



HAL
open science

Reproducible molecular networking of untargeted mass spectrometry data using GNPS

Allegra Aron, Emily Gentry, Kerry Mcphail, Louis-Félix Nothias, Mélissa Nothias-Esposito, Amina Bouslimani, Daniel Petras, Julia Gauglitz, Nicole Sikora, Fernando Vargas, et al.

► **To cite this version:**

Allegra Aron, Emily Gentry, Kerry Mcphail, Louis-Félix Nothias, Mélissa Nothias-Esposito, et al.. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols*, 2020, 15 (6), pp.1954-1991. 10.1038/s41596-020-0317-5 . hal-03015872

HAL Id: hal-03015872

<https://hal.science/hal-03015872>

Submitted on 1 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Title:** Reproducible Molecular Networking Of Untargeted Mass Spectrometry Data Using
2 GNPS.

3
4 **Authors:** Allegra T. Aron¹ #, Emily C. Gentry¹ #, Kerry L. McPhail² #, Louis Felix Nothias¹,
5 Mélissa Nothias-Esposito¹, Amina Bouslimani¹, Daniel Petras^{1,5}, Julia M. Gauglitz¹, Nicole
6 Sikora¹, Fernando Vargas¹, Justin J. J. van der Hooft³, Madeleine Ernst¹, Kyo Bin Kang⁴,
7 Christine M. Aceves¹, Andrés Mauricio Caraballo-Rodríguez¹, Irina Koester^{1,5}, Kelly C.
8 Weldon¹, Samuel Bertrand^{6,7}, Catherine Roullier⁶, Kunyang Sun¹, Richard M. Tehan²,
9 Christopher A. Boya P.^{8,9}, Christian Martin H.⁸, Marcelino Gutiérrez⁸, Aldo Moreno Ulloa¹⁰,
10 Javier Andres Tejeda Mora¹⁰, Randy Mojica-Flores^{8,11}, Johant Lakey-Beitia⁸, Victor
11 Vásquez-Chaves¹², Angela I. Calderon¹³, Nicole Tayler^{8,9}, Robert A. Keyzers¹⁴, Fidele
12 Tugizimana¹⁵, Nombuso Ndlovu¹⁵, Alexander A. Aksenov¹, Alan Jarmusch¹, Robin
13 Schmid¹⁶, Andrew W. Truman¹⁷, Nuno Bandeira^{18*}, Mingxun Wang^{1*}, Pieter C Dorrestein¹.
14 ^{19-21*}

15
16 **Affiliations:** ¹Skaggs School of Pharmacy and Pharmaceutical Sciences, University of
17 California San Diego, La Jolla, California, USA. ²Department of Pharmaceutical Sciences,
18 College of Pharmacy, Oregon State University, Corvallis, Oregon, USA. ³Bioinformatics
19 Group, Wageningen University, Wageningen 6708 PB, The Netherlands. ⁴College of
20 Pharmacy, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul
21 04310, Korea. ⁵Scripps Institution of Oceanography, University of California San Diego,
22 La Jolla, California, USA. ⁶Groupe Mer, Molécules, Santé-EA 2160, UFR des Sciences
23 Pharmaceutiques et Biologiques, Université de Nantes, 44035 Nantes, France.
24 ⁷ThalassOMICS Metabolomics Facility, Plateforme Corsaire, Biogenouest, 44035 Nantes,
25 France. ⁸Centro de Biodiversidad y Descubrimiento de Drogas, Instituto de
26 Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), Panamá,
27 Apartado 0843-01103, República de Panamá. ⁹Department of Biotechnology, Acharya
28 Nagarjuna University, Guntur, Nagarjuna Nagar-522 510, India. ¹⁰Biomedical Innovation
29 Department, CICESE, México. ¹¹Universidad Autónoma de Chiriquí (UNACHI), Mexico.
30 ¹²Centro de Investigaciones en Productos Naturales (CIPRONA), Universidad de Costa
31 Rica, San José, Costa Rica. ¹³Harrison School of Pharmacy, Auburn University, Auburn,
32 Alabama, USA. ¹⁴School of Chemical & Physical Sciences, Victoria University of
33 Wellington, Wellington, New Zealand. ¹⁵Centre for Plant Metabolomics Research,
34 Department of Biochemistry, University of Johannesburg, Auckland Park 2006, South
35 Africa. ¹⁶Institute of Inorganic and Analytical Chemistry, University of Münster, 48149
36 Münster, Germany. ¹⁷Department of Molecular Microbiology, John Innes Centre, Norwich,
37 NR4 7UH, U.K. ¹⁸Computer Science and Engineering, University of California San Diego,
38 La Jolla, California, USA. ¹⁹Center for Computational Mass Spectrometry, University of
39 California San Diego, La Jolla, California, USA. ²⁰Department of Pharmacology, University
40 of California San Diego, La Jolla, California, USA. ²¹Department of Pediatrics, University
41 of California San Diego, La Jolla, California, USA. #These authors contributed equally to
42 this work. Correspondence should be addressed to N.B.(nbandeira@ucsd.edu, M.W.
43 (miw023@ucsd.edu) or P.C.D. (pdorrestein@ucsd.edu)

44
45 **Author contributions:** Design and oversight of the project: P.C.D., M.W., N.B. Instrument
46 acquisition parameters: A.T.A., E.C.G., K.L.M., R.M.T., K.B.K., S.B., C.R., A.W.T., F.T.,
47 N.N., A.M.U. Data conversion and upload: K.L.M., E.C.G., A.T.A., J.J.J. v.d.H., M.E. GNPS
48 documentation: M.W., L.F.N., E.C.G., A.T.A., K.L.M., J.J.J.v.d.H., M.E, M.N.-E. Cytoscape

49 documentation: M.N-E., F.V., I.K., A.M.C-R. Metadata curation: J.M.G., C.M.A., F.V.,
50 A.M.C-R. Mass spectra annotations: D.P, R.S., M.E. Theoretical tools and advanced
51 features, statistical analysis: L.F.N., A.A. Supplementary information: A.T.A., N.S., E.C.G.,
52 K.L.M., M.E. Testing the workflows described and improving the descriptions: A.I.C,
53 A.M.U, J.A.T.M, C.M.H., C.A.B.P., M.G., V.V-C., J.L-B., R.M-F., M.E.

54

55 Pieter C Dorrestein (pdorrestein@ucsd.edu)

56

57 **Abstract:** Global Natural Product Social (GNPS) Molecular Networking is an interactive
58 online chemistry-focused mass spectrometry data curation and analysis infrastructure.
59 The goal of GNPS is to provide as much chemical insight for an untargeted tandem mass
60 spectrometry data set as possible and to connect this chemical insight to the underlying
61 biological questions a user wishes to address. This can be performed within one
62 experiment or at the repository scale. GNPS not only serves as a public data repository
63 for untargeted tandem mass spectrometry data with the sample information (metadata), it
64 also captures community knowledge that is disseminated *via* living data across all public
65 data. One of the main analysis tools used by the GNPS community is molecular
66 networking. Molecular networking creates a structured data table that reflects the chemical
67 space from tandem mass spectrometry experiments *via* computing the relationships of the
68 tandem mass spectra through spectral similarity. This protocol provides step-by-step
69 instructions for creating reproducible high-quality molecular networks. For training
70 purposes, the reader is led through the protocol from recalling a public dataset and its
71 sample information to creating and interpreting a molecular network. Each data analysis
72 job can be shared or cloned to disseminate the knowledge gained, thus propagating
73 information that can lead to the discovery of molecules, metabolic pathways and
74 ecosystem/community interactions.

75

76 **1.0 Introduction:** Molecular networking for the analysis of tandem mass spectra of small
77 molecules was introduced in 2012¹. Upon its introduction, molecular networking was
78 compared to sequencing of environmental DNA to study the microbial communities
79 present in diverse ecosystems². For the first time we were able to get a map of the
80 chemical diversity that is observed in an untargeted mass spectrometry experiment. In
81 addition to providing unprecedented systems-level views of the chemical space in various
82 environments, molecular networking has aided structure elucidation of many compounds<sup>3-
83 9</sup>.

84 The foundation of molecular networking is pairwise spectral alignment using a
85 modified cosine spectral similarity algorithm originally intended to discover modified forms
86 of peptides and proteins¹⁰. In a modified spectral similarity search, not only are
87 fragmentation spectra (MS²) from ions at identical *m/z* compared, but also MS² spectra
88 that are offset by the same *m/z* difference as the precursor ion. By eliminating the amino
89 acid filtering from the original spectral alignment algorithms, it became possible to extend
90 spectral similarity to any set of MS² spectra, including those from small molecules and
91 natural products. When a pairwise spectral similarity search/alignment is performed, each
92 MS² spectrum in a given dataset is compared against every other, and a network of MS²
93 spectral relations is obtained, from which molecular networks are created (**Fig. 1**).
94 Molecular networking build on the fundamental observation that two structurally related
95 molecules share fragment ion patterns when subjected to MS² fragmentation methods
96 such as collision induced dissociation (CID). In order to make the molecular networking

97 algorithm accessible to the scientific community, its script was converted to a web-based
 98 platform backed by a supercomputer. This enabled the creation of a community
 99 infrastructure supporting both a database and knowledge-base around the needs of the
 100 community. The result was the Global Natural Products Social (GNPS) Molecular
 101 Networking community effort that started in 2014 and was published in 2016. The user
 102 base has expanded to 49 of 50 states in the United States and worldwide to over 150
 103 countries¹¹. GNPS is currently widely used by scientists working in industry, academia and
 104 government in the fields of biomedical research, environmental science, ecology,
 105 forensics, microbiology, chemistry, and others. This crowdsourced, community-driven
 106 analysis infrastructure not only facilitates data and knowledge storage but also enables
 107 knowledge capture, sharing, dissemination and data driven social networking while
 108 promoting reproducible data analysis. Moreover, GNPS can be accessed on a computer
 109 or on any mobile device connected to the internet making any public data set readily
 110 accessible for analysis. While there are many analysis tools available within the GNPS
 111 infrastructure, molecular networking is the most frequently used tool. Other tools available
 112 on GNPS such as network annotation propagation (NAP) briefly discussed in section 3.5.

113

114 To create a molecular network, GNPS first aligns each MS² spectrum in a dataset to each
 115 of the others, and assigns a *cosine score* to each combination to describe their similarity
 116 (**Fig. 1**). Identical mass are collapsed based on a hierarchical cosine clustering algorithm
 117 into a single *node* or *consensus cluster* due to the high similarity of their fragment ions.
 118 This is accomplished using the MS-Cluster algorithm¹². Structurally related molecules yield
 119 comparable MS² spectra due to commonalities in their gas phase chemistry¹³, and are
 120 represented by separate nodes that connect within the network via *edges*. Each
 121 consensus spectrum (node) is then queried against spectral library databases to assign
 122 putative known molecules within a network.

123

124 **Table 1.** Terminology

Term	Definition
<i>annotation</i>	The process of attributing a putative chemical structure to a detected molecule. The level of annotations from spectral matches are considered level 2 or 3 according to the 2007 Metabolomics Standards Initiative ¹⁴ .
<i>bucket table</i>	A tab separated table (.tsv file format) downloadable from the GNPS interface, which shows per sample summed precursor ion intensities per MS ² ion. Pie charts generated in visualization tools are based off of intensities in the bucket table.
<i>cluster index</i>	Reference identification number for a MS ² consensus cluster. In Cytoscape this identification number is also called 'shared name'.
<i>consensus cluster</i>	A grouping of MS ² spectra that are considered identical based on the MS-Cluster algorithm ^{10,12} . Since GNPS brings together approaches from different scientific communities, there are terms

	such as “cluster” that have different meanings. Thus, the context in which the term is used should be considered. The term ‘consensus cluster’ refers to the grouping of MS ² spectra into a node and is different from clusters of nodes in molecular networks as visualized in Cytoscape ^{15,16} .
<i>cosine score</i>	A value that represents the MS ² spectral similarity between two nodes in the molecular network, where a cosine score of 1 represents identical spectra and a cosine score of 0 denotes no similarity at all. The cosine score takes into account precursor ion, fragment ions as well as peak intensities ¹ .
<i>DDA</i>	Abbreviation for data-dependent acquisition; a method for tandem mass spectrometry data collection where the most intense MS ¹ ions are iteratively selected for MS ² fragmentation ¹⁷ .
<i>dereplication</i>	Rapid identification of previously characterized (known) molecules ¹⁸ .
<i>edge</i>	A line connecting nodes that represents related but not identical MS ² spectra based on a cosine similarity score.
<i>identification</i>	Validation of a molecular assignment using an authentic chemical standard analyzed under the same experimental conditions as the sample containing the unknown compound. Molecular identification requires matching at least one physical characteristic, e.g. retention time, exact <i>m/z</i> , and MS ² fragmentation pattern ^{14, 19} .
<i>LC</i>	Abbreviation for liquid chromatography; a method used to separate molecules in a mixture using a liquid mobile phase.
<i>natural product</i>	A small molecule (< 2000 Da) produced by a biological source ²⁰ .
<i>m/z</i>	Mass-to-charge ratio, a dimensionless quantity resulting from dividing the mass number of an ion by its charge number. ²¹
<i>molecular network</i>	A map of all nodes illustrating connectivity that represents the chemical space detected in the experiment.
<i>molecular networking</i>	A computational approach that organizes MS ² data based on spectral similarity, from which we can infer relationships in chemical structures ¹ .

<i>MSCluster</i>	An algorithm used by GNPS to collapse nearly identical MS ² spectra with the same precursor ion <i>m/z</i> into a single consensus spectrum.
<i>MS¹</i>	The collection of all precursor ions (<i>m/z</i>) and associated abundancies in a sample. MS ¹ is the first stage of tandem mass spectrometry, where compounds can be further fragmented ^{22, 23} . See also <i>tandem MS</i> , <i>MS²</i> , <i>MS/MS</i> .
<i>node</i>	A consensus cluster of identical MS ² spectra that represent one molecule, or a single MS ² spectrum if cluster size is 1.
<i>precursor ion (parent ion)</i>	The ionized form of a molecule that is selected for tandem MS fragmentation. In electrospray ionization, the parent ion is a synonym of precursor ion ²¹ .
<i>product ion (fragment ion)</i>	An ion originating from a gas-phase reaction of the precursor ion ¹³ .
<i>sample information (metadata)</i>	Data that provide basic information about the sample and descriptions to facilitate data analysis and interpretation. Examples of sample information include: the identification number, the source and origin of the sample collected, time, age, sex and date of collection.
<i>small molecule</i>	This protocol considers a molecule with a molecular weight < 1500 Da a small molecule.
<i>spectral alignment</i>	An algorithmic approach that aligns related spectra. This is the basis of molecular networking which relies on the assumption that two structurally related molecules share similarity in their MS ² spectra ¹ .
<i>spectral similarity</i>	The likeness of MS ² spectra based on all or some of the following: precursor ion, fragment ions, and relative intensities of these peaks. Structurally related molecules tend to exhibit similar fragmentation ¹³ . In molecular networking spectral similarity is calculated through a modified cosine score.
<i>summed ion intensities</i>	Sum of precursor ion intensities in the MS ² spectra for all ions with the same associated MS ² detected by the mass spectrometer.
<i>tandem MS, MS/MS, MS²</i>	Abbreviations for tandem mass spectrometry, which defines a technique where mass-selected ions are subjected to a second mass spectrometric analysis. In the first stage, also referred to as MS ¹ , precursor ions are formed and detected. In the second stage, also referred to as MS ² or MS/MS, precursor ions are fragmented resulting in a spectral fingerprint ^{22, 23} .

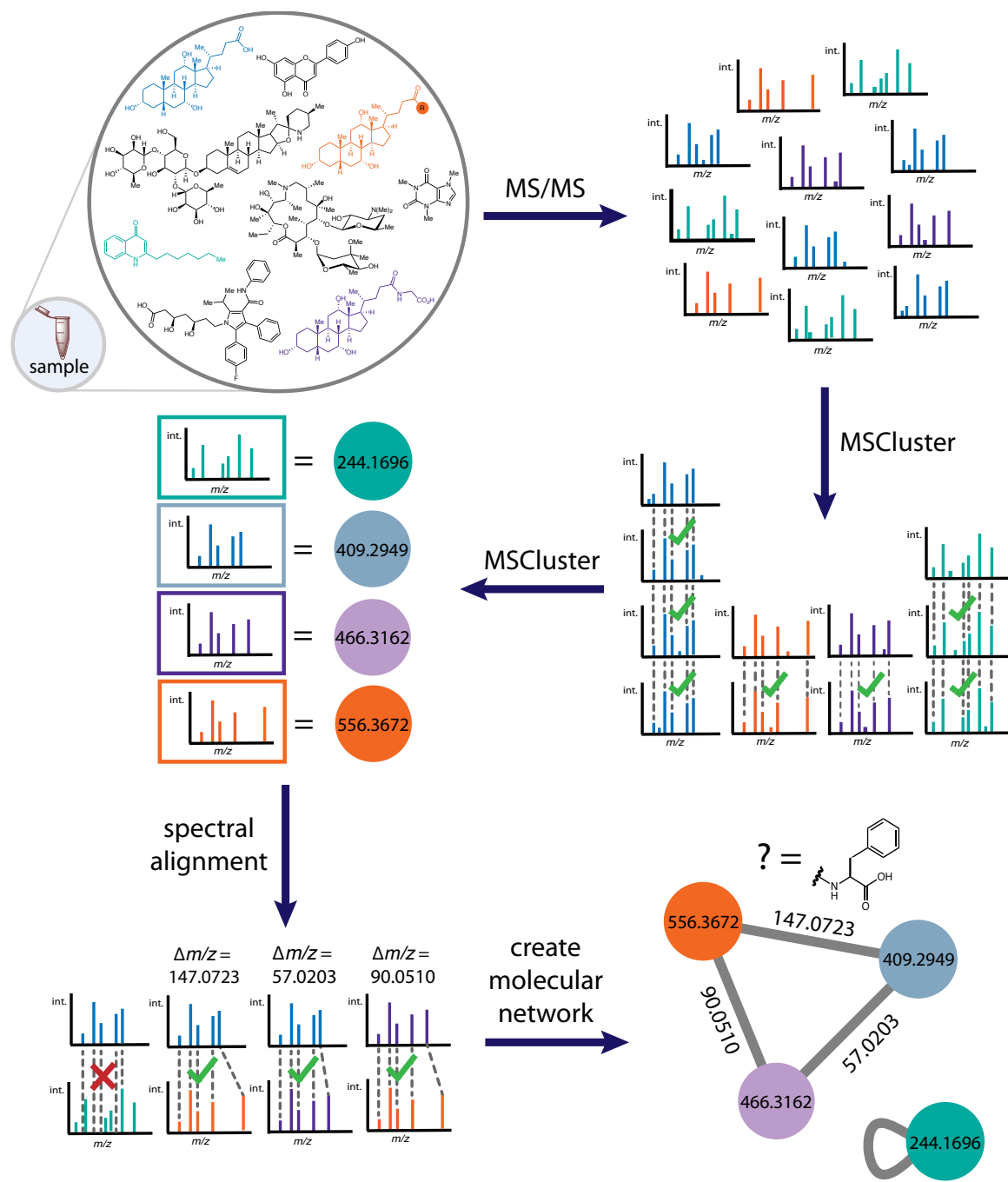
125

126

127

All mass spectrometry data used in GNPS, both in the private user workspace or data that are made public, is stored in Massive - an interactive virtual environment developed

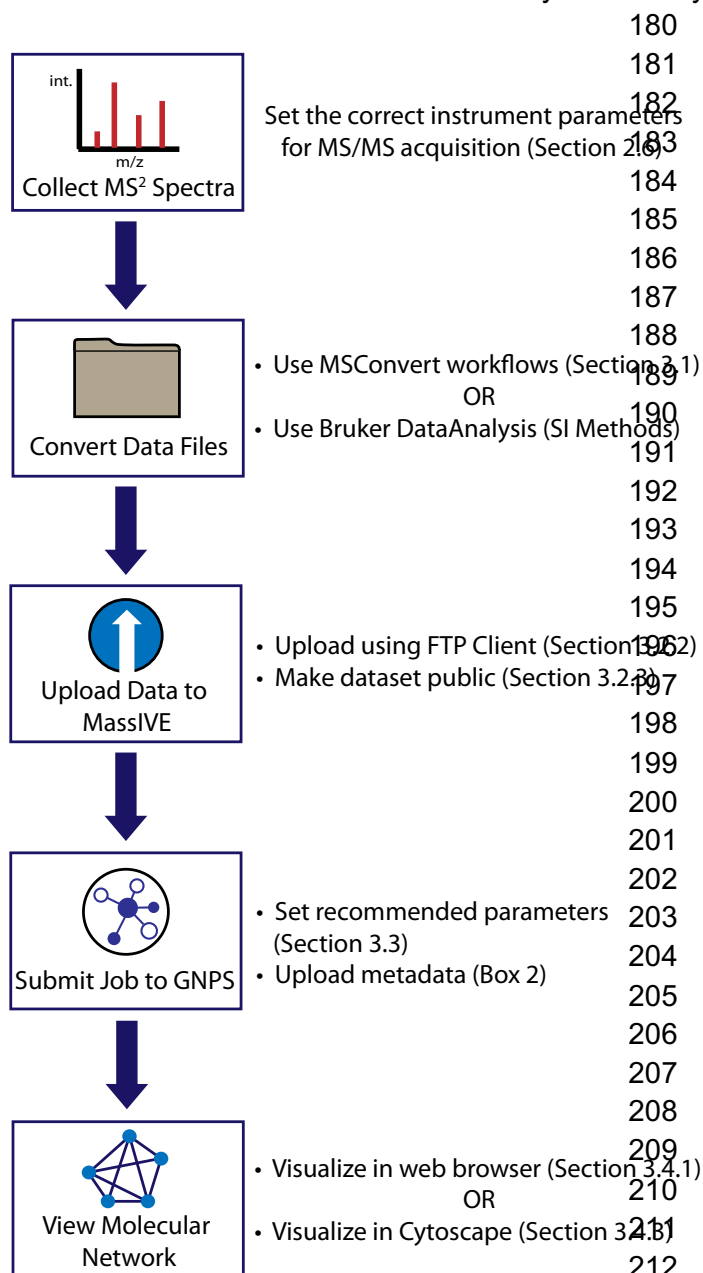
128 to facilitate and encourage the exchange of mass spectrometry data. MassIVE accepts
129 data files (organized as datasets) and facilitates the sharing of datasets with a unique
130 identifier; one can use this unique identifier as an accession number for publications. In
131 addition, public datasets that the user publishes can, by choice of the depositor, have an
132 associated DOI. Currently, MassIVE is an approved repository for the Journal of
133 Proteome Research (<https://pubs.acs.org/journal/jprobs>) and Nature Partner Journals
134 (<https://www.nature.com/sdata/policies/repositories#chem>) and is widely used as a
135 repository for other journals²⁴⁻³³. GNPS-MassIVE has more than a thousand public
136 metabolomics datasets. The GNPS knowledge base includes 221,083 reference MS²
137 spectra, provided by the GNPS community, spectral libraries generated for GNPS
138 (GNPS-collections) and third party libraries¹¹. Examples are LDB Lichen Database,
139 MIADB Spectral Library, Sumner Spectral Library, CASMI Spectral Library, and
140 Massbank, a large MS data library that is directly synced with GNPS. There are also tags
141 and sample information (metadata) entries provided by the community in the GNPS
142 knowledge base. Furthermore, all public data is periodically searched against the NIST
143 2017 spectral library and high confidence spectral matches are annotated. GNPS-
144 MassIVE now performs more than 6,000 analysis jobs a month and more than 200,000
145 page views (excluding developers), with the predominant analysis being molecular
146 networking. As a result, GNPS based analysis has been used for the discovery of
147 hundreds of new molecules in the last few years, ranging from immune regulators to
148 antimicrobials, including antiviral agents and protease inhibitors^{9, 34-36}. Here we provide a
149 detailed protocol on how to generate a publishable and reproducible molecular network
150 from a mass spectrometry dataset. This protocol will take the reader through the
151 following steps: how to upload data, how to make the data public, how to subscribe to
152 public data for living data updates, and how to reproducibly create publishable molecular
153 networks using standardized sample information (metadata) through the GNPS
154 infrastructure (**Fig. 1**).
155



156
 157
 158
 159
 160
 161
 162
 163
 164
 165
 166
 167

Figure 1. Schematic representation of the process for creating a molecular network from tandem mass spectra acquired for metabolites in complex sample mixtures. The colors are used to track how we go from molecules in a sample to nodes in the molecular network. We start by obtaining MS² spectra of all ionized molecules in the sample. MS-Cluster first aligns each MS² spectrum in a dataset to each of the others. Mass spectra from identical compounds coalesced using MS-Cluster¹² into a single *node* or *consensus cluster* due to the high similarity of their precursor ion and fragment ions. Subsequently a spectral alignment is performed enabling for similarity searches even when the precursor ion masses are not identical. This is accomplished using a modified cosine score where all the ions that differ by the mass difference of the two precursor ions are also considered.

168 Structurally-related molecules yield comparable MS² spectra due to commonalities in their
 169 gas phase chemistry, and are represented by separate nodes that connect within the
 170 network via *edges*. Each node is then queried against spectral libraries to assign putative
 171 known molecules within a molecular network and unknowns can be propagated using
 172 chemical rationale. For illustration purposes, the blue node with *m/z* 409.2949 is cholate,
 173 *m/z* 446.3162 in purple is glycocholic acid (the user would discover this based on MS²
 174 matches to a reference library) while the orange one is unknown but has a mass shift of
 175 147.0723 Da. This is a typical mass shift of phenylalanine and thus a prediction can be
 176 made that this is a phenylalanine conjugate of cholic acid. The difference between the
 177 glycine and phenylalanine conjugate is 90.0510 Da and supports such structural
 178 hypothesis. The self looped green node *m/z* 244.1696 is attributed to an unrelated
 179 molecule and therefore does not have any structurally related molecule in the sample.



216

217 Data collection and processing procedures will vary depending on the instrument
218 available to the user. Although the user can modify any procedure to fit their specific goals,
219 this protocol specifies a set of starting parameters for acquiring and converting data with
220 various mass spectrometers, including AB Sciex, Agilent, Bruker, Shimadzu, Thermo
221 Scientific, and Waters instruments. We also provide a protocol for the conversion of the
222 data from each of these vendors to an open format (.mzXML, .mzML or MGF) that is usable
223 within the GNPS-MassIVE infrastructure. Once the data is converted to the proper open
224 format, the protocol describes how to upload data files to MassIVE, a public repository that
225 enables community sharing of mass spectrometry data, using either a web browser or FTP
226 client. The resulting datasets can then be subsequently submitted to GNPS for molecular
227 networking analysis, wherein MS² spectra are organized in a network according to
228 similarity and compared against a reference database to identify putative known molecules
229 and 'molecular families' in the samples. Finally, visualization and analysis of GNPS-
230 generated molecular networks can be performed either in the web browser itself or in
231 [Cytoscape](#), an open-source software for visualizing complex networks³⁷.

232

233 **1.2 Applications of the method**

234 GNPS molecular networking provides the ability to analyze and compare MS² spectra in
235 one or more datasets acquired within the scope of a specific study, across datasets from
236 multiple studies, and also to compare those datasets to all publicly available GNPS-
237 MassIVE datasets, including community curated spectral libraries. In addition, ongoing
238 contributions to spectral libraries and submissions of new public datasets enable
239 continuous identification: the periodic and automated reanalysis of all public datasets.
240 GNPS is being used to network data acquired on a number of different mass
241 spectrometers in a wide variety of exploratory studies, with samples originating from
242 diverse environments and used for varying purposes. These range from the indoor
243 environment³⁸⁻⁴⁰ to dissolved organic matter in the oceans⁴¹, from microbes in culture<sup>9, 42-
244 45</sup> to mouse⁴⁶ or human microbiomes^{47, 48} or infections⁴⁹⁻⁵¹, from clinical samples^{32, 52, 53} to
245 plants⁵⁴, algae⁵⁵, sponges^{5, 56} and corals⁵⁷, as well as a number of other sample types<sup>26,
246 58</sup>. Additionally, molecular networking has been applied to natural products discovery from
247 a variety of organisms⁵⁹⁻⁶², forensics⁶³, small molecule identification⁶⁴ and biological
248 discovery in hypothesis-driven research⁶⁵. Furthermore, GNPS facilitates large-scale
249 meta-analyses that can compare and potentially link studies from different laboratories by
250 enabling rapid comparisons across multiple public datasets. Finally, to promote data
251 analysis reproducibility, all analysis jobs are saved together with their parameters, which
252 can be shared or cloned for reanalysis; no other platform provides this service.

253

254 **1.3 Alternative methods to this protocol**

255 Several aspects of the GNPS-based molecular networking protocol are provided
256 elsewhere, but not previously as a coherent workflow in one package. There are several
257 repositories where metabolomics data can be uploaded⁶⁶⁻⁶⁸. According to the OMICS
258 discovery index, the most widely used are GNPS-MassIVE, Metabolomics workbench¹²
259 and MetaboLights^{69, 70}.

260

261 Mass spectral library searching, or comparing MS² spectra of compounds in a
262 sample to reference data in order to annotate metabolites⁷¹, has been implemented
263 extensively, and successfully, for decades. Numerous commercial and non-commercial
MS² reference databases exist, such as the NIST/EPA/NIH Mass Spectral Library⁷²,

264 METLIN⁷³, MassBank of Japan (<http://massbank.jp>)⁷⁴, EU
265 (<https://massbank.eu/MassBank/>)⁷⁵ and North America
266 (<http://mona.fiehnlab.ucdavis.edu/>), mzCloud^{76, 77}, and ReSpec⁷⁸, which potentially
267 provides users with access to around 2.4 million MS² reference spectra, when GC-MS and
268 LC-MS reference spectra are both considered⁶⁶. Many of these reference databases have
269 an integrated spectral matching tool for compound identification, including mzCloud,
270 METLIN/XCMS Online^{79, 80}, Metabox⁸¹, MassBank). The goal of GNPS is not only to
271 provide a spectral matching tool, but also to serve as a data storage and knowledge
272 capture and dissemination platform, and to provide access to a host of other analysis tools
273 not covered in detail here, such as *in silico*-based dereplication⁸²⁻⁸⁴, network annotation
274 propagation⁸⁵, genome mining tools⁸⁶, and MASST searches.

275 GNPS is currently the only online platform that provides molecular networking, a
276 computational tool that compares pairs of MS² spectra based on their similarities and
277 connects them to MS² reference spectral libraries. Molecular networking enables further
278 propagation of annotations through mass spectral relations. MetGem⁸⁷ is a standalone
279 software package that can be used for the generation of molecular networks which works
280 well for smaller data sets, it is not connected to a knowledge base, repository wide analysis
281 tools and additional computational resources that GNPS provides.

282

283 **1.4 Expertise needed to implement the protocol**

284 Sampling and sample preparation, including sample extraction, should be
285 performed by a trained analytical chemist, and mass spectrometry data should be acquired
286 by a trained mass spectrometrist. It is imperative that the parameters for mass
287 spectrometry be suitably optimized for the experimental conditions and sample type in
288 order to generate meaningful molecular networks. Important instrument parameters to
289 consider may include precursor isolation window, mass resolution, collision energy, data
290 dependent acquisition settings (e.g. duty cycle time and dynamic exclusion parameters),
291 and the mass spectrometer has to be properly calibrated before use. While an expert user
292 will have preferred instrument parameters, recommended data acquisition parameters
293 from major instrument manufacturers are provided below (section 2.6) for newer mass
294 spectrometry users who aim to create molecular networks in GNPS. Basic knowledge of
295 tandem mass spectrometry fundamentals as well as knowledge of sample handling and
296 preparation are required to further optimize the data analysis parameters appropriate to
297 the instrument used and the experimental design.

298

299 **1.5 Experimental design**

300 After running the molecular networking algorithm, GNPS creates a data table that
301 provides as much chemical insight into the data as possible in relation to the metadata
302 (associated sample information) provided by the user. Such data tables can be viewed as
303 networks directly in the GNPS website or exported and manipulated in other data
304 visualization tools and statistical analysis packages. Here we provide a GNPS-based
305 molecular networking tutorial in which we import the table into a third party tool called
306 Cytoscape, a powerful network visualization software. Notably, the information
307 represented in and inferred from a molecular network is dependent on the input, including
308 both the mass spectrometry data⁸⁸ and networking parameters selected.

309

310 **1.5.1 Reproducibility, blanks, and controls**

311 A well organized and well thought-out experimental plan is essential for the
312 successful creation of useful molecular networks, since molecular networks are only as
313 meaningful as the experiment and data from which they originate. This includes providing
314 sample information (metadata) tables and raw data files for the sample set; metadata
315 tables aid the creation of molecular networks that have increased interpretative value. In
316 order to avoid pitfalls associated with large-scale mass spectrometry experiments, e.g.
317 batch effects⁸⁹, sample carryover and/or contamination⁹⁰, and high background signal⁹¹,
318 and to maximize reproducibility and signal-to-noise ratio⁹², a dataset should include
319 blanks, quality control (QC) samples, and experimental replicates. Dunn et al.⁹³ describe
320 an appropriate representative experimental design in detail that includes blanks, quality
321 control mixtures, and samples plus internal standards. Petras et. al.³⁸ provide an example
322 that illustrates control metrics, including evaluation of quality control mixtures and signal
323 deviation of the internal standard.

324 We recommend preparing control samples using exactly the same protocols and
325 experimental conditions used to prepare test samples (i.e. the same types of tubes, the
326 same batches of tubes, the same extraction solvent, extraction time, sonication time/power
327 and so on). These blank samples inform which ions come from the experimental conditions
328 and they can be subtracted from test sample signals in the molecular networking analysis
329 (see section 3.4.3). The requirements for QC associated with a broad assessment of the
330 natural product composition of an extract library used in bioactivity screens is different from
331 a detailed clinical study for biomarker discovery. When possible, one should add internal
332 standard(s) to each sample to ensure that the system performs consistently. If the internal
333 standard(s) do not match the user-defined acceptable chromatography variations, the
334 sample needs to be either removed from downstream analysis or rerun. This is particularly
335 useful in applications where thousands of samples, such as natural product extract
336 libraries, are screened. Further, when acquiring data for a large number of samples,
337 especially when multiple batches are used, we suggest acquiring data for additional QC
338 samples to monitor batch and plate effects throughout the experiment in order to assess
339 instrumental variations over time, such as retention time drift. QC samples may either
340 consist of aliquots from a subset of test samples pooled together (pooled QC) or be
341 mixtures of molecules specifically defined for quality assurance. For example, it is common
342 to use the last column of a 96-well plate for the QC mixture to ensure that the instrument
343 and chromatography behave in an identical fashion throughout an experiment. Finally,
344 data from experimental replicates, including both technical and biological replicates should
345 be acquired in a randomized fashion. This is especially important for large-scale population
346 studies to ensure minimized bias. One common problem in metabolomics and LC-MS
347 analysis is sample carryover, caused by residual compound(s) from a previous run. One
348 way to reduce this issue is to insert a wash routine between samples followed by a blank
349 to ensure that no carryover is observed.

351 **1.5.2 Molecular networking parameters**

352 GNPS-based molecular networking parameters may be varied significantly and need to
353 be set appropriately for the acquired dataset, based on sample (anticipated molecular
354 masses and types of molecules), instrument resolution and collision energies used for MS
355 acquisition. Networking parameters are described in detail in Section 3.3, Table 1, and
356 should be considered and selected carefully in order to obtain useful networks, which
357 ultimately depend on the quality and quantity of MS² spectra.

358

359 1.6 Limitations and challenges

360 Since GNPS-based molecular networking utilizes MS² data, it is susceptible to the
361 same challenges encountered in any mass spectrometry data acquisition experiment,
362 such as low signal-to-noise, insufficient separation of analytes, or poor peak shape.^{94, 95} In
363 addition, classical molecular networking can provide only qualitative information about the
364 experiment because only MS² scans are considered in the analysis. While feature-based
365 molecular networking (Box 4) incorporates MS¹ and chromatographic data, which
366 approximates quantitation, it is still not strictly quantitative. If calibrated quantitative
367 information is needed to answer the scientific question, follow-up experiments should be
368 performed using targeted LC-MS.

369 Additionally, one should consider potential issues that accompany metabolomics
370 experiments, such as sample extraction efficiency and reproducibility, as well as unwanted
371 metabolite degradation. While avoiding degradation or modification of all molecules in a
372 sample is impossible, it is important that all samples for comparison are prepared and
373 analyzed in an identical manner, unless the goal is to understand the effects of sample
374 preparation conditions⁹⁶. While a few publications describe the impact of storage on the
375 detectable metabolome, these are sample type-specific and there is currently no
376 consensus for a “gold standard”⁹⁷⁻⁹⁹. Ultimately, sample preparation is highly dependent
377 on the type of sample collected, and includes drying, homogenization, and extraction
378 steps¹⁰⁰. Although every lab has their own preferences for sample treatment, we strongly
379 advocate for samples to be collected and extracted with solvent as soon as possible. The
380 speed of this is dependent on the experimental environment. For example, samples
381 collected in remote areas, at sea using a small boats, or often even in a clinical setting,
382 may be stored for hours or days before they can be extracted, given that some solvents
383 are not easily brought into a clinical setting or used while out at sea. In contrast, samples
384 from a cultured system in a lab or an enzymatic reaction, for example, can be halted in
385 milliseconds using a rapid quench system and can then be extracted in seconds. The
386 choice of solvent and extraction protocol is dictated by the experimentalist’s interests and
387 questions. Although there is always overlap among the molecules from even very different
388 extraction protocols, more polar metabolites are extracted with ethanol, methanol and
389 butanol while more hydrophobic metabolites are extracted with benzene, ethyl acetate or
390 chloroform⁹⁶. The samples can then be introduced into the mass spectrometer using front-
391 end separation techniques, most often liquid chromatography or ion mobility. If mass
392 spectrometry cannot be performed immediately, we recommend completely drying the
393 samples before storage at cryogenic temperatures.

394 To annotate unknown molecules, GNPS queries MS² spectra against MS² data in
395 reference libraries and assigns a cosine score based on their similarity. For the GNPS
396 spectral library, MS² spectra are acquired from laboratories around the world using a
397 variety of mass spectrometers and sample preparation protocols. Therefore, mass spectra
398 submitted to GNPS can differ in terms of both quality and content. For instance, MS²
399 fragment ions and their intensities can vary significantly between instruments, and even
400 on the same instrument if the experimental setup is changed¹⁰¹. GNPS requires that the
401 instrument and ion source be specified with each reference spectrum submitted and it is
402 recommended that this be taken into account when assessing the quality of a library hit.
403 Along these lines, annotations of unknown molecules are not all accurate and should be
404 considered putative until confirmed with an authentic chemical standard.

405 On average, in 2016 when GNPS was published, only 2% of spectra in an
406 untargeted mass spectrometry metabolomics experiment were annotated¹⁰². Although this

407 percentage has grown to an average of 5-6% annotations, a large percentage of MS²
408 spectra typically remain unannotated. The structures of these unannotated molecules or
409 “dark matter”¹⁰³ might be known, but their identity is not revealed because no reference
410 spectra exist in library databases, against which to compare. To improve annotation rates,
411 *in silico* tools have been developed to match unknown MS² spectra to putative chemical
412 structures¹⁰⁴. Several of these computational tools, which include MetFrag¹⁰⁵,
413 MetFusion¹⁰⁶, SIRIUS^{107, 108}, CSI:FingerID¹⁰⁹, MS-Finder¹¹⁰, Network Annotation
414 Propagation (NAP)⁸⁵, and Dereplicator^{82, 83} can be integrated into GNPS molecular
415 networking workflows to provide insight into the annotation; the application of such tools is
416 beyond the immediate scope of the networking protocol presented here.

417

418 2.0 Materials

419 2.1 REAGENTS

420 **CRITICAL** For specific storage and handling instructions, consult the manufacturer of each
421 reagent. Although high grade solvents are used, different batches of the same solvents
422 (even purchased from the same vendors), can give rise to different background
423 contaminants in the experiment. There are also many possible substitutes for the reagents
424 and consumables listed below.

- 425 • Water of LC–MS (Optima) grade (Thermo Fisher Scientific, cat. no. W6-4)
- 426 • Acetonitrile (ACN), LC–MS (Optima) grade (Thermo Fisher Scientific, catalogue
427 number A955-4) ! **CAUTION** Acetonitrile is highly flammable, and the bottles
428 should be stored in a flammable-liquid cabinet.
- 429 • Methanol (MeOH), LC-MS grade ! **CAUTION** Methanol is highly flammable, and
430 the bottles should be stored in a flammable-liquid cabinet.
- 431 • Formic acid (FA). LC–MS grade, Optima grade (Thermo Fisher Scientific,
432 catalogue number A117-50) ! **CAUTION** Formic acid is highly corrosive. It should
433 be handled in a flow cabinet while wearing eye protection and gloves.
- 434 • LC-MS calibration solutions, e.g. for the Bruker MaXis II QTOF mass spectrometer:
435 ESI-TOF Low Concentration Tuning Mix (Agilent Technologies, catalogue number
436 G1969-85000) for external calibration and Hexakis(1H,1H,3H-
437 tetrafluoropropoxy)phosphazene (Synquest Laboratories, catalogue number
438 8H79-3-08), *m/z* 922.009798 for internal calibration (lock mass) ! **CAUTION** This
439 compound is irritating to the eyes and the skin. It should be handled wearing eye
440 protection and gloves; for the Q-Exactive mass spectrometer: Pierce LTQ Velos
441 ESI Positive Ion Calibration Solution (Thermo Fisher Scientific, catalogue number
442 88323) and ESI Negative Ion Calibration Solution (Thermo Fisher Scientific,
443 catalogue number 88324).

444

445 2.2 EQUIPMENT

- 446 • Microtiter plates (e.g. Nunc 96-Well Round Bottom Polypropylene Storage
447 Microplates, Thermo Fisher Scientific, catalogue number 267245) containing
448 samples of interest at, e.g., 1 mg/mL concentration.
- 449 • Benchtop vacuum concentrator compatible with 96-well microplate evaporation
450 (Centrivap; Labconco)
- 451 • Reversed phase C18 LC column, 1.7- μ m particle size, 50 \times 2.1-mm (Phenomenex,
452 part number 00B-4475-AN or equivalent)
- 453 • UHPLC system coupled to a tandem mass spectrometer with an ESI source; e.g.
454 a 1260 HPLC (Agilent) coupled to a QTOF 6530 mass spectrometer (Agilent),

455 UltiMate 3000 UHPLC system (Dionex) coupled to a MaXis II QTOF system
456 (Bruker Daltonics), a Vanquish UHPLC system coupled to a Q-Exactive mass
457 spectrometer (Thermo Fisher Scientific), an Acquity UHPLC I coupled to a Xevo
458 G2-XS QTOF (Waters), a Nexera X2 UHPLC (or a Prominence UFLC) coupled to
459 an IT-TOF mass spectrometer (Shimadzu) or an AB Sciex 5600 TripleTOF mass
460 spectrometer.

461

462 2.3 SOFTWARE

- 463 • MSConvert tool from the ProteoWizard
464 (<http://proteowizard.sourceforge.net/downloads.shtml>)
- 465 • AB Sciex MS Data Converter (Beta 1.3) is freely available for download from the
466 AB Sciex website <https://sciex.com/software-support/software-downloads>
- 467 • AB Sciex Analyst Software 1.7 is available for download, trial license use and
468 purchase from the AB Sciex website [https://sciex.com/products/software/analyst-](https://sciex.com/products/software/analyst-software)
469 [software](https://sciex.com/products/software/analyst-software)
- 470 • Agilent MassHunter software can be obtained from the Agilent website:
471 [https://www.agilent.com/en/products/software-informatics/masshunter-](https://www.agilent.com/en/products/software-informatics/masshunter-suite/masshunter/masshunter-software)
472 [suite/masshunter/masshunter-software](https://www.agilent.com/en/products/software-informatics/masshunter-suite/masshunter/masshunter-software)
- 473 • Bruker DataAnalysis is available for download from the Bruker website
474 ([www.bruker.com/service/support-upgrades/software-downloads/mass-](http://www.bruker.com/service/support-upgrades/software-downloads/mass-spectrometry.html)
475 [spectrometry.html](http://www.bruker.com/service/support-upgrades/software-downloads/mass-spectrometry.html))
- 476 • Shimadzu LabSolutions can be obtained from the Shimadzu website:
477 [https://www.ssi.shimadzu.com/products/liquid-chromatography-mass-](https://www.ssi.shimadzu.com/products/liquid-chromatography-mass-spectrometry/lcms-software.html)
478 [spectrometry/lcms-software.html](https://www.ssi.shimadzu.com/products/liquid-chromatography-mass-spectrometry/lcms-software.html)
- 479 • Thermo Scientific Xcalibur software can be obtained at:
480 <https://www.thermofisher.com/order/catalog/product/OPTON-30801>
- 481 • Waters MassLynx MS software can be obtained at:
482 [http://www.waters.com/waters/en_US/MassLynx-MS-](http://www.waters.com/waters/en_US/MassLynx-MS-Software/nav.htm?locale=en_US&cid=513662)
483 [Software/nav.htm?locale=en_US&cid=513662](http://www.waters.com/waters/en_US/MassLynx-MS-Software/nav.htm?locale=en_US&cid=513662)
- 484 • FTP Client (e.g. WinSCP for Windows; Cyberduck for Macintosh)
- 485 • Web Browser, Firefox or Google Chrome to access GNPS
- 486 • Cytoscape for data visualization: <https://cytoscape.org/> ([current version at the time](https://cytoscape.org/)
487 [of publication is 3.7.1](https://cytoscape.org/)).
- 488 • Software relevant to optional pipelines, e.g. 2D or 3D Visualization¹¹¹; Feature-
489 based molecular networking, see **Box 4**.

490

491 2.4 EXAMPLE DATASETS

492 **CRITICAL** All LC–MS data used in this paper are publicly available at the GNPS-MassIVE
493 repository under the following accession numbers.

- 494 • [MSV000083437](https://massive.ucsd.edu/MSV000083437) (Germ Free and Specific Pathogen Free Mice, unpublished)
- 495 • [MSV000083359](https://massive.ucsd.edu/MSV000083359) (3D Cartography of Diseased Human Lung⁵⁰)
- 496 • [MSV000083381](https://massive.ucsd.edu/MSV000083381) (Stenothricin-GNPS analogues¹¹)

497

498 2.5 REAGENT SETUP

499 **Aqueous LC–MS mobile phase, Solvent A** Prepare the aqueous mobile phase (Solvent
500 A) for LC–MS by adding LC–MS-grade formic acid to LC–MS-grade water to make a
501 100:0.1 (vol/vol) water/formic acid mixture. The LC solvents can be stored at room
502 temperature for up to 1 week.

503 **! CAUTION** Formic acid is highly corrosive. **CRITICAL** The aqueous mobile phase for LC–
504 MS should not be stored for more than a week because of the potential for microbial
505 growth.

506 **Organic LC–MS mobile phase, Solvent B** Prepare the organic mobile phase (Solvent B)
507 for LC–MS by adding LC–MS-grade formic acid to LC–MS-grade acetonitrile to make a
508 100:0.1 (vol/vol) acetonitrile/formic acid mixture.

509 The LC solvents can be stored at room temperature for up to 1 week.

510 **! CAUTION** Formic acid is highly corrosive and should be handled in a flow cabinet while
511 wearing eye protection and gloves.

512

513 **2.6 EQUIPMENT SETUP**

514 **Mass spectrometry**

515 Both ion source parameters and data dependent acquisition (DDA) parameters are
516 essential for obtaining quality MS² spectra to be used for molecular networking. Although
517 many instrument configurations exist, several representative ion source and DDA
518 parameters are described below. Relevant to these MS parameters is the LC method used,
519 an example of which is a gradient profile from 10 to 100% ACN + 0.1% FA in H₂O + 0.1%
520 FA (for 12 min), followed by isocratic 100% ACN + 0.1% FA (for 3 min), and 5% ACN +
521 0.1% FA (3 min) re-equilibration phase, with a flow rate of 400 µL/min.

522

523 Suggested instrument parameters for ABSciex, Agilent, Bruker, Shimadzu, Thermo
524 Scientific, and Waters are provided in the supporting information.

525

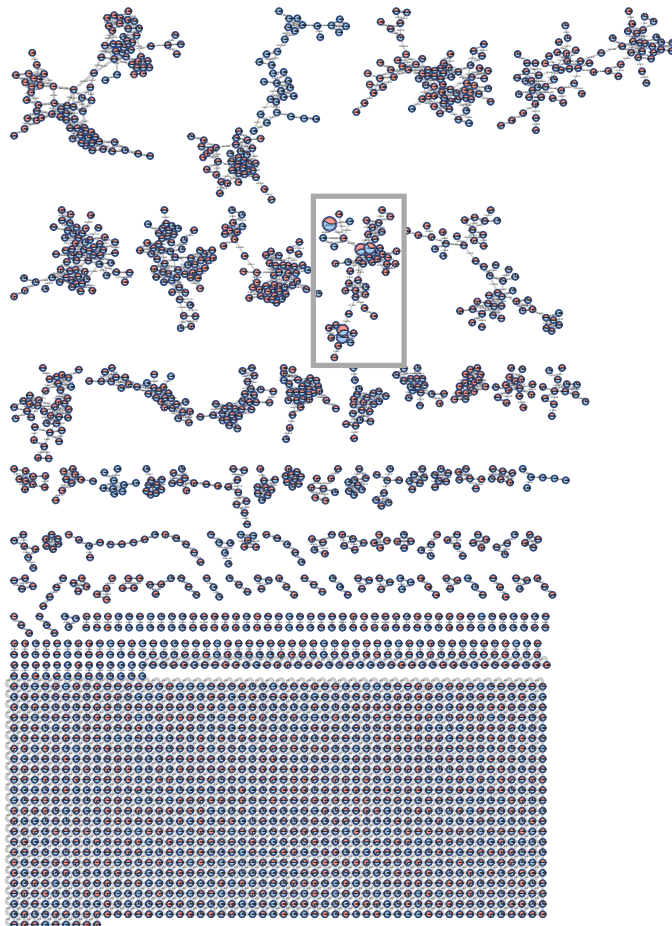
526 **3.0 Procedure**

527

528 In addition to the protocol described in the following, all steps, albeit in less detail, are also
529 described and continuously updated and maintained in the online GNPS documentation
530 at: <https://ccms-ucsd.github.io/GNPSDocumentation/>

531

532 The data submission and molecular networking workflow (section 3.2 onwards) may be
533 followed as a tutorial using an untargeted metabolomics dataset for 3D molecular
534 cartography of the mouse duodenum (paper is in review, Massive dataset
535 [MSV000083437](https://doi.org/10.26434/chemrxiv-2023-08343)). This dataset is a subset of a collection of metabolomes analysed from
536 organs of germ free (GF) and specific pathogen free (SPF) mice that led to the discovery
537 of new amide conjugated bile acids made by bacteria that affect host metabolism *via*
538 farnesoid X receptor (FXR) agonism. The following procedure will take the reader through
539 submission of dataset [MSV000083437](https://doi.org/10.26434/chemrxiv-2023-08343) to the molecular networking workflow in GNPS,
540 through the molecular networking workflow in GNPS (including input parameters), and
541 through visualization of the generated network using both in browser and Cytoscape-
542 based visualization (**Fig. 3**).



● = Germ free (GF) mice
 ● = Specific Pathogen Free (SPF) mice

543
 544

545 **Figure 3.** The readers will recreate the mouse duodenum global molecular network
 546 depicted above, created from MassIVE dataset [MSV000083437](https://massive.ucsd.edu/MSV000083437) and visualized in
 547 Cytoscape. Pie charts represent relative summed precursor ion intensities per MS² spectra
 548 detected within each metadata group: red for germ-free (GF) and blue for specific-
 549 pathogen free (SPF) mice. The Box highlights a cluster we will examine below in terms of
 550 chemical interpretation.

551

552 **3.1 Data conversion** - Timing 1 hour up to a few days (varies depending on size of dataset
 553 and computer set-up)

554

555 GNPS-MassIVE converts raw data formats after upload to .mzML format (stored in the
 556 ccms_peak folder) for GNPS processing. Nevertheless, to enable immediate use of the
 557 data, it is recommended to manually convert the raw data to open file formats prior to
 558 uploading to GNPS-MassIVE. The protocol for data conversion depends on the instrument
 559 used for mass spectrometry acquisition. MSConvert can be used for the conversion to a
 560 GNPS-compatible format of mass spectrometry data acquired on AB Sciex, Agilent,
 561 Shimadzu (after initial conversion, SI Methods), Thermo Scientific and Waters instruments.
 562 Although one of the most common formats used in GNPS, Bruker files (.d format), at this
 563 time, are still not MSConvert compatible. For Bruker files, a separate workflow must be

564 utilized, which applies internal lockmass calibration to the output file. This Bruker workflow
565 is described in more detail in the SI methods. Alternatively, for AB Sciex, raw files (.wiff)
566 could be converted into .mzML format using the AB MS Data Converter (AB Sciex version
567 1.3 beta, freely available at <https://sciex.com/software-support/software-downloads>).

568

569 1) MSConvert can be downloaded freely from ProteoWizard at:
570 <http://proteowizard.sourceforge.net/download.html>. This software is compatible with
571 Windows and Linux operating systems but is not supported for Mac OS. When
572 downloading ProteoWizard, the version of Windows must be specified and .NET
573 Framework 3.5 SP1 and 4.0 must be installed. Then either a traditional workflow or an
574 easy workflow can be used for the file conversion. These two workflows are detailed below.
575 The “traditional” workflow, outlined below, is the manual workflow.

576

577 2) Mass spectrometry files must be converted to open file formats such as .mzXML,
578 .mzML, and .mgf formats for analysis in GNPS, with the preferred formats being .mzXML
579 and .mzML. Although it is encouraged to co-submit the raw data to MassIVE, GNPS does
580 not support .mzData, .xml, .raw, .wiff, .scan, .d, and .cdf formats.

581

582 3) MSConvert is the recommended software for conversion of data acquired on AB Sciex,
583 Agilent, Thermo Scientific and Waters instruments. Conversion can be performed following
584 the steps outlined below:

585 a. In the Start Menu, the ProteoWizard folder can be selected and MSConvert can be
586 opened.

587 b. To select file(s) for conversion, click Browse; then click ‘Add’ to add file(s) to the
588 workflow and select a directory for the output.

589 c. To convert the vendor file format to an .mzXML file, select .mzXML under Options;
590 32-bit should be selected for binary encoding precision and Use zlib compression
591 should be unchecked.

592 d. Choose Peak Picking under the Filters heading and under Algorithm check Vendor,
593 then write in MS-Levels 1-2 and finally add the filter by clicking Add. ! **CRITICAL**
594 **STEP** Move the peakPicking filter to the top of filter list. The peakPicking filter must
595 be the first filter in the list or the output file will not be centroided.

596 e. Click Start then check the folder for the .mzXML files in the Output Directory. These
597 files can be opened in SeeMS (Installed with MSConvert), OpenMS TOPPView
598 (<https://github.com/OpenMS/OpenMS/releases>)¹¹² or MZmine2
599 (<https://github.com/mzmine/mzmine2/releases>)¹¹³ to verify that the conversion
600 worked properly.

601

602 An “easy” workflow is also available. This simple batch conversion method includes a
603 complete package for Windows users to convert vendor formats to GNPS compatible
604 format (mzXML, mzML, MGF) and is described in the SI Methods. An online data
605 conversion tutorial can be accessed at: [https://ccms-](https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/)
606 [ucsd.github.io/GNPSDocumentation/fileconversion/](https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/).

607

608 3.2 Data submission to GNPS / MassIVE

609 It is necessary to create an account with GNPS in order to submit datasets and create
610 workflows, as well as to receive emails about the outcomes. Making a GNPS account
611 automatically sets up a MassIVE account that uses the same login and password. To

612 manipulate MS data files in GNPS, they must first be uploaded to MassIVE, which is an
613 online repository for mass spectrometry datasets hosted by the UCSD Center for
614 Computational Mass Spectrometry (CCMS). The user workspace in GNPS / MassIVE
615 provides a personalized location for researchers to curate mass datasets, submit and
616 monitor GNPS workflows, subscriptions to datasets that have been made publicly
617 available by others, or clone and reanalyze either their own or other public datasets. More
618 information on subscriptions to data can be found in sections 3.6.

619

620 **3.2.1 Create a GNPS / MassIVE account (SI Fig. 1):**

621

622 1) Open up a web browser. GNPS is designed to work with Firefox or Google Chrome
623 but also works in Microsoft Edge, Safari, and Opera.

624 2) Navigate to the GNPS home page by using this link
625 <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash2.jsp>

626 3) Towards the top center of the page, above the large GNPS logo, click on “Register
627 New Account” (right hand grey box).

628 4) On the new page that loads, enter a username, name (optional), organization
629 (optional), email, and password (twice for confirmation) in the spaces provided.

630 4) Click submit.

631 5) Sign-in to your new GNPS account <http://massive.ucsd.edu/ProteoSAFe/> and
632 check that your GNPS credentials work for logging in to MassIVE.

633

634 **Box 1: Navigating the User Workspace Portal**



635

636 At the top of the GNPS website, users will find a banner that allows them to navigate their
637 personal workspace and access additional resources such as the help forum and
638 molecular networking documentation. Within this space, the ‘My User’ tab provides a way
639 to view all MassIVE datasets and reference spectra deposited by the user, and the ‘Jobs’
640 button allows easy access to all jobs submitted by the user through the GNPS and
641 MassIVE interfaces. Clicking on ‘MassIVE datasets’ allows the user to browse and
642 subscribe (section 3.7.3) to all public MassIVE datasets with GNPS in the title. Additionally,
643 this banner is a portal to all resources for help using GNPS. The ‘Documentation’ link in
644 the banner takes the user to the GNPS documentation website, which has step-by-step
645 instructions and links to tutorial videos as well as access to the ‘legacy’ documentation
646 (from a menu on the right-hand side of the page) that can provide additional information
647 to the user. The ‘Forum’ button opens a Google groups forum where users can post
648 questions, have discussions and report potential bugs. The corresponding online tutorial
649 can be accessed at: <https://ccms-ucsd.github.io/GNPSDocumentation/quickstart/>

650

651 **3.2.2 Deposit data files by submitting a dataset (SI Fig. 3) :**

652 An online tutorial on how to submit a dataset to MassIVE can be accessed at: <https://ccms-ucsd.github.io/GNPSDocumentation/datasets/#submitting-gnps-massive-datasets>

653

654 There are two steps to submitting a dataset to the GNPS-MassIVE repository:
655

656 **Step 1 (SI Fig. 3a).** Upload your data files to the MassIVE web server using an
657 FTP client - Timing 10 min to get the upload process started.

658 Of the many free dedicated FTP clients, the following are more popular ones that have
659 been tested with MassIVE: WinSCP, CoreFTP, and CoffeeCup Free FTP for Windows,
660 and Cyberduck or FileZilla for Macintosh. Caution: when downloading an FTP client for
661 use, make sure it comes from a trusted source to avoid malware. Data files transferred to
662 MassIVE should be in .mzXML, .mzML, .mgf formats. The data that is uploaded should
663 **not be in a file archive (e.g. zip, tar) format.** It is also encouraged that the original vendor
664 raw data files (e.g. .wiff for AB Sciex, .yep for Agilent, .d for Bruker, .lcd for Shimadzu, .raw
665 for Thermo Scientific) are uploaded together with the open formats as described below.

666
667 1) Change file protocol to FTP and log onto the FTP server with the host name
668 *massive.ucsd.edu* using your MassIVE web account username and password in the FTP
669 client program for FTP file transfer. Most FTP clients use this "Quick Connect" feature.
670 Alternatively, type in the FTP server name, username and password, and then connect
671 directly.

672
673 **Step 2 (SI Fig. 3b).** Run the MassIVE dataset submission workflow on the
674 uploaded files as follows:

675
676 1) Load the home page for MassIVE from the GNPS home page by scrolling down to
677 the GNPS-MassIVE datasets section and click on the 'Deposit dataset' bar in the 'Create
678 Public datasets' block. Alternatively, click on the 'Submit your data' link in the paragraph
679 titled 'Submit Data' on the MassIVE home page. A direct way to deposit the data is to
680 navigate directly to the MassIVE home page (<http://massive.ucsd.edu/ProteoSAFe/>). This
681 will bring up the Dataset Submission workflow input form, on which there are varying
682 numbers of fillable fields under each of the following sections described below.

683
684 The reader can follow along (**SI Fig. 3**), as this has already been completed for the
685 MassIVE dataset [MSV000083437](#).

686
687 2) In the 'Workflow Selection' section:
688 Enter a title for your dataset, **noting that GNPS datasets must have a 'GNPS'**
689 **prefix in the title** in order for these GNPS-MassIVE datasets to be visible to GNPS users.
690 **Adding GNPS in the title is therefore absolutely !IMPORTANT! for the dataset to**
691 **become a part of the community and ensures that the data becomes alive (Section**
692 **3.6) and enables subscriptions and other analysis features specifically used for the**
693 **GNPS community (Section 3.6).** If a "GNPS" tag is not added at the beginning of the title
694 it will not be part of the GNPS analysis infrastructure. Currently all of MassIVE has almost
695 ~11,000 public mass spectrometry datasets (mostly proteomics), ~1,100 of which are also
696 part of GNPS. If GNPS is not added from the beginning it is possible to go to MassIVE,
697 log-in and edit the title at a later time.

698
699 To satisfy this requirement for the dataset that reader will use in this tutorial,
700 MassIVE dataset [MSV000083437](#) has been titled "GNPS Example Dataset_GF vs. SPF
701 Mouse Duodenum."
702

703 3) In the 'Dataset Metadata' section:

704 To minimize the burden to make datasets for GNPS analysis and to enable as
 705 much flexibility in what additional information the user wants to make available, very few
 706 metadata fields are absolutely required, although the user is encouraged to provide as
 707 much metadata as possible. It should be noted that the datasets that have the most
 708 information associated with it are also the datasets that are the most visible to the
 709 community. Fields for metadata relevant to the dataset being submitted are listed in the
 710 table below. The first three fields ('Species', 'Instrument' and 'Post-Translational
 711 Modifications') are backed by lists of standardized controlled vocabulary (CV) terms,
 712 maintained by organizations such as the [HUPO Proteomics Standards Initiative](#)¹¹⁴ and
 713 many others CVs that the user can implement^{114, 115}. To search these terms, type at least
 714 3 characters into any of these text boxes, and a drop-down list of supported terms that
 715 match your query will be displayed. To select a term, click on it in the drop-down list and it
 716 will be added to your dataset. **Using the official CV to tag your dataset greatly
 717 increases the likelihood that it will be found and processed correctly by any
 718 automated software that may interface with the MassIVE repository.** If the term you
 719 want is not present in the list, you can type your custom text in the text box and click the
 720 adjacent 'Add' button to tag your dataset.

721

722 **Table 2.** Metadata Categories for Data Upload to MassIVE

Metadata Category	Required	Notes	Example Dataset MSV000083437
Species	Yes	Enter custom text if the correct species for your dataset is not supported in the list or if you sample is not a specific species (e.g. environmental sample or community of organisms).	<i>Mus musculus</i> (house mouse)
Instrument	Yes	Enter custom text if the correct instrument for your dataset is not supported in the list.	maXis
Post-Translational Modifications	Yes	For small molecule metabolomics datasets the appropriate entry in the drop-down list is: 'PRIDE:0000398, No PTMs are included in the dataset'.	No PTMs included in the dataset
Keywords to assign to your dataset	Yes	Your dataset must be tagged with at least one keyword - there is no limit. Keywords are custom text, so you must click the 'Add' button after entering text.	mouse duodenum
Principal	Yes	To identify the lab providing	Pieter Dorrestein

Investigator		the data.	(pdorrestein@ucsd.edu) UCSD, United States
Description	No	Recommended to provide as much detail as possible	N/A

723

724 Metadata (sample information) for MassIVE dataset [MSV000083437](#) has been added as
725 shown in **SI Fig. 2b** and is tabulated above.

726

727 4) In the 'Dataset File Selection' section there are eleven different file types that can
728 be added and these are organized into three different categories - required, recommended
729 or optional. **Most of these file categories are not strictly required. The only official**
730 **file requirement for a MassIVE dataset is that at least one file is submitted in either**
731 **the 'Raw Spectrum Files' or 'Peak List Files' categories. If a submitted dataset does**
732 **not meet the additional requirements for a '[complete](#)' submission, then it is**
733 **considered 'partial', which is currently standard for small molecule datasets that are**
734 **a part of GNPS.**

735

a) *Recommended for all submissions*

736

i) Raw Spectrum Files – Raw mass spectrum files in a non-standard or
737 instrument-specific format, such as AB Sciex .wiff files, Agilent .yep files,
738 Shimadzu .lcd files, Bruker .d files Thermo Scientific .raw files, Waters .raw
739 files.

740

ii) Peak List Files – Processed mass spectrum files in a standardized format.
741 The following formats are recognized by MassIVE as valid for this category:
742 .mzXML, .mzML, and .mgf. This is the file from which GNPS analysis is
743 enabled.

744

b) *Strongly encouraged for submissions to improve the ability to interpret the final
745 molecular networks.*

746

i) Supplementary Files – All remaining files relevant to this dataset that do not
747 properly fit into any of the other listed file categories. **A metadata file**
748 **(sample information in a tab delimited text format) with relevant**
749 **attributes that can be used for visualizing the data in networks should**
750 **be included here (see Box 3).**

751

c) *Required for "Complete" Submission* Result Files – **Not necessary for small
752 molecule workflows - although possible and encouraged.** Spectrum
753 identifications in a standardized format. The following formats are recognized by
754 MassIVE as valid for this category: mzIdentML¹¹⁶ and mzTab¹¹⁷, mzTab-M¹¹⁸.

755

i) Search Engine Files – The output of any search engine or data analysis
756 tools or pipelines that were used to analyze this dataset, unless provided in
757 a standardized format recognized by the 'Result Files' category (see
758 above).

759

d) *Optional*

760

i) License Files – Specifying how and under what conditions the dataset files
761 may be downloaded and used. Multiple license files may be uploaded, if
762 appropriate. By default, you can simply leave the 'Standard License'
763 checkbox checked and your dataset will be submitted under the default
764 [Creative Commons CC0 1.0 Universal](#) license. However, if you wish to

- 765 provide your own license, then you can uncheck this box and then assign
766 your own file to the 'License Files' category.
- 767 ii) Spectral Libraries – Any custom spectral library files that were searched
768 against in the analysis of this dataset, or that were generated using the
769 spectrum files provided in this dataset, if applicable.
 - 770 iii) Methods and Protocols – Any open-format files containing explanations or
771 discussions of the experimental procedures used to obtain or analyze this
772 dataset.
- 773 e) *Optional, mostly relevant to peptidomics and proteomics projects*
- 774 i) Quantification Results – Any data and metadata generated by the analysis
775 software used. Typically applied to the quantification analysis of peptides
776 and proteins.
 - 777 ii) Gel Images – Any gel image files generated, in the event that two-
778 dimensional gel electrophoresis has been used as a separation method.
 - 779 iii) Sequence Databases – Any files from protein or other sequence databases
780 that were associated with or searched against in the analysis of this dataset,
781 if applicable (usually .fasta format).
- 782

783 For readers that are following the example, peak List files were uploaded previously for
784 dataset [MSV000083437](#), as illustrated in **SI Fig 3b**, where nine folders (Control, GF1,
785 GF2, GF3, GF4, SPF1, SPF2, SPF3, SPF4) have been added.

786

787 5) 'Mapping Spectrum Files to Identification Files' is **not necessary for small**
788 **molecule workflows**. In order for a submission to qualify as 'complete', each spectrum
789 (data) file referenced within a "Result File" must be associated with a file from the "Peak
790 List Files" category. This section is where these two types of files are associated with each
791 other as appropriate.

792

793 6) The 'Dataset Publication' section has three optional fields to:

- 794 a) 'Enter a Password' (e.g. to share selectively with collaborators and
795 manuscript reviewers),
- 796 b) 'Share on ProteomeXchange' is **not applicable to small molecule**
797 **workflows**: checking the box will submit and announce the dataset via the
798 ProteomeXchange consortium at the time that it is made public on
799 MassIVE. The dataset will not appear publicly in either repository until you
800 click the 'Make Public' button on your dataset's status page (see below).
- 801 c) 'Generate a DOI' if you want a Digital Object Identifier to be generated and
802 assigned to this dataset. This is encouraged for all public datasets and can
803 be used in publications.

804

805 7) The section titled 'Advanced Global FDR Settings' is **not applicable to small**
806 **molecule workflows**. It is currently for global False Discovery Rates across submitted
807 files in proteomics datasets.

808

809 8) In the 'Workflow submission' section, enter an email address at which you will
810 receive notifications when workflow jobs are completed.

811

812 9) **!CRITICAL STEP** *Making your dataset public: this is not automatic and must be*
813 *done explicitly after submitting data and generating a dataset MSV accession number.*

814

815 Once a dataset is submitted to MassIVE, it will have an MSV accession number, and will
816 be a private dataset in the repository, accessible only to the submitter through their
817 personal user interface or via a user approved password protected link (e.g. perhaps
818 during a review for publications). To make a dataset public, first select the 'Jobs' tab of the
819 user workspace portal (**Box 1**) to find the dataset. In the list of all job submissions,
820 MassIVE dataset submissions will appear as 'MASSIVE-COMplete' workflows. Click on
821 'DONE' next to the MassIVE dataset to be made public and choose 'Make Dataset
822 Public'. On the MassIVE website, to enable immediate use of the MassIVE dataset
823 for GNPS workflows click on the „Convert Spectra“ tab. This converts the uploaded
824 files to .mzML in a new folder called „ccms peak“. Otherwise, the uploaded data
825 will be queued for this conversion and will not be immediately available.

826

827 The dataset [MSV000083437](#) has been made public, as illustrated in **SI Fig. 3**; this feature
828 enables any reader to interact with the data and follow along with this workflow.

829

830

831 **BOX 2. The Importance of making your GNPS-MassIVE data public.**

832 Many GNPS users do not realize that when they have a dataset with MSV accession
833 number their data is not yet public and thus remains in their private space, in accordance
834 with GNPS-MassIVE philosophy that the data depositor should define how much and when
835 they want to share their data in the public domain. Alternatively, upon submission, users
836 can choose to make a dataset entirely available or 'public' to the GNPS community for
837 browsing, commenting, subscribing, and/or downloading. This not only promotes
838 robustness and reproducibility in MS data analysis, but also provides the user with access
839 to the knowledge of the entire community. Indeed, the utility of GNPS for all users
840 increases as more data becomes public, and the information and knowledge gained by
841 any one user from this free service to the community derives from contributions made by
842 the rest of the GNPS community. Thus, if you are a GNPS user benefiting from community
843 contributions, by making your datasets public (and contributing network annotations,
844 section 3.5), you are giving back to the community. It is encouraged that all users make
845 their data public as early as possible, which provides the depositor with access to
846 advanced features that are not available for private datasets. These features include being
847 able to subscribe to the dataset, find related datasets, share datasets with collaborators,
848 access living data, and utilize emerging features such as Mass Spectrometry Search Tool
849 or MASST (the equivalent of BLAST for small molecules¹¹⁹). It is expected that features
850 will continue to be developed further, thereby continually increasing the value for the end
851 user, of both their own and other public datasets.

852

853 **3.3 Molecular networking in GNPS (SI Fig. 4) - Few minutes to several hours/days** 854 **(depending on dataset size, user expertise)**

855

856 Once MS data files are uploaded as datasets in GNPS-MassIVE, they are available to use
857 for analysis workflows within GNPS. Here we highlight how to execute the molecular
858 networking workflow. A dataset can be recalled from either private or public domains in

859 MassIVE for networking analysis. Once data files have been added, they will be populated
860 in the 'Basic Options' section of the workflow selection. The user must then input a number
861 of parameters before running the GNPS job in both the 'Basic Options' section and in a
862 number of 'Advanced Options' sections. The advanced parameters are dependent on
863 analysis platform, experimental setup and conditions for acquisition of mass spectra, and
864 will require the user to understand their ionization methods, fragmentation conditions and
865 energies, mobile and stationary phases, and the fragmentation behavior of molecules of
866 interest. Suggested settings for a variety of platforms are provided in the experimental
867 section (Equipment Setup, Mass Spectrometry). A GNPS job will take approximately 10
868 min for small datasets (up to 4 LC/MS files), 1 hr for medium datasets (5 to 400 LC/MS
869 files), and several hrs (to days) for larger datasets (400+ LC/MS files).

870

871

872 **Molecular networking workflow**

873

- 874 1) Log in to GNPS (refer to section 3.2.1 for information about how to set up account).
875 The GNPS website banner contains tabs to navigate the platform, including tabs
876 to navigate to MassIVE datasets, help Documentation and Forum, along with
877 Contact information (**SI Fig. 1, Box 1**).
- 878 2) Upload desired dataset(s) to MassIVE (section 3.2.2). This step can be skipped if
879 importing existing data files from MassIVE. Readers following the tutorial can omit
880 this step because the GNPS-MassIVE dataset [MSV000083437](#) already exists.
- 881 3) From the GNPS splash screen (home page), start a molecular networking job by
882 clicking the 'Create Molecular Network' button (**SI Fig. 4a**). This will bring up the
883 main workflow input page which has a number of fillable fields to complete under
884 each of ten sections (**SI Fig. 4b**).
- 885 4) In the 'Networking Parameter Presets' section, one of three options may be
886 selected to set the networking parameters to approximately appropriate values
887 depending on the size of your dataset. Clicking on one of these three options will
888 open a workflow input form in a new tab. The default workflow settings are for
889 'medium data'. 'Small data' refers to a dataset of up to 4 LC-MS files, 'medium data'
890 corresponds to datasets of 5 to 400 LC-MS files, and 'large data' is applicable to
891 datasets of more than 400 LC-MS files (e.g. [MSV000083437](#) is a medium dataset
892 with 113 files in total). Since readers following the tutorial on the dataset
893 [MSV000083437](#) are guided through selection of parameters, no Parameter Preset
894 should be chosen for this example.
- 895 5) In the 'Workflow Selection' section, enter a descriptive name for the job into the
896 'Title' field to facilitate retrieval of the workflow upon its completion. Readers
897 following the tutorial can type 'GF/SPF Mouse Duodenum Example' in the 'Title'
898 field (**SI Fig. 4c**).
- 899 6) Under 'Basic Options', the user will input the LC-MS files for the molecular
900 networking workflow by choosing the 'Select Input Files' tab next to the 'Spectrum
901 Files (Required)' field. A pop-up window with three tabs will appear: 'Select Input
902 Files', 'Upload Files', 'Share Files' (**SI Fig. 4d**). If you are interested in analyzing
903 multiple datasets together, you will have to repeat the above procedure with the
904 other MSV numbers to import them into your user space.

905

906 For readers following the dataset [MSV000083437](#) tutorial, files can be imported by
907 selecting the 'Share Files' tab. In the 'Share Files' window enter the MassIVE
908 accession number for the dataset ([MSV000083437](#)) in the 'Import Data Share' box
909 (**SI Fig. 4e**). After clicking 'Import', the dataset will appear in your GNPS user
910 workspace and files can be selected for the GNPS networking workflow under the
911 'Select Input Files' tab as described below.

- 912 7) For inputting mass spectrometry files already in your user workspace choose the
913 'Select Input Files' tab (**SI Fig. 4f**). From the list of datasets towards the lower left
914 of the window, select all of the files you want to analyze by clicking on individual
915 files or an entire folder. For readers following the tutorial, GF1, GF2, GF3, GF4,
916 SPF1, SPF2, SPF3, and SPF4 should be selected from the folder labeled 'peak'.
917 8) Next click on the 'Spectrum Files G1' button (top of left-hand column list, with green
918 arrow) to mark this folder / files for analysis. Your selection(s) should appear in the
919 'Selected Spectrum Files G1' folder in the right-hand column of the window. For
920 readers following the tutorial, folders containing data for GF1, GF2, GF3, GF4,
921 SPF1, SPF2, SPF3, and SPF4 should now be under 'Selected Spectrum Files G1'
922 (**SI Fig. 4g**).
- 923 9) Load the associated metadata file (see **Box 3** for format) separately into the
924 'Selected Metadata File' folder. To do this, select the file from your workspace list
925 (often within a MassIVE dataset in the folder labeled as 'other'), click on the
926 'Metadata File' tab with the green arrow, and check that the file appears in the right-
927 hand 'Selected Metadata File' folder. For readers following the tutorial,
928 '3DMouse_duodenum_metadata.txt' can be selected from the folder labeled
929 'other' (**SI Fig. 4h**).
- 930 10) Once files have been selected, the popup window can be closed by clicking on
931 'Finish Selection'. Datasets from both your private workspace and the public
932 domain can be recalled using either strategy. For readers following the tutorial, the
933 final data input is shown in **SI Fig. 4i**.
- 934 11) In the 'Basic Options' section, fill in the 'Precursor Ion Mass Tolerance' (PIMT) and
935 'Fragment Ion Mass Tolerance' (FIMT) fields taking into consideration the
936 instrument resolution and calibration, as well as the acquisition parameters and the
937 targeted/anticipated molecular masses (see definitions and **Table 2** below). The
938 default is ± 2.0 Da for PIMT and ± 0.5 Da for FIMT because the reference libraries
939 also contain spectra from low resolution instruments (e.g. ion traps of QqQ). These
940 can be adjusted to any appropriate value. For high resolution instruments the
941 values commonly used are ± 0.01 Da (Orbitrap) and ± 0.02 Da (qTOF) for both
942 PIMT and FIMT.

943

944 For readers following the tutorial example, data were acquired on Bruker MaXis
945 qTOF instrument using ± 0.02 Da. The 0.02 Da value translates into a maximum
946 error of 40 ppm at m/z 500, 20 ppm at m/z 1000 for the precursor ion, and 13 ppm
947 at m/z 1500, which is consistent with the typical m/z range for small molecules.
948 (Note that peptidic small molecules may be 2000 Da or more, although multiply
949 charged, and thus PIMT and FIMT values of 0.03 Da should be used.) Therefore,
950 readers should use ± 0.02 Da for both PIMT and FIMT for the example dataset (**SI**
951 **Fig. 3c**).

952

953 **CRITICAL NOTE:** The default parameters recommended above for high resolution
 954 mass spectrometers will not result in comprehensive searches of the spectral
 955 libraries generated on low resolution mass spectrometers, such as ReSpect⁷⁸,
 956 large portions of MassBanks⁷⁴, GNPS community contributed; a significant portion
 957 of spectra that were annotated by matching to the NIST Mass Spectral Library with
 958 Search Program Data Version: NIST v17 ([https://www.nist.gov/srd/nist-standard-
 959 reference-database-1a-v17](https://www.nist.gov/srd/nist-standard-reference-database-1a-v17)) are also low resolution. In addition the natural
 960 products community contributes annotated spectra that may be high or low
 961 resolution, from a range of different spectrometers.
 962 **!CAUTION!** Though using low resolution parameters may increase the number of
 963 annotations, it will also increase the number of false positive annotations.

964
 965 PIMT: This parameter is used for MS-Cluster^{10,12} and spectral library searching, and the
 966 value influences the clustering of nearly identical MS² spectra via MS-Cluster.
 967 FIMT: For every group of MS² spectra being considered for clustering (consensus
 968 spectrum creation), this value specifies how much fragment ions can be shifted from their
 969 expected *m/z* values.

970
 971 **Table 3.** Absolute mass differences (Da) and associated mass error (parts-per-million,
 972 ppm) for illustrative *m/z* values

	2.0 Da	0.5 Da	0.1 Da	0.05 Da	0.03 Da	0.025 Da	0.02 Da	0.0175 Da	0.015 Da	0.01 Da	0.0075 Da
<i>m/z</i> 200	10000 ppm	2500 ppm	500 ppm	250 ppm	150 ppm	250 ppm	100 ppm	87.5 ppm	75 ppm	50 ppm	37.5 ppm
<i>m/z</i> 500	4000 ppm	1000 ppm	200 ppm	100 ppm	60 ppm	49 ppm	40 ppm	35 ppm	29 ppm	20 ppm	15 ppm
<i>m/z</i> 1000	2000 ppm	500 ppm	100 ppm	50 ppm	30 ppm	25 ppm	20 ppm	17.5 ppm	15 ppm	10 ppm	7.5 ppm
<i>m/z</i> 1500	1333 ppm	333 ppm	66 ppm	33 ppm	20 ppm	16 ppm	13 ppm	11.6 ppm	10 ppm	6.6 ppm	5.0 ppm
<i>m/z</i> 2000	1000 pm	250 pm	50 ppm	25 ppm	15 ppm	12.5 ppm	10 ppm	8.75 ppm	7.4 ppm	5.0 ppm	3.75 ppm

973
 974 **For advanced users:**

975 12) The user should complete the remaining fillable fields in 'Advanced Network
 976 Options', 'Advanced Library Search Options', and 'Advanced Filtering Options'
 977 according to their experimental design. Recommendations and values used for the
 978 example dataset are provided in Table 4 below and **SI Fig. 4j**.

- 979 13) Use the default parameters for 'Advanced GNPS Repository Search Options',
 980 'Advanced Annotation Options', and 'Advanced Output Options'. The option
 981 'Create Cluster Buckets and BioM/PCoA Plots Output' must be enabled in the
 982 'Advanced Output Option' to generate bucket tables and PCoA plots from the
 983 'Export' and 'Advanced Views' options on the job status page (**SI Fig. 4j**).
- 984 14) Finally, under 'Workflow Submission', the user should enter an email address to
 985 receive notifications when workflow jobs are completed. Readers following the
 986 tutorial should do this to receive notification when the example job is completed.
- 987 15) Click 'Submit' to begin the job. The molecular networking job for the example
 988 dataset ([MSV000083437](#)) should take about 20 minutes.

989
 990 **Table 4.** Parameters for Molecular Networking in GNPS

Advanced Network Options		
Fillable Field	Definition	Recommended User Input
Min Pairs Cos	Minimum cosine score required for an edge to be formed between nodes	Most commonly set to 0.7 when a minimum of 6 ions are matched. When fewer ions are used, it is better to be more stringent and increase this value (e.g. 0.8) but when more ions are required, one can relax this value (e.g. 0.6) ¹²⁰ (Use 0.7 for example MSV000083437)
Minimum Matched Fragment Ions	Minimum number of common fragments that must be matched by two nodes for an edge to be formed	<i>Highly</i> dependent on the experiment – While 6 is listed as default, a lower value could be used if the user wants to be less restrictive or if the sample largely contains molecules with a small number of fragment ions. The maximum number of significant annotations are found when this value is set to 4 or 5 ¹²⁰ . (Use 4 for example MSV000083437)
Network TopK	Maximum number of neighbor nodes for one single node. The edges between two nodes are kept only if both nodes are within each other's TopK most similar nodes. If this value is set at 10, a single node may be connected to up to 10 other nodes.	Default is set to 10. Adjusting this value enables the network to be more or less stringent. Keeping this value low makes very large networks (many nodes) much easier to visualize. (Use 10 for example MSV000083437)
Minimum Cluster Size	Minimum number of identical MS ² spectra that are merged by MS-Cluster	This is a very important parameter as it is a very good quality of spectra filter. If this is set to 1 then each MS ² spectrum is compared to

	for the consensus spectrum to be represented as a node	all other MS ² spectra, including MS ² spectra of noise thus increasing the computational time and exploding the final molecular network. By requiring more identical spectra to be merged (clustered) before considering the MS ² spectral alignments it will ensure that only reproducible and higher quality data is used in the final molecular network. The default is set to two but if it is a very large dataset (hundreds to thousands of files) one may use 5 or more while for smaller datasets (e.g. 1 or 2 files) it may be set to 1 or 2. (Use 4 for example MSV000083437)
Run MSCluster	Clusters MS ² spectra and creates consensus MS ² spectra using the specified mass tolerance settings	Set to 'yes' for classical molecular networking (Set to 'yes' for example MSV000083437)
Maximum Connected Component Size (Beta)	Maximum number of nodes that can be connected in a single component (molecular family) of a molecular network. This process iteratively breaks up large 'hairball' networks (of false positives) by removing the lowest scoring alignments (by cosine score) first until the resulting pieces fall below the maximum size.	Default setting is 100 – this value can be set to 0 to allow for an unlimited number of nodes or a higher setting can be used for larger datasets or for datasets containing many structurally-related molecules. (Use 100 for example MSV000083437)
Metadata File (= sample information file)	File added to the analysis that describes the experimental setup and details to allow for better downstream data visualization, analysis and interpretation	Add as a .txt file that follows the template and instructions available in the supporting information (Metadata file uploaded is described in step 9, section 3.3. Example metadata can be found in SI Tables 10 and 11, and a description of how to create a metadata file can be found in Box 3 .)
Group Mapping and Attribute Mapping	Legacy version of metadata file	It is encouraged to use the metadata table instead
Advanced Library Search Options		
Library Search	Minimum number of shared	The default value is 6. Dependent on the aim

Min Matched Peaks	fragment ions to make a library match.	of the experiment: a lower value may yield more tenuous matches to library spectra, suitable for exploratory structure searching; a higher value, selecting for closer matches, facilitates dereplication of putative known compounds. The impact of this parameter is discussed in Scheubert et al. ¹²⁰ (Use 4 for example MSV000083437)
Score Threshold	Minimum cosine similarity score to make a library match.	The default setting is 0.7. Dependent on the aim of the experiment: a lower value may yield more tenuous matches to library spectra, suitable for exploratory structure searching; a higher value, selecting for closer matches, facilitates dereplication of putative known compounds. (Use 0.7 for example MSV000083437)
Search Analogs	Matches query spectra against library spectra with a modification tolerant search within a specified range for mass differences. Precursor ion <i>m/z</i> are allowed to deviate up to a user-defined maximum. Fragment ions that differ by the mass difference of the two parent ions are also considered.	Dependent on the user's preferences, selecting 'Do Search' requires more computing time but also the results are more exploratory. It allows for dereplication not only of identical molecules but also related molecules.
Maximum Analog Search Mass Difference	Maximum mass shift allowed between the query spectra and library spectra <i>m/z</i> values to make a library match.	Use default parameter of 100 Da: may increase or decrease the value depending on properties such as anticipated molecular mass shift of related molecules in the samples. (e.g. 162 Da is a common mass shift for oligosaccharides). The larger this value the more likely spurious matches will be found.
Advanced Filtering Options		
Filter Below Std Dev	Applied before MS-Cluster. For each MS ² spectrum, the 25% least intense fragment ions are collected and the std-dev is calculated, as well as the mean. A minimum peak intensity is calculated as mean + k * std-dev where k is user-selectable. All peaks below this threshold are deleted. By default, this filter is	<i>Using this filter is not recommended.</i> A default value of 0 should be used so that no filter is applied.

	inactive (value is set to 0).	
Minimum Peak Intensity	All fragment ions in the MS ² spectrum below this raw intensity will be deleted.	This filter is infrequently used. Use a default value of 0 so that no filter is applied, especially if the raw intensities of your data are very low.
Filter Precursor Ion Window	All peaks in a +/- 17 Da around precursor ion mass are deleted. This removes the residual precursor ion, which is frequently observed in MS ² spectra in the comparison of all spectra for molecular networking.	Apply filter, which is the default option.
Filter Library	Applies the above precursor ion window filter to the library as well.	Apply filter, which is the default option
Filter Peaks in 50 Da Window	Removes peaks that are not one of the top 6 most intense within a +/- 50 Da window.	This is commonly turned on. Dependent on the dataset: if samples contain a large number of low mass molecules or are complex mixtures containing compounds of low titer, this filtering should be turned off, as it may filter out relevant peaks that could be signals.

991

992

993 **BOX 3: Sample information (metadata) collation and input** - Timing typically 1-2 hours
 994 for a small dataset; up to a few days for large complex metadata entries of large
 995 datasets¹²¹.

996 The inclusion of a metadata (sample information) table is extremely valuable for
 997 interpreting the molecular network that is generated using the data. Although a time
 998 consuming step, it is also one of the most valuable steps for interpreting the final molecular
 999 network. The more time spent on curating sample information (metadata), the more useful
 1000 the resulting molecular network will be. The metadata table links the MS files uploaded
 1001 and selected for molecular networking analysis in GNPS with various attributes of the
 1002 collated data based on the filename (such as "Filename.mzXML"). For instance, the
 1003 metadata table provides the necessary information to visualize the "origin" of the detected
 1004 metabolites when "origin" is one of the attributes used in the metadata table (e.g. column
 1005 heading: ATTRIBUTE_Origin). A metadata file can be created as follows:

- 1006 1) The metadata table must be provided as a text file (tab separated) and can be
 1007 prepared in a text editor of choice (e.g. Microsoft Excel, Notepad++ for Windows,
 1008 gedit for Linux, and TextEdit or TextWrangler for Mac OS) .
- 1009 a) When uploading metadata associated with a GNPS job, specifically
 1010 formatted column headers are required. The first column header must be
 1011 "filename" (no capitals, case-sensitive and no unusual characters such as
 1012 @, #, !). **Important:** The filenames must be the filenames of the data (to
 1013 be) uploaded to GNPS-MassIVE otherwise the metadata cannot be linked

1014 to the data. We recommend not to use any special characters such as @,
1015 #, ! or spaces in any of the metadata fields.

1016 b) Each other column must begin with the phrase "ATTRIBUTE_" before any
1017 header description (e.g. ATTRIBUTE_Origin)

1018 2) In order for sample information (metadata) to be incorporated into global
1019 metaanalyses, the template provided in SI Table 10 should be utilized and labeled
1020 "gnps_metadata.tsv".

1021 There are a number of advantages to uploading a metadata table associated with a GNPS
1022 job. When the network generated after data processing is subsequently opened in
1023 Cytoscape, the nodes of sub-networks can be visualized based on their associated
1024 metadata. This can be represented as a pie chart contained within each node. Additionally,
1025 metadata can be used to color-code categories of samples when visualizing the MS²-
1026 based statistics, such as principal coordinates analysis (PCoA) in browser using the
1027 EMPEROR package¹²² available in Qiime2¹²³. This allows the user to quickly attribute the
1028 molecular differences of the samples to certain characteristics found in the metadata. For
1029 example, if two distinct groups appeared in the PCoA plot, it would then be possible to
1030 color all samples of type one blue and all samples of type two red in order to determine if
1031 this attribute could be responsible for the separation. However, it is important to note that
1032 PCoA is only visual and doesn't give any statistical support; a PERMANOVA analysis
1033 would have to be performed in order to actually test whether an attribute is responsible for
1034 separation. Finally, data sharing is a vital part of modern science because it gives
1035 opportunities for collaboration, wider scope analyses, and transparency promotes
1036 reproducibility and thus scientific rigor. Without metadata attached, public data has less
1037 value, will not be discovered as easily by others, and will not provide meaningful results
1038 with MASST¹¹⁹. A metadata text-based search is being engineered in GNPS so that all
1039 public data files with specific metadata entries may be re-analyzed together. When no
1040 metadata is available, these public data will not be included in such searches. In short, the
1041 visibility and value of data goes up by improving the amount of metadata that is uploaded.
1042 Therefore, uploading metadata associated with the MS data to GNPS promotes a more
1043 universal approach to science.

1044 3) In cases where you want to add a new/external metadata file (tab delimited text
1045 format) into your workspace, under the 'Upload Files' tab: select the destination folder for
1046 the upload on the left and drag the file for upload to the 'File Drag and Drop' box on the
1047 right before following the same actions listed in this step. The online tutorial on metadata
1048 formatting, including a template file, can be accessed at: [https://ccms-
1049 ucsd.github.io/GNPSDocumentation/networking/#metadata](https://ccms-ucsd.github.io/GNPSDocumentation/networking/#metadata).

1050

1051 *Metadata format for 'ili'*¹¹¹

1052 *For 2D or 3D molecular cartography using 'ili', metadata must contain the following*
1053 *additional information. The spatial coordinates that dictate the spatial distribution of a*
1054 *detected metabolite in a 2D (.PNG format) or 3D image (.STL format) must be included.*
1055 *In addition to the column "filename", extra columns containing the following information:*
1056 *"COORDINATE_x", "COORDINATE_y", "COORDINATE_z", "COORDINATE_radius"*
1057 *have to be added. The x, y and z correspond to the 3D coordinates and the radius*
1058 *corresponds to the approximate values of radii of the sampling points. An image viewer*
1059 *can be used to estimate this value; for example, half of the difference between boundaries*
1060 *of a sampling point in a horizontal or vertical dimension can be estimated. Additional*

1061 information related to 'ili can be obtained through
1062 <https://github.com/MolecularCartography/ili>.

1063

1064

1065 **3.4 Visualization of the molecular network**

1066 To visualize molecular networks generated, the user can either (1) directly visualize their
1067 network in the GNPS web browser for exploratory purposes, or (2) import data tables
1068 generated for viewing in third party software, such as Cytoscape³⁷, which is a free software
1069 tool that enables visualization of the entire molecular network. These methods are
1070 complementary to one another and the user should choose the preferred visualization
1071 strategy based on their data analysis needs. The GNPS in-browser visualization tool is a
1072 quick, simple way to begin analyzing data, particularly if the user wants to view and
1073 compare MS² spectra within the network. However, in-browser visualization only allows
1074 the user to view one molecular family (sub-network) at a time. For more advanced data
1075 analysis and formatting options, the user can visualize their network offline in Cytoscape,
1076 a program originally introduced by the systems biology community to allow visualization of
1077 the complex relationships in biological sequence data. With Cytoscape, one can visualize
1078 the chemical space that was detected in the mass spectrometry experiment as a molecular
1079 network and provides a way to encode any property of the network (i.e. node label, shape,
1080 color or size as well as edge label, thickness, etc.) with a metadata category (i.e. cohort,
1081 cosine score, compound source). An online tutorial can be accessed at: [https://ccms-
1082 ucسد.github.io/GNPSDocumentation/networking/#online-exploration-of-molecular-
1083 networks](https://ccms-ucsd.github.io/GNPSDocumentation/networking/#online-exploration-of-molecular-networks).

1084

1085

1086 **3.4.1 Molecular network visualization in browser**

1087 After completing the above molecular networking workflow, data analysis can be
1088 performed directly in the GNPS web interface. The user can access the in-browser data
1089 analysis options from the job status page (**Fig. 4**), several of which are described in **Table**
1090 **5**.

1091

Job Status

Workflow METABOLOMICS-SNETS-V2
 DONE [Clone] [Restart][Delete]

Default Molecular Networking Results Views
 [View All Library Hits | View Unique Library Compounds | View All Clusters With IDs]

Network Visualizations
 [View Spectral Families (In Browser Network Visualizer) | Network Summarizing Graphs]

Methods and Citation for Manuscripts
 [Networking Parameters and Written Network Description]

Export/Download Network Files
 [Download Clustered Spectra as MGF | Download GraphML for Cytoscape | Download Bucket Table | Download BioM For Qiime/Qiita | Download Metadata For Qiime | Download ili Data]

Status

Advanced Views - Global Public Dataset Matches
 [View Matches to All Public Datasets]

Advanced Views - Third Party Visualization
 [View Emporer PCoA Plot in GNPS | View ili in GNPS]

Advanced Views - Networking Graphs/Histograms
 [Nodes, MZ Histogram | Edges, MZ Delta Histogram | Edges, Score vs MZ Delta Plot | Library Search, PPM Error Histogram]

Advanced Views - Misc Views
 [View Network, Node Centric | View Network Pairs | Networking Statistics | View Raw/Unclustered Spectra | View Compounds and File Occurrence]

Advanced Views - Make Dataset Public Documentation
 [Make Public Dataset]

Advanced Views - Experimental Views
 [Direct Cytoscape Preview/Download]

User emgentry (emgentry.nc@gmail.com), UC San Diego

Title GF/SPF Mouse Duodenum Example

1092
 1093
 1094
 1095
 1096

Figure 4. GNPS Job Status Page.

Table 5. Data analysis options

Data Analysis Option	Description
View all library hits (SI Fig. 5a)	View all spectra with reference database matches and assess the quality of the MS ² match using the 'View Mirror Match' option. Readers following the tutorial example can view the mirror plot for cholic acid (SI Fig. 5a) in order to compare experimental spectra with library annotation. Readers can investigate mirror plots for other bile acids, as bile acid discovery is the focus of this example.
View unique library compounds (SI Fig. 5b)	View all <i>unique</i> spectral matches to the reference database and perform side-by-side comparison between the query spectrum and reference spectrum. Readers following the tutorial can view query and reference spectra for cholic acid (SI Fig. 5b).
View all clusters with IDs (SI Fig. 5c)	View all consensus MS ² spectra that make up a node.
View spectral families (SI Fig. 5d)	List of all spectral families (nodes that are connected to one another) and view individual sub-networks using in browser visualization
View EMPeror PCoA plot	Measures the binary Jaccard distance between samples based on presence/absence of molecular

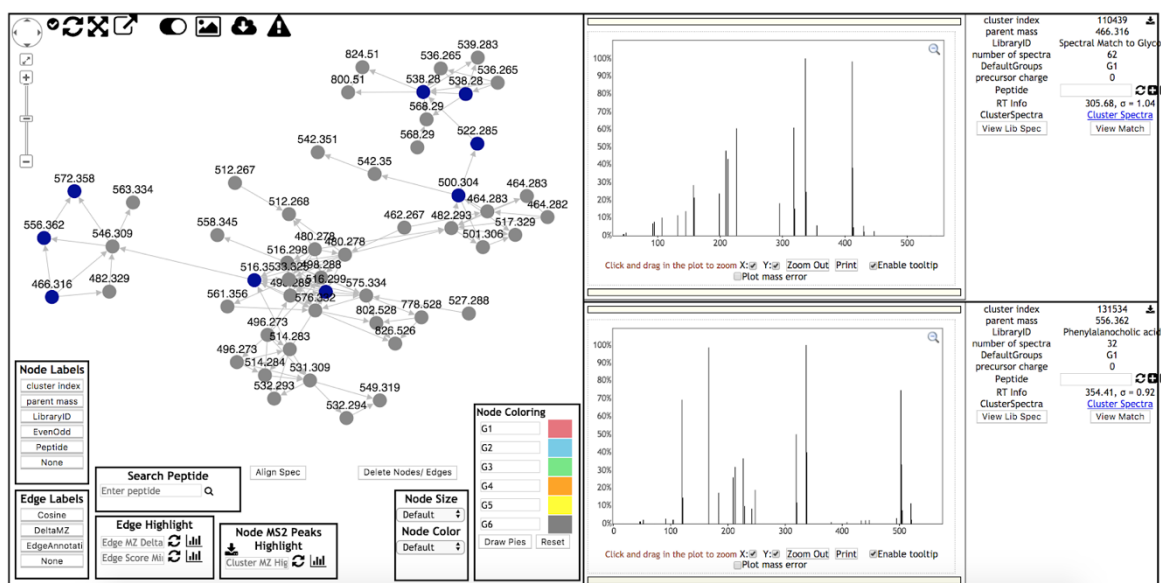
	features with associated MS ² spectra as defined by the mass spectral molecular network. Interactive Principal Coordinates Analysis (PCoA) visualization is enabled through EMPeror ¹²² .
--	---

1097

1098 The “View spectral families” option lists each individual molecular family that contributes
1099 to the entire molecular network and displays the number of MS² spectra and spectral
1100 matches to the reference library that contribute to a given sub-network. This function also
1101 allows users to visualize each sub-network individually in the web browser by selecting the
1102 “Visualize network” link. Once the in browser network is displayed, the user can
1103 immediately distinguish between nodes with library matches (blue circles) and
1104 unannotated nodes (gray circles). Edges are represented by gray arrows that point from
1105 the low mass spectra to the high mass spectra. Further data analysis can be performed in
1106 this online interface as described below:

- 1107 • *Node Labels* - Nodes can be labeled by their index number given by MS-cluster,
1108 parent mass, or library annotation name. Additionally, the node can be labeled by
1109 a binary system to denote if the parent mass is even (1) or odd (0) to assist in
1110 visualizing the nitrogen-rule ¹²⁴, or with a peptide annotation label (see Search
1111 Peptide below). If no node label is desired, select ‘None’.
- 1112 • *Node Coloring* - This legacy feature creates pie charts to visualize mapping of
1113 metabolites into different groups. However, this option does not use the sample
1114 information (metadata) table and will work only if files were inputted into different
1115 groups by the user.
1116 !CAUTION! Note also that this is not a quantitative representation of the data
1117 because it relies only on MS² spectral counts. Rather this feature can be used to
1118 understand presence versus absence of compounds in specific groups.
- 1119 • *Edge Labels* - Edges connecting two nodes can be labeled with either the cosine
1120 score or the mass difference between the parent *m/z* values (‘DeltaMZ’). If no edge
1121 label is desired, select ‘None’.
- 1122 • *Edge Highlights* - Edges by default are represented as arrows pointing from low
1123 mass spectra to high mass spectra, and can be colored. Users are able to enter a
1124 mass difference (*m/z* delta) of their choice in the ‘Edge MZ Delta’ field, causing
1125 those edges to be highlighted in red. Clicking on the graph icon next to ‘Edge MZ
1126 Delta’ opens a new window containing a graph that shows the distribution of all
1127 edge *m/z* delta values in the sub-network. Selecting a peak in this ‘Network MZ
1128 Delta Histogram’ highlights the corresponding edges in red. The same function can
1129 be performed for ‘Edge Score Minimum’ to highlight edges that have a cosine score
1130 greater than what the user enters.
- 1131 • *Node size/color* - The size and color of nodes can be adjusted based on spectral
1132 counts, precursor intensity, number of files, parent mass, even/odd mass, or
1133 precursor charge.
- 1134 • *Node MS² Peaks Highlight* - This option allows users to search the sub-network for
1135 molecules that contain an MS² fragment of interest. To perform this query, first click
1136 the download button within this box to pull all of the MS² spectra into the browser.
1137 The desired *m/z* value can then be entered into the field to highlight the nodes
1138 comprising spectra which contain the desired product ion. Alternatively, the

- 1139 histogram icon can be selected to visualize all product ions from the MS² spectra
 1140 in the sub-network.
 1141 • *Align Spectra* - This function enables direct comparison between the spectra of two
 1142 connected nodes at the peak level. To perform this analysis, the user should first
 1143 select an edge connecting two nodes, which pulls up the spectra for each node in
 1144 the right display window. Clicking the “align spec” button overlays the spectra,
 1145 where red peaks represent peaks of the exact same masses shared between the
 1146 top and bottom spectra and blue peaks denote peaks matching at shifted masses.
 1147 • *Search Peptide* - This is a function added to GNPS to support proteomic and
 1148 peptidomic dataset analysis. If a peptide sequence is found to be associated with
 1149 the molecular family and was found through automated peptide mining in MASSIVE
 1150 then the amino acid sequence entered here will be searched.
 1151



1152
 1153
 1154 **Figure 5.** In browser visualization of the bile acid spectral family from dataset
 1155 MSV000083437.
 1156

1157 **3.4.2 Assessing the quality of a library hit.** All spectral matches are putative
 1158 annotations⁶ until experimentally validated. Spectral matches from molecular networking
 1159 analysis are annotations at level 2 (compounds that have been putatively annotated e.g.
 1160 no reference standards) or 3 (compounds that can be putatively assigned to a chemical
 1161 class based on physicochemical properties and/or spectral similarity) before validation with
 1162 chemical standards. For level 1 annotation, the molecules would have to be isolated and
 1163 structures elucidated or confirmed with other techniques such as NMR or X-ray analysis,
 1164 or matching MS² and retention times, together with co-analysis with pure standards, ideally
 1165 under more than one chromatographic condition. All non-annotated molecules in a
 1166 molecular network are level 4 unless they are part of a molecular family containing a library
 1167 match. Levels were defined by the 2007 Metabolomics Initiative¹⁴, and subsequently
 1168 refined by the Compound Identification work group of the Metabolomics Society at the
 1169 2017 annual meeting of the Metabolomics Society¹²⁵. In order to judge the quality of a
 1170 match, it is important to consider the mass accuracy of the reference spectra (resolution
 1171 and calibration of the instrument) as compared with that of the experimental spectra. The
 1172 sample type, experimental setup, and associated sample information (metadata) should

1173 also be taken into account when judging the accuracy of the matches. Notably,
1174 MS² spectra typically cannot differentiate regio- or stereo-isomers and additional
1175 experiments, including comparison with standards, are required to assign the absolute
1176 structure.

1177 To decrease the impact of this variation all spectra, when compared, are subjected
1178 to a square root conversion. This decreases the high intensity ions and increases the low
1179 intensity ions. Furthermore, to address variability in data quality and source of the
1180 reference spectra, GNPS utilizes a ranking system for submitted reference spectra, to
1181 enable filtering of the reference library either before performing molecular networking or
1182 afterwards, which is the default approach. Similarly the instrument that the reference data
1183 were collected on can be considered after doing the analysis in GNPS using post-
1184 molecular networking filtering capabilities. 'Gold' reference spectra can only be submitted
1185 by approved users and must originate from fully characterized synthetic or purified
1186 compounds. This is the same gold standard by which other metabolomics reference
1187 libraries such as NIST17⁷², METLIN⁷³ mzCloud (<https://www.mzcloud.org/>)⁷⁶, WeizMass
1188 ([https://www.weizmann.ac.il/LS_CoreFacilities/weizmass-spectral-library-high-
1189 confidence-metabolite-identification](https://www.weizmann.ac.il/LS_CoreFacilities/weizmass-spectral-library-high-confidence-metabolite-identification))¹²⁶ libraries are curated. Gold level spectra comprise
1190 83% of the MS² spectra provided to GNPS as libraries. A 'silver' rating signifies that the
1191 spectrum was submitted with an associated publication. However, GNPS also curates
1192 crowdsourced knowledge from users in the community. All remaining reference spectra
1193 provided by the user community receive a 'bronze' rating to denote that the annotation is
1194 contributed by users including partial or putative annotations. The annotation within GNPS
1195 can be made directly from the data and thus relies on the expertise of the experimentalist
1196 and purification of the molecules is not required. This gives access to a curated reference
1197 database that is crowdsourced and does not rely on commercially available standards. For
1198 example, most natural products from microbes, food and plants are not commercially
1199 available and thus the crowdsourced knowledge capture provides a resource of
1200 information that is inaccessible any other way. The only other resource that currently
1201 accepts putative and partial annotations is MassBank EU
1202 (<https://massbank.eu/MassBank/>). Examples of useful but partial annotations include
1203 modifications of molecules, such as oxidation of a molecule in which the site of oxidation
1204 is unknown¹²⁷ and thus a SMILES or InChI cannot be drawn but the partial annotation
1205 provides valuable insight to the end user. Additional partial annotations would include
1206 adduct clusters such as sodium formate clusters or polymeric substances, including
1207 oligosaccharides, commonly detected in mass spectrometry where a structure cannot be
1208 drawn but is useful knowledge for the community when performing an untargeted LC-
1209 MS/MS experiment. Users can use the above information along with the corresponding
1210 cosine score, which takes into account the number of matching fragment ions and
1211 differences in peak intensities, and parent mass accuracy to assess the quality of
1212 annotation. An empirical cut-off for cosine scoring of 0.7 with 6 MS² ions matching is the
1213 default setting in GNPS. On average this gives rise to 91% accurate annotations, and ~1%
1214 incorrect annotations, with the remainder being attributed to possible isomers (4%) or
1215 having not enough information by the user to judge (4%)¹²⁰. However, using a target decoy-
1216 based method to estimate confidence measures of annotations and false discovery rates
1217 (FDR) in large scale metabolomics experiments, revealed that the annotation quality is
1218 dataset-dependent and dependent on analysis settings such as number of ions that are
1219 required to match. The general trend was that when few MS² ions are required to match,
1220 a much higher cosine is required and fewer matches will be obtained at the same FDR

1221 compared to when more MS² ions are required to match the reference spectra. When more
1222 ions are matched, the cosine score can be lowered. There is an dataset-dependent
1223 optimum for the maximum number of spectral library matches at a specific FDR that is
1224 typically around 4 to 6 minimum matched peaks¹²⁰. Although the confidence of the spectral
1225 matches increase when more MS² fragment peaks are required, there are fewer spectra
1226 that have a larger number of ions, resulting in a diminished number of annotations,
1227 especially for low MW compounds.

1228

1229 **3.4.3 Molecular network visualization in Cytoscape** - Timing 1-4 hours

1230 In addition to in-browser visualization, networks can be visualized using third party tools.
1231 One popular GNPS-derived molecular network visualization tool is Cytoscape³⁷, a
1232 convenient software tool to use for data visualization. The steps outlined below provide
1233 the user with a working knowledge on how to configure a network in Cytoscape. Readers
1234 following the tutorial example can not only reproduce the same properties described in the
1235 the steps below to generate a publishable network but also use this network to specifically
1236 focus on the cluster containing bile acids in order to discover novel compounds.

1237

1238 There are a few options for exporting molecular networks for visualization in Cytoscape.
1239 Once molecular networks generated from GNPS are imported into Cytoscape, a number
1240 of simple commands can be used to make the network generated more informative,
1241 visually appealing, and accessible (**SI Fig. 6**). [Documentation](#) on how to use Cytoscape
1242 (versions after 3.7 release) and a [Cytoscape community forum](#) are available to assist with
1243 troubleshooting and to learn about the latest plugins (also called Cytoscape Apps):
1244 https://cytoscape.org/documentation_users.html, <https://cytoscape.org/community.html>.
1245 An online version of this tutorial is accessible at: [https://ccms-](https://ccms-ucsd.github.io/GNPSDocumentation/cytoscape/)
1246 [ucsd.github.io/GNPSDocumentation/cytoscape/](https://ccms-ucsd.github.io/GNPSDocumentation/cytoscape/).

1247

- 1248 1. To begin using Cytoscape, download the latest version of the software from:
1249 <https://cytoscape.org/> according to their instructions (**SI Fig. 6a**).
- 1250 2. Once Cytoscape has been downloaded, molecular networks can be imported and
1251 visualized using two different strategies. The first option (a) will show a network
1252 with no preset layout, while the second option (b) will show a network with default
1253 layout settings.
 - 1254 a. In order to import data for a network with no layout present (option 1), click on
1255 “Download GraphML for Cytoscape” in the GNPS Job status window (**SI Fig.**
1256 **6b**). This will prompt an immediate download of a compressed folder containing
1257 the .graphML file of interest; after uncompressing this folder using a variety of
1258 programs, Cytoscape can be opened. The import network button (three nodes
1259 connected by edges, **SI Fig. 6c**) in Cytoscape can be selected, permitting
1260 selection of the .graphml file to load the network of interest.
 - 1261 b. The second option for opening a network in Cytoscape is to click on “Direct
1262 Cytoscape Preview/Download” in the GNPS Job Status window (**SI Fig. 6d**).
1263 This will direct the user to a new window where a pre-configured version of the
1264 molecular network will be displayed. In this window, click on “Download
1265 Cytoscape File” to download the file as a Cytoscape session file (.cys file) with
1266 the visualization parameters already defined. Cytoscape can then be opened
1267 by double clicking on the downloaded .cys file and this network will come
1268 preloaded with GNPS default layout.

1269 c. Readers following the tutorial can use either strategy to open the completed
1270 GNPS job run on dataset MSV000083437.

1271

1272

3. Once the molecular network has been loaded into Cytoscape, it can be customized
1273 for viewing. By altering many properties of nodes, edges, and networks such as
1274 colors, sizes, shapes, and labels, the default network can be transformed into a
1275 chemically informative molecular network. Readers following the tutorial example
1276 are guided through this process in steps 3a-3j. In the Control panel window, located
1277 on the left side of the screen, the style and select tabs offer many options.

1278

1279

To alter a node style, click on the “Style” tab at the top of the Control Panel, then
1280 click on the “Node” tab at the bottom of this window (**SI Fig. 6e**).

1281

1282

- a. The node labels can be changed in Cytoscape by selecting the dropdown arrow
1283 next to the “Label” tab. Readers following the tutorial example can label nodes
1284 by selecting “precursor mass” as column and “Passthrough Mapping” for
1285 mapping type (**SI Fig. 6f**).

1286

1287

- b. Node shape can also be changed. Readers following the tutorial example can
1288 click directly on the “Shape” symbol button and select “Ellipse” shape or change
1289 to another desired shape (**SI Fig. 6g**). If using ellipse, the shape can be
1290 converted into a circle by checking the box labeled ‘lock node width and height’
(**SI Fig 6h**).

1291

1292

- c. To change the node color, click on “Fill Color” dropdown. Under this column,
1293 readers following the tutorial example can select the desired value (i.e.
1294 “ATTRIBUTE_host_microbiome”) and use this to discriminate groups (i.e. germ
1295 free vs. specific pathogen free) from one another. Readers can select “Discrete
1296 Mapping” under the “Mapping Type” column, which allows for the selection of
a color to be associated with each group (**SI Fig. 6i**).

1297

1298

- d. Alternatively at the “Fill Color” option, the “Image/Chart 1” tab can be used to
1299 visualize the relative ion distribution from each chosen group in the nodes as a
1300 pie chart. Readers following the tutorial can perform this type of visualization
1301 by clicking on the “Image/Chart 1” button, selecting the “Charts” tab, and
1302 choosing a chart type (the pie chart is chosen in this example). The spectral
1303 count information from groups defined in the metadata file can then be selected
1304 from the “Available columns” to the “Selected columns” (**SI Fig. 6j**) and the user
1305 can edit the chart color scheme using the “Options” tab. In this example, “Germ
1306 free” and “Specific Pathogen free” can be selected and colored pink and blue,
respectively.

1307

1308

- e. To visualize the variation in the occurrence of each ion across samples (e.g
1309 count of 1 if not zero) as a function of the node size, go to Size option, select
1310 “number of spectra” or “sum(precursor intensity)” as *Column* and “Continuous
1311 Mapping” as *Mapping Type*. The opened window allows to modify the node
1312 size in function of the node metadata column chosen. Begin by setting the value
1313 for minimum and maximum node size value with the button *Set Min and Max*,
1314 and then *OK*. Then move the cursor at each extremities. For readers following
the tutorial example, set to the min size at 92 and the max at 362 (**SI Fig. 6k**).

1315

1316

- f. Edge style can also be altered by clicking on the “Edge” tab at the bottom of
the Control Panel (next to the “Node” tab) (**SI Fig. 6l**). Readers following the

- 1317 tutorial example can select this tab to make alterations in edge color and width,
1318 in addition to other settings.
- 1319 g. To change an edge label, readers following the tutorial can click on the “Label”
1320 dropdown arrow then select desired value. For example, mass_difference can
1321 be selected as “Column” in the “Passthrough Mapping mode (SI Fig. 6m).
- 1322 h. Edge width can be altered by clicking on the dropdown arrow next to “Width.”
1323 Under the “select value” tab next to the “Column” tab, the desired value used
1324 for scaling edges (such as cosine_score) can be selected. At this point,
1325 “Continuous Mapping” can be selected under “Mapping Type” (SI Fig. 6n).
1326 Cosine_score can be selected in the column tab and “continuous mapping” can
1327 be chosen under mapping type to easily visualize the approximate cosine score
1328 of all edges.
- 1329 i. The ions from experimental conditions present in the blank sample can be
1330 subtracted from the molecular networks. In the table panel, readers following
1331 the tutorial example can go to the column GNPSGROUP:blank, select every
1332 rows with ion occurrence (>0), then click on the right mouse button and “select
1333 nodes from selected rows” can be choose (SI Fig. 6o). The selected nodes
1334 were automatically highlighted in yellow in the network. Then, do a right click
1335 to choose in the select row “hide selected nodes and edges” (SI Fig. 6p).
1336 However, it is possible to remove the ions from experimental conditions before
1337 generating a molecular network by data processing¹²⁸.
- 1338 j. To separate one or some specific desired network(s), press “ctrl” or “command”
1339 (windows or MacOS, respectively) at the same time selecting the network(s)
1340 with the mouse. Then, click on the bottom as shown in SI Fig. 6q.
1341 Automatically, the sub-network is created. For going back to the main network,
1342 go into the Control Panel by selecting Network, then click on the main network
1343 bottom.
- 1344 4. At this point, readers following the tutorial example have generated a publishable
1345 network in Cytoscape from the output of molecular networking in GNPS. This
1346 network should look like that shown in Fig. 3. Interested readers can look more
1347 closely at the sub-network containing key bile acids in order to practice manual
1348 propagation of annotations throughout a sub-network (Fig. 3). Style options are
1349 described in more detail in the Cytoscape manual:
1350 <http://manual.cytoscape.org/en/stable/Styles.html>.

1351 1352 3.5 How to propagate annotations through manual interpretation of the networks

1353 A molecular network can be very useful in propagating annotations through manual
1354 interpretation of networks in parallel with raw MS² spectra. Manual annotation can be
1355 performed by looking at mass differences (deltas) in the molecular network and assigning
1356 the source of these deltas, i.e. charge retention fragmentations such as retro-Diels Alder
1357 reactions or McLafferty rearrangements and charge migration fragmentations such as
1358 simple inductive cleavages or α - or β -eliminations¹²⁹. The novel bile acids found in the
1359 mouse duodenum provide an example of the utility of manual interpretation of networks
1360 (SI Fig 7b). One can use the mass deltas between unknown nodes and neighboring library
1361 hits to determine new structures. In the above example, three unknown nodes were
1362 determined to be novel bile acids conjugated with phenylalanine, leucine, and tyrosine
1363 based on their mass deltas with respect to glycocholic or glycomuricholic acid. A

1364 description of how manual propagation of annotations can be performed in the context of
1365 the example is given below:

1366

- 1367 1) The Cytoscape's toolbar can be used to search nodes or edge metadata (e.g.,
1368 "shared name"). Readers following the tutorial example can enter "glycocholic acid"
1369 with the quotation marks. The node of interest at m/z 466.316 that matches
1370 glycocholic acid in the GNPS library are automatically selected and highlighted in
1371 yellow in the network (**SI Fig. 6g**).
- 1372 2) Manually propagate annotation based on mass shifts. In **SI Fig. 7a**, glycocholic
1373 acid connects to a node with m/z 556.363. Based on the mass shift of 90.047, the
1374 unknown node can be manually annotated as glycocholic acid conjugated with
1375 phenylalanine. Analogously, nodes with m/z 572.358 and 522.379 could be
1376 manually annotated as glycocholic acid conjugated with tyrosine and leucine
1377 respectively, accounting for mass shifts of 106.042 and 56.063 Da.
- 1378 3) The select function is helpful to find the annotated nodes within the network with a
1379 m/z error from 0 to 10 ppm between precursor ions. This tool is available in Control
1380 Panel at the Select tab, and can be used to create a selection of nodes and/or
1381 edges based on their metadata and/or network topology. Readers following the
1382 tutorial example can click on the "+" button and choose "MZErrorPPM" as column
1383 filter and move the cursor from 0 to 10, then click on Apply (**SI Fig. 7b**). These
1384 nodes are automatically selected and highlighted in yellow in the network.
- 1385 4) Advanced computational tools can also be used for automated annotation
1386 propagation, such as the Network Annotation Propagation (NAP) tool⁸⁵, or manual
1387 annotation can be performed using the results of Dereplicator^{82, 83} and
1388 Mass2Motifs,¹³⁰ which can be accessed through GNPS at
1389 <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp>.

1390

1391 3.6 Capturing information by adding reference spectra from your data

1392 Once an MS² spectra has been fully annotated, it can be added as a reference
1393 spectrum to GNPS. Because the GNPS library database is crowd-sourced, users are
1394 encouraged to submit spectral annotations because knowledge they have is captured
1395 through these annotations of reference spectra and reusable by others. This enables the
1396 creation of reference spectra from MS² spectra in the dataset without needing to purify the
1397 molecule. The assumption is made that the people who collected the data are experts in
1398 their samples and thus are in the best position to curate. Additionally, if the same user or
1399 lab then uploads another related dataset, and it contains the same molecule, it will be
1400 automatically annotated. Users can upload a single reference spectrum by first clicking on
1401 "View All Clusters With IDs" in the job status page, then selecting the cluster desired for
1402 annotation from the "ClusterIdx" column. Once the cluster is selected, the
1403 "AnnotatetoGNPS" button can be selected. This button brings up the workflow for
1404 annotation, where input files, sample parameters, desired annotation, advanced
1405 annotations and library selections can be added and the job can be submitted. Users can
1406 also add a known spectrum to the library from a file uploaded to MassIVE by selecting
1407 "Contribute" under the "Add Your Spectrum" heading on the main page, even if molecular
1408 networking has not been performed on this file. Additionally, if the user wishes to upload
1409 >50 reference spectra to GNPS, a separate batch upload can be performed to streamline
1410 the process as detailed in the online help [documentation at https://ccms-](https://ccms-ucsd.github.io/GNPSDocumentation/batchupload/)
1411 [ucsd.github.io/GNPSDocumentation/batchupload/](https://ccms-ucsd.github.io/GNPSDocumentation/batchupload/). All annotations can be refined at a later

1412 step, and the provenance of each curation is retained within the GNPS-MassIVE
1413 environment. For example one person may annotate that they think it is a lipid, the next
1414 person may update and specify it is a phosphatidylcholine and the next person may refine
1415 this to be 1-oleoyl-2-palmitoyl-phosphatidylcholine and this is all logged in the CCMS
1416 spectral library for each MS² spectrum.

1417

1418 **3.7 Data sharing & reproducibility of molecular networking**

1419 GNPS users are encouraged to share both the raw mass spectrometry data and
1420 associated molecular networking jobs that contributed to peer-reviewed publications by
1421 providing the MassIVE accession number (e.g. MSV000083437) and a hyperlink to the
1422 GNPS job in the methods or experimental details section of the publication. Datasets
1423 uploaded to MassIVE ideally include all raw and peak picked mass spectrometry data and
1424 associated sample information (metadata). GNPS records all data inputs, manipulations
1425 and analyses of the data, providing a historical record of the data and its origins. This data
1426 provenance promotes reproducibility and ultimately quality of the data and its annotations.

1427

1428 **3.7.1. Cloning a job**

1429 Once a job's URL address is shared, any GNPS user can clone the job by following
1430 the provided link and clicking 'clone' on the job status page (**SI Fig. 8**). Cloning a job allows
1431 users to view all parameters and files that were used to create the existing network and
1432 easily rerun the molecular networking job with the same (or adjusted) parameters and files.
1433 Cloning a GNPS job is an extremely useful tool that promotes reproducibility and scientific
1434 rigor. This is a feature many users use to submit multiple molecular networking jobs with
1435 modified parameters. Note that if data were imported from your private user workspace
1436 and not from within MASSIVE, other users will not have access to the mass data and
1437 consequently will not alter the analysis in GNPS. If a job has been run in the previous V1
1438 version of GNPS (i.e. it ran using the 'METABOLOMICS-SNETS' workflow), it can be
1439 cloned and re-run in version 2 (V2) of GNPS by simply clicking 'Clone Job to Latest
1440 Molecular Networking V2 Workflow' on the job status page (**SI Fig 8b**).

1441

1442 **3.7.2. Accessing a dataset**

1443 If a dataset is public, users are able to download all files for reanalysis, including
1444 raw data and the sample information table (metadata). To access a MassIVE dataset of
1445 interest, users should select 'MassIVE Datasets' in the GNPS workspace portal (**Box 1**)
1446 and enter the MassIVE accession number or defining keywords into the search bar. The
1447 user can then click on the MassIVE accession number highlighted in green to link to the
1448 'MassIVE dataset information page', and select the 'FTP Download' link to download files.
1449 Alternatively, this link can be pasted into the quick connect box of an FTP client.

1450 In contrast, private datasets can only be viewed by the user who uploaded the data
1451 and anyone who has a link to the job status page. The user can create a password
1452 protected link. When downloading data from a private dataset, you will be prompted to
1453 enter a password for that MassIVE dataset ID. If using an FTP client, you will need to enter
1454 the MassIVE ID as the username, followed by a password. If the submitter did not specify
1455 a password, then it should be accessible using the password 'a'.

1456

1457 **3.7.3. Subscribing to a dataset and living data**

1458 Public datasets remain alive long after publication: for example, they will be
1459 searched periodically against the ever growing annotated GNPS spectral libraries,

1460 potentially yielding new putative annotations within those datasets. Beyond new
1461 identifications within a dataset, subscribers will receive email notifications of other datasets
1462 that exhibit chemical similarities to the subscribed dataset. This allows for users to be
1463 connected via their research interest to similar datasets. Updates are sent out about once
1464 a month and only when there is new information associated with the dataset. To subscribe
1465 to a dataset, the user should navigate to the 'MassIVE dataset information' page as
1466 described above in section 3.7.2 and click 'Subscribe'. This feature changes the way we
1467 interact with data. Previously, data was periodically reanalyzed by the submission of new
1468 jobs, but in GNPS, data is automatically reanalyzed and updates are sent to the
1469 subscribers. Therefore, data may give rise to useful results a few weeks or even a few
1470 years later after it is uploaded or it may enable the dissemination of all the knowledge of
1471 this dataset to all lab members or collaborators.

1472

1473 **BOX 4: Feature-based molecular networking (FBMN)**

1474 The above described molecular networking analysis represents the type of
1475 molecular networking that is most widely used currently. This workflow connects clustered
1476 MS² spectra as nodes based on spectral similarities and makes use of MS² data only, even
1477 for quantitation. The chromatographic dimension and MS¹ data are not considered in
1478 classical molecular networking. However, in MS-based metabolomics studies, statistical
1479 analysis is done predominantly from MS¹-based peak abundances from extracted ion
1480 chromatograms (XIC). These chromatographic peaks with a specific accurate mass-to-
1481 charge ratio are described as features. In order to bridge this gap between MS¹ abundance
1482 and MS² qualitative information, there is a workflow to link MS¹ intensities derived from
1483 LC-MS features with MS² information from molecular networking^{131, 132}. This workflow is
1484 called feature-based molecular networking ([https://ccms-
1485 ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/)) and can be
1486 performed using open access mass spectrometry processing tools such as MZmine 2¹¹³,
1487 XCMS⁷⁹, MS-DIAL¹³³, or OpenMS¹¹². In this workflow, feature finding is the computational
1488 process of selecting and identifying features in the MS¹ across multiple samples and must
1489 be performed prior to generating a network. These tools allow the export of a feature table
1490 and corresponding MS² scans for each feature, which can be submitted to feature-based
1491 molecular networking through GNPS. Furthermore, the integration in MZmine 2 allows a
1492 direct submission to GNPS even without being a registered GNPS user. However, by
1493 providing the username and password, the new networking job is directly created in the
1494 specified user space.

1495

1496 **4.0 Troubleshooting**

1497 Table 6 below lists some more common scenarios or questions encountered when using
1498 GNPS. We also recommend to check the forum link from the banner in GNPS where users
1499 can post questions to the GNPS community.

1500

1501 **Table 6.** Troubleshooting

This protocol does not address the issues that the user faces.	Check the GNPS forum and post questions.
Job fails with the	Check that your data are in a supported file format; check that the

message 'Empty MS/MS'	submitted files are centroided and have MS ² data, check that filtering criteria are not too aggressive; check that raw files are not included in the file selection.
Job fails with the message 'spectral library search exceeded memory'	This means that the spectral library search step used too much memory and had to be terminated. This is likely caused by changing the set of spectral libraries used in search (such as removing the spectra filtering). This issue can potentially be resolved by increasing the maximum cluster size value to reduce the number of searched spectra. It is not recommended to change the set of libraries included unless you are an advanced user. Please remove all libraries except for the default "speclibs" and rerun.
Network is too large to view in Cytoscape	If a dataset cannot be loaded into Cytoscape, a sub-network of interest can be opened. Alternatively, larger networks can be opened on a computer with more RAM.
I do not know how to include / exclude blanks	This is most easily addressed if the blanks are included in the metadata. Then the user can opt to visualize spectra found in blanks using discrete mapping in Cytoscape or other visualization tool.
Metadata does not sync with Cytoscape	The metadata (sample information) table must be formatted correctly. In particular, check whether the first column is named 'filename', whether all file names match exactly the files uploaded to GNPS and have '.mzXML' extensions (or other compatible file format), and whether each metadata column uses the prefix "ATTRIBUTE_" and there are no trailing spaces in any of the headings.
GNPS job fails due to improper metadata format	The metadata file must be formatted as a tab-separated .txt file.
I cannot see my file(s) after drag and drop upload to GNPS workspace	Check that the targeted folder is highlighted before dragging and dropping file.
My GNPS network is much smaller (fewer nodes) than expected	Check that you selected the mzXML peaklist files from the 'ccms_peak' folder of your MassIVE dataset for the GNPS workflow, not the mzXML files generated directly from the raw data files in the 'raw' folder. The value of the minimum cluster size can be reduced. The minimum cosine score can also be decreased to increase the number of edges in the networks.
Cannot convert Waters .raw files to .mzXML / .mzML from data acquired in the MSE mode of Waters mass	Datasets acquired on a Waters mass spectrometer using the MSE mode can currently only be converted to .mzML using the vendors UNIFI platform. Alternatively, data need to be collected in DDA or MS ² mode, for which data conversion to .mzXML/.mzML is enabled through ProteoWizard.

spectrometers using ProteoWizard	
Molecules that I know are structurally similar do not appear to form a cluster	Check consensus spectra for the molecules of interest. It is possible that low abundance noisy spectra are included which results in poor consensus. For some classes of compounds that do not fragment efficiently, e.g. certain lipids, the MS ² spectra are not informative enough to build meaningful network.

1502

1503

5.0 Anticipated Results

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

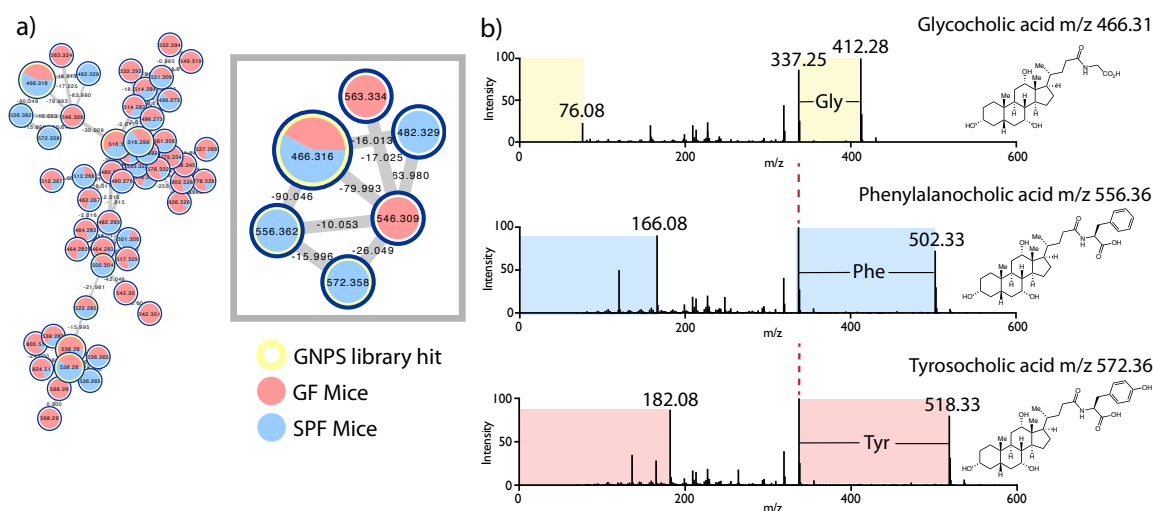
1524

1525

1526

1527

Molecular networking of LC–MS/MS data according to the protocol described herein integrates an associated sample information table (metadata file) with the latest molecular networking workflow, to yield a network (.graphml file) that may be visualized directly in GNPS or imported into Cytoscape. The tutorial example followed throughout the protocol demonstrates how contemporary GNPS molecular networking can be used to discover a new set of conjugated bile acids from the mouse gut microbiome as described in section 3.5.⁶⁵ The network produced from the protocol should contain a molecular family of conjugated bile acids that includes a library hit for glycocholic acid (Figure 5a). This annotation can be propagated to identify new bile acids by converting the mass differences of the edges into structural motifs. For instance, the user can identify the *m/z* 546.309 node as a sulfated cholic acid by using its mass difference of 79.993. This strategy was key in determining the structures for the new phenylalanine (*m/z* 556.362) and tyrosine (*m/z* 572.358) conjugated cholic acids. This example also showcases how manual comparison of the MS/MS spectra that make up the conjugated bile acid molecular family can also be critical for structural annotation. For example, spectra of Gly-, Phe-, and Tyr-conjugated cholic acid all contain fragment ions identical in mass to their respective amino acid conjugates (Figure 5b). Furthermore, the mass difference between the precursor ion and the common peak at *m/z* 337.25, which corresponds to amide bond cleavage, matches the exact mass of the conjugated amino acid. In addition to the conjugated bile acids, the user can also find hits for cholic acid and deoxycholic acid in the network. These compounds are present only in colonized mice, as microbes deconjugate tauro- and glyco-conjugated bile acids in the duodenum.

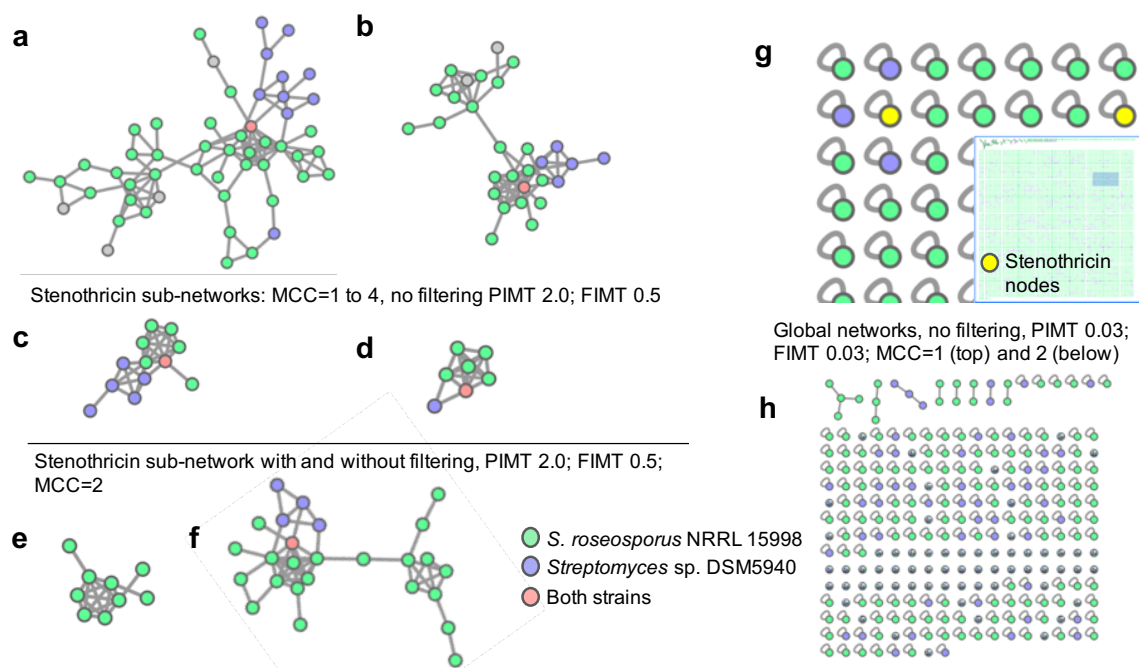


1528
 1529 **Figure 5.** (a) The molecular family of conjugated bile acids from the duodenum of germ
 1530 free (GF) (red) vs. specific pathogen free (SPF) (blue) mice in [MSV000083437](#)
 1531 dataset. As shown in the inset, a library hit for glycocholic acid (m/z 466.316) is present in both GF
 1532 and SPF mice while the new phenylalanine (m/z 556.362) and tyrosine (m/z 572.358)
 1533 conjugated bile acids are seen only in colonized mice. (b) Comparison of MS² spectra for
 1534 Gly-, Phe-, and Tyr-conjugated bile acids.

1535

1536 In addition to the tutorial example, which highlights how molecular networking can be used
 1537 for the discovery of new endogenous metabolites related to human health, two more
 1538 examples are presented from published studies^{11, 50}. One highlights the use of molecular
 1539 networking in natural products discovery and the other integrates metabolomic and
 1540 microbiome data into 3D maps. It is worth noting that the molecular networking workflow
 1541 in GNPS continues to be updated and additional reference library entries are continually
 1542 added by the GNPS community, which may result in some new network annotations since
 1543 the original publication. The current reference libraries used (curated in speclibs,
 1544 December 2018) are listed in the supporting information (SI Table 11). To illustrate the
 1545 utility of GNPS in revealing the extent of suites of related natural products, the discovery
 1546 of new stenothricins-GNPS 1-5 from *Streptomyces* strains reported in Wang et al.¹¹ is
 1547 revisited here. The dataset MSV000083381 comprises MS² data for *n*-butanol and
 1548 methanol extracts from each of *Streptomyces* sp. DSM5940 and *S. roseosporus* NRRL
 1549 15998 cultures grown on solid agar, together with a metadata table that links each of the
 1550 four MS² data files with the originating *Streptomyces* strain. In reproducing the observation
 1551 of a distinct sub-network comprising the MS² data from *Streptomyces* sp. DSM5940
 1552 connected to known *S. roseosporus* stenothricin analogs, we highlight the effect of
 1553 minimum consensus cluster size, PIMT and FIMT settings, and advanced filtering options
 1554 (**Fig. 6**). Importantly, the choice of low resolution settings for PIMT (2.0) and FIMT (0.5) to
 1555 facilitate library searching enables annotation of multiple stenothricin analogues in an
 1556 expansive sub-network, which is otherwise lost with more stringent mass tolerance
 1557 settings of 0.03. Minimum consensus cluster size also has a pronounced effect on the
 1558 range of stenothricin analogues detected. As is common for many natural product
 1559 molecular families, a few major stenothricin analogues are likely accompanied by
 1560 numerous minor stenothricins, for which the MS² spectra generated readily fall below the
 1561 threshold for representation as a node. The distinct clustering of stenothricins from
 1562 *Streptomyces* sp. DSM5940 in **Fig. 6a** is because the parent ion m/z values for these

1563 nodes are 41 Da less than the corresponding values for the known *S. roseosporus*
 1564 stenothricin compounds, consistent with the substitution of serine for lysine in stenothricin-
 1565 GNPS 1-5¹¹.

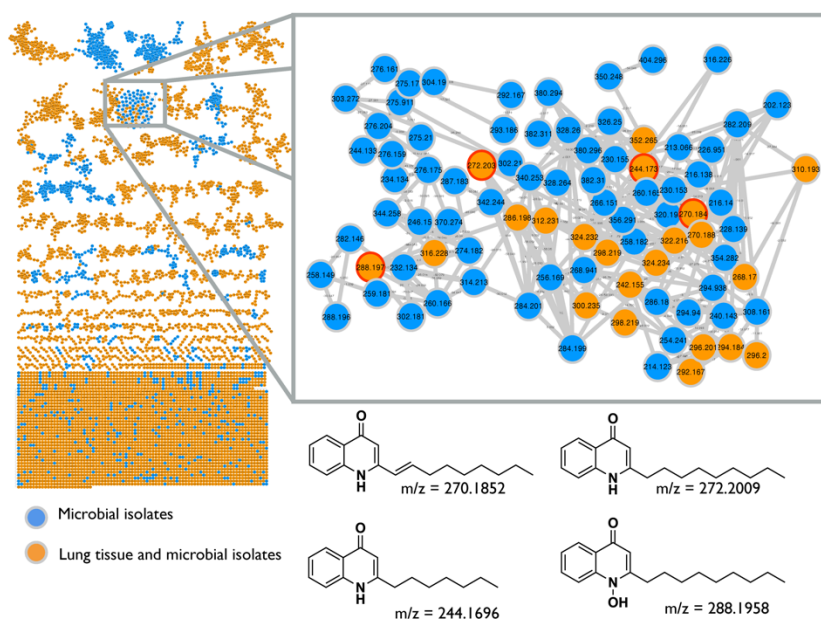


1566
 1567

1568 **Figure 6.** Networking of the stenothricin natural product molecular family
 1569 ([MSV000083381](#)) detected in *Streptomyces* sp. DSM5940 (purple nodes), *S. roseosporus*
 1570 NRRL 15998 (green nodes) or both strains (yellow nodes). Variation in number of nodes
 1571 and spectra with 'Minimum Cluster Size' (MCS) yields sub-networks (a) MCC=1, 52 nodes,
 1572 169 spectra, (b) MCC=2, 29 nodes, 144 spectra, (c) MCC=3, 12 nodes, 89 spectra, (d)
 1573 MCC=4, 7 nodes, 73 spectra (no filtering). Selecting advanced filtering options results in
 1574 (e) 9 nodes, compared to (f) 26 nodes. High resolution settings for PIMT (0.03) and FIMT
 1575 (0.03) reduce stenothricin annotations with (g) MCC = 1 providing two stenothricin nodes
 1576 of 7642 total, and (h) MCC = 2 giving no stenothricin annotations and only 192 nodes.
 1577 Parent ion mass tolerance = PIMT and fragment ion mass tolerance = FIMT.

1578

1579 To further illustrate that molecular networking in GNPS can be used for a diverse range of
 1580 applications, we highlight that molecular networking can be used to visualize quinolones
 1581 produced by *Pseudomonas* isolated from a patient lung⁵⁰. **Fig. 7** reproduces the previous
 1582 analysis ([MSV000083359](#)), where the orange nodes represent quinolones detected in both
 1583 lung tissue extracts and cultured microbial isolates, while cyan nodes represent those only
 1584 detected in cultured microbial isolates.



1585

1586

1587

1588 **Figure 7.** Molecular family (a sub-network) of quinolones detected in lung tissue extracts

1589 (orange nodes) and cultured *Pseudomonas* isolates (cyan nodes), created from MASSIVE

1590 dataset [MSV000083359](https://massive.ucsf.edu/MSV000083359). 2-heptyl-4-quinolone (HHQ), 2-nonyl-4-quinolone (NHQ) and its

1591 unsaturated derivative (NHQ-C9:1 db), and 2-nonyl-4-quinolone-N-oxide (NQNO) were

1592 found in lung tissue, and are highlighted by a red node border.

1593 With a network in hand, there are a number of data analysis tools and experimental

1594 validation steps that may be performed. As discussed in section 3.4.2, to legitimize a library

1595 annotation beyond inspecting mirror plots, the user should verify molecular formula and

1596 identify associated adducts using MS¹ data. Additionally, rationalization based on

1597 biological source is recommended. Ideally, an annotation is authenticated by comparison

1598 with a known standard compound or isolation and full characterization. In the example

1599 followed throughout the protocol, the molecular structures of the new conjugated bile acids

1600 from the mouse duodenum were confirmed by comparison with synthetic standards. For

1601 more complex structures such as those in the stenothricin example¹¹ (Figure 6), the most

1602 abundant analog, stenothricin-GNPS 2, was purified for acquisition. The structure was

1603 assigned from 1D and 2D NMR data, Marfey's analysis¹³⁴, and manual comparison of the

1604 MS² spectra with MS² spectra for previously reported stenothricin D. Genome mining

1605 further supported the conclusion that the -41 Da mass shift observed for stenothricin-

1606 GNPS 1-5 is due to a Lys to Ser substitution. For nodes that are not annotated, the *in*

1607 *silico* Dereplicator may predict peptidic natural products, while NAP (Network Annotation

1608 Propagation) can use annotated nodes to predict related metabolites. Molecular formulas

1609 may be generated using additional tools, one of which is SIRIUS¹⁰⁸. This software uses

1610 MS² features to arrive at the best molecular formula for the precursor MS¹ ion, and works

1611 best for smaller molecules (<600 Da).

1612

1613 In the example of the human lung colonized by *Pseudomonas* bacteria (Figure 7)⁵⁰, the

1614 authors use spatial mapping to visualize annotated molecules on an exploded lung, and

1615 then correlate the distribution of molecules to microbiome maps generated from 16S rRNA

1616 gene amplicon sequencing. This study shows how molecular networking can be used to

1617 elucidate spatial variation in chemical profile and how this can be correlated with microbial
1618 makeup using 3D maps. Statistical analyses of microbiome sequence data were
1619 performed in QIIME2; a number of additional statistical tools as well. Ongoing
1620 developments in GNPS include the integration of some of these statistical analysis tools
1621 into GNPS. Ultimately, it is envisioned that streamlined integration of pre- and post-
1622 networking tools with the GNPS platform will facilitate both creation and mining of
1623 molecular networks.

1624

1625

1626 **Acknowledgements:**

1627 National Research System (SNI) of SENACYT Panama funded CABP, CMH, JL-B, MG;
1628 Gordon and Betty Moore Foundation (PD, NB, KLM), National Institutes of Health
1629 (GM122016-01: KLM), National Science Foundation (DEB1354944: RMT); AKJ
1630 recognizes the American Society for Mass Spectrometry 2018 Postdoctoral Career
1631 Development Award. DP was supported through Deutsche Forschungsgemeinschaft
1632 (DFG) with grant PE 2600/1. R03 CA211211 (PD) on reuse of metabolomics data and
1633 P41 GM103484 (PD, NB) Center for Computational Mass Spectrometry as well as
1634 Instrument support through NIH S10RR029121 (PD).

1635

1636

1637

1638 **References:**

- 1639 1. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies.
1640 *Proc Natl Acad Sci U S A* **109**, E1743-1752 (2012).
- 1641 2. Traxler, M.F. & Kolter, R. A massively spectacular view of the chemical lives of
1642 microbes. *Proc Natl Acad Sci U S A* **109**, 10128-10129 (2012).
- 1643 3. Ramos, A.E.F., Evanno, L., Poupon, E., Champy, P. & Beniddir, M.A. Natural
1644 products targeting strategies involving molecular networking: different manners,
1645 one goal. *Natural Product Reports* **Advance article** (2019).
- 1646 4. Teta, R. et al. A joint molecular networking study of a Smenospongia sponge and
1647 a cyanobacterial bloom revealed new antiproliferative chlorinated polyketides. *Org.*
1648 *Chem. Front.* **6**, 1762-1774 (2019).
- 1649 5. Kalinski, J.J. et al. Molecular Networking Reveals Two Distinct Chemotypes in
1650 Pyrroloiminoquinone-Producing Tsitsikamma favus Sponges. *Mar Drugs* **17**
1651 (2019).
- 1652 6. Raheem, D.J., Tawfike, A.F., Abdelmohsen, U.R., Edrada-Ebel, R. & Fitzsimmons-
1653 Thoss, V. Application of metabolomics and molecular networking in investigating
1654 the chemical profile and antitrypanosomal activity of British bluebells
1655 (*Hyacinthoides non-scripta*). *Sci Rep* **9**, 2547 (2019).
- 1656 7. Trautman, E.P., Healy, A.R., Shine, E.E., Herzon, S.B. & Crawford, J.M. Domain-
1657 Targeted Metabolomics Delineates the Heterocycle Assembly Steps of Colibactin
1658 Biosynthesis. *J Am Chem Soc* **139**, 4195-4201 (2017).
- 1659 8. Vizcaino, M.I., Engel, P., Trautman, E. & Crawford, J.M. Comparative
1660 metabolomics and structural characterizations illuminate colibactin pathway-
1661 dependent small molecules. *J Am Chem Soc* **136**, 9244-9247 (2014).
- 1662 9. Nguyen, D.D. et al. Indexing the *Pseudomonas* specialized metabolome enabled
1663 the discovery of poaeamide B and the bananamides. *Nature Microbiology* **2**, 16197
1664 (2016).
- 1665 10. Frank, A.M. et al. Clustering millions of tandem mass spectra. *J Proteome Res* **7**,
1666 113-122 (2008).

- 1667 11. Wang, M. et al. Sharing and community curation of mass spectrometry data with
1668 Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**, 828-837
1669 (2016).
- 1670 12. Frank, A.M. et al. Spectral archives: extending spectral libraries to analyze both
1671 identified and unidentified spectra. *Nat Methods* **8**, 587-591 (2011).
- 1672 13. De Vijlder, T. et al. A tutorial in small molecule identification via electrospray
1673 ionization-mass spectrometry: The practical art of structural elucidation. *Mass*
1674 *Spectrom Rev* **37**, 607-629 (2018).
- 1675 14. Sumner, L.W. et al. Proposed minimum reporting standards for chemical analysis
1676 Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative
1677 (MSI). *Metabolomics* **3**, 211-221 (2007).
- 1678 15. Su, G., Morris, J.H., Demchak, B. & Bader, G.D. Biological network exploration with
1679 Cytoscape 3. *Curr Protoc Bioinformatics* **47**, 8 13 11-24 (2014).
- 1680 16. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new
1681 features for data integration and network visualization. *Bioinformatics* **27**, 431-432
1682 (2011).
- 1683 17. Sandhu, C. et al. Evaluation of Data-Dependent versus Targeted Shotgun
1684 Proteomic Approaches for Monitoring Transcription Factor Expression in Breast
1685 Cancer. *Journal of Proteome Research* **7**, 1529-1541 (2008).
- 1686 18. Hubert, J., Nuzillard, J.-M. & Renault, J.-H. Dereplication strategies in natural
1687 product research: How many tools and methodologies behind the same concept?
1688 **16**, 55-95 (2017).
- 1689 19. Rochat, B. Proposed Confidence Scale and ID Score in the Identification of Known-
1690 Unknown Compounds Using High Resolution MS Data. *J Am Soc Mass Spectrom*
1691 **28**, 709-723 (2017).
- 1692 20. All natural. *Nature Chemical Biology* **3**, 351 (2007).
- 1693 21. in The "Gold Book", Edn. 2nd. (eds. A.D. McNaught & A. Wilkinson) (Blackwell
1694 Scientific Publications, Oxford; 1997).
- 1695 22. McLafferty, F.W. Tandem mass spectrometry. *Science* **214**, 280-287 (1981).
- 1696 23. Gross, J.H. in *Mass Spectrometry: A Textbook* 415-478 (Springer Berlin
1697 Heidelberg, Berlin, Heidelberg; 2011).
- 1698 24. Artyukhin, A.B. et al. Metabolomic "Dark Matter" Dependent on Peroxisomal β -
1699 Oxidation in *Caenorhabditis elegans*. *Journal of the American Chemical Society*
1700 **140**, 2841-2852 (2018).
- 1701 25. Edwards, E.D., Woolly, E.F., McLellan, R.M. & Keyzers, R.A. Non-detection of
1702 honeybee hive contamination following *Vespula* wasp baiting with protein
1703 containing fipronil. *PLoS One* **13**, e0206385 (2018).
- 1704 26. Hoffmann, T. et al. Correlating chemical diversity with taxonomic distance for
1705 discovery of natural products in myxobacteria. *Nature Communications* **9**, 803
1706 (2018).
- 1707 27. Leipoldt, F. et al. Warhead biosynthesis and the origin of structural diversity in
1708 hydroxamate metalloproteinase inhibitors. *Nat Commun* **8**, 1965 (2017).
- 1709 28. Kang, K.B., Gao, M., Kim, G.J., Choi, H. & Sung, S.H. Rhamnelloides A and B,
1710 omega-Phenylpentaene Fatty Acid Amide Diglycosides from the Fruits of
1711 *Rhamnella franguloides*. *Molecules* **23** (2018).
- 1712 29. Remy, S. et al. Structurally Diverse Diterpenoids from *Sandwithia guyanensis*.
1713 *Journal of Natural Products* **81**, 901-912 (2018).
- 1714 30. Riewe, D., Wiebach, J. & Altmann, T. Structure Annotation and Quantification of
1715 Wheat Seed Oxidized Lipids by High-Resolution LC-MS/MS. *Plant Physiol* **175**,
1716 600-618 (2017).
- 1717 31. Senges, C.H.R. et al. The secreted metabolome of *Streptomyces*
1718 *chartreusis* and implications for bacterial chemistry. *Proceedings of the*
1719 *National Academy of Sciences* **115**, 2490-2495 (2018).

- 1720 32. van der Hooff, J.J.J. et al. Unsupervised Discovery and Comparison of Structural
1721 Families Across Multiple Samples in Untargeted Metabolomics. *Anal Chem* **89**,
1722 7569-7577 (2017).
- 1723 33. Wolff, H. & Bode, H.B. The benzodiazepine-like natural product tilivalline is
1724 produced by the entomopathogenic bacterium *Xenorhabdus eapokensis*. *PLoS*
1725 *One* **13**, e0194297 (2018).
- 1726 34. von Eckardstein, L. et al. Total Synthesis and Biological Assessment of Novel
1727 Albicidins Discovered by Mass Spectrometric Networking. *Chemistry* **23**, 15316-
1728 15321 (2017).
- 1729 35. Vizcaino, M.I. & Crawford, J.M. The colibactin warhead crosslinks DNA. *Nat Chem*
1730 **7**, 411-417 (2015).
- 1731 36. Saleh, H. et al. Deuterium-Labeled Precursor Feeding Reveals a New pABA-
1732 Containing Meroterpenoid from the Mango Pathogen *Xanthomonas citri* pv.
1733 *mangiferaeindicae*. *J Nat Prod* **79**, 1532-1537 (2016).
- 1734 37. Shannon, P. et al. Cytoscape: a software environment for integrated models of
1735 biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).
- 1736 38. Petras, D. et al. Mass Spectrometry-Based Visualization of Molecules Associated
1737 with Human Habitats. *Anal Chem* **88**, 10775-10784 (2016).
- 1738 39. Kapon, C.A. et al. Creating a 3D microbial and chemical snapshot of a human
1739 habitat. *Sci Rep* **8**, 3669 (2018).
- 1740 40. Adams, R.I. et al. Microbes and associated soluble and volatile chemicals on
1741 periodically wet household surfaces. *Microbiome* **5**, 128 (2017).
- 1742 41. Petras, D. et al. High-Resolution Liquid Chromatography Tandem Mass
1743 Spectrometry Enables Large Scale Molecular Characterization of Dissolved
1744 Organic Matter. *Frontiers in Marine Science* **4** (2017).
- 1745 42. Trautman, E.P. & Crawford, J.M. Linking Biosynthetic Gene Clusters to their
1746 Metabolites via Pathway- Targeted Molecular Networking. *Curr Top Med Chem* **16**,
1747 1705-1716 (2016).
- 1748 43. Luzzatto-Knaan, T., Melnik, A.V. & Dorrestein, P.C. Mass Spectrometry Uncovers
1749 the Role of Surfactin as an Interspecies Recruitment Factor. *ACS Chemical Biology*
1750 (2019).
- 1751 44. Machushynets, N.V., Wu, C., Elsayed, S.S., Hankemeier, T. & van Wezel, G.P.
1752 Discovery of novel glycerolated quinazolinones from *Streptomyces* sp. MBT27. *J*
1753 *Ind Microbiol Biotechnol* (2019).
- 1754 45. Yao, L. et al. Discovery of novel xylosides in co-culture of basidiomycetes *Trametes*
1755 *versicolor* and *Ganoderma applanatum* by integrated metabolomics and
1756 bioinformatics. *Sci Rep* **6**, 33237 (2016).
- 1757 46. Tripathi, A. et al. Intermittent Hypoxia and Hypercapnia, a Hallmark of Obstructive
1758 Sleep Apnea, Alters the Gut Microbiome and Metabolome. *mSystems* **3** (2018).
- 1759 47. Smits, S.A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-
1760 gatherers of Tanzania. *Science* **357**, 802-806 (2017).
- 1761 48. McDonald, D. et al. American Gut: an Open Platform for Citizen Science
1762 Microbiome Research. *mSystems* **3**, e00031-00018 (2018).
- 1763 49. Edlund, A. et al. Metabolic Fingerprints from the Human Oral Microbiome Reveal
1764 a Vast Knowledge Gap of Secreted Small Peptidic Molecules. *mSystems* **2**,
1765 e00058-00017 (2017).
- 1766 50. Garg, N. et al. Three-Dimensional Microbiome and Metabolome Cartography of a
1767 Diseased Human Lung. *Cell Host Microbe* **22**, 705-716 e704 (2017).
- 1768 51. McCall, L.I. et al. Mass Spectrometry-Based Chemical Cartography of a Cardiac
1769 Parasitic Infection. *Anal Chem* **89**, 10414-10421 (2017).
- 1770 52. Watrous, J.D. et al. Directed Non-targeted Mass Spectrometry and Chemical
1771 Networking for Discovery of Eicosanoids and Related Oxylipins. *Cell Chemical*
1772 *Biology* (2019).

- 1773 53. Allard, S., Allard, P.M., Morel, I. & Gicquel, T. Application of a molecular networking
1774 approach for clinical and forensic toxicology exemplified in three cases involving 3-
1775 MeO-PCP, doxylamine, and chlormequat. *Drug Test Anal* (2018).
- 1776 54. Ernst, M. et al. Did a plant-herbivore arms race drive chemical diversity in
1777 Euphorbia? *bioRxiv*, 323014 (2018).
- 1778 55. Philippus, A.C. et al. Molecular networking prospection and characterization of
1779 terpenoids and C15-acetogenins in Brazilian seaweed extracts. *RSC Advances* **8**,
1780 29654-29661 (2018).
- 1781 56. Li, F., Janussen, D., Peifer, C., Perez-Victoria, I. & Tasdemir, D. Targeted Isolation
1782 of Tsitsikammamines from the Antarctic Deep-Sea Sponge *Latrunculia biformis* by
1783 Molecular Networking and Anticancer Activity. *Mar Drugs* **16** (2018).
- 1784 57. Hartmann, A.C. et al. Meta-mass shift chemical profiling of metabolomes from coral
1785 reefs. *Proc Natl Acad Sci U S A* **114**, 11685-11690 (2017).
- 1786 58. Tobias, N.J. et al. Natural product diversity associated with the nematode
1787 symbionts *Photobacterium* and *Xenorhabdus*. *Nature Microbiology* **2**, 1676-1685
1788 (2017).
- 1789 59. Nothias, L.F. et al. Bioactivity-Based Molecular Networking for the Discovery of
1790 Drug Leads in Natural Product Bioassay-Guided Fractionation. *J Nat Prod* **81**, 758-
1791 767 (2018).
- 1792 60. Zou, Y. et al. Computationally Assisted Discovery and Assignment of a Highly
1793 Strained and PANC-1 Selective Alkaloid from Alaska's Deep Ocean. *Journal of the*
1794 *American Chemical Society* (2019).
- 1795 61. Parkinson, E.I. et al. Discovery of the Tyrobetaine Natural Products and Their
1796 Biosynthetic Gene Cluster via Metabologenomics. *ACS Chemical Biology* **13**,
1797 1029-1037 (2018).
- 1798 62. Naman, C.B. et al. Integrating Molecular Networking and Biological Assays To
1799 Target the Isolation of a Cytotoxic Cyclic Octapeptide, Samoamide A, from an
1800 American Samoan Marine Cyanobacterium. *Journal of Natural Products* **80**, 625-
1801 633 (2017).
- 1802 63. Bouslimani, A. et al. Lifestyle chemistries from phones for individual profiling. *Proc*
1803 *Natl Acad Sci U S A* **113**, E7645-E7654 (2016).
- 1804 64. Schymanski, E.L. et al. Critical Assessment of Small Molecule Identification 2016:
1805 automated methods. *Journal of Cheminformatics* **9**, 22 (2017).
- 1806 65. Quinn, R.A. et al. Niche partitioning of a pathogenic microbiome driven by chemical
1807 gradients. *Sci Adv* **4**, eaau1908 (2018).
- 1808 66. Aksenov, A.A., da Silva, R., Knight, R., Lopes, N.P. & Dorrestein, P.C. Global
1809 chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry* **1**,
1810 0054 (2017).
- 1811 67. Tsugawa, H. Advances in computational metabolomics and databases deepen the
1812 understanding of metabolisms. *Current Opinion in Biotechnology* **54**, 10-17 (2018).
- 1813 68. Johnson, S.R. & Lange, B.M. Open-Access Metabolomics Databases for Natural
1814 Product Research: Present Capabilities and Future Potential. *Frontiers in*
1815 *Bioengineering and Biotechnology* **3** (2015).
- 1816 69. Haug, K. et al. MetaboLights--an open-access general-purpose repository for
1817 metabolomics studies and associated meta-data. *Nucleic Acids Res* **41**, D781-786
1818 (2013).
- 1819 70. Perez-Riverol, Y. et al. Discovering and linking public omics data sets using the
1820 Omics Discovery Index. *Nat Biotechnol* **35**, 406-409 (2017).
- 1821 71. Stein, S.E. & Scott, D.R. Optimization and testing of mass spectral library search
1822 algorithms for compound identification. *Journal of the American Society for Mass*
1823 *Spectrometry* **5**, 859-866 (1994).
- 1824 72. NIST Standard Reference Database 1A v17.
- 1825 73. Guijas, C. et al. METLIN: A Technology Platform for Identifying Knowns and
1826 Unknowns. *Anal Chem* **90**, 3156-3164 (2018).

- 1827 74. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for
1828 life sciences. *J Mass Spectrom* **45**, 703-714 (2010).
- 1829 75. Stravs, M.A., Schymanski, E.L., Singer, H.P. & Hollender, J. Automatic
1830 recalibration and processing of tandem mass spectra using formula annotation. *J*
1831 *Mass Spectrom* **48**, 89-99 (2013).
- 1832 76. Wang, J., Peake, D.A., Mistrik, R., Huang, Y. & Araujo, G.D.
1833 ([http://www.unifylabservices.eu/content/dam/tfs/ATG/CMD/CMD%20Documents/](http://www.unifylabservices.eu/content/dam/tfs/ATG/CMD/CMD%20Documents/posters/PN-ASMS13-a-platform-to-identify-endogenous-metabolites-using-a-novel-high-performance-orbitrap-and-the-mzcloud-library-E.pdf)
1834 [posters/PN-ASMS13-a-platform-to-identify-endogenous-metabolites-using-a-](http://www.unifylabservices.eu/content/dam/tfs/ATG/CMD/CMD%20Documents/posters/PN-ASMS13-a-platform-to-identify-endogenous-metabolites-using-a-novel-high-performance-orbitrap-and-the-mzcloud-library-E.pdf)
1835 [novel-high-performance-orbitrap-and-the-mzcloud-library-E.pdf](http://www.unifylabservices.eu/content/dam/tfs/ATG/CMD/CMD%20Documents/posters/PN-ASMS13-a-platform-to-identify-endogenous-metabolites-using-a-novel-high-performance-orbitrap-and-the-mzcloud-library-E.pdf); 2013).
- 1836 77. Sheldon, M.T., Mistrik, R. & Croley, T.R. Determination of ion structures in
1837 structurally related compounds using precursor ion fingerprinting. *J Am Soc Mass*
1838 *Spectrom* **20**, 370-376 (2009).
- 1839 78. Sawada, Y. et al. RIKEN tandem mass spectral database (ReSpect) for
1840 phytochemicals: a plant-specific MS/MS-based data resource and database.
1841 *Phytochemistry* **82**, 38-45 (2012).
- 1842 79. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS:
1843 Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak
1844 Alignment, Matching, and Identification. *Analytical Chemistry* **78**, 779-787 (2006).
- 1845 80. Tautenhahn, R., Patti, G.J., Rinehart, D. & Siuzdak, G. XCMS Online: a web-based
1846 platform to process untargeted metabolomic data. *Anal Chem* **84**, 5035-5039
1847 (2012).
- 1848 81. Wanichthanarak, K., Fan, S., Grapov, D., Barupal, D.K. & Fiehn, O. Metabox: A
1849 Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration.
1850 *PLOS ONE* **12**, e0171046 (2017).
- 1851 82. Mohimani, H. et al. Dereplication of microbial metabolites through database search
1852 of mass spectra. *Nature Communications* **9**, 4035 (2018).
- 1853 83. Mohimani, H. et al. Dereplication of peptidic natural products through database
1854 search of mass spectra. *Nat Chem Biol* **13**, 30-37 (2017).
- 1855 84. Gurevich, A. et al. Increased diversity of peptidic natural products revealed by
1856 modification-tolerant database search of mass spectra. *Nat Microbiol* **3**, 319-327
1857 (2018).
- 1858 85. da Silva, R.R. et al. Propagating annotations of molecular networks using in silico
1859 fragmentation. *PLoS computational biology* **14**, e1006089 (2018).
- 1860 86. Mohimani, H. et al. Automated genome mining of ribosomal peptide natural
1861 products. *ACS Chem Biol* **9**, 1545-1551 (2014).
- 1862 87. Olivon, F. et al. MetGem Software for the Generation of Molecular Networks Based
1863 on the t-SNE Algorithm. *Anal Chem* (2018).
- 1864 88. Olivon, F., Roussi, F., Litaudon, M. & Touboul, D. Optimized experimental workflow
1865 for tandem mass spectrometry molecular networking in metabolomics. *Anal*
1866 *Bioanal Chem* **409**, 5767-5778 (2017).
- 1867 89. Wehrens, R. et al. Improved batch correction in untargeted MS-based
1868 metabolomics. *Metabolomics* **12**, 88 (2016).
- 1869 90. Koal, T. & Deigner, H.P. Challenges in mass spectrometry based targeted
1870 metabolomics. *Curr Mol Med* **10**, 216-226 (2010).
- 1871 91. Bylda, C., Thiele, R., Kobold, U. & Volmer, D.A. Recent advances in sample
1872 preparation techniques to overcome difficulties encountered during quantitative
1873 analysis of small molecules from biofluids using LC-MS/MS. *Analyst* **139**, 2265-
1874 2276 (2014).
- 1875 92. Vuckovic, D. Current trends and challenges in sample preparation for global
1876 metabolomics using liquid chromatography-mass spectrometry. *Anal Bioanal*
1877 *Chem* **403**, 1523-1548 (2012).
- 1878 93. Dunn, W.B. et al. Procedures for large-scale metabolic profiling of serum and
1879 plasma using gas chromatography and liquid chromatography coupled to mass
1880 spectrometry. *Nature Protocols* **6**, 1060 (2011).

- 1881 94. Taylor, P.J. Matrix effects: the Achilles heel of quantitative high-performance liquid
1882 chromatography-electrospray-tandem mass spectrometry. *Clin Biochem* **38**, 328-
1883 334 (2005).
- 1884 95. Annesley, T.M. Ion suppression in mass spectrometry. *Clin Chem* **49**, 1041-1044
1885 (2003).
- 1886 96. Crüsemann, M. et al. Prioritizing Natural Product Diversity in a Collection of 146
1887 Bacterial Strains Based on Growth and Extraction Protocols. *Journal of Natural*
1888 *Products* **80**, 588-597 (2017).
- 1889 97. Wandro, S., Carmody, L., Gallagher, T., LiPuma, J.J. & Whiteson, K. Making It
1890 Last: Storage Time and Temperature Have Differential Impacts on Metabolite
1891 Profiles of Airway Samples from Cystic Fibrosis Patients. *mSystems* **2** (2017).
- 1892 98. Zhao, J., Evans, C.R., Carmody, L.A. & LiPuma, J.J. Impact of storage conditions
1893 on metabolite profiles of sputum samples from persons with cystic fibrosis. *J Cyst*
1894 *Fibros* **14**, 468-473 (2015).
- 1895 99. Hirayama, A. et al. Effects of processing and storage conditions on charged
1896 metabolomic profiles in blood. *ELECTROPHORESIS* **36**, 2148-2155 (2015).
- 1897 100. Mushtaq, M.Y., Choi, Y.H., Verpoorte, R. & Wilson, E.G. Extraction for
1898 metabolomics: access to the metabolome. *Phytochem Anal* **25**, 291-306 (2014).
- 1899 101. Bazsó, F.L. et al. Quantitative Comparison of Tandem Mass Spectra Obtained on
1900 Various Instruments. *J Am Soc Mass Spectrom* **27**, 1357-1365 (2016).
- 1901 102. Bowen, B.P. & Northen, T.R. Dealing with the unknown: metabolomics and
1902 metabolite atlases. *J Am Soc Mass Spectrom* **21**, 1471-1476 (2010).
- 1903 103. da Silva, R.R., Dorrestein, P.C. & Quinn, R.A. Illuminating the dark matter in
1904 metabolomics. *Proc Natl Acad Sci U S A* **112**, 12549-12550 (2015).
- 1905 104. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for
1906 Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8**
1907 (2018).
- 1908 105. Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J. & Neumann, S. MetFrag
1909 relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*
1910 **8**, 3 (2016).
- 1911 106. Gerlich, M. & Neumann, S. MetFusion: integration of compound identification
1912 strategies. *J Mass Spectrom* **48**, 291-298 (2013).
- 1913 107. Böcker, S., Letzel, M.C., Liptak, Z. & Pevukhin, A. SIRIUS: decomposing isotope
1914 patterns for metabolite identification. *Bioinformatics* **25**, 218-224 (2009).
- 1915 108. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into
1916 metabolite structure information. *Nat Methods* **16**, 299-302 (2019).
- 1917 109. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Bocker, S. Searching molecular
1918 structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad*
1919 *Sci U S A* **112**, 12580-12585 (2015).
- 1920 110. Tsugawa, H. et al. Hydrogen Rearrangement Rules: Computational MS/MS
1921 Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem*
1922 **88**, 7946-7958 (2016).
- 1923 111. Protsyuk, I. et al. 3D molecular cartography using LC-MS facilitated by Optimus
1924 and 'ili software. *Nat Protoc* **13**, 134-154 (2018).
- 1925 112. Röst, H.L. et al. OpenMS: a flexible open-source software platform for mass
1926 spectrometry data analysis. *Nat Methods* **13**, 741-748 (2016).
- 1927 113. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular
1928 framework for processing, visualizing, and analyzing mass spectrometry-based
1929 molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
- 1930 114. Deutsch, E.W. et al. Proteomics Standards Initiative: Fifteen Years of Progress and
1931 Future Work. *Journal of Proteome Research* **16**, 4288-4298 (2017).
- 1932 115. Brooksbank, C., Cameron, G. & Thornton, J. The European Bioinformatics
1933 Institute's data resources. *Nucleic Acids Res* **38**, D17-25 (2010).
- 1934 116. Jones, A.R. et al. The mzIdentML data standard for mass spectrometry-based
1935 proteomics results. *Mol Cell Proteomics* **11**, M111 014381 (2012).

- 1936 117. Griss, J. et al. The mzTab data exchange format: communicating mass-
1937 spectrometry-based proteomics and metabolomics experimental results to a wider
1938 audience. *Mol Cell Proteomics* **13**, 2765-2775 (2014).
- 1939 118. Hoffmann, N. et al. mzTab-M: A Data Standard for Sharing Quantitative Results in
1940 Mass Spectrometry Metabolomics. *Analytical Chemistry* (2019).
- 1941 119. Wang, M. et al. MASST: A Web-based Basic Mass Spectrometry Search Tool for
1942 Molecules to Search Public Data. *bioRxiv*, 591016 (2019).
- 1943 120. Scheubert, K. et al. Significance estimation for large scale metabolomics
1944 annotations by spectral matching. *Nat Commun* **8**, 1494 (2017).
- 1945 121. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I
1946 learned to stop worrying and love the ome-ome. *GigaScience* **1**, 7 (2012).
- 1947 122. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPERor: a tool for
1948 visualizing high-throughput microbial community data. *GigaScience* **2**, 16 (2013).
- 1949 123. Bolyen, E. et al. QIIME 2: Reproducible, interactive, scalable, and extensible
1950 microbiome data science. *PeerJ Preprints* (2018).
- 1951 124. McLafferty, F.W. & Tureček, F.e. Interpretation of mass spectra, Edn. 4th.
1952 (University Science Books, Mill Valley, Calif.; 1993).
- 1953 125. Viant, M.R., Kurland, I.J., Jones, M.R. & Dunn, W.B. How close are we to complete
1954 annotation of metabolomes? *Curr Opin Chem Biol* **36**, 64-69 (2017).
- 1955 126. Shahaf, N. et al. The WEIZMASS spectral library for high-confidence metabolite
1956 identification. *Nature Communications* **7**, 12423 (2016).
- 1957 127. Schymanski, E.L. et al. Identifying small molecules via high resolution mass
1958 spectrometry: communicating confidence. *Environ Sci Technol* **48**, 2097-2098
1959 (2014).
- 1960 128. Cleary, J.L., Luu, G.T., Pierce, E.C., Dutton, R.J. & Sanchez, L.M. BLANKA: an
1961 Algorithm for Blank Subtraction in Mass Spectrometry of Complex Biological
1962 Samples. *Journal of The American Society for Mass Spectrometry* (2019).
- 1963 129. Demarque, D.P., Crotti, A.E.M., Vessecchi, R., Lopes, J.L.C. & Lopes, N.P.
1964 Fragmentation reactions using electrospray ionization mass spectrometry: an
1965 important tool for the structural elucidation and characterization of synthetic and
1966 natural products. *Natural Product Reports* **33**, 432-455 (2016).
- 1967 130. van der Hoof, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V. & Rogers, S. Topic
1968 modeling for untargeted substructure exploration in metabolomics. *Proceedings of
1969 the National Academy of Sciences* **113**, 13738-13743 (2016).
- 1970 131. Olivon, F., Grelier, G., Roussi, F., Litaudon, M. & Touboul, D. MZmine 2 Data-
1971 Preprocessing To Enhance Molecular Networking Reliability. *Analytical Chemistry*
1972 **89**, 7836-7840 (2017).
- 1973 132. Winnikoff, J.R., Glukhov, E., Watrous, J., Dorrestein, P.C. & Gerwick, W.H.
1974 Quantitative molecular networking to profile marine cyanobacterial metabolomes.
1975 *J Antibiot (Tokyo)* **67**, 105-112 (2014).
- 1976 133. Tsugawa, H. et al. MS-DIAL: data-independent MS/MS deconvolution for
1977 comprehensive metabolome analysis. *Nat Methods* **12**, 523-526 (2015).
- 1978 134. Marfey, P. Determination of D-Amino Acids .2. Use of a Bifunctional Reagent, 1,5-
1979 Difluoro-2,4-Dinitrobenzene. *Carlsberg Res Commun* **49**, 591-596 (1984).
- 1980