Data Quality as a Bottleneck in AI-Driven Natural Product Discovery: Bias, Completeness, Assay Reliability, Chemical Annotation, and Model Generalizability

Victor Santos[1]*, Ana Beatriz[2], Joaquim Dias[1]

[1]Department of AI for Phytopharmaceutical Quality Control and Fingerprinting, Faculty of Pharmacy, University of Porto, Porto, Portugal.
[2]Department of Computational Metabolomics for Herbal Standardization, Faculty of Pharmaceutical Sciences, University of Campinas, Campinas, Brazil.

ABSTRACT

Artificial intelligence is increasingly used to support natural product discovery through structure-based prediction, bioactivity modeling, target inference, omics interpretation, ADMET estimation, toxicity prediction, and candidate prioritization. However, the reliability of these approaches depends on the quality of the chemical, biological, assay, source, and annotation data from which models learn. This article develops an original data quality framework for AI-driven natural product discovery. The framework treats data quality as a multidimensional bottleneck involving provenance, completeness, representativeness, assay reliability, chemical annotation, biological context, harmonization, leakage prevention, external validation, domain applicability, uncertainty communication, and auditability. Particular emphasis is placed on sampling bias, publication bias, database bias, missingness, class imbalance, label noise, conflicting records, incomplete metadata, assay variability, uncertain compound identity, stereochemical ambiguity, target annotation quality, and model generalizability. The central argument is that AI-driven natural product discovery is constrained not only by model architecture but also by the reliability, provenance, completeness, and domain relevance of the data used to train, validate, and interpret models. The proposed framework supports cautious research prioritization by defining quality-control gates across raw data, curated data, and model-ready datasets. It is conceptual rather than empirically validated and is intended to guide transparent dataset evaluation, model interpretation, and uncertainty-aware discovery workflows.

Key Words: *Natural product discovery, Artificial intelligence, Data quality, Chemical annotation, Assay reliability, Bias*

eIJPPR 2025; 15(6):23-33

HOW TO CITE THIS ARTICLE: Santos V, Beatriz A, Dias J. Data Quality as a Bottleneck in AI-Driven Natural Product Discovery: Bias, Completeness, Assay Reliability, Chemical Annotation, and Model Generalizability. Int J Pharm Phytopharmacol Res. 2025;15(6):23-33. https://doi.org/10.51847/OfW9SyZSje

INTRODUCTION

AI-driven drug discovery is fundamentally data-dependent: molecular representations, bioactivity labels, target annotations, omics profiles, ADMET endpoints, toxicity records, and literature-derived evidence all shape what models can learn and how their outputs can be interpreted. Machine learning has become embedded across drug discovery workflows, but its usefulness depends on data suitability, validation design, and the degree to which training information reflects the intended prediction problem [1].

Natural product discovery intensifies these data-quality requirements because natural products often involve complex structures, stereochemical diversity, variable biological sources, incomplete source metadata, and

Corresponding author: Victor Santos
Address: Department of AI for Phytopharmaceutical Quality Control and Fingerprinting, Faculty of Pharmacy, University of Porto, Porto, Portugal.
E-mail: ✉ victor.santos@ff.up.pt
Received: 10 August 2025; **Revised:** 17 November 2025; **Accepted:** 19 November 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.

uneven analytical or bioactivity evidence. The value of large-scale natural product data therefore depends not only on volume but also on preserving chemical, biological, and provenance information that permits careful reuse [2].

Recent AI applications in natural product discovery connect chemical structure prediction, biosynthetic inference, dereplication, activity modeling, metabolomics, genomics, and candidate prioritization. These approaches create opportunities for discovery acceleration, yet they also expose bottlenecks when datasets contain uncertain annotations, sparse metadata, heterogeneous assays, biased sampling, or limited external validation [3].

This article proposes an original data quality framework for AI-driven natural product discovery. The framework focuses on bias, completeness, assay reliability, chemical annotation, model-ready harmonization, domain applicability, external validation, uncertainty communication, and auditability, while avoiding the assumption that deep learning architectures alone can overcome weak or poorly documented datasets [4].

Data quality in natural product discovery

Data quality in natural product discovery should be understood as a multidimensional property rather than a single completeness score or a post hoc model-performance indicator. Relevant dimensions include provenance, source transparency, compound identity, structural correctness, stereochemical clarity, biological source metadata, assay metadata, bioactivity annotation, target annotation, harmonization, versioning, and external validity. Natural product databases vary in scope, coverage, curation depth, and field completeness, which means reuse requires explicit assessment of what each resource can and cannot support [5].

Integrated natural product repositories can improve accessibility by aggregating compound records across sources, but aggregation also increases the need for identifier harmonization, duplicate management, standardization of molecular structures, and transparent handling of uncertain or conflicting records. In this context, model-ready data are not merely raw database exports; they are curated, documented, and quality-controlled representations of natural product identity and associated evidence [6].

Source metadata are especially important because natural product meaning is linked to the producing organism, ecological context, isolation history, and chemical evidence. Microbial natural product resources illustrate how structured source and compound records can support reuse, but they also show that database value depends on traceable structure information, update practices, and provenance-aware documentation [7].

Bioactivity data add another layer of complexity because activity labels are meaningful only when interpreted together with assay format, target identity, organism or cell system, endpoint definition, units, confidence level, and curation history. Public bioactivity resources support AI-driven discovery when they preserve assay, compound, and target context, but these records should not be treated as context-free labels for model training [8].

Bias, completeness, and assay reliability

Bias in AI-driven natural product discovery can arise from the way compounds, organisms, targets, assays, diseases, and publication records enter databases. Sampling bias may overrepresent well-studied species, chemical classes, geographic regions, or therapeutic areas, while publication bias may privilege positive or unusual findings. Weak chemical and biological data can propagate through machine-learning pipelines and create misleading candidate rankings if data quality is not evaluated before modeling [9, 10].

Assay reliability is a central bottleneck because bioactivity values may vary across laboratories, measurement formats, endpoints, biological systems, and reporting conventions. Combining potency labels such as IC50 or Ki values from different sources without sufficient assay-context control can introduce substantial noise, especially when measurements are treated as interchangeable evidence for the same biological effect [10].

Assay metadata therefore function as a quality-control layer rather than as optional descriptive information. Bioactivity modeling should account for biological context, assay conditions, target information, endpoint definitions, and experimental setting before interpreting labels as comparable training data [11]. Missing bioactivity values also require explicit treatment, because missingness may reflect untested combinations, reporting gaps, assay selection bias, or structured absence rather than random data loss [12, 13].

Model evaluation can be distorted when redundancy, class imbalance, incomplete metadata, or inappropriate split strategies allow models to memorize chemical similarity patterns rather than generalize to new discovery settings. Ligand-based classification benchmarks can reward memorization when related compounds are distributed across training and test data [13]. Molecular benchmark design also shows that task definition, split strategy, and dataset imbalance influence how model performance should be interpreted [14].

Table 1 summarises major data-quality bottlenecks related to bias, completeness, and assay reliability in AI-driven natural product discovery.

Table 1. Data-Quality Bottlenecks in AI-Driven Natural Product Discovery: Sampling Bias, Publication Bias, Database Bias, Missingness, Class Imbalance, Assay Variability, Label Noise, Metadata Gaps, and Validation Requirements

Data-quality bottleneck	Core problem	Typical source	AI modeling consequence	Risk of false inference	Quality-control requirement	Validation need	Framework relevance
Sampling bias	Dataset overrepresents certain compounds, species, targets, diseases, or assay types	Literature-derived records, selective screening campaigns, specialized databases	Model learns overrepresented discovery spaces	A candidate appears broadly promising when evidence is domain-narrow	Profile chemical, biological, taxonomic, and assay coverage before modeling	Test on underrepresented or independent domains	Bias detection gate
Publication bias	Positive or unusual findings are more visible than negative or routine results	Published pharmacology and natural product reports	Training labels overrepresent active or notable compounds	Model overestimates activity likelihood	Track evidence source and absence of negative data	Compare with independent or prospective evidence where available	Evidence-balance assessment
Database bias	Database scope and curation choices shape available records	Curated repositories, aggregation pipelines, field-specific databases	Model inherits resource-specific coverage and curation assumptions	Database coverage is mistaken for biological reality	Document database source, version, scope, and curation rules	Evaluate against data from another source or time period	Source transparency gate
Geographic or species bias	Natural product sources are unevenly represented	Regional sampling, organism-focused studies, accessible species	Model prioritizes well-sampled taxa or locations	Underexplored sources are falsely deprioritized	Capture source organism, geography where available, and sampling context	Test transfer to less represented source groups	Representativeness check
Missingness	Essential fields are absent or incomplete	Incomplete reporting, database extraction gaps, untested endpoints	Reduced feature reliability and biased imputation	Missing records are treated as negative or irrelevant	Classify missingness by field and likely mechanism	Sensitivity analysis or missingness-aware modeling	Completeness gate
Class imbalance	Active, inactive, toxic, non-toxic, or target classes are unevenly distributed	Selective assays, positive-result enrichment, rare endpoints	Model favors majority classes or overfits minority labels	High apparent accuracy hides poor minority-class behavior	Report label distribution and endpoint balance	Class-sensitive evaluation	Label-distribution gate
Assay variability	Values differ across assay formats, systems, or laboratories	Heterogeneous bioassays, different protocols, endpoint definitions	Labels become non-comparable across contexts	Potency differences are misread as biological truth	Preserve assay format, endpoint, units, conditions, and source	Assay-stratified validation	Assay reliability gate
Label noise	Activity labels contain measurement, extraction, or harmonization errors	Cross-source potency aggregation, inconsistent units, conflicting records	Model learns unstable or contradictory labels	False target, activity, or toxicity inference	Detect duplicates, conflicts, unit inconsistencies, and outliers	Noise-robust validation and curator review	Bioactivity annotation gate

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

Metadata gaps	Biological, chemical, or assay context is missing	Incomplete records, minimal reporting, legacy datasets	Model inputs lack explanatory context	Prediction is interpreted beyond supported context	Require minimum metadata for source, structure, assay, target, and endpoint	External review of metadata sufficiency	Model-readiness gate
Conflicting records	Multiple entries disagree on compound identity, activity, or target	Database merging, repeated assays, inconsistent curation	Ambiguous labels reduce interpretability	Framework treats disagreement as certainty	Keep conflict flags and resolution rules	Independent evidence review	Auditability and conflict-resolution gate
Leakage-prone redundancy	Related or duplicate compounds appear across model splits	Random splitting, duplicated records, shared scaffolds	Inflated internal performance	Model appears generalizable when it memorizes similarity	Apply duplicate, scaffold, and similarity-aware split checks	External or leakage-aware validation	Train-test integrity gate
Limited reproducibility evidence	Results lack replication or independent confirmation	Single-source assay records, sparse experimental metadata	Model confidence exceeds evidence strength	One record drives candidate prioritization	Flag single-source labels and unsupported endpoints	Independent confirmation or cautious ranking	Evidence-strength gate

Chemical annotation and model generalizability

Chemical annotation is a decisive data-quality layer in AI-driven natural product discovery because model inputs often depend on whether compound identity, structure representation, stereochemistry, and analytical evidence are reliable enough for computational reuse. Tandem mass spectrometry workflows can support metabolite structure annotation, but annotation should be interpreted as evidence-linked rather than automatically definitive, particularly when structural assignments are computationally inferred or partially supported [15].

Molecular networking can strengthen comparative interpretation by linking related spectral features and supporting dereplication, but its usefulness depends on feature quality, preprocessing consistency, spectral evidence, and transparent annotation confidence. In natural product informatics, feature-level workflows should therefore be connected to data-quality controls that distinguish raw spectral signals, processed features, tentative annotations, and curated compound identities [16].

Structure standardization is also required before chemical records become model-ready. Chemical datasets may contain salts, mixtures, tautomers, charge-state inconsistencies, duplicate identifiers, ambiguous stereochemistry, or structure-format errors, and curation pipelines can help make these records more consistent for computational modeling [17].

Generalizability depends on whether chemical, biological, and contextual representations remain valid beyond the dataset used for model development. Multi-level biological representations can connect molecules with bioactivity, target, pathway, and phenotypic information, but these layers require harmonization before they can support interpretable AI workflows [18]. Toxicity models illustrate how temporal or domain-related data drift may reduce predictive reliability when new compounds or endpoints differ from the training domain [18, 19]. Molecular property predictions should therefore be accompanied by uncertainty estimates when models are applied to unfamiliar chemical spaces, shifted data distributions, or weakly represented natural product classes [20].

Data quality criteria

An AI-ready natural product discovery dataset should be evaluated through explicit criteria before it is used for model training, validation, interpretation, or candidate prioritization. FAIR-oriented dataset selection supports provenance, documentation, transparency, reuse, and auditability, but FAIR alignment should be treated as a starting point rather than a substitute for chemical, biological, and assay-level quality assessment [21].

Model-ready therapeutic datasets also require task clarity, benchmark transparency, endpoint definition, and evaluation designs that reflect realistic discovery use. Benchmark infrastructure can support consistent comparison, yet its scientific value depends on whether tasks, labels, splits, and metadata are appropriate for the intended prediction setting [22].

ADMET and toxicity data require additional scrutiny because endpoint definitions, assay systems, biological context, and external validation strongly influence whether

predictions are useful for discovery decisions. Prospective ADME validation shows why prediction reliability should be assessed beyond internal retrospective evaluation [23]. Toxicity prediction similarly depends on careful endpoint definition and validation because noisy, imbalanced, or context-poor toxicity labels may produce misleading safety interpretations [24].

Uncertainty reporting is a required criterion for responsible use of AI-ready natural product datasets because predictions should not be interpreted as equally reliable across all compounds, endpoints, assays, or source domains. Scalable uncertainty estimation methods support more cautious interpretation of molecular property predictions and can help separate higher-confidence prioritization from predictions requiring further review [25].

Table 2 presents data-quality criteria for evaluating AI-ready natural product discovery datasets.

Table 2. Data-Quality Criteria for AI-Ready Natural Product Discovery Datasets: Provenance, Completeness, Representativeness, Chemical Annotation, Assay Metadata, Target Annotation, Harmonization, Leakage Prevention, Domain Applicability, External Validation, and Uncertainty Reporting

Quality criterion	Core assessment question	Required metadata or evidence	Failure risk	AI consequence	Quality-control gate	Validation requirement	Framework role
Provenance	Can each record be traced to its source?	Source database, publication, extraction method, curation history, version	Unverifiable reuse	Irreproducible training and weak interpretation	Provenance documentation gate	Source audit before modeling	Foundation layer
Source transparency	Is the origin and scope of the dataset clear?	Database scope, inclusion logic, update record, field definitions	Hidden database bias	Overgeneralized model claims	Source-scope review	Cross-source comparison where possible	Bias-awareness layer
Completeness	Are essential fields present?	Compound identity, structure, source, assay, target, endpoint, units	Missing context	Feature gaps and biased learning	Completeness screening	Missingness sensitivity review	Data-readiness layer
Representativeness	Does the dataset reflect the intended discovery domain?	Chemical class, organism, geography, target, assay, endpoint distribution	Narrow applicability	Weak transfer to new domains	Domain-coverage audit	External-domain testing	Generalizability layer
Bias detection	What systematic biases are present?	Sampling source, label distribution, publication pattern, redundancy	Inflated evidence confidence	Misleading prioritization	Bias-mapping gate	Bias-sensitive evaluation	Interpretation safeguard
Chemical annotation quality	Is compound identity sufficiently reliable?	Structure, identifiers, InChIKey, analytical support, confidence flag	Incorrect structure-label mapping	Invalid chemical learning	Chemical identity check	Curator or orthogonal evidence review	Chemical-data gate
Stereochemical clarity	Is stereochemistry known and encoded?	Chiral centers, stereochemical notation, isomer confidence	Isomer ambiguity	Incorrect similarity or activity inference	Stereochemistry flagging	Ambiguous-record review	Natural product identity gate
Bioactivity annotation quality	Are activity labels interpretable?	Endpoint, value, unit, relation symbol, assay system, source	Incomparable labels	Label noise and endpoint confusion	Bioactivity annotation review	Endpoint-specific validation	Bioactivity gate
Assay metadata quality	Is biological context documented?	Assay type, organism or cell line, target,	Context-free labels	Poor assay transfer	Assay metadata requirement	Assay-aware evaluation	Reliability layer

		protocol, detection method					
Target annotation reliability	Is the target assignment credible?	Target ID, species, isoform, confidence, mapping evidence	Mechanistic misassignment	False target inference	Target mapping review	Orthogonal evidence check	Mechanism layer
Pathway annotation reliability	Are pathway links evidence-based?	Pathway database, identifier mapping, evidence type	Overinterpreted biology	Misleading pathway ranking	Pathway evidence grading	Biological plausibility review	Biological-context layer
Missingness management	Why are values missing?	Missing-field patterns, imputation status, absence reason where known	Biased deletion or imputation	Distorted training signal	Missingness classification	Sensitivity analysis	Completeness safeguard
Label-noise detection	Are labels internally consistent?	Duplicate records, conflicting values, units, assay source	Unstable labels	Reduced calibration and false confidence	Conflict and noise detection	Noise-aware validation	Reliability safeguard
Conflict resolution	How are contradictory records handled?	Resolution rule, retained alternatives, curator notes	Arbitrary data cleaning	Hidden curation bias	Conflict-resolution protocol	Curator review	Auditability criterion
Data harmonization	Are units, identifiers, and structures standardized?	Unit conversion, ontology mapping, identifier crosswalk, structure normalization	Non-comparable records	Feature-label mismatch	Harmonization gate	Cross-database consistency check	Model-readiness layer
Leakage prevention	Are training and test data independent?	Split rule, duplicate check, scaffold or similarity grouping	Inflated internal performance	Memorization mistaken for generalization	Leakage audit	Split-stress testing	Validation gate
External validation	Does the model hold outside internal data?	Independent dataset, temporal split, prospective evidence, external source	Local-only performance	Weak translational reliability	External validation requirement	Independent evaluation	Generalizability gate
Domain applicability	Where should predictions be trusted?	Applicability domain, similarity range, source-domain match	Out-of-domain use	Unreliable prioritization	Applicability-domain flag	Domain-shift evaluation	Use-condition layer
Uncertainty reporting	Is confidence communicated?	Calibration, uncertainty score, confidence interval, abstention rule	Overconfident decisions	Misleading candidate ranking	Uncertainty output gate	Calibration assessment	Interpretation safeguard
Versioning and auditability	Can the dataset and model be reconstructed?	Dataset version, curation log, split record, code version, change history	Irreproducible workflow	Non-auditable claims	Version-control gate	Reproducibility audit	Governance layer

Proposed framework

The proposed data quality framework begins with data source identification and provenance documentation. At this stage, raw natural product records, chemical structures, source metadata, bioactivity evidence, assay records, target annotations, omics profiles, ADMET endpoints, toxicity data, and literature-derived evidence are mapped to their origin, version, and curation status. Biomedical dataset integration and knowledge-graph perspectives highlight the need to connect identifiers, entities, relationships, and

provenance before downstream inference is attempted [26].

The second stage evaluates completeness, bias, assay reliability, and annotation confidence before data are converted into model-ready form. This proposed stage requires missingness classification, source transparency review, class-balance inspection, conflict detection, assay-context assessment, chemical annotation review, stereochemistry flags, and target or pathway mapping checks. Predictions derived from records that pass these gates should still be interpreted as confidence-qualified research priorities, and conformal prediction offers one route for communicating prediction confidence in drug discovery contexts [27].

The third stage concerns model-readiness and validation. Model-ready datasets should include standardized structures, harmonized identifiers, documented endpoint definitions, assay metadata, train-test split integrity checks, leakage-prevention logic, applicability-domain flags, and uncertainty outputs. Evidential deep learning illustrates how molecular property prediction can integrate uncertainty estimates into guided discovery, supporting cautious ranking rather than definitive claims [28].

The fourth stage introduces external validation, domain applicability, and auditability as required safeguards before AI outputs are used for discovery interpretation. Leakage-prevention methods emphasize that split design is itself a data-quality issue because inappropriate splitting can allow information from related records to influence both training and evaluation [29].

The final stage defines decision boundaries for AI-driven natural product discovery. Candidate prioritization should be accompanied by documentation of data provenance, quality-control gates passed, unresolved uncertainty, applicability limits, expert-review requirements, and validation plans. Confidence assignment in molecular property prediction supports the principle that AI outputs should guide prioritization under uncertainty rather than replace chemical, biological, assay, and translational validation [30].

Table 3 presents the proposed data quality framework for AI-driven natural product discovery.

Table 3. Proposed Data Quality Framework for AI-Driven Natural Product Discovery: Data Provenance, Bias Detection, Completeness Assessment, Assay Reliability, Chemical Annotation, Model-Ready Harmonization, External Validation, Generalizability, Uncertainty, and Auditability

Framework stage	Core purpose	Required input	Quality-control action	High-confidence signal	Caution signal	Validation requirement	Failure risk	Discovery implication
Data source identification	Define what data are being used	Database records, literature-derived evidence, assay records, omics files, ADMET and toxicity records	Catalogue source, scope, extraction route, and intended use	Source is traceable and relevant	Source scope is unclear or mismatched	Source audit	Hidden provenance gaps	Discovery claims cannot be traced reliably
Provenance documentation	Preserve origin and version history	Source identifiers, version labels, curation notes, access date, update record	Create record-level provenance fields	Each record has transparent lineage	Aggregated data lack source separation	Provenance review	Irreproducibility	Dataset can or cannot support audit
Completeness assessment	Identify missing critical context	Compound, source, assay, target, endpoint, unit, annotation, metadata fields	Classify missingness and minimum fields	Essential fields present or flagged	Missingness affects core interpretation	Missingness sensitivity review	Biased modeling	Model-readiness is limited
Bias detection	Detect systematic skew	Chemical class, species, geography, target, assay, label, publication patterns	Map sampling, publication, database, species, and class bias	Coverage matches intended domain	Narrow or overrepresented domain	Bias-aware evaluation	Inflated discovery confidence	Prioritization may exclude underexplored areas

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

Assay reliability evaluation	Assess whether labels are comparable	Assay format, conditions, endpoint, target, organism, units, replicate or source evidence	Preserve assay context and detect conflicting labels	Assay metadata supports interpretation	Cross-source labels lack context	Assay-stratified validation	Label noise	Candidate ranking may reflect assay artifacts
Chemical annotation review	Verify compound identity	Structure, identifiers, stereochemistry, spectral or analytical support, duplicate flags	Standardize and flag uncertain structures	Curated structure and identity evidence	Ambiguous stereochemistry or tentative annotation	Chemical curation review	Wrong structure-label mapping	Chemical inference becomes unreliable
Biological annotation review	Evaluate target and pathway context	Target IDs, organism, isoform, pathway mapping, evidence type	Grade annotation evidence and mapping confidence	Target and pathway links are traceable	Mechanistic claims are weakly mapped	Biological plausibility review	False mechanism inference	Target prediction requires caution
Harmonization	Convert curated records into comparable format	Units, identifiers, ontologies, structures, endpoints, assay terms	Standardize units, structures, and vocabulary	Records are comparable across sources	Non-standard fields remain unresolved	Cross-source consistency check	Feature-label mismatch	Data may not be suitable for modeling
Model-readiness assessment	Decide whether data can support AI training	Curated chemical, biological, assay, source, and endpoint records	Apply minimum metadata and quality gates	Dataset is documented and fit-for-purpose	Essential quality gates fail	Pre-model review	Unreliable training inputs	Model development should be delayed or restricted
Leakage prevention	Protect evaluation integrity	Split plan, duplicate list, similarity or scaffold relationships	Check duplicate, scaffold, temporal, and source leakage	Train-test separation is defensible	Related records cross evaluation boundary	Leakage-aware split testing	Inflated internal performance	Generalizability may be overstated
External validation	Test performance beyond internal data	Independent dataset, temporal holdout, external assay, prospective evidence where available	Evaluate outside development data	Performance remains interpretable externally	Performance drops under domain shift	Independent evaluation	Local-only prediction	Discovery use should be limited
Domain applicability	Define valid prediction space	Applicability-domain model, similarity range, source-domain mapping	Flag out-of-domain predictions	Compound lies within supported domain	Novel scaffold or weakly represented source	Domain-shift assessment	Overextension	Prediction requires expert caution
Uncertainty communication	Report confidence and limitations	Calibration metrics, uncertainty estimate, confidence	Attach uncertainty to outputs	Prediction is confidence-qualified	Uncertainty is high or unreported	Calibration review	Overconfident decisions	Prioritization remains transparent

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

		flag, abstention rule						
Expert review	Integrate domain knowledge	Chemist, pharmacologist, data curator, assay expert review	Review uncertain annotations, mechanisms, and rankings	Expert review supports interpretation	Automated output lacks biological plausibility	Multidisciplinary review	Automation bias	Candidate selection becomes more defensible
Auditability	Preserve reproducibility and accountability	Curation logs, versions, split records, model input files, decision notes	Maintain audit trail	Workflow can be reconstructed	Changes are undocumented	Reproducibility audit	Non-reproducible claims	Framework supports transparent reuse

Figure 1 illustrates the proposed data quality framework for AI-driven natural product discovery, linking provenance, bias detection, completeness, assay reliability, chemical annotation, model generalizability, validation, and uncertainty communication.

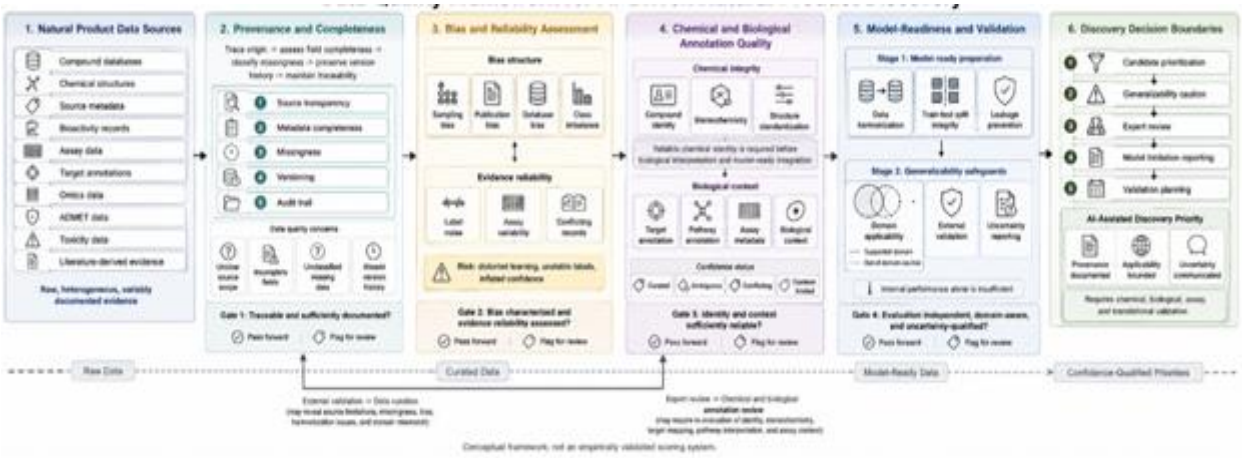

Figure 1. Data Quality Framework for AI-Driven Natural Product Discovery

Alt-text description
A conceptual framework diagram showing natural product data sources flowing through data provenance, bias detection, completeness assessment, assay reliability review, chemical annotation, harmonization, model-ready data preparation, external validation, domain applicability assessment, uncertainty communication, and auditability.

CONCLUSION

This article proposed a data quality framework for AI-driven natural product discovery that links provenance, completeness, bias detection, assay reliability, chemical annotation, model-ready harmonization, train-test split integrity, external validation, domain applicability, uncertainty communication, expert review, and auditability. The framework argues that reliable AI-driven discovery requires not only advanced computational models but also trustworthy, well-documented, representative, chemically accurate, assay-aware, externally validated, and uncertainty-aware data. By distinguishing raw data, curated data, and model-ready data, the framework supports cautious discovery prioritization while avoiding overconfident interpretation of predictions derived from incomplete, biased, poorly annotated, or weakly validated datasets.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

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

[2] Cech NB, Medema MH, Clardy J. Benefiting from big data in natural products: Importance of preserving foundational skills and prioritizing data quality. Nat Prod Rep. 2021;38(11):1947-53. doi:10.1039/D1NP00061F

[3] Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJJ, Martin NI, et al. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023;22(11):895-916. doi:10.1038/s41573-023-00774-7

[4] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241-50. doi:10.1016/j.drudis.2018.01.039

[5] 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

[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, 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

[8] Zdrazil B, Felix 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

[9] Rodrigues T. The good, the bad, and the ugly in chemical and biological data for machine learning. Drug Discov Today Technol. 2019;32-33:3-8. doi:10.1016/j.ddtec.2020.07.001

[10] 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

[11] 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

[12] Whitehead TM, Irwin BWJ, Hunt P, Segall MD, Conduit GJ. Imputation of assay bioactivity data using deep learning. J Chem Inf Model. 2019;59(3):1197-204. doi:10.1021/acs.jcim.8b00768

[13] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403

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

[15] Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019;16(4):299-302. doi:10.1038/s41592-019-0344-8

[16] 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

[17] 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

[18] Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, et al. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol. 2020;38(9):1087-96. doi:10.1038/s41587-020-0502-7

[19] Morger A, Garcia de Lomana M, Norinder U, Svensson F, Kirchmair J, Mathea M, et al. Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. Sci Rep. 2022;12(1):7244. doi:10.1038/s41598-022-09309-3

[20] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502

[21] Alharbi E, Gadiya Y, Henderson D, Zaliani A, Delfin-Rossaro A, Cambon-Thomsen A, et al. Selection of data sets for FAIRification in drug discovery and development: Which, why, and how? Drug Discov Today. 2022;27(8):2080-5. doi:10.1016/j.drudis.2022.05.010

[22] Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, et al. Artificial intelligence foundation for therapeutic science. Nat Chem Biol. 2022;18(10):1033-6. doi:10.1038/s41589-022-01131-2

[23] Fang C, Wang Y, Grater R, Kapadnis S, Black C, Trapa P, et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. J Chem Inf Model. 2023;63(11):3263-74. doi:10.1021/acs.jcim.3c00160

[24] Cai C, Guo P, Zhou Y, Zhou J, Wang Q, Zhang F, et al. Deep learning-based prediction of drug-induced

cardiotoxicity. J Chem Inf Model. 2019;59(3):1073-84. doi:10.1021/acs.jcim.8b00769

[25] Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975

[26] Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Brief Bioinform. 2022;23(6):bbac404. doi:10.1093/bib/bbac404

[27] Alvarsson J, Arvidsson McShane S, Norinder U, Spjuth O. Predicting with confidence: Using conformal prediction in drug discovery. J Pharm Sci. 2021;110(1):42-9. doi:10.1016/j.xphs.2020.09.055

[28] Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546

[29] Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. Nat Commun. 2025;16(1):3337. doi:10.1038/s41467-025-58606-8

[30] Nigam AK, Pollice R, Hurley MFD, Hickman RJ, Aldeghi M, Yoshikawa N, et al. Assigning confidence to molecular property prediction. Expert Opin Drug Discov. 2021;16(9):1009-23. doi:10.1080/17460441.2021.1925247