Global Journal of Research in
Chemistry and Pharmacy
Journal homepage: https://gsjournals.com/gjrcp/
ISSN: 2980-4213 (Online)

(RESEARCH ARTICLE)

Check for updates

THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN NATURAL PRODUCT CHEMISTRY RESEARCH

Ramesh Reddy Gayam *

Independent Researcher, USA.
* Corresponding Author

Global Journal of Research in Chemistry and Pharmacy, 2024, 03(01), 001–007

Publication history: Received on 02 January 2024; revised on 23 January 2024; accepted on 25 January 2024

Article DOI: https://doi.org/10.58175/gjrcp.2024.3.1.0004

ABSTRACT

The integration of artificial intelligence (AI) into natural product chemistry represents a transformative paradigm shift in drug discovery and chemical research. Traditional approaches to natural product discovery—involving extraction, isolation, bioactivity screening, and structure elucidation—are time-consuming, resource-intensive, and limited by the vast chemical complexity of natural sources. AI technologies, including machine learning (ML) and deep learning (DL), are emerging as powerful tools that significantly enhance the speed, efficiency, and accuracy of natural compound discovery. This review critically examines current AI applications in natural product chemistry research, focusing on three key domains: bioactive compound prediction, complex spectral data interpretation, and de novo structure elucidation. Developments in computational omics technologies have provided new means to access the hidden diversity of natural products, unearthing new potential for drug discovery. AI-driven tools enable the exploration of natural products from diverse sources, assist in optimizing high-throughput screening processes, and facilitate the identification of novel bioactive molecules. Machine learning has emerged as a popular tool for analyzing the structures of natural products. However, integrating AI in natural product research presents challenges including data quality issues, model interpretability, and the need for standardized benchmarks. This review synthesizes recent advances, critically appraises evidence on model generalizability, and outlines future directions for AI-driven natural product discovery.

Keywords: Artificial Intelligence; Machine Learning; Natural Products; Bioactive Compound Prediction; Structure Elucidation; Drug Discovery

1 INTRODUCTION

Natural products have historically been a cornerstone of drug discovery, with numerous therapeutics derived from or inspired by compounds from plants, fungi, bacteria, and marine organisms (Newman and Cragg, 2020). Despite the remarkable chemical diversity and structural complexity of natural products—often surpassing synthetic chemical libraries—traditional discovery methods are constrained by significant limitations. The extraction, isolation, and bioactivity screening processes are laborious and time-consuming (Atanasov et al., 2021). Additionally, biodiversity loss and overexploitation threaten valuable sources of bioactive compounds (Harvey et al., 2015).

Artificial intelligence has emerged as a transformative tool in natural product research, offering new means to access the hidden chemical diversity of natural sources (Mullowney et al., 2023). Computational approaches to drug development are rapidly growing in popularity and have been used to produce significant results. Recent developments in information science have expanded databases and chemical informatics knowledge relating to natural products (Ogawa et al., 2023). Natural products have long been well-studied, and a large number of unique structures and remarkable active substances have been reported. Analyzing accumulated natural product knowledge using emerging computational science techniques is expected to yield more new discoveries (Ogawa et al., 2023).

Developments in computational omics technologies have provided new means to access the hidden diversity of natural products, unearthing new potential for drug discovery. In parallel, artificial intelligence approaches such as machine learning have led to exciting developments in the computational drug design field, facilitating biological activity prediction and de novo drug design for molecular targets of interest (Mullowney et al., 2023). This review provides a comprehensive examination of AI applications in natural product chemistry research, with emphasis on bioactive compound prediction, spectral data interpretation, and de novo structure elucidation.

2 AI TECHNIQUES IN NATURAL PRODUCT CHEMISTRY

2.1 Machine Learning and Deep Learning Approaches

Machine learning algorithms, including random forests, support vector machines, and neural networks, have been widely applied to predict the bioactivity of natural compounds (Mullowney et al., 2023). These approaches utilize chemical descriptors and molecular fingerprints to establish quantitative structure-activity relationships (QSAR) that enable the prediction of pharmacological properties (Ogawa et al., 2023). The basic concepts and frameworks of machine learning have been summarized in the context of natural product research (Ogawa et al., 2023).

Deep learning, including convolutional neural networks and graph neural networks, has further advanced predictive capabilities by automatically learning relevant features from molecular structures (Mullowney et al., 2023). In recent years, ML and DL methods have been applied to assist in the analysis of MS and NMR spectra, forming a series of valuable research outcomes (Hu and Qiu, 2023). The workflow for building a machine learning model consists of four main parts: dataset preparation, molecular representations and descriptors, model training, and model evaluation (Ogawa et al., 2023). Dataset preparation is crucial to generate a successful ML model. A high-quality NP dataset is a prerequisite and leads to better model performance (Ogawa et al., 2023).

Natural product research that utilizes machine learning is described in terms of the exploration of active compounds, automatic compound design, and application to spectral data (Ogawa et al., 2023). Key considerations for applying machine learning in this field include data quality, model selection, and validation strategies (Mullowney et al., 2023).

2.2 Cheminformatics and Virtual Screening

Cheminformatics approaches integrate chemical information with computational methods to accelerate drug discovery from natural sources (Ogawa et al., 2023). AI-driven virtual screening enables the rapid evaluation of large compound libraries, identifying potential bioactive molecules for experimental validation (Mullowney et al., 2023). A deep learning and docking simulation-based virtual screening strategy has enabled the rapid identification of HIF-1α pathway activators from a marine natural product database (Deep learning and docking simulation, 2023).

Machine learning applications in chemoinformatic analysis and virtual screening for natural products-based drug discovery have been extensively explored (Ogawa et al., 2023). Integrated molecular modeling and machine learning for drug design has been reviewed, demonstrating the synergy between computational approaches. Computational science and chemoinformatics approaches have been applied in natural products research in terms of their applications, strengths, limitations, and implications for the field (Ogawa et al., 2023).

2.3 Spectral Data Interpretation

The interpretation of complex spectral data, including mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectra, has been significantly enhanced by AI approaches (Hu and Qiu, 2023). In the structural study of natural products, MS and NMR technologies are considered the most insightful tools (Hu and Qiu, 2023). Deep learning algorithms can identify patterns in spectral data, enabling rapid compound identification and dereplication (Mullowney et al., 2023).

Computer-assisted structure elucidation (CASE) has greatly accelerated the process of NP structure elucidation (Hu and Qiu, 2023). However, CASE methods typically follow strict programming rules and lack the adaptive and cognitive functions of human intelligence, making them inflexible for all NP structure elucidation scenarios (Hu and Qiu, 2023).

ML and DL methods have been applied to assist structural studies of natural products based on NMR from four perspectives: NMR prediction, functional group identification, structural categorization, and quantum chemical calculation (Hu and Qiu, 2023).

Table 1: AI Techniques and Their Applications in Natural Product Chemistry

Technique	Application	Key Features	Current Status
Machine Learning	Bioactive compound prediction	QSAR modelling, molecular fingerprints, classification algorithms	Established
Deep Learning	Spectral interpretation & feature extraction	Automated pattern recognition, deep neural networks, complex feature learning	Rapidly advancing
Cheminformatics	Virtual screening & target prediction	Multi-omics data integration, molecular docking, database mining	Established
Generative Models	De novo structure design & molecular optimization	Chemical language models (e.g., Transformers), Variational Autoencoders (VAEs)	Emerging
Graph Neural Networks (GNNs)	Structure-activity relationship (SAR) prediction	Direct molecular graph representation, atom/bond node embeddings	Rapidly advancing

3 APPLICATIONS IN BIOACTIVE COMPOUND PREDICTION

3.1 Antimicrobial Compound Discovery

AI-driven approaches have successfully identified antimicrobial compounds from natural sources (Mullowney et al., 2023). Machine learning models trained on known antimicrobial compounds can predict the activity of untested natural products, prioritizing candidates for experimental validation (Ogawa et al., 2023). Studies have demonstrated the utility of AI in identifying novel antimicrobial agents from diverse natural sources, addressing the growing challenge of antimicrobial resistance (Machine learning-enabled genome mining, 2023).

Machine learning has been combined with atomistic simulations to predict SARS-CoV-2 Mpro inhibitors from natural compounds. Natural product research that utilizes machine learning has been applied to the exploration of active compounds (Ogawa et al., 2023).

3.2 Anticancer Compound Discovery

The application of AI in phytochemical-based anticancer drug discovery has gained significant momentum (Mullowney et al., 2023). AI-driven phytochemical screening represents a promising path toward accelerating the development of next-generation cancer therapies. Deep learning approaches for predicting bioactivity have been applied to target prediction, virtual screening, and poly-pharmacology studies of natural substances.

Graph Attention Network based mapping of knowledge relations between chemical spaces has been applied to study Nuclear factor kappa B and Centella asiatica (Graph Attention Network, 2023). Deep learning models have been used to predict the affinity between natural products and therapeutic drug targets, exploring the potential target space for natural products.

3.3 Genomic Mining and Bioactivity Prediction

AI has been employed in genome mining and bioactivity prediction of natural products (Machine learning-enabled genome mining, 2023). Artificial intelligence has been used to predict the bioactivities of diverse fungal biosynthetic pathways. However, it has been noted that the accuracies of such predictions are generally low, between 51% and 68%, likely because the natural products and bioactivities of only very few fungal pathways are known (Machine learning-enabled genome mining, 2023).

Table 2: Machine Learning Models for Bioactive Compound Prediction

Model Type	Algorithm Class	Primary Application	Strengths	Limitations
Random Forest (RF)	Ensemble learning *(Decision trees)*	Bioactivity prediction & toxicity scoring	Robust against overfitting; handles high-dimensional, non-linear data well	Limited interpretability relative to single trees; sensitive to imbalanced datasets
Support Vector Machine (SVM)	Kernel methods	Binary & multi-class bioactivity classification	Highly effective for high-dimensional, small-to-medium datasets	Computationally intensive for large datasets; highly sensitive to hyperparameter tuning
Neural Networks (NNs)	Deep learning *(Multi-layer Perceptrons / DNNs)*	Complex property & bioactivity predictions	Learns non-linear hierarchical features without manual feature engineering	Requires very large datasets; prone to overfitting on small chemical libraries
Graph Neural Networks (GNNs)	Graph-based representation learning	Molecular property & activity prediction	Directly captures molecular graph topology (atoms as nodes, bonds as edges)	High training complexity; requires significant computational resources
QSAR Models	Statistical regression & classification	Quantitative structure-activity relationship prediction	Well-established methodology; highly interpretable regression metrics	Applicability domain is limited to structural analogs of known training compounds

4 DE NOVO STRUCTURE ELUCIDATION

4.1 Machine Learning-Assisted Structure Elucidation (MLASE)

AI approaches are increasingly applied to the elucidation of natural product structures from spectral data (Hu and Qiu, 2023). The concept of MLASE has been proposed as a branch of CASE, representing a significant advancement in natural product structure determination (Hu and Qiu, 2023). ML algorithms have been applied to assist structural studies of natural products based on NMR from four perspectives: NMR prediction, functional group identification, structural categorization, and quantum chemical calculation (Hu and Qiu, 2023). These approaches address the limitations of traditional CASE methods, which typically follow strict programming rules and lack the adaptive and cognitive functions of human intelligence (Hu and Qiu, 2023).

A summary of the recent advancements in ML-assisted MS and NMR data analysis to establish the chemical structures of natural products has been presented (Hu and Qiu, 2023). ML-based MS/MS analyses that rely on library matching have been discussed, involving the utilization of ML algorithms to calculate similarity, predict MS/MS fragments, and form molecular fingerprints (Hu and Qiu, 2023). ML-assisted MS/MS structural annotation without library matching has also been reviewed (Hu and Qiu, 2023).

4.2 Generative Models for Natural Product Discovery

Deep generative models have been employed to explore novel natural product chemical space (Ochiai et al., 2023). A variational autoencoder-based deep learning method called NP-VAE (Natural Product-oriented Variational Autoencoder) has been developed for managing hard-to-analyze datasets from natural products (Ochiai et al., 2023). This method can handle natural products and demonstrate its utility as an in silico drug discovery tool (Ochiai et al., 2023).

A database of 67 million natural product-like molecules has been generated using a recurrent neural network trained on known natural products (67 million natural product-like compound database, 2023). This study highlights the potential of using deep generative models to explore novel natural product chemical space for high throughput in silico discovery (67 million natural product-like compound database, 2023).

4.3 MS/MS Structural Annotation

ML-based MS/MS analyses have been categorized into two approaches: library matching and non-library matching (Hu and Qiu, 2023). Library matching involves the utilization of ML algorithms to calculate similarity, predict MS/MS fragments, and form molecular fingerprints (Hu and Qiu, 2023). Non-library matching approaches for MS/MS structural annotation have also been systematically reviewed (Hu and Qiu, 2023). A transformer model called MS2Mol has been developed for illuminating dark chemical space from mass spectra (MS2Mol, 2023).

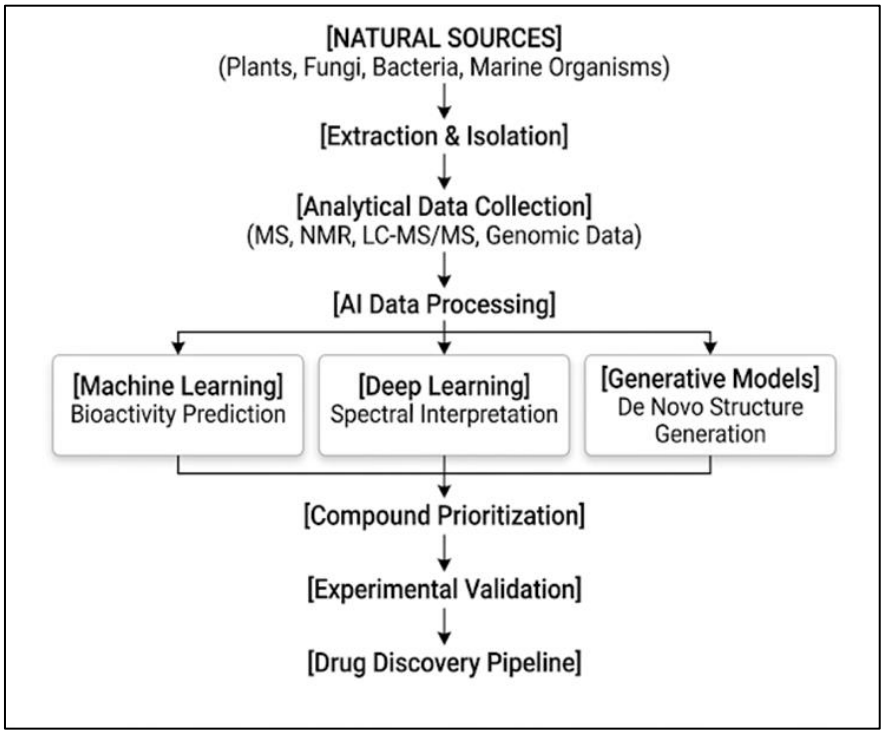

Figure 1: AI-Driven Natural Product Discovery Workflow

5 CHALLENGES AND LIMITATIONS

5.1 Data Quality and Availability

The limited availability of high-quality, standardized natural product data constrains model development (Mullowney et al., 2023). The accuracies of bioactivity predictions are often low because the natural products and bioactivities of only very few pathways are known (Machine learning-enabled genome mining, 2023). Recent developments in information science have expanded databases and chemical informatics knowledge relating to natural products, but significant gaps remain (Ogawa et al., 2023).

5.2 Model Interpretability

The "black box" nature of many deep learning models limits confidence in predictions (Mullowney et al., 2023). Traditional CASE methods follow strict programming rules and lack human-like adaptive and cognitive functions (Hu and Qiu, 2023). While ML approaches offer improvements, the interpretability of ML models remains a challenge (Ogawa et al., 2023).

5.3 Standardization

The absence of standardized benchmarks and evaluation protocols hinders comparative evaluation of AI methodologies (Mullowney et al., 2023). Key considerations for applying machine learning in this field include data quality, model selection, and validation strategies (Ogawa et al., 2023).

5.4 Computational Challenges

The conformational flexibility and stereochemical complexity of natural products present computational challenges (Ogawa et al., 2023). Deep learning methods must be adapted to handle the unique characteristics of natural product chemical space (Mullowney et al., 2023).

Table 3: Challenges and Potential Solutions in AI-Driven Natural Product Research

Challenge	Impact	Potential Solutions
Data Quality & Sparsity	Limits robust model development and leads to poor predictive generalization	Integration of standardized, curated public databases (e.g., COCONUT, GNPS); rigorous data curation protocols
Model Interpretability	Restricts trust and adoption by medicinal chemists due to "black-box" predictions	Implementation of Explainable AI (XAI) frameworks (e.g., SHAP, LIME, attention mechanisms)
Lack of Standardization	Leads to inconsistent performance evaluation across different computational pipelines	Development of open-access benchmarking datasets, unified evaluation protocols, and FAIR data principles
Computational Complexity	Increases resource requirements and limits real-time high-throughput virtual screening	Optimization of algorithms, utilization of graph compression methods, and leveraging hardware acceleration (GPUs/TPUs)
Data Imbalance	Causes biased predictions toward over-represented compound classes and false negatives	Synthetic data augmentation, active learning, transfer learning from large chemical language models, and few-shot learning

6 FUTURE PERSPECTIVES

The future of AI in natural product chemistry lies in the integration of diverse data types, including genomics, metabolomics, and chemical databases. Cross-disciplinary collaboration will be essential for leveraging AI capabilities to drive sustainable and efficient drug discovery. The integration of AI with laboratory automation promises to accelerate the experimental validation of AI-generated hypotheses.

Emerging trends include the development of explainable AI approaches to enhance model interpretability and user confidence in predictions. The application of AI to personalized medicine and precision bioprospecting represents a promising direction. The continued exploration of traditional medicine knowledge using AI approaches may reveal novel therapeutic agents and mechanisms.

The development of more sophisticated generative models for natural product design and the integration of multi-omics data with AI approaches will further accelerate natural product discovery. The potential of using deep generative models to explore novel natural product chemical space for high throughput in silico discovery has been demonstrated. Continued investment in natural product databases and computational resources is essential for advancing the field.

7 CONCLUSION

Artificial intelligence is transforming natural product chemistry research, enabling faster, more efficient discovery of bioactive compounds. Machine learning and deep learning approaches facilitate bioactive compound prediction, spectral data interpretation, and de novo structure elucidation. The concept of MLASE represents a significant advancement in natural product structure determination. Developments in computational omics technologies have provided new means to access the hidden diversity of natural products. While challenges including data quality, model interpretability, and standardization persist, the continued development of AI approaches promises to accelerate natural product drug discovery and expand access to the chemical diversity of natural sources.

REFERENCES

[1] Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., & Supuran, C. T. (2021). Natural products in drug discovery: Advances and opportunities. *Nature Reviews Drug Discovery*, *20*(3), 200–216. https://doi.org/10.1038/s41573-020-00114-z

[2] Hamdy, R., Altaie, A. M., El-Labbad, E. M., Ashmawy, N. S., Ozsahin, D. U., Husseiny, M. I., & Soliman, S. S. M. (2023). Artificial intelligence and discovery of microbial natural products. In J. N. Cruz (Ed.), *Drug discovery and design using natural products* (pp. 37–78). Springer Nature. https://doi.org/10.1007/978-3-031-35205-8_3

[3] Harvey, A. L., Edrada-Ebel, R., & Quinn, R. J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, *14*(2), 111–129. https://doi.org/10.1038/nrd4510

[4] Hu, G., & Qiu, M. (2023). Machine learning-assisted structure annotation of natural products based on MS and NMR data. *Natural Product Reports*, *40*(11), 1735–1753. https://doi.org/10.1039/D3NP00010E

[5] Mullowney, M. W., Duncan, K. R., Elsayed, S. S., Garg, N., van der Hooft, J. J. J., Martin, N. I., & Moore, B. S. (2023). Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery*, *22*(11), 895–916. https://doi.org/10.1038/s41573-023-00774-7

[6] Neelakandan, A. R., & Rajanikant, G. K. (2024). A deep learning and docking simulation-based virtual screening strategy enables the rapid identification of HIF-1α pathway activators from a marine natural product database. *Journal of Biomolecular Structure and Dynamics*, *42*(2), 629–651. https://doi.org/10.1080/07391102.2023.2194997

[7] Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, *83*(3), 703–721. https://doi.org/10.1021/acs.jnatprod.9b01285

[8] Ochiai, T., Takahashi, K., & Tanaka, K. (2023). Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. *Communications Chemistry*, *6*, Article 150. https://doi.org/10.1038/s42004-023-00951-y

[9] Ogawa, K., Sakamoto, D., & Hosoki, R. (2023). Computer science technology in natural products research: A review of its applications and implications. *Chemical and Pharmaceutical Bulletin*, *71*(7), 486–494. https://doi.org/10.1248/cpb.c23-00171

[10] Tay, D. W. P., Yeo, N. Z. X., Adaikkappan, K., Lim, Y. H., & Ang, S. J. (2023). 67 million natural product-like compound database generated via molecular language processing. *Scientific Data*, *10*, Article 296. https://doi.org/10.1038/s41597-023-02207-x

[11] Yuan, Y., Shi, C., & Zhao, H. (2023). Machine learning-enabled genome mining and bioactivity prediction of natural products. *ACS Synthetic Biology*, *12*(9), 2650–2662. https://doi.org/10.1021/acssynbio.3c00234