Open Access Repository-Scale Propagated Nearest Neighbor Suspect Spectral Library for Untargeted Metabolomics

Wout Bittremieux[1,2], Nicole E. Avalon[3], Sydney P. Thomas[1,2], Sarvar A. Kakhkhorov[4,5], Alexander A. Aksenov[1,2,6], Paulo Wender P. Gomes[1,2], Christine M. Aceves[7], Andrés Mauricio Caraballo Rodríguez[1,2], Julia M. Gauglitz[1,2], William H. Gerwick[2,3], Alan K. Jarmusch[1,2,8], Rima F. Kaddurah-Daouk[9,10,11], Kyo Bin Kang[12], Hyun Woo Kim[13], Todor Kondić[14], Helena Mannochio-Russo[1,2,15], Michael J. Meehan[1,2], Alexey V. Melnik[6], Louis-Felix Nothias[16,17], Claire O'Donovan[18], Morgan Panitchpakdi[1,2], Daniel Petras[1,2,19], Robin Schmid[1,2], Emma L. Schymanski[14], Justin J. J. van der Hooft[1,20], Kelly C. Weldon[1,2], Heejung Yang[21], Jasmine Zemlin[1,2], Mingxun Wang[1,2], Pieter C. Dorrestein[1,2,*]

1. Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA 92093, USA
2. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla CA 92093, USA
3. Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA
4. Laboratory of Physical and Chemical Methods of Research, Center for Advanced Technologies, Tashkent 100174, Uzbekistan
5. Laboratory of Biotechnology, Center for Advanced Technologies, Tashkent 100174, Uzbekistan
6. Department of Chemistry, University of Connecticut, Storrs, CT 06269, USA
7. Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA
8. Immunity, Inflammation, and Disease Laboratory, Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA
9. Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, NC 27701, USA
10. Department of Medicine, Duke University, Durham, NC 27710, USA
11. Duke Institute of Brain Sciences, Duke University, Durham, NC 27710, USA
12. College of Pharmacy and Research Institute of Pharmaceutical Sciences, Sookmyung Women's University, Seoul 04310, Korea
13. College of Pharmacy and Integrated Research Institute for Drug Development, Dongguk University, Goyang 10326, Korea
14. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-4367 Belvaux, Luxembourg
15. Department of Biochemistry and Organic Chemistry, Institute of Chemistry, São Paulo State University, Araraquara, 14800-901, Brazil

16. Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, CMU, 1211 Geneva, Switzerland
17. School of Pharmaceutical Sciences, University of Geneva, CMU, 1211 Geneva, Switzerland
18. European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
19. CMFI Cluster of Excellence, Interfaculty Institute of Microbiology and Infection Medicine, University of Tuebingen, 72076 Tuebingen, Germany
20. Bioinformatics Group, Wageningen University, 6708PB Wageningen, the Netherlands
21. Laboratory of Natural Products Chemistry, College of Pharmacy, Kangwon National University, Chuncheon 24341, Korea

Corresponding author: pdorrestein@health.ucsd.edu

Abstract

Despite the increasing availability of tandem mass spectrometry (MS/MS) community spectral libraries for untargeted metabolomics over the past decade, the majority of acquired MS/MS spectra remain uninterpreted. To further aid in interpreting unannotated spectra, we created a nearest neighbor suspect spectral library, consisting of 87,916 annotated MS/MS spectra derived from hundreds of millions of public MS/MS spectra. Annotations were propagated based on structural relationships to reference molecules using MS/MS-based spectrum alignment. We demonstrate the broad relevance of the nearest neighbor suspect spectral library through representative examples of propagation-based annotation of acylcarnitines, bacterial and plant natural products, and drug metabolism. Our results also highlight how the library can help to better understand an Alzheimer's brain phenotype. The nearest neighbor suspect spectral library is openly available through the GNPS platform to help investigators hypothesize candidate structures for unknown MS/MS spectra in untargeted metabolomics data.

Introduction

When searching untargeted tandem mass spectrometry (MS/MS) metabolomics data using spectral libraries, on average only ~5% of the data can be annotated (~10% for human datasets). Unannotated spectra can arise due to incomplete coverage of the reference MS/MS spectral libraries of known compounds, including missing MS/MS spectra of different ion species, such as different ion forms, in-source fragments, and formation of multimers.[1–3] We hypothesized that many of the unidentified ions originate from different but related known molecules. Those molecules could be a result of host or microbial metabolism or promiscuous enzymes that accept various analogous substrates during biosynthesis.[4] To find related candidate ion species or to discover analogous MS/MS spectra from ions that originate from related molecules, strategies such as molecular networking[5] and other analog searching strategies[3,6–9] can be employed, for which molecular networking—a data visualization and interpretation strategy of MS/MS spectral alignment—in the Global Natural Products Social Molecular Networking (GNPS) environment[10] is one of the most widely used tools.[11]

These strategies can also be used to generate new libraries of MS/MS reference spectra of potentially related MS/MS annotations from analog molecules that can subsequently be reused by the community. For example, small reference spectral libraries of human milk oligosaccharides[12] and urine acylcarnitines[13] were produced using an analog searching strategy (although the user licenses of these libraries restrict their redistribution). We hypothesized that the benefits of this approach could be further increased by considering analog matches across extremely large collections of MS/MS spectra to maximize the number of relevant spectrum links that can be found. Therefore, we have created a freely accessible and reusable MS/MS spectral library of MS/MS spectra related to identifiable molecules using molecular networking at the repository scale and created a nearest neighbor suspect spectral library to facilitate the annotation of mass spectrometry features that are present in public data.

Results

Nearest neighbor suspect spectral library creation

Using molecular networking, we have created a freely available and open-access mass spectral library of chemical analogs, referred to as the "nearest neighbor suspect spectral library." The library was created from compatible public datasets deposited to GNPS/MassIVE,[10] MetaboLights,[14] and Metabolomics Workbench.[15] In total, 521 million MS/MS spectra in 1335 public projects, with data from thousands of different organisms from diverse sources, including microbial culture collections, food, soil, dissolved organic matter, marine invertebrates, and humans, were used to compile the nearest neighbor suspect spectral library. Entries in this library, or "suspects," were derived from unannotated spectra that were linked in a molecular network (based on spectral similarity) to an annotated spectrum by MS/MS spectral library searching and where the precursor ion mass difference between the two spectra was non-zero.

Because we do not yet have the computational resources to process 521 million MS/MS spectra simultaneously, a hierarchical processing strategy was employed (**Figure 1a**). First, separate molecular networks were created from each dataset individually, while merging near-identical spectra and only keeping spectra that occur at least twice within the dataset to eliminate non-reproducible MS/MS spectra (**Figure 1a, step 1**). Spectrum annotations were obtained at the individual dataset level by matching against 221,224 reference spectra available in the GNPS community spectral libraries (June 2021) using parameters consistent with a false discovery rate (FDR) <1%.[10] The cosine similarity was calculated using filtered spectra (the precursor *m*/*z* peak was removed and only the top 6 most intense ions in every 50 *m*/*z* window were included), and spectrum matches with a cosine score of 0.8 or higher and a minimum of 6 matching ions were accepted. Second, a global molecular network was created from all of the individual networks using the GNPS modified cosine similarity (**Figure 1a, step 2**). Finally, annotation propagations to the nearest neighbors were extracted from all molecular networks to create the library of nearest neighbor suspects (**Figure 1a, step 3**). To maximize the quality of the suspect annotations, suspects with infrequent mass offsets that occur fewer than ten times were excluded, as these are considered to be less-reproducible mass differences (**Supplementary Figure 1**). Finally, a representative number of the annotation propagations were validated through expert manual inspection.

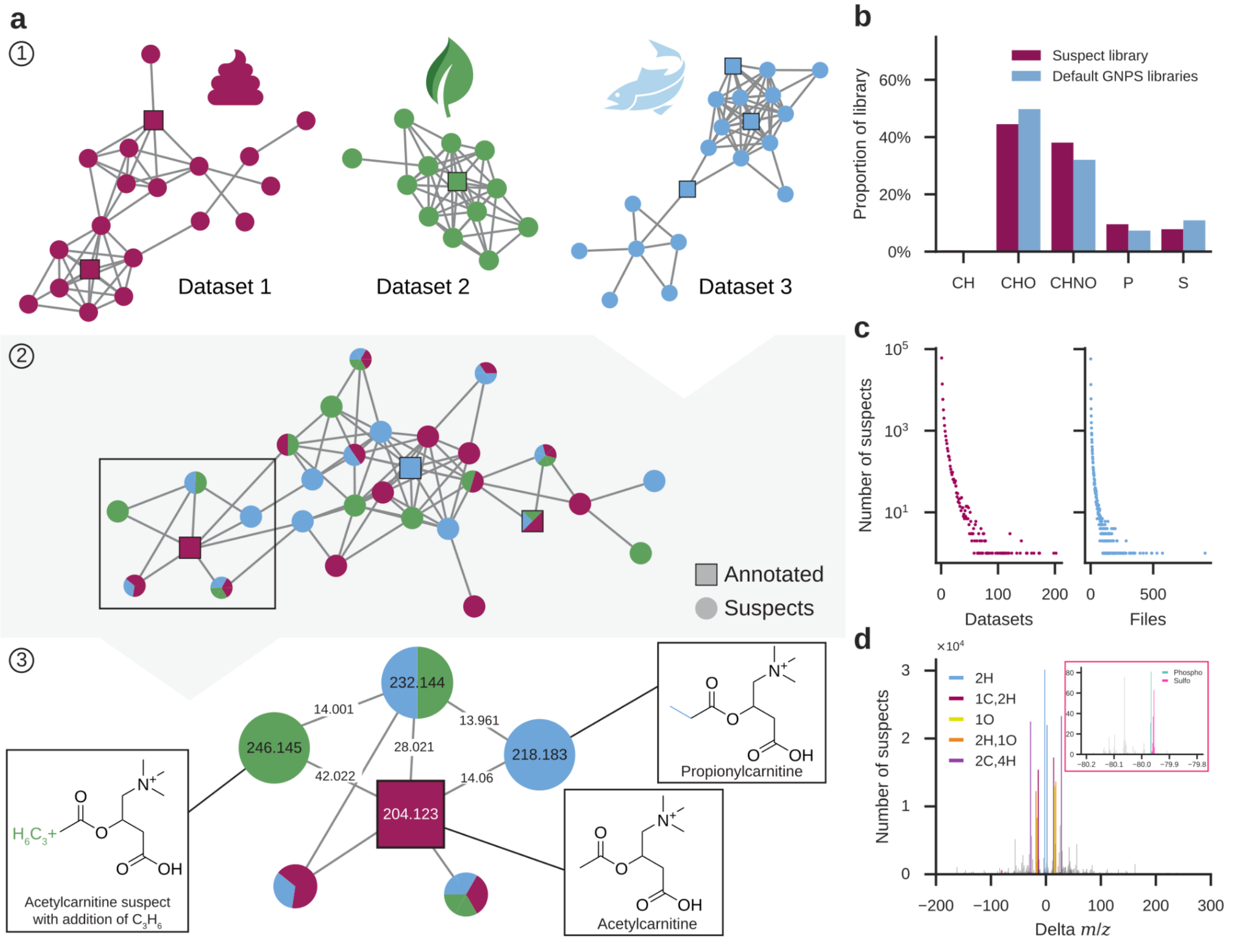


Figure 1. Creation and composition of the nearest neighbor suspect spectral library. a. Overview of how the suspect library was created. Step 1: molecular networking of individual datasets. Step 2: co-networking of the 1335 datasets to create a global molecular network. Step 3: extract nearest neighbor suspects through annotation propagation to create the library. **b.** The composition of suspects that exclusively exist of CH, CHO, CHNO, or contain P or S compared to the reference libraries. **c.** Repeated occurrences of the suspects across datasets and files (i.e., individual LC-MS runs). **d.** Frequently observed mass offsets (delta masses between pairs of spectra) associated with the suspect library. The inset shows the mass offset around a nominal mass of -80 Da.

In total, 87,916 unique MS/MS spectra and provenance to their matching analogs in the GNPS spectral libraries are included in the nearest neighbor suspect spectral library. Importantly, all of the nearest neighbor suspects are real spectra that occur in experimental data, whereas only a small portion (less than 10%) of reference MS/MS spectra in public and commercially available MS/MS spectral libraries have been observed in public data. To homogenize and extend the information available for the suspects, molecular formulas for 64,810 suspects were determined using SIRIUS.[16] The elemental composition of the suspects reflects the characteristics of known reference libraries (**Figure 1b**). For example, molecules that exclusively contain CH are poor ionizers and are observed very rarely for both library types. Some suspects, such as common contaminants from sample vials, skin, or sodium formate clusters, as well as those related to endogenous molecules, such as fatty acids (e.g. vaccenic acid), bile acids (e.g. cholic acid), and lipids (e.g. phosphatidylcholines), are found in hundreds of public datasets and mass spectrometry files. In contrast, others, such as the natural products apratoxin, chelidonine, or marrubiin are observed less frequently (**Figure 1c, Supplementary Table 1**).

There are 1350 frequent delta masses that occur in the nearest neighbor suspect spectral library (**Figure 1d, Supplementary Table 2**). When possible, the elemental composition of the delta masses and potential explanations, sourced from UNIMOD[17]—as many post-translational modifications or adducts that are observed in proteomics can also be found for small molecules—and a community-curated list of delta masses (**Supplementary Table 3**) are provided. The most common mass offsets observed in the suspect library correspond to a gain or loss of 2.016 Da, which can be interpreted as the gain or loss of 2H (e.g., a double bond or ring structure), followed by a gain or loss of 28.031 Da, 14.016 Da, 18.011 Da, and 15.995 Da, corresponding to C_2H_4 (e.g., di(de)methylation or (de)ethylation), CH_2 (e.g., (de)methylation), H_2O (e.g., water gain/loss), and O (e.g., (de)oxidation or (de)hydroxylation), respectively. However, 852 out of the 1350 mass offsets have not yet been explained (**Supplementary Figure 1**). For example, although these mass offsets occur less frequently, there are at least seven repeatedly observed offsets with a nominal delta mass of -80 Da (**Figure 1d, inset**), of which only phosphate loss (-79.966 Da) and sulfate loss (-79.957 Da) could currently be explained.

Spectral libraries are typically created by acquiring spectral data for pure standards, and reference MS/MS spectra have associated information on the precursor ions, compound names, and, when available, the molecular structures. In contrast, because the nearest neighbor suspect spectral library was compiled in a data-driven fashion, exact molecular structures are not known. Instead, the provenance of the suspect MS/MS spectra is described by their relationships to spectra that have an annotation, including the name and structure of the nearest neighbor MS/MS annotations and the observed pairwise delta masses. This is complemented by molecular formulas computed by SIRIUS, if available, and the elemental composition and potential explanation of the delta masses, as determined by matching against a curated list of delta masses. Suspects thus represent unknown molecules that are likely structurally related to reference molecules annotated using spectral library searching, with the location of the structural modification generally unspecified. Without any additional information, this is in agreement with a level 3 annotation (family level match) according to the Metabolomics Standards Initiative guidelines.[18]

Suspects provide structural hypotheses for observed molecules

The nearest neighbor suspect spectral library covers various classes of molecules arising from both primary and specialized metabolism, including lipids, flavonoids, and peptides. A fundamental understanding of organic chemistry, mass spectral fragmentation, and awareness of the information that mass spectrometry can or cannot provide is key to achieve the deepest possible structural insights from the suspect library. To demonstrate how the nearest neighbor annotations can be used to propagate structural information, we highlight below representative examples of acylcarnitines, apratoxin natural products, drug metabolism, flavonoids, and polymers in greater detail. Note that we do not discuss the stereochemistry of the suspect examples, as this information generally cannot be determined using mass spectrometry.

The first example involves several acylcarnitines, a group of molecules that plays a key role in mammalian—including human—energy cycling (**Figure 2a**).[19] Hexanoylcarnitine, C6:0, is formed from the condensation of carnitine with hexanoic acid, a linear fatty acid with six carbons and zero double bonds. The first suspect was initially annotated as a hexanoylcarnitine but with a loss of 12.036 Da. Although close in value, based on accurate mass defects, the observed mass difference does not correspond to a loss of C (12.000 Da), but rather a combination that corresponds to the loss of C_2H_2 and gain of O. As the typical carnitine fragmentation pattern is conserved,[13] we can determine that these changes occur in the fatty acid portion of the molecule, and thus, that this is likely a hydroxybutanoic acid carnitine derivative. Additionally, we can observe a characteristic 3-hydroxy fragment ion with a mass of 145.050 Da,[20] leading to the final interpretation of 3-hydroxybutyrylcarnitine. The second acylcarnitine suspect example was initially annotated as a hexanoylcarnitine but with a loss of 2.016 Da. This indicates that this suspect is likely a hexenoylcarnitine derived from a C6:1 fatty acid. Thus, the six carbon fatty acid tail now has one double bond, but the location of the double bond and its configuration (*E* vs *Z*) cannot be determined. The third suspect example was annotated as a hexanoylcarnitine with a loss of 5.955 Da, which corresponds to a gain of one C along with the loss of six hydrogens. The only structure that can match the acyl side chain is a planar benzoyl ester. The fourth suspect derivative of hexanoylcarnitine showed an addition of 15.995 Da, representing a possible oxidation. Based again on the characteristic 3-hydroxy fragment,[20] the oxidation could be localized, resulting in a spectrum annotation of 3-hydroxyhexanoylcarnitine. The final hexanoylcarnitine suspect showed an addition of 114.068 Da, representing a carnitine with an acyl side chain that has two oxygens and twelve carbons, as derived from the mass difference and characteristic neutral losses for carnitine conjugates of dicarboxylic acids (179.121 Da and 207.130 Da),[21] which is consistent with dodecanedioylcarnitine.

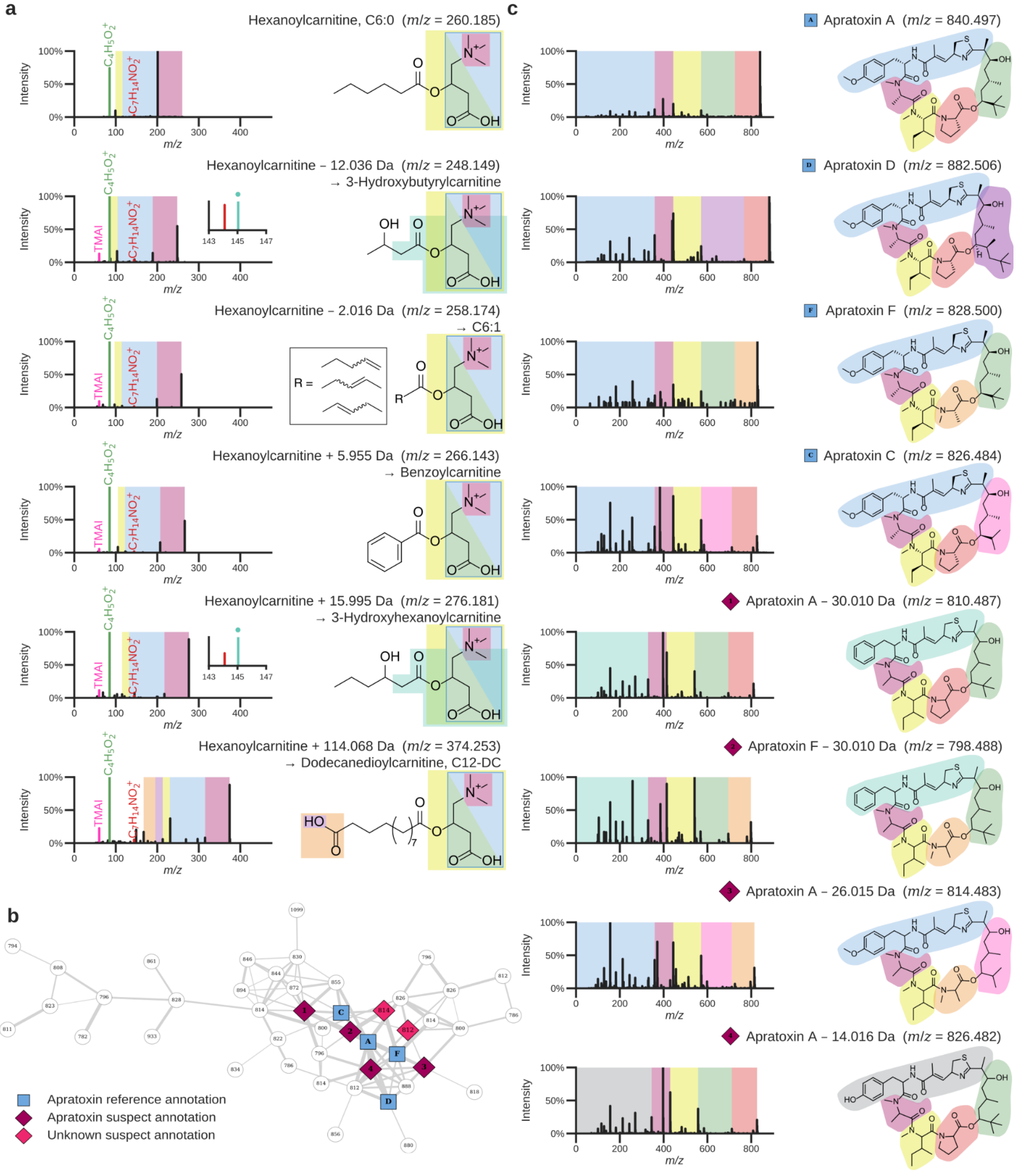


Figure 2. Examples of structural insights that can be obtained using the nearest neighbor suspect spectral library. a. The reference MS/MS spectrum for hexanoylcarnitine and five nearest neighbor suspects related to hexanoylcarnitine. **b.** Apratoxin cluster in a molecular network created from *Moorena bouillonii* crude extracts. The reference spectral library hits are shown by the blue squares. The purple and pink diamond nodes represent matches to the nearest neighbor suspect spectral library, with the purple diamonds matching the MS/MS

spectra shown in panel c for which structures could be proposed. The white nodes are additional MS/MS spectra within the apratoxin molecular family that remained unannotated, even when including the suspect library. **c.** MS/MS spectra and structural hypotheses for four apratoxin suspects, compared to the reference MS/MS spectra of known apratoxins.

The suspect library is also informative for the analysis of more complex molecules. The apratoxin class of natural products was isolated from filamentous cyanobacteria, and has been investigated in a number of biological systems due to its potent antineoplastic activities.[22,23] Using the suspect library to analyze a *Moorena bouillonii* cyanobacterial dataset achieved six additional spectrum annotations in the apratoxin molecular family (**Figure 2b**). A structural annotation can be determined for four of these based on comparisons to the MS/MS spectra of apratoxin standards (**Figure 2c**). The first four MS/MS spectra in **Figure 2c** show standards of purified apratoxin A, D, F, and C, followed by four apratoxin suspects with proposed structures. Some key substitutions observed are proline for *N*-methylalanine, methoxytyrosine for tyrosine, and dimethyl versus trimethyl polyketide initiating units. These substitutions are likely generated due to biosynthetic promiscuity commonly associated with multimodular hybrid non-ribosomal peptide synthetases-polyketide synthases.[24] The apratoxin suspects that were observed are apratoxin A and F with loss of 30.010 Da, apratoxin A with loss of 26.015 Da, apratoxin A with loss of 28.031 Da, and apratoxin A and F with loss of 14.016 Da; corresponding to CH_2O (e.g., methoxy loss), C_2H_2 or CH_2 + C loss, C_2H_4 loss (e.g., dimethylation), and CH_2 loss (e.g., methyl), respectively. The MS/MS spectra for four of the apratoxin suspects are shown in **Figure 2c**. Based on the fragmentation, the *m*/*z* difference corresponding to CH_2 loss in apratoxin A is due to unmethylated tyrosine, which could be explained by inactivity of an *O*-methyl transferase during biosynthesis. The fragmentation for both apratoxins with the 30.010 Da loss supports that the *m*/*z* difference corresponding to methoxy loss is a result of phenylalanine incorporation by the associated adenylation domain rather than the methylated tyrosine observed in previously published apratoxin structures. Finally, although the loss of 26.015 Da is more complex, the other known apratoxins, together with their fragmentation, can be used to formulate a refined structural hypothesis. Compared to apratoxin A, the proline is likely substituted by an *N*-methylalanine, corresponding to a loss of C, and the trimethyl initiating unit is replaced by an isopropyl initiating unit. To obtain support for these modifications, this suspect (apratoxin A - 26.015 Da) was purified from an extract of *Moorena bouillonii* and subjected to nuclear magnetic resonance (NMR) analysis (**Supplementary Figures 2–3**). Compared to NMR analysis of apratoxin A and consistent with the mass spectrometry interpretation, the NMR correlations associated with proline are lost and the NMR signals corresponding to *N*-methylation of alanine are now observed. Substructure analysis based on the MS/MS data revealed that the polyketide synthase portion of the apratoxin suspect differs by one methyl group. This is consistent with the suspect containing an isopropyl group, as observed in apratoxin C, rather than the *tert*-butyl group observed in nearly all of the other apratoxins.

The suspect library also contains modified versions of known drugs that can arise due to in-source fragmentation, the formation of different ion species, incomplete synthesis or biosynthesis of the active ingredient that arises during manufacturing of the drug, or modifications introduced due to metabolism. An example is a suspect found in a human breast milk dataset matching the antibiotic azithromycin (**Supplementary Figure 4**).[25] The suspect is 14.015 Da lighter, consistent with a CH_2 loss. Based on inspection of the MS/MS data, it is possible to tentatively assign this loss of CH_2 to the methoxy group in the cladinose sugar, based on the presence of a hydroxy loss and absence of a methoxy loss.

Next, MS/MS data from medicinal plants listed in the Korean Pharmacopeia were analyzed using molecular networking. Several flavonoid diglycosides containing pentoses and hexoses were detected using MS/MS spectral library searching, with the default GNPS libraries providing ten spectrum annotations in this molecular family and the suspect library contributing annotations for 27 diglycoside analogs (**Supplementary Figure 5**). Visual inspection of the MS/MS spectra indicated several modifications to formulate structural hypotheses for these suspects. For example, the apigenin-8-*C*-hexosylhexoside suspect with a delta mass of -30.019 Da corresponds to apigenin-8-*C*-pentosylhexoside. The presence of a pentose, instead of a hexose, is indeed consistent with the loss of CH_2O.

Finally, analysis of closely related polymeric substances resulted in a substantial increase in annotations (**Supplementary Figure 6**). In an indoor chemistry environmental study,[26] where a house was sampled before and after a month of human occupancy, there was a single spectrum match using the default GNPS libraries, to *p*-tert-octylphenol pentaglycol ether. Incorporating the suspect library added 55 matches that are related to polyethers, and that could be interpreted as part of a molecular family containing polymers. Thus, matching to the default GNPS spectral libraries alone gave the erroneous impression that there were only a few octylphenol-polyethylene glycol molecules detectable within the house, while the suspect library revealed that there is a large and diverse group of them.

In conclusion, these examples highlight how annotations provided by MS/MS spectral libraries, including the nearest neighbor suspect spectral library, can assist in providing structural hypotheses at the molecular family level for observed molecules.

MS/MS spectrum annotation increases provide new biomedical insights

To evaluate the spectrum annotation performance of the nearest neighbor suspect spectral library, we performed spectral library searching of public untargeted metabolomics data on GNPS/MassIVE (**Figure 3a**). For the 1335 public datasets included during the creation of the suspect library, the default GNPS libraries resulted in an average MS/MS spectrum match rate of 5.5% (median 3.6%). Inclusion of the suspect library boosted the MS/MS spectrum match rate to 9.3% (median 6.4%), corresponding to 19 million additional spectrum matches. While these datasets were used to generate the suspect library, a similar increase in spectrum match rate was achieved for independent test data that were not part of the molecular networks from which the suspect library was compiled. For 72 datasets that were publicly deposited after the creation of the suspect library, the average spectrum match rate using the default GNPS libraries was 5.7% (median 4.7%), which increased to 8.9% (median 7.5%) when including the suspect library. Furthermore, we evaluated the performance of the suspect library for samples of different origins as recorded using the ReDU metadata system (**Figure 3b**).[27] For 45,845 raw files from 179 datasets with controlled vocabularies for sample information, such as animal (including human), bacterial, fungal, environmental, food, and plant samples, the suspect library consistently achieved an increased spectrum match rate, ranging from a 1.7 ± 0.3 fold increase in interpreted spectra for food data to 3.0 ± 0.7 fold increase for environmental samples (mean and standard deviation).

To further demonstrate the utility of the suspect library, we focused on untargeted metabolomics data from 514 human brains with and without Alzheimer's disease.[28] Using the default GNPS libraries there were 248,317 MS/MS spectral library matches, corresponding to 1305 unique molecule annotations. Including the suspect library increased the number of spectrum matches to 401,039, covering 5184 unique molecule annotations (**Figure 3c**). One specific class of molecules that saw a particularly large increase in the number of annotations in this cohort was the acylcarnitines. There were 942 spectrum matches to 12 unique molecule annotations before the suspect library was included, but 1896 spectrum matches to 104 unique molecule annotations after inclusion of the suspect library.

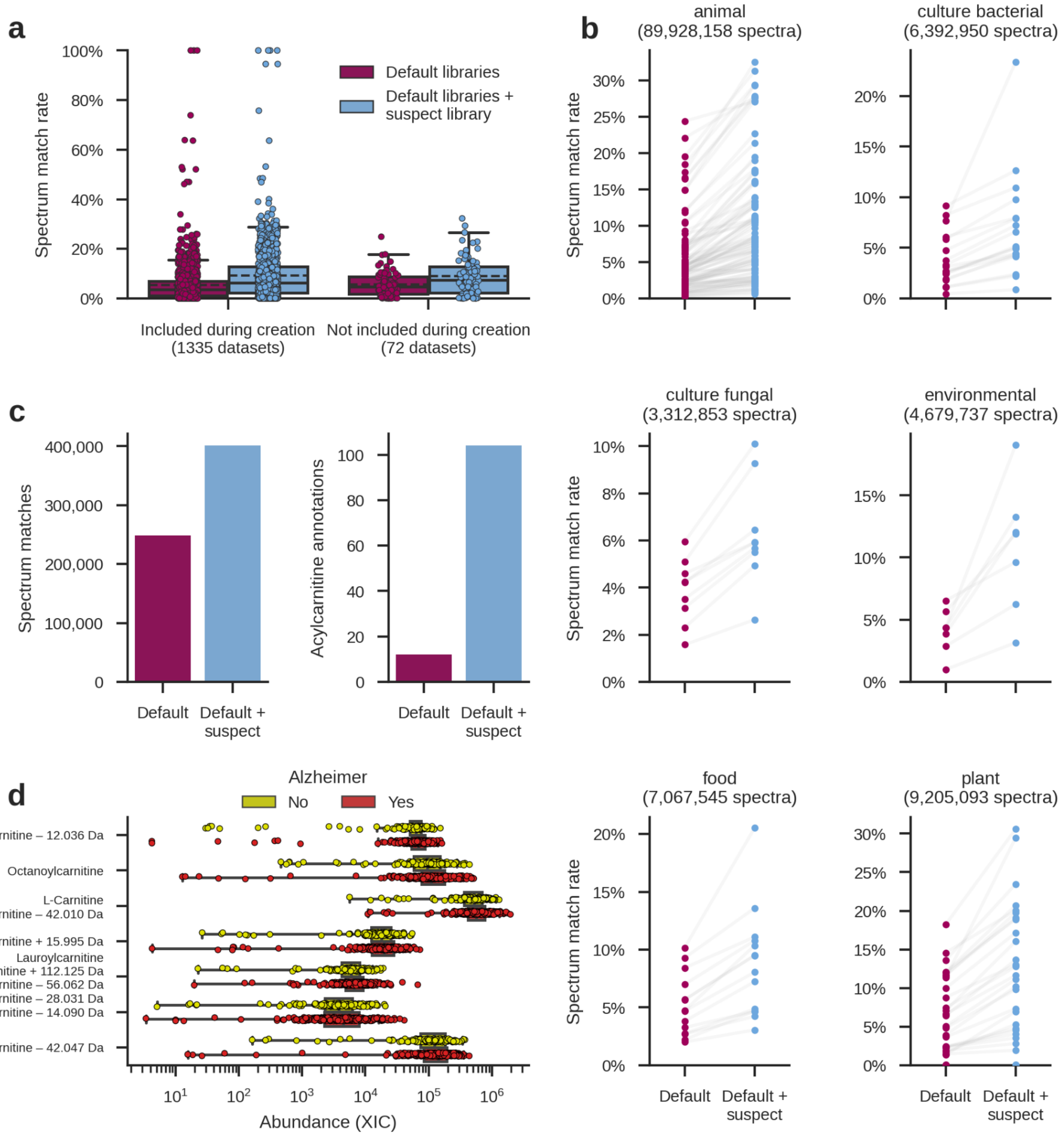


Figure 3. Impact of the nearest neighbor suspect spectral library on spectrum matches to enable the formulation of structural hypotheses. a. The MS/MS spectrum match rate with and without the suspect library for 1407 public datasets on GNPS/MassIVE. **b.** The MS/MS spectrum match rate for different types of datasets with and without the suspect library. The data comes from 45,845 raw files in 179 datasets with known sample types recorded using the ReDU metadata system.[27] **c.** MS/MS matches to an untargeted metabolomics human brain dataset from Alzheimer's disease patients (n=360) and healthy subjects (n=154) with and without the suspect library. **d.** Differentially abundant carnitines for Alzheimer's disease patients (Benjamini-Hochberg corrected p-value < 0.05). The suspect library was able to identify four additional mass spectrometry features as acylcarnitines, which would have remained unannotated matched only against the default GNPS libraries.

We observed significant abundance differences for six acylcarnitines, as well as carnitine, when comparing brain metabolites between groups with and without Alzheimer's diagnosis (**Figure 3d**). Three of those—carnitine, octanoylcarnitine, and lauroylcarnitine—could be annotated using the default GNPS libraries, while the remaining four metabolites could only be identified as acylcarnitines using the suspect library. The annotations of these seven metabolites were covered by three spectral library matches and eight suspect matches. When multiple spectra matched against the same metabolite, these annotations reinforced each other. For example, the carnitine annotation co-occurs with a match to the acetylcarnitine suspect with a loss of 42.010 Da. In this case, 42.010 Da corresponds to the mass of acetylation, which is lost in the suspect annotation, and therefore the suspect MS/MS spectrum represents carnitine itself. The four suspects that would otherwise remain unassigned as potential acylcarnitines are hexanoylcarnitine with loss of 12.036 Da (3-hydroxybutyrylcarnitine, **Figure 2a**), hexanoylcarnitine with addition of 15.995 Da (3-hydroxyhexanoylcarnitine, **Figure 2a**), decanoyl-L-carnitine with loss of 14.090 Da (-3C,-10H,+2O in the acyl chain), and decanoyl-L-carnitine with loss of 42.047 Da (-3C,-6H in the acyl chain). The first two suspects are related 3-hydroxy acylcarnitines that have 3-hydroxy-butyrate and 3-hydroxy-hexonate as the acyl side chain. The other two suspects are consistent with DC7:1 and C7:0 fatty acids.[29] These observations provide additional support that there are different fatty acids—now also including 3-hydroxy and odd-chain fatty acids—that are transported as carnitine derivatives in Alzheimer's disease brains in comparison to healthy brains.[30,31]

Discussion

The annotation of untargeted metabolomics data is based on reference spectral libraries. However, because many known compounds and previously undiscovered analogs of compounds are unavailable as reference standards, alternative approaches are required to interpret such MS/MS spectra. Here we have introduced a data-driven approach to compile an extensive nearest neighbor suspect spectral library. This library consists of 87,916 unique MS/MS spectra and can be freely downloaded as a Mascot generic format (MGF) file from the GNPS website. Additionally, through its direct integration in the spectral library searching and molecular networking functionality on the GNPS platform, the scientific community can incorporate the nearest neighbor suspect spectral library in their data analyses to formulate structural hypotheses.

Entries in the nearest neighbor suspect spectral library are not obtained by measuring pure reference standards. Therefore, it is important to consider that, initially, the exact molecular structure of the suspects is undetermined. Nevertheless, the suspect library includes essential information that can help to interpret MS/MS data that would otherwise remain entirely unexplored. Additionally, all of the spectra that are part of the suspect library have been detected experimentally and occur in biological data. In contrast, only a minority of the compounds contained in reference spectral libraries are actually observed in public data, indicating a mismatch between the laborious reference library creation efforts and the practical needs of metabolomics researchers. Consequently, incorporating the nearest neighbor suspect spectral library significantly increases the spectrum match rate across a wide variety of sample types. We have demonstrated how careful investigation of the suspects can provide highly detailed interpretations, and we anticipate that similar community contributions will be used to add and confirm further suspect annotations. Finally, when future studies uncover biologically relevant suspects, their molecular identities, including the location of modifications and stereochemical features, might be refined by measuring orthogonal properties, such as collision cross-section by ion mobility spectrometry or using genome mining, when possible. Ultimately, as is the case for all spectrum annotations, experimental validation of the complete molecular stereostructure requires either a reference standard or further isolation followed by structure elucidation by NMR, X-ray crystallography, or cryogenic electron microscopy experiments.

Acknowledgements

This research was supported in part by BBSRC-NSF award 2152526. This research was supported in part by National Institutes of Health awards R01 GM107550, U19 AG063744, U01AG061359, R03 CA211211, P41 GM103484, T32 HD123456. This research was supported in part by federal award DE-SC0021340 subaward 1070261-436503. This research was supported in part by the Gordon and Betty Moore Foundation through grant GBMF7622. This research was supported in part by the Intramural Research Program of National Institute of Environmental Health Sciences of the National Institutes of Health (ES103363-01). ELS & TK acknowledge funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006. MW was partially supported by the US Department of Energy Joint Genome Institute operated under Contract No. DE-AC02-05CH11231. DP was supported by the Deutsche Forschungsgemeinschaft (DFG) through the CMFI Cluster of Excellence (EXC 2124). SAK was supported by the Fund for Financing Science and Supporting Innovation under the Ministry of Innovative Development of the Republic of Uzbekistan. KBK was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (NRF-2020R1C1C1004046). HM-R acknowledges the Brazilian National Council for Scientific and Technological Development (CNPq, #142014/2018-4) and the Brazilian Fulbright Commission for the scholarships provided. JJJvdH was supported by an ASDI eScience grant from the Netherlands eScience Center (ASDI.2017.030). COD was supported by EMBL core funds. The Alzheimer's disease metabolomics data was funded wholly or in part by the Alzheimer's Gut Microbiome Project (AGMP) NIH grant U19AG063744 awarded to Dr. Kaddurah-Daouk at Duke University in partnership with a large number of academic institutions. More information about the project and the institutions involved can be found at https://alzheimergut.org/meet-the-team/.

Author contributions statement

PCD conceptualized and supervised the work. COD and CMA helped transfer and convert data from MetaboLights. WB, MW, and PCD created the methodology to compile the nearest neighbor suspect spectral library from molecular networking data. WB, JMG, and MW developed the software. WB, NEA, SPT, SAK, AAA, PWPG, AMCR, JMG, AKJ, TK, HM-R,

MJM, LFN, MP, DP, RS, RS, ELS, and JJJvdH validated entries in the suspect spectral library and evaluated its identification performance. NEA, SPT, SAK, AAA, and PWPG provided case studies to demonstrate the utility of the tool. MW provided computational resources. CMA, CO, MP, and JZ performed data curation. WHG supervised acquisition of the *Moorena bouillonii* MS/MS data. KBK, HWK, and HY acquired the medicinal plants from the Korean Pharmacopeia MS/MS data. AAA and AM acquired the HomeCHEM MS/MS data. RFKD supervised acquisition of the ROSMAP metabolomics data and links to ADMC and AMP-AD consortia. MJM, MP, KCW, and JZ processed and prepared the ROSMAP samples and acquired the MS/MS data. WB, NEA, SPT, AAA, PWPG, CMA, MJM, and PCD wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests statement

PCD is an advisor to Cybele and co-founder and scientific advisor to Ometa and Enveda, with prior approval by UC San Diego. MW is a co-founder of Ometa Labs LLC. AAA, AVM are founders of Arome Science Inc. CMA is a consultant for Nuanced Health. JJJvdH is a member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy. RKD is an inventor on several patents in the metabolomics field and holds founder equity in Metabolon, Chymia, and PsyProtix.

Methods

Integration of Metabolights into GNPS/MassIVE

As a joint effort of the European Bioinformatics Institute (EMBL-EBI) and the GNPS/MassIVE teams, approximately 10,000 LC-MS/MS samples acquired in positive ion mode were imported from MetaboLights[14] into the GNPS/MassIVE repository by mirroring relevant files from MetaboLights on GNPS/MassIVE. These files represent over a hundred studies containing data from biologically diverse backgrounds, including but not limited to human, fungus, various bacterial and microbial species, and ecological samples. The data consist of both metabolomics and lipidomics samples.

GNPS living data molecular networking

The nearest neighbor suspect spectral library was derived from molecular networking results as performed by GNPS's "living data" functionality, which periodically reanalyses all publicly available untargeted metabolomics data on GNPS/MassIVE.[10] The living data analysis (https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=25cc4f9135c6428aabe1f41a9e54c369&view=advanced_view; update performed on November 17, 2020) includes results for 1335 datasets, corresponding to 438,703 annotations from spectral library searching against the default GNPS spectral libraries, and a molecular network consisting of 13,003,591 spectrum pair edges.

The living data analysis consists of two phases of spectrum clustering using MS-Cluster[32] and molecular networking. First, spectra were networked within each individual dataset. Per-dataset molecular networking outputs are available on the MassIVE repository with dataset identifier MSV000084314 (https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=25cc4f9135c6428aabe1f41a9e54c369&view=advanced_view). Next, a second round of molecular networking was performed on the combined consensus spectra for all datasets generated from the first molecular network (GNPS task: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4f69e11bfb544010b2c4225a255f17ba).

Spectra were preprocessed by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m*/*z.* Only the top 6 most abundant ions in every 50 *m*/*z* window were retained. The first round of molecular networking used a precursor mass tolerance of 2.0 *m*/*z*, a fragment mass tolerance of 0.5 *m*/*z*, and three rounds of MS-Cluster clustering with mixture probability threshold 0.05. Thresholds for the second round of molecular networking were modified due to computational and memory constraints, and consisted of a precursor mass tolerance of 0.1 *m*/*z* and a fragment mass tolerance of 0.1 *m*/*z*. The molecular networking used a minimum cosine similarity of 0.8, minimum six matched peaks, only considered clusters that consist of at least two MS/MS spectra, and retained the ten strongest edges for each node in the molecular network.

Spectral annotations were obtained through spectral library searching against the default GNPS spectral libraries (GNPS Collections Bile Acid Library 2019,[10] CASMI,[33] Dereplicator Identified MS/MS Spectra,[34] GNPS Collections Miscellaneous,[10] Pesticides, EMBL Metabolomics Core Facility,[35] Faulkner Legacy Library provided by Sirenas MD, GNPS Library,[10] NIH Clinical Collection 1, NIH Clinical Collection 2, NIH Natural Products Library Round 1,[36] NIH Natural Products Library Round 2,[36] Pharmacologically Active Compounds in the NIH Small Molecule Repository, GNPS Matches to NIST14,[10] PhytoChemical Library, FDA Library Pt 1, FDA Library Pt 2, HMDB,[37] LDB Lichen Database,[38] Massbank Spectral Library,[39] Massbank EU Spectral Library, MIADB Spectral Library,[40] Medicines for Malaria Venture Pathogen Box, Massbank NA Spectral Library, Pacific Northwest National Lab Lipids,[41] ReSpect Spectral Library,[42] Sumner/Bruker), which contained 221,224 reference MS/MS spectra (June 2021). Settings for the living data spectral library searching step included a precursor ion tolerance of 2.0 *m*/*z*, a fragment ion tolerance of 0.5 *m*/*z*, a minimum cosine similarity of 0.7, and minimum six matched peaks.

Nearest neighbor suspect spectral library creation

High-quality MS/MS spectra were extracted from the GNPS living data molecular network to compile the nearest neighbor suspect spectral library. Suspects were derived from spectrum pairs for which only one of the spectra was identified during spectral library searching and both spectra have a non-zero precursor mass difference. In this case, the unidentified spectrum was included in the nearest neighbor suspect spectral library, as it corresponds to a previously unknown molecule that is structurally related to the reference molecule identified using spectral library searching. Strict filtering thresholds were used to avoid inclusion of incorrect entries: spectrum–spectrum matches required a maximum precursor mass tolerance of 20 ppm, a minimum cosine similarity of 0.8, and minimum six matched ions.

To homogenize and extend the information available for the suspects, their molecular formulas were determined using SIRIUS (version 4.5.2)[16] with a precursor mass tolerance of 10 ppm for Orbitrap spectra and 25 ppm for Q-TOF spectra. Additionally, the observed precursor mass differences were calibrated and matched to putative modification explanations contained in the UNIMOD database[17] and a manually compiled list of modifications and their mass differences (**Supplementary Table 3**). Suspects whose delta mass occurred fewer than ten times were discarded, as true modifications are expected to occur repeatedly for different molecules, while suspects with infrequent mass differences more likely correspond to spurious matches.

To ensure that the provenance of the suspects to the matched reference molecules compared to which they are annotated based on spectral similarity is properly understood, their names are of the form: “Suspect related to [*compound name*] (predicted molecular formula: [*molecular formula*]) with delta *m*/*z* [*positive (addition) or negative (loss) delta m/z*] (putative explanation: [*modification*]).” In case multiple propagations to different reference spectra are available, information for all matches is included.

Spectrum annotation using the nearest neighbor suspect spectral library

The spectrum annotation performance of the nearest neighbor suspect spectral library was assessed by large-scale spectral library searching using the default GNPS spectral libraries excluding and including the nearest neighbor suspect spectral library on 1407 public datasets available on GNPS/MassIVE, consisting of a combined 592 million MS/MS spectra. Of these datasets, 1335 datasets were also included in the GNPS living data analysis from which the nearest neighbor suspect spectral library was compiled (521 million MS/MS spectra; see above) and 72 datasets were deposited at a later date and can be considered a completely independent test set (72 million MS/MS spectra).

Spectral library searching was performed on the GNPS platform, with the following task identifiers (200 datasets per task):

- Using the default GNPS spectral libraries only:
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=308b3393a2b2401e8c9b562152531b4c
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=18cf4e521f9b4124af54d7e3d837a888
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c0249eb6a52e4ea993b03de90a509b35
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=debd3bbb51f6490394e905e13779f295
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8cdb4d7d1a784f5bb4f99e4c31564cd1

    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a9e7e4b1b8104416a39142fd6072e02a
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=334ed0d944844e90b71d6151d4e74263
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b55aef34c0bd4d78a1f3952f7c49a52c
- Using the default GNPS spectral libraries and the nearest neighbor suspect spectral library:
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=064be855f46e407f9f5fcbe652c8b9d5
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d243afb8f233490886bb8ab5eedcf8b8
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=febab54db7a14af6b451ab5e5789785f
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=eba0dfe63a464b0a924fd5e373917b37
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=95b541cb3be54d08a0b14367554630ca
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1df48f2dc7c443fc9364dfc8b28f6b47
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b7f8c3d47a464b53ab94f1780f56c893
    - https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=50e3d8ae4e004f989862fcc9d1353534

All searches used a precursor mass tolerance of 2.0 *m*/*z*, a fragment mass tolerance of 0.5 *m*/*z*, a minimum cosine similarity of 0.8, and minimum 6 matched peaks. Other options were kept at their default values.

Evaluation of apratoxin suspects

Mass spectrometry analysis. Apratoxin suspects were investigated in the context of *Moorena bouillonii*, a tropical marine benthic filamentous cyanobacterium. The mass spectrometry data were derived from both field-collected and laboratory-cultured biomass of *Moorena bouillonii* (MassIVE dataset identifier MSV000086109). A number of collections are represented in this dataset, including those originating from sites around Guam, Saipan (Commonwealth of the Northern Mariana Islands), Palmyra Atoll, Papua New Guinea, American Samoa, Kavaratti (Lakshadweep, India), the Paracel Islands (Xisha, China), the Solomon Islands, and the Red Sea (Egypt). The biomass from each of the samples was extracted using 2:1 dichloromethane and methanol. The crude extracts were concentrated and resuspended in acetonitrile, followed by a desalting protocol using C18 SPE with acetonitrile. Samples were then resuspended in methanol containing 2 µM sulfamethazine as an internal standard. Untargeted metabolomics was performed using an UltiMate 3000 liquid chromatography system (Thermo Scientific) coupled to a Maxis Q-TOF (Bruker Daltonics) mass spectrometer with a Kinetex C18 column (Phenomenex). Data were collected in positive ion mode using data-dependent acquisition. All solvents used were LC-MS grade.

Molecular networking. Molecular networking and spectral library searching were performed using the GNPS platform as described above (GNPS task identifier https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=5c41693f607d4b4cabbcfbbf5b9bcf86). Settings included a precursor mass tolerance of 2.0 *m*/*z*, fragment mass tolerance of 0.5 *m*/*z*, minimum cosine similarity of 0.7, and minimum 6 matched peaks. Data visualization was performed using the Metabolomics USI interface[43] and spectrum–spectrum matches were evaluated manually to develop hypotheses regarding the structure of apratoxin analogs that were annotated using the nearest neighbor suspect spectral library.

Cyanobacterial culture. *Moorena bouillonii* PNG05-198 was initially collected by SCUBA in 3–10 m of water off the coast of Pigeon Island, Papua New Guinea (S4 16.063' E152 20.266’) in May 2005. Live cultures have been maintained in SWBG-11 media under laboratory conditions at 27°C and a 16/8 h light/dark schedule. Biomass for *Moorena bouillonii* was obtained through ongoing laboratory culture.

Extraction and isolation of apratoxins. The cultured biomass was extracted using 2:1 CH_2Cl_2/MeOH affording 241.4 g of organic extract. The extract was then subjected to vacuum liquid chromatography (VLC) on silica gel (type H, 10–40 μm) using normal phase solvents in a stepwise gradient of hexanes/EtOAc and EtOAc/MeOH, resulting in nine fractions (A-I). The fraction eluting with 25% MeOH/75% EtOAc (fraction H) had a mass of 21.6 mg. This fraction was found to have the characteristic MS/MS signatures of the apratoxins and was selected for further purification using reversed-phase HPLC. A Phenomenex Kinetex C18 5*μ*m 100Å 100 x 4.6 mm column with a 3 mL/min was used to obtain 1.2 mg of semipure suspect (apratoxin A - 26.015 Da) and 2.1 mg of semipure apratoxin A.

NMR spectroscopy: 1H NMR and 2D NMR spectra were obtained on a Bruker Advance III DRX-600 NMR with a 1.7 mm dual tune TCI cryoprobe (600 MHz and 150 Mhz for ^{1}H and ^{13}C, respectively). NMR spectra were referenced to residual solvent $CDCl_3$ signals as internal standards. NMR spectra were processed using MestReNova (Mnova 14.2.3, Mestrelab Research).

Evaluation of azithromycin suspects

Mass spectrometry analysis. The presence of azithromycin suspects was investigated using human breast milk data (MassIVE dataset identifier MSV000081432).[25] Human milk samples were extracted using 80:20 methanol and water. Untargeted metabolomics was performed using an UltiMate 3000 liquid chromatography system (Thermo Scientific) coupled to a Maxis Q-TOF (Bruker Daltonics) mass spectrometer with a Kinetex C18 column (Phenomenex). Samples were run using a linear gradient of mobile phase A (water 0.1% formic acid (v/v)) and phase B (acetonitrile 0.1% formic acid (v/v)). A representative linear gradient consisted of 0-0.5 min isocratic at 5% B, 0.5-8.5 min 100% B, 8.5-11 min isocratic at 100% B, 11-11.5 min 5% B, and 11.5-12 min 5% B. Data were collected in positive ion mode using data-dependent acquisition. All solvents used were LC-MS grade.

Molecular networking. Molecular networking and spectral library searching were performed using the GNPS platform as described above (GNPS task identifier https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e91e2e44e3234f08bb3d7f3f16d5f782). Settings included a precursor mass tolerance of 0.02 *m*/*z*, fragment mass tolerance of 0.02 *m*/*z*, minimum cosine similarity of 0.6, and minimum 5 matched peaks. Data visualization was performed using the Metabolomics USI interface[43] and spectrum–spectrum matches were evaluated manually to interpret the azithromycin suspect.

Evaluation of flavonoid suspects

Mass spectrometry analysis. Untargeted metabolomics data for medicinal plants listed in the Korean Pharmacopeia were used to investigate flavonoid suspects (MassIVE dataset identifier MSV000086161). Samples were extracted using methanol. Untargeted metabolomics was performed using an Acquity liquid chromatography system coupled to a Xevo G2 Q-TOF (Waters) mass spectrometer with a BEH C18 column at 40°C (Waters Corp.; 50 mm; 2.1 mm; 1.7 µm particle size). Water (solvent A) and acetonitrile (solvent B) were used as mobile phase, both with 0.1% formic acid, and a method of 20 min (linear gradient), flow 0.3 mL/min was performed using the following settings: 0-14 min. from 5 to 95% B; 14-17 min, 95% B; 17-17.1 min from 95% to 5% B; 17.1-20 min, 5% B for equilibration of the column for the next sample. Data were collected in positive and negative ion modes using data-dependent acquisition. All solvents used were LC-MS grade.

Molecular networking. Molecular networking and spectral library searching were performed in the GNPS platform as described above (GNPS task identifier https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=38a1bd60bd094c8a97cf49d822e7f853). Settings included a precursor ion mass tolerance of 2.0 *m/z*, fragment ion mass tolerance of 0.5 *m/z*, minimum cosine similarity of 0.7, and minimum 6 matched peaks. Data visualization was performed using the Metabolomics USI interface[43] and spectrum–spectrum matches were evaluated manually to interpret the flavonoid suspects.

Home environment personal care products

Mass spectrometry analysis. The presence of polymeric suspects was investigated in the context of the HOMEChem project, a study of the indoor chemical environment (MassIVE dataset identifier MSV000083320).[26] For full details on the experimental set-up, see Aksenov et al. (2021).[26] Briefly, scripted activities, including cleaning and cooking, were performed in a controlled home environment. Sample collection consisted of swabbing different locations in the test house. Untargeted metabolomics was performed using a Vanquish liquid chromatography system (Thermo Scientific) coupled to a QExactive Orbitrap (Thermo Scientific) mass spectrometer with a Kinetex C18 column (Phenomenex). The mobile phase used was water (phase A) and acetonitrile (phase B), both containing 0.1% formic acid (Fisher Scientific, Optima LC/MS), employing the following gradient: 0-1 min 5% B, 1-8 min 100% B, 8-10.9 min 100% B, 10.9-11 min 5% A, 11-12 min 5% B. Data were collected in positive ion mode using data-dependent acquisition. All solvents used were LC-MS grade.

Molecular networking. Molecular networking and spectral library searching were performed using the GNPS platform as described above (GNPS task identifier https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=890e39f28140470ab0598c77cc5c048e). Settings included a precursor mass tolerance of 0.02 *m*/*z*, fragment mass tolerance of 0.02 *m*/*z*, minimum cosine similarity of 0.7, and minimum 6 matched peaks.

Alzheimer's disease acylcarnitine analysis

Mass spectrometry analysis. The presence of acylcarnitine suspects was investigated in the context of the Religious Orders Study/Memory and Aging Project (ROSMAP) to study Alzheimer's disease (MassIVE dataset identifier MSV000086415).[28] Untargeted metabolomics was performed on human brain samples from 514 individuals with and without Alzheimer's disease (360 Alzheimer's disease patients, 154 healthy subjects). Human brain tissue samples were placed into tubes with 800 µl of a 1:1 mixture of H_2O (Optima LC-MS grade W64) and MeOH (100%) containing 1 µM of sulfamethazine. The samples were homogenized using a Qiagen TissueLyser II at 25 Hz for 5 minutes, then centrifuged at 14,000 relative centrifugal force for 5 minutes before being incubated for a period of 30 minutes at -20 °C. A 200 µl aliquot of supernatant from each sample was transferred into a 96-well plate and vacuum concentrated to dryness via centrifugal lyophilization (Labconco Centrivap). Once dried, the samples were stored at -80 °C until LC-MS was performed. Untargeted metabolomics was performed using a Vanquish liquid chromatography system (Thermo Scientific) coupled to a QExactive (Thermo Scientific) mass spectrometer with a C18 column (Phenomenex Kinetex 1.7 µm C18 100 Å LC Column 50 x 2.1 mm). The mobile phase used was LC-MS grade water (phase A) and LC-MS grade acetonitrile (phase B), both containing 0.1% formic acid (Fisher Scientific, Optima LC-MS), with a flow rate set to 0.5 mL/min. Samples were injected at 95%A:5%B, which was held for 1 minute, before ramping up to 100%B over 7 minutes, which was held for 0.5 minutes before returning to starting conditions. Data were collected in positive ion mode using data-dependent acquisition to acquire MS full scan spectra, followed by MS/MS spectra of the top 5 most abundant ions. Precursor ions were fragmented once before being added to an exclusion list for 30 seconds.

Data analysis. Spectral library searching was performed using the GNPS platform as described above using the default GNPS spectral libraries only (GNPS task identifier https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b55aef34c0bd4d78a1f3952f7c49a52c) and including the nearest neighbor suspect spectral library (GNPS task identifier https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=50e3d8ae4e004f989862fcc9d1353534). Settings included a precursor mass tolerance of 2.0 *m*/*z*, fragment mass tolerance of 0.5 *m*/*z*, minimum cosine similarity of 0.8, and minimum 6 matched peaks. Raw MS data visualization was performed using the GNPS Dashboard.[44] Spectrum annotations corresponding to carnitines were extracted by filtering on "carnitine" in the compound name. Different spectrum annotations with near-identical precursor *m*/*z* (precursor *m*/*z* tolerance 100 ppm) and retention time (retention time tolerance 20 seconds) were merged. Feature abundances were obtained by computing extracted ion chromatograms (XICs) with *m*/*z* tolerance 100 ppm and retention time tolerance 20 seconds for all uniquely annotated acylcarnitines across all 514 raw files. Next, the Spearman correlations between all acylcarnitine XICs and the subjects' CERAD scores (a measure of Alzheimer's disease progression, with 1 indicating "definite" Alzheimer's disease and 4 indicating "no" Alzheimer's disease) were calculated and the correlation coefficients and associated p-values were recorded. Multiple testing correction of the p-values was performed using the Benjamini-Hochberg procedure, and acylcarnitines with a corrected p-value below 0.05 were considered to be significantly associated with Alzheimer's disease. For visualization purposes the four-scale CERAD score was binarized by considering a CERAD score of 1 or 2 to correspond to positive Alzheimer's disease patients, and a CERAD score of 3 or 4 to correspond to healthy individuals.

Data availability

All of the data used to compile the nearest neighbor suspect spectral library are publicly available through GNPS/MassIVE. The living data covering the 1335 public datasets from which the nearest neighbor suspect spectral library was derived are available at https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=25cc4f9135c6428aabe1f41a9e54c369&view=advanced_view (update performed on November 17, 2020). Spectrum annotations for the ROSMAP study are available via the AD Knowledge Portal (https://adknowledgeportal.org). The AD Knowledge Portal is a platform for accessing data, analyses, and tools generated by the Accelerating Medicines Partnership (AMP-AD) Target Discovery Program and other National Institute on Aging (NIA)-supported programs to enable open-science practices and accelerate translational learning. The data, analyses, and tools are shared early in the research cycle without a publication embargo on secondary use. Data is available for general research use according to the following requirements for data access and data attribution (https://adknowledgeportal.org/DataAccess/Instructions). For access to content described in this manuscript see: https://doi.org/10.7303/syn30255033.1.

The nearest neighbor suspect spectral library is freely available under the CC0 license at https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=GNPS-SUSPECTLIST and archived on Zenodo at https://doi.org/10.5281/zenodo.6512084. Additionally, it can be used for any data analysis task on GNPS by selecting it from the CCMS_SpectralLibraries > GNPS_Propogated_Libraries > GNPS-SUSPECTLIST > GNPS-SUSPECTLIST.mgf path in the GNPS file selector dialog. Step-by-step instructions are also provided on the GNPS Documentation website at https://ccms-ucsd.github.io/GNPSDocumentation/browselibraries/#nearest-neighbor-suspect-spectral-library.

Individual spectra are accessible by their Universal Spectrum Identifiers (USIs).[43,45] The spectra displayed in **Figure 2**, **Supplementary Figure 4**, and **Supplementary Figure 5** are:

- Hexanoylcarnitine, C6:0: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00003135669
- 3-Hydroxybutyrylcarnitine: mzspec:MSV000082049:20_51:scan:106
- Hexenoylcarnitine, C6:1: mzspec:MSV000085561:011c:scan:2864
- Benzoylcarnitine: mzspec:MSV000085561:010c:scan:2829
- 3-Hydroxyhexanoylcarnitine: mzspec:MSV000085561:018b:scan:2609
- Dodecanedioylcarnitine, C12-DC: mzspec:MSV000082650:M031_48:scan:1501
- Apratoxin A: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00000424840
- Apratoxin D: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00000424841
- Apratoxin F: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00000070287
- Apratoxin C: mzspec:MSV000086109:BF9_BF9_02_57124.mzML:scan:722
- Apratoxin A – 30.010 Da: mzspec:MSV000086109:BD5_dil2x_BD5_01_57213:scan:760
- Apratoxin F – 30.010 Da: mzspec:MSV000086109:BC11_dil2x_BC11_02_57176:scan:736
- Apratoxin A – 26.015 Da: mzspec:MSV000086109:BD5_dil2x_BD5_01_57213:scan:614
- Apratoxin A – 14.016 Da: mzspec:MSV000086109:BD11_BD11_02_57022:scan:591
- Azithromycin: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00005434451
- 3'-O(desmethyl)azithromycin: mzspec:MSV000084132:Pos_C18_Aq7:scan:977
- Apigenin-8-*C*-hexosylhexoside: mzspec:GNPS:GNPS-LIBRARY:accession:CCMSLIB00004698180
- 7-*O*-methylapigenin-6-*C*-hexoside + 132.042 Da: mzspec:GNPS:TASK-38a1bd60bd094c8a97cf49d822e7f853-spectra/specs_ms.mgf:scan:1573560
- Apigenin-8-*C*-hexosylhexoside – 30.010 Da: mzspec:GNPS:TASK-38a1bd60bd094c8a97cf49d822e7f853-spectra/specs_ms.mgf:scan:1559636
- Apigenin-8-*C*-hexosylhexoside – 31.991 Da: mzspec:GNPS:TASK-38a1bd60bd094c8a97cf49d822e7f853-spectra/specs_ms.mgf:scan:1559563
- Apigenin-8-*C*-hexosylhexoside – 46.005 Da: mzspec:GNPS:TASK-38a1bd60bd094c8a97cf49d822e7f853-spectra/specs_ms.mgf:scan:1543689

Code availability

Code to extract spectra from the molecular networks and compile the nearest neighbor suspect spectral library, as well as code notebooks to generate the figures and analyses presented in this manuscript are freely available on GitHub at https://github.com/bittremieux/gnps_suspect_library under the open source BSD license. A permanent code archive is available on Zenodo at https://doi.org/10.5281/zenodo.6459282.

All code was implemented in Python 3.8, and uses NumPy (version 1.19.2),[46] SciPy (version 1.5.2),[47] Pandas (version 1.1.3),[48] and statsmodels (version 0.13.1)[49] for scientific data processing, Pyteomics (version 4.4.0)[50] to interface the UNIMOD repository,[17] and matplotlib (version 3.5.1),[51] Seaborn (version 0.11.0),[52] spectrum_utils (version 0.3.4),[53] Jupyter notebooks,[54] and Cytoscape[55] for visualization purposes.

References

1. Sindelar, M. & Patti, G. J. Chemical discovery in the era of metabolomics. *J. Am. Chem. Soc.* **142**, 9097–9105 (2020).
2. Schmid, R. *et al.* Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat. Commun.* **12**, 3832 (2021).
3. Chen, L. *et al.* Metabolite discovery through global annotation of untargeted metabolomics data. *Nat. Methods* **18**, 1377–1385 (2021).
4. Djoumbou-Feunang, Y. *et al.* BioTransformer: A comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminformatics* **11**, (2019).
5. Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
6. Burke, M. C. *et al.* The hybrid search: A mass spectral library search method for discovery of modifications in proteomics. *J. Proteome Res.* (2017) doi:10.1021/acs.jproteome.6b00988.
7. Huber, F. *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Comput. Biol.* **17**, e1008724 (2021).
8. Aisporna, A. *et al.* Neutral loss mass spectral data enhances molecular similarity analysis in METLIN. *J. Am. Soc. Mass Spectrom.* **33**, 530–534 (2022).
9. Treen, D. G., Northen, T. R. & Bowen, B. P. SIMILE enables alignment of fragmentation mass spectra with statistical significance. *bioRxiv* (2021) doi:10.1101/2021.02.24.432767.
10. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
11. 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.* **36**, 960–980 (2019).
12. Remoroza, C. A., Mak, T. D., De Leoz, M. L. A., Mirokhin, Y. A. & Stein, S. E. Creating a mass spectral reference library for oligosaccharides in human milk. *Anal. Chem.* **90**, 8977–8988 (2018).
13. Yan, X. *et al.* Mass spectral library of acylcarnitines derived from human urine. *Anal. Chem.* **92**, 6521–6528 (2020).
14. Haug, K. *et al.* MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, D781–D786 (2013).

15. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2015).
16. Dührkop, K. *et al.* SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
17. Creasy, D. M. & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *PROTEOMICS* **4**, 1534–1536 (2004).
18. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
19. McCann, M. R., George De la Rosa, M. V., Rosania, G. R. & Stringer, K. A. L-Carnitine and acylcarnitines: Mitochondrial biomarkers for precision medicine. *Metabolites* **11**, 51 (2021).
20. Su, X., Han, X., Mancuso, D. J., Abendschein, D. R. & Gross, R. W. Accumulation of long-chain acylcarnitine and 3-hydroxy acylcarnitine molecular species in diabetic myocardium: Identification of alterations in mitochondrial fatty acid processing in diabetic myocardium by shotgun lipidomics. *Biochemistry* **44**, 5234–5245 (2005).
21. Zuniga, A. & Li, L. Ultra-high performance liquid chromatography tandem mass spectrometry for comprehensive analysis of urinary acylcarnitines. *Anal. Chim. Acta* **689**, 77–84 (2011).
22. Luesch, H., Yoshida, W. Y., Moore, R. E., Paul, V. J. & Corbett, T. H. Total structure determination of apratoxin A, a potent novel cytotoxin from the marine cyanobacterium *Lyngbya m ajuscula*. *J. Am. Chem. Soc.* **123**, 5418–5423 (2001).
23. Gutiérrez, M. *et al.* Apratoxin D, a potent cytotoxic cyclodepsipeptide from Papua New Guinea collections of the marine cyanobacteria *Lyngbya majuscula* and *Lyngbya sordida*. *J. Nat. Prod.* **71**, 1099–1103 (2008).
24. Fischbach, M. A. & Clardy, J. One pathway, many products. *Nat. Chem. Biol.* **3**, 353–355 (2007).
25. Thomas, S. P. *et al.* An untargeted metabolomics analysis of exogenous chemicals in human milk and transfer to the infant. *bioRxiv* (2022) doi:10.1101/2022.03.31.486633.
26. Aksenov, A. A. *et al.* The molecular impact of life in an indoor environment. *ChemRxiv* (2021).
27. Jarmusch, A. K. *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).
28. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
29. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.* **50**, S9–S14 (2009).
30. Horgusluoglu, E. *et al.* Integrative metabolomics‐genomics approach reveals key metabolic pathways and regulators of Alzheimer's disease. *Alzheimers Dement.* alz.12468 (2021) doi:10.1002/alz.12468.
31. Jia, L. *et al.* A metabolite panel that differentiates Alzheimer's disease from other dementia types. *Alzheimers Dement.* alz.12484 (2021) doi:10.1002/alz.12484.
32. Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122

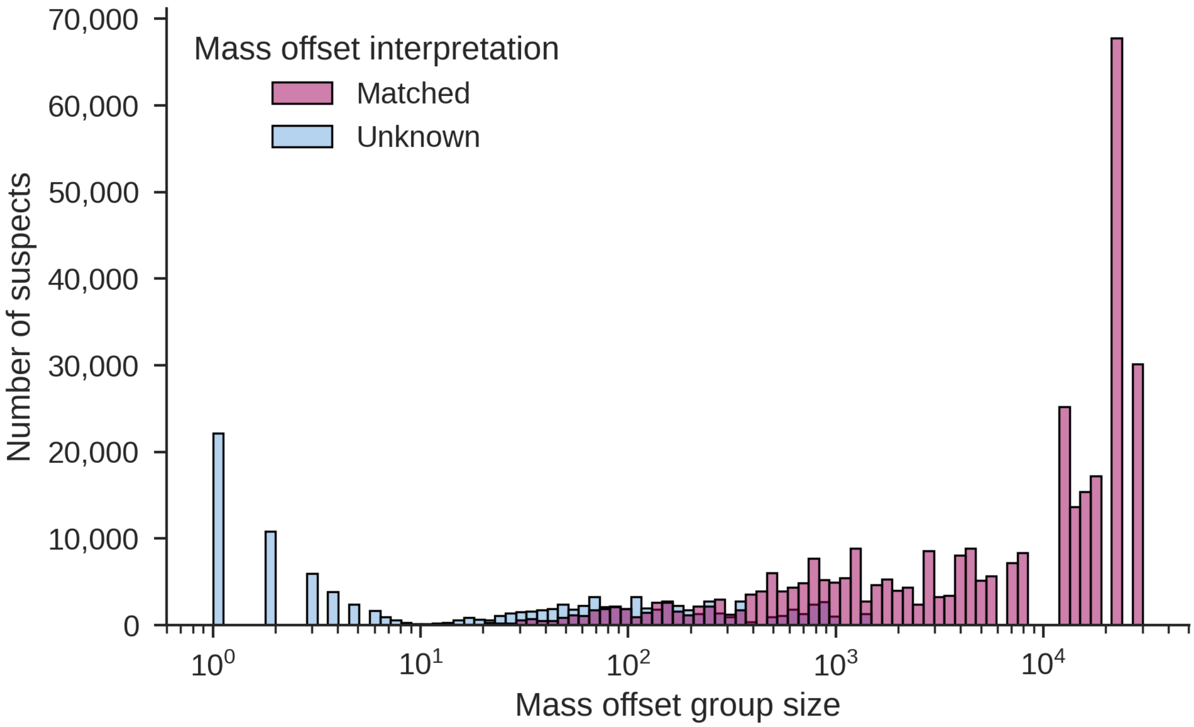
(2008).

33. Schymanski, E. & Neumann, S. The Critical Assessment of Small Molecule Identification (CASMI): Challenges and solutions. *Metabolites* **3**, 517–538 (2013).
34. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2016).
35. Phapale, P. *et al.* Public LC-Orbitrap Tandem Mass Spectral Library for Metabolite Identification. *J. Proteome Res.* **20**, 2089–2097 (2021).
36. Huang, R. *et al.* The NCATS Pharmaceutical Collection: a 10-year update. *Drug Discov. Today* **24**, 2341–2349 (2019).
37. Wishart, D. S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2021).
38. Olivier-Jimenez, D. *et al.* A database of high-resolution MS/MS spectra for lichen metabolites. *Sci. Data* **6**, 294 (2019).
39. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
40. Fox Ramos, A. E. *et al.* Collected mass spectrometry data on monoterpene indole alkaloids from natural product chemistry research. *Sci. Data* **6**, 15 (2019).
41. Kyle, J. E. *et al.* LIQUID: an-open source software for identifying lipids in LC-MS/MS-based lipidomics data. *Bioinformatics* **33**, 1744–1746 (2017).
42. Sawada, Y. *et al.* RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. *Phytochemistry* **82**, 38–45 (2012).
43. Bittremieux, W. *et al.* Universal MS/MS visualization and retrieval with the Metabolomics Spectrum Resolver web service. *bioRxiv* (2020) doi:10.1101/2020.05.09.086066.
44. Petras, D. *et al.* GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nat. Methods* (2021) doi:10.1038/s41592-021-01339-5.
45. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).
46. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
47. SciPy 1.0 Contributors *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* (2020) doi:10.1038/s41592-019-0686-2.
48. McKinney, W. Data structures for statistical computing in Python. in *Proceedings of the 9th Python in Science Conference* (eds. van der Walt, S. & Millman, J.) 51–56 (2010).
49. Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with Python. in *Proceedings of the 9th Python in Science Conference (SciPy 2010)* 92096 (2010). doi:10.25080/Majora-92bf1922-011.
50. Levitsky, L. I., Klein, J. A., Ivanov, M. V. & Gorshkov, M. Pyteomics 4.0: Five years of development of a Python proteomics framework. *J. Proteome Res.* **18**, 709–714 (2019).
51. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
52. Waskom, M. L. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
53. Bittremieux, W. spectrum_utils: A Python package for mass spectrometry data processing and visualization. *Anal. Chem.* **92**, 659–661 (2020).
54. Thomas, K. *et al.* Jupyter Notebooks -- A publishing format for reproducible computational

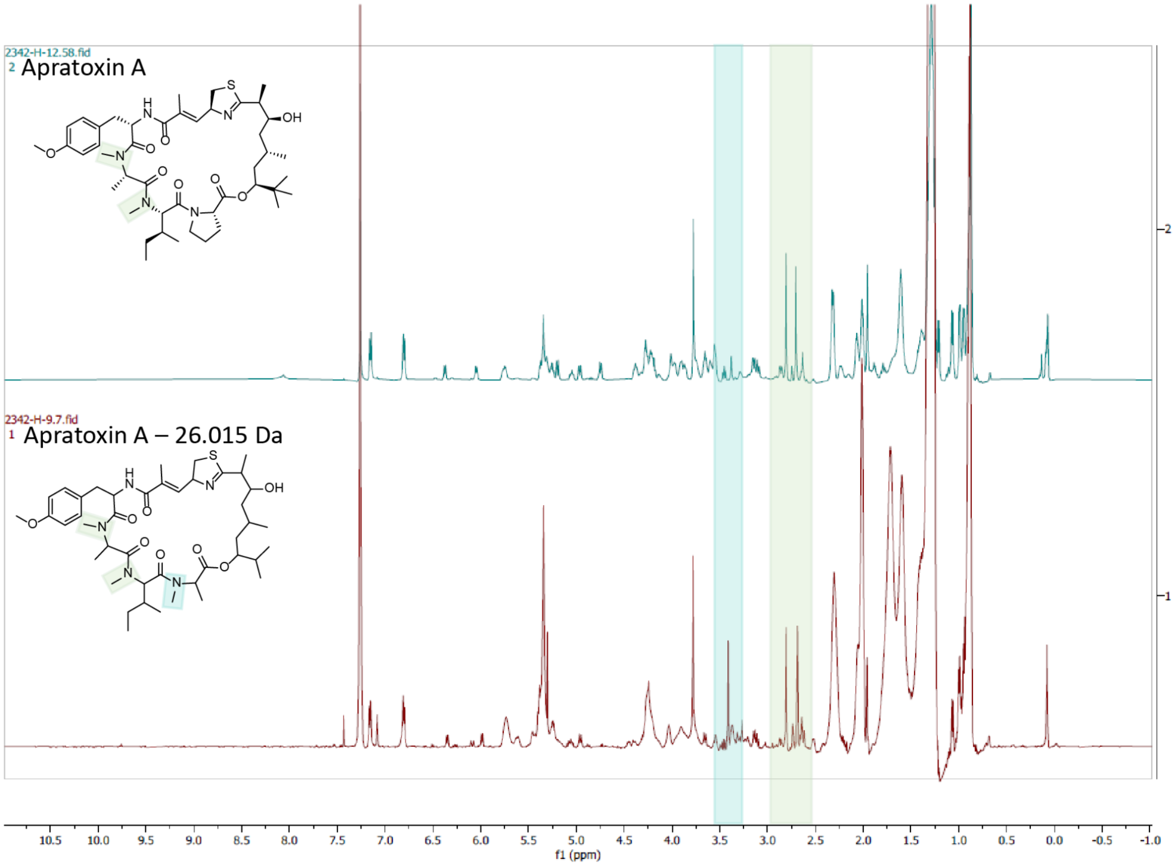
workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90 (IOS Press, 2016).

55. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
56. Aversa, Z. *et al.* Association of infant antibiotic exposure with childhood health outcomes. *Mayo Clin. Proc.* **96**, 66–77 (2020).
57. Hunter, R. P., Koch, D. E., Coke, R. L., Goatley, M. A. & Isaza, R. Azithromycin metabolite identification in plasma, bile, and tissues of the ball python (*Python regius*). *J. Vet. Pharmacol. Ther.* **26**, 117–121 (2003).

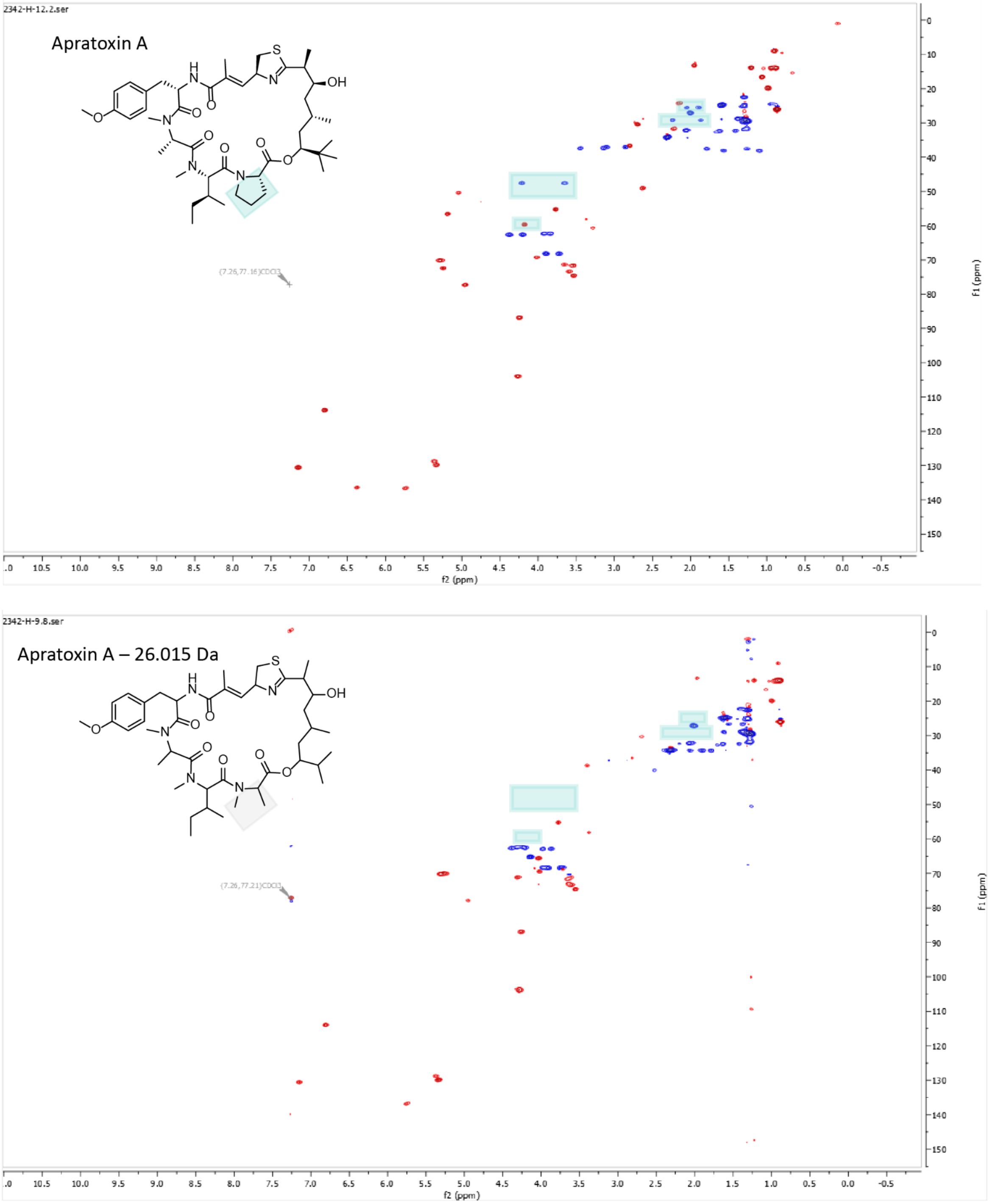
Supporting Information



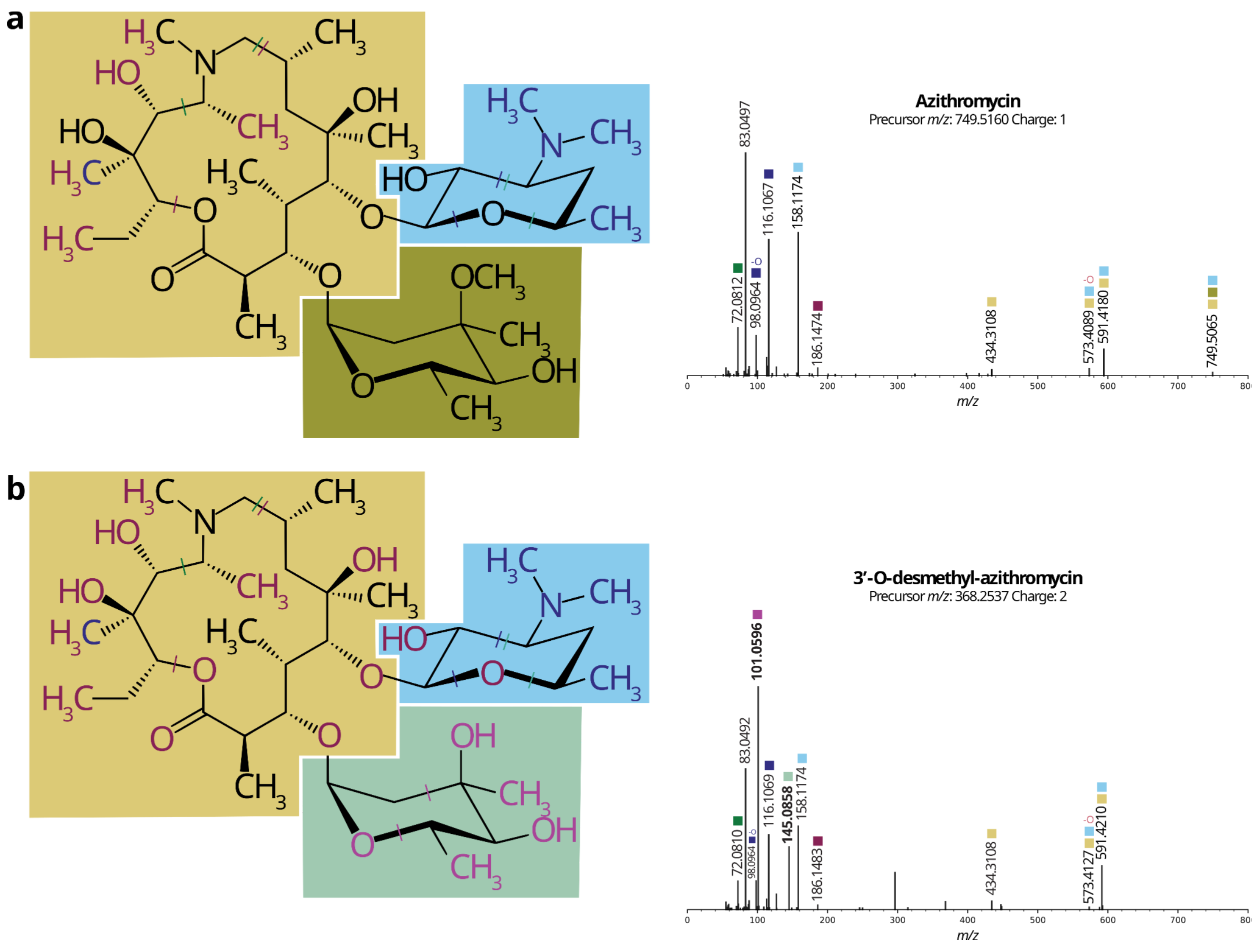
Supplementary Figure 1. Frequency of the observed mass offsets. Several mass offsets occur hundreds to thousands of times, whereas less frequent mass offsets occur only a handful of times. Spectra with delta masses that occur fewer than ten times were not included in the final suspect library. These mass offsets could not be interpreted by matching against modifications in the UNIMOD database[17] and a community curated list of delta masses, and are considered to be non-reproducible mass differences that likely do not correspond to real modifications.



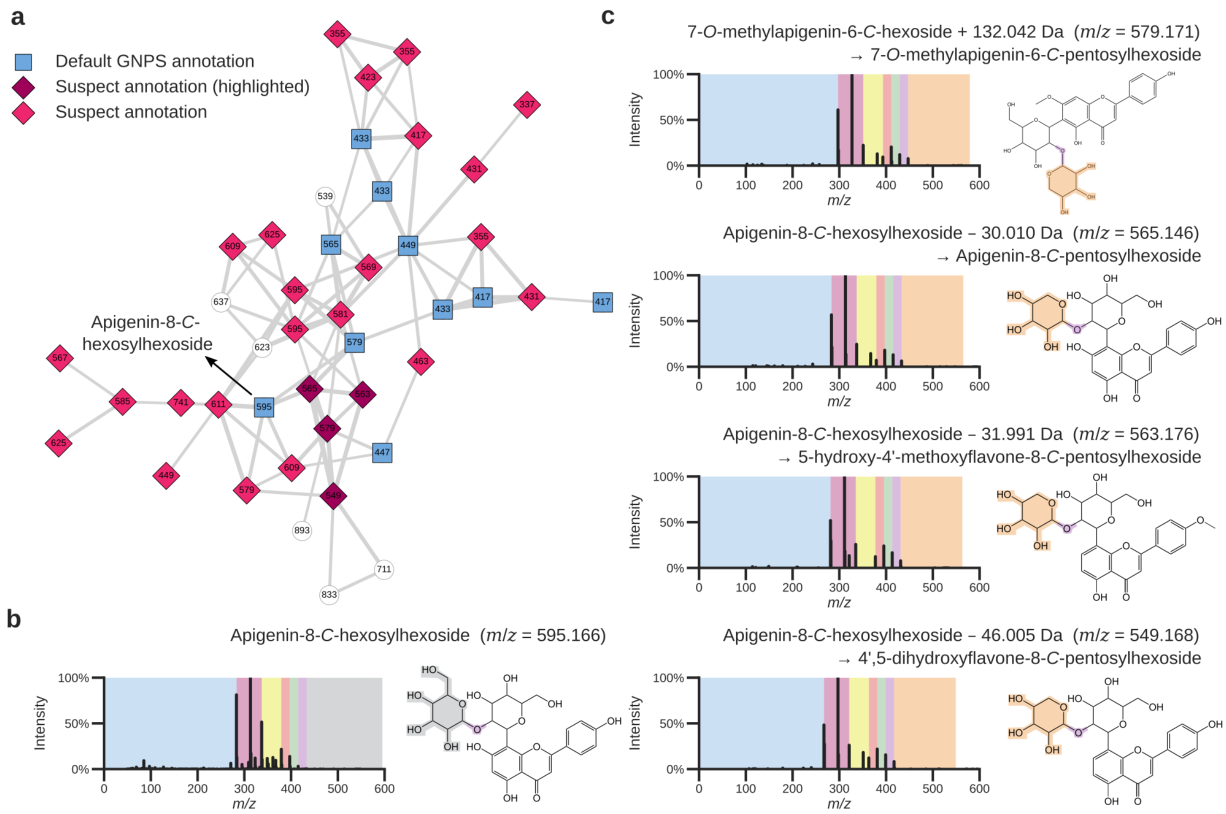
Supplementary Figure 2. Comparison of ^{1}H NMR spectra (600 MHz, $CDCl_3$) of apratoxin A (top) and its related suspect (apratoxin A - 26.015 Da; bottom). Indicated by green shading are the proton signals for the *N*-methyl groups on the *N*-methyl-isoleucine and adjacent *N*-methyl-alanine at 2.71 ppm and 2.81 ppm, respectively. In the suspect there is an additional singlet proton signal observed at 3.41 ppm corresponding to the *N*-methyl-alanine adjacent to the ester bond (turquoise shading).



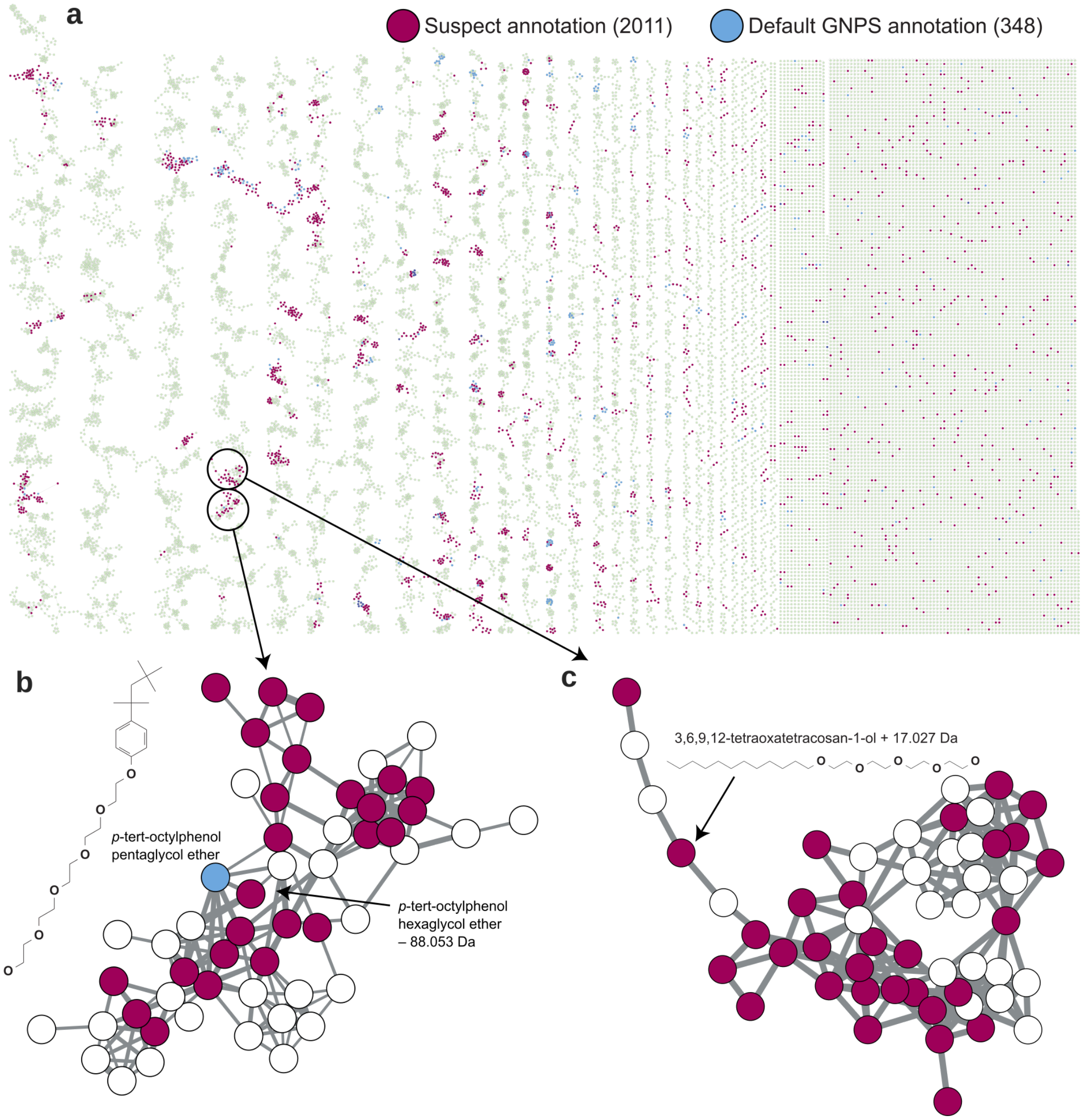
Supplementary Figure 3. Comparison of ^{1}H-^{13}C HSQC spectra (600 MHz, $CDCl_3$) of apratoxin A (top) and its related suspect (apratoxin A - 26.015 Da; bottom). The ^{1}H-^{13}C correlations associated with the proline ring (turquoise boxes) are notably absent in the suspect. Based on the MS/MS fragmentation pattern, the suspect also possesses one less methyl group in the polyketide portion of the molecule: this is possibly explained by an isopropyl rather than a *tert*-butyl group at the initiating terminus, as seen in apratoxin C.



Supplementary Figure 4. Although human breast milk is the gold standard of infant nutrition, the presence of exogenous metabolites—such as food and drugs consumed by the mother—therein is not well understood. This is especially pressing in the case of antibiotics in breast milk, as it is known that antibiotic administration in infancy can cause lasting changes in microbial colonization and host health.[56] A public human breast milk dataset was searched for suspects related to the antibiotic azithromycin (**a**) and found specific azithromycin metabolites, including 3'-*O*-desmethyl-azithromycin (**b**), an azithromycin metabolite previously identified only in snakes.[57]



Supplementary Figure 5. Investigation of suspects from a dataset of medicinal plants listed in the Korean Pharmacopeia. **a.** Flavonoids cluster in a molecular network created from the Korean Pharmacopeia medicinal plants dataset. The reference library hits are shown by the blue squares. The purple and pink diamonds are nodes that represent matches to the nearest neighbor suspect spectral library, with the purple diamonds matching the MS/MS spectra shown in panel c for which structures could be proposed. The white nodes are additional MS/MS spectra within the flavonoids molecular family that could not be annotated, even when utilizing the suspect library. **b.** Reference library annotation of an MS/MS spectrum matching to apigenin-8-*C*-hexosylhexoside. **c.** MS/MS spectra and structural hypotheses of apigenin-8-*C*-hexosylhexoside suspects.


a
● Suspect annotation (2011)
● Default GNPS annotation (348)
b
p-tert-octylphenol
pentaglycol ether
p-tert-octylphenol
hexaglycol ether
– 88.053 Da
c
3,6,9,12-tetraoxatetracosan-1-ol + 17.027 Da

Supplementary Figure 6. Molecular networking of the HOMEChem study to explore the chemistry of a house and how it relates to human activities within.[26] **a.** Inclusion of the suspect library revealed a large portion of the otherwise hidden chemistry, including multiple newly annotated clusters that were found to originate from various skincare-related chemistries, in particular polyether variants. As MS/MS libraries are far from comprehensive, they contain spectra for only a small subset of possible variants of these molecules. This is especially problematic for molecules such as polyethers, as the likelihood of encountering any one particular isomer of many possible variants of polyethers, and related molecules, is very low. **b.** Example of a cluster in the molecular network where multiple spectra could be interpreted based on suspect annotations, while only a single spectrum could be annotated with conventional libraries. **c.** In the majority of cases no annotations were possible at all for skincare ingredient molecules. In contrast, using the suspect library these molecules could be readily identified. All annotations in the cluster are concordant with each other, reinforcing the suspect annotations.