Bootstrap resampling of mass spectral pairs with SpecReBoot reveals hidden molecular relationships

Esteban Charria-Girón[a,b*†], Laura Rosina Torres-Ortega[a†], Joelle Mergola Greef[a,c], Yasmina Marin Felix[b,d], Nelson H. Caicedo[e,f], Frank Surup [b,d], Marnix H. Medema[a], Justin J.J. van der Hooft[a,g*]

[a]Bioformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands
[b]Department of Microbial Drugs, Helmholtz Centre for Infection Research GmbH (HZI) and German Centre for Infection Research (DZIF), Inhoffenstraße 7, 38124 Braunschweig, Germany
[c]Institute of Biology, Leiden University, Sylviusweg 72, Leiden 2333 BE, The Netherlands
[d]Institute of Microbiology, Technische Universität Braunschweig, Spielmannstraße 7, 38106 Braunschweig, Germany
[e]Departamento de Ciencias Biológicas, Bioprocesos y Biotecnología, Facultad de Ingeniería, Diseño y Ciencias Aplicadas, Universidad Icesi, Calle 18 No. 122-135 Pance, Cali, Colombia
[f]Centro BioInc, Universidad Icesi, Cali, Colombia
[g]Department of Biochemistry, University of Johannesburg, 2006 Johannesburg, South Africa

[†]These authors contributed equally

*Correspondence: esteban.charriagiron@wur.nl; justin.vanderhooft@wur.nl

Abstract

Mass spectral molecular networking (MN) has emerged as a key computational approach to organize and analyze the vast volumes of tandem mass spectrometry (MS/MS) data generated in natural product research. MN connections are based on mass spectral similarities derived from cosine-based scores or machine learning-based scores such as Spec2Vec and MS2DeepScore. These similarity scores are single, deterministic values and provide no estimate of the statistical robustness of the inferred mass spectral connections. As a result, molecular networks frequently contain edges arising from noise, missing fragments, or experimental variability, while, simultaneously, authentic chemical relationships remain hidden. To remedy this situation, here, we introduce SpecReBoot, a statistical framework that adapts Felsenstein's bootstrap principle from phylogenetics to metabolomics. Within this framework, mass fragmentation peaks are treated as resampling units with replacement to generate pseudo-replicate spectra. Spectral similarities are recalculated across replicates, and the robustness of each edge between a pair of spectra is quantified by how frequently they appear as mutual top-*k* neighbors across bootstrap replicates. This approach generates bootstrap-derived confidence scores for every spectral connection, transforming mass spectral similarity from an absolute score into a distribution-based, confidence-aware measure. We show how, across public GNPS spectral library and natural products discovery case study datasets on bioactive metabolites produced by bacteria and fungi, SpecReBoot reliably identifies high-confidence spectral connections, filters unstable or noise-driven edges, and rescues chemically meaningful relationships that conventional metrics systematically miss. Applying SpecReBoot to study the polyketide-lactones produced by the endophytic fungus *Diaporthe caliensis* revealed previously hidden spectral relationships leading to the discovery of the novel caliensomycin macrolactone scaffold, biosynthetically and biochemically related to the bioactive phomol known polyketide present in *D. caliensis*. In conclusion, this study provides the first statistical framework for quantifying uncertainty in MS/MS mass spectral similarity. Due to its context-agnostic nature, we anticipate that our computational metabolomics framework can also be adopted across other disciplines like clinical and environmental metabolomics. Altogether, SpecReBoot introduces statistical rigor, improves reproducibility, thus enhancing molecular networking-based natural product discovery.

Keywords: Natural products discovery, spectral similarity, computational metabolomics, molecular networking, RiPPs, endophytes, fungi, bootstrapping, statistics.

1. Main

Small molecules of natural origin, commonly referred to as natural products (NP), play important roles in diverse ecosystems, where they mediate intra- and interspecies interactions through their role in chemical communication, competition, and defense (Demian and Fang 2000, Keller et al 2005). Beyond their ecological relevance, NPs have long been exploited by society as a rich source of therapeutics and other beneficial bioactive molecules (Clardy et al. 2009, Atanasov et al. 2021). Tandem mass spectrometry (MS/MS), often coupled to liquid chromatography (LC), has become the primary analytical technique to explore the diversity of these molecules (Alseekh and Fernie 2018, de Jonge et al. 2022), owing to its sensitivity, high throughput, and ability to capture structural information through distinct fragmentation patterns.

The widespread adoption of molecular networking (MN) within the GNPS infrastructure has further advanced the analysis of MS/MS spectral data by enabling large-scale organization, visualization, and annotation of spectral relationships across complex biological datasets (Wang et al. 2016, Nothias et al. 2020). In these networks, MS/MS spectra are represented as nodes connected by similarity-based edges, allowing researchers to efficiently explore the chemical space and prioritize unknown metabolite families for downstream investigation (Aron et al. 2020, Beniddir et al. 2021). A variety of mass spectral similarity scores have been introduced and can broadly be classified in two categories: cosine-based scores, which compare aligned fragment ion intensities, and machine learning–based approaches that embed spectra into a latent space to capture higher-order fragment context such as Spec2Vec, MS2DeepScore, and most recently dreaMS (van der Hooft et al. 2016, Huber et al. 2021, Bittremieux et al. 2022, de Jonge et al. 2024, Bushuiev et al. 2025). In both cases, users must define a similarity threshold to prune weak spectral relationships and enable network construction, a choice that is often arbitrary and strongly influences the resulting topology and downstream interpretation.

Despite the broad acceptance of MN as a cornerstone method in natural products research, the underlying mass spectral similarity scores on which these networks are built are treated as absolute. These scores remain deterministic and lack any explicit evaluation of the reproducibility or stability of inferred spectral connections under perturbations of the underlying data. Regardless of the choice of similarity measure, users must rely on fixed similarity thresholds, often without a clear understanding of how individual connections are built. Consequently, molecular networks often require careful manual inspection to distinguish true biochemical relationships between connected spectral nodes from spurious or noise-driven connections (van der Hooft et al. 2016b, Beniddir et al. 2021, Sachsenberg et al. 2025).

Indeed, alongside the several success stories following the introduction of GNPS, multiple studies have highlighted fundamental challenges inherent to the study of MS/MS data, including spectral noise, incomplete or missing fragment peaks, instrument- and acquisition-dependent variability, and, more recently, the impact of unintended or in-source

fragmentation (Stein 2012, Clark et al 2021, El Abiead et al. 2025, Sachsenberg et al. 2025). Remarkably, no general statistical framework currently exists to quantify the reproducibility or uncertainty of spectral connections within molecular networks. This lack of confidence-aware metrics limits the reliability of MN-derived hypotheses and highlights a clear methodological gap in metabolomics analyses.

Other disciplines have long recognized the value of quantifying confidence in inferred relationships. In phylogenetics, for example, the Felsenstein's bootstrap algorithm introduced a principled framework for assessing the robustness of evolutionary trees by resampling alignment columns and evaluating the consistency of inferred relationships across pseudo-replicates (Felsenstein 1985, Efron and Tibshirani 1994). This concept has since become a standard component of evolutionary analyses, enabling the distinction of well-supported relationships from those driven by stochastic variation (Efron et al. 1996).

To address the lack of uncertainty estimation in molecular networking (Fig. 1a), we developed SpecReBoot, a statistical framework that introduces bootstrap-based confidence measures into spectral similarity analysis. SpecReBoot draws inspiration from phylogenetic bootstrapping by treating fragment peaks as resampling units (Fig. 1b), generating pseudo-replicate spectra from which spectral similarities are recalculated. This procedure yields two complementary quantities for each spectral pair, a mean similarity, and an edge support value that reflects the stability of their mutual nearest-neighbor relationship across bootstrap replicates (Fig. 1c).

In practice, these measures allow molecular networks to be interpreted not solely based on similarity scores, but also on the robustness of spectral connections. Together, these quantities enable the construction of confidence-aware MNs, either by retaining only edges with both high similarity and high support (core edges), or by rescuing edges with low mass spectral similarity but consistently high support across bootstrap replicates (Fig. 1d). As demonstrated in the following sections, SpecReBoot improves the recovery of chemically meaningful spectral connections across different similarity scores, including cases where just using similarity scores fail. By linking bootstrap-derived edge support to fragment-level perturbations, our computational metabolomics framework enhances network interpretability and provides an explainable layer for machine learning-based similarity scores, ultimately bringing forward metabolomics-based discovery workflows.

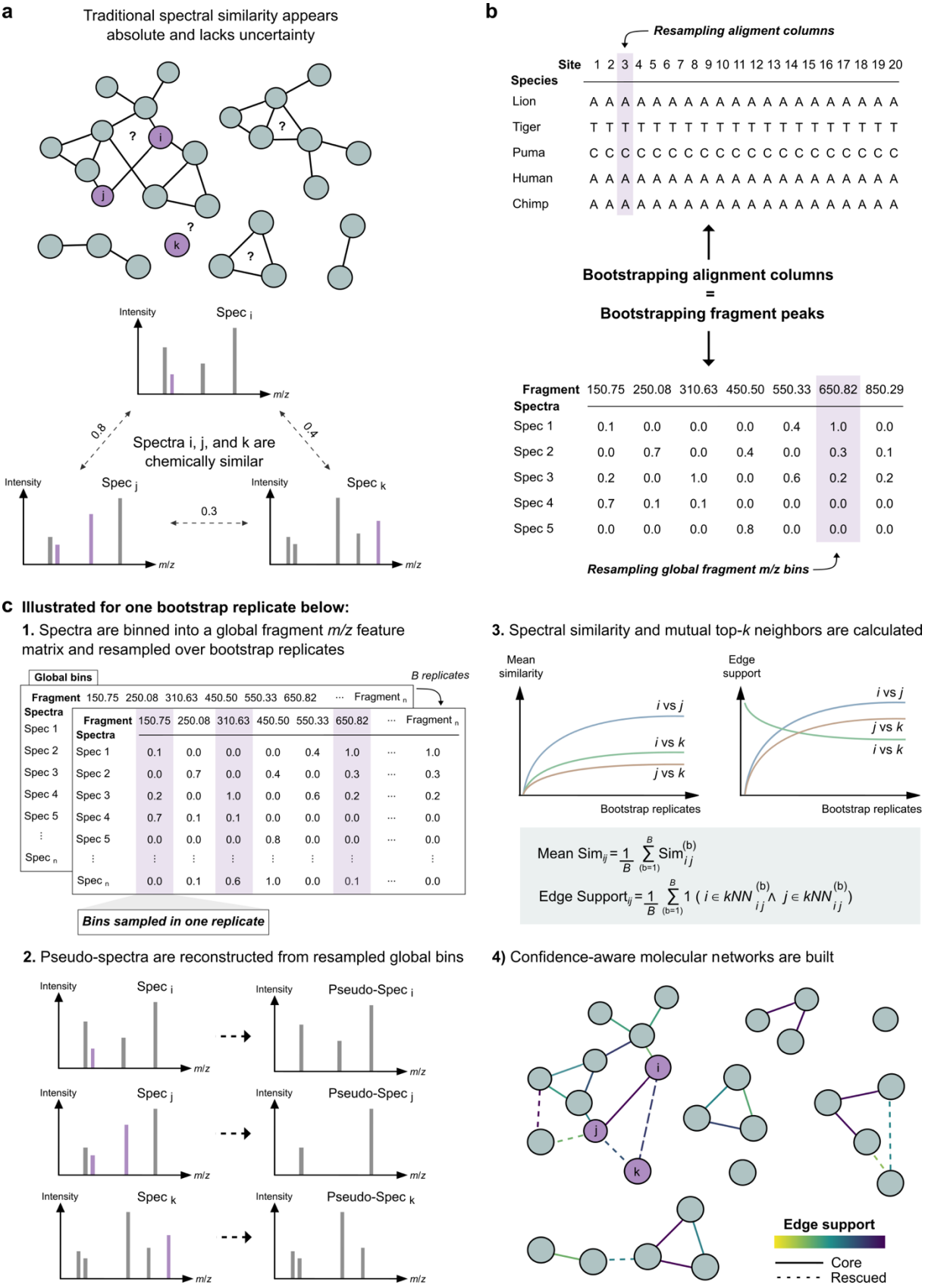

Fig. 1. Bootstrapping fragment peaks introduces confidence-aware spectral similarity in metabolomics. a, In conventional molecular networking, MS/MS spectra are represented as nodes connected by edges constructed from pairwise similarity scores. These values are typically treated as absolute, even though they lack any estimate of stability under perturbations of the underlying spectra. **b**, SpecReBoot adapts nonparametric bootstrapping to MS/MS data. In analogy to phylogenetics, where alignment columns are resampled across replicates, fragment

peaks are first binned into a global fragment *m/z* features matrix that includes all observed fragments across spectra. In each bootstrap replicate, fragment *m/z* bins are resampled with replacement (on average ~37% of global bins are sampled per replicate), and pseudo-replicate spectra are reconstructed from resampled bins, altering which peaks contribute to similarity calculations. **c**, The workflow is illustrated for a single bootstrap replicate. For each replicate, spectral similarities are recalculated and mutual top-*k* nearest neighbors are identified. Repeating this process over *B* replicates produces two complementary measures: (i) a mean similarity, defined as the average spectral similarity across replicates, and (ii) an edge support, defined as the frequency with which two spectra remain mutual nearest neighbors. These confidence-aware measures are then used to "reboot" molecular networks, highlighting stable connections (core), filtering unstable edges, and rescuing previously missed relationships (rescued), enabling more reliable and interpretable MS/MS-based metabolomics analyses.

2. Results

2.1. Mean similarity and edge support reveal stable and chemically meaningful spectral relationships across metrics

To introduce the SpecReBoot framework and illustrate the concept of bootstrap-derived confidence estimation applied to mass spectral links in practice, we first applied the method to a small, curated subset of MS/MS spectra derived from ribosomally synthesized and post-translationally modified peptides (RiPPs). Compounds belonging to this biosynthetic class are particularly intriguing because, despite the first RiPP pathways being discovered relatively recently (Arnison et al. 2013, Zdouc et al. 2023), they are now known to be far more widespread than initially anticipated (Skinnider et al. 2016, Nie et al. 2025). Even with increased research into RiPPs for different applications (Arnison et al. 2013, Montalbán-López et al. 2021), the availability of well-curated MS/MS reference spectra for this compound class remains limited. RiPPs exhibit extensive structural and chemical diversity, as they may contain any canonical amino acid sequence, vary in length, and undergo a wide variety of post-translational modifications (Arnison et al. 2013). The ongoing discovery of new RiPP families further highlights their molecular diversity (Montalbán-López et al. 2021), while the presence of extensive and heterogeneous modifications often results in complex and unpredictable fragmentation patterns. These properties make RiPPs a suitable test case for evaluating the robustness of inferred spectral relationships.

Tandem mass spectrometry data for compounds belonging to the RiPPs compound class were retrieved from public spectral repositories in the GNPS platform (Wang et al. 2016) through targeted manual curation of spectra annotated by NPClassifier as a peptide. Retrieved spectra were subsequently cross validated against the original literature reports to ensure data integrity, spectral quality, and consistency with a RiPP biosynthetic origin. This procedure resulted in a curated set of 24 RiPP-associated MS/MS spectra. From this collection, we selected the aerucyclamides as a representative case study to illustrate the ability of SpecReBoot to recover chemically meaningful relationships that are missed by conventional molecular networking approaches. Aerucyclamides are a family of RiPPs first reported from the cyanobacterium *Microcystis aeruginosa* as hexacyclopeptides (Portmann et al. 2008ab). Their biosynthetic origin was later reassigned to a ribosomal pathway by comparative analyses to the patellamide pathway characterized in cyanobacterial symbionts of ascidians (Ziemert et al. 2008). MS/MS spectral data for aerucyclamides A–C were extracted from the curated RiPP dataset and further validated against the originally published spectra (Portmann et al. 2008ab) to ensure spectral concordance. Despite their high structural similarity, these molecules exhibit extremely low spectral similarity when calculated using the modified cosine score. Specifically, similarity values of 0.0004 (A–B), 0.000 (A–C), and 0.021 (B–C) are obtained, causing these relationships to be missed by conventional MN approaches relying solely of spectral similarity thresholds.

Using SpecReBoot, all curated RiPP-associated spectra were first represented as global *m/z* bins, after which fragment bins were resampled with replacement across bootstrap replicates to generate pseudo-spectra. The general procedure is illustrated for the aerucyclamides in Fig.

2a. For each bootstrap replicate, pairwise spectral similarities were recalculated, and mutual top-*k* nearest-neighbor relationships were identified. Each time a pair of spectra *i* and *j* appeared as mutual neighbors, the corresponding edge support value was incremented. Repeating this process across *B* bootstrap replicates results in two complementary quantities for each spectral pair, (i) a mean similarity, defined as the average similarity across replicates, and (ii) an edge support value reflecting the frequency with which the spectral relationship is recovered.

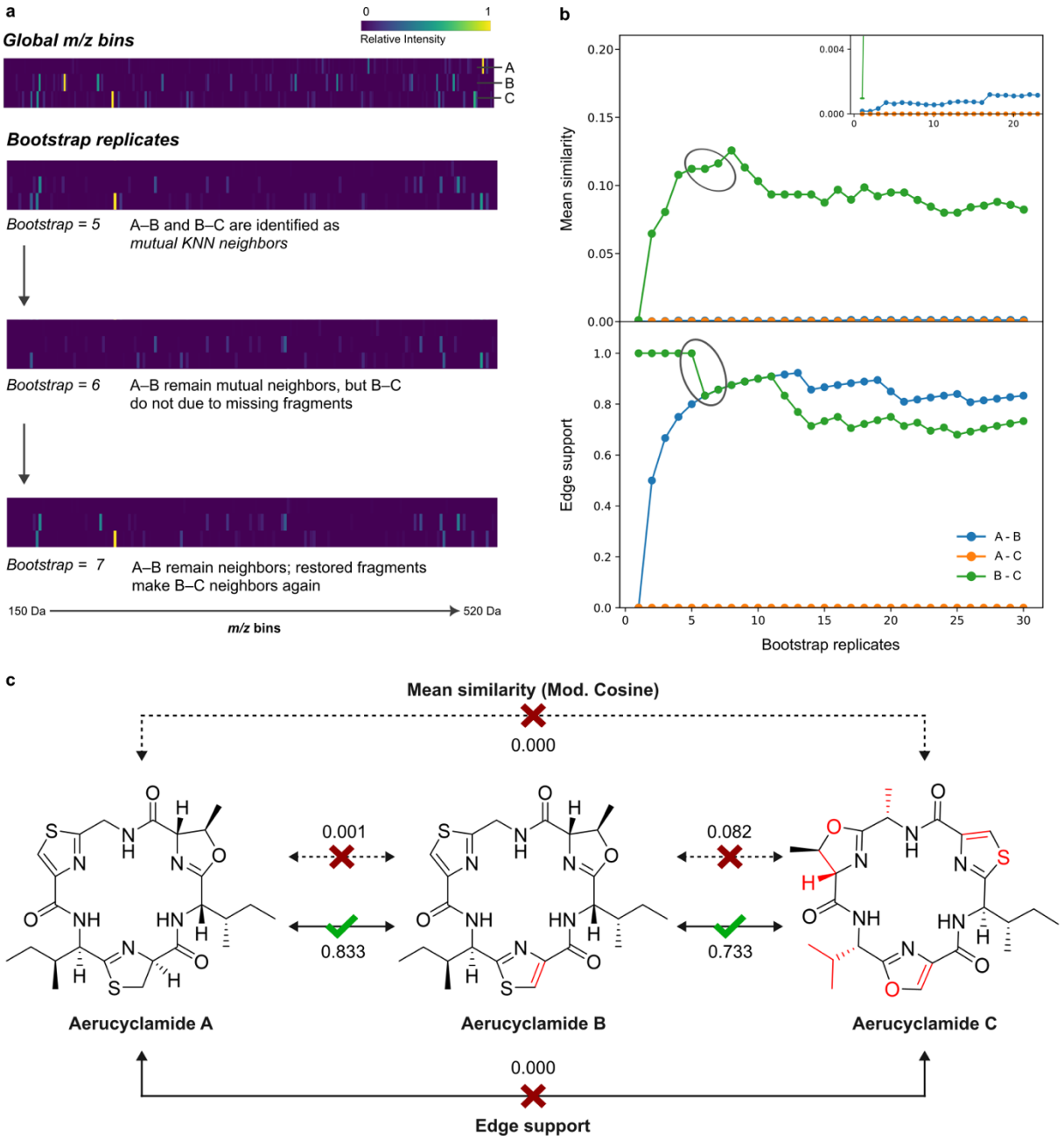

Fig. 2. Fragment resampling uncovers confidence-aware spectral relationships in aerucyclamides. a, Schematic illustration of bootstrap resampling for aerucyclamides A–C. MS/MS spectra are first represented as global *m/z* bins, after which bins are resampled with replacement to generate pseudo-replicate spectra. Example bootstrap replicates (5–7) show how changes in fragment composition affect *k*-nearest-neighbor (KNN) relationships between spectra. Across replicates, aerucyclamides A and B are consistently identified as mutual neighbors, whereas the relationship between B and C is variably present depending on the resampled fragment set. **b,**

Trajectories of mean spectral similarity (top) and edge support (bottom) across increasing numbers of bootstrap replicates for all pairwise comparisons among aerucyclamides A–C. Values are shown as a function of the number of bootstrap replicates. The inset on the top right corner highlights the low-amplitude variation in mean similarity for the A–B pair at an expanded scale. **c,** Comparison of mean similarity (dashed lines) and edge support (solid lines) values for aerucyclamides A–C using the modified cosine. While the mean similarity results in a near-zero (≤ 0.01) score for these structurally related compounds, SpecReBoot recovers stable relationships between A–B and B–C based on high edge support values. Structural differences between aerucyclamides are highlighted in red.

For aerucyclamides A–C, these quantities revealed distinct stability patterns across spectral pairs (Fig. 2b). The A–B pair, which exhibits near-zero similarity in individual conventional comparisons, showed a slight increase in the mean similarity across bootstrap replicates (see inset in Fig. 2b), indicating that similarity emerges only under specific fragment subsets and is otherwise masked by spectral variability. In contrast, the A–C pair consistently showed both mean similarity and edge support equal to zero across all replicates. Analysis of the edge support further demonstrated that the A–B and B–C spectral pairs were repeatedly recovered across as mutual nearest neighbors, whereas the A–C relationship remained unsupported. Importantly, these patterns reflect the stability of spectral relationships under fragment-level perturbation rather than the magnitude of chemical similarity observed in any single comparison. As the number of bootstrap replicates increased (B = 30), both mean similarity and edge support converge, providing stable and interpretable measures of the robustness of spectral connections.

Furthermore, because SpecReBoot explicitly records the *m/z* bins sampled across bootstrap replicates, it enables direct correlation of changes in both mean similarity and edge support trajectories to specific fragment ions. Informative fragments can therefore be mapped back onto reference MS/MS spectra, providing a mechanistic explanation of why particular relationships are stabilized or lost under resampling. As illustrated for the aerucyclamides (Fig. S1–S3), *m/z* bins that are repeatedly associated with gains or losses of mutual kNN relationships correspond to structurally meaningful fragment ions arising from conserved backbone substructures.

For instance, fragment ions at *m/z* 425.2 and 436.2 Da showed strong positive correlations with the high edge support observed between aerucyclamides A and B and can be traced back to differences in the amino acid composition along their peptide backbones. Manual strucural assignment reveals that the fragment at *m/z* 425.2 corresponds to a motif uniquely present in aerucyclamide A, comprising thiazoline, *L*-Ile, thiazole, and *D*-allo-Ile residues arranged along the peptide backbone. In contrast, the fragment at *m/z* 436.2, observed exclusively in aerucyclamide B, arises from a closely related substructure containing *L*-Ile, thiazoline, *D*-allo-Ile, and an oxazoline ring (Fig. S3).

Interestingly, MS2DeepScore captures spectral connections between aerucyclamides with mean similarity values of 0.676 (A–B), 0.769 (A–C), and 0.571 (B–C), which will be as well missed when applying the commonly recommended similarity threshold of 0.8 for this metric. Moreover, MS2DeepScore does not rely on the same set of *m/z* bins as the modified cosine

score to recover these relationships (Fig. S2b). Nevertheless, partial overlap in informative fragment evidence is observed. For example, in the spectral comparison between aerucyclamide A and B, both MS2DeepScore and modified cosine identify the fragment at *m/z* 332.1 Da as positively correlated with the high edge support between the corresponding spectra. Beyond this shared fragment, however, the two similarity measures diverge in the fragment-level evidence underlying their spectral relationships. Specifically, for this spectral pair, fragment ions positively correlated with high edge support above *m/z* 475 Da exhibited predominantly even nominal masses when using modified cosine, whereas for MS2DeepScore the highest-mass positively correlated fragments corresponded to odd nominal masses. This contrast suggests that chemically distinct fragment subsets contribute to the similarity captured by each metric, highlighting that embedding-based and alignment-based approaches can converge on the same stable spectral relationships while relying on different underlying fragment evidence.

Together, these results demonstrate how SpecReBoot enhances the explainability of spectral similarity by directly linking fragment-level chemical features to edge support, thereby rationalizing the recovery of reproducible spectral relationships despite near-zero conventional spectral similarity. In this context, edge support is particularly informative, as it enables the reconnection of aerucyclamide spectra into a coherent molecular family within a larger network comprising RiPP-associated spectra obtained from our manual curation effort (Fig. S4). By creating repository-wide molecular networks with SpecReBoot edge support, our confidence-aware frameworks provide an alternative strategy for the targeted mining of specific RiPP families in public, untargeted mass spectrometry datasets, ultimately facilitating a more comprehensive assessment of their chemical diversity and distribution across microbial sources.

2.2. SpecReBoot provides scalable and interpretable spectral similarity across large-scale MS/MS mass spectral library datasets

We next evaluated the scalability and insights provided by SpecReBoot across publicly available MS/MS reference datasets that span different sizes and various degrees of chemical diversity they contain. Specifically, we analyzed three representative spectral collections, a pesticide dataset previously introduced as part of the *matchms* framework (Huber et al., 2020), the NIH Natural Products Library, comprising approximately 1,200 curated spectra, and the MSn-COCONUT subset from MSn-Lib, containing approximately 13,000 spectra, after filtering (see Methods). As a structural reference, we calculated pairwise Tanimoto similarities based on Morgan fingerprints (4,096 bits) derived from annotated SMILES for each dataset. SpecReBoot was applied using widely adopted spectral similarity metrics with a fast implementation from the *matchms* framework, including Cosine, Modified Cosine, Spec2Vec, and MS2DeepScore (de Jonge et al. 2024). This allowed us to systematically evaluate how bootstrap-derived confidence estimates behave across dataset scales and similarity scores.

Despite fundamental differences in how these methods compute mass spectral similarity, ranging from direct peak alignment to learned mass spectral embeddings, edge support values

exhibited consistent stability patterns across evaluated metrics. As illustrated for the pesticide dataset (Fig. 3a), selected here because its relatively small size enables direct visualization of full pairwise similarity matrices. We observed that mean spectral similarity matrices differed substantially between methods, whereas the edge support structure was markedly more conserved across metrics and, to a certain extent, reflected the underlying structure of pairwise chemical similarity (Fig. 3a–c). Importantly, however, the edge support should not be interpreted as a proxy for the magnitude of chemical similarity itself. While edge support visually aligns with our "ground-truth" chemical relationships, it performs poorly as a standalone metric, as demonstrated by receiver operating characteristic and precision–recall analyses for all datasets (Fig. S5–S10). Instead, these initial results motivated us to assess the use of edge support as a complementary filter applied on top of conventional similarity scores, with the aim of refining chemically meaningful spectral relationships.

This complementary behavior was subsequently evaluated in detail using the NIH Natural Products Library (Fig. 3b). At this scale, the modified cosine score is highly stringent, as most pairwise comparisons result in low similarity values, and applying similarity threshold of 0.7 retains only 2,626 edges. Nevertheless, many edges still correspond to pairs with low actual chemical similarity. When the edge support (≥ 0.5) is incorporated as a secondary filter together with the same spectral similarity threshold, the number of retained edges is reduced to 1,597. We define this subset as core edges, which exhibit a clear enrichment toward the region of high chemical similarity (Fig. 3b). A similar pattern is observed for MS2DeepScore. However, due to the embedding-based nature of this method, similarity values are shifted toward higher ranges (Fig. S13), resulting in 8,378 retained edges when applying the same threshold of 0.7, including many chemically unrelated pairs (Fig. S15). Consequently, a more stringent similarity threshold of 0.8 was applied for MS2DeepScore, which still retained 2,629 edges that also included a fraction of pairs with low chemical similarity. Applying edge support as an additional constraint (≥ 0.5) reduces this set to 1,589 core edges, preferentially enriched in connections within the high chemical and high spectral similarity regime (Fig. 3b). A similar scenario is observed in the case of Spec2Vec, however, in contrast to MS2DeepScore, fewer unsupported edges are retained when using similarity thresholds alone, consistent with the more conservative behavior of this machine learning model generally observed (Fig. S16). Yet, applying the SpecReBoot dual-filter strategy further refines the set of retained edges obtained for this similarity metric.

In general, edges with low edge support are widespread across the full range of chemical similarity values but are particularly enriched among pairs with lower structural similarity (Tanimoto < 0.7) across all evaluated metrics. This chemical enrichment can be quantified by comparing the proportion of edges connecting chemically similar pairs before and after applying the SpecReBoot dual-filter strategy. Specifically, the fraction of retained edges above the Tanimoto threshold increased by 1.32-fold, 1.33-fold, and 1.31-fold for modified cosine, MS2DeepScore, and Spec2Vec, respectively, relative to using spectral similarity thresholds alone for the NIH Natural Products Library dataset. In contrast, evaluating chemical

enrichment at a lower Tanimoto threshold of 0.5 did not yield substantial gains, as similarity metrics already perform well under this more permissive threshold.

To evaluate the performance of SpecReBoot in large-scale metabolomic analyses, we applied it to the MSn-COCONUT dataset. Using a flash cosine similarity threshold of 0.7 alone, 76,076 edges were retained, spanning a broad range of chemical similarities, including a substantially larger fraction of chemically unrelated pairs compared to the previously analyzed datasets. Incorporating an additional edge support threshold (≥ 0.5) reduced the number of retained edges to 34,874 and selectively enriched for chemically meaningful connections, indicating improved chemical specificity rather than indiscriminate network sparsification at this scale (Fig. S17). This observation is particularly relevant for large and chemically diverse datasets, where increased scaffold diversity and recurrent fragment motifs across unrelated molecules can inflate similarity-based relationships. For this dataset, the fraction of retained edges above the Tanimoto similarity threshold increased by 2.00-fold, significantly exceeding the enrichment observed for the NIH Natural Products Library. In this context, SpecReBoot effectively compensates for the challenges posed by large chemical diversity by prioritizing reproducible spectral relationships over deterministic similarity scores.

To gain further insights into how the edge support affects agreement across different spectral similarity metrics, we compared retained edges obtained using modified cosine, MS2DeepScore, and Spec2Vec for the NIH Natural Products Library (Table S1). When spectral similarity thresholds were applied alone, all three metrics retained a large number of edges (~2,360 per metric), but only moderate pairwise agreement was observed (Jaccard index = 0.52–0.61). Although a substantial number of edges were shared across all of them (n = 1,444), only 515 of these corresponded to chemically similar pairs (Tanimoto similarity ≥ 0.7), indicating that agreement between metrics does not necessarily imply chemical coherence. Incorporating edge support (≥ 0.5) as a secondary filter reduced the total number of shared edges (n = 783), but substantially increased their chemical consistency, with approximately 50% of the retained shared connections now linking chemically similar molecules. Notably, this enrichment occurred without an increase in the overall proportion of shared edges across metrics, suggesting that edge support does not artificially enforce the agreement between similarity measures. Instead, it refines the metric-agnostic core connections by preferentially removing unstable connections and enriching reproducible, chemically meaningful spectral relationships.

Building on this observation, we next asked whether this refined, metric-agnostic core of connections translates into more consistent and chemically coherent molecular network topologies. To this end, we generated molecular networks for this dataset using our dual-filter strategy, in which edges are built by retaining only those spectral connections that satisfy both a spectral similarity threshold (0.7 for cosine-based metrics and Spec2Vec, and 0.8 for MS2DeepScore) and a minimum edge support of 0.5, and compared them to networks constructed using solely the spectral similarity threshold (Fig. S18–S25). In MNs constructed using spectral similarity thresholds alone, together with standard topology-based parameters, the largest molecular families are dominated by low-support edges, whereas smaller

components tend to exhibit higher internal stability. Applying an edge support threshold of 0.5 in addition to spectral similarity effectively “reboots” the network topology by removing unstable connections, resulting in more balanced molecular networks composed of chemically coherent families. Importantly, the enrichment of chemically similar pairs is observed accompanied by a substantial reduction in the total number of edges, indicating improved chemical specificity rather than generally increased network density.

In addition, all “rebooted“ networks showed similar degree distributions, characterized by a predominance of low-degree nodes and a limited number of moderately connected molecular families (Fig. S26). Despite clear metric-specific differences, most notably for Spec2Vec, which produced a denser and more highly clustered MN, global topological statistics revealed comparable overall organization across all three metrics (Table S2). Average degree (3.3–5.0), network centralization (0.13–0.28), and the number of connected molecular families (140–157) were of similar magnitude across the obtained networks. Importantly, none of the MNs exhibited highly centralized or “hairball“-like structures, and the proportion of singletons remained in average around approximately 40% across all networks. Together, these results demonstrate that incorporating edge support stabilizes global network organization across similarity metrics, supporting our observation that SpecReBoot reveals a robust, chemically meaningful backbone of spectral relationships.

In Fig. 3c, an example is shown for a molecular family identified within the NIH Natural Products Library, where spectra *i* and *j* correspond to natural products of distinct biosynthetic origin, a linear tetrapeptide and a hybrid phenylpropanoid, respectively. Despite sharing only a substructure corresponding to *O*-methyl-tyrosine, not reflecting their main chemical scaffolds, these compounds and their derivatives are grouped into the same molecular family based on a single edge exceeding the spectral similarity threshold. In contrast, applying SpecReBoot resolves these relationships into two distinct molecular families, each exhibiting coherent internal chemistry (Fig. S27–S29). This refinement is quantitatively supported by an increase in the mean intra-family chemical similarity of the resulting molecular families compared to the original molecular family before removing unstable connections (Fig. 3c).

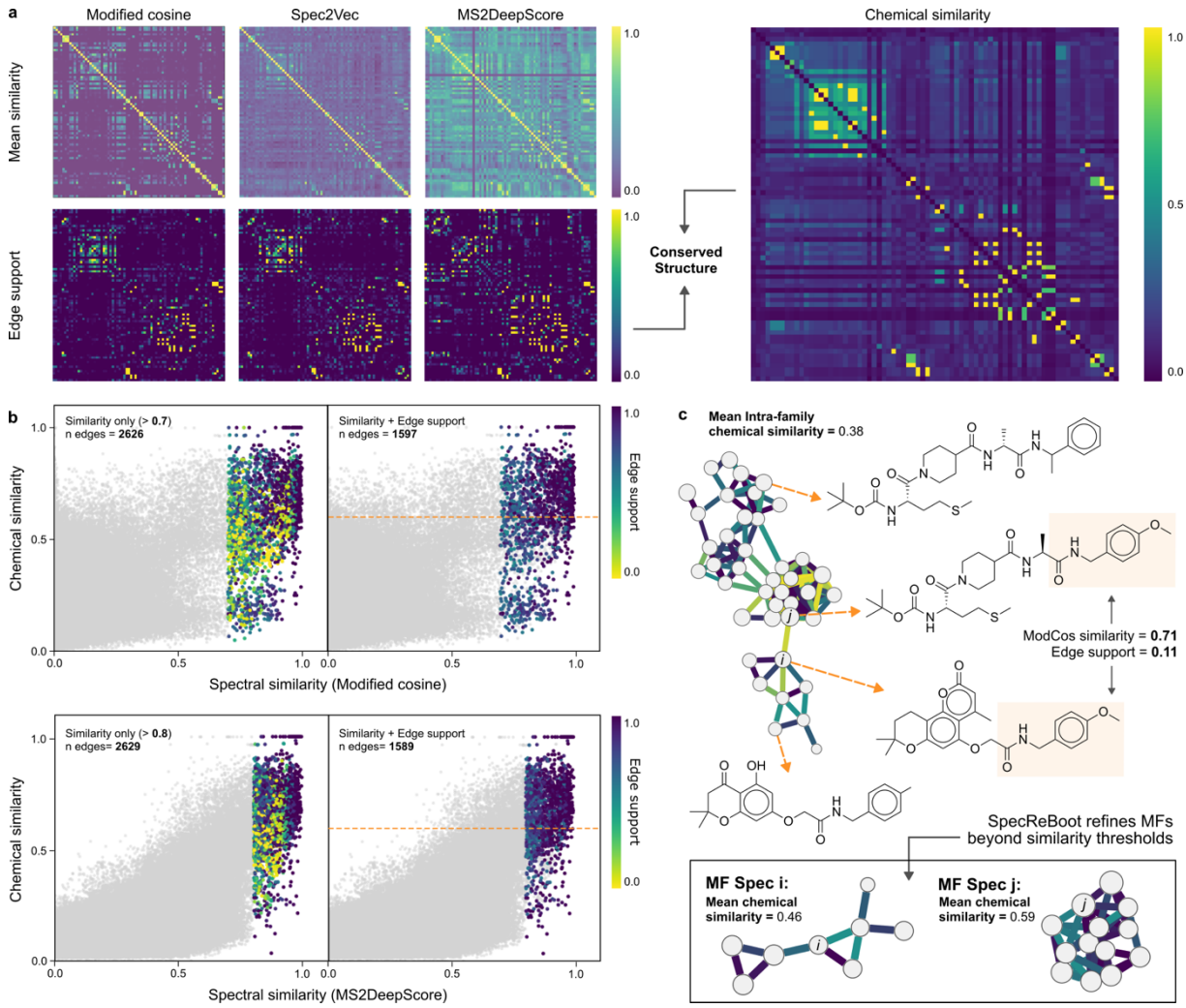

Fig. 3. Edge support captures metric-independent spectral relationships and enriches chemically meaningful connections. a, Pairwise similarity matrices for the pesticide dataset computed using modified cosine, Spec2Vec, and MS2DeepScore (top row), displayed using an identical, fixed scan-ID ordering and without method-specific clustering. The corresponding edge support matrices (bottom row) are shown under the same ordering and compared to the pairwise chemical similarity matrix (right). Despite substantial differences in raw similarity structure across metrics, edge support matrices are highly conserved and show a strong correspondence with the underlying chemical similarity matrix, highlighting stable spectral relationships that are largely independent of the chosen similarity metric. **b,** Chemical versus spectral similarity for edges retained using similarity thresholds alone (left panels) or in combination with edge support (right panels). The horizontal dashed line indicates a chemical similarity threshold above 0.7. Applying edge support as a secondary filter substantially reduces the number of retained edges (n indicated in each panel) and enriches connections in regions of higher chemical similarity across both similarity metrics. Color indicates edge support values. **c,** Representative molecular family showing the impact of edge support on network refinement. Even though the edge between spectra *i* and *j* exceeded the modified cosine similarity threshold, their low edge support reflects inconsistent recovery across bootstrap replicates. Applying edge support as an additional constraint removes unstable connections and yields molecular families that are more chemically coherent as illustrated by the increased intra-family mean chemical similarity.

2.3. Rescuing hidden polyketide-lactone diversity in *Diaporthe caliensis* reveals an unprecedented macrolactone scaffold

To further assess the applicability of SpecReBoot in untargeted metabolomics for natural products discovery, we re-examined a previously published MN dataset derived from the endophytic fungus *Diaporthe caliensis*, a prolific producer of structurally diverse polyketide-lactones isolated from the medicinal plant *Otoba gracilipes* in Colombia (Charria-Giron et al. 2021, Charria-Girón et al. 2023ab, Hoyos et al. 2024). Among the specialized metabolites produced by this fungus are the bioactive 10-membered macrolactone polyketide phomol and the polyketide butanolide derivatives caliensolide A and B. In the original study, conventional cosine-based molecular networking clustered phomol and caliensolide B within the same molecular family, while caliensolide A belonged to a distinct cluster (Charria-Girón et al. 2023). Subsequent application of MS2LDA partially recovered these relationships through shared substructure Mass2Motifs mainly comprising of concurring neutral losses (van der Hooft et al. 2016), highlighting the limitations of similarity-based networking alone and suggesting complex underlying mass fragmental relationships between these molecules.

We therefore applied SpecReBoot to re-examine the spectral relationships among polyketide-lactone derivatives and evaluate if we could reconcile the molecular relationships missed by our previous MN analysis. As a starting point, we focused on recovering the molecular relationship between caliensolide A and B, which belong to the same compound series but remained systematically disconnected under conventional MN. Bootstrap analysis (B = 100) revealed a mean spectral similarity of 0.32 between their MS/MS spectra using the modified cosine metric, accompanied by an edge support value of 0.22. Although the similarity value is below the commonly used threshold of 0.7 as expected, the non-zero edge support value indicates that this relationship is repeatedly recovered across some bootstrap replicates. This behavior reflects fragment-level information that was not captured in our former study, suggesting that fragment perturbations increase the overall spectral similarity across specific replicates.

Consequently, caliensolides A and B now appeared as neighboring nodes within the same molecular family as phomol, in agreement with earlier substructure-level insights obtained using MS2LDA (Charria-Girón et al. 2023). Motivated by this finding, we selected an edge support threshold of 0.2 as an exploratory criterion to “reboot” the molecular network topology and assess whether bootstrap-derived confidence could improve the connectivity of known polyketide-lactones produced by *D. caliensis*. Applying this threshold resulted in a substantially different network topology compared to the original MN (Fig. 4). By filtering unstable edges and rescuing low-similarity connections with and edge support over 0.2, SpecReBoot produced a coherent network of polyketide-lactones in which multiple molecular families that were previously disconnected became linked. Notably, several rescued edges connected spectra that exhibited near-zero spectral similarity but consistently high edge support, indicating robust fragment-level relationships that were systematically missed in the first place.

Based on these observations, we hypothesized that additional, chemically uncharacterized polyketide-derived metabolites associated with this molecular family may have been overlooked in the initial analysis. Guided by the SpecReBoot-derived, "rebooted" network topology, we reinvestigated fractions obtained after the scaled-up solid cultivation of *D. caliensis* and flash chromatographic pre-fractionation. Within this molecular family, comprising 179 features associated with known polyketide-lactones produced by *D. caliensis*, we identified a previously uncharacterized feature with molecular weight of 468 Da and a molecular formula of $C_{25}H_{40}O_8$. This feature originated from a fraction distinct from the one yielding phomol and caliensolides, yet all belonged to the same molecular family after the rescue of connections with high edge support. This compound differed from phomol by an additional C_3H_2O unit, suggesting a closely related carbon skeleton and supporting its annotation as a putative polyketide-lactone. Subsequently, the corresponding molecule was purified by preparative HPLC, and its structure was elucidated by HRESI-MS and comprehensive 1D- and 2D-NMR spectroscopic analyses (Table S3, Fig. S30–35).

^{1}H and HSQC NMR spectra displayed five methyl, seven methylene and eight methine groups, including four oxymethines and one olefinic methine. The ^{13}C spectrum further indicated the presence of two ketone carbonyls, one carboxyl carbon, as well as one olefinic and one dioxygenated sp^3 hybridized carbons devoid bound protons. The C–1' to C–8' side chain and the C-11 to C–17 moieties were readily identified by COSY and HMBC correlations, including their mutual connection via C–11, confirming the structural similarity of **1** to phomol. However, pronounced differences relative to phomol were observed for the remaining carbon framework. A large spin system extending from H_2–3 to H–9 was established by COSY, TOCSY and H2BC correlations. HMBC correlations from H_3–1, H_2–3 and H–4 located one carbonyl at C–2. Additional key HMBC correlations from H–11 to C–9 and C–10 from H–9 to C–8 and C–10, from 8–OH to C–8 and C–9 as well as from H–12 to C–8, enabled the assignment of the planar structure of this compound. The relative configuration of the bicyclic core was determined by ROESY correlations. Strong ROESY cross-peaks between H–4/H–6, H–5/H–9, H–6/8–OH, 8–OH/H–12, and H–9/H–11 indicated diaxial relationships among these protons, which were further supported by large $^3J_{H,H}$ coupling constants. This polyketide, representing a currently unprecedented macrolactone scaffold was named caliensomycin, honoring Cali, the city where its fungal producer was originally isolated.

Up to this point, the biosynthetic basis of phomol had remained unresolved. However, the close structural resemblance between phomol and caliensomycin, together with the concordance of their relative and absolute configurations, suggested that these compounds share a common biosynthetic origin. Based on this structural relationship, we formulated the hypothesis that phomol and caliensomycin originate from the same, or closely related, biosynthetic gene clusters (BGCs). Specifically, we proposed that their shared macrolactone scaffold is assembled by a highly reducing polyketide synthase (HR-PKS), while a second PKS contributes the aliphatic side chain esterified at C11, a feature conserved in both molecules. This hypothesis motivated a deeper investigation into the biosynthetic potential of *D. caliensis*, prompting us to sequence its genome and systematically analyze its predicted BGC repertoire.

From the 90 biosynthetic gene cluster (BGC) regions predicted by antiSMASH 8.0.4, only five corresponded to putative clusters encoding two PKSs within the same genomic region. Comparative synteny analysis of these candidate BGCs with known macrolactone-producing clusters (Fig. S36), including those responsible for aspinolide and brefeldin A biosynthesis, led to discovery of a candidate BGC with high similarity to these (Fig. 4b), which supports our initial hypothesis. In particular, the highly reducing PKS shared across all three clusters shows high sequence identity, consistent with a conserved role in macrolactone formation. In contrast, the second PKS is shared only between the aspinolide and candidate phomol clusters, in agreement with the presence of O-substitution at the macrolactone ring observed in these compounds but absent in brefeldin A (Trenti et al. 2020, Cardoza et al. 2022).

Notably, the candidate BGC in *D. caliensis* is larger than the corresponding aspinolide and brefeldin A BGCs and encodes an expanded tailoring machinery. It contains multiple cytochrome P450 monooxygenases and oxidoreductases, providing a plausible mechanistic basis for the late-stage oxidative rearrangements observed in caliensomycin, with phomol as a plausible precursor scaffold. This extensive scaffold reorganization is consistent with oxidative transformations previously described for related macrolactones, such as brefeldin A and SCH metabolites (Trenti et al. 2020). Together, these findings demonstrate that SpecReBoot does not only improve confidence in molecular networking but can directly facilitate biological discovery. By prioritizing stable yet low-similarity spectral relationships, SpecReBoot enabled the identification of cryptic chemical connections that guided both compound isolation and biosynthetic hypothesis generation, insights that would likely remain inaccessible under conventional similarity-based molecular networking alone.

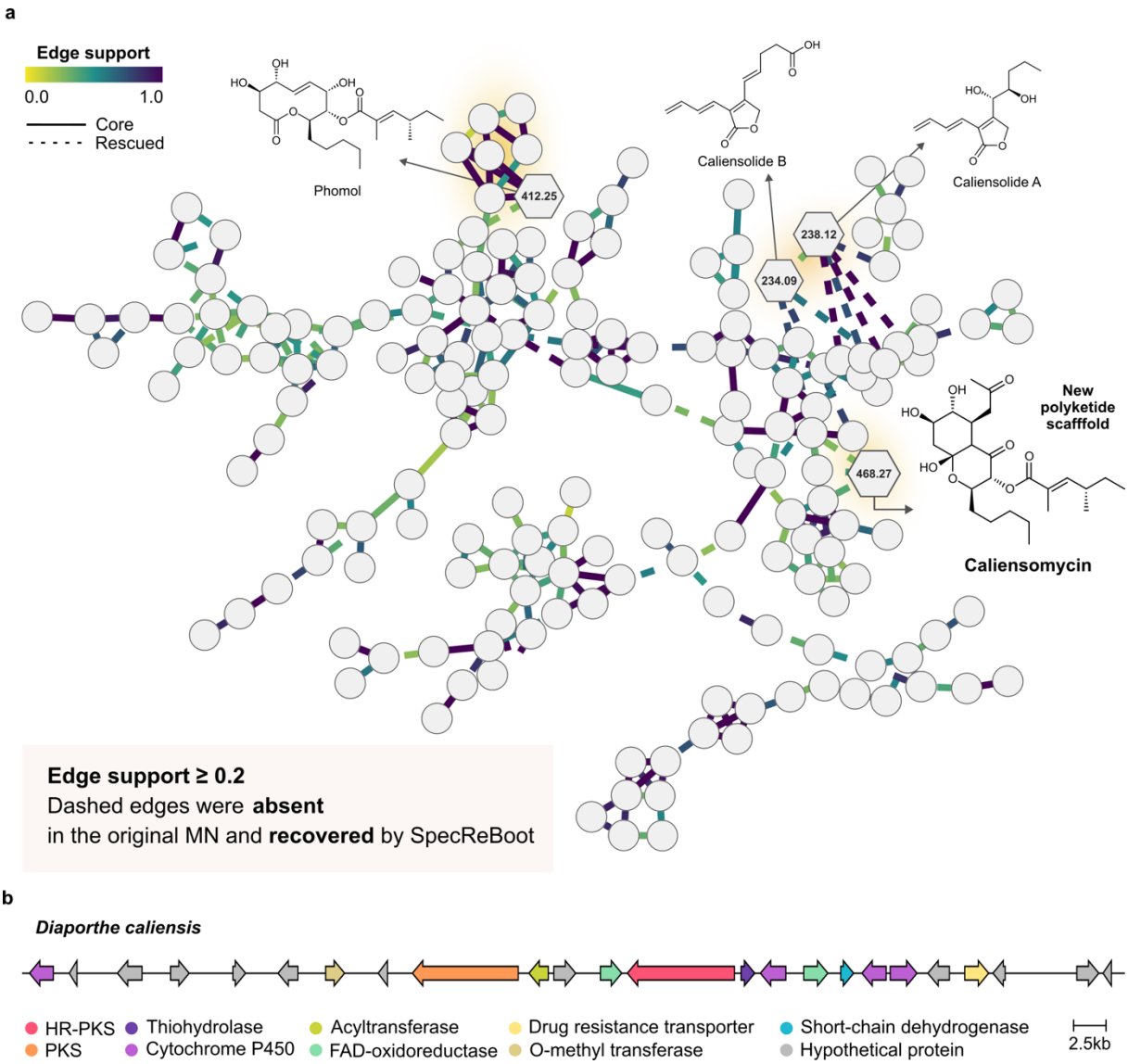

Fig. 4. SpecReBoot rescues hidden polyketide-lactone diversity in *Diaporthe caliensis* and enables discovery of a novel macrolactone scaffold. a, Molecular network reconstructed using bootstrap-derived edge support. Nodes represent MS/MS spectra and edges indicate spectral relationships. Solid edges denote core connections present in the original molecular network, whereas dashed edges correspond to relationships rescued by SpecReBoot that were absent under conventional similarity-based networking. Application of an edge-support threshold reconnects phomol with caliensolide A and B and reveals additional, previously disconnected chemistry. Guided by rescued connections, an uncharacterized node (*m/z* 468.27) was prioritized, leading to the isolation and structural elucidation of caliensomycin, an unprecedented polyketide-lactone scaffold. **b,** Candidate biosynthetic gene cluster (BGC) region for phomol and caliensomycin production in *Diaporthe caliensis*, predicted by antiSMASH v8.0.4. The BGC spans 15,623–100,409 nt (~85 kb) and was classified as a type I polyketide synthase (T1PKS) together with an expanded set of tailoring enzymes, including multiple cytochrome P450 monooxygenases and oxidoreductases, consistent with late-stage oxidative scaffold remodeling.

3. Discussion

The introduction of molecular networking into the natural products field has boosted the integration of metabolomics-based approaches as an essential toolkit of modern discovery pipelines (Alseekh and Fernie 2018, de Jonge et al. 2022). The widespread adoption of MN has also contributed to a rapid expansion of publicly available spectral data over the last decade, enabling large-scale comparative analyses and community-driven annotation efforts aimed at depicting the boundaries of chemical space (Wang et al. 2016, Aron et al. 2020). Despite the methodological advances such as Feature-Based MN and Ion Identity MN (Nothias et al. 2020, Schmid et al. 2021, Gentry et al. 2024), the core principle of organizing spectra by pairwise spectral similarity, remains sensitive to noise, missing fragments and experimental variability. As a result, most networks still lack an explicit, edge-level notion of confidence (Gloaguen et al. 2022, Sachsenberg et al. 2025). This gap contrasts with other data-driven fields in which uncertainty estimates for similarity-based inference are a standard practice, such as ultrafast approximations for phylogenetic bootstrap and the confidence index pLDDT recently introduced for AlphaFold 2.0 (Minh et al. 2013, Vander Meersche et al. 2025).

3.1. From deterministic similarity scores to confidence-aware connections

SpecReBoot addresses this challenge by quantifying the stability of spectral relations under fragment-level perturbations. By adapting the Felsenstein's bootstrap resampling principle from phylogenetics to MS/MS spectral data, SpecReBoot converts a single similarity score into a distribution-aware assessment and reports edge support as a confidence layer for molecular network construction and downstream interpretation. This framework consistently recovered chemically meaningful relationships that are missed by single similarity calculations, particularly for pairs with low or near-zero similarity (Fig. 2). Such "rescue" links arise when informative subsets of fragments repeatedly support mutual kNN neighbor relationships across bootstrap replicates, even when full-spectrum comparisons are diluted by noise, missing peaks, or divergent fragmentation patterns. In practice, using edge support as additional criterion reduces unstable connections and enriches networks for chemically related spectral pairs, including when starting from machine-learning similarities such as MS2DeepScore. The successful recovery of structurally related compounds in benchmark datasets and real-world natural products case studies demonstrates that stable fragment-level relationships can exist even when similarity scores alone fail to identify them.

3.2. Metric-agnostic stability and fragment-level interpretability enrich molecular networks

SpecReBoot is agnostic to the choice of the underlying spectral similarity metric. Across cosine-based and embedding models, such as Spec2Vec and MS2DeepScore, edge support exhibited highly conserved structural patterns despite substantial differences in the corresponding spectral similarity matrices, and these patterns closely resembled the organization of the chemical similarity based on Tanimoto scores of molecular fingerprints of the library structures (Fig. 3a). For instance, in the NIH Natural Products Library, applying spectral similarity thresholds alone retained 2,626 edges using modified cosine and 2,629 edges using MS2DeepScore, whereas incorporating an edge support-based threshold reduced these

numbers to 1,597 and 1,589 edges, respectively, producing comparable and chemically enriched network structures despite fundamentally different similarity models (Fig. 3b). Importantly, this enrichment did not increase cross-metric edge overlap, supporting the idea that edge support captures a complementary property, reproducibility under perturbation, rather than forcing agreement between dissimilar scoring methods. The persistence of stable yet metric-specific connections after dual filtering further suggests that current similarity measures provide complementary but incomplete views of spectral relatedness, motivating continued development of improved scoring methods.

Beyond cross-metric consistency, SpecReBoot enables fragment-level interpretability by documenting the bootstrap histories (e.g., bin-level resampling records). This allows direct inspection of which fragment bins repeatedly support and edge and how supporting evidence varies across datasets and scoring methods, exemplified by the aerucyclamide case study (Fig. 2). This perspective reconnects network edges to the underlying spectral evidence and can partially mitigate the limited explainability of embedding based similarities. More broadly, fragment-level insights are naturally compatible with substructure-oriented approaches such as MS2LDA and consensus-based strategies such as Multiple Mass Spectral Alignment (MMSA) (Torres-Ortega et al. 2025, Lau et al. 2026).

3.3. Confidence-aware molecular networking enables discovery of new chemistry

SpecReBoot can also reshape network topology in ways that directly facilitate natural products discovery. When re-examining the polyketide-lactone diversity produced by the endophytic fungus *Diaporthe caliensis*, the bootstrapping framework reconciled compound-class relationships that were previously only evident through shared MS2LDA-derived Mass2Motif (van der Hooft et al. 2016a, Charria-Girón et al. 20213). By removing unstable connections while rescuing low-similarity but high-support edges, SpecReBoot produced a coherent molecular network that connected polyketide-lactone molecular families previously fragmented under conventional molecular networking (Fig. 4).

This confidence-aware “rebooting” of the chemical diversity guided the prioritization of features outside the originally defined phomol- and caliensolide-like molecular families, leading to the isolation and structural characterization of an unprecedent macrolactone scaffold, caliensomycin. The identification of a candidate biosynthetic gene cluster shared between phomol and caliensomycin is consistent with oxidative remodeling routes described for macrolactones, in which cytochrome P450-initiated cascades drive extensive scaffold reorganization of pre-assembled macrolactone frameworks rather than de novo polyketide assembly (Zabala et al. 2014, Guengerich 2018, Trenti et al. 2020)

By analogy to brefeldin A and related SCH metabolites, we propose that caliensomycin arises from late-stage oxidative remodeling of phomol as a biosynthetic intermediate. In contrast to previously characterized systems, however, the rearrangement in this case appears to operate on a C10 macrolactone substrate, yielding an unprecedented scaffold. The expanded set of oxidative tailoring enzymes encoded within the *D. caliensis* candidate gene cluster, including multiple cytochrome P450 monooxygenases and oxidoreductases, provides a plausible

enzymatic basis for such transformations. Together, these observations suggest that phomol is not the terminal product of this pathway but a branching intermediate from which alternative oxidative routes generate structurally divergent macrolactones. Future biosynthetic studies will be required to resolve the enzymatic steps underlying the transannulation mechanism and the origin of the additional C3 unit in caliensomycin.

3.4. Adoption in existent ecosystems

From a practical perspective, SpecReBoot is designed for compatibility with widely used metabolomics infrastructures. On the one hand, it is built around *matchms*, a well-maintained software infrastructure for mass spectrometry, supports the usage of a broad range of spectral similarity metrics beyond classical cosine-based approaches, including embedding-based models such as Spec2Vec and MS2DeepScore. This is particularly relevant in cases where strict cosine-based comparisons fail to capture relationships between spectra derived from molecules differing by multiple or complex modifications, a well-known pitfall of peak-alignment similarity measures (van der Hooft et al. 2016, Huber et al. 2020, Beniddir et al. 2021, Huber et al. 2021, De Jonge et al. 2024ab).

On the other hand, although SpecReBoot is not yet integrated into GNPS2, it directly accepts GNPS2-based molecular networking outputs as inputs, enabling confidence-aware refinement of existing networks and the rescue of reproducible, low-similarity edges, while fully preserving GNPS-derived metadata, library annotations, and provenance information. Together, these design choices facilitate straightforward adoption of SpecReBoot within established metabolomics workflows and lower the barrier to incorporating confidence-aware analyses into routine molecular networking practice. Looking ahead, SpecReBoot could be integrated into community platforms such as GNPS2, mzmine, or SIRIUS, which would further increase accesibility for users without programming expertise and align with ongoing efforts to unifiy confidence-aware analyses within widely used metabolomics software ecosystems.

3.5. Limitations and practical considerations

SpecReBoot introduces several user-defined parameters, including the number of nearest neighbors (k) used for mutual kNN neighbors detection and the number of bootstrap replicates (B). Although these parameters exhibited stable behavior in our analyses, their optimal selection may depend on dataset size, chemical diversity, and the intended downstream application, we therefore recommend reporting *(k, B)* alongside networking settings and providing sensitivity checks in representative datasets. Here, for the largest dataset (MSn-dataset) we used a k = 20 and B = 100, while for the smallest dataset we used k = 3 and B = 30 (see Table S4 for more details). As with other resampling-based approaches, bootstrap analysis increases computational cost relative to single similarity calculations, although ongoing improvements in the *matchms* framework, algorithmic optimization and parallelization strategies are expected to further mitigate these challenges, facilitating efficient application to increasingly large-scale metabolomics datasets. More broadly, while the direct transfer of nonparametric bootstrapping concepts from phylogenetics to tandem mass spectral similarity relationships proved effective relatively 'out of the box', this

implementation represents only a first step toward a confidence-aware framework tailored to tandem mass spectrometry. Further testing and optimization of various resampling strategies and bootstrapping parameter settings are part of future research avenues we foresee, possibly further improving and fine-tuning our approach. Finally, edge support is not intended to replace spectral similarity scores or machine-learning-based embeddings; rather, it provides a complementary confidence layer that filters unstable edges and prioritizes reproducible relationships. As demonstrated by our benchmarking analyses, edge support alone does not capture the magnitude of chemical similarity and should therefore be interpreted in conjunction with conventional mass spectral similarity measures.

3.6. Outlook

Taken together, SpecReBoot shifts molecular networking from a purely deterministic framework toward one that explicitly accounts for confidence and reproducibility, thereby enabling more reliable interpretation of spectral relationships in complex natural product mixtures. By identifying a robust, metric-agnostic core of spectral connections that is consistently recovered across fundamentally different similarity models, SpecReBoot provides means to harmonize molecular network topology without artificially forcing agreement between similarity scores. Importantly, our approach introduces a new avenue for interpreting ML-based methods, such as Spec2Vec and MS2DeepScore have expanded the scope of detectable spectral relationships, yet their limited explainability had remained a key challenge. By edge support to fragment-level resampling information, SpecReBoot reconnects learned similarity relationships to underlying spectral evidence, offering a confidence-aware and interpretable layer that complements AI-driven similarity scoring and could guide the development of improved embedding-based approaches.

At the same time, our results demonstrate that individual similarity metrics retain stable, metric-specific connections that are not captured by others, highlighting that any single similarity score provides an incomplete and potentially biased view of spectral relatedness. This observation opens new opportunities for consensus network construction, in which a shared core of reproducible edges is complemented by metric-specific insights to more fully capture chemical diversity. Looking forward, bootstrap-derived confidence measures open new opportunities for uncertainty-aware metabolomics analyses, including principled network comparison, consensus building across datasets and similarity metrics, and improved integration with downstream annotation, prioritization, and substructure discovery strategies. Beyond natural product discovery, SpecReBoot provides a general computational metabolomics framework for assessing the robustness of MS/MS-derived spectral relationships, with potential applications in exposomics, clinical metabolomics, and large-scale exploration of public spectral repositories.

4. Methods

4.1 Datasets and accessions

4.1.1 RiPP reference set construction. MS/MS spectra were mined from the GNPS platform using publicly available MGF files with associated structural identifiers (SMILES, InChIKey, InChI or InChIAux). Duplicate entries were removed based on identical structural identifiers and/or redundant spectrum metadata. For spectra lacking SMILES, structural representations were generated from available InChI-based identifiers where possible. Compounds were classified using NPClassifier, and peptide-class candidates were manually curated against the literature to confirm RiPP biosynthetic origin. The final RiPP reference library comprised 24 curated MS/MS spectra with literature-supported structures and biosynthetic context.

4.1.2 Benchmark datasets. The pesticides dataset, introduced by the *matchms* team, comprised 76 reference spectra (pesticides.mgf). The NIH Natural Products Library dataset comprised 1,267 reference spectra (GNPS-NIH-NATURALPRODUCTSLIBRARY.mgf). The MSn-COCONUT benchmark set was derived from a harmonized collection of MS/MS spectra assembled from GNPS and MSn-Lib (Brungs et al. 2025). From the initial 1,017,531 spectra, only MS2 spectra acquired in positive ionization mode were retained. To restrict the dataset to natural products, annotated spectra were matched by InChIKey connectivity block (first 14 characters since it contains the structural information) against the COCONUT database, retaining only spectra whose compounds were present in COCONUT (407,029 registered compounds). This yielded 15,947 spectra, of which 12,929 remained after preprocessing and were used for downstream analyses.

4.2 Implementation, required inputs and outputs

SpecReBoot operates on standard GNPS2-compatible data formats and supports two execution modes: (i) full_matchms mode, which starts from a MGF file (can be derived from mzmine for example) and exports new networks, and (ii) gnps_workflow mode, which starts from a GNPS2 GraphML network flus its merged MGF and appends the bootstrap-derived supported edges.

Required inputs

1. **MS/MS spectra (both modes):** an MGF file containing MS/MS spectra and associated metadata. For datasets assembled from multiple sources or libraries, each spectrum must include a stable, unique identifier (e.g., feature_id) or be added. This is required to ensure consistent spectrum tracking and unambiguous mapping across intermediate outputs and molecular network nodes.
2. **GNPS2 workflow mode only (gnps_workflow):** a GraphML network generated by a GNPS2 molecular job, plus the corresponding merged MGF used to construct that GNPS2 network.

Key user parameters

- Binning resolution (number of decimal places; default: 2 decimals), number of bootstrap replicates B, mutual kNN parameter k, similarity metric (cosine, modified cosine, Spec2Vec, MS2DeepScore), and optional dual thresholds: similarity threshold (τ_{sim}) and support threshold (τ_{sup}).

Primary outputs

- **Pairwise tables (CSV):** (i) bootstrap-mean similarity and (ii) edge support for spectrum pairs.
- **Networks (GraphML):** depending on mode, either newly constructed GraphML networks or a GNPS2 GraphML with appended bootstrap-derived edge support attributes.
- **Optional provenance/history file:** bootstrap replicate-level diagnostic information (e.g., sampled/missing bin sets) when enabled for reproducibility and debugging.

4.3 MS/MS preprocessing

All datasets were imported from MGF format. For datasets assembled from multiple libraries, a unique identifier field (feature_id) was added to the metadata to ensure stable spectrum tracking and unambiguous mapping across intermediate outputs. Spectra were processed using *matchms* (v0.31.0) default harmonization and cleaning pipelines (DEFAULT_FILTERS and CLEAN_PEAKS). By default, intensities were normalized to unit height (scaled by the maximum intensity). Cleaning steps followed the *matchms* pipeline implementation and included precursor-window peak removal (±17 Da around the precursor, excluding the precursor peak), retention of peaks within ±50 Da of at least one of the six most intense peaks, and an upper bound of 1,000 fragment peaks per spectrum. Users can modify these steps depending on dataset characteristics and metadata completeness.

4.4 Peak binning and global fragment-bin space

Fragment *m/z* values were discretized by rounding to a user-defined number of decimal places (default: two decimals; bin width 0.01 Da). A dataset-wide global bin list G was constructed as the union of all discretized fragment m/z values observed across all spectra. Each spectrum was represented its discretized fragment *m/z* values and the corresponding normalized intensities.

4.5 SpecReBoot algorithm

4.5.1 Bootstrap resampling over global bins

SpecReBoot applies bootstrap perturbations by repeatedly masking spectra using replicate-specific subsets of the dataset-wide global bins G. For each bootstrap replicate, $|G|$ bin indices are sampled uniformly with replacement from the global bin list, and duplicates are collapsed to obtain a unique sampled bin subset $G^{(b)} \subseteq G$. Each original spectrum is converted into a pseudo-spectrum by retaining only those peaks whose discretized m/z value is contained in $G^{(b)}$; all other peaks per pseudo-spectrum are removed. Importantly, this process does not

shuffle or exchange intensities across spectra: intensities remain attached to the original spectrum and are only retained or removed in each bootstrap replicate.

Bins are sampled with replacement and duplicates are collapsed, and each global bin list replicate retains on average $1-(1-1/|G|)^{|G|} \approx 0.6321$ of the global bins, thus per replicate ~36.8 of the bins are retained and ~63.2% of the bins are excluded. Consequently, depending on the fragment composition, some spectra may be only weakly perturbed by a given replicate, motivating the use of multiple bootstrap replicates to robustly estimate edge stability. Doing multiple bootstraps is further motivated by the fact that each spectrum occupies only a small subset of the global bin space (i.e., binned spectra are sparse), a given bootstrap replicate may remove few or no peaks from some spectra and many peaks from others.

If a spectrum contains no peaks after masking, a dummy peak (*m/z* set to the first global bin and intensity 0.0) is inserted to ensure similarity computation remains well-defined; such spectra effectively contribute zero similarity and do not form mutual kNN relationships for that replicate. For each replicate, pairwise similarities are computed using the selected similarity metric, producing a replicate similarity matrix. Two outputs are then aggregated across replicates:

1. **Mean similarity**: defined for each spectrum pair (*i*, *j*) as the mean of replicate similarity values across B bootstraps.

2. **Edge support**: defined for each spectrum pair (*i*, *j*) as the fraction of bootstraps in which spectra i and j are recovered as **mutual** top-k nearest neighbors. Specifically, for each spectrum i, the top-k neighbors are obtained by ranking similarity values within replicate b (excluding self), and an undirected edge is counted once per replicate when *i* is in the top-*k* neighbors of *j* and *j* is in the top-*k* neighbors of *i*. Edge support is reported as $\text{support}_{ij} = c_{ij}/B$, where c_{ij} is the number of replicates in which the mutual kNN relationship is registered (default: $k = 5$).

4.5.2 Spectral similarity metrics and configurations

Cosine and modified cosine similarities were computed using *matchms* similarity implementations with user-defined precursor *m/z* tolerance where applicable. Spec2Vec similarities were computed using a model retrained for positive ionization mode spectra using curated positive-mode training spectra from [https://zenodo.org/records/12543129], files associated can be downloaded here: https://zenodo.org/records/15857387; and MS2DeepScore similarities were computed using a publicly available pretrained model [https://zenodo.org/records/17826815]. Model sources and versions are reported in the Code availability section.

For network construction, SpecReBoot supports dual filtering, i.e., applying thresholds on both the bootstrap-mean similarity and edge support. In this way, an edge between spectra *i* and *j* is retained only if bootstrap-mean similarity $\geq \tau_{sim}$ and edge support $\geq \tau_{sup}$, where τ_{sim} and τ_{sup} are user-defined.

4.6 Network construction and graph parameters

SpecReBoot supports two execution modes:

(i) full_matchms mode. Starting from an MGF file, SpecReBoot performs preprocessing, binning, and bootstrap similarity estimation, and can export up to three GraphML networks:

1. a **base network** thresholded on bootstrap-mean similarity (bootstrap-mean similarity ≥ τ_{sim});
2. a **dual-threshold network** requiring both bootstrap-mean similarity and edge support (bootstrap-mean similarity ≥ τ_{sim} and edge support ≥ τ_{sup}); and
3. a **core/rescue network** that labels edges as "core" (high similarity and high support) or *rescued* (lower similarity but high support) based on user-defined criteria.

(ii) gnps_workflow mode. SpecReBoot loads a GNPS-exported GraphML network and the corresponding merged MGF, maps spectra to nodes via identifiers, preserves all original node and edge attributes, and appends bootstrap-derived edge attributes (mean similarity and edge support). Optionally, it generates rescued-edge annotations following the same core/rescue logic.

4.7 Benchmarking against chemical similarity

To evaluate whether edge support enriches chemically meaningful relationships beyond similarity alone, structural similarity was computed for annotated compounds using RDKit Morgan fingerprints (radius 2, 4,096 bits) and Tanimoto similarity. Pairwise similarity tables were generated by merging bootstrap-mean similarity, edge support, and Tanimoto similarity by spectrum identifiers.

4.8 Statistical analysis

Binary classification performance was assessed using receiver operating characteristic (ROC) and precision–recall (PR) analyses, defining "chemically similar" pairs as those with Tanimoto threshold ≥ 0.7. In addition, we visualized the relationship between mean spectral similarity (x-axis) and structural similarity (y-axis), with points colored by edge support to highlight stable spectral relationships across bootstrap replicates.

4.9 Experimental procedures for *Diaporthe caliensis* and structure elucidation

4.9.1 Fungal genomic DNA extraction, and genome sequencing

For genomic DNA isolation, *D. caliensis* was cultivated on sterile 50 µm nylon mesh placed on the surface of Petri dishes containing potato dextrose agar (PDA). Cultures were prepared in duplicate to obtain sufficient biomass for high-quality DNA extraction. After 7 days of cultivation, the mycelia developing on the mesh was harvested by scraping with a sterile scalpel, snap-frozen in liquid nitrogen, and stored at −80 °C until further processing. Genomic DNA was extracted using the NucleoBond HMW DNA kit following the manufacturer's protocol. Briefly, frozen mycelia were ground under liquid nitrogen and lysed in the presence

of proteinase K. DNA was subsequently purified using the kit's affinity columns and eluted in HE buffer. DNA integrity and purity were assessed by electrophoresis on a 1% agarose gel prior to sequencing. Whole-genome sequencing was performed using Illumina technology. Raw sequencing reads were subjected to quality assessment and adapter trimming using FASTQ, and de novo genome assembly was carried out with SPAdes version 3.15.3.

4.9.2 Cultivation of *D. caliensis* and isolation of caliensomycin

Diaporthe caliensis CBS 149729, originally isolated from the medicinal plant *Otoba gracilipes* in Cali, Colombia, was cultivated as previously described (Charria-Girón *et al.* 2023). Briefly, seed cultures were prepared in 200 mL yeast malt extract broth (YM broth) by transferring colonies from fresh cultures in yeast malt agar (YM agar) and incubating at 23 °C with agitation at 140 rpm for 7 days. Aliquots (6 mL) of this seed culture were then used to inoculate 12 flasks containing solid oat medium (OFT), which were incubated statically at 23 °C for 15 days. Metabolite extraction was performed following the procedure reported in Charria-Girón *et al.* (2023), yielding 1,153 mg of crude extract. The crude extract was pre-fractionated by flash chromatography (Grace Reveleris, Columbia, MD, USA) using a 40 g silica cartridge. The mobile phase consisted of solvent A (DCM), solvent B (acetone), and solvent C [(DCM/acetone 8:2):MeOH], with the following gradient: 100% A for 5 min, increasing to 100% B over 20 min, followed by a transition to 100% C over 15 min and a final hold at 100% C for 5 min. Five fractions (F1–F5) were collected and subsequently analyzed for the presence of previously uncharacterized metabolites. Fraction F3 was purified by preparative RP-HPLC (Synergi Polar-RP, 250 × 50 mm, 10 µm; Phenomenex) with H_2O (A) and ACN (B) (0.1% formic acid) at 45 mL/min (UV 210, 240, and 300 nm), using a 5–32% B (15 min), 32–36% B (40 min), 36–100% B (25 min) gradient and 100% B (5 min), affording caliensomycin (1.4 mg, *t*R = 21–23 min).

4.9.3. Spectral data

Optical rotations were recorded employing an MCP 150 circular polarimeter (Anton Paar, Seelze, Germany) at 20°C. UV/Vis spectra were recorded with a UV-2450 spectrophotometer (Shimadzu, Kyoto, Japan). Spectral data were measured in MeOH (Uvasol, Merck, Darmstadt, Germany) for all compounds. The 1D- and 2D-nuclear magnetic resonance (NMR) spectra were recorded with an Avance III 700 spectrometer with a 5 mm TCI cryoprobe (^{1}H NMR: 700 MHz, ^{13}C: 175 MHz, Bruker, Billerica, MA, USA) and an Avance III 500 spectrometer (^{1}H NMR: 500 MHz, ^{13}C: 125 MHz, Bruker, Billerica, MA, USA). The chemical shifts δ were referenced to the solvents DMSO-d_6 (^{1}H, δ = 2.50; ^{13}C, δ = 39.51).

Caliensomycin: White powder; UV (MeOH) λ_{max} (log ε) 263 (3.35), (log ε) 202.5 (3.06); HRESI-MS: *m/z* 469.2719 $[M + H]^+$ (calculated for $C_{24}H_{39}O_8^+$: 469.5882 Da).

4. 10. Data and Code Availability

SpecReBoot codebase is publicly available in our GitHub repository (https://github.com/ECharria/SpecReBoot). The documentation is available in the README.md file for installation and execution. The models for Spec2Vec and MS2DeepScore, datasets used for benchmarking and

case studies, as well as the produced networks for the NIH Natural Products Library are available at https://zenodo.org/records/18466976.

5. Author contributions

E.C.-G and L.R.T.-O conceived the concept behind SpecReBoot and developed the methodology. Both implemented the computational framework, performed data analysis, and led the manuscript writing. J.M.G. contributed the RiPP-associated case study and performed manual curation of public spectral repositories. Y.M.F. supervised and supported the initial analysis of the experimental part of the study. N.H.C. contributed the genome sequencing data and project coordination. F.S. performed structural elucidation analysis and supervised the experimental part of the study. M.H.M. contributed to the initial development of the SpecReBoot concept, interpretation of results, and critical revision of the manuscript. J.J.J.v.d.H. supervised the project, contributed to conceptualization, interpretations of results, and manuscript writing and editing. All authors reviewed and approved the final version of the manuscript.

6. Acknowledgements

We thank Esther Surges for recording NMR data and Ulrike Beutling for performing tandem mass spectrometry measurements related to the *D. caliensis* case study. Access to genetic resources for the study of *D. caliensis* was granted under contract RGE244-44#190 by the Autoridad Nacional de Licencias Ambientales (ANLA), Ministerio de Ambiente y Desarrollo Sostenible, Colombia. Whole-genome sequencing was conducted within the framework of a BiGTREE mobility scholarship awarded to Rafael Góngora. Laboratory work, including molecular biology procedures, sequence analysis, and genome assembly, was carried out at the University of Oslo under the supervision of Professor Inger Skrede, with results reported in the corresponding BiGTREE scholarship activity report. Transport of fungal cultures was conducted under export and import authorization for specimens of biological diversity listed in the appendices of the CITES Convention (Permit No. 3018), issued by ANLA. L.R.T.O gratefully recognizes the Marie Skłodowska-Curie grant under the European Union's Horizon Europe programme MAGiC-MOLFUN (grant no. No. 101072485).

7. Conflict of interests

M.H.M. is a member of the scientific advisory boards of Hexagon Bio and Hothouse Therapeutics Ltd. J.J.J.v.d.H. is member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy and consults for Corteva Agriscience, Indianapolis, IN, USA.

8. References

Alseekh S, Fernie AR. Metabolomics 20 years on: what have we learned and what hurdles remain?. *Plant J.* **94**, 933–942 (2018). doi:10.1111/tpj.13950.

Arnison PG, Bibb MJ, Bierbaum G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep*. **30**, 108–160 (2013). doi:10.1039/c2np20085f.

Aron AT, Gentry EC, McPhail KL, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc*. **15**, 1954–1991 (2020). doi:10.1038/s41596-020-0317-5.

Atanasov, A.G., Zotchev, S.B., Dirsch, V.M. et al. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug. Discov.* **20**, 200–216 (2021). doi:10.1038/s41573-020-00114-z.

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.* **38**, 1967–1993 (2021). doi:10.1039/d1np00023c.

Bittremieux W, Schmid R, Huber F, van der Hooft JJJ, Wang M, Dorrestein PC. Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment For Discovery of Structurally Related Molecules. *J. Am. Soc. Mass. Spectrom*. **33**, 1733–1744 (2022). doi:10.1021/jasms.2c00153.

Brungs C, Schmid R, Heuckeroth S, et al. MS^nLib: efficient generation of open multi-stage fragmentation mass spectral libraries. *Nat. Methods* **22**, 2028–2031 (2025). doi:10.1038/s41592-025-02813-0.

Bushuiev R, Bushuiev A, Samusevich R, et al. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat. Biotechnol.* (2025). doi: 10.1038/s41587-025-02663-3.

Cardoza RE, McCormick SP, Izquierdo-Bueno I, et al. Identification of polyketide synthase genes required for aspinolide biosynthesis in *Trichoderma arundinaceum*. *Appl. Microbiol. Biotechnol*. **106**, 7153–7171 (2022). doi:10.1007/s00253-022-12182-9.

Charria-Girón E, Espinosa MC, Zapata-Montoya A, et al. Evaluation of the antibacterial activity of crude extracts obtained from cultivation of native endophytic fungi belonging to a tropical montane rainforest in Colombia. *Front. Microbiol*. **12**, 716523 (2021). doi:10.3389/fmicb.2021.716523.

Charria-Girón E, Marin-Felix Y, Beutling U, et al. Metabolomics insights into the polyketide-lactones produced by *Diaporthe caliensis* sp. nov., an endophyte of the medicinal plant *Otoba gracilipes*. *Microbiol. Spectr*. **11**, e0274323 (2023a). doi:10.1128/spectrum.02743-23.

Charria-Girón E, Vasco-Palacios AM, Moncada B, Marin-Felix Y. Colombian fungal diversity: Untapped potential for diverse applications. *Microbiol. Res*. **14**, 2000–2021 (2023b). doi:10.3390/microbiolres14040135.

Clardy J, Fischbach MA, Currie CR. The natural history of antibiotics. *Curr. Biol.* **19**, R437–R441 (2009). doi:10.1016/j.cub.2009.04.001.

Clark TN, Houriet J, Vidar WS, et al. Interlaboratory Comparison of untargeted mass spectrometry data uncovers underlying causes for variability. *J. Nat. Prod*. **84**, 824–835 (2021). doi:10.1021/acs.jnatprod.0c01376.

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*. **16**, 88 (2024). doi:10.1186/s13321-024-00878-1.

de Jonge NF, Joas D, Truong L-J, et al. Reliable cross-ion mode chemical similarity prediction between MS2 spectra. Preprint at https://www.biorxiv.org/content/10.1101/2024.03.25.586580v5 (2024).

de Jonge NF, Mildau K, Meijer D, et al. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* **18**, 103 (2022). doi:10.1007/s11306-022-01963-y.

Demain AL, Fang A. The natural functions of secondary metabolites. *Adv. Biochem. Eng. Biotechnol.* **69**,1–39 (2000). doi:10.1007/3-540-44964-7_1.

Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U S A* **93**, 13429–13434 (1196). doi:10.1073/pnas.93.23.13429.

Efron, B., & Tibshirani, R.J. *An Introduction to the Bootstrap* (Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593.

El Abiead Y, Mohanty I, Xing S, et al. A Perspective on Unintentional Fragments and Their Impact on the Dark Metabolome, Untargeted Profiling, Molecular Networking, Public Data, and Repository Scale Analysis. *JACS Au* **5**, 5828–5850 (2025). doi:10.1021/jacsau.5c01063.

Felsenstein J. Confidence Limits On Phylogenies: An Approach Using The Bootstrap. *Evolution* **39**, 783–791 (1985). doi:10.1111/j.1558-5646.1985.tb00420.x.

Gentry EC, Collins SL, Panitchpakdi M, et al. Reverse metabolomics for the discovery of chemical structures from humans. *Nature*. **626**, 419–426 (2024). doi:10.1038/s41586-023-06906-8.

Gloaguen Y, Kirwan JA, Beule D. Deep learning-assisted peak curation for large-scale LC-MS metabolomics. *Ana.l Chem.* **94,** 4930–4937 (2022). doi:10.1021/acs.analchem.1c02220.

Guengerich FP. Mechanisms of cytochrome P450-catalyzed oxidations. *ACS Catal*. **8**, 10964–10976 (2018). doi:10.1021/acscatal.8b03401.

Hoyos LV, Vasquez-Muñoz LE, Osorio Y, et al. Tailored culture strategies to promote antimicrobial secondary metabolite production in *Diaporthe caliensis*: a metabolomic approach. *Microb. Cell Fact*. **23**, 328 (2024). doi:10.1186/s12934-024-02567-y.

Huber F, Ridder L, Verhoeven S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol*. **17**, e1008724 (2021). doi:10.1371/journal.pcbi.1008724.

Huber F, Verhoeven S, Meijer C, et al. matchms-processing and similarity evaluation of mass spectrometry data. *J. Open Source Softw.* **5**, 2411 (2020). doi:10.21105/joss.02411.

Keller NP, Turner G, Bennett JW. Fungal secondary metabolism - from biochemistry to genomics. *Nat. Rev. Microbiol*. **3**, 937–947 (2005). doi:10.1038/nrmicro1286.

Lau A, Wang X, Xu T, et al. Multiple spectrum alignment for molecular networking exploration and discovery. *J. Am. Soc. Mass Spectrom*. **37**, 126–134 (2026). doi:10.1021/jasms.5c00237.

Minh BQ, Nguyen MA, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol*. **30**, 1188–1195 (2013). doi:10.1093/molbev/mst024.

Montalbán-López M, Scott TA, Ramesh S, et al. New developments in RiPP discovery, enzymology and engineering. *Nat Prod Rep*. **38**, 130–239 (2021). doi:10.1039/d0np00027b.

Nie Q, Sun C, Liu S, Gao X. Correction to "Exploring bioactive fungal RiPPs: Advances, challenges, and future prospects". *Biochemistry* **64**, 3485 (2025). doi:10.1021/acs.biochem.5c00383.

Nothias LF, Petras D, Schmid R, et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods*. **17**, 905–908 (2020). doi:10.1038/s41592-020-0933-6.

Portmann C, Blom JF, Gademann K, Jüttner F. Aerucyclamides A and B: isolation and synthesis of toxic ribosomal heterocyclic peptides from the cyanobacterium *Microcystis aeruginosa* PCC 7806. *J. Nat. Prod*. **71**, 1193-1196 (2008). doi:10.1021/np800118g.

Portmann C, Blom JF, Kaiser M, Brun R, Jüttner F, Gademann K. Isolation of aerucyclamides C and D and structure revision of microcyclamide 7806A: heterocyclic ribosomal peptides from *Microcystis aeruginosa* PCC 7806 and their antiparasite evaluation. *J. Nat. Prod*. **71**, 1891-1896 (2008). doi:10.1021/np800409z.

Sachsenberg T, Pino LK, Brunet M, Bludau I, Kohlbacher O, Vizcaino JA, Bittremieux W. Perspectives in computational mass spectrometry: recent developments and key challenges. *Bioinformatics Advances* **5**, vbaf301 (2025). doi:10.1093/bioadv/vbaf301.

Schmid R, Petras D, Nothias LF, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat. Commun*. **12**, 3832 (2021). doi:10.1038/s41467-021-23953-9.

Skinnider MA, Johnston CW, Edgar RE, et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc. Natl. Acad. Sci. U S A*. **113**, E6343–E6351 (2016). doi:10.1073/pnas.1609014113.

Snider BB, Zhou J. Synthesis of (+)-sch 642305 by a biomimetic transannular Michael reaction. *Org. Lett*. **8**, 1283–1286 (2006). doi:10.1021/ol052948+.

Stein S. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem*. **84**, 7274–7282 (2012). doi:10.1021/ac301205z.

Torres-Ortega LR, Dietrich J, Wandy J, Mol H, van der Hooft JJJ. Large-scale discovery and annotation of hidden substructure patterns in mass spectrometry profiles. Preprint at https://doi.org/10.1101/2025.06.19.659491.

Trenti F, Lebe K, Adelin E, et al. Investigating the biosynthesis of Sch-642305 in the fungus *Phomopsis* sp. CMU-LMA. *RSC Advances* **10**, 27369–27376 (2020). doi: 10.1039/D0RA05311B.

van der Hooft JJJ, Padmanabhan S, Burgess KE, Barrett MP. Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics*. **12**, 125 (2016B). doi:10.1007/s11306-016-1064-z.

van der Hooft JJJ, Wandy J, Barrett MP, Burgess KE, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci. USA* **113**, 13738–13743 (2016a). doi:10.1073/pnas.1608041113.

Vander Meersche Y, Diharce J, Gelly JC, Galochkina T. Flexibility or uncertainty? A critical assessment of AlphaFold 2 pLDDT. *Structure*. **33**, 2157–2163.e2. doi:10.1016/j.str.2025.09.001.

Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol*. **34**, 828-837 (2016). doi:10.1038/nbt.3597.

Zabala AO, Chooi YH, Choi MS, Lin HC, Tang Y. Fungal polyketide synthase product chain-length control by partnering thiohydrolase. *ACS Chem. Biol*. **9**, 1576–1586 (2014). doi:10.1021/cb500284t.

Zdouc MM, van der Hooft JJJ, Medema MH. Metabolome-guided genome mining of RiPP natural products. *Trends Pharmacol. Sci*. **44**, 532-541 (2023). doi:10.1016/j.tips.2023.06.004.

Ziemert N, Ishida K, Quillardet P, et al. Microcyclamide biosynthesis in two strains of *Microcystis aeruginosa*: from structure to genes and vice versa. *Appl. Environ. Microbiol*. **74**, 1791-1797 (2008). doi:10.1128/AEM.02392-07.