



Rebuilding SAR from Fragmented Natural Product Evidence across Chemical Series, Assay Systems, Targets, and Contexts

David Osei¹, Akua Afriyie¹, Kofi Adu²

¹Department of Structure–Activity Relationship Reconstruction, Faculty of Pharmacy, University of Ghana, Accra, Ghana.

²Department of Fragmented Evidence Integration for SAR, Faculty of Pharmacy, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.

ABSTRACT

Structure–activity relationships are usually interpreted within deliberately constructed chemical series measured under reasonably controlled conditions. Natural-product evidence rarely has this architecture. Parent metabolites, congeners, biosynthetic variants, degradation products, semisynthetic derivatives, and mechanistic probes may instead be distributed across different publications, databases, laboratories, assay formats, target constructs, and endpoint definitions. An apparently sparse SAR may therefore represent a fragmented evidence problem in which observations cannot be pooled safely before chemical and experimental context are resolved. This article develops an original evidence-integration framework for rebuilding comparable local SAR from such records. The analysis separates chemical identity, scaffold relationship, assay state, target identity, mechanistic endpoint, potency representation, provenance, inactivity evidence, contradiction, and uncertainty rather than collapsing them into a single activity matrix. The central proposal is that structural relatedness and evidential comparability should be evaluated separately: close analogs can provide weak SAR evidence when experimental contexts differ materially, whereas records from different sources may support a defensible local relationship when identity and context are sufficiently aligned. Reconstructed SAR is therefore treated as context-tagged evidence rather than as a universally harmonized potency surface. The framework is intended to support auditable medicinal-chemistry reasoning and more defensible model training. It is not a validated reconstruction algorithm, does not define universal comparability thresholds, and cannot recover missing experimental states. Genuine activity cliffs, uncertain identities, incomplete metadata, context-dependent pharmacology, and missing negative evidence remain substantive boundaries.

Key Words: Natural products, Structure–activity relationships, Bioactivity curation, Analogue series, Assay context, Molecular machine learning

eIJPPR 2025; 15(2):80-89

HOW TO CITE THIS ARTICLE: Osei D, Afriyie A, Adu K. Rebuilding SAR from Fragmented Natural Product Evidence across Chemical Series, Assay Systems, Targets, and Contexts. Int J Pharm Phytopharmacol Res. 2025;15(2):80-9. <https://doi.org/10.51847/RfYdbfnUqf>

INTRODUCTION

Natural products remain an important source of pharmacologically informative chemical matter, but their evidence base is usually assembled through discovery histories that differ from prospective medicinal-chemistry campaigns. Molecules may be reported because they were isolated from different organisms, discovered years apart, obtained during biosynthetic studies, synthesized for

structure confirmation, or modified later for mechanistic or optimization questions. Contemporary analyses of natural products in drug discovery accordingly describe a chemically and biologically heterogeneous field rather than a uniform collection of matched analogue series [1]. The implication for SAR analysis is that absence of a single curated series does not necessarily imply absence of related chemical evidence.

Corresponding author: David Osei

Address: Department of Structure–Activity Relationship Reconstruction, Faculty of Pharmacy, University of Ghana, Accra, Ghana.

E-mail: ✉ david.osei@ug.edu.gh

Received: 21 November 2024; **Revised:** 28 February 2025; **Accepted:** 02 March 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.



Modern bioactivity resources make this fragmentation increasingly visible. ChEMBL integrates compounds, targets, assays, activity types, and source-linked records across long time periods, enabling relationships to be recovered that were not originally generated as one experiment [2]. Yet discoverability is not comparability. Two values associated with the same target name may differ in assay format, protein construct, endpoint semantics, or measurement convention; conversely, data originating from different publications may still form a meaningful local comparison when those dimensions align. The central problem is therefore not merely finding more records, but determining what each record is evidence of and at what level it can legitimately be compared.

Natural-product repositories expose a second difficulty upstream of activity interpretation: the chemical entities themselves require substantial reconciliation. The recent COCONUT overhaul illustrates the amount of curation required to create a coherent open collection of natural-product structures from distributed sources [3]. For reconstructed SAR, errors in identity are not peripheral metadata defects. Duplicate structures, unresolved stereochemistry, inconsistent component handling, ambiguous mixtures, and name–structure mismatches can create artificial analogies or conceal real ones before pharmacological comparison begins.

This article therefore treats fragmented natural-product SAR as an evidence-integration problem before treating it as a modelling problem. The proposed task is not to maximize the number of compounds assigned to a family. It is to determine which chemically related observations retain enough identity, context, provenance, and uncertainty information to support a bounded local comparison. “Scattered rather than absent” is consequently a reconstruction hypothesis: some evidence chains may become interpretable after context is restored, whereas others should remain explicitly unresolved.

Why published analogs rarely form a clean chemical series
A conventional medicinal-chemistry series is usually prospective: a scaffold is varied intentionally, closely related compounds are tested under relatively stable conditions, and inactive observations remain visible beside active ones. Natural-product analogs often arise through different routes. Congeners may be reported because they co-occur in one organism, appear in related taxa, arise from biosynthetic variation, or are discovered independently in later isolation studies. LOTUS demonstrates why provenance belongs to this evidence structure by linking natural-product molecular records to organisms, publications, and source information rather than treating structures as context-free entities [4].

The Natural Products Atlas similarly curates microbial natural products reported across dispersed literature [5]. Such integration can reveal candidate chemical neighborhoods that no individual paper presented as a medicinal-chemistry series, but common publication, organism, or biosynthetic origin does not establish assay comparability.

Reconstruction therefore requires explicit chemical relations. Reproducible grouping can identify structural neighborhoods independently of naming, while context-dependent analogue-series analysis shows that similarity useful for SAR transfer depends on series and activity context [6, 7]. The framework consequently distinguishes publication-defined sets, provenance-linked sets, structural neighborhoods, and context-conditioned analogue series; none automatically establishes pharmacological equivalence.

The goal is therefore not the largest possible series but the smallest defensible local neighborhood that preserves meaningful chemical variation for later experimental qualification. **Table 1** formalizes these constructs so that chemical connection is not confused with activity comparability.

Table 1. Construct definitions for recovering candidate local natural-product SAR before pharmacological comparison

Construct	Definition in the proposed framework	Analytical use	Evidential boundary
Publication-defined set	Compounds reported within one article or experimental programme	Locates observations generated or discussed together	Does not establish complete analogue coverage or uniform assay conditions
Provenance-linked set	Molecules connected through organism, source, or curated natural-product provenance	Recovers chemically relevant records dispersed across sources	Common provenance does not establish structural proximity, equivalent potency, or common mechanism
Structural neighborhood	Compounds linked by an explicit reproducible structural grouping criterion	Identifies candidate analogs independently of naming or publication boundaries	Structural similarity does not establish transferable SAR
Context-conditioned analogue series	Local structural neighborhood whose usefulness for SAR transfer depends on comparison context	Focuses analysis on chemically related observations relevant to a defined activity question	Series membership and transferability are not globally fixed

Comparable local-SAR candidate	Proposed: a chemically defined local set retained for subsequent experimental-context qualification	Creates the chemical substrate for evidence reconstruction	It is neither a validated SAR nor a claim of potency equivalence
---------------------------------------	--	--	--

note: “Comparable local-SAR candidate” is a proposed evidence-integration construct. Whether it becomes interpretable SAR depends on later identity, assay, target, endpoint, provenance, activity-state, and uncertainty qualification.

Identity resolution before activity comparison

Once candidate chemical relationships have been recovered, identity resolution becomes the first comparability gate. Structure-standardization workflows used for QSAR modelling show that source records often require explicit treatment of components, charge states, aromaticity, stereochemical information, and related representation choices before modelling [8]. In fragmented natural-product evidence, however, the objective should not be to force every record into a maximally simplified representation. It should be to determine which transformations are safe for the comparison being attempted and which details must remain visible because they may alter biological interpretation.

Open-source curation pipelines show that normalization can be implemented as inspectable, reproducible operations rather than undocumented preprocessing [9]. This supports a distinction between source identity, preserving what was originally reported, and comparison identity, recording the resolved entity used for deduplication and structural linking.

PubChem likewise generates standardized representations through explicit rules and validation procedures [10]. A standardized record is therefore a transformation output, not proof that samples from separate experiments were materially identical; stereochemical certainty, salt state, purity, formulation, or analytical confirmation may still matter.

Identity confidence consequently remains part of the evidence record. An observation that cannot support pairwise inference may remain linked at a broader scaffold or provenance level while being excluded from narrower SAR interpretation. This preserves useful evidence without equating cheminformatic deduplication with experimental identity confirmation.

Connecting parent scaffolds, congeners, derivatives, and semisynthetic analogs

Natural-product relationships extend beyond conventional substituent series. NPClassifier demonstrates that natural products can be organized through hierarchical structural and biosynthetic categories, providing a principled way to locate compounds within broader chemical families [11]. Such classification is valuable for evidence discovery, but it is too coarse to serve as a direct SAR rule. Molecules within the same class may differ at several stereochemical, conformational, functional-group, or appendage positions

that independently alter target interaction or assay behaviour.

The more informative question is how one molecule is chemically related to another. Matched molecular pair analysis provides an established framework for localized structural transformations and associated property changes [12]. Applied cautiously to natural products, it distinguishes relatively local changes from congeneric relationships containing several simultaneous modifications.

The proposed model therefore records relationship type and locality separately. Broad scaffold links find candidate neighborhoods; congeneric links capture naturally sampled variation; localized synthetic or semisynthetic changes can support stronger contrasts when experimental contexts align. Parent–derivative terminology alone is insufficient because synthetic chronology does not quantify chemical locality.

This hierarchy concerns chemistry, not pharmacological equivalence. A local transformation measured under incompatible conditions may provide weaker SAR evidence than a broader relation measured under tightly aligned conditions. Reconstruction therefore proceeds from chemical linkage to experimental qualification rather than allowing similarity itself to carry the inference.

Assay context can reverse apparent sar

Potency values can appear directly comparable because they share units even when they measure materially different experimental states. IC₅₀ and K_i estimates vary with laboratory procedures, substrates, target preparations, incubation conditions, detection systems, and analysis conventions. Direct analysis of potency values collected from different sources has shown that combining IC₅₀ or K_i measurements can introduce substantial noise [13]. For retrospective natural-product SAR, converting heterogeneous measurements to a common logarithmic scale is therefore a numerical transformation, not sufficient evidence of experimental equivalence.

Assay context should instead remain part of the activity observation. Assay-aware bioactivity modelling explicitly incorporates biological context rather than treating it as disposable metadata [14]. Relevant descriptors can include biochemical versus cellular format, target construct, species, substrate state, exposure duration, detection technology, and endpoint type.

Different assays should not be declared incomparable by default. Matched-pair analysis shows that some local SARs

remain robust despite inter-assay variability [15]. Experimental mismatch should therefore reduce confidence only when it plausibly changes interpretation of the structural comparison.

The proposed framework uses graded comparability: closely aligned conditions may support quantitative comparison, related settings may support directional corroboration, and materially divergent systems may require separate context-specific SARs. Apparent reversals deserve scrutiny because they may reflect noise, altered target state, changed biological context, or genuine context-dependent pharmacology rather than a single smooth trend.

Target identity and mechanistic endpoint as separate axes

Assay matching remains insufficient when a common target label is treated as evidence that two observations ask the same pharmacological question. Contemporary bioactivity resources explicitly preserve different layers of target-associated evidence: ChEMBL spans multiple assay and activity types, BindingDB focuses quantitative protein–small-molecule binding data, and the IUPHAR/BPS Guide to PHARMACOLOGY provides expert-curated functional pharmacology [16–18]. Their coexistence supports a crucial reconstruction principle: target identity and endpoint identity are related, but they are not interchangeable.

Target resolution may require species, construct boundaries, mutation state, complex composition, or domain restriction, while endpoint resolution asks what was actually measured. Binding can support physical interaction without establishing the same functional outcome observed elsewhere; a functional response likewise does not by itself establish direct engagement.

The framework therefore stores target state and endpoint as separate tags. Multiple local SARs may coexist for one chemical neighborhood—for binding, biochemical function, and cellular response—without being forced into one potency continuum.

This avoids false equivalence, in which records are pooled merely because they share a target identifier, and unnecessary fragmentation, in which related evidence is discarded because endpoint labels differ. Context-preserving reconstruction instead asks what level of comparison each evidence pair supports.

Normalizing evidence without erasing experimental meaning

Once identity, chemical relation, assay state, target, and endpoint have been separated, normalization becomes a controlled evidence operation rather than simple unit conversion. PubChem demonstrates that standardized chemical records can remain connected to source and assay

information [19]. This is important because the reconstructed dataset should retain a route back to the original experimental meaning even when identifiers, units, or structural representations are harmonized.

Papyrus shows that multi-source bioactivity data can be curated specifically for predictive modelling [20]. Yet the resulting comparability depends on explicit inclusion, standardization, and filtering rules. The proposed framework therefore stores raw activity values and relations alongside normalized representations, rather than allowing the normalized value to overwrite source semantics. Inequality signs, censored values, endpoint-specific units, and qualitative activity statements should remain explicit whenever conversion would imply more precision than the source supports.

Borderline active/inactive labels require particular care because experimental uncertainty around a classification threshold can materially affect prediction [21]. “Inactive” should consequently be interpreted as failure to meet a defined criterion under specified conditions, not as universal absence of biological activity. A compound inactive in one biochemical assay may remain relevant in a cellular or alternative target-state context, and a weakly tested compound should not be treated as equivalent to a rigorously established negative.

Normalization therefore serves comparability only when it is reversible and context preserving. The proposed workflow distinguishes representation harmonization from evidential harmonization: the former makes records computationally alignable; the latter asks whether their experimental meaning is sufficiently compatible for a given SAR claim. This distinction also constrains aggregation. Replicate measurements within a sufficiently matched context may justify summary statistics, whereas values from materially different assay states should remain separate even when they share units. The objective is not to preserve every historical formatting difference, but to preserve every difference that could change the interpretation of a structural comparison.

Building context-tagged sar networks

A flat potency matrix cannot preserve the distinctions required by fragmented evidence. Therapeutic-target databases demonstrate that target context can be represented independently from raw activity, while compound–target resources show that dataset composition and provenance differ across classes of chemical evidence. Large chemogenomic datasets also demonstrate the importance of retaining active and inactive observations as explicit components of the evidence base [22–24]. Together, these precedents motivate a local representation in which chemistry, experiment, and provenance remain separable.

The proposed network represents identity-resolved compounds connected by typed chemical relations while activity records retain assay, target, endpoint, provenance, activity-state, and uncertainty tags. Contradictions remain visible because they may indicate context dependence, noise, or a genuine activity cliff. Records derived from the

same experiment also remain distinguishable from independent corroboration.

The network is an interpretive layer over source-linked observations, not a causal graph. Its purpose is to show which evidence dimensions agree, differ, or remain unresolved. **Table 2** separates the relationship types that must remain visible

Table 2. Relationship types retained in a context-tagged natural-product SAR network

Relationship	Function in reconstruction	What must remain separate	Evidential boundary
Chemical relation	Links resolved molecules through scaffold, congener, stereochemical, or localized transformation relationships	Chemical similarity and pharmacological comparability	Does not establish shared target, mechanism, or potency trend
Experimental relation	Links observations with sufficiently aligned assay conditions	Assay implementation and activity value	Comparability remains graded rather than binary
Target/endpoint relation	Preserves target state and measured pharmacological meaning	Binding, biochemical function, and cellular response	Shared target name is insufficient for equivalence
Provenance relation	Retains source, experiment lineage, and record independence	Replication and duplicated reporting	Common origin does not strengthen activity validity automatically
Contradiction relation	Marks incompatible activity patterns without forced reconciliation	Genuine context effect, activity cliff, and possible error	Discordance is not automatically artefact
Reconstruction relation	Proposed: connects evidence-qualified observations into a local SAR interpretation	Evidence record and interpretive claim	Remains conditional, revisable, and testable

note: The taxonomy is a proposed organizational synthesis. It does not define a numerical confidence score or causal network

84

Figure 1 makes visible how these evidence layers converge without becoming equivalent: identity resolution and typed chemical relationships establish the chemical

substrate, experimental-context tags condition local comparison, and uncertainty remains external to the reconstructed core as a limit on chemistry guidance.

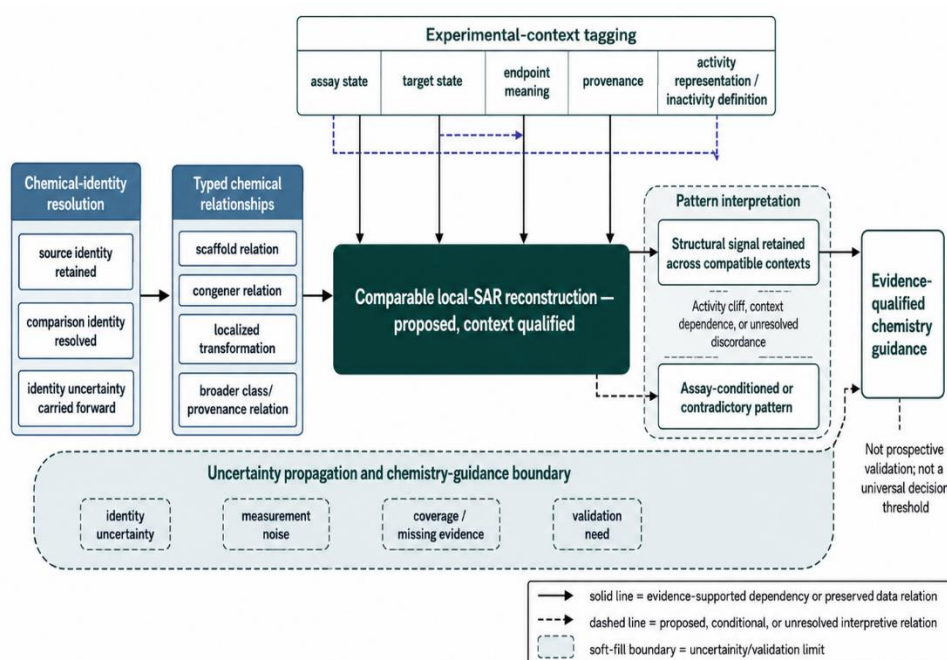


Figure 1. Context-preserving reconstruction of comparable local SAR

A framework for reconstructing comparable local SAR

Locality is essential because structural similarity does not guarantee smooth potency. Activity cliffs show that closely related molecules can exhibit large activity differences that challenge both conventional SAR intuition and molecular machine learning [25]. Such discontinuities should not be “corrected” merely because they complicate a reconstructed trend; they may represent some of the most informative observations in the series.

Analogue-focused prediction studies likewise show that increasing potency divergence creates difficult local prediction regimes [26]. The proposed framework therefore asks three separate questions for every candidate relation: Is the chemical relation sufficiently interpretable? Are the experimental observations sufficiently comparable for the intended claim? Is the observed activity pattern coherent after contextual differences are retained? These questions should not be collapsed into one similarity score. A reconstructed relation can therefore be quantitatively comparable, directionally comparable, context specific, contradictory, or unresolved. Quantitative comparability supports bounded local potency comparison; directional comparability supports a consistent effect without pooled absolute values; context-specific evidence supports separate SARs; contradictory evidence preserves disagreement; unresolved evidence remains connected without supporting structural attribution.

This classification is proposed, not validated. Its purpose is to prevent false precision while allowing evidence states to change when identity information, assay metadata, or matched experiments become available. Importantly, the state belongs to a defined comparison, not permanently to a molecule or series. The same pair may support quantitative SAR for one biochemical endpoint yet remain only directionally comparable for a cellular phenotype. Local SAR is therefore indexed to the experimental question being reconstructed.

Distinguishing structural signal from assay-induced pattern

Data curation and confidence levels can materially alter molecular predictions [27], indicating that apparent structure–activity signal partly depends on how observations are selected and represented. Experimental and label noise further constrain achievable QSAR performance [28]. A reconstructed SAR should therefore infer structural signal only after identity, experimental context, and evidence provenance have been qualified.

Concordant effects across compatible contexts strengthen a local structural pattern; uncertain identity, incompatible assays, duplicated low-confidence observations, or threshold-sensitive labels weaken it. Yet disagreement cannot simply be assigned to artefact: a genuine activity

cliff is structural information, and a context-specific reversal may reveal target-state or endpoint dependence.

The framework therefore separates “structural signal retained across compatible contexts” from “assay-conditioned or contradictory pattern.” This is interpretive, not a claim that one category is biologically true and the other false.

Competing explanations should remain visible until discriminating evidence is available. Early averaging can manufacture smoothness, whereas early exclusion can erase informative context dependence. The output should state the supported pattern, plausible alternative, and evidence needed to discriminate between them.

Uncertainty propagation across fragmented evidence chains

Uncertainty accumulates before any model is trained. Chemical-space coverage can inflate apparent predictive performance when test compounds are well represented by available training chemistry [29]. Representation learning does not, by itself, repair limitations arising from dataset composition, missing labels, or poorly defined prediction tasks [30]. These findings are directly relevant to reconstructed SAR because evidence integration can increase record count without necessarily increasing independent chemical or experimental coverage.

The framework therefore treats uncertainty as multidimensional. Identity uncertainty concerns whether records refer to the intended molecular entity. Measurement uncertainty concerns assay variability, censoring, or threshold dependence. Coverage uncertainty concerns missing analogs, absent negative data, and underrepresented regions of the local series. Interpretation uncertainty concerns whether a discordant observation reflects real pharmacological context or error. Transfer uncertainty concerns whether a reconstructed relation remains informative beyond the represented chemical and biological neighborhood.

These components need not be collapsed into a single numerical probability. Doing so would imply a calibration that the available evidence does not provide. Instead, uncertainty propagates as a boundary on the strength and domain of the reconstructed claim. Uncertainty can also be asymmetric: uncertainty in a weakly characterized inactive may matter differently from uncertainty in the identity of the apparent lead compound. Evidence gaps can therefore become operationally useful. The most informative next experiment may be the one that resolves identity, aligns assay context, tests a missing inactive analogue, or deliberately probes a contradictory local transformation.

When a reconstructed sar is strong enough to guide chemistry

Molecular prediction depends jointly on data, representation, algorithm, and evaluation design rather than on any single modelling component [31]. Evaluation guidance in the chemical sciences further emphasizes that validation should match the intended inference and deployment setting [32]. Best-practice recommendations for chemistry machine learning similarly stress transparent baselines, splits, reproducibility, and careful interpretation of performance [33]. Reconstructed SAR should be held to an analogous standard before it guides synthesis.

Chemistry guidance is therefore strongest when identities are resolved; structural changes are local; assay, target, and endpoint contexts are compatible; inactive observations

have clear meaning; contradictions are explained or retained; and the proposed next compound remains within the represented neighborhood. No dimension is sufficient alone.

Guidance should remain proportional to distance from evidence. A close analogue inside a supported neighborhood is a smaller extrapolation than a new scaffold, species, target state, or cellular context. The framework avoids a universal “SAR confidence” score because no validated common calibration exists. **Table 3** summarizes the dimensions without converting them into a threshold.

Table 3. Evidence dimensions governing chemistry use of reconstructed SAR

Dimension	Stronger evidential state	Weakening condition	Consequence for chemistry guidance
Identity	Relevant molecular state is resolved and source identity is traceable	Ambiguous stereochemistry, composition, or material identity	Restrict or defer pairwise interpretation
Chemical relation	Structural change is local and chemically interpretable	Only broad similarity or multiple simultaneous changes	Treat effect attribution cautiously
Experimental context	Assay, target state, and endpoint are sufficiently aligned	Material biological or methodological mismatch	Prefer separate context-specific SAR
Activity evidence	Trend includes interpretable active, weak, and inactive observations	Sparse, censored, or threshold-ambiguous measurements	Limit quantitative claims
Contradiction	Discordance is explained, reproduced, or retained as a genuine cliff	Conflicting records remain unresolvable	Preserve competing interpretations
Transfer boundary	Proposed chemistry remains near represented chemical and biological evidence	Extrapolation extends beyond local coverage	Require stronger prospective validation

note: These dimensions constitute a proposed evidence-integration synthesis. They are not a validated score, probability, or prospective decision rule.

Implications for machine learning and lead optimization

Accurate bioactivity modelling depends on problem definition and evidence quality as well as algorithm choice [34]. Reconstructed SAR should therefore not be exported as a single pooled potency column if doing so discards the very context needed to interpret the labels. Identity confidence, relationship type, assay state, target state, endpoint, provenance, inactivity semantics, and uncertainty can instead become explicit model inputs, stratification variables, pairwise labels, multitask contexts, or exclusion criteria.

Natural-compound modelling provides evidence that relational target information can support bioactivity prediction [35]. This does not establish that reconstructed local SAR will necessarily improve every natural-product model, but it supports the broader premise that biologically structured relationships can carry predictive information beyond an isolated compound label. Low-data methods likewise show that contextual prediction is possible when task-specific evidence is limited [36], making the quality and composition of a local support set especially important.

For model training, near-duplicate records from one experiment can remain in one evaluation partition, incompatible observations can become separate tasks, activity cliffs can remain deliberate hard cases, and ambiguous labels can remain uncertain rather than forced into binary classes.

For medicinal chemistry, rebuilt SAR can identify which transformation is supported, which context produced the evidence, and which experiment would most efficiently challenge the interpretation. It may also distinguish a chemistry question from a data question: sometimes the next useful step is synthesis of a matched analogue, whereas in other cases repeating an existing pair under harmonized conditions is more informative. It should prioritize hypotheses and matched testing, not replace synthesis or validation.

Limits of retrospective evidence reconstruction

Retrospective reconstruction cannot recover experimental facts that were never recorded. Missing target constructs, uncertain stereochemistry, undocumented purity, absent negative results, and poorly described assays can leave

observations permanently noncomparable. Cross-source potency variation remains a substantive limitation even when values have been standardized [13]. No evidence-integration framework can reconstruct the exact experimental state from a numerical endpoint alone.

Genuine activity cliffs create a second boundary because chemical proximity does not guarantee potency continuity [25]. Coverage bias creates a third because a dense reconstructed network may still represent a narrow chemical region [29]. These are not necessarily data-cleaning failures; they can be intrinsic limits of local inference.

Evaluation discipline cannot convert historical evidence into prospective validation [32]. Publication bias also creates asymmetry because weak or inactive analogs may never have been reported.

The appropriate endpoint is explicit bounding, not elimination, of uncertainty. Unresolved records should remain visible. The framework should be judged by whether it clarifies provenance, comparability, contradiction, and transfer limits rather than whether it produces an artificially complete SAR map.

CONCLUSION

Natural-product SAR can be fragmented simultaneously across molecular identity, chemical relationships, assays, target states, endpoints, publications, and database representations. Pooling these records indiscriminately risks manufacturing regularity, whereas discarding them because they were not generated as one conventional series can waste chemically informative evidence.

This article proposes an intermediate reconstruction layer: resolve identity, type chemical relationships, retain experimental context, preserve inactive and contradictory observations, and state only the local SAR supported by the resulting evidence. The representation is deliberately context tagged, local, revisable, and uncertainty bounded. It separates structural relatedness from pharmacological comparability and structural signal from assay-conditioned pattern, while allowing several endpoint-specific local SARs to coexist when the evidence does not justify their collapse.

Rebuilt SAR can provide a more defensible substrate for medicinal-chemistry reasoning and molecular machine learning, but it is not a validated universal algorithm and does not substitute for prospective experiments. Its principal value is to make the path from fragmented observations to a testable SAR claim explicit.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, et al. Natural products in drug discovery: Advances and opportunities. *Nat Rev Drug Discov.* 2021;20(3):200-16. doi:10.1038/s41573-020-00114-z
- [2] 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
- [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] 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
- [5] 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
- [6] Moreira-Filho JT, Ranganath D, Conway M, Schmitt C, Kleinstreuer N, Mansouri K. Democratizing cheminformatics: Interpretable chemical grouping using an automated KNIME workflow. *J Cheminform.* 2024;16(1):101. doi:10.1186/s13321-024-00894-1
- [7] Yoshimori A, Bajorath J. Context-dependent similarity analysis of analogue series for structure–activity relationship transfer based on a concept from natural language processing. *J Cheminform.* 2025;17(1):5. doi:10.1186/s13321-025-00951-3
- [8] 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
- [9] Bento AP, Hersey A, Félix E, Landrum GA, 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

- [10] Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardization. *J Cheminform.* 2018;10(1):36. doi:10.1186/s13321-018-0293-8
- [11] Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, et al. NPClassifier: A deep neural network-based structural classification tool for natural products. *J Nat Prod.* 2021;84(11):2795-807. doi:10.1021/acs.jnatprod.1c00399
- [12] Yang Z, Shi S, Fu L, Lu A, Hou T, Cao D. Matched molecular pair analysis in drug discovery: Methods and recent applications. *J Med Chem.* 2023;66(7):4361-77. doi:10.1021/acs.jmedchem.2c01787
- [13] 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
- [14] 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
- [15] Nelen J, Pérez-Sánchez H, De Winter H, Van Rompaey D. Matched pairs demonstrate robustness against inter-assay variability. *J Cheminform.* 2025;17(1):8. doi:10.1186/s13321-025-00956-y
- [16] Zdravil B, Felix E, Hunter F, Mannes 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
- [17] 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
- [18] Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Res.* 2024;52(D1):D1438-49. doi:10.1093/nar/gkad944
- [19] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. *Nucleic Acids Res.* 2023;51(D1):D1373-80. doi:10.1093/nar/gkac956
- [20] 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
- [21] Mervin LH, Trapotsi MA, Afzal AM, Barrett IP, Bender A, Engkvist O. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *J Cheminform.* 2021;13(1):62. doi:10.1186/s13321-021-00539-7
- [22] Zhou Y, Zhang Y, Zhao D, Yu X, Shen X, Zhou Y, et al. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res.* 2024;52(D1):D1465-77. doi:10.1093/nar/gkad751
- [23] Heinzke AL, Zdravil B, Leeson PD, Young RJ, Pahl A, Waldmann H, et al. A compound-target pairs dataset: Differences between drugs, clinical candidates and other bioactive compounds. *Sci Data.* 2024;11(1):1160. doi:10.1038/s41597-024-03582-9
- [24] Sun J, Jeliaskova N, Chupakhin V, Golib-Dzib JF, Engkvist O, Carlsson L, et al. ExCAPE-DB: An integrated large scale dataset facilitating big data analysis in chemogenomics. *J Cheminform.* 2017;9(1):17. doi:10.1186/s13321-017-0203-5
- [25] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. *J Chem Inf Model.* 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
- [26] Janela T, Bajorath J. Anatomy of potency predictions focusing on structural analogues with increasing potency differences including activity cliffs. *J Chem Inf Model.* 2023;63(22):7032-44. doi:10.1021/acs.jcim.3c01530
- [27] 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
- [28] Kolmar SS, Grulke CM. The effect of noise on the predictive limit of QSAR models. *J Cheminform.* 2021;13(1):92. doi:10.1186/s13321-021-00571-7
- [29] 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
- [30] Dias AL, Bustillo L, Rodrigues T. Limitations of representation learning in small molecule property prediction. *Nat Commun.* 2023;14(1):6394. doi:10.1038/s41467-023-41967-3
- [31] Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. *Nat Commun.* 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
- [32] Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. *Nat Rev Chem.* 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
- [33] Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for

- chemistry. *Nat Chem.* 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z
- [34] Martinez-Mayorga K, Rosas-Jiménez JG, Gonzalez-Ponce K, López-López E, Neme A, Medina-Franco JL. The pursuit of accurate predictive models of the bioactivity of small molecules. *Chem Sci.* 2024;15(6):1938-52. doi:10.1039/D3SC05534E
- [35] Periwal V, Bassler S, Andrejev S, Gabrielli N, Patil KR, Typas A, et al. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. *PLoS Comput Biol.* 2022;18(4):e1010029. doi:10.1371/journal.pcbi.1010029
- [36] Schimunek J, Luukkonen S, Klambauer G. MHNfs: Prompting in-context bioactivity predictions for low-data drug discovery. *J Chem Inf Model.* 2025;65(9):4243-50. doi:10.1021/acs.jcim.4c02373