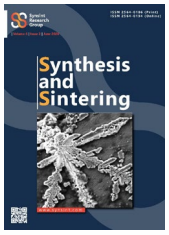
Synsint Research Group

Available online at www.synsint.com

Synthesis and Sintering

ISSN 2564-0186 (Print), **ISSN** 2564-0194 (Online)



Review article

Artificial intelligence-guided biosynthesis and retrosynthesis in pharmacognosy: Toward the synthesis-oriented discovery of natural-product therapeutics

Kiarash Solouki [a], **Niloufar Moharrer Navaei** [a,*], **Ayla Balkan** [b]

[a] *Faculty of Pharmacy, Cyprus International University, Nicosia 99258, Northern Cyprus, Mersin 10, Turkey*
[b] *Faculty of Pharmacy, Bahçeşehir University, Nicosia 99258, Northern Cyprus, Mersin 10, Turkey*

ABSTRACT

Artificial intelligence is reshaping Pharmacognosy by connecting ethnobotanical knowledge, multi-omic data, Biosynthetic pathway prediction, Retrosynthetic planning, and medicinal chemistry optimization. Particular attention is given to AI-driven tools, including BioNavi-NP, graph-sequence-enhanced transformers, NAG2G, RSGPT, RetroExplainer, and human-in-the-loop systems such as DeepRetro. These platforms can reconstruct natural-product biosynthesis, predict plausible precursors, preserve molecular topology, suggest multi-step disconnections, and explore broad reaction spaces. They are especially relevant for metabolites with dense stereochemistry, unusual ring systems, multifunctional scaffolds, and enzyme-guided biosynthetic logic. Beyond route design, AI may help prioritize biosynthetic genes, optimize scarce plant-derived compounds, and guide the development of more drug-like analogues with improved potency, selectivity, pharmacokinetic behavior, and synthetic accessibility. However, important limitations remain, including limited plant-specific reaction datasets, weak reaction-condition prediction, incomplete stereochemical and regioselective modeling, benchmark weaknesses, and the need for expert validation. Overall, AI is best understood as a decision-support layer linking biodiversity, traditional knowledge, biosynthetic logic, and experimental synthesis for responsible future therapeutic discovery and validation across modern natural-product-based drug discovery pipelines.

KEYWORDS

Natural products
Pharmacognosy
Artificial intelligence
Biosynthetic pathway prediction
Retrosynthetic planning

OPEN ACCESS

1. Introduction

Pharmacognosy, broadly understood as the study of plant-derived medicines, has always been an interdisciplinary field. It brings together ethnobotanical knowledge, biochemical investigation, and analytical science in a way that few other areas of pharmaceutical research do [1, 2]. In recent years, its scope has expanded further. Quantitative ethnopharmacology [3] and biodiversity-informed models [4] have added more structured and data-driven layers to traditional knowledge systems. At the same time, advances in analytical chemistry and multi-omics technologies are allowing researchers to map plant metabolite diversity and biosynthetic pathways with a level of detail that was not previously possible [5, 6]. These developments show how closely plant chemical diversity and human medicinal experience are connected. Machine learning is now becoming part of this shift as well, especially through its use in large plant datasets [7] and in the prediction of metabolic phenotypes [8]. Taken together, these changes suggest the emergence of a more precise form of herbal medicine, in which AI helps refine traditional remedies through systems-level biological

* Corresponding author. *E-mail address:* Niloufar.navaei@yahoo.com (N. Moharrer Navaei)
Received 13 May 2026; Received in revised form 2 June 2026; Accepted 13 June 2026.

https://doi.org/10.53063/synsint.2026.62341

insight. Recent studies provide clear examples of how AI is already influencing natural-product research. Large-scale neural network screening, for instance, has led to the identification of new antibiotic candidates, supporting the value of AI-guided discovery for finding novel chemical scaffolds [9, 10]. In the area of synthesis planning, AI-based retrosynthetic tools, including transformer and graph-based models, have reached strong performance in route prediction [11, 12]. Importantly, this progress is not limited to conventional synthetic chemistry. Natural-product-focused platforms such as BioNavi-NP use deep learning to predict multi-step plant biosynthetic pathways [13], which is especially relevant for complex plant metabolites. On a broader systems level, AI models that combine genomics and metabolomics have been used to infer enzyme networks involved in the production of specialized metabolites [6]. Network embedding approaches have also shown that different plant lineages can independently evolve similar chemical traits, and that ethnobotanical concept maps often reflect these underlying metabolic patterns [14]. These findings, taken as a whole, show that AI is beginning to reshape both pharmacognosy and medicinal chemistry in practical and conceptually important ways. Still, several important gaps remain. Many current AI models are not trained on sufficiently plant-specific datasets, and they often fail to account for the traditional medicinal context in which plant remedies are used. This is a real limitation, because the chemistry and therapeutic use of medicinal plants can shift with climate, geography, and ecological conditions [15]. As a result, static datasets may become less reliable over time. When these factors are not integrated, ethnobotanical knowledge, ecological variation, plant genomics, and chemical data remain separated from one another, even though they are closely connected in real medicinal practice. For this reason, a more critical and integrated review is needed to explain how current AI tools and methods connect plant genomics, phytochemistry, ecological context, ethnobotanical knowledge, biosynthetic pathway prediction, retrosynthetic planning, and medicinal chemistry. This review therefore examines existing AI-assisted approaches in pharmacognosy and synthesis-oriented natural-product discovery, with attention to their applications, underlying methodological assumptions, practical limitations, and translational value. Rather than proposing a new computational method or framework, the review focuses on evaluating how available AI-based tools can help interpret the chemical diversity of plant-derived compounds, support biosynthetic and retrosynthetic reasoning, and assist expert-guided decision-making in the development of natural-product-based therapeutics.

2. Artificial intelligence as a systems-level catalyst for scalable natural-product discovery and pharmacognosy

AI is needed in pharmacognosy because the discipline now operates at the intersection of chemically complex natural extracts, multimodal omics, and massive but unevenly curated knowledge sources, while its classical workflow still suffers from slow structure elucidation, low-abundance actives, frequent rediscovery of known compounds, and poor scalability; recent reviews describe natural-product data as multimodal, unbalanced, unstandardized, and scattered, which makes conventional manual analysis increasingly inadequate. On the dereplication and annotation side, the need is quantifiable: in untargeted metabolomics, only about 2% of MS/MS spectra are typically annotatable by reference spectral libraries and usually less than 10% even with current machine-learning tools, which is why transformer-scale models such as DreaMS were trained on GNPS/MassIVE data at repository scale; DreaMS used up to 700 million experimental MS/MS spectra, achieved state-of-the-art performance across annotation tasks, and enabled a 201 million-spectrum atlas, while CANOPUS used deep neural networks to predict 2,497 compound classes directly from fragmentation spectra with an average cross-validation accuracy of 99.7%, including compounds lacking spectral or structural reference matches [16]. For structure elucidation and chemotaxonomic triage, ML also helps convert spectroscopy into fast decision support: a 13C-NMR-based study trained classifiers for eight common natural-product classes and reported best F1 scores above 0.82, while recent reviews show ML-assisted MS and NMR workflows are becoming central to library-based and library-independent structure annotation. For genome-enabled pharmacognosy, AI is needed because biosynthetic potential vastly exceeds what can be inspected manually: DeepBGC's BiLSTM plus pfam2vec workflow improved validated BGC detection over ClusterFinder from AUC 0.847 to 0.923, improved leave-class-out generalization from 0.865 to 0.946, and reached a >4-fold precision gain under one evaluation setting, while the later self-supervised BiGCARP model, trained on roughly 127,000 BGC sequences derived from 142,821 antiSMASH-identified clusters, further outperformed DeepBGC on BGC detection and average product-class AUROC. AI is equally needed for prioritization and design: ML models have now classified natural products into 23 bioactive drug classes using LC–MS/MS-derived fingerprints with >93% accuracy on experimental spectra, and the natural-product-inspired generative transformer NIMO achieved 99.9% reconstruction accuracy and enriched predicted antimalarial actives while preserving natural-product-like ring systems and functional groups, showing why generative models matter for scaffold hopping and pseudo-natural-product design [17, 18]. NLP and knowledge-graph methods are also important because ethnobotanical and traditional-medicine knowledge remains locked in text; resources such as COCONUT 2.0, which curates 695,133 unique natural-product structures from 63 source collections, and IMPPAT 2.0, which links 4,010 medicinal plants to 17,967 phytochemicals and 1,095 therapeutic uses, create the machine-readable substrate needed for linking species, metabolites, bioactivities, and traditional uses. Downstream, AI is needed for developability because natural compounds are often scarce, unstable, or poorly soluble: Recent ADME reviews emphasize that in silico ADME/ADMET (absorption, distribution, metabolism, excretion, and toxicity) methods are especially valuable for natural compounds because they require no physical sample and can screen pharmacokinetic risk early, while formulation-focused reviews now describe AI-assisted optimization of herbal nanocarriers and phytomedicine delivery as an emerging translation layer. The main caveat is that AI in pharmacognosy is only as reliable as its data: COCONUT 2.0 still documents residual non-natural or misclassified entries despite extensive curation, DeepBGC explicitly warns that training data are biased toward "workhorse" taxa such as Streptomyces, and ADME reviews stress persistent dependence on data quality, quantity, and wet-lab confirmation. Taken together, the evidence suggests that AI is needed in pharmacognosy not to replace botanical, chemical, or pharmacological expertise, but because it is the only credible way to scale discovery, dereplication, bioactivity inference, biosynthetic mining, ADMET triage, and ultimately formulation across the true size and complexity of natural-product chemical space [19].

Recent studies demonstrate that artificial intelligence is reshaping pharmacognosy and plant gene research into a systems-level discipline, in which biosynthetic gene clusters, metabolite outputs, ecological interactions, and ethnobotanical knowledge are interpreted as interconnected information-processing layers rather than isolated datasets. For instance, multi-omic modeling of Arabidopsis using large-scale integration of genome, transcriptome, proteome, and metabolome matrices has shown that AI can infer pathway logic by identifying transcriptional co-activation signatures, enzyme-substrate dependencies, and latent regulatory modules that jointly determine specialized-metabolite phenotypes [6]. This work highlights that plant biosynthetic systems behave as modular networks, in which clusters of co-evolving genes exhibit predictable interaction patterns, and where AI-driven inference can distinguish between constitutive, stress-induced, and tissue-specific pathway activity. Extending this mechanistic perspective, recent reviews emphasize that next-generation natural-product discovery requires multimodal deep learning architectures capable of fusing orthogonal data types including MS/MS fragmentation patterns, isotopic distribution profiles, NMR descriptors, phylogenetic embeddings, gene-expression tensors, and genomic neighborhood signatures to reconstruct metabolic pathways with far greater accuracy than single-omic approaches [20]. These multimodal systems do not merely correlate features across omics: they learn shared latent manifolds where chemically similar metabolites cluster near their biosynthetic gene clusters, enabling AI to predict uncharacterized metabolites, annotate cryptic pathways, and identify regulatory bottlenecks controlling metabolite flux [21].

3. AI driven retrosynthetic planning for complex natural products

AI-powered retrosynthesis tools now span template-based and template-free models (from neural-symbolic template retrievers to graph/transformer generators) trained on large reaction corpora, yielding state-of-the-art single-step accuracies, yet applying these methods to highly complex natural products (NPs) still poses challenges in stereochemistry, conditions, and multi-step route assembly. Recent studies have leveraged advanced model architectures and expanded datasets (including synthetic

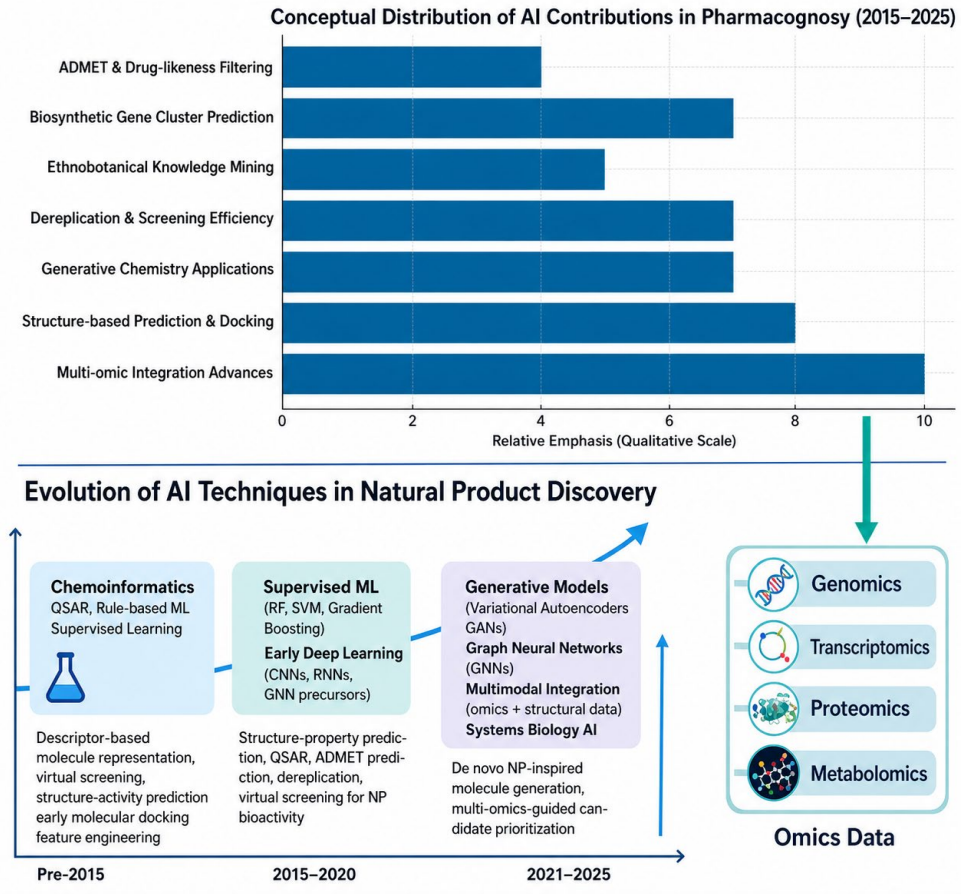


Fig. 1. Temporal progression and comparative prominence of artificial intelligence methodologies applied in pharmacognosy from 2015 to 2025, highlighting the shift from earlier cheminformatics and supervised machine learning approaches toward generative models, multi-omic integration, and systems-level AI strategies in natural product discovery.

augmentation) to improve predictions for complex targets [12, 13, 22]. Template-based and template-free deep-learning models both underpin modern retrosynthesis planners. Template-based methods (e.g., GLN, RetroComposer) use curated reaction templates and often couple neural networks with fingerprint encodings. Semi-template or latent-variable methods predict key intermediates. In contrast, template-free approaches generate reactant structures directly via sequence or graph models. For example, the node-aligned graph-to-graph model NAG2G uses a transformer on 2D/3D graphs with atom-mapping alignment to autoregressively generate reactant graphs, achieving top-1 accuracies exceeding prior graph/sequential models on USPTO-50k/FULL datasets. Similarly, sequence-to-sequence LSTM and transformer networks (e.g., RetroExplainer) have been developed: Wang et al. [23] introduce RetroExplainer, a graph transformer with multi-scale contrastive and multi-task learning, which outperformed prior single-step models across 12 benchmarks, and successfully planned multi-step routes (86.9% of steps matched literature). Recent work also exploits very large synthetic data. Deng et al. [12] generated approximately 10 billion reactions via templates, pre-trained a GPT-style SMILES transformer (RSGPT), and then fine-tuned with reinforcement learning. RSGPT achieved 63.4% top-1 accuracy on USPTO-50k (vs. ≈55% for earlier models). Notably, RSGPT still uses only reaction SMILES (no conditions) and relies on post-hoc human validity checks, pointing to limitations in capturing context. In all cases, models are trained on large patent datasets (USPTO-50k/FULL, Pistachio) and evaluated via Top-k metrics, "All-correct/Any-correct" reactant matches, or fragment metrics. These systems exhibit high SMILES validity and improved generalization on larger corpora, but their gains largely reflect massive data and advanced architectures [11, 12, 23].

Specific NP-relevant applications have emerged by tailoring models or workflows to complex molecules. Bio-retrosynthesis tools use enzyme reaction data: Zheng et al. [13] developed BioNavi-NP, a transformer trained on both general organic and curated biosynthetic reactions, coupled with an and–or tree search. It successfully recovered pathways for 90.2% of 368 test NPs and retrieved known NP building blocks for 72.8% of cases (vs. approximately 42% for rule-based methods). Cong et al. [22] similarly target NP biosynthesis: their graph-sequence enhanced transformer integrates graph neural nets (which preserve molecular topology/stereochemistry) with sequential transformers, achieving state-of-the-art accuracy on NP pathway benchmarks. For synthetic NPs, hybrid large language model (LLM) frameworks have shown promise. Sathyanarayana et al. [24] introduced DeepRetro, an iterative pipeline combining a template-based solver (AiZynthFinder), large LLM proposals (e.g., Anthropic Claude), and human-in-the-loop validation. DeepRetro solved 96–97% of targets in challenging benchmark sets (solving 183/190 and 168/172 targets on USPTO-190 and a "Drug Hunter" NP set) – far above traditional tools – and discovered novel synthetic routes for case-study NPs (Erythromycin B, Reserpine, and Ohauamine C, which is a complex marine natural product molecule with a difficult three-ring peptide-like structure). For example, DeepRetro planned a concise route to Ohauamine C (a tricyclic peptide with four contiguous stereocenters) by strategically assembling amino acid fragments and introducing an early esterification step. Despite these advances, significant limitations remain. Most models do not model reaction conditions (no solvents, catalysts, yields) and often ignore nuanced stereochemical/regioselective outcomes. As Cong et al. [22] note, conventional retrosynthesis tools were "not effective in predicting the enzyme reactions" of NPs, motivating graph encodings that explicitly preserve stereochemistry. Moreover, multi-step planning for NPs still typically requires human guidance: DeepRetro's NP routes often needed 6–14 iterative passes with expert intervention to finalize a route. Finally, benchmark metrics are imperfect: many valid alternative disconnections go uncounted by strict Top-k accuracy, and pathway "success" rates (e.g., % solved) do not reflect route optimality or experimental feasibility. In sum, modern AI retrosynthesis increasingly handles large reaction spaces and leverages LLM reasoning, but predicting truly novel, stereochemically precise syntheses of complex natural products remains an open challenge requiring integrated chemical knowledge and human insight [13, 22, 24]. One major reason this gap remains is the limited availability of plant-specific reaction datasets for training and testing AI models. Most retrosynthetic systems are developed using broad organic chemistry datasets or patent-derived reaction collections. These resources are useful for learning common laboratory transformations, but they do not fully reflect the biosynthetic logic that shapes plant-specialized metabolism. Plant natural products are often built through enzyme-directed processes such as oxidation, glycosylation, methylation, prenylation, terpene cyclization, phenylpropanoid coupling, alkaloid rearrangement, and other pathway-specific transformations. Many of these reactions are poorly represented in conventional reaction databases. This creates a domain-shift problem. In other words, a model may suggest reactions that look chemically reasonable from a synthetic chemistry perspective, but that do not match well with how plant metabolites are formed, modified, or diversified in living systems. The issue is made even more difficult by the incomplete annotation of many reported plant biosynthetic reactions. Important details may be missing, including enzyme identity, organismal or tissue source, stereochemical outcome, substrate scope, reaction conditions, yield, pathway position, or the level of experimental validation. As a result, AI tools may learn molecular connectivity without also learning the biological and stereochemical constraints that make plant-derived natural products distinctive. This limitation also affects benchmarking. Strong performance on USPTO-like datasets does not necessarily mean that a model can reliably predict reactions for rare medicinal plants, poorly studied biosynthetic pathways, or structurally complex phytochemicals. Future progress will therefore depend on curated plant-reaction repositories that combine metabolite structures, enzyme annotations, gene or pathway information, taxonomic source, stereochemical outcomes, ecological or tissue context, and experimentally verified transformations. Such datasets would help AI systems move beyond generic reaction prediction and toward biosynthetic and retrosynthetic reasoning that is more specific to pharmacognosy [22, 25, 26]. A particularly important limitation in AI-guided retrosynthesis of natural products is not just that these molecules contain stereochemistry. The harder problem is that stereochemical information has to be translated into synthetic decisions that are chemically realistic. Many plant-

derived metabolites contain several neighboring stereocenters, fused or bridged rings, spirocyclic motifs, conformationally restricted frameworks, and enzyme-shaped three-dimensional architectures. These features often determine both biological activity and whether a proposed synthetic route is actually feasible [22]. Current AI models struggle with this issue in several ways. First, stereochemical information in training datasets is often incomplete, inconsistently reported, or simplified in SMILES-based representations. As a result, models may learn molecular connectivity more reliably than absolute or relative configuration. Second, even when stereochemical labels are included, many benchmarks reward correct reactant prediction or top-k structural matching without strongly penalizing wrong stereochemical outcomes. Third, stereoselective synthesis depends on many experimental and mechanistic factors that are rarely captured in retrosynthesis models. These include catalyst choice, solvent, temperature, substrate conformation, protecting-group strategy, neighboring-group effects, enzyme active-site geometry, and kinetic versus thermodynamic control. This is especially problematic for natural products. A disconnection may look reasonable in a two-dimensional structure, but it may not be synthetically realistic if it fails to preserve the required stereochemical relationships or does not offer a plausible way to install them. Some recent methodological developments are beginning to address this weakness. Graph-based and graph-to-graph models can preserve atom-level connectivity and stereochemical annotations more effectively than purely sequence-based systems. Hybrid graph-sequence transformers also help by combining molecular topology with reaction-sequence learning. Three-dimensional and conformer-aware models may improve how spatial constraints are represented, while biosynthesis-oriented platforms can use enzyme logic to restrict predictions to transformations that better match natural stereochemical assembly. Human-in-the-loop approaches remain important as well. Expert chemists can often recognize when a predicted route is stereochemically ambiguous, lacks stereocontrol, or depends on unrealistic selectivity. Even so, these advances are still only partial solutions. For AI to become more reliable in natural-product synthesis planning, future systems will need stereochemically explicit and plant-relevant reaction datasets, condition-aware prediction, benchmarks that evaluate stereochemical correctness rather than only connectivity, and experimental validation of key stereoselective steps. This distinction is essential in pharmacognosy, because the therapeutic and synthetic value of a natural product often depends not only on its scaffold, but also on the precise three-dimensional arrangement produced by biosynthesis [22, 24, 26].

Table 1. Recent AI-based retrosynthesis and biosynthetic pathway prediction models relevant to natural-product discovery and synthesis planning.

Model/approach	Dataset(s)	Key result /natural-product application	Main limitations	Ref.
Graph-Sequence Enhanced Transformer; hybrid GNN and SMILES Transformer; template-free biosynthetic prediction	Curated enzymatic and organic reaction datasets related to natural-product biosynthesis	Achieved state-of-the-art single-step and multi-step prediction of natural-product biosynthetic transformations, with improved retention of molecular topology and stereochemical information.	Primarily focused on biosynthetic and enzymatic reactions; predictive scope remains constrained by the availability and diversity of curated enzyme-reaction data.	[22]
BioNavi-NP; Transformer-based sequence-to-sequence model combined with AND–OR tree search	33,710 biosynthetic and organic reactions compiled from MetaCyc, KEGG, and patent-derived reaction data	Predicted plausible biosynthetic routes for 90.2% of tested natural products and recovered 72.8% of reported natural-product building blocks.	Performance depends strongly on known biosynthetic pathway information; sparse enzymatic reaction data and absence of reaction-condition prediction limit practical implementation.	[13]
DeepRetro; hybrid LLM-assisted retrosynthesis framework integrating template-based CASP tools and human-in-the-loop refinement	USPTO and Pistachio patent-derived reaction datasets; evaluated using USPTO-50k and selected complex natural-product/drug-like case studies	Solved approximately 97–98% of targets across two natural-product/drug-related test sets and proposed new retrosynthetic routes for complex case molecules such as Ohauamine C.	Requires substantial computational resources and expert human guidance; detailed reagent, solvent, and reaction-condition optimization is not fully addressed; iterative refinement is often necessary.	[24]
RSGPT; GPT-style generative Transformer pre-trained on large-scale template-generated reactions	More than 10 billion synthetic reactions generated from USPTO-derived templates; fine-tuned on USPTO-50k, MIT, and USPTO-FULL benchmarks	Reached 63.4% top-1 accuracy on USPTO-50k, outperforming previous state-of-the-art models, and achieved high top-10 SMILES validity of 97.7%.	Synthetic pre-training data may introduce template-derived bias; reaction conditions are not explicitly encoded; mechanistic interpretability remains limited.	[12]
RetroExplainer; interpretable Graph Transformer using multi-scale representation learning, contrastive learning, and multi-task training	Twelve public retrosynthesis benchmark datasets, including USPTO-derived variants	Outperformed state-of-the-art single-step retrosynthesis models across all evaluated benchmarks; in multi-step planning, 86.9% of predicted single reactions matched literature-reported reactions.	Requires extensive multi-task training; stereochemical resolution may remain incomplete in some cases; validation on natural-product-specific retrosynthetic problems is still limited.	[23]
NAG2G; Node-Aligned Graph-to-Graph Transformer incorporating 2D graph and 3D conformational information	USPTO-50k, USPTO-MIT, and USPTO-FULL	Demonstrated strong accuracy across USPTO retrosynthesis benchmarks and successfully predicted synthesis pathways for drug-like molecules.	Relies on 3D conformer generation and patent-derived training data, which may limit generalization to structurally complex natural products; reaction conditions are not directly modeled.	[11]

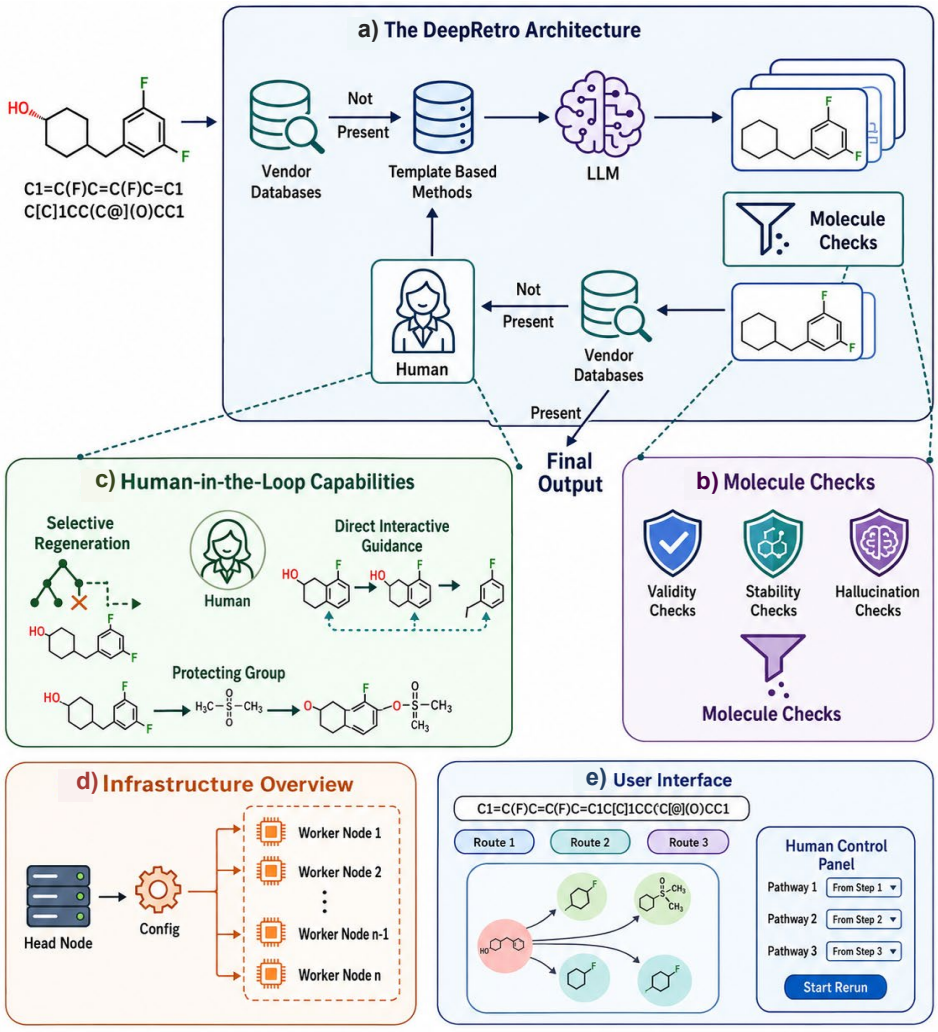


Fig. 2. Overview of the DeepRetro retrosynthesis framework: a) DeepRetro begins with a vendor database search and template-based retrosynthesis; when these steps fail, an LLM proposes single-step routes that are filtered through molecule validation checks before further recursive evaluation, with optional human input. b) The incorporated molecule checks assess chemical validity, stability, and hallucination risk in LLM-generated outputs. c) Human-in-the-loop functions support selective pathway regeneration, direct chemist-guided edits, and protecting-group addition. d) The system uses a scalable head-node/worker-node infrastructure for complex syntheses. e) The graphical interface enables pathway visualization, node selection for regeneration, and direct molecular editing.

4. Emerging applications and limitations of AI in medicinal chemistry

Artificial intelligence (AI) methods are now transforming many stages of drug discovery. For example, Bhat and Ahmed [27] review how advanced AI techniques – including machine learning, deep learning, and reinforcement learning – are accelerating tasks such as target identification, de novo lead generation, drug repurposing, lead optimization, and toxicity prediction. Deep generative models in particular (e.g., graph neural nets, diffusion models, and transformer-based systems) can propose novel, property-optimized molecules beyond existing screening libraries. Structural AI models are also emerging: AlphaFold and its successors now predict protein structures with near-experimental accuracy, enabling structure-based drug design and a deeper understanding of ligand–target interactions. AI is similarly being applied to related areas such as automated retrosynthesis planning and reaction prediction. For instance, the recent SynthSense framework combines reinforcement learning with retrosynthetic feedback to bias molecular generation toward synthetically accessible compounds, yielding an approximately 6-fold increase in easily-made hits compared to naïve generative models. In summary, AI-driven workflows – from large-scale virtual screening and generative design to predictive ADMET modeling and protein modeling – are rapidly emerging to accelerate medicinal chemistry discovery [27–29]. Although these approaches are often discussed within the wider field of medicinal chemistry, they are highly relevant to pharmacognosy when they are applied to plant-derived and natural-product-inspired chemical space. Natural products often have stereochemically complex scaffolds, dense functional groups,

uncommon ring systems, and structural features shaped by biosynthesis. These characteristics can make their therapeutic development more challenging than that of many conventional synthetic screening hits. In this context, protein-structure prediction and structure-based modeling can help evaluate whether natural-product scaffolds or semi-synthetic analogues are likely to interact with disease-relevant targets. Virtual screening and target-prediction tools can also help prioritize potentially bioactive metabolites from phytochemical, metabolomic, or ethnopharmacological datasets. ADMET and toxicity models are similarly important in pharmacognosy, since many plant-derived compounds show promising biological activity but still face practical barriers such as poor solubility, metabolic instability, off-target effects, herb-drug interactions, or narrow safety margins. Generative and synthesizability-aware models may also support natural-product lead optimization. They can suggest analogues that retain important pharmacophoric or biosynthetically meaningful features while improving drug-likeness, synthetic accessibility, and overall developability. For this reason, the relevance of these general AI methods in the present review is not that pharmacognosy should be treated as ordinary small-molecule discovery. Rather, it is that computational drug-discovery tools can extend the value of natural-product research beyond compound identification, helping with target interpretation, safety evaluation, structural optimization, and expert-guided therapeutic development [14, 30, 31]. Despite this promise, several inherent limitations constrain AI's impact in practice. Bhat and Ahmed [27] caution that AI in pharma still faces methodological and data-related challenges. In particular, training datasets often contain biases or gaps (e.g., under-represented chemotypes or endpoints) that can lead to overfitting and poor generalization. A concrete example is synthetic feasibility: Dekleva et al. [29] demonstrate that many molecules suggested by generative models are practically "difficult or impossible to produce" because the models lack built-in chemistry knowledge. Similarly, Yoo [32] emphasizes the need for rigorous validation of AI-generated compounds – for instance, a de novo AI-designed TNIK inhibitor did reach Phase II trials, but Yoo notes that "broader validation, mechanistic understanding, and regulatory alignment remain essential" to translate such successes into reliable therapeutics. More broadly, issues like model interpretability (the "black box" problem), reproducibility between labs, and alignment with regulatory standards all remain hurdles [27, 32, 33]. In short, while AI opens powerful new tools for medicinal chemistry, its outputs must be treated with caution: careful curation of data, integration of chemical knowledge, and extensive experimental benchmarking are required to overcome AI's current limitations and realize its full potential.

5. Case studies of AI-powered systems for medicinal chemistry, modeling plant biosynthetic networks, modification and lead optimization of natural products

Deep learning has dramatically expanded the search for new antibiotics. Stokes et al. [9] trained a neural network on 2,300 known antibacterial compounds and screened over 100 million structures from a drug repurposing library. This led to halicin, a chemically novel broad-spectrum antibiotic active against E. coli, M. tuberculosis, and carbapenem-resistant Enterobacteriaceae in mice [9]. In the same study, eight additional compounds with distinct scaffolds were experimentally validated. More recently, Liu et al. [10] used deep learning on a smaller dataset (approximately 7,500 molecules screened against A. baumannii) to find abaucin, a narrow-spectrum antibiotic active specifically against multidrug-resistant Acinetobacter baumannii [10]. Abaucin's mechanism (lipoprotein transport inhibition) was elucidated via machine learning–suggested targets and lab assays, and it controlled A. baumannii infection in mice. These cases illustrate how AI-powered virtual screening can rapidly prioritize structurally novel natural-product–like compounds with potent activity, greatly accelerating early lead identification [9, 10]. AI has optimized both production and therapeutic use of complex anticancer natural products. For example, AI-assisted modeling (adaptive neuro-fuzzy inference systems with genetic algorithms) improved paclitaxel yields in hazel cell cultures by optimizing culture media and conditions beyond classical statistical designs. Machine learning also guided reinforcement learning models to simulate optimal dosing in lung cancer patients by balancing efficacy and toxicity. Upstream, genome mining coupled with AI has identified missing enzymes in Taxus pathways, enabling semi-synthetic paclitaxel production. In drug repurposing, a hybrid AI–experimental pipeline predicted covalent-binding natural products for polo-like kinase 1 (PLK1), a cancer target. By screening natural-product libraries in silico for molecules predicted to covalently bind PLK1's active site, researchers pinpointed a compound from Scutellaria baicalensis (baicalein/baicalin family) that was experimentally confirmed as a selective PLK1 inhibitor. These efforts show AI's role in refining natural product leads: improving yields of scarce plant drugs (paclitaxel) and identifying NP scaffolds as targeted cancer therapies (PLK1 inhibitors) [31]. AI can reconstruct entire plant biosynthetic networks from endpoints to known precursors. Zheng et al. [13] developed BioNavi-NP, a deep-learning retrosynthesis tool that predicts natural product biosynthetic routes. A transformer neural network was trained on approximately 33,000 known enzymatic reactions and organic transformations to propose possible precursor steps. An and–or tree search then assembles multi-step pathways. BioNavi-NP correctly generated full biosynthetic paths for 90.2% of 368 test natural products, recovering known precursor building blocks in 72.8% of cases. In other words, it outperformed rule-based methods by >70% (72.8% vs. 42%) in identifying valid NP biosynthesis routes. Fig. 3 illustrates the vast NP space vs. known pathways; BioNavi-NP bridges this gap by efficiently "back-tracking" complex molecules to simple precursors.

By automating pathway prediction, BioNavi-NP enables researchers to engineer plant pathways in silico and plan semi-synthetic production of valuable NPs [13]. Integrating multi-omics data with AI can predict the genes that make specific plant metabolites. Bai et al. [6] assembled multi-layer data for Arabidopsis (genome sequence, proteomics, epigenetics, expression, etc.) and used the AutoGluon AutoML framework to classify which genes encode enzymes of specialized metabolism. Focusing on alkaloids, terpenoids, and phenolics, they found that genomic and proteomic features (e.g., gene length, domain counts) were most predictive, outstripping transcriptomic or epigenetic features. The resulting model showed excellent accuracy, which transferred surprisingly well: when the Arabidopsis model was applied to maize, tomato, grape, and poppy, it predicted pathway genes with equivalent or better accuracy than models trained de novo in those species. In short, this ML approach effectively mapped "wiring diagrams" of plant metabolic networks: it pinpointed candidate biosynthetic

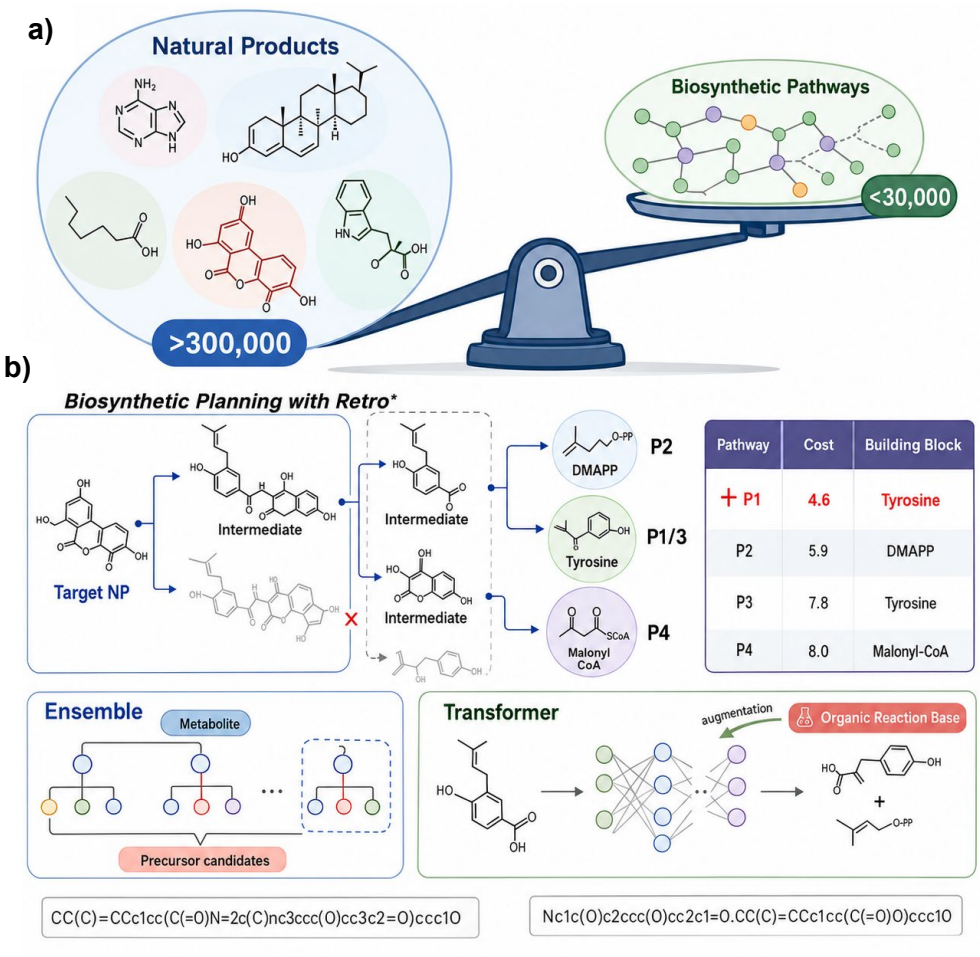


Fig. 3. BioNavi-NP reconstructs natural product biosynthesis. a) Over 300,000 known natural products derive from a few simple building blocks, but only about 30,000 enzymatic reactions are characterized. b) BioNavi-NP uses deep learning and search to identify plausible precursor pathways.

genes for high-value phytochemicals across diverse plants. This case demonstrates that AI can integrate genomic and proteomic patterns to infer entire biosynthetic pathways, guiding metabolic engineering of medicinal plants [6]. Generative AI can optimize natural product structures for potency and drug-likeness. One compelling example is the marinopyrrole A scaffold (a marine Streptomyces metabolite). Using AI-driven molecular generation, researchers created dozens of marinopyrrole-inspired analogs and screened them computationally as COX-1 inhibitors. The top AI-designed analog showed >100-fold improved potency (IC_{50} ≈0.1 μM) compared to the parent marinopyrrole (IC_{50} ≈16.6 μM). This analog retained the complex core features of marinopyrrole but incorporated synthetic modifications suggested by the model to enhance binding. Such AI-assisted design distilled key pharmacophores from the natural scaffold and grafted them into more drug-like structures. This case (from Muthuraj and Chandrasekaran) highlights how machine learning can navigate NP chemical space to generate superior leads from traditional scaffolds. As a result, AI-enabled analog design can convert a barely active natural molecule into a potent lead with optimized efficacy and safety. Advanced AI systems also integrate multiple sources of information to optimize complex NP-based therapies. For instance, graph-based network pharmacology models have been used to analyze herb–ingredient–target networks. Although specific examples are emerging, general studies note that AI tools (e.g., random forests, graph neural networks, and embeddings) are successfully predicting bioactivities of natural compounds in oncological and anti-inflammatory assays. In silico work combining plant metabolomics and structural modeling can propose synergistic NP combinations or prioritize analogs with improved ADME properties. For example, deep learning models have been reported to predict anticancer and anti-inflammatory actions of phytochemicals by learning from databases of traditional remedies and clinical data. These AI-powered integrative approaches promise to optimize “leads” by considering multiple compounds or formulation factors simultaneously [34]. In summary, modern medicinal chemistry increasingly relies on AI pipelines that not only predict individual compound potency, but also design and tune entire natural-product–based regimens through multi-target optimization [34–36].

6. Conclusions and future directions

This review shows that the strongest contribution of artificial intelligence to pharmacognosy is its ability to turn scattered natural-product information into decisions that can be tested experimentally. Across the areas discussed, AI appears most useful when it supports four connected tasks: improving dereplication and metabolite annotation, prioritizing biosynthetic genes and pathway modules, proposing plausible biosynthetic or retrosynthetic routes, and guiding medicinal chemistry optimization of natural-product scaffolds. At the same time, the evidence reviewed here makes it clear that AI has not reached the same level of maturity across all of these tasks. Spectral learning, molecular networking, multi-omic integration, and early ADMET prediction are already helpful for prioritization and hypothesis generation. Fully reliable synthesis planning for complex plant metabolites, however, remains more difficult. The main obstacles are not only molecular complexity. They also include limited plant-specific reaction data, weak prediction of reaction conditions, inconsistent stereochemical annotation, poor modeling of enzyme selectivity, and benchmark systems that often reward structural matching without showing that a predicted route is experimentally realistic. Future work should therefore begin with natural-product classes that are most suitable for AI-guided synthesis or semi-synthesis. The most realistic near-term candidates are compounds with defined structures, known or partly known biosynthetic origins, available metabolomic or genomic support, and clear pharmacological value, but also practical problems related to supply or optimization. These include taxane-type diterpenoids, benzylisoquinoline and indole alkaloids, phenylpropanoids, flavonoids, terpenoid glycosides, polyketide-like scaffolds, and other structurally defined plant metabolites for which semi-synthetic modification, pathway engineering, or analogue design is feasible. In such cases, AI can help identify likely precursors, prioritize enzymes, suggest synthetically accessible analogues, and reduce the number of low-value experimental routes. By contrast, poorly characterized extracts, metabolites with unresolved stereochemistry, compounds without validated bioactivity, or pathways lacking genetic or enzymatic evidence should be treated mainly as discovery-stage problems. For these targets, AI is better used for annotation, clustering, dereplication, and hypothesis generation, rather than for confident route prediction. The most urgent need, then, is the expansion of curated plant-specific databases. General reaction datasets and patent-derived corpora are useful, but they do not sufficiently represent the enzyme-guided logic of plant-specialized metabolism. Priority should be given to databases that record plant enzymatic oxidations, glycosylations, methylations, prenylations, terpene cyclizations, phenylpropanoid couplings, alkaloid rearrangements, and stereoselective tailoring reactions. These resources should not contain structures alone. Ideally, each reaction or metabolite should be linked to taxonomic source, tissue or developmental origin, ecological condition, traditional preparation method, spectral evidence, enzyme or gene identity, substrate scope, stereochemical outcome, reaction condition, yield or conversion data, bioactivity, toxicity, herb-drug interaction evidence, and level of experimental validation. This is especially important because plant-specific reaction scarcity and incomplete annotation are major reasons for the current domain-shift problem in AI retrosynthesis. Benchmarking also needs to become more specific to pharmacognosy. Future models should be assessed not only by top-k accuracy or by whether predicted reactants match a known answer, but also by whether the suggested route preserves stereochemistry, uses plausible enzymatic or chemical transformations, includes feasible conditions, avoids unrealistic protecting-group or selectivity assumptions, and can be reproduced experimentally. Human-in-the-loop validation should remain central, especially for dense stereochemistry, unusual ring systems, and multifunctional scaffolds. In practical terms, the next stage of the field should not aim for fully autonomous natural-product synthesis. A more realistic goal is an expert-guided AI workflow in which computational systems rank compounds, genes, reactions, disconnections, analogues, and safety risks, while pharmacognosists, synthetic chemists, medicinal chemists, and plant biologists judge which predictions are chemically meaningful and worth testing. Finally, AI-integrated pharmacognosy must remain ethically and biologically grounded. Traditional medicinal knowledge, regional biodiversity, and plant materials connected to specific communities should not be treated simply as data to be extracted. Future platforms should include transparent attribution, culturally responsible data use, benefit-sharing considerations, and careful documentation of geographical and ecological context.

CRediT authorship contribution statement

Kiarash Solouki: Investigation, Writing – original draft, Writing – review & editing.
Niloufar Moharrer Navaei: Investigation, Supervision, Writing – review & editing.
Ayla Balkan: Investigation, Supervision, Writing – review & editing

Declaration of AI Assistance

Artificial intelligence tools were used solely to improve the grammar, academic clarity, overall language quality and figure design of this manuscript. Specifically, OpenAI and QuillBot were employed for linguistic refinement and polishing and designing purposes only. These tools were not used for generating scientific content, designing experiments, interpreting results, or contributing to the intellectual or analytical aspects of the research. The authors reviewed and edited the content as needed and took full responsibility for the content of the published article. The graphical abstract and figures were prepared with the assistance of an AI-based image-designing tool (OpenAI). The authors carefully reviewed, edited, and verified all scientific content, labels, and visual information presented in the graphical abstract and confirm that it accurately reflects the concepts and findings discussed in the manuscript. The authors take full responsibility for the accuracy, integrity, and final content of the graphical abstract. All scientific ideas, analyses, and conclusions presented in this manuscript are the sole work of the authors.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and acknowledgment

No funding was received to conduct this study.

References

[1] T. Adhikary, P. Basak, Interdisciplinary approaches incorporating computational intelligence in modern pharmacognosy to address biological problems, Electronic Systems and Intelligent Computing: Proceedings of ESIC 2020, Springer, Singapore. (2020) 11–19. https://doi.org/10.1007/978-981-15-7031-5_2.

[2] S. Banerjee, Introduction to Ethnobotany and Traditional Medicine, Traditional Resources and Tools for Modern Drug Discovery: Ethnomedicine and Pharmacology, Singapore: Springer Nature Singapore. (2024) 1–30. https://doi.org/10.1007/978-981-97-4600-2_1.

[3] M. Leonti, The relevance of quantitative ethnobotanical indices for ethnopharmacology and ethnobotany, J. Ethnopharmacol. 288 (2022) 115008. https://doi.org/10.1016/j.jep.2022.115008.

[4] C.C. Davis, P. Choisy, Medicinal plants meet modern biodiversity science, Curr. Biol. 34 (2024) R158–R173. https://doi.org/10.1016/j.cub.2023.12.038.

[5] P.M. Allard, T. Péresse, J. Bisson, K. Gindro, L. Marcourt, et al., Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication, Anal. Chem. 88 (2016) 3317–3323. https://doi.org/10.1021/acs.analchem.5b04804.

[6] W. Bai, C. Li, W. Li, H. Wang, X. Han, et al., Machine learning assists prediction of genes responsible for plant specialized metabolite biosynthesis by integrating multi-omics data, BMC Genom. 25 (2024) 418. https://doi.org/10.1186/s12864-024-10258-6.

[7] H. Ali, Artificial intelligence in multi-omics data integration: Advancing precision medicine, biomarker discovery and genomic-driven disease interventions, Int. J. Sci. Res. Arch. 8 (2023) 1012–1030. https://doi.org/10.30574/ijsra.2023.8.1.0189.

[8] F.C. Wolters, E. Del Pup, K.S. Singh, K. Bouwmeester, M.E. Schranz, et al., Pairing omics to decode the diversity of plant specialized metabolism, Curr. Opin. Plant Biol. 82 (2024) 102657. https://doi.org/10.1016/j.pbi.2024.102657.

[9] J.M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, et al., A deep learning approach to antibiotic discovery, Cell. 180 (2020) 688–702. https://doi.org/10.1016/j.cell.2020.01.021.

[10] G. Liu, D.B. Catacutan, K. Rathod, K. Swanson, W. Jin, et al., Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii, Nat. Chem. Biol. 19 (2023) 1342–1350. https://doi.org/10.1038/s41589-023-01349-8.

[11] L. Yao, W. Guo, Z. Wang, S. Xiang, W. Liu, G. Ke, Node-aligned graph-to-graph: elevating template-free deep learning approaches in single-step retrosynthesis, JACS Au. 4 (2024) 992–1003. https://doi.org/10.1021/jacsau.3c00737.

[12] Y. Deng, X. Zhao, H. Sun, Y. Chen, X. Wang, et al., RSGPT: a generative transformer model for retrosynthesis planning pre-trained on ten billion datapoints, Nat. Commun. 16 (2025) 7012. https://doi.org/10.1038/s41467-025-62308-6.

[13] S. Zheng, T. Zeng, C. Li, B. Chen, C.W. Coley, et al., Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP, Nat. Commun. 13 (2022) 3342. https://doi.org/10.1038/s41467-022-30970-9.

[14] D. Meijer, M.A. Beniddir, C.W. Coley, Y.M. Mejri, M. Öztürk, et al., Empowering natural product science with AI: leveraging multimodal data and knowledge graphs, Nat. Prod. Rep. 42 (2025) 654–662. https://doi.org/10.1039/d4np00008k.

[15] P. Palit, S.C. Mandal, Climate change, geographical location, and other allied triggering factors modulate the standardization and characterization of traditional medicinal plants: a challenge and prospect for phyto-drug development, Evidence based validation of traditional medicines: a comprehensive approach, Springer Singapore, Singapore. (2021) 359–369. https://doi.org/10.1007/978-981-15-8127-4_18.

[16] R. Bushuiev, A. Bushuiev, R. Samusevich, C. Brungs, J. Sivic, T. Pluskal, Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS, Nat. Biotechnol. (2025) 1–11. https://doi.org/10.1038/s41587-025-02663-3.

[17] N.J. Brittin, J.M. Anderson, D.R. Braun, S.R. Rajski, C.R. Currie, T.S. Bugni, Machine learning-based bioactivity classification of natural products using LC-MS/MS metabolomics, J. Nat. Prod. 88 (2025) 361–372. https://doi.org/10.1021/acs.jnatprod.4c01123.

[18] G.D. Hannigan, D. Prihoda, A. Palicka, J. Soukup, O. Klempir, et al., A deep learning genome-mining strategy for biosynthetic gene cluster prediction, Nucleic Acids Res. 47 (2019) e110–e110. https://doi.org/10.1093/nar/gkz654.

[19] R. Ancuceanu, B.E. Lascu, D. Drăgănescu, M. Dinu, In silico ADME methods used in the evaluation of natural products, Pharmaceutics. 17 (2025) 1002. https://doi.org/10.3390/pharmaceutics17081002.

[20] J.K. Reinhardt, D. Craft, J.K. Weng, Toward an integrated omics approach for plant biosynthetic pathway discovery in the age of AI, Trends Biochem. Sci. 50 (2025) 311–321. https://doi.org/10.1016/j.tibs.2025.01.010.

[21] B. Behsaz, E. Bode, A. Gurevich, Y.N. Shi, F. Grundmann, et al., Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery, Nat. Commun. 12 (2021) 3225. https://doi.org/10.1038/s41467-021-23502-4.

[22] S. Cong, M. Zhang, Y. Song, S. Chang, J. Tian, et al., Graph-sequence enhanced transformer for template-free prediction of natural product biosynthesis, Patterns. 6 (2025) 101259. https://doi.org/10.1016/j.patter.2025.101259.

[23] Y. Wang, C. Pang, Y. Wang, J. Jin, J. Zhang, et al., Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks, Nat. Commun. 14 (2023) 6155. https://doi.org/10.1038/s41467-023-41698-5.

[24]S.V. Sathyanarayana, S.D. Hiremath, S. Rahil Kirankumar, R. Panda, R. Jana, et al., DeepRetro discovers retrosynthetic pathways through iterative large language model reasoning, Sci. Rep. 16 (2026) 8448. https://doi.org/10.1038/s41598-026-38821-z.

[25] Y. Kwak, T. Kim, S.G. Kim, J. Park, READRetro web: A user-friendly platform for predicting plant natural product biosynthesis, Mol. Cells. 48 (2025) 100235. https://doi.org/10.1016/j.mocell.2025.100235.

[26] M. Orsi, J.L. Reymond, Assigning the stereochemistry of natural products by machine learning, J. Cheminform. 18 (2026) 76. https://doi.org/10.1186/s13321-026-01205-6.

[27] A.R. Bhat, S. Ahmed, Artificial intelligence (AI) in drug design and discovery: A comprehensive review, Silico Res. Biomed. 1 (2025) 100049. https://doi.org/10.1016/j.insi.2025.100049.

[28] E.S. Ozdemir, H. Jang, O. Keskin, A. Gursoy, R. Nussinov, Deep generative molecular design and its value in modern drug discovery, Expert Opin. Drug Discov. 21 (2026) 273–287. https://doi.org/10.1080/17460441.2026.2636192.

[29] D. Dekleva, A. Voronov, J.P. Janet, A. Ekborg, J. Borišek, et al., Synthesizability via reward engineering: expanding generative molecular design into synthetic space, Chem. Sci. 17 (2026) 10015–10028. https://doi.org/10.1039/D5SC09263A.

[30] A. Gangwal, A. Lavecchia, Artificial intelligence in natural product drug discovery: current applications and future perspectives, J. Med. Chem. 68 (2025) 3948–3969. https://doi.org/10.1021/acs.jmedchem.4c01257.

[31] R. Muthuraj, J. Chandrasekaran, Nature meets machine: the AI renaissance in natural product drug discovery, Nat. Prod. Bioprospect. 16 (2026) 37. https://doi.org/10.1007/s13659-025-00589-6.

[32] W. Yoo, Precision oncology in the age of AI: lessons from AI-driven drug discovery and clinical translation, BJC Rep. 4 (2026) 18. https://doi.org/10.1038/s44276-026-00221-1.

[33] J.J. Louwen, M.H. Medema, J.J. van der Hooft, Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching, Microbiome. 11 (2023) 13. https://doi.org/10.1186/s40168-022-01444-3.

[34] Z.K. Othman, M.M. Ahmed, O. Kasimieh, S.S. Musa, F. Branda, et al., Artificial intelligence for natural product drug discovery and development: current landscape, applications, and future directions, Intell.-Based Med. 12 (2025) 100316. https://doi.org/10.1016/j.ibmed.2025.100316.

[35] B.K. Babu, Revolutionizing Herbal Medicine: The Role of E-health Informatics and Network Pharmacology in Personalized Herbal Therapies, Int. J. Pharm. Investig. 15 (2025) 219–227. https://doi.org/10.5530/ijpi.20251754.

[36] A. Mukherjee, S. Abraham, A. Singh, S. Balaji, K.S. Mukunthan, From data to cure: A comprehensive exploration of multi-omics data analysis for targeted therapies, Mol. Biotechnol. 67 (2025) 1269–1289. https://doi.org/10.1007/s12033-024-01133-6.