Biotechnology Frontiers
An International Journal
OPEN ACCESS
ISSN 3117-5791
Review Article
Check for updates

Artificial Intelligence-Driven Discovery of Medicinal Plant Bio-actives Integrating Ethnobotanical Databases, Metabolomics, and Molecular Docking

Talari Parameshwari*[1,2]

[1]*Department of Botany, Telangana Tribal Welfare Residential Degree College for Women (TGTWRDC-W), Shadnagar, Rangareddy, Telangana - 509216, India*
[2]*Department of Botany, UCS, Saifabad, Osmania University, Hyderabad, Telangana - 500007, India*

Corresponding author: **Talari Parameshwari** | E-mail: **parameshwaritalari@gmail.com**

Citation: Talari Parameshwari (2024). Artificial Intelligence-Driven Discovery of Medicinal Plant Bio-actives Integrating Ethnobotanical Databases, Metabolomics, and Molecular Docking. *Biotechnology Frontiers: An International Journal.*
DOI: https://doi.org/10.51470/BF.2024.4.2.03

20 May 2024: Received | 13 June 2024: Revised | 10 July 2024: Accepted | 02 August 2024: Available Online

Abstract

Medicinal plants represent an important reservoir of structurally diverse bioactive compounds with potential applications in drug discovery, disease prevention, and therapeutic development. However, the identification and characterization of pharmacologically relevant phytochemicals remain challenging because of the chemical complexity of plant extracts, variations in phytochemical composition, incomplete documentation of traditional medicinal knowledge, and limitations associated with conventional screening approaches. Recent advances in artificial intelligence (AI), machine learning (ML), metabolomics, and molecular modelling have created opportunities to accelerate the systematic discovery of plant-derived bioactive compounds. This examines an integrated computational and experimental framework for medicinal plant bioactive discovery, emphasizing the convergence of ethnobotanical databases, metabolomics profiling, artificial intelligence, and molecular docking. Ethnobotanical information provides a knowledge-based foundation for prioritizing medicinal species, plant parts, traditional applications, and therapeutic indications. Metabolomics technologies, particularly liquid chromatography–mass spectrometry (LC–MS), gas chromatography–mass spectrometry (GC–MS), and nuclear magnetic resonance (NMR) spectroscopy, facilitate the comprehensive identification and characterization of plant metabolites. Artificial intelligence approaches, including supervised learning, unsupervised clustering, natural language processing, deep learning, and network-based prediction, can integrate heterogeneous datasets and identify relationships between phytochemical structures, biological targets, and pharmacological activities. Molecular docking and molecular dynamics simulations further support the assessment of predicted ligand–target interactions and their structural stability. The integration of these technologies offers a promising strategy for prioritizing candidate bioactive compounds, identifying potential molecular targets, and developing testable hypotheses for experimental investigation and challenges associated with data standardization, ethnobotanical knowledge protection, metabolite annotation, model interpretability, computational reproducibility, and experimental validation remain important considerations. An integrated and ethically responsible AI-assisted discovery framework could strengthen evidence-based medicinal plant research and contribute to the development of natural-product-inspired therapeutic candidates.

Keywords: *Artificial intelligence; medicinal plants; ethnobotany; machine learning; metabolomics; phytochemicals; molecular docking; natural products; drug discovery; bioactive compounds.*

1. Introduction

Medicinal plants have played a substantial role in traditional healthcare systems and continue to contribute to the identification of pharmacologically important natural products. Across different geographical regions, indigenous and traditional communities have developed extensive knowledge concerning the therapeutic applications of roots, leaves, flowers, fruits, seeds, bark, and other plant-derived materials.

This knowledge represents an important starting point for scientific investigations into plant-associated secondary metabolites and their potential biological activities. Numerous medicinally relevant compounds belong to diverse chemical classes, including alkaloids, flavonoids, terpenoids, phenolic acids, glycosides, tannins, and other specialized metabolites. Their structural diversity provides opportunities for investigating multiple biological pathways and identifying new molecular scaffolds for pharmaceutical development [1], the considerable

© ***Authors:*** *Published in* ***Biotechnology Frontiers: An International Journal*** *under the CC BY-NC-ND 4.0 license (http://creativecommons.org/licenses/by-nc-nd/4.0/). No commercial use or modifications permitted.*

chemical and pharmacological diversity of medicinal plants, the systematic identification of active constituents remains a complex process. Traditional approaches generally involve plant selection, extraction, preliminary phytochemical screening, chromatographic separation, compound characterization, and biological evaluation. Although these methods have generated important discoveries, they may require substantial time, laboratory resources, and repeated experimental procedures. Furthermore, the biological activity of a crude extract is not necessarily attributable to a single compound. Different metabolites may exhibit additive, synergistic, antagonistic, or metabolism-dependent effects, making the identification of biologically relevant constituents particularly challenging [2]. Variations in geographical origin, environmental conditions, plant developmental stages, extraction procedures, and post-harvest handling further influence the chemical composition of medicinal plant materials.

The development of high-throughput analytical technologies has substantially improved the characterization of plant metabolites. LC–MS, GC–MS, and NMR-based metabolomics can generate extensive chemical datasets that describe the metabolic composition of plant samples. These technologies facilitate the identification of known metabolites, the detection of previously uncharacterized chemical features, and the comparison of metabolic profiles across species, tissues, developmental stages, and environmental conditions. However, metabolomics experiments frequently generate large and complex datasets that require advanced computational approaches for data processing, feature selection, statistical interpretation, and biological annotation [3]. Artificial intelligence has emerged as a valuable computational approach for addressing these challenges. Machine learning algorithms can recognize patterns within large datasets, identify relationships between chemical structures and biological activities, classify plant metabolites, and predict potential protein targets. Deep learning approaches can further analyse complex molecular representations, spectral information, and heterogeneous biological datasets. Natural language processing offers additional opportunities to extract medicinal plant information from ethnobotanical literature, historical records, scientific publications, and biodiversity databases [4]. Consequently, AI-assisted approaches can connect traditional medicinal knowledge with contemporary chemical and molecular investigations.

Molecular docking represents another important component of this integrated discovery process. It provides a computational framework for estimating the preferred orientation and potential interactions of small molecules within the binding sites of biological macromolecules. When combined with predicted bioactivities, metabolomics-derived chemical identities, and disease-associated molecular targets, docking can help prioritize candidate compounds for further investigation. Nevertheless, docking scores and predicted binding poses should not be interpreted as direct evidence of biological efficacy or therapeutic potential. Additional computational and experimental validation is necessary to establish the reliability of these predictions [5]. The integration of ethnobotanical knowledge, metabolomics, artificial intelligence, and molecular docking therefore offers an emerging multidisciplinary framework for medicinal plant research. Rather than treating these technologies as independent analytical methods, their coordinated application can facilitate the prioritization of medicinal species, identification of candidate phytochemicals, prediction of molecular mechanisms, and selection of compounds for experimental evaluation. The scientific foundations, computational methodologies, potential applications, limitations, and future perspectives of this integrated approach, with particular emphasis on its contribution to natural-product-based drug discovery.

2. Ethnobotanical Databases as a Foundation for Medicinal Plant Bioactive Discovery

Ethnobotanical research documents the relationships between human communities and plant resources, including traditional medicinal applications, preparation methods, plant parts utilized, administration practices, and cultural significance. Such information provides a valuable knowledge base for identifying medicinal species that warrant further pharmacological investigation. Traditional plant-use records can help researchers prioritize species based on historical therapeutic applications, frequency of reported use, geographical distribution, and the consistency of traditional knowledge across communities [6]. The scientific value of ethnobotanical information extends beyond the identification of medicinal plant species. Documentation of the plant part used, preparation technique, route of administration, and traditional indication can provide important contextual information for experimental design. For example, traditional preparations involving aqueous decoctions may have substantially different chemical profiles from organic-solvent extracts. Similarly, the medicinal properties attributed to leaves may not be chemically or pharmacologically equivalent to those of roots, seeds, or bark. Integrating these observations into bioactive discovery workflows can improve the biological relevance of sample selection and extraction procedures.

Several digital resources support the documentation and retrieval of ethnobotanical and botanical information. The World Flora Online, Plants of the World Online, and the Global Biodiversity Information Facility provide taxonomic, nomenclatural, and biodiversity information that can assist in verifying plant identities and distributions. The Kew Medicinal Plant Names Services supports the standardization of medicinal plant nomenclature, while specialized ethnobotanical repositories and published literature provide information on traditional uses.

The NAPRALERT database is an important resource for natural-product and ethnomedical literature, subject to its access conditions. These resources differ in scope and data structure, and no single database should be assumed to contain comprehensive information on all traditional medicinal practices [7]. An important challenge in integrating ethnobotanical information is the inconsistency of plant nomenclature. The same species may be recorded under different scientific synonyms, local names, vernacular terms, or historical taxonomic classifications. Common names may also refer to multiple unrelated species across different geographical regions. Incorrect taxonomic matching can introduce substantial errors into subsequent chemical and pharmacological analyses. Therefore, accepted scientific names, botanical author citations, taxonomic identifiers, and voucher specimen information should be incorporated into digital datasets wherever possible.

Artificial intelligence, particularly natural language processing, can improve the extraction of ethnobotanical information from unstructured scientific literature. Text-mining systems can identify scientific plant names, traditional therapeutic indications, reported plant parts, preparation methods, and associated disease terminology. Named entity recognition and relation extraction can be used to construct structured datasets connecting plant species with their reported traditional applications.

For example, a computational system may identify statements describing the use of a particular leaf preparation for inflammatory conditions and convert the information into a structured plant–part–preparation–indication relationship [8]. Knowledge graphs provide another approach for organizing ethnobotanical information. In such systems, medicinal plants, phytochemicals, traditional applications, biological targets, and disease pathways can be represented as interconnected entities. Relationships between these entities may facilitate the identification of medicinal species that share chemical constituents or are traditionally used for similar conditions. Graph-based machine learning can subsequently identify potentially important connections that may not be apparent through conventional literature searches [9]. However, the digitization of traditional medicinal knowledge requires careful consideration of intellectual property, community consent, access restrictions, and benefit-sharing. Traditional knowledge is not simply an unrestricted scientific dataset. Some medicinal practices are community-specific, confidential, or governed by customary systems of knowledge ownership. Researchers should therefore follow applicable national and international requirements, including the principles of access and benefit-sharing under the Convention on Biological Diversity and the Nagoya Protocol. AI-assisted ethnobotanical research should incorporate transparent data provenance and appropriate recognition of knowledge-holding communities.

Table 1: Major databases and computational resources supporting AI-driven medicinal plant bioactive discovery

Database/resource	Primary information	Application in medicinal plant research
IMPPAT	Indian medicinal plants, phytochemicals and therapeutic associations	Identification of medicinal plants and their reported chemical constituents
LOTUS	Natural products, chemical structures and taxonomic associations	Linking plant species with reported natural products
GNPS	Tandem mass spectrometry data and molecular networking	Metabolite annotation, spectral matching and identification of related compounds
MassBank	Experimental mass spectrometry reference spectra	Identification and confirmation of plant metabolites
PubChem	Chemical structures, molecular properties and bioassay information	Retrieval of phytochemical structures and molecular descriptors
NPASS	Natural products and associated biological activities	Screening and prioritization of natural compounds for pharmacological investigation
KNApSAcK	Organism–metabolite relationships	Linking plant species with their reported metabolites
GBIF	Global biodiversity and species occurrence records	Verification of geographical distribution and biodiversity information
Plants of the World Online	Plant taxonomy, accepted names and distribution	Botanical authentication and taxonomic standardization
Protein Data Bank	Three-dimensional macromolecular structures	Retrieval of protein structures for molecular docking
ChEMBL	Bioactive molecules, drug-like compounds and experimental bioassays	Training and validation of machine learning models for bioactivity prediction

Source: *Compiled from the respective database platforms. Database coverage, access conditions and available records may vary.*

3. Metabolomics-Based Characterization of Medicinal Plant Bioactives

Metabolomics involves the comprehensive analysis of small molecules present within biological systems and provides an important analytical foundation for medicinal plant research. Unlike targeted phytochemical analysis, which focuses on predetermined compounds, untargeted metabolomics aims to characterize a broad range of metabolites within a sample. This approach is particularly useful for medicinal plants because their biological properties frequently arise from complex mixtures of primary and secondary metabolites.

3.1 Major analytical platforms

Liquid chromatography–mass spectrometry is widely employed for the separation and characterization of plant metabolites with different physicochemical properties. Reversed-phase LC–MS is particularly useful for analysing moderately polar and nonpolar compounds, including many flavonoids, phenolic compounds, alkaloids, and terpenoids. Hydrophilic interaction liquid chromatography can complement reversed-phase methods by improving the detection of highly polar metabolites. High-resolution mass spectrometry provides accurate mass measurements and isotopic information that support molecular formula prediction and candidate annotation [10].

Gas chromatography–mass spectrometry is particularly suitable for volatile and thermally stable compounds, as well as metabolites that can be converted into volatile derivatives through chemical derivatization. It is commonly applied to the characterization of essential oils, fatty acids, organic acids, sugars, and other low-molecular-weight metabolites. GC–MS provides reproducible fragmentation patterns that can be compared with established spectral libraries to support compound identification. Nuclear magnetic resonance spectroscopy offers complementary structural information and is particularly valuable for confirming molecular structures. Although NMR may have lower sensitivity than mass spectrometry for certain low-abundance metabolites, it provides highly reproducible spectral measurements and can support quantitative analysis without extensive compound-specific calibration. Combining NMR with LC–MS and GC–MS can provide a more comprehensive understanding of medicinal plant metabolomes.

3.2 Data processing and metabolite annotation

Metabolomics datasets require multiple processing steps before biological interpretation. These generally include baseline correction, peak detection, chromatographic alignment, feature extraction, normalization, missing-value assessment, and signal correction. In LC–MS experiments, a metabolomic feature is commonly represented by a mass-to-charge ratio and retention time, while tandem mass spectrometry provides additional fragmentation information [10]. Metabolite annotation remains one of the major challenges in untargeted metabolomics. A mass feature may correspond to different structural isomers, adducts, isotopes, or in-source fragments. Therefore, accurate mass alone is generally insufficient for definitive structural identification. Spectral matching, retention-time information, fragmentation interpretation, reference standards, and orthogonal analytical techniques should be used to improve confidence. Public resources such as MassBank, GNPS, METLIN, and the Human Metabolome Database can support spectral comparison and metabolite annotation, although their coverage and suitability differ according to the sample type and analytical platform. The Metabolomics Standards Initiative provides important recommendations for reporting metabolomics experiments and describing metabolite identification confidence.

3.3 AI-assisted metabolomics interpretation

Artificial intelligence can improve the interpretation of complex metabolomics datasets by identifying patterns, reducing dimensionality, and prioritizing features associated with particular biological conditions. Principal component analysis and hierarchical clustering are widely used exploratory techniques, while supervised machine learning approaches can distinguish samples according to species, plant parts, environmental conditions, or biological activity [11].

Random forest, support vector machines, gradient boosting, and artificial neural networks can be used to identify metabolite features that contribute to sample classification or predicted biological responses. Deep learning models may also assist with the interpretation of tandem mass spectra and the prediction of molecular structures from spectral information. The integration of metabolomic profiles with biological assay results is particularly important. For example, a study may compare the chemical profiles of several plant extracts with their measured antioxidant, antimicrobial, or enzyme-inhibitory activities. Machine learning can then identify metabolite features associated with variation in biological responses. These associations can guide targeted isolation and subsequent experimental investigation [12]. However, metabolomics-based AI models require appropriate statistical design. Small sample sizes, batch effects, confounding variables, and data leakage can generate misleading predictions. Feature selection and model validation should be performed using appropriate training, validation, and independent test datasets. Cross-validation must also respect biological replication and avoid treating technical replicates as independent biological samples.

4. Artificial Intelligence and Machine Learning in Medicinal Plant Research

Artificial intelligence encompasses computational approaches designed to perform tasks that typically require aspects of human intelligence, including pattern recognition, prediction, classification, and information extraction. Within medicinal plant research, AI can support the analysis of botanical, chemical, genomic, proteomic, pharmacological, and ethnomedical datasets. Machine learning provides algorithms that identify relationships within data, while deep learning employs multilayer neural networks to model complex patterns.

4.1 Supervised learning for bioactivity prediction

Supervised learning algorithms are trained using datasets containing input variables and corresponding known outcomes. In medicinal plant research, molecular descriptors, chemical fingerprints, spectral features, and biological assay measurements can be used to train models for predicting potential bioactivity [13]. Quantitative structure–activity relationship models are particularly relevant because they associate molecular characteristics with biological responses. Molecular descriptors may include molecular weight, lipophilicity, hydrogen-bond donors and acceptors, topological polar surface area, and structural fingerprints. These characteristics can be used to predict potential pharmacological properties or prioritize compounds for experimental screening. Random forest algorithms are useful for identifying nonlinear relationships and ranking important molecular features. Support vector machines can classify compounds according to their predicted activity, while gradient-boosting approaches can improve predictive performance in appropriately designed datasets.

Nevertheless, predictions remain dependent on the quality, chemical diversity, and biological relevance of the training data.

4.2 Unsupervised learning and metabolite clustering

Unsupervised learning identifies patterns in datasets without requiring predefined outcome labels. Principal component analysis, hierarchical clustering, and other dimensionality-reduction methods can reveal similarities and differences between medicinal plant samples [14]. These methods can be used to compare metabolomic profiles across plant species, geographical locations, tissues, developmental stages, and extraction conditions. Clustering may also reveal groups of metabolites that consistently occur together, suggesting shared biosynthetic pathways or common responses to environmental conditions. More advanced approaches, including t-distributed stochastic neighbor embedding and uniform manifold approximation and projection, can assist in visualizing high-dimensional datasets, these visualization techniques should not be interpreted as direct evidence of biological relationships without appropriate statistical evaluation.

4.3 Deep learning and molecular representation

Deep learning methods are increasingly relevant to molecular informatics because they can learn complex representations from molecular structures, chemical fingerprints, and spectral data. Graph neural networks represent molecules as graphs in which atoms correspond to nodes and chemical bonds correspond to edges. These representations can capture structural relationships that may be less directly represented by conventional molecular descriptors [15]. Convolutional neural networks and transformer-based architectures can also be applied to chemical and spectral data. Their potential applications include metabolite classification, molecular property prediction, compound prioritization, and the identification of chemical patterns associated with particular biological activities.

4.4 Natural language processing and knowledge extraction

Natural language processing offers an important bridge between ethnobotanical literature and computational drug discovery. Scientific publications, pharmacopoeias, ethnomedicinal surveys, and historical botanical records contain substantial information that may not be available in structured databases [16]. Named entity recognition can identify medicinal plant species, phytochemicals, diseases, and biological targets. Relation extraction can determine whether a particular compound was reported in a specific plant or whether a particular species was traditionally used for a defined therapeutic purpose. Large language models may assist with document screening, terminology harmonization, and preliminary knowledge extraction, but their outputs require verification against original sources.

Table 2: Artificial intelligence and computational approaches for medicinal plant bioactive discovery

Computational approach	Representative algorithms/tools	Major application	Expected research outcome
Natural language processing	Named entity recognition, BERT, BioBERT	Extraction of medicinal plant names, traditional uses and phytochemical information from scientific literature	Structured ethnobotanical knowledge
Supervised machine learning	Random forest, support vector machine, XGBoost	Prediction of biological activities using molecular descriptors and chemical fingerprints	Prioritization of potentially bioactive phytochemicals
Unsupervised learning	PCA, hierarchical clustering, k-means	Classification of metabolomic profiles and identification of chemical patterns	Identification of metabolite clusters and chemotypic variation
Deep learning	Convolutional neural networks, recurrent neural networks, transformers	Analysis of complex spectral, chemical and biological datasets	Improved molecular and spectral pattern recognition
Graph neural networks	Graph convolutional networks, message-passing neural networks	Molecular representation and ligand–target interaction prediction	Identification of promising molecular structures and potential targets
Molecular networking	GNPS molecular networking, feature-based molecular networking	Visualization of structurally related metabolites using MS/MS fragmentation patterns	Identification and dereplication of related natural products
Quantitative structure–activity relationship	2D/3D-QSAR, molecular fingerprints	Prediction of relationships between molecular structures and biological activities	Bioactivity prediction and chemical prioritization
Molecular docking	AutoDock Vina, AutoDock4, Glide	Prediction of potential phytochemical–protein interactions	Candidate ligand–target interaction models
Molecular dynamics simulation	GROMACS, AMBER, NAMD	Evaluation of protein–ligand complex stability and conformational changes	Dynamic structural assessment
Network pharmacology	Cytoscape, STRING, target–pathway networks	Integration of phytochemicals, molecular targets and biological pathways	Hypothesis generation for multi-target mechanisms
Explainable AI	SHAP, LIME, feature attribution methods	Interpretation of machine learning predictions and influential molecular features	Improved model transparency and interpretability

Abbreviations: *AI, artificial intelligence; PCA, principal component analysis; QSAR, quantitative structure–activity relationship; MS/MS, tandem mass spectrometry; SHAP, Shapley additive explanations; LIME, local interpretable model-agnostic explanations.*

5. Integration of Ethnobotanical and Metabolomic Databases

The integration of ethnobotanical and metabolomic datasets provides an opportunity to move from traditional plant-use documentation toward chemically informed bioactive discovery. Ethnobotanical data can guide the selection of species, plant parts, and biological activities, whereas metabolomics can provide detailed information about the chemical constituents of selected materials. A practical integration framework begins with botanical authentication and taxonomic standardization. Each plant specimen should be associated with an accepted scientific name, collection location, collection date, plant part, and voucher specimen identifier.

Ethnobotanical information can then be connected to the corresponding taxonomic record using standardized identifiers.

Metabolomics data should subsequently be linked to the authenticated plant material. Each metabolite record should include the analytical platform, mass-to-charge ratio, retention time, fragmentation information, annotation confidence, and reference database where applicable. When available, experimentally confirmed chemical structures should be distinguished from tentative annotations [17]. An integrated database can then connect several information categories, including plant species, geographical origin, traditional uses, plant parts, extraction methods, chemical constituents, biological assay results, molecular targets, and computational predictions. Knowledge graphs can provide a useful architecture for representing these relationships.

For example, a medicinal plant traditionally used for inflammatory conditions may be prioritized for metabolomic analysis. LC–MS profiling may identify several candidate flavonoids and phenolic compounds. Bioactivity data from related compounds may then be used to train or apply machine learning models to prioritize metabolites associated with inflammatory pathways. Molecular docking can subsequently evaluate the predicted interactions of selected compounds with relevant protein targets.

6. Molecular Docking in the Identification of Potential Plant-Derived Bio-actives

Molecular docking is a computational method used to investigate the potential binding orientation of small molecules within the active or binding sites of macromolecular targets. In natural-product research, docking can help prioritize phytochemicals for further investigation by assessing their possible interactions with proteins associated with selected biological pathways.

6.1 Selection and preparation of molecular targets

The selection of a protein target should be based on a clearly defined biological hypothesis. Targets may be selected from established disease-associated pathways, experimental literature, or validated molecular mechanisms. Protein structures can be obtained from the Protein Data Bank, which provides experimentally determined macromolecular structures, including X-ray crystallographic, NMR, and cryo-electron microscopy structures. Protein preparation generally involves assessing structural integrity, identifying relevant binding sites, addressing missing residues or atoms where necessary, assigning appropriate protonation states, and preparing the structure for docking. The presence of cofactors, metal ions, water molecules, and other structural components should be considered according to the biological system and docking protocol.

6.2 Preparation of phytochemical ligands

Candidate phytochemicals identified through metabolomics can be converted into suitable molecular structures for docking. Chemical structures may be obtained from resources such as PubChem, provided that the database record corresponds to the experimentally supported metabolite identity [7]. Ligand preparation involves checking molecular connectivity, assigning stereochemistry, generating relevant protonation states, and performing energy minimization. Particular attention should be given to natural products containing multiple stereo-centres, glycosidic linkages, or flexible substituent groups because incorrect structural representation can substantially affect docking outcomes.

6.3 Docking algorithms and interaction analysis

Software packages such as AutoDock Vina are widely used for molecular docking. Docking algorithms explore possible ligand orientations and estimate binding affinity using scoring functions. Predicted binding poses can be analysed to identify hydrogen bonds, hydrophobic interactions, electrostatic contacts, aromatic interactions, and other noncovalent associations.

The predicted binding energy can provide a comparative measure within a consistently designed docking experiment. The docking scores are approximate computational estimates and should not be treated as experimentally measured binding affinities. Comparisons should ideally include known ligands, suitable negative controls, and validated docking protocols [8]. Redocking a crystallographically observed ligand into its original binding site is one approach for evaluating whether the docking procedure can reproduce an experimentally established binding orientation. Additional validation may include comparison with known active and inactive compounds or enrichment analysis using an appropriately curated dataset.

6.4 Molecular dynamics and binding stability

Molecular dynamics simulations can provide additional information about the structural stability of protein–ligand complexes over time. Unlike conventional docking, which generally produces static binding poses, molecular dynamics simulations investigate the movement of atoms under a specified force field and simulation environment [9]. The analysed parameters include root mean square deviation, root mean square fluctuation, radius of gyration, hydrogen-bond persistence, and solvent-accessible surface area. Binding free-energy estimation methods, including molecular mechanics Poisson–Boltzmann surface area and molecular mechanics generalized Born surface area, may also be applied with appropriate methodological consideration. Molecular dynamics results should be interpreted within the limitations of the simulation model, force field, solvent representation, and simulation duration. A stable simulated complex does not independently establish biological activity, target engagement, or therapeutic efficacy.

7. AI-Assisted Integration of Molecular Docking and Bioactivity Prediction

The integration of AI with molecular docking can improve the prioritization of candidate phytochemicals by combining chemical, biological,

and structural information. Instead of relying exclusively on docking scores, an integrated prediction framework may consider molecular descriptors, metabolite annotation confidence, predicted target associations, known bioactivity information, and structural interaction patterns. Machine learning models can be trained using experimentally characterized ligand–target datasets to predict potential interactions between phytochemicals and proteins. Chemical fingerprints, molecular graphs, and protein sequence or structural representations may serve as model inputs. Deep learning approaches can further identify complex relationships between ligand structures and target proteins [12]. Network pharmacology offers a complementary approach by connecting medicinal plants, phytochemicals, molecular targets, and disease-associated pathways. Plant may contain several compounds that potentially interact with multiple proteins, while a particular protein may be influenced by metabolites from different species. Such networks can help generate hypotheses about multi-component and multi-target effects. However, network connectivity and predicted target associations should not be interpreted as proof of pharmacological synergy. Demonstrating synergy requires appropriately designed experimental studies, including dose–response analyses and suitable combination-effect models. Similarly, computational target prediction should be followed by biochemical, cellular, or other relevant biological assays.

8. Potential Applications in Medicinal Plant-Based Drug Discovery

The integrated application of ethnobotanical databases, metabolomics, AI, and molecular docking has potential applications across several therapeutic research areas. These applications should be understood as research opportunities rather than established clinical outcomes.

8.1 Antioxidant and anti-inflammatory bioactive discovery

Medicinal plants contain numerous phenolic acids, flavonoids, and other secondary metabolites that have been investigated for antioxidant and anti-inflammatory activities. Metabolomics can identify compounds associated with variation in antioxidant capacity, while machine learning can prioritize features associated with measured biological responses. Molecular docking may subsequently investigate their potential interactions with proteins involved in inflammatory signalling, including cyclooxygenases, lipoxygenases, and selected inflammatory regulatory proteins. Nevertheless, antioxidant activity measured through chemical assays does not necessarily translate into antioxidant effects in living systems. Cellular uptake, metabolism, bioavailability, and the biological context must be considered before drawing conclusions about therapeutic relevance.

8.2 Antimicrobial and antifungal compound discovery

The increasing concern regarding antimicrobial resistance has strengthened interest in natural products as sources of potential antimicrobial agents. AI-assisted metabolomic analysis can help prioritize plant-derived compounds associated with antimicrobial activity. Docking may be used to investigate potential interactions with microbial enzymes, membrane-associated proteins, or other validated molecular targets [13]. Experimental evaluation should include standardized antimicrobial susceptibility testing, appropriate reference compounds, and assessment of cytotoxicity in relevant host cells. Predicted interactions alone cannot establish antimicrobial efficacy.

8.3 Antidiabetic and metabolic disease research

Medicinal plants have been investigated for their potential influence on carbohydrate metabolism, insulin signalling, lipid regulation, and oxidative stress. Computational screening may prioritize phytochemicals for investigation against targets such as alpha-glucosidase, alpha-amylase, and selected proteins involved in metabolic regulation [15]. The integration of ethnobotanical reports with metabolomic profiling may help identify species and chemical constituents that warrant further investigation. However, the clinical management of diabetes requires evidence from controlled pharmacological and clinical studies, and traditional use or computational prediction should not be considered evidence of clinical effectiveness.

8.4 Neuroprotective and neuropharmacological research

Plant-derived compounds are also investigated for their potential interactions with molecular pathways associated with neuroinflammation, oxidative stress, neurotransmission, and neurodegeneration. AI-assisted target prediction can help prioritize phytochemicals for selected neurological targets, while molecular docking can provide an initial assessment of possible ligand–protein interactions [2]. Further evaluation should consider blood–brain barrier permeability, metabolic stability, toxicity, and pharmacokinetic properties. These factors are particularly important because computational binding predictions do not establish that a compound can reach the relevant biological compartment.

8.5 Anticancer natural-product research

Natural products have historically contributed important molecular scaffolds to anticancer drug discovery. Metabolomics and AI can support the identification of candidate plant metabolites associated with cytotoxic or antiproliferative activity. Molecular docking may be used to investigate their potential interactions with proteins involved in cell-cycle regulation, apoptosis, DNA repair, and oncogenic signalling [3]. However, the distinction between general cytotoxicity and selective anticancer activity is essential.

Candidate compounds should be evaluated using appropriate cancer and noncancer cell models, concentration–response experiments, and mechanistic assays before any claims regarding anticancer potential are made.

9. Proposed Integrated Workflow for AI-Driven Medicinal Plant Bioactive Discovery

A systematic workflow can improve the reproducibility and scientific reliability of AI-assisted medicinal plant research. The proposed framework integrates botanical authentication, ethnobotanical knowledge extraction, metabolomic profiling, computational prioritization, and experimental validation.

Stage 1: Medicinal plant selection and authentication. Candidate species are selected using ethnobotanical evidence, biodiversity information, literature reports, and research objectives. Botanical identity is confirmed through taxonomic verification and voucher specimen documentation.

Stage 2: Ethnobotanical data extraction. Traditional uses, plant parts, preparation methods, geographical information, and reported therapeutic indications are collected from reliable sources. NLP-based approaches can assist in extracting and structuring information from scientific publications and other authorized resources.

Stage 3: Plant extraction and metabolomic profiling. Authenticated plant materials are processed using standardized extraction procedures. LC–MS, GC–MS, and/or NMR are employed to characterize chemical constituents. Appropriate analytical quality-control samples and reference standards are included.

Stage 4: Metabolite identification and data curation. Raw analytical data are processed, and chemical features are annotated using spectral libraries and relevant chemical databases. Metabolite identification confidence is recorded, distinguishing confirmed structures from tentative assignments.

Stage 5: AI-assisted data integration. Ethnobotanical information, metabolomic profiles, chemical structures, and biological assay results are integrated. Machine learning models are used to identify patterns, prioritize compounds, and predict potential bioactivities.

Stage 6: Molecular target identification. Biological targets are selected using established literature, curated target databases, and validated pathway information. Predicted target associations are evaluated for biological relevance.

Stage 7: Molecular docking and structural evaluation. Selected phytochemicals are prepared as ligands and evaluated against suitable protein structures. Docking protocols are validated using appropriate controls, and predicted interactions are analysed.

Stage 8: Molecular dynamics and computational assessment. Selected protein–ligand complexes may be subjected to molecular dynamics simulations and additional structural or energetic analyses.

Stage 9: Experimental validation. Prioritized compounds are isolated or obtained as reference materials and evaluated using relevant biochemical, cellular, or microbiological assays. Mechanistic studies and toxicity assessment are conducted where appropriate.

Stage 10: Reproducibility and data sharing. Analytical methods, computational parameters, software versions, model performance, source datasets, and experimental findings are documented to support independent verification.

10. Major Challenges and Methodological Limitations

Although AI-assisted discovery offers substantial opportunities, several methodological challenges may affect the reliability and interpretation of results.

10.1 Incomplete and heterogeneous ethnobotanical information

Ethnobotanical databases vary in geographical coverage, documentation quality, taxonomic accuracy, and reporting standards. Traditional medicinal applications may be described using local terminology that cannot always be directly translated into modern biomedical disease classifications. Furthermore, historical records may contain incomplete information about plant identity, dosage, preparation, or administration. Standardized data collection and the preservation of original terminology are therefore important. AI-based extraction should retain links to the original sources and should not convert traditional use claims into established therapeutic indications.

10.2 Uncertainty in metabolite identification

Metabolomic analysis frequently produces large numbers of molecular features, but not all features can be confidently identified. Isomeric compounds, overlapping signals, adduct formation, and incomplete reference libraries can complicate annotation. AI-based structural prediction may help prioritize candidate identities, but experimental confirmation using authentic standards and orthogonal analytical methods remains important.

10.3 Limited training datasets and model generalizability

Machine learning models require sufficiently large, diverse, and accurately labelled datasets. Natural-product datasets often contain imbalanced chemical classes, inconsistent bioactivity measurements, and limited information on inactive compounds. Models trained on synthetic pharmaceutical compounds may not generalize effectively to structurally complex natural products. Independent validation, scaffold-based dataset splitting, uncertainty estimation, and applicability-domain assessment can improve confidence in predictive performance.

10.4 Interpretability and reproducibility

Complex deep learning models may provide predictions without clearly explaining the chemical or biological characteristics responsible for their outputs. Explainable AI methods, including feature attribution and model interpretation techniques, can help researchers evaluate predictive relationships. However, feature importance should not automatically be interpreted as evidence of a causal biological mechanism. Reproducibility also requires detailed reporting of data sources, preprocessing procedures, model architecture, hyperparameters, software versions, and validation methods.

10.5 Limitations of molecular docking

Molecular docking provides an approximation of potential ligand–protein interactions and is influenced by protein flexibility, protonation states, solvent effects, binding-site selection, and scoring-function limitations. Large and flexible natural products may be particularly challenging to model accurately. Therefore, docking should be treated as a prioritization technique rather than a definitive measure of binding affinity or pharmacological activity.

10.6 Ethical and regulatory considerations

The use of traditional medicinal knowledge in computational drug discovery raises questions concerning consent, intellectual property, community recognition, and equitable benefit-sharing. Researchers should comply with applicable national biodiversity legislation and access-and-benefit-sharing frameworks. AI systems should also be designed to minimize the unauthorized disclosure of confidential traditional knowledge and should clearly identify the provenance of their training data.

12. Conclusion

Artificial intelligence-driven medicinal plant research offers an emerging framework for integrating traditional knowledge with contemporary analytical and computational technologies. Ethnobotanical databases provide a foundation for selecting medicinal species and understanding their traditional applications, while metabolomics facilitates the systematic characterization of their chemical constituents. Machine learning and deep learning can identify complex relationships among plant metabolites, biological activities, and molecular targets, whereas molecular docking and molecular dynamics simulations provide additional structural evidence for prioritizing candidate compounds. The strength of this integrated framework lies in its capacity to connect different levels of evidence within a coordinated discovery process. However, computational predictions must be interpreted cautiously, particularly when metabolite identities are tentative, training datasets are limited, or molecular docking results lack experimental confirmation. When supported by rigorous experimental investigation, AI-assisted approaches may substantially improve the efficiency, reproducibility, and scientific value of medicinal plant bioactive discovery and contribute to the development of natural-product-inspired therapeutic candidates.

References

1. Zhang, S., Wang, W., Pi, X., He, Z., & Liu, H. (2023). Advances in the application of traditional Chinese medicine using artificial intelligence: a review. *The American journal of Chinese medicine*, *51*(05), 1067-1083.
2. Barlow, D. J., Buriani, A., Ehrman, T., Bosisio, E., Eberini, I., & Hylands, P. J. (2012). In-silico studies in Chinese herbal medicines' research: evaluation of in-silico methodologies and phytochemical data sources, and a review of research to date. *Journal of ethnopharmacology*, *140*(3), 526-534.
3. Anand, U., Jacobo-Herrera, N., Altemimi, A., & Lakhssassi, N. (2019). A comprehensive review on medicinal plants as antimicrobial therapeutics: potential avenues of biocompatible drug discovery. *Metabolites*, *9*(11), 258.
4. Lephatsi, M. M., Choene, M. S., Kappo, A. P., Madala, N. E., & Tugizimana, F. (2023). An integrated molecular networking and docking approach to characterize the metabolome of Helichrysum splendidum and its pharmaceutical potentials. *Metabolites*, *13*(10), 1104.
5. Adhikary, T., & Basak, P. (2020). Interdisciplinary approaches incorporating computational intelligence in modern pharmacognosy to address biological problems. In *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2020* (pp. 11-19). Singapore: Springer Singapore.
6. Sherlin, Rosita A., Pushparaj Annadurai, Thomas Jebastin, Sriram Srinivasan, Priscilla Arockiasamy, S. Danushri, Vijay Nirusimhan, Abhinav Parashar, and Daniel A. Gideon. "281 Workflow and Methods of In Silico Analysis of Compounds Isolated From Medicinal Plants: An Overview." *Phytopharmaceuticals and Biotechnology of Herbal Plants* (2023): 281-305.
7. Liu, Y., Sun, W., Shen, N., Hao, W., Xin, H., Che, F., & Cui, Y. (2023). Network pharmacology and molecular docking combined with widely targeted metabolomics to elucidate the potential compounds and targets of Euphorbia helioscopia seeds for the treatment of pulmonary fibrosis. *Computers in Biology and Medicine*, *160*, 107007.
8. Khairy, A., Ghareeb, D. A., Celik, I., Hammoda, H. M., Zaatout, H. H., & Ibrahim, R. S. (2023). Forecasting of potential anti-inflammatory targets of some immunomodulatory plants and their constituents using in vitro, molecular docking and network pharmacology-based analysis. *Scientific reports*, *13*(1), 9539.
9. Sharma, V., & Sarkar, I. N. (2013). Bioinformatics opportunities for identification and study of medicinal plants. *Briefings in bioinformatics*, *14*(2), 238-250.
10. Tung, C. W. (2014). Public databases of plant natural products for computational drug discovery. *Current computer-aided drug design*, *10*(3), 191-196.
11. Hussain, Shajaat, Tania Sagar, Sandeep Kaur, Nipunta, Nisha Kapoor, and Ritu Mahajan. "Omics approaches to study the biosynthesis of bioactive compounds in medicinal and aromatic plants." In *Biosynthesis of Bioactive Compounds in Medicinal and Aromatic Plants: Manipulation by Conventional and Biotechnological Approaches*, pp. 41-64. Cham: Springer Nature Switzerland, 2023.

12. Ma, X., Meng, Y., Wang, P., Tang, Z., Wang, H., & Xie, T. (2020). Bioinformatics-assisted, integrated omics studies on medicinal plants. *Briefings in bioinformatics*, *21*(6), 1857-1874.
13. Khurshid, Umair, Saeed Ahmad, Hammad Saleem, Arslan Hussain LodhI, Irfan Pervaiz, Mohsin Abbas Khan, Haroon Khan et al. "Multifaced assessment of antioxidant power, phytochemical metabolomics, in-vitro biological potential and in-silico studies of Neurada procumbens L.: An important medicinal plant." *Molecules* 27, no. 18 (2022): 5849.
14. Babar, M. M., Zaidi, N. U. S. S., Pothineni, V. R., Ali, Z., Faisal, S., Hakeem, K. R., & Gul, A. (2017). Application of bioinformatics and system biology in medicinal plant studies. In *Plant Bioinformatics: Decoding the Phyta* (pp. 375-393). Cham: Springer International Publishing.
15. Sharma, B., & Yadav, D. K. (2022). Metabolomics and network pharmacology in the exploration of the multi-targeted therapeutic approach of traditional medicinal plants. *Plants*, *11*(23), 3243.
16. Singh, K., Kumar, A., Kumar, S., & Gairola, S. (2023). Bioprospecting of plants for phytochemicals: Important for drugs. In *Phytochemical genomics: Plant metabolomics and medicinal plant genomics* (pp. 69-83). Singapore: Springer Nature Singapore.
17. Palermo, A. (2023). Metabolomics-and systems-biology-guided discovery of metabolite lead compounds and druggable targets. *Drug Discovery Today*, *28*(2), 103460.