Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2024, Volume 4, Issue 2, Page No: 70-80

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

Galaxy Publication

Two Decades of Natural-Product Dereplication Research: A Scientometric Review of Analytical Platforms, Molecular Networking, Databases, and Shifting Discovery Bottlenecks

Rafael Mendoza[1]*, Isabela Cruz[1], Carlos Lopez[2]

[1] Department of Dereplication and Scientometric Analysis, Faculty of Pharmacy, National Autonomous University of Mexico, Mexico City, Mexico.
[2] Department of Molecular Networking and Discovery Bottlenecks, Faculty of Pharmacy, Monterrey Institute of Technology, Monterrey, Mexico.

***E-mail** ✉ rafael.mendoza@unam.mx

Received: 20 July 2024; **Revised:** 18 August 2024; **Accepted:** 19 August 2024

ABSTRACT

Natural-product dereplication was established to reduce repeated isolation of known chemistry, but its practical meaning has broadened as mass spectrometry, molecular networking, public repositories, structure-prediction tools, and specialized databases have become integrated into discovery workflows. This review examines how the intellectual and technological organization of dereplication changed and whether the principal constraint shifted from detecting known compounds toward interpreting and validating increasingly abundant candidate annotations. A bibliometric/scientometric design was combined with evidence-bounded interpretation of peer-reviewed dereplication methods, databases, analytical platforms, and computational resources published through 2024. Bibliographic performance, network, and thematic analyses were specified as the appropriate corpus-level tools, while a verified evidence layer was used to interpret analytical and computational change. Citation or network prominence was not treated as a measure of chemical validity or methodological quality. The mapped literature supports a progression from preprocessing and spectral database search toward feature-aware molecular networking, repository-scale searching, machine-learning-assisted class prediction, in-silico structure annotation, and increasingly interoperable knowledge resources. The central synthesis proposed here is that dereplication has become less constrained by the simple availability of candidate matches and more constrained by annotation confidence, context, interoperability, unknown chemical space, and the distinction between analytical association and structural proof. Dereplication has evolved from a comparatively discrete filtering task into a distributed evidence-integration problem. Scientometric interpretation is useful for reconstructing this transition, but bibliographic prominence cannot establish analytical superiority, causal influence, or structural correctness, and corpus-derived quantitative claims require verified database-level records.

Keywords: Natural-product dereplication, Molecular networking, Mass spectrometry, Scientometrics, Spectral libraries, Metabolomics

How to Cite This Article: Mendoza R, Cruz I, Lopez C. Two Decades of Natural-Product Dereplication Research: A Scientometric Review of Analytical Platforms, Molecular Networking, Databases, and Shifting Discovery Bottlenecks. Spec J Pharmacogn Phytochem Biotechnol. 2024;4(2):70-80. https://doi.org/10.51847/yDApKRl4Oy

Introduction

Natural-product discovery repeatedly confronts a practical asymmetry: complex extracts may contain large numbers of detectable chemical features, but only a fraction justify isolation, complete structure elucidation, and biological follow-up. Dereplication developed as a means of reducing resources spent rediscovering known compounds, yet contemporary practice no longer consists of a single comparison between an observed signal and a reference. It may incorporate chromatographic and mass-spectral preprocessing, database interrogation, spectral matching, network organization, taxonomic or occurrence context, computational structure inference, and orthogonal confirmation. Contemporary accounts accordingly position dereplication within a broader natural-

product discovery architecture linking analytical, informatic, genomic, and structural resources rather than treating it as an isolated terminal lookup [1].

This expansion is inseparable from changes in mass-spectrometry practice. Improvements in data acquisition, storage, sharing, and mining have transformed spectra from experiment-specific outputs into reusable analytical objects that can be searched, recombined, and interpreted across studies [2]. The scale of the interpretive problem consequently changed. A researcher may now retrieve numerous structurally related spectra, predicted chemical classes, molecular-formula candidates, database records, or network neighbors before obtaining evidence sufficient for definitive structure assignment. The question is therefore no longer simply whether a compound has previously been reported, but how confidently heterogeneous evidence can be linked to the measured feature and how much uncertainty is tolerable before experimental escalation is required.

Specialized natural-product knowledge resources reinforce this transition. The Natural Products Atlas demonstrated how curated microbial natural-product structures and associated information could be organized as an openly accessible discovery resource [3]. Such databases strengthen recognition of known chemistry and provide valuable contextual knowledge, but database occurrence is not equivalent to experimental confirmation of identity in a new biological sample. Similarity, occurrence, and prediction therefore occupy different evidential positions.

This review treats dereplication as an evolving evidence-management problem. Its principal proposed interpretation is that the field's bottleneck has progressively moved from obtaining sufficient analytical discrimination and locating candidate matches toward managing uncertainty among increasingly numerous plausible annotations, connecting heterogeneous information resources, and deciding what level of evidence is sufficient either to stop redundant isolation or to justify further structural work. Scientometric reconstruction provides a means of examining that transition while retaining a strict distinction between scholarly prominence, analytical capability, and chemical validation.

Scientometric objectives

The scientometric objective is not to rank dereplication technologies according to citation frequency. It is to reconstruct how the scholarly organization of the field changed across two decades and to use that organization, together with method-specific evidence, to interpret shifts in analytical emphasis. Bibliometrix provides an established framework for performance analysis and science mapping from bibliographic records, including source, author, citation, collaboration, keyword, and conceptual-structure analyses [4]. Those operations are appropriate for characterizing the organization of a research field; they do not determine whether a spectral annotation is correct or whether one dereplication technology is experimentally superior to another.

Three linked questions organize the reconstruction. First, how did dereplication research diversify from analytical preprocessing and reference-library search into molecular networking, public-data reuse, computational annotation, and database-centered knowledge infrastructures? Second, which concepts appear to constitute recurrent intellectual groupings, and how did their relative emphasis change as platforms matured? Third, is the later methodological literature consistent with a redistribution of the central discovery bottleneck—from obtaining candidate matches toward evaluating annotation confidence, unexplored chemical space, contextual relevance, and interoperability?

The third question is explicitly interpretive. Even an apparent thematic transition cannot demonstrate that one technology caused another to emerge, that publication attention corresponds to analytical performance, or that contemporary computational methods have eliminated classical structure-elucidation problems. Scientometric relationships are therefore interpreted as patterns in the literature rather than proxies for chemical truth.

The verified selection pathway supporting the interpretive evidence layer is shown in **Figure 1**. It intentionally separates evidence-selection counts from unverified corpus-level bibliometric counts so that the former cannot be mistaken for the latter.

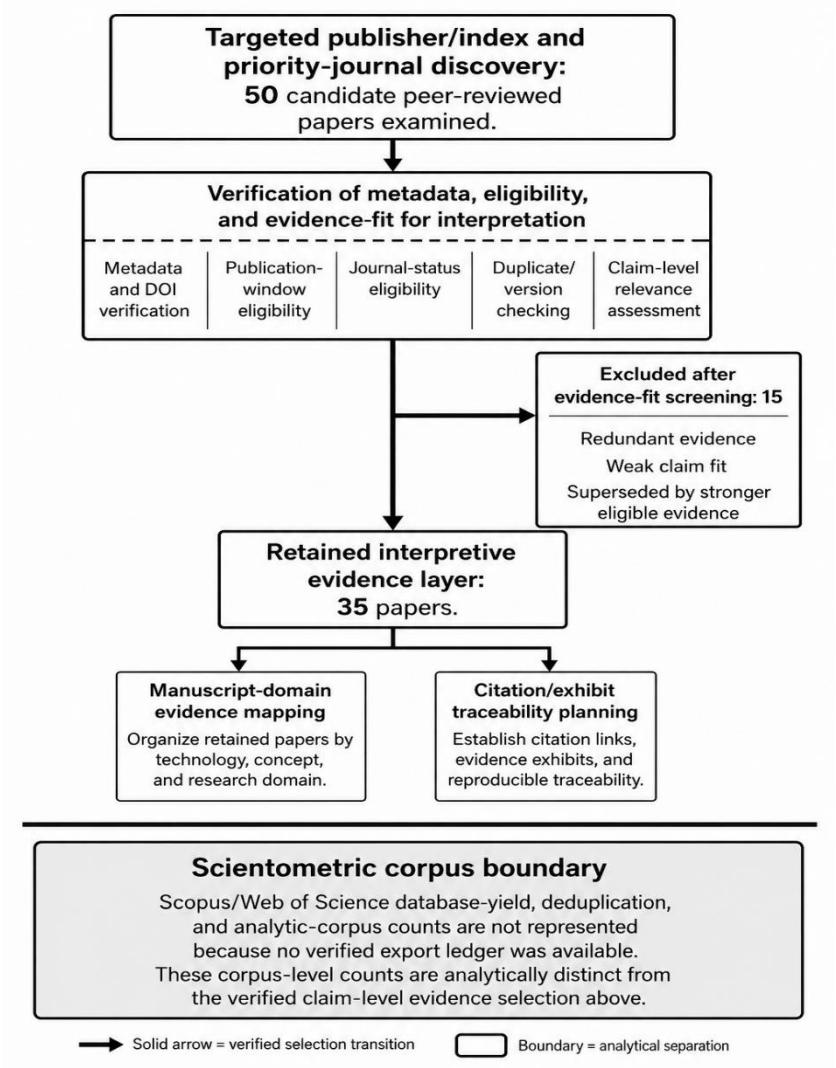

Figure 1. Verified evidence-selection pathway supporting the scientometric interpretation of natural-product dereplication

Methods and Dataset Construction

The review was designed around two analytically distinct evidence layers. The first is the bibliographic corpus required for formal evaluation of publication growth, source distribution, collaboration, co-citation, co-word structure, and thematic evolution. The second is a claim-level scholarly evidence layer used to interpret what particular analytical platforms, databases, and computational approaches actually contribute to dereplication. Keeping these layers separate is essential because a publication may occupy a prominent position in a citation network without providing stronger evidence for chemical identification than a less frequently cited analytical study.

For the corpus layer, the intended retrieval frame covered 2005–2024 and targeted dereplication in natural products, secondary or specialized metabolites, phytochemical investigations, extracts, and metabolomics. Scopus and Web of Science Core Collection were specified as complementary bibliographic sources. Eligible document types were peer-reviewed journal articles and reviews, with duplicate resolution based primarily on DOI and subsequently on normalized title, authorship, and publication year. Bibliometrix was selected as the science-mapping environment for the intended performance, network, and thematic analyses [4].

A verified database-export ledger was not available for this reconstruction. Consequently, database yields, duplicate-removal totals, annual publication counts, country or institutional rankings, source frequencies, network centralities, and thematic frequencies are not reported as observed numerical results. This prevents evidence-selection numbers from being repurposed as fictitious corpus statistics.

The interpretive evidence layer was evaluated separately for peer review, publication period, DOI integrity, journal standing, direct relevance to a planned claim, and limits on transferability. **Table 1** summarizes the evidential roles that follow from this separation.

Table 1 summarizes the distinct evidence layers used in the review and, critically, the types of inference that each layer can support.

Table 1. Evidence layers and admissible inference in the reconstruction of dereplication research

Evidence layer	Principal information	Appropriate analytical use	Interpretation boundary
Bibliographic metadata	Publication, source, author, keyword and citation records	Map field structure and temporal distribution	Does not establish chemical correctness
Network outputs	Co-citation, coupling and co-word relationships	Identify scholarly or thematic organization	Centrality is not methodological quality
Analytical/method papers	Platform-specific experimental or computational evidence	Interpret capabilities and limitations	Findings may remain platform- or dataset-dependent
Database/resource papers	Coverage, curation and search infrastructure	Characterize knowledge organization and reuse	Database occurrence is not sample confirmation
Scholarly reviews	Cross-method conceptual synthesis	Connect developments across domains	Cannot substitute for direct validation
Proposed reconstruction	Integration of temporal and methodological changes	Interpret movement of discovery bottlenecks	Remains inferential unless independently tested

Growth and distribution of the research field

The verified evidence chronology demonstrates technological diversification, but it must not be mistaken for a quantitative publication-growth curve. One early concern in the selected period was whether data preprocessing preserved meaningful relationships in downstream molecular networking. MZmine-based preprocessing was shown to influence molecular-network reliability, establishing that the quality of a dereplication result can depend on decisions made before spectral relationships are interpreted [5]. Database searching then demonstrated how microbial metabolite spectra could be dereplicated systematically against reference information [6], while SIRIUS 4 extended tandem-mass-spectral interpretation toward molecular-formula and structure information rather than requiring exact reference-spectrum recognition alone [7].

By 2020, dereplication was increasingly embedded in connected analytical infrastructure. Feature-based molecular networking integrated chromatographic feature information with tandem-mass-spectral relationships in the GNPS environment, improving correspondence between detected LC-MS features and network organization [8]. MASST reframed an individual spectrum as a query that could be searched across public mass-spectrometry datasets [9]. ReDU complemented this capability by emphasizing standardized sample metadata and systematic reanalysis of public data [10]. In parallel, detailed GNPS protocols formalized molecular-networking procedures into reproducible analytical workflows [11].

Taken together, these developments support a qualitative shift in the distribution of technological functions. Dereplication expanded from local data processing and comparison against known standards toward network-level organization, public-data searching, reusable metadata, and computational structure inference. The change widened both the evidence available for candidate prioritization and the number of analytical decisions separating a measured feature from an accepted structure.

The chronology does not establish which countries, institutions, journals, or individual years dominated publication output. Those are corpus-level scientometric claims that require verified bibliographic records. Method chronology can explain how capabilities accumulated and interacted; it cannot substitute for a validated quantitative map of the literature. Maintaining that distinction is necessary to avoid turning a selectively interpreted technological history into an unsupported claim about the statistical growth or geographic distribution of the field.

Intellectual structure of dereplication research

The intellectual structure represented by the approved evidence is better understood as a set of interacting methodological families than as a single progressive lineage. One family concerns organization of measured ions

and spectral relationships. Ion identity molecular networking addressed ambiguity created when different ion forms derive from the same underlying molecule, allowing those relationships to be incorporated into network interpretation without equating ion association with definitive structural identification [12]. This refinement illustrates an important theme in dereplication: improving organization of analytical evidence does not itself eliminate uncertainty about molecular identity.

A second family shifts attention from exact reference matching toward computational organization of unknown chemical space. CANOPUS used high-resolution fragmentation information to assign compound classes to metabolites even when exact spectral-library identities were unavailable [13]. NPClassifier approached the organizational problem from the natural-product structure side, providing a deep-learning-based classification system tailored to natural-product chemical space [14]. Both approaches support chemical interpretation beyond binary known-versus-unknown matching, but predicted class membership remains different from complete constitutional and stereochemical proof.

A third family concerns accessibility and reuse of analytical evidence. The GNPS Dashboard enabled interactive browser-based exploration of mass-spectrometry data, allowing investigators to interrogate automated analyses rather than treating them as inaccessible computational outputs [15]. This human-facing layer is significant because visual inspection and contextual exploration can support interpretation while still remaining distinct from independent chemical validation.

Knowledge infrastructure forms a fourth family. By 2022, the Natural Products Atlas 2.0 expanded curated microbial natural-product information, NP-MRD established a natural-product-centered NMR resource, and LOTUS advanced an open linked-knowledge model connecting natural-product structures, organisms, and literature [16–18]. Their coexistence indicates diversification rather than convergence on one universal repository: structural records, orthogonal spectroscopic information, biological occurrence, and literature provenance answer different dereplication questions.

This structure suggests that contemporary dereplication distributes evidence across several layers—signal organization, spectral similarity, computational classification, database knowledge, metadata, and researcher interpretation. The important analytical issue is therefore not merely which technology produces the largest candidate list, but what uncertainty each technology resolves and what uncertainty remains afterward.

Table 2 compares these principal methodological families by the analytical object they organize, the contribution they make to dereplication, and the uncertainty they leave unresolved.

Table 2. Method and technology families shaping the intellectual structure of natural-product dereplication

Technology family	Primary analytical object	Principal dereplication contribution	Residual uncertainty
Preprocessing	LC-MS/MS features and spectra	Improves data quality before matching or networking	Parameter dependence and feature-assignment error
Spectral database search	Experimental query spectra	Recognizes known or related reference spectra	Library coverage and match specificity
Formula/structure inference	Fragmentation patterns	Generates molecular-formula or structural candidates	Candidate ranking is not structural proof
Molecular networking	Similarity among fragmentation spectra	Organizes related features and chemical families	Network proximity does not establish identity
Repository search and reanalysis	Public spectra with associated metadata	Adds cross-study occurrence and reuse context	Metadata, instrument and sampling heterogeneity
Structural/class prediction	Unknown spectra or molecular structures	Places incompletely characterized chemistry into classes	Training-space and ontology limitations
Specialized knowledge resources	Structures, spectra, taxa and literature records	Connects prior chemical, spectroscopic and biological knowledge	Curation completeness and transfer to a new sample

Evolution of analytical and computational themes

The most consequential thematic change is the expansion from recognizing reference-like spectra toward managing compounds for which exact references are absent. COSMIC addressed this space by assigning confidence to structural annotations generated without direct spectral-library matches, while MSNovelist

demonstrated de novo structure generation from tandem mass spectra [19, 20]. Both approaches broaden the candidate space, but neither collapses computational plausibility into experimentally established identity. Their importance lies in moving dereplication from binary lookup toward graded hypothesis generation.

The evidential substrate also became more specialized. A reusable mass-spectrometry dataset for medicinal plants used in East Asian traditional medicine expanded plant-specific spectral information [21], whereas automated mining of diagnostic fragmentation patterns enabled targeted discovery of lindenane sesquiterpenoids [22]. Integration subsequently became more explicit: molecular-network topology and structural-similarity fingerprints were combined to annotate natural-product families [23], and repository-scale nearest-neighbor propagation generated suspect spectral libraries from public data [24]. Such propagation increases searchable chemical context, but suspect entries remain categorically different from spectra anchored to authentic standards.

Search and library construction are evolving in parallel. Flash entropy search offered a scalable alternative for querying mass-spectral libraries [25], while LibGen addressed sparse natural-product reference coverage by generating libraries across fragmentation modes [26]. Building Block Extractor further used diagnostic substructural evidence to prioritize natural products with specified features [27]. Across these developments, machine-learning-assisted interpretation increasingly links MS and NMR evidence, although heterogeneous training spaces and validation regimes still limit generalization [28].

Importantly, the newer methods do not solve the same analytical problem. Faster search expands the number of comparisons that can be made; generated libraries expand the reference space; class prediction reduces the informational void surrounding unmatched spectra; and de novo generation expands structural hypotheses. Their outputs therefore should not be placed on one undifferentiated accuracy ladder. A useful dereplication workflow must preserve the provenance of each inference so that confidence gained at one layer is not silently transferred to another.

The synthesis is therefore not a simple replacement of older platforms by newer algorithms. It is an accumulation of complementary evidence types with different confidence ceilings. **Table 3** distinguishes the problems these newer approaches reduce from the gaps they leave open.

Table 3. Analytical advances and the unresolved evidence gaps they redistribute

Advance	Problem reduced	New analytical opportunity	Persistent gap
Beyond-library annotation	Missing exact reference spectra	Rank plausible structures	Confidence remains method- and candidate-space dependent
De novo generation	Database-limited candidate retrieval	Explore structures absent from candidate libraries	Generated candidates require orthogonal confirmation
Targeted spectral mining	Large untargeted feature spaces	Prioritize specified structural motifs	Diagnostic fragments may not uniquely define a molecule
Network–structure integration	Isolated feature interpretation	Annotate families using multiple relationships	Family membership is not exact structural identity
Repository propagation	Sparse experimental libraries	Extend searchable spectral neighborhoods	Propagated suspects lack standard-level certainty
Generated libraries	Limited reference coverage	Expand instrument-specific comparison space	Transfer depends on fragmentation regime
Multimodal machine learning	Fragmented MS/NMR interpretation	Integrate complementary structural evidence	Benchmarks and training domains remain heterogeneous

Emergence of new research fronts

Recent work indicates that the frontier is increasingly defined by integration, context, and scale rather than by one new measurement modality. IMN4NPD integrates molecular-networking operations specifically around natural-product dereplication [29], while NP3 links untargeted metabolomics processing with natural-product prioritization in an open workflow [30]. These systems suggest movement toward end-to-end analytical environments, but workflow integration alone does not validate an annotation.

Infrastructure itself has become a research problem. FragHub addresses spectral-library integration across heterogeneous resources [31], and SpecXplore adds interactive exploration of spectral relationships that can expose patterns or ambiguities hidden by automated scores [32]. Biological context is also entering the search process: microbeMASST connects mass-spectral queries with taxonomic information in microbial metabolomics

[33]. Taxonomic co-occurrence can refine interpretation, but it cannot establish biosynthetic origin without additional evidence.

The direction of discovery can also be reversed. Reverse metabolomics starts from candidate chemical structures and searches public biological datasets for matching signals, demonstrating how repository-scale data can support structure-to-occurrence investigation [34]. MetaboAnalystR 4.0 similarly reflects broader consolidation of LC-MS processing and analysis stages [35]. Together, these fronts support the proposed interpretation that the limiting problem is shifting toward deciding which linked signals, predictions, contextual records, and database relationships deserve confidence.

This convergence also changes what counts as useful context. Earlier dereplication could often be framed around chemical similarity to a known compound. Repository-scale and taxonomically aware tools make occurrence, organism, sample metadata, and prior observation increasingly available as secondary evidence. Such context can prioritize a hypothesis or expose implausibility, but it remains logically separate from direct structural evidence. The frontier is therefore partly an information-architecture problem: heterogeneous evidence must become interoperable without becoming evidentially indistinguishable.

Figure 2 integrates this mature argument while separating evidence-supported technological changes from the proposed reconstruction of the changing bottleneck.

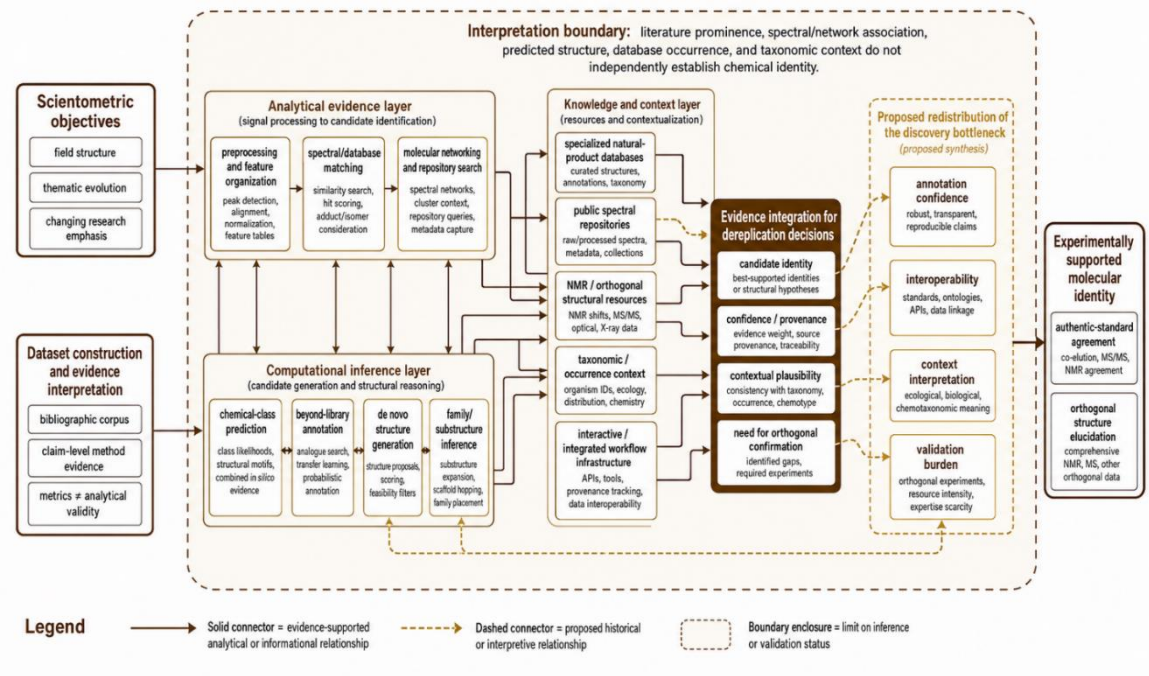

Figure 2. From dereplication tools to an evidence-bounded interpretation of the changing natural-product discovery bottleneck

How the central discovery bottleneck has changed

The earlier methodological emphasis represented in this evidence base concerned obtaining analytically useful signals and organizing them sufficiently well to recognize known or related chemistry. Preprocessing, database search, feature-based networking, public-spectrum search, metadata standardization, and reproducible workflows progressively reduced barriers between raw LC-MS/MS data and candidate interpretation [5]. Yet each solved problem exposed another: better feature organization created more relationships to interpret, and greater access to public spectra increased both contextual information and the possibility of heterogeneous or weakly supported matches [9].

Later approaches attack a different constraint. Class prediction, confidence-aware in-silico annotation, de novo generation, structural-family inference, and repository propagation make it possible to reason about compounds that are poorly represented in conventional libraries [19]. The resulting challenge is no longer merely obtaining a candidate. It is determining whether a candidate is sufficiently supported to justify dereplication, further isolation,

or structure elucidation. Repository-scale suspect libraries make this boundary particularly visible because broad coverage can be gained without conferring authentic-standard certainty [24].

This interpretation has a practical consequence for prioritization. The threshold for stopping an isolation campaign cannot be assumed to be fixed across evidence types. An authentic-standard match, a high-quality library spectrum, a predicted chemical class, a propagated neighbor, and a taxonomically plausible occurrence record answer different questions. Combining them may strengthen a decision, but only if their dependencies and confidence limits remain visible. The changing bottleneck is thus partly a problem of evidence governance: deciding which combinations reduce uncertainty enough to redirect scarce isolation and structure-elucidation effort.

The central bottleneck shift proposed here is therefore a redistribution rather than a substitution. Signal quality, chromatography, spectral quality, and library coverage remain important, but they coexist with newer constraints involving evidence calibration, resource interoperability, biological context, and orthogonal confirmation [31]. **Table 4** translates this interpretation into decision priorities without presenting the sequence as a causally proven historical progression.

Table 4. Proposed interpretation of shifting dereplication bottlenecks and associated priorities

Bottleneck domain	Earlier dominant question	Emerging question	Priority for interpretation
Signal organization	Is the feature analytically credible?	Are related ion forms and spectra organized correctly?	Preserve preprocessing and ion-level provenance
Reference coverage	Is the compound already represented?	How should absent-library chemistry be handled?	Separate exact matches from inferred candidates
Candidate generation	Can a plausible identity be found?	How many plausible alternatives remain?	Report confidence and candidate-space limits
Knowledge integration	Which database contains relevant information?	Can heterogeneous resources be reconciled?	Maintain provenance across resources
Biological context	Where has related chemistry been reported?	Does occurrence context alter prioritization?	Treat taxonomy or occurrence as contextual evidence
Validation	Is rediscovery sufficiently likely to stop isolation?	What evidence warrants accepting or escalating an annotation?	Preserve orthogonal confirmation thresholds

Interpretation of the field's development

The development of dereplication is most coherently interpreted as a change in the distribution of evidential work. Earlier workflows concentrated much of that work at the stages of chromatographic separation, spectral acquisition, and comparison with known compounds. Contemporary workflows distribute it among preprocessing software, similarity measures, network topology, public repositories, machine-learning models, curated databases, taxonomic metadata, and human interpretation. This distribution increases analytical reach but also creates more interfaces at which uncertainty can enter.

Accordingly, the proposed bottleneck reconstruction should not be read as a claim that dereplication has become intrinsically harder or that computational approaches have displaced classical natural-product chemistry. Rather, the object of uncertainty has changed. When a reference spectrum and authentic standard are available, identification may remain comparatively direct. Difficulty grows when evidence consists of predicted classes, propagated spectral neighbors, generated structures, database occurrence, or contextual similarity. These layers can be useful precisely because they allow prioritization before full elucidation, but their usefulness depends on keeping confidence levels distinct.

The development is also non-linear. Mature spectral matching remains useful alongside molecular networking; public repositories depend on preprocessing and metadata standards; machine-learning models depend on representative training data; and contextual searches depend on accurate taxonomic or sample annotations. New layers therefore inherit weaknesses from older ones while adding their own. A chronological narrative should not imply technological obsolescence where the field instead shows stacking, recombination, and recursive reuse of prior analytical capabilities.

The scientometric implication is equally bounded. Increasing visibility of terms associated with molecular networking, machine learning, public repositories, or integrated workflows would be consistent with

methodological diversification, but it would not prove that those approaches improved discovery success. The strongest interpretation arises when literature-level patterns are read alongside method-specific capabilities and limitations rather than used as substitutes for them.

Limitations

This review has an important corpus-level limitation. A verified database export ledger was unavailable for the present reconstruction, so database yields, deduplication totals, annual publication trajectories, geographical rankings, source rankings, centrality measures, and observed thematic frequencies cannot be reported as quantitative scientometric results. The review consequently emphasizes an evidence-bounded reconstruction of intellectual and technological development rather than claiming a complete numerical map of the field.

Additional limitations would remain even with a complete export. Database coverage differs across indexing systems; terminology has shifted from conventional dereplication language toward metabolomics, molecular networking, annotation, and repository mining; and author-keyword normalization can alter apparent thematic structure. Citation measures are sensitive to publication age and community size, while co-citation and co-word networks reflect scholarly attention rather than analytical validity.

The selected method literature is also heterogeneous in chemical space, instrument platform, organismal source, benchmark design, and validation depth. Results from microbial metabolites, medicinal plants, general metabolomics, or particular fragmentation regimes cannot automatically be transferred across natural-product domains. Finally, the proposed changing-bottleneck model is an interpretive synthesis. Testing it requires a reproducible full corpus and comparative analytical studies that measure annotation correctness, false-novelty decisions, confidence calibration, and cross-platform transfer against orthogonal structural evidence.

The claim-level evidence layer is necessarily selective rather than exhaustive. It was constructed to support specific historical, analytical, computational, and interpretive propositions, not to estimate prevalence of every dereplication technology. Absence of a method from the retained evidence set therefore cannot be interpreted as absence from the wider field. This distinction further limits any attempt to convert the present qualitative chronology into publication-frequency estimates.

Conclusion

Natural-product dereplication has developed from a comparatively localized task of recognizing known chemistry into a distributed process for integrating analytical signals, spectral relationships, computational hypotheses, databases, public repositories, and biological context. The central contribution of this review is a proposed reconstruction of that development: as candidate-generation and information-retrieval capabilities expanded, the limiting problem increasingly moved toward determining what those candidates mean, how strongly they are supported, and when an annotation is sufficient to prevent redundant work or justify further structural investigation.

This interpretation does not imply that classical analytical bottlenecks have disappeared. Feature quality, reference coverage, chromatographic resolution, instrument behavior, and authentic-standard evidence remain consequential. Nor does a visible or highly connected scientometric theme establish methodological superiority. Molecular-network proximity, predicted class, generated structure, database occurrence, and taxonomic association occupy different evidential levels and should remain distinguishable from structural confirmation.

Future scientometric work should combine a reproducibly exported bibliographic corpus with transparent terminology normalization and longitudinal mapping. Parallel analytical benchmarking should test how different dereplication strategies alter correct identification, false novelty, reproducibility, and escalation decisions under shared validation standards. Such work would determine whether the proposed bottleneck shift represents a durable restructuring of natural-product discovery or primarily a change in its computational vocabulary and infrastructure.

The most productive next step is consequently not simply more automation. It is stronger linkage between automated candidate generation and explicit evidence status, together with interoperable provenance and orthogonal validation. Dereplication is most valuable when it prevents redundant work without prematurely converting analytical similarity into chemical certainty.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Gaudêncio SP, Bayram E, Lukić Bilela L, Cueto M, Díaz-Marrero AR, Haznedaroglu BZ, et al. Advanced methods for natural products discovery: Bioactivity screening, dereplication, metabolomics profiling, genomic sequencing, databases and informatic tools, and structure elucidation. Mar Drugs. 2023;21(5):308. doi:10.3390/md21050308
2. 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
3. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An open access knowledge base for microbial natural products discovery. ACS Cent Sci. 2019;5(11):1824-33. doi:10.1021/acscentsci.9b00806
4. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959-75. doi:10.1016/j.joi.2017.08.007
5. Olivon F, Grelier G, Roussi F, Litaudon M, Touboul D. MZmine 2 data-preprocessing to enhance molecular networking reliability. Anal Chem. 2017;89(15):7836-40. doi:10.1021/acs.analchem.7b01563
6. Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, et al. Dereplication of microbial metabolites through database search of mass spectra. Nat Commun. 2018;9(1):4035. doi:10.1038/s41467-018-06082-8
7. 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
8. 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
9. Wang M, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, Weldon K, et al. Mass spectrometry searches using MASST. Nat Biotechnol. 2020;38(1):23-6. doi:10.1038/s41587-019-0375-9
10. Jarmusch AK, Wang M, Aceves CM, Advani RS, Aguirre S, Aksenov AA, et al. ReDU: A framework to find and reanalyze public mass spectrometry data. Nat Methods. 2020;17(9):901-4. doi:10.1038/s41592-020-0916-7
11. 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
12. 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
13. 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
14. 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
15. 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

16. van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, et al. The Natural Products Atlas 2.0: A database of microbially-derived natural products. Nucleic Acids Res. 2022;50(D1). doi:10.1093/nar/gkab941
17. Wishart DS, Sayeeda Z, Budinski Z, Guo A, Lee BL, Berjanskii M, et al. NP-MRD: The Natural Products Magnetic Resonance Database. Nucleic Acids Res. 2022;50(D1). doi:10.1093/nar/gkab1052
18. 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. doi:10.7554/eLife.70780
19. 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
20. 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
21. Kang KB, Jeong E, Son S, Lee E, Lee S, Choi SY, et al. Mass spectrometry data on specialized metabolome of medicinal plants used in East Asian traditional medicine. Sci Data. 2022;9(1):528. doi:10.1038/s41597-022-01662-2
22. Li Y, Zhao S, Sun Y, Li J, Wang Y, Xu W, et al. Automatic MS/MS data mining strategy for discovering target natural products: A case of lindenane sesquiterpenoids. Anal Chem. 2022;94(23):8514-22. doi:10.1021/acs.analchem.2c01559
23. 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
24. Bittremieux W, Avalon NE, Thomas SP, Kakhkhorov SA, Aksenov AA, Gomes PWP, et al. Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. Nat Commun. 2023;14(1):8488. doi:10.1038/s41467-023-44035-y
25. 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
26. Kong F, Keshet U, Shen T, Rodriguez E, Fiehn O. LibGen: Generating high quality spectral libraries of natural products for EAD-, UVPD-, and HCD-high resolution mass spectrometers. Anal Chem. 2023;95(46):16810-8. doi:10.1021/acs.analchem.3c02263
27. Zhu D, Yuan J, Yue Q, Zhao Y, Yu S, Abilova Z, et al. Building Block Extractor: An MS/MS data mining tool for targeted discovery of natural products with specified features. Anal Chem. 2023;95(29):10939-46. doi:10.1021/acs.analchem.3c00744
28. Hu G, Qiu M. Machine learning-assisted structure annotation of natural products based on MS and NMR data. Nat Prod Rep. 2023;40(11):1735-53. doi:10.1039/D3NP00025G
29. Sheng Y, Wang J, Liu S, Jiang Y. IMN4NPD: An integrated molecular networking workflow for natural product dereplication. Anal Chem. 2024;96(7):2990-7. doi:10.1021/acs.analchem.3c04746
30. Bazzano CF, de Felicio R, Alves LFG, Costa JH, Ortega R, Vieira BD, et al. NP3 MS Workflow: An open-source software system to empower natural product-based drug discovery using untargeted metabolomics. Anal Chem. 2024;96(19):7460-9. doi:10.1021/acs.analchem.3c05829
31. Dablanc A, Hennechart S, Perez A, Cabanac G, Guitton Y, Paulhe N, et al. FragHub: A mass spectral library data integration workflow. Anal Chem. 2024;96(30):12489-96. doi:10.1021/acs.analchem.4c02219
32. Mildau K, Ehlers H, Oesterle I, Pristner M, Warth B, Doppler M, et al. Tailored mass spectral data exploration using the SpecXplore interactive dashboard. Anal Chem. 2024;96(15):5798-806. doi:10.1021/acs.analchem.3c04444
33. Zuffa S, Schmid R, Bauermeister A, Gomes PWP, Caraballo-Rodriguez AM, El Abiead Y, et al. microbeMASST: A taxonomically informed mass spectrometry search tool for microbial metabolomics data. Nat Microbiol. 2024;9(2):336-45. doi:10.1038/s41564-023-01575-9
34. Gentry EC, Collins SL, Panitchpakdi M, Belda-Ferre P, Stewart AK, Carrillo Terrazas M, et al. Reverse metabolomics for the discovery of chemical structures from humans. Nature. 2024;626(7998):419-26. doi:10.1038/s41586-023-06906-8
35. Pang Z, Xu L, Viau C, Lu Y, Salavati R, Basu N, et al. MetaboAnalystR 4.0: A unified LC-MS workflow for global metabolomics. Nat Commun. 2024;15(1):3675. doi:10.1038/s41467-024-48009-6