numerical indicators unless generated from actual data
Thematic synthesis	Integration of data sources, methods, tasks, disease areas, barriers, and frontiers	Use charted evidence to synthesize patterns and limitations	Charting table; coded domains	Link each theme to supporting references	Thematic interpretation may be shaped by selected literature	Use cautious language and avoid unsupported generalization
Reporting safeguards	Rules preventing unsupported claims	Do not invent counts, rankings, clusters, docking values, ADMET values, assay results, or clinical outcomes	All reported fields	Cross-check claims against charted evidence	Missing full dataset limits quantitative bibliometric analysis	Explicitly state limitations and separate mapped evidence from inference

Figure 1 presents the bibliometric workflow used to identify, screen, clean, map, and thematically synthesize literature on AI-driven natural product drug discovery.

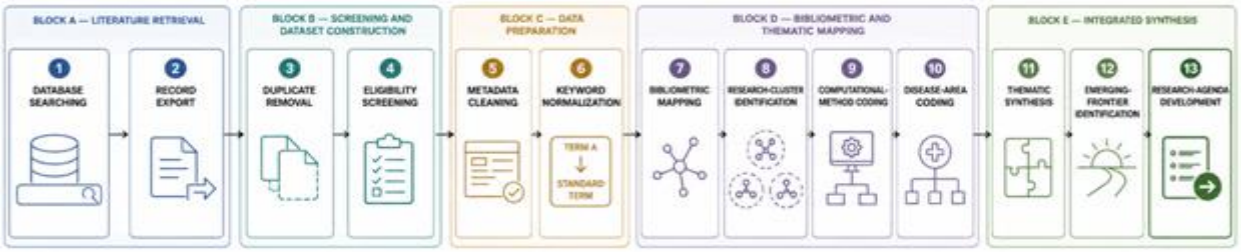

Figure 1. Bibliometric Workflow for AI-Driven Natural Product Drug Discovery

Alt-text description

A bibliometric review workflow diagram showing database searching, record export, duplicate removal, eligibility screening, metadata cleaning, keyword normalization, bibliometric mapping, thematic cluster interpretation, computational-method classification, disease-priority mapping, emerging-frontier identification, and research-agenda development. Count boxes must use only verified numbers from the bibliometric search process.

Publication trends and research clusters

Because verified export-based bibliometric counts were not generated in the present execution, publication trends are interpreted qualitatively rather than numerically. The mapped evidence indicates that data infrastructure is a central knowledge-structure theme, especially through open natural product databases that make molecular records more accessible for computational workflows [11]. This cluster is important because AI-driven discovery depends on searchable, standardized, and sufficiently annotated chemical information before model development or downstream prioritization can be reliable.

A second visible cluster concerns microbial natural product data infrastructure. Curated microbial natural product resources support computational exploration of biosynthetically diverse chemical space and can contribute to discovery prioritization, dereplication, and annotation workflows [12]. In thematic terms, this cluster links natural product chemistry with bioinformatics and cheminformatics, while also highlighting a limitation: database scope influences which chemical families and biosynthetic origins are visible to computational models.

A third cluster centers on molecular representation and structural classification. Deep learning approaches for natural product structural classification suggest that AI can help organize natural product chemical space into interpretable categories when training data and structural annotations are adequate [13]. This cluster is not evidence that classification alone identifies drug candidates; rather, it shows how representation learning can support compound organization, similarity assessment, and downstream prioritization.

A fourth cluster concerns database curation, chemical standardization, and metadata quality. Updated open natural product database resources show that curation is not a peripheral technical detail but a foundation for reliable AI-enabled discovery [14]. Across the mapped literature, publication trends and research clusters should therefore be read as signals of field organization: data infrastructure, microbial natural product discovery, molecular representation, and database curation appear as major thematic anchors, while claims about dominance, impact, or therapeutic success require verified quantitative mapping and experimental evidence.

Computational methods and disease priorities

The mapped literature shows that microbial natural product discovery is one of the clearest areas where AI and computational prioritization intersect with natural product research. AI methods are used to support discovery workflows that involve biosynthetic interpretation, chemical-space exploration, dereplication, structural annotation, and prioritization of candidates for downstream testing [15]. This pattern is particularly important for antimicrobial and infectious-disease discovery, where microbial metabolites remain relevant but experimental search space is large, costly, and difficult to navigate without computational triage.

Machine learning applications in microbial natural product discovery further illustrate the value and limits of computational prioritization. ML can help rank candidate biosynthetic products, connect structural features to predicted activities, and reduce the number of compounds requiring immediate experimental evaluation, but these predictions remain dependent on training-set composition, annotation quality, and validation design [16]. The disease-area implications are therefore best interpreted as research prioritization signals rather than direct evidence of efficacy against infectious or antimicrobial targets.

Beyond microbial natural products, machine learning is also used to infer biological effects, mechanisms, and potential therapeutic relevance of natural products. These approaches can connect phytochemical and natural product structures with predicted targets, pathways, bioactivities, or phenotypic effects, making them relevant to cancer, inflammation, metabolic disease, neurodegenerative disease, and other disease domains where mechanism-based prioritization is needed [17]. However, target prediction and pathway inference should not be treated as target validation, because computational associations require biological confirmation.

Structure-based ML provides an additional bridge between molecular representation and disease-oriented prioritization. By using molecular structures as inputs for biochemical and physiological effect prediction, these approaches can support the organization of natural product chemical space around predicted biological functions [18]. In the mapped literature, this creates a methodological link between molecular fingerprints, graph-like representations, similarity learning, target inference, and disease-area mapping, while reinforcing the need for careful interpretation of predictions derived from heterogeneous natural product data.

Emerging frontiers in AI-natural product discovery

Explainable AI and generative molecular design represent two major frontier directions for AI-enabled natural product discovery, particularly because the field requires both interpretable prioritization and chemically plausible molecular innovation [19, 20]. Explainability is relevant when models are used to support target prediction, toxicity estimation, or lead prioritization, because domain scientists need to understand which structural features, data relationships, or learned representations contribute to model outputs. Generative design is relevant where natural product-inspired scaffolds are used to explore new analogues, but generated molecules must still be evaluated for synthetic feasibility, biological plausibility, novelty, safety, and experimental tractability.

Graph neural network models represent another frontier because they encode molecular structure in a way that can support drug–target affinity prediction and related interaction-learning tasks [21]. For natural product discovery, graph-based methods are conceptually attractive because many natural products contain complex ring systems, stereochemical features, and scaffold architectures that may not be adequately captured by simple descriptors alone. Yet natural-product-specific use of graph models must be assessed cautiously because many graph drug–target datasets are not built primarily from natural product chemical space.

Deep generative models expand the frontier of molecular design by enabling algorithmic exploration of chemical space under selected constraints [22]. In natural product-inspired discovery, these models could support analogue generation, scaffold hopping, fragment recombination, or property-guided exploration when trained and constrained appropriately. Their value depends on the quality of source compounds, the realism of molecular representations, the avoidance of benchmark leakage, and the ability to connect generated structures with synthesis, assay testing, and pharmacological interpretation.

Other emerging frontiers include multimodal learning, knowledge graphs, transformer-based molecular representations where sufficiently supported, open database integration, reproducible workflows, and uncertainty-aware prediction. These areas are especially relevant because AI-driven natural product discovery depends on linking chemical structures with biological targets, pathway annotations, assay results, omics data, pharmacokinetic information, toxicity data, and disease contexts. The field's next phase will likely depend less on isolated model accuracy claims and more on transparent, interpretable, and experimentally grounded workflows that connect prediction to validation.

Thematic synthesis

The thematic synthesis indicates that AI-driven natural product discovery is organized around a central tension: natural products offer chemically rich and biologically meaningful starting points, but the data required to model them are fragmented, unevenly curated, and inconsistently annotated. Reviews of natural product databases show that the availability, structure, standardization, and interoperability of data sources strongly shape what computational models can learn and how reproducible discovery workflows can become [23]. Database quality is therefore not merely a technical background issue; it is a core determinant of model reliability and field maturity.

Virtual screening is another major synthesis domain because it provides a practical route for exploring large natural product libraries and prioritizing candidates for experimental testing. In natural product bioprospecting, virtual screening can connect chemical diversity with target hypotheses, docking workflows, similarity searches, and lead prioritization [24]. However, screening outputs should be interpreted as hypotheses rather than biological proof, because docking scores, ranking outputs, and computational enrichment do not establish pharmacological activity, selectivity, toxicity profile, or therapeutic relevance.

ADMET and toxicity modeling form a complementary prioritization layer in AI-driven natural product discovery. These approaches can help identify compounds with potentially unfavorable absorption, distribution, metabolism, excretion, or toxicity profiles before more resource-intensive testing is pursued [25]. For natural products, this is particularly important because structural complexity, reactive moieties, stereochemistry, and limited annotated pharmacokinetic data can complicate prediction. ADMET models may support triage, but they do not replace experimental pharmacokinetics, toxicology, or formulation studies.

The translational synthesis also requires distinguishing computational discovery from preclinical validation. Machine learning can support preclinical discovery by improving prediction, prioritization, and decision-making, but model outputs must be evaluated against experimental reproducibility, external validation, assay quality, and biological relevance [26]. In AI-natural product research, this distinction is essential because predicted targets, network pharmacology pathways, docking poses, and toxicity estimates often represent early-stage prioritization rather than validated mechanisms or therapeutic outcomes.

Overall, the mapped evidence suggests that the field is moving from isolated computational applications toward more integrated workflows that combine natural product databases, molecular representation learning, target and mechanism inference, virtual screening, ADMET triage, and frontier AI methods. The most productive

interpretation is not that AI replaces natural product chemistry or pharmacology, but that it can help organize chemical space, generate hypotheses, identify candidates for testing, and clarify where experimental resources may be focused. The field's maturity will depend on whether these computational workflows become more transparent, benchmark-aware, interpretable, reproducible, and connected to experimental validation.

Table 2 synthesizes the main computational methods, disease priorities, emerging frontiers, translational barriers, and future research needs in AI-driven natural product discovery.

Table 2. Thematic Synthesis of AI-Driven Natural Product Drug Discovery: Computational Methods, Natural Product Data Sources, Disease Priorities, Discovery Tasks, Emerging Frontiers, Translational Barriers, and Future Research Needs

Synthesis domain	Dominant thematic pattern	Computational method involved	Natural product data source	Disease or therapeutic area	Drug discovery task	Main limitation	Emerging frontier	Future research priority
Natural product data infrastructure	Databases are foundational for AI-ready discovery workflows	Cheminformatics; bioinformatics; database curation	Open natural product databases; microbial natural product databases	Cross-cutting	Data mining; compound retrieval; dereplication; annotation	Fragmented metadata, inconsistent structure standardization, uneven biological annotation	Open curated databases; interoperable data standards	Improve curation, provenance tracking, stereochemical annotation, and reproducible database access
Molecular representation	Structural representation connects natural product chemistry to model prediction	Molecular fingerprints; SMILES; molecular graphs; deep learning classifiers	Natural product structures; phytochemical libraries; curated compound sets	Cross-cutting	Classification; similarity assessment; bioactivity prediction	Complex scaffolds, stereochemistry, and scarce annotations may limit model generalizability	Graph learning; multimodal representation; transformer-based molecular encoders where supported	Develop natural-product-aware benchmarks and representation methods
Microbial natural product discovery	Microbial metabolites are a strong AI-linked discovery domain	Machine learning; deep learning; bioinformatics	Microbial natural product databases; biosynthetic and structural records	Infectious disease; antimicrobial discovery; cross-cutting	Candidate prioritization; dereplication; biosynthetic interpretation	Predictions depend on training-set coverage and experimental follow-up	Biosynthetic AI; knowledge graphs; integrated omics	Link computational prioritization to experimentally verified activity and biosynthetic context
Target and mechanism prediction	ML supports hypothesis generation for biological effects	Machine learning; target prediction; network pharmacology	Natural product structures; pathway databases; target annotations	Cancer; inflammation; metabolic disease; neurodegenerative disease where supported	Target inference; pathway prioritization; mechanism hypothesis generation	Predicted targets are not validated targets	Explainable AI; causal pathway modeling	Require external validation, biological assays, and transparent confidence reporting
Virtual screening and docking-linked prioritization	Screening explores chemical diversity and ranks candidates	Virtual screening; molecular docking; similarity search; ML-	Natural product libraries; phytochemical datasets	Cross-cutting; antimicrobial; antiviral; cancer where supported	Hit prioritization; lead discovery; compound ranking	Docking scores and screening ranks do not prove	Hybrid ML–docking workflows; active learning	Connect screening outputs to assay validation and

		assisted prioritization				biological efficacy		reproducible protocols
ADMET and toxicity triage	Predictive models support early safety and developability assessment	ADMET prediction; toxicity prediction; machine learning	Natural product structures; property datasets; toxicity annotations	Cross-cutting	Safety triage; developability prioritization	Limited natural-product-specific pharmacokinetic and toxicity data	Uncertainty-aware ADMET; interpretable toxicity models	Build curated natural-product ADMET datasets and report uncertainty
Disease-priority mapping	Disease areas appear as thematic contexts rather than proof of clinical impact	ML; network pharmacology; virtual screening; target prediction	Natural product and phytochemical datasets	Cancer; infectious disease; inflammation; metabolic disease; neurodegenerative disease; antimicrobial and antiviral discovery where supported	Disease-focused prioritization; mechanism mapping	Disease mentions may reflect computational interest, not validated efficacy	Multimodal disease modeling; patient-relevant data integration	Distinguish computational disease relevance from pharmacological and clinical validation
Generative natural product-inspired design	Generative AI can propose novel or analogue-like structures	Deep generative models; molecular design algorithms	Natural product-inspired scaffolds; curated compound libraries	Cross-cutting	De novo design; analogue generation; scaffold exploration	Generated molecules may lack synthesis feasibility or biological relevance	Constraint-aware generative AI; synthesis-aware design	Integrate generative design with synthetic feasibility and experimental testing
Explainability and reproducibility	Interpretability and workflow transparency are central to trust	Explainable AI; reproducible pipelines; benchmark-aware validation	Model inputs, feature attributions, curated datasets	Cross-cutting	Model interpretation; decision support; validation planning	Black-box outputs, benchmark leakage, poor reporting	Explainable and auditable AI workflows	Require transparent datasets, code, validation splits, and interpretable outputs
Translation and validation	Computational prioritization must connect to pharmacological confirmation	Integrated ML, screening, ADMET, and experimental workflows	Natural product databases plus assay and pharmacology data	Cross-cutting	Lead prioritization; preclinical decision support	Computational outputs alone cannot establish efficacy, safety, or clinical readiness	Design-make-test cycles; translational readiness frameworks	Build workflows that connect prediction to synthesis, assays, pharmacology, and reproducibility

Figure 2 illustrates the integrated thematic synthesis linking natural product data sources, AI methods, drug discovery tasks, disease priorities, emerging frontiers, and translational barriers.

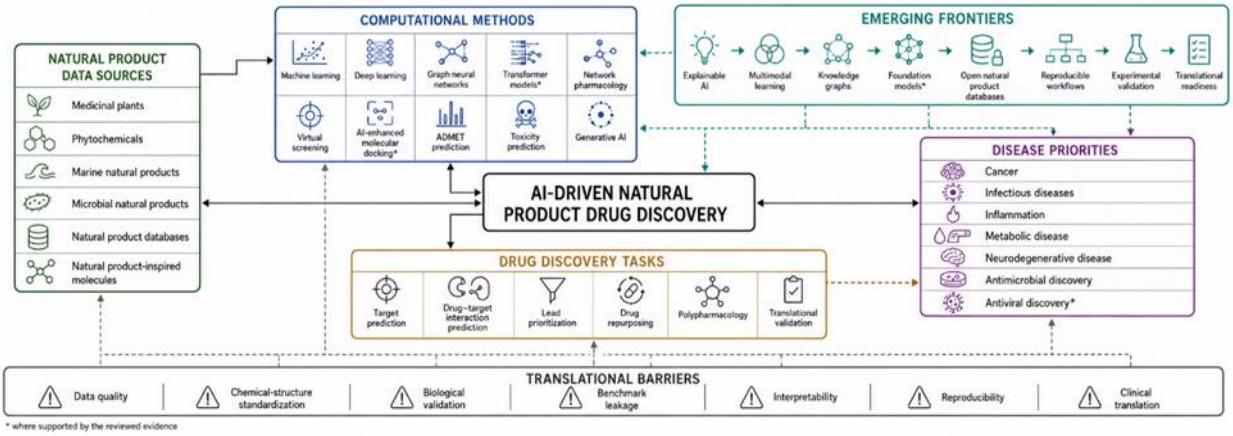

Figure 2. Integrated Thematic Map of AI-Driven Natural Product Drug Discovery

Alt-text description

A conceptual thematic map showing AI-driven natural product drug discovery at the center, connected to natural product databases, phytochemicals, medicinal plants, marine and microbial natural products, machine learning, deep learning, graph models, network pharmacology, virtual screening, target prediction, ADMET prediction, disease priorities, emerging frontiers, validation challenges, and translational barriers.

Research agenda

A stronger research agenda for AI-driven natural product discovery should begin with chemistry-aware modeling and responsible drug-design criteria. AI in chemistry and drug design can support compound prioritization, representation learning, and design exploration, but future workflows should report molecular-standardization rules, dataset provenance, validation splits, uncertainty estimates, and applicability domains [27]. For natural products, this means preserving stereochemical information, biosynthetic context, structural complexity, and metadata provenance rather than forcing chemically distinctive compounds into overly generic model pipelines. The next priority is to connect AI prediction with synthesis feasibility, analogue generation, and iterative design-make-test workflows. AI has an expanding role in medicinal chemistry synthesis planning and workflow integration, which is relevant for natural product-inspired discovery when computationally prioritized or generated molecules must be made, modified, tested, and optimized [28]. Future studies should therefore integrate generative design, retrosynthetic feasibility, assay planning, ADMET triage, and experimental validation rather than presenting predicted molecules as discovery endpoints.

Finally, the field needs validation standards tailored to natural product data. These should include leakage-aware benchmarking, external validation, negative controls, assay-quality reporting, open code where possible, transparent database versions, interpretable model outputs, and independent experimental confirmation. A robust agenda should also promote multimodal datasets linking structures, biosynthetic information, omics, target annotations, bioactivity, toxicity, pharmacokinetics, and disease biology, while maintaining a clear distinction between computational prioritization and therapeutic evidence.

CONCLUSION

This bibliometric and thematic review maps natural products in AI-driven drug discovery as a field shaped by data infrastructure, molecular representation, target and mechanism prediction, virtual screening, network pharmacology, ADMET and toxicity modeling, disease-area prioritization, generative design, explainability, and validation constraints. The analysis emphasizes that AI can support natural product discovery by organizing chemical space, improving prioritization, generating hypotheses, and identifying candidates for further study, but reliable translation requires curated data, chemical-structure standardization, benchmark transparency, interpretability, reproducibility, experimental validation, and pharmacological confirmation. The central contribution of the review is a structured field-level synthesis that distinguishes bibliometric mapping from thematic interpretation and positions AI as a decision-support framework within natural product drug discovery rather than a replacement for experimental chemistry, biology, and translational pharmacology.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod. 2020;83(3):770-803. doi:10.1021/acs.jnatprod.9b01285

[2] Othman ZK, Ahmed MM, Kasimieh O, Musa SS, Branda F, Cue EG, et al. Artificial intelligence for natural product drug discovery and development: Current landscape, applications, and future directions. Intell Based Med. 2025;12:100316. doi:10.1016/j.ibmed.2025.100316

[3] Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023;22(11):895-916. doi:10.1038/s41573-023-00774-7

[4] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[5] Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. J Bus Res. 2021;133:285-96. doi:10.1016/j.jbusres.2021.04.070

[6] Aria M, Cuccurullo C. Bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959-75. doi:10.1016/j.joi.2017.08.007

[7] van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics. 2017;111(2):1053-70. doi:10.1007/s11192-017-2300-7

[8] Moral-Muñoz JA, Herrera-Viedma E, Santisteban-Espejo A, Cobo MJ. Software tools for conducting bibliometric analysis in science: An up-to-date review. Prof Inf. 2020;29(1):e290103. doi:10.3145/epi.2020.ene.03

[9] Karger E, Kureljusic M. Using artificial intelligence for drug discovery: A bibliometric study and future research agenda. Pharmaceuticals (Basel). 2022;15(12):1492. doi:10.3390/ph15121492

[10] He B, Guo J, Tong HHY, To WM. Artificial intelligence in drug discovery: A bibliometric analysis and literature review. Mini Rev Med Chem. 2024;24(14):1353-67. doi:10.2174/0113895575271267231123160503

[11] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021;13(1):2. doi:10.1186/s13321-020-00478-9

[12] van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, et al. The Natural Products Atlas 2.0: A database of microbially-derived natural products. Nucleic Acids Res. 2022;50(D1):D1317-23. doi:10.1093/nar/gkab941

[13] Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, et al. NPClassifier: A deep neural network-based structural classification tool for natural products. J Nat Prod. 2021;84(11):2795-807. doi:10.1021/acs.jnatprod.1c00399

[14] Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res. 2025;53(D1):D634-43. doi:10.1093/nar/gkae1063

[15] Sahayasheela VJ, Lankadasari MB, Dan VM, Dastager SG, Pandian GN, Sugiyama H. Artificial intelligence in microbial natural product drug discovery: Current and emerging role. Nat Prod Rep. 2022;39(12):2215-30. doi:10.1039/D2NP00035K

[16] Arnold A, Alexander J, Liu G, Stokes JM. Applications of machine learning in microbial natural product drug discovery. Expert Opin Drug Discov. 2023;18(11):1259-72. doi:10.1080/17460441.2023.2251400

[17] Zhang R, Li X, Zhang X, Qin H, Xiao W. Machine learning approaches for elucidating the biological effects of natural products. Nat Prod Rep. 2021;38(2):346-61. doi:10.1039/D0NP00043D

[18] Jeon J, Kang S, Kim HU. Predicting biochemical and physiological effects of natural products from molecular structures using machine learning. Nat Prod Rep. 2021;38(11):1954-66. doi:10.1039/D1NP00016K

[19] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4

[20] Meyers J, Fabian B, Brown N. De novo molecular design and generative models. Drug Discov Today. 2021;26(11):2707-15. doi:10.1016/j.drudis.2021.05.019

[21] Nguyen T, Le H, Quinn TP, Nguyen TM, Le TD, Venkatesh S. GraphDTA: Predicting drug-target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140-7. doi:10.1093/bioinformatics/btaa921

[22] Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: A comprehensive review of deep generative models. Brief Bioinform. 2021;22(6):bbab344. doi:10.1093/bib/bbab344

[23] Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1):20. doi:10.1186/s13321-020-00424-9

[24] Santana K, do Nascimento LD, Lima e Lima A, Damasceno V, Nahum C, Braga RC, et al. Applications of virtual screening in bioprospecting: Facts, shifts, and perspectives to explore the chemo-structural diversity of natural products. Front Chem. 2021;9:662688. doi:10.3389/fchem.2021.662688

[25] Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. Drug Discov Today. 2019;24(5):1157-65. doi:10.1016/j.drudis.2019.03.015

[26] Catacutan DB, Alexander J, Arnold A, Stokes JM. Machine learning in preclinical drug discovery. Nat Chem Biol. 2024;20(8):960-73. doi:10.1038/s41589-024-01679-1

[27] Brown N, Ertl P, Lewis R, Luksch T, Reker D, Schneider N. Artificial intelligence in chemistry and drug design. J Comput Aided Mol Des. 2020;34(7):709-15. doi:10.1007/s10822-020-00317-x

[28] Struble TJ, Alvarez JC, Brown SP, Chytil M, Cisar J, DesJarlais RL, et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. J Med Chem. 2020;63(16):8667-82. doi:10.1021/acs.jmedchem.9b02120