Advances in Basic and Life Sciences
Vol. 1, No. 1 (2026), 1–17
© RK Scientific Publishers
Online Published on 28.11.2026
Review Article

AI-Driven Single-Cell and Spatial Multi-Omics for Medicinal Plant Drug Discovery

Romeo Khongbantabam[1,†]

[1]Poincare Institute of Advanced Mathematical Sciences,
Rohini, New Delhi-110085. India.
romeokhongbantabam@yahoo.in

Submitted: 02.10.2026 | Revised: 20.11.2026 | Accepted: 21.11.2026

Abstract. Medicinal plants provide structurally diverse natural products, but conventional bulk-omics and phytochemical workflows often obscure the rare cells, transport processes and regulatory mechanisms responsible for specialized metabolism. This critical narrative review evaluates single-cell transcriptomics, chromatin accessibility, proteomics, spatial transcriptomics and metabolomics alongside machine learning, graph models and emerging foundation models. Particular attention is given to experimentally demonstrated monoterpene indole alkaloid biosynthesis in *Catharanthus roseus*, contrasted with proposed applications to paclitaxel, artemisinin, curcuminoids and withanolides. We compare measurement limitations, distinguish computational association from causal validation and propose an evidence-based workflow connecting molecular maps to enzyme assays, genetic perturbation, compound authentication and pharmacological testing. Although AI-assisted multi-omics offers substantial opportunities, generalization across plant species, chemical identification and reproducibility remain central challenges. A staged translational roadmap is presented to support rigorous, sustainable natural-product discovery.

Keywords: medicinal plants, single-cell RNA sequencing, spatial transcriptomics, metabolomics, natural products, artificial intelligence, biosynthetic pathways, multi-omics integration.
[†]**Corresponding author**

1. Introduction

The integration of artificial intelligence (AI) with single-cell and spatial multi-omics technologies has opened new avenues for understanding the molecular complexity of medicinal plants and accelerating natural product drug discovery. Medicinal plants represent a rich source of structurally diverse bioactive compounds with considerable therapeutic potential. However, conventional experimental approaches often fail to capture the cellular heterogeneity and spatial organization underlying specialized metabolite biosynthesis. Recent advances in single-cell transcriptomics, spatial transcriptomics, proteomics, metabolomics, and epigenomics have enabled the investigation of complex biological processes at unprecedented resolution. Artificial intelligence, particularly machine learning and deep learning, provides powerful computational tools for integrating these high-dimensional datasets, identifying biosynthetic genes, reconstructing metabolic pathways, and predicting biologically active

compounds. The convergence of these technologies facilitates a comprehensive understanding of cell-specific metabolic regulation and the discovery of previously unexplored therapeutic molecules. Furthermore, AI-assisted multi-omics approaches offer promising opportunities for precision metabolic engineering, sustainable production of valuable phytochemicals, and rational drug development. This review critically examines recent advances, emerging applications, methodological challenges, and future perspectives of AI-driven single-cell and spatial multi-omics in medicinal plant drug discovery.

Scope and critical perspective. Medicinal plants have supplied pharmacologically important natural products, but the transition from a promising extract to a validated drug requires chemical identification, mechanism-of-action evidence, pharmacokinetics, safety testing and clinical evaluation [24, 3]. Genome sequencing and bulk transcriptomics identify plausible pathway components, yet bulk measurements conflate specialized cells with abundant background tissue. Single-cell and spatial methods make it possible to ask not merely which genes are present, but in which cells biosynthesis, transport and storage occur. The strongest medicinal-plant demonstration is the work on *Catharanthus roseus*, where single-cell transcriptomic and metabolomic evidence identified cell-partitioned alkaloid biosynthesis and facilitated gene discovery [18]. This review distinguishes established experimental results, transferable computational methods and proposals still requiring validation.

Review approach and evidentiary boundaries. This is a critical narrative review rather than a systematic review or meta-analysis. We prioritize original methodological articles, experimentally validated medicinal-plant studies and foundational computational papers. Biomedical and model-plant methods are discussed as potentially transferable, not as proof that they work without modification in medicinal plants. No quantitative claims of sensitivity, cost reduction or drug-development success are made without direct comparative evidence. The manuscript focuses on the chain from spatially resolved pathway discovery to biochemical confirmation and preclinical lead assessment.

2. Evolution of Omics Technologies

The evolution of omics technologies has revolutionized modern biological research by enabling comprehensive investigations of the molecular components and complex regulatory networks governing living systems. Beginning with genomics, which focuses on the structure and function of entire genomes, the field has progressively expanded to include transcriptomics, proteomics, metabolomics, epigenomics, and other specialized disciplines. Advances in high-throughput sequencing, mass spectrometry, and computational biology have substantially improved the accuracy, sensitivity, and efficiency of molecular analyses. More recently, the emergence of single-cell and spatial omics technologies has overcome several limitations of conventional bulk analyses by revealing cellular heterogeneity and the spatial organization of molecular processes within biological tissues. In medicinal plant research, these developments have facilitated the identification of biosynthetic genes, characterization of specialized metabolic pathways, and investigation of regulatory mechanisms responsible for producing therapeutically valuable phytochemicals. Furthermore, the integration of multiple omics platforms with artificial intelligence and advanced computational approaches has enabled a more comprehensive and predictive understanding of complex biological systems. This section provides an overview of the evolution, fundamental principles, major technological advances,

and applications of omics technologies, with particular emphasis on their contributions to medicinal plant research and natural product drug discovery.

Genomics and transcriptomics. Long-read assemblies, chromosome-scale scaffolding and transcriptome annotation establish the gene catalog needed to distinguish paralogous biosynthetic enzymes. RNA sequencing then reveals inducible and tissue-dependent expression, while single-cell sequencing resolves rare producer cells [33, 23, 18]. Transcript abundance is not equivalent to enzyme activity: post-transcriptional regulation, transport and turnover may uncouple expression from metabolite accumulation.

Proteomics and epigenomics. Proteomics measures enzymes and their modifications more directly than RNA sequencing, but plant secondary metabolites and cell walls complicate extraction. Single-cell proteomics remains substantially less mature in plants than scRNA-seq; the SCoPE-MS study illustrates a transferable mass-spectrometry strategy, not a demonstrated medicinal-plant pipeline [6, 29]. Chromatin accessibility, measured using ATAC-seq, can prioritize transcription-factor binding sites and regulatory elements, but motif enrichment alone does not establish binding or causal regulation [4, 5].

Metabolomics and chemical annotation. LC–MS/MS, GC–MS and NMR characterize specialized metabolites and help connect pathway hypotheses to chemical products. Molecular networking and in silico fragmentation facilitate candidate annotations, but isomeric compounds often require authentic standards and orthogonal structure determination [35, 10]. Spatial metabolite patterns reflect synthesis, storage, movement and degradation; they cannot be interpreted as biosynthetic sites without complementary evidence.

3. Single-Cell Omics

Single-cell omics has emerged as a transformative approach in modern molecular biology, enabling the comprehensive characterization of biological processes at the resolution of individual cells. Unlike conventional bulk omics technologies, which provide averaged molecular profiles across heterogeneous cell populations, single-cell approaches reveal cellular diversity, rare cell populations, and cell-specific regulatory mechanisms that would otherwise remain undetected. Recent advances in single-cell RNA sequencing (scRNA-seq), single-cell chromatin accessibility sequencing (scATAC-seq), single-cell proteomics, and metabolomics have significantly enhanced our understanding of cellular functions, developmental processes, and molecular interactions. In plant biology, these technologies provide valuable insights into cell differentiation, tissue organization, stress responses, and specialized metabolic pathways. Particularly in medicinal plants, single-cell omics facilitates the identification of metabolite-producing cell types, characterization of biosynthetic gene expression, and elucidation of regulatory networks associated with therapeutically important phytochemicals. Furthermore, the integration of single-cell datasets with artificial intelligence and machine learning enables advanced cell-type classification, developmental trajectory analysis, gene regulatory network reconstruction, and prediction of biosynthetic pathways. Despite challenges associated with plant cell walls, cell isolation, technical variability, and computational complexity, single-cell omics offers promising opportunities for advancing medicinal plant biotechnology and natural product drug discovery. This section discusses the fundamental principles,

technological developments, major applications, and emerging challenges of single-cell omics, with particular emphasis on their relevance to medicinal plant research.

Single-cell RNA sequencing. Droplet-based assays enabled large-scale single-cell transcriptome profiling [23]. A defensible workflow includes independent biological replicates, tissue-specific dissociation optimization, empty-droplet filtering, doublet assessment, normalization, batch-aware integration, marker-based annotation and pseudobulk differential-expression testing [21, 13, 31]. Plant roots have served as model systems for cell-state reconstruction [9, 28]. In medicinal plants, protoplasting can selectively lose large, lignified or fragile cells and induce stress responses; matched single-nucleus and in situ experiments are therefore valuable.

Single-cell chromatin accessibility. Single-cell ATAC-seq reveals regulatory heterogeneity by assaying accessible chromatin, while integrated RNA–ATAC analysis can connect candidate transcription factors to cell-specific pathway genes [5, 32]. Such links are hypotheses until supported by perturbation, reporter assays or biochemical binding experiments. In *C. roseus*, integration of expression, accessibility and binding information supported identification and functional validation of the idioblast regulator CrIDM1 [19].

Single-cell proteomics and metabolomics. Protein and metabolite measurements provide complementary evidence beyond RNA abundance. Low analyte quantities, ion suppression and variable extraction recovery make single-cell mass spectrometry difficult; cross-platform calibration and cell-type verification are essential [6, 29]. The *C. roseus* study combined single-cell metabolite profiling with transcriptomic maps to identify a reductase in anhydrovinblastine biosynthesis [18]. This is a stronger functional case than a gene–metabolite correlation alone.

4. Spatial Omics

Spatial omics has emerged as a transformative approach in modern molecular biology, enabling the comprehensive investigation of molecular processes while preserving their spatial organization within biological tissues. Unlike conventional bulk and dissociation-based single-cell omics technologies, which often lose information about the original tissue architecture, spatial omics provides valuable insights into the localization, distribution, and interactions of genes, proteins, and metabolites within their native cellular environments. Recent advances in spatial transcriptomics, spatial proteomics, and spatial metabolomics have significantly enhanced our understanding of tissue heterogeneity, cell–cell communication, and spatially regulated biological processes. In medicinal plant research, these technologies offer promising opportunities for identifying specialized metabolite-producing regions, investigating tissue-specific biosynthetic pathways, and characterizing the spatial distribution of therapeutically important phytochemicals. Furthermore, the integration of spatial omics with single-cell sequencing and artificial intelligence facilitates the reconstruction of complex molecular networks, identification of spatially associated gene expression patterns, and prediction of metabolic interactions across different plant tissues. Advanced computational approaches, including machine learning, deep learning, and graph-based models, further support the analysis and interpretation of high-dimensional spatial datasets. Despite challenges associated with spatial resolution, analytical sensitivity, tissue preparation, data integration,

and computational complexity, spatial omics holds considerable potential for advancing medicinal plant biotechnology and natural product drug discovery. This section discusses the fundamental principles, recent technological developments, major applications, and emerging challenges of spatial transcriptomics, spatial metabolomics, and spatial proteomics, with particular emphasis on their contributions to understanding specialized metabolism and accelerating medicinal plant-derived drug discovery.

Spatial transcriptomics. Spatial transcriptomics retains tissue context while measuring gene expression, an important advantage for vascular tissues, epidermal cells, idioblasts and secretory structures. Early array-based methods established spatially indexed expression profiling [30]; later bead-based approaches increased spatial resolution [26]. Platform choice involves trade-offs among spatial resolution, sensitivity, gene coverage, cost and compatibility with thick, pigmented or lignified plant sections. Spot-level measurements may mix cells; cell-type deconvolution should be evaluated against microscopy and marker localization.

Spatial metabolomics. Mass-spectrometry imaging can localize molecular ions without tissue dissociation, potentially revealing alkaloid and terpenoid gradients. However, ion images do not by themselves identify molecular structures, prove flux direction or distinguish biosynthesis from storage. Orthogonal LC–MS/MS, stable-isotope tracing and laser-capture sampling should be combined with anatomical registration and standards [35, 10, 18].

Spatial proteomics and cross-modal alignment. Imaging-based protein localization and spatially resolved mass spectrometry can validate candidate enzymes, transporters and transcription factors. Cross-modal registration must account for adjacent-section differences, image distortion and incompatible spatial scales. Integration should preserve the original measurement resolution and report uncertainty rather than treating imputed cellular values as direct observations [14, 1].

5. Artificial Intelligence in Plant Biology

Artificial intelligence (AI) has emerged as a powerful interdisciplinary approach in modern plant biology, transforming the analysis, interpretation, and prediction of complex biological processes. Recent advances in machine learning, deep learning, neural networks, and transformer-based architectures have significantly enhanced the ability to analyze large-scale genomic, transcriptomic, proteomic, metabolomic, and phenotypic datasets. AI-driven computational methods facilitate the identification of complex molecular patterns, prediction of gene functions, reconstruction of regulatory networks, and characterization of metabolic pathways that are difficult to investigate using conventional analytical techniques. In medicinal plant research, these approaches offer promising opportunities for identifying biosynthetic genes, predicting bioactive compounds, understanding specialized metabolite production, and accelerating natural product drug discovery. Furthermore, the integration of AI with single-cell and spatial multi-omics technologies enables the investigation of cellular heterogeneity, tissue-specific gene expression, and spatially regulated metabolic processes with greater precision. Emerging computational frameworks, including graph neural networks, explainable AI, and foundation models, are further expanding the possibilities for predictive modeling and knowledge discovery in plant systems. Despite challenges related to data quality, model interpretability, computational requirements, and experimental validation,

AI-driven approaches have considerable potential to advance plant biotechnology, precision metabolic engineering, and sustainable phytochemical production. This section provides an overview of major AI methodologies, their applications in plant biology, and their emerging contributions to medicinal plant research and drug discovery.

Classical machine learning. Random forests, support-vector machines and gradient boosting can rank pathway candidates from sequence features, expression, phylogenetic context and metabolite covariation. These methods are useful baselines, especially for modest sample sizes. Training and test partitions must be separated by species, pathway family or experiment when the intended deployment requires extrapolation; random splitting of highly similar paralogs can inflate apparent accuracy.

Deep learning and graph neural networks. Variational autoencoders can represent sparse single-cell counts while modeling technical variation [20]. Graph neural networks provide a language for gene regulatory, enzyme–reaction and molecule–target networks [17, 11]. The graph itself is a biological assumption: missing edges, incorrect homology and biased interaction databases can propagate errors into predictions.

Transformers, foundation models and large language models. Transformer attention mechanisms support representation learning across long sequences and sets [34]. scGPT demonstrates foundation-model approaches to human-dominated single-cell datasets [8]; direct transfer to underrepresented medicinal-plant taxa requires orthology-aware gene vocabularies, benchmarking and fine-tuning. Large language models can assist literature triage and annotation but may fabricate citations or mechanisms; every asserted gene function and chemical identity requires primary-source checking.

6. AI-Driven Multi-Omics Integration

AI-driven multi-omics integration has emerged as a powerful approach for understanding complex biological systems by combining molecular information from genomics, transcriptomics, proteomics, metabolomics, epigenomics, and other omics platforms within a unified computational framework. Conventional single-omics analyses often provide incomplete insights into biological processes, whereas integrated multi-omics approaches enable a more comprehensive understanding of molecular interactions, regulatory mechanisms, and metabolic pathways. Recent advances in artificial intelligence, particularly machine learning, deep learning, graph neural networks, and transformer-based architectures, have significantly improved the analysis and interpretation of heterogeneous, high-dimensional biological datasets. In medicinal plant research, AI-driven multi-omics integration facilitates the identification of biosynthetic genes, reconstruction of specialized metabolic pathways, characterization of gene regulatory networks, and prediction of therapeutically important secondary metabolites. Furthermore, integrating single-cell and spatial omics data enables researchers to investigate cell-specific metabolic activities, spatial gene expression patterns, and molecular interactions within complex plant tissues. Advanced computational frameworks also provide opportunities for discovering previously unrecognized relationships among genes, proteins, metabolites, and their associated biological functions. Despite challenges related to data heterogeneity, batch effects, computational complexity, model interpretability, and experimental validation, AI-driven multi-omics integration holds considerable promise for

accelerating natural product drug discovery and precision metabolic engineering. This section discusses the major computational strategies, emerging AI methodologies, and applications of multi-omics integration, with particular emphasis on computational pipelines, graph neural networks, and explainable artificial intelligence in medicinal plant research.

Computational pipeline and data provenance. A practical pipeline begins with taxonomically authenticated specimens, voucher records, matched tissue collection, replicate-aware sequencing and metabolite measurements. After modality-specific quality control, cells or tissue regions are annotated using morphology and validated markers. Shared latent-factor approaches such as MOFA+ and multimodal integration methods can identify common and modality-specific variation [1, 14, 2]. Crucially, unpaired modalities require uncertainty-aware mapping rather than an assumption of one-to-one cell correspondence.

Gene–metabolite networks and candidate ranking. A candidate biosynthetic gene should be supported by several independent signals: sequence-domain compatibility, co-localization with pathway intermediates, cell-type-specific coexpression, inducible response, phylogenetic placement and perturbation evidence. Correlation alone cannot distinguish biosynthesis from co-regulation or transport. Graph-based prioritization can incorporate known enzyme reactions and pathway topology [11, 18].

Explainability, uncertainty and validation. Feature attribution methods such as SHAP can expose influential input features but do not establish causal biological mechanisms [22]. Model calibration, out-of-distribution detection, confidence intervals and negative controls should accompany candidate rankings. Functional validation should include recombinant enzyme assays, CRISPR or silencing perturbations, isotope-labeled precursor feeding and independent metabolite identification.

7. Drug Discovery Applications

The integration of artificial intelligence (AI) with advanced omics technologies has opened new avenues for accelerating drug discovery by enabling the systematic identification, characterization, and optimization of biologically active compounds. Medicinal plants represent an invaluable source of structurally diverse natural products with significant therapeutic potential, although conventional drug discovery approaches are often time-consuming, expensive, and experimentally demanding. Recent advances in machine learning, deep learning, computational chemistry, and multi-omics integration have provided powerful tools for predicting bioactive compounds, identifying molecular targets, elucidating biosynthetic pathways, and evaluating pharmacological properties. Furthermore, single-cell and spatial omics technologies enable the identification of specialized metabolite-producing cell populations and the characterization of tissue-specific biosynthetic processes, thereby facilitating the discovery of previously unexplored phytochemicals. AI-assisted molecular docking, virtual screening, quantitative structure–activity relationship modeling, and toxicity prediction further support the prioritization of promising drug candidates for experimental evaluation. In addition, the integration of AI with metabolic engineering and precision breeding offers opportunities to enhance the production, quality, and sustainability of therapeutically valuable plant-derived compounds. Despite these advances, challenges associated with data reliability, biological validation, pharmacokinetic properties, and clinical

translation remain significant. This section examines the major applications of AI-driven multi-omics approaches in medicinal plant drug discovery, with particular emphasis on bioactive compound prediction, metabolic engineering, and precision breeding, highlighting their potential to accelerate the development of novel plant-derived therapeutics.

Prediction of bioactive compounds. AI may prioritize structures for target binding, permeability, toxicity and novelty, while structure prediction tools such as AlphaFold can help assess protein folds [15]. These predictions do not demonstrate therapeutic efficacy. Hits require chemical purity assessment, orthogonal target-engagement assays, selectivity profiling, ADME studies and in vivo validation before claims of drug potential [3].

Metabolic engineering. Engineering aims to increase pathway flux, remove competing reactions and improve compartment-specific precursor supply. Engineered yeast production of artemisinic acid and later semisynthetic artemisinin illustrates a well-established translational pathway [27, 25]. For multicellular plant pathways, transport and cell-specific regulation may be as important as enzyme expression; computational flux models should be tested experimentally.

Precision breeding and sustainability. Genome-informed selection and metabolite phenotyping can identify chemotypes with favorable production characteristics. However, field conditions, genotype-by-environment interactions, conservation status, and equitable sourcing influence feasibility. Species identification, voucher deposition and transparent reporting of cultivation conditions are necessary for reproducible medicinal-plant studies.

8. Case Studies

Case studies provide valuable insights into the practical applications of artificial intelligence (AI), single-cell technologies, and spatial multi-omics in understanding the biosynthesis and pharmaceutical potential of medicinal plant-derived natural products. Several therapeutically important phytochemicals, including paclitaxel (Taxol), artemisinin, curcumin, and withanolides, serve as representative examples for investigating the complex molecular mechanisms underlying specialized metabolite production. Advances in genomics, transcriptomics, proteomics, and metabolomics have significantly improved our understanding of the biosynthetic pathways, regulatory networks, and metabolic diversity associated with these compounds. More recently, single-cell and spatial omics approaches have provided opportunities to investigate cell-specific gene expression, tissue-specific metabolite accumulation, and the spatial organization of biosynthetic processes. The integration of these technologies with AI-driven computational methods offers promising strategies for identifying previously uncharacterized biosynthetic genes, predicting enzyme functions, reconstructing metabolic pathways, and optimizing metabolite production. Furthermore, machine learning and deep learning approaches can facilitate the prioritization of candidate genes and bioactive molecules for subsequent experimental validation. However, the extent of experimental evidence supporting these integrated approaches varies considerably among different medicinal plant species and their metabolites. This section presents selected case studies involving paclitaxel, artemisinin, curcumin, and withanolides, highlighting established scientific advances, emerging computational applications, existing research gaps, and future opportunities for AI-assisted medicinal plant drug discovery.

Catharanthus roseus: a directly demonstrated single-cell multi-omics case. The clearest direct evidence is *C. roseus*, whose monoterpene indole alkaloid pathway is partitioned among internal phloem-associated parenchyma, epidermal cells and idioblasts. Li and colleagues integrated a chromosome-scale genome, chromatin conformation, single-cell RNA sequencing and single-cell metabolomics to identify transport and enzymatic candidates, including a reductase involved in anhydrovinblastine formation [18]. Subsequent cell-type-aware regulatory analysis identified and functionally tested CrIDM1 [19]. These studies show how cell-type resolution can produce testable mechanistic discoveries; they should not be described as proof that AI autonomously discovered an approved drug.

Paclitaxel (Taxol). Paclitaxel biosynthesis in *Taxus* involves complex diterpenoid transformations, cytochrome P450 reactions and tissue-dependent regulation [7]. High-quality genomes, single-nucleus atlases and spatial metabolite imaging could prioritize unresolved pathway components. This is a proposed integrative strategy rather than a claim that a complete AI–spatial pipeline has already resolved the pathway.

Artemisinin. *Artemisia annua* produces artemisinin through specialized terpene metabolism, with glandular trichomes playing important roles. The production of artemisinic acid in engineered yeast and subsequent semisynthetic manufacture are landmark examples of pathway engineering [27, 25]. Spatial profiling may help investigate precursor partitioning and environmental regulation, but clinical effectiveness and industrial viability must not be inferred from expression patterns.

Curcuminoids. In *Curcuma longa*, biochemical work identified type III polyketide synthases involved in curcuminoid biosynthesis [16]. Integrating rhizome cell atlases, metabolite imaging and expression studies could clarify cell-specific accumulation. Curcumin itself is a bioactive research compound, not an established clinical drug across the wide range of diseases sometimes claimed in review literature.

Withanolides. Withanolides from *Withania somnifera* illustrate a high-priority but incompletely resolved specialized-metabolism problem. A robust future study would compare authenticated chemotypes and organs, map candidate steroid-modifying enzymes at cellular resolution and validate intermediates using isotope tracing and targeted gene perturbations. In the absence of direct studies for a specific enzyme or tissue, predictions must remain explicitly hypothetical.

9. Challenges and Reporting Standards

Despite the remarkable progress in artificial intelligence (AI), single-cell technologies, and spatial multi-omics, several technical, computational, and methodological challenges continue to limit their widespread application in medicinal plant research and natural product drug discovery. The structural complexity of plant tissues, difficulties in isolating intact cells and nuclei, limited availability of high-quality reference genomes, and incomplete molecular annotations present significant obstacles to generating reliable omics datasets. Furthermore, high dimensionality, data sparsity, batch effects, platform-specific variability, and the heterogeneous nature of multi-omics data complicate their integration and biological interpretation. AI-driven computational approaches introduce additional challenges related

to model interpretability, algorithmic bias, overfitting, computational scalability, and the generalizability of predictions across different plant species and experimental conditions. The absence of standardized experimental protocols, comprehensive reference databases, and independent biological validation further restricts the reproducibility and reliability of published findings. Addressing these limitations requires the development of robust quality-control procedures, transparent computational workflows, standardized metadata reporting, and rigorous statistical validation strategies. Particular emphasis should be placed on data accessibility, reproducible analyses, independent experimental verification, and adherence to the FAIR principles of data management, namely findability, accessibility, interoperability, and reusability. Establishing comprehensive reporting standards is therefore essential for ensuring the scientific credibility, comparability, and translational relevance of AI-assisted multi-omics investigations. This section critically examines the principal experimental and computational challenges and discusses recommended reporting practices to enhance the reliability, reproducibility, transparency, and practical applicability of AI-driven single-cell and spatial multi-omics research in medicinal plant drug discovery.

Experimental bias and limited references. Plant cell walls, vacuoles, phenolics and polysaccharides create isolation and extraction biases. Sparse genomes, incomplete orthology assignments and limited standards impede annotation. Biological replicates should be collected across plants and batches, not treated as independent merely because many cells were measured [21, 18].

Statistical leakage and reproducibility. Data leakage occurs when cells from the same biological replicate or highly similar genes appear in both training and evaluation sets. Independent-species and independent-laboratory tests are especially important for claims of transferability. Studies should release raw and processed matrices, quality-control thresholds, accession identifiers, model code, random seeds, software versions and evaluation splits.

Evidence ladder for pathway claims. We recommend distinguishing (i) computational association, (ii) spatial or cell-type co-localization, (iii) biochemical activity, (iv) in planta genetic causality and (v) reproducible production or pharmacological activity. Each step answers a different question. A predicted pathway is not a validated biosynthetic route, and a cytotoxic extract is not a clinically effective medicine [3, 18].

10. Future Perspectives

The next phase of AI-enabled medicinal-plant research should be judged by experimentally validated biosynthetic mechanisms, reproducible compound production, and credible pharmacological leads rather than the number of computational predictions. The following directions are promising, but their levels of technical maturity differ substantially [3, 18].

Plant-specific foundation models and transfer learning. Foundation models learn transferable representations from large collections of sequences, molecular profiles, or cellular measurements. Transformer architectures provide a general computational foundation [34], while single-cell models such as scGPT illustrate how generative pretraining can support cell-level prediction and data integration [8]. Their direct application to medicinal plants,

however, is constrained by incomplete reference genomes, uneven species representation, inconsistent tissue annotation, and scarce experimentally validated pathway labels.

A useful plant-specific model should represent gene families and paralogy, distinguish developmental stage from stress response, and accommodate both single-cell and single-nucleus assays. Orthology-aware transfer learning may be useful for prioritizing candidate enzymes in under-studied species, but transfer across divergent taxa should not be assumed. Benchmarking should hold out entire species, plant families, experimental laboratories, and biosynthetic pathways, rather than randomly partitioning cells from the same experiments. Uncertainty calibration and comparison with simpler baselines are essential before model predictions can guide costly experiments [21, 8].

Spatially informed digital twins of specialized metabolism. A digital twin is a computational representation updated with measurements from a corresponding biological system. For medicinal plants, an aspirational digital twin would connect cell-type abundance, enzyme localization, precursor transport, metabolite turnover, environmental regulation, and tissue development. Such a model should combine mechanistic reaction constraints with transcriptomic, proteomic, and metabolomic observations, while explicitly representing uncertainty in fluxes and transport rates.

Spatial expression alone is insufficient to estimate metabolic flux. For example, the tissue in which an alkaloid accumulates need not be the tissue in which it was synthesized. Stable-isotope labeling, enzyme kinetics, genetic perturbations, and time-resolved metabolite measurements are required to distinguish synthesis from transport and storage. The cell-partitioned metabolism demonstrated in *Catharanthus roseus* provides a biological motivation for such models [18], but fully validated medicinal-plant digital twins remain a future research objective.

Explainable and uncertainty-aware artificial intelligence. Explainability is important when computational models are used to select enzymes, transcription factors, or candidate therapeutic compounds. Feature-attribution methods, including SHAP, can help identify which measured variables influence a prediction [22]. Nevertheless, an attribution score is not evidence of causal regulation. Highly correlated gene expression, shared ancestry, and batch-specific technical effects can create plausible but misleading explanations.

We recommend reporting calibration, predictive intervals where appropriate, out-of-distribution behavior, ablation experiments, and external validation. Candidate pathway edges should be ranked by convergent evidence: co-expression, subcellular localization, spatial co-occurrence, compatible biochemical function, and genetic perturbation. When evidence conflicts, models should preserve that conflict instead of forcing a single pathway assignment. This approach makes AI useful for hypothesis prioritization without conflating prediction with mechanistic proof [18, 3].

Graph learning and multimodal molecular representations. Biosynthetic pathways are naturally represented as graphs connecting genes, enzymes, metabolites, reactions, cellular compartments, and tissues. Graph neural networks and message-passing models offer a way to combine local network structure with sequence-derived and chemical features [17, 11]. In principle, these models could prioritize missing pathway reactions or transporters when some components have already been characterized.

The key obstacle is incomplete and biased ground-truth networks. A missing reaction in a database is not necessarily a negative example, and graph connectivity can reflect literature popularity rather than biology. Models should be evaluated on prospectively tested candidates and against enzyme-family annotation, homology searches, and established pathway-reconstruction approaches. Cross-modal integration should also retain the spatial resolution and detection limits of each experimental modality.

AI-guided synthetic biology and sustainable production. Once a pathway is validated, metabolic engineering may increase product yield through enzyme optimization, precursor supply, pathway balancing, transport engineering, or expression in an alternative host. The engineered production of artemisinic acid and subsequent semisynthetic artemisinin manufacture illustrate the potential of pathway-based production [27, 25]. However, success with one pathway does not establish feasibility for complex plant metabolites that require multiple compartments, specialized transport, or poorly characterized tailoring reactions.

AI could help prioritize enzyme variants and combinations of pathway modifications, but predictions must be tested against measured titer, rate, yield, product purity, and process stability. Sustainability assessment should include biomass sourcing, energy and solvent consumption, biodiversity conservation, and the ecological consequences of cultivation or harvesting. A promising compound is not automatically an economically or environmentally viable product.

Active learning and autonomous laboratories. Active learning can select experiments expected to reduce uncertainty most efficiently. In a medicinal-plant setting, a closed-loop workflow might propose a small set of enzyme assays, gene knockdowns, or isotope-labeling experiments; incorporate the resulting measurements; and update the next experimental priorities. This could be particularly valuable when authentic standards and plant material are scarce.

A credible implementation requires predefined stopping rules, biological replication, quality controls, negative results, and traceable laboratory metadata. Automation should support expert interpretation, not replace botanical authentication, analytical chemistry, or toxicological assessment. Autonomous discovery remains a research direction rather than an established routine pipeline for medicinal plants.

A staged roadmap from discovery to translation. We propose a five-stage framework with explicit evidence gates. **Stage 1: Biological resources.** Authenticate plant material, document provenance and chemotype, generate adequate genomic references, and establish reproducible sampling protocols. **Stage 2: Cell-resolved molecular mapping.** Obtain matched single-cell or single-nucleus profiles, spatial expression maps, and targeted or untargeted chemical measurements across biological replicates. **Stage 3: Computational integration.** Build uncertainty-aware models that nominate candidate enzymes, regulators, transporters, and metabolites, with rigorous external evaluation. **Stage 4: Mechanistic validation.** Confirm enzymatic activity, molecular identity, localization, and in planta causality using orthogonal experiments. **Stage 5: Translational evaluation.** Assess pharmacological mechanism, selectivity, safety, pharmacokinetics, production feasibility, and sustainability [3, 18, 25].

At each gate, evidence should be evaluated independently. For example, spatial co-localization is not proof of an enzyme's catalytic role; a reconstructed pathway is not proof of scalable yield; and an in vitro bioactivity signal is not evidence of clinical efficacy. Future community benchmarks should publish both successful and unsuccessful pathway predictions to reduce reporting bias.

Research priorities for the coming decade. Three priorities are particularly consequential. First, the community needs reference-quality, openly documented cell atlases for taxonomically diverse medicinal plants, with tissue-level chemical validation. Second, shared benchmarks should measure cross-species generalization and reproducibility under realistic data scarcity. Third, prospective studies should follow AI-nominated hypotheses through biochemical and genetic tests to compound-level and, where appropriate, preclinical evaluation. These steps would transform multi-omics integration from a descriptive exercise into a testable and cumulative framework for natural-product discovery.

11. Comparative Methodological Framework

A comparative methodological framework is essential for systematically evaluating the capabilities, limitations, and complementary roles of different experimental and computational approaches in AI-driven medicinal plant drug discovery. The rapid development of single-cell sequencing, spatial transcriptomics, proteomics, metabolomics, and artificial intelligence has introduced a wide range of technologies that differ considerably in their resolution, sensitivity, data requirements, computational complexity, and biological applications. Selecting an appropriate methodology therefore requires careful consideration of the underlying research objectives, the nature of the biological samples, and the specific molecular processes under investigation. In medicinal plant research, comparative evaluation of these technologies is particularly important for identifying specialized metabolite-producing cell populations, elucidating biosynthetic pathways, and characterizing spatially regulated metabolic activities. Furthermore, integrating complementary omics platforms with machine learning and deep learning approaches can enhance the reliability of molecular predictions and facilitate the reconstruction of complex biological networks. However, differences in experimental design, data quality, analytical assumptions, and validation strategies may substantially influence the interpretation and reproducibility of research findings. A systematic comparison of methodological strengths, limitations, and evidence requirements is therefore necessary to establish robust and scientifically reliable analytical workflows. This section presents a comparative methodological framework emphasizing technology selection, evidence-based biosynthetic pathway identification, multimodal data integration, uncertainty assessment, reproducibility, and the translation of computational predictions into experimentally validated drug discovery outcomes.

Selecting technologies according to the biological question. No single omics platform can establish a complete specialized-metabolite pathway. Genome and bulk-transcriptome resources identify the candidate gene repertoire, whereas single-cell or single-nucleus RNA sequencing resolves transcriptional heterogeneity and nominates cell populations likely to contribute to synthesis. Spatial transcriptomics then tests whether these assignments are consistent with tissue anatomy; metabolite imaging or laser-microdissection-assisted chemical profiling provides complementary evidence on chemical distribution. These measurements

are not interchangeable: RNA abundance is neither a direct measure of enzyme activity nor a reliable proxy for local metabolite concentration [21, 30, 18]. The appropriate platform therefore depends on whether the primary question concerns candidate discovery, regulation, transport, storage or pharmacological activity.

TABLE 1. Strengths, limitations and validation needs of complementary technologies.

Method	Primary contribution	Principal caveat / validation
scRNA-seq	Cell-state and candidate pathway-gene expression	Dissociation bias; verify with in situ localization
scATAC-seq	Candidate regulatory elements and motifs	Accessibility does not prove TF binding; perturbation required
Spatial RNA	Tissue-contextualized transcription	Mixed spots and section artifacts; microscopy controls
LC–MS/MS	Chemical profiles and putative structures	Isomers and adducts; standards and NMR as needed
MS imaging	Ion localization in tissue	Ion suppression and ambiguous identity; orthogonal MS/MS
AI integration	Cross-modal hypotheses and prioritization	Leakage and domain shift; external validation

An evidence hierarchy for biosynthetic pathway assignment. We recommend distinguishing five levels of evidence. *Level I* comprises genome annotation, sequence similarity and predicted protein structure. *Level II* adds differential or cell-type-enriched transcription, coexpression and chromatin accessibility. *Level III* requires spatial agreement between candidate transcripts and relevant metabolite patterns, with appropriate resolution and identity controls. *Level IV* includes biochemical conversion of a chemically authenticated substrate to a characterized product, supported where possible by kinetic analysis. *Level V* requires genetic perturbation or complementation demonstrating the proposed role in planta, ideally together with isotope-tracing evidence and rescue experiments. These levels are an analytical proposal rather than an existing universal standard. The *Catharanthus roseus* investigations illustrate why integrated cellular and chemical observations are particularly informative for pathways distributed across cell types [18]. Importantly, even Level V pathway evidence does not establish drug efficacy or safety [3].

Multimodal alignment, uncertainty and reproducibility. Cross-platform integration should begin with sample provenance, developmental stage, growth conditions, tissue position and biological replicate identity. Cells from distinct preparations cannot automatically be treated as paired observations. Computational alignment may use shared genes, latent factors or spatial neighborhood information, but integration can also erase real rare-cell differences or introduce spurious correspondences. Accordingly, every analysis should report raw-data quality controls, filtering thresholds, batch correction strategy, model hyperparameters, uncertainty estimates and sensitivity to alternative integration methods. Evaluation should hold out biological replicates and, when transferability is claimed, species or chemotypes rather than randomly dividing cells from the same plant between training and test sets

[21, 20]. Spatial resolution must be reported at the actual measurement scale; inferred single-cell assignments from mixed spots should be described as computational estimates, not directly observed measurements [30, 26].

From computational nomination to drug-discovery decisions. A practical prioritization scheme can rank candidates using convergent, independently measured evidence: biosynthetic plausibility, expression specificity, spatial concordance, metabolite identity confidence, feasibility of enzyme assays and genetic tractability. Ranking scores should be accompanied by their uncertainty and should not be interpreted as calibrated probabilities unless this has been demonstrated. After pathway validation, the translational pipeline requires purified compound identity, reproducible supply, orthogonal target-engagement assays, selectivity, pharmacokinetic and toxicological characterization, and eventually clinical evidence. Metabolic engineering can address supply limitations, but productivity, ecological sustainability and genetic stability require independent assessment [24, 3, 25].

Recommended reporting checklist. For future studies we propose reporting: (i) taxonomic authentication and voucher information; (ii) plant genotype, tissue, developmental stage and cultivation conditions; (iii) biological and technical replication; (iv) cell or nucleus isolation and potential dissociation biases; (v) sequencing depth, mapping and filtering criteria; (vi) spatial resolution and registration uncertainty; (vii) metabolite identification confidence and reference-standard use; (viii) model architecture, baselines, data splits and external validation; (ix) orthogonal biochemical and genetic confirmation; and (x) accession numbers, executable analysis code and declared limitations. Such reporting would make cross-study comparisons more reliable and clarify whether an apparent computational advance produces a genuinely new and experimentally testable biological insight.

12. Limitations of This Review

This article is a critical narrative synthesis and does not claim exhaustive database coverage or quantitative meta-analysis. The underlying literature is uneven: high-resolution cell atlases and multimodal studies are concentrated in a limited number of well-characterized plants, while some methodological evidence is extrapolated from model organisms or mammalian datasets. Consequently, cross-species applicability, spatial co-registration, metabolite annotation and predictive performance must be evaluated independently in each target medicinal plant. The proposed workflows and evidence hierarchy are recommendations rather than experimentally validated universal standards. Any subsequent systematic review should report databases, exact search strings, dates, eligibility criteria and screening decisions prospectively.

13. Priority Research Questions

(1) Which cell types synthesize, transport and store pathway intermediates, and how can these roles be separated experimentally?
(2) Which cross-species model features remain reliable under incomplete genomes and sparse training data?
(3) How can spatial transcript and metabolite measurements be registered without overstating resolution?

(4) What minimum biochemical and genetic evidence should be required before assigning an enzyme to a pathway?
(5) Can cell-resolved insights improve reproducible yields while preserving biodiversity and equitable sourcing?

Declarations

Funding: The authors report no external funding for this review.
Conflicts of interest: The author declares no competing interests.
Data availability: No new datasets were generated or analyzed in this narrative review.
Review methodology: Critical narrative synthesis; not a systematic review.
Ethics approval: Not applicable.

References

[1] Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21 (2020), 111.
[2] Ashuach, T. et al. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20 (2023), 1222–1231.
[3] Atanasov, A. G. et al. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery*, 20 (2021), 200–216.
[4] Buenrostro, J. D. et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10 (2013), 1213–1218.
[5] Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523 (2015), 486–490.
[6] Budnik, B. et al. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology*, 19 (2018), 161.
[7] Croteau, R. et al. Taxol biosynthesis and molecular genetics. *Phytochemistry Reviews*, 5 (2006), 75–97.
[8] Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21 (2024), 1470–1480.
[9] Denyer, T. et al. Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Developmental Cell*, 48 (2019), 840–852.e5.
[10] Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16 (2019), 299–302.
[11] Gilmer, J. et al. Neural message passing for quantum chemistry. *International Conference on Machine Learning*, (2017), 1263–1272.
[12] Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53 (2021), 403–411.
[13] Hafemeister, C. and Satija, R.. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20 (2019), 296.
[14] Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell*, 184 (2021), 3573–3587.e29.
[15] Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (2021), 583–589.
[16] Katsuyama, Y. et al. Curcuminoid biosynthesis by two type III polyketide synthases in the herb Curcuma longa. *Journal of Biological Chemistry*, 284 (2009), 11160–11170.
[17] Kipf, T. N. and Welling, M.. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, (2017).

[18] Li, C. et al. Single-cell multi-omics in the medicinal plant Catharanthus roseus. *Nature Chemical Biology*, 19 (2023), 1031–1041.
[19] Li, C. et al. Cell-type-aware regulatory landscapes governing monoterpene indole alkaloid biosynthesis in the medicinal plant Catharanthus roseus. *New Phytologist*, 245 (2025), 347–362.
[20] Lopez, R. et al. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15 (2018), 1053–1058.
[21] Luecken, M. D. and Theis, F. J.. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15 (2019), e8746.
[22] Lundberg, S. M. and Lee, S.-I.. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30 (2017).
[23] Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161 (2015), 1202–1214.
[24] Newman, D. J. and Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, 83 (2020), 770–803.
[25] Paddon, C. J. et al. High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, 496 (2013), 528–532.
[26] Rodriques, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363 (2019), 1463–1467.
[27] Ro, D.-K. et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440 (2006), 940–943.
[28] Ryu, K. H. et al. Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiology*, 179 (2019), 1444–1456.
[29] Slavov, N.. Single-cell protein analysis by mass spectrometry. *Current Opinion in Chemical Biology*, 60 (2021), 1–9.
[30] Stähl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353 (2016), 78–82.
[31] Stuart, T. et al. Comprehensive integration of single-cell data. *Cell*, 177 (2019), 1888–1902.e21.
[32] Stuart, T. et al. Single-cell chromatin state analysis with Signac. *Nature Methods*, 18 (2021), 1333–1341.
[33] Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6 (2009), 377–382.
[34] Vaswani, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30 (2017).
[35] Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34 (2016), 828–837.

©RK Scientific Publishers, India