Mapping Natural Product Data Ecosystems for Computational Pharmacology: Phytochemical Libraries, Bioassays, Targets, Pathways, Omics Data, and Clinical Signals

Kristian Eriksen[1]*, Solveig Berg[1], Magnus Dahl[2]

[1]Department of AI for Marine and Terrestrial Natural Products, Faculty of Pharmaceutical Sciences, University of Oslo, Oslo, Norway.
[2]Department of Computational Metabolomics, Faculty of Pharmacy, University of Bergen, Bergen, Norway.

ABSTRACT

Natural-product computational pharmacology increasingly depends on heterogeneous data that extend from biological source identity and phytochemical composition to bioactivity, molecular targets, pathways, omics states, pharmacokinetics, safety observations, and clinical signals. Yet these data are generated under different experimental conditions, represented through incompatible identifiers and schemas, and supported by evidence of unequal maturity. This Original Taxonomy Article maps that ecosystem and proposes a multi-dimensional classification framework designed specifically for natural-product computational pharmacology. The analysis distinguishes phytochemical records from experimentally supported occurrence and quantitative composition, structure-resolved identities from names and synonyms, and spectral or computational annotations from confirmed structures. Bioassay data are classified according to assay type, endpoint, concentration context, metadata completeness, and evidential status. Target evidence is separated into measured, functional, curated, predicted, and text-derived relations, while network and pathway associations are distinguished from direct target engagement. Transcriptomic, proteomic, metabolomic, multi-omics, single-cell, and spatial data are treated as context-dependent molecular-state evidence rather than automatic mechanistic confirmation. Pharmacokinetic, toxicity, adverse-event, and clinical data are further separated according to exposure context, provenance, causal limitations, and translational proximity. The proposed taxonomy organizes data through nine top-level classes and cross-cutting dimensions of evidence mode, biological scale, computational role, provenance, evidence maturity, interoperability readiness, uncertainty, validation status, and translational proximity. It further identifies interoperability gates involving identity resolution, metadata harmonization, semantic alignment, version control, and uncertainty preservation. The framework is intended to support more auditable databases, knowledge graphs, multimodal models, candidate-prioritization systems, and translation-aware computational platforms without assuming that data aggregation removes upstream uncertainty or makes heterogeneous evidence types equivalent.

Key Words: *Natural products, Computational pharmacology, Phytochemical databases, Bioassays, Omics data, Data interoperability*

eIJPPR 2025; 15(3): 1-18

HOW TO CITE THIS ARTICLE: Eriksen K, Berg S, Dahl M. Mapping Natural Product Data Ecosystems for Computational Pharmacology: Phytochemical Libraries, Bioassays, Targets, Pathways, Omics Data, and Clinical Signals. Int J Pharm Phytopharmacol Res. 2025;15(3):1-18. https://doi.org/10.51847/ngYmZOExlc

INTRODUCTION

Computational pharmacology has become increasingly important to natural-product discovery because candidate identification, mechanism inference, prioritization, and translation now draw on data distributed across chemical, biological, systems-level, exposure, safety, and clinical

Corresponding author: Kristian Eriksen
Address: Department of AI for Marine and Terrestrial Natural Products, Faculty of Pharmaceutical Sciences, University of Oslo, Oslo, Norway.
E-mail: ✉ kristian.eriksen@farmasi.uio.no
Received: 26 January 2025; **Revised:** 28 March 2025; **Accepted:** 03 May 2025

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

contexts. The expansion of computational opportunity nevertheless occurs within a fragmented resource landscape in which natural-product records differ in scope, provenance, representation, and evidential meaning [1, 2]. A chemically represented molecule, a reported constituent of a medicinal plant, an activity value from a screening assay, a predicted target, an enriched pathway, and a clinical mention may all enter the same computational workflow, yet they do not constitute equivalent observations. The central challenge is therefore not simply insufficient data volume. It is the absence of a coherent classification logic for determining what type of data object is being used, at which biological scale, under which evidence mode, for what computational purpose, with what degree of traceability, and at what distance from pharmacological or translational validation.

Natural-product informatics compounds this problem because computational workflows must connect highly heterogeneous entities. These may include source organisms, taxonomic names, herbal preparations, extracts, fractions, compounds, metabolites, chemical structures, assays, targets, genes, proteins, pathways, diseases, omics features, tissues, patients, interventions, and adverse events. Computational methods can support similarity analysis, target prediction, network construction, mechanism inference, candidate prioritization, and other analytical tasks, but their outputs remain conditional on the structure and quality of the underlying information [3]. A pipeline that begins with unresolved compound names and then performs structure lookup, target prediction, pathway enrichment, and disease-network ranking can appear analytically sophisticated while carrying unresolved identity uncertainty through every stage. Conversely, a smaller but provenance-rich dataset may be more suitable for a particular pharmacological question than a larger integrated collection containing duplicated, weakly mapped, or semantically heterogeneous records.

Natural products create additional representational challenges that are not fully captured by generic drug-informatics models. Taxonomic ambiguity can disconnect a compound record from the organism or preparation in which it was reported; geographic origin, harvest conditions, plant part, extraction procedure, and batch variation can alter composition; complex mixtures can obscure the relationship between source material and specific molecular entities; stereochemistry, tautomerism, protonation state, salts, and incomplete structures can complicate cross-database identity resolution; and multiple common, systematic, or historical names can refer to the same or different entities. Open natural-product knowledge initiatives demonstrate the importance of making source–compound relations explicit and traceable rather than treating occurrence claims as context-free facts [4]. Such distinctions are particularly consequential when downstream analyses involve multi-target activity, metabolism, pathway interpretation, or cross-layer integration, because ambiguity at the source or structure layer can propagate into apparently independent computational results.

A dedicated taxonomy is therefore needed to classify natural-product data according to more than subject matter alone. The present article maps the natural-product data ecosystem and proposes a multi-dimensional taxonomy for computational pharmacology. Its organizing principle is that an evidence-bearing data object should be characterized through the conjunction of data domain, primary entity, evidence mode, biological scale, computational role, provenance status, granularity, standardization, interoperability readiness, uncertainty, temporal stability, validation status, evidence maturity, and translational proximity. This orientation is deliberately different from a database catalogue: it asks how data classes relate, where category boundaries should be drawn, how ambiguous records should be handled, and what is lost when measured, curated, predicted, derived, or clinically observed evidence is flattened into a common computational representation.

Natural product data ecosystems

A natural-product data ecosystem can be defined as an interconnected but heterogeneous arrangement of data sources, biological and chemical entities, evidence types, standards, mapping relationships, computational uses, and governance practices that collectively support pharmacological inference. This ecosystem is not reducible to a set of databases because the meaning of a record depends on its relation to other entities and on the evidence process through which it was created. A natural-product resource that links species sources, compounds, and activity information illustrates why these components must remain analytically separable: a species-source record, a chemical identity, and an activity observation are linked but distinct evidence units [5]. The ecosystem perspective therefore treats source and identity data, phytochemical data, bioassay evidence, target interactions, pathways, omics states, pharmacokinetic and exposure information, safety observations, and clinical signals as separate but potentially connected layers.

Integration within this ecosystem is not equivalent to simple aggregation. A structural collection may combine records from multiple upstream sources, normalize some representations, and expose them through a common interface, but the resulting harmonization does not prove that every compound occurs in a particular organism or preparation, nor that duplicated records constitute independent support. Open natural-product collections

demonstrate both the value of large-scale structural integration and the importance of retaining source awareness when heterogeneous records are assembled [6]. At minimum, integration should distinguish the database in which a record is encountered from the primary or secondary source from which the underlying assertion originated. Without this distinction, repeated copies of the same structure, name, or occurrence relation can be mistaken for corroboration, while normalization procedures may conceal whether stereochemistry, salt form, or other identity features changed during data processing.

Source-organism information forms a distinct ecosystem layer because the same molecular entity can be reported across different taxa, ecological contexts, preparations, or biosynthetic settings, whereas closely related names may refer to different organisms. Microbial natural-product knowledge resources make this separation especially visible by linking compounds to producing organisms and discovery contexts rather than treating the chemical entity as the sole unit of organization [7]. Comparable issues arise for medicinal plants, fungi, marine organisms, and mixed preparations. Taxonomic identity should therefore be represented through accepted names, synonym relationships, and stable identifiers where possible, while geographic provenance, biological material, preparation type, extraction context, and batch information should be retained as additional attributes rather than compressed into a generic "source" field. Source annotation does not by itself establish that a compound is present in every specimen, batch, formulation, or extraction condition associated with that organism.

Medicinal-plant resources further demonstrate that plant, phytochemical, and therapeutic associations occupy different semantic layers. Curated systems can connect a medicinal plant with compounds and therapeutic contexts, but these edges need not share the same evidence mode or maturity [8]. A plant–compound link may derive from analytical reporting, a compound–target relation may be predicted or curated, and a plant–disease association may arise from traditional use, literature curation, or clinical observation. Treating such edges as equivalent can create circular evidence pathways in which an integrated database is used to generate a hypothesis that is later "validated" against another database reproducing the same upstream source. The ecosystem must therefore preserve relation type, source lineage, evidence mode, version, and confidence where available, while acknowledging that database overlap may reflect shared provenance rather than independent replication.

The resulting architecture is best understood as a set of conditional transitions rather than a frictionless pipeline. Source and identity information may connect to phytochemical records; phytochemical entities may connect to assays, targets, exposure data, or metabolomic features; assay results may support target hypotheses; targets may be contextualized through networks and pathways; pathways and targets may be compared with omics states; and exposure, safety, and clinical signals may inform translation. At each transition, however, identifier mismatch, synonym ambiguity, structure normalization, taxonomic uncertainty, assay heterogeneity, inconsistent units, version drift, conflicting annotations, unclear provenance, and predicted-to-predicted evidence chaining can disrupt the inference. Data lineage is therefore not an administrative add-on but a pharmacologically relevant property: when an upstream identity or annotation error is propagated through several computational layers, apparent cross-layer convergence may be generated by shared error rather than independent evidence.

Table 1 maps the major data layers that constitute the natural-product computational pharmacology ecosystem and distinguishes their primary entities, evidence modes, computational roles, provenance requirements, and integration risks.

Table 1. Natural Product Data Ecosystems for Computational Pharmacology: Core Data Layers, Primary Entities, Evidence Modes, Biological Scales, Computational Roles, Provenance Requirements, and Integration Risks

Data layer	Primary entity	Typical evidence mode	Biological scale	Primary computational role	Provenance requirement	Interoperability need	Major uncertainty	Integration risk
Source and identity	Plant, organism, preparation, extract, batch	Observed, recorded, curated, text-derived	Organism; preparation	Identification; provenance filtering	Taxonomic source, accepted name, synonym history, specimen or source citation,	Taxonomic identifiers; source vocabularies; preparation descriptors	Misidentification, synonyms, incomplete provenance, batch variability	Wrong source assignment propagates into constituent and mechanism layers

					preparation context			
Phytochemical and structural	Compound, metabolite, structure, spectral feature	Measured, curated, inferred, predicted, derived	Molecular	Representation; similarity; dereplication	Structure source, analytical support, stereochemical status, occurrence source	Structure normalization; persistent chemical identifiers; synonym control	Tautomers, protonation, stereochemistry, tentative annotation	Name-based or normalized identity collapse creates false equivalence
Bioassay and phenotypic	Assay, compound, sample, endpoint, cell system	Experimentally measured, curated	Molecular; cellular; organism	Activity analysis; hit selection; dose-response modeling	Protocol, endpoint, units, concentration, controls, biological system	Assay ontology; unit harmonization; endpoint mapping	Protocol dependence, interference, missing context	Non-comparable values are pooled or transferred across incompatible assays
Target and interaction	Compound, target, protein, interaction edge	Measured, curated, predicted, text-derived	Molecular; target	Target inference; interaction analysis	Interaction source, assay type, prediction method, curation lineage	Chemical–protein identifier mapping; evidence typing	Isoforms, indirect effects, prediction uncertainty	Predicted association is upgraded to validated engagement
Network and pathway	Protein, gene, disease, pathway, module, relation	Curated, measured, predicted, derived	Interaction; pathway; cellular	Network construction; mechanism inference	Edge source, relation semantics, pathway version, enrichment inputs	Gene–protein mapping; pathway vocabularies; relation typing	Indirect association, hub bias, version drift	Membership or enrichment is misread as pathway modulation
Omics and molecular-state	Transcript, protein, metabolite, feature, cell, tissue	Measured, derived, model-derived	Molecular; cellular; tissue	Contextualization; biomarker analysis; multimodal integration	Sample origin, platform, preprocessing, batch, tissue and cell context	Feature mapping; modality alignment; sample linkage	Batch effects, missingness, confounding, cell composition	Correlation or latent structure is interpreted as causality
Pharmacokinetic and exposure	Compound, metabolite, dose, concentration, subject	Experimentally measured, curated, model-derived	Organism; population	Exposure assessment; PK modeling	Dose, route, formulation, matrix, time, units, analytical method	Unit harmonization; parent–metabolite mapping	Formulation effects, interindividual variability, incomplete sampling	In vitro activity is assumed to imply achievable exposure
Safety and toxicity	Compound, preparation, organ, event, interaction	Measured, predicted, curated, clinically observed signal	Molecular; organism; population	Risk prioritization; safety assessment	Identity, exposure context, endpoint, report source, model provenance	Product normalization; event terminology; exposure linkage	Confounding, underreporting, prediction error, product heterogeneity	Safety signal or prediction is treated as established causation
Clinical and translational signal	Patient, intervention, biomarker, response, adverse event	Clinically observed, curated, real-world where supported, derived	Organism; population	Clinical contextualization; translation	Study design, intervention identity, population, outcome definition,	Clinical terminologies; intervention mapping; outcome alignment	Confounding, indication bias, reporting bias, temporal ambiguity	Clinical mention or association is treated as validated efficacy

	temporal context

Phytochemical and bioassay data

Phytochemical data occupy a critical interface between source identity and computational representation. A phytochemical record may contain a compound name, synonym, canonical structure, stereochemical specification, taxonomic association, reported occurrence, analytical annotation, or quantitative measurement, but these attributes should not be collapsed into a single notion of "compound presence." Enhanced medicinal-plant phytochemical resources illustrate the value of organizing plant–compound relationships while also showing why the plant entity, chemical entity, and contextual association remain distinct [9]. For computational purposes, the same named constituent may map to multiple structures because of ambiguous naming, or several names may map to the same normalized structure. Canonicalization can facilitate deduplication, yet it can also remove distinctions that matter pharmacologically if stereochemistry, protonation state, tautomeric form, isotopic labeling, salt form, or mixture context is not preserved.

Analytical evidence introduces a second axis of distinction. A reported compound in the literature is not necessarily analytically confirmed in the material under study; a detected feature is not necessarily an exact structure; a quantified analyte carries different evidential content from a tentative annotation; and a predicted constituent remains computationally inferred. Feature-based molecular networking and computational metabolite classification provide complementary forms of derived analytical evidence, but neither should be interpreted as independent confirmation of exact chemical structure [10, 11]. LC–MS and LC–MS/MS data can support feature detection, fragmentation-based similarity, annotation propagation, and chemical-class inference, whereas GC–MS may be appropriate for volatile constituents and NMR can provide stronger structural evidence when sufficient data and interpretation are available. The taxonomy must therefore retain whether a record is reported, detected, quantified, structure-confirmed, spectrally annotated, class-predicted, or otherwise inferred.

Compound occurrence is likewise context-dependent. A constituent reported from one specimen, geographic origin, plant part, growth condition, extraction procedure, or analytical platform should not automatically be assigned to all preparations associated with the same taxon. Quantitative composition requires explicit units, calibration context, analytical method, matrix, and sampling conditions; otherwise values may be incomparable even when they refer to the same nominal compound. This distinction matters computationally because source-specific candidate generation, mixture modeling, similarity analysis, and downstream exposure inference can all be distorted by unqualified occurrence records. At the structural level, persistent identifiers are valuable only when they preserve the intended entity. A generic identifier mapped to an unspecified stereoisomer or normalized parent structure may be suitable for some similarity tasks but inadequate for target binding, metabolism, or pharmacokinetic reasoning.

Bioassay data add another layer of heterogeneity because activity is conditional on experimental design. Structured bioactivity resources demonstrate the importance of representing assay descriptions, endpoints, compounds, targets, and measurements rather than storing isolated activity values [12]. Natural-product computational pharmacology may draw on biochemical assays, direct binding measurements, functional target assays, cell-based assays, phenotypic screens, high-throughput screening outputs, and dose-response studies. Each class answers a different question. A biochemical inhibition value concerns a defined molecular or enzymatic context; a binding measurement may quantify physical interaction without establishing downstream functional modulation; a functional assay measures a response associated with target activity; and a phenotypic effect may occur without a known molecular target. Consequently, assay type, endpoint, concentration range, units, controls, replicate design, protocol, species, cell line, readout technology, and experimental context are integral parts of the evidence.

Target-oriented binding evidence provides an especially important boundary between the bioassay and interaction layers. FAIR-oriented binding resources emphasize explicit protein–small-molecule measurement data, making clear why experimentally measured binding should remain distinguishable from a curated association, predicted target, or text-mined relation [13]. Even measured values require contextual interpretation: affinity and potency endpoints are not interchangeable; conditions can differ; concentration units may vary; censored or threshold values require care; and proteins, constructs, isoforms, cofactors, or assay formats may not be equivalent. Natural products can also create assay-specific complications through aggregation, nonspecific reactivity, autofluorescence, optical interference, poor solubility, or mixture effects. Bioassay values are therefore not directly comparable merely because they can be transformed into the same numerical unit. Computational integration should first determine whether the underlying measurement semantics and experimental contexts are sufficiently aligned.

Table 2 compares major phytochemical and bioassay data classes according to identity resolution, experimental context, metadata requirements, computational utility, and risks of misinterpretation.

Table 2. Phytochemical and Bioassay Data for Natural Product Computational Pharmacology: Identity Resolution, Structural Representation, Experimental Context, Metadata Requirements, Computational Uses, Validation Needs, and Failure Risks

Data class	Primary entity	Identity or measurement basis	Critical metadata	Typical computational use	Evidence maturity	Validation requirement	Major interoperability challenge	Failure risk
Reported phytochemical	Compound name linked to source	Literature or curated report	Original source, taxon, material, preparation, extraction context	Candidate generation; occurrence mapping	Exploratory to curated	Trace to primary occurrence evidence	Name and taxon normalization	Database report is treated as universal occurrence
Structure-resolved phytochemical	Chemical structure	Explicit molecular representation	Stereochemistry, tautomer/protonation policy, salt handling, structure source	Similarity; descriptors; modeling	Curated to experimentally supported	Confirm structure provenance and representation	Cross-resource structure normalization	Distinct entities collapse into one normalized record
Experimentally detected constituent	Compound or analytical feature	Instrumental observation	Platform, sample, retention information, spectral evidence, annotation level	Dereplication; sample profiling	Experimentally supported at detection level	Orthogonal confirmation proportional to claim	Feature-to-identity mapping	Tentative feature is treated as confirmed compound
Quantified constituent	Compound and abundance value	Calibrated or semi-quantitative measurement	Units, calibration, matrix, method, sample context	Composition modeling; exposure hypotheses	Experimentally supported	Analytical method and calibration review	Unit and matrix harmonization	Values from incompatible matrices or methods are pooled
Spectral annotation	Spectrum or feature	Similarity to reference or library evidence	Spectrum source, match criteria, confidence level	Annotation; dereplication	Exploratory to supported	Reference spectrum or orthogonal structural evidence	Spectral identifiers and confidence representation	Similarity is treated as exact structural identity
Molecular-network relation	Feature-to-feature relation	Fragmentation similarity and feature correspondence	Network parameters, feature processing, sample linkage	Chemical-family exploration	Hypothesis-generating	Orthogonal annotation for exact identity claims	Feature alignment across experiments	Network proximity is treated as shared structure or bioactivity
Predicted constituent	Compound hypothesis	Computational inference	Model, input data, training scope, confidence	Candidate prioritization	Exploratory	Analytical verification	Predicted and observed occurrence states	Predicted presence is flattened into measured composition
Biochemical assay	Compound–endpoint measurement	Defined biochemical response	Target/system, endpoint, units, concentration, controls, protocol	Activity modeling; hit ranking	Experimentally supported	Reproducibility and assay-context assessment	Endpoint and unit harmonization	Incompatible assay outputs are directly compared
Binding assay	Compound–protein measurement	Physical interaction or affinity endpoint	Protein construct, assay format, endpoint type, conditions	Interaction modeling; target evidence	Experimentally supported	Independent or orthogonal confirmation	Protein identity and endpoint semantics	Binding is treated as functional modulation

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

						where consequential		
Functional target assay	Compound–target-linked response	Functional readout	Biological system, pathway context, endpoint, controls	Mechanism assessment	Experimentally supported	Target dependence and reproducibility	Mapping readout to target mechanism	Indirect response is treated as direct engagement
Cell-based assay	Compound/sample–cell response	Cellular phenotype or functional endpoint	Cell line, species, exposure time, concentration, readout, medium	Activity prediction; contextual screening	Exploratory to experimentally supported	Cross-model and mechanistic validation	Cell-system and endpoint mapping	Cell-context-specific effects are generalized
Phenotypic screen	Compound/sample–phenotype relation	Multicomponent phenotype	Model system, phenotype definition, imaging/readout, exposure	Hit discovery; mechanism hypothesis generation	Hypothesis-generating to supported	Target deconvolution or orthogonal phenotype confirmation	Phenotype ontology and protocol heterogeneity	Phenotypic activity is assigned to an unvalidated target
Dose-response evidence	Compound–response series	Concentration-dependent measurement	Concentration range, curve model, replicates, censoring, endpoint	Potency estimation; comparative ranking	Experimentally supported	Curve-quality and reproducibility assessment	Unit, endpoint, and model consistency	Summary value obscures incomplete or non-comparable curves

Target, pathway, and omics data

Target data require an explicit evidence hierarchy because the term "target" can refer to materially different relations. Experimentally measured binding, functional target evidence, curated interactions, predicted compound–target links, and text-mined associations should be encoded as distinct evidence modes rather than merged into a single edge type. Integrated target-prioritization resources illustrate how target–disease evidence can be assembled from heterogeneous sources, but they also make clear that evidence items differ in origin and inferential role [14]. In natural-product computational pharmacology, this distinction is essential because compounds are frequently assigned multi-target profiles through prediction or network methods. A predicted target may be useful for hypothesis generation, yet it does not demonstrate binding, target engagement in cells, causal mediation of phenotype, or relevance at achievable exposure. Target records should therefore preserve the compound identity, target identifier, evidence mode, source, assay or prediction context, direction where meaningful, confidence, and validation state.

Interaction networks introduce a separate semantic problem. Protein–protein resources can integrate experimental, curated, predicted, co-expression, text-derived, and other association channels, but an association edge is not necessarily a direct physical interaction and is not a compound–target interaction [15]. For natural-product analysis, drug–target interactions, compound–protein measurements, protein–protein relations, disease-gene associations, co-expression links, and network modules should remain typed because they answer different questions. A compound may have measured evidence against one protein, predicted associations with several others, and network proximity to a disease module. Combining these relations can support prioritization, but their convergence should not be mistaken for repeated experimental verification. Network topology can also be shaped by literature bias, highly studied hubs, incomplete coverage, and dependencies among source databases.

Pathway data sit at a higher contextual level than direct target evidence. Curated pathway knowledgebases organize reactions, molecular participants, and biological processes, providing a structured basis for functional interpretation [16]. However, the presence of a target in a pathway does not show that a natural product modulates that pathway, and pathway membership should not be treated as equivalent to target engagement. Similar caution applies to functional annotations, disease pathways, network modules, and enrichment outputs. A pathway enrichment result is a derived statistical relation contingent on the input feature list, background universe, identifier mappings, pathway version, significance procedure, and biological context. It can suggest that a set of observations is non-randomly associated with a functional category, but it does not independently establish mechanism. The

taxonomy therefore separates T4 target and interaction data from T5 network and pathway data while preserving explicit cross-layer links.

Omics data contribute context by measuring molecular states at scales that may include transcripts, proteins, metabolites, cells, tissues, and integrated modalities. Transcriptomics can characterize expression patterns; proteomics can provide protein-level abundance or state information; metabolomics can measure chemical features associated with endogenous or exogenous processes; and multi-omics methods can integrate partially overlapping views of biological systems. Statistical integration frameworks show that shared or modality-specific structure can be extracted across complex datasets, including multimodal single-cell and multi-block settings, while remaining dependent on model assumptions and integration design [17, 18]. Accordingly, a latent factor, selected feature set, integrated component, or cross-modal correlation is classified as model-derived evidence rather than direct proof of causality. Single-cell and spatial data may add cell-type or location-specific resolution where supported, but they also introduce additional requirements for sample provenance, feature mapping, batch correction, cell-state definition, and spatial context.

The principal uncertainties in omics data arise not only from measurement error but from study design and preprocessing. Batch effects can correlate with biological groups; platforms may measure different feature spaces; tissue specificity can make a molecular association irrelevant outside the sampled organ; cell-type composition can create apparent expression changes; temporal sampling can alter observed states; normalization choices can reshape effect estimates; missingness may be non-random; multiple testing can generate unstable discoveries; and integration methods can be sensitive to scaling, feature selection, sample overlap, or modality imbalance. For natural products, mixture composition and metabolism further complicate interpretation because the molecular state observed after exposure may reflect parent compounds, active metabolites, indirect responses, toxicity, or adaptive processes. Omics evidence should therefore retain sample context, biological material, exposure design, preprocessing, analytical platform, batch structure, feature identifiers, and validation status.

Cross-layer integration is most informative when the relation being tested is explicit. Target-to-omics integration can ask whether a proposed target is expressed or perturbed in a relevant tissue or cell state; pathway-to-omics analysis can compare curated functional structures with measured molecular changes; compound-to-metabolomics integration can relate exposures or chemical features to molecular-state changes; and phenotype-to-transcriptomics analysis can connect an observed response with context-specific expression patterns. None of these mappings removes upstream uncertainty. An incorrectly normalized compound can be linked to an incorrect predicted target; that target can map to a pathway; the pathway can overlap an omics signature; and the resulting chain can appear coherent despite originating from a flawed identity assertion. Robust integration should therefore preserve evidence mode at every edge, distinguish observed from derived relations, maintain versioned identifier mappings, expose missingness and conflicts, and require orthogonal validation before a cross-layer association is upgraded to a mechanistic claim.

Table 3 compares target, network, pathway, and omics data layers by evidence status, biological scale, inferential role, integration opportunity, and major uncertainty.

Table 3. Target, Network, Pathway, and Omics Data Layers in Natural Product Computational Pharmacology: Evidence Status, Biological Scale, Mechanistic Role, Integration Opportunities, Validation Requirements, and Uncertainty Sources

Data layer	Primary entity	Evidence status	Biological scale	Typical analytical output	Mechanistic role	Integration opportunity	Validation requirement	Major uncertainty	Interpretive safeguard
Experimentally measured binding	Compound; protein/target; binding value	Measured	Molecular; target	Affinity or binding endpoint	Supports direct physical interaction under assay conditions	Connect phytochemical identity to target evidence	Orthogonal binding or context-appropriate replication where consequential	Assay conditions, protein construct, endpoint comparability	Do not equate binding with cellular engagement or functional effect
Functional target evidence	Compound; target; functional readout	Measured	Target; cellular	Target-linked response	Supports functional relation in a defined system	Link target hypothesis with phenotype or pathway response	Demonstrate target dependence and reproducibility	Indirect effects, system dependence, off-	Preserve assay context and distinguish direct from indirect evidence

								target activity	
Curated interaction	Compound; protein; relation	Curated secondary	Molecular; target	Typed interaction edge	Provides organized prior knowledge	Knowledge graphs; evidence retrieval; hypothesis support	Trace to underlying source and evidence mode	Curation lag, source heterogeneity, copied assertions	Retain primary-source lineage and curation status
Predicted target	Compound; target; probability/rank	Predicted or model-derived	Molecular; target	Candidate target set	Hypothesis generation	Link structures to network and pathway analyses	Experimental target validation	Domain shift, training bias, uncertain structure input	Keep "predicted" status visible through all downstream layers
Text-mined association	Compound; gene/protein; textual relation	Text-derived	Molecular to disease	Extracted relation	Hypothesis generation and literature navigation	Expand candidate relation space	Manual or experimental confirmation	Entity ambiguity, negation, context loss	Preserve sentence/document provenance and extraction confidence
Protein interaction network	Protein; interaction/association edge	Mixed: measured, curated, predicted, text-derived	Interaction; cellular	Network topology; neighborhoods; modules	Contextualizes targets within molecular systems	Connect target evidence to functional neighborhoods	Evidence-channel review; context-specific validation	Association semantics, hub bias, coverage bias	Type edge sources and distinguish physical from functional association
Disease-gene relation	Disease; gene/protein	Curated, observed, inferred	Target to organism	Disease-associated gene set	Links molecular entities to disease context	Target prioritization; module analysis	Disease-context and evidence-source validation	Pleiotropy, study bias, indirect association	Avoid treating association as causal disease mechanism
Curated pathway	Pathway; reaction; molecular participant	Curated	Pathway; cellular	Pathway membership and reaction structure	Functional contextualization	Target-to-pathway and omics-to-pathway mapping	Version and context review	Incomplete knowledge, pathway boundaries, version drift	Treat membership as context, not demonstrated modulation
Enrichment output	Feature set; pathway/category	Derived	Pathway	Enriched categories or scores	Generates higher-level functional hypotheses	Integrate targets or omics features with pathways	Independent biological or experimental contextualization	Input list dependence, background choice, multiple testing	Classify as derived evidence and report analytical assumptions
Transcriptomics	Transcript/gene; sample; cell/tissue	Measured plus derived preprocessing	Molecular; cellular; tissue	Differential expression, signatures, trajectories	Contextualizes molecular response	Phenotype ↔ expression; target ↔ tissue context	Independent cohort or orthogonal validation where needed	Batch effects, tissue specificity, cell composition, normalization	Preserve sample context and avoid equating expression with target activity
Proteomics	Protein/peptide; sample	Measured plus derived	Molecular; cellular; tissue	Abundance or protein-state patterns	Adds protein-level context	Target ↔ protein state; pathway ↔	Orthogonal measurement and context validation	Coverage, missingness, peptide mapping,	Retain protein isoform and measurement context

		preprocessing				measured response		platform effects	
Metabolomics	Metabolite/feature; sample	Measured, annotated, inferred, derived	Molecular; cellular; organism	Metabolic features, signatures, networks	Characterizes molecular-state and exposure-related changes	Compound ↔ metabolome; pathway ↔ metabolites	Structural confirmation proportional to claim	Annotation ambiguity, matrix effects, feature matching	Separate detected feature, annotated metabolite, and confirmed identity
Multi-omics integration	Multi-modal feature sets; latent factors/components	Model-derived	Molecular to tissue	Integrated components, factors, selected features	Cross-modal contextualization	Link transcriptomic, proteomic, metabolomic, and other states	External replication and modality-specific interpretation	Integration instability, scaling, missing modalities, model assumptions	Treat integrated factors as derived representations, not causal mechanisms
Single-cell omics	Cell; transcript/protein/feature	Measured plus model-derived	Cellular; tissue	Cell states, clusters, trajectories	Resolves cellular heterogeneity	Target/pathway ↔ cell type; intervention ↔ state	Replication and biological annotation validation	Cell-state definitions, batch, dropout, dissociation artifacts	Preserve cell provenance and avoid overgeneralizing clusters
Spatial omics	Spatial unit; feature; tissue region	Measured plus derived	Cellular; tissue	Spatial patterns and neighborhoods	Adds localization to molecular-state interpretation	Pathway/target ↔ tissue architecture	Orthogonal spatial and biological validation	Resolution, segmentation, platform effects, compositional bias	Keep spatial scale and analytical uncertainty explicit

Clinical and safety signals

The transition from preclinical computational evidence to human-level decision contexts requires a distinct pharmacokinetic and exposure layer. A compound can be reported in a source organism, structurally represented, active in vitro, and associated with plausible targets while remaining irrelevant to human pharmacology if exposure is insufficient or contextually mismatched. Structured pharmacokinetic resources illustrate why dose, route, formulation, biological matrix, sampling time, concentration units, subject characteristics, and modeling context must accompany exposure observations [19]. For natural products, these requirements are intensified by complex preparations, variable composition, parent–metabolite relationships, active metabolites, transformation by host or microbial metabolism, and formulation-dependent bioavailability. Compound occurrence therefore does not establish systemic exposure, and in vitro potency should not be interpreted as evidence that an active concentration can be achieved in a relevant tissue. T7 Pharmacokinetic and Exposure Data consequently forms a separate class rather than an extension of phytochemical occurrence or bioassay activity.

Safety data are similarly heterogeneous and should not be compressed into a single toxicity label. Experimental toxicity, predicted toxicity, organ-specific toxicity, genotoxicity where relevant, interaction risk, and adverse-event observations arise from different evidence-generating processes. Integrated chemical and toxicity-oriented resources demonstrate the importance of resolvable chemical identity when linking structures to safety information, while also showing that aggregation can connect data generated under different contexts [20]. A predicted hepatotoxicity output, an experimentally measured cytotoxic response, an animal toxicology observation, a documented pharmacokinetic interaction, and a spontaneously reported adverse event differ in biological scale, provenance, uncertainty, and causal interpretation. Natural-product safety assessment must additionally account for mixtures, adulteration, contamination, dose uncertainty, preparation differences, co-medication, and the possibility that a named herbal product does not uniquely identify its botanical or chemical composition.

Pharmacovigilance creates a particularly difficult interoperability problem because spontaneous-reporting systems are organized around product names, event terms, and reporting practices that may not align with taxonomic

or phytochemical representations. Taxonomic name-resolution methods applied to herbal adverse-event retrieval show that synonyms, spelling variation, vernacular names, and botanical naming inconsistencies materially affect which reports can be identified [21]. A single plant may appear under accepted scientific names, historical synonyms, common names, preparation names, or commercial descriptors, while a common name may refer to more than one taxon. Botanical misidentification and incomplete product composition further limit interpretation. Consequently, an observed reporting association should be classified as a safety signal with explicit source and entity-resolution provenance rather than as proof that a chemically defined constituent caused the reported event.

Natural-product mention recovery in spontaneous-reporting data further demonstrates that entity normalization is itself an inferential step. Expanded matching methods can retrieve reports that exact-string searches miss, but the resulting increase in capture does not remove reporting bias, confounding, co-medication, duplicate information, incomplete exposure histories, or uncertainty about product identity [22]. An adverse-event signal may support surveillance, prioritization, or hypothesis generation, yet it does not independently establish causation. Similar caution applies to observational clinical associations, real-world signals where peer-reviewed evidence is available, and literature-derived clinical mentions. A clinical mention is not equivalent to demonstrated efficacy; a biomarker association is not necessarily predictive; treatment response may depend on indication and population; and an adverse-event report does not show that the product or constituent was the causal agent.

Clinical and safety data should therefore be treated as active contextual layers rather than terminal endpoints appended to a computational pipeline. Human-level evidence can challenge upstream assumptions by showing that an active in vitro compound lacks exposure, that a predicted target is not relevant in the treated population, that metabolism produces different active entities, or that formulation and co-medication alter the safety profile. Conversely, clinical signals can generate new computational hypotheses, but reverse translation should retain the same evidential discipline as forward integration. The relevant safeguard is to preserve study design, intervention identity, population, outcome definition, temporal context, exposure information, co-medications, and provenance while distinguishing exploratory signals from externally supported or translation-relevant evidence. No single clinical or safety layer should automatically validate preceding computational predictions.

Table 4 synthesizes the principal clinical, pharmacokinetic, safety, and translational signal classes relevant to connecting computational natural-product evidence with human-level decision contexts.

Table 4. Clinical and Safety Signal Layers for Natural Product Computational Pharmacology: Evidence Sources, Translational Roles, Causal Limitations, Provenance Requirements, Integration Opportunities, and Decision Safeguards

Signal class	Primary entity	Typical evidence source	Translational role	Evidence maturity	Major confounder	Causal limitation	Integration opportunity	Validation requirement	Decision safeguard
Pharmacokinetic observation	Compound, metabolite, dose, concentration, subject	Preclinical or clinical PK study; curated PK resource	Establishes exposure context	Experimentally supported to clinically contextualized	Population variability, formulation, sampling design	Exposure does not prove efficacy or toxicity	Link constituent identity to achievable concentration and time course	Verify route, dose, matrix, units, analytical method, and population	Keep measured and model-derived exposure distinct
Bioavailability evidence	Compound, metabolite, formulation	Clinical or preclinical exposure study	Assesses fraction and context of systemic availability	Experimentally supported	Formulation effects, food effects, metabolism	Bioavailability does not establish target engagement	Filter candidates by exposure plausibility	Replication and formulation-specific assessment	Do not transfer across preparations without justification
Metabolism evidence	Parent compound,	PK, metabolite-identification,	Identifies transformed active or	Experimentally supported	Species differences, incomplete	Detected metabolite does not	Connect phytochemical,	Structural confirmation and	Preserve parent–

	metabolite, enzyme, pathway	clinical pharmacology studies	inactive entities	to clinically contextualized	metabolite coverage	prove causal pharmacological role	exposure, target, and safety layers	exposure relevance	metabolite lineage
Drug–herb or compound–drug interaction evidence	Product, compound, drug, enzyme/transporter	Clinical pharmacology, experimental interaction study	Supports interaction-risk assessment	Hypothesis-generating to clinically contextualized	Co-medication, product heterogeneity, dose	Association or in vitro inhibition does not prove clinically meaningful interaction	Link exposure, metabolism, and safety evidence	Context-appropriate clinical or mechanistic validation	Distinguish predicted, experimental, and clinically observed interactions
Predicted toxicity	Compound, endpoint, organ/system	Computational model	Early safety prioritization	Exploratory to hypothesis-generating	Training-domain bias, uncertain structure input	Prediction is not observed toxicity	Integrate with candidate ranking	Experimental evaluation	Preserve model provenance and applicability limits
Experimental toxicity	Compound, extract, preparation, biological system	In vitro or in vivo toxicology study	Supports hazard characterization	Experimentally supported	Dose, species, model system, preparation	Model-system toxicity may not establish human harm	Connect exposure and mechanism hypotheses	Replication and exposure relevance	Avoid direct extrapolation across species or formulations
Organ-specific toxicity evidence	Compound, preparation, organ endpoint	Experimental or clinical investigation	Contextualizes tissue-specific risk	Experimentally supported to clinically contextualized	Comorbidities, co-exposures, endpoint definition	Organ association may not identify causal constituent	Connect target, pathway, omics, and exposure layers	Orthogonal evidence and causal assessment	Preserve organ, dose, timing, and identity context
Genotoxicity evidence	Compound, preparation, genotoxic endpoint	Validated experimental assay where relevant	Supports specific hazard evaluation	Experimentally supported	Assay conditions, cytotoxicity, metabolic activation	Positive assay does not automatically establish human clinical risk	Connect chemical identity with safety prioritization	Assay confirmation and contextual interpretation	Keep endpoint-specific evidence separate from generic toxicity
Spontaneous adverse-event signal	Product, exposure mention, adverse event	Pharmacovigilance reporting system	Surveillance and hypothesis generation	Exploratory to clinically contextualized signal	Underreporting, stimulated reporting, co-medication, duplicates	Reporting association does not establish causation	Link products or constituents to safety hypotheses	Case assessment, independent evidence, exposure clarification	Label explicitly as signal; retain report provenance
Clinical study evidence	Intervention, patient, outcome	Interventional or observational clinical study	Human-level contextualization	Clinically contextualized to translation-relevant	Study design, confounding, selection, product heterogeneity	Clinical association depends on design and bias control	Connect candidate and mechanism hypotheses to human outcomes	Design-appropriate replication and critical appraisal	Do not infer beyond intervention, population, and endpoint studied
Clinical pharmacology evidence	Intervention, exposure, biomarker, physiological response	Human PK/PD or mechanistic clinical study	Connects exposure with biological response	Clinically contextualized	Interindividual variability, formulation, timing	Association between exposure and response	Link T7 exposure to target, biomarker, and	Replication and mechanistic coherence	Preserve dose–exposure–response context

						may not establish therapeutic efficacy	response layers		
Biomarker signal	Biomarker, patient, intervention, outcome	Clinical or translational study	Supports stratification or mechanistic contextualization	Hypothesis-generating to externally validated	Confounding, assay variability, population shift	Association does not establish predictive or causal value	Link omics evidence to clinical context	Independent validation in relevant population	Distinguish prognostic, predictive, pharmacodynamic, and exploratory roles
Treatment-response signal	Intervention, patient, response endpoint	Clinical study or peer-reviewed real-world analysis	Supports translation-aware prioritization	Clinically contextualized	Indication bias, adherence, co-treatment, endpoint heterogeneity	Observed response may not establish mechanism or generalizability	Compare upstream computational hypotheses with human response	Replication and study-design assessment	Retain population and intervention specificity
Peer-reviewed real-world signal	Exposure/intervention, patient, outcome	Observational clinical or healthcare data	Complements controlled-study evidence where supported	Hypothesis-generating to clinically contextualized	Confounding, misclassification, missingness	Association is not automatically causal	Test transportability and detect uncommon contexts	Robust design, sensitivity analysis, external replication	Keep real-world source, temporality, and confounding assumptions explicit

Proposed data taxonomy

The proposed taxonomy is designed to classify evidence-bearing natural-product data objects rather than merely enumerate databases or modalities. Its objective is to make explicit what an item represents, how it was generated, how it may be used computationally, and what transformations are required before it can be connected to another layer. Interoperability-oriented infrastructure demonstrates that standards and resource metadata can improve discoverability and coordination, but technical accessibility should not be conflated with pharmacological validity [23]. Accordingly, the taxonomy treats interoperability as one cross-cutting dimension among several rather than as an endpoint. The core unit of classification is an entity or relation in context, characterized by data domain, primary entity, evidence mode, biological scale, computational role, provenance status, evidence maturity, interoperability readiness, uncertainty, validation state, temporal stability, and translational proximity.

Exactly nine top-level classes define the domain axis: T1. Source and Identity Data covers taxonomic identity, geographic provenance, botanical material, preparation type, extraction context, and batch metadata; T2. Phytochemical and Structural Data covers compound identity, chemical structure, stereochemistry, spectral evidence, occurrence, and quantitative composition; T3. Bioassay and Phenotypic Data covers biochemical, cell-based, phenotypic, dose-response, endpoint, and screening evidence; T4. Target and Interaction Data covers measured binding, functional evidence, curated interactions, predicted targets, and text-derived associations; T5. Network and Pathway Data covers protein interactions, disease associations, pathway membership, functional annotations, network modules, and enrichment outputs; T6. Omics and Molecular-State Data covers transcriptomics, proteomics, metabolomics, multi-omics, and supported single-cell or spatial data; T7. Pharmacokinetic and Exposure Data covers absorption, bioavailability, distribution, metabolism, exposure, and active metabolites; T8. Safety and Toxicity Data covers experimental and predicted toxicity, organ toxicity, interaction risk, and adverse-event signals; and T9. Clinical and Translational Signal Data covers clinical studies, clinical pharmacology, biomarkers, treatment response, supported real-world signals, and translation-relevant evidence. Reliable cross-class mapping requires stable, namespace-aware identifiers because syntactic identifier resolution is a prerequisite for joining records without silently merging unrelated entities [24].

Figure 1 presents the proposed multi-dimensional taxonomy linking natural-product data layers, evidence modes, biological scales, computational roles, provenance requirements, interoperability transitions, and translational proximity.

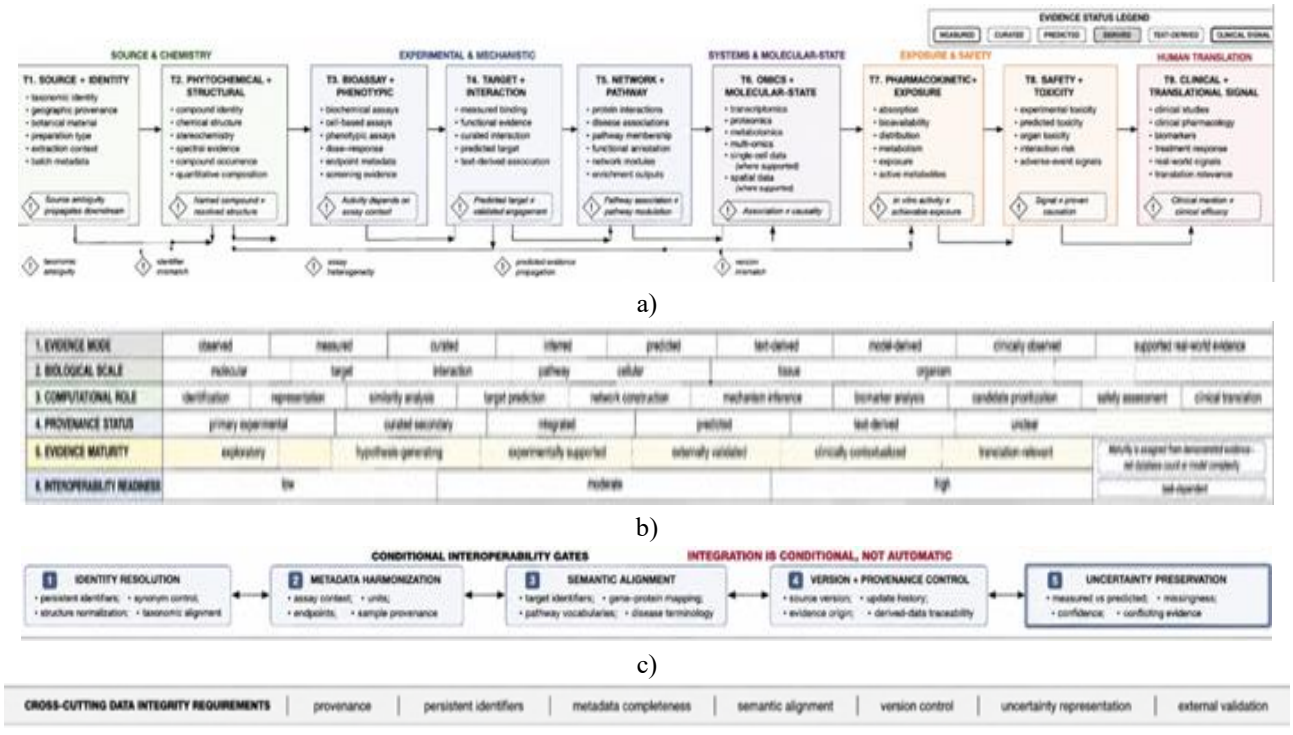

Figure 1. Multi-Dimensional Data Taxonomy for Natural Product Computational Pharmacology.

The taxonomy organizes natural-product data into source and identity, phytochemical and structural, bioassay and phenotypic, target and interaction, network and pathway, omics and molecular-state, pharmacokinetic and exposure, safety and toxicity, and clinical and translational signal classes. Cross-cutting dimensions distinguish observed, curated, inferred, predicted, text-derived, and clinically observed evidence; molecular-to-population biological scales; computational functions; provenance status; evidence maturity; and interoperability readiness. Directional connectors show that cross-layer integration is conditional on identity resolution, semantic alignment, metadata quality, version control, and uncertainty-aware provenance.

Alt-text description

A wide hierarchical systems taxonomy with nine major natural-product data classes arranged from source and identity data through phytochemical, bioassay, target, network, omics, pharmacokinetic, safety, and clinical signal layers. Cross-cutting bands classify each layer by evidence mode, biological scale, computational role, provenance, maturity, and interoperability readiness. Arrows show transitions between layers, while caution markers identify identifier mismatch, taxonomic ambiguity, assay heterogeneity, version mismatch, and propagation of predicted evidence.

Class boundaries are defined by the primary evidential object and inferential function, not merely by the database in which a record appears. A predicted target belongs in T4 because its primary assertion concerns a compound–target relationship, but it must retain the evidence status "predicted"; it does not move into a measured category when imported into a knowledge graph. An enriched pathway belongs in T5 because it is a derived pathway-level result and must not be reclassified as direct target evidence. A metabolomics measurement belongs in T6 when it represents molecular state, whereas a measured metabolite concentration used specifically to characterize systemic exposure may connect to T7 without erasing its analytical origin. An adverse-event signal belongs principally in T8 as safety evidence but may connect to T9 when interpreted within a clinically contextualized human evidence layer. Registry-based entity identification can support these mappings, yet entity resolution still requires explicit namespace and semantic control rather than assuming that matching labels denote the same object [25].

The taxonomy is multi-dimensional because a strictly hierarchical tree cannot represent evidence mode, biological scale, computational function, or maturity without losing information. Each top-level class is therefore cross-classified by evidence mode—observed, measured, curated, inferred, predicted, text-derived, model-derived, clinically observed, or supported real-world evidence where appropriate—and by biological scale from molecular to population. Computational roles include identification, representation, similarity analysis, target prediction, network construction, mechanism

inference, biomarker analysis, candidate prioritization, safety assessment, and clinical translation. Provenance is separately coded as primary experimental, curated secondary, integrated, predicted, text-derived, or unclear. Semantic alignment requires more than identical labels: mapping assertions should preserve relation semantics, provenance, and confidence context because ontology and entity mappings can themselves be uncertain or method-derived [26].

Evidence maturity is also orthogonal to data class. A T4 target record may be exploratory if predicted, experimentally supported if direct interaction has been measured, externally validated if replicated independently, or clinically contextualized if linked to relevant human evidence. A T6 omics observation may be experimentally measured yet remain hypothesis-generating with respect to mechanism. A T9 clinical signal may be close to human decision contexts while still having weak causal support because of confounding or reporting bias. The proposed maturity categories—exploratory, hypothesis-generating, experimentally supported, externally validated, clinically contextualized, and translation-relevant—should therefore be assigned from the evidence actually demonstrated rather than from the prestige of a source, the number of databases reproducing an assertion, or the downstream sophistication of a model. Aggregation must not automatically upgrade maturity.

Cross-layer integration is governed by five interoperability gates. Identity resolution requires persistent identifiers, synonym control, structure normalization, and taxonomic alignment. Metadata harmonization requires assay context, units, endpoints, and sample provenance. Semantic alignment requires explicit target identifiers, gene–protein mappings, pathway vocabularies, disease terminology, and typed relations. Version and provenance control requires source versions, update histories, evidence origins, and traceability of derived transformations. Uncertainty preservation requires measured and predicted evidence to remain distinguishable while recording missingness, confidence, conflict, and unresolved ambiguity. These gates explain why integration is conditional rather than automatic. A common graph schema can make heterogeneous data technically co-located while still preserving—or concealing—substantial semantic and evidential incompatibility.

The resulting taxonomy is intended as a reproducible conceptual framework, not as a universally validated ontology or fixed scoring system. Interoperability readiness can be classified as low, moderate, or high only in relation to a specified integration task and available metadata; no inherent numeric score is proposed. Ambiguous records should be assigned to the narrowest defensible class while retaining all known evidence modes and cross-class relations. Where provenance cannot be resolved, "unclear provenance" is preferable to inferred certainty. Where two classes legitimately overlap, the relation should be represented explicitly rather than forcing the record into one class and discarding context. This approach preserves the distinction between data availability, quality, interoperability, completeness, validity, provenance, relevance, and evidence maturity, thereby reducing the risk that integration itself becomes mistaken for verification.

Research and platform implications

For computational pharmacology platforms, the primary implication is that heterogeneous evidence should be integrated through typed, provenance-aware transformations rather than undifferentiated data pooling. Biomedical network integration has demonstrated that heterogeneous knowledge can support candidate prioritization, but downstream rankings remain dependent on the source relations and assumptions encoded upstream [27]. A natural-product platform should therefore retain the path by which a candidate moved from source identity to structure, assay, target, pathway, omics context, exposure, safety, and clinical signal. Every transformation should be auditable: a predicted compound occurrence should remain predicted; a target inferred from similarity should remain linked to its method; a pathway enrichment should retain its input features and version; and a clinical signal should preserve study or reporting context. This architecture supports prioritization without allowing computational depth to masquerade as independent validation.

Knowledge graphs offer one practical representation for such ecosystems because they can encode heterogeneous entities and typed relations while making cross-domain connections queryable. Pharmacological knowledge-graph resources demonstrate the utility of explicit drug, gene, disease, pathway, and relation structures for computational repurposing and related inference tasks [28]. For natural products, however, graph architecture must extend beyond generic drug nodes to include taxa, specimens where appropriate, preparations, extracts, fractions, structures, stereochemical states, analytical features, occurrence assertions, assays, exposures, metabolites, and signals. Edge typing is equally important: "reported in," "analytically detected in," "quantified in," "measured to bind," "predicted to target," "member of pathway," "enriched in," and "associated with adverse event" should not collapse into a generic association. Provenance graphs and evidence lineage should accompany the domain graph. Multimodal machine learning, graph neural networks, and other AI-enabled approaches can potentially exploit combinations of structures, spectra, text, targets, networks, omics profiles, and translational data, but natural-product-

specific analyses emphasize that multimodal integration requires careful data and knowledge representation [29]. Foundation-model approaches may become relevant where supported by sufficiently appropriate training data, but scale alone does not solve taxonomic ambiguity, uncertain structures, missing assay metadata, label leakage, source duplication, or non-independent evidence. A model trained on integrated resources can learn artefacts of database construction as readily as pharmacological regularities. Predicted-to-predicted chains are especially problematic: a predicted constituent may feed a predicted target model, whose outputs generate pathway features, which then become labels for another learning system. Without explicit lineage, such a system can present internally recycled hypotheses as multimodal convergence.

Future platform requirements therefore include persistent and namespace-aware identifiers, taxonomic provenance, explicit source-material and preparation descriptors, chemical structure standardization with stereochemical status, assay metadata, evidence typing, unit harmonization, version control, uncertainty fields, provenance graphs, cross-layer mappings, and auditable transformation histories. Silent identifier collapse should be prevented when different names, salts, stereoisomers, metabolites, or taxonomic concepts are merged. Database duplication should be examined at the assertion level rather than inferred from resource count. Label leakage should be assessed when knowledge used to construct training features overlaps with benchmark labels. Version drift should be visible because updated pathway memberships, taxonomic concepts, or database mappings can change model inputs without a change in model code. Mapping provenance and relation semantics should remain first-class data objects rather than hidden preprocessing artefacts.

The strategic shift is therefore from database aggregation toward provenance-aware evidence ecosystems, typed evidence graphs, uncertainty-preserving integration, versioned multimodal data layers, and translation-aware computational platforms. Research priorities within this direction include evaluating how upstream identity uncertainty changes downstream predictions; developing benchmark tasks that preserve evidence provenance; distinguishing independent corroboration from duplicated source lineage; testing model robustness to version changes and identifier remapping; representing mixtures without falsely assigning all effects to individual constituents; integrating exposure constraints into candidate prioritization; and designing validation pathways that require an explicit maturity transition before mechanistic or translational claims are upgraded. The proposed taxonomy provides a structure for these developments by making evidence type, provenance, interoperability, and translational proximity visible at the point of computation rather than only during retrospective interpretation.

CONCLUSION

This Original Taxonomy Article proposes a framework for mapping natural-product data ecosystems in computational pharmacology across source and identity, phytochemical and structural, bioassay and phenotypic, target and interaction, network and pathway, omics and molecular-state, pharmacokinetic and exposure, safety and toxicity, and clinical and translational signal layers. Its central premise is that computational value depends not only on data availability or volume but on identity resolution, provenance, metadata quality, evidence typing, interoperability, uncertainty preservation, validation, and translational context. The taxonomy organizes heterogeneous data through nine top-level classes and cross-cutting dimensions of evidence mode, biological scale, computational role, provenance, evidence maturity, interoperability readiness, uncertainty, validation status, and translational proximity. It also identifies conditional interoperability gates involving identity resolution, metadata harmonization, semantic alignment, version and provenance control, and uncertainty preservation. By maintaining distinctions among measured, curated, predicted, derived, and clinically observed evidence, the framework is intended to reduce false equivalence, circular evidence propagation, and silent upgrading of weak associations during cross-layer integration. The taxonomy is not presented as universally validated or exhaustive; rather, it provides an extensible conceptual architecture for designing, auditing, and interpreting natural-product databases, knowledge graphs, multimodal models, candidate-prioritization systems, safety analyses, and translation-aware computational platforms.

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] 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

[3] Romano JD, Tatonetti NP. Informatics and computational methods in natural product drug discovery: A review and perspectives. Front Genet. 2019;10:368. doi:10.3389/fgene.2019.00368

[4] Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen EL, Gaudry A, et al. The LOTUS initiative for open knowledge management in natural products research. Elife. 2022;11:e70780. doi:10.7554/eLife.70780

[5] Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 2018;46(D1):D1217-22. doi:10.1093/nar/gkx1026

[6] 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

[7] van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An open access knowledge base for microbial natural products discovery. ACS Cent Sci. 2019;5(11):1824-33. doi:10.1021/acscentsci.9b00806

[8] Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, et al. IMPPAT: A curated database of Indian medicinal plants, phytochemistry and therapeutics. Sci Rep. 2018;8(1):4329. doi:10.1038/s41598-018-22631-z

[9] Vivek-Ananth RP, Mohanraj K, Sahoo AK, Samal A. IMPPAT 2.0: An enhanced and expanded phytochemical atlas of Indian medicinal plants. ACS Omega. 2023;8(9):8827-45. doi:10.1021/acsomega.3c00156

[10] Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based molecular networking in the GNPS analysis environment. Nat Methods. 2020;17(9):905-8. doi:10.1038/s41592-020-0933-6

[11] Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol. 2021;39(4):462-71. doi:10.1038/s41587-020-0740-8

[12] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930-40. doi:10.1093/nar/gky1075

[13] Liu T, Hwang L, Burley SK, Nitsche CI, Southan C, Walters WP, et al. BindingDB in 2024: A FAIR knowledgebase of protein-small molecule binding data. Nucleic Acids Res. 2025;53(D1):D1633-44. doi:10.1093/nar/gkae1075

[14] Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: Supporting systematic drug-target identification and prioritisation. Nucleic Acids Res. 2021;49(D1):D1302-10. doi:10.1093/nar/gkaa1027

[15] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-13. doi:10.1093/nar/gky1131

[16] Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The Reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687-92. doi:10.1093/nar/gkab1028

[17] Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111. doi:10.1186/s13059-020-02015-1

[18] Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752. doi:10.1371/journal.pcbi.1005752

[19] Grzegorzewski J, Brandhorst J, Green K, Eleftheriadou D, Duport Y, Barthorscht F, et al. PK-DB: Pharmacokinetics database for individualized and stratified computational modeling. Nucleic Acids Res. 2021;49(D1):D1358-64. doi:10.1093/nar/gkaa990

[20] Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. J Cheminform. 2017;9(1):61. doi:10.1186/s13321-017-0247-6

[21] Sharma V, Gelin LFF, Sarkar IN. Identifying herbal adverse events from spontaneous reporting systems using taxonomic name resolution approach. Bioinform Biol Insights. 2020;14:1177932220921350. doi:10.1177/1177932220921350

[22] Dilán-Pantojas IO, Boonchalermvichien T, Taneja SB, Li X, Chapin MR, Karcher S, et al. Broadening the capture of natural products mentioned in FAERS using fuzzy string-matching and a Siamese neural network. Sci Rep. 2024;14(1):1272. doi:10.1038/s41598-023-51004-4

[23] Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al. FAIRsharing as a community approach to standards, repositories and

policies. Nat Biotechnol. 2019;37(4):358-67. doi:10.1038/s41587-019-0080-8

[24] Wimalaratne SM, Juty N, Kunze J, Janée G, McMurry JA, Beard N, et al. Uniform resolution of compact identifiers for biomedical data. Sci Data. 2018;5:180029. doi:10.1038/sdata.2018.29

[25] Hoyt CT, Balk M, Callahan TJ, Domingo-Fernández D, Haendel MA, Hegde HB, et al. Unifying the identification of biomedical entities with the Bioregistry. Sci Data. 2022;9(1):714. doi:10.1038/s41597-022-01807-3

[26] Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database (Oxford). 2022;2022:baac035. doi:10.1093/database/baac035

[27] Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017;6:e26726. doi:10.7554/eLife.26726

[28] Boudin M, Diallo G, Drancé M, Mougin F. The OREGANO knowledge graph for computational drug repurposing. Sci Data. 2023;10(1):871. doi:10.1038/s41597-023-02757-0

[29] Meijer D, Beniddir MA, Coley CW, Mejri YM, Öztürk M, van der Hooft JJJ, et al. Empowering natural product science with AI: Leveraging multimodal data and knowledge graphs. Nat Prod Rep. 2025;42(4):654-62. doi:10.1039/D4NP00008K