Deep Learning for Natural Product Bioactivity Prediction: A Critical Review of Molecular Representation, Model Reliability, and Therapeutic Relevance

Chinedu Okafor[1]*, Amina Bello[1], Ibrahim Musa[2]

[1]Department of AI for Ethnopharmacology and Tropical Medicinal Plants, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka, Nigeria.
[2]Department of Computational Screening for Antimalarial Agents, Faculty of Pharmacy, University of Ibadan, Ibadan, Nigeria.

ABSTRACT

Natural products remain a major source of chemically diverse and pharmacologically informative molecules, yet their structural complexity, sparse annotation, assay heterogeneity, and uneven database coverage create persistent barriers for computational bioactivity prediction. Deep learning has expanded the analytical toolkit for natural product research by enabling representation learning from molecular structures, bioactivity records, genomic context, and drug-discovery datasets. This critical review evaluates how deep learning has been applied to natural product bioactivity prediction, with particular attention to molecular representation, feature learning, model performance, reliability, reproducibility, and therapeutic relevance. The review critically appraises evidence across natural product databases, deep neural networks, graph-based models, SMILES-based models, transformer-style architectures, biosynthetic-gene-cluster learning, drug–target interaction prediction, ADMET prediction, uncertainty estimation, explainability, and reproducible computational workflows. The synthesis emphasizes that descriptors, fingerprints, SMILES strings, molecular graphs, and learned embeddings do not merely encode molecules differently; they shape the kinds of structure–activity relationships that models can detect and the errors that may remain hidden. The review also distinguishes retrospective benchmark performance from pharmacological meaning, arguing that predicted bioactivity should be interpreted as prioritization or hypothesis generation rather than therapeutic evidence. Major limitations include dataset imbalance, benchmark leakage, activity cliffs, limited external validation, weak interpretability, incomplete reproducibility, and uncertain translation from predicted activity to biological mechanism or therapeutic value. Future progress requires natural product-specific benchmarks, leakage-aware validation, uncertainty reporting, interpretable modeling, reproducible workflows, and integration with experimental pharmacology, medicinal chemistry, toxicology, and translational validation.

Key Words: *Deep learning, Natural products, Bioactivity prediction, Molecular representation, Model reliability, Reproducibility*

eIJPPR 2025; 15(1):47-57

HOW TO CITE THIS ARTICLE: Okafor C, Bello A, Musa I. Deep Learning for Natural Product Bioactivity Prediction: A Critical Review of Molecular Representation, Model Reliability, and Therapeutic Relevance. Int J Pharm Phytopharmacol Res. 2025;15(1):47-57. https://doi.org/10.51847/MHJQ75shnT

INTRODUCTION

Natural products provide a distinctive chemical foundation for drug discovery because they occupy biologically shaped regions of chemical space and often contain stereochemical, scaffold, and functional-group complexity not fully represented in conventional synthetic libraries. At

Corresponding author: Chinedu Okafor
Address: Department of AI for Ethnopharmacology and Tropical Medicinal Plants, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka, Nigeria.
E-mail: ✉ chinedu.okafor@unn.edu.ng
Received: 15 November 2024; **Revised:** 24 February 2025; **Accepted:** 26 February 2025

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

the same time, the computational use of natural products is constrained by uneven database coverage, inconsistent bioactivity annotations, and variable links between structures, biological sources, and assay metadata. Artificial intelligence has therefore become attractive not because it removes these constraints, but because it can help organize, prioritize, and interrogate structurally complex natural product evidence when the underlying data are sufficiently curated and biologically interpretable [1, 2].

Natural product bioactivity prediction depends heavily on the quality of activity records, taxonomic or species-source information, structural identifiers, and assay context. Databases that connect natural products with source organisms and activity records create important infrastructure for computational screening, but they also expose a central challenge: bioactivity data are not uniform measurements of therapeutic value, and many records reflect heterogeneous endpoints, assay formats, concentrations, and biological systems [3]. This matters because deep learning models trained on mixed activity labels may learn dataset-specific regularities rather than transferable pharmacological relationships.

Open natural product structure collections have expanded the chemical space available for cheminformatics and machine learning, especially by making diverse molecular structures more accessible for representation learning and virtual screening. However, structural availability alone does not guarantee reliable bioactivity prediction, because many natural products remain incompletely annotated, and predicted activity cannot be separated from the biological and experimental conditions under which activity labels were generated [4]. For critical review purposes, this distinction is essential: molecular computability is not equivalent to pharmacological interpretability.

This review critically evaluates deep learning for natural product bioactivity prediction through four linked questions. First, how have deep learning and related machine learning approaches been applied to natural product research? Second, how do molecular descriptors, fingerprints, SMILES strings, molecular graphs, and learned embeddings shape prediction quality and interpretability? Third, how reliable are reported model-performance claims when dataset imbalance, assay heterogeneity, benchmark leakage, scaffold novelty, activity cliffs, and weak external validation are considered? Fourth, how should predicted bioactivity be interpreted in relation to therapeutic relevance, experimental validation, medicinal chemistry, ADMET behavior, and translational pharmacology?

Deep learning in natural product research

Deep learning has entered natural product research through several routes, including structural classification, bioactivity prediction, genome mining, molecular prioritization, drug–target interaction modeling, ADMET estimation, and natural product-inspired screening. One important early use is structural classification, where deep neural networks can learn patterns that distinguish natural product classes and thereby support annotation, dereplication, and chemical-space organization [5]. Such models are valuable because natural product libraries are structurally heterogeneous, but classification should not be confused with bioactivity confirmation; a compound class may suggest plausible biological behavior without proving target engagement, potency, selectivity, or therapeutic relevance.

Machine learning has also been used to connect biosynthetic logic with predicted biological activity, especially in microbial and biosynthetic natural product discovery. Genome mining approaches can use biosynthetic-gene-cluster features to prioritize candidate metabolites or infer likely activity classes, which is useful when molecular structures are unknown, incompletely characterized, or difficult to isolate [6]. The critical limitation is that biosynthetic features are indirect evidence: they may suggest chemical families or biosynthetic capacity, but they do not by themselves establish molecular identity, assay-confirmed activity, or pharmacological mechanism.

A related line of work predicts biological activity from biosynthetic gene clusters, using genomic context as a proxy for chemical and functional potential. This strategy is particularly relevant for microbial natural products, where the chemical product of a gene cluster may be unknown or silent under standard culture conditions [7]. The strength of this approach is its ability to support hypothesis generation beyond structure-only screening, but its reliability depends on the quality of gene-cluster annotation, the completeness of known biosynthetic–activity links, and the representativeness of training data.

Deep learning screens for bioactive molecules illustrate the difference between computational prediction and experimentally supported prioritization. In antimicrobial discovery, deep learning has been used to identify candidates for experimental testing, showing that predictive models can guide screening decisions when paired with biological assays [8]. However, this example also clarifies the interpretive boundary that is central to natural product bioactivity prediction: computational activity scores are not therapeutic claims. They become meaningful only when connected to experimental validation, mechanism-of-action evidence, toxicity assessment, pharmacokinetic considerations, and medicinal chemistry feasibility.

Table 1 maps the major deep learning applications in natural product research across data sources, prediction tasks, model families, validation approaches, and critical limitations.

Table 1. Deep Learning Applications in Natural Product Bioactivity Prediction: Data Sources, Model Families, Prediction Tasks, Validation Approaches, Strengths, Limitations, and Critical Appraisal Dimensions

Application domain	Natural product data source	AI or deep learning model family	Prediction task	Bioactivity endpoint	Validation approach	Main strength	Main limitation	Critical appraisal concern
Natural product structural classification	Curated natural product structures and class labels	Deep neural networks; representation learning models	Chemical class assignment and annotation support	Indirect; class-associated biological plausibility	Internal classification evaluation and label-based testing	Helps organize structurally diverse natural products	Does not directly predict pharmacological activity	Risk of treating structural class as biological evidence
Biosynthetic genome mining	Biosynthetic gene clusters and microbial genomic features	Machine learning and bioinformatics classifiers	Candidate prioritization and activity hypothesis generation	Predicted biological activity class	Model-based validation against known gene-cluster annotations	Extends prediction to metabolites without fully resolved structures	Depends on known gene-cluster–activity associations	Indirect link between genomic signal and compound-level activity
Natural product activity prediction	Natural product structures with curated activity records	QSAR models; deep neural networks; graph-based models	Classification or regression of bioactivity	Antimicrobial, anticancer, enzyme inhibition, or other assay-defined activity	Internal validation, cross-validation, or dataset-specific testing	Supports prioritization of compounds for further study	Sensitive to endpoint heterogeneity and class imbalance	Model may learn assay or dataset artifacts rather than transferable pharmacology
Drug–target interaction prediction	Natural product-like ligands, target sequences, and interaction datasets	CNN, sequence-based deep learning, multimodal encoders	Compound–target affinity or interaction prediction	Target binding or interaction probability	Benchmark evaluation and task-specific validation	Links compounds to possible molecular targets	Often trained on general drug-like datasets rather than natural product-specific data	Predicted target interaction does not establish mechanism in biological systems
ADMET and toxicity prediction	Molecular structures with absorption, distribution, metabolism, excretion, and toxicity annotations	Machine learning and deep learning property predictors	Safety and developability estimation	Toxicity, metabolism, permeability, and related ADMET endpoints	Platform or benchmark validation	Adds translational context beyond potency	Applicability domain may be weak for complex natural products	Safety prediction uncertainty can be underreported
Natural product-inspired screening	Natural product structures, analogues, and screening libraries	Deep learning screeners; representation-learning models	Compound prioritization for experimental assays	Assay-specific activity such as antimicrobial or cytostatic effects	Retrospective testing and, where available, prospective experimental validation	Can reduce screening burden and guide experimental selection	Performance depends on training-library composition	Predicted hits require experimental confirmation and medicinal chemistry follow-up

Molecular representation and feature learning

Molecular representation is a core determinant of model behavior in natural product bioactivity prediction because each representation encodes some chemical features while simplifying or omitting others. Descriptors and fingerprints can capture predefined physicochemical properties or substructure patterns, whereas learned representations and graph models can infer task-relevant features from data; these differences influence performance, interpretability, and generalizability [9–11].

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

For natural products, representation choice is especially consequential because stereochemistry, glycosylation, macrocycles, fused ring systems, unusual heteroatom patterns, and scaffold novelty may not be adequately captured by representations optimized for simpler drug-like molecules.

Unsupervised molecular embeddings provide one route for learning chemically meaningful features without relying entirely on handcrafted descriptors. Mol2vec, for example, treats molecular substructures analogously to words and learns vector representations that can encode recurring chemical environments [12]. This approach is attractive for natural product modeling because substructure patterns can reflect biosynthetic logic and scaffold families, but its value depends on whether the training corpus contains enough natural product-like chemistry to represent rare scaffolds, stereochemical complexity, and underrepresented metabolite classes.

SMILES-based models, including recurrent, convolutional, and transformer-style architectures, offer a text-like route to molecular learning. Their appeal lies in the ability to process molecular strings directly, often with augmentation strategies such as SMILES enumeration or multitask learning to improve model robustness [13]. Yet SMILES is not a neutral molecular language: different strings can represent the same molecule, stereochemical details may be inconsistently encoded, and sequence models may learn syntax-associated shortcuts rather than chemically generalizable structure–activity relationships. For complex natural products, this creates a need for careful canonicalization, augmentation control, and external testing.

Graph representations are conceptually well aligned with chemical structure because atoms and bonds can be modeled as nodes and edges, allowing message-passing or attention mechanisms to learn local and extended molecular environments. This is useful for natural products with dense stereochemical and scaffold information, but graph learning also faces limitations when bioactivity depends on conformational behavior, three-dimensional shape, macrocyclic flexibility, solubility, metabolism, or assay-specific biological context. Multimodal representations that integrate structure, bioactivity, biosynthetic origin, spectral data, omics information, or target context may improve biological relevance, but they also increase dependence on harmonized, high-quality metadata and transparent validation.

Table 2 compares molecular representation strategies used for natural product bioactivity prediction, highlighting their advantages, constraints, and implications for model reliability.

Table 2. Molecular Representation and Feature Learning Strategies for Natural Product Bioactivity Prediction: Descriptors, Fingerprints, SMILES, Graphs, Embeddings, Multimodal Inputs, Strengths, Constraints, and Reliability Implications

Representation strategy	Input information captured	Typical model compatibility	Feature learning mechanism	Strength for natural product modeling	Limitation for structural diversity	Reliability concern	Interpretability implication	Best-use scenario
Molecular descriptors	Physicochemical, topological, electronic, or structural summary features	QSAR models, random forests, support vector machines, neural networks	Mostly predefined features with model-based weighting	Provides interpretable chemical summaries and supports small datasets	May miss rare scaffolds, stereochemistry, macrocycles, and local substructure context	Descriptor selection can bias conclusions	Often more interpretable than learned embeddings	Small or moderately sized curated datasets with clear endpoints
Molecular fingerprints	Presence or absence of substructures or atom environments	Classical ML, neural networks, similarity models	Encodes predefined local structural patterns	Useful for scaffold comparison and activity-class modeling	Sparse encoding may underrepresent unusual natural product motifs	Similarity-based learning may fail on novel scaffolds	Substructure bits can be partially interpretable	Baseline bioactivity classification and similarity-guided prioritization
SMILES strings	Linearized molecular syntax, atoms, bonds, branches, and	CNNs, RNNs, transformers, sequence models	Learns syntax and sequence patterns from	Easy to scale and compatible with language-model architectures	Multiple valid strings, stereochemical ambiguity, and	Model may learn string artifacts rather than chemistry	Usually difficult to interpret without	Large datasets with careful standardization and augmentation

	stereochemical tokens when encoded		molecular strings		syntax shortcuts		attribution methods	
Molecular graphs	Atom-bond topology, neighborhoods, and connectivity patterns	Graph neural networks, graph attention networks, message-passing models	Learns atom and bond features through neighborhood aggregation	Chemically natural representation for complex scaffolds	May not capture 3D conformation, tautomerism, or assay context	Scaffold generalization remains uncertain	Node and edge attributions can aid interpretation but require validation	Structure-rich datasets where topology is central to activity
Learned embeddings	Latent vector representations of molecules or substructures	Neural networks, transfer learning, clustering, downstream predictors	Learns features from molecular corpora or supervised tasks	Can capture recurring chemical environments and transfer features	Training corpus may underrepresent natural product-like chemistry	Out-of-domain embeddings may appear confident but fail	Latent dimensions are often difficult to explain	Transfer learning or feature extraction when labeled NP data are sparse
Multimodal representations	Molecular structures plus target, genomic, omics, spectral, source, or assay metadata	Multimodal deep learning, multitask models, attention models	Joint learning across heterogeneous data types	Can connect structure with biosynthesis, source, and biological context	Requires harmonized metadata and consistent identifiers	Missing data and modality imbalance can distort learning	Interpretation becomes more complex across modalities	Mechanism-informed prediction and integrated prioritization workflows
Three-dimensional or conformer-aware representations	Molecular geometry, conformers, stereochemical arrangement, and spatial features	3D graph models, geometric deep learning, docking-informed ML	Learns from spatial molecular arrangements	Relevant for stereochemically rich and macrocyclic natural products	High computational cost and conformer uncertainty	Incorrect conformers can mislead predictions	Potentially mechanistically informative but harder to validate	Target-aware modeling where geometry is biologically important
Biosynthetic or source-based representations	Gene clusters, taxonomy, producing organism, or biosynthetic context	ML classifiers, genomic models, multimodal frameworks	Learns associations between biosynthetic features and activity labels	Useful when structure is unknown or incomplete	Indirect relationship between gene cluster and final compound	Annotation bias and incomplete activity linkage	Biological context may improve interpretability if curated	Microbial natural product discovery and genome-mining prioritization

Bioactivity prediction and model performance

Bioactivity prediction in molecular machine learning is often framed through benchmark datasets that allow comparison across models, representations, and validation strategies. Benchmarks such as MoleculeNet have helped standardize molecular prediction tasks and made it easier to compare descriptors, fingerprints, graph models, and neural architectures across property and activity endpoints [14]. For natural product research, however, benchmark performance should be interpreted cautiously because common benchmarks are not necessarily representative of natural product chemical space, biosynthetic complexity, assay heterogeneity, or scaffold novelty. A model that performs well on a general molecular benchmark may still fail when applied to stereochemically rich, macrocyclic, glycosylated, or otherwise underrepresented natural product scaffolds.

Natural product-specific bioactivity prediction studies demonstrate the practical value and limits of machine learning. Predictive classifiers for antimalarial natural products show how curated activity records and molecular features can be used to prioritize compounds for further investigation [15]. Such models are useful when the prediction task is clearly defined, the endpoint is biologically meaningful, and the training labels are sufficiently consistent. The critical issue is that endpoint-specific classifiers cannot be generalized automatically to broader therapeutic claims: a predicted antimalarial label is not equivalent to validated potency, selectivity, mechanism of action, safety, pharmacokinetics, or clinical relevance.

Drug–target interaction modeling provides another bioactivity-prediction route by linking compounds to protein targets or binding-affinity estimates. Deep learning models that encode compounds and protein sequences can predict drug–target binding affinity and therefore support target-hypothesis generation for natural products or natural product-like molecules [16]. Yet these models often depend on benchmark interaction datasets dominated by synthetic or drug-like compounds, which may limit their applicability to natural product scaffolds. For natural products, target prediction should therefore be treated as a hypothesis-generating layer that requires biochemical, cellular, and mechanistic validation.

Scalable drug–target prediction frameworks can compare multiple encoders and model families, making them useful for computational screening workflows and reproducible experimentation [17]. Bioactivity prediction also needs ADMET and toxicity context, because activity without acceptable absorption, distribution, metabolism, excretion, and safety properties has limited translational value [18]. Recent natural product-focused deep learning work on cytostatic activity patterns further illustrates that models can detect biologically relevant activity signatures, but these signatures remain endpoint- and context-dependent rather than direct evidence of therapeutic efficacy [19]. Overall, model performance should be reported with clear endpoint definitions, validation design, applicability-domain analysis, and caution against interpreting numerical accuracy as pharmacological success.

Reliability, bias, and reproducibility concerns

Reliability is a central concern in deep learning-based natural product bioactivity prediction because models can appear accurate while learning non-generalizable patterns. Ligand-based benchmark studies have shown that some classification settings may reward memorization rather than true generalization, especially when chemically similar compounds appear across training and test splits [20]. Reanalyses of bioactivity prediction comparisons further show that conclusions about model superiority can change when evaluation choices, data handling, or benchmarking assumptions are reconsidered [21]. These findings are highly relevant to natural products because structurally related analogues, duplicated records, unevenly curated activity labels, and scaffold-biased datasets may inflate apparent performance.

Uncertainty estimation is essential because a model's point prediction does not reveal whether the prediction is reliable for a specific natural product scaffold, endpoint, or assay context [22]. Benchmark performance can be distorted by memorization, evaluation design, and local structure–activity discontinuities, including activity cliffs where structurally similar compounds display sharply different activities [20, 21, 23]. For natural products, this problem is especially important because minor stereochemical, oxidation-state, glycosylation, or substitution differences may substantially alter biological activity. Reliability assessment should therefore include scaffold-aware validation, uncertainty reporting, applicability-domain analysis, and careful treatment of structurally close analogues.

Interpretability is often presented as a solution to black-box modeling, but explanations can themselves be unreliable if they are not tested against chemically meaningful failure cases. Feature-attribution methods benchmarked with activity cliffs show that explanation quality must be evaluated under difficult structure–activity conditions rather than accepted at face value [24]. In natural product research, interpretability should help chemists and pharmacologists understand which structural motifs, physicochemical features, or biological-context signals influence predictions. However, visually appealing attribution maps or attention weights should not be treated as mechanistic evidence unless they align with experimental chemistry and pharmacology.

Reproducibility affects every stage of computational natural product bioactivity prediction, from data cleaning and molecular standardization to model training, validation, reporting, and reuse. Reproducible computational drug discovery requires transparent workflows, accessible data where possible, clear software environments, versioned code, explicit hyperparameters, and reporting practices that allow independent checking [25]. Natural product studies face additional reproducibility challenges because compound identity, stereochemistry, source organism, isolation conditions, assay protocol, and endpoint definition may be inconsistently reported. Without reproducible workflows and transparent validation, deep learning predictions may be difficult to audit, compare, or translate into experimental research.

Table 3 synthesizes the main reliability, bias, reproducibility, and validation concerns that affect deep learning-based natural product bioactivity prediction.

Table 3. Reliability, Bias, and Reproducibility Concerns in Deep Learning-Based Natural Product Bioactivity Prediction: Dataset Limitations, Benchmark Risks, Validation Requirements, Interpretability Challenges, and Mitigation Strategies

Concern	Where it appears in the modeling pipeline	Underlying cause	Effect on model performance interpretation	Validation requirement	Reproducibility implication	Therapeutic relevance risk	Mitigation strategy	Future research need
Dataset imbalance	Dataset assembly and training	Overrepresentation of common scaffolds, endpoints, or active compounds	Inflates performance on majority classes while hiding poor minority-class prediction	Class-stratified and endpoint-aware evaluation	Requires transparent label distributions	May overlook rare but important natural product chemotypes	Balanced sampling, class weighting, and transparent reporting	Natural product-specific balanced benchmark datasets
Assay heterogeneity	Bioactivity-label generation	Different assay formats, concentrations, organisms, cell lines, and readouts	Makes labels appear comparable when they may not be biologically equivalent	Endpoint harmonization and assay-specific validation	Requires detailed assay metadata	Predicted activity may not transfer across biological contexts	Separate endpoints by assay type and biological system	Standardized bioactivity annotation for natural products
Duplicate or near-duplicate compounds	Data preprocessing and splitting	Repeated records, analogues, salts, stereochemical variants, or database overlaps	Can cause memorization and overoptimistic test results	Duplicate removal and scaffold-aware splitting	Requires documented molecular standardization	Apparent activity may reflect repeated data rather than generalizable learning	Deduplication by structure, identifiers, and stereochemical checks	Shared preprocessing protocols
Benchmark leakage	Model evaluation	Chemically similar or identical compounds across training and test sets	Overstates generalization capacity	Leakage-aware train-test separation	Requires public split definitions	Misleads compound prioritization decisions	Scaffold splits, temporal splits, and external datasets	Benchmark designs aligned with natural product novelty
Activity cliffs	Prediction and interpretation	Small structural changes causing large bioactivity changes	High average performance may hide local prediction failure	Activity-cliff-sensitive testing	Requires reporting difficult local SAR cases	May misprioritize close analogues with different effects	Local uncertainty estimation and SAR-aware evaluation	Natural product activity-cliff datasets
Weak external validation	Post-model evaluation	Reliance on internal cross-validation or random splits	Limits confidence in real-world deployment	Independent datasets and prospective testing	Requires shareable validation sets	Predicted hits may fail experimentally	External validation and prospective assay follow-up	Community external-validation challenges
Poor interpretability	Model explanation stage	Black-box architectures or unvalidated attribution methods	Makes predictions difficult to chemically assess	Explanation benchmarking and expert review	Requires reproducible explanation workflows	May create false mechanistic confidence	Use interpretable baselines and validated attribution methods	Natural product-specific explainability evaluation
Limited uncertainty reporting	Prediction deployment	Focus on point estimates rather than calibrated confidence	Makes unreliable predictions appear equally credible	Calibration testing and uncertainty estimates	Requires reporting confidence and calibration metrics	High-risk predictions may be overinterpreted	Ensembles, conformal methods, Bayesian approximations, and	Standard uncertainty-reporting expectations

							applicability-domain flags	
Incomplete code or data availability	Reproducibility and reuse	Closed workflows, missing hyperparameters, unavailable preprocessing scripts	Prevents independent verification	Open code, versioned environments, and detailed methods	Directly limits reproducibility	Undermines confidence in prioritization claims	FAIR data practices and executable workflows	Reproducible natural product ML repositories

Therapeutic relevance and translational limitations

The central translational limitation of deep learning-based natural product bioactivity prediction is that predicted activity is not equivalent to therapeutic relevance. Machine learning can support multiple stages of drug discovery and development, including screening, target hypothesis generation, lead prioritization, and property prediction, but each stage requires different evidence standards [26]. In natural product research, a model may identify a compound as potentially active, yet therapeutic relevance still depends on isolation or synthesis feasibility, structural confirmation, dose–response behavior, mechanism of action, selectivity, toxicity, pharmacokinetics, pharmacodynamics, and biological context.

Critical discussions of artificial intelligence in drug discovery warn against confusing model output with real-world impact. High predictive performance, elegant architectures, or large datasets do not automatically solve the problems of biological causality, clinical translation, assay validity, or medicinal chemistry optimization [27]. This caution is particularly important for natural products because complex mixtures, stereochemical variants, source-dependent composition, low compound availability, and difficult synthesis can interrupt the path from computational prioritization to experimental validation. A predicted natural product hit should therefore be described as a candidate for investigation, not as a therapeutic lead unless supported by appropriate experimental evidence.

Prospective AI-enabled discovery studies show that computational models can contribute meaningfully when predictions are followed by synthesis or procurement, biochemical testing, and experimental validation [28]. The lesson for natural product bioactivity prediction is not that deep learning alone establishes drug-like value, but that computational prioritization can become useful when embedded in an iterative discovery loop. Such a loop should include compound verification, orthogonal assays, mechanism testing, cytotoxicity or selectivity assessment, ADMET profiling, and medicinal chemistry evaluation before translational claims are made.

Translational implementation also requires attention to model lifecycle, deployment context, and domain expertise. Machine learning for small-molecule discovery must be integrated with data preparation, validation strategy, uncertainty analysis, user interpretation, and experimental decision-making [29]. In natural product research, this integration is especially demanding because computational predictions must interface with natural product chemistry, dereplication, compound isolation, structural elucidation, supply constraints, biological assay design, and pharmacological interpretation. The most defensible role for deep learning is therefore to support prioritization and hypothesis generation while preserving the authority of experimental pharmacology and medicinal chemistry in judging therapeutic promise.

Critical synthesis

The reviewed evidence suggests that deep learning for natural product bioactivity prediction is most mature where tasks are clearly bounded, data are curated, endpoints are well defined, and validation is aligned with the intended use. Natural product virtual screening and bioprospecting benefit from computational prioritization because chemical diversity is large and experimental resources are limited [30]. However, screening value depends on whether molecular representations capture relevant natural product chemistry and whether predicted outputs are connected to realistic experimental follow-up. The strongest applications are therefore not fully automated discovery pipelines, but decision-support systems that help researchers choose which compounds, extracts, targets, or hypotheses deserve closer investigation.

Across the evidence base, molecular representation emerges as a major source of both opportunity and fragility. Descriptors and fingerprints remain useful because they are transparent and effective for small or curated datasets, but they may underrepresent complex scaffold features. SMILES and transformer-style models scale well but can learn syntax-associated shortcuts. Graph models align more directly with atom-bond structure, yet they may still miss conformational, stereochemical, and biological-context effects. Learned embeddings can transfer information across tasks, but their reliability depends on the chemical space represented during training. Explainable AI can support model interpretation, but explanation methods must produce chemically meaningful insight rather than post hoc visual reassurance [31].

Figure 1 presents a critical synthesis framework linking molecular representation, deep learning model design, bioactivity prediction, reliability assessment, and therapeutic relevance in natural product research.

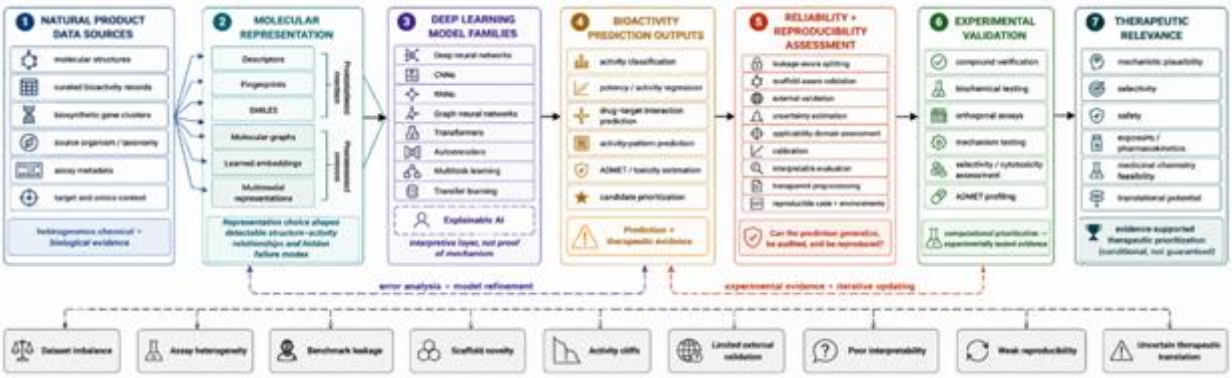

Figure 1. Critical Synthesis Framework for Deep Learning-Based Natural Product Bioactivity Prediction

Alt-text description

A conceptual framework showing natural product datasets and molecular structures flowing into representation strategies such as descriptors, fingerprints, SMILES, graphs, and learned embeddings. These representations feed into deep learning model families such as neural networks, graph neural networks, transformers, autoencoders, and multitask models. The framework connects model outputs to bioactivity prediction, reliability assessment, reproducibility safeguards, experimental validation, and therapeutic relevance, with feedback loops for model refinement.

A second synthesis point is that bioactivity prediction should be separated into endpoint prediction, biological interpretation, and therapeutic relevance. Endpoint prediction asks whether a model can reproduce or infer activity labels under defined conditions. Biological interpretation asks whether predicted activity is plausible in relation to target engagement, pathway modulation, selectivity, or phenotype. Therapeutic relevance asks whether the compound can plausibly advance through toxicity, exposure, efficacy, formulation, and validation requirements. Deep learning can contribute to the first layer and sometimes support the second, but it cannot independently establish the third.

A third synthesis point is that reliability safeguards should be treated as core components of model design rather than optional reporting additions. Natural product bioactivity models should include leakage-aware splitting, external validation where possible, uncertainty estimation, applicability-domain assessment, endpoint-specific reporting, and transparent preprocessing. Reproducibility should also be defined broadly: it includes not only code availability but also molecular standardization rules, stereochemistry handling, database versions, assay-label definitions, hyperparameter settings, model seeds, and split files. Without these safeguards, claims of model performance may be difficult to interpret and impossible to verify.

Finally, future research should move from architecture-centered claims toward evidence-centered workflows. Broad machine learning reviews of drug discovery show that AI methods can contribute across hit discovery, lead optimization, target prediction, and property prediction, but they also show that successful use depends on task definition, validation, data quality, and integration into scientific decision-making [32]. For natural product bioactivity prediction, the most important agenda is therefore not simply to apply larger models, but to build better natural product-specific datasets, evaluate models against scaffold novelty and activity cliffs, report uncertainty, improve interpretability, and connect predictions to experimental pharmacology.

CONCLUSION

Deep learning can support natural product bioactivity prediction by improving compound prioritization, molecular screening, representation learning, and hypothesis generation across structurally diverse chemical spaces. Its value, however, depends on the quality and biological meaning of the data used to train and validate models. Molecular representations shape what models can learn, and benchmark performance must be interpreted through the lens of dataset bias, assay heterogeneity, scaffold novelty, activity cliffs, uncertainty, interpretability, and reproducibility. Predicted bioactivity should not be treated as therapeutic evidence unless it is supported by experimental validation, mechanistic plausibility, ADMET and toxicity assessment, medicinal chemistry evaluation, and translational pharmacology. A credible future for deep learning in natural product research requires leakage-aware benchmarks, transparent workflows, uncertainty-aware prediction, chemically

meaningful explanations, natural product-specific validation datasets, and close integration with natural product chemistry, pharmacology, and medicinal chemistry.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

1. 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
2. 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
3. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 2018;46(D1):D1217-22. doi:10.1093/nar/gkx1026
4. 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
5. 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
6. Yuan Y, Shi C, Zhao H. Machine learning-enabled genome mining and bioactivity prediction of natural products. ACS Synth Biol. 2023;12(9):2650-62. doi:10.1021/acssynbio.3c00234
7. Walker AS, Clardy J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. J Chem Inf Model. 2021;61(6):2560-71. doi:10.1021/acs.jcim.0c01304
8. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.e13. doi:10.1016/j.cell.2020.01.021
9. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide. J Cheminform. 2020;12(1):56. doi:10.1186/s13321-020-00460-5
10. Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237
11. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2020;63(16):8749-60. doi:10.1021/acs.jmedchem.9b00959
12. Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition. J Chem Inf Model. 2018;58(1):27-35. doi:10.1021/acs.jcim.7b00616
13. Zhang XC, Wu CK, Yi JC, Zeng XX, Yang CQ, Lu AP, et al. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. Research (Wash D C). 2022;2022:0004. doi:10.34133/research.0004
14. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
15. Egieyeh S, Syce J, Malan SF, Christoffels A. Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. PLoS One. 2018;13(9). doi:10.1371/journal.pone.0204644
16. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17). doi:10.1093/bioinformatics/bty593
17. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: A deep learning library for drug-target interaction prediction. Bioinformatics. 2020;36(22-23):5545-7. doi:10.1093/bioinformatics/btaa1005
18. Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1). doi:10.1093/nar/gkab255
19. Gahl M, Kim HW, Glukhov E, Gerwick WH, Cottrell GW. PECAN predicts patterns of cancer cell cytostatic activity of natural products using deep learning. J Nat Prod. 2024;87(3):567-75. doi:10.1021/acs.jnatprod.3c00879
20. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403

21. Robinson MC, Glen RC, Lee AA. Validating the validation: Reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. J Comput Aided Mol Des. 2020;34(7):717-30. doi:10.1007/s10822-019-00274-0
22. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
23. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
24. Jiménez-Luna J, Skalic M, Weskamp N, Schneider G. Benchmarking molecular feature attribution methods with activity cliffs. J Chem Inf Model. 2022;62(2):274-83. doi:10.1021/acs.jcim.1c01163
25. Schaduangrat N, Lampa S, Simeon S, Gleeson MP, Spjuth O, Nantasenamat C. Towards reproducible computational drug discovery. J Cheminform. 2020;12(1):9. doi:10.1186/s13321-020-0408-x
26. 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
27. Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
28. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x
29. Volkamer A, Riniker S, Nittinger E, Lanini J, Grisoni F, Evertsson E, et al. Machine learning for small molecule drug discovery in academia and industry. Artif Intell Life Sci. 2023;3:100056. doi:10.1016/j.ailsci.2022.100056
30. 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
31. 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
32. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: A review. Artif Intell Rev. 2022;55(3):1947-99. doi:10.1007/s10462-021-10058-4