Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 21 March 2026
Revised 25 April 2026
Accepted 21 May 2026

Natural Resources for Human Health 2026, Vol. 6 (3): 593–601

KEYWORDS:
retrobiosynthesis; cryptic metabolites; secondary metabolism; artificial intelligence; machine learning; biosynthetic pathway prediction; genome mining; multi-omics

Retrobiosynthesis-Guided Discovery of Cryptic Secondary Metabolites: An AI-Assisted Reconstruction of Plant Metabolic Pathways

Saraswati B[1], Koushik Kumar Nandhagopal[2], Abhishek Bhati[3], Komil Mukhammadiev[4], C. Selvakumar[5], Jahnavi Pandya[6], Megha Jagga[7]

[1]Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research. saraswatib@maher.ac.in
[2]Meenakshi College of Physiotherapy, Meenakshi Academy of Higher Education and Research. koushik@maher.ac.in
[3]Department of Forensic Medicine, Noida International University, Greater Noida, Uttar Pradesh 203201, India. bhatiforensic@gmail.com
[4]Tashkent State Transport University, Tashkent, Uzbekistan / University of Tashkent for Applied Sciences, Tashkent, Uzbekistan. komilmuhammadiev@yandex.ru
[5]Department of Biochemistry, School of Allied Health Sciences, VIMS Hospital, Salem Campus, Vinayaka Mission's Research Foundation (DU), Tamil Nadu, India. selvaihasc@yahoo.com
[6]Department of Pharmaceutics, Parul Institute of Pharmacy, Parul University, Vadodara, Gujarat, India. jahnavi.pandya37595@paruluniversity.ac.in
[7]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India. megha.jagga.orp@chitkara.edu.in

ABSTRACT: Plants synthesise an enormous chemical repertoire—more than 200,000 known specialised (secondary) metabolites and, by most estimates, a far larger number still undetected. Yet the biosynthetic pathways of the great majority remain unresolved, and many metabolites are "cryptic": encoded in genomes but unexpressed under standard conditions, or predicted to exist but never isolated. This constitutes a vast reservoir of untapped chemistry of high therapeutic value, given that natural products underlie more than 70% of antibiotics and over half of anticancer agents. Conventional, gene-by-gene pathway elucidation cannot scale to this challenge—a difficulty compounded in plants, whose biosynthetic genes, unlike those of microbes, are frequently dispersed rather than clustered. This review examines an emerging solution: retrobiosynthesis, the application of retrosynthetic logic to enzyme-catalysed reactions, accelerated by artificial intelligence. We describe how AI-driven bio-retrosynthesis engines (READRetro, BioNavi-NP, graph-transformer models) predict biosynthetic routes backward from a target metabolite to primary-metabolite building blocks; how machine learning and genome mining (plantiSMASH, multi-omics ML) nominate the candidate genes and enzymes for each predicted step; and how protein-structure prediction (AlphaFold, ESMFold) and sequence-similarity networks refine enzyme-function assignment before heterologous expression validates the reconstructed pathway.

We propose an integrated, iterative workflow uniting these components, illustrate it with a benzylisoquinoline-alkaloid example, and critically assess the principal limitations—

Natural Resources for Human Health
eISSN: 2583-1194

enzyme promiscuity, data scarcity, plant genome complexity and the gap between prediction and experimental proof. Retrobiosynthesis-guided, AI-assisted reconstruction offers a scalable strategy for illuminating plant metabolic dark matter and unlocking cryptic secondary metabolites for drug discovery and synthetic biology.

1. INTRODUCTION

Plants are the most prolific chemists in nature. As sessile organisms, they have evolved an immense arsenal of specialised (secondary) metabolites—more than 200,000 structures are already known—to defend against pathogens and herbivores, attract pollinators and adapt to their environment. [1,2] These compounds are also the foundation of much of medicine: natural products and their derivatives underlie over 70% of antibiotics and more than half of anticancer agents (Figure 2). [3] Despite this importance, the biosynthetic pathways of the overwhelming majority of plant metabolites remain unresolved—a vast "metabolic dark matter" whose systematic illumination is one of the central challenges of modern natural-products science. [2,4]

A large part of this dark matter is *cryptic*. In the genomic sense, biosynthetic gene clusters (BGCs) and pathways are frequently "silent"—present in the genome but transcriptionally inactive under standard laboratory conditions—and in microbes such silent clusters outnumber active ones by roughly five- to tenfold. [5,6] In the chemical sense, many metabolites are predicted from genomic or metabolomic data but have never been isolated or structurally confirmed. Accessing these hidden products—by predicting them, linking them to their genes, and reconstructing their biosynthesis—would dramatically expand the accessible chemical space for drug discovery. [5,6]

Plants pose a distinctive difficulty. Unlike bacteria and fungi, in which biosynthetic genes are typically co-localised in compact clusters that can be found by genome mining, plant biosynthetic genes are often dispersed across the genome and embedded in large, highly duplicated gene families, so cluster-centric strategies alone are insufficient. [2,7] Traditional, gene-by-gene elucidation—expression, purification and assay of each candidate enzyme—is accurate but far too slow to keep pace with the flood of genome and metabolome data. What is required is a scalable, predictive framework, and this review argues that *retrobiosynthesis*, accelerated by artificial intelligence, provides one. We introduce the retrobiosynthetic concept (Figure 1), survey the AI methods and computational tools that enable it, propose an integrated discovery workflow, illustrate it with a worked example, and appraise the field's limitations and prospects.

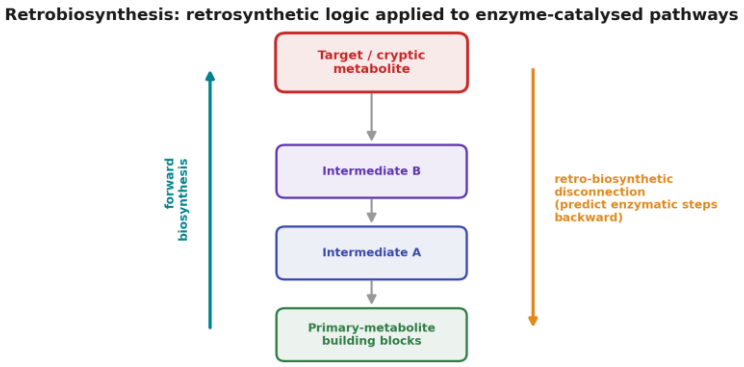

Figure 1. Retrobiosynthesis applies the retrosynthetic logic of synthetic chemistry to enzyme-catalysed pathways. Beginning from a target or cryptic metabolite, successive retro-biosynthetic disconnections predict the intermediate metabolites and the enzymatic steps (read backward) that connect the target to primary-metabolite building blocks—the reverse of the forward biosynthetic route the plant actually uses.

Natural Resources for Human Health
eISSN: 2583-1194

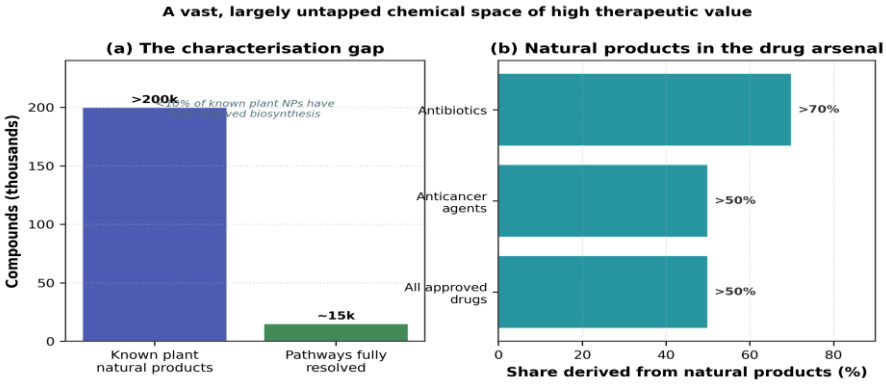

Figure 2. The scale of the opportunity. (a) The characterisation gap: more than 200,000 plant natural products are known, but fewer than ~10% have fully resolved biosynthetic pathways. (b) Natural products contribute disproportionately to the drug arsenal, underlying more than 70% of antibiotics and over half of anticancer and all approved drugs. Values are indicative estimates compiled from the cited literature.

2. FROM RETROSYNTHESIS TO RETROBIOSYNTHESIS

Retrosynthetic analysis, introduced by Corey for organic synthesis, works backward from a target molecule through a series of "disconnections" to progressively simpler precursors. Retrobiosynthesis transposes this logic onto biology: instead of chemical reactions, the disconnections correspond to enzyme-catalysed transformations, and the goal is to trace a target metabolite back to the primary-metabolite building blocks (amino acids, acetyl-CoA, isoprenoid units, sugars) from which the plant assembles it (Figure 1). [8,9] Each predicted retro-step implies a class of enzyme—a methyltransferase, oxidoreductase, glycosyltransferase and so on—and thus a hypothesis about the genes involved.

The value of the approach is twofold. It converts an open-ended search ("which of thousands of genes make this molecule?") into a structured, tractable one ("which genes catalyse this specific predicted step?"), and it can operate on metabolites that have only been predicted, not yet isolated—making it naturally suited to cryptic chemistry. Because the plant metabolome is built from a limited set of biosynthetic scaffolds and building blocks (Table 4), retrobiosynthetic prediction is constrained and therefore feasible, even for complex products. [8,10]

3. AI AND COMPUTATIONAL METHODS

The practicality of retrobiosynthesis at scale rests on recent advances in machine learning (Table 1). Bio-retrosynthesis engines are the core: BioNavi-NP couples deep-learning single-step reaction prediction with tree-search route planning to navigate biosynthetic pathways for natural products, and READRetro, developed specifically for plants, combines dual-view neural architectures with retrieval augmentation and an ensemble strategy to predict multi-step routes from a target to its building blocks. [9,11] Complementary models classify metabolites and predict precursors: graph-transformer–CNN models (GTC) trained on KEGG and fine-tuned on the Plant Metabolic Network predict pathway membership, and automated-ML tools such as DeepMol predict the primary-metabolite precursors of specialised metabolites. [12,13] Newer transformer- and graph-based systems (NAG2G, RetroExplainer, and human-in-the-loop planners) continue to improve accuracy and interpretability. [14]

Tool / model	Type / architecture	Primary task
READRetro	Retrieval-augmented dual-view neural ensemble	Multi-step plant bio-retrosynthesis
BioNavi-NP	Deep single-step prediction + tree search	Biosynthetic route navigation

Tool / model	Type / architecture	Primary task
GTC	Graph transformer + CNN (transfer learning)	Pathway-class prediction
DeepMol	Automated ML (interpretable classifiers)	Precursor prediction
plantiSMASH / antiSMASH	Rule-based genome mining	BGC / cluster detection
AlphaFold, ESMFold	Deep-learning structure prediction	Enzyme structure modelling
CoExpPhylo	Co-expression + phylogeny	Candidate biosynthetic-gene discovery

Table 1. Representative AI and computational tools spanning the retrobiosynthesis-guided discovery workflow, from route prediction and precursor inference to genome mining, enzyme-structure modelling and candidate-gene discovery.

These methods are only as good as the data that train them, and a set of curated resources underpins the field (Table 2): reaction and pathway databases (KEGG, MetaCyc, the Plant Metabolic Network), the MIBiG repository of characterised BGCs, and mass-spectral libraries organised through molecular-networking platforms such as GNPS. Figure 3 maps the principal tools onto the capabilities they provide, making clear that no single tool spans the whole workflow—integration is essential.

Resource	Content	Role in workflow
KEGG / MetaCyc	Reference reactions & pathways	Training data; route templates
Plant Metabolic Network	Plant-specific pathways & compounds	Plant fine-tuning; annotation
MIBiG	Characterised biosynthetic gene clusters	Genome-mining reference
GNPS / molecular networking	Tandem-MS spectral libraries	Metabolite dereplication & discovery
UniProt / structural databases	Enzyme sequences & structures	Enzyme-function assignment

Table 2. Key data resources supporting AI-driven retrobiosynthesis. Model performance and coverage depend directly on the breadth and quality of these curated genomic, metabolic and spectral databases.

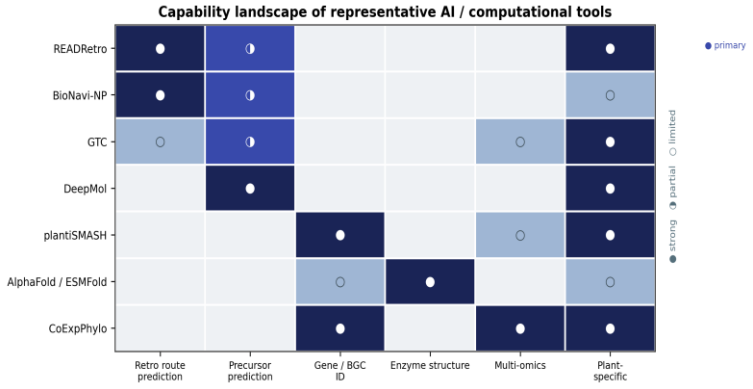

Figure 3. Capability landscape of representative computational tools across six functions of the discovery workflow (filled = strong, half = partial, open = limited). Bio-retrosynthesis engines excel at route and precursor prediction; genome-mining and co-expression tools at gene identification; structure predictors at enzyme modelling. No single tool spans the workflow, underscoring the need for integration.

Natural Resources for Human Health
eISSN: 2583-1194

4. AN INTEGRATED AI-ASSISTED DISCOVERY WORKFLOW

Because the individual methods are complementary, their power is realised only when they are combined into a single iterative pipeline (Figure 4, Table 3). We propose a five-stage workflow. First, a target is defined—either a compound of interest or a mass-spectral feature flagged as novel by untargeted metabolomics and molecular networking. Second, a bio-retrosynthesis engine predicts one or more candidate routes from the target back to primary-metabolite building blocks, decomposing the problem into a sequence of defined enzymatic steps. Third, each predicted step is matched to candidate genes and enzymes by genome mining and multi-omics machine learning—integrating genomic, transcriptomic (co-expression) and metabolomic evidence to nominate the genes most likely to catalyse it. Fourth, protein-structure prediction (AlphaFold, ESMFold), sequence-similarity-network analysis and molecular docking refine and rank these candidates by structural and mechanistic plausibility. Fifth, the reconstructed pathway is tested experimentally—typically by heterologous expression in a microbial or plant chassis and enzymatic assay. [2,7,10,15] Crucially, the workflow is iterative: experimental results feed back to retrain and refine the models, an active-learning loop that progressively improves prediction.

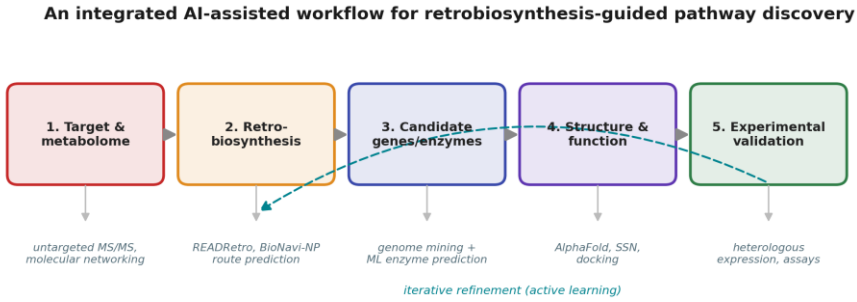

Figure 4. An integrated AI-assisted workflow for retrobiosynthesis-guided pathway discovery. A target or metabolomic feature (1) is decomposed by bio-retrosynthesis into candidate routes (2); each step is matched to candidate genes/enzymes by genome mining and machine learning (3); structure prediction and network analysis refine enzyme assignments (4); and heterologous expression validates the pathway (5). Experimental results feed back to refine the models (active learning).

Stage	Principal methods / tools	Output
1 Target definition	Untargeted MS/MS; molecular networking (GNPS)	Target structure or novel feature
2 Retrobiosynthesis	READRetro, BioNavi-NP; precursor ML	Candidate biosynthetic route(s)
3 Gene / enzyme nomination	plantiSMASH; multi-omics ML; co-expression	Ranked candidate genes
4 Structure & function	AlphaFold/ESMFold; SSN; docking	Refined enzyme assignments
5 Validation	Heterologous expression; enzyme assays	Confirmed pathway; model feedback

Table 3. The five stages of the integrated retrobiosynthesis-guided discovery workflow, their methods and outputs. The pipeline is iterative, with experimental validation feeding back to improve the predictive models.

4.1 Multi-omics integration

The third stage—linking predicted reactions to real genes—is where plant-specific difficulty is greatest and where multimodal AI is most transformative. Rather than treating genomics, transcriptomics and metabolomics separately, modern systems learn a shared latent representation in which chemically related metabolites cluster near their biosynthetic genes (Figure 5). [2,10] This enables the model not merely to correlate features but to predict uncharacterised metabolites, annotate cryptic pathways and even identify the regulatory bottlenecks that control metabolite flux. Co-expression analysis across conditions and tissues (exploited by tools such as CoExpPhylo) is especially powerful in plants, because genes in a shared pathway tend to be co-regulated even when they are not physically clustered—precisely the situation that defeats cluster-based genome mining. [7,10]

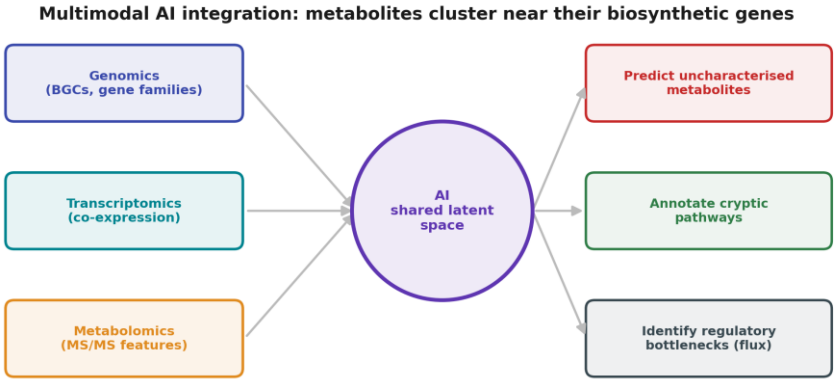

Figure 5. Multimodal AI integration of genomics, transcriptomics and metabolomics. By learning a shared latent space in which metabolites cluster near their biosynthetic genes, integrated models can predict uncharacterised metabolites, annotate cryptic pathways and identify regulatory bottlenecks controlling flux—capabilities unattainable from any single omics layer alone.

5. ENZYME-FUNCTION PREDICTION AND A WORKED EXAMPLE

Assigning function to candidate enzymes is the linchpin of validation, and here structural AI has been decisive. AlphaFold, ESMFold and RoseTTAFold now predict enzyme structures at scale, and—combined with sequence-similarity-network (SSN) analysis, which groups enzymes by relatedness to reveal functional sub-families, and molecular docking of predicted substrates—they allow candidate enzymes for each retrobiosynthetic step to be ranked before any experiment. [2,16] This structure-informed triage is what makes the workflow tractable: it narrows highly duplicated plant gene families down to a testable handful.

Figure 6 illustrates the logic with the benzylisoquinoline alkaloids (BIAs), a pharmacologically important class (the group that includes morphine and berberine precursors) and a model for metabolic engineering. Working backward from a target BIA such as (S)-reticuline, retrobiosynthesis disconnects it to the central intermediate (S)-norcoclaurine, then to the condensation of dopamine and 4-hydroxyphenylacetaldehyde (4-HPAA), and finally to the primary-metabolite precursor L-tyrosine. Each disconnection nominates an enzyme class—norcoclaurine synthase for the key Pictet–Spengler condensation, decarboxylases and transaminases upstream—that machine learning and structural modelling can then match to specific genes. Notably, machine-learning approaches have recently helped uncover "missing link" enzymes hidden within duplicated plant gene families in exactly such alkaloid pathways, demonstrating the strategy in practice. [15] The example is deliberately illustrative rather than a novel result, but it shows how an abstract retrobiosynthetic tree becomes a concrete, testable set of gene hypotheses.

Natural Resources for Human Health
eISSN: 2583-1194

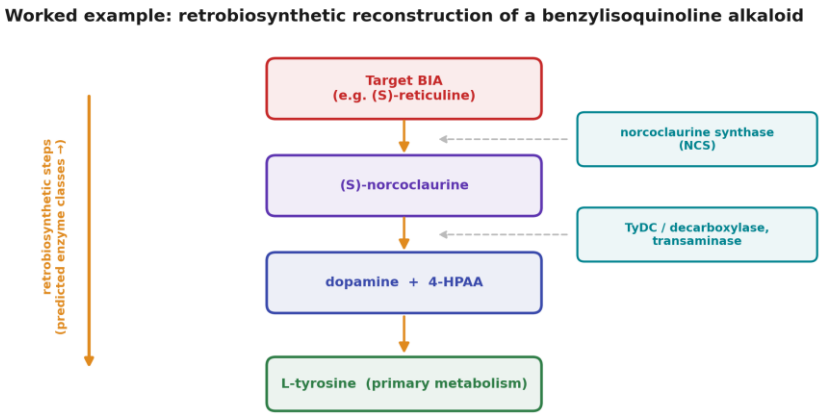

Figure 6. Worked conceptual example: retrobiosynthetic reconstruction of a benzylisoquinoline alkaloid. Successive disconnections trace the target (e.g. (S)-reticuline) through (S)-norcoclaurine to dopamine and 4-HPAA and, ultimately, to L-tyrosine; each step nominates an enzyme class (dashed) that machine learning and structural modelling can match to candidate genes. The scheme is illustrative of the workflow, not a novel finding.

Metabolite class	Building block / precursor	Representative products	Core scaffold
Alkaloids	Amino acids (Tyr, Trp, Lys, Orn)	Morphine, vinblastine, berberine	N-heterocyclic
Terpenoids	Isoprenoid units (IPP/DMAPP)	Artemisinin, taxol, menthol	Isoprene-derived
Phenylpropanoids / polyphenols	Phenylalanine → cinnamate	Flavonoids, lignans, stilbenes	C6–C3 / C6–C3–C6
Polyketides	Acetyl-/malonyl-CoA	Anthraquinones, curcuminoids	Poly-β-keto
Glucosinolates / others	Amino acids + sugar/sulfur	Sulforaphane precursors	Thioglucoside

Table 4. Major classes of plant secondary metabolites and their primary-metabolite building blocks. Because the plant metabolome is assembled from this limited set of precursors and scaffolds, retrobiosynthetic prediction is constrained and therefore computationally tractable.

6. CHALLENGES AND FUTURE PERSPECTIVES

Substantial obstacles remain (Table 5). Enzyme promiscuity and novelty: many plant enzymes are catalytically promiscuous or genuinely novel, so models trained on known reactions may miss the "atypical" chemistry that makes cryptic metabolites interesting; predictions biased toward the known are a real risk. Data scarcity and imbalance: high-quality, plant-specific training data are limited and skewed toward well-studied pathways, degrading performance on under-represented classes. Plant genome complexity: large, polyploid, repeat-rich genomes with dispersed, highly duplicated biosynthetic genes complicate every stage from mining to gene assignment. The prediction–validation gap: computational routes and enzyme assignments are hypotheses, and bioinformatic predictions still frequently diverge from experimentally determined structures, so experimental confirmation remains indispensable and is the throughput bottleneck. Interpretability: black-box predictions are hard to trust and act upon, motivating explainable and human-in-the-loop systems (RetroExplainer, DeepRetro). [2,4,10,14] Progress is likely to come from larger and better-curated multi-omics datasets, foundation models that integrate sequence, structure and chemistry, tighter

coupling of prediction with automated, high-throughput experimental validation ("self-driving" discovery), and generative approaches that propose genuinely novel enzymes and pathways rather than retrieving known ones. As these mature, retrobiosynthesis-guided AI could shift natural-product discovery from serendipity to systematic design.

Challenge	Emerging solution
Enzyme promiscuity / novel chemistry	Generative models; structure-based function prediction (AlphaFold, SSN)
Data scarcity & imbalance	Curated multi-omics datasets; transfer & few-shot learning
Plant genome complexity (dispersed genes)	Co-expression + phylogeny (CoExpPhylo); multimodal integration
Prediction–validation gap	High-throughput heterologous expression; active-learning loops
Interpretability / trust	Explainable AI; human-in-the-loop planners

Table 5. Principal challenges of retrobiosynthesis-guided, AI-assisted discovery and the emerging approaches addressing them. The prediction–validation gap remains the central bottleneck and the focus of "self-driving" discovery efforts.

7. CONCLUSION

The chemical space of plant secondary metabolism is vast, therapeutically rich and largely unexplored, and much of it is cryptic—hidden in silent genes or predicted but never isolated. Conventional pathway elucidation cannot scale to illuminate it, particularly in plants whose biosynthetic genes are dispersed rather than clustered. Retrobiosynthesis, which applies retrosynthetic logic to enzyme-catalysed reactions, reframes the problem into a tractable, predictive one, and artificial intelligence now makes it practical at scale. AI bio-retrosynthesis engines predict biosynthetic routes backward from a target; machine learning and genome mining nominate the responsible genes; structural AI refines enzyme-function assignment; and heterologous expression validates the reconstruction—ideally within an iterative, self-improving loop. Integrated into the multi-omics workflow proposed here, these components can predict uncharacterised metabolites, annotate cryptic pathways and reconstruct the routes that produce them. Formidable challenges persist—enzyme promiscuity, data scarcity, genome complexity and, above all, the gap between prediction and experimental proof—but the trajectory is clear. Retrobiosynthesis-guided, AI-assisted reconstruction promises to transform the discovery of plant secondary metabolites from a slow, serendipitous art into a systematic, scalable science, with far-reaching implications for drug discovery, synthetic biology and our understanding of plant chemical diversity.

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 data and tools discussed are available in the cited references.

REFERENCES

[1] Yonekura-Sakakibara K, Saito K. Review: genetically modified plants for the promotion of human health. Biotechnology Letters. 2009;31:1983–1991.
[2] Toward an integrated omics approach for plant biosynthetic pathway discovery in the age of AI. (Review). 2025.
[3] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 1981 to 2019. Journal of Natural Products. 2020;83(3):770–803.

[4] From data to discovery: leveraging big data in plant natural products biosynthesis research. New Phytologist / (Review). 2025.
[5] Wang Y, Mao XM. Mining silent biosynthetic gene clusters for natural products in filamentous fungi. Chemistry & Biodiversity. 2025;22:e202402715.
[6] Cryptic metabolites and methods for activating silent biosynthetic gene clusters in diverse microorganisms. (Patent / review). 2023.
[7] Grünig N, Pucker B. CoExpPhylo — a novel pipeline for biosynthesis gene discovery. BMC Genomics. 2025;26:807. doi:10.1186/s12864-025-12061-3
[8] Corey EJ, Cheng XM. The Logic of Chemical Synthesis. Wiley; 1989.
[9] Kim T, et al. READRetro: natural product biosynthesis prediction with retrieval-augmented dual-view retrosynthesis. New Phytologist. 2024;243(6):2512–2527.
[10] Artificial intelligence-guided biosynthesis and retrosynthesis in pharmacognosy: toward synthesis-oriented discovery of natural-product therapeutics. (Review). 2025.
[11] Zheng S, Zeng T, Li C, et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nature Communications. 2022;13:3342.
[12] Bao H, Zhao J, Zhao X, et al. Prediction of plant secondary metabolic pathways using deep transfer learning. BMC Bioinformatics. 2023;24:348.
[13] Predicting precursors of plant specialized metabolites using DeepMol automated machine learning. (Journal). 2025.
[14] Toward synthesis-oriented discovery: NAG2G, RetroExplainer, RSGPT and human-in-the-loop retrosynthesis (DeepRetro). (Review). 2025.
[15] Vavricka CJ, Takahashi S, Watanabe N, et al. Machine learning discovery of missing links that mediate alternative branches to plant alkaloids. Nature Communications. 2022;13:1405. doi:10.1038/s41467-022-28883-8
[16] Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.
[17] Bai W, Li C, Li W, et al. Machine learning assists prediction of genes responsible for plant specialized metabolite biosynthesis by integrating multi-omics data. BMC Genomics. 2024;25:394.