



Machine Learning in Tandem Mass Spectrometry for Natural Product Discovery Through Spectral Annotation, Molecular Networking, Dereplication, and Structure Prioritization

Pieter Botha^{1*}, Anel Van Wyk¹, Johan Marais²

¹Department of ML for Tandem Mass Spectrometry, Faculty of Pharmacy, Stellenbosch University, Stellenbosch, South Africa.

²Department of Molecular Networking and Dereplication, Faculty of Pharmacy, University of Pretoria, Pretoria, South Africa.

ABSTRACT

Tandem mass spectrometry has shifted from a predominantly library-centered identification technique toward a computational discovery environment in which spectra can be embedded, networked, propagated, classified, ranked, and used to generate structural hypotheses. For natural products, this expansion is attractive because chemical diversity is high and authenticated reference spectra remain incomplete, yet the resulting outputs span substantially different levels of structural certainty. A systematic mapping review was conducted across analytical chemistry, cheminformatics, metabolomics, and natural-product informatics literature. DOI-unique records published from 2017 through 2025 were screened for peer-reviewed evidence in which MS/MS materially contributed to spectral annotation, molecular networking, dereplication, formula or substructure inference, candidate ranking, de novo generation, or contextual prioritization. Evidence was charted by discovery task, representation, validation design, inference level, uncertainty source, and transferability boundary. Forty-two unique records entered screening and 30 studies were included after title/abstract and full-text assessment. The evidence map shows movement from database search and fragmentation-based inference toward learned spectral representations, analogue retrieval, graph and transformer models, joint spectrum-molecule embeddings, and generative approaches. Better retrieval or ranking nevertheless remains conditional on library composition, candidate-space construction, acquisition conditions, network anchors, and independent structural confirmation. Machine learning broadens the structural hypotheses extractable from MS/MS, but its outputs should be interpreted as task-specific evidence rather than interchangeable forms of identification. The review proposes a task-and-confidence mapping that separates computational reach from structural confirmation and identifies cross-instrument, chemically disjoint, prospective validation as a major priority.

Key Words: Tandem mass spectrometry, Machine learning, Natural products, Molecular networking, Dereplication, Structure

eIJPPR 2025; 15(6):105-114

HOW TO CITE THIS ARTICLE: Botha P, Van Wyk A, Marais J. Machine Learning in Tandem Mass Spectrometry for Natural Product Discovery Through Spectral Annotation, Molecular Networking, Dereplication, and Structure Prioritization. Int J Pharm Phytopharmacol Res. 2025;15(6):105-14. <https://doi.org/10.51847/1PW0moEsyp>

INTRODUCTION

Natural-product mass spectrometry no longer functions only as a terminal comparison between an unknown spectrum and an authenticated library entry. Tandem spectra can now anchor chemical-family annotation,

organize extract-derived features, and connect unknowns to structurally related compounds through computational neighborhoods [1]. Integrated platforms such as MZmine 3 further position MS/MS interpretation within feature processing, alignment, metadata handling, and multimodal analysis rather than treating spectral matching as an

Corresponding author: Pieter Botha

Address: Department of ML for Tandem Mass Spectrometry, Faculty of Pharmacy, Stellenbosch University, Stellenbosch, South Africa.

E-mail: ✉ pieter.botha@sun.ac.za

Received: 28 July 2025; **Revised:** 08 November 2025; **Accepted:** 15 November 2025

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



isolated endpoint [2]. This transition matters for natural products because complex extracts contain many signals for compounds without exact reference spectra, making discovery increasingly dependent on how computational procedures transform incomplete spectral evidence into testable structural hypotheses.

Repository-scale propagation illustrates both the opportunity and the evidential problem. Nearest-neighbor suspect libraries can extend searchable chemical context beyond directly annotated reference spectra and provide plausible structural neighborhoods for previously unmatched features [3]. Yet a propagated analogue, molecular-network neighbor, predicted formula, chemical class, ranked database candidate, and de novo generated structure are not equivalent observations. This review therefore treats machine learning in MS/MS not as one identification technology but as a collection of operations acting at different stages of inference. In natural-product dereplication, the useful question is often whether evidence is sufficient to deprioritize known chemistry, elevate an analogue family, or justify isolation and orthogonal structure elucidation. The review consequently maps where computational methods contribute, what structural claim each output permits, and where uncertainty should remain explicit.

Mapping questions and review scope

The review asks how MS/MS-centered machine learning and adjacent computational methods redistribute work

across spectral annotation, molecular networking, dereplication, formula or substructure inference, and structure prioritization. It further asks which representations are used, whether methods retrieve, propagate, classify, rank, predict, or generate structural hypotheses, and what validation designs support those outputs. A third mapping question concerns confidence: which workflows remain dependent on authenticated references, which infer analogues or structural classes, and which attempt candidate-level or de novo structure assignment. These questions compare inferential roles rather than pooling performance measures that are not directly commensurate across tasks.

The scope is narrower than metabolomics informatics as a whole. Eligible studies had to use tandem mass-spectral information materially in computational structural inference; chromatographic preprocessing alone, molecular-property prediction, and structure classification without MS/MS input were outside the analytic core. Natural-product relevance included direct natural-product applications and broader metabolite-identification methods whose representations or validation designs bear directly on natural-product annotation, but transfer into chemically complex extracts was not presumed. The review therefore preserves distinctions among spectral resemblance, molecular formula, chemical class, candidate rank, proposed structure, and experimentally supported identity. The resulting search and selection boundaries are summarized in **Table 1**.

Table 1. Search and selection architecture used to define the mapped MS/MS evidence base

Review element	Operational boundary	Evidence retained	Excluded adjacency
Publication scope	Peer-reviewed journal articles published 2017–2025	DOI-verifiable analytical and computational studies	Preprints, conference-only papers, reports
Analytical core	MS/MS materially influences structural inference	Annotation, networking, dereplication, formula inference, ranking	Processing without structural inference
Structural output	Evidence ranges from family to candidate or generated structure	Formula, class, analogue, ranked candidate, generated hypothesis	Property prediction without MS/MS
Natural-product relevance	Direct NP studies or transferable metabolite methods	Extract, microbial, spectral-library, benchmark evidence	Unrelated forensic applications
Synthesis role	Map task, representation, validation and uncertainty	Study characteristics and evidence boundaries	Pooled effectiveness or meta-analysis

Note: Search-selection logic is manuscript-derived. Verified screening counts are reported separately from the evidence-selection ledger.

Protocol and reporting framework

The review was designed as a systematic mapping review because its central objective is to characterize the distribution and relationships of methods across discovery tasks rather than estimate one intervention effect. Reporting follows a PRISMA-ScR-compatible logic for eligibility, information sources, selection, data charting, and synthesis [4]. This framework was adapted to computational evidence in which an eligible study may be

an algorithm paper, workflow validation, spectral-resource contribution, or repository-scale analysis, and where conventional intervention risk-of-bias instruments are not directly transferable.

Publication eligibility was separated from evidential usefulness. A bibliographically eligible article could still be excluded if MS/MS was peripheral to its structural inference, if a stronger retained source represented the same methodological role, or if application-specific

conclusions could not be carried into the review without overgeneralization. Search records were normalized at DOI level to avoid duplicate versions. Claim appraisal was correspondingly task-specific: evidence of analogue retrieval could support an analogue-search claim but not authenticated molecular identification; formula accuracy could support formula inference but not stereochemical resolution.

Uncertainty and transferability were charted alongside positive results. Validation design, candidate-set definition, library coverage, acquisition conditions, and chemical-space similarity were treated as part of the evidential interpretation rather than as secondary reporting details. This approach allows computational capability to be mapped without converting benchmark performance into claims of universal identification accuracy.

Eligibility criteria

Studies were eligible when they were peer-reviewed journal articles published from 2017 through 2025, had a verifiable DOI, met the prespecified journal-quality requirement, and contributed directly to at least one MS/MS discovery function. Evidence could be methodological, computational, analytical, resource-oriented, or application-based, provided that an exact claim could be assigned concerning representation, annotation, networking, dereplication, formula or substructure inference, candidate ranking, generation, contextualization, confidence, or validation.

Studies were excluded when they were duplicate versions, secondary reviews replaceable by primary evidence, non-journal outputs, adjacent informatics without material MS/MS structural inference, or application domains whose interpretation would require unsupported transfer. Recency alone did not justify inclusion. A study demonstrating analogue retrieval, for example, could support related-structure searching but not exact identification.

Information sources and search strategy

Searches combined publisher platforms, PubMed-indexed records, and DOI-level metadata verification across analytical chemistry, computational metabolomics, cheminformatics, bioinformatics, and natural-product journals. Query families combined “tandem mass spectrometry” or “MS/MS” with machine learning, transformer, embedding, spectral similarity, molecular networking, dereplication, formula prediction, substructure prediction, candidate ranking, de novo generation, biosynthetic context, taxonomy, library curation, instrument transfer, collision energy, uncertainty, and benchmark terminology.

Searches and verification were completed on 10 August 2026 while bibliographic eligibility was capped at 31

December 2025. Search logic was task-oriented rather than exhibit-oriented. Synonym expansion included latent representation, contrastive learning, graph neural networks, fragmentation trees, domain shift, library bias, candidate-space restriction, collision-energy dependence, and uncertainty calibration.

Evidence selection

Screening proceeded after DOI normalization. Title and abstract assessment removed records without sufficient MS/MS-driven structural inference, secondary evidence where a primary paper was retained, and applications outside the mapped chemical domain. Full-text or complete-record assessment then evaluated whether each study supplied a distinct discovery function, validation design, uncertainty source, or transferability boundary.

Redundancy exclusions were conservative. A method was retained when it introduced a substantively different representation, inference task, validation strategy, or natural-product application. Model complexity was not treated as an inclusion advantage. The purpose was to construct an evidence map of discovery functions rather than a catalogue of increasingly elaborate architectures.

Data charting and taxonomy development

Each included study was charted for publication type, chemical space, analytical setting, data source, reported sample or structure count where relevant, focal task, operational definition, representation, algorithm, comparator, validation design, evaluation metric, uncertainty source, stated limitation, and transferability boundary. Missing information was recorded as not reported rather than inferred.

Taxonomy development was iterative and evidence-led. Initial codes followed observable computational operations—matching, embedding, propagating, classifying, formula inference, candidate ranking, spectrum prediction, structure generation, and contextualization—before being assembled into broader discovery-task families. The resulting taxonomy is proposed as a review-level synthesis rather than an established ontology.

A second coding layer distinguished library-dependent, network-dependent, candidate-set-dependent, and open-ended generative approaches. This distinction is important because freedom from exact library matching does not necessarily mean independence from existing chemical databases or training distributions.

Mapping and synthesis approach

The synthesis used descriptive evidence mapping rather than pooled effect estimation. Studies were positioned according to the computational operation performed on

MS/MS evidence and the structural level of the resulting hypothesis. A learned spectral-similarity model and a molecular-formula predictor were therefore treated as different inferential operations even if both ultimately contributed to structure prioritization.

Validation evidence was interpreted in relation to train–test construction, spectral-library composition, candidate-space definition, acquisition platform, collision conditions, network dependence, and availability of authenticated standards. Cross-study synthesis emphasized convergences and discontinuities rather than global algorithm ranking.

The review additionally separates expansion of the hypothesis space from structural confirmation. A method may generate more plausible structures while increasing downstream discrimination burden; conversely, a restrictive candidate universe can produce high ranking accuracy while excluding unusual chemistry. Disagreement among spectral, formula, network, taxonomic, or candidate-ranking signals was therefore considered informative rather than something that should automatically be collapsed into a single score.

Evidence-selection results

Forty-two DOI-unique candidate records entered screening. Five were excluded during title and abstract assessment because they lacked sufficient MS/MS-driven structural inference, represented secondary or infrastructure-adjacent evidence, or concerned an application domain outside scope. Thirty-seven records proceeded to full-text or complete-record assessment, where seven additional records were excluded principally for methodological redundancy or insufficient centrality of MS/MS.

Thirty studies entered data charting and evidence synthesis. The included evidence was heterogeneous, spanning analytical search methods, machine-learning architectures, molecular-network workflows, natural-product applications, and spectral resources. Study counts were therefore used descriptively rather than to infer pooled comparative performance.

Evolution of machine learning in natural product ms/ms

Early computational expansion focused on extracting additional structural information from fragmentation patterns and repositories. DEREPLICATOR+ demonstrated database-driven dereplication across multiple natural-product classes [5], while SIRIUS 4 established fragmentation-tree-based routes from high-resolution MS/MS toward molecular-formula and structure prioritization [6]. Spec2Vec subsequently showed that structural relationships between spectra could be learned

instead of being represented exclusively through predetermined peak-overlap rules [7].

CANOPUS extended inference toward compound-class prediction [8], whereas COSMIC explicitly addressed confidence when ranking structures absent from spectral libraries [9]. MSNovelist demonstrated that MS/MS-conditioned models could generate structures outside a fixed candidate database [10]. Formula-aware transformer models such as MIST further shifted the field toward learned spectrum–structure representations [11]. This sequence should not be interpreted as linear replacement: database search, fragmentation-based inference, learned similarity, ranking, and generation solve different parts of the annotation problem.

Spectral representation and embedding

Representation determines which spectral relationships a computational model can recognize. MS2Query combines learned spectral information with complementary evidence to retrieve structural analogues [12]. Graph-transformer models invert the mapping by predicting tandem spectra from molecular graphs, creating a forward-model route for comparing structural candidates [13].

MSBERT uses masking and contrastive learning to organize spectra in chemically informative latent space [14], while TransExION learns structure-relevant similarity and exposes fragment and neutral-loss relationships contributing to that score [15]. Learned representations can therefore encode structural relationships beyond fixed peak-overlap scoring [7, 14, 15]. Their latent organization nevertheless remains a statistical representation: proximity in embedding space does not establish a unique fragmentation mechanism or molecular identity.

Library matching beyond exact spectral identity

Exact-reference absence does not terminate computational searching. Hybrid Search combines conventional fragment matching with mass-shift relationships to retrieve metabolites related to, but absent from, reference libraries [16]. Library quality nevertheless remains an upstream determinant of both conventional and learned matching. Reproducible cleaning with matchms demonstrates how duplicated spectra, inconsistent metadata, and heterogeneous processing alter the usable spectral evidence [17].

Propagated suspect libraries, learned analogue retrieval, and shift-aware searching consequently broaden structural hypothesis generation beyond exact references [3, 12, 16]. Their output is still an analogue-level hypothesis. Structural resemblance does not automatically resolve positional isomers, stereochemistry, or other alternatives requiring orthogonal evidence.

Molecular networking and chemical-family discovery

Molecular networking changes the unit of interpretation from an isolated tandem spectrum to relationships among spectra. Network Annotation Propagation combines network proximity with in-silico fragmentation to distribute candidate hypotheses across connected features [18]. Feature-based molecular networking retains chromatographic feature information while organizing MS/MS relationships [19], whereas IMN4NPD adapts integrated networking specifically to natural-product dereplication [20].

These methods collectively support network-based organization and propagation of annotations [18-20]. Their evidential boundary is equally important. A network edge indicates a computationally defined relationship between spectra; it does not inherently establish an identical scaffold, shared biosynthetic origin, or common pharmacological effect.

Machine-assisted dereplication

Dereplication becomes most useful when computational evidence changes experimental priority rather than merely adding a provisional name. Repository-scale approaches can recognize known or related natural products across chemically diverse classes [5], and integrated natural-product networking can combine spectral organization with annotation evidence to reduce repeated investigation of familiar chemistry [20].

The appropriate output is therefore a decision about whether a feature appears sufficiently explained by known

chemical space to warrant lower isolation priority. Failure to obtain a library hit should not be inverted into evidence of novelty. An unmatched feature can arise from missing references, unusual fragmentation, acquisition differences, incomplete candidate coverage, or genuinely new chemistry. Computational dereplication can prioritize novelty assessment; it cannot alone establish novelty.

Formula and substructure prediction

Formula and substructure inference occupy an intermediate level between spectral resemblance and complete molecular structure. BUDDY reconstructs molecular-formula evidence through bottom-up interrogation of fragment relationships [21]. MIST-CF instead learns molecular-formula and adduct ranking directly from tandem spectra [22]. Autoregressive fragmentation-graph models provide another intermediate representation by predicting molecular fragmentation structure rather than treating spectra as undifferentiated vectors [23].

Together these studies demonstrate that formula, class, fragment, and candidate evidence can be computationally separated. This separation prevents a common interpretive collapse: accurate formula prediction does not resolve constitutional isomers, while class prediction does not define the precise molecular structure.

The characteristics of representative studies spanning these intermediate inference levels are summarized in **Table 2**.

Table 2. Representative MS/MS approaches linking spectra to formula-, class-, and fragmentation-level structural evidence

Study	Principal task	Representation or method	Structural output	Important evidential boundary
SIRIUS 4 [6]	Formula and structure prioritization	Fragmentation trees	Formula and ranked candidates	Isomer discrimination remains candidate-dependent
CANOPUS [8]	Compound-class prediction	Learned fragmentation fingerprints	Chemical class	Class assignment is not exact structure
BUDDY [21]	Molecular-formula discovery	Bottom-up fragment formulas	Ranked formula hypotheses	Noise and adduct ambiguity affect inference
MIST-CF [22]	Formula/adduct ranking	Formula-aware transformer	Ranked formula hypotheses	Validation remains distribution-dependent
Fragmentation graphs [23]	Fragmentation prediction	Autoregressive graph generation	Predicted fragment relationships	Predicted graph is not measured mechanism

Note: Study characteristics reflect reported methods. The separation of formula-, class-, fragment-, and candidate-level outputs is the proposed evidence organization used in this review.

The verified evidence-selection process underlying the mapped literature is shown in **Figure 1**.

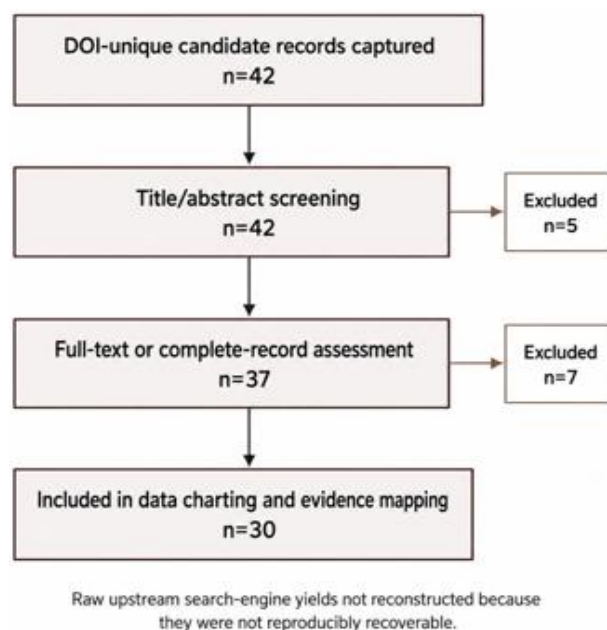


Figure 1. Evidence-selection pathway for the systematic mapping review

Candidate structure ranking

Once candidate structures have been assembled, the problem changes from generation to discrimination. JESTR places spectra and candidate molecules within a joint embedding space so that candidates can be ranked through learned spectrum–molecule correspondence [24]. FIORA supplies a complementary route by predicting spectra from local fragmentation events, enabling experimental and predicted spectral information to contribute to candidate comparison [25]. Such methods can make prioritization substantially more computationally expressive, but the meaning of ranking remains conditional on the candidate universe. A top-ranked structure cannot be considered an identification when the correct structure was never present among the candidates. Validation design must therefore distinguish ranking within a known candidate set from recovery of genuinely unseen chemistry.

De novo annotation and generative approaches

Generative methods address the fixed-database constraint by proposing structures rather than only retrieving or ranking them. Optimized de novo molecular generation uses transfer and reinforcement learning to generate molecules conditioned on mass-spectral evidence [26]. Such methods expand the structural hypothesis space and could be useful where natural-product scaffolds are poorly represented in conventional libraries.

The same openness increases uncertainty. A generated molecule may be chemically valid and spectrally plausible while being absent from the biological sample. De novo annotation should consequently be viewed as hypothesis

production followed by discrimination and orthogonal validation, not as database-free molecular confirmation.

Integrating biosynthetic and taxonomic information

MS/MS interpretation can be strengthened by biological context when contextual evidence remains distinguishable from structural evidence. microbeMASST connects spectral occurrence to microbial taxonomy, enabling a query spectrum to be examined against taxonomically organized public metabolomics data [27]. Biotransformation-aware global molecular networking can use plausible transformations among related features to generate and rank candidate structures [28]. Knowledge- and data-driven two-layer networking further demonstrates how biochemical relationships and empirical spectral associations can be combined while remaining conceptually distinct [29]. Taxonomic occurrence does not establish which organism produced a molecule, and a plausible biotransformation does not prove an enzymatic pathway. These information sources constrain candidate interpretation rather than replacing chemical validation.

Confidence levels across annotation workflows

The mapped literature supports a distinction between computational reach and structural confirmation. Learned similarity can retrieve analogues, formula models can reduce elemental possibilities, class predictors can delimit chemical families, and candidate rankers can organize plausible structures. Confidence-aware ranking explicitly addresses uncertainty when direct library evidence is absent [9].

The proposed review synthesis therefore treats confidence as evidence-type dependent rather than as a universal

score. An authenticated spectral match, analogue relationship, molecular formula, chemical class, ranked candidate, and generated structure answer different questions. Combining them can improve prioritization, but a numerically high model score cannot remove missing

candidates, ambiguous isomers, or absent experimental confirmation.

Figure 2 visualizes this proposed distinction by separating the breadth of computational hypothesis generation from the degree of structural confirmation supported by each evidence type.

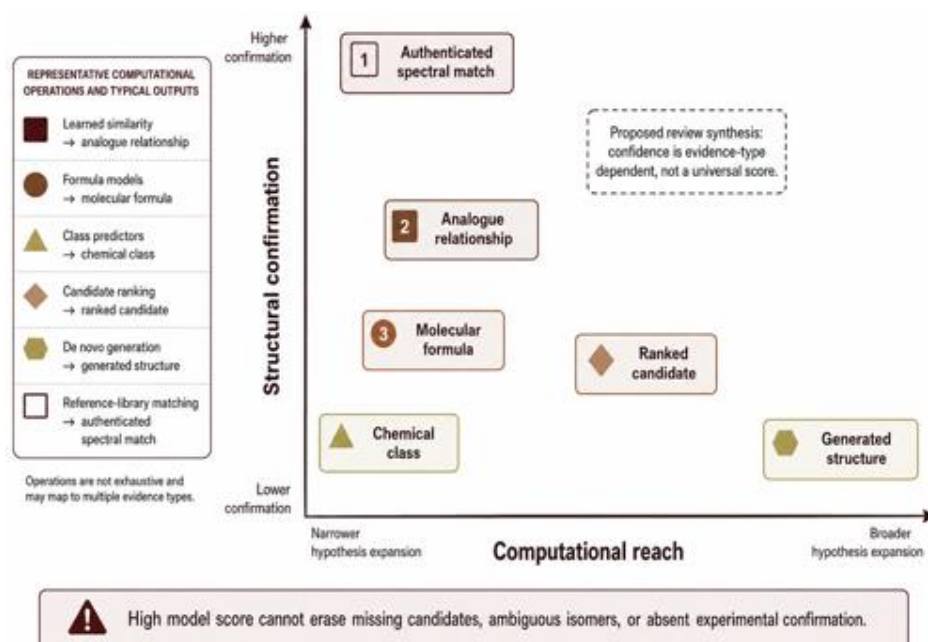


Figure 2. Computational reach versus structural confirmation in MS/MS annotation

111

Dataset concentrations, blind spots, and benchmark limitations

Model performance cannot be separated from the data defining the task. Graph-transformer spectrum-prediction studies illustrate how chemical similarity and train-test construction influence forward prediction [13], while spectral-library cleaning demonstrates how metadata quality, duplication, and preprocessing alter the usable evidence base [17]. MSnLib expands open multistage fragmentation resources to more than two million spectra across 30,008 unique small molecules [30], but a large standardized collection does not automatically reproduce the structural diversity encountered in natural-product extracts.

Benchmark interpretation should therefore remain conditional on compound coverage, acquisition platform, ionization mode, collision regime, candidate construction, and similarity between training and test chemical spaces. Improvements obtained within one benchmark do not establish transfer to another instrument, taxon, compound class, or natural-product discovery setting.

Coverage gaps across instruments, taxa, and compound classes

The principal coverage gaps are multidimensional. Learned embeddings trained on particular collections may

transfer differently across acquisition regimes [14], and explainable similarity learned from GNPS-derived data cannot automatically be assumed to retain identical calibration on another instrument or fragmentation protocol [15]. Library-cleaning decisions also affect which compounds and metadata remain available for downstream training or retrieval [17].

Taxonomy-aware searching introduces a different boundary because public microbial datasets unevenly represent organisms, environments, and sampling programs [27]. Large standardized spectral resources improve reference availability [30], but absence from a repository remains ambiguous. Coverage should therefore be characterized across acquisition context, taxonomic representation, and compound class rather than reduced to library size alone.

Evidence map of methods versus discovery tasks

Across the mapped literature, computational methods occupy overlapping but non-equivalent positions along the natural-product discovery process. Database and analogue searches retrieve existing chemical knowledge; molecular networks organize relationships among observed features; formula and class predictors constrain structural space; joint embeddings and forward-spectrum models prioritize candidates; generative methods create additional

hypotheses; and taxonomic or biotransformation layers add biological context.

The principal synthesis proposed here is that methods are more usefully mapped according to what operation they perform on evidence and what structural claim their output permits than according to architecture alone. A transformer may perform similarity learning, formula inference, or spectrum prediction. Molecular networking may support family discovery, annotation propagation, or dereplication. Architectural labels therefore communicate less about evidential meaning than task and confidence level.

This evidence map also clarifies why “annotation accuracy” can conceal important heterogeneity. Correctly predicting a molecular formula and correctly ranking an exact structure are different achievements, while recovering an analogue may be scientifically useful despite not providing molecular identity. Comparative evaluation should therefore preserve the resolution level at which success is defined.

Research priorities

Future benchmarking should emphasize conditions most likely to change during real natural-product discovery. Priority designs include scaffold-disjoint and temporally separated test sets, cross-instrument and cross-collision-energy evaluation, and explicit stratification by natural-product class and taxon. Prospective blinded analysis of complex extracts would test whether performance established on retrospective libraries survives realistic unknown chemistry.

Candidate-ranking studies should state whether the correct structure is guaranteed to appear in the candidate set. Generative studies should separate chemical validity, spectral plausibility, and recovery of independently supported structures. Confidence calibration should operate against evidence levels rather than relying only on top-k ranking accuracy.

A further priority is deliberate preservation of disagreement. Conflicting spectral similarity, molecular-formula, network, taxonomic, or candidate-ranking signals can reveal where models are extrapolating or where the evidence base is internally inconsistent. These priorities are proposed from the mapped limitations and require prospective evaluation rather than being treated as validated benchmarking standards.

RESULTS AND DISCUSSION

The evidence map shows that machine learning has expanded tandem mass spectrometry from a comparison technique into a layered hypothesis-generation environment. The major transition is not elimination of reference dependence but redistribution of it. Exact

spectral libraries can be supplemented by analogue neighborhoods, learned embeddings, fragmentation models, molecular networks, candidate databases, generative models, and biological-context layers. Each can recover information unavailable to straightforward exact matching, yet each introduces its own dependency on training distributions, candidate construction, network anchors, or contextual assumptions.

For natural-product discovery, computational breadth should therefore not be mistaken for structural certainty. A workflow can increase the number of chemically plausible candidates while simultaneously making discrimination harder. Conversely, a restrictive candidate space may produce impressive ranking performance while excluding structurally unusual natural products.

The proposed task-and-confidence mapping places dereplication, family discovery, formula inference, candidate ranking, and de novo generation on related but distinct evidential tracks. **Figures 1 and 2** address complementary questions: the first makes the review-selection process auditable, whereas the second makes visible the conceptual distinction between computational reach and structural confirmation. Neither implies a quantitative performance hierarchy.

The practical value of these methods lies in prioritizing subsequent chemical work. Computational evidence can determine which spectra merit isolation, which compounds are probably familiar, which belong to related chemical families, and where structural alternatives remain unresolved. Orthogonal analytical evidence nevertheless remains necessary whenever the scientific question requires exact molecular identity rather than computational plausibility.

Review limitations

This mapping review is limited by terminology that differs across metabolomics, cheminformatics, analytical chemistry, and natural-product informatics, making complete retrieval difficult despite synonym expansion. Restriction to peer-reviewed journal literature meeting the stated selection requirements may omit technically important preprints or conference contributions.

The included studies also use heterogeneous spectral collections, validation splits, candidate spaces, instruments, collision regimes, and metrics, so direct global performance ranking would be misleading. Natural-product-specific validation is uneven because several influential machine-learning methods were developed on broader metabolomics collections. Their transfer to structurally unusual natural-product extracts therefore remains conditional.

Finally, the review maps published evidence and authors' stated limitations; it does not independently reproduce

every computational benchmark or prospectively test the proposed evidence organization.

CONCLUSION

Machine learning has transformed tandem mass spectrometry into a multi-stage environment for natural-product hypothesis generation, connecting spectra to analogues, molecular families, formulas, structural classes, ranked candidates, generated molecules, and biological context. These outputs are complementary rather than interchangeable.

The central distinction emerging from this mapping review is between computational reach and structural confirmation. Expanding searchable or generatable chemical space can improve discovery potential without necessarily increasing certainty about molecular identity. Conversely, narrowly constrained workflows may rank candidates accurately while overlooking unexpected chemistry.

Reliable natural-product annotation therefore depends on preserving evidence levels, exposing uncertainty, and validating computational outputs under chemically disjoint, cross-instrument, and prospective conditions. Machine learning can substantially reorganize which experimental questions are asked next, but it should not erase the boundary between computationally plausible structure and experimentally supported structural identity.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] 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
- [2] Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrland 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
- [3] 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
- [4] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann Intern Med.* 2018;169(7):467-73. doi:10.7326/M18-0850
- [5] 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
- [6] 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
- [7] 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). doi:10.1371/journal.pcbi.1008724
- [8] 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
- [9] 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
- [10] 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
- [11] Goldman S, Wohlwend J, Stražar M, Haroush G, Xavier RJ, Coley CW. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat Mach Intell.* 2023;5(9):965-79. doi:10.1038/s42256-023-00708-3
- [12] de Jonge NF, Louwen JJR, Chekmeneva E, Camuzeaux S, Vermeir FJ, Jansen RS, et al. MS2Query: Reliable and scalable MS2 mass spectra-based analogue search. *Nat Commun.* 2023;14(1):1752. doi:10.1038/s41467-023-37446-4
- [13] Young A, Röst H, Wang B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat Mach Intell.* 2024;6(4):404-16. doi:10.1038/s42256-024-00816-8
- [14] Zhang H, Yang Q, Xie T, Wang Y, Zhang Z, Lu H. MSBERT: Embedding tandem mass spectra into

- chemically rational space by mask learning and contrastive learning. *Anal Chem.* 2024;96(42):16599-608. doi:10.1021/acs.analchem.4c02426
- [15] Bui-Thi D, Liu Y, Lippens JL, Laukens K, De Vijlder T. TransExION: A transformer based explainable similarity metric for comparing IONS in tandem mass spectrometry. *J Cheminform.* 2024;16(1):61. doi:10.1186/s13321-024-00858-5
- [16] Cooper BT, Yan X, Simón-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] de Jonge NF, Hecht H, Strobel M, Wang M, van der Hooft JJJ, Huber F. Reproducible MS/MS library cleaning pipeline in matchms. *J Cheminform.* 2024;16(1):88. doi:10.1186/s13321-024-00878-1
- [18] 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). doi:10.1371/journal.pcbi.1006089
- [19] 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
- [20] 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
- [21] Xing S, Shen S, Xu B, Li X, Huan T. BUDDY: Molecular formula discovery via bottom-up MS/MS interrogation. *Nat Methods.* 2023;20(6):881-90. doi:10.1038/s41592-023-01850-x
- [22] Goldman S, Xin J, Provenzano J, Coley CW. MIST-CF: Chemical formula inference from tandem mass spectra. *J Chem Inf Model.* 2024;64(7):2421-31. doi:10.1021/acs.jcim.3c01082
- [23] Goldman S, Li J, Coley CW. Generating molecular fragmentation graphs with autoregressive neural networks. *Anal Chem.* 2024;96(8):3419-28. doi:10.1021/acs.analchem.3c04654
- [24] Kalia A, Chen YZ, Krishnan D, Hassoun S. JESTR: Joint embedding space technique for ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics.* 2025;41(7). doi:10.1093/bioinformatics/btaf354
- [25] Nowatzky Y, Russo FF, Lisec J, Kister A, Reinert K, Muth T, et al. FIORA: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nat Commun.* 2025;16(1):2298. doi:10.1038/s41467-025-57422-4
- [26] Martin MR, Hassoun S. Optimized de novo molecular generation (OMG) for mass spectra annotation using transfer and reinforcement learning. *Anal Chem.* 2025;97(38):20734-42. doi:10.1021/acs.analchem.5c01770
- [27] Zuffa S, Schmid R, Bauermeister A, Gomes PWP, Caraballo-Rodríguez 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
- [28] Martin MR, Bittremieux W, Hassoun S. Molecular structure discovery for untargeted metabolomics using biotransformation rules and global molecular networking. *Anal Chem.* 2025;97(6):3213-9. doi:10.1021/acs.analchem.4c01565
- [29] Zhang H, Zeng X, Yin Y, Zhu ZJ. Knowledge and data-driven two-layer networking for accurate metabolite annotation in untargeted metabolomics. *Nat Commun.* 2025;16(1):8118. doi:10.1038/s41467-025-63536-6
- [30] Brungs C, Schmid R, Heuckeroth S, Mazumdar A, Drexler M, Šácha P, et al. MSnLib: Efficient generation of open multi-stage fragmentation mass spectral libraries. *Nat Methods.* 2025;22(10):2028-31. doi:10.1038/s41592-025-02813-0