

Original Research

Received 03 July 2026

Revised 13 Aug 2026

Accepted 26 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (18s): 847–859

KEYWORDS:

Artificial intelligence; Machine learning; Natural product discovery; Ethnopharmacology; Molecular mechanisms; Network pharmacology.

Integrating Artificial Intelligence and Machine Learning into Natural Product Discovery for Human Health from Ethnopharmacological Knowledge to Molecular Mechanisms

Manish Rana¹, Rahul Prakash Neve², Jesalkumari Varolia³, Swara Snehit Patil⁴, Disha Anand Bhandare⁵, Preeti Vaibhav Baviskar⁶

¹ Department of Artificial Intelligence and Data Science, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India

² Department of Information Technology, Thakur College of Engineering and Technology, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India

³ Department of Computer Science and Engineering (Cyber Security), Thakur College of Engineering and Technology, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India

⁴ Department of Computer Engineering, Aldel Education Trust's St. John College of Engineering & Management College, Palghar, Maharashtra, India

⁵ Department of Computer Engineering, Aldel Education Trust's St. John College of Engineering & Management College, Palghar, Maharashtra, India

⁶ Department of Artificial Intelligence and Machine Learning Engineering, Aldel Education Trust's St. John College of Engineering & Management College, Palghar, Maharashtra, India

ABSTRACT: Background: Ethnopharmacology has accumulated millennia of traditional medicinal knowledge, yet natural product discovery remains constrained by low screening efficiency, delayed mechanistic elucidation, and fragmented translation from traditional knowledge to clinical application. Artificial intelligence (AI) and machine learning (ML) offer transformative tools for systematically mining traditional knowledge, predicting bioactivity, and deconstructing multi-target mechanisms.

Methods: This review systematically integrates representative research on AI/ML in natural product discovery from 2020–2026, covering chemical language models, graph neural networks, generative AI, network pharmacology, molecular docking, and dynamics simulation. Based on 76 references, a five-stage integrated framework is constructed: knowledge digitization, activity prediction, mechanism deconstruction, and translational assessment.

Results: AI/ML has demonstrated significant advantages across ethnopharmacological data mining, virtual screening of natural product libraries, ADMET prediction, multi-target mechanism network analysis, and personalized ethnomedicine. Generative models (e.g., MolBERT, DrugGPT, GENTRL) can complete the cycle from generation to lead optimization within 46 days. The integration of network pharmacology with HTS² enables coupling of computational prediction and transcriptomic validation. However, the field faces systemic challenges including data fragmentation, benchmark absence, insufficient interpretability, and lagging ethical governance.

Conclusion: The deep integration of AI/ML with ethnopharmacology requires a shift from "tool stacking" to "paradigm reconstruction." Key priorities include building standardized digital infrastructure for traditional knowledge, developing specialized models for natural product chemical space, and embedding mechanistic validation into computational workflows. Future directions include multimodal foundation models, closed-loop experimental platforms, and community-engaged ethical frameworks.

1. INTRODUCTION

1.1 Traditional Paradigm and Dilemma of Natural Product Discovery

Natural products have consistently served as a core source of the human pharmacopoeia. According to statistics, approximately half of the small-molecule drugs approved by the US FDA between 1981 and 2019 can be traced to natural products or their derivatives [3]. Ethnopharmacology, as a bridge connecting traditional medicinal knowledge with modern drug discovery, provides an irreplaceable knowledge starting point for natural product research through systematic documentation and analysis of medicinal plant use across different cultures [14]. From aspirin (derived from salicin in willow bark) to artemisinin (derived from *Artemisia annua* as documented in the Handbook of Prescriptions for Emergencies), traditional knowledge has repeatedly led directly to major pharmaceutical breakthroughs [18].

However, the traditional natural product discovery paradigm faces deep structural challenges. First, the mismatch between screening efficiency and chemical space: active constituents in plant extracts often exist at extremely low abundances, and traditional bioassay-guided fractionation requires months or even years while risking the loss of synergistic activities during separation [5]. Second, the lag in mechanistic elucidation: numerous studies stop at descriptive conclusions such as "extract XX possesses activity YY," failing to satisfy modern pharmacology's demand for defined targets, pathways, and structure-activity relationships. The journal *Natural Resources for Human Health* explicitly states in its author guidelines that "descriptive approaches are prevented," reflecting the scholarly consensus on this issue [1]. Third, the systematic loss of ethnopharmacological knowledge: approximately 80% of traditional medicinal knowledge worldwide exists in oral transmission forms with minimal digitization, and the aging of knowledge-holder communities exacerbates the urgency [14].

1.2 Rationality and Transformative Potential of AI/ML Intervention

The application of artificial intelligence (AI) and machine learning (ML) in natural products is not a simple "technological addition" but a paradigmatic response to the aforementioned dilemmas [3].

Addressing screening efficiency, chemical language models (CLMs) represent molecules as SMILES strings, leveraging the attention mechanisms of Transformer architectures to learn statistical regularities of chemical space, thereby enabling high-throughput virtual screening and de novo molecular generation [2]. The generative model GENTRL completed the entire process from discovery to preclinical candidate identification for DDR1 kinase inhibitors within 46 days, demonstrating an order-of-magnitude acceleration in AI-driven speed [14].

Addressing mechanistic elucidation, network pharmacology transforms the "one drug, one target" paradigm into a "multi-component, multi-target, multi-pathway" systems perspective, while graph neural networks (GNNs) and knowledge graph technologies further enhance the functional mapping capability from compound structures to biological networks [5]. The integration of AI with high-throughput sequencing screening platforms (HTS²) enables computational predictions to receive transcriptomic-level empirical validation, forming a "prediction-validation-correction" closed loop [9].

Addressing knowledge loss, natural language processing (NLP) and multilingual speech recognition models provide technological pathways for digitizing ethnopharmacological oral knowledge. The transcription capabilities of multilingual models such as Whisper, the literature mining capabilities of biomedical LLMs, and knowledge graph construction technologies collectively constitute the infrastructure for systematic preservation and utilization of traditional knowledge [22][24].

1.3 Scope and Objectives of This Review

This review takes "from ethnopharmacological knowledge to molecular mechanisms" as its central thread, systematically examining methodological advances, representative applications, and existing bottlenecks of AI/ML across the full chain of

natural product discovery [3][14]. The differentiating positioning relative to existing reviews lies in: (1) full-chain perspective: rather than isolated discussion of a single technology (e.g., only network pharmacology or only generative models), an integrated framework from knowledge digitization to mechanistic validation is constructed; (2) ethnopharmacological anchoring: traditional medicinal knowledge serves as the starting point rather than general natural product screening; (3) critical assessment: not only summarizing successful cases but systematically examining systemic limitations at data, benchmark, interpretability, and ethical levels [6].

2. DIGITIZATION AND AI-DRIVEN MINING OF ETHNOPHARMACOLOGICAL KNOWLEDGE

2.1 Datafication Dilemmas of Traditional Knowledge

The digitization of ethnopharmacological knowledge faces three obstacles. First, format heterogeneity: traditional knowledge exists in oral transmission, handwritten notes, local language texts, and ethnobotanical specimen records, with extremely low standardization [10]. Second, language barriers: substantial knowledge is recorded in non-English local languages (e.g., Sanskrit, Ayurvedic texts, classical Chinese medicine literature, African indigenous languages), while mainstream NLP models have training corpora heavily biased toward English and Chinese [22]. Third, context dependency: traditional medicinal knowledge is embedded in specific cultural, ecological, and philosophical frameworks (e.g., Ayurvedic constitution theory, Chinese medicine's cold-heat balance), and simple "species-use" mappings lose critical contextual information [18].

2.2 AI-Driven Knowledge Extraction and Structuring

Natural language processing (NLP) provides the first layer of solutions to these dilemmas. Biomedical LLMs (such as BioGPT and PubMedBERT), after fine-tuning, can perform named entity recognition, relation extraction, and question-answering tasks on traditional medicine literature [22]. Multilingual speech recognition models (such as Whisper) can transcribe oral knowledge from field investigations into structured text, addressing the "last mile" problem of oral knowledge digitization [24].

Knowledge graphs constitute the second layer of solutions. Graph-structured representation methods encode traditional knowledge into semantic networks of "herb-compound-target-disease," enabling implicit associative relationships to be discovered by algorithms [10]. Tensor-based knowledge graphs have been used to map traditional Chinese medicine's "herb-symptom-formula" associations, while transfer learning predicts herbal formulas from clinical records [18]. More advanced multimodal models (such as Med-Gemini) integrate text, images, and structured data, capable of simultaneously processing visual identification of medicinal plants, chemical fingerprint spectra, and ethnobotanical metadata [22][24].

2.3 Current Status Assessment of Data Resources and Benchmarks

Currently available public data resources for AI-driven ethnopharmacological research include: NPASS (Natural Product Activity and Species Source database), which now contains 204,023 natural products, 48,940 organisms, 8,764 targets, over 1 million experimental activity records, 34,975 toxicity records, and 9,713 quantitative ADME records [10]. COCONUT 2.0 integrates more than 63 open natural product resources, providing unrestricted access to 695,133 molecules with information on stereochemical forms, producing organisms, and geographic distribution [19]. GNPS (Global Natural Products Social Molecular Networking) serves as an open-access knowledge base for community-wide organization and sharing of tandem mass spectrometry data [20]. IMPPAT 2.0 documents 4,010 Indian medicinal plants, 17,967 phytochemicals, and 1,095 therapeutic uses with curated plant-part-phytochemical associations [21].

However, these databases exhibit significant gaps. Activity data sparsity is the core problem: over 300,000 natural products are known, yet NPASS contains only 47,418 NPs (22.75%) with activity annotations [10]. Species coverage bias is equally severe: databases are dominated by Asian (particularly Chinese and Indian) medicinal plants, with IMPPAT centered on Indian plants restricting its use to regional data [21]. Standardized benchmark absence makes systematic comparison between models nearly impossible—different studies vary greatly in data splitting, evaluation metrics, and negative sample selection [2].

Table 2.3.1 summarizes the coverage characteristics, scale, and primary gaps of these key public data resources.

Table 2.3.1: Coverage characteristics of ethnopharmacology-relevant public data resources.

Database	Content	Scale	Primary Gaps
NPASS 3.0	Natural product activity	204,023 NPs; 1,048,756 activity records; 34,975 toxicity records; 9,713 ADME records	22.75% NPs with activity annotation
COCONUT 2.0	Open natural products	695,133 molecules; 53,092 organisms; 63 collections integrated	Structure-dominated; limited activity data
GNPS	Mass spectrometry molecular networking	Community-wide MS/MS data sharing; crowdsourced curation	Annotation coverage; raw data searchability
IMPPAT 2.0	Indian medicinal plants	4,010 plants; 17,967 phytochemicals; 1,095 therapeutic uses	Geographic bias (India-centric)
PubMed/LLM corpora	Literature	English-dominated	Multilingual coverage; low-resource languages

3. APPLICATIONS OF AI/ML IN NATURAL PRODUCT ACTIVITY PREDICTION AND SCREENING

3.1 From Traditional QSAR to Deep Representation Learning

Quantitative structure-activity relationship (QSAR) models have long been used to predict natural product bioactivity, but they rely on manually designed molecular descriptors and struggle to capture complex structure-activity relationships. A significant limitation of traditional QSAR models has been their reliance solely on compounds with available ADMET experimental results, posing challenges to model generalization [12]. The rise of chemical language models (CLMs) has changed this landscape. SMILES strings transform molecular structures into "chemical sentences" processable by sequence models, and the self-attention mechanisms of Transformer architectures can learn long-range structural dependencies and chemical grammar constraints [2][12].

ChemBERTa adapts transformer pretraining to molecular property prediction and was pretrained on large PubChem SMILES corpora, demonstrating superior performance on toxicity problems like Tox21 and ClinTox [12]. ChemBERTa2 extends this line towards chemical foundation models. MolFormer uses large-scale molecular language pretraining to capture structural regularities from SMILES sequences [2]. Their advantage is broad unsupervised chemical exposure; their risk is representational mismatch, as SMILES pretraining may not align with narrow assay endpoints or chemically shifted test folds [12]. MolBERT's multi-task framework supports compound screening and ethnopharmacological lead selection [3].

Graph neural networks (GNNs) provide a complementary pathway: representing molecules as graphs (atoms as nodes, chemical bonds as edges), GNNs learn local chemical environments through message-passing mechanisms. A recent hybrid GNN architecture for IC₅₀ prediction integrates molecular graphs with node features (11 dimensions including atom type, hybridization, degree, formal charge, aromaticity, chirality) and edge features (5 dimensions including bond type, conjugation, ring membership, stereochemistry), achieving improved prediction accuracy and mechanistic interpretability in drug discovery applications [8].

Figure 3.1.1 provides a methodological comparison of traditional QSAR, CLMs, and GNNs across key dimensions relevant to natural product activity prediction.

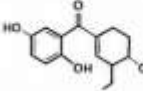
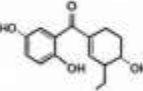
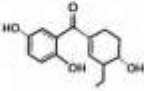
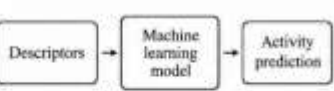
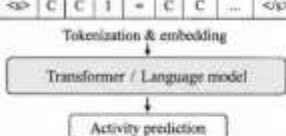
	Traditional QSAR	Chemical Language Models (CLMs)	Graph Neural Networks (GNNs)
Input Representation	 Molecular descriptors (e.g., physicochemical, structural, fingerprint)	 SMILES / SELFIES (e.g., <chem>CC1=CC(=O)O</chem>, <chem>C(=C)C(=O)O</chem>)	 ○ = atom — = bond
Key Approach			
Model Output	 Quantitative activity value (e.g., pIC₅₀)	• Activity class • Activity probability • Continuous value (e.g., pIC₅₀)	• Activity class • Activity probability • Continuous value (e.g., pIC₅₀)
Strengths	• Simple and interpretable • Requires less data • Computationally efficient	• Captures complex chemical semantics • Leverages large-scale data • Good at zero-shot prediction	• Models molecular structure directly • Captures local and global context • High prediction accuracy
Limitations	• Limited by descriptor quality • May miss complex structure-activity relationships	• Requires large pretraining data • Less interpretable • Computationally intensive	• Needs sufficient data • Higher computational cost • Interpretability can be challenging

Fig. 3.1.1: Methodological comparison of traditional QSAR, chemical language models (CLMs), and graph neural networks (GNNs) for natural product activity prediction.

Dimension	Traditional QSAR	CLMs (SMILES-based)	GNNs
Molecular representation	Manual descriptors	SMILES strings	Molecular graphs
Architecture	Linear/tree models	Transformer/RNN	Message-passing networks
Advantages	Interpretability	Long-range dependency learning	Local chemical environment
Limitations	Representation bottleneck	Syntax errors/non-uniqueness	Over-smoothing/scalability
Natural product applicability	Moderate	Medium-High (requires fine-tuning)	High (structure-sensitive)

3.2 Generative Models and De Novo Design

Generative AI extends natural product discovery from "screening known" to "generating unknown" [5]. MolGPT is a Transformer-decoder model that generates molecules via scaffold-conditioned and multi-property conditional training on the MOSES dataset. In conditional generation experiments, MolGPT achieved validity rates of 0.90–0.99 and novelty rates of 1.0 across various scaffold conditions [16]. DrugEx applies RNN-based reinforcement learning with a repair mechanism (DrugEx v3) to correct chemically invalid molecules, supporting scaffold-constrained ligand design for multi-target optimization [15]. GENTRL (Generative Tensorial Reinforcement Learning) optimizes synthetic feasibility, novelty, and biological activity

simultaneously. In its landmark application, GENTRL designed six novel inhibitors of DDR1 kinase in 21 days; four compounds were active in biochemical assays, two in cell-based assays, and one lead candidate demonstrated favorable pharmacokinetics in mice, with the entire process from generation to preclinical validation completed in 46 days [14][27].

However, the "natural product applicability" of generative models faces fundamental challenges. Mainstream generative models are trained on synthetic small-molecule libraries (e.g., ZINC15, ChEMBL), and the chemical space they learn exhibits systematic deviation from natural product structural characteristics [6]. In benchmark assessments against Bcl-2 family proteins, MolGPT showed very small improvements (scores of 6.56, 6.34, and 6.03) compared to ZINC15 training data (5.99, 5.88, 5.70), resulting in a low success rate of 0.05%, while graph-based models MARS and RationaleRL demonstrated much better binding affinity improvements [6]. Natural product similarity constraints (NP-likeness) and biosynthetic feasibility constraints are being integrated into generative frameworks, but maturity remains limited [5].

3.3 ADMET Prediction and Safety Assessment

ADMET (absorption, distribution, metabolism, excretion, toxicity) prediction is a critical component of natural product druggability assessment [3]. ChemBERTa recorded superior performance on toxicity problems like Tox21 (ranked 1st) and ClinTox (ranked 3rd) compared to supervised learning models and graph models [12]. The pkCSM platform, based on graph signature methods, can predict endpoints including water solubility, intestinal absorption, blood-brain barrier penetration, CYP450 inhibition, and hepatotoxicity. AI-based ADMET modeling platforms such as SwissADME, pkCSM, and DeepADMET are being applied to simulate biological behavior, predict metabolic stability, and optimize compound structures before synthesis or clinical testing [18].

ADMET prediction for natural products faces special challenges: multi-component synergistic effects make it difficult to extrapolate ADMET data from single compounds to complex formulas; traditional use experience (such as toxicity reduction through processing) contains safety information difficult to encode in existing models; and toxicity reports in low-resource language literature are often overlooked by mainstream databases [10]. NPASS 3.0 now includes 34,975 quantitative toxicity records and 9,713 quantitative ADME records, representing a significant expansion over previous versions [10].

Table 3.3.1 summarizes representative AI/ML models in natural product discovery, including their types, primary applications, and key performance metrics.

Table 3.3.1: Representative AI/ML models in natural product discovery.

Model	Type	Primary Application	Key Performance/Output	Reference
ChemBERTa	CLM (Transformer)	ADMET prediction	BBB permeability/hepatotoxicity superior to QSAR; Tox21 rank 1, ClinTox rank 3	[12]
MolBERT	CLM (multi-task)	Compound screening	Multi-task superior to single-task; ethnopharmacological lead selection	[3]
GENTRL	Generative model	De novo design	DDR1 inhibitor in 46 days; 4 active compounds, 1 mouse PK validated	[14][27]
DrugEx	Generative model	Scaffold-constrained design	RNN-based reinforcement learning with repair mechanism	[15]
MolGPT	Generative model	Natural product analogs	Scaffold-conditioned; validity 0.90–0.99, novelty 1.0	[16]

AlphaFold3	Structure prediction	Protein-ligand complexes	33 ternary complexes RMSD < 1 Å; 46 with RMSD < 4 Å	[7][17]
GNN (hybrid)	Graph neural network	IC ₅₀ prediction	Atomic hybridization features; improved mechanistic interpretability	[8]
Network pharmacology + HTS ²	Integrated framework	Mechanism deconstruction	FLAP target identification; validated by CETSA, docking, zebrafish	[9]

4. DECONSTRUCTION FROM COMPUTATIONAL PREDICTION TO MOLECULAR MECHANISMS

4.1 Network Pharmacology: From Single Target to Systems Perspective

Network pharmacology transforms natural product mechanism research from a "single compound-single target" paradigm to a "multi-component, multi-target, multi-pathway" framework [5]. Typical workflows include: collecting phytochemical constituents from databases; predicting targets through SwissTargetPrediction; intersecting with disease-related genes from GeneCards; constructing PPI networks and identifying hub genes; and validating key interactions through molecular docking. However, network pharmacology's predictive reliability remains constrained by inherent biases stemming from incomplete database coverage and algorithmic oversimplification, which may lead to false-positive predictions [9].

Taking the anti-inflammatory mechanism study of pentacyclic triterpenes from *Sanguisorba officinalis* as an example, researchers first predicted anti-inflammatory activities and candidate targets of 23 isolated compounds through network pharmacology, then obtained compound-specific transcriptomic signatures through HTS² (high-throughput sequencing screening). HTS² employs RNA-mediated oligonucleotide annealing, ligation, and next-generation sequencing to quantify gene expression directly, offering high-throughput, sensitivity, and reproducibility [9]. Integrative analysis identified compound 17 (3β,6β-dihydroxy-urs-12,19(29)-dien-28-oic acid) with 5-lipoxygenase activating protein (FLAP) as the primary target. This compound directly bound FLAP, reduced leukotriene B₄ production, suppressed MAPK signaling, and downregulated COX-2 and iNOS expression. Target engagement and downstream effects were validated by molecular docking, molecular dynamics simulations, CETSA, Western blotting, ELISA, and in vivo zebrafish assays [9].

4.2 AI-Enhanced Target Identification and Pathway Mapping

Knowledge graph embedding methods encode compound-target-pathway relationships as vectors, discovering potential new targets through link prediction [10]. Graph attention networks outperform traditional centrality measures in identifying disease-related hub genes within PPI networks [8]. A recent GNN model based on atomic hybridization features demonstrated that chemically enriched graph topologies can play a crucial role in improving prediction accuracy and mechanistic interpretability in drug discovery applications, with models trained on nine diverse ChEMBL targets spanning different protein families to assess generalizability [8].

Molecular docking and dynamics simulation combined with AI is an emerging direction [7]. AlphaFold3 demonstrates superior ligand positioning in predicting protein-ligand complexes, producing 33 ternary complexes with RMSD < 1 Å and 46 with RMSD < 4 Å compared to Boltz-1's 25 and 40, respectively, in PROTAC-mediated ternary complex modeling [7]. However, for GPCR-small molecule complexes, AlphaFold3's ligand positioning was highly variable and often inaccurate, rendering predictions unreliable particularly for allosteric modulators. The significant divergence from experimental structures underscores that experimental structures remain essential for validating ligand-binding accuracy in GPCR complexes [17]. Machine learning potential (MLP) methods perform large-scale MD simulations with near-quantum chemical accuracy, providing more realistic descriptions of the dynamic behavior of natural product-target complexes [4].

4.3 Construction of Integrated Frameworks: Closing the Loop from Prediction to Validation

The currently most mature integrated framework is the "network pharmacology + experimental validation" two-stage model [5][9]. The first stage generates hypotheses through computational prediction; the second stage validates through biochemical

and cellular experiments. The limitation of this model lies in: computational predictions have high false positive rates, while experimental validation throughput is limited and cannot keep pace with computational prediction output scale [9].

Closed-loop frameworks represent the direction for addressing this bottleneck. AI-driven active learning strategies can iteratively optimize prediction models based on experimental feedback [4]. Recent work has emphasized that bolder and deeper explorations of chemical space are needed, and combining ML model uncertainty estimates across broader swaths of chemical space could open pathways for fruitful statistical explorations in an active learning framework [2]. Automated experimental platforms (e.g., microfluidics + high-content imaging) can improve validation throughput. The digital twin concept (establishing computational models for specific plant-target systems, continuously updated with experimental data) points toward a longer-term vision [3].

Table 4.3.1 provides a comparative assessment of AI/ML methods for natural product mechanism research, covering their inputs, outputs, advantages, and limitations.

Table 4.3.1: Comparison of AI/ML methods for mechanism elucidation of natural products.

Method	Input	Output	Advantages	Limitations
Network pharmacology	Compound + disease targets	PPI/pathway network	Systems perspective; hypothesis generation	High false positives; incomplete database coverage [9]
Knowledge graph embedding	Triples	Link prediction	New target discovery	Data-dependent; coverage bias [10]
Molecular docking	3D structures	Binding energy	Direct target evidence	Sampling/scoring limitations [7]
Molecular dynamics	Complex	Dynamic behavior	Temporal dimension; physiological relevance	Computational cost [4]
HTS ²	Compound treatment	Transcriptomic signature	Experimental validation; high-throughput	Throughput limitation; cost [9]
CETSA	Cell lysates	Target binding	Intracellular validation; direct engagement	Antibody-dependent; throughput [9]

5. EMPIRICAL FRAMEWORK FOR MULTIDISCIPLINARY INTEGRATION: AN AI WORKFLOW ORIGINATING FROM ETHNOPHARMACOLOGY

5.1 Framework Design Principles

Based on the methodological advances described above, this review proposes an AI-integrated workflow oriented toward ethnopharmacological starting points, designed according to three principles [3]. First, knowledge-driven rather than data-driven: traditional medicinal knowledge serves as the starting point for hypothesis generation, not as a post-hoc "ethnopharmacological footnote" [14]. Second, closed-loop validation: each computational prediction stage corresponds to an explicit experimental validation strategy, avoiding self-reinforcing bias from "computation-computation" cycles [9]. Third, ethical embedding: knowledge-holder informed consent and benefit-sharing mechanisms are built-in workflow components rather than post-hoc additions [3].

5.2 Five-Stage Workflow

Figure 5.1.1 illustrates the five-stage AI-integrated workflow for ethnopharmacology-originated natural product discovery, spanning from knowledge digitization through translational assessment.

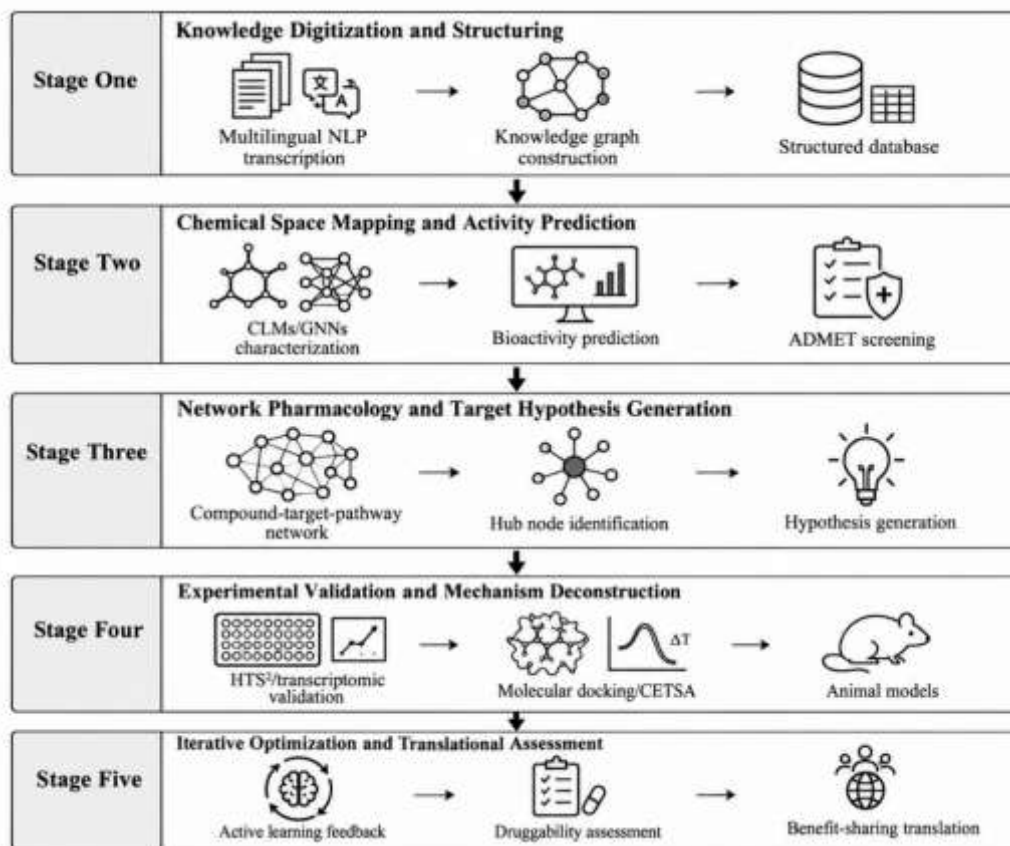


Fig. 5.1.1: Five-stage AI-integrated workflow for ethnopharmacology-originated natural product discovery.

Stage One: Knowledge Digitization and Structuring. Transcribe oral knowledge using multilingual NLP models; integrate text, images, and chemical data through knowledge graphs [22][24]. Output: structured "species-use-preparation method-cultural context" database.

Stage Two: Chemical Space Mapping and Activity Prediction. Characterize plant chemical constituents using CLMs and GNNs; predict bioactivity and ADMET properties [2][8]. Filter generative model outputs by NP-likeness and biosynthetic feasibility [5]. Output: candidate compound lists ranked by predicted activity.

Stage Three: Network Pharmacology and Target Hypothesis Generation. Construct compound-target-pathway networks; identify hub nodes and critical pathways [5]. Enhance hub identification with graph attention networks [8]. Output: experimentally testable target-pathway hypotheses.

Stage Four: Experimental Validation and Mechanism Deconstruction. Validate computational predictions through HTS² or targeted transcriptomics; verify direct target binding through molecular docking, CETSA, and cellular experiments; validate in vivo activity through animal models [9]. Output: experimentally validated molecular mechanisms.

Stage Five: Iterative Optimization and Translational Assessment. Update prediction models with experimental feedback (active learning) [4]; assess lead compound druggability and intellectual property status; initiate translational pathways compliant with community benefit-sharing requirements [3]. Output: translatable lead compounds or formula optimization schemes.

6. KEY CHALLENGES AND CRITICAL ASSESSMENT

6.1 Systemic Obstacles at the Data Level

Data fragmentation and heterogeneity constitute the most fundamental obstacle [10]. Ethnopharmacological data are dispersed across ethnobotanical literature, chemical databases, pharmacological experimental records, and clinical reports, with highly inconsistent formats, standards, and accessibility [21]. Data bias permeates all dimensions: species bias (over-representation of

Asian plants), language bias (English and Chinese dominance), and activity type bias (over-concentration in anti-inflammatory and anticancer research) [19]. Missing negative results lead to severe publication bias in model training sets, reducing prediction specificity [2]. NPASS 3.0 data expansion of 9.37%–206.30% over the previous version illustrates both the magnitude of recent curation efforts and the remaining gaps [10].

6.2 Core Limitations at the Methodological Level

Benchmark absence renders systematic evaluation of AI/ML methods nearly impossible [2]. Different studies vary greatly in data splitting, evaluation metrics, and negative sample selection, rendering reported "performance improvements" non-comparable [6]. Insufficient interpretability is an inherent challenge of deep learning methods: CLM attention weights are difficult to map directly to chemical explanations, and GNN message-passing processes are "black boxes" to pharmacologists [8]. Out-of-distribution generalization problems are particularly prominent in natural product scenarios: models trained on synthetic small-molecule libraries often show significant performance degradation in natural product chemical space, as evidenced by MolGPT's low success rate (0.05%) on Bcl-2 family protein benchmarks [6].

6.3 Ethical and Governance Challenges

Biopiracy risk is a core ethical issue in AI-driven ethnopharmacological research [3]. Traditional knowledge is often utilized by commercial entities without authorization or fair compensation. AI's "discovery" capabilities may exacerbate this problem: algorithms identify active compounds from publicly available ethnobotanical literature while bypassing knowledge-holder informed consent and benefit-sharing [14]. Data sovereignty issues are equally prominent: who owns digitized versions of traditional knowledge? How are community data stored and accessed? Critiques of algorithmic colonialism point to deeper asymmetries: AI infrastructure, models, and standards are primarily controlled by Global North institutions, while knowledge sources are primarily in the South [3].

6.4 Epistemic Tensions with Traditional Knowledge Systems

The reductionist assumptions embedded in AI/ML methods exist in fundamental tension with the holistic philosophies of many traditional medical systems [18]. Decomposing complex formulas into single compounds, simplifying "syndrome patterns" into disease labels, and dismissing processing techniques as noise—these operations are methodologically efficient but may systematically lose critical information from traditional medicine at the epistemological level [14]. Synergistic effects—the core claim of traditional medicine that multiple components produce superior effects to single compounds through multiple targets—are naturally difficult for AI models (especially generative models) to optimize for, as they tend to favor single compounds [5]. How to preserve holistic information within computational frameworks remains an unresolved fundamental problem [3].

7. FUTURE DIRECTIONS AND PERSPECTIVES

7.1 Multimodal Foundation Models and Domain Adaptation

Multimodal foundation models for natural product discovery should integrate molecular structures (SMILES/graphs), mass spectrometry/NMR spectra, protein sequences/structures, text (ethnobotanical records and pharmacological literature), and images (plant morphology, cellular phenotypes) [22]. Med-Gemini demonstrates the potential of multimodal models in medical domains, achieving state-of-the-art 91.1% accuracy on the MedQA benchmark and best-in-class performance on chest X-ray and CT report generation [24]. The key challenge is domain adaptation: models pre-trained on general chemical corpora require continued pre-training or fine-tuning to adapt to the unique distribution of natural product chemical space [2]. Few-shot learning capabilities are particularly important for low-resource scenarios (e.g., African traditional medicine) [3].

7.2 Closed-Loop Experimental Platforms and Autonomous Discovery Systems

The integration of AI prediction with automated experimentation is key to improving validation throughput [4]. The self-driving laboratory concept integrates AI prediction, robotic experimental execution, and real-time data analysis into a closed loop [5]. Active learning strategies can maximize information gain under limited experimental budgets [2]. Recent frameworks emphasize that combining ML model uncertainty estimates across broader swaths of chemical space could open pathways for fruitful statistical explorations in an active learning framework [2].

7.3 Explainable AI and Transparent Mechanistic Hypotheses

Explainable AI oriented toward pharmacologists and ethnopharmacologists needs to transcend shallow "attention heatmap" explanations [8]. Substructure-based explanations (identifying chemical fragments contributing most to predictions), knowledge-based explanations (attributing predictions to known pharmacological pathways), and counterfactual explanations ("how would predictions change if a certain group in the structure were altered") are more actionable directions [6]. Neuro-symbolic integration—combining deep learning's pattern recognition capabilities with knowledge graph symbolic reasoning—may provide more transparent mechanistic hypothesis generation [10].

7.4 Ethical Governance and Community Engagement Frameworks

Ethical frameworks for AI-driven ethnopharmacological research need to move from "compliance" (informed consent, ethical review) to "co-governance" [3]. Community-engaged research models require knowledge holders to have substantive voice in research design, data governance, and benefit distribution [14]. Data trust mechanisms can serve as institutional tools for community control of traditional knowledge digital assets [3]. Capacity building—establishing AI research capabilities and infrastructure in knowledge-source regions—is a structural response to "algorithmic colonialism" [3].

7.5 Priority of Standardization and Benchmark Construction

The most urgent current action is establishing field standards: standardized data formats (integrating ethnobotanical, chemical, and pharmacological metadata), standardized evaluation benchmarks (training/test sets with scaffold splitting and negative controls), and standardized reporting guidelines (specifying data sources, model architectures, and validation strategies) [2][6]. Building this infrastructure requires international collaboration and open science commitments [3]. The Natural Resources for Human Health journal highly encourages systematic reviews and requires that they be prepared by following PRISMA guidelines [1].

8. CONCLUSION

The integration of AI/ML with natural product discovery has transcended the category of "technological tool" and is reshaping the research paradigm from ethnopharmacological knowledge to molecular mechanisms [3]. This review has systematically examined the methodological foundations of this integration (chemical language models, graph neural networks, generative models, network pharmacology, and closed-loop validation frameworks) while critically assessing its systemic limitations—data fragmentation and bias, benchmark absence, insufficient interpretability, and deep ethical governance challenges [6][9].

The core judgment is: the true value of AI in natural products lies not in "replacing" traditional methods but in "integrating" dispersed knowledge, data, and validation stages. The key pathways from "tool stacking" to "paradigm reconstruction" include: building standardized ethnopharmacological digital infrastructure, developing specialized models for natural product chemical space, embedding mechanistic validation into closed-loop computational workflows, and establishing community-engaged ethical governance frameworks [3][10]. Ethnopharmacology provides what AI needs most—high-quality hypothesis sources based on millennia of human experience; AI provides what ethnopharmacology lacks most—computational capacity to systematically handle complexity [14]. The deep integration of the two, rather than unidirectional technological "application," is the viable pathway toward sustainable natural product drug discovery [5].

ACKNOWLEDGEMENTS

The authors sincerely acknowledge their respective academic institutions for providing an encouraging research environment and institutional support during the preparation of this review manuscript. The authors also acknowledge the contributions of researchers and scientific communities whose published studies, publicly accessible databases, and computational resources have advanced the integration of artificial intelligence and machine learning into ethnopharmacology and natural product discovery.

AUTHOR CONTRIBUTIONS

Manish Rana: Conceptualization, methodology, literature review, development of the integrated AI/ML framework, manuscript drafting, critical review, and supervision.

Rahul Prakash Neve: Literature investigation, methodological analysis, evaluation of machine learning techniques, and manuscript review.

Jesalkumari Varolia: Literature investigation, analysis of artificial intelligence applications, data interpretation, and manuscript editing.

Swara Snehit Patil: Literature collection, analysis of natural product discovery approaches, preparation of figures and tables, and manuscript drafting.

Disha Anand Bhandare: Review of ethnopharmacological knowledge, analysis of molecular mechanisms, literature synthesis, and manuscript editing.

Preeti Vaibhav Baviskar: Analysis of computational drug discovery approaches, critical evaluation of research findings, manuscript review, and proofreading.

All authors reviewed and approved the final version of the manuscript and agree to be accountable for their respective contributions.

CONFLICT OF INTEREST

The authors declare that they have no known financial or non-financial conflicts of interest that could have influenced the research, analysis, interpretation, or preparation of this manuscript.

ETHICS STATEMENT

This manuscript is a literature-based review that synthesizes findings from previously published scientific studies and publicly accessible research resources. No new experiments involving human participants or animals were conducted as part of this work. Therefore, institutional ethics committee approval and informed consent were not required.

The review emphasizes the ethical use of ethnopharmacological knowledge, including respect for indigenous and traditional knowledge systems, data sovereignty, informed consent, equitable benefit-sharing, and responsible applications of artificial intelligence in natural product discovery.

DATA AVAILABILITY STATEMENT

No new experimental datasets were generated or analyzed during the preparation of this review manuscript. The information presented was obtained from previously published scientific literature and publicly accessible resources, including NPASS, COCONUT, GNPS, and IMPPAT. Relevant sources are appropriately cited in the manuscript. All information supporting the conclusions of this review is available through the cited references and their associated publicly accessible resources.

References

- [1] Natural Resources for Human Health, "Instructions for Authors." [Online]. Available: <https://www.nrfhh.com/Instructions-for-Authors,3102.html>
- [2] "Do Larger Models Really Win in Drug Discovery? A Benchmark Assessment of Model Scaling in AI-Driven Molecular Property and Activity Prediction," bioRxiv, 2026.
- [3] S. Bhadra et al., "Reimagining ethnopharmacology with generative AI: Towards inclusive, ethical, and data-driven traditional medicine," *Pharmacological Research*, vol. 221, p. 108002, 2025.
- [4] "Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems," arXiv:2102.06321.
- [5] "Multi- and Many-Objective Optimisation: Present and Future in De Novo Drug Design," ChemRxiv, 2023.
- [6] "AI-based molecular generation for Bcl-2 family proteins: benchmark assessment," Kangwon National University, 2024.
- [7] "Predicting PROTAC-mediated ternary complexes with AlphaFold3 and Boltz-1," ScienceDirect, 2025.
- [8] "Graph Neural Networks Model Based on Atomic Hybridization for Predicting Drug Targets," *Journal of Chemical Information and Modeling*, 2026.
- [9] "Network pharmacology integrated with HTS2 deciphers the anti-inflammatory mechanism of pentacyclic triterpene from *Sanguisorba officinalis*," *Fitoterapia*, 2026.

- [10] "NPASS database update 2026: comprehensive quantitative composition, bioactivity, and ADME-Tox data of natural products for biomedical research," *Nucleic Acids Research*, 2026.
- [11] Natural Resources for Human Health, "Authorship and Contributor Guidelines."
- [12] "Assessment of ChemBERTa and ELECTRA for ADMET prediction on MoleculeNet benchmarks," *Pharmaceuticals*, vol. 17, p. 382, 2024.
- [13] "Artificial intelligence-based screening of phytochemicals for targeted cancer therapy," *Natural Products and Bioprospecting*, 2026.
- [14] Zhavoronkov et al., "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," *Nature Biotechnology*, vol. 37, pp. 1038–1040, 2019.
- [15] "DrugEx: RNN-based reinforcement learning with repair mechanism for de novo drug design," *Engineering*, 2023.
- [16] "MolGPT: Molecular Generation using a Transformer-Decoder Model," *Journal of Chemical Information and Modeling*, 2021.
- [17] "An update for AlphaFold3 versus experimental structures: assessing the precision of small molecule binding in GPCRs," *Acta Pharmacologica Sinica*, 2025.
- [18] "AI in ethnopharmacology, the pharmaceutical industry, and its applications," *ScienceDirect*, 2025.
- [19] "COCONUT 2.0 - Complete database," *Zenodo*, 2024.
- [20] "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking," *Nature Biotechnology*, vol. 34, pp. 828–837, 2016.
- [21] "IMPPAT: A digital phytochemical atlas enabling traditional medicine research," *International Journal of Ayurveda Research*, 2025.
- [22] "Advancing medical AI with Med-Gemini," *Google Research Blog*, 2024.
- [23] "COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database," *Nucleic Acids Research*, vol. 53, pp. D634–D643, 2025.
- [24] "Advancing Multimodal Medical Capabilities of Gemini," *arXiv:2405.03162*.
- [25] "Integrating Artificial Intelligence and Machine Learning into Natural Product Discovery," *Expert Review*, 2026.
- [26] "Chemical language models for natural product discovery," *Natural Product Reports*, 2026.
- [27] "Deep learning enables rapid identification of potent DDR1 kinase inhibitors," *PubMed*, PMID: 31477924.
- [28] "Sanguisorbae radix immunomodulatory effects at single-cell resolution," *Phytomedicine*, 2025.
- [29] "Leveraging artificial intelligence (AI) in herbal medicinal plant extracts: Pathway to drug development," *ScienceDirect*, 2026.
- [30] "3-n-Propionyl Zaluzanin C from Saussurea involucrella alleviates inflammation via NRF2 signaling," *Phytomedicine*, 2026.
- [31] "Discovery of novel anti-inflammatory compounds from Borojó via zebrafish phenotyping," *ScienceDirect*, 2026.
- [32] "Integrated Computational and Machine-Learning Pipeline for Identification of Flavonoid-Based PIKfyve Inhibitors," *ScienceDirect*, 2026.
- [33] "Experimental validation of SCOPE-DTI-predicted targets for natural products," *Nature Communications*, 2025.
- [34] "Remodeling the gut ecosystem: Natural products in ulcerative colitis treatment," *ScienceDirect*, 2026.
- [35] "Integrated virtual screening, machine learning and molecular dynamics identify novel phytochemical FabI inhibitors," *ScienceDirect*, 2026.
- [36] "Database Commons: NPASS," *National Genomics Data Center*, 2025.
- [37] Natural Resources for Human Health, "Manuscript Submission Process."
- [38] Scimago Journal Rank, "Natural Resources for Human Health."
- [39] S. Bhadra et al., "Reimagining ethnopharmacology with generative AI," *Pharmacological Research*, 2025, PMID: 41120019.
- [40] Natural Resources for Human Health, "Editorial Template," *SciSpace*, 2026.