Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2026, Volume 6, Issue 1, Page No: 12-21

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Natural-Product Models Learn What Databases Choose to Record: A Data-Governance Architecture for Structures, Spectra, Bioassays, Taxonomy, Provenance, and Negative Results

Gabriel Costa[1]*, Lucas Ribeiro[1], Ricardo Alves[2], Mariana Lopes[3]

[1] Department of Natural-Product Data Governance and Provenance, Faculty of Pharmacy, Federal University of Minas Gerais, Belo Horizonte, Brazil.
[2] Department of Spectral and Bioassay Database Architecture, Faculty of Pharmacy, University of Coimbra, Coimbra, Portugal.
[3] Department of Taxonomy and Negative-Result Reporting, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil.

***E-mail** ✉ gabriel.costa@ufmg.br

Received: 15 October 2025; **Revised:** 08 January 2026; **Accepted:** 10 January 2026

ABSTRACT

Natural-product machine-learning models do not encounter molecules, spectra, organisms, or biological responses directly; they encounter database records produced through choices about what is deposited, normalized, linked, corrected, omitted, and retained. This article develops an original data-governance architecture for natural-product modeling that treats these choices as part of the inferential system rather than as background data-management operations. The analysis integrates structural identity, spectral evidence, bioassay context, taxonomy, provenance, missingness, duplication, record quality, correction, versioning, and negative results while preserving distinctions between chemical identity and material provenance, analytical association and structural proof, assay observations and context-independent biological properties, and observed negatives and untested or generated examples. The central contribution is a proposed governance architecture in which record lineage, evidence state, contextual metadata, correction history, and task-specific fitness remain separable but interoperable. Quality is therefore treated not as a single database-wide score but as a conditional relationship between a record state and the modeling task for which it is used. The framework further argues that model evaluation should account for coverage, redundancy, source dependence, assay context, and version structure because nominally independent test data may reproduce the same curation choices as training data. The architecture is conceptual rather than prospectively validated. Its practical value therefore depends on record-level auditability, reproducible state assignment, cross-resource reconciliation, and future benchmarking against external and temporally separated data.

Keywords: Natural products, Data governance, Chemical databases, Provenance, Bioactivity data, Spectral data

How to Cite This Article: Costa G, Ribeiro L, Alves R, Lopes M. Natural-Product Models Learn What Databases Choose to Record: A Data-Governance Architecture for Structures, Spectra, Bioassays, Taxonomy, Provenance, and Negative Results. Spec J Pharmacogn Phytochem Biotechnol. 2026;6(1):12-21. https://doi.org/10.51847/jujwJdusFg

Introduction

Machine learning is increasingly positioned to combine molecular structures, spectra, biological measurements, and contextual knowledge in natural-product research, yet these inputs do not enter computational systems as untouched observations. They arrive as records whose content and relationships have already been shaped by deposition, curation, standardization, annotation, and database design. Contemporary natural-product informatics therefore confronts a problem that is broader than model architecture: heterogeneous analytical and biological evidence must first be made computable, linkable, and sufficiently explicit for reuse [1]. Database integrity and provenance are consequently not merely administrative concerns. They determine whether a modeled object can

be traced to an experimental source, whether transformations applied during curation remain visible, and whether later users can distinguish primary observations from derived representations [2].

Natural-product databases themselves demonstrate that database content is dynamic rather than fixed. COCONUT 2.0 required a substantial overhaul and renewed curation of aggregated natural-product records, illustrating that large collections depend on continuing normalization and quality-control decisions [3]. The Natural Products Atlas 3.0 likewise incorporates revised and reassigned structures alongside database expansion, showing that a chemically plausible and previously published entry may later acquire a different evidential state [4]. Such revisions should not be interpreted as evidence that databases are generally unreliable; they demonstrate instead that record identity, confidence, and version history are scientifically relevant properties.

The problem for modeling follows from this mediation. A molecular representation may be internally standardized while its source material remains poorly identified; an assay value may be numerically precise while its biological context is incompletely represented; and an analytical match may support an annotation without constituting definitive structural proof. Conversely, a record excluded because it lacks a preferred field may still contain useful evidence for another task. This article therefore proposes that natural-product data governance should operate at the level of record states and relationships, not only database-level quality. The aim is to establish an evidence-bounded architecture in which structures, spectra, bioassays, taxonomy, provenance, negative results, corrections, and modeling fitness can be governed without collapsing their distinct meanings.

How natural-product data become training data

Natural-product information becomes model-ready through a sequence of transformations rather than a single act of database retrieval. NPASS illustrates how natural-product structures can be linked with composition, bioactivity, source, and ADME–Tox information within one resource [5]. Broader infrastructures such as PubChem further distinguish depositor-level substances from standardized compound entities and associated assay records, making clear that source identity and normalized chemical identity are not interchangeable objects [6].

Bioactivity data undergo additional mediation. ChEMBL integrates measurements from heterogeneous experimental sources into curated activity records that retain assay and endpoint information [7]. BindingDB similarly organizes quantitative protein–small-molecule binding measurements while preserving links to experimental provenance [8]. These infrastructures support reuse, but neither implies that values produced under different protocols are automatically commensurable.

Analytical data introduce another layer. Pan-repository metabolomics reanalysis requires harmonized sample and experimental metadata before spectra can be compared or reused across collections [9]. NP-MRD, meanwhile, brings together natural-product NMR information with evidence modes that may include experimentally acquired and computationally generated spectral information [10]. A governance system should therefore preserve not only a spectrum but also how that spectrum was produced and what inferential role it can legitimately play.

Traceability mechanisms reinforce this distinction. Structured deposition of natural-product NMR data permits analytical objects to remain accessible beyond prose descriptions [11]. Universal Spectrum Identifiers provide persistent references to individual mass spectra without asserting that the associated interpretation is correct [12]. ReDU demonstrates how controlled metadata can make public mass-spectrometry data discoverable and reanalyzable across studies [13].

Taxonomic and bibliographic relationships add further transformations. LOTUS explicitly links natural-product occurrence information to organismal and source provenance [14], whereas the original COCONUT resource shows the integration burden created when structures are aggregated from multiple predecessor databases [15]. Successive Natural Products Atlas releases demonstrate that such collections evolve through versioned curation rather than simple accumulation [16]. Machine-learning-assisted structure annotation further illustrates that spectral evidence may narrow candidate structures without automatically converting an annotation into structural proof [17]. Finally, Papyrus shows explicitly that prediction-ready bioactivity datasets are constructed through curation and standardization of heterogeneous source records [18]. Training data are therefore not raw observations; they are governed derivatives of observations.

Bias across chemical, biological, and provenance records

Bias enters natural-product modeling before an algorithm is fitted because different record dimensions are curated with different completeness and precision. Assay records are a prominent example. Recent work combining artificial intelligence with manual curation of ChEMBL assays demonstrates that biological context itself requires interpretation and annotation rather than being fully captured by a molecular structure and an activity number [19]. Measurement provenance introduces a related but distinct problem: combining IC50 or Ki values from different sources can introduce substantial noise, even when the endpoint labels appear superficially compatible [20]. The relevant bias is therefore not simply "bad data"; it may arise from valid measurements being treated as more interchangeable than their experimental contexts justify.

Negative evidence is similarly heterogeneous. The European Chemical Biology Database preserves experimentally generated screening information that includes inactive as well as active observations [21]. InertDB, by contrast, expands an inactive-molecule resource through a combination of curated information and generative methods [22]. These resources may each be useful, but they do not represent the same evidential state. A measured inactive result is conditional on a defined assay, whereas a generated inactive example has computational provenance and should not silently inherit the status of an experimental observation.

Chemical-space representation produces a third form of bias. Coverage bias can make performance appear stronger for molecules that occupy well-represented regions of the data distribution than for poorly covered regions [23]. Binding-affinity prediction studies likewise show that dataset bias and redundancy can alter apparent generalization and that controlling these properties changes the difficulty of the prediction problem [24]. Such findings do not establish a universal magnitude of bias for natural-product models, but they support treating coverage, redundancy, and source dependence as properties requiring explicit governance rather than as nuisances discoverable only after model failure.

The distinctions established in this section are organized in **Table 1** to separate what is recorded from the different modeling distortions that can arise when those record types are treated as equivalent.

Table 1. Record dimensions that can generate distinct pre-modeling biases in natural-product datasets

Record dimension	What the record represents	Principal distortion if collapsed	Governance requirement
Chemical identity	Standardized representation of a molecular entity	Different source materials or uncertain assignments appear chemically equivalent	Preserve links between normalized structure and source-level identity
Assay context	Measurement conditions, target system, endpoint, and protocol	Numerically similar activities are treated as directly comparable	Retain assay and measurement provenance with the value
Spectral evidence	Analytical observation or computationally generated spectrum	Evidence-generation modes become indistinguishable	Preserve modality, acquisition status, and analytical lineage
Taxonomic occurrence	Association between a chemical record and a biological source	Synonyms, reclassification, or weak source identification propagate as fixed labels	Maintain organismal provenance separately from chemical identity
Negative evidence	Inactive, failed, absent, untested, or generated state	Distinct evidential states become a single negative class	Encode how the negative state was established
Dataset coverage	Distribution of represented chemistry and contexts	Dense regions dominate learning and evaluation	Audit coverage and redundancy relative to the intended modeling task

Missingness, duplication, and unequal data quality

Natural-product datasets also differ in what is absent, repeated, or represented with unequal confidence. These properties should not be merged into a generic "data quality" label. Comparative machine-learning work shows that changing curation stringency and confidence levels can alter resulting predictions, demonstrating that record-selection rules become part of the computational experiment [25]. The appropriate response is not necessarily maximal exclusion: aggressive filtering can remove uncertain observations, but it may also narrow chemical or biological coverage. Quality must therefore be interpreted in relation to the intended task.

Cheminformatics curation encompasses multiple operations, including normalization, validation, deduplication, and treatment of inconsistent records [26]. Reproducible structure-curation pipelines make these transformations

explicit and computationally repeatable [27]. However, normalization and truth are different concepts. A record can be syntactically standardized yet still reflect a disputed stereochemical assignment, incorrect source attribution, or incomplete evidence. Conversely, two records that standardize to the same molecular graph may remain scientifically nonredundant when they represent different organisms, specimens, assay conditions, or analytical observations.

Missingness is equally heterogeneous. An unreported assay condition, absent stereochemical specification, unavailable raw spectrum, unlinked taxonomic identifier, and compound never tested in an assay are not equivalent missing values. Their mechanisms may differ, and the available evidence does not justify assuming that all natural-product missingness is missing not at random. Governance should instead encode what field is missing, whether absence reflects nonmeasurement or nonreporting, and whether the missing field is required for the intended model.

Structured, machine-readable chemical reporting provides one route for reducing ambiguity introduced during later extraction and reuse [28]. Nevertheless, structured representation cannot recover evidence that was never recorded. The proposed architecture therefore treats missingness, duplication, and confidence as separable record properties whose relevance depends on downstream use.

Table 2 distinguishes these defect classes and the corresponding actions that are justified without implying that every uncertain record should be discarded.

Table 2. Analytical distinctions among missingness, duplication, and unequal record quality

Data condition	Diagnostic question	Why it matters for modeling	Appropriate governance response	What should not be inferred
Missing field	Was the property unmeasured, unreported, unavailable, or lost during integration?	Different mechanisms of absence imply different risks	Record the missing field and, where known, its reason	Missingness is automatically informative or random
Apparent duplicate	Do records share a normalized structure but differ in source, assay, spectrum, or version?	Naive deduplication can erase legitimate contextual evidence	Deduplicate identity only at the appropriate analytical level	Same molecular graph means scientifically identical record
Conflicting records	Do sources assign different structures, activities, taxa, or interpretations?	A single consensus value can hide real evidential disagreement	Preserve alternatives, provenance, and resolution status	One source must be correct without further evidence
Standardized but uncertain structure	Has formatting been normalized without resolving assignment confidence?	Models may treat a clean representation as established identity	Keep normalization state separate from evidential confidence	Standardization constitutes structural validation
Unequal assay quality	Are measurements supported by different contextual completeness or confidence?	Filtering choices can change model behavior	Preserve quality attributes for task-specific selection	Stricter filtering is always superior
Incomplete provenance	Can the record be traced to its analytical, biological, or bibliographic source?	Corrections and contextual interpretation become difficult	Retain lineage identifiers and unresolved provenance explicitly	Lack of provenance proves the underlying observation is false

A governance architecture for natural-product data

The preceding evidence suggests that natural-product data governance cannot be reduced to a single cleaning step applied before modeling. Database integrity, natural-product curation, and cross-repository metadata harmonization instead indicate linked problems of traceability, transformation, and reuse [2, 3, 9]. This article therefore proposes a governance architecture organized around four separable record properties: identity, evidence, context, and lineage. Identity specifies what chemical or biological entity a record claims to represent; evidence specifies what observation or derivation supports it; context records the analytical, assay, taxonomic, or

experimental conditions under which that evidence was produced; and lineage records how the entry was imported, standardized, linked, corrected, or superseded.

These properties should remain interoperable without being collapsed. Taxonomic occurrence information can connect molecules to organisms and publications, but those links are distinct from assay context and should remain separately queryable [14]. Likewise, assay annotation is a contextual property rather than an intrinsic property of the molecular graph [19]. At the chemical layer, reproducible normalization is valuable because transformations can be made explicit and repeatable [27], but a normalized structure can still later require reassignment when new evidence changes the accepted interpretation [4]. The architecture therefore treats normalization state and evidence confidence as different dimensions.

A second principle is lineage preservation. Structured, machine-readable chemistry facilitates reuse, but downstream usability should not erase the source record or transformation history from which a standardized object was derived [28]. The proposed architecture consequently favors appendable governance events—such as normalization, linkage, confidence reassessment, correction, or deprecation—over silent replacement. This is not a claim that every database must expose an identical technical schema. It is a conceptual requirement that a model-ready object remain traceable to the evidence and transformations that made it model-ready.

A third principle is conditional transfer. Records may move from deposition to database integration and then into a modeling dataset only when the receiving task can interpret their evidential state correctly. A spectrum with incomplete acquisition metadata, an activity value without sufficient assay context, or a source organism recorded only at a high taxonomic rank may be suitable for some tasks and unsuitable for others. Governance therefore controls transfer between evidence states rather than merely separating "clean" from "unclean" data.

Quality states and data fitness for modeling

The proposed architecture treats quality as task conditional. Comparative bioactivity modeling shows that the balance between data quantity and curation quality can change predictive behavior rather than following a universal rule that either larger or more stringently filtered datasets are always superior [29]. Correspondingly, a QSAR claim is bounded by the dataset used to define and evaluate it [30]. These observations support a distinction between record quality and data fitness: quality describes properties of the record and its evidence, whereas fitness asks whether those properties are adequate for a specified modeling use.

No single scalar quality score is assumed to be sufficient. Model-free approaches to dataset quality provide useful formal alternatives for describing dataset composition [31], while assay-aware modeling emphasizes that biological context can itself be prediction relevant [32]. The proposed framework therefore uses multidimensional quality states. A record can be structurally well standardized but analytically weakly supported; biologically informative but incompletely contextualized; taxonomically precise but spectrally absent; or highly traceable yet outside the chemical space required by a particular model. Such states should be queryable rather than collapsed into an overall "high-quality" label.

Fitness decisions should also remain reversible. When a task changes—from spectral retrieval to bioactivity prediction, for example—the relevant evidence requirements change with it. Cross-source measurement heterogeneity can alter the reliability of apparently equivalent activity labels [20], and confidence-based curation can affect resulting model behavior [25]. The architecture therefore proposes that dataset assembly declare both an inclusion state and a reason: accepted for the current task, retained but excluded, unresolved pending additional evidence, or unsuitable because a required evidence dimension is absent. These are governance decisions, not claims about the intrinsic scientific worth of a record.

Table 3 converts this task-conditional logic into a compact decision workflow that keeps record state, modeling use, and the permissible inference separate.

Table 3. Proposed task-conditional workflow for deciding natural-product data fitness for modeling

Decision step	Question asked	Possible governance state	Modeling consequence
Identity check	Is the modeled entity sufficiently specified for the task?	Accepted; ambiguous; conflicting; unresolved	Use only when the representation matches the intended prediction unit

Evidence check	What observation or derivation supports the record?	Direct observation; analytical assignment; curated derivation; generated state	Preserve evidence type so downstream labels are not treated as equivalent
Context check	Are assay, spectral, taxonomic, or experimental conditions adequate?	Context sufficient; partial; missing; not applicable	Restrict or stratify use when context affects label meaning
Lineage check	Can source and transformations be reconstructed?	Traceable; partially traceable; provenance gap	Flag records whose correction or duplication history cannot be audited
Coverage check	Does the record occupy the domain relevant to the model?	In-domain; boundary; out-of-domain; unknown	Separate fitness from intrinsic quality and avoid unsupported extrapolation
Task decision	Are required states adequate for this use?	Include; retain but exclude; unresolved; unsuitable	Record the decision and reason so dataset assembly remains reproducible

The interactions among record formation, bias, governance, quality states, evaluation, and implementation are summarized in **Figure 1** as a proposed evidence-transfer architecture.

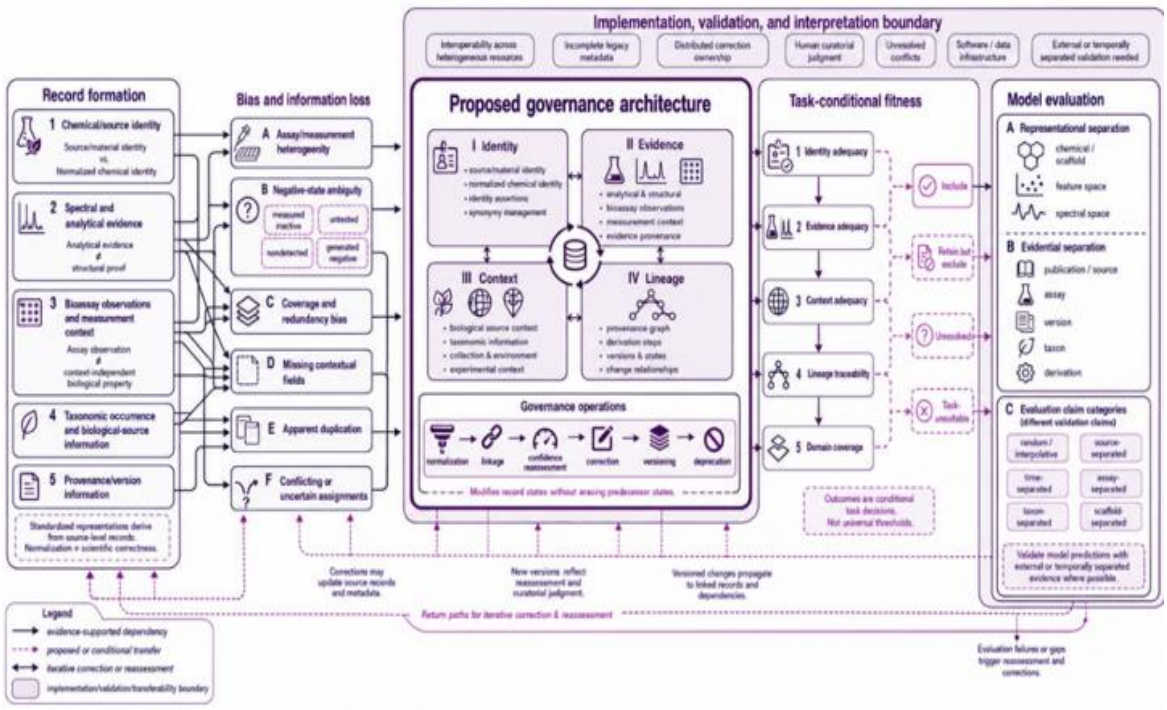

Figure 1. From recorded evidence to model claims: a proposed nested governance architecture for natural-product data

Correction, versioning, and negative evidence

Correction should be represented as a scientific event, not simply as replacement of one database value by another. Successive natural-product database releases demonstrate that curation can alter structural or associated records over time [3]. The Natural Products Atlas provides a particularly direct example because later releases incorporate revised or reassigned structures [4]. The proposed architecture therefore stores the prior state, correction event, rationale or source where available, and successor state as linked objects. This preserves the ability to reconstruct which version of a record was available when a model was trained or evaluated.

Negative evidence requires an equally explicit state model. Experimentally recorded inactive screening results and generatively expanded inactive resources differ in provenance and should not be collapsed into one undifferentiated negative class [21, 22]. An experimentally inactive compound is inactive only under the conditions actually tested. An untested compound has no such observation. A nondetected metabolite concerns an

analytical measurement rather than necessarily a biological phenotype. A generated inactive example is a computationally derived training object and must remain distinguishable from experimental evidence.

Negative observations should nevertheless not be discarded by default. In reaction prediction, inclusion of negative chemical data has been shown to improve performance under the evaluated setting [33]. That result does not demonstrate equivalent benefit for natural-product bioassays, but it supports the narrower principle that failed or negative observations can contain learnable information when their provenance and meaning are explicit.

Table 4 defines the boundaries that must be maintained when correction history and negative evidence enter modeling datasets.

Table 4. Proposed evidence boundaries for correction, versioning, and negative natural-product records

Evidence state	What is known	Required boundary	Validation need
Corrected record	A prior representation was superseded or revised	Preserve historical and successor states with lineage	Verify correction propagation across linked resources
Conflicting record	More than one supported interpretation remains	Do not force consensus without additional evidence	Reconcile against primary analytical or source evidence
Measured inactive	No activity met the assay criterion under tested conditions	Do not generalize beyond assay context	Replicate or test under relevant alternative conditions when required
Untested	No relevant experiment is recorded	Do not encode as inactive	Obtain measurement or retain as unknown
Analytical nondetection	Signal was not detected under defined analytical conditions	Do not equate with chemical absence or biological inactivity	Assess detection limits, sampling, and acquisition context
Generated negative	Negative-like example was computationally derived	Keep generative provenance visible	Evaluate calibration and transfer against experimental negatives

Consequences for model evaluation

A governance architecture changes model evaluation because test-set independence cannot be assessed from row separation alone. Chemical-space coverage and dataset redundancy or related bias can affect apparent generalization [23, 24]. A test molecule may be absent from training while remaining surrounded by close analogues, sharing the same measurement source, inheriting a duplicated literature record, or reflecting the same curation rule. Evaluation therefore needs to ask not only whether records were split, but which dependencies survived the split.

The proposed approach distinguishes representational separation from evidential separation. Representational separation concerns molecular, scaffold, spectral, or other feature similarity. Evidential separation concerns shared publications, assays, source laboratories, database ancestors, correction histories, taxonomic sources, or derived labels. These dependencies need not all be removed for every benchmark; instead, the evaluation design should state which form of generalization is being tested. A random split may estimate interpolation within a curated data environment, whereas a source-held-out or time-aware test addresses a different claim.

A second consequence concerns benchmark construction from merged resources. If a standardized compound appears through several source databases or publications, a row-level split can conceal shared ancestry even after obvious duplicates are removed. The proposed architecture therefore treats provenance overlap as a potential dependency alongside structural similarity. This does not mean that every shared source invalidates a benchmark; it means that the permissible claim should match the separation actually achieved. For example, a chemistry-held-out test addresses a different uncertainty from an assay-held-out or publication-held-out test.

Validation strategy should consequently resemble intended use. SIMPD demonstrates one method for constructing simulated time splits that better approximate historical medicinal-chemistry progression than convenient random partitions in its evaluated context [34]. For natural-product models, analogous strategies could hold out later database versions, newly deposited spectra, source publications, taxa, assay families, or chemical scaffolds, depending on the scientific question. These adaptations are proposed and require testing rather than being assumed valid.

Performance should finally be interpreted alongside the state of the underlying data. Curation strategy can affect useful training information [29], while assay context can determine whether the same molecular label means the same biological task [32]. A governance-aware evaluation report should therefore specify dataset version,

inclusion rules, unresolved evidence states, provenance separation, coverage limitations, and the validation claim being made. High internal performance without these boundaries may still be useful, but it should not be presented as evidence of unrestricted external generalization.

Implementation challenges

Implementation need not require a single centralized natural-product database. A federated design could allow resources to retain domain-specific schemas while exposing compatible identity, evidence-state, provenance, and version information. Such an approach is consistent with broader chemical-database integrity principles [2] and with cross-repository metabolomics workflows that depend on harmonized metadata rather than elimination of repository boundaries [9]. The practical challenge is to define interoperability tightly enough for machine use without pretending that heterogeneous evidence has become homogeneous.

Curation burden is another constraint. Explicit workflows improve reproducibility, but different resources have different staffing, evidence access, and update cycles [26]. Structured chemical reporting can reduce later extraction ambiguity [28], yet legacy natural-product literature will continue to contain unstructured or incompletely digitized evidence. A realistic architecture must therefore tolerate partial states and provenance gaps rather than requiring complete metadata before any record can be retained.

A further challenge is governance ownership. Corrections can originate with depositors, database curators, later structural studies, or downstream users who identify conflicts. The architecture therefore requires attributable correction proposals and resolvable successor states rather than assuming one institution can adjudicate every dispute. Community-specific authority can remain distributed while the history of disagreement stays machine-readable and auditable.

Computational metabolomics continues to face interoperability, software, annotation, and data-management challenges despite rapid methodological development [35]. Similar constraints are likely to affect implementation of the proposed architecture. Persistent identifiers, controlled vocabularies, correction logs, and machine-readable provenance can make governance more tractable, but their adoption does not itself establish scientific correctness. Human curatorial judgment remains necessary for disputed structures, ambiguous taxonomy, conflicting assays, and evidence that cannot be reduced to a deterministic rule.

The architecture therefore should be tested incrementally. Priorities include intercurator agreement for governance states, cross-database entity reconciliation, correction-propagation audits, experimentally anchored evaluation of negative-data states, and benchmarks in which source, time, taxon, scaffold, or assay context is deliberately separated. Until such tests are performed, the framework should be regarded as an auditable design proposal rather than an implementation standard.

Conclusion

Natural-product machine learning begins long before model training. Decisions about which structures are registered, which spectra remain traceable, which assay conditions are represented, which taxonomic links are preserved, which corrections remain visible, and which negative observations are retained define the informational environment from which prediction is possible. This article proposes a data-governance architecture that makes those decisions explicit by separating identity, evidence, context, lineage, quality state, and task-specific fitness.

The central implication is that data quality should not be reduced to a single score or a universal filtering rule. A record may be useful for one computational purpose and unsuitable for another, while apparently clean records may still conceal uncertain assignments, missing context, inherited duplication, or weak provenance. Likewise, negative evidence is useful only when its origin and meaning remain distinguishable from absence, nondetection, or generated examples. This shifts the central question from whether a database is globally “good” to whether its records are sufficiently characterized for the claim a model is asked to make.

The architecture is intentionally evidence bounded. It does not establish that governance improvements will necessarily increase predictive accuracy, eliminate bias, or guarantee external validity. Its value instead lies in making otherwise hidden data transformations inspectable and testable. Future work should determine whether provenance-aware, version-aware, context-aware, and task-conditional dataset construction produces more reliable evaluation and more transferable natural-product models across independent resources, assays, taxa, chemical spaces, and time.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Meijer D, Beniddir MA, Coley CW, Mejri YM, Ozturk 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
2. Williams AJ, Richard AM. Three pillars for ensuring public access and integrity of chemical databases powering cheminformatics. J Cheminform. 2025;17(1):40. doi:10.1186/s13321-025-00983-9
3. Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res. 2025;53(D1):D634-43. doi:10.1093/nar/gkae1063
4. Poynton EF, van Santen JA, Pin M, Macias Contreras M, McMann E, Parra J, et al. The Natural Products Atlas 3.0: Extending the database of microbially derived natural products. Nucleic Acids Res. 2025;53(D1):D691-9. doi:10.1093/nar/gkae1093
5. Lin H, Hou D, Jiang X, Yin R, Yu S, Lu S, et al. NPASS database update 2026: Comprehensive quantitative composition, bioactivity, and ADME-Tox data of natural products for biomedical research. Nucleic Acids Res. 2026;54(D1):D1519-27. doi:10.1093/nar/gkaf1196
6. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2025 update. Nucleic Acids Res. 2025;53(D1):D1516-25. doi:10.1093/nar/gkae1059
7. Zdrazil B, Félix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1):D1180-92. doi:10.1093/nar/gkad1004
8. 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
9. El Abiead Y, Strobel M, Payne T, Fahy E, O'Donovan C, Subramamiam S, et al. Enabling pan-repository reanalysis for big data science of public metabolomics data. Nat Commun. 2025;16(1):4838. doi:10.1038/s41467-025-60067-y
10. Wishart DS, Sajed T, Pin M, Poynton EF, Goel B, Lee BL, et al. The Natural Products Magnetic Resonance Database (NP-MRD) for 2025. Nucleic Acids Res. 2025;53(D1):D700-8. doi:10.1093/nar/gkae1067
11. Pin M, Poynton EF, Jordan T, Kim J, Ledingham B, van Santen JA, et al. A data deposition platform for sharing nuclear magnetic resonance data. J Nat Prod. 2023;86(11):2554-61. doi:10.1021/acs.jnatprod.3c00795
12. Deutsch EW, Perez-Riverol Y, Carver J, Kawano S, Mendoza L, Van Den Bossche T, et al. Universal Spectrum Identifier for mass spectra. Nat Methods. 2021;18(7):768-70. doi:10.1038/s41592-021-01184-6
13. Jarmusch AK, Wang M, Aceves CM, Advani RS, Aguirre S, Aksenov AA, et al. ReDU: A framework to find and reanalyze public mass spectrometry data. Nat Methods. 2020;17(9):901-4. doi:10.1038/s41592-020-0916-7
14. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al. The LOTUS initiative for open knowledge management in natural products research. Elife. 2022;11:e70780. doi:10.7554/eLife.70780
15. 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
16. van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, et al. The Natural Products Atlas 2.0: A database of microbially-derived natural products. Nucleic Acids Res. 2022;50(D1):D1317-23. doi:10.1093/nar/gkab941

17. Hu G, Qiu M. Machine learning-assisted structure annotation of natural products based on MS and NMR data. Nat Prod Rep. 2023;40(11):1735-53. doi:10.1039/D3NP00025G
18. Béquignon OJM, Bongers BJ, Jespers W, IJzerman AP, van de Water B, van Westen GJP. Papyrus: A large-scale curated dataset aimed at bioactivity predictions. J Cheminform. 2023;15(1):3. doi:10.1186/s13321-022-00672-x
19. Smit I, Adasme MF, Manners EJ, Corbett S, Bosc N, Do HMA, et al. Integrating artificial intelligence and manual curation to enhance bioassay annotations in ChEMBL. J Cheminform. 2026;18(1):24. doi:10.1186/s13321-026-01165-x
20. Landrum GA, Riniker S. Combining IC50 or Ki values from different sources is a source of significant noise. J Chem Inf Model. 2024;64(5):1560-7. doi:10.1021/acs.jcim.4c00049
21. Škuta C, Müller T, Voršilák M, Popr M, Epp T, Skopelitou KE, et al. ECBD: European chemical biology database. Nucleic Acids Res. 2025;53(D1):D1383-92. doi:10.1093/nar/gkae904
22. An S, Lee Y, Gong J, Hwang S, Park IG, Cho J, et al. InertDB as a generative AI-expanded resource of biologically inactive small molecules from PubChem. J Cheminform. 2025;17(1):49. doi:10.1186/s13321-025-00999-1
23. Kretschmer F, Seipp J, Ludwig M, Klau GW, Böcker S. Coverage bias in small molecule machine learning. Nat Commun. 2025;16(1):554. doi:10.1038/s41467-024-55462-w
24. Graber D, Stockinger P, Meyer F, Mishra S, Horn C, Buller R. Resolving data bias improves generalization in binding affinity prediction. Nat Mach Intell. 2025;7(10):1713-25. doi:10.1038/s42256-025-01124-5
25. Xerxa E, Vogt M, Bajorath J. Influence of data curation and confidence levels on compound predictions using machine learning models. J Chem Inf Model. 2024;64(24):9341-9. doi:10.1021/acs.jcim.4c01573
26. Esaki T, Ikeda K. Data curation in cheminformatics: Importance and implementation. J Cheminform. 2026;18(1):43. doi:10.1186/s13321-026-01174-w
27. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12(1):51. doi:10.1186/s13321-020-00456-1
28. Mercado R, Kearnes SM, Coley CW. Data sharing in chemistry: Lessons learned and a case for mandating structured reaction data. J Chem Inf Model. 2023;63(14):4253-65. doi:10.1021/acs.jcim.3c00607
29. Schiebroek CCG, Landrum GA, Riniker S. Balancing data quantity and quality: Evaluating curation strategies for bioactivity prediction in lead optimization. J Chem Inf Model. 2026;66(13):7446-52. doi:10.1021/acs.jcim.6c01018
30. Nael MA, Alakonda LM, Elokely KM. Defining the data set defines the QSAR claim. J Chem Inf Model. 2026;66(6):2951-4. doi:10.1021/acs.jcim.6c00514
31. Xie C, Li Y. DCC: A model-free frame to evaluate data set quality. J Chem Inf Model. 2026;66(3):1457-65. doi:10.1021/acs.jcim.5c02429
32. Schoenmaker L, Sastrokarijo EG, Heitman LH, Beltman JB, Jespers W, van Westen GJP. Toward assay-aware bioactivity model(er)s: Getting a grip on biological context. J Chem Inf Model. 2025;65(13):7013-23. doi:10.1021/acs.jcim.5c00603
33. Toniato A, Vaucher AC, Laino T, Graziani M. Negative chemical data boosts language models in reaction outcome prediction. Sci Adv. 2025;11(24):eadt5578. doi:10.1126/sciadv.adt5578
34. Landrum GA, Beckers M, Lanini J, Schneider N, Stiefl N, Riniker S. SIMPD: An algorithm for generating simulated time splits for validating machine learning approaches. J Cheminform. 2023;15(1):119. doi:10.1186/s13321-023-00787-9
35. Domingo-Fernández D, Healey D, Kind T, Allen A, Colluru V, Misra BB. Trends in computational metabolomics: A perspective on five years of software development, challenges, and opportunities (2021-2025). Anal Chem. 2026;98(27):19962-74. doi:10.1021/acs.analchem.6c00361