Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2023, Volume 3, Issue 2, Page No: 31-40

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

Galaxy Publication

Designing Dereplication Before Isolation: A Decision Architecture for Combining LC–HRMS, MS/MS Similarity, NMR Evidence, and Taxonomic Context in Natural-Product Discovery

Kenji Tanaka[1]*, Yui Kobayashi[1], Takeshi Nakamura[2]

[1] Department of Dereplication and Analytical Strategy, Graduate School of Pharmaceutical Sciences, University of Tokyo, Tokyo, Japan.
[2] Department of LC–HRMS and NMR Integration, Graduate School of Pharmacy, Kyoto University, Kyoto, Japan.

***E-mail** ✉ tanaka.kenji@mol.f.u-tokyo.ac.jp

Received: 10 April 2023; **Revised:** 30 July 2023; **Accepted:** 02 August 2023

ABSTRACT

Dereplication is often described as the recognition of known natural products before resources are committed to repeated isolation and structure elucidation. In contemporary discovery, however, consequential decisions are increasingly made while chemical identity remains incomplete. LC–HRMS features, molecular formulas, MS/MS similarities, predicted compound classes, ranked structural candidates, mixture-level NMR evidence, and taxonomic occurrence each provide different forms of information, with different dependencies and failure modes. This article develops an original methodological decision architecture for using such evidence before full isolation without converting provisional inference into premature identification. The central proposal is to treat dereplication as an action-guiding process in which evidence is retained at its native analytical resolution, dependencies among evidence sources are made explicit, and contextual information modifies rather than substitutes for structural evidence. The architecture distinguishes analyte consolidation from annotation, relational similarity from identity, candidate generation from validation, and taxonomic plausibility from direct chemical observation. It further separates decisions about whether to deprioritize, re-interrogate, enrich, isolate, or escalate a feature from claims that its structure has been established. The framework is intended to improve transparency and resource allocation rather than provide a universal scoring rule. Its principal limitations are the incompleteness of spectral and occurrence databases, metric-dependent similarity, isomeric and stereochemical ambiguity, mixture complexity, correlated computational evidence, and variable taxonomic expression. Prospective testing against independently established structures will be required before claims of decision superiority are justified.

Keywords: Dereplication, Nnatural products, LC–HRMS, Molecular networking, NMR, Chemotaxonomy

How to Cite This Article: Tanaka K, Kobayashi Y, Nakamura T. Designing Dereplication Before Isolation: A Decision Architecture for Combining LC–HRMS, MS/MS Similarity, NMR Evidence, and Taxonomic Context in Natural-Product Discovery. Spec J Pharmacogn Phytochem Biotechnol. 2023;3(2):31-40. https://doi.org/10.51847/kezMDOqZta

Introduction

Natural-product discovery has long used dereplication to avoid spending isolation and structure-elucidation effort on constituents that are already known. The analytical setting in which dereplication now occurs is considerably richer than a comparison between an isolated compound and a reference spectrum. LC–HRMS/MS, NMR profiling, computational databases, and chemometric methods can contribute complementary information while the material remains an extract, fraction, or partially resolved mixture [1]. This expansion changes the methodological question. The relevant problem is no longer only whether a compound can eventually be identified, but also what decision can reasonably be made before complete isolation when the available evidence is heterogeneous, incomplete, and potentially dependent.

Profiling-based strategies have already shifted natural-product workflows toward targeted isolation by using chemical information to determine which regions of an extract merit further work [2]. That development is

important because purification is not analytically neutral: every additional chromatographic step consumes material, time, and attention, and may narrow subsequent investigation toward what is readily detected or recoverable. Yet a decision to isolate a feature and a claim that the feature has been structurally identified are different propositions. Treating them as equivalent encourages either excessive conservatism, in which potentially informative evidence is ignored until purification is complete, or excessive confidence, in which provisional computational annotation is treated as structural closure.

The intermediate evidential space is therefore methodologically important. Complex-mixture analysis can yield compound-class information, substructural evidence, analogue relationships, and network membership even when exact structures remain unresolved [3]. Such outputs are not failed identifications; they are observations at different resolutions. Their value depends on whether a discovery workflow asks them to do more than they can support. A class prediction may be sufficient to recognize chemical redundancy in one context but inadequate to distinguish congeners in another. A molecular-network relationship may justify targeted comparison across fractions but cannot, by itself, settle stereochemistry. Taxonomic occurrence may increase the plausibility of one candidate while remaining vulnerable to incomplete literature coverage, misidentification of source organisms, or genuine chemotypic variation.

This article therefore treats pre-isolation dereplication as a proposed decision architecture rather than a single annotation algorithm. Its central methodological claim is that analytical observations should be translated into actions only after their resolution, provenance, dependence, and uncertainty have been made explicit. The architecture developed below separates where dereplication decisions currently enter a workflow, what evidence is available before full isolation, how complementary evidence can be combined without double counting or premature naming, and what kinds of downstream actions can be justified. The objective is not to replace structure elucidation. It is to design the reasoning that determines when further isolation is warranted, when existing evidence is already sufficient for a limited decision, and when discordance should trigger reassessment rather than forced identification.

Where dereplication decisions are currently made

Dereplication decisions now arise well before a chemist inspects the final spectrum of an isolated compound. Large-scale capture and mining of mass-spectrometric data allow features to be interpreted against project-local and external chemical context, moving judgment upstream into data processing and annotation [4]. Feature-based molecular networking makes this shift explicit by retaining chromatographic feature information while organizing MS/MS relationships, thereby coupling decisions about abundance, retention behavior, and spectral relatedness [5]. The resulting network is useful precisely because it is not a list of established structures: it is an organized field of observations in which some nodes are annotated, others remain unknown, and edges express defined spectral relationships.

Methodological choices within that field already influence what appears chemically important. Reproducible GNPS workflows demonstrate that preprocessing, filtering, networking parameters, and metadata handling are part of the analytical provenance of a molecular network [6]. Ion identity molecular networking further shows that multiple LC–MS features may arise from adducts, isotopes, or related ion species of one analyte, so treating every feature as a separate compound can inflate apparent chemical diversity [7]. An early dereplication decision therefore occurs before candidate naming: signals must first be interpreted as possible observations of underlying analytes.

Other decisions occur at higher structural resolution. Molecular-network topology and structural-similarity fingerprinting can support annotation of natural-product compound families without resolving every member to an exact structure [8]. This family-level information can still influence whether a chemical region appears redundant, unusual, or worthy of targeted follow-up. Similarly, coupling molecular-network information with fraction bioactivity can help prioritize which features or fractions proceed to purification [9]. The association is operationally useful, but it does not prove that a correlated feature is the causal bioactive constituent.

These examples reveal a recurrent pattern: dereplication is already distributed across feature construction, analyte consolidation, spectral comparison, candidate assignment, contextual interpretation, and isolation prioritization. What is missing is not evidence but an explicit rule for preserving the distinction between these decision stages. A defensible architecture must therefore identify what object is being judged—signal, analyte, family, candidate structure, or contextual hypothesis—and what consequence follows from that judgment.

Evidence available before full isolation

Pre-isolation evidence differs not only in strength but in the kind of proposition it can support. Best-practice guidance for mass-spectrometry-based metabolomics emphasizes that annotation language should reflect the evidence actually obtained rather than implying a higher identification level than the data justify [10]. High-resolution MS/MS can substantially constrain molecular formulas and candidate structures through fragmentation-based computation [11], but candidate-space reduction is not equivalent to complete structural proof.

A second evidential layer concerns chemical class and ranked structural hypotheses. CANOPUS can assign systematic compound classes to spectra for which exact structures are unavailable [12]. COSMIC extends candidate ranking with confidence-oriented structural annotation, while its evaluation also demonstrates that incorrect high-confidence assignments remain possible when structurally close alternatives are difficult to distinguish [13]. Class assignment and candidate ranking should therefore be retained as different analytical states rather than collapsed into "identified."

Relational evidence introduces another distinction. Spec2Vec learns MS/MS relationships from spectral representations and can recover structural relatedness beyond conventional similarity scoring [14]. Spectral entropy provides an alternative similarity formulation that outperformed dot-product approaches in the evaluated small-molecule identification setting [15]. Because different similarity functions capture different spectral properties, a similarity edge is partly a product of the comparison model. Hybrid Search further shows that useful analogue evidence may be recovered when an exact reference spectrum is absent, but the result remains an analogue-level inference rather than unique structure confirmation [16].

NMR can contribute orthogonal information before conventional full purification. Direct mixture annotation, diffusion-ordered spectroscopy, and NMR-based compound networking demonstrate that structural or dereplication evidence can be extracted from complex natural-product mixtures without requiring each constituent first to be isolated individually [17-19]. Taxonomic context constitutes a different category again: taxonomically informed scoring can improve candidate ranking by incorporating known organism–compound relationships [20]. Such priors require traceable occurrence knowledge, for which LOTUS provides structured natural-product provenance [21], while the Natural Products Atlas supplies curated organism-linked information for microbial natural products [22].

The distinctions that matter for a pre-isolation decision are summarized in **Table 1.**

Table 1. Pre-isolation evidence classes and the claims they can legitimately support

Evidence class	Primary observation	Useful decision contribution	What it does not establish	Key boundary condition
LC–HRMS feature evidence	Accurate mass, isotope pattern, retention behavior, ion relationships	Consolidates signals and constrains plausible analytes	Complete molecular structure	Adducts, in-source products, coelution, response differences
Formula or candidate evidence	Fragmentation-consistent formulas or ranked structures	Narrows candidate space and guides targeted comparison	Unique identity or stereochemistry	Database coverage and spectral informativeness
Class-level evidence	Predicted chemical superclass or subclass	Recognizes broad redundancy or unexpected chemistry	Specific congener	Model and ontology domain
MS/MS relational evidence	Similarity to reference or neighboring spectra	Detects analogues, families, and related unknowns	Structural equivalence	Metric choice, fragmentation behavior, isomerism
Mixture-level NMR evidence	Correlations, diffusion behavior, recurrent NMR signal patterns	Adds orthogonal substructural or compound-level constraints	Universal resolution of every mixture component	Sensitivity, overlap, exchange, concentration
Taxonomic-context evidence	Documented organism–compound occurrence	Re-ranks plausibility and highlights discordance	Direct observation in the analyzed specimen	Coverage, provenance, chemotype, taxonomy

Note: The classification is proposed here as a decision-oriented organization of evidence. Entries describe distinct evidential roles rather than a universal hierarchy of accuracy.

Combining complementary evidence without premature identification

Combining evidence should not mean converting every concordant signal into another vote for the same structure. Spectral similarity illustrates why. Learned embeddings, entropy-based similarity, and significance-aware spectral alignment recover relationships according to different mathematical definitions and can therefore produce overlapping but nonidentical views of chemical relatedness [14, 15, 23]. Agreement among several MS/MS algorithms may increase confidence that a spectrum occupies a particular structural neighborhood, yet those outputs can remain statistically and chemically dependent because they originate from the same fragmentation event. The number of agreeing tools is consequently not equivalent to the number of independent observations.

Candidate generation requires the same restraint. MSNovelist demonstrates that structures can be generated de novo from tandem-MS information, extending interpretation beyond compounds already present in a candidate database [24]. The important methodological consequence is increased hypothesis space, not automatic structural discovery. Several constitutionally or stereochemically distinct structures may remain compatible with the observed spectrum, and generated candidates should therefore trigger comparison or validation rather than immediate naming.

Statistical calibration also solves only part of the problem. Significance estimation for large-scale spectral matching addresses the risk of accepting high similarity scores merely because large libraries contain many opportunities for chance matches [25]. A statistically meaningful match is still evidence about a spectrum-to-spectrum relationship, not direct proof that every structural feature of the library compound is present in the unknown. Likewise, network-based annotation propagation can use neighboring information and in-silico fragmentation to improve candidate ranking [26], but the propagated result inherits assumptions from its seed annotations, candidate generator, and network topology.

The proposed combination principle is therefore evidence convergence without evidential duplication. Each contribution should retain four attributes: its analytical source, the proposition it directly supports, its dependence on other contributions, and the unresolved alternatives that remain after it is considered. LC–HRMS formula evidence and MS/MS similarity may converge because both constrain plausible chemistry, but they are not fully independent when derived from the same precursor and fragmentation experiment. A taxonomic prior can modify the relative plausibility of candidates, yet it cannot repair an analytically incompatible formula. Mixture-level NMR may provide a genuinely orthogonal constraint, although signal overlap or insufficient sensitivity can leave alternatives unresolved.

Premature identification occurs when concordance is allowed to erase these distinctions. A more defensible workflow asks a narrower question: what action becomes justified after convergence? The answer may be targeted reinjection, comparison with a reference, selective enrichment, NMR interrogation, continued isolation, or temporary deprioritization. Structural naming should remain proportional to the strongest independent evidence available, not to the visual persuasiveness of a network or the number of computational labels attached to one feature.

A pre-isolation decision architecture

A useful pre-isolation architecture should begin with the inferred analyte, not the raw feature list. Automated composition assessment shows why this distinction matters: ion redundancy and detector-response differences can distort interpretations based only on LC–MS feature counts or intensity [27]. The proposed architecture therefore inserts an analyte-consolidation gate before any dereplication conclusion. Mass, retention behavior, isotope patterns, ion relationships, and fragmentation are retained as observations of an analyte hypothesis rather than automatically counted as independent evidence.

Evidence is then organized into four analytically separable channels: physicochemical constraints from LC–HRMS; relational and candidate evidence from MS/MS; orthogonal structural constraints from NMR; and contextual priors from taxonomic occurrence. These channels are not assigned universal numerical weights. Instead, each is recorded with its provenance, resolution, plausible alternatives, and dependence on other evidence.

Taxonomy illustrates the importance of conditional weighting. Chemotaxonomic surveys can reveal meaningful lineage-associated patterns, yet candidate interpretation remains vulnerable to structurally similar fragmentation, incomplete elucidation, and biological variation [28]. Context should therefore modify plausibility only when it does not contradict stronger direct analytical evidence.

The central decision gate asks two separate questions: how constrained is the chemical hypothesis, and what action is justified at that level of constraint? Concordant formula, MS/MS family, NMR, and taxonomic evidence may

justify targeted isolation even if an exact structure cannot yet be named. Conversely, a familiar database candidate accompanied by analytical discordance should trigger reassessment rather than automatic deprioritization. Evidence disagreement is therefore informative: it can indicate isomerism, an unrepresented analogue, incorrect ion consolidation, incomplete taxonomic knowledge, or a genuinely unusual metabolite.

Table 2 translates this proposed architecture into analytically distinct decision conditions without imposing unsupported numerical thresholds.

Table 2. Proposed analytical conditions for pre-isolation dereplication decisions

Evidence condition	Interpretation	Appropriate action	Boundary
Multiple concordant direct observations	Chemical hypothesis is increasingly constrained	Target confirmation or selective isolation	Concordance does not itself prove complete structure
Strong relational evidence, weak direct structure evidence	Family or analogue is plausible	Compare standards, acquire orthogonal data, or enrich	Similarity remains metric- and fragmentation-dependent
Context agrees with analytical candidate	Prior plausibility increases	Retain candidate while preserving provisional status	Occurrence is not direct sample evidence
Context conflicts with strong analytical evidence	Existing prior may be incomplete or misleading	Reassess taxonomy, candidate space, and data quality	Do not force evidence toward expected chemistry
Evidence channels disagree materially	Alternative hypotheses remain viable	Re-interrogate acquisition or proceed to discriminating analysis	Disagreement is not equivalent to analytical failure
Evidence remains nonunique after available tests	Identity remains open	Isolate when the discovery value justifies escalation	Action should be separated from naming

Note: The decision conditions are proposed methodological categories, not validated thresholds or universal confidence levels.

Figure 1 integrates the architecture with the downstream outcomes, uncertainty pathways, and implementation boundary developed across the manuscript.

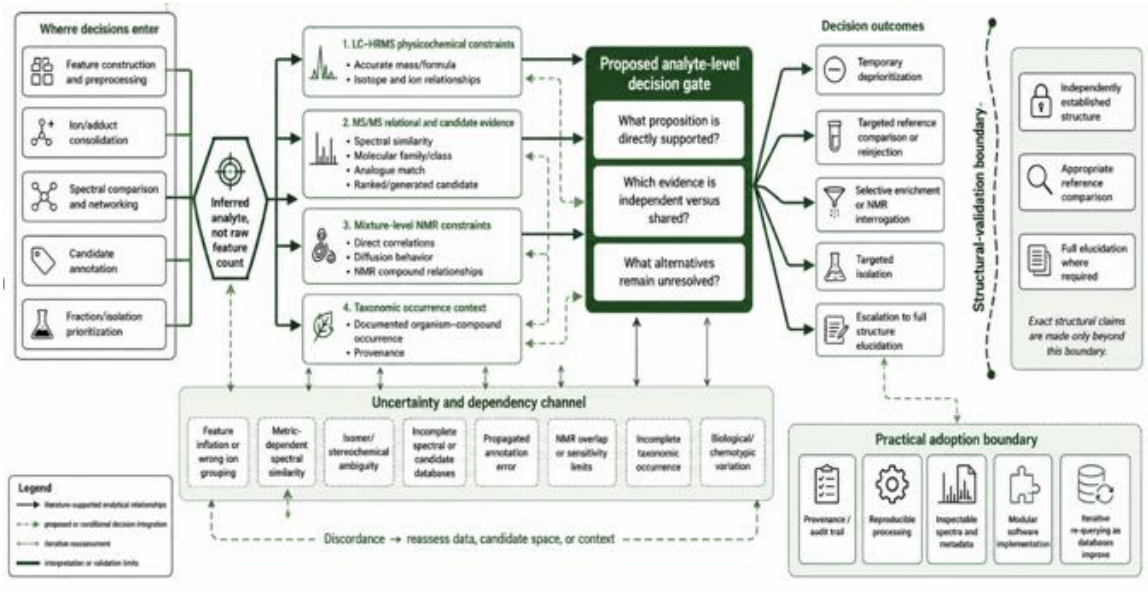

Figure 1. Evidence-bounded pre-isolation dereplication as a conditional decision architecture

Caption. Original proposed synthesis. Distinct evidence channels remain separable before converging on an analyte-level decision gate. Solid connectors represent literature-supported analytical relationships; dashed connectors represent proposed decision integration or reassessment. Discordance feeds back to data and candidate review rather than forcing identification. Downstream actions remain separated from the structural-validation boundary, while practical adoption requires provenance, inspectability, and reproducible processing.

Alt text. Landscape decision architecture in which upstream feature processing and analyte consolidation feed four evidence lanes for LC–HRMS constraints, MS/MS relationships and candidates, mixture-level NMR, and taxonomic context. The lanes converge conditionally on a central decision gate, branch to several discovery actions, and connect to an uncertainty channel that returns discordant cases for reassessment. A final boundary separates pre-isolation decisions from stronger structural validation and routine implementation.

Decision outcomes across natural-product discovery

The proposed architecture deliberately returns actions, not merely names. A feature associated with a well-characterized molecular family may be provisionally deprioritized when the discovery objective is chemical novelty, while the same family-level assignment may justify targeted isolation when biological activity, uncommon substitution, or unresolved analogue chemistry is important. Partial structural knowledge can therefore be decision-sufficient without becoming identification-sufficient, consistent with the broader observation that complex-mixture analysis often yields useful substructural or family information before full structures are available [3].

Targeted isolation is one particularly important outcome. Bioactivity-based molecular networking demonstrates that integrated chemical and fraction-level information can help direct purification toward candidate features [9]. The proposed architecture extends that logic beyond bioactivity: orthogonal evidence may justify selective enrichment, reference comparison, additional NMR acquisition, alternative fragmentation, or complete purification according to what uncertainty must be resolved. Mixture-level NMR can sometimes supply enough additional constraint to avoid indiscriminate isolation, but difficult cases still require conventional structure elucidation.

Taxonomic context should affect these outcomes asymmetrically. Documented occurrence can lower the novelty priority of an otherwise compatible candidate, but absence from a database should not be treated as proof of novelty because occurrence resources remain bounded by published and curated knowledge [20]. Provenance-rich resources make contextual reasoning more auditable, not exhaustive [21]. Accordingly, strong direct analytical discordance should outweigh an expected taxonomic association, whereas a taxonomically unexpected but analytically coherent candidate should normally be escalated rather than rejected.

Table 3 summarizes how evidence states translate into actions while preserving the distinction between discovery decisions and structural claims.

Table 3. Proposed decision outcomes for unresolved natural-product features

Decision state	Typical evidential situation	Action	Identity status
Apparent redundancy	Familiar candidate with coherent supporting evidence	Deprioritize provisionally or archive	Putative unless independently confirmed
Related but nonidentical chemistry	Strong family/analogue relationship with unresolved substitution	Enrich, compare, or selectively isolate	Family or analogue level
Contextual discordance	Analytical evidence conflicts with expected occurrence	Recheck source, candidate set, and acquisition	Open
Orthogonal convergence	MS-derived hypothesis gains independent NMR or reference support	Escalate to focused confirmation	Increasingly constrained
Persistent ambiguity	Several chemically plausible alternatives survive	Full isolation/elucidation if decision value warrants	Unresolved
High discovery value despite weak annotation	Unusual feature, activity, distribution, or chemistry	Retain and prioritize investigation	Unknown by design

Note: These states are proposed action categories. They do not constitute validated probabilities, universal novelty criteria, or substitutes for structural confirmation.

Failure modes and uncertainty

The most consequential failure is premature closure: a candidate becomes treated as an identity because several outputs appear concordant. Even confidence-oriented computational annotation can produce incorrect high-confidence structures when close alternatives are difficult to distinguish [13]. De novo structure generation

broadens the candidate space but does not eliminate structural degeneracy [24]. The architecture therefore requires unresolved alternatives to remain visible rather than disappearing once a preferred candidate has been selected.

A second problem is dependent evidence masquerading as replication. A high candidate rank, statistically significant spectral match, and network-propagated annotation may all arise from overlapping information in the same MS/MS observation. These methods can strengthen interpretation, but none independently converts a shared spectral premise into complete structural proof [13, 25, 26]. The proposed framework therefore asks whether evidence sources are analytically independent before treating convergence as substantially stronger support.

Feature-level errors can enter even earlier. Incorrect adduct grouping or failure to recognize multiple ion forms may inflate compound counts [7]. Conversely, automated composition analysis shows that abundance interpretation remains sensitive to ionization, detector response, and feature redundancy [27]. A prominent feature is not necessarily the most abundant constituent in molar terms, and multiple prominent features are not necessarily multiple compounds.

Contextual evidence has analogous limitations. Taxonomic scoring is useful only to the extent that organism–compound knowledge is accurate and sufficiently complete [20]. Empirical chemotaxonomic work also shows that biological variation and alternative structures compatible with diagnostic fragmentation can complicate interpretation [28]. Taxonomy should therefore function as a conditional prior, not a veto against unexpected chemistry. NMR provides an orthogonal route, but overlapping resonances, limited sensitivity, or similar diffusion behavior can still leave mixtures unresolved [17]. The appropriate response to uncertainty is consequently not always “more algorithms”; it is often acquisition of a measurement capable of discriminating the surviving alternatives.

Practical adoption

The proposed architecture does not require a single new software platform. Its minimum practical requirement is that evidence remain inspectable. The GNPS Dashboard demonstrates how raw and processed mass-spectrometric information can be explored interactively rather than reduced to an opaque terminal annotation [29]. For dereplication decisions, this means retaining spectra, chromatographic context, parameters, candidate alternatives, and metadata alongside the eventual action.

Integrated processing environments can provide much of the necessary infrastructure. MZmine 3 supports multimodal mass-spectrometry processing and annotation within a reproducible analytical environment [30]. The architecture could therefore be implemented as an additional decision layer over existing feature tables, ion-identity relationships, molecular networks, candidate lists, and NMR or taxonomic evidence rather than as a replacement for established tools. The methodological requirement is provenance: a reviewer should be able to reconstruct why a feature was deprioritized, retained, or escalated.

Scalable searching also makes decisions revisable. Flash entropy search demonstrates that very large spectral resources can be queried rapidly [31]. A feature classified as unresolved today may therefore become interpretable when libraries, occurrence databases, or reference spectra expand. Dereplication records should consequently be versioned rather than regarded as immutable conclusions.

Prospective validation should focus on the decision architecture itself, not only on annotation accuracy. A rigorous test would freeze pre-isolation data and candidate information, record the action selected by the proposed architecture, and then compare those decisions with structures established independently after isolation or appropriate orthogonal confirmation. Relevant outcomes could include erroneous deprioritization of genuinely distinct chemistry, unnecessary repeated isolation, retained novel chemistry, and calibration between claimed evidence level and later structural resolution. Until such testing is performed across chemically and taxonomically diverse collections, the framework should be regarded as an explicit, testable method for organizing decisions under uncertainty rather than a validated optimization algorithm.

Conclusion

Dereplication before isolation should be designed around the distinction between what the data indicate and what action those indications justify. LC–HRMS constraints, MS/MS similarity and candidate ranking, mixture-level NMR, and taxonomic occurrence provide complementary but non-equivalent evidence. Their value is lost when heterogeneous observations are collapsed prematurely into a single compound name or when correlated computational outputs are counted as independent confirmation.

The methodological architecture proposed here instead retains evidence at its native resolution, consolidates signals before making compound-level judgments, records dependencies among analytical outputs, uses taxonomic information as a conditional prior, and treats discordance as a reason for reassessment. Its endpoints are decision states—such as provisional deprioritization, targeted comparison, selective enrichment, isolation, or escalation to full elucidation—not claims that structural validation has already occurred.

This approach is intentionally conservative about identity while remaining permissive about action. It allows incomplete evidence to guide resource allocation without pretending that every useful inference is a solved structure. The principal unresolved challenges are spectral and occurrence-database incompleteness, isomeric and stereochemical ambiguity, metric-dependent similarity, mixture complexity, feature inflation, correlated annotation evidence, and biological variability. Practical implementation is feasible with existing analytical infrastructure, but methodological usefulness is not equivalent to validated superiority. Prospective, blinded comparison against independently established structures will be needed to determine whether the architecture actually reduces redundant work while preserving chemically important unknowns. Until then, its contribution is a transparent and falsifiable way to make pre-isolation dereplication decisions without turning plausibility into proof.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Wolfender JL, Nuzillard JM, van der Hooft JJJ, Renault JH, Bertrand S. Accelerating metabolite identification in natural product research: Toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. Anal Chem. 2019;91(1):704-42. doi:10.1021/acs.analchem.8b05112
2. Wolfender JL, Litaudon M, Touboul D, Queiroz EF. Innovative omics-based approaches for prioritisation and targeted isolation of natural products-new strategies for drug discovery. Nat Prod Rep. 2019;36(6):855-68. doi:10.1039/C9NP00004F
3. Beniddir MA, Kang KB, Genta-Jouve G, Huber F, Rogers S, van der Hooft JJJ. Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. Nat Prod Rep. 2021;38(11):1967-93. doi:10.1039/D1NP00023C
4. Jarmusch SA, van der Hooft JJJ, Dorrestein PC, Jarmusch AK. Advancements in capturing and mining mass spectrometry data are transforming natural products research. Nat Prod Rep. 2021;38(11):2066-82. doi:10.1039/D1NP00040C
5. 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
6. Aron AT, Gentry EC, McPhail KL, Nothias LF, Nothias-Esposito M, Bouslimani A, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc. 2020;15(6):1954-91. doi:10.1038/s41596-020-0317-5
7. Schmid R, Petras D, Nothias LF, Wang M, Aron AT, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nat Commun. 2021;12(1):3832. doi:10.1038/s41467-021-23953-9
8. Morehouse NJ, Clark TN, McMann EJ, van Santen JA, Haeckl FPJ, Gray CA, et al. Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting. Nat Commun. 2023;14(1):308. doi:10.1038/s41467-022-35734-z

9. Nothias LF, Nothias-Esposito M, da Silva R, Wang M, Protsyuk I, Zhang Z, et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. J Nat Prod. 2018;81(4):758-67. doi:10.1021/acs.jnatprod.7b00737
10. Alseekh S, Aharoni A, Brotman Y, Contrepois K, D'Auria J, Ewald J, et al. Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. Nat Methods. 2021;18(7):747-56. doi:10.1038/s41592-021-01197-1
11. 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
12. Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol. 2021;39(4):462-71. doi:10.1038/s41587-020-0740-8
13. Hoffmann MA, Nothias LF, Ludwig M, Fleischauer M, Gentry EC, Witting M, et al. High-confidence structural annotation of metabolites absent from spectral libraries. Nat Biotechnol. 2022;40(3):411-21. doi:10.1038/s41587-021-01045-9
14. Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLoS Comput Biol. 2021;17(2):e1008724. doi:10.1371/journal.pcbi.1008724
15. Li Y, Kind T, Folz J, Vaniya A, Mehta SS, Fiehn O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nat Methods. 2021;18(12):1524-31. doi:10.1038/s41592-021-01331-z
16. Cooper BT, Yan X, Simon-Manso Y, Tchekhovskoi DV, Mirokhin YA, Stein SE. Hybrid Search: A method for identifying metabolites absent from tandem mass spectrometry libraries. Anal Chem. 2019;91(21):13924-32. doi:10.1021/acs.analchem.9b03415
17. Egan JM, van Santen JA, Liu DY, Linington RG. Development of an NMR-based platform for the direct structural annotation of complex natural products mixtures. J Nat Prod. 2021;84(4):1044-55. doi:10.1021/acs.jnatprod.0c01076
18. Kleks G, Holland DC, Porter J, Carroll AR. Natural products dereplication by diffusion ordered NMR spectroscopy (DOSY). Chem Sci. 2021;12(32):10930-43. doi:10.1039/D1SC02940A
19. Flores-Bocanegra L, Al Subeh ZY, Egan JM, El-Elimat T, Raja HA, Burdette JE, et al. Dereplication of fungal metabolites by NMR-based compound networking using MADByTE. J Nat Prod. 2022;85(3):614-24. doi:10.1021/acs.jnatprod.1c00841
20. Rutz A, Dounoue-Kubo M, Ollivier S, Bisson J, Bagheri M, Saesong T, et al. Taxonomically informed scoring enhances confidence in natural products annotation. Front Plant Sci. 2019;10:1329. doi:10.3389/fpls.2019.01329
21. 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
22. 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
23. Treen DGC, Wang M, Xing S, Louie KB, Huan T, Dorrestein PC, et al. SIMILE enables alignment of tandem mass spectra with statistical significance. Nat Commun. 2022;13(1):2510. doi:10.1038/s41467-022-30118-9
24. Stravs MA, Dührkop K, Böcker S, Zamboni N. MSNovelist: De novo structure generation from mass spectra. Nat Methods. 2022;19(7):865-70. doi:10.1038/s41592-022-01486-3
25. Scheubert K, Hufsky F, Petras D, Wang M, Nothias LF, Dührkop K, et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun. 2017;8(1):1494. doi:10.1038/s41467-017-01318-5
26. da Silva RR, Wang M, Nothias LF, van der Hooft JJJ, Caraballo-Rodríguez AM, Fox E, et al. Propagating annotations of molecular networks using in silico fragmentation. PLoS Comput Biol. 2018;14(4):e1006089. doi:10.1371/journal.pcbi.1006089
27. Rutz A, Wolfender JL. Automated composition assessment of natural extracts: Untargeted mass spectrometry-based metabolite profiling integrating semiquantitative detection. J Agric Food Chem. 2023;71(46):18010-23. doi:10.1021/acs.jafc.3c03099

28. Barny LA, Tasca JA, Sanchez HA, Smith CR, Koptur S, Livshultz T, et al. Chemotaxonomic investigation of Apocynaceae for retronecine-type pyrrolizidine alkaloids using HPLC-MS/MS. Phytochemistry. 2021;185:112662. doi:10.1016/j.phytochem.2021.112662
29. Petras D, Phelan VV, Acharya D, Allen AE, Aron AT, Bandeira N, et al. GNPS Dashboard: Collaborative exploration of mass spectrometry data in the web browser. Nat Methods. 2022;19(2):134-6. doi:10.1038/s41592-021-01339-5
30. Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrlund TS, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol. 2023;41(4):447-9. doi:10.1038/s41587-023-01690-2
31. Li Y, Fiehn O. Flash entropy search to query all mass spectral libraries in real time. Nat Methods. 2023;20(10):1475-8. doi:10.1038/s41592-023-02012-9