Multimodal Molecular Intelligence for Natural Product Drug Discovery: Integrating Chemical Structures, Bioassays, Omics Profiles, Text Evidence, and Clinical Signals

Olivia Bennett[1]*, Harry Collins[2], Jack Foster[1]

[1]Department of AI for Respiratory and Anti-asthmatic Herbal Formulations, Faculty of Pharmacy, University of Nottingham, Nottingham, United Kingdom
[2]Department of Computational Pharmacology for Histamine Receptors, Faculty of Pharmacy, University of Southampton, Southampton, United Kingdom.

ABSTRACT

Natural product drug discovery increasingly draws on heterogeneous evidence spanning chemical structures, biological assays, molecular profiles, published literature, and clinical observations, yet these evidence streams are commonly stored, modelled, and interpreted in isolation. This article proposes a multimodal molecular intelligence framework for organizing and integrating such evidence without treating computational convergence as proof of mechanism, efficacy, safety, or clinical utility. The framework begins with traceable natural product identity and botanical-source records, followed by complementary representations of chemical structures, stereochemistry, molecular descriptors, fingerprints, and molecular graphs. Bioassay evidence is incorporated with assay metadata, exposure context, target annotations, and explicit confidence boundaries. Transcriptomic, proteomic, metabolomic, genomic, microbiomic, pathway, and network evidence are positioned as context-dependent sources of mechanistic hypotheses. Literature mining, named entity recognition, relation extraction, knowledge graphs, and citation traceability provide structured text evidence while preserving source provenance. Clinical literature, real-world observations, pharmacovigilance reports, adverse-event information, ADMET evidence, and toxicity signals contribute translational and safety context but are not treated as causal confirmation. The proposed architecture combines modality alignment, cross-modal consistency assessment, conflict detection, missing-data handling, uncertainty estimation, bias assessment, expert review, candidate-prioritization boundaries, validation planning, and iterative feedback. Its principal contribution is a provenance-preserving and validation-gated framework that supports research prioritization while maintaining explicit no-clinical-claim boundaries.

Key Words: *Multimodal AI, Molecular intelligence, Natural product drug discovery, Chemical structures, Bioassays, Omics profiles*

eIJPPR 2026; 16(2):85-105

HOW TO CITE THIS ARTICLE: Bennett O, Collins H, Foster J. Multimodal Molecular Intelligence for Natural Product Drug Discovery: Integrating Chemical Structures, Bioassays, Omics Profiles, Text Evidence, and Clinical Signals. Int J Pharm Phytopharmacol Res. 2026;16(2):85-105. https://doi.org/10.51847/CFx9mQDN3R

INTRODUCTION

Natural product drug discovery remains a major source of chemically diverse scaffolds, pharmacologically active compounds, and therapeutic inspiration. Its scientific value, however, is accompanied by substantial evidentiary complexity arising from uncertain botanical identities, variable preparations, structurally related constituents, incomplete stereochemical annotations, heterogeneous assays, and uneven translational documentation. Natural products and their derivatives have contributed meaningfully to drug discovery, but natural origin alone

Corresponding author: Olivia Bennett
Address: Department of AI for Respiratory and Anti-asthmatic Herbal Formulations, Faculty of Pharmacy, University of Nottingham, Nottingham, United Kingdom.
E-mail: ✉ olivia.bennett@nottingham.ac.uk
Received: 26 January 2026; **Revised:** 05 April 2026; **Accepted:** 09 April 2026

does not establish pharmacological relevance, safety, or therapeutic potential [1, 2]. A rigorous discovery process must therefore distinguish the existence of a natural compound from the quality of evidence supporting its biological activity and possible development value.

Modern discovery programmes generate information across chemical structures, molecular descriptors, fingerprints, molecular graphs, bioassays, target annotations, transcriptomics, proteomics, metabolomics, pathway analyses, literature corpora, clinical studies, and safety-surveillance systems. Machine learning can assist molecular-property prediction, target prioritization, screening, biomarker analysis, and development planning, yet its usefulness depends on the appropriateness of the data, endpoints, validation strategy, and intended decision context [3]. When these evidence sources remain disconnected, potentially informative relationships may be overlooked; when they are combined without adequate controls, incompatible observations may be mistaken for mutually reinforcing evidence.

Multimodal integration therefore creates both an opportunity and a methodological risk. It may improve evidence organization, reveal cross-modal patterns, highlight contradictions, and support the prioritization of testable hypotheses. However, artificial intelligence does not independently establish target validity, mechanism, therapeutic efficacy, safety, or clinical utility, particularly when its outputs depend on incomplete, biased, heterogeneous, or proxy-labelled datasets [4]. Provenance gaps, inconsistent identifiers, assay-context loss, missing modalities, publication bias, distribution shift, uncertain clinical coding, and uncalibrated confidence may produce an appearance of evidentiary coherence that is not supported by the underlying observations.

This article proposes an original multimodal molecular intelligence framework for natural product drug discovery. The framework connects natural product identity, botanical provenance, chemical representations, bioassay evidence, omics profiles, pathway and network information, text-mined literature, clinical signals, pharmacovigilance evidence, ADMET and toxicity information, and multidisciplinary expert interpretation. Its central premise is that multimodal AI should support evidence organization, mechanistic hypothesis formation, safety-aware prioritization, and validation planning while preserving modality-specific limitations, source traceability, uncertainty, conflicts, and explicit boundaries between research prioritization and clinical claims.

Multimodal molecular intelligence

Multimodal molecular intelligence can be defined as the structured computational integration of heterogeneous chemical, biological, textual, and clinical evidence for a clearly bounded discovery question. It differs from single-modality machine learning because its purpose is not merely to predict an endpoint from one feature space. It also differs from simple database aggregation, which co-locates records without necessarily aligning their meanings, contexts, quality levels, or evidentiary relationships. Multimodal integration may involve early fusion of harmonized features, late fusion of modality-specific outputs, intermediate representation fusion, graph-based integration, or coordinated embeddings, but each strategy makes distinct assumptions about comparability, missingness, and information sharing [5, 6].

Early fusion combines modality-derived variables before model fitting and may be useful when evidence objects are well aligned, sufficiently complete, and measured at compatible levels. Its limitations include dimensional imbalance, scale differences, sparse modalities, and the possibility that high-volume data streams dominate lower-dimensional but more reliable evidence. Late fusion preserves separate modality-specific analyses and combines their outputs after individual evaluation, making it easier to inspect modality contributions but potentially limiting interaction discovery. Joint embeddings and cross-modal representations can capture relationships among molecular, omics, pathological, textual, or clinical features, although their performance and interpretability remain dependent on the population, task, data-generating process, and validation design [7].

Multimodal molecular intelligence should also be distinguished from network pharmacology alone, omics analysis alone, and text mining alone. Network pharmacology can organize compound–target–pathway relationships but may inherit unverified database edges and centrality biases. Omics analysis can identify response signatures and enriched pathways but cannot independently determine whether observed changes are causal, specific, or therapeutically favourable. Text mining can retrieve entities, relations, and reported mechanisms but cannot convert published assertions into validated biological facts. A credible integration architecture must preserve modality identity, experimental context, data lineage, and evidentiary independence rather than merging every available observation into a single undifferentiated score [8].

The proposed concept therefore treats each modality as a bounded evidence channel with its own representation, validation procedure, uncertainty estimate, and permitted interpretation. Modality alignment should verify that records refer to compatible compounds, organisms, preparations, targets, tissues, doses, time points, phenotypes, and outcomes. Missing data should remain visible rather than being interpreted automatically as negative evidence. Cross-modal agreement should increase

the priority of a hypothesis only when the supporting evidence is contextually compatible and sufficiently independent, whereas disagreement should generate a conflict record requiring expert review or additional experimentation. Source provenance, bias assessment, expert interpretation, and no-clinical-claim controls are consequently structural requirements rather than optional reporting additions.

Chemical structures and bioassays

Chemical evidence begins with reliable identification of the natural product object under investigation. A record should distinguish an isolated compound from a standardized preparation, complex extract, fraction, mixture, salt, derivative, or incompletely characterized material. Natural product databases vary in coverage, curation practices, identifiers, structural completeness, biological annotations, and source documentation, making cross-database reconciliation necessary before evidence integration [9]. Aggregated resources can improve access to natural product structures, but their records may inherit inconsistencies from upstream databases; consequently, canonical identifiers, source links, version information, alternative structures, and normalization histories should be retained [10]. Botanical or microbial provenance should include taxonomic identity, organism part or strain, extraction and preparation context, and voucher or source information where available.

Chemical structure representation should be plural rather than singular. Molecular descriptors provide interpretable physicochemical properties, fingerprints encode predefined structural neighbourhoods, and molecular graphs capture atom–bond topology for representation learning. Stereochemistry requires separate attention because enantiomers, diastereomers, and geometrical isomers may differ in binding, metabolism, selectivity, and toxicity. Computational assignment may help identify unresolved stereochemical relationships, but predicted configurations should remain explicitly labelled as uncertain until supported by appropriate structural evidence [11]. Molecular benchmarks further indicate that representation quality must be assessed in relation to the endpoint, dataset composition, evaluation metric, and splitting strategy rather than assumed from model complexity alone [12].

Bioassay evidence should not be reduced to a compound–activity label. Each observation requires the assay type, biological system, species, cell line or tissue, endpoint, units, exposure concentration, duration, controls, curve quality, replication status, laboratory source, and experimental conditions. Target-prediction performance can depend strongly on assay-derived labels, class imbalance, endpoint heterogeneity, and evaluation design, so predicted target associations should be separated from direct binding evidence and functional target confirmation [13]. Similarly, no single molecular representation performs optimally across all activity and property-prediction tasks; comparisons among descriptors, fingerprints, graphs, and learned representations should therefore use scaffold-aware or otherwise task-appropriate validation [14].

Compound–target evidence should be represented as a traceable relationship connecting a verified compound identity, a normalized target, a defined assay, an evidence type, and a confidence category. Graph-based models may support structure–property and structure–activity learning by encoding atomic and bond relationships, but their outputs remain dataset-dependent predictions and should not be interpreted as mechanistic explanations merely because attention or feature-attribution values are available [15]. ADMET and toxicity information should integrate experimental evidence, exposure context, structural alerts, metabolism hypotheses, and uncertainty without equating an absence of recorded toxicity with safety. Within the proposed framework, chemical and bioassay modalities provide evidence for research prioritization and validation planning; they do not independently establish target validity, drug potential, safety, or translational readiness.

Table 1 summarises chemical structure and bioassay components required for multimodal molecular intelligence in natural product drug discovery.

Table 1. Chemical Structure and Bioassay Components for Multimodal Molecular Intelligence in Natural Product Drug Discovery: Natural Product Identity, Botanical Source, Stereochemistry, Molecular Descriptors, Molecular Graphs, Bioassay Evidence, Assay Metadata, Bioactivity Context, Target Annotation, ADMET Evidence, and Validation Needs

Data component	Core drug discovery question	Required input	AI or integration function	Potential output	Reliability concern	Validation requirement	Framework role
Natural product identity	Is the evidence linked to a distinct and traceable natural	Stable identifier, preferred name, synonyms, preparation	Entity resolution, deduplication, and synonym reconciliation	Canonical identity record with linked alternatives	Extracts, mixtures, salts, derivatives, and isolated compounds	Manual adjudication of ambiguous identities and reconciliation	Root evidence object for downstream integration

	product object?	type, source record, and version			may be conflated	across sources	
Botanical source	Which organism, tissue, strain, or preparation produced the material?	Taxonomic name, organism part or strain, geographical context, voucher information where available, and preparation details	Taxonomic normalization and source linking	Botanical or microbial provenance profile	Misidentification, outdated taxonomy, absent voucher records, and batch variability	Taxonomic verification and source-document review	Prevents transfer of evidence between non-equivalent sources
Chemical structure	Which molecular entity is represented?	Standardized connection table, charge, tautomer, isotope, salt status, and mixture status	Structure standardization and representation generation	Canonical structure with preserved alternatives	Incorrect records, unresolved mixtures, and normalization-induced information loss	Independent structure review and comparison with primary sources	Primary chemical evidence modality
Stereochemistry	Are chiral centres and geometric configurations resolved?	Isomeric structure, stereochemical annotations, evidence source, and confidence status	Stereochemical encoding or prediction	Resolved, partial, or uncertain stereochemical record	Missing or incorrectly inferred configurations	Spectroscopic, synthetic, crystallographic, or other appropriate confirmation when decision-relevant	Mandatory uncertainty-bearing structural attribute
Molecular descriptor	Which interpretable physicochemical characteristics describe the compound?	Standardized structure and documented calculation settings	Descriptor calculation, normalization, and feature selection	Interpretable descriptor vector	Software dependence, correlated variables, and invalid calculations	Recalculation, range checks, and cross-tool comparison	Interpretable representation and baseline modelling channel
Molecular graph	How are atoms, bonds, charges, and local environments connected?	Atom features, bond features, aromaticity, formal charge, and stereochemical information	Graph representation learning	Molecular embedding or task-specific prediction	Structural graphs omit preparation, exposure, assay, and biological context	External evaluation, scaffold-aware testing, and applicability-domain assessment	Structure-learning channel
Molecular fingerprint	Which predefined substructures or neighbourhoo	Standardized structure, fingerprint type, radius,	Similarity search, clustering, and conventional	Similarity profile or predictive feature vector	Hash collisions and representation-specific blind spots	Comparison across fingerprint types and external	Retrieval, baseline prediction, and chemical-

	ds are present?	and bit settings	machine learning			chemical-space testing	space mapping
Bioassay evidence	Has the material produced a measured biological response?	Endpoint, value, units, concentration, protocol, controls, replication, and source	Assay harmonization and evidence grading	Context-qualified activity record	Heterogeneous protocols, selective reporting, and incomparable endpoints	Replication, control review, unit verification, and protocol appraisal	Empirical biological evidence
Assay metadata	Under what experimental conditions was the observation generated?	Assay type, organism, cell line, tissue, readout, duration, laboratory, plate, and batch information	Metadata normalization and assay grouping	Assay-context profile	Missing context may make activity records uninterpretable or non-comparable	Minimum metadata completeness gate	Preserves experimental meaning
Bioactivity context	Is the measured activity relevant to plausible exposure, selectivity, and biological conditions?	Potency, efficacy, dose, duration, exposure estimate, comparator, matrix, and selectivity data	Cross-assay contextualization and exposure-aware interpretation	Qualified bioactivity hypothesis	Activity may occur only at implausible concentrations or under artificial conditions	Orthogonal assays, concentration–response testing, and exposure-relevance assessment	Distinguishes measured signals from development-relevant evidence
Target annotation	Which molecular target is associated with the observation?	Standardized target identifier, species, isoform, assay modality, and evidence directness	Target normalization and relationship aggregation	Target-evidence record	Predicted, indirect, and directly measured targets may be conflated	Direct binding or functional confirmation when required	Connects chemical and biological evidence
Target confidence	How strong, reproducible, and direct is the compound–target relationship?	Evidence type, assay quality, replication, selectivity, source independence, and uncertainty	Evidence grading and confidence calibration	Traceable target-confidence category	Aggregate confidence may conceal heterogeneous or dependent evidence	Predefined grading rubric and multidisciplinary review	Prevents target prediction from being described as target validation
Compound–target evidence	Are compound identity, assay findings, and target annotations mutually compatible?	Verified compound record, assay evidence, target identity, provenance, and	Cross-modal matching, consistency assessment, and conflict detection	Supported, conflicting, or unresolved compound–target hypothesis	Entity mismatch, promiscuity, assay artefacts, and publication bias	Orthogonal assays, identity audit, and independent replication	Core cross-modal evidence unit

		contextual metadata					
ADMET evidence	Are absorption, distribution, metabolism, excretion, and transport characteristics plausible?	Experimental or predicted ADMET endpoints, biological system, exposure conditions, and method metadata	Property modelling and evidence aggregation	ADMET hypothesis profile with limitations	In vitro and in silico endpoints may not translate across species or exposure settings	Relevant experimental confirmation and applicability assessment	Development-risk prioritization and test planning
Toxicity evidence	Are there structural, experimental, metabolic, or observed safety concerns?	Cytotoxicity, organ-toxicity evidence, structural alerts, reactive-metabolite data, dose, and exposure context	Safety-evidence integration and alert generation	Toxicity concern profile	Alerts may be nonspecific, labels may be incomplete, and unreported toxicity may appear absent	Dose-aware experimental assessment, mechanistic follow-up, and expert toxicology review	Exclusion, caution, and validation-planning input

Omics profiles and text evidence

Omics evidence can extend natural product investigation beyond isolated compound–target associations by characterizing biological responses across molecular layers. Transcriptomic profiles may identify genes and cellular programmes that change after exposure, proteomics may reveal altered proteins or signalling states, and metabolomics may connect exposure with metabolic pathways, biotransformation products, or downstream physiological responses. In natural product research, these modalities can support target discovery and pathway-hypothesis generation when compound identity, dose, time, tissue, model system, and analytical platform are adequately documented [16]. Such profiles nevertheless remain context-dependent observations: a differentially expressed gene, altered protein, or enriched metabolite set does not independently prove direct target engagement or a therapeutically relevant mechanism.

Single-cell and cell-resolved multi-omics approaches may provide additional information about heterogeneous responses that are obscured in bulk measurements. These methods can help distinguish cell-type-specific effects, identify rare responding populations, and examine whether apparent pathway activity originates from a restricted biological compartment. Their use in natural product research may be particularly valuable when extracts or compounds produce divergent responses across immune, stromal, epithelial, or disease-associated cellular populations [17]. However, cell isolation procedures, batch effects, sequencing depth, temporal sampling, tissue dissociation, and population-composition differences can alter the resulting profiles. The framework therefore treats cell-resolved signatures as qualified mechanistic hypotheses whose relevance depends on replication, orthogonal confirmation, and compatibility with exposure and assay evidence.

Integrated analysis can identify variation shared across omics layers while preserving modality-specific signals. Latent-factor approaches, for example, may organize coordinated transcriptomic, proteomic, or metabolomic variation and accommodate partially observed data without requiring all modalities to contain identical information [18]. These representations are useful for identifying patterns, patient or sample groupings, and candidate biological processes, but latent factors remain mathematical summaries rather than validated mechanisms. More broadly, multi-omics interpretation requires alignment of tissue, sampling time, disease state, analytical platform, cohort composition, and preprocessing because apparent agreement across modalities can arise from correlated technical effects or shared confounding rather than biological convergence [19]. Cross-modal validation should therefore test whether a hypothesis suggested in one layer is supported by an independently generated and biologically appropriate layer.

Genomic, transcriptomic, proteomic, and metabolomic evidence may also support the investigation of biosynthetic pathways and natural product origins, particularly for

microbial secondary metabolites [20]. Genomics and transcriptomics can suggest biosynthetic gene clusters or expression programmes, whereas metabolomics may help connect those programmes with candidate products; experimental confirmation remains necessary when pathway assignments or compound identities are uncertain. Text evidence provides a complementary means of organizing published observations. Biomedical language models can support named entity recognition and relation extraction for chemicals, genes, proteins, diseases, and biological effects, but extracted relations remain source-dependent assertions rather than validated facts [21]. Automated annotation resources can improve retrieval and normalization while retaining links to the source article [22]. Within the proposed framework, literature mining, relation extraction, and knowledge graphs must preserve DOI-level and passage-level provenance, distinguish primary from secondary evidence, capture negation and speculation where possible, and expose contradictory reports for expert assessment.

Table 2 maps omics profile and text evidence modalities that can support multimodal natural product drug discovery.

Table 2. Omics Profiles and Text Evidence for Multimodal Natural Product Drug Discovery: Transcriptomics, Proteomics, Metabolomics, Genomics, Microbiomics, Pathway Enrichment, Network Pharmacology, Literature Mining, Named Entity Recognition, Relation Extraction, Knowledge Graphs, and Citation Traceability

Evidence modality	Core biological question	Required input	Computational function	Potential output	Interpretation risk	Validation requirement	Framework role
Transcriptomics	Which genes and cellular programmes change after exposure or between relevant biological states?	Raw or processed expression data, tissue or cell type, dose, duration, controls, batch, and platform metadata	Differential expression, signature matching, clustering, and latent representation	Context-qualified response signature or pathway hypothesis	Batch effects, cell-composition differences, nonspecific stress responses, and temporal ambiguity	Independent experiment or cohort, quantitative confirmation, and dose–time assessment	Hypothesis-generating biological-response modality
Proteomics	Which proteins, complexes, or post-translational states change in the relevant system?	Protein or peptide measurements, platform, sample preparation, controls, and identification confidence	Differential abundance, module detection, interaction mapping, and pathway analysis	Protein-response profile or candidate functional module	Incomplete coverage, peptide ambiguity, abundance–activity mismatch, and platform bias	Targeted protein measurement, functional assay, or orthogonal confirmation	Adds functional proximity without independently validating mechanism
Metabolomics	Which metabolic pathways, exposure markers, or biotransformation products are altered?	Spectral features, retention information, metabolite annotations, standards, matrix, dose, and time	Feature annotation, pathway mapping, and exposure-response analysis	Metabolic signature or biotransformation hypothesis	Unknown-feature misannotation, ion suppression, matrix effects, dietary confounding, and microbial contributions	Authentic standards, targeted assays, replication, and exposure-aware interpretation	Connects compound exposure to metabolic response
Genomics where supported	Do inherited, acquired, or biosynthetic genomic features	Variant, copy-number, genome-assembly,	Variant annotation, genomic-context analysis, and	Genomic-context or biosynthetic-origin hypothesis	Population structure, multiple testing, assembly	Independent replication, functional testing, or	Conditional context modality rather than universal

	modify target or response context?	or biosynthetic-cluster data with sample metadata	biosynthetic inference		errors, and weak causal inference	biosynthetic confirmation	mechanism evidence
Microbiomics where supported	Could microbial composition or metabolism modify natural product exposure or response?	Microbial sequencing or functional data, host metadata, sampling context, diet, geography, and contamination controls	Community profiling, functional inference, and host–microbe association analysis	Microbiome-associated metabolism or response hypothesis	Compositionality, contamination, geographic effects, dietary confounding, and reverse causation	Longitudinal or perturbational study and metabolite-level confirmation	Optional contextual modality when biologically justified
Pathway enrichment	Which predefined biological processes are statistically overrepresented?	Ranked genes, proteins, or metabolites; pathway database, version, and background set	Overrepresentation, gene-set enrichment, or topology-aware analysis	Ranked pathway hypotheses	Database bias, overlapping pathways, arbitrary thresholds, and circular interpretation	Multiple methods, sensitivity analysis, permutation controls, and orthogonal evidence	Organizes molecular responses without establishing mechanism
Network pharmacology	How might compounds, targets, pathways, and phenotypes be connected?	Curated compound–target evidence, interaction data, pathway annotations, disease associations, and provenance	Network construction, module detection, traversal, and centrality analysis	Exploratory compound–target–pathway map	Predicted edges, database popularity, circular reuse of sources, and centrality inflation	Edge-level evidence grading, source-independence assessment, and experimental perturbation	Systems-level hypothesis organization
Literature mining	What published evidence addresses a defined compound, target, pathway, phenotype, or safety question?	Reproducible search strategy, peer-reviewed sources, identifiers, and inclusion logic	Retrieval, relevance ranking, evidence classification, and passage extraction	Source-linked evidence set	Retrieval bias, publication bias, context loss, and overreliance on abstracts	Reproducible retrieval, primary-source verification, and expert appraisal	Evidence discovery and organization rather than truth adjudication
Named entity recognition	Which chemicals, species, genes, proteins, diseases,	Full text or abstracts, dictionaries, ontologies, and	Entity detection, normalization, and synonym resolution	Candidate entity mentions linked to standardized identifiers	Abbreviations, nested entities, taxonomic ambiguity, and	Manual validation of decision-relevant entities and	Populates auditable evidence objects

	outcomes, and preparations are mentioned?	annotated corpora			preparation–compound conflation	error analysis	
Relation extraction	What relationship is asserted between identified entities?	Source sentences or passages containing normalized entities	Rule-based, statistical, or neural relation extraction	Candidate compound–target, compound–effect, or safety relation	Negation, speculation, temporality, dose, study design, and directionality may be lost	Sentence-level review, evidence-type classification, and primary-source verification	Generates reviewable relation hypotheses
Knowledge graph evidence	Which typed and traceable relations connect products, compounds, targets, pathways, diseases, assays, and safety events?	Normalized entities, typed edges, source identifiers, timestamps, and confidence attributes	Graph integration, query, traversal, relation aggregation, and link prediction	Provenance-rich evidence subgraph	False entity resolution, predicted-edge inflation, and hidden source dependence	Relation-level provenance, confidence separation, and expert review of decision-relevant paths	Cross-modal organizational and retrieval layer
Text-mined mechanism evidence	Does the literature describe a plausible sequence from exposure to target, pathway, and phenotype?	Source-linked entities, relations, directionality, study design, dose, and model context	Multi-sentence evidence assembly and argument mapping	Mechanistic hypothesis with supporting passages	Narrative coherence may be mistaken for causal validation	Experimental-design appraisal, source independence, and orthogonal corroboration	Mechanism-hypothesis support only
Citation traceability	Can every text-derived assertion be traced to an identifiable source and passage?	DOI, article identifier, section, passage, extraction method, and timestamp	Provenance attachment, versioning, and audit logging	Citation-linked evidence object	Secondary citations may be mistaken for primary evidence, and corrected articles may remain in pipelines	Primary-source verification, version checks, and correction monitoring	Mandatory safeguard for text-derived evidence
Conflicting evidence detection	Do sources or modalities disagree on identity, target, effect direction, pathway, or safety?	Normalized claims with direction, context, source, confidence, and evidence type	Contradiction detection, contextual comparison, and conflict clustering	Conflict record requiring adjudication	Apparent contradictions may arise from different doses, tissues, preparations, or time points	Contextual expert review and targeted experimental resolution	Prevents forced consensus and reduces overconfident synthesis

Clinical signals and evidence integration

Clinical signals occupy a fundamentally different evidentiary position from molecular structures, laboratory assays, and omics profiles. A clinical signal may arise from a controlled study, observational cohort, electronic health record, spontaneous report, case series, or published clinical narrative, and each source has distinct limitations regarding population selection, exposure definition,

ISSN (Online) 2249-6084 (Print) 2250-1029

outcome ascertainment, denominators, confounding, and missingness. Pharmacovigilance is designed to detect and assess possible medicine-related harms, but a detected signal is an invitation to investigate rather than a definitive causal conclusion [23]. Within natural product research, this distinction is particularly important because product identity, formulation, co-exposures, contamination, variable composition, and incomplete documentation can complicate attribution.

Artificial intelligence may assist pharmacovigilance by supporting case intake, duplicate detection, coding, information extraction, signal triage, and pattern identification. Such automation requires validated workflows, transparent governance, and sustained human oversight because errors can be propagated rapidly when source records are incomplete or ambiguous [24]. Deep-learning approaches may also support adverse-event prediction or drug–event association analysis, but their outputs are shaped by the composition of the training data, reporting practices, label definitions, and representation of exposure and patient context [25]. These methods can therefore generate safety hypotheses or review priorities, but they do not establish individual risk, incidence, causality, or clinical actionability.

Clinical text can provide information that is absent from structured fields, including symptom timing, suspected exposure, co-medication, dechallenge, rechallenge, and clinician interpretation. Natural language processing has been used to identify adverse-event information in electronic health records, although portability across institutions remains difficult because documentation practices, coding systems, patient populations, and annotation standards differ [26]. Real-world data may complement clinical evidence by describing routine-care populations and outcomes under defined analytical conditions, but observational estimates remain vulnerable to confounding, selection bias, measurement error, missing exposure details, and incomplete outcome capture [27]. Consequently, clinical and real-world signals should enter the framework with explicit study-design metadata, temporal information, source quality, and uncertainty rather than being converted directly into a generalized effectiveness or safety score.

Traditional use may provide historical exposure context, preparation information, or hypotheses regarding possible effects, but duration of use is not equivalent to proof of efficacy or absence of harm. Herb-related adverse outcomes illustrate the need for verified product identity, exclusion of alternative causes, dose and temporal assessment, evaluation of co-medications, and structured causality review [28]. Cross-modal integration should therefore examine whether chemical identity, assay findings, omics responses, literature assertions, and clinical observations refer to compatible products, doses, biological contexts, and effect directions. Agreement among genuinely independent and compatible sources may justify higher research priority, whereas conflict, missingness, uncertain provenance, or weak clinical attribution should lower confidence and trigger further review. Expert clinical, pharmacological, toxicological, botanical, and methodological judgment is required before any translational interpretation, and candidate prioritization must remain separated from therapeutic recommendation.

Table 3 summarises clinical signal and evidence integration domains required for cautious translational interpretation in multimodal natural product drug discovery.

Table 3. Clinical Signals and Evidence Integration in Multimodal Natural Product Drug Discovery: Pharmacovigilance Signals, Real-World Evidence, Safety Signals, Adverse-Event Evidence, Clinical Literature, Cross-Modal Consistency, Conflict Detection, Expert Review, Uncertainty, and Decision Boundaries

Integration domain	Core translational question	Evidence needed	Integration function	Potential output	Risk of overinterpretation	Validation requirement	Decision boundary
Clinical signal	Is there an observed patient-level pattern relevant to exposure, response, or harm?	Defined population, product or compound identity, exposure, comparator, outcome, timing, and data source	Signal detection, temporal linkage, and contextual evidence mapping	Clinical research signal	Association may be described as treatment effect, safety, or causality	Epidemiological appraisal, replication, confounder assessment, and intervention-identity verification	Does not establish efficacy, safety, causality, or clinical utility alone
Pharmacovigilance signal	Is an adverse-event pattern sufficiently	Spontaneous reports, case narratives,	Disproportionality analysis, case	Safety-review priority	Reporting bias, stimulated reporting,	Case validation, product-identity review, and	Signal detection is not causality determination

	unusual or important to justify investigation?	literature reports, duplicate status, seriousness, timing, and exposure estimates where available	aggregation, NLP, and signal triage		duplicates, missing denominators, and notoriety effects	structured signal assessment	
Real-world evidence where supported	Does routine-care evidence inform effectiveness or safety under a prespecified design?	Fit-for-purpose records, registries, claims, patient-generated data, exposure definitions, outcomes, covariates, and follow-up	Cohort construction, adjustment, outcome analysis, and sensitivity testing	Context-bound effectiveness or safety estimate	Residual confounding, selection bias, misclassification, missingness, and data drift	Prespecified protocol, robustness analyses, negative controls where appropriate, and external assessment	Cannot automatically substitute for prospective or randomized validation
Safety signal	Is there convergent evidence suggesting a possible hazard?	Structural alerts, experimental toxicity, metabolites, clinical cases, dose, route, and exposure context	Cross-source safety aggregation and alerting	Safety concern requiring investigation	Repeated dependent evidence may appear independent, and nonspecific alerts may appear definitive	Source-dependence assessment, exposure relevance, replication, and expert toxicology review	Supports caution or testing rather than a definitive safety label
Adverse-event evidence	What adverse outcome was observed, under what exposure, and with what alternatives?	Event definition, seriousness, onset, duration, dechallenge, rechallenge, co-medication, comorbidity, and product details	Event normalization, temporal analysis, and alternative-cause mapping	Structured adverse-event record	Attribution may ignore contamination, substitution, underlying disease, or concurrent treatment	Clinical case review and structured causality assessment	No automatic attribution to a natural product or constituent
Traditional use context where supported	Does documented traditional use inform exposure history, preparation, or hypothesis formation?	Preparation, route, dose, duration, population, indication, and source quality	Contextual evidence mapping	Traditional-use context record	Historical use may be misrepresented as modern efficacy or safety evidence	Product-composition verification and contemporary pharmacological and safety evaluation	Supports context and hypothesis generation only
Clinical literature evidence	What peer-reviewed clinical	Study design, intervention	Study classification, intervention	Design-qualified clinical	Claims may be transferred across non-equivalent	Intervention-identity audit, study appraisal,	No cross-product or cross-

	evidence exists for the product, preparation, extract, or constituent?	identity, comparator, population, dose, outcomes, duration, and risk-of-bias information	matching, and evidence extraction	evidence profile	extracts, preparations, or isolated compounds	and replication where relevant	preparation therapeutic inference
Evidence triangulation	Do independently generated evidence streams support the same bounded hypothesis?	Chemical, bioassay, omics, textual, and clinical evidence with explicit provenance	Structured convergence assessment	Triangulated research hypothesis	Shared databases, duplicated publications, or circular citations may create false independence	Source-dependence analysis, compatibility checks, and expert adjudication	Convergence prioritizes testing but does not prove the hypothesis
Cross-modal consistency	Are identity, direction, dose, target, model, tissue, time, and outcome compatible?	Harmonized entities and contextual metadata across modalities	Semantic, temporal, and biological consistency checking	Compatible evidence bundle	Superficial agreement may conceal incompatible experiments or populations	Context-level comparison and independent verification of decisive links	Only compatible evidence contributes to convergence
Cross-modal conflict detection	Where and why do modalities disagree?	Signed assertions, effect direction, experimental context, source, and uncertainty	Contradiction localization, context comparison, and conflict classification	Conflict map with possible explanations	Automated systems may suppress negative or minority evidence	Expert review and targeted experimental or clinical clarification	Unresolved conflict lowers confidence and blocks strong interpretation
Uncertainty estimation	How uncertain are individual signals and the integrated interpretation?	Measurement error, sampling uncertainty, model uncertainty, missingness, extraction confidence, and provenance quality	Uncertainty decomposition, calibration, and propagation	Modality-specific and integrated uncertainty profile	A single confidence score may conceal severe uncertainty in one critical source	Calibration, sensitivity analysis, and transparent component reporting	High uncertainty restricts prioritization and requires further evidence
Bias assessment	Could sampling, reporting, measurement, publication, or model bias explain the apparent signal?	Population composition, data-generation process, missingness, source selection, and model-development details	Structured bias appraisal and stress testing	Bias-risk profile	Apparent cross-modal consistency may reflect shared systematic bias	Independent review, subgroup assessment, and external validation	High bias risk prevents strong translational interpretation
Expert clinical or pharmacology review	Are integrated outputs chemically, biologically,	Evidence graph, source passages, conflicts,	Multidisciplinary adjudication	Accepted, revised, deferred, or	Automation bias, selective acceptance, and	Independent reviewers, documented rationale, and	Human review is mandatory before

	and clinically coherent?	uncertainty, dose, exposure, and intended use		rejected interpretation	disciplinary blind spots	conflict-of-interest disclosure	translational interpretation
Candidate prioritization boundary	Which bounded hypotheses merit additional laboratory or translational testing?	Evidence quality, uncertainty, feasibility, novelty, safety concerns, and validation burden	Multi-criteria prioritization and sensitivity analysis	Research-priority tier with rationale	Rank order may be mistaken for therapeutic superiority or probability of success	Alternative weighting, committee review, and prespecified advancement gates	Supports resource allocation only
No clinical-claim boundary	Which outputs are prohibited without direct clinical validation?	Intended-use statement, validation status, claim taxonomy, and user role	Output restriction, labelling, and escalation	Research-only output with explicit limitations	Suggestive wording or visual design may imply clinical actionability	Governance, editorial review, and interface testing	No diagnosis, prescribing, treatment, efficacy, or safety recommendation

Figure 1 illustrates how chemical structures, bioassays, omics profiles, text evidence, and clinical signals can be integrated into a multimodal evidence logic for natural product drug discovery.

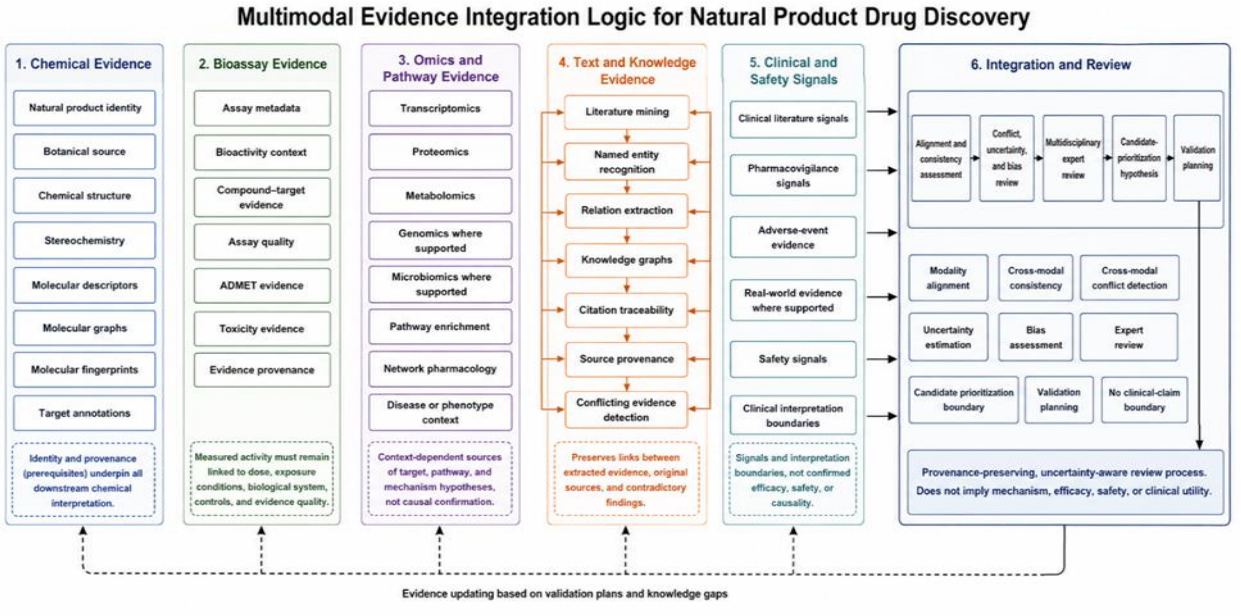

Figure 1. Multimodal Evidence Integration Logic for Natural Product Drug Discovery

Alt-text description

A conceptual evidence integration diagram showing chemical structures, bioassays, omics profiles, text evidence, clinical signals, pathway knowledge, safety evidence, provenance, uncertainty, expert review, and candidate prioritization boundaries.

Proposed multimodal framework

The proposed framework begins with a versioned natural product data foundation. Each evidence object should include natural product identity, botanical or microbial source, preparation type, structural status, stereochemical completeness, assay links, omics context, literature provenance, clinical-signal source, safety information, and transformation history. Dataset documentation principles support the recording of composition, collection processes,

intended use, exclusions, known limitations, maintenance, and potential sources of bias [29]. Applied here, these principles require every compound, assay, omics profile, extracted relation, and clinical record to retain its original source and subsequent processing steps. Records with unresolved identity, unclear provenance, or non-equivalent preparations should not be fused merely because they share a common name.

The second stage comprises modality-specific representation and quality control. Chemical structures may be represented simultaneously through descriptors, fingerprints, graphs, and stereochemical annotations; bioassays should be encoded with endpoints and experimental context; omics profiles require platform, tissue, dose, time, batch, and cohort metadata; text evidence should retain source passages and extraction confidence; and clinical signals require population, exposure, outcome, temporality, and study-design information. Missingness should be represented explicitly because absence of a measurement, missing documentation, an untested endpoint, and a true negative result are not equivalent states. Reviews of machine-learning prediction studies have shown that missing data are often handled or reported inadequately, creating avoidable risks of biased estimates and irreproducible conclusions [30]. The proposed system therefore assigns a typed missingness status to each unavailable field and requires sensitivity analysis when missingness can influence integration or prioritization.

The third stage performs modality alignment, fusion, and uncertainty-aware integration. Entity alignment reconciles compounds, targets, species, tissues, preparations, phenotypes, and clinical outcomes; contextual alignment compares dose, time, biological system, and direction of effect. Early fusion may be used only when inputs are sufficiently harmonized, whereas late fusion may be preferable when modality-specific models require independent validation. Joint embeddings and cross-modal representations may support pattern discovery, but no integrated representation should erase the source, confidence, or validation status of its constituent evidence. Uncertainty should be distinguished across measurement noise, sampling variability, model uncertainty, extraction uncertainty, and distribution shift. Methods for uncertainty estimation can support calibration and error communication, but their reliability depends on the estimation strategy and the data distribution in which they are evaluated [31]. The proposed framework therefore preserves modality-level uncertainty and propagates it to integrated outputs without collapsing it into an unexplained confidence score.

The fourth stage applies cross-modal validation, bias assessment, and transparent reporting. A hypothesis generated from chemical structure should be assessed against assay evidence; an omics-derived pathway hypothesis should be examined for target, dose, and temporal compatibility; a text-derived relation should be checked against the primary experiment; and a clinical signal should be evaluated for exposure identity, confounding, and alternative explanations. Studies of supervised machine-learning prediction models have identified recurring risks involving participant selection, predictor handling, outcome definition, analysis, overfitting, and weak external validation [32]. Transparent reporting should therefore specify intended use, data sources, preprocessing, model development, evaluation strategy, limitations, and the role of human reviewers, consistent with minimum-information principles for clinical AI modelling [33]. In this framework, any evidence chain that fails source-independence, identity, compatibility, bias, or validation checks is marked as unresolved rather than being forced into a positive or negative conclusion.

The fifth stage combines multidisciplinary expert review, candidate prioritization, validation planning, and feedback updating. Experts should review the evidence graph, modality-specific limitations, conflicts, uncertainty, safety concerns, and intended decision boundary before a candidate or mechanistic hypothesis is prioritized. Research outputs may include a candidate-priority tier, target-context hypothesis, mechanism hypothesis, safety-risk hypothesis, or recommended validation experiment, but these outputs must be accompanied by the evidence and uncertainty that produced them. Reviews of clinical AI studies have shown that weak design and reporting can support claims that exceed the underlying evidence [34]. The proposed architecture therefore includes an explicit no-clinical-claim boundary: it cannot generate diagnoses, treatment recommendations, efficacy conclusions, safety conclusions, or prescribing advice. New assays, omics results, literature corrections, negative findings, safety reports, and expert decisions should enter an append-only feedback process that updates the evidence state without erasing prior versions.

Table 4 presents the proposed multimodal molecular intelligence framework for natural product drug discovery.

Table 4. Proposed Multimodal Molecular Intelligence Framework for Natural Product Drug Discovery: Data Foundation, Chemical Structure Representation, Bioassay Integration, Omics Integration, Text Evidence Retrieval, Clinical Signal Integration, Provenance Mapping, Cross-Modal Validation, Expert Review, Candidate Prioritization, Feedback, and Claim Boundaries

Framework stage	Core purpose	Required input	Multimodal function	Potential output	Validation need	Failure risk	Feedback mechanism	Decision boundary
Natural product data foundation	Establish traceable product, organism, preparation, and compound identities	Taxonomy, source records, voucher or strain information where available, preparation details, structures, identifiers, and database versions	Cross-source entity resolution and versioned evidence-object construction	Auditable natural product record	Identity, taxonomy, preparation, and provenance audit	Conflation of species, extracts, mixtures, derivatives, and isolated compounds	Corrections propagate through linked evidence while preserving prior versions	No downstream fusion until minimum identity and provenance requirements are met
Chemical structure representation	Encode complementary structural information without discarding uncertainty	Standardized structures, stereochemistry, descriptors, fingerprints, graphs, salts, tautomers, and confidence status	Parallel chemical encoding and applicability-domain assessment	Structure-representation bundle	Independent structure and stereochemistry verification	Information loss, invalid normalization, and representation-specific bias	New structural evidence updates the bundle and affected predictions	Structure-derived outputs remain research hypotheses
Bioassay integration	Connect measured biological responses to explicit experimental conditions	Endpoint, protocol, biological system, dose, duration, controls, replication, units, and source	Assay normalization, evidence grading, and cross-assay comparison	Context-qualified bioactivity profile	Protocol review, replication, orthogonal testing, and exposure assessment	Incomparable endpoints, assay artefacts, and potency inflation	New experiments revise evidence grade and conflict status	No target-validation or efficacy claim from isolated assay evidence
Omics integration	Characterize biological responses across molecular layers	Transcriptomic, proteomic, metabolomic, genomic, or microbiomic profiles with context metadata	Latent-factor, pathway, network, and response-signature integration	Context-bound target, pathway, or response hypothesis	Batch, tissue, dose, time, cohort, and platform validation	Correlation, technical variation, or nonspecific stress interpreted as mechanism	Replication and perturbation studies update evidence and uncertainty	Omics signatures do not independently establish mechanism
Text evidence retrieval	Retrieve, normalize, and structure published claims	Reproducible searches, peer-reviewed sources, full text or abstracts,	NER, relation extraction, evidence classification, and	Citation-traceable evidence record	Primary-source verification and extraction-error review	Decontextualization, negation errors, duplication, and	Reviewer corrections and literature updates revise	Text-mined evidence cannot establish biological or

		identifiers, and passages	source linking			publication bias	extraction records	clinical validity alone
Clinical signal integration	Add bounded patient, effectiveness, and safety context	Clinical studies, EHRs, registries, reports, pharmacovigilance records, exposure identity, outcomes, and timing	Clinical-signal linkage, case aggregation, and contextual evidence integration	Clinical or safety research signal	Epidemiological review, causality assessment, and product verification	Confounding, reporting bias, missing denominators, and exposure misclassification	Adjudicated cases and new studies update signal status	No causal, effectiveness, safety, or treatment conclusion
Evidence provenance mapping	Preserve origin, transformation, version, and dependency of every evidence object	DOI, dataset, database, assay, model, preprocessing, extraction, timestamp, and reviewer metadata	Provenance-graph construction and dependency tracking	Auditable evidence lineage	Completeness, integrity, source-dependence, and version checks	Circular evidence, duplicated sources, and untraceable transformations	Corrections create append-only lineage updates	Evidence without adequate provenance is excluded or down-weighted
Modality alignment	Ensure that modalities refer to compatible entities and contexts	Normalized compounds, preparations, targets, organisms, tissues, doses, times, phenotypes, and outcomes	Semantic, unit, temporal, and biological alignment	Aligned multimodal evidence bundle	Identity and context verification	False fusion of non-equivalent objects or experiments	Alignment conflicts trigger remapping or separation	Unaligned evidence cannot contribute to integrated inference
Data fusion and cross-modal representation	Combine compatible evidence while retaining modality identity	Validated modality-specific representations and alignment metadata	Early fusion, late fusion, joint embedding, graph integration, and modality weighting	Integrated research representation	Comparative evaluation, ablation, leakage testing, and distribution-shift assessment	Dominant modalities, hidden leakage, and opaque weighting	Validation results recalibrate fusion strategy and modality contribution	Integrated representation is not proof of convergence
Cross-modal validation	Determine whether one modality's hypothesis is corroborated independently	Aligned evidence and predefined validation questions	Out-of-modality corroboration, consistency checking, and conflict analysis	Supported, contradicted, or unresolved hypothesis status	Source-independence testing and orthogonal validation	Shared sources create false corroboration	New independent evidence updates hypothesis status	Convergence supports prioritization rather than proof
Missing data handling	Preserve distinctions among	Typed missingness metadata,	Missingness modelling,	Missingness-aware	Mechanism-of-missingnes	Silent imputation, negative-	New measurements replace	High-impact unresolve

	unavailable, unmeasured, censored, and negative evidence	observation process, and modality availability	masking, sensitivity analysis, and partial-modality inference	evidence profile	s assessment and robustness analysis	evidence misclassification, and biased ranking	missing states while retaining history	d missingness limits integration
Uncertainty estimation	Quantify and communicate uncertainty within and across modalities	Measurement errors, calibration data, extraction confidence, sample size, and distribution information	Aleatoric and epistemic uncertainty estimation and propagation	Modality-specific and integrated uncertainty profile	Calibration, stress testing, and external-distribution assessment	Overconfident predictions and misleading aggregate confidence	Validation outcomes recalibrate uncertainty estimates	High uncertainty prevents strong prioritization or interpretation
Bias assessment	Identify sampling, reporting, measurement, annotation, and model bias	Data-generation history, population or chemical-space composition, missingness, labels, and source selection	Structured bias appraisal, subgroup analysis, and stress testing	Bias-risk record	Independent methodological review and external testing	Biased evidence appears consistent across modalities	Audit results trigger data revision, model retraining, or restricted use	High bias risk blocks advancement
Expert review	Apply chemical, botanical, pharmacological, omics, toxicological, clinical, and methodological judgment	Full evidence graph, source passages, conflicts, uncertainty, and intended use	Structured multidisciplinary adjudication	Accepted, revised, deferred, or rejected interpretation	Independent review, rationale documentation, and conflict management	Automation bias, selective interpretation, and disciplinary blind spots	Reviewer rationale updates rules, evidence status, and validation plans	Expert review cannot replace experimental or clinical validation
Candidate prioritization	Allocate research resources among bounded hypotheses	Evidence quality, source independence, uncertainty, safety concerns, novelty, feasibility, and validation burden	Multi-criteria prioritization with sensitivity analysis	Research-priority tier and rationale	Alternative weighting and committee review	False precision, unstable rankings, and therapeutic overstatement	New evidence and completed validation revise priority	No therapeutic-superiority or clinical-utility claim
Validation planning	Convert unresolved questions into decisive	Conflicts, missing modalities, uncertainty, safety concerns,	Evidence-gap mapping and test selection	Orthogonal validation plan	Endpoint relevance, controls, feasibility, and prospective	Convenient proxy testing that does not resolve the key uncertainty	Completed studies update the evidence graph	Advancement requires prespecified validatio

	experiments or studies	and mechanistic gaps			specification			n-gate completion
Feedback updating	Maintain a learning evidence system while preserving history	New bioassays, omics profiles, literature, safety reports, corrections, and expert decisions	Versioned updating, recalibration, and change control	Updated evidence state and decision record	Temporal validation, leakage prevention, and auditability	Retrospective rewriting, loss of negative evidence, and unstable outputs	Append-only updates and periodic multidisciplinary review	Updated outputs retain prior versions and rationale
No clinical-claim boundary	Prevent research outputs from being interpreted as patient-care decisions	Intended use, validation status, user role, claim taxonomy, and communication template	Claim filtering, warning labels, and escalation	Research-only output with explicit limitations	Governance, editorial audit, and user-comprehension testing	Clinical implication through wording, ranking, or interface design	No diagnosis, treatment, prescribing, efficacy, safety, or clinical recommendation	

Figure 2 presents the proposed multimodal molecular intelligence framework for natural product drug discovery, integrating chemical, bioassay, omics, text, clinical, safety, provenance, validation, and expert-review layers.

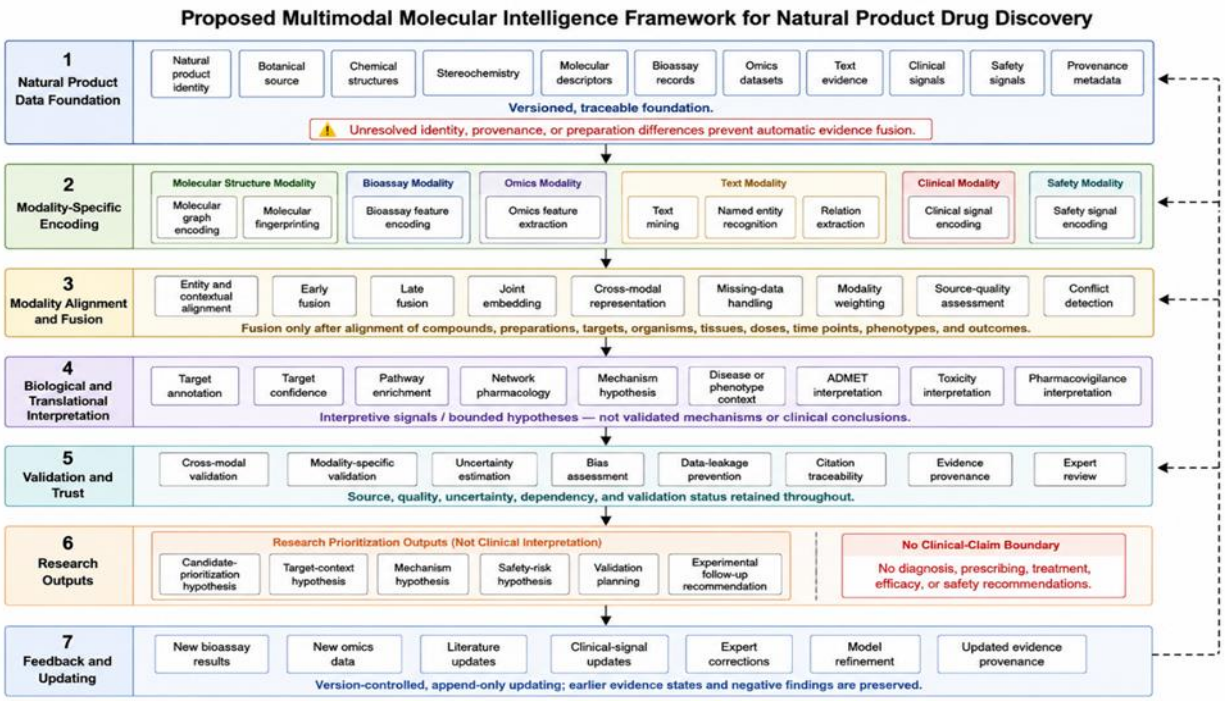

Figure 2. Proposed Multimodal Molecular Intelligence Framework for Natural Product Drug Discovery

Alt-text description

A conceptual multimodal AI framework showing natural product data flowing through chemical structure representation, bioassay integration, omics integration, text evidence retrieval, clinical signal integration, provenance mapping, cross-modal validation, expert review, candidate prioritization, validation planning, and feedback loops.

Research implications

The proposed framework has implications for how research questions are selected and how computational outputs are evaluated in natural product drug discovery. Realistic AI contributions depend on addressing biologically and translationally meaningful problems,

using endpoints that correspond to the intended decision, and planning the experimental work needed after computational prioritization [35, 36]. For natural product informatics, this requires movement away from isolated structure–activity prediction toward traceable evidence objects that connect compound identity, botanical or microbial source, preparation, stereochemistry, assay conditions, omics context, literature claims, and safety information. It also requires recognition that integrated evidence remains limited by the validity of the chemical and biological data from which it is constructed.

Multimodal model development should be accompanied by benchmark designs that reflect natural product chemical space, scaffold diversity, preparation heterogeneity, assay variability, missing modalities, and source-domain shift. Random data splits alone are unlikely to demonstrate whether a representation generalizes to unseen scaffolds, species, preparations, assay systems, or clinical contexts. Benchmarks should therefore distinguish molecular prediction, assay-context interpretation, cross-modal retrieval, conflict detection, missing-modality robustness, uncertainty calibration, and provenance reconstruction as separate tasks. Dataset documentation, interoperable identifiers, versioned ontologies, assay metadata standards, and passage-level citation links should be treated as model-development infrastructure rather than administrative additions. Negative findings, ambiguous structures, failed assays, contradictory omics signatures, and weak clinical attributions should remain available because removing them can produce misleadingly coherent evidence spaces.

For phytopharmacology and translational research, the framework supports safety-aware prioritization by making uncertainty, conflict, product identity, dose relevance, and clinical-signal quality visible before experimental resources are committed. It can also help multidisciplinary teams identify which uncertainty is most decision-relevant and which orthogonal experiment could meaningfully change the evidence state. The framework should not be used to automate therapeutic claims or to substitute computational agreement for pharmacological replication, toxicological evaluation, clinical investigation, or expert judgment. Its most appropriate role is to improve the traceability and structure of research decisions: why a hypothesis was prioritized, which evidence supported it, which evidence conflicted, what remained unknown, and what validation would be required before stronger conclusions could be considered.

CONCLUSION

This article proposes a multimodal molecular intelligence framework for natural product drug discovery that integrates natural product identity, botanical provenance, chemical structures, stereochemistry, molecular descriptors, fingerprints, graphs, bioassays, omics profiles, pathway and network evidence, text-mined literature, clinical signals, pharmacovigilance information, ADMET and toxicity evidence, source provenance, uncertainty, bias assessment, expert review, validation planning, and iterative feedback. Its central contribution is not an implemented predictive system but a structured architecture for organizing heterogeneous evidence without erasing modality-specific limitations or converting computational convergence into unsupported claims. Multimodal integration may strengthen hypothesis prioritization when identities and contexts are aligned, supporting evidence is sufficiently independent, conflicts and missingness remain visible, and uncertainty is communicated transparently. The framework therefore positions candidate prioritization as a research decision-support activity bounded by experimental validation, clinical evaluation, safety review, and multidisciplinary interpretation. Explicit provenance, validation gates, feedback updating, and a no-clinical-claim boundary are essential for ensuring that multimodal molecular intelligence improves evidentiary discipline rather than amplifying overconfidence.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Atanasov AG, Zotchev SB, Dirsch VM, International Natural Product Sciences Taskforce, Supuran CT. Natural products in drug discovery: Advances and opportunities. Nat Rev Drug Discov. 2021;20(3):200-16. doi:10.1038/s41573-020-00114-z

[2] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod. 2020;83(3):770-803. doi:10.1021/acs.jnatprod.9b01285

[3] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[4] Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: Recent advances and future perspectives. Expert Opin Drug

Discov. 2021;16(9):949-59. doi:10.1080/17460441.2021.1909567

[5] Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion. 2019;50:71-91. doi:10.1016/j.inffus.2018.09.012

[6] Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773-84. doi:10.1038/s41591-022-01981-2

[7] Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40(10):1095-110. doi:10.1016/j.ccell.2022.09.012

[8] Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer. 2022;22(2):114-26. doi:10.1038/s41568-021-00408-3

[9] Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1):20. doi:10.1186/s13321-020-00424-9

[10] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021;13(1):2. doi:10.1186/s13321-020-00478-9

[11] Orsi M, Reymond JL. Assigning the stereochemistry of natural products by machine learning. J Cheminform. 2026;18:76. doi:10.1186/s13321-026-01205-6

[12] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A

[13] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/C8SC00148K

[14] Yang K, Swanson K, Jin W, Coley CW, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237

[15] Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2020;63(16):8749-60. doi:10.1021/acs.jmedchem.9b00959

[16] Zhang HW, Lv C, Zhang LJ, Guo X, Shen YW, Nagle DG, et al. Application of omics- and multi-omics-based techniques for natural product target discovery. Biomed Pharmacother. 2021;141:111833. doi:10.1016/j.biopha.2021.111833

[17] Zhu Y, Ouyang Z, Du H, Wang M, Wang J, Sun H, et al. New opportunities and challenges of natural products research: When target identification meets single-cell multiomics. Acta Pharm Sin B. 2022;12(11):4011-39. doi:10.1016/j.apsb.2022.08.022

[18] Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6). doi:10.15252/msb.20178124

[19] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83. doi:10.1186/s13059-017-1215-1

[20] Palazzotto E, Weber T. Omics and multi-omics approaches to study the biosynthesis of secondary metabolites in microorganisms. Curr Opin Microbiol. 2018;45:109-16. doi:10.1016/j.mib.2018.03.004

[21] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi:10.1093/bioinformatics/btz682

[22] Wei CH, Allot A, Leaman R, Lu Z. PubTator Central: Automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1). doi:10.1093/nar/gkz389

[23] Beninger P. Pharmacovigilance: An overview. Clin Ther. 2018;40(12):1991-2004. doi:10.1016/j.clinthera.2018.07.012

[24] Bate A, Hobbiger SF. Artificial intelligence, real-world automation and the safety of medicines. Drug Saf. 2021;44(2):125-32. doi:10.1007/s40264-020-01001-7

[25] Lee CY, Chen YPP. Prediction of drug adverse events using deep learning in pharmaceutical discovery. Brief Bioinform. 2021;22(2):1884-901. doi:10.1093/bib/bbaa040

[26] Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: A structured review. Drug Saf. 2017;40(11):1075-89. doi:10.1007/s40264-017-0558-6

[27] Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. J Natl Cancer Inst. 2017;109(11). doi:10.1093/jnci/djx187

[28] Jing J, Teschke R. Traditional Chinese medicine and herb-induced liver injury: Comparison with drug-induced liver injury. J Clin Transl Hepatol. 2018;6(1):57-68. doi:10.14218/JCTH.2017.00033

[29] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723

[30] Nijman SWJ, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs JJL, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. J Clin Epidemiol. 2022;142:218-29. doi:10.1016/j.jclinepi.2021.11.023

[31] Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. Artif Intell Rev. 2023;56(Suppl 1):1513-89. doi:10.1007/s10462-023-10562-9

[32] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. BMJ. 2021;375. doi:10.1136/bmj.n2281

[33] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y

[34] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368. doi:10.1136/bmj.m689

[35] Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009

[36] Bender A, Cortes-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037