Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 20 March 2026
Revised 24 April 2026
Accepted 26 May 2026

Natural Resources for Human Health 2026, Vol. 6 (3): 584–592

KEYWORDS:
natural product-likeness; drug-likeness; cheminformatics; machine learning; molecular descriptors; QED; phytochemical databases; chemical space

Natural Product-Likeness Scores Revisited: Machine Learning Models for Predicting Drug-Like Properties from Phytochemical Databases

Nirali Patel[1], Neha P Singh[2], Manvinder Brar[3], K.V. Navin Raja[4], Satish N. Pawar[5], Dilnavoz Jalilova[6], Vinitha M[7], Seethaladevi S[8]

[1]Department of Pharmaceutical Chemistry, Parul Institute of Pharmacy, Parul University, Vadodara, Gujarat, India. Nirali.patel16204@paruluniversity.ac.in
[2]School of Pharmacy and Research, Dev Bhoomi Uttarakhand University. nehusingh2711@gmail.com
[3]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India. manvinder.brar.orp@chitkara.edu.in
[4]Department of Pharmacology, Vinayaka Missions Medical College, Karaikal, Vinayaka Mission's Research Foundation (DU), Tamil Nadu, India. kothengudidoctor@gmail.com
[5]Department of Commerce & Management, Sri Balaji Society's Balaji College of Arts, Commerce & Science, Tathawade, Pune, Maharashtra 411033, India. satishnpawar2024@gmail.com
[6]Kimyo International University in Tashkent, Uzbekistan. d.jalilova@kiut.uz
[7]Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research. vinitham@maher.ac.in
[8]Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research. seethaladevi@maher.ac.in

ABSTRACT: Natural products remain among the richest sources of drug leads, yet the cheminformatic tools used to prioritise molecules were largely developed with synthetic compounds in mind. Natural products occupy a distinct region of chemical space—characteristically higher in molecular weight, oxygen content, fraction of sp^3 carbons and stereocentres, and lower in nitrogen, lipophilicity and aromaticity than synthetic drugs—and roughly one in five violate Lipinski's "Rule of Five" while remaining bioactive. This mismatch motivates a fresh look at the two families of scores that guide natural-product-based discovery: drug-likeness metrics (the Rule of Five, Veber, Ghose, Egan, Muegge and the quantitative estimate of drug-likeness, QED) and the natural product-likeness (NP-likeness) score of Ertl and colleagues, a Bayesian fragment-based measure that cleanly separates natural products from synthetic molecules. This review revisits these scores in the machine-learning era.

We summarise the classical rules and their limitations for natural products; describe the NP-likeness score and its open-source and neural-network reimplementations; quantify the systematic physicochemical differences between natural-product and synthetic chemical space; survey the major phytochemical databases (COCONUT, LOTUS, NPASS, KNApSAcK) that now supply training data at scale; and examine how molecular representations (physicochemical descriptors, fingerprints such as ECFP/Morgan, and graph- and SMILES-based encodings) feed machine-learning models—random forests,

Natural Resources for Human Health
eISSN: 2583-1194

gradient boosting, deep and graph neural networks—that predict drug-likeness, ADMET and bioactivity. We conclude that data-driven, natural-product-aware models are superseding rigid rule-based filters, but that their reliability depends on data quality, an honestly defined applicability domain, interpretability, and resistance to the historical bias toward synthetic, Rule-of-Five-compliant chemistry.

1. INTRODUCTION

Natural products (NPs)—the secondary metabolites of plants and other organisms—have been optimised over evolutionary time for interaction with biological macromolecules, and they remain a pre-eminent source of drug leads and validated pharmacophores. [1,2] Modern drug discovery, however, is dominated by computational triage: before a molecule is synthesised or assayed, its "druggability" is estimated by cheminformatic scores. The difficulty is that most of these tools were calibrated on synthetic compounds, whose chemical space differs systematically from that of natural products (Figure 1). [2,3] Natural products tend to be larger and more three-dimensional—richer in sp^3 carbons, stereocentres and oxygen, poorer in nitrogen and aromatic rings—and, as a consequence, roughly 20% of them fall outside Lipinski's Rule of Five, the most widely used drug-likeness filter, even though many such molecules are potent, orally active drugs. [3,4]

This mismatch has real cost: applying synthetic-centric filters to natural-product libraries risks discarding exactly the complex, bioactive chemistry that makes NPs valuable. [3,5] Two families of score bear on the problem. Drug-likeness metrics—the Rule of Five and its successors—estimate whether a molecule resembles known oral drugs; the natural product-likeness (NP-likeness) score, introduced by Ertl and colleagues, instead measures how closely a molecule resembles the structural space of natural products. [6,7] Both were devised in an era of rule-based cheminformatics, and both merit revisiting now that large phytochemical databases and machine learning allow drug-like properties to be *learned* from data rather than encoded as fixed thresholds.

This review undertakes that revisitation. We first summarise the classical drug-likeness rules and their limitations for natural products; we then describe the NP-likeness score and its modern reimplementations, quantify the physicochemical differences between natural-product and synthetic chemical space, and survey the databases that now supply training data at scale. Finally, we examine how molecular representations and machine-learning models predict drug-likeness, ADMET and bioactivity from phytochemical data, and we assess the challenges that determine whether such models can be trusted.

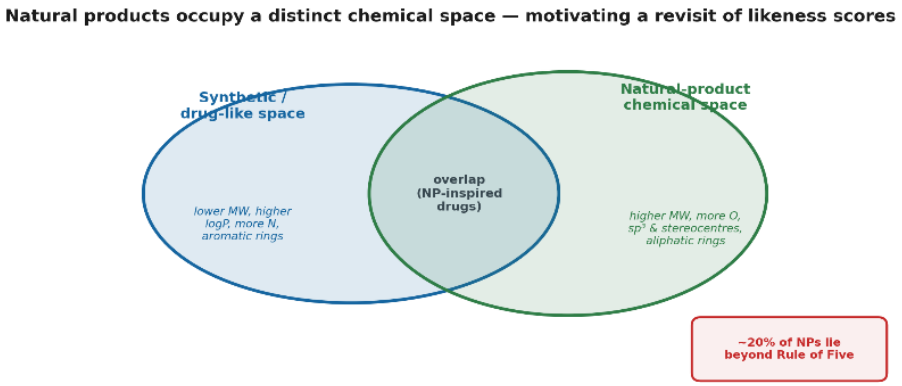

Figure 1. Natural products occupy a distinct, though overlapping, region of chemical space relative to synthetic drug-like molecules. They tend toward higher molecular weight, oxygen content, sp^3 character and stereochemical complexity and lower nitrogen, lipophilicity and aromaticity; approximately 20% lie beyond Lipinski's Rule of Five. This mismatch with synthetic-centric filters motivates revisiting likeness scores in the machine-learning era.

Natural Resources for Human Health
eISSN: 2583-1194

2. CLASSICAL DRUG-LIKENESS RULES AND THEIR LIMITS

The paradigm of rule-based drug-likeness began with Lipinski's Rule of Five (Ro5), which flags poor oral absorption when a molecule has more than five hydrogen-bond donors, ten acceptors, a molecular weight above 500 Da or a calculated logP above 5. [8] Successive refinements added complementary constraints (Table 1): Veber's rules on rotatable bonds and polar surface area, and the Ghose, Egan and Muegge filters, while the quantitative estimate of drug-likeness (QED) replaced hard thresholds with a continuous desirability score integrating several properties. [9,10] These rules brought welcome discipline to synthetic library design.

Rule / score	Key criteria	Note
Lipinski Rule of Five	MW ≤ 500; logP ≤ 5; HBD ≤ 5; HBA ≤ 10	Oral absorption; most widely used
Veber	Rotatable bonds ≤ 10; TPSA ≤ 140 Å²	Oral bioavailability
Ghose	MW 160–480; logP −0.4–5.6; atoms 20–70	Drug-like range filter
Egan	logP and TPSA boundaries	Absorption model
Muegge	Pharmacophore / property filter	Stringent drug-likeness
QED	Weighted desirability (0–1)	Continuous, not pass/fail

Table 1. Classical drug-likeness rules and scores. Threshold-based filters (Ro5, Veber, Ghose, Egan, Muegge) give binary pass/fail verdicts, whereas QED provides a continuous desirability score. All were calibrated chiefly on synthetic, orally administered drugs.

For natural products, however, these rules are an imperfect fit. Lipinski himself noted that the Ro5 does not apply to substrates of biological transporters, and orally active natural products such as paclitaxel, rapamycin and vinorelbine violate it substantially—some exceeding 750 Da—yet reach their targets. [4,5] Because the rules encode the properties of synthetic drugs, they penalise the very features (size, sp^3 richness, stereochemical complexity) that distinguish natural products and that correlate with binding selectivity and clinical success. [3,11] A likeness framework tuned to natural-product chemistry is therefore needed.

3. THE NATURAL PRODUCT-LIKENESS SCORE

The NP-likeness score, introduced by Ertl, Roggo and Schuffenhauer in 2008, provides exactly such a framework. [6] It is a Bayesian measure built from the frequencies of structural fragments in a reference set of natural products versus a reference set of synthetic molecules: a molecule composed of fragments common in natural products scores positive, one dominated by synthetic-typical fragments scores negative. The score efficiently separates the two classes in cross-validation (Figure 2), and—notably—molecules with an NP-likeness near zero (roughly between −1 and +1) are enriched in known drugs, marking a productive middle ground between synthetic simplicity and natural-product complexity. [6,7]

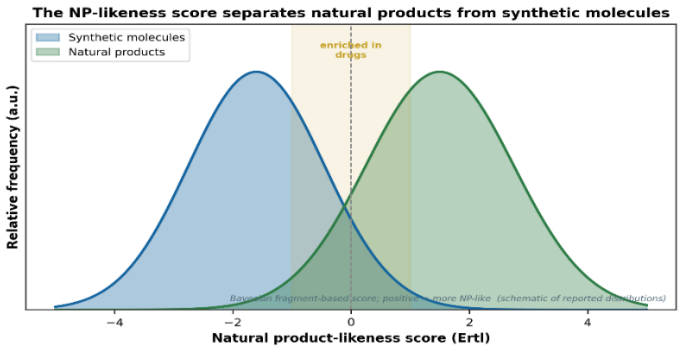

Figure 2. The natural product-likeness (NP-likeness) score, a Bayesian fragment-based measure, separates natural products (positive scores) from synthetic molecules (negative scores). The intermediate region (approximately −1 to +1, shaded) is enriched in approved drugs. The curves are a schematic of the distributions reported for the score.

The score has since been made more accessible and more powerful. An open-source, open-data reimplementation reproduced the original with high fidelity (correlation r ≈ 0.94–0.97), removing dependence on proprietary training sets and toolkits, and web microservices now expose the calculation without local installation. [7] More recently, neural-network approaches have learned NP-likeness directly from data: scores extracted from artificial neural networks agree closely with the fragment-based score on databases such as ZINC and COCONUT while capturing additional structural nuance. [12] The trajectory—from a hand-crafted Bayesian statistic to a learned representation—mirrors the broader shift this review documents.

4. NATURAL-PRODUCT VERSUS SYNTHETIC CHEMICAL SPACE

Quantifying how natural products differ from synthetic drugs is the empirical foundation for any natural-product-aware model. Cheminformatic analyses across many databases converge on a consistent picture (Figure 3, Table 2): relative to synthetic drug-like compounds, natural products have higher molecular weight, more oxygen atoms but fewer nitrogen and halogen atoms, more hydrogen-bond donors and acceptors, lower calculated logP, and greater molecular rigidity, with a marked preference for aliphatic over aromatic rings. [2,3] Two descriptors of three-dimensional complexity are especially discriminating: the fraction of sp^3-hybridised carbons (Fsp^3) and the number of stereocentres are both substantially higher in natural products. [11] These are not merely descriptive curiosities—Fsp^3 and stereochemical content correlate with successful progression from lead to approved drug, which is one reason natural-product complexity is prized rather than penalised in modern medicinal chemistry. [11]

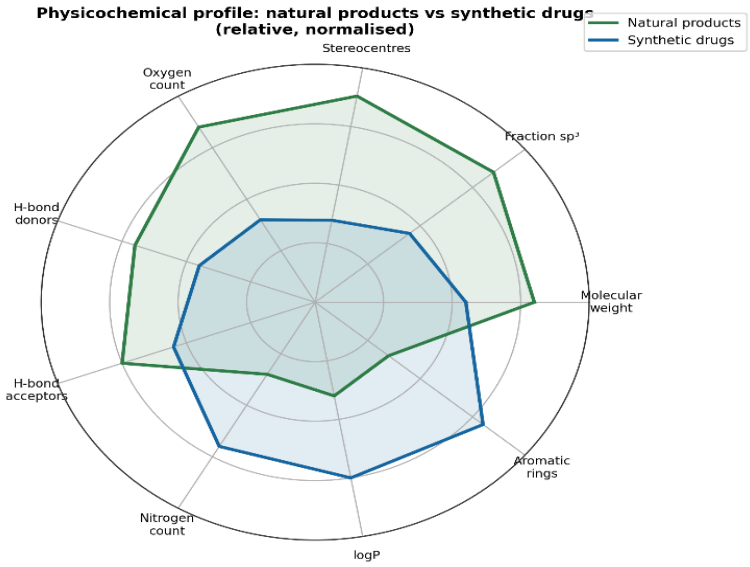

Figure 3. Physicochemical profile of natural products versus synthetic drugs across nine descriptors (relative, normalised). Natural products score higher on molecular weight, fraction sp^3, stereocentres, oxygen count and hydrogen-bonding, and lower on nitrogen count, logP and aromatic rings—a systematic divergence that synthetic-centric filters fail to capture.

Descriptor	Natural products vs synthetic drugs	Relevance
Molecular weight	Higher (wider range)	Often exceeds Ro5 limit
Fraction sp^3 (Fsp^3)	Higher	3-D complexity; predicts approval
Stereocentres	More	Selectivity; clinical success
Oxygen atoms	More	H-bonding, polarity
Nitrogen / halogen atoms	Fewer	Synthetic-typical features
H-bond donors / acceptors	More	Can exceed Ro5 limits

Descriptor	Natural products vs synthetic drugs	Relevance
logP	Lower	Polarity / solubility
Aromatic vs aliphatic rings	Fewer aromatic; more aliphatic	Rigidity; 3-D shape

Table 2. Systematic physicochemical differences between natural products and synthetic drugs. Several of these features cause natural products to fail synthetic-calibrated drug-likeness filters despite retaining—or because of—favourable pharmacological properties.

5. PHYTOCHEMICAL DATABASES AS TRAINING DATA

Machine learning needs data, and the past decade has produced natural-product databases of unprecedented scale and openness (Table 3). COCONUT, which aggregates dozens of open collections, contains more than 400,000 structures; NPASS catalogues fewer than 100,000 NPs with associated bioactivity—about a quarter of the known total—while LOTUS, KNApSAcK and NuBBEDB provide curated, often organism-linked phytochemical data. [1,2] These resources, combined with descriptor calculation in toolkits such as RDKit, make it possible to train and validate models on natural-product chemistry directly rather than by extrapolation from synthetic libraries. [2] They also enable large-scale druggability profiling: in one collection of over 100,000 phytochemicals, about one-third of compounds satisfied all major drug-likeness rules simultaneously while roughly three-quarters satisfied at least one, and about one in five fell beyond the Rule of Five (Figure 4)—quantifying both the reach and the limits of rule-based filtering. [13]

Database	Approx. size	Focus
COCONUT	> 400,000	Aggregated open NPs (dozens of sources)
NPASS	< 100,000	NPs with bioactivity data
LOTUS	~ 750,000 entries	Structure–organism pairs
KNApSAcK	tens of thousands	Species–metabolite (phytochemicals)
NuBBEDB	thousands	Brazilian biodiversity NPs
Dictionary of Natural Products	hundreds of thousands	Curated commercial reference

Table 3. Major natural-product and phytochemical databases used as sources of training and validation data. Sizes are approximate and evolving; open resources such as COCONUT and LOTUS have been transformative for data-driven natural-product cheminformatics.

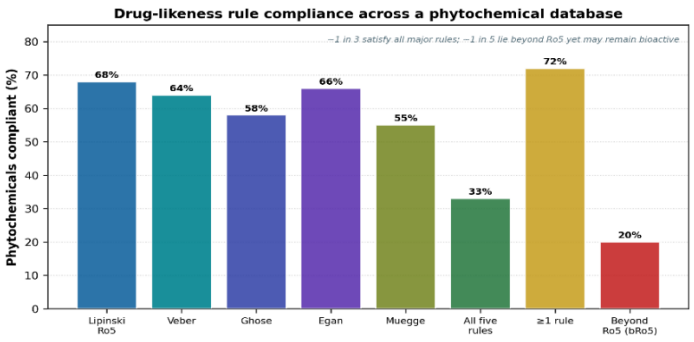

Figure 4. Drug-likeness rule compliance across a large phytochemical collection. Individual filters (Lipinski, Veber, Ghose, Egan, Muegge) pass a majority of compounds, but only about one-third satisfy all major rules simultaneously, roughly three-quarters satisfy at least one, and about one in five lie beyond the Rule of Five (bRo5) yet may remain bioactive. Values are representative figures compiled from database analyses.

Natural Resources for Human Health
eISSN: 2583-1194

6. MACHINE-LEARNING MODELS FOR PROPERTY PREDICTION

Given natural-product-scale data, machine learning can predict drug-like properties directly, replacing rigid thresholds with learned functions (Figure 5). The pipeline has three decisions: how to represent molecules, which model to train, and which property to predict.

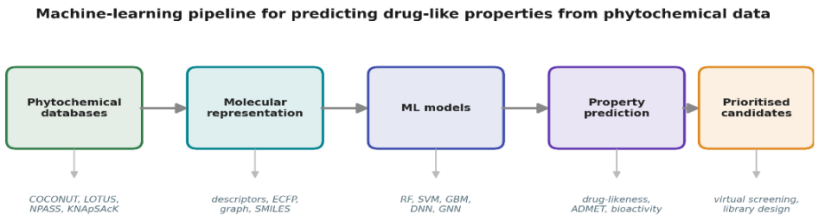

Figure 5. A machine-learning pipeline for predicting drug-like properties from phytochemical data. Structures drawn from natural-product databases are converted to a molecular representation (descriptors, fingerprints, graph or SMILES), used to train a model (from linear methods to graph neural networks), and applied to predict drug-likeness, ADMET or bioactivity—ultimately prioritising candidates for screening and library design.

6.1 Molecular representation

Representation is decisive, and natural products are notoriously hard to encode because of their size and complexity (Table 4). Physicochemical-descriptor vectors (from RDKit and similar tools) are interpretable and capture the properties the rules use. Fingerprints encode substructure presence: circular fingerprints (Morgan/ECFP) generally perform best on natural-product space, while key-based fingerprints (MACCS) are simpler but less expressive, and substructure fingerprints can behave atypically on the larger, more complex NP scaffolds. [14] Graph-based representations (learned by graph neural networks) and SMILES-based sequence models (transformers) learn task-specific features directly from structure and increasingly outperform fixed fingerprints, at some cost in interpretability (Figure 6a). [2,14]

Representation	Type	Notes for natural products
Physicochemical descriptors	Interpretable vectors	Capture rule properties; transparent
MACCS keys	Key-based fingerprint	Simple; limited expressiveness
Morgan / ECFP	Circular fingerprint	Strong general performance on NP space
Graph (GNN)	Learned from molecular graph	Task-specific; high performance
SMILES (transformer)	Learned from sequence	Flexible; large-data hungry

Table 4. Molecular representations for machine learning on natural products. Circular fingerprints are robust general performers, while graph- and sequence-based learned representations achieve the highest accuracy at the expense of interpretability.

6.2 Models and tasks

On these representations, a spectrum of models is deployed (Table 5): interpretable linear and tree-based methods (logistic regression, random forests, gradient boosting) remain strong baselines and are often competitive with deep learning on tabular descriptor data, while deep and graph neural networks excel when large datasets and learned representations are available (Figure 6b). [2,15] The prediction targets span the discovery pipeline: binary or continuous drug-likeness (including QED and NP-likeness themselves), ADMET endpoints (solubility, permeability, blood–brain-barrier penetration, hepatotoxicity), and biological activity against specific targets. Trained on natural-product data, such models prioritise phytochemical libraries in a manner that respects, rather than penalises, natural-product chemistry—the central advantage over fixed rules.

Natural Resources for Human Health
eISSN: 2583-1194

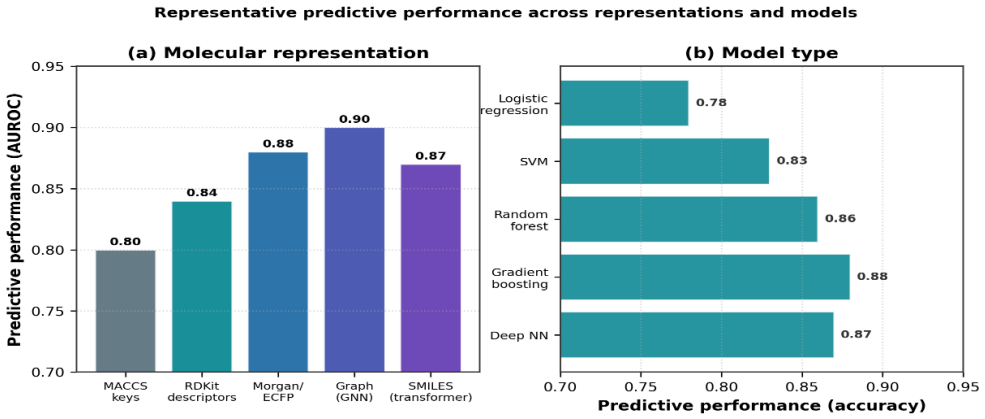

Figure 6. Representative predictive performance for drug-like property prediction. (a) Across molecular representations, learned graph and circular-fingerprint encodings tend to outperform simple key-based fingerprints. (b) Across model types, ensemble (random forest, gradient boosting) and deep methods outperform linear baselines. Values are representative and illustrative; absolute performance depends on task, dataset and validation protocol.

Model class	Typical strength	Prediction tasks
Logistic regression / linear	Interpretable baseline	Drug-likeness classification
SVM	Robust on medium data	Bioactivity, drug-likeness
Random forest / gradient boosting	Strong on descriptor data	Drug-likeness, ADMET
Deep neural networks	High capacity, large data	ADMET, multi-task
Graph neural networks	Learned structural features	Bioactivity, property prediction

Table 5. Machine-learning model classes and their typical roles in predicting drug-like properties, ADMET and bioactivity from natural-product data. Ensemble methods remain strong, interpretable baselines; graph neural networks lead on learned-representation tasks.

7. CHALLENGES AND FUTURE PERSPECTIVES

Several issues determine whether natural-product-aware models can be trusted. Data quality and coverage: natural-product databases contain errors, duplicates and inconsistent stereochemistry, and are biased toward well-studied taxa; models inherit these flaws. Applicability domain: because natural products are structurally diverse and often novel, predictions for molecules outside the training distribution are unreliable, and an honestly defined applicability domain (with calibrated uncertainty) is essential rather than optional. Historical bias: models trained on data curated under Rule-of-Five conventions can perpetuate the very synthetic-centric bias this review criticises, systematically down-ranking legitimate beyond-Ro5 natural products. Interpretability: high-capacity models are hard to trust and act upon in medicinal chemistry, motivating explainable methods and the retention of transparent descriptor-based baselines. Validation rigour: performance must be judged with scaffold- or cluster-based splits rather than random splits, which flatter models by leaking similar structures between training and test sets.[2,11,14] Progress is likely to come from larger and cleaner open databases, natural-product-specific representations and pretrained "foundation" chemical models, multi-task and uncertainty-aware learning, and the tight coupling of prediction with experimental feedback. The broad direction is clear: from rigid, synthetic-calibrated rules toward flexible, data-driven, natural-product-aware models.

8. CONCLUSION

Natural products are a distinct chemical world, and the scores used to navigate them deserve to be revisited in the light of that distinctiveness and of modern machine learning. Classical drug-likeness rules—the Rule of Five and its successors—remain useful heuristics but are calibrated on synthetic chemistry and systematically misjudge the large, sp^3-rich, stereochemically complex molecules that natural products so often are; roughly one in five natural products lie beyond the Rule of Five yet retain drug-like activity. The natural product-likeness score complements these rules by measuring resemblance to natural-product space, and its evolution from a Bayesian fragment statistic to open-source and neural-network implementations tracks the field's broader shift. Powered by large open phytochemical databases and expressive molecular representations, machine-learning models now predict drug-likeness, ADMET and bioactivity directly from natural-product data, replacing fixed thresholds with learned, natural-product-aware functions. Their promise is real but conditional: it depends on data quality, honestly bounded applicability domains, interpretability, rigorous validation and vigilance against inherited synthetic bias. Addressed together, these requirements point toward cheminformatic tools that finally do justice to the chemistry of natural products—and thereby to their enduring value in drug discovery.

Declarations

Funding: No external funding was received for this work.

Conflicts of interest: The authors declare no conflict of interest.

Author contributions: Conceptualisation, A.O.; investigation and writing—original draft, A.O. and A.T.; visualisation, A.T.; writing—review and editing, A.T.; supervision, A.O. All authors approved the final manuscript.

Data availability: No new primary data were generated; all databases and tools discussed are available in the cited references.

REFERENCES

[1] Natural product databases for drug discovery: features and applications. (Journal). 2024.
[2] Exploring chemical space for "druglike" small molecules in the age of AI. Frontiers in Molecular Biosciences. 2025;12:1553667.
[3] Cheminformatic comparison of approved drugs from natural product versus synthetic origins. Bioorganic & Medicinal Chemistry Letters. 2015;25(19):4258–4262.
[4] Doak BC, Over B, Giordanetto F, Kihlberg J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. Chemistry & Biology. 2014;21(9):1115–1142.
[5] Rodrigues T, Reker D, Schneider P, Schneider G. Counting on natural products for drug design. Nature Chemistry. 2016;8(6):531–541.
[6] Ertl P, Roggo S, Schuffenhauer A. Natural product-likeness score and its application for prioritization of compound libraries. Journal of Chemical Information and Modeling. 2008;48(1):68–74.
[7] Jayaseelan KV, Moreno P, Truszkowski A, Ertl P, Steinbeck C. Natural product-likeness score revisited: an open-source, open-data implementation. BMC Bioinformatics. 2012;13:106.
[8] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. Advanced Drug Delivery Reviews. 1997;23(1–3):3–25.
[9] Veber DF, Johnson SR, Cheng HY, et al. Molecular properties that influence the oral bioavailability of drug candidates. Journal of Medicinal Chemistry. 2002;45(12):2615–2623.
[10] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs (QED). Nature Chemistry. 2012;4(2):90–98.
[11] Lovering F, Bikker J, Humblet C. Escape from flatland: increasing saturation (Fsp^3) as an approach to improve clinical success. Journal of Medicinal Chemistry. 2009;52(21):6752–6756.
[12] Menke J, Koch O. Natural product scores and fingerprints extracted from artificial neural networks. Computational and Structural Biotechnology Journal. 2021;19:4593–4602.
[13] Sorokina M, Merseburger P, Rajan K, et al. COCONUT online: the Collection of Open Natural Products database. Journal of Cheminformatics. 2021;13:2.
[14] Capecchi A, Reymond JL. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. Journal of Cheminformatics. 2024;16:33.

[15] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463–477.