



From Misassigned Structures to Misleading Models: How Structural Annotation Errors Propagate Through Natural Product Informatics and Computational Pharmacology

Ahmed Ben Ali^{1*}, Samira Gharbi¹, Nabil Jebali²

¹Department of Structural Annotation and Error Propagation, Faculty of Pharmacy, University of Tunis, Tunis, Tunisia.

²Department of Natural Product Informatics and Model Reliability, Faculty of Pharmacy, University of Sfax, Sfax, Tunisia.

ABSTRACT

Natural-product informatics often treats a molecular structure as a stable input, yet the structure itself may remain an evidential claim. A misassigned constitution, configuration, regiochemistry, or molecular state can therefore enter databases and computational workflows before any model is fitted, scored, or validated. This methodological framework examines structural annotation error as an upstream modeling hazard and distinguishes experimental structure assignment from database standardization, molecular-state preparation, descriptor generation, docking, similarity analysis, and machine-learning prediction. Evidence from natural-product structure revision, chemical database curation, stereochemical representation, docking benchmarks, and molecular machine learning is integrated to identify where structural uncertainty can be introduced, transformed, or hidden. The article proposes a structural-provenance framework in which computational reuse is conditioned on traceable assignment evidence, explicit representation policy, state specification, versioned correction, and model-sensitivity analysis. The framework does not assume that every alternative tautomer, protomer, stereoisomer, or database representation is erroneous; rather, it separates genuine assignment error from legitimate molecular multiplicity and deliberate harmonization choices. It further distinguishes chemical-identity confidence from predictive uncertainty and numerical model confidence. The principal implication is methodological: model validation cannot compensate for an unexamined molecular identity claim when the model consumes that claim as data. The proposed framework is not a validated quality standard and does not establish the prevalence or quantitative effect of misassigned natural products in current computational datasets. Prospective testing with paired corrected and uncorrected structures is required.

Key Words: *Natural products, Structure elucidation, Chemical provenance, Molecular representation, Cheminformatics, Computational pharmacology*

eIJPPR 2025; 15(4):85-93

HOW TO CITE THIS ARTICLE: Ben Ali A, Gharbi S, Jebali N. From Misassigned Structures to Misleading Models: How Structural Annotation Errors Propagate Through Natural Product Informatics and Computational Pharmacology. Int J Pharm Phytopharmacol Res. 2025;15(4):85-93. <https://doi.org/10.51847/IOjLj8ndu>

INTRODUCTION

Cheminformatics workflows routinely standardize molecular records before descriptors are calculated or predictive models are trained, because inconsistent structures can undermine reproducibility and downstream comparability [1, 2]. Curated bioactivity resources likewise depend on sustained integration of chemical

structures, assay records, and biological annotations [3]. Yet standardization answers a different question from structure elucidation. A canonicalized record may be internally consistent while still representing the wrong constitution, regioisomer, stereoisomer, or experimentally relevant molecular state. Computational precision therefore begins downstream of an epistemic commitment: that the encoded structure is an adequate representation of

Corresponding author: Ahmed Ben Ali

Address: Department of Structural Annotation and Error Propagation, Faculty of Pharmacy, University of Tunis, Tunis, Tunisia.

E-mail: ✉ ahmed.benali@fpharm.u-tunis.tn

Received: 25 March 2025; **Revised:** 22 July 2025; **Accepted:** 23 July 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.



the compound that was isolated, measured, or pharmacologically tested.

This distinction matters acutely for natural products because structural assignments are often assembled from incomplete, indirect, or chemically constrained evidence rather than read directly from a single measurement. Once a proposed structure is converted to a machine-readable graph, identifiers, descriptors, conformers, similarities, docking poses, and training examples can all inherit that representation. None of those downstream operations independently establishes that the upstream assignment was correct. The methodological problem is thus not that computational models are inherently unreliable, but that they can become highly reproducible with respect to an incorrect or insufficiently qualified structural assertion.

This article develops a proposed framework for treating structural provenance as part of model validity. It separates four questions that are frequently conflated: whether the molecular assignment is supported by experimental evidence; whether the database record is standardized and traceable; whether the modeled molecular state matches the computational task; and whether the model is robust to plausible structural alternatives. The framework does not presume that every representation difference is an error, nor that every corrected structure changes a model output. Instead, it asks where structural uncertainty enters, when it survives computational preprocessing, and how it should constrain confidence in downstream pharmacological inference.

Where natural product structural misassignment begins

Natural-product misassignment is rarely reducible to a single clerical mistake. Reanalysis of oxirane-containing natural products has shown that apparently plausible published structures can become unstable when stereochemical and spectroscopic evidence is reconsidered computationally [4]. In a broader alkaloid-focused study, machine-learning-augmented DFT NMR calculations were used to screen reported structures and identify assignments requiring revision, demonstrating that inconsistencies can remain latent until many compounds are evaluated under a common computational criterion [5]. These findings support targeted re-examination, but they do not provide a population estimate for natural-product misassignment.

The trigger may instead be chemical plausibility or evidential incompleteness. Penicitone was reassigned after scrutiny of an unusual proposed structure using computer-assisted and DFT-supported analysis [6]. Caylobolide A illustrates a different problem: an isolated natural product may be real while its stereochemical description remains incomplete until orthogonal high-resolution NMR, derivatization, and synthesis resolve the remaining configuration [7]. Misassignment should therefore be distinguished from underassignment. A record can be wrong, partially specified, historically superseded, or simply unable to express the degree of uncertainty present in the underlying evidence.

For computational reuse, the critical upstream question is consequently not merely whether a structure exists in a database, but what evidential pathway produced that structural assertion and which dimensions remain unresolved. The recurring components of that interrogation are summarized in **Table 1**.

Table 1. Method components for interrogating an upstream natural-product structural assignment before computational reuse

Component	Diagnostic question	Failure signal	Containment response
Assignment completeness	Are constitution and relevant stereochemical elements specified?	Unresolved connectivity or configuration	Preserve the unresolved state
Spectral concordance	Does the proposal remain compatible with available spectra?	Persistent unexplained disagreement	Recalculate or reassess the assignment
Chemical plausibility	Does an exceptional motif require stronger scrutiny?	Weak support for unusual chemistry	Use plausibility only as a review trigger
Orthogonal challenge	Has material uncertainty been tested independently?	Dependence on one analytical route	Seek complementary spectroscopy, computation, or synthesis
Record status	Is the assertion current, provisional, or superseded?	Competing structures without lineage	Preserve version and interpretive status

The same evidence also reveals a process problem that prose alone obscures: a reported structure should not pass directly into computational reuse when its evidential status remains unresolved. **Figure 1** therefore represents a

proposed upstream triage workflow derived from the revision patterns described above.

Figure 1 From a reported natural-product assignment to a computationally reusable structural assertion

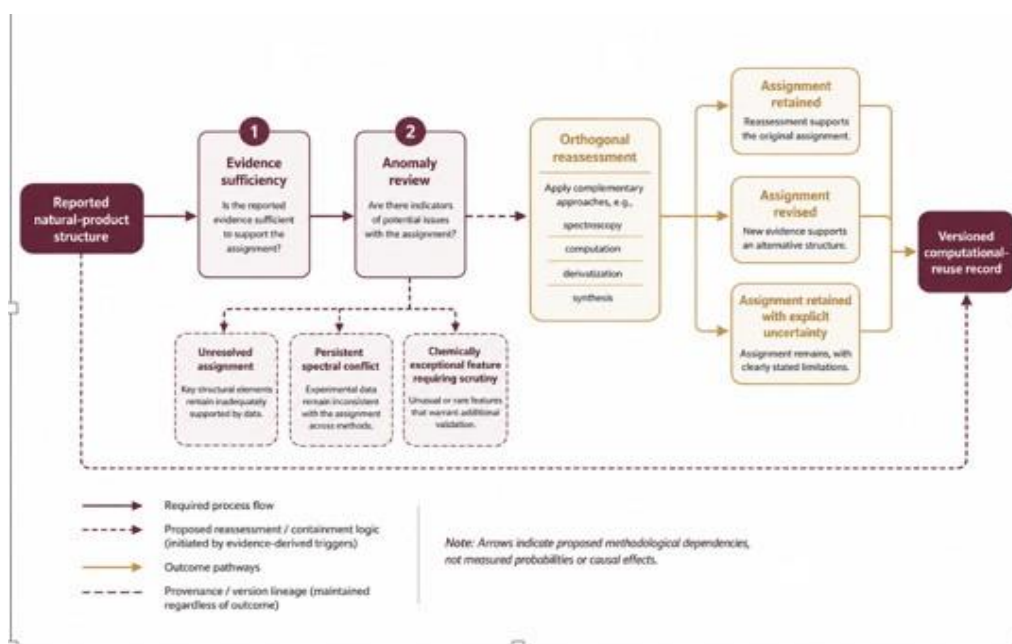


Figure 1. From a reported natural-product assignment to a computationally reusable structural assertion

Stereochemical, regiochemical, tautomeric, and connectivity errors

Structural uncertainty must next be decomposed by chemical type because constitution, regiochemistry, configuration, tautomerism, and protonation do not have the same evidential meaning. Specialized NMR approaches can provide relative configurational information from very small quantities of natural products, while the existence of genuine racemic natural products cautions against treating every stereochemical multiplicity as an annotation failure; combined ECD, NMR, and DP4+ analysis can further discriminate among candidate structures [8–10]. The methodological consequence is that “stereochemical error” should be reserved for an unsupported or incorrect stereochemical assertion, not for legitimate racemism, unresolved configuration, or an intentionally unspecified center.

Regiochemical and connectivity errors concern the molecular graph itself. Database-assisted and DFT-supported reassessment of reported natural products has demonstrated that apparently small positional changes can resolve systematic spectral inconsistencies [11]. Computational reassignment gains substantially greater evidential weight when it is challenged experimentally, as illustrated by synthesis-based confirmation of the revised structure of peganutonin A [12]. Series-level analysis of pyranoxanthones likewise shows how repeated positional assignments can be interrogated through diagnostic carbon-NMR behavior within a defined chemical family [13]. These studies justify vigilance but not extrapolation of class-specific revision rates to natural products as a whole.

Tautomeric and protonation states require a different logic. Virtual-screening preparation tools explicitly enumerate ionization, tautomeric, stereochemical, cis/trans, and ring states because the structure submitted to a docking workflow is a modeled state rather than an unqualified synonym for experimental identity [14]. A tautomer or protomer may therefore be an appropriate task-specific representation rather than a correction. The framework proposed here treats assignment error, unresolved identity, and state-selection uncertainty as separate provenance fields so that a model does not silently convert one category into another.

Database replication turns local error into distributed error

Once a structural assertion enters a public or institutional resource, its representation is commonly transformed before reuse. PubChem applies explicit structure-standardization rules, open RDKit-based curation pipelines normalize and check molecular records, and natural-product databases differ substantially in coverage, metadata, and curation practice [15–17]. These processes are essential for interoperability, but they operate on submitted or aggregated records; they do not independently reproduce the original spectroscopic reasoning that established the compound. Standardization can therefore make representations consistent without resolving whether an inherited assignment was evidentially sound.

Cross-source registration also changes how apparently related records are grouped. The canSAR chemistry pipeline demonstrates that tautomer handling, parent extraction, and canonicalization influence whether

alternative forms are linked while also documenting edge cases that automated checks may not fully resolve [18]. Papyrus provides a complementary boundary condition: its principal standardized version deliberately removed stereochemistry to improve consistency across sources, while a separate stereo-preserving version retained less standardization [19]. Loss of structural detail is therefore not always an error; sometimes it is an explicit harmonization decision whose appropriateness depends on the modeling task.

Modern curated resources continue to revise how structures and bioactivity records are integrated. ChEMBL exemplifies ongoing multi-source curation rather than a static archive [20], whereas COCONUT 2.0 emphasizes large-scale natural-product curation and source provenance [21]. The methodological hazard arises when downstream users cannot distinguish an original assignment, a standardized derivative, a later correction, and a representation chosen for modeling convenience. The framework therefore treats provenance as a chain of transformations rather than a single database identifier: each reusable record should preserve where the structural claim originated, how it was normalized, whether it was superseded, and which chemically meaningful distinctions were intentionally retained or discarded.

Descriptor and fingerprint corruption

A structural correction becomes computationally consequential only through the representation actually consumed by a model. Stereochemistry illustrates this dependency clearly. Molecular-learning experiments show that stereoisomers remain difficult to distinguish when the chosen representation or model does not preserve the relevant configurational information [22]. String-based Transformer models can also learn chirality less readily than other structural regularities, making successful molecular-language modeling compatible with incomplete stereochemical discrimination [23]. In chiral-property prediction, alternative descriptor constructions likewise differ in their ability to encode enantiomeric distinctions [24]. These findings support representation sensitivity, not a universal claim that any stereochemical correction must change prediction.

The term “descriptor corruption” is therefore used here conditionally. If a corrected assignment changes atoms, bonds, stereochemistry, charge, or another feature encoded by a fingerprint or descriptor, the resulting vector may change; if the representation intentionally removes that dimension, two chemically distinct records may collapse to the same computational object. Conversely, a structural revision can be invisible to a representation that does not encode the revised feature. A QSAR-ready standardization workflow can improve consistency of the encoded

structure [1], but it cannot restore chemical information that was absent from the underlying assignment or deliberately discarded during representation.

For natural-product informatics, this creates a two-stage integrity requirement: verify what chemical distinction is evidentially justified, then verify whether the chosen representation preserves that distinction when the endpoint could depend on it. Model sensitivity should therefore be evaluated against chemically plausible alternative representations rather than inferred from nominal descriptor completeness alone.

Docking the wrong structure with high numerical confidence

Docking adds three-dimensional preparation and scoring to this problem but does not change its epistemic order. A docking engine receives a prepared ligand, not the original spectroscopic evidence from which the ligand identity was inferred. Large-scale redocking demonstrates that pose reproduction and docking outcomes depend on preparation, protocol, and the structural context supplied to the calculation [25]. Ligand-preparation workflows also explicitly enumerate protonation, tautomeric, stereochemical, and ring states before virtual screening [14]. A well-optimized pose can therefore be internally coherent for the modeled state while remaining silent about whether that state corresponds to the correctly assigned natural product.

This distinction is especially important when numerical ranking is interpreted as chemical validation. A favorable score cannot repair an incorrect connectivity, establish an unresolved configuration, or determine which protomer dominated under the experimental assay conditions. Nor does a poor score prove that a reported natural-product structure is wrong, because receptor preparation, scoring-function limitations, conformational sampling, and state selection provide competing explanations. The proposed methodological rule is consequently to separate “identity confidence” from “pose or score confidence.” Docking may help compare explicitly defined structural alternatives, but adjudicating an assignment requires evidence capable of bearing on chemical identity.

Misassigned structures inside machine-learning training sets

A machine-learning model can only learn from the structures and labels presented to it. Benchmark analyses have shown that strong ligand-classification performance may partly reflect train-test redundancy rather than transferable chemical inference [26]. Structure-based virtual-screening datasets can likewise contain biases that permit models to obtain apparently successful results for reasons that do not correspond to the intended physical

discrimination [27]. These findings do not demonstrate that natural-product misassignment is a measured source of benchmark bias; they establish the broader methodological point that numerical performance can remain high when the dataset encodes shortcuts or distortions.

Structural annotation error should therefore be treated as a distinct, testable data-quality mechanism rather than assumed from disappointing model behavior. If an incorrect structure is duplicated across sources, associated with an activity label, and retained through preprocessing, the model may learn features of the encoded structure even though those features do not belong to the assayed compound. If a later correction changes the molecular

representation while the activity label remains attached to the historical record, retraining may alter local neighborhoods, learned associations, or uncertainty. Predictive uncertainty itself does not quantify the probability that a chemical structure is correctly assigned; chemistry-focused uncertainty methods characterize uncertainty in model prediction under their own assumptions [28].

The resulting workflow problem is summarized in **Table 2**, which separates proposed containment steps from the evidence-supported computational operations that motivate them.

Table 2. Proposed workflow for containing structural-annotation error before model fitting and evaluation

Workflow step	Question to resolve	Evidence-supported concern	Proposed containment action
Identity intake	What structure was assigned and on what evidence?	Standardized records do not independently validate assignment	Link model input to assignment source and status
Molecular-state preparation	Which protomer, tautomer, stereoisomer, or conformer is modeled?	Screening workflows enumerate alternative states	Record state-selection policy separately from identity
Representation generation	Does the descriptor preserve the relevant distinction?	Stereo sensitivity depends on representation	Recompute features for plausible structural alternatives
Dataset assembly	Are historical and corrected records merged, duplicated, or split?	Dataset construction can preserve redundancy and bias	Track version lineage and apply provenance-aware deduplication
Model evaluation	Does performance survive structurally meaningful perturbation?	Benchmark success may reflect dataset bias	Add correction-sensitive and externally separated tests
Confidence interpretation	Is prediction uncertainty mistaken for assignment confidence?	Predictive uncertainty is model-specific	Report structural and model uncertainty separately

89

Figure 2 extends this workflow into a methodological architecture by showing where structural evidence and

model validation must remain connected but non-interchangeable.

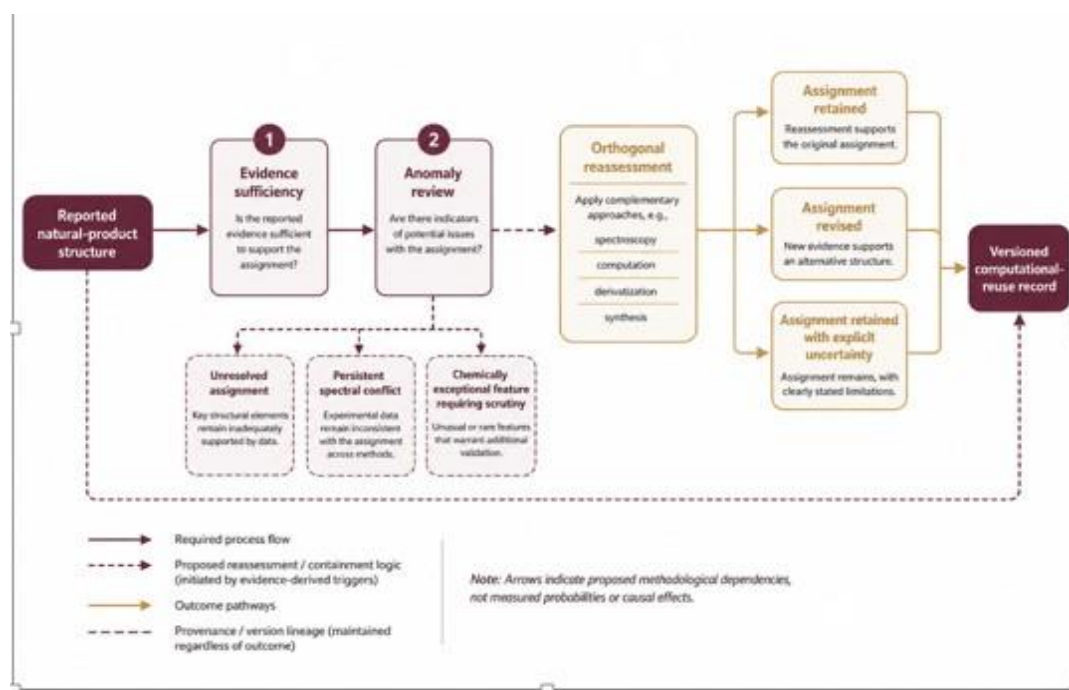


Figure 2. Structural-provenance gates between natural-product evidence and model validation

Error propagation into similarity search and sar

Similarity search converts a molecular representation into a neighborhood, not into a statement of chemical or pharmacological equivalence. Similarity-based read-across and related predictive methods depend on the selected descriptors, fingerprints, metrics, and endpoint context [29]. A structural revision can therefore move a natural product into or out of a local neighborhood when the corrected feature is encoded. Stereo-sensitive machine-learning studies reinforce that configurational distinctions may be either retained or obscured depending on representation [22], while chiral descriptor comparisons show that different encodings can support different levels of discrimination [24].

The downstream implication for structure-activity relationships is asymmetric. A corrected structure may invalidate an apparent local SAR explanation if the feature believed to accompany activity was never present in the assayed compound. Yet unchanged similarity after correction does not prove that the revision is pharmacologically irrelevant, because the chosen representation may be insensitive to the altered dimension. The proposed framework therefore requires SAR statements to specify whether they concern experimental identity, encoded structural similarity, or observed pharmacological response. These are connected levels of analysis, but none can be substituted for another.

A Structural-provenance framework for natural product informatics

The central contribution of this article is a proposed structural-provenance quality-control framework that positions chemical identity as a versioned evidence object rather than a static structure string. Public-database integrity arguments support explicit provenance [2], while PubChem and RDKit workflows demonstrate the importance of transparent structure transformation and curation [15]. ChEMBL shows why chemical structures must remain linked to heterogeneous assay and bioactivity records [20], and COCONUT 2.0 illustrates the value of source-aware natural-product curation [21]. These sources establish infrastructure needs; they do not establish the integrated framework proposed here.

The framework separates four records that are commonly collapsed. The first is the assignment record: what structure was proposed, from which material, and with what unresolved dimensions. The second is the provenance record: where the assertion originated, which identifiers referred to it, and whether it was later revised. The third is the computational representation record: what standardization, stereochemical handling, tautomer/protomer policy, and descriptor construction were applied. The fourth is the model-use record: which

version entered docking, similarity search, QSAR, or machine learning and which outputs depended on it.

Two controls follow from this separation. First, provenance should distinguish a changed scientific assertion from a merely changed serialization: an updated SMILES string caused by canonicalization is not equivalent to a revised constitution or configuration. Second, computational dependencies should be reversible. If a structural version is superseded, datasets, descriptors, docking jobs, similarity neighborhoods, and trained models that consumed that version should be identifiable so that reanalysis can be prioritized rather than indiscriminately repeated. This dependency tracing is proposed here as error containment, not as evidence that every structural correction materially changes downstream conclusions.

A structural correction should therefore propagate forward as a version event rather than overwrite history. The old representation remains necessary for reproducibility because it may explain previously published model inputs, while the corrected representation becomes the preferred object for new analysis. This proposed architecture converts “data cleaning” into traceable model governance without implying that every record requires full spectroscopic re-elucidation.

Spectroscopic revalidation before computational reuse

Revalidation should be proportional to structural uncertainty and downstream sensitivity. Ultra-high-resolution NMR and synthesis resolved a substantially incomplete stereochemical assignment in caylobolide A [7], while integrated ECD, NMR, and DP4+ analysis discriminated candidate diphenazine structures [10]. Database-assisted carbon-NMR analysis has also exposed conflicting natural-product assignments [11], and synthesis provided orthogonal confirmation for the revised peganutonin A structure [12]. Together, these cases support using complementary evidence when a structural question is consequential, but they do not establish a universal ranking of analytical methods.

The proposed trigger for revalidation is therefore not database age alone. Re-examination becomes methodologically important when there is unresolved configuration or connectivity, persistent spectral discordance, conflict among source records, unusual chemistry supported by weak evidence, or a modeling task that is demonstrably sensitive to the disputed structural dimension. Where the assignment is mature and the computational task insensitive to the uncertainty, preserving provenance may be sufficient. Where the disputed feature determines the modeled graph, stereochemistry, or molecular state, stronger orthogonal

evidence is warranted before treating the structure as an unqualified input.

Confidence labels for structure assignments

Chemical-identity confidence should be reported separately from model uncertainty. Database provenance can indicate how an assertion entered and changed within a resource [2], while revised natural-product series demonstrate that apparently established structures can later acquire stronger or contradictory evidence [13]. Predictive-uncertainty methods, by contrast, quantify

uncertainty associated with a model and its data-generating assumptions rather than the probability that the molecular identity itself is correct [28]. Conflating these quantities would give a numerical model output epistemic authority it does not possess.

This article therefore proposes qualitative structural-confidence labels based on evidence status, not fabricated probabilities. **Table 3** compares that proposal with neighboring practices and makes explicit which question each method can and cannot answer.

Table 3. Comparison of proposed structural-confidence labeling with existing computational data-quality practices

Practice	Primary question answered	What it cannot establish	Relationship to structural confidence
Structure standardization	Is the record normalized consistently?	Experimental correctness	Transformation metadata
Database provenance	Where did the record originate and change?	Chemical truth by itself	Core confidence evidence
Ligand-state preparation	Which state enters screening?	Dominant experimental state	Separate state-confidence field
Model uncertainty	How uncertain is the prediction?	Correctness of structure assignment	Report independently
Proposed structural-confidence label	How well is identity supported for intended reuse?	Validated probability or universal grade	Qualitative, revisable, provenance-linked

A practical label should state whether the assignment is experimentally corroborated, computationally supported but incompletely resolved, contested by conflicting evidence, or superseded by a revised structure. These categories are proposed reporting aids rather than calibrated probabilities, consensus grades, or regulatory classifications.

Correcting legacy records without losing provenance

Legacy correction should preserve the record that was wrong as well as the record considered current. Chemical-database stewardship depends on traceability [2], and automated curation pipelines illustrate how records may undergo explicit normalization steps over time [16]. ChEMBL's continuing curation demonstrates that chemical and bioactivity resources evolve rather than remain fixed [20], while source-aware natural-product databases make lineage increasingly practical [21]. Erasing the superseded structure may therefore improve apparent cleanliness while damaging reproducibility.

The proposed correction pattern is append-only at the provenance level: preserve the historical structure, identifier relationships, source citation, transformation history, reason for revision, and the date or version at which the corrected structure became preferred. Downstream models should then identify which structural version they consumed. This approach allows previously reported docking poses, descriptors, or predictions to be reconstructed without treating the superseded assignment

as currently valid. It also permits uncertainty to persist where evidence supports more than one plausible representation.

Minimum structural integrity requirements for modeling

Before a natural-product structure is reused computationally, the minimum integrity check proposed here asks five questions. First, is the molecular assignment sufficiently specified for the intended endpoint, or are constitution or stereochemical elements unresolved? Second, can the record be traced to its source and revision history, as modern provenance-aware resources increasingly permit [21]? Third, has the modeled state been declared rather than silently selected, particularly where protonation or tautomerism matters [14]? Fourth, does the chosen representation preserve the structural dimensions that may affect the task, given the documented sensitivity of stereochemical learning [22]? Fifth, has the modeling workflow tested whether plausible structural alternatives materially affect its conclusions?

For docking, the last requirement means recognizing that prepared-input choices condition pose and score interpretation [25]. For machine learning, uncertainty estimates should remain distinct from structural confidence [28]. For similarity and SAR, conclusions should disclose the representation and metric from which the neighborhood was derived [29]. These requirements are proposed minimum reporting and sensitivity principles, not validated thresholds. Their intensity should scale with

the consequence of the structural uncertainty and with the model's demonstrated dependence on the disputed feature. The framework also has important scope limits. Natural-product structure revision is unevenly documented across chemical classes, and the available evidence consists of case studies, database methods, and computational benchmarks rather than a prospective dataset in which the same compounds were modeled before and after confirmed structural correction. Consequently, no error frequency, expected performance decrement, or universal revalidation threshold can be inferred. Model sensitivity will also vary with endpoint and representation: a correction invisible to one fingerprint may be decisive for a stereochemistry-aware model or a three-dimensional docking calculation. Prospective evaluation should therefore compare paired historical and corrected records across multiple representations, preserve assay labels and provenance, and test whether conclusions change under external or temporally separated validation.

The framework is deliberately permissive about legitimate multiplicity. A true racemate, a biologically interconverting protomer, or a deliberately stereo-agnostic representation is not automatically defective. What is unacceptable methodologically is silent substitution: allowing a provisional assignment, a database-normalized form, and a task-specific modeled state to appear as though they were the same evidential object.

CONCLUSION

Computational validation begins too late when molecular identity has already been accepted without qualification. Natural-product structures are evidence-bearing assertions that may be complete, partial, contested, revised, or transformed before they reach a computational workflow. The proposed framework therefore places structural provenance upstream of descriptor generation, docking, similarity analysis, and machine learning and keeps assignment confidence distinct from database consistency, modeled molecular state, predictive uncertainty, and numerical model performance.

Its central methodological principle is containment rather than presumed causation: identify where structural uncertainty enters, preserve how records change, test whether the disputed feature affects the model, and prevent downstream confidence from being misread as evidence that the starting structure was correct. The framework does not estimate the prevalence or quantitative impact of natural-product misassignment, and it is not a validated quality standard. Its value must be tested prospectively by comparing corrected and uncorrected structural records across representations, models, and endpoints. Until then,

chemical identity should be treated as an input requiring evidence, not as a premise made true by computation.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Mansouri K, Moreira-Filho JT, Lowe CN, Charest N, Martin T, Tkachenko V, et al. Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. *J Cheminform.* 2024;16(1):19. doi:10.1186/s13321-024-00814-3
- [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] Zdrazil B. Fifteen years of ChEMBL and its role in cheminformatics and drug discovery. *J Cheminform.* 2025;17(1):32. doi:10.1186/s13321-025-00963-z
- [4] Kutateladze AG, Kuznetsov DM, Beloglazkina AA, Holt T. Addressing the challenges of structure elucidation in natural products possessing the oxirane moiety. *J Org Chem.* 2018;83(15):8341-52. doi:10.1021/acs.joc.8b01027
- [5] Novitskiy IM, Kutateladze AG. DU8ML: Machine learning-augmented density functional theory nuclear magnetic resonance computations for high-throughput in silico solution structure validation and revision of complex alkaloids. *J Org Chem.* 2022;87(7):4818-28. doi:10.1021/acs.joc.2c00169
- [6] Novitskiy IM, Elyashberg M, Bates RW, Kutateladze AG, Williams CM. Penicitone: Structural reassignment of a proposed natural product acid chloride. *Org Lett.* 2023;25(43):7796-9. doi:10.1021/acs.orglett.3c02859
- [7] George MRP, Deering M, Fiorito D, Solomon K, Tidgewell K, Noble A, et al. The combination of synthesis and ultra-high-resolution NMR spectroscopy reveals the correct structure of caylobolide A. *Nat Synth.* 2025;4(7):859-68. doi:10.1038/s44160-025-00762-2
- [8] Nath N, Fuentes-Monteverde JC, Pech-Puch D, Rodríguez J, Jiménez C, Noll M, et al. Relative configuration of micrograms of natural compounds using proton residual chemical shift anisotropy. *Nat*

- Commun. 2020;11(1):4372. doi:10.1038/s41467-020-18093-5
- [9] Bitchagno GTM, Nchiozem-Ngnitedem VA, Melchert D, Fobofou SA. Demystifying racemic natural products in the homochiral world. *Nat Rev Chem*. 2022;6(11):806-22. doi:10.1038/s41570-022-00431-4
- [10] Zhuang Y, Yang F, Menon AR, Song JM, Espinoza RV, Schultz PJ, et al. An ECD and NMR/DP4+ computational pipeline for structure revision and elucidation of diphenazine-based natural products. *J Nat Prod*. 2023;86(7):1801-14. doi:10.1021/acs.jnatprod.3c00306
- [11] Sánchez-Martínez HA, Morán-Pinzón JA, Del Olmo Fernández E, López Eguliz D, Adserias Vistué JF, López-Pérez JL, et al. Synergistic combination of NAPROC-13 and NMR 13C DFT calculations: A powerful approach for revising the structure of natural products. *J Nat Prod*. 2023;86(10):2294-303. doi:10.1021/acs.jnatprod.3c00437
- [12] Saranya NSS, Choudhury R, Supekar PR, Beduru S, Williams CM, Elyashberg ME, et al. Peganutinin A: Computationally informed synthesis of its revised structure. *J Nat Prod*. 2025;88(5):1253-8. doi:10.1021/acs.jnatprod.5c00322
- [13] Sánchez-Martínez HA, Morán-Pinzón JA, del Olmo E, Adserias Vistué JF, Guerrero De León E, López-Pérez JL. Structure revision of pyranoxanthones via DFT-assisted 13C NMR analysis and NAPROC-13 platform: Diagnostic markers and discovery of unreported natural products. *J Org Chem*. 2025;90(49):17240-51. doi:10.1021/acs.joc.5c01371
- [14] Ropp PJ, Spiegel JO, Walker JL, Green H, Morales GA, Milliken KA, et al. Gypsum-DL: An open-source program for preparing small-molecule libraries for structure-based virtual screening. *J Cheminform*. 2019;11(1):34. doi:10.1186/s13321-019-0358-3
- [15] Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardization. *J Cheminform*. 2018;10(1):36. doi:10.1186/s13321-018-0293-8
- [16] 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
- [17] 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
- [18] Dolciemi D, Villascaras-Fernandez E, Kannas CC, Meniconi M, Al-Lazikani B, Antolin AA. canSAR chemistry registration and standardization pipeline. *J Cheminform*. 2022;14(1):28. doi:10.1186/s13321-022-00606-7
- [19] Béquignon OJM, Bongers BJ, Jespers W, IJzerman AP, van der 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
- [20] 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). doi:10.1093/nar/gkad1004
- [21] 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). doi:10.1093/nar/gkae1063
- [22] Tahlil G, Delorme F, Le Berre D, Monflier É, Sayede A, Tilloy S. Stereoisomers are not machine learning's best friends. *J Chem Inf Model*. 2024;64(14):5451-69. doi:10.1021/acs.jcim.4c00318
- [23] Yoshikai Y, Mizuno T, Nemoto S, Kusuhara H. Difficulty in chirality recognition for Transformer architectures learning chemical structures from string representations. *Nat Commun*. 2024;15(1):1197. doi:10.1038/s41467-024-45102-8
- [24] Baimacheva N, Gao X, Aires-de-Sousa J. Evaluation of chirality descriptors derived from SMILES heteroencoders. *J Cheminform*. 2025;17(1):137. doi:10.1186/s13321-025-01080-7
- [25] Flachsenberg F, Ehrt C, Gutermuth T, Rarey M. Redocking the PDB. *J Chem Inf Model*. 2024;64(1):219-37. doi:10.1021/acs.jcim.3c01573
- [26] 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
- [27] Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. *J Chem Inf Model*. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712
- [28] Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing uncertainty in machine learning for chemistry. *J Chem Inf Model*. 2023;63(13):4012-29. doi:10.1021/acs.jcim.3c00373
- [29] Banerjee A, Kar S, Roy K, Patlewicz G, Charest N, Benfenati E, et al. Molecular similarity in chemical informatics and predictive toxicity modeling: From quantitative read-across (q-RA) to quantitative read-across structure-activity relationship (q-RASAR) with the application of machine learning. *Crit Rev Toxicol*. 2024;54(9):659-84. doi:10.1080/10408444.2024.2386260

