

MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools

Madeleine Ernst^{1,2*}, Kyo Bin Kang^{1,3}, Andrés Mauricio Caraballo-Rodríguez¹, Louis-Felix Nothias¹, Joe Wandy⁴, Mingxun Wang¹, Simon Rogers⁵, Marnix H. Medema⁶, Pieter C. Dorrestein^{1,7,8}, Justin J.J. van der Hooft^{1,6*}

¹ Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA

² Present address: Department of Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark

³ College of Pharmacy, Sookmyung Women's University, Seoul, Republic of Korea

⁴ Glasgow Polyomics, University of Glasgow, Glasgow, United Kingdom

⁵ School of Computing Science, University of Glasgow, Glasgow, United Kingdom

⁶ Bioinformatics Group, Department of Plant Sciences, Wageningen University, Wageningen, The Netherlands

⁷ Department of Pediatrics, University of California San Diego, La Jolla, CA, USA

⁸ Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA

* Correspondence: maet@ssi.dk (M.E.), justin.vanderhooft@wur.nl (J.J.J.v.d.H.)

Abstract:

Metabolomics has started to embrace computational approaches for chemical interpretation of large data sets. Yet, metabolite annotation remains a key challenge. Recently, molecular networking and MS2LDA emerged as molecular mining tools that find molecular families and substructures in mass spectrometry fragmentation data. Moreover, *in silico* annotation tools obtain and rank candidate molecules for fragmentation spectra. Ideally, all structural information obtained and inferred from these computational tools could be combined to increase the resulting chemical insight one can obtain from a data set. However, integration is currently hampered as each tool has its own output format and efficient matching of data across these tools is lacking. Here, we introduce MolNetEnhancer, a workflow that combines the outputs from molecular networking, MS2LDA, *in silico* annotation tools (such as Network Annotation Propagation or DEREPLICATOR) and the automated chemical classification through ClassyFire to provide a more comprehensive chemical overview of metabolomics data whilst at the same time illuminating structural details for each fragmentation spectrum. We present examples from four plant and bacterial case studies and show how MolNetEnhancer enables the chemical annotation, visualization, and discovery of the subtle substructural diversity within molecular families. We conclude that MolNetEnhancer is a useful tool that greatly assists the metabolomics researcher in deciphering the metabolome through combination of multiple independent *in silico* pipelines.

Keywords: chemical classification; *in silico* workflows; metabolite annotation; metabolite identification; metabolome mining; molecular families; networking; substructures.

1. Introduction

Metabolomics has matured into a research field generating increasing amounts of metabolome profiles of complex metabolite mixtures aiming to provide biochemical insights. Mass spectrometry has become the workhorse of metabolomics and typical untargeted experiments currently result in qualitative and semi-quantitative information on several thousands of molecular ions across tens to hundreds of samples. Technical advances in the last decade have allowed researchers to fragment increasing amounts of mass peaks that result in mass fragmentation spectra (MS/MS or MS2). Metabolite annotation and identification tools have benefited from these advances as now more MS2 spectra per sample can be queried in reference libraries in order to find candidate structures or submitted to *in silico* tools that propose a putative structure [1–9].

Despite these tremendous advances, a key challenge remaining for metabolomics researchers is to biochemically interpret large-scale untargeted metabolomics studies due to the complexity of the metabolomes represented by mass fragmentation spectra to which actual chemical structures need to be assigned, and for which reference spectra are not available. In biological samples, many metabolites share molecular substructures and form structurally related molecular families (MFs) of various chemical classes, which has

inspired metabolome mining tools exploiting these biochemical relationships. Indeed, since the molecular networking approach was proposed in 2012 [10], numerous complementary molecular mining workflows as well as annotation and classification tools have been introduced including SIRIUS [3], CSI:FingerID [4], MetFusion [11], MetFamily [12], and many others [1,2,7,8,13–22] and their combined use for natural product discovery was very recently reviewed [23]. Where tandem mass spectral molecular networking efficiently can group molecular features in molecular families [10], MS2LDA can discover substructures that aid in further annotation of subfamilies and shared modifications [14]. Furthermore, recently introduced tools such as Network Annotation Propagation (NAP) [8], DEREPLICATOR [1], VarQuest [2], or SIRIUS+CSI:FingerID [4] allow for effective searching in chemical databases for candidate structures. These candidate structures can now be automatically chemically classified using the ClassyFire tool [16] which takes molecular descriptors as SMILES or InchiKeys as input and outputs hierarchical chemical ontology terms. Taken together, these developments enable the discovery of relations between millions of spectra and the listing of candidate structures from various spectral libraries or alternatively from compound libraries using *in silico* approaches.

Whilst each of those tools produce useful structural information, their combined application has been hampered by the use of different file formats, platforms, and the challenge to match molecular features across the outputs of these tools. We postulate that whilst each tool provides complementary insights, their combined use allows an increased level of biochemical interpretation: i.e., the sum becomes greater than the individual parts. Furthermore, it would be practically advantageous to combine all these results in one place. We have previously described the integration of Mass2Motifs and chemical classifications with molecular networks to assess the chemical diversity within a subset of species of the plant genus *Euphorbia* [24] and the plant family Rhamnaceae [25]. However, in those studies, integration was achieved using custom in-house scripts in R, hampering adoption by the community. Moreover, results of the peptide annotation tools DEREPLICATOR and VarQuest were not included in those custom scripts.

Here, we introduce MolNetEnhancer a software package available in Python and R that unites many of the above mentioned metabolome mining and annotation tools independent of what dataset it processes, thus making the algorithm accessible in an easy-to-use format to the community (Figure 1). MolNetEnhancer discovers molecular families (MFs), subfamilies, and subtle structural differences between family members. The workflow enhances both currently available molecular networking methods based on either MS-Cluster [26] (classical) or MZmine2 [27] (also called “feature-based molecular networking”) and results in annotated molecular networks that can be explored in Cytoscape [28]. We applied MolNetEnhancer to publicly available mass spectrometry fragmentation data ranging from marine-sediment and nematode-related bacteria, to *Euphorbia* and Rhamnaceae plants. Illustrated by four case studies, we demonstrate how our integrative workflow discovers dozens of MFs in large-scale metabolomics studies of these plant and bacterial extracts. Moreover, discovered MFs can be divided into subfamilies using the mapped MS2LDA results. Structural annotation of Mass2Motifs is facilitated by having chemical and structural annotations at hand, for example by recognizing substructures in peptidic molecules. We conclude that our workflow provides chemical refinement of metabolomics results beyond spectral matches through large-scale MF and substructure discovery and annotation by integrating outputs of various tools in one place allowing for enhanced visualization. This also guides the metabolomics researcher in prioritizing MFs to explore and in structurally annotating molecules.

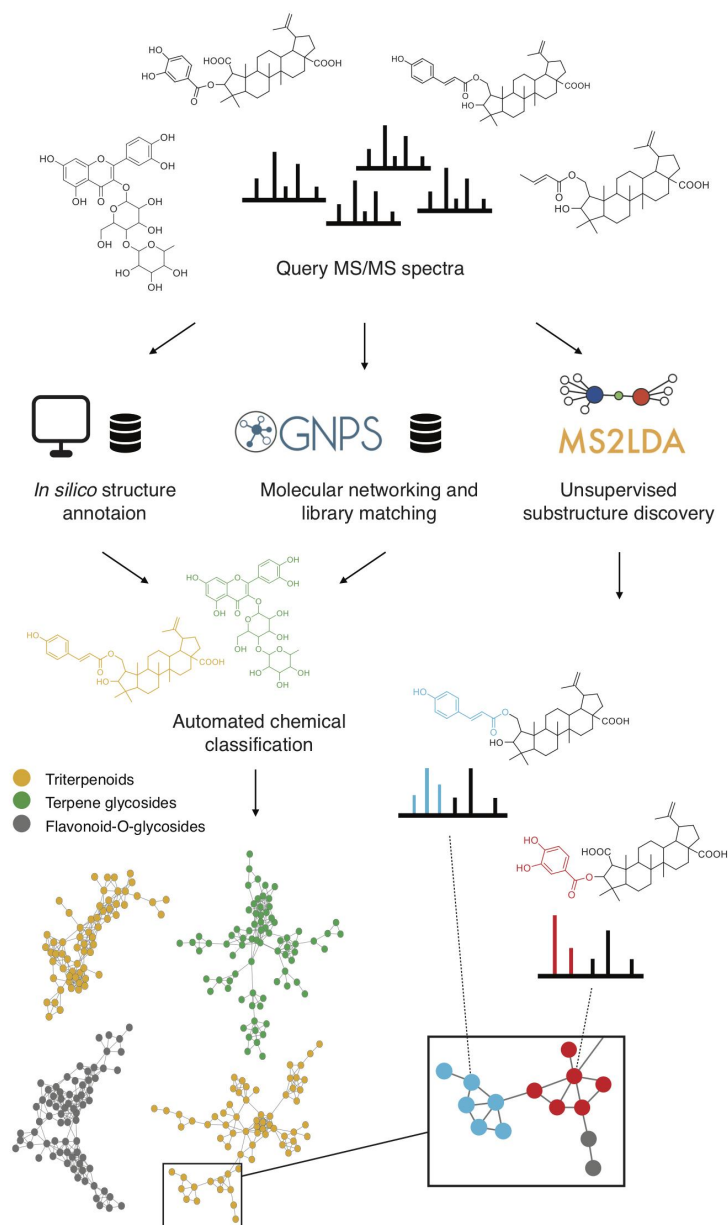


Figure 1. Schematic overview of the MolNetEnhancer workflow. Starting with mass spectrometry data obtained from complex metabolic mixtures mass spectral molecular networks are built by GNPS molecular networking and substructure patterns (Mass2Motifs) are discovered by MS2LDA. These information layers can be mapped on top of each other resulting in more detailed substructure information within molecular families (as exemplified for the organic acid conjugates in the enlarged part of the triterpenoid molecular family on the right). Another route obtains candidate structures for as many nodes as possible through GNPS library matches, or *in silico* annotation such as Network Annotation Propagation (NAP) or DEREPLICATOR and VarQuest for peptidic molecules. All the obtained candidate molecules are fed into the ClassyFire tool to obtain the most abundant annotated chemical classes per molecular family, resulting in a quick overview of the measured chemistry by colouring molecular families according to most occurring chemical class terms in candidate structures of their molecular features as exemplified for three different chemical classes in the molecular network on the left. Structural information obtained through both routes can be combined into one molecular network.

2. Results

2.1. MolNetEnhancer workflow

MolNetEnhancer is a software package available both in R and Python, which enables straightforward integration of mass spectral molecular networks with substructure information, *in silico* structural annotations and chemical classifications and is available at <https://github.com/madeleineernst/pyMolNetEnhancer> and <https://github.com/madeleineernst/RMolNetEnhancer>. MolNetEnhancer consists of two independent steps. During the first step, molecular substructures detectable by co-occurring fragment ions or neutral losses, so called Mass2Motifs, are mapped onto a Molecular Network. Each node in the network represents a molecular feature, whereas Mass2Motifs represent substructural features. Most fragmented mass peaks (precursor ions) represent molecular ions, although fragmented mass peaks may also represent adducts of one and the same molecule, in source fragments or doubly-charged peaks [29]. For simplicity, we will refer to any fragmented mass peak as molecular feature throughout the manuscript. Mass2Motifs contained within each molecular feature can be visualized as pie charts on the nodes. Alternatively, Mass2Motifs shared across multiple molecular features can be visualized as multiple lines (edges) connecting the nodes. In a second step, most abundant chemical classes per molecular family based on candidate structures from *in silico* annotation tools as well as GNPS library matches can be mapped through chemical classification using ClassyFire [Feunang et al., 2016]. A Chemical Classification score is calculated representing what percentage of nodes within a molecular family are attributed to a given chemical class (see Material and Methods section and Figure 8 therein). Mapping Mass2Motifs onto a molecular network is possible through both the Python module as well as the R package, whereas mapping of chemical classes is available through the Python module only. In sections 2.2. to 2.5. we show how MolNetEnhancer can accelerate and enrich chemical information retrieval in 4 case studies, comprising two plant and two bacterial publicly accessible datasets. The MolNetEnhancer workflow results in one graphml network file that contains all the structural information obtained from the individual tools. Such a file can be easily imported into network visualization tools such as Cytoscape [28], an environment where additional metadata on the molecular features can be added. In addition, all structural information is also available as tab delimited text files.

2.2. Case study 1: Annotation of *Euphorbia* specialized metabolites using MolNetEnhancer

With more than 2000 species worldwide, the plant genus *Euphorbia* is among the most species-rich and diverse flowering plants on earth [30,31]. Besides exhibiting an extreme diversity in its growth forms and habitat types, the genus has also attracted interest within natural products drug discovery [32,33]. *Euphorbia* species are chemically highly diverse, particularly within macro- and polycyclic diterpenoids, biosynthetically derived from a head-to-tail cyclization of the tetraprenyl pyrophosphate precursor, which have been found to exhibit a range of biological activities with pharmaceutical interest, such as antitumor, antimicrobial or immunomodulatory activity [Vasas and Hohmann, 2014]. Ingenol mebutate for example, a diterpenoid originally isolated from *Euphorbia peplus* L. is marketed for the topical treatment of actinic keratosis, a precancerous skin condition [34], however production through plant extraction or chemical synthesis is inefficient and expensive [35,36].

A key interest is therefore to find species within the genus producing higher quantities of ingenol mebutate or other close diterpenoid analogs exhibiting biological activities with pharmaceutical interest. We have previously assessed chemical diversity within a representative subset of species of the plant genus *Euphorbia* [Ernst et al., 2018]. A major challenge is the rapid identification of known and unknown *Euphorbia* diterpenoid structures. Using MolNetEnhancer, we were able to significantly accelerate manual annotation of diterpenoids and retrieve chemical structural information, even for molecular families with no structural matches in the GNPS spectral libraries.

An example of how MolNetEnhancer increases chemical structural information throughout two molecular families is highlighted in Figure 2. Using GNPS spectral library matching, chemical structural information for only one molecular feature was obtained, and manual propagation of the annotation throughout molecular family (i) was limited given that the annotated ion exhibited one neighbor only. No structural information could be retrieved for family (ii), where no chemical structural information was retrieved through GNPS library matching (Figure 2a).

Using MolNetEnhancer however, we were able to highlight substructural Mass2Motifs within both molecular families (Figure 2b). Substructural Mass2Motifs, putatively annotated as a *Euphorbia* diterpenoid backbone skeleton with mass peaks at m/z 313, 295, and 285 were found both in molecular families (i) and (ii)

(Figure 2b). Manual annotation of these Mass2Motifs was possible by comparing mass fragments of the library spectrum to mass fragments contained in the Mass2Motifs. A mirror plot comparing the GNPS reference spectrum to the unknown spectrum found in our samples is shown in Supplementary Figure 1. The exact *Euphorbia* backbone skeleton type could not be identified unambiguously, as many *Euphorbia* diterpenoid skeletons are isomeric, and their respective MS² spectra identical or very similar. A *Euphorbia* backbone skeleton with masses at m/z 313, 295, 285 can either result from a jatrophone, deoxy tiglane, or ingenane ester like skeleton [37,38]. Furthermore, we were able to see that molecular family (ii) contains substructural Mass2Motifs related to a nicotinoyl side chain. Manual annotation of these Mass2Motifs was possible by comparing chemical structures retrieved through NAP *in silico* structure annotation with mass fragments found in the Mass2Motifs. Motifs 432 and 180, were both found to contain mass peaks at m/z 106 and 124, possibly resulting from a nicotinoyl side chain and a hydroxylation (Figure 2b). Chemical structures retrieved through *in silico* annotation or library matching can aid the manual annotation of Mass2Motifs and vice versa annotated Mass2Motifs can aid the propagation of chemical structural information throughout the network. Additionally, chemical structural hypotheses can be reinforced by taking into consideration both substructural information as well as chemical class information obtained through *in silico* annotation and library matching. Most chemical structures retrieved for molecular family (i) and (ii) were diterpenoids of the jatrophone, tiglane or ingenane type and substructures related to these *Euphorbia* diterpenoid backbone skeletons were also found within the Mass2Motifs (Figure 2c).

In conclusion, using MolNetEnhancer we were able to significantly increase chemical structural information from retrieving chemical structural information of one molecular feature through GNPS library matching (Figure 2a), to retrieving chemical structural information at an annotation level 3 (putatively characterized compound classes) according to the Metabolomics Standard Initiative's reporting standards [39] of 2 molecular families comprising 73 molecular features (Figure 2b-d). Finally, this information allowed us to conclude that *Euphorbia* diterpenoid skeletons of the jatrophone, deoxy tiglane, or ingenane ester type are found within all *Euphorbia* subgeneric clades, whereas nicotinoyl sidechain modifications are unique to subgenus *Esula* (Figure 2d).

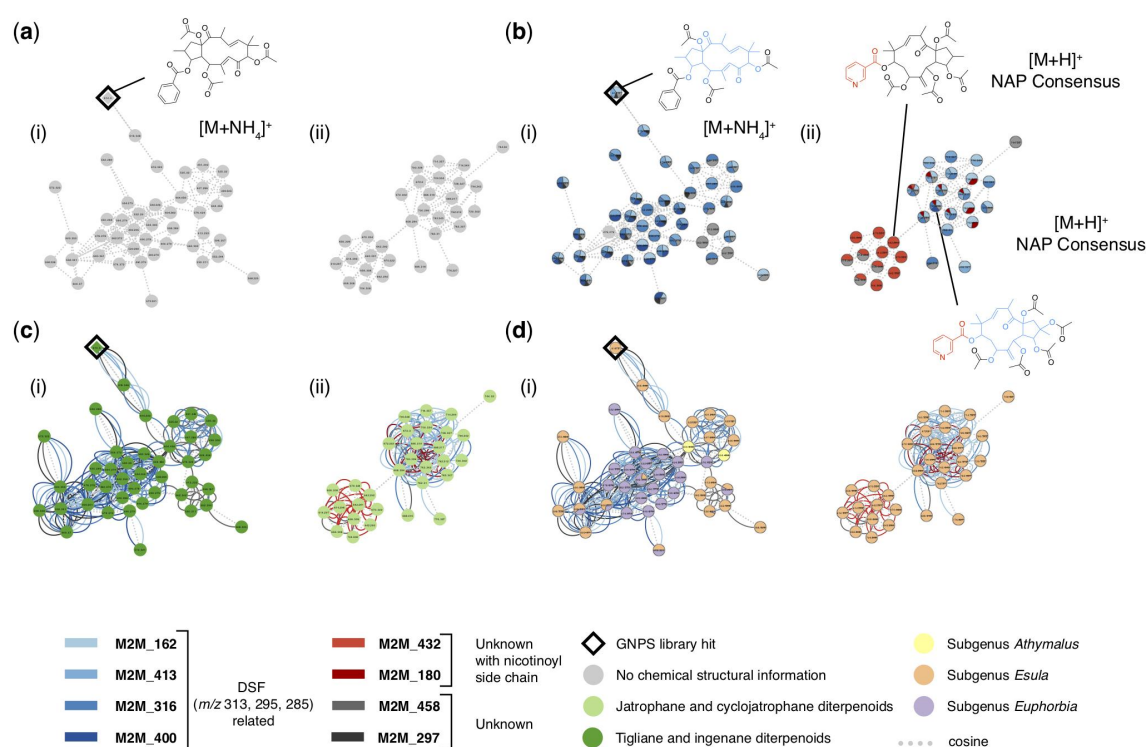


Figure 2. MolNetEnhancer increases chemical structural information obtained for *Euphorbia* specialized metabolites. (a) Mass spectral molecular network showing two molecular families of *Euphorbia* specialized metabolites. Using GNPS library matching only one molecular feature could be putatively annotated, manual annotation propagation is limited for family (i) and none for family (ii). (b) Using MolNetEnhancer,

substructural Mass2Motifs can be visualized within the network; both molecular family (i) and (ii) contain Mass2Motifs related to a *Euphorbia* diterpene spectral fingerprint (DSF) and molecular family (ii) contains Mass2Motifs related to a nicotinoyl side chain. Mass2Motifs are mapped on the nodes as pie charts with an area proportional to their overlap score, a score measuring how much of the Mass2Motif is present in the spectrum, whereas dotted lines connecting the nodes represent features with a MS2 spectral similarity of a cosine score over 0.6 (c) Most chemical structures retrieved for molecular family (i) and (ii) are diterpenoids of the jatrophane, tiglane or ingenane type, which both can result in a DSF with m/z 313, 295, 285. Substructures with mass fragments characteristic of these *Euphorbia* DSFs were also found within the Mass2Motifs. Node colors represent most abundant chemical classes, colored lines connecting the nodes represent shared Mass2Motifs, and dotted lines connecting the nodes represent features with a MS2 spectral similarity of a cosine score over 0.6 (d) *Euphorbia* diterpenoid skeletons of the jatrophane, deoxy tiglane, or ingenane ester type are found within all *Euphorbia* subgeneric clades, whereas nicotinoyl sidechain modifications are unique to subgenus *Esula*. Node colors represent summed peak area per *Euphorbia* subgeneric clade, colored lines connecting the nodes represent shared Mass2Motifs, and dotted lines connecting the nodes represent features with a MS2 spectral similarity of a cosine score over 0.6.

2.3. Case study 2: Annotation of Rhamnaceae specialized metabolites

Another case where we demonstrate the efficiency of MolNetEnhancer for enhancing the chemical annotation of metabolomics data is our previous study on the plant family Rhamnaceae [25]. Rhamnaceae is a cosmopolitan family including about 900 species, and Rhamnaceae species are known for their exceptional morphological and genetic diversity, which are thought to be caused by the wide geographic distribution and different habitats [40]. We applied an MS2-based untargeted metabolomics approach to get insights on the metabolomic diversity of this highly-diversified family, and MolNetEnhancer was used as a key to provide fundamental annotations for MS2 spectra.

As shown in Figure 3a, MolNetEnhancer provided the putative chemical classification of molecular families within the Rhamnaceae molecular network. After combining this chemical class annotations with taxonomic information of each molecular feature, the normalized distribution pattern of different classes of metabolites were analyzed. This revealed that the taxonomic clade Rhamnoid exhibits more diversified flavonoids, carbohydrate, and anthraquinones, while the Ziziphoid clade produces various triterpenoids and triterpenoid glycosides [25].

MolNetEnhancer allowed us to visualize and discover the subtle substructural diversity within the molecular families. In the molecular family of triterpenoid esters, for example, substructural differences of phenolic moieties such as protocatechuate, vanillate, and coumarate were easily recognized by analyzing the distribution of Mass2Motifs 28, 117, 120, and 191 (Figure 3b). Two flavonoid aglycone substructures, kaempferol and quercetin, were also distinguished by analyzing the distribution of Mass2Motifs 86, 130, and 149 in the molecular family of flavone 3-*O*-glycosides (Figure 3c). Mass2Motif 130 contained mass peaks at m/z 284, 255, and 227, while Mass2Motifs 86 and 149 covered mass peaks at m/z 300, 271, and 255. These fragment ions are well-known as characteristic fragments of kaempferol 3-*O*-glycosides and quercetin 3-*O*-glycosides [41–43], so these Mass2Motifs could be easily annotated. This case study shows how MolNetEnhancer facilitates the interpretation process and our knowledge on MS2 fragmentation, previously mainly applied manually by experts.

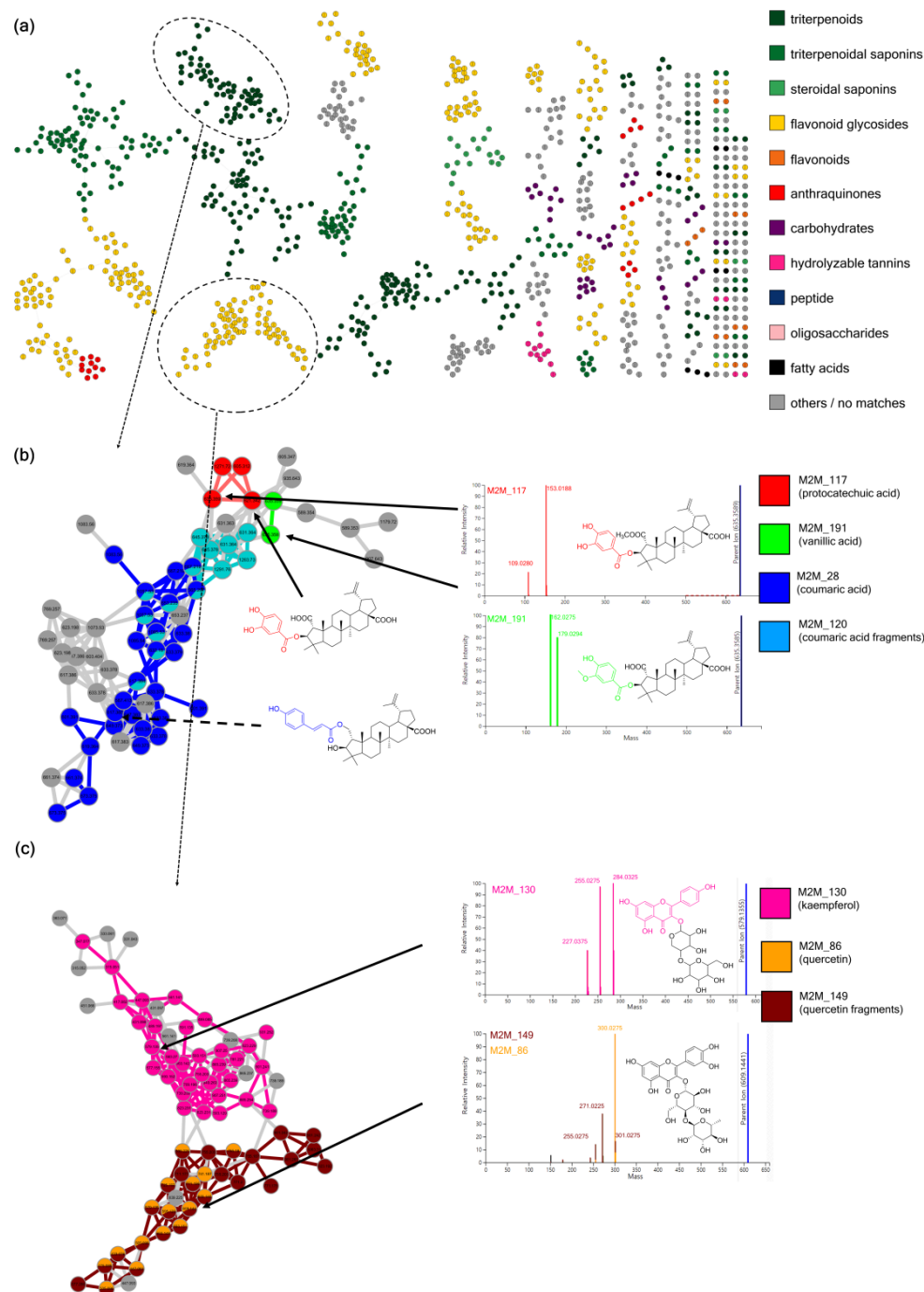


Figure 3. MolNetEnhancer increases chemical structural information obtained for Rhamnaceae specialized metabolites. **(a)** Structural annotation for molecular families was suggested based on consensus-based classification of NAP *in silico* structure annotation. **(b)** Subtle chemical differences of phenolic acid moieties can be visualized within the molecular family of triterpenoid esters based on Mass2Motifs. **(c)** Molecular family annotated as flavonoid glycosides reveals two subfamilies by Mass2Motif mapping: the pink Mass2Motif is related to the kaempferol core structure, whereas the orange and brown Mass2Motifs are related to the quercetin core structure - two related yet distinct flavonoid structures.

2.4. Case study 3: Large chemical diversity uncovered by annotating specialized metabolites in marine sediment *Streptomyces* and *Salinispora* bacterial extracts

The MolNetEnhancer workflow was also applied to bacterial data sets to gain more detailed insights into their chemical richness. Crüsemann and coworkers created a molecular network of extracts of the marine sediment bacteria *Salinispora* and *Streptomyces* that formed the basis for this case study [44]. Figure 4 displays the molecular network coloured by the most prevalent chemical class annotations. Whilst we can observe that the bacteria also produce a structurally diverse arsenal of molecules, its composition is clearly different from that of the Rhamnaceae plants in Fig. 3a. The most prevalent chemical class annotations are “Carboxylic acid and derivatives” and “Prenol lipids” with the first containing peptide-related molecules and the latter containing terpenoid molecules. Both these classes of molecules are known to be produced by *Salinispora* and *Streptomyces* bacteria. The Chemical Classification Scores (see Methods section) for the ClassyFire class and kingdom terms are presented in Supporting Figure S2. These scores aid in assessing chemical novelty and also provide information on the consistency of the chemical class annotations of the structural candidates.

From the 5930 network nodes, we discovered 300 Mass2Motifs using MS2LDA. From those, we could annotate 40 with structural information at various levels of structural details gained from spectral matching with the GNPS libraries, or from the *in silico* annotation tools NAP, DEREPLICATOR, and VarQuest. For example, we could annotate an aminosugar-related Mass2Motif with fragment ions related to two known N,N-dimethyl amino sugars present in known specialized molecules from the bacteria under study [44]: dimethylamino-β-D-xylo-hexopyranoside (rosamicin) and N,N-dimethyl-pyrrolosamine (lomaiviticin) which have overlapping fragment ions and are therefore characterized by the same Mass2Motif. With a frequency of more than 70 throughout the entire molecular network (using probability and overlap score thresholds of 0.1 and 0.3, respectively, for the molecular feature - Mass2Motif connections), the aminosugar Mass2Motif can be used as a handle to identify known and potential novel natural products throughout network. Indeed, the Mass2Motif was found in all members of the Rosamicin MF (Figure 5A) and the Lomaiviticin MF (Supporting Fig. S3-A). Moreover, the same aminosugar-related Mass2Motif was also found in all members of two yet unknown MFs (Figure 5B, Supporting Fig. S3-B). In addition, the Mass2Motif was also found in a number of singletons not connected to any MF, often in combination with Mass2Motif 66 as well like we see for the rosamicin-related MF. Mass2Motif 66 represents the presence of an *m/z* 116 fragment which is likely also generated by the dimethylated amino sugar; in fact it may point to the dimethylamino-β-D-xylo-hexopyranoside moiety or something very similar as this fragment is absent in spectra from the lomaiviticin MF which contains the different dimethylated aminosugar N,N-dimethyl-pyrrolosamine. In most singletons, no other Mass2Motifs were discovered that could provide clues on the complete structures of these molecules; however, given the presence of the aminosugar moiety they are likely natural products and not core metabolites or contaminants - something that we could not confidently state without using the MolNetEnhancer workflow.

Another MF displayed in Figure 5C did not return any GNPS library hits; however, all its members shared Mass2Motif 154. Due to its indicative fragment ions, we could annotate this Mass2Motif as tryptophan-related, indicating that all these molecules contain a tryptophan core structure. Based on their shared Mass2Motif, the masses of the molecular features, and their fragmentation patterns, with help of MolNetEnhancer we could now tentatively annotate this MF as tryptophan-related containing molecules such as small peptides or N-acyltryptophans. Figure 5D shows the peptidic MF of actinomycin-related molecules. The annotation of this MF was guided by DEREPLICATOR and VarQuest annotations as well as the Mass2Motif that 10 of its members shared. We could annotate this Mass2Motif as the amino acid lactone loop present twice in actinomycins using reference data from literature [45]. The unique combination of four actinomycin-related mass fragments was only present in the 10 MF members, thereby reinforcing the DEREPLICATOR and VarQuest annotations.

Furthermore, mapping the Mass2Motifs on the molecular network means that we can more easily track neutral loss-based motifs such as the loss of an acetyloxy group that was only found in *Streptomyces* MFs. Moreover, inspection of the MFs without annotated chemical classes revealed that they contained some Mass2Motifs with relatively low frequency throughout the data set - something that could point to a unique substructure or scaffold possibly from a unique biosynthesis enzymatic function. For example, Mass2Motif 35 has a frequency of 43 and was present in all four members of the MF in Supporting Fig. S3-C. It is a mass-fragment-based Mass2Motif and with masses of 142, 100, and 58 Da it could be related to a polyamine-like structural feature. Finally, the MF in Supporting Fig. S3-D shares the two still unknown loss-based Mass2Motifs 250 and 261 that have frequencies of 26 and 50, respectively. These are examples of Mass2Motifs representing potential novel chemistry that can now be easily tracked in the molecular network.

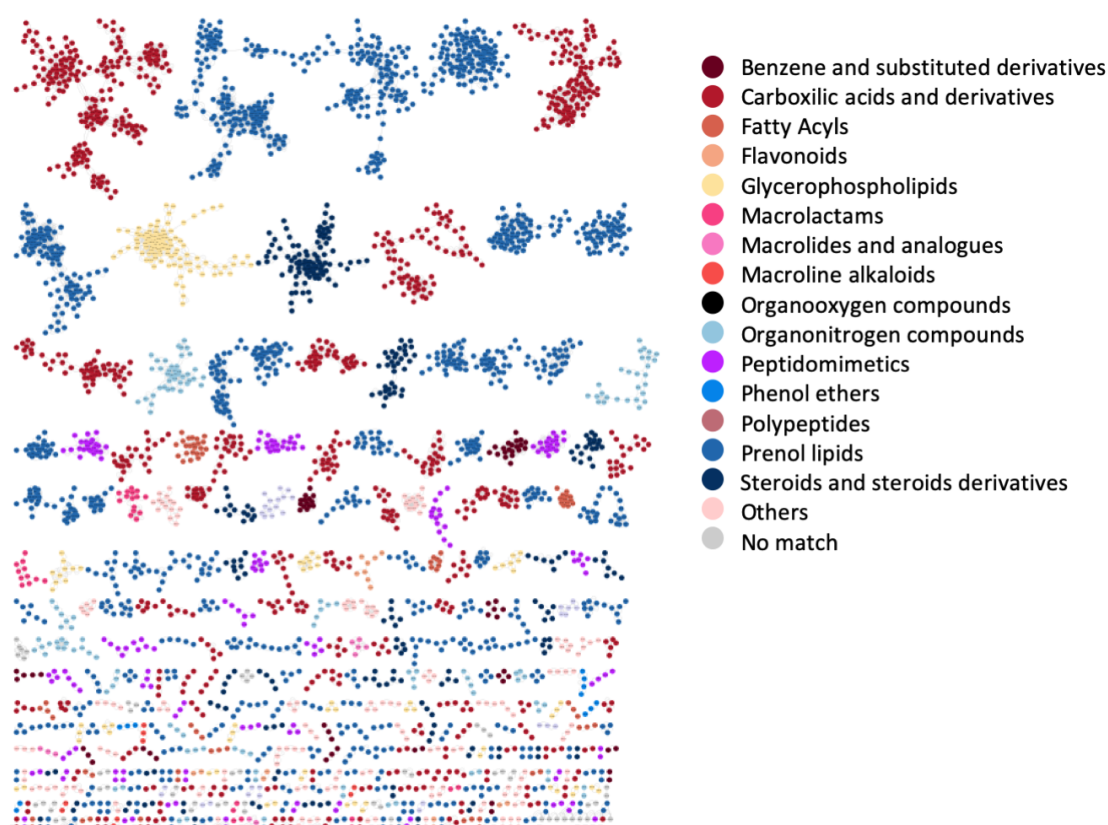


Figure 4. Marine sediment *Salinispora/Streptomyces* molecular network colored by 15 selected chemical class terms as indicated in the legend. In total, 50 different class terms were annotated in the network using MolNetEnhancer, indicating that the metabolic output of the *Salinispora/Streptomyces* strains is chemically very diverse. We can observe that the larger molecular families are mostly annotated with prenol lipids (blue) and carboxylic acids and derivatives (red). Furthermore, for a couple of MFs no chemical class annotations were obtained as no candidate structures were retrieved through any of the annotation tools.

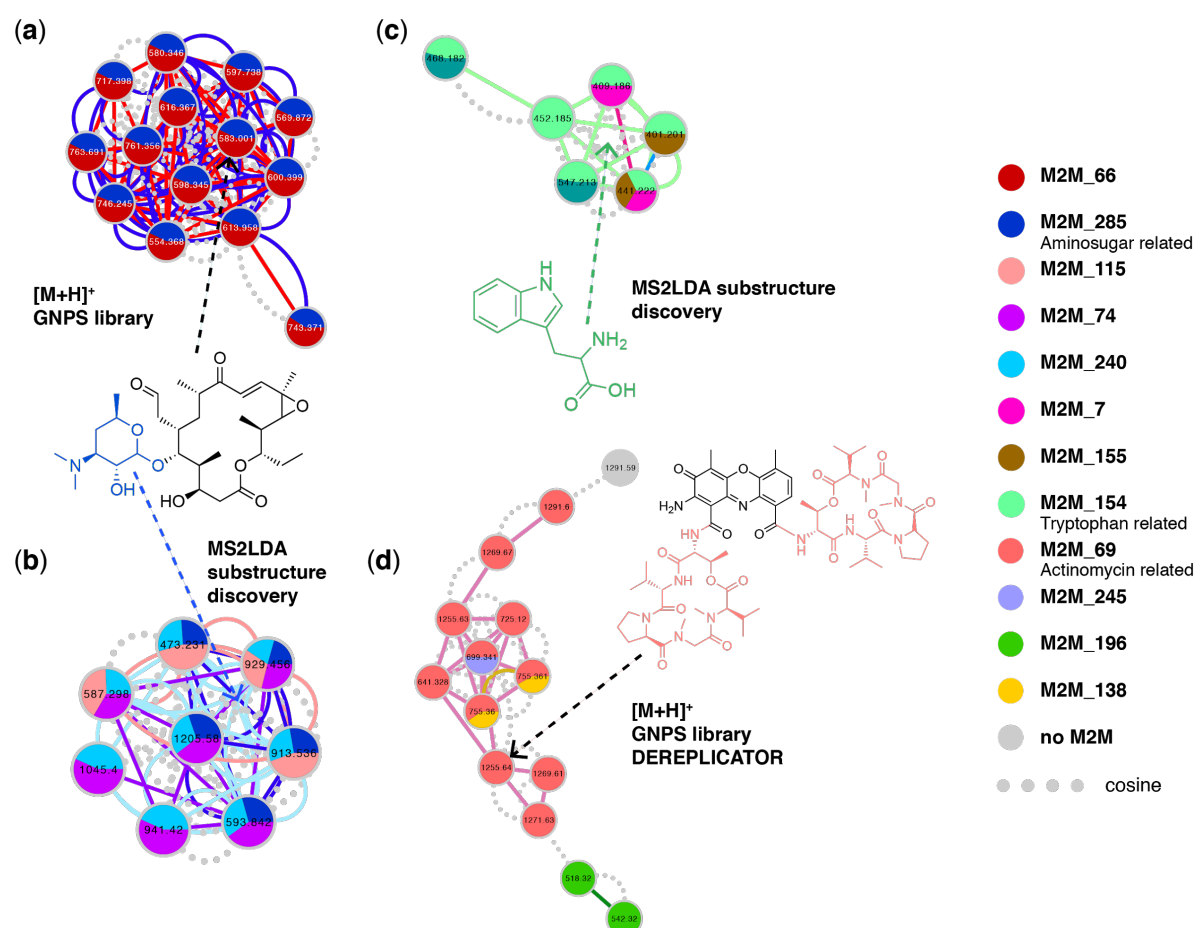


Figure 5. Molecular families from marine sediment bacteria with color coded Mass2Motif substructure information mapped on them, with **(a)** Rosamicin-related molecular family found through GNPS library hits where all members contain an amino sugar-related motif as coloured in blue in its depicted structure - substructures or motifs found within each molecular feature are mapped on the nodes as pie charts, where the relative abundance of each motif represents the overlap score, a score measuring how much of the motif is present in the spectrum. Furthermore, motifs shared between two nodes are visualized as coloured continuous lines (edges) connecting the nodes whereas dashed lines (edges) represent a cosine score of over 0.6, **(b)** Yet unknown molecular family that shares an amino sugar-related motif connecting this MF to **(a)** by sharing a substructure, **(c)** Tryptophan-related molecular family sharing the Tryptophan Mass2Motif, and **(d)** Actinomycin-related molecular family - found through GNPS library hits and further validated with help of DEREPLICATOR results - sharing an Actinomycin related motif across most of its members. The Actinomycin D (Daptomycin) structure is depicted with the Mass2Motif substructure highlighted in colour: the peptide lactone ring present twice in the molecule. In all MFs, nodes are coloured based on Mass2Motif overlap scores and the edges show if cosine score-connected nodes share similar Mass2Motifs. It can be seen that in all families multiple motifs are shared across some of its members.

2.5. Case study 4: Annotating peptidic motifs in peptide-rich *Xenorhabdus*/*Photorhabdus* extracts

Xenorhabdus and *Photorhabdus* are Gammaproteobacteria that live in symbiotic association with soil-dwelling nematodes of the genus *Steinernema* [46,47]. Eventually as a consequence thereof, they spend a large amount of their resources to the production of specialized metabolites, in particular non-ribosomal peptides and polyketides. Tobias and coworkers recently published metabolomics data of 25 *Xenorhabdus* and 5 *Photorhabdus* strains to explore metabolic diversity amongst these strains [46]. Here, we applied MolNetEnhancer on this publicly available molecular networking data to further probe the chemical diversity previously found. The 6228 network nodes were analysed with MS2LDA to discover 300 Mass2Motifs. Furthermore, we also submitted the *Xenorhabdus*/*Photorhabdus* molecular networking data to NAP, DEREPLICATOR, and VarQuest to run the MF chemical class annotation pipeline. By far the majority of the 46 annotated motifs were peptide, amino acid, or likely to be peptidic-related which fits with the ClassyFire predicted peptide-related MFs present in the *Xenorhabdus*/*Photorhabdus* extracts with “Carboxylic acids and derivatives” and “Peptidomimetics” as most frequently occurring annotations (see Figure 6 - with corresponding Chemical Classification Scores in Figure S4). We could also annotate an indole-related Mass2Motif which can be part of peptides/amino acids. An exception is the ethylphenyl-related Mass2Motif that was found in 478 molecules (out of 6228 nodes, corresponding to 7.7%) of the *Xenorhabdus*/*Photorhabdus* extracts. This can be explained by the reported production of phenylethylamides, dialkylresorcinols, and cyclohexadien derivatives by the studied strains [48].

Annotations included Mass2Motifs that form peptidic substructures related to well-known *Xenorhabdus* peptidic families such as the commonly found bioactive rhabdopeptides and the related xenorhides [48,49]. We could annotate two rhabdopeptide-related motifs with frequencies of 231 and 186 (3.7% and 3.0% of nodes, respectively). Compared to the structurally less diverse xentrivalpeptides [50] which the Mass2Motif had a frequency of 28, corresponding to 0.45% of the nodes, we can conclude that rhabdopeptide-related molecules are widespread in the *Xenorhabdus*/*Photorhabdus* extracts. The PAX peptides constitute another well-known *Xenorhabdus*/*Photorhabdus* lysine-rich peptide class [51]. The corresponding MF consisted of 13 members; indeed, they shared a Mass2Motif related to lysine (lys) and lys-lys fragments. Similarly, a leucine-leucine Mass2Motif was found in molecules annotated as xenobovoid. This motif occurred in 110/6228 (1.8%) nodes pointing to several peptidic families that contain this amino acid motif - this in contrast to the lys-lys amino acid motif that is very wide-spread in *Xenorhabdus*/*Photorhabdus* molecules, being present in 1500 (24%) nodes. In total, using the MolNetEnhancer workflow we could annotate 32 peptidic motifs of which we could link 11 to peptides known to be produced by *Xenorhabdus*/*Photorhabdus* strains whilst the other 21 Mass2Motifs represent substructures not yet elucidated. The peptidic nature of these Mass2Motifs was assessed by recognition of typical fragment ion patterns as seen for known peptides as well as doubly charged precursor ions that are often a sign of peptides in these extracts.

With the help of the integrative display of DEREPLICATOR and VarQuest annotation results, we could also annotate two xenoamicin-related peptidic MFs (Figure 7 A-B). Xenoamicins are known to be produced by *Xenorhabdus* and eight variants have been described in detail with variants A and B present in peptidic databases [52]. Xenoamicin is a cyclic peptide consisting of a peptidic ring and peptidic tail (see Figure 7D). Interestingly, in one of the annotated MFs, not one but two Mass2Motifs were shared between most of its members (see Figure 7A). With help of DEREPLICATOR-predicted annotations of the fragment ions, we could annotate the Mass2Motif shared by almost the entire MF as being related to the xenoamicin A peptidic ring, whereas the other more abundant Mass2Motif was related to the xenoamicin peptidic tail (Figure 7C, and Supplementary Figure S5 A-B). These Mass2Motifs are quite specific as we observed that 9 and 6 mass fragments, respectively, were consistently present in more than 75% of the molecular features to which the ring and tail Mass2Motifs were linked. A third Mass2Motif could be putatively annotated as xenoamicin B peptidic ring-related as its masses are +14 Da as compared to the ring A motif and xenoamicin B differs from A with an isobutyl replacing an isopropyl group. Based on the Mass2Motif presence/absence analysis in the larger MF of 32 members, we observe that 4 have links (overlap score > 0.3) to both ring A and tail motifs, 10 just have the ring A motif, 3 have only links to the peptidic tail motif, 2 share both ring A and putative ring B together with the tail Mass2Motif, and 2 share the putative ring B with the tail Mass2Motif (Figure 7A). Thus, this indicates how MolNetEnhancer increases the resolution in molecular networks by highlighting structural differences in between MF members.

We could also find additional MFs and singletons in which the xenoamicin ring or tail Mass2Motif was present, pointing to related peptidic molecules not linked through the modified cosine score. Further inspection with help of VarQuest annotations strengthened these annotations as VarQuest annotated modified amino acids in both rings (Figure 7, Supplementary Figure S5 E-F) and the tail region (Supplementary Figure S5 C-D) of xenoamicin many of which, to our knowledge, have not been reported yet such as the one highlighted

in Figure 7D where the ring-proline is likely methylated (the ring A motif is not linked to this molecular feature). In fact, xenoamicin A was annotated as variant from xenoamicin B (Supplementary Figure S5-F) where the modified amino acid (demethylation) corresponds to previous literature findings [52], further increasing our trust in these *in-silico* approaches. The smaller MF of 22 nodes consisted of doubly-charged precursor ions where no ring-related Mass2Motifs were assigned. Some members like xenoamicin A appeared in both MFs as singly and doubly charged precursor ions; the differences in motif distributions between the two MFs indicates that the initial charge has an impact on the fragmentation pathways and thus the acquired spectra given that we know the ring A is part of xenoamicin A.

Altogether, this example highlights how the MolNetEnhancer approach facilitates fragmentation based metabolomics analysis workflows by increasing the “structural resolution”, the discovery of more xenoamicin variants than previously described, and highlighting previously unseen connections between MFs and molecules. Furthermore, the integrative approach enabled straightforward annotation of Mass2Motifs found in the xenoamicin MF by using the VarQuest fragment ion annotations as guide for Mass2Motif feature annotation. Both Mass2Motif and VarQuest results strengthened each other since when predicted amino acid changes occurred in the peptidic ring, the corresponding ring-related Mass2Motif was absent, and vice versa - made possible by combining the outputs of several *in silico* tools together.

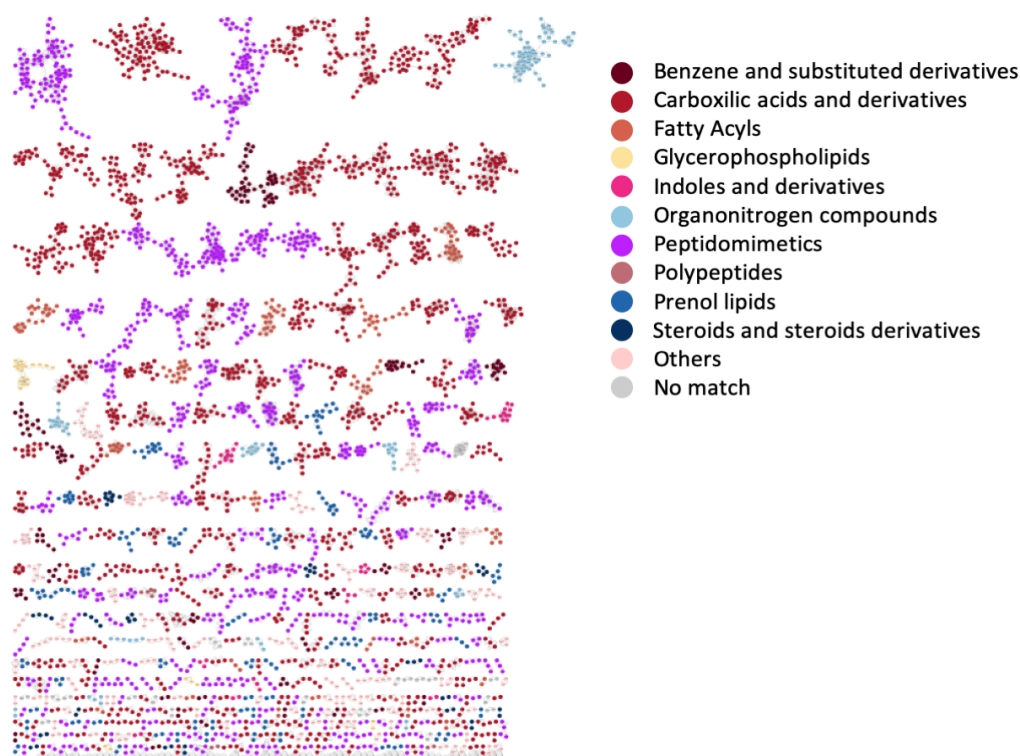


Figure 6. Nematode symbionts *Photorhabdus/Xenorhabdus* network colored by 10 selected chemical class terms as indicated in the legend. In total, 49 different class terms were annotated in the network using MolNetEnhancer. We can observe that the larger molecular families as well as many smaller molecular families are mostly annotated with peptidomimetics (purple) and carboxylic acids and derivatives (red). This is consistent with earlier findings that these nematode symbionts produce a wide array of peptidic products.

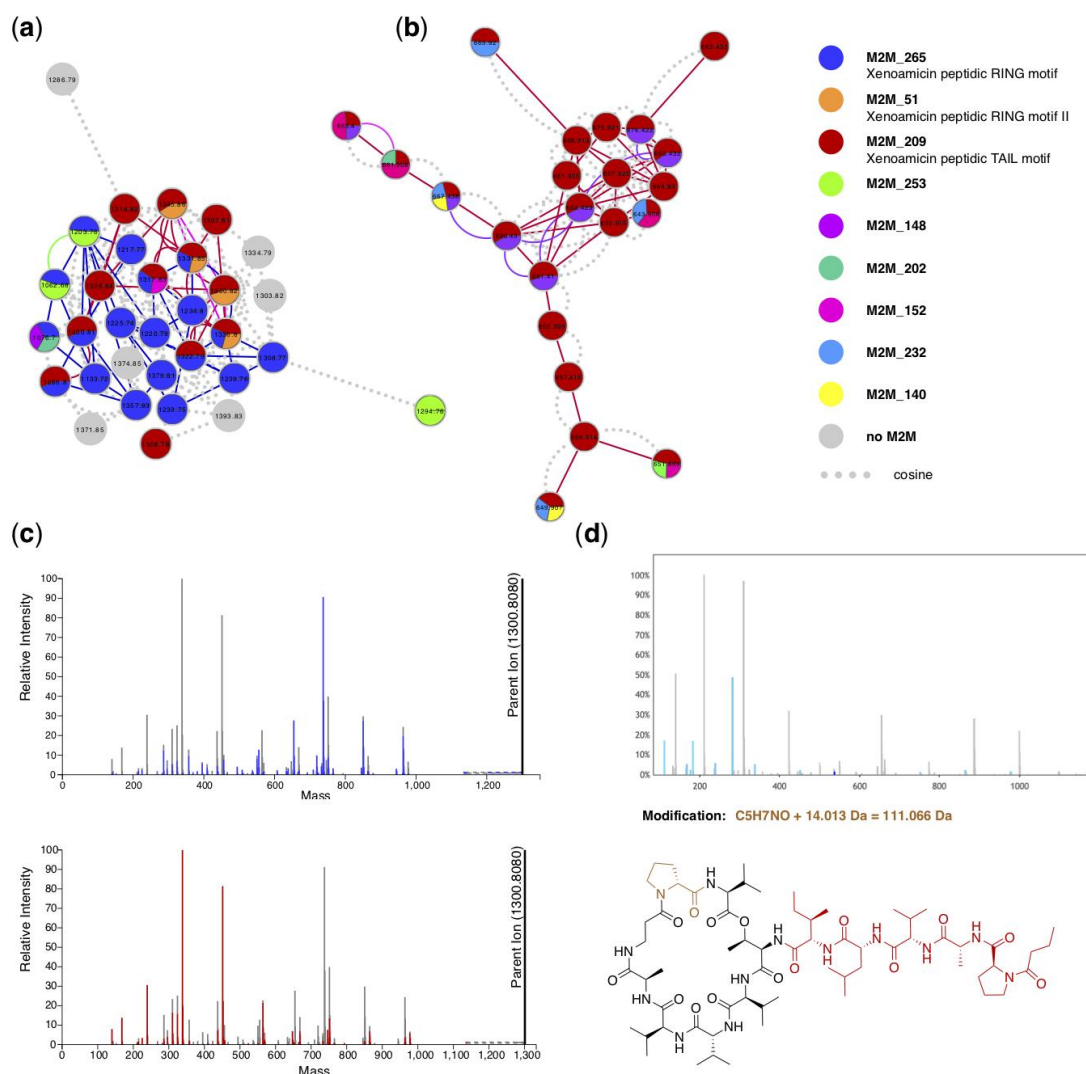


Figure 7. Xenoamicin-related molecular families annotated by MolNetEnhancer with **(a)** MF of 32 nodes of which 23 were annotated with at least one xenoamicin modified structure (xenoamicin A or B) by either VarQuest or DEREPLICATOR with VarQuest using 0.005 Da fragment binning assigning most xenoamicin structures (FDRs mostly < 2.5). This MF also contains nodes sharing all Mass2Motifs related to xenoamicin structures with two ring and tail-related Mass2Motif. Mass2Motif 265 contains mass fragments related to xenoamicin A, whereas masses in Mass2Motif 51 are shifted with 14 Da pointing towards xenoamicin B. The MF consists of singly charged molecular features. **(b)** Related MF of which 20 out of 22 nodes were annotated with xenoamicin modified structures (FDRs mostly < 2.5). This MF only shares the Mass2Motif annotated as xenoamicin tail-related and consists of doubly-charged precursor ions. **(c)** Xenoamicin A spectrum in the ms1da.org environment with (top) ring-related Mass2Motif highlighted and (bottom) tail-related Mass2Motif highlighted with the corresponding blue and red colors as in (a) and (b). **(d)** VarQuest annotation of xenoamicin modified peptide where a ring proline indicated in brown is likely methylated. All light blue peaks in the mass spectrum were annotated by VarQuest. The red part in the xenoamicin structure corresponds to the selected fragment of m/z 537.348 which includes the tail part, whereas the light blue amino acid is annotated to be modified with a mass shift of 14.013 Da that likely corresponds to a methylation. Indeed, the Mass2Motif related to the xenoamicin tail is found in this fragmentation spectrum, whereas the ring Mass2Motif is absent.

3. Discussion

Although significant advances have been made in molecular mining workflows, chemical annotation as well as classification tools [1–4,7,8,10,14–16], chemical structure annotation remains the major and most challenging bottleneck in mass spectrometry-based metabolomics as most of our biological interpretations rely on annotated structures [8,53,54]. MolNetEnhancer is a workflow that combines chemical structural information retrieved from different *in silico* tools, thus increasing structural information retrieved and enhancing biological interpretation. Here, we have chosen a representative number of *in silico* tools covering mining, annotation, and chemical annotation to provide the user with different chemical insights. Although we used DEREPLICATOR and NAP to exemplify *in silico* annotation tools here, MolNetEnhancer is platform independent, meaning that chemical structures retrieved from any *in silico* annotation platform could be used given the molecular feature identities correspond across all molecular mining and annotation tools.

Particularly in natural products research, the rapid annotation of known (i.e., dereplication) as well as unknown specialized metabolites from complex metabolic mixtures hinders interpretation in an ecological, agricultural or pharmaceutical context. Many specialized metabolites from natural sources are used as pharmaceuticals [55], in agriculture [56] or nutrition [57]; however, their discovery is inherently slow due to the above-mentioned limitations. To highlight how MolNetEnhancer can accelerate chemical structural annotation in complex metabolic mixtures from natural sources, we exemplified its use on four plant and bacterial datasets.

In the plant genus *Euphorbia*, we were able to retrieve chemical structural information of previously described pharmaceutically highly valuable diterpenoid skeletons corresponding to an annotation level 3 according to the Metabolomics Standard Initiative's reporting standards [39]. The use of different tools combined in one data format with MolNetEnhancer allowed both for the retrieval of complementary information as well as the reinforcement of putative annotations, in cases where two independent tools pointed to the same chemical structural conclusion. Used separately, none of the tools were able to retrieve as much chemical structural information as when combined in MolNetEnhancer. Likewise, MolNetEnhancer allowed for the annotation of triterpenoids chemistries with several distinct phenolic acid modifications (e.g., vanillate, protocatechuate) in the plant family Rhamnaceae. In *Salinispora* and *Streptomyces* bacterial extracts, MolNetEnhancer aided the annotation of a previously unreported tryptophan-based MF, and a xenoamycin-related MF in the Gammaproteobacteria of the genus *Xenorhabdus* and *Photorhabdus* could be studied in more detail than in previous studies.

It is of utmost importance to note that results retrieved from MolNetEnhancer summarize results retrieved from third-party software and manual inspection and validation of all structural hypotheses remain essential. However, MolNetEnhancer significantly aids the manual inspection and validation process conducted by the expert, by making substructural as well as chemical class information readily available and visible within one data resource. As exemplified in the case studies, MolNetEnhancer can for example help in prioritizing molecular families within a molecular network, which consists of many hundred to thousands of molecular features, be it by highlighting different chemical classes of interest or molecular families, for which only very few structural hypotheses could be retrieved, potentially highlighting novel chemistry.

Limitations introduced through data acquisition on different mass spectrometric instrument types do also apply to MolNetEnhancer. Acquiring data on different instruments can cause different MS2 fragmentation patterns, thus in some cases leading to different structural hypotheses through library matching or *in silico* structure prediction [58]. Also, the presence of low quality and/or chimeric MS2 spectra is a challenge for mass spectrometry annotation tools as the one described here, and methods that are capable of filtering-out these spectra before proceeding with *in silico* annotation tools will improve our confidence in *in silico* spectral annotation [59].

These limitations highlight the importance of good practices during data acquisition and processing to minimize the time spending analyzing mass spectrometry artefacts and improving the confidence in any downstream annotations. Here, the use of feature-based molecular networking could also help to focus the analysis on those molecular features that are very likely molecular ions [60] - and it has the added benefit that MS1 differential abundance information from LC-MS peak picking is available on the molecular features as well.

Apart from limitations caused by experimental conditions, analysis bias can be introduced for structural predictions based on chemical structures available in public databases, which are still limited especially for particular compound classes. This is in particular true for the chemical class annotations provided through

ClassyFire, which rely on collecting correct or structurally closely related candidate structures from compound databases. The chemical annotation score was implemented to guide the researcher in assessing how consistent the chemical annotations are and for how many molecular features at least one candidate structure is found. The peptidic annotations by DEREPLICATOR and VarQuest come with scores, p-values, and false discovery rates to assess confidence in the annotations. Using MolNetEnhancer, it is now also possible to explore the consistency in peptidic annotations within MFs, along with their associated Mass2Motifs, which also assist in improving confidence in the annotations, as we have shown for the xenoamicin MFs in the nematode symbiont bacteria where the majority of the MFs were annotated with xenoamicin variants.

One limitation of the use of MS2LDA on the bacterial datasets is that most non-cyclic peptidic molecular families do not share any motifs as typically analogues differ by modifications such as methylation or hydroxylation causing a shift in m/z in most of their mass fragment peaks. Incorporation of amino acid-related mass differences as features for MS2LDA could be a route to also discover Mass2Motifs for non-cyclic peptides. As it is, cyclic peptides do often contain one or more Mass2Motifs and peptides containing positively charged amino acids such as lysine and leucine have this structural information represented by Mass2Motifs. Furthermore, many Mass2Motifs are currently still unannotated, which hampers fast structural analysis. To partially solve this bottleneck, MotifDB (www.ms2lda.org/motifdb) was recently introduced [61] and the here annotated Mass2Motif sets from the four case studies are made available through MotifDB for matching against Mass2Motifs found in other MS2LDA experiments. Furthermore, this will allow to use a combination of “supervised” (annotated) Mass2Motifs and “unsupervised” (free) Mass2Motifs in future MS2LDA experiments on data of related samples thereby accelerating structural annotation since part of the motifs already discovered do not need to be re-annotated.

Despite the limitations discussed above, MolNetEnhancer assists in metabolite annotations by its combined analysis of chemical class annotations, structural annotations, and Mass2Motif annotations. If these annotations support each other, as for example for the actinomycin MF in the marine sediment bacteria, there is more confidence that these *in silico* annotations will indeed be correct. It is noteworthy that the modularity of MolNetEnhancer allows for complementary sources of structural information to be added on in future. We showed that MolNetEnhancer is a practical tool to annotate the chemical space of complex metabolic mixtures using a panel of complementary *in silico* annotation tools for mass-spectrometry based metabolomics experiments. Although we have highlighted the use of MolNetEnhancer using two plant and bacterial datasets, MolNetEnhancer is sample-type-independent and may be used for any mass-spectrometry-based metabolomics experiment, where chemical structural annotation and interpretation is of interest. Future work will focus on making the MolNetEnhancer workflow available within the GNPS platform in order to further increase its user-friendliness. Furthermore, the integration of other existing and future metabolome mining and annotation tools in the output of MolNetEnhancer is also planned to extend on the initial set of *in silico* tools that it currently can combine.

4. Materials and Methods

Currently, two distinct methods from raw data to MNs exist. One method takes all MS2 spectra found in the input files and uses MS-Cluster to prepare a set of representative “consensus” MS2 spectra for molecular networking, and the other method uses MZmine2 for data preprocessing, which performs molecular feature detection at the MS1 level and associates each MS1 feature with its respective MS2 spectra to send off to GNPS Molecular Networking. The here proposed MolNetEnhancer workflow can enrich both these molecular networking methods with Mass2Motif presence and chemical class annotations.

The MolNetEnhancer workflow comprises the following main steps:

1. Perform mass spectral molecular networking analysis through the Global Natural Products Social Molecular Networking platform (<https://gnps.ucsd.edu/>)
2. Perform unsupervised substructure discovery using MS2LDA (<http://ms2lda.org/>)
3. Perform *in silico* chemical structural annotation using for example Network Annotation Propagation (NAP) and DEREPLICATOR through the GNPS platform. Alternatively, other *in silico* tools for putative chemical structural annotation (e.g. SIRIUS+CSI:FingerID) [Dührkop et al., 2015, 2019] can also be used.
4. Run MolNetEnhancer to:
 - a. Combine substructure information retrieved through MS2LDA with mass spectral molecular network created through GNPS.

- b. Retrieve most abundant chemical classes per molecular family based on GNPS structural library hits and *in silico* chemical structural annotation and integrate this information with mass spectral molecular network created through GNPS.

5. Visualize enhanced mass spectral molecular network in Cytoscape.

A step by step tutorial on how to use MolNetEnhancer can be accessed at

<https://github.com/madeleineernst/RMolNetEnhancer> and
<https://github.com/madeleineernst/pyMolNetEnhancer>.

Substructural information retrieved through MS2LDA is integrated in two ways within the mass spectral molecular networks. Shared substructures or motifs between two molecular features are visualized as multiple edges connecting the nodes. Furthermore, motifs found within a molecular feature can be visualized as pie charts, where the relative abundance of each motif represents the overlap score, a score measuring how much of the motif is present in the spectrum [62]. Furthermore, for each molecular family, the x most shared motifs are shown, where x is defined by the user. An example of such a molecular family with motifs mapped is shown in Figure 5.

To retrieve the most abundant chemical classes per molecular family, all chemical structures obtained through GNPS library matching, and *in silico* chemical structural annotation are submitted to automated chemical classification and taxonomy structure using ClassyFire [16]. This retrieves chemical classes for each of the putative structures submitted organized in 5 hierarchical levels of a chemical taxonomy (kingdom, superclass, class, subclass, direct parent). For each level of the chemical ontology, a score is calculated, which represents the most abundant chemical class found for the structural matches within the molecular family at each hierarchical level. It is important to note that a high score does not represent a higher confidence in the true identity of the chemical structures found within the molecular family, but indicates more consistency as more structural matches obtained for this molecular family fall within the same chemical class. Figure 4 exemplifies how the score is calculated. Given a molecular family consisting of 6 molecular features (nodes), the percentage of nodes classified as cinnamaldehydes, coumarins and derivatives, flavonoids and macrolactams at the chemical class level respectively is calculated. Each molecular feature can have multiple structural matches with multiple (e.g. node 2) or identical (e.g. node 3) chemical classes. A majority of the structural matches obtained in the network shown in Figure 4 were classified as flavonoids (2.25 out of 6 nodes), thus the molecular family was classified as flavonoids with a chemical classification score at the class level of 0.375 (2.25/6). For single nodes (molecular features which show no spectral similarity with any other molecular features found in the dataset) the chemical classes are retrieved analogously, however, it should be noted that single nodes often result in a very high score, as only one structural match is retrieved, resulting in a score of 1 (1 node out of 1).

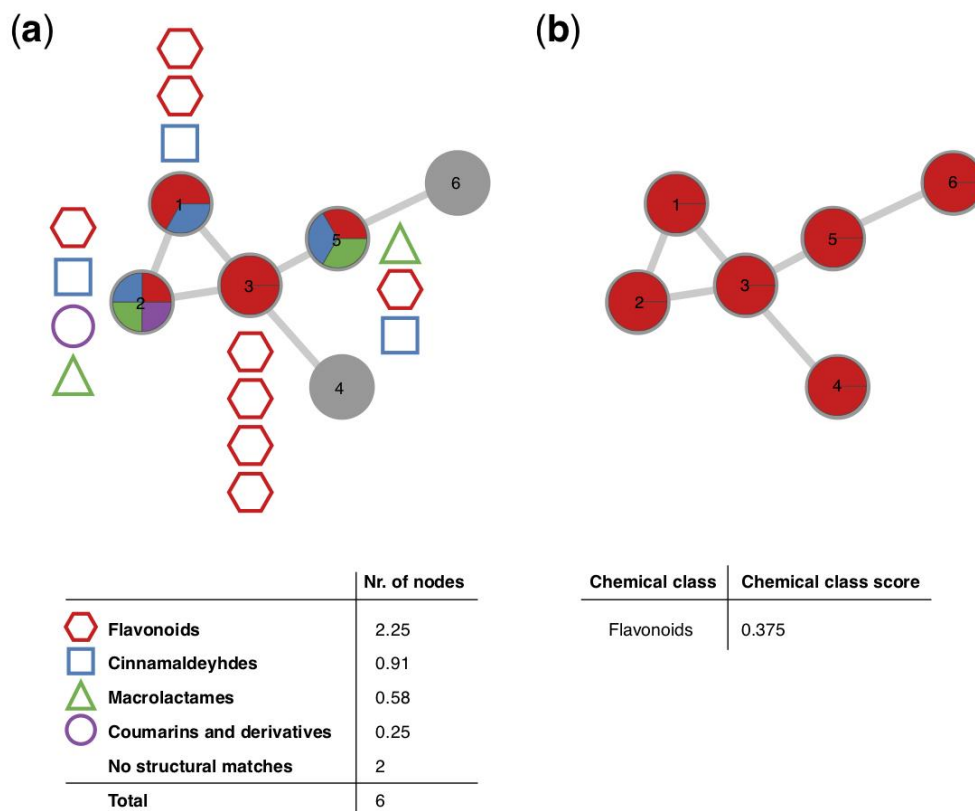


Figure 8. Schematic overview of how the Chemical Classification Score is calculated and visualized within a molecular family. (a) Schematic overview of hypothetical structural annotations within a molecular family consisting of 6 nodes. Out of the 6 nodes, chemical structural information could be retrieved for 4, where each node can consist of structural annotations to multiple (e.g. node 2) or identical (e.g. node 3) chemical classes. The total number of nodes per chemical class retrieved is calculated and the most abundant chemical class is assigned to the molecular family, resulting in (b) Schematic overview of the molecular family shown in (a), classified as ‘flavonoids’ at the chemical class level by MolNetEnhancer, with a score of 0.375, translating to the majority of the putative structural annotations within this molecular family (2.25) belong to the flavonoid structural class.

Publicly available mass spectrometry fragmentation data sets from four studies were used for this study. Details on how samples and data were collected can be found in the original studies [24,25,44,46]. Here, we list links to the different analyses that were done on each of the studies. Through these links, all used settings and parameters can be retrieved.

Data illustrating MolNetEnhancer applied to feature-based molecular networking are publicly accessible through the links listed below:

Case study 1: *Euphorbia* study - combined analysis of 43 *Euphorbia* plant extracts

- MASSIVE: MSV000081082
<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=c9f09d31a24c475e87a0a11f6e8889e7>
- GNPS Molecular Networking job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=26326c233918419f8dc80e8af984cdae>
- GNPS NAP jobs:
<https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=2cfddd3b8b1e469181a13e7d3a867a6f> and
<https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=184a80db74334668ae1d0c0f852cb77c>
- MS2LDA experiment: <http://ms2lda.org/basicviz/summary/390/>

Case study 2: Rhamnaceae study - combined analysis of 71 Rhamnaceae plant extracts

- MASSIVE: MSV000081805
<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=36f154d1c3844d31b9732fbaa72e9284>

- GNPS Molecular Networking job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e9e02c0ba3db473a9b1ddd36da72859b>
- GNPS NAP job:
<https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=6b515b235e0e4c76ba539524c8b4c6d8>
- MS2LDA experiment: <http://ms2lda.org/basicviz/summary/566>

GNPS example study used in Jupyter notebook to show MolNetEnhancer based on feature-based molecular networking - subset of American Gut Project

- MASSIVE: MSV000082678
<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=de2d18fd91804785bce8c225cc94a44>
- GNPS Molecular Networking job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b817262cb6114e7295fee4f73b22a3ad>
- GNPS NAP job:
<https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=c4bb6b8be9e14bdebe87c6ef3abe11f6>
- MS2LDA experiment: <http://ms2lda.org/basicviz/summary/907>

Data illustrating MolNetEnhancer applied to classical molecular networking are publicly accessible through the links listed below:

Case study 3: Marine-sediment bacteria study - combined analysis of 120 *Salinospira* and 26 *Streptomyces* bacterial strain extracts

- MASSIVE: MSV000078836, MSV000078839
<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=9277186021274990a5e646874a435c0d>
<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=a507232a787243a5afd69a6c6fa1e508>
- GNPS Molecular Networking job:
<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c36f90ba29fe44c18e96db802de0c6b9>
- GNPS NAP job:
<https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=60925078e0c148cbaba3593569e983d6>
- GNPS DEREPLICATOR 0.005 job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0ad6535e34d449788f297e712f43068a>
- GNPS DEREPLICATOR 0.05 job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e494a63be6d34747a4b8cdfb838ef96e>
- GNPS VARQUEST 0.005 job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f1f00c1c20ba4f61ad471d340066df76>
- GNPS VARQUEST 0.05 job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f5ffcc8f63ab4e6f96a97caabc11048b>
- MS2LDA experiment: <http://ms2lda.org/basicviz/summary/551/>
- MS2LDA MolNetEnhancer workflow experiment: <http://ms2lda.org/basicviz/summary/912/>

Case study 4: Nematode symbionts study - combined analysis of 25 *Xenorhabdus* and 5 *Photorhabdus* bacterial strain extracts

- MASSIVE: MSV000081063
<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=dc30b777c344d668a5626d01f26c9a0>
- GNPS Molecular Networking job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=aaff4721951b4d92b54ecbd2fe4b9b4f>
- GNPS NAP job:
<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=677f076eb04b4518958ca8cd56b4c753>
- GNPS DEREPLICATOR 0.005 job:
<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=338b422483d1432e82afd1bf848f1292>
- GNPS DEREPLICATOR 0.05 job:
<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=83bca3c45665470891d41ead275dcae7>
- GNPS VARQUEST 0.005 job:
<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=20cfb9af4a244feca102aa9c9da2651c>
- GNPS VARQUEST 0.05 job:
<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a4ffda169823476a9b1e81616aaccbda>
- MS2LDA annotation experiment: <http://ms2lda.org/basicviz/summary/570/>

- MS2LDA MolNetEnhancer workflow experiment: <http://ms2lda.org/basicviz/summary/917/>

GNPS example study used in Jupyter notebook to show MolNetEnhancer based on classical molecular networking - drug metabolism in set of sputum samples

- MASSIVE: MSV000081098
https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7c4b25d21a6348df9a6942d3071a4b1f&view=advanced_view
- GNPS Molecular Networking job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b76dd5a123e54a7eb42765499f9163a5>
- GNPS NAP job:
<https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=cb63770fe307410492468f62f9edb8f3>
- VarQuest job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4d971b8162644e869a68faa35f01b915>
- DEREPLICATOR job:
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c62d3283752f4f98b1720d0a6d1ee65b>
- MS2LDA experiment: <http://ms2lda.org/basicviz/summary/909/>

The MolNetEnhancer package in R including Jupyter notebooks with an exemplary analysis workflow for mapping Mass2Motifs onto classical and feature-based molecular networking is publicly accessible at: <https://github.com/madeleineernst/RMolNetEnhancer> and the MolNetEnhancer package in python including Jupyter notebooks with an exemplary analysis workflow for mapping Mass2Motifs and chemical class annotations onto classical and feature-based molecular networking is publicly accessible at: <https://github.com/madeleineernst/pyMolNetEnhancer>

5. Conclusions

MolNetEnhancer is a powerful tool to accelerate chemical structural annotation within complex metabolic mixtures through the combined use of mass spectral molecular networking, substructure discovery, *in silico* annotation as well as chemical classifications provided by ClassyFire. The MolNetEnhancer workflow is presented both as an open-source python module and R package, allowing easy access and usability by the community as well as the possibility for customization and further development by integration into future collaborative modular tools and by integration of other existing or future metabolome mining and annotation tools. Whilst its use was showcased using natural product examples, we expect that MolNetEnhancer will also enhance biological and chemical interpretations in other scientific fields such as clinical and environmental metabolomics.

Supplementary Materials: The following are available:

Figure S1. Mirror plot comparing molecular feature with m/z 614.30 and RT 373.17 (black) to GNPS reference spectrum of a jatrophone diterpenoid (green). A total of 289 shared peaks were found. Mass peaks at m/z 313, 295, 285 are characteristic for a *Euphorbia* diterpenoid backbone skeleton, however spectral similarity (cosine score) was only found to be 0.71. The unknown molecular feature is thus likely a close structural analogue of the jatrophone diterpenoid. The GNPS reference spectrum as well as the mirror plot is publicly accessible at https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=26326c233918419f8dc80e8af984cdac&view=view_all_annotations_D_B#%7B%22main.%23Scan%23_lowerinput%22%3A%223633%22%2C%22main.%23Scan%23_upperinput%22%3A%223633%22%7D.

Figure S2. (a) Marine sediment *Salinispora*/*Streptomyces* molecular network colored by chemical classification scores for annotated chemical class terms, and (b) same molecular network colored by chemical classification scores for annotated chemical kingdom terms. Light grey means no database matches were found. The higher the class score, the more consistent the chemical annotations are. The kingdom scores represent the database coverage of nodes across a molecular family with scores closer to zero representing families with fewer nodes that have at least one database hit. Whilst most MFs do have database matches for all or most nodes, the consistency in chemical class annotations is - apart from some exceptions - less (indicated by the more orange/pink colors in the left panel). This indicates that for many MF family members the right molecular structures might not yet be present in the structural databases used.

Figure S3. Molecular families from marine sediment bacteria with color coded Mass2Motif substructure information mapped on them, with (a) Lomaiviticin related molecular family where members contain an amino sugar related motif, (b) yet unknown molecular family that shares an amino sugar related motif, (c) yet unknown molecular family sharing an unknown fragment-based motif occurring 0.7% in the marine sediment data set, and (d) yet unknown molecular family sharing unknown loss-based motifs occurring 0.4% (Mass2Motif 250) and 0.8% (Mass2Motif 261) in the marine sediment data. In all MFs, nodes are coloured based on motif overlap scores and the edges present similar colours to show if cosine score-connected nodes share similar Mass2Motifs. It can be seen that in most families multiple motifs are shared across some of its members.

Figure S4. (a) Nematode symbionts *Photorhabdus*/*Xenorhabdus* network colored by chemical classification scores for annotated chemical class terms, and (b) same molecular network colored by chemical classification scores for annotated chemical kingdom terms. Light grey means no database matches were found. The higher the class score, the more consistent the chemical annotations are. The kingdom scores represent the database coverage of nodes across a molecular family with scores closer to zero representing families with fewer nodes that have at least a database hit. We observe database coverages of close to 1 for most molecular families; however, some molecular families have a lower coverage with a few nodes that return candidate structures. Furthermore, we observe that the chemical class annotation is not always consistent indicating that manual inspection and validation of those hits remains essential.

Figure S5. Xenoamicin Mass2Motif mass feature frequency plots for (a) Mass2Motif related to Xenoamicin peptidic ring and (b) Xenoamicin peptidic tail. It can be observed that many mass fragments are present in at least 75% of the associated molecular features (9 and 6 for ring and tail Mass2Motif, respectively) with a few mass fragments present in nearly all associated molecular features. (c) and (d) Examples of annotated Xenoamicin A modified structures in which only the ring Mass2Motif was found. Indeed, we observe that VarQuest annotates a modified amino acid (addition and loss of) in the tail region of Xenoamicin A indicated in orange. (e) and (f) Examples of annotated Xenoamicin B modified structures in which only the ring Mass2Motif was found. Indeed, we observe that VarQuest annotates a modified amino acid (double water addition, loss of methyl) in the ring region of Xenoamicin B indicated in orange. The structures of Xenoamicin A and B differ in one methyl group on the amino acid highlighted in orange in (f) where B has an isobutyl group and A an isopropyl group. In fact, the structure of Xenoamicin A is correctly annotated by VarQuest to this fragmented doubly charged ion.

Author Contributions: Conceptualization, M.E., S.R., J.J.J.v.d.H.; methodology, J.J.J.v.d.H. and M.E.; software M.E., M.W., J.W., S.R.; validation, J.J.J.v.d.H., K.B.K., A.M.C.R., L.-F.N.; formal analysis, J.J.J.v.d.H., M.E., K.B.K.; supervision, J.J.J.v.d.H.; writing—original draft preparation, J.J.J.v.d.H. and M.E.; writing—review and editing, J.J.J.v.d.H., M.E., K.B.K., A.M.C.R., L.-F.N., J.W., S.R., M.M., P.C.D.; visualization, M.E., J.J.J.v.d.H., K.B.K.; funding acquisition, J.J.J.v.d.H., P.C.D., M.M.

Funding: J.J.J.v.d.H. was funded by an ASDI eScience grant, ASDI.2017.030, from the Netherlands eScience Center - NLeSC. AMCR and PCD were supported by US National Science Foundation grant IOS-1656481. SR and JW were supported by EPSRC EP/R018634/1. SR was supported by BBSRC BB/R022054/1.

Acknowledgments: The authors thank all research groups that made their metabolomics data publicly available so it could be reused in the current study. Dr. Yannick Djoumbou Feunang (University of Alberta, Canada) is thanked for his support with the use of ClassyFire and Dr. Ricardo da Silva for scientific discussions and feedback on the methodology and workflow.

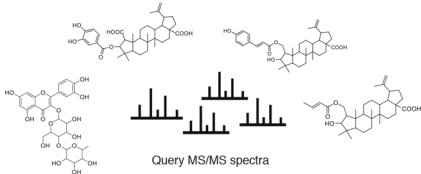
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada, K.; Dorrestein, P.C.; Pevzner, P.A. Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **2017**, *13*, 30–37.
2. Gurevich, A.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Mohimani, H.; Pevzner, P.A. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat Microbiol* **2018**, *3*, 319–327.
3. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A.A.; Melnik, A.V.; Meusel, M.; Dorrestein, P.C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **2019**, *16*, 299–302.
4. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12580–12585.
5. Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.* **2014**, *42*, W94–9.
6. Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F.; Wishart, D.S. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. *Metabolites* **2019**, *9*, 72.
7. Wang, M.; Carver, J.J.; Phelan, V.V.; Sanchez, L.M.; Garg, N.; Peng, Y.; Nguyen, D.D.; Watrous, J.; Kapon, C.A.; Luzzatto-Knaan, T.; et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828–837.
8. da Silva, R.R.; Wang, M.; Nothias, L.-F.; van der Hooft, J.J.J.; Caraballo-Rodríguez, A.M.; Fox, E.; Balunas, M.J.; Klassen, J.L.; Lopes, N.P.; Dorrestein, P.C. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **2018**, *14*, e1006089.
9. Ridder, L.; van der Hooft, J.J.J.; Verhoeven, S.; de Vos, R.C.H.; Vervoort, J.; Bino, R.J. In silico prediction and automatic LC-MS(n) annotation of green tea metabolites in urine. *Anal. Chem.* **2014**, *86*, 4767–4774.
10. Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B.S.; Yang, J.Y.; Kersten, R.D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J.M.; et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1743–E1752.
11. Gerlich, M.; Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **2013**, *48*, 291–298.
12. Treutler, H.; Tsugawa, H.; Porzel, A.; Gorzolka, K.; Tissier, A.; Neumann, S.; Balcke, G.U. Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. *Anal. Chem.* **2016**, *88*, 8082–8090.
13. Hooft, J.J.J. van der; van der Hooft, J.J.J.; Padmanabhan, S.; Burgess, K.E.V.; Barrett, M.P. Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics* **2016**, *12*.
14. van der Hooft, J.J.J.; Wandy, J.; Barrett, M.P.; Burgess, K.E.V.; Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 13738–13743.
15. Wandy, J.; Zhu, Y.; van der Hooft, J.J.J.; Daly, R.; Barrett, M.P.; Rogers, S. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **2018**, *34*, 317–318.
16. Feunang, Y.D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **2016**, *8*, 61.
17. Olivon, F.; Elie, N.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D. MetGem Software for the Generation of Molecular Networks Based on the t-SNE Algorithm. *Anal. Chem.* **2018**, *90*, 13900–13908.
18. Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. MS2Analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Anal. Chem.* **2014**, *86*, 10724–10731.
19. Laponogov, I.; Sadawi, N.; Galea, D.; Mirnezami, R.; Veselkov, K.A. ChemDistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics* **2018**, *34*, 2096–2102.
20. Edmands, W.M.B.; Petrick, L.; Barupal, D.K.; Scalbert, A.; Wilson, M.J.; Wickliffe, J.K.; Rappaport, S.M. compMS2Miner: An Automatable Metabolite Identification, Visualization, and Data-Sharing R Package for High-Resolution LC–MS Data Sets. *Analytical Chemistry* **2017**, *89*, 3919–3928.
21. Ruttkies, C.; Schymanski, E.L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **2016**, *8*, 3.
22. Naake, T.; Gaquerel, E. MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. *Bioinformatics* **2017**, *33*, 2419–2420.
23. Fox Ramos, A.E.; Evanno, L.; Poupon, E.; Champy, P.; Beniddir, M.A. Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat. Prod. Rep.* **2019**.
24. Ernst, M.; -F. Nothias, L.; van der Hooft, J.J.J.; Silva, R.R.; Saslis-Lagoudakis, C.H.; Grace, O.M.;

- Martinez-Swatson, K.; Hassemer, G.; Funez, L.A.; Simonsen, H.T.; et al. Did a plant-herbivore arms race drive chemical diversity in Euphorbia? *bioRxiv* 2018, 323014.
25. Kang, K.B.; Ernst, M.; van der Hooft, J.J.J.; da Silva, R.R.; Park, J.; Medema, M.H.; Sung, S.H.; Dorrestein, P.C. Comprehensive mass spectrometry-guided phenotyping of plant specialized metabolites reveals metabolic diversity in the cosmopolitan plant family Rhamnaceae. *Plant J.* **2019**.
26. Frank, A.M.; Bandeira, N.; Shen, Z.; Tanner, S.; Briggs, S.P.; Smith, R.D.; Pevzner, P.A. Clustering millions of tandem mass spectra. *J. Proteome Res.* **2008**, *7*, 113–122.
27. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 395.
28. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504.
29. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.
30. Govaerts, R.; Fernández Casas, F.J.; Barker, C.; Carter, S.; Davies, S.; Esser, H.-J.; Gilbert, M.; Hoffmann, P.; Radcliffe-Smith, A.; Steinmann, V.; et al. World Checklist of Euphorbiaceae. Facilitated by the Royal Botanic Gardens, Kew Available online: <http://apps.kew.org/wcsp/> (accessed on Jul 25, 2014).
31. Horn, J.W.; van Ee, B.W.; Morawetz, J.J.; Riina, R.; Steinmann, V.W.; Berry, P.E.; Wurdack, K.J. Phylogenetics and the evolution of major structural characters in the giant genus Euphorbia L. (Euphorbiaceae). *Mol. Phylogenet. Evol.* **2012**, *63*, 305–326.
32. Vasas, A.; Hohmann, J. Euphorbia Diterpenes: Isolation, Structure, Biological Activity, and Synthesis (2008–2012). *Chemical Reviews* 2014, *114*, 8579–8612.
33. Shi, Q.-W.; Su, X.-H.; Kiyota, H. Chemical and pharmacological research of the plants in genus Euphorbia. *Chem. Rev.* **2008**, *108*, 4295–4327.
34. Berman, B. New developments in the treatment of actinic keratosis: focus on ingenol mebutate gel. *Clin. Cosmet. Investig. Dermatol.* **2012**, *5*, 111–122.
35. Luo, D.; Callari, R.; Hamberger, B.; Wubshet, S.G.; Nielsen, M.T.; Andersen-Ranberg, J.; Hallström, B.M.; Cozzi, F.; Heider, H.; Möller, B.L.; et al. Oxidation and cyclization of casbene in the biosynthesis of Euphorbia factors from mature seeds of Euphorbia lathyris L. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E5082–E5089.
36. Appendino, G. Ingenane Diterpenoids. In *Progress in the Chemistry of Organic Natural Products 102*; Springer, Cham, 2016; pp. 1–90.
37. Nothias-Scaglia, L.-F.; Schmitz-Afonso, I.; Renucci, F.; Roussi, F.; Touboul, D.; Costa, J.; Litaudon, M.; Paolini, J. Insights on profiling of phorbol, deoxyphorbol, ingenol and jatrophone diterpene esters by high performance liquid chromatography coupled to multiple stage mass spectrometry. *J. Chromatogr. A* **2015**, *1422*, 128–139.
38. Nothias, L.-F.; Boutet-Mercey, S.; Cachet, X.; De La Torre, E.; Laboureur, L.; Gallard, J.-F.; Retaillieu, P.; Brunelle, A.; Dorrestein, P.C.; Costa, J.; et al. Environmentally Friendly Procedure Based on Supercritical Fluid Chromatography and Tandem Mass Spectrometry Molecular Networking for the Discovery of Potent Antiviral Compounds from Euphorbia semiperfoliata. *J. Nat. Prod.* **2017**, *80*, 2620–2629.
39. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; -M. Fan, T.W.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics* 2007, *3*, 211–221.
40. Onstein, R.E.; Carter, R.J.; Xing, Y.; Richardson, J.E.; Linder, H.P. Do Mediterranean-type ecosystems have a common history?--insights from the Buckthorn family (Rhamnaceae). *Evolution* **2015**, *69*, 756–771.
41. March, R.E.; Lewars, E.G.; Stacey, C.J.; Miao, X.-S.; Zhao, X.; Metcalfe, C.D. A comparison of flavonoid glycosides by electrospray tandem mass spectrometry. *International Journal of Mass Spectrometry* 2006, *248*, 61–85.
42. van der Hooft, J.J.J.; Vervoort, J.; Bino, R.J.; de Vos, R.C.H. Spectral trees as a robust annotation tool in LC–MS based metabolomics. *Metabolomics* 2012, *8*, 691–703.
43. van der Hooft, J.J.J.; Vervoort, J.; Bino, R.J.; Beekwilder, J.; de Vos, R.C.H. Polyphenol identification based on systematic and robust high-resolution accurate mass spectrometry fragmentation. *Anal. Chem.* **2011**, *83*, 409–416.
44. Crüsemann, M.; O'Neill, E.C.; Larson, C.B.; Melnik, A.V.; Floros, D.J.; da Silva, R.R.; Jensen, P.R.; Dorrestein, P.C.; Moore, B.S. Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. *J. Nat. Prod.* **2017**, *80*, 588–597.

45. Crnovčić, I.; Semsary, S.; Vater, J.; Keller, U. Biosynthetic rivalry of o-aminophenol-carboxylic acids initiates production of hemi-actinomycins in *Streptomyces antibioticus*. *RSC Advances* 2014, 4, 5065.
46. Tobias, N.J.; Wolff, H.; Djahanschiri, B.; Grundmann, F.; Kronenwerth, M.; Shi, Y.-M.; Simonyi, S.; Grün, P.; Shapiro-Ilan, D.; Pidot, S.J.; et al. Natural product diversity associated with the nematode symbionts *Photorhabdus* and *Xenorhabdus*. *Nature Microbiology* 2017, 2, 1676.
47. Shi, Y.-M.; Bode, H.B. Chemical language and warfare of bacterial natural products in bacteria–nematode–insect interactions. *Natural Product Reports* 2018, 35, 309–335.
48. Tobias, N.; Parra-Rojas, C.; Shi, Y.-N.; Shi, Y.-M.; Simonyi, S.; Thanwisai, A.; Vitta, A.; Chantratita, N.; Hernandez-Vargas, E.A.; Bode, H.B. Focused natural product elucidation by prioritizing high-throughput metabolomic studies with machine learning. *bioRxiv* 2019, 535781.
49. Zhao, L.; Kaiser, M.; Bode, H.B. Rhabdopeptide/Xenortide-like Peptides from *Xenorhabdus innexi* with Terminal Amines Showing Potent Antiprotozoal Activity. *Org. Lett.* 2018, 20, 5116–5120.
50. Zhou, Q.; Dowling, A.; Heide, H.; Wöhnert, J.; Brandt, U.; Baum, J.; French-Constant, R.; Bode, H.B. Xentrivalpeptides A–Q: Dipeptide Diversification in *Xenorhabdus*. *Journal of Natural Products* 2012, 75, 1717–1722.
51. Fuchs, S.W.; Proschak, A.; Jaskolla, T.W.; Karas, M.; Bode, H.B. Structure elucidation and biosynthesis of lysine-rich cyclic peptides in *Xenorhabdus nematophila*. *Org. Biomol. Chem.* 2011, 9, 3130–3132.
52. Zhou, Q.; Grundmann, F.; Kaiser, M.; Schiell, M.; Gaudriault, S.; Batzer, A.; Kurz, M.; Bode, H.B. Structure and biosynthesis of xenoamicins from entomopathogenic *Xenorhabdus*. *Chemistry* 2013, 19, 16772–16779.
53. da Silva, R.R.; Dorrestein, P.C.; Quinn, R.A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* 2015, 112, 12549–12550.
54. Metabolomics: Dark matter. *Nature* 2008, 455, 698.
55. Newman, D.J.; Cragg, G.M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* 2016, 79, 629–661.
56. Crupi, P.; Antonacci, D.; Savino, M.; Genghi, R.; Perniola, R.; Coletta, A. Girdling and gibberellic acid effects on yield and quality of a seedless red table grape for saving irrigation water supply. *European Journal of Agronomy* 2016, 80, 21–31.
57. Pan, A.; Chen, M.; Chowdhury, R.; Wu, J.H.Y.; Sun, Q.; Campos, H.; Mozaffarian, D.; Hu, F.B. α -Linolenic acid and risk of cardiovascular disease: a systematic review and meta-analysis. *Am. J. Clin. Nutr.* 2012, 96, 1262–1273.
58. Oberacher, H.; Reinstadler, V.; Kreidl, M.; Stravs, M.A.; Hollender, J.; Schymanski, E.L. Annotating Nontargeted LC-HRMS/MS Data with Two Complementary Tandem Mass Spectral Libraries. *Metabolites* 2018, 9, 3.
59. Scheubert, K.; Hufsky, F.; Petras, D.; Wang, M.; Nothias, L.-F.; Dührkop, K.; Bandeira, N.; Dorrestein, P.C.; Böcker, S. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* 2017, 8, 1494.
60. Olivon, F.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D. MZmine 2 Data-Preprocessing To Enhance Molecular Networking Reliability. *Anal. Chem.* 2017, 89, 7836–7840.
61. Rogers, S.; Ong, C.W.; Wandy, J.; Ernst, M.; Ridder, L.; van der Hooft, J.J.J. Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra.
62. van der Hooft, J.J.J.; Wandy, J.; Young, F.; Padmanabhan, S.; Gerasimidis, K.; Burgess, K.E.V.; Barrett, M.P.; Rogers, S. Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Anal. Chem.* 2017, 89, 7569.



In silico structure
annotaion



Molecular networking and
library matching

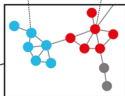
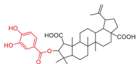
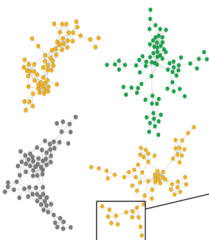


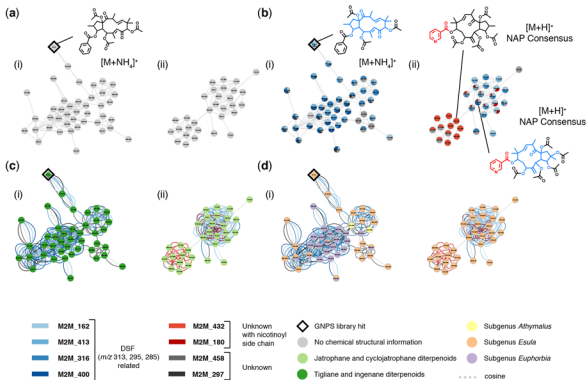
MS2LDA

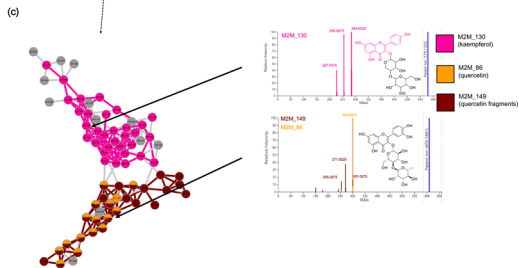
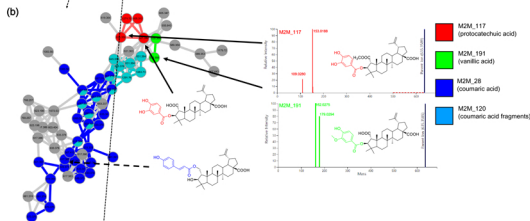
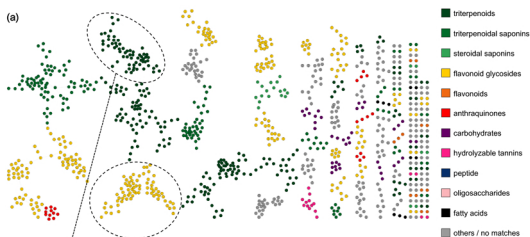
Unsupervised
substructure discovery

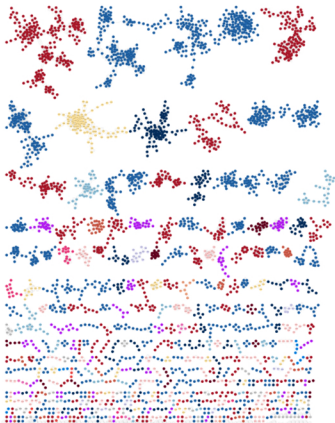


- Triterpenoids
- Terpene glycosides
- Flavonoid-O-glycosides



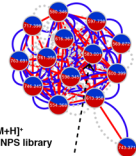




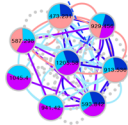


- Benzene and substituted derivatives
- Carboxylic acids and derivatives
- Fatty Acyls
- Flavonoids
- Glycerophospholipids
- Macrolactams
- Macrolides and analogues
- Macroline alkaloids
- Organooxygen compounds
- Organonitrogen compounds
- Peptidomimetics
- Phenol ethers
- Polypeptides
- Prenol lipids
- Steroids and steroids derivatives
- Others
- No match

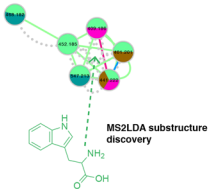
(a)

[M+H]⁺
GNPS library

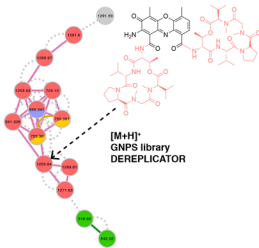
(b)

MS2LDA
substructure
discovery

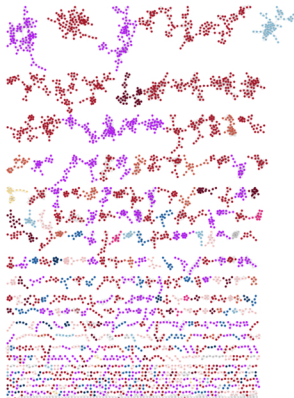
(c)



(d)

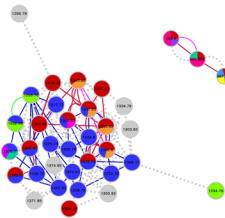
[M+H]⁺
GNPS library
DEREPLICATOR

- M2M_66
- M2M_285
Aminosugar related
- M2M_115
- M2M_74
- M2M_240
- M2M_7
- M2M_155
- M2M_154
Tryptophan related
- M2M_69
Actinomycin related
- M2M_245
- M2M_196
- M2M_138
- no M2M
- ● ● ● cosine

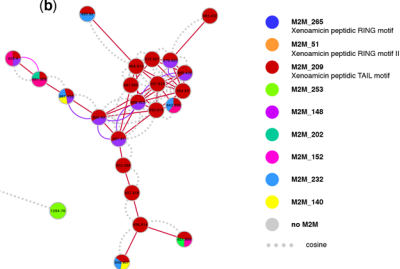


- Benzene and substituted derivatives
- Carboxylic acids and derivatives
- Fatty Acyls
- Glycerophospholipids
- Indoles and derivatives
- Organonitrogen compounds
- Peptidomimetics
- Polypeptides
- Prenol lipids
- Steroids and steroids derivatives
- Others
- No match

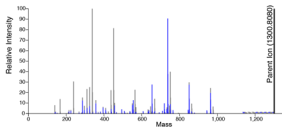
(a)



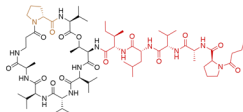
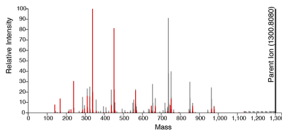
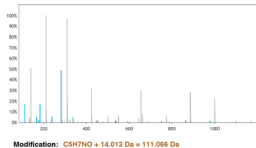
(b)



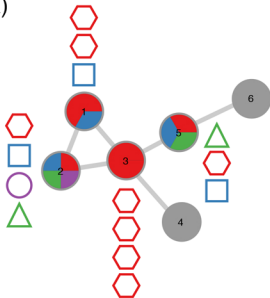
(c)



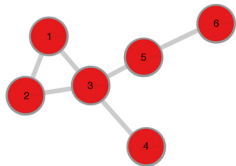
(d)







(a)



(b)



	Nr. of nodes
 Flavonoids	2.25
 Cinnamaldehydes	0.91
 Macrolactams	0.58
 Coumarins and derivatives	0.25
No structural matches	2
Total	6

Chemical class	Chemical class score
Flavonoids	0.375