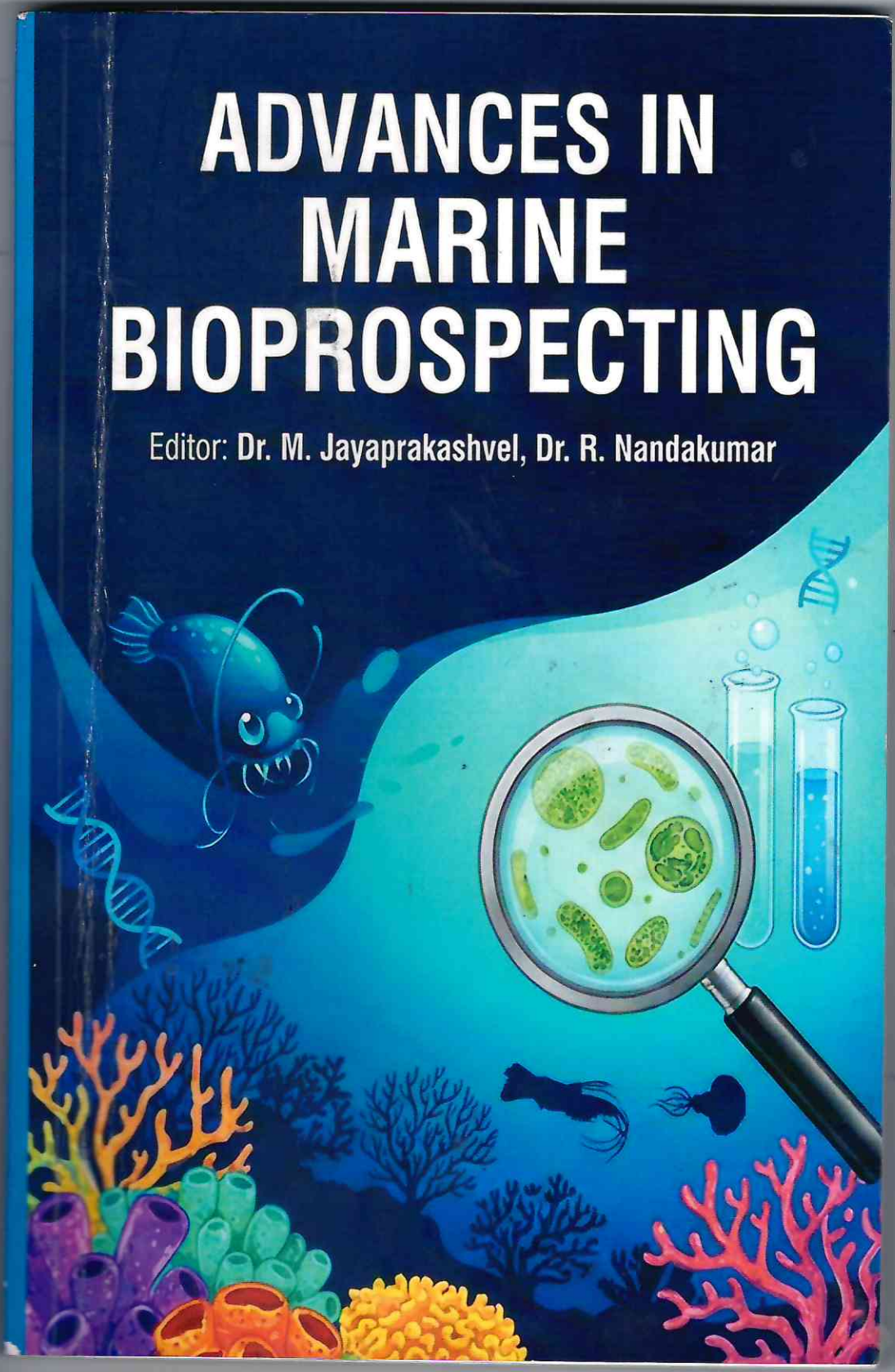
ADVANCES IN MARINE BIOPROSPECTING

Editor: **Dr. M. Jayaprakashvel, Dr. R. Nandakumar**



Advances in Marine Bioprospecting

Dr. M. Jayaprakashvel
Dr. R. Nandakumar

HSRA
PUBLICATIONS

Published by
HSRA Publications 2026
#02, Sri Annapoorneshwari Nilaya,
1st Main, Byraveshwara Nagar, Laggere,
Bangalore – 560058
Sales Headquarters – Bangalore

ISBN: 978-93-6850-536-5

First Edition 2026

No. of Pages – 160

Size : 1/8 Demi

Artificial Intelligence–Driven Natural Product Drug Discovery: Integrating Machine Learning with Metabolite Databases

Shiammala P N[a,*] and V. Raghavendran

Department of Computer Applications, VELS Institute of Science Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu-600117

Corresponding author mail I.D: shiammala@gmail.com

Abstract

Natural products have been a key driver of drug discovery through their diverse chemical nature and wide range of biological activities. However, conventional drug discovery based on natural products is often limited by issues such as re, discovery of known compounds, complicated structure determination, constrained scalability, and lengthy development times. The integration of artificial intelligence (AI), machine learning (ML), and extensive metabolite databases is changing the way drug discovery works by allowing data, driven, predictive, and high, throughput discovery strategies. This chapter reviews the impact of machine learning algorithms combined with natural product and metabolite databases on various stages of drug discovery such as dereplication, bioactivity prediction, virtual screening, structure activity relationship analysis, and mechanism of action inference. Ocean and microbial natural products are primarily focused on as typical examples of chemically rich resources that are less explored. Besides that, the chapter outlines present issues regarding data quality, model interpretability, and database interoperability and points out potential future developments like generative AI and autonomous discovery platforms. Altogether, this integration illustrates the way AI, driven informatics frameworks are revolutionizing natural product research to be quicker, more efficient, and more environmentally friendly pharmaceutical discovery.

Keywords: Artificial intelligence; Machine learning; Natural products; Drug discovery; Metabolite databases; Marine bioactives

Introduction

Natural products (NPs) have been at the core of the evolution of modern therapeutics, and have been one of the main sources of approved drugs for multiple therapeutic areas, including oncology, infectious diseases, immunology, and neurology. The success of natural products lies in their wide diversity, high stereochemical complexity, and evolutionary optimization for a biological function. These characteristics allow natural products to be in the unique chemical space and impossible to find in purely synthetic compound libraries, therefore they offer privileged scaffolds for drug discovery (Aarthy et al., 2022; Gaudncio et al., 2023). In spite of their value, the traditional natural product based drug discovery has been less preffered choice by industries in the last twenty years. The decrease is attributed to the low throughput of bioassay, guided fractionation, frequent rediscovery of known metabolites, difficulties in structure elucidation, limited availability of source organisms and problems in sustainability, as well as rediscovery of known metabolites, for which industrial adoption has been reluctant. Moreover, natural products often have complicated pharmacokinetic and toxicity profiles that make their progression through the drug development pipeline more difficult (Vijayraj et al., 2023: Romanelii et al., 2023;Thirumalai swamy et al., 2024; Gangwal et al., 2025).

One of the most essential barriers in natural product research is dereplication, the process of identifying previously known compounds in complex extracts at an early stage. Poor dereplication results in repetitions, waste of resources, and extension of discovery timelines. Despite the improvements of metabolite profiling brought by the hyphenated analytical techniques such as LCMS/MS and NMR spectroscopy, the handling of large, multidimensional datasets still poses a significant challenge when performed manually or through rule, based methods only (Gaudncio et al., 2023).

The fast development of artificial intelligence (AI) and machine learning (ML) technologies is turning to be a game changer

for many long standing issues. Machine learning algorithms can reveal non-linear relationships in extensive chemical, biological, and omics datasets, which opens the way for predictive modeling at various stages of drug discovery. In the case of natural products, ML may be helpful in bioactivity prediction, toxicity assessment, virtual screening, structureactivity relationship (SAR) analysis, and target identification. The availability of such tools makes it possible to use natural product discovery not as a blindly trial, and, error empirical method, but as a rational, data, driven framework (Gangwal et al., 2025).

Metabolite and natural product databases have grown in parallel and now offer a unique possibility for AI-powered discovery. The modern databases combine chemical structures, spectroscopic fingerprints, genomic and biosynthetic gene cluster annotations, and experimentally verified bioactivities. These databases, when coupled with machine learning models, become capable of automated dereplication, chemical space navigation, and rapid prioritization of novel compounds that are biologically relevant. The experimental burden is greatly mitigated through such integrative strategies while the efficiency of the discovery process is enhanced (Gaudncio et al., 2023). The marine and microbial natural products are a source of both challenges and opportunities for AI, enabled discovery. The marine ecosystems, especially, constitute a huge and mostly untapped reservoir of secondary metabolites with unprecedented structural features and high bioactivities. Nevertheless, the difficulties in sampling, cultivation, and compound isolation have, for the most part, hindered their exploitation. Machine learningguided virtual screening of marine metabolite libraries to pinpoint candidate antiviral and anticancer compounds is what recent studies have shown, thus, demonstrating the practical utility of AI, based methodologies in tackling complex and unexplored chemical spaces (Zhang et al., 2024; Albukhari et al., 2025).

AI-driven frameworks are not only used for compound identification but are also being extended to understand the mechanisms of action, predict resistance pathways, and support precision drug delivery strategies. These changes emphasize how the three fields of natural product chemistry, systems biology, and artificial intelligence are increasingly merging into one integrated discovery ecosystem. With continuous improvements in data quality,

model interpretability, and database interoperability, AI-assisted natural product research is set to become a major component of next, generation drug discovery. The chapter explores the use of machine learning techniques in combination with metabolite and natural product databases for speeding up and making drug discovery more logical. By looking at methodological innovations, representative case studies, and new trends, the chapter intends to offer a comprehensive and future, oriented view of how AI is transforming the field of natural product based therapeutics.

2. Bottlenecks in Traditional Natural Product Drug Discovery and the Rationale for AI Integration

Traditional natural product (NP) drug discovery pipelines that have brought forth numerous successful drugs over the years are increasingly facing structural, methodological, and economic bottlenecks. These issues stem from the complexity of natural matrices, the limitations of experimental workflows, and the rising demand for faster and more affordable drug development. Consequently, comprehending these bottlenecks is a prerequisite for grasping the reason for the AI/ML-powered revolution in NP research that is ongoing.

2.1 Rediscovery and Inefficient Dereplication

Naturally, one of the most severe problems in NP discovery is the frequent rediscovery of known compounds. Biological extracts can be extremely complex and contain hundreds to thousands of metabolites, of which many have already been reported. Therefore, in the absence of effective dereplication, a large part of the time and money are unnecessarily spent on isolating and characterizing the already known molecules. Metabolite detection has been revolutionized by the improvements in analytical techniques such as LCMS/MS, high, resolution mass spectrometry, and NMR spectroscopy. However, the interpretation and cross, referencing of large spectral datasets are still very demanding in terms of labor and are often influenced by human bias (Gaudncio et al., 2023). Moreover, traditional dereplication workflows are heavily dependent on the manual comparison against reference libraries, thus, limiting scalability and throughput. As the natural product libraries get bigger, these methods become less and less viable, making the transition to

automated data-driven dereplication frameworks inevitable (Singh et al., 2025).

2.2 Low Throughput and Resource, Intensive Screening

Bioassay-guided fractionation has been the major method used in NP drug discovery. Although this technique is still efficient, it is low-throughput by nature and demands going through the cycles of extraction, purification, and biological testing multiple times. Such procedures are not appropriate for the large-scale exploration of natural product collections and, as a result, are usually incompatible with the timelines of modern drug discovery programs (Gangwal et al., 2025; ihedioha et al., 2023; Jayaprakashvel et al., 2023). Furthermore, numerous bioassays are target, specific and thus have a limited capacity to detect multi, target or system, level activities that are typical of many natural products. This restriction can lead to the elimination of compounds with therapeutic potential beyond the initial screening context.

2.3 Structural Complexity and SAR Challenges

Natural products often have high molecular weights, are densely functionalized, and contain several stereocenters. While these characteristics are the source of biological potency, they also make it difficult to understand the structure, carry out chemical synthesis, and perform structureactivity relationship (SAR) analysis (Massad et al., 2022). Traditional SAR studies involve analogue synthesis and testing on a large scale, which is rarely feasible for complex NP scaffolds. In addition, the typical non, linear and multi, target interactions of NPs are hard to explain when using classical medicinal chemistry frameworks. This drawback slows down lead optimization and lowers the level of trust in the decisions made regarding further development.

2.4 Limited Accessibility and Sustainability Issues

A large proportion of bioactive natural products are derived from rare, slow-growing, or organisms that cannot be cultivated, especially those in marine and extreme environments. The scarcity of biomass and the risk of ecological damage due to overharvesting are major obstacles to sustainable drug development. These limitations escalate the cost and risk associated with NP-based pipelines (Albukhari et al., 2025).

2.5 Rationale for AI and Machine Learning Integration

The use of AI and ML is a logical and effective response to the problems described above. Machine learning algorithms can quickly process high-dimensional chemical and biological datasets to achieve automated dereplication, bioactivity prediction, toxicity profiling, and identification of novel compounds. By utilizing information from metabolite databases and experimental annotations, ML models minimize redundancy and maximize the number of experimental candidates (Gaudncio et al., 2023). Moreover, AI-based virtual screening and predictive modeling facilitate the rapid and extensive natural product chemical space exploration without the need for laborious laboratory experiments. Besides, the discovery timelines are radically shortened by these strategies, and the decision-making process is also improved since the compound performance at an early stage can be inferred probabilistically (Gangwal et al., 2025).

The integration of AI, machine learning, and metabolite databases is reshaping how new drugs are discovered from natural products. These advanced, AI-powered systems that are able to learn and interpret vast datasets in metabolomics, fundamentally change the way we overcome the main obstacles in efficiency, scalability, and interpretability. As a result, they not only pave the way for a more rational and ethical use of the immense chemical diversity of nature but also significantly increases the overall productivity of the drug discovery process.

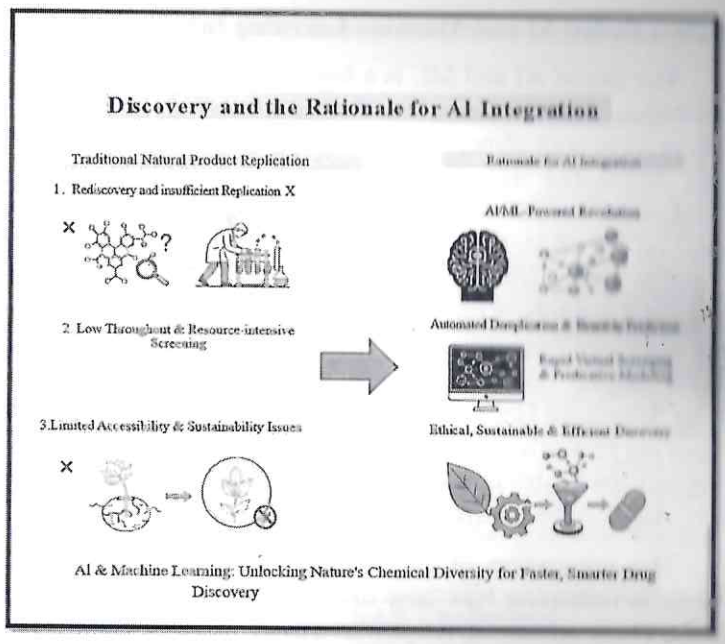


Figure 1: Bottlenecks in Natural Product Drug Discovery and the Rationale for AI Integration

3. Machine Learning Paradigms and Algorithms Applied to Natural Product Drug Discovery

Machine learning (ML) has become one of the key enabling technologies in contemporary drug discovery. It provides computational strategies capable of learning complex patterns from large and heterogeneous datasets. ML methods in natural product (NP) research are extremely beneficial because NP-derived compounds exhibit high dimensionality, structural diversity, and non-linear biological interactions. The current section highlights the main ML paradigms and algorithmic approaches that have been effectively used in natural product drug discovery.

3.1 Supervised Learning Approaches

Supervised learning is the ML paradigm most commonly employed in drug discovery. Here, models are trained on labeled datasets that link chemical structures with biological activities, targets, or pharmacokinetic properties. The most common supervised

algorithms are random forests, support vector machines, k-nearest neighbors, and artificial neural networks. Supervised learning models have found extensive application in natural product discovery to predict bioactivity, toxicity, and drug-likeness. These models, by learning from annotated metabolite databases, can rapidly prioritize compounds with a high likelihood of biological relevance, thus, do the necessary experimental screening only on a small fraction of the compounds. Random forest and support vector machine models are two particularly favored approaches for smaller datasets due to their robustness and interpretability, while deep neural networks show better performance when large, high-quality datasets are available (Elbadawi et al., 2021).

3.2 Unsupervised Learning and Chemical Space Exploration

Unsupervised learning methods are instrumental in chemically exploring natural products' space without the need for labeled data. K-means clustering, hierarchical clustering, principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE) are some of the techniques that allow the visualization and grouping of compounds based on structural similarity or physicochemical properties. Such approaches enable scaffold discovery, chemical diversity assessment, and the recognition of less populated NP libraries' regions. Unsupervised clustering of mass spectrometric features, in metabolomics-driven workflows, helps metabolite annotation and dereplication by disclosing the relationships between known and unknown compounds (Gaudncio et al., 2023).

3.3 Deep Learning and Neural Network Models

Deep learning is a significant extension of classical ML, with multi-layer neural network architectures capable of automatically extracting hierarchical feature representations. Deep learning models in NP drug discovery have been used for bioactivity prediction, absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties prediction, as well as molecular target inference. Convolutional neural networks and recurrent neural networks have been implemented to obtain features directly from molecular descriptors, fingerprints, or sequence-based representations. The models are very good at capturing subtle, non-linear relationships which are hard to specify even with the most

advanced handcrafted features. However, their performance is highly dependent on data quality and quantity, highlighting the importance of well-curated metabolite databases (Gangwal et al., 2025).

3.4 Graph-Based and Message-Passing Neural Networks

Graph-based learning has been a significant driver of innovations in natural product research. These models reinterpret molecules as graphs where atoms are nodes and bonds are edges. Message-passing neural networks and graph convolutional networks extract the features of molecules from these graphs themselves, thus forgoing the traditional descriptors. The complexity of natural products with their multiple ring systems, stereochemistry, and functional group variation makes graph-based models an ideal choice. Examples of the use of these models are virtual screening, bioactivity classification, and prediction of molecular interactions with protein targets. By leveraging message-passing neural networks, studies have achieved in identifying biologically active natural products with high accuracy from large metabolite libraries, including marine-derived compounds (Zhang et al., 2024).

3.5 Hybrid and Ensemble Learning Strategies

Hybrid and ensemble methods can either combine several ML algorithms or use ML with physics-based methods such as molecular docking. By an ensemble model output is decided from the outputs of several learners, therefore the model is more robust and has higher predictive accuracy. Hybrid workflows in NP discovery generally utilize ML for quick prescreening, then molecular docking or molecular dynamics simulations for mechanistic confirmation. These integrative tactics give users the benefit of computational speed and biological interpretability, which is why they are used in large-scale virtual screening campaigns (Gaudncio et al., 2023)

3.6 Challenges in Model Development and Deployment

Machine learning models hold great promise for natural product drug discovery. However, they are still hampered by data imbalance, limited availability of labeled datasets, and lack of model interpretability. Natural product libraries tend to have significantly fewer annotated examples than synthetic compound databases, which greatly increases the risk of overfitting. In addition, the "black-box" nature of deep learning models may pose a problem for mechanistic

understanding and acceptance by regulators. Overcoming these obstacles involves not only carefully curating datasets and using domain knowledge but also implementing explainable AI methods to guarantee that predictions are dependable and can be acted upon.

4. Metabolite and Natural Product Databases as Foundations for AI-Driven Discovery

The success of machine learning-driven natural product (NP) drug discovery is heavily reliant on the existence, quality, and compatibility of metabolite and natural product databases. These databases are the definable vessels of chemical, spectroscopic, biological, and genomic data and they are the fundamental training material on which AI models learn to detect patterns, predict activities, and recognize novelties. Hence, databases constitute the core of data-driven NP discovery.

4.1 Evolution of Natural Product and Metabolite Databases

Initially, natural product databases were largely catalog-based, recording compounds' names, source organisms, and providing minimal structural information. These resources were handy but had limitations in terms of their computational analysis. Over the last ten years, significant improvements in analytical chemistry, metabolomics, and bioinformatics have led to the development of the next-generation databases that interact with multi-dimensional data, such as high-resolution mass spectra, NMR fingerprints, biological activities, and biosynthetic annotations (Gaudncio et al., 2023). Currently, databases are more focused on open access, standardization, and machine-readability, which facilitates their integration with AI and chemoinformatics pipelines without any barriers. This progression has turned databases from being mere passive reference instruments to the active ones that can generate hypotheses and perform predictive modeling.

4.2 Types of Databases Relevant to AI-Based Drug Discovery

Almost every natural product drug discovery requires the synchronized application of various types of databases, each of which offers discrete, yet harmonized, layers of information. The layering of databases creates an ecosystem, with structural databases at the lowest tier, which include collections of chemical structure representations, molecular fingerprints, and physicochemical

parameters (descriptors) like molecular weight, hydrophilicity, and hydrogen-bonding capability. These databases provide the means to perform similarity searches, identify chemical scaffolds, and analyze chemical spaces. These actions are necessary for the machine learning-driven virtual screening and prioritization of chemical compounds.

Another important database type includes spectral databases which provide high-quality datasets for mass spectrometry (MS/MS) and nuclear magnetic resonance (NMR). These databases help to detect and remove duplicate metabolites (or at least homologous structures) in metabolomics workflows through rapid spectral matching to reference libraries. In combination with machine learning, spectral databases perform automated pattern recognition, improve confidence in metabolite identification, and reduce the time spent on identifying known entities to help identify novel structures. Bioactivity databases help link chemical structures to biological functions by describing natural products with experimentally verified bioactivities, molecular targets, and therapeutic applications. These datasets are used as training data for supervised machine learning techniques that predict biological activity, toxicity, and mechanisms of action. Quantitative bioactivity data makes models more robust and allows for the analysis of structure-activity relationships among various classes of natural products (parvatikar et al.,2023).

Finally, genomic and biosynthetic databases link metabolites to the organisms that produce them, as well as to the corresponding biosynthetic gene clusters. These databases support genome-guided discovery methods by associating chemical variation with corresponding biological information and enzymatic systems. The blending of genomic data with information about the chemistry and biological activity of a compound enables machine learning methods to incorporate biosynthetic information to enhance predictions about metabolite novelty, functional importance, and evolutionary relationships (Chen et al., 2023).

The synthesis of these various types of databases establishes a system for mapping multidimensional data that is able to simultaneously correlate chemical structure, biological activity, spectral features, and biosynthetic sources. This type of comprehensive data synthesis improves the interpretability, predictive precision, and translational

applicability of machine learning models for natural products. This highlights the importance of machine learning models in natural product drug discovery.

4.3 Database-Driven Dereplication and Novelty Assessment

Among the various applications of AI-supported databases in NP discovery, dereplication stands out for its straightforwardness and practicality. Using a hybrid approach of spectral matching, molecular networking, and ML similarity scoring, one can quickly catalog known compounds and distinguish them from prospective metabolites. In dereplication, ML improves upon rule-base methodologies by analyzing more complex spectral or structural designs, thereby dramatically decreasing both the false positive and rediscovery rates (Gaudêncio et al., 2023). Database-driven novelty scoring facilitates the ranking of compounds located in less populated regions of chemical space, thereby heightening the potential for novel discoveries of innovative chemical and/or biological structures.

4.4 Training and Validation of Machine Learning Models

High-quality databases for annotated datasets are essential to build reliable ML models. For instance, the bioactivity and toxicity models are supervised learning models and require labeled examples for each of the learning tasks, and the diversity and balance of the entries in a given database influence the generalizability of the model and its performance. The research of natural products is still faced with a serious paucity of data and bias in its annotations. This is predominantly due to the lack of a thorough biological characterization of a given metabolite, thereby rendering it less useful for supervised learning. In such circumstances, the implementation of semi-supervised and transfer learning methodologies seems to be the most promising. These methodologies utilize a well-annotated dataset of synthetic compounds to improve predictions in underrepresented classes of natural products (Gangwal et al., 2025).

4.5 Interoperability and Data Integration Challenges

Even with a lot of progress, the fragmentation of databases and lack of standardization still significantly impede AI-driven NP discovery. Variations in data formats, different annotation standards, and diverse curation practices make it very challenging to integrate data from different databases and train models. Initiatives to

standardize metadata, implement FAIR (Findable, Accessible, Interoperable, Reusable) principles, and encourage open data sharing are absolutely necessary for the AI to have its full effect in this domain (Gaudncio et al., 2023).

4.6 Strategic Importance of Marine and Microbial Databases

The metabolite databases for marine and microbial sources hold a key significance because these sources offer an exceptionally high level of chemical novelty. AI-enabled mining of such databases has led to the discovery of antiviral and anticancer lead compounds in which the identification would have been very challenging by the traditional methods (Zhang et al., 2024; Albukhari et al., 2025). Moreover, by associating chemical data with ecological and genomic context, these databases enable the development of more targeted exploration strategies, thus lowering experimental risk and enhancing sustainability (Selvaraj et al., 2025).

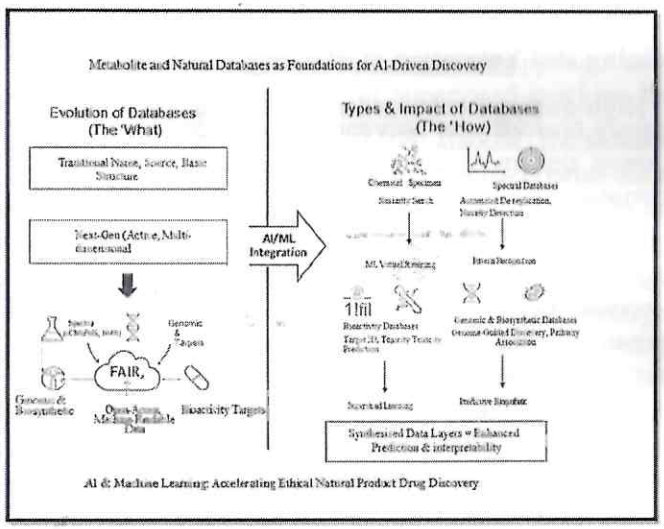


Figure 2. Metabolite and Natural Product Databases as Foundations for AI-Driven Discovery

5. AI-Integrated Workflows for Virtual Screening, Dereplication, and Hit Prioritization

The use of artificial intelligence (AI) and machine learning (ML) in natural product (NP) drug discovery has led to the creation

of efficient, end-to-end computational workflows that not only save time but also greatly improve decision-making. These AI-powered pipelines integrate metabolite databases, chemoinformatics, and predictive modeling to solve the challenges of virtual screening, dereplication, and hit prioritization, which are the three stages most critical to the discovery process.

5.1 Virtual Screening of Natural Product Libraries

Virtual screening is perhaps the major one among the AI applications in NP research. In contrast to traditional high-throughput experimental screening, AI-driven virtual screening permits extremely rapid in silico testing of thousands or even millions of natural products against specific biological targets or phenotypic endpoints. Machine learning models trained on curated bioactivity datasets generate probability scores that a given compound will manifest the desired biological effects, thus allowing the discarding of low-probability candidates at an early stage. Virtual screening in natural product drug discovery is extremely useful because metabolites are structurally complex and diverse. ML-based screening methods can handle non-linear dependencies between structure and activity and can also find compounds with novel scaffolds that a rule-based filter may reject without further consideration. These models are often combined with molecular docking or pharmacophore modeling to refine predictions and provide mechanistic insights, thereby increasing confidence in selected hits (Lyu et al., 2023).

5.2 AI-Assisted Dereplication and Redundancy Reduction

Dereplication represents a major source of delay in natural product research and is responsible for a significant fraction of discovery timelines being extended. AI-assisted dereplication workflows leverage spectral databases, molecular networking, and machine learningbased similarity analysis to achieve the purpose in the shortest possible time and distinguish metabolites known from those novel ones. Machine learning models identify differences in mass spectrometric and structural data and so can find those compounds which were reported in the literature even if spectral variation or incomplete reference data are present. Redundancy is brought down to a negligible level by such automated dereplication systems thus releasing researchers' experimental resources to unique

chemical entities. Besides, AI-powered dereplication is scalable and adaptable, so it is highly compatible with metabolomics datasets of any size, which are derived from complex biological samples (Gupta et al., 2024).

5.3 Hit Prioritization and Lead Selection

Besides the identification of active compounds, AI-integrated workflows have pivotal functions in hit prioritization. Machine learning models simultaneously take into account the evaluation of predicted bioactivity, toxicity, drug-likeness, novelty, and biosynthetic feasibility, among others, to finally produce the ranking of candidate molecules. Such a multi-parameter optimization is of great help especially for natural products which are characterized by trade-offs between potency, complexity, and developability. Moreover, with the help of AI-based prioritization frameworks, which employ biosynthetic and genomic information, the aspect of production sustainability can also be accounted for, e.g., whether it can be done through cultivation, fermentation, or synthetic biology. As a result, prioritized hits are more likely to progress successfully through downstream validation and development stages (Periwal et al., 2022).

5.4 Iterative and Adaptive Discovery Pipelines

One of the AI-integrated workflows features most prominently is their potential for iterative learning. Once experimental data from bioassays, structural elucidation, and mechanistic studies are obtained, they can be reintroduced into the model that constantly improves its predictive performance. Such an adaptive cycle turns natural product drug discovery from a traditional linear workflow into a dynamic, self-optimizing process. In sum, AI-integrated virtual screening, dereplication, and hit prioritization pipelines to be considered as a paradigm shift in natural product drug discovery. By alleviating the experimental burden, reducing redundancy, and allowing rational decision-making at the earliest stages, these workflows accelerate to a great extent the identification of high-value lead compounds and enhance the overall productivity of NP-based therapeutic research (Nashaat and Ebeni, 2025).

6. Challenges, Limitations, and Future Perspectives of AI-Driven Natural Product Drug Discovery

The incorporation of artificial intelligence (AI) and machine learning (ML) into the natural product (NP) drug discovery process has attracted a lot of attention and has been very promising. However, the integration of these technologies is still constrained by various scientific, technical, and practical challenges (Bresciani et al., 2023). It is absolutely necessary to overcome these limitations if we want to see the outcomes of AI-driven discoveries being translated successfully into therapeutics that work in the clinic and are commercially viable, as well as influencing the directions of the future research. Arguably, the most critical issue is that of data availability, quality, and bias. Natural product datasets are less complete, more diverse, and have fewer annotations as opposed to synthetic compound libraries. For a large number of metabolites, there are no experimentally validated data on bioactivity, toxicity, or targets, which is why the datasets are so unbalanced and this can bias machine learning models and also limit their generalizability (Huang et al., 2021). The lack of consistency in data curation standards for different databases further complicates model training and cross-study comparison, thus pointing to the necessity of standardized, high-quality, and FAIR-compliant data infrastructures.

Another significant limitation is model interpretability. Although deep learning and graph-based neural networks show impressive predictive performance, they are often criticized for being "black boxes" which provide very limited mechanistic insight as to why a compound is predicted to be active or inactive. This absence of transparency can biologically limit the understanding, experimental researchers' confidence, and also pose problems for regulatory acceptance. Therefore, the creation of explainable AI techniques specifically designed for chemical and biological data is a very important and progressing area of research. Biological complexity is also a constant problem. Natural products most of the time have polypharmacology, which means they can interact with multiple molecular targets and pathways simultaneously. Although this feature can be used as a therapy advantage, it makes the prediction of efficacy, toxicity, and resistance mechanisms very difficult. Most AI models today are normally dependent on biologically simplified system representations in which the biological

systems may be difficult to comprehend at the system level without the integration of multi-omics, network biology, and phenotypic data.

From a practical perspective, sustainability and scalability are still major issues. In some cases, AI may even identify extremely promising natural products as drug leads; however, development may be limited due to the availability of the source organism, low yields of metabolites, or the ecological impact. While genome mining and synthetic biology may provide ways around these obstacles, the AI-driven pipeline in which they need to be integrated is still at its infancy and needs further methodological refinement (Devi et al., 2025). Several emerging trends could, in fact, reshape the landscape of AI-enabled natural product drug discovery in the near future. Among others, generative AI and deep generative models are anticipated to have an increasingly significant role in the design of NP-inspired analogues that not only retain biological activity but also have improved drug-like properties and synthetic accessibility. These models are not limited to prediction but rather to creative molecular design, thus greatly broadening the chemical space that can be derived from natural scaffolds.

Another potential and fascinating direction is that of the development of autonomous and closed-loop discovery platforms. In such platforms, AI models can not only plan experiments but also, upon analyzing the results obtained, decide on the next steps and, thus, by continuous updating of hypotheses, approach the solution in a never-ending learning cycle. In fact, when combined with high-throughput metabolomics, robotics, and automated synthesis, these platforms may have the potential of turning discovery timelines upside down and thus, to a great extent, human intervention can be taken out of the equation. The success of AI-powered NP discovery in the future is going to be heavily influenced by collaboration across different disciplines as well as open science (Chakraborthy et al., 2024). The seamless interplay of chemistry, biology, data science, and engineeringfacilitated by common databases and open methodswill be the main factors that lead to breaking through the present constraints. When these issues are gradually being solved, AI-assisted systems will be termed as ushering in a new era in natural product drug discovery which will be more predictive, environmentally friendly, and innovation-driven.

9. Conclusion

Natural products are an important source of drug leads due to their structurally diverse compounds, evolutionary refinement, and wide range of biological activities. However, the conventional methods of identifying drug leads from natural products are affected by inefficiencies, including rediscovery, low output, structural sophistication, and sustainability. In particular, the combination of artificial intelligence (AI) and machine learning (ML) with natural products and metabolites databases holds great promise to overcome many of these challenges.

This chapter has demonstrated how AI-based techniques allow for the significant and transformative change from empirical and labor-intensive drug discovery processes to sophisticated AI-based techniques. ML can streamline drug discovery to focus on dereplication, virtual screens, and multi-parameter selection of compounds to make hit lists. AI-based techniques also are able to analyze structural, bioactivity, and genomic databases, and to simultaneously analyze and optimize structural, functional, and biosynthetic hypotheses. The combined approaches improve the interpetative ability to construct hypotheses and the overall efficacy of the predictions. AI integrated workflows have significantly contributed to the study of novel, sparse and often overlooked, chemical fields, including those fields that come from the sea and from micro-organisms. AI systems enhance the efficiency of the discovery of compounds and focus the research on the sustainable production of compounds. AI based systems helps to align the research on ecological sustainable and feasible production with the research on natural compounds.

Despite the advances, there are still many challenges related to the quality of the data that have been collected, the transparency of the analytical tools used, the complexity of the research subjects and the interoperability of the various research databases. The provision of standard data curation, the use of transparent tools that deliver results in an explainable manner and the cooperation between various disciplines are the means to overcome those challenges. Emerging tools, including generative AI, autonomous discovery systems and cross-omics analytics, will further transform the natural product drug discovery field.

In conclusion, the integration of artificial intelligence, machine learning, and metabolite databases represent an outstanding opportunity to advance the field of natural product research. The use of computation systems to methodically explore the diverse chemicals present in the environment offers the opportunity to reposition natural sources as a core element in the development of modern pharmaceuticals that include products to treat diseases that are lacking effective pharmaceuticals.

References (APA Style)

Aarthi, G., Sarath, A., Harikrishnan, S., Sudarshan, S., Muthezhilan, R., & Jayalakshmi, S. (2022). bioactive potential of aspergillus sp. isolated from uppanar estuarine sediment samples against plant pathogens. Biochemical & Cellular Archives, 22(2).

Albukhari, A. F. (2025). AI-driven marine bioactives for cancer therapy: A systematic review of drug discovery, resistance overcoming strategies, and precision drug delivery. *Pharmacognosy Reviews, 19*(38), 175–185. https://doi.org/10.5530/phrev.20252348

Bresciani, S., Dabić, M., & Bertello, A. (2022). Collaborative technological development for addressing grand challenges: Opportunities, limitations, and new frameworks. *Technology in society*, *71*, 102063.

Chakraborty, C., Bhattacharya, M., Lee, S. S., Wen, Z. H., & Lo, Y. H. (2024). The changing scenario of drug discovery using AI to deep learning: Recent advancement, success stories, collaborations, and challenges. *Molecular Therapy Nucleic Acids*, *35*(3).

Chen, W., Liu, X., Zhang, S., & Chen, S. (2023). Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular therapy Nucleic acids*, *31*, 691-702.

Devi, N. B., Sen, S., & Pakshirajan, K. (2025). Artificial Intelligence in Synthetic Biology. In *Artificial Intelligence and Biological Sciences* (pp. 278-300). CRC Press.

Elbadawi, M., Gaisford, S., & Basit, A. W. (2021). Advanced machine-learning techniques in drug discovery. *Drug Discovery Today*, *26*(3), 769-777.

Gangwal, A., & Lavecchia, A. (2025). Artificial intelligence in natural product drug discovery: Current applications and future perspectives. *Journal of Medicinal Chemistry, 68*(6), 3948–3969. https://doi.org/10.1021/acs.jmedchem.4c01257

Gaudêncio, S. P., & Pereira, F. (2023). Marine drug discovery through computer-aided approaches. *Marine Drugs, 21*(8), 452. https://doi.org/10.3390/md21080452

Gaudêncio, S. P., Bayram, E., Lukić Bilela, L., Cueto, M., Díaz-Marrero, A. R., Haznedaroglu, B. Z., Jimenez, C., Mandalakis, M., Pereira, F., Reyes, F., & Tasdemir, D. (2023). Advanced methods for natural products discovery: Bioactivity screening, dereplication, metabolomics profiling, genomic sequencing, databases and informatic tools, and structure elucidation. *Marine Drugs, 21*(5), 308. https://doi.org/10.3390/md21050308

Gupta, S., Kashyap, S., & Kumar, D. (2024). AI-driven Analysis of Environmental Metagenomic Data. In *Genomic Intelligence* (pp. 188-206). CRC Press.

Huang, D. Z., Baber, J. C., & Bahmanyar, S. S. (2021). The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction. *Expert opinion on drug discovery*, *16*(9), 1045-1056.

Ihedioha, T. E., Asuzu, I. U., Anaga, A. O., Ihedioha, J. I., & Nnadi, C. O. (2023). Bioassay guided fractionation, isolation and characterization of hepatotherapeutic 1, 3-di-ortho-galloyl quinic acid from the methanol extract of the leaves of Pterocarpus santalinoides. *Journal of Ethnopharmacology*, *301*, 115864.

Jayaprakashvel, M., Thamada, S., Gunaswetha, K., & Pallaval, V. B. (2023). Microbiome Therapeutics: Emerging Concepts and Challenges in Translational Microbial Research. Human Microbiome in Health, Disease, and Therapy, 287-300

Lyu, J., Irwin, J. J., & Shoichet, B. K. (2023). Modeling the expansion of virtual screening libraries. *Nature chemical biology*, *19*(6), 712-718.

Massad, I., Suresh, R., Segura, L., & Marek, I. (2022). Stereoselective synthesis through remote functionalization. *Nature Synthesis*, *1*(1), 37-48.

Nashaat, B., & Elzeni, M. M. (2025). Reviewing AI in architectural computational design: Applications, opportunities, and the AI-ACD workflow for improved design integration. *International Journal of Architectural Computing*, 14780771251405443.

Parvatikar, P. P., Patil, S., Khaparkhuntikar, K., Patil, S., Singh, P. K., Sahana, R., ... & Raghu, A. V. (2023). Artificial intelligence: Machine learning approach for screening large database and drug discovery. *Antiviral Research*, *220*, 105740.

Periwal, V., Bassler, S., Andrejev, S., Gabrielli, N., Patil, K. R., Typas, A., & Patil, K. R. (2022). Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. *PLoS computational biology*, *18*(4), e1010029.

Romanelli, V.; Cerchia, C.; Lavecchia, A. Unlocking the Potential of Generative Artificial Intelligence in Drug Discovery. In *Applications of Generative AI*; Springer, 2024; pp 37−63.

Selvaraj, C., Desai, D., de los Santos-Villalobos, S., Jayaprakashvel, M., Muthezhilan, R., & Singh, S. K. (2025). Marine-derived antimicrobial peptides (AMPs): Blue biotechnological assets for sustainable healthcare and circular bioeconomy.

Singh, K., Gupta, J. K., Chanchal, D. K., Shinde, M. G., Kumar, S., Jain, D., ... & Tripathi, A. (2025). Natural products as drug leads: Exploring their potential in drug discovery and development. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *398*(5), 4673-4687.

Thirumalaiswamy, V., Vaishali, C. V., Sundararaju, S., Ramakritinan, C. M., Thillaichidambaram, M., & Quero, F. (2024). Nanotechnological strategy for the diagnosis of infectious diseases: recent developments and opportunities. Nanoscience and nanotechnology for smart prevention, diagnostics and therapeutics: fundamentals to applications, 143-181.

Vijayaraj, R., Altaff, K., Jayaprakashvel, M., Muthezhilan, R., Saran, B., Kurinjinathan, P., ... & Govindan, L. (2023, November). Chitin-Derived Silver Nanoparticles for Enhanced Food Preservation: Synthesis, Characterization, and Antimicrobial Potential. In Micro (Vol. 3, No. 4, pp. 912-929). MDPI.

Zhang, T., Sun, G., Cheng, X., Cao, C., Cai, Z., & Zhou, J. (2024). Screening for potential antiviral compounds from cyanobacterial secondary metabolites using machine learning. *Marine Drugs*, *22*(11), 501. https://doi.org/10.3390/md22110501

Dr. M. Jayaprakashvel is a distinguished academician and researcher serving as a Professor in the Department of Marine Biotechnology at AMET (Academy of Maritime Education and Training) Deemed to be University, Chennai, India. He holds a Ph.D. in Industrial Microbiology from the University of Madras and has long-standing expertise in marine microbiology, natural products, environmental biotechnology, plant pathology, and food technology. With over 90 research publications, multiple book chapters, and editorial contributions, Dr. M. Jayaprakashvel has significantly advanced the understanding of microbial secondary metabolites, coastal ecosystem microbiology, and biotechnological applications of marine organisms. His work has attracted hundreds of citations, reflecting its impact in both academic and applied research communities.

Beyond research, he has played key roles in academic leadership and institutional development, contributing to accreditation, policy implementation, and quality assurance at AMET University. Dr. Jayaprakashvel has mentored numerous Ph.D. and M.Phil. scholars, served as an examiner and doctoral committee member across institutions, and delivered invited lectures at national and international forums. His interdisciplinary work bridges fundamental microbiology and practical solutions for environmental and industrial challenges, making him a respected author and educator in marine biotechnology and related fields.

HSRA PUBLICATIONS

ISBN 978-93-6050-536-5

9 789368 505365

300 /-