

Reproducible Molecular Networking Of Untargeted Mass Spectrometry Data Using GNPS.

Allegra T. Aron, Emily Gentry, Kerry L. McPhail, Louis Felix Nothias, Mélissa Nothias-Esposito, Amina Bouslimani, Daniel Petras, Julia M. Gauglitz, Nicole Sikora, Fernando Vargas, Justin J. J. van der Hooft, Madeleine Ernst, Kyo Bin Kang, Christine M. Aceves, Andrés Mauricio Caraballo-Rodríguez, Irina Koester, Kelly C. Weldon, Samuel Bertrand, Catherine Roullier, Kunyang Sun, Richard M. Tehan, Cristopher A. Boya, Christian Martin H., Marcelino Gutiérrez, Aldo Moreno Ulloa, Javier Andres Tejeda Mora, Randy Mojica-Flores, Johant Lakey-Beitia, Victor Vásquez-Chaves, Angela I. Calderón, Nicole Tayler, Robert A. Keyzers, Fidele Tugizimana, Nombuso Ndlovu, Alexander A. Aksenov, Alan K. Jarmusch, Robin Schmid, Andrew W. Truman, Nuno Bandeira, Mingxun Wang, **Pieter Dorrestein**

Submitted date: 07/08/2019 • Posted date: 08/08/2019

Licence: CC BY-NC-ND 4.0

Citation information: Aron, Allegra T.; Gentry, Emily; McPhail, Kerry L.; Nothias, Louis Felix; Nothias-Esposito, Mélissa; Bouslimani, Amina; et al. (2019): Reproducible Molecular Networking Of Untargeted Mass Spectrometry Data Using GNPS.. ChemRxiv. Preprint.

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

GNPSProtocols_Paper_08062019_Submission.pdf (4.24 MiB)

[view on ChemRxiv](#) • [download file](#)

SI_Document_20190806_submission.pdf (16.11 MiB)

[view on ChemRxiv](#) • [download file](#)

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

Authors: Allegra T. Aron¹ #, Emily C. Gentry¹ #, Kerry L. McPhail² #, Louis Felix Nothias¹, Mélissa Nothias-Esposito¹, Amina Bouslimani¹, Daniel Petras^{1,5}, Julia M. Gauglitz¹, Nicole Sikora¹, Fernando Vargas¹, Justin J. J. van der Hooft³, Madeleine Ernst¹, Kyo Bin Kang⁴, Christine M. Aceves¹, Andrés Mauricio Caraballo-Rodríguez¹, Irina Koester^{1,5}, Kelly C. Weldon¹, Samuel Bertrand^{6,7}, Catherine Roullier⁶, Kunyang Sun¹, Richard M. Tehan², Christopher A. Boya^{8,9}, Christian Martin H.⁸, Marcelino Gutiérrez⁸, Aldo Moreno Ulloa¹⁰, Javier Andres Tejeda Mora¹⁰, Randy Mojica-Flores^{8,11}, Johant Lakey-Beitia⁸, Victor Vásquez-Chaves¹², Yilue Zhang¹³, Angela I. Calderon¹³, Nicole Tayler^{8,9}, Robert A. Keyzers¹⁴, Fidele Tugizimana¹⁵, Nombuso Ndlovu¹⁵, Alexander A. Aksenov¹, Alan Jarmusch¹, Robin Schmid¹⁶, Andrew W. Truman¹⁷, Nuno Bandeira^{18*}, Mingxun Wang^{1*}, Pieter C Dorrestein^{1, 19-21*}

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

Author contributions: Design and oversight of the project: P.C.D., M.W., N.B. Instrument 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., 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 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

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

Authors names and emails:

Allegra T. Aron (alargon@ucsd.edu)
Emily C. Gentry (emgentry@ucsd.edu)
Kerry L. McPhail (kerry.mcphail@oregonstate.edu)
Louis-Felix Nothias (lnothiasscaglia@ucsd.edu)
Julia M. Gauglitz (jgauglitz@ucsd.edu)
Christine M. Aceves (caceves@ucsd.edu)
Fernando Vargas (fernando.vargas0341@gmail.com)
Amina Bouslimani (abouslimani@ucsd.edu)
Justin J. J. van der Hooft (justin.vanderhooft@wur.nl)
Kyo Bin Kang (kbkang@sookmyung.ac.kr)
Andrés Mauricio Caraballo-Rodríguez (acaraballorodriguez@ucsd.edu)
Irina Koester (ikoester@ucsd.edu)
Kelly C. Weldon (kcweldon@ucsd.edu)
Daniel Petras (dpetras@ucsd.edu)
Samuel Bertrand (Samuel.Bertrand@univ-nantes.fr)
Catherine Roullier (Catherine.Roullier@univ-nantes.fr)
Madeleine Ernst (mernst@ucsd.edu)
Kunyang Sun (ksun@ucsd.edu)
Richard M. Tehan (tehanr@oregonstate.edu)
Cristopher A. Boya P. (c.boyas@indicasat.org.pa)
Christian Martin H. (cmartin@indicasat.org.pa)
Marcelino Gutiérrez (mgutierrez@indicasat.org.pa)
Aldo Moreno Ulloa (amoreno@cicese.mx)
Javier Andres Tejeda Mora (andres.android@gmail.com)
Randy Mojica-Flores (wendel2506@gmail.com)
Johant Lakey-Beitia (jlakey@indicasat.prg.pa)
Victor Vásquez-Chaves (vvasquezch@gmail.com)
Angela I. Calderon (aic0001@auburn.edu)
Nicole Tayler (ntayler@indicasat.org.pa)
Robert A. Keyzers (robert.Keyzers@vuw.ac.nz)
Fidele Tugizimana (ftugizimana@uj.ac.za)
Nombuso Ndlovu (nndlovu@uj.ac.za)
Nicole Sikora (nsikora@ucsd.edu)
Alexander Aksenov (aaaksenov@ucsd.edu)
Alan Jarmusch (ajarmusch@ucsd.edu)
Robin Schmid (robinschmid@uni-muenster.de)
Andrew W. Truman (Andrew.Truman@jic.ac.uk)
Nuno Bandeira (bandeira@ucsd.edu)
Mingxun Wang (miw023@ucsd.edu)
Pieter C Dorrestein (pdorrestein@ucsd.edu)

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

1.0 Introduction: Molecular networking for the analysis of tandem mass spectra of small molecules was introduced in 2012¹. Upon its introduction, molecular networking was compared to sequencing of environmental DNA to study the microbial communities present in diverse ecosystems². For the first time we were able to get a map of the chemical diversity that is observed in an untargeted mass spectrometry experiment. In addition to providing unprecedented systems-level views of the chemical space in various environments, molecular networking has aided structure elucidation of many compounds³⁻⁹.

The foundation of molecular networking is pairwise spectral alignment using a modified cosine spectral similarity algorithm originally intended to discover modified forms of peptides and proteins¹⁰. In a modified spectral similarity search, not only are fragmentation spectra (MS²) from ions at identical *m/z* compared, but also MS² spectra that are offset by the same *m/z* difference as the precursor ion. By eliminating the amino acid filtering from the original spectral alignment algorithms, it became possible to extend spectral similarity to any set of MS² spectra, including those from small molecules and natural products. When a pairwise spectral similarity search/alignment is performed, each MS² spectrum in a given dataset is compared against every other, and a network of MS² spectral relations is obtained, from which molecular networks are created (**Fig. 1**). Molecular networking build on the fundamental observation that two structurally related molecules share fragment ion patterns when subjected to MS² fragmentation methods such as collision induced dissociation (CID). In order to make the molecular networking algorithm accessible to the scientific community, its script was converted to a web-based platform backed by a supercomputer. This enabled the creation of a community infrastructure supporting both a database and knowledge-base around the needs of the community. The result was the Global Natural Products Social (GNPS) Molecular Networking community effort that started in 2014 and was published in 2016. The user base has expanded to 49 of 50 states in the United States and worldwide to over 150 countries¹¹. GNPS is currently widely used by scientists working in industry, academia and government in the fields of biomedical research, environmental science, ecology,

forensics, microbiology, chemistry, and others. This crowdsourced, community-driven analysis infrastructure not only facilitates data and knowledge storage but also enables knowledge capture, sharing, dissemination and data driven social networking while promoting reproducible data analysis. Moreover, GNPS can be accessed on a computer or on any mobile device connected to the internet making any public data set readily accessible for analysis. While there are many analysis tools available within the GNPS infrastructure, molecular networking is the most frequently used tool. Other tools available on GNPS such as network annotation propagation (NAP) briefly discussed in section 3.5.

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

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</i> (parent ion)	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</i> (fragment ion)	An ion originating from a gas-phase reaction of the precursor ion ¹³ .
<i>sample information</i> (metadata)	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} .

165

166 All mass spectrometry data used in GNPS, both in the private user workspace or data
167 that are made public, is stored in MassIVE - an interactive virtual environment developed
168 to facilitate and encourage the exchange of mass spectrometry data. MassIVE accepts
169 data files (organized as datasets) and facilitates the sharing of datasets with a unique
170 identifier; one can use this unique identifier as an accession number for publications. In
171 addition, public datasets that the user publishes can, by choice of the depositor, have an
172 associated DOI. Currently, MassIVE is an approved repository for the Journal of
173 Proteome Research (<https://pubs.acs.org/journal/jprobs>) and Nature Partner Journals
174 (<https://www.nature.com/sdata/policies/repositories#chem>) and is widely used as a
175 repository for other journals²⁴⁻³³. GNPS-MassIVE has more than a thousand public

metabolomics datasets. The GNPS knowledge base includes 221,083 reference MS² spectra, provided by the GNPS community, spectral libraries generated for GNPS (GNPS-collections) and third party libraries¹¹. Examples are LDB Lichen Database, MIADB Spectral Library, Sumner Spectral Library, CASMI Spectral Library, and Massbank, a large MS data library that is directly synced with GNPS. There are also tags and sample information (metadata) entries provided by the community in the GNPS knowledge base. Furthermore, all public data is periodically searched against the NIST 2017 spectral library and high confidence spectral matches are annotated. GNPS-MassIVE now performs more than 6,000 analysis jobs a month and more than 200,000 page views (excluding developers), with the predominant analysis being molecular networking. As a result, GNPS based analysis has been used for the discovery of hundreds of new molecules in the last few years, ranging from immune regulators to antimicrobials, including antiviral agents and protease inhibitors^{9, 34-36}. Here we provide a detailed protocol on how to generate a publishable and reproducible molecular network from a mass spectrometry dataset. This protocol will take the reader through the following steps: how to upload data, how to make the data public, how to subscribe to public data for living data updates, and how to reproducibly create publishable molecular networks using standardized sample information (metadata) through the GNPS infrastructure (**Fig. 1**).

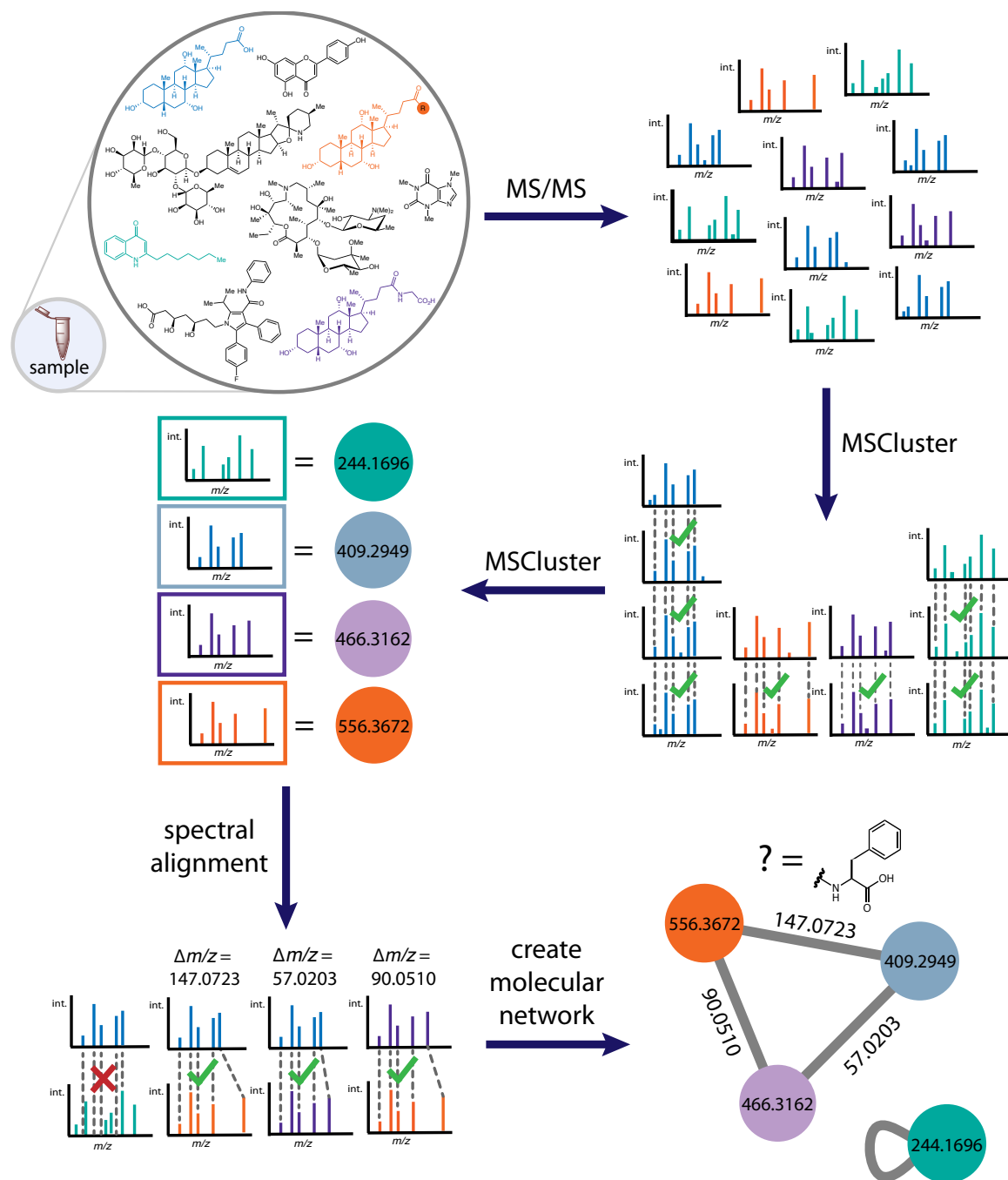
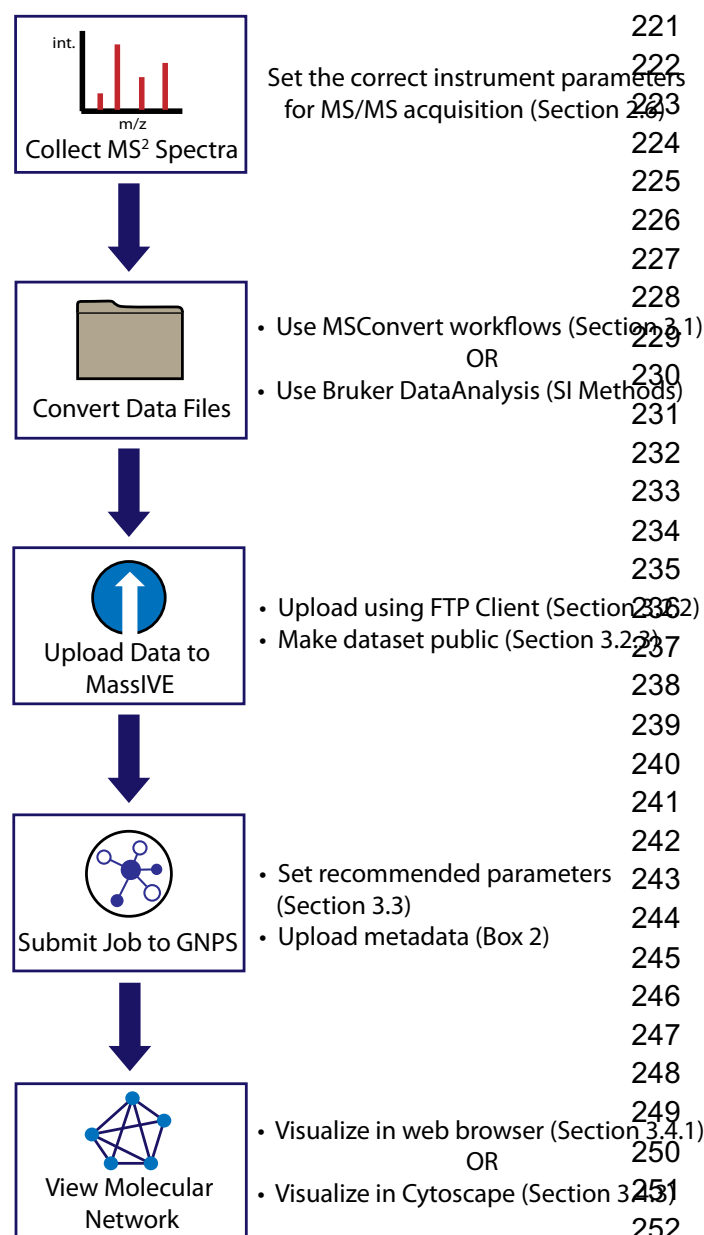


Figure 1. Schematic representation of the process for creating a molecular network from tandem mass spectra acquired for metabolites in complex sample mixtures. The colors are used to track how we go from molecules in a sample to nodes in the molecular network. We start by obtaining MS² spectra of all 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.

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



1.1 Overview of the Method

This protocol aims to provide researchers with a basic workflow for reliably and reproducibly creating molecular networks from mass spectrometry data using GNPS. The workflows for GNPS molecular networking described in this protocol can be performed in any laboratory with access to a tandem mass spectrometer, usually connected to an HPLC system. The overall procedure consists of five main steps (**Fig. 2**): 1) collecting MS² spectra, 2) converting instrument-specific raw data files to an open format, 3) uploading data to the MassIVE public repository, 4) submitting a job to GNPS, 5) visualizing the resulting molecular network. Additional topics include how to navigate 'My User' and 'Jobs', including how to share the links of a job, how to clone a job, (section 6) and how to subscribe to datasets in the public domain to obtain living data updates (section 7).

Figure 2. Flowchart of the protocol, delineating the workflow from tandem mass spectrometry

data acquisition, conversion, upload and networking to visualization. Readers following the tutorial example can follow these steps to generate a publishable network.

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

1.2 Applications of the method

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

1.3 Alternative methods to this protocol

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

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

METLIN⁷³, MassBank of Japan (<http://massbank.jp>)⁷⁴, EU (<https://massbank.eu/MassBank/>)⁷⁵ and North America (<http://mona.fiehnlab.ucdavis.edu/>), mzCloud^{76, 77}, and ReSpect⁷⁸, which potentially provides users with access to around 2.4 million MS² reference spectra, when GC-MS and LC-MS reference spectra are both considered⁶⁶. Many of these reference databases have an integrated spectral matching tool for compound identification, including mzCloud, METLIN/XCMS Online^{79, 80}, Metabox⁸¹, MassBank). The goal of GNPS is not only to provide a spectral matching tool, but also to serve as a data storage and knowledge capture and dissemination platform, and to provide access to a host of other analysis tools not covered in detail here, such as *in silico*-based dereplication⁸²⁻⁸⁴, network annotation propagation⁸⁵, genome mining tools⁸⁶, and MASST searches.

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

1.4 Expertise needed to implement the protocol

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

1.5 Experimental design

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

1.5.1 Reproducibility, blanks, and controls

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

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

1.5.2 Molecular networking parameters

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

1.6 Limitations and challenges

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

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

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

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

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

2.0 Materials

2.1 REAGENTS

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

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

2.2 EQUIPMENT

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

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

2.3 SOFTWARE

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

2.4 EXAMPLE DATASETS

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

- [MSV000083437](#) (Germ Free and Specific Pathogen Free Mice, unpublished)
- [MSV000083359](#) (3D Cartography of Diseased Human Lung⁵⁰)
- [MSV000083381](#) (Stenothricin-GNPS analogues¹¹)

2.5 REAGENT SETUP

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

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

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

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

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

2.6 EQUIPMENT SETUP

Mass spectrometry

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

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

3.0 Procedure

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

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

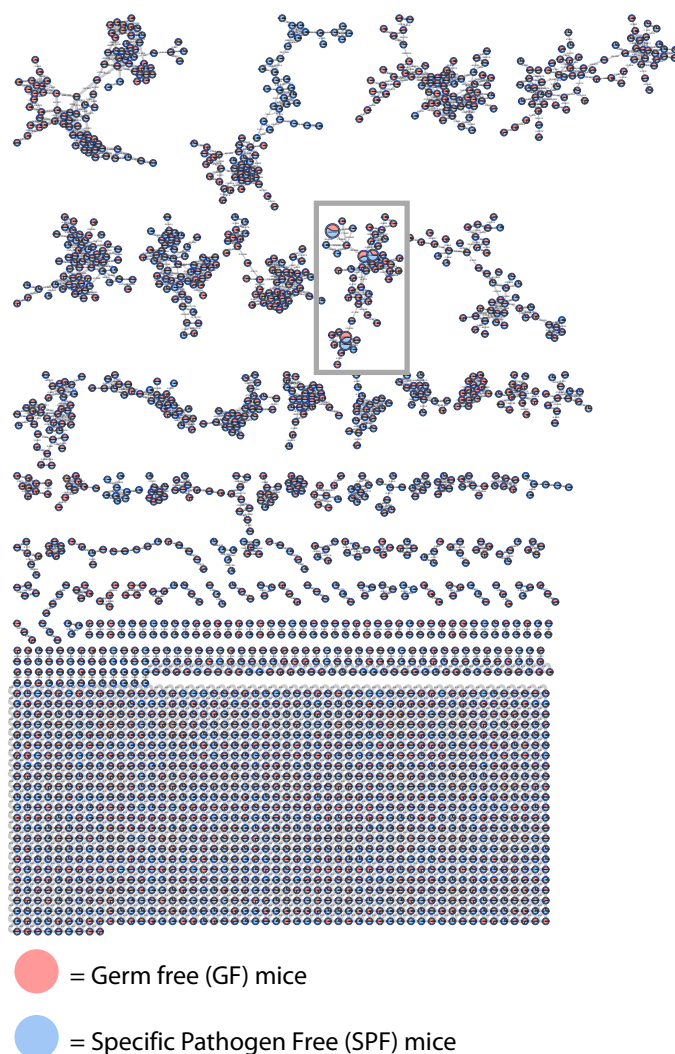


Figure 3. The readers will recreate the mouse duodenum global molecular network depicted above, created from MassIVE dataset [MSV000083437](https://massive.ucsd.edu/MSV000083437) and visualized in Cytoscape. Pie charts represent relative summed precursor ion intensities per MS² spectra detected within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

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

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

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

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

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

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

- a. In the Start Menu, the ProteoWizard folder can be selected and MSConvert can be opened.
- b. To select file(s) for conversion, click Browse; then click ‘Add’ to add file(s) to the workflow and select a directory for the output.
- c. To convert the vendor file format to an .mzXML file, select .mzXML under Options; 32-bit should be selected for binary encoding precision and Use zlib compression should be unchecked.
- d. Choose Peak Picking under the Filters heading and under Algorithm check Vendor, then write in MS-Levels 1-2 and finally add the filter by clicking Add. **! CRITICAL STEP** Move the peakPicking filter to the top of filter list. The peakPicking filter must be the first filter in the list or the output file will not be centroided.
- e. Click Start then check the folder for the .mzXML files in the Output Directory. These files can be opened in SeeMS (Installed with MSConvert), OpenMS TOPPView (<https://github.com/OpenMS/OpenMS/releases>)¹¹² or MZmine2 (<https://github.com/mzmine/mzmine2/releases>)¹¹³ to verify that the conversion worked properly.

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

3.2 Data submission to GNPS / MassIVE

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

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

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

- 1) Open up a web browser. GNPS is designed to work with Firefox or Google Chrome but also works in Microsoft Edge, Safari, and Opera.
- 2) Navigate to the GNPS home page by using this link <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash2.jsp>
- 3) Towards the top center of the page, above the large GNPS logo, click on "Register New Account" (right hand grey box).
- 4) On the new page that loads, enter a username, name (optional), organization (optional), email, and password (twice for confirmation) in the spaces provided.
- 4) Click submit.
- 5) Sign-in to your new GNPS account <http://massive.ucsd.edu/ProteoSAFe/> and check that your GNPS credentials work for logging in to MassIVE.

Box 1: Navigating the User Workspace Portal



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

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

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>

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

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

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

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

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

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

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

2) In the 'Workflow Selection' section:

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

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

3) In the 'Dataset Metadata' section:

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

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

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

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

a) Recommended for all submissions

- i) Raw Spectrum Files – Raw mass spectrum files in a non-standard or instrument-specific format, such as AB Sciex .wiff files, Agilent .yep files, Shimadzu .lcd files, Bruker .d files Thermo Scientific .raw files, Waters .raw files.
- ii) Peak List Files – Processed mass spectrum files in a standardized format. The following formats are recognized by MassIVE as valid for this category: .mzXML, .mzML, and .mgf. This is the file from which GNPS analysis is enabled.

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

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

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

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

d) Optional

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

provide your own license, then you can uncheck this box and then assign your own file to the 'License Files' category.

ii) Spectral Libraries – Any custom spectral library files that were searched against in the analysis of this dataset, or that were generated using the spectrum files provided in this dataset, if applicable.

iii) Methods and Protocols – Any open-format files containing explanations or discussions of the experimental procedures used to obtain or analyze this dataset.

e) *Optional, mostly relevant to peptidomics and proteomics projects*

i) Quantification Results – Any data and metadata generated by the analysis software used. Typically applied to the quantification analysis of peptides and proteins.

ii) Gel Images – Any gel image files generated, in the event that two-dimensional gel electrophoresis has been used as a separation method.

iii) Sequence Databases – Any files from protein or other sequence databases that were associated with or searched against in the analysis of this dataset, if applicable (usually .fasta format).

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

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

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

a) 'Enter a Password' (e.g. to share selectively with collaborators and manuscript reviewers),

b) 'Share on ProteomeXchange' is **not applicable to small molecule workflows**: checking the box will submit and announce the dataset via the ProteomeXchange consortium at the time that it is made public on MassIVE. The dataset will not appear publicly in either repository until you click the 'Make Public' button on your dataset's status page (see below).

c) 'Generate a DOI' if you want a Digital Object Identifier to be generated and assigned to this dataset. This is encouraged for all public datasets and can be used in publications.

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

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

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

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

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

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

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

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

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

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

Molecular networking workflow

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

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

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

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

CRITICAL NOTE: The default parameters recommended above for high resolution mass spectrometers will not result in comprehensive searches of the spectral libraries generated on low resolution mass spectrometers, such as ReSpect⁷⁸, large portions of MassBanks⁷⁴, GNPS community contributed; a significant portion of spectra that were annotated by matching to the NIST Mass Spectral Library with Search Program Data Version: NIST v17 (<https://www.nist.gov/srd/nist-standard-reference-database-1a-v17>) are also low resolution. In addition the natural products community contributes annotated spectra that may be high or low resolution, from a range of different spectrometers.

!CAUTION! Though using low resolution parameters may increase the number of annotations, it will also increase the number of false positive annotations.

PIMT: This parameter is used for MS-Cluster^{10,12} and spectral library searching, and the value influences the clustering of nearly identical MS² spectra via MS-Cluster.

FIMT: For every group of MS² spectra being considered for clustering (consensus spectrum creation), this value specifies how much fragment ions can be shifted from their expected *m/z* values.

Table 3. Absolute mass differences (Da) and associated mass error (parts-per-million, 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

For advanced users:

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

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

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 m/z 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 m/z 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.

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

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

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

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

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

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

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

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

*Metadata format for 'ili'*¹¹

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

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

3.4 Visualization of the molecular network

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

3.4.1 Molecular network visualization in browser

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

Job Status	
Workflow	<p>METABOLOMICS-SNET5-V2</p> <p>DONE [Clone] [Restart][Delete]</p> <p>Default Molecular Networking Results Views [View All Library Hits View Unique Library Compounds View All Clusters With IDs]</p> <p>Network Visualizations [View Spectral Families (In Browser Network Visualizer) Network Summarizing Graphs]</p> <p>Methods and Citation for Manuscripts [Networking Parameters and Written Network Description]</p> <p>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]</p>
Status	<p>Advanced Views - Global Public Dataset Matches [View Matches to All Public Datasets]</p> <p>Advanced Views - Third Party Visualization [View Emporer PCoA Plot in GNPS View ili in GNPS]</p> <p>Advanced Views - Networking Graphs/Histograms [Nodes, MZ Histogram Edges, MZ Delta Histogram Edges, Score vs MZ Delta Plot Library Search, PPM Error Histogram]</p> <p>Advanced Views - Misc Views [View Network, Node Centric View Network Pairs Networking Statistics View Raw/Unclustered Spectra View Compounds and File Occurrence]</p> <p>Advanced Views - Make Dataset Public Documentation [Make Public Dataset]</p> <p>Advanced Views - Experimental Views [Direct Cytoscape Preview/Download]</p>
User	emgentry (egentry.nc@gmail.com), UC San Diego
Title	GF/SPF Mouse Duodenum Example

Figure 4. GNPS Job Status Page.

Table 5. Data analysis options

<i>Data Analysis Option</i>	<i>Description</i>
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 .
--	---

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

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

histogram icon can be selected to visualize all product ions from the MS² spectra in the sub-network.

- *Align Spectra* - This function enables direct comparison between the spectra of two connected nodes at the peak level. To perform this analysis, the user should first select an edge connecting two nodes, which pulls up the spectra for each node in the right display window. Clicking the “align spec” button overlays the spectra, where red peaks represent peaks of the exact same masses shared between the top and bottom spectra and blue peaks denote peaks matching at shifted masses.
- *Search Peptide* - This is a function added to GNPS to support proteomic and peptidomic dataset analysis. If a peptide sequence is found to be associated with the molecular family and was found through automated peptide mining in MassIVE then the amino acid sequence entered here will be searched.

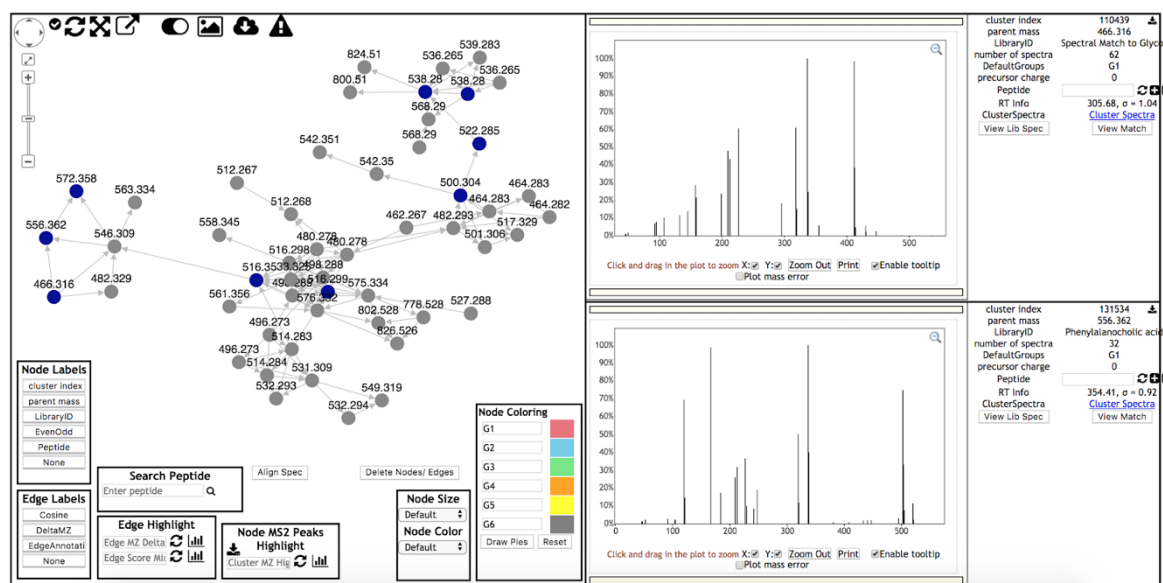


Figure 5. In browser visualization of the bile acid spectral family from dataset MSV000083437.

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

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

To decrease the impact of this variation all spectra, when compared, are subjected to a square root conversion. This decreases the high intensity ions and increases the low intensity ions. Furthermore, to address variability in data quality and source of the reference spectra, GNPS utilizes a ranking system for submitted reference spectra, to enable filtering of the reference library either before performing molecular networking or afterwards, which is the default approach. Similarly the instrument that the reference data were collected on can be considered after doing the analysis in GNPS using post-molecular networking filtering capabilities. 'Gold' reference spectra can only be submitted by approved users and must originate from fully characterized synthetic or purified compounds. This is the same gold standard by which other metabolomics reference libraries such as NIST⁷², METLIN⁷³ mzCloud (<https://www.mzcloud.org/>)⁷⁶, WeizMass (https://www.weizmann.ac.il/LS_CoreFacilities/weizmass-spectral-library-high-confidence-metabolite-identification)¹²⁶ libraries are curated. Gold level spectra comprise 83% of the MS² spectra provided to GNPS as libraries. A 'silver' rating signifies that the spectrum was submitted with an associated publication. However, GNPS also curates crowdsourced knowledge from users in the community. All remaining reference spectra provided by the user community receive a 'bronze' rating to denote that the annotation is contributed by users including partial or putative annotations. The annotation within GNPS can be made directly from the data and thus relies on the expertise of the experimentalist and purification of the molecules is not required. This gives access to a curated reference database that is crowdsourced and does not rely on commercially available standards. For example, most natural products from microbes, food and plants are not commercially available and thus the crowdsourced knowledge capture provides a resource of information that is inaccessible any other way. The only other resource that currently accepts putative and partial annotations is MassBank EU (<https://massbank.eu/MassBank/>). Examples of useful but partial annotations include modifications of molecules, such as oxidation of a molecule in which the site of oxidation is unknown¹²⁷ and thus a SMILES or InChI cannot be drawn but the partial annotation provides valuable insight to the end user. Additional partial annotations would include adduct clusters such as sodium formate clusters or polymeric substances, including oligosaccharides, commonly detected in mass spectrometry where a structure cannot be drawn but is useful knowledge for the community when performing an untargeted LC-MS/MS experiment. Users can use the above information along with the corresponding cosine score, which takes into account the number of matching fragment ions and differences in peak intensities, and parent mass accuracy to assess the quality of annotation. An empirical cut-off for cosine scoring of 0.7 with 6 MS² ions matching is the default setting in GNPS. On average this gives rise to 91% accurate annotations, and ~1% incorrect annotations, with the remainder being attributed to possible isomers (4%) or having not enough information by the user to judge (4%)¹²⁰. However, using a target decoy-based method to estimate confidence measures of annotations and false discovery rates (FDR) in large scale metabolomics experiments, revealed that the annotation quality is dataset-dependent and dependent on analysis settings such as number of ions that are required to match. The general trend was that when few MS² ions are required to match, a much higher cosine is required and fewer matches will be obtained at the same FDR

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

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

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

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

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

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

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

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

- a. The node labels can be changed in Cytoscape by selecting the dropdown arrow next to the “Label” tab. Readers following the tutorial example can label nodes by selecting “precursor mass” as column and “Passthrough Mapping” for mapping type (**SI Fig. 6f**).
- b. Node shape can also be changed. Readers following the tutorial example can click directly on the “Shape” symbol button and select “Ellipse” shape or change to another desired shape (**SI Fig. 6g**). If using ellipse, the shape can be converted into a circle by checking the box labeled ‘lock node width and height’ (**SI Fig 6h**).
- c. To change the node color, click on “Fill Color” dropdown. Under this column, readers following the tutorial example can select the desired value (i.e. “ATTRIBUTE_host_microbiome”) and use this to discriminate groups (i.e. germ free vs. specific pathogen free) from one another. Readers can select “Discrete Mapping” under the “Mapping Type” column, which allows for the selection of a color to be associated with each group (**SI Fig. 6i**).
- d. Alternatively at the “Fill Color” option, the “Image/Chart 1” tab can be used to visualize the relative ion distribution from each chosen group in the nodes as a pie chart. Readers following the tutorial can perform this type of visualization by clicking on the “Image/Chart 1” button, selecting the “Charts” tab, and choosing a chart type (the pie chart is chosen in this example). The spectral count information from groups defined in the metadata file can then be selected from the “Available columns” to the “Selected columns” (**SI Fig. 6j**) and the user can edit the chart color scheme using the “Options” tab. In this example, “Germ free” and “Specific Pathogen free” can be selected and colored pink and blue, respectively.
- e. To visualize the variation in the occurrence of each ion across samples (e.g. count of 1 if not zero) as a function of the node size, go to Size option, select “number of spectra” or “sum(precursor intensity)” as *Column* and “Continuous Mapping” as *Mapping Type*. The opened window allows to modify the node size in function of the node metadata column chosen. Begin by setting the value for minimum and maximum node size value with the button *Set Min and Max*, 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**).
- 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

tutorial example can select this tab to make alterations in edge color and width, in addition to other settings.

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

3.5 How to propagate annotations through manual interpretation of the networks

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

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

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

3.6 Capturing information by adding reference spectra from your data

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

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

3.7 Data sharing & reproducibility of molecular networking

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

3.7.1. Cloning a job

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

3.7.2. Accessing a dataset

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

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

3.7.3. Subscribing to a dataset and living data

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

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

BOX 4: Feature-based molecular networking (FBMN)

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

4.0 Troubleshooting

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

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.

5.0 Anticipated Results

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.

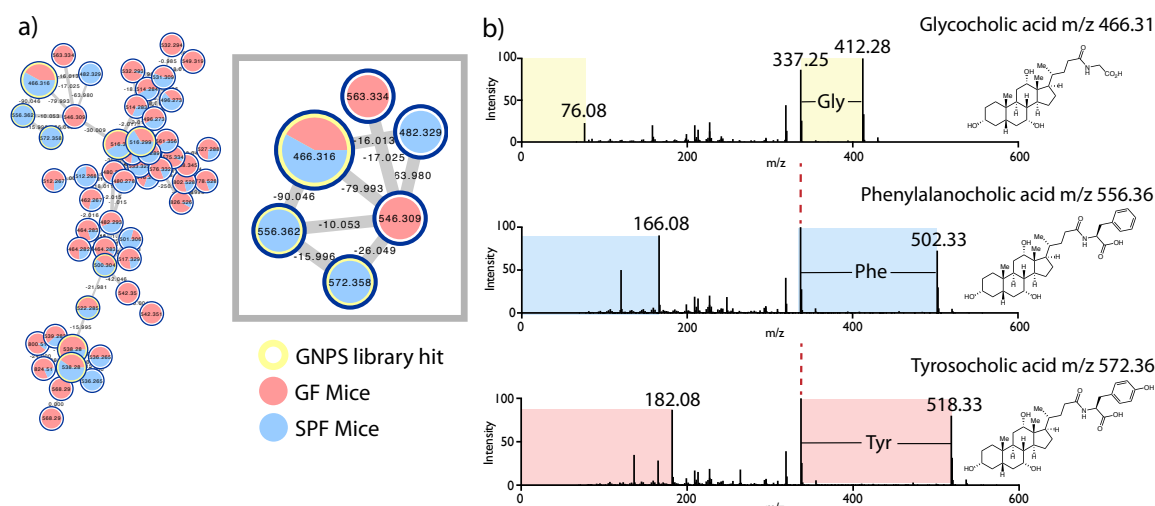


Figure 5. (a) The molecular family of conjugated bile acids from the duodenum of germ free (GF) (red) vs. specific pathogen free (SPF) (blue) mice in [MSV000083437](#) dataset. As shown in the inset, a library hit for glycocholic acid (m/z 466.316) is present in both GF and SPF mice while the new phenylalanine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

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

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

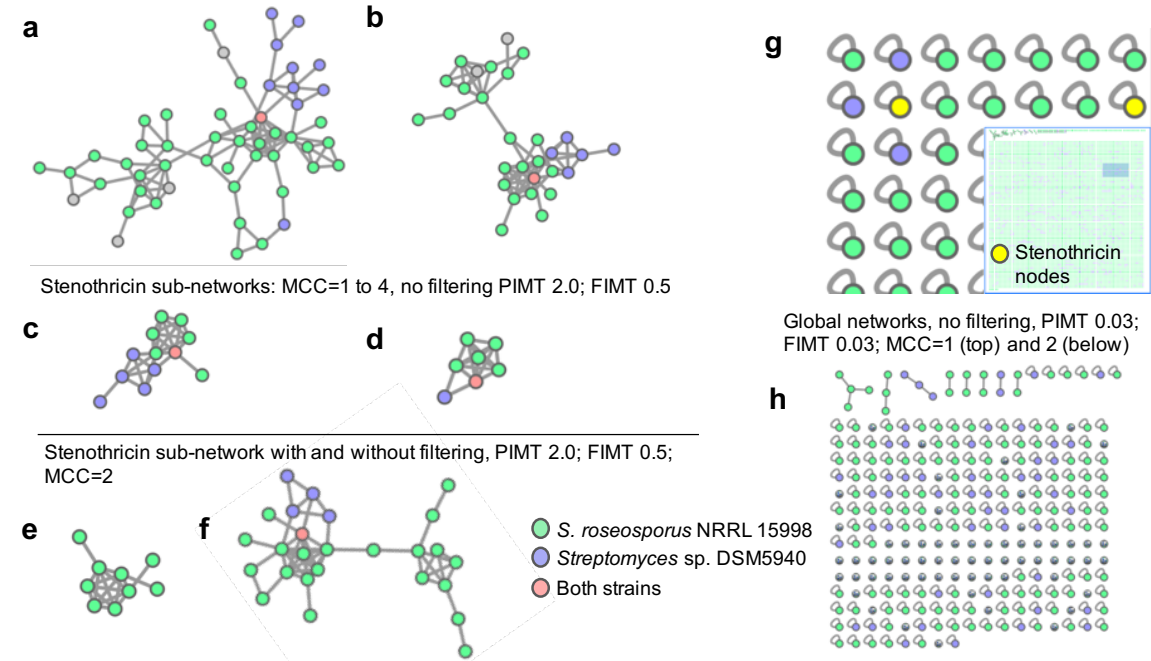


Figure 6. Networking of the stenothricin natural product molecular family ([MSV000083381](#)) detected in *Streptomyces* sp. DSM5940 (purple nodes), *S. roseosporus* NRRL 15998 (green nodes) or both strains (yellow nodes). Variation in number of nodes and spectra with 'Minimum Cluster Size' (MCS) yields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

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

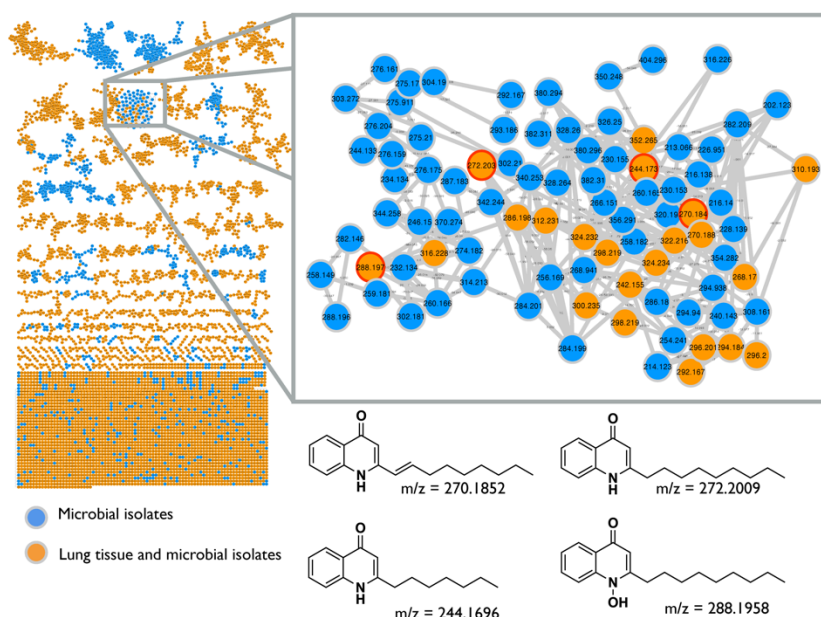


Figure 7. Molecular family (a sub-network) of quinolones detected in lung tissue extracts (orange nodes) and cultured *Pseudomonas* isolates (cyan nodes), created from MassIVE dataset [MSV000083359](https://massive.ucsf.edu/MSV000083359). 2-heptyl-4-quinolone (HHQ), 2-nonyl-4-quinolone (NHQ) and its unsaturated derivative (NHQ-C9:1 db), and 2-nonyl-4-quinolone-N-oxide (NQNO) were found in lung tissue, and are highlighted by a red node border.

With a network in hand, there are a number of data analysis tools and experimental validation steps that may be performed. As discussed in section 3.4.2, to legitimize a library annotation beyond inspecting mirror plots, the user should verify molecular formula and identify associated adducts using MS¹ data. Additionally, rationalization based on biological source is recommended. Ideally, an annotation is authenticated by comparison with a known standard compound or isolation and full characterization. In the example followed throughout the protocol, the molecular structures of the new conjugated bile acids from the mouse duodenum were confirmed by comparison with synthetic standards. For more complex structures such as those in the stenothricin example¹¹ (Figure 6), the most abundant analog, stenothricin-GNPS 2, was purified for acquisition. The structure was assigned from 1D and 2D NMR data, Marfey's analysis¹³⁴, and manual comparison of the MS² spectra with MS² spectra for previously reported stenothricin D. Genome mining further supported the conclusion that the -41 Da mass shift observed for stenothricin-GNPS 1-5 is due to a Lys to Ser substitution. For nodes that are not annotated, the *in silico* Dereplicator may predict peptidic natural products, while NAP (Network Annotation Propagation) can use annotated nodes to predict related metabolites. Molecular formulas may be generated using additional tools, one of which is SIRIUS¹⁰⁸. This software uses MS² features to arrive at the best molecular formula for the precursor MS¹ ion, and works best for smaller molecules (<600 Da).

In the example of the human lung colonized by *Pseudomonas* bacteria (Figure 7)⁵⁰, the authors use spatial mapping to visualize annotated molecules on an exploded lung, and then correlate the distribution of molecules to microbiome maps generated from 16S rRNA gene amplicon sequencing. This study shows how molecular networking can be used to

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

Acknowledgements:

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

References:

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

- 1707 11. Wang, M. et al. Sharing and community curation of mass spectrometry data with
1708 Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**, 828-837
1709 (2016).
- 1710 12. Frank, A.M. et al. Spectral archives: extending spectral libraries to analyze both
1711 identified and unidentified spectra. *Nat Methods* **8**, 587-591 (2011).
- 1712 13. De Vijlder, T. et al. A tutorial in small molecule identification via electrospray
1713 ionization-mass spectrometry: The practical art of structural elucidation. *Mass*
1714 *Spectrom Rev* **37**, 607-629 (2018).
- 1715 14. Sumner, L.W. et al. Proposed minimum reporting standards for chemical analysis
1716 Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative
1717 (MSI). *Metabolomics* **3**, 211-221 (2007).
- 1718 15. Su, G., Morris, J.H., Demchak, B. & Bader, G.D. Biological network exploration with
1719 Cytoscape 3. *Curr Protoc Bioinformatics* **47**, 8 13 11-24 (2014).
- 1720 16. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new
1721 features for data integration and network visualization. *Bioinformatics* **27**, 431-432
1722 (2011).
- 1723 17. Sandhu, C. et al. Evaluation of Data-Dependent versus Targeted Shotgun
1724 Proteomic Approaches for Monitoring Transcription Factor Expression in Breast
1725 Cancer. *Journal of Proteome Research* **7**, 1529-1541 (2008).
- 1726 18. Hubert, J., Nuzillard, J.-M. & Renault, J.-H. Dereplication strategies in natural
1727 product research: How many tools and methodologies behind the same concept?
1728 **16**, 55-95 (2017).
- 1729 19. Rochat, B. Proposed Confidence Scale and ID Score in the Identification of Known-
1730 Unknown Compounds Using High Resolution MS Data. *J Am Soc Mass Spectrom*
1731 **28**, 709-723 (2017).
- 1732 20. All natural. *Nature Chemical Biology* **3**, 351 (2007).
- 1733 21. in The "Gold Book", Edn. 2nd. (eds. A.D. McNaught & A. Wilkinson) (Blackwell
1734 Scientific Publications, Oxford; 1997).
- 1735 22. McLafferty, F.W. Tandem mass spectrometry. *Science* **214**, 280-287 (1981).
- 1736 23. Gross, J.H. in *Mass Spectrometry: A Textbook* 415-478 (Springer Berlin
1737 Heidelberg, Berlin, Heidelberg; 2011).
- 1738 24. Artyukhin, A.B. et al. Metabolomic "Dark Matter" Dependent on Peroxisomal β -
1739 Oxidation in *Caenorhabditis elegans*. *Journal of the American Chemical Society*
1740 **140**, 2841-2852 (2018).
- 1741 25. Edwards, E.D., Woolly, E.F., McLellan, R.M. & Keyzers, R.A. Non-detection of
1742 honeybee hive contamination following *Vespula* wasp baiting with protein
1743 containing fipronil. *PLoS One* **13**, e0206385 (2018).
- 1744 26. Hoffmann, T. et al. Correlating chemical diversity with taxonomic distance for
1745 discovery of natural products in myxobacteria. *Nature Communications* **9**, 803
1746 (2018).
- 1747 27. Leipoldt, F. et al. Warhead biosynthesis and the origin of structural diversity in
1748 hydroxamate metalloproteinase inhibitors. *Nat Commun* **8**, 1965 (2017).
- 1749 28. Kang, K.B., Gao, M., Kim, G.J., Choi, H. & Sung, S.H. Rhamnelloides A and B,
1750 omega-Phenylpentaene Fatty Acid Amide Diglycosides from the Fruits of
1751 *Rhamnella franguloides*. *Molecules* **23** (2018).
- 1752 29. Remy, S. et al. Structurally Diverse Diterpenoids from *Sandwithia guyanensis*.
1753 *Journal of Natural Products* **81**, 901-912 (2018).
- 1754 30. Riewe, D., Wiebach, J. & Altmann, T. Structure Annotation and Quantification of
1755 Wheat Seed Oxidized Lipids by High-Resolution LC-MS/MS. *Plant Physiol* **175**,
1756 600-618 (2017).
- 1757 31. Senges, C.H.R. et al. The secreted metabolome of *Streptomyces*
1758 *chartreusis* and implications for bacterial chemistry. *Proceedings of the*
1759 *National Academy of Sciences* **115**, 2490-2495 (2018).

- 1760 32. van der Hooft, J.J.J. et al. Unsupervised Discovery and Comparison of Structural
1761 Families Across Multiple Samples in Untargeted Metabolomics. *Anal Chem* **89**,
1762 7569-7577 (2017).
- 1763 33. Wolff, H. & Bode, H.B. The benzodiazepine-like natural product tilivalline is
1764 produced by the entomopathogenic bacterium *Xenorhabdus eapokensis*. *PLoS*
1765 *One* **13**, e0194297 (2018).
- 1766 34. von Eckardstein, L. et al. Total Synthesis and Biological Assessment of Novel
1767 Albicidins Discovered by Mass Spectrometric Networking. *Chemistry* **23**, 15316-
1768 15321 (2017).
- 1769 35. Vizcaino, M.I. & Crawford, J.M. The colibactin warhead crosslinks DNA. *Nat Chem*
1770 **7**, 411-417 (2015).
- 1771 36. Saleh, H. et al. Deuterium-Labeled Precursor Feeding Reveals a New pABA-
1772 Containing Meroterpenoid from the Mango Pathogen *Xanthomonas citri* pv.
1773 *mangiferaeindicae*. *J Nat Prod* **79**, 1532-1537 (2016).
- 1774 37. Shannon, P. et al. Cytoscape: a software environment for integrated models of
1775 biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).
- 1776 38. Petras, D. et al. Mass Spectrometry-Based Visualization of Molecules Associated
1777 with Human Habitats. *Anal Chem* **88**, 10775-10784 (2016).
- 1778 39. Kapon, C.A. et al. Creating a 3D microbial and chemical snapshot of a human
1779 habitat. *Sci Rep* **8**, 3669 (2018).
- 1780 40. Adams, R.I. et al. Microbes and associated soluble and volatile chemicals on
1781 periodically wet household surfaces. *Microbiome* **5**, 128 (2017).
- 1782 41. Petras, D. et al. High-Resolution Liquid Chromatography Tandem Mass
1783 Spectrometry Enables Large Scale Molecular Characterization of Dissolved
1784 Organic Matter. *Frontiers in Marine Science* **4** (2017).
- 1785 42. Trautman, E.P. & Crawford, J.M. Linking Biosynthetic Gene Clusters to their
1786 Metabolites via Pathway- Targeted Molecular Networking. *Curr Top Med Chem* **16**,
1787 1705-1716 (2016).
- 1788 43. Luzzatto-Knaan, T., Melnik, A.V. & Dorrestein, P.C. Mass Spectrometry Uncovers
1789 the Role of Surfactin as an Interspecies Recruitment Factor. *ACS Chemical Biology*
1790 (2019).
- 1791 44. Machushynets, N.V., Wu, C., Elsayed, S.S., Hankemeier, T. & van Wezel, G.P.
1792 Discovery of novel glycerolated quinazolinones from *Streptomyces* sp. MBT27. *J*
1793 *Ind Microbiol Biotechnol* (2019).
- 1794 45. Yao, L. et al. Discovery of novel xylosides in co-culture of basidiomycetes *Trametes*
1795 *versicolor* and *Ganoderma applanatum* by integrated metabolomics and
1796 bioinformatics. *Sci Rep* **6**, 33237 (2016).
- 1797 46. Tripathi, A. et al. Intermittent Hypoxia and Hypercapnia, a Hallmark of Obstructive
1798 Sleep Apnea, Alters the Gut Microbiome and Metabolome. *mSystems* **3** (2018).
- 1799 47. Smits, S.A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-
1800 gatherers of Tanzania. *Science* **357**, 802-806 (2017).
- 1801 48. McDonald, D. et al. American Gut: an Open Platform for Citizen Science
1802 Microbiome Research. *mSystems* **3**, e00031-00018 (2018).
- 1803 49. Edlund, A. et al. Metabolic Fingerprints from the Human Oral Microbiome Reveal
1804 a Vast Knowledge Gap of Secreted Small Peptidic Molecules. *mSystems* **2**,
1805 e00058-00017 (2017).
- 1806 50. Garg, N. et al. Three-Dimensional Microbiome and Metabolome Cartography of a
1807 Diseased Human Lung. *Cell Host Microbe* **22**, 705-716 e704 (2017).
- 1808 51. McCall, L.I. et al. Mass Spectrometry-Based Chemical Cartography of a Cardiac
1809 Parasitic Infection. *Anal Chem* **89**, 10414-10421 (2017).
- 1810 52. Watrous, J.D. et al. Directed Non-targeted Mass Spectrometry and Chemical
1811 Networking for Discovery of Eicosanoids and Related Oxylipins. *Cell Chemical*
1812 *Biology* (2019).

- 1813 53. Allard, S., Allard, P.M., Morel, I. & Gicquel, T. Application of a molecular networking
1814 approach for clinical and forensic toxicology exemplified in three cases involving 3-
1815 MeO-PCP, doxylamine, and chlormequat. *Drug Test Anal* (2018).
- 1816 54. Ernst, M. et al. Did a plant-herbivore arms race drive chemical diversity in
1817 Euphorbia? *bioRxiv*, 323014 (2018).
- 1818 55. Philippus, A.C. et al. Molecular networking prospection and characterization of
1819 terpenoids and C15-acetogenins in Brazilian seaweed extracts. *RSC Advances* **8**,
1820 29654-29661 (2018).
- 1821 56. Li, F., Janussen, D., Peifer, C., Perez-Victoria, I. & Tasdemir, D. Targeted Isolation
1822 of Tsitsikammamines from the Antarctic Deep-Sea Sponge *Latrunculia biformis* by
1823 Molecular Networking and Anticancer Activity. *Mar Drugs* **16** (2018).
- 1824 57. Hartmann, A.C. et al. Meta-mass shift chemical profiling of metabolomes from coral
1825 reefs. *Proc Natl Acad Sci U S A* **114**, 11685-11690 (2017).
- 1826 58. Tobias, N.J. et al. Natural product diversity associated with the nematode
1827 symbionts *Photobacterium* and *Xenorhabdus*. *Nature Microbiology* **2**, 1676-1685
1828 (2017).
- 1829 59. Nothias, L.F. et al. Bioactivity-Based Molecular Networking for the Discovery of
1830 Drug Leads in Natural Product Bioassay-Guided Fractionation. *J Nat Prod* **81**, 758-
1831 767 (2018).
- 1832 60. Zou, Y. et al. Computationally Assisted Discovery and Assignment of a Highly
1833 Strained and PANC-1 Selective Alkaloid from Alaska's Deep Ocean. *Journal of the*
1834 *American Chemical Society* (2019).
- 1835 61. Parkinson, E.I. et al. Discovery of the Tyrobetaine Natural Products and Their
1836 Biosynthetic Gene Cluster via Metabologenomics. *ACS Chemical Biology* **13**,
1837 1029-1037 (2018).
- 1838 62. Naman, C.B. et al. Integrating Molecular Networking and Biological Assays To
1839 Target the Isolation of a Cytotoxic Cyclic Octapeptide, Samoamide A, from an
1840 American Samoan Marine Cyanobacterium. *Journal of Natural Products* **80**, 625-
1841 633 (2017).
- 1842 63. Bouslimani, A. et al. Lifestyle chemistries from phones for individual profiling. *Proc*
1843 *Natl Acad Sci U S A* **113**, E7645-E7654 (2016).
- 1844 64. Schymanski, E.L. et al. Critical Assessment of Small Molecule Identification 2016:
1845 automated methods. *Journal of Cheminformatics* **9**, 22 (2017).
- 1846 65. Quinn, R.A. et al. Niche partitioning of a pathogenic microbiome driven by chemical
1847 gradients. *Sci Adv* **4**, eaau1908 (2018).
- 1848 66. Aksenov, A.A., da Silva, R., Knight, R., Lopes, N.P. & Dorrestein, P.C. Global
1849 chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry* **1**,
1850 0054 (2017).
- 1851 67. Tsugawa, H. Advances in computational metabolomics and databases deepen the
1852 understanding of metabolisms. *Current Opinion in Biotechnology* **54**, 10-17 (2018).
- 1853 68. Johnson, S.R. & Lange, B.M. Open-Access Metabolomics Databases for Natural
1854 Product Research: Present Capabilities and Future Potential. *Frontiers in*
1855 *Bioengineering and Biotechnology* **3** (2015).
- 1856 69. Haug, K. et al. MetaboLights--an open-access general-purpose repository for
1857 metabolomics studies and associated meta-data. *Nucleic Acids Res* **41**, D781-786
1858 (2013).
- 1859 70. Perez-Riverol, Y. et al. Discovering and linking public omics data sets using the
1860 Omics Discovery Index. *Nat Biotechnol* **35**, 406-409 (2017).
- 1861 71. Stein, S.E. & Scott, D.R. Optimization and testing of mass spectral library search
1862 algorithms for compound identification. *Journal of the American Society for Mass*
1863 *Spectrometry* **5**, 859-866 (1994).
- 1864 72. NIST Standard Reference Database 1A v17.
- 1865 73. Guijas, C. et al. METLIN: A Technology Platform for Identifying Knowns and
1866 Unknowns. *Anal Chem* **90**, 3156-3164 (2018).

- 1867 74. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for
1868 life sciences. *J Mass Spectrom* **45**, 703-714 (2010).
- 1869 75. Stravs, M.A., Schymanski, E.L., Singer, H.P. & Hollender, J. Automatic
1870 recalibration and processing of tandem mass spectra using formula annotation. *J*
1871 *Mass Spectrom* **48**, 89-99 (2013).
- 1872 76. Wang, J., Peake, D.A., Mistrik, R., Huang, Y. & Araujo, G.D.
1873 ([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)
1874 [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)
1875 [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).
- 1876 77. Sheldon, M.T., Mistrik, R. & Croley, T.R. Determination of ion structures in
1877 structurally related compounds using precursor ion fingerprinting. *J Am Soc Mass*
1878 *Spectrom* **20**, 370-376 (2009).
- 1879 78. Sawada, Y. et al. RIKEN tandem mass spectral database (ReSpect) for
1880 phytochemicals: a plant-specific MS/MS-based data resource and database.
1881 *Phytochemistry* **82**, 38-45 (2012).
- 1882 79. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS:
1883 Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak
1884 Alignment, Matching, and Identification. *Analytical Chemistry* **78**, 779-787 (2006).
- 1885 80. Tautenhahn, R., Patti, G.J., Rinehart, D. & Siuzdak, G. XCMS Online: a web-based
1886 platform to process untargeted metabolomic data. *Anal Chem* **84**, 5035-5039
1887 (2012).
- 1888 81. Wanichthanarak, K., Fan, S., Grapov, D., Barupal, D.K. & Fiehn, O. Metabox: A
1889 Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration.
1890 *PLOS ONE* **12**, e0171046 (2017).
- 1891 82. Mohimani, H. et al. Dereplication of microbial metabolites through database search
1892 of mass spectra. *Nature Communications* **9**, 4035 (2018).
- 1893 83. Mohimani, H. et al. Dereplication of peptidic natural products through database
1894 search of mass spectra. *Nat Chem Biol* **13**, 30-37 (2017).
- 1895 84. Gurevich, A. et al. Increased diversity of peptidic natural products revealed by
1896 modification-tolerant database search of mass spectra. *Nat Microbiol* **3**, 319-327
1897 (2018).
- 1898 85. da Silva, R.R. et al. Propagating annotations of molecular networks using in silico
1899 fragmentation. *PLoS computational biology* **14**, e1006089 (2018).
- 1900 86. Mohimani, H. et al. Automated genome mining of ribosomal peptide natural
1901 products. *ACS Chem Biol* **9**, 1545-1551 (2014).
- 1902 87. Olivon, F. et al. MetGem Software for the Generation of Molecular Networks Based
1903 on the t-SNE Algorithm. *Anal Chem* (2018).
- 1904 88. Olivon, F., Roussi, F., Litaudon, M. & Touboul, D. Optimized experimental workflow
1905 for tandem mass spectrometry molecular networking in metabolomics. *Anal*
1906 *Bioanal Chem* **409**, 5767-5778 (2017).
- 1907 89. Wehrens, R. et al. Improved batch correction in untargeted MS-based
1908 metabolomics. *Metabolomics* **12**, 88 (2016).
- 1909 90. Koal, T. & Deigner, H.P. Challenges in mass spectrometry based targeted
1910 metabolomics. *Curr Mol Med* **10**, 216-226 (2010).
- 1911 91. Bylda, C., Thiele, R., Kobold, U. & Volmer, D.A. Recent advances in sample
1912 preparation techniques to overcome difficulties encountered during quantitative
1913 analysis of small molecules from biofluids using LC-MS/MS. *Analyst* **139**, 2265-
1914 2276 (2014).
- 1915 92. Vuckovic, D. Current trends and challenges in sample preparation for global
1916 metabolomics using liquid chromatography-mass spectrometry. *Anal Bioanal*
1917 *Chem* **403**, 1523-1548 (2012).
- 1918 93. Dunn, W.B. et al. Procedures for large-scale metabolic profiling of serum and
1919 plasma using gas chromatography and liquid chromatography coupled to mass
1920 spectrometry. *Nature Protocols* **6**, 1060 (2011).

- 1921 94. Taylor, P.J. Matrix effects: the Achilles heel of quantitative high-performance liquid
1922 chromatography-electrospray-tandem mass spectrometry. *Clin Biochem* **38**, 328-
1923 334 (2005).
- 1924 95. Annesley, T.M. Ion suppression in mass spectrometry. *Clin Chem* **49**, 1041-1044
1925 (2003).
- 1926 96. Crüsemann, M. et al. Prioritizing Natural Product Diversity in a Collection of 146
1927 Bacterial Strains Based on Growth and Extraction Protocols. *Journal of Natural*
1928 *Products* **80**, 588-597 (2017).
- 1929 97. Wandro, S., Carmody, L., Gallagher, T., LiPuma, J.J. & Whiteson, K. Making It
1930 Last: Storage Time and Temperature Have Differential Impacts on Metabolite
1931 Profiles of Airway Samples from Cystic Fibrosis Patients. *mSystems* **2** (2017).
- 1932 98. Zhao, J., Evans, C.R., Carmody, L.A. & LiPuma, J.J. Impact of storage conditions
1933 on metabolite profiles of sputum samples from persons with cystic fibrosis. *J Cyst*
1934 *Fibros* **14**, 468-473 (2015).
- 1935 99. Hirayama, A. et al. Effects of processing and storage conditions on charged
1936 metabolomic profiles in blood. *ELECTROPHORESIS* **36**, 2148-2155 (2015).
- 1937 100. Mushtaq, M.Y., Choi, Y.H., Verpoorte, R. & Wilson, E.G. Extraction for
1938 metabolomics: access to the metabolome. *Phytochem Anal* **25**, 291-306 (2014).
- 1939 101. Bazsó, F.L. et al. Quantitative Comparison of Tandem Mass Spectra Obtained on
1940 Various Instruments. *J Am Soc Mass Spectrom* **27**, 1357-1365 (2016).
- 1941 102. Bowen, B.P. & Northen, T.R. Dealing with the unknown: metabolomics and
1942 metabolite atlases. *J Am Soc Mass Spectrom* **21**, 1471-1476 (2010).
- 1943 103. da Silva, R.R., Dorrestein, P.C. & Quinn, R.A. Illuminating the dark matter in
1944 metabolomics. *Proc Natl Acad Sci U S A* **112**, 12549-12550 (2015).
- 1945 104. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software Tools and Approaches for
1946 Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **8**
1947 (2018).
- 1948 105. Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J. & Neumann, S. MetFrag
1949 relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*
1950 **8**, 3 (2016).
- 1951 106. Gerlich, M. & Neumann, S. MetFusion: integration of compound identification
1952 strategies. *J Mass Spectrom* **48**, 291-298 (2013).
- 1953 107. Böcker, S., Letzel, M.C., Liptak, Z. & Pervukhin, A. SIRIUS: decomposing isotope
1954 patterns for metabolite identification. *Bioinformatics* **25**, 218-224 (2009).
- 1955 108. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into
1956 metabolite structure information. *Nat Methods* **16**, 299-302 (2019).
- 1957 109. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Bocker, S. Searching molecular
1958 structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad*
1959 *Sci U S A* **112**, 12580-12585 (2015).
- 1960 110. Tsugawa, H. et al. Hydrogen Rearrangement Rules: Computational MS/MS
1961 Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal Chem*
1962 **88**, 7946-7958 (2016).
- 1963 111. Protsyuk, I. et al. 3D molecular cartography using LC-MS facilitated by Optimus
1964 and 'ili software. *Nat Protoc* **13**, 134-154 (2018).
- 1965 112. Röst, H.L. et al. OpenMS: a flexible open-source software platform for mass
1966 spectrometry data analysis. *Nat Methods* **13**, 741-748 (2016).
- 1967 113. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular
1968 framework for processing, visualizing, and analyzing mass spectrometry-based
1969 molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
- 1970 114. Deutsch, E.W. et al. Proteomics Standards Initiative: Fifteen Years of Progress and
1971 Future Work. *Journal of Proteome Research* **16**, 4288-4298 (2017).
- 1972 115. Brooksbank, C., Cameron, G. & Thornton, J. The European Bioinformatics
1973 Institute's data resources. *Nucleic Acids Res* **38**, D17-25 (2010).
- 1974 116. Jones, A.R. et al. The mzIdentML data standard for mass spectrometry-based
1975 proteomics results. *Mol Cell Proteomics* **11**, M111 014381 (2012).

1976 117. Griss, J. et al. The mzTab data exchange format: communicating mass-
1977 spectrometry-based proteomics and metabolomics experimental results to a wider
1978 audience. *Mol Cell Proteomics* **13**, 2765-2775 (2014).
1979 118. Hoffmann, N. et al. mzTab-M: A Data Standard for Sharing Quantitative Results in
1980 Mass Spectrometry Metabolomics. *Analytical Chemistry* (2019).
1981 119. Wang, M. et al. MASST: A Web-based Basic Mass Spectrometry Search Tool for
1982 Molecules to Search Public Data. *bioRxiv*, 591016 (2019).
1983 120. Scheubert, K. et al. Significance estimation for large scale metabolomics
1984 annotations by spectral matching. *Nat Commun* **8**, 1494 (2017).
1985 121. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I
1986 learned to stop worrying and love the ome-ome. *GigaScience* **1**, 7 (2012).
1987 122. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPERor: a tool for
1988 visualizing high-throughput microbial community data. *GigaScience* **2**, 16 (2013).
1989 123. Bolyen, E. et al. QIIME 2: Reproducible, interactive, scalable, and extensible
1990 microbiome data science. *PeerJ Preprints* (2018).
1991 124. McLafferty, F.W. & Tureček, F.e. Interpretation of mass spectra, Edn. 4th.
1992 (University Science Books, Mill Valley, Calif.; 1993).
1993 125. Viant, M.R., Kurland, I.J., Jones, M.R. & Dunn, W.B. How close are we to complete
1994 annotation of metabolomes? *Curr Opin Chem Biol* **36**, 64-69 (2017).
1995 126. Shahaf, N. et al. The WEIZMASS spectral library for high-confidence metabolite
1996 identification. *Nature Communications* **7**, 12423 (2016).
1997 127. Schymanski, E.L. et al. Identifying small molecules via high resolution mass
1998 spectrometry: communicating confidence. *Environ Sci Technol* **48**, 2097-2098
1999 (2014).
2000 128. Cleary, J.L., Luu, G.T., Pierce, E.C., Dutton, R.J. & Sanchez, L.M. BLANKA: an
2001 Algorithm for Blank Subtraction in Mass Spectrometry of Complex Biological
2002 Samples. *Journal of The American Society for Mass Spectrometry* (2019).
2003 129. Demarque, D.P., Crotti, A.E.M., Vessecchi, R., Lopes, J.L.C. & Lopes, N.P.
2004 Fragmentation reactions using electrospray ionization mass spectrometry: an
2005 important tool for the structural elucidation and characterization of synthetic and
2006 natural products. *Natural Product Reports* **33**, 432-455 (2016).
2007 130. van der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V. & Rogers, S. Topic
2008 modeling for untargeted substructure exploration in metabolomics. *Proceedings of*
2009 *the National Academy of Sciences* **113**, 13738-13743 (2016).
2010 131. Olivon, F., Grelier, G., Roussi, F., Litaudon, M. & Touboul, D. MZmine 2 Data-
2011 Preprocessing To Enhance Molecular Networking Reliability. *Analytical Chemistry*
2012 **89**, 7836-7840 (2017).
2013 132. Winnikoff, J.R., Glukhov, E., Watrous, J., Dorrestein, P.C. & Gerwick, W.H.
2014 Quantitative molecular networking to profile marine cyanobacterial metabolomes.
2015 *J Antibiot (Tokyo)* **67**, 105-112 (2014).
2016 133. Tsugawa, H. et al. MS-DIAL: data-independent MS/MS deconvolution for
2017 comprehensive metabolome analysis. *Nat Methods* **12**, 523-526 (2015).
2018 134. Marfey, P. Determination of D-Amino Acids .2. Use of a Bifunctional Reagent, 1,5-
2019 Difluoro-2,4-Dinitrobenzene. *Carlsberg Res Commun* **49**, 591-596 (1984).
2020

GNPSProtocols_Paper_08062019_Submission.pdf (4.24 MiB)

[view on ChemRxiv](#) • [download file](#)

Supplementary Figures

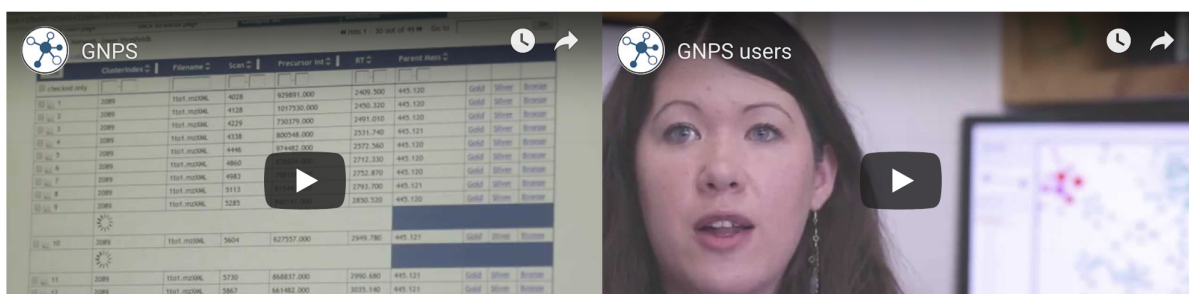
SI Figure 1. Navigate to the new GNPS homepage, and register an account.



Please Login to Use Workflows

Checkout the upcoming new GNPS homepage [here!](#)

Join our bi-weekly virtual video chat [office hours.](#)



Please Login to Analyze Data at GNPS

Login to Existing Account

Register New Account

OR



GNPS is a web-based mass spectrometry ecosystem that aims to be an open-access knowledge base for community-wide organization and sharing of raw, processed or identified tandem mass (MS/MS) spectrometry data. GNPS aids in identification and discovery throughout the entire life cycle of data; from initial data acquisition/analysis to post publication.

SI Figure 2. Create the metadata file to submit with a MassIVE dataset.

3DMouseoduodenum_metadata

HomeInsertPage LayoutFormulasDataReviewView

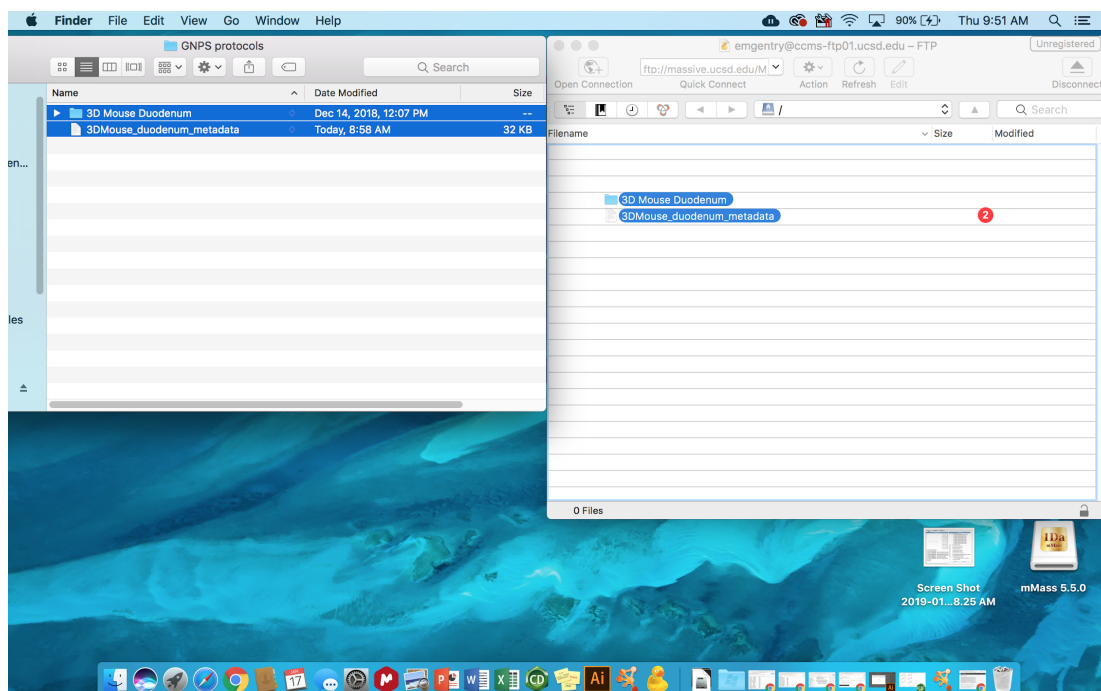
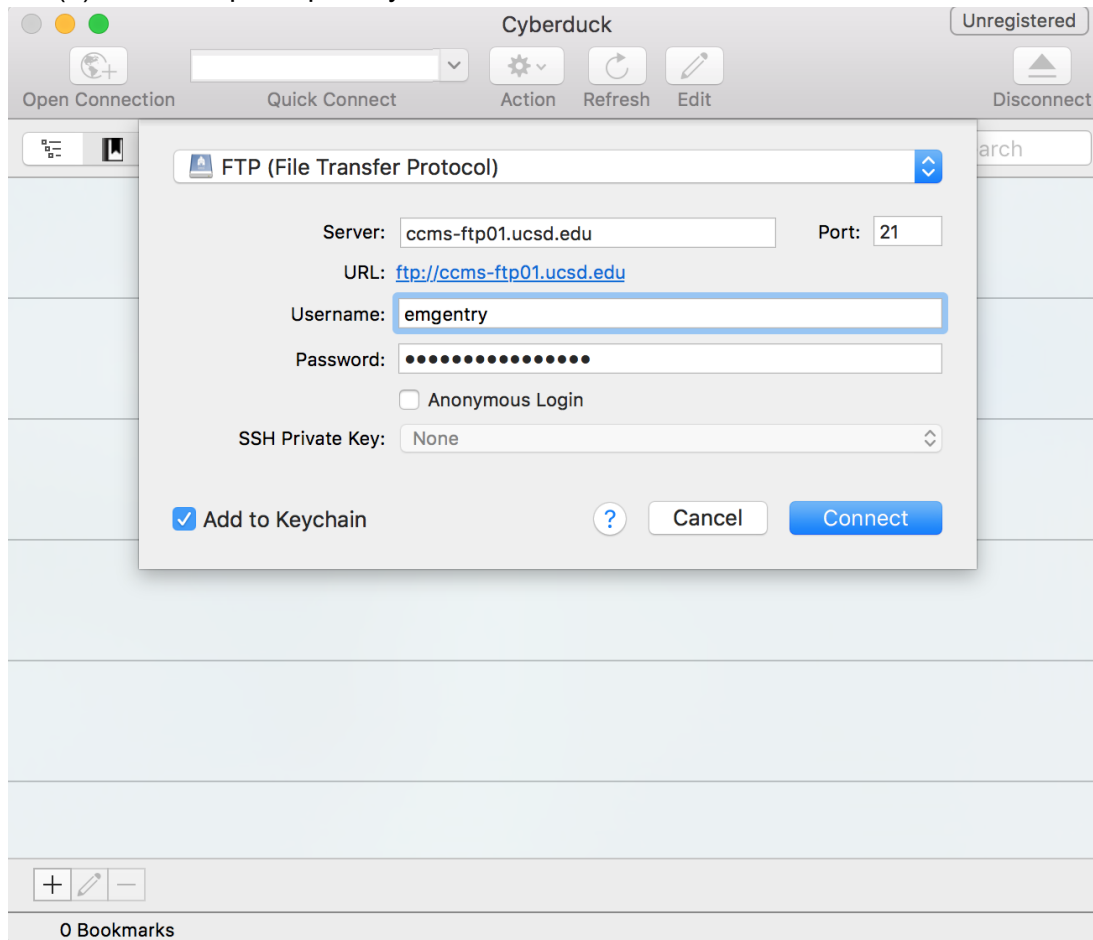
A1

	A	B	J	K	Q	U	V	W	X	Y
1	filename	sample_name	ATTRIBUTE_host_microbiome	ATTRIBUTE_host_subject_id	ATTRIBUTE_host_scientific_name	ATTRIBUTE_	ATTRIBUTE_	ATTRIBUTE_	body_product	
2	GF1_Duo1_GC4_01_21761.mzXML	Duodenum sampleid. 28	Germ free	GF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
3	GF1_Duo2_GC5_01_21762.mzXML	Duodenum sampleid. 29	Germ free	GF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
4	GF1_Duo3_GC6_01_21763.mzXML	Duodenum sampleid. 30	Germ free	GF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
5	GF1_Duo4_GC7_01_21764.mzXML	Duodenum sampleid. 31	Germ free	GF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
6	GF1_Duo5_GC8_01_21765.mzXML	Duodenum sampleid. 32	Germ free	GF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
7	GF1_Duo6_GC9_01_21766.mzXML	Duodenum sampleid. 33	Germ free	GF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
8	GF2_Duo1_BC4_01_22006.mzXML	Duodenum sampleid. 124	Germ free	GF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
9	GF2_Duo2_BC5_01_22007.mzXML	Duodenum sampleid. 125	Germ free	GF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
10	GF2_Duo3_BC6_01_22008.mzXML	Duodenum sampleid. 126	Germ free	GF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
11	GF2_Duo4_BC7_01_22009.mzXML	Duodenum sampleid. 127	Germ free	GF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
12	GF2_Duo5_BC8_01_22010.mzXML	Duodenum sampleid. 128	Germ free	GF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
13	GF2_Duo6_BC9_01_22011.mzXML	Duodenum sampleid. 129	Germ free	GF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
14	GF3_Duo1_BC4_01_22232.mzXML	Duodenum sampleid. 220	Germ free	GF3	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
15	GF3_Duo2_BC5_01_22233.mzXML	Duodenum sampleid. 221	Germ free	GF3	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
16	GF3_Duo3_BC6_01_22234.mzXML	Duodenum sampleid. 222	Germ free	GF3	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
17	GF3_Duo4_BC7_01_22235.mzXML	Duodenum sampleid. 223	Germ free	GF3	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
18	GF3_Duo5_BC8_01_22236.mzXML	Duodenum sampleid. 224	Germ free	GF3	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
19	GF3_Duo6_BC9_01_22237.mzXML	Duodenum sampleid. 225	Germ free	GF3	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
20	GF4_Duo1_BC4_01_22458.mzXML	Duodenum sampleid. 316	Germ free	GF4	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
21	GF4_Duo2_BC5_01_22459.mzXML	Duodenum sampleid. 317	Germ free	GF4	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
22	GF4_Duo3_BC6_01_22460.mzXML	Duodenum sampleid. 318	Germ free	GF4	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
23	GF4_Duo4_BC7_01_22461.mzXML	Duodenum sampleid. 319	Germ free	GF4	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
24	GF4_Duo5_BC8_01_22462.mzXML	Duodenum sampleid. 320	Germ free	GF4	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
25	GF4_Duo6_BC9_01_22463.mzXML	Duodenum sampleid. 321	Germ free	GF4	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
26	SPF1_Duo1_BC4_01_21878.mzXML	Duodenum sampleid. 412	Specific Pathogen Free	SPF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
27	SPF1_Duo2_BC5_01_21879.mzXML	Duodenum sampleid. 413	Specific Pathogen Free	SPF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
28	SPF1_Duo3_BC6_01_21880.mzXML	Duodenum sampleid. 414	Specific Pathogen Free	SPF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
29	SPF1_Duo4_BC7_01_21881.mzXML	Duodenum sampleid. 415	Specific Pathogen Free	SPF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
30	SPF1_Duo5_BC8_01_21882.mzXML	Duodenum sampleid. 416	Specific Pathogen Free	SPF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
31	SPF1_Duo6_BC9_01_21883.mzXML	Duodenum sampleid. 417	Specific Pathogen Free	SPF1	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
32	SPF2_Duo1_GC4_01_22119.mzXML	Duodenum sampleid. 508	Specific Pathogen Free	SPF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
33	SPF2_Duo2_GC5_01_22120.mzXML	Duodenum sampleid. 509	Specific Pathogen Free	SPF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
34	SPF2_Duo3_GC6_01_22121.mzXML	Duodenum sampleid. 510	Specific Pathogen Free	SPF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
35	SPF2_Duo4_GC7_01_22122.mzXML	Duodenum sampleid. 511	Specific Pathogen Free	SPF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
36	SPF2_Duo5_GC8_01_22123.mzXML	Duodenum sampleid. 512	Specific Pathogen Free	SPF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		
37	SPF2_Duo6_GC9_01_22124.mzXML	Duodenum sampleid. 513	Specific Pathogen Free	SPF2	Mus musculus	UBERON:gas	UBERON:duc	UBERON:wall of small intestine		

3DMouseoduodenum_metadata

SI Figure 3. Step-by-step screenshots for submission of a dataset from MassIVE into GNPS.

(a) **3.2.2.** Step 1. Upload your data files to the MassIVE web server



(b) 3.2.2. Step 2: Run the MassIVE dataset submission workflow

Workflow Selection

Workflow: MassIVE Dataset Submission (v1.2.8) Search Protocol: None Reset Form Save as Protocol

Title: GNPS Example Dataset_GF vs. SPF Mouse Duodenum

Dataset Metadata

[See here for detailed instructions on how to submit a MassIVE dataset.](#)

Enter at least 3 characters to see options from the controlled vocabulary.
If what you want isn't available, then you can just type free text and add it by clicking the "Add" button.

Species:

Mus musculus (house mouse) NCBITaxon:10090

Add

Instrument:

maXis (Bruker Daltonics maXis series) MS:1001541

Add

Post-Translational Modifications:

No PTMs are included in the dataset MS:1002864

Add

Keywords:

Dataset Keyword mouse

Dataset Keyword duodenum

Add

Principal Investigator:

Name:

Email:

Institution:

Country:

Principal Investigator Pieter Dorrestein (pdorrestein@ucsd.edu), UCSD, United States

Select Input Files

Upload Files

Share Files

Select Input Files

Peak List Files

Result Files

Raw Spectrum Files

Search Engine Files

License Files

Sequence Databases

Spectral Libraries

Quantification Results

Statistical Analysis of Quantified Analytes

Methods and Protocols

Supplementary Files

Gel Images

CCMS_ProteomeDatabases

CCMS_SpectralLibraries

[Dataset MSV000078556] - "GNPS - Topobiographical molecular

[Dataset MSV000079949] - "GNPS 3D Germ Free and Specific Pa

[Dataset MSV000081098] - "GNPS Drug Metabolism Demo Data"

[Dataset MSV000081171] - "GNPS_Danone_fermentation_study"

[Dataset MSV000081777] - "GNPS - AMG_Bloom_Attack"

[Dataset MSV000082045] - "GNPS_HMP_cultures_full_set"

[Dataset MSV000082610] - "GNPS Novel Conjugate Bile Acids"

emgentry

speclibs

Selected Files

Selected Peak List Files

emgentry/3D Mouse Duodenum/Controls

emgentry/3D Mouse Duodenum/GF1

emgentry/3D Mouse Duodenum/GF2

emgentry/3D Mouse Duodenum/GF3

emgentry/3D Mouse Duodenum/GF4

emgentry/3D Mouse Duodenum/SPF1

emgentry/3D Mouse Duodenum/SPF2

emgentry/3D Mouse Duodenum/SPF3

emgentry/3D Mouse Duodenum/SPF4

Selected Result Files

Selected Raw Spectrum Files

Selected Search Engine Files

Selected License Files

Selected Sequence Databases

Selected Spectral Libraries

Selected Quantification Results

Selected Statistical Analysis of Quantified Analytes

Selected Methods and Protocols

Selected Supplementary Files

emgentry/3DMouse_duodenum_metadata.txt

Selected Gel Images

Clear Selection

Finish Selection

MassIVE Dataset Information

Title

Description

MassIVE Accession

Dataset Type

Make Public

Principal Investigators

Username

Contact Email

Species

Instrument

Post-Translational Modifications

Keywords

Number of Files

Total Size

Subscribers

Subscription Status

Analyze Data

FTP Download

Other Dataset Actions

Add Comment

Publications

GNPS_GF vs. SPF Mouse Duodenum

metabolomics data for duodenum samples of germ free (GF) vs specific pathogen free (SPF) mice

MSV000083437

Private | Partial

Make Public

Pieter Dorrestein (pdorrestein@ucsd.edu), UCSD, United States

emeentry

egentry.nc@gmail.com

Mus musculus

maXis

MS:1002864

mouse | duodenum

113

4.30 GB

1

Subscribe

Analyze Submitted Spectra

Import and Analyze Dataset with Networking Now!

ftp://MSV000083437@massive.ucsd.edu

Add Files

Add/Update Metadata

Add Publication

Comment on Dataset

Attach Reanalysis Results

SI Fig 4. Step-by-step screenshots of molecular networking workflow in GNPS.

(a)

GNPS: Global Natural Products(Social)Molecular Networking

Logout | My User | Update Profile | Jobs | MassIVE Datasets | Documentation | Forum | Contact

GNPS

GNPS is a web-based mass spectrometry ecosystem that aims to be an open-access knowledge base for community-wide organization and sharing of raw, processed or identified tandem mass (MS/MS) spectrometry data. GNPS aids in identification and discovery throughout the entire life cycle of data; from initial data acquisition/analysis to post publication.

Learn More

GNPS Forums

GNPS Quick Start Guide

GNPS Super-Quick Start Interface

Data Analysis

Molecular Networking

Organize Spectra by Family
Visualize Large Datasets
Discover New Molecules

Create Molecular Network

Help

Library Search

High-Throughput Dereplication
Comprehensive MS/MS Libraries
Find Analogs of Knowns

Match Libraries

Help

Molecular BLAST

Put Spectrum in Context
Query Across Public Spectra
Infer Metadata of Spectrum

Query Spectrum

Help

(b)

Workflow Selection

Title:

Search Protocol:

Networking Parameter Presets

Basic Options

Spectrum files must be centroided and be in an open spectrum format (mzXML, mzML, or mgf)

Spectrum Files (Required): [See here for further documentation about molecular networking.](#)

Spectrum Files G2: [Click Here here to run a demo molecular network.](#)

Spectrum Files G3:

Spectrum Files G4:

Spectrum Files G5:

Spectrum Files G6: [For custom group/attribute documentation click here](#)

(c)

Share Files With User

Import Data Share

Shared Users

Imported Data Shares

- ✖ CCMS_ProteomeDatabases
- ✖ CCMS_SpectralLibraries
- ✖ speclibs

(d)

Share Files With User

Import Data Share

Shared Users

Imported Data Shares

- ✖ [Dataset MSV000083437] - "GNPS_GF vs. SPF Mouse Duodenum"
- ✖ CCMS_ProteomeDatabases
- ✖ CCMS_SpectralLibraries
- ✖ speclibs

(e)

Interface (e) shows the "Select Input Files" window. The "Selected Files" list includes:

- Selected Spectrum Files G1
- Selected Spectrum Files G2
- Selected Spectrum Files G3
- Selected Spectrum Files G4
- Selected Spectrum Files G5
- Selected Spectrum Files G6
- Selected Metadata File
- Selected Group Mapping (Legacy)
- Selected Attribute Mapping (Legacy)
- Selected Library Files
 - spectlibs
- Selected STL Model for ili

The "Select Input Files" list includes:

- Spectrum Files G1
- Spectrum Files G2
- Spectrum Files G3
- Spectrum Files G4
- Spectrum Files G5
- Spectrum Files G6
- Metadata File
- Group Mapping (Legacy)
- Attribute Mapping (Legacy)
- Library Files
- STL Model for ili

The "Dataset MSV000083437" - "GNPS_GF vs. SPF Mouse Duodenum" folder is expanded, showing subfolders: ccms_parameters, ccms_statistics, other, peak, Controls, GF1, GF2, GF3, GF4, SPF1, SPF2, SPF3, SPF4, emgentry, and spectlibs.

(f)

Interface (f) shows the "Select Input Files" window. The "Selected Files" list includes:

- Selected Spectrum Files G1
 - MSV000083437/peak/Controls
 - MSV000083437/peak/GF1
 - MSV000083437/peak/GF2
 - MSV000083437/peak/GF3
 - MSV000083437/peak/GF4
 - MSV000083437/peak/SPF1
 - MSV000083437/peak/SPF2
 - MSV000083437/peak/SPF3
 - MSV000083437/peak/SPF4
- Selected Spectrum Files G2
- Selected Spectrum Files G3
- Selected Spectrum Files G4
- Selected Spectrum Files G5
- Selected Spectrum Files G6
- Selected Metadata File
- Selected Group Mapping (Legacy)
- Selected Attribute Mapping (Legacy)
- Selected Library Files
 - spectlibs
- Selected STL Model for ili

The "Select Input Files" list includes:

- Spectrum Files G1
- Spectrum Files G2
- Spectrum Files G3
- Spectrum Files G4
- Spectrum Files G5
- Spectrum Files G6
- Metadata File
- Group Mapping (Legacy)
- Attribute Mapping (Legacy)
- Library Files
- STL Model for ili

The "Dataset MSV000083437" - "GNPS_GF vs. SPF Mouse Duodenum" folder is expanded, showing subfolders: ccms_parameters, ccms_statistics, other, peak, Controls, GF1, GF2, GF3, GF4, SPF1, SPF2, SPF3, SPF4, emgentry, and spectlibs.

(g)

Select Input Files

- Spectrum Files G1
- Spectrum Files G2
- Spectrum Files G3
- Spectrum Files G4
- Spectrum Files G5
- Spectrum Files G6
- Metadata File
- Group Mapping (Legacy)
- Attribute Mapping (Legacy)
- Library Files
- STL Model for ili

Selected Files

- Selected Spectrum Files G1
 - MSV000083437/peak/Controls
 - MSV000083437/peak/GF1
 - MSV000083437/peak/GF2
 - MSV000083437/peak/GF3
 - MSV000083437/peak/GF4
 - MSV000083437/peak/SPF1
 - MSV000083437/peak/SPF2
 - MSV000083437/peak/SPF3
 - MSV000083437/peak/SPF4
- Selected Spectrum Files G2
- Selected Spectrum Files G3
- Selected Spectrum Files G4
- Selected Spectrum Files G5
- Selected Spectrum Files G6
- Selected Metadata File
- Selected Group Mapping (Legacy)
- Selected Attribute Mapping (Legacy)
- Selected Library Files
 - speclibs
- Selected STL Model for ili

Clear Selection Finish Selection

File Tree:

- CCMS_ProteomeDatabases
- CCMS_SpectralLibraries
- [Dataset MSV000083437] - "GNPS_GF vs. SPF Mouse Duodenum"
 - ccms_parameters
 - ccms_statistics
 - other
 - 3DMouse_duodenum_metadata.txt
 - peak
 - Controls
 - GF1
 - GF2
 - GF3
 - GF4
 - SPF1
 - SPF2
 - SPF3
 - SPF4
- emgentry
- speclibs

(h)

Select Input Files

- Spectrum Files G1
- Spectrum Files G2
- Spectrum Files G3
- Spectrum Files G4
- Spectrum Files G5
- Spectrum Files G6
- Metadata File
- Group Mapping (Legacy)
- Attribute Mapping (Legacy)
- Library Files
- STL Model for ili

Selected Files

- Selected Spectrum Files G1
 - MSV000083345/peak/GF1
 - MSV000083345/peak/GF2
 - MSV000083345/peak/GF3
 - MSV000083345/peak/SPF1
 - MSV000083345/peak/SPF3
 - MSV000083345/peak/SPF2
 - MSV000083345/peak/GF4
 - MSV000083345/peak/SPF4
- Selected Spectrum Files G2
- Selected Spectrum Files G3
- Selected Spectrum Files G4
- Selected Spectrum Files G5
- Selected Spectrum Files G6
- Selected Metadata File
 - MSV000083345/other/3DMouse_duodenum_metadata.txt
- Selected Group Mapping (Legacy)
- Selected Attribute Mapping (Legacy)
- Selected Library Files
 - speclibs
- Selected STL Model for ili

Clear Selection Finish Selection

File Tree:

- CCMS_ProteomeDatabases
- CCMS_SpectralLibraries
- [Dataset MSV000083345] - "GNPS Example Dataset_GF vs. SPF /"
 - ccms_parameters
 - ccms_statistics
 - other
 - 3DMouse_duodenum_metadata.txt
 - peak
 - Controls
 - GF1
 - GF2
 - GF3
 - GF4
 - SPF1
 - SPF2
 - SPF3
 - SPF4

(i)

Basic Options

Spectrum files must be centroided and be in an open spectrum format (mzXML, mzML, or mgf)

Spectrum Files (Required): 0 files and 9 folders are selected

[See here for further documentation about molecular networking.](#)

Spectrum Files G2:

[Click Here here to run a demo molecular network.](#)

Spectrum Files G3:

Spectrum Files G4:

Spectrum Files G5:

Spectrum Files G6:

[For custom group/attribute documentation click here](#)

Precursor Ion Mass Tolerance: Da

Fragment Ion Mass Tolerance: Da

Advanced Network Options

Min Pairs Cos:

Minimum Matched Fragment Ions:

Network TopK:

Minimum Cluster Size:

Run MScCluster:

Maximum Connected Component Size:

Metadata File: 1 file and 0 folders are selected

Group Mapping (Legacy):

Attribute Mapping (Legacy):

Hide Fields

Advanced Library Search Options

Spectral Library: 0 files and 1 folder are selected

[To import libraries for search click here](#)

Library Search Min Matched

Score Threshold:

Peaks:

Search Analogs:

Maximum Analog Search Mass

Difference:

Hide Fields

Advanced Filtering Options

Filter below Std Dev:

Minimum Peak Intensity:

Filter Precursor Window:

Filter Library:

Filter peaks in 50Da Window:

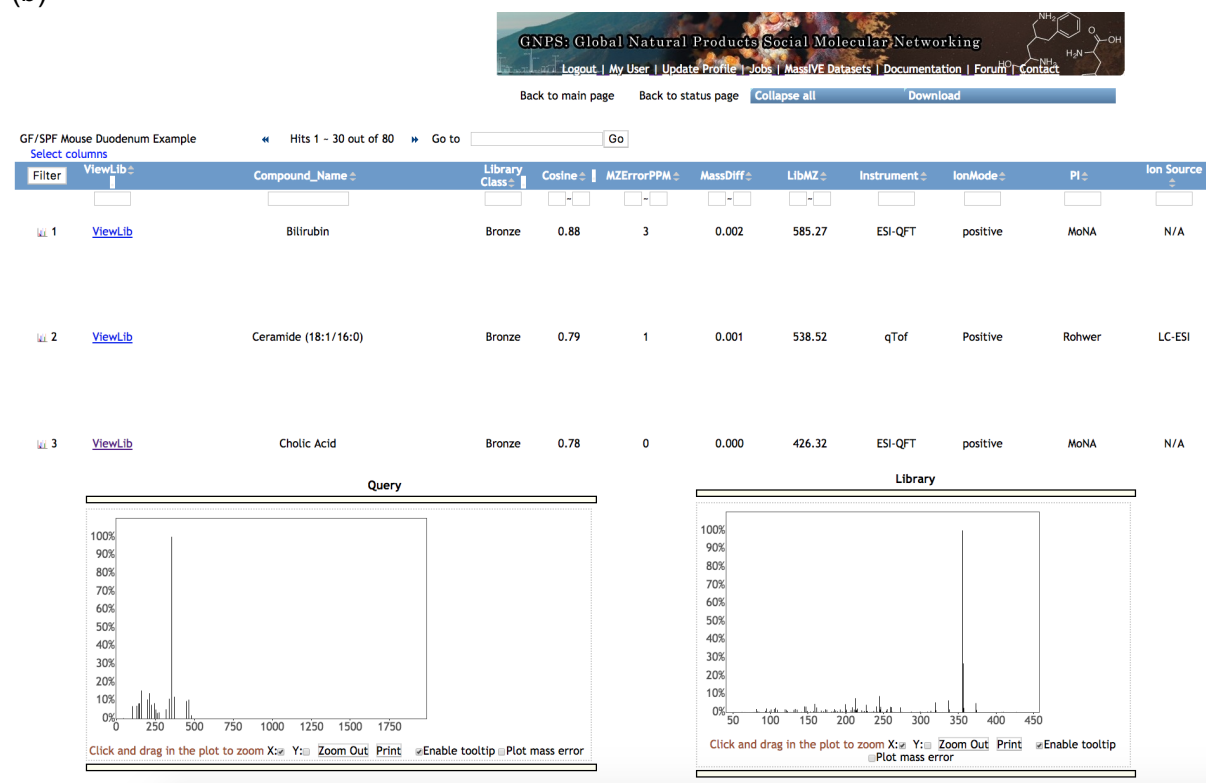
Hide Fields

SI Figure 5

(a)



(b)



(c)

GNPS: Global Natural Products Social Molecular Networking

[Logout](#) | [My User](#) | [Update Profile](#) | [Jobs](#) | [MassIVE Datasets](#) | [Documentation](#) | [Forum](#) | [Contact](#)

Back to main page

Back to status page

Collapse all

Download

GF/SPF Mouse Duodenum Example

Hits 1 - 30 out of 2543

Go to

Go

Select columns

Filter	Clusteridx	Spec Family	AnnotatetoGNPS	AddtoChallenge	NumSpectra	NumFiles	PrecursorMZ	PrecursorInt	RTMean		
View Analogs	Search Spec	1	Cluster - 4	AnnotatetoGNPS	AddtoChallenge	15	13	108.52000	55884.00000	826.2784	
View Analogs	Search Spec	2	Cluster - 66	View Network	AnnotatetoGNPS	AddtoChallenge	27	22	120.08400	8706950.00000	61.3616296296

(d)

GNPS: Global Natural Products Social Molecular Networking

[Logout](#) | [My User](#) | [Update Profile](#) | [Jobs](#) | [MassIVE Datasets](#) | [Documentation](#) | [Forum](#) | [Contact](#)

Back to main page

Back to status page

Collapse all

Download

GF/SPF Mouse Duodenum Example

Hits 1 - 30 out of 183

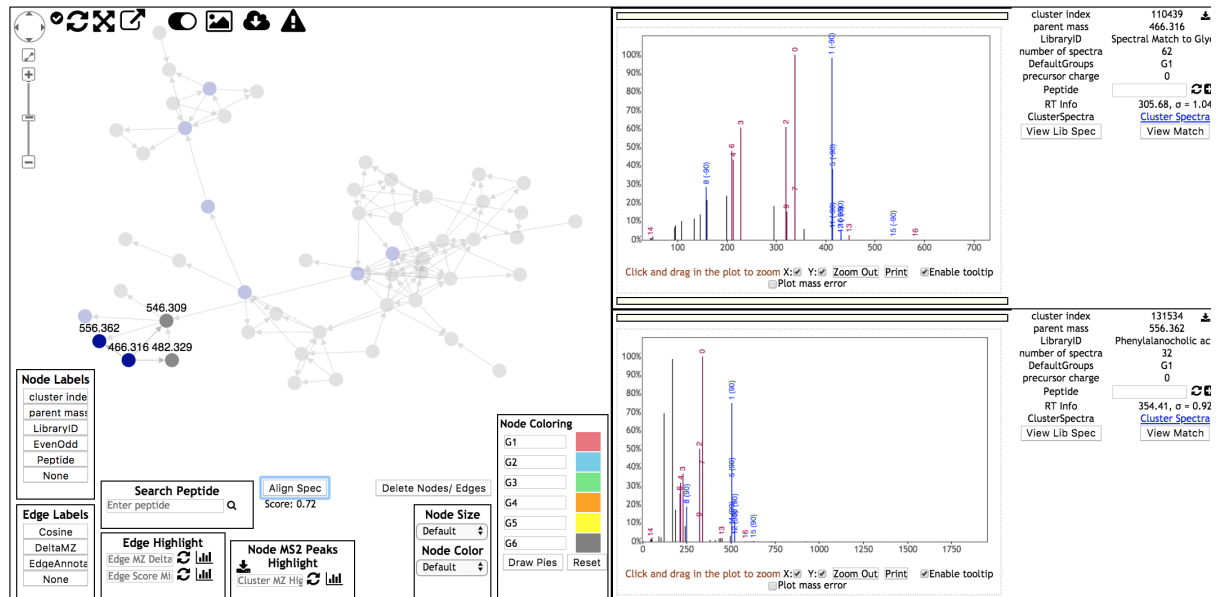
Go to

Go

Select columns

Filter	Visualize Network	View Network Nodes	NodeCount	%ID	#Spectra	AllIDs	DefaultGroups
1	Visualize Network	Network Nodes	97	0.320	4390	<div>Show</div>	G1
2	Visualize Network	Network Nodes	75	0.093	2824	<div>Show</div>	G1
3	Visualize Network	Network Nodes	73	0.110	2370	<div>Show</div>	G1
4	Visualize Network	Network Nodes	71	0.000	1858	N/A	G1
5	Visualize Network	Network Nodes	61	0.016	2290	<div>Show</div>	G1
6	Visualize Network	Network Nodes	58	0.017	1481	<div>Show</div>	G1
7	Visualize Network	Network Nodes	53	0.000	1637	N/A	G1
8	Visualize Network	Network Nodes	52	0.173	2800	<div>Show</div>	G1

(e) Align spectra feature of the 'view spectral families' option:



SI Figure 6. Step-by-step workflow for data visualization in Cytoscape.



(b)

Job Status

Workflow METABOLOMICS-SNETS-V2

Status

[DONE](#)
[\[Clone\]](#)

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
[\[Download Clustered Spectra as MGF \]](#) [\[Download GraphML for Cytoscape \]](#) [\[Download Bucket Table \]](#) [\[Download BioM For Qiime/Qiita \]](#) [\[Download Metadata For Qiime \]](#) [\[Download RT Data \]](#)

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
[\[Topology Signatures \]](#) [\[Topology Signatures Histogram \]](#) [\[Direct Cytoscape Preview/Download \]](#)

Option 1

(c)



Session: /Users/emilygentry/Google Drive (

Control Panel

Network **Style** Select Annotation

default

Properties

Def.	Map.	Byp.	
			Border Paint
0.0			Border Width
			Fill Color

(d)

Job Status	
Workflow	METABOLOMICS-SNETS-V2
Status	DONE [Clone]
	<p>Default Molecular Networking Results Views [View All Library Hits View Unique Library Compounds View All Clusters With IDs]</p> <p>Network Visualizations [View Spectral Families (In Browser Network Visualizer) Network Summarizing Graphs]</p> <p>Methods and Citation for Manuscripts [Networking Parameters and Written Network Description]</p> <p>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]</p> <p>Advanced Views - Global Public Dataset Matches [View Matches to All Public Datasets]</p> <p>Advanced Views - Third Party Visualization [View Emporer PCoA Plot in GNPS View Ili in GNPS]</p> <p>Advanced Views - Networking Graphs/Histograms [Nodes, MZ Histogram Edges, MZ Delta Histogram Edges, Score vs MZ Delta Plot Library Search, PPM Error Histogram]</p> <p>Advanced Views - Misc Views [View Network, Node Centric View Network Pairs Networking Statistics View Raw/Unclustered Spectra View Compounds and File Occurrence]</p> <p>Advanced Views - Make Dataset Public Documentation [Make Public Dataset]</p> <p>Advanced Views - Experimental Views [Topology Signatures Topology Signatures Histogram Direct Cytoscape Preview/Download]</p>

Option 2

Cytoscape Downloader

[Download Cytoscape File](#)



(e)

Control Panel

Network **Style** Select Annotation

default

Properties

Def.	Map.	By.
		Border Paint
0.0		Border Width
		Fill Color
35.0		Height
		Image/Chart 1
		Label
		Label Color
12		Label Font Size
		Shape
		Size
255		Transparency
75.0		Width

☐ Lock node width and height

Node Edge Network

METABOLOMICS-SNETS-V2-6a

Table Panel

Adduct Compound_Nam

(f)

Control Panel

Network **Style** Select Annotation

default

Properties

Def.	Map.	By.
		Border Paint
0.0		Border Width
		Fill Color
35.0		Height
		Image/Chart 1
		Label

Column precursor mass

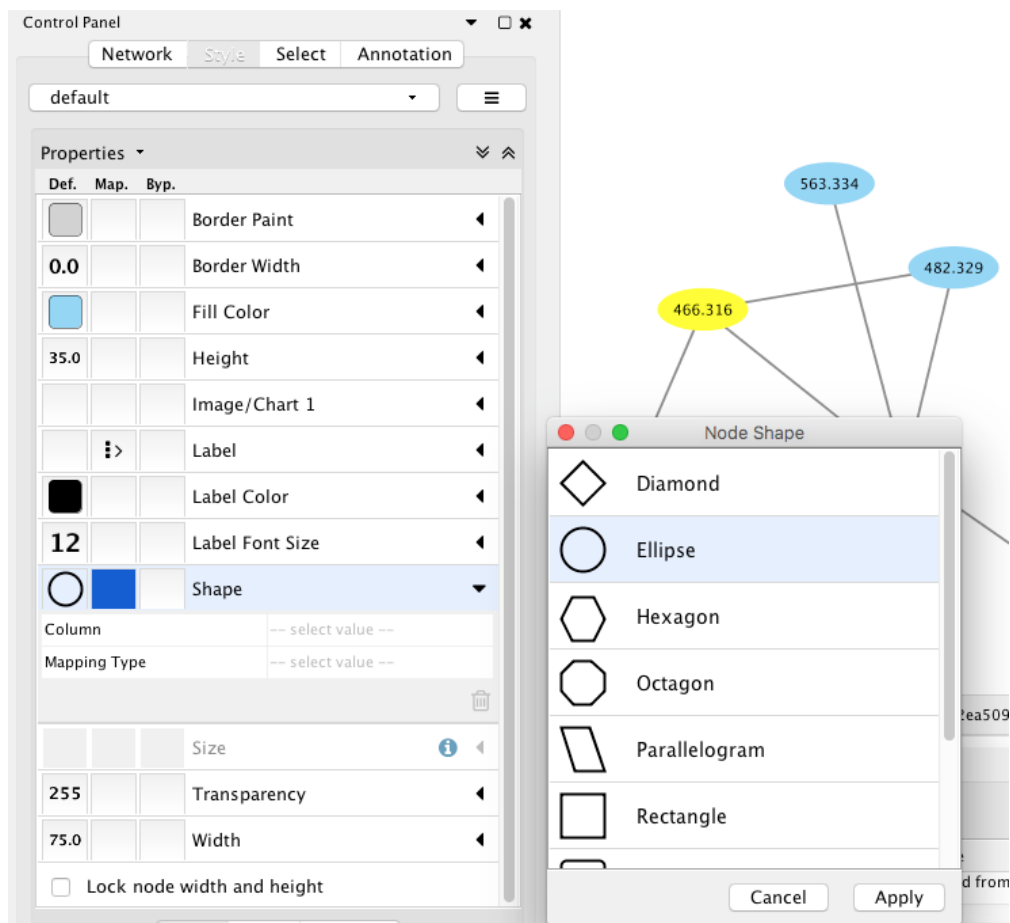
Mapping Type Passthrough Mapping

	Label Color
12	Label Font Size
	Shape
	Size
255	Transparency
75.0	Width

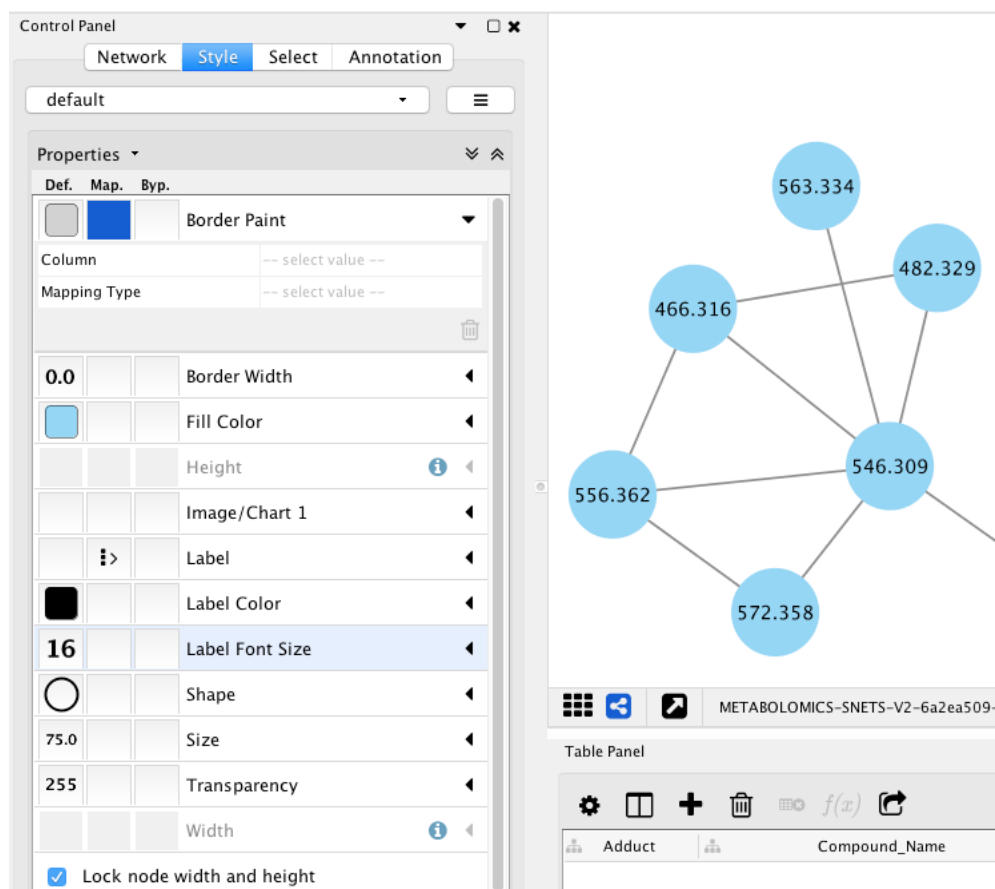
☐ Lock node width and height

Node Edge Network

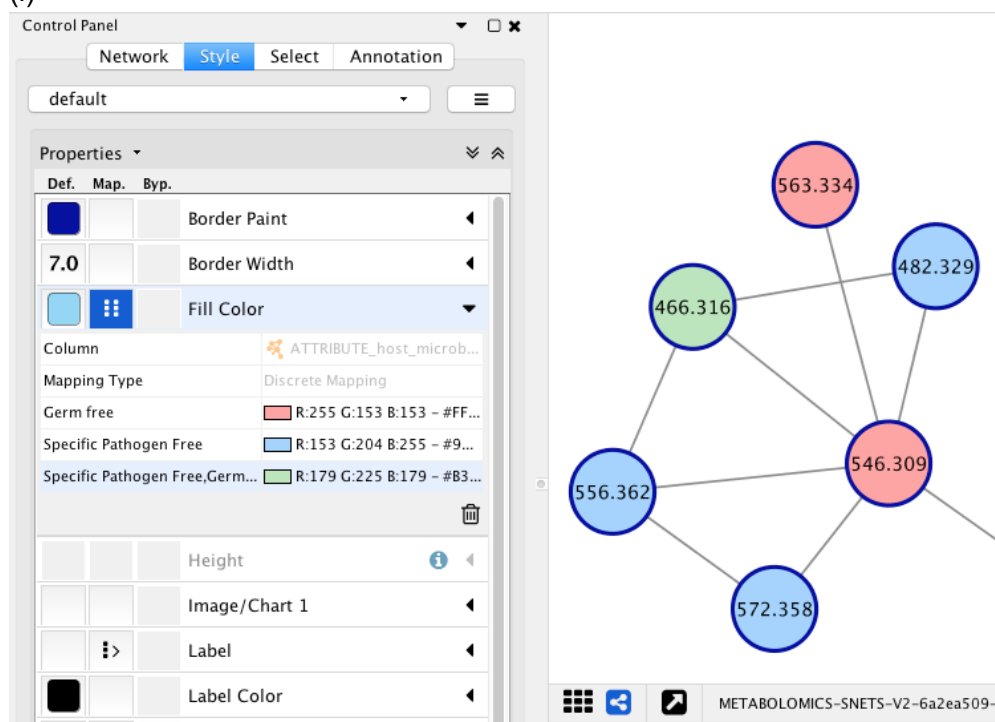
(g)



(h) Here, the size of the node and font size of the label were also increased for clearer visualization.



(i)



(j)

Network Style Select Annotation

BlueMustard

Properties

Def.	Map.	By.	
			Border Paint
15.0			Border Width
			Fill Color
			Height
			Image/Chart 1

Column: -- select value --

Mapping Type: -- select value --

Graphics

Images Charts Gradients

Available Columns:

- 23. GNPSGROUP:host-associated
- 24. GNPSGROUP:Mus musculus
- 25. GNPSGROUP:Specific Pathoge
- 26. GNPSGROUP:SPF1
- 27. GNPSGROUP:SPF2
- 28. GNPSGROUP:SPF3
- 29. GNPSGROUP:SPF4

Selected Columns:

- 1. GNPSGROUP:Germ free

Remove Graphics Cancel Apply

Graphics

Images Charts Gradients

Color Scheme: Contrasting

1 2

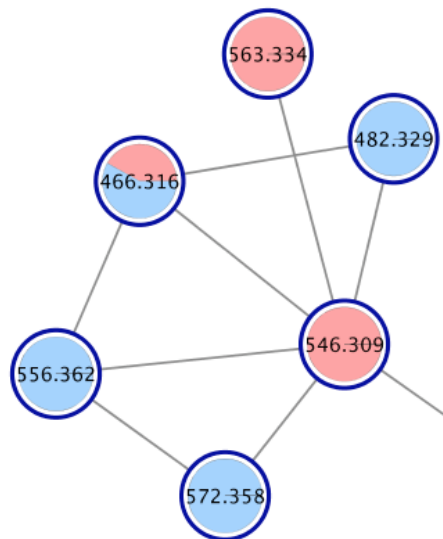
Show Value Labels: ☐ Column: -- none -- Font Size: 1

Border Width: 0.25

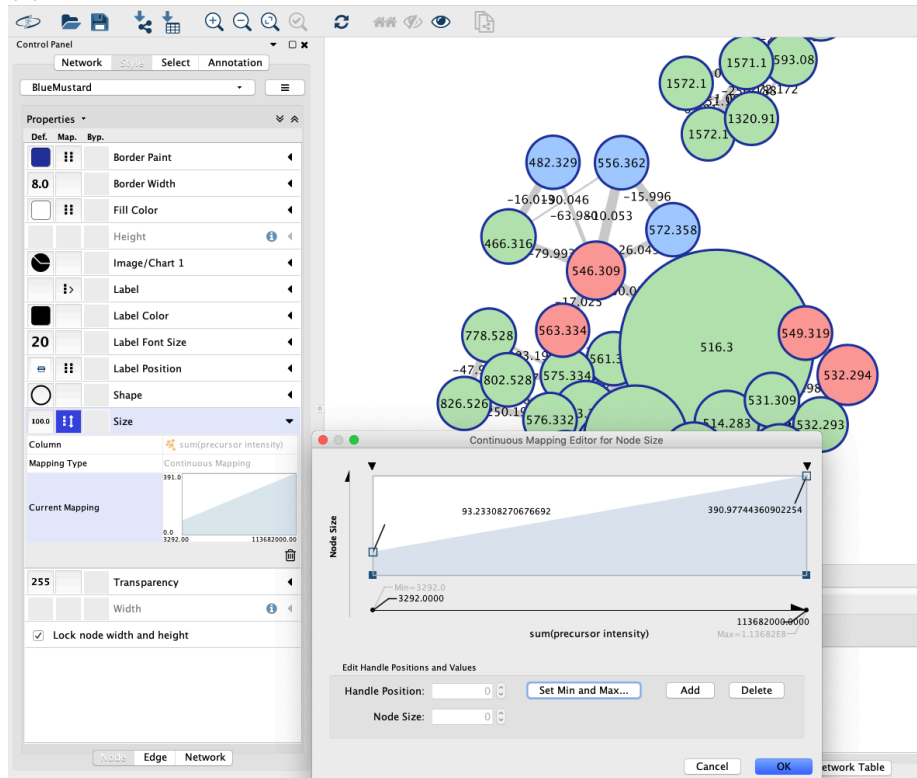
Border Color:

Start Angle (degrees): 0.0

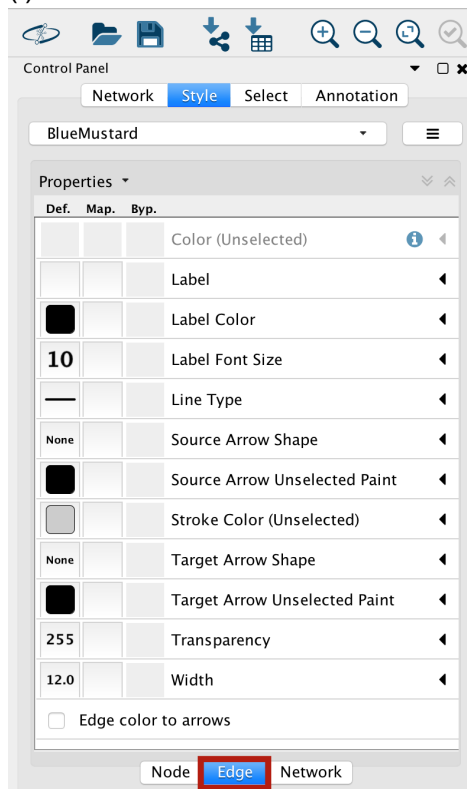
Remove Graphics Cancel Apply



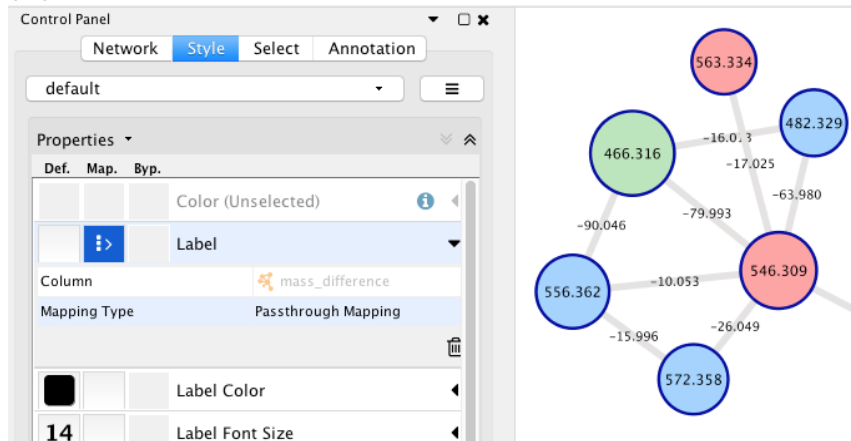
(k)



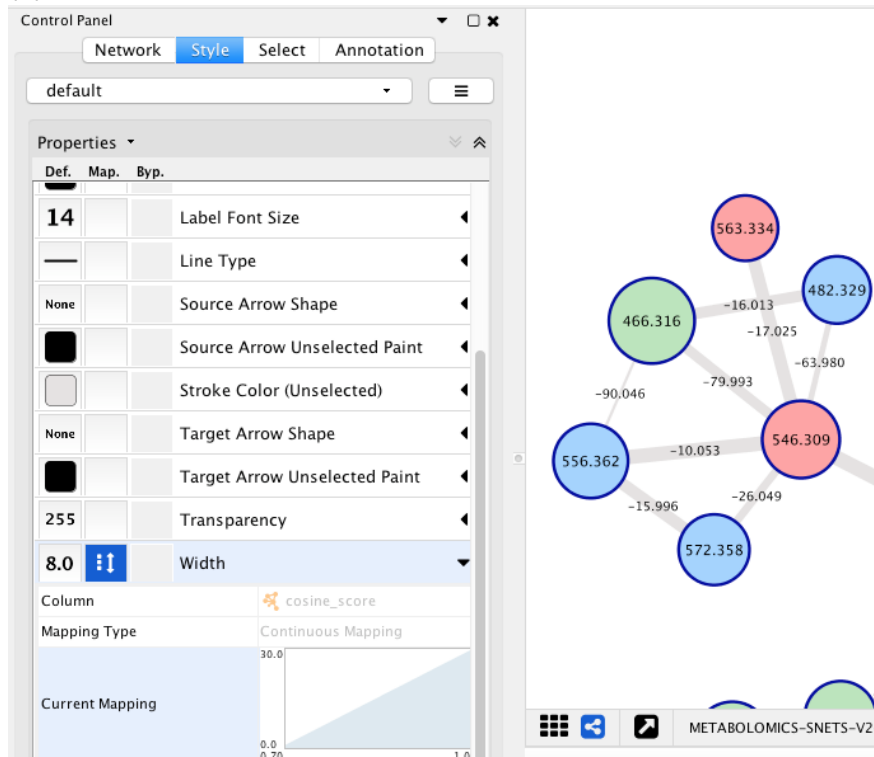
(l)



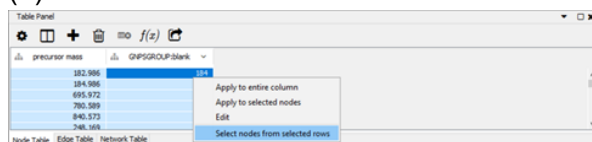
(m)



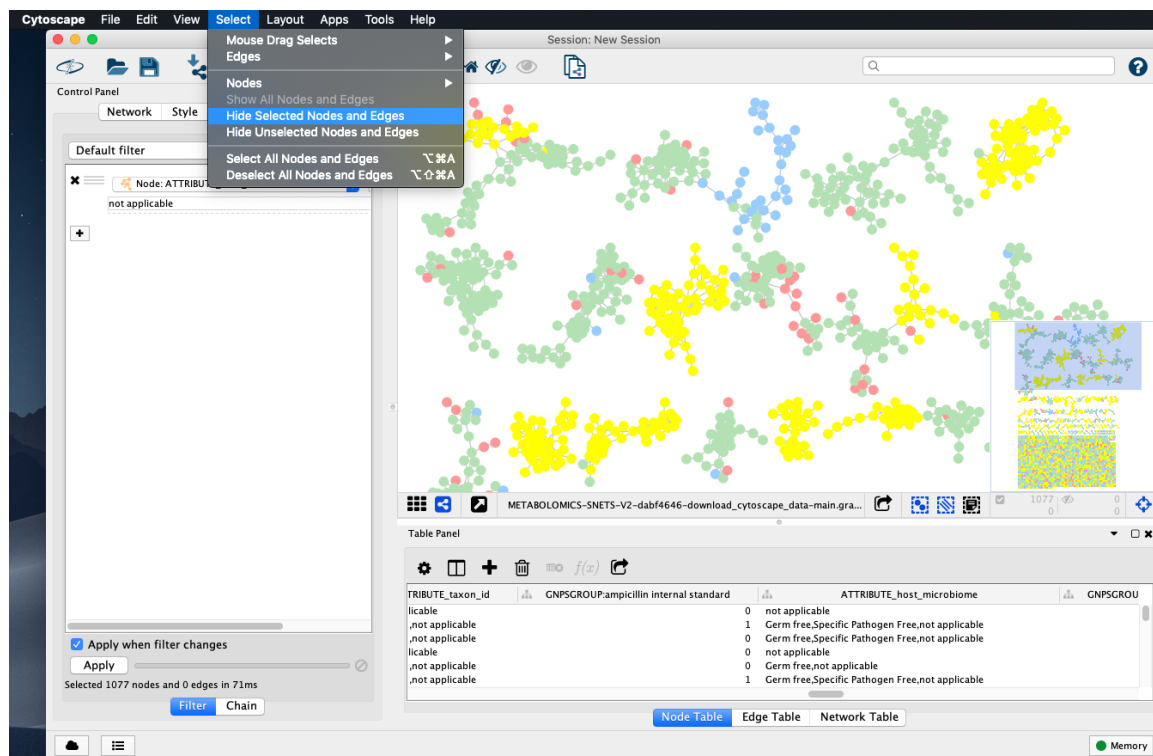
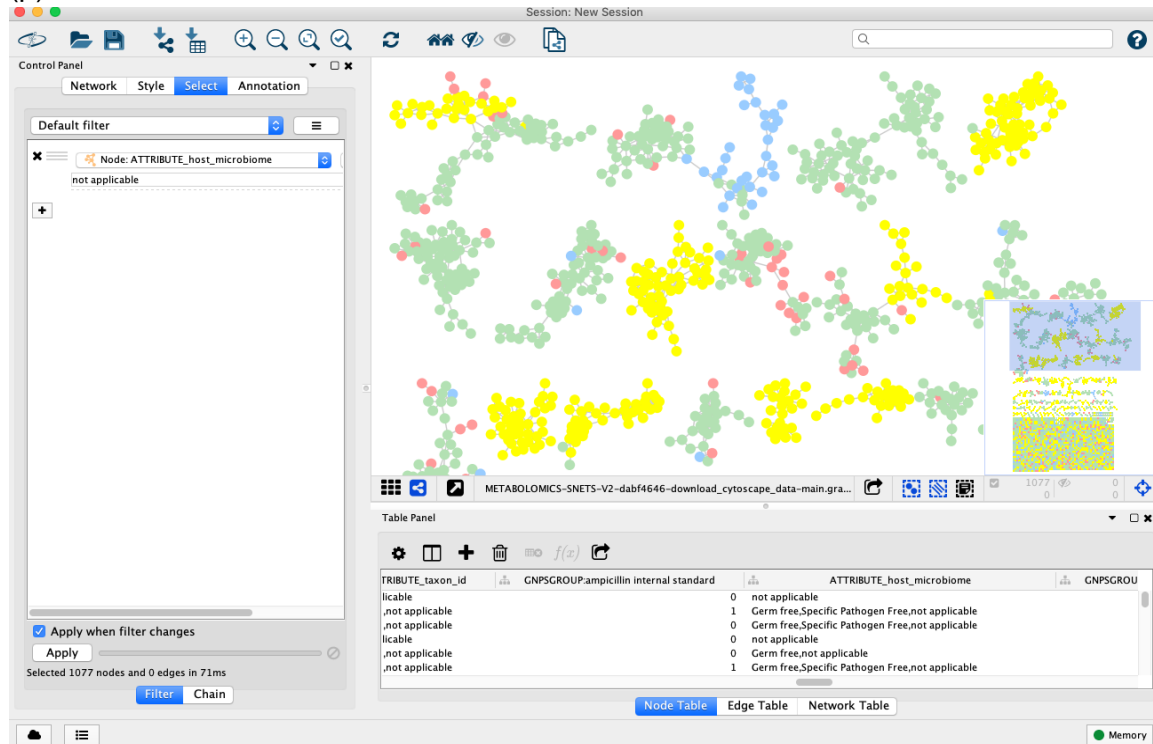
(n)

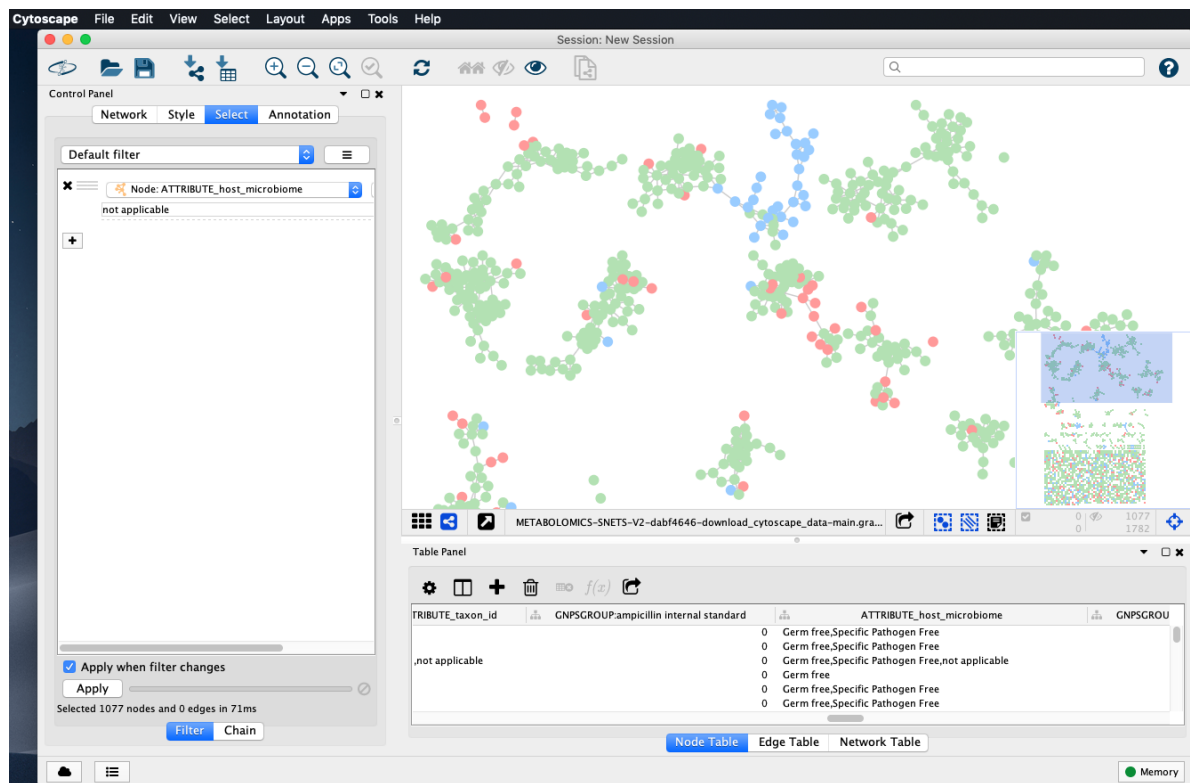


(o)

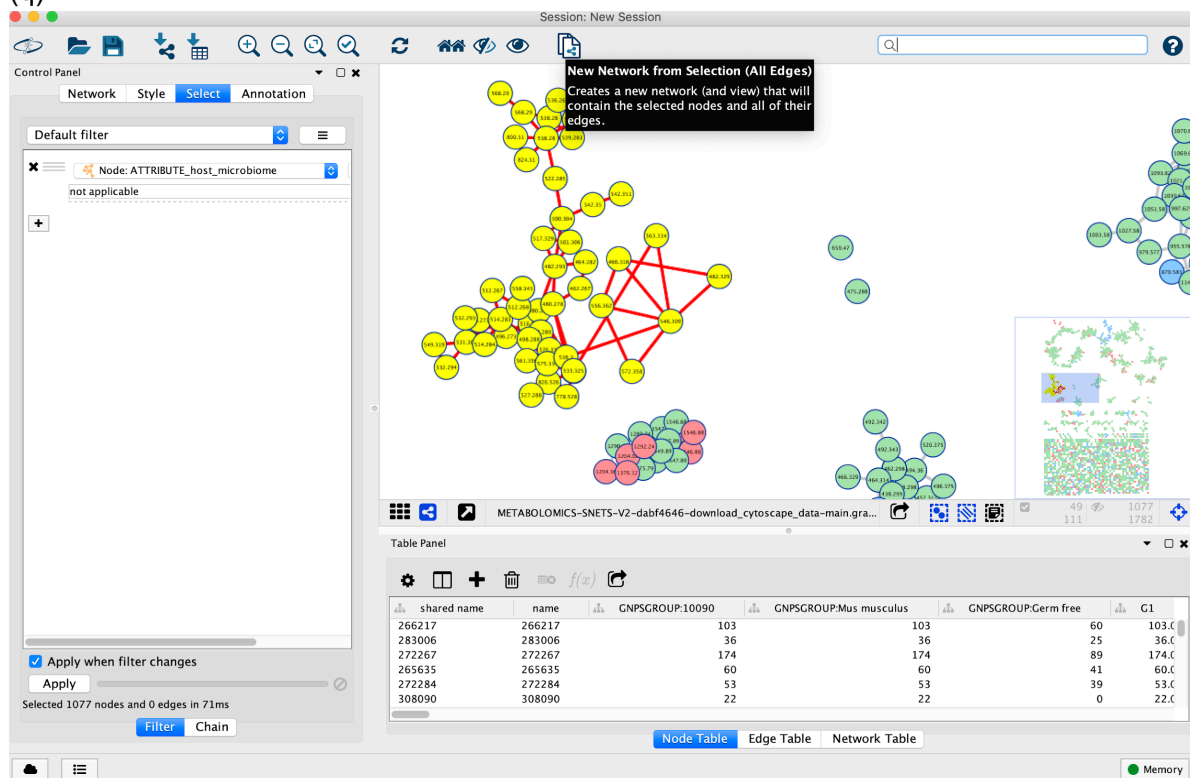


(p)



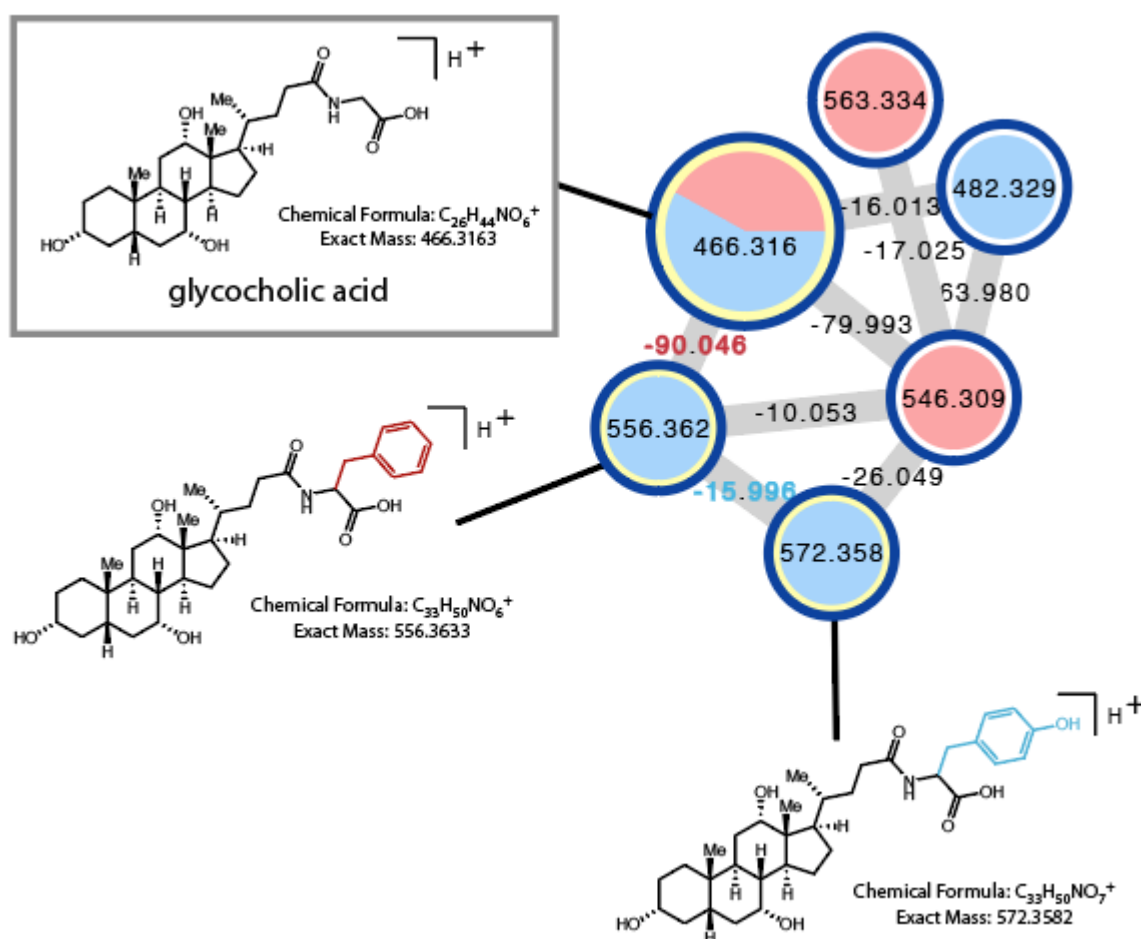


(q)

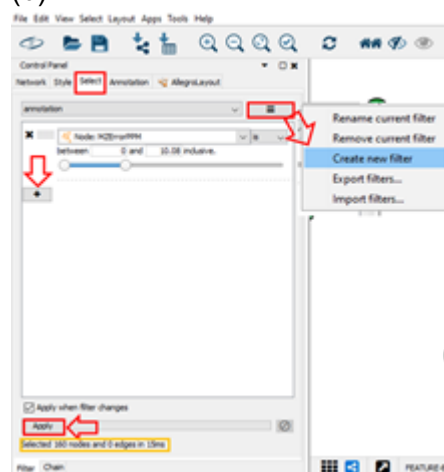


SI Figure 7. Step-by-step workflow for propagation of structural information in Cytoscape.

(a) Example of manual propagation of structural information. Glycocholic acid was putatively annotated through GNPS library matching. Based on characteristic mass shifts of 90.047, 106.042 and 56.063 neighboring nodes could be putatively annotated as glycocholic acid conjugated with phenylalanine (red), tyrosine (blue) and leucine (green) respectively.



(b)



SI Figure 8

(a) Cloning a job from a previous GNPS V2 networking workflow:

Job Status	
Workflow	METABOLOMICS-SNETS-V2
	<div> <div>DONE</div> <div>[Clone]</div> <div>[Restart][Delete]</div> </div>
Status	<p>Default Molecular Networking Results Views [View All Library Hits] View Unique Library Compounds View All Clusters With IDs]</p> <p>Network Visualizations [View Spectral Families (In Browser Network Visualizer)] Network Summarizing Graphs]</p> <p>Methods and Citation for Manuscripts [Networking Parameters and Written Network Description]</p> <p>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]</p> <p>Advanced Views - Global Public Dataset Matches [View Matches to All Public Datasets]</p> <p>Advanced Views - Third Party Visualization [View Emporer PCoA Plot in GNPS] View ili in GNPS]</p> <p>Advanced Views - Networking Graphs/Histograms [Nodes, MZ Histogram] Edges, MZ Delta Histogram Edges, Score vs MZ Delta Plot Library Search, PPM Error Histogram]</p> <p>Advanced Views - Misc Views [View Network, Node Centric] View Network Pairs Networking Statistics View Raw/Unclassified Spectra View Compounds and File Occurrence]</p> <p>Advanced Views - Make Dataset Public [Make Public Dataset Documentation]</p> <p>Advanced Views - Experimental Views [Direct Cytoscape Preview/Download] Analyze with MS2LDA]</p>

(b) Cloning a GNPS V1 job to GNPS V2 networking workflow

Job Status	
Workflow	METABOLOMICS-SNETS
	<div> <div>DONE</div> <div>[Clone]</div> <div>[Restart][Delete]</div> </div>
Status	<p>[View All Library Hits] View All Clusters With IDs View All Compounds]</p> <p>Clone to Latest Networking Workflow [Clone Job to Latest Molecular Networking V2 Workflow]</p> <p>Methods and Citation for Manuscripts [Networking Parameters and Written Network Description]</p> <p>Experimental Views [Reanalyze Cluster Spectra] Topology Signatures Topology Signatures Histogram View Raw Spectra]</p> <p>Auxiliary Views [View Network, Node Centric] View Network Pairs Networking Statistics View Compounds and File Occurrence]</p> <p>Advanced Views - Networking Graphs [Nodes, MZ Histogram] Edges, MZ Delta Histogram Edges, Score vs MZ Delta Plot Library Search, PPM Error Histogram]</p> <p>Community Matches [Dataset Matches]</p> <p>Advanced Views - Third Party Visualization [View Emporer PCoA Plot in GNPS] View ili in GNPS]</p> <p>Network Visualizations [View Spectral Families (In Browser Network Visualizer)] Network Summarizing Graphs]</p> <p>Export [Download Clustered Spectra as MGF] Download Cytoscape Data Download Bucket Table Download BioM For Qiime/Qiita Download Metadata For Qiime Make Public Dataset Download ili Data]</p>

Supplementary Methods

Instrument parameters for AB Sciex mass spectrometer

For the AB Sciex TripleTOF 5600+ mass spectrometer data can be acquired in the positive or negative ion mode using ESI parameters as follows:

SI Table 1. AB Sciex parameters

Ion Source Parameter	Value
Ion source temp	450 °C
Curtain gas	35 L/h
GS1	50 L/h
GS2	40 L/min
Capillary voltage	5200 V

Data are acquired from MS¹ survey over ranges from m/z 100 to 1500 or 2000 Da (at 0.25 s/spectra) if there is an expectation that there are many large natural products (e.g., many lipopeptides, oligosaccharides and ribosomally encoded peptides with m/z >1500) in information dependent acquisition (IDA) mode (high sensitivity mode), automatically switching between full-scan MS and MS². The suggested IDA settings are: charge state +1 to +2, intensity 150 cps, exclude isotopes within 6 Da, mass tolerance 50 mDa, and a maximum number of candidate ions 20 (over the m/z range of 50-1500/2000 Da). Under IDA settings, select "exclude former target ions" to 15-20 s after two occurrences. In advanced settings, "Dynamic background subtraction" is selected. Importantly, collision energy (CE) is ramped, with a collision energy spread (CES) ranging from 21 – 53 eV, enabling more complete fragmentation than a single CE value affords. To ensure a mass accuracy of <5 ppm for both MS¹ and MS² the instrument should continuously be calibrated (e.g., using the batch mode or calibration delivery system [CDS]) every 3-4 working hours using appropriate positive or negative calibration solutions (or in-house calibrants).

SI Table 2. Additional AB Sciex parameters

For survey ion scans:		For dependent ion scans:	
Parameter	Value	Parameter	Value
CE	10	CE	37

DP	80	CES	16
IDlx	0	DP	80
IDUx	5	IDlx	0
IRDx	15000	IDUx	5
IRWx	10000	IRD	66.633
IWlx	0	IRDx	15000
IWUx	5	IRW	24.917
XA1	70.16	IRWx	10000
		IWlx	0
		IWUx	5
		XA1	70.16

Instrument parameters for *Agilent* mass spectrometer

For the *Agilent 6530 QTOF mass spectrometer* using MassHunter Data Acquisition software, data can be acquired in either positive or negative ion mode (F. Olivon, Roussi, Litaudon, & Touboul, 2017), for example, with ESI parameters as follows:

SI Table 3. Agilent parameters

Ion Source Parameter	Value
Gas temp	270 °C
Drying gas	10 L/min
Nebulizer	30 psig
Sheath gas temp	300 °C
Sheath gas flow	7 L/min

Capillary voltage	3500 V
Nozzle voltage	500 V
Fragmentor	140 V
Skimmer	45 V

The Auto-MS² (AMS) algorithm in the MassHunter software may be used with an MS range of m/z 100-1500 and an MS² range of m/z 50-1500, both at 5 spectra/s. The medium ($\sim m/z$ 4) isolation width is used unless significant numbers of halogen atoms are expected per molecule in which case the wide ($\sim m/z$ 9) isolation width can be used. The collision energy gradient should be set to be automatic according to m/z values of precursor ions, and the values should be optimized considering the chemical properties of analytes. For example, under the collision energy tab, use Formula with two lines, charge set to "All", "Slope" is 2.6 and "Offset" is 14.75, and charge set to "All", "Slope" is 3.9 and "Offset" is 22.13. The maximum precursors per cycle is set to 5, with the Absolute Precursor Threshold set to 15000 (Relative Threshold 0.015%) and Active Exclusion enabled; ions are excluded after 3 scans and released after 0.3 min. Additionally, under the Precursor II tab, Isotope model is set to "Common Organic Molecules" with the Active Precursor Charge state set to "1". If using reference solution HP-921, under the Ref Mass tab, enable "Reference Mass Correction" and the appropriate reference solution source bottle. Set "Auto Recalibration" with a detection window of 100 ppm and a minimum height of 3000 counts, and select m/z 121.0509 and m/z 922.0098 (positive ion mode) or m/z 112.9856 and m/z 1033.9881 (negative ion mode). These calibrant masses need to be listed for exclusion under the Preferred/Exclude tab.

Instrument parameters for *Bruker Daltonics* mass spectrometer

For the *Bruker Daltonics Maxis QTOF* mass spectrometer, spectra can be acquired in the positive and/or negative ionization mode with the following parameters:

SI Table 4. Bruker parameters

Ion Source Parameter			Value
Capillary Voltage			4500 V
Ion Source Temperature			200 °C
Dry Gas Flow			9 L/min
Nebulizer Pressure	Gas	(nitrogen)	2 bar

Data are acquired from an MS¹ survey over m/z 100 - 2,000 in data-dependent acquisition mode, where MS² active exclusion is set after 2 spectra and released after 30 s. Up to five scans in CID mode are acquired per cycle, where the top five most abundant MS1 ions are selected for MS² fragmentation. If current intensity/previous intensity = 5.0, reconsider precursor. Ramped collision energies are utilized to fragment ions, ranging from 18 to 50 eV, as shown in the table below:

SI Table 5. Additional Bruker parameters

Isolation mass value (m/z)	Charge state value (z)	Isolation width value (m/z)	Collision energy value (eV)
100	1	4	22
300	1	5	27
500	1	6	35
1000	1	8	45
2000	1	10	50
100	2	4	18
300	2	5	22
500	2	6	30
1000	2	8	35
2000	2	10	50

Instrument parameters for Shimadzu mass spectrometer

For the *Shimadzu IT-TOF mass spectrometer*, spectra can be acquired in the positive or negative ionization mode using the following parameters:

SI Table 6. Shimadzu parameters

Ion Source Parameter	Value
Capillary Voltage	Relative to the Tuning results

CDL Temperature	200–250 °C
Dry Gas	On
Nebulizer Gas (nitrogen) flow	1.5 L/min
Ion accumulation (ESI+)	15 ms
Ion accumulation (ESI-)	5 ms

Data is acquired from MS¹ survey over m/z 100 – 1,000 in data-dependent acquisition mode. The most abundant MS¹ ion in the m/z 150 – 1,000 Da range is selected for fragmentation and one MS² is acquired per cycle. The MS² execution trigger is set to 100,000 count (depending on the noise level observed on the instrument and the desired exhaustivity of MS² peak fragmentations) with a dynamic exclusion period of 3 s along with isotope exclusion (additionally, blank-related ions are added to a static exclusion list). Precursor ion isolation is performed on a 3 Da width with ion accumulation during 30 ms and 10 ms for ESI+ and ESI-, respectively. CID parameters are set to 50% energy, 50% collision gas and 0.251(45.0KHz) for q(Frequency).

Centroiding Shimadzu Data

For conversion of Shimadzu IT-TOF mass spectrometry data to a GNPS-compatible format, LcSolution (3.6) should be used for complete data export as .mzXML format. As data are exported in profile mode, it is important to convert to centroid mode using MSConvert (ProteoWizard) [Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 2008, 24 (21), 2534-2536]. Using these programs, data export can be achieved as batch or single files; data conversion can also be performed using both modes. Step-by-step explanation is provided as follows:

- 1) In LcSolution/Postrun, open any .lcd file.
- 2) Calibrate the file using 'Qualitative/Postrun calibration' if needed
- 3) to export the file as mzXML format, right-click on the file and select 'File Conversion/LcSolution Data file to mzXML'. In the pop-up window, select the target folder for export and start the conversion.
- 4) Open MSconvert (ProteoWizard).
- 5) Add the file(s) to be converted to centroid mode
- 6) Add the centroid filter, by selecting "Peak picking" and providing appropriate parameters as follows: MS Level = 1-2, Signal-to-noise = 0 and PeakSpace = 0.
- 7) select the appropriate file format as mzXML or .MGF (see above)
- 8) Start the conversion.

Instrument parameters for *Thermo* mass spectrometer

For the *Thermo Q-Exactive* mass spectrometer, ESI parameters can be the following:

SI Table 7. Thermo parameters

Ion Source Parameter	Value
Sheath gas flow	52 U
Auxiliary gas flow	14 U
Auxiliary gas temperature	400 °C
Sweep gas flow	0 U
Spray voltage	3.5 kV
Inlet capillary temperature	320 °C

Data are acquired from an MS¹ survey over m/z 100 - 1,500 at 35,000 resolution in data-dependent acquisition mode, where up to five MS² scans in CID/HCD mode are acquired per cycle (where the TOP5 most abundant MS¹ ions can be selected) with 17,500 resolution and one microscan in positive or negative mode. The maximum injection time should be set to 100 ms, and the MS² precursor window should be set to m/z 1.5 (offset m/z 0.5). S-lens radio frequency (RF) level should be set to 50 V, and apex trigger minimum (5 s) and apex trigger maximum (10 s). The normalized collision energy should be set to *stepped energy mode* where normalized collision energies (NCEs) from 20, 30 and 40% are combined, with the default charge state set as $z=1$. MS² experiments should be set to be automatically triggered within 2–15 s of their first occurrence at the apex of a peak. The dynamic exclusion time should be set to a value representing just over half the peak width (typically between 5 and 15 s), and the apex triggering mode should be enabled in the dd settings. Ion species with unassigned charge states and isotope peaks should be excluded. Note that the MS² settings described above are highly sample dependent, and we recommend optimizing based on sample type. For short gradients and UHPLC, fast duty cycles (lower resolution and lower number of MS² scans) will most likely be desired. For more complex samples, such as dissolved organic matter or other environmental samples on the other hand, high resolutions especially on MS¹ level might be important to resolve almost isobaric compounds and will increase duty cycle time. However, the cycle time should be adjusted together with gradient length and duty cycle time in order to archive enough data points (i.e. scans) to represent the chromatographic peak shape appropriately.

Instrument parameters for Waters mass spectrometer

For the Waters Xevo QTOF mass spectrometer, ESI parameters can be the following:

SI Table 8. Waters parameters

Ion Source Parameter	Value
Acquisition mode	Resolution

Capillary voltage	2.5 kV
Cone voltage	20 V
Source temperature	120 °C
Desolvation temperature	350 °C
Cone gas flow	50 L/h
Desolvation gas flow	800 L/h

Data are acquired from MS¹ survey over m/z 100 - 1,500 with the scan time of 150 ms in data-dependent acquisition mode, and up to five MS² scans in CID mode are acquired per cycle (where the TOP 5 most abundant MS¹ ions can be selected) with 100 ms scan time in positive or negative mode. MS survey is set to switch to MS2 acquisition when TIC rises above, and MS² acquisition switch back to MS survey after 0.25 s have elapsed. If it is desired, the de-isotope peak selection filter can be enabled with a window of 3.0 Da (this depends on sample properties and is especially important if samples contain intended isotopic chemicals). The collision energy gradient should be set to be automatic according to m/z values of precursor ions, and the values should be optimized considering the chemical properties of analytes. Gradient parameters for general samples can be set as follows: 20 to 40 V for 100 Da to 60 to 80 V for 1500 Da.

“Easy” Workflow for batch conversion from vendor format to GNPS-compatible format

Windows users can convert files from vendor formats to GNPS compatible format (mzXML) by simply placing files into a folder and batch converting with minimal installation.

- 1) The zip file can be downloaded from the following link: <https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/>, the the contents can be unzipped into a folder present on the Desktop.
- 2) Windows libraries in "pwizLibraries-and-Installation" can be installed then the appropriate program for 32-bit (32-Bit_Double-Click_To_Install.bat) or 64-bit system (64-Bit_Double-Click_To_Install.bat) can be run.
- 3) Vendor formats can be placed in in "Input_Files" not embedded in other folders.
- 4) Double-Click_To-Convert.bit can be double clicked on; download zip includes demo files for major vendor formats as a test. (This will take a minute.)
- 5) All converted files will be present in Output_Folder. If there are errors, log.txt can be inspected or the “traditional” method (described in the main text) can be performed.

Conversion of Bruker mass spectrometry data to GNPS-compatible format.

The conversion can be performed as follows:

DataAnalysis and CompassXport programs must be installed. Both software packages are available to download for free from the Bruker website (www.bruker.com/service/support-upgrades/software-downloads/mass-spectrometry.html) and are only compatible with

Windows operating systems. Additionally, Insilicos (www.insilicos.com), a free viewer for mass spectrometry data, should be downloaded to open the converted .mzXML files. Using these programs, data conversion and export can be performed on single files or on large batch datasets.

Before data conversion, the user must first apply lock mass calibration to the data, if required. To streamline this process, a custom method is written to automate lock mass calibration and data conversion:

- 1) In DataAnalysis, open any .d file.
- 2) Under 'Calibrate', click on 'parameters'.
- 3) In the pop-up window, select the 'mass list' tab and ensure that the 'peak finder' is set to 'sum peak'.
- 4) In the same window, select the 'calibration' tab then 'lock mass calibration' and choose 'ESI' from the 'calibration group' drop-down menu, then click 'Edit'.
- 5) The 'Compass Reference Mass List Editor' window should open. To create a reference mass list, the name, ion formula, and charge of each reference compound should be added (Figure SXX).
- 6) Click 'Save' and name the reference list file appropriately, ensuring that the file extension is .ref (e.g. LockMass Tuning Mix ESI-TOF (922 Da).ref)
- 7) Once the reference list is saved, close the window.
- 8) In the *Method parameters* window, select the new reference mass list in the 'Calibration List' drop-down menu and change the 'intensity threshold' to 500 and click 'Apply' then 'OK'.
- 9) Under 'Method', select 'Script'.
- 10) Type the following exactly as written:

```
option explicit
Analysis.ApplyLockMassCalibration true
Analysis.Export "C:\Users\username\userfolder", daMzXML, daLine
Form.close
```
- 11) Under 'Method', select 'Save As' and save the new method with a descriptive name in an appropriate folder.

To export files as 32-bit spectra, the Registry on the user's computer needs to be edited as follows:

- 12) In Registry Editor, locate the folder 'HKEY_CURRENT_USER\Software\Bruker Daltonik\CompassXport'
- 13) Open 'ExportPrecision64bit' and change the value data to '0' to specify that .mzXML will be exported with 32-bit precision, then select OK
- 14) Open 'UseRecalibratedSpectra' and ensure that the value data is set to '1' to export recalibrated spectra acquired after lock mass calibration, then press OK.

To convert a large batch dataset from .d files to .mzXML files:

- 15) Under 'Tools', choose 'Process with method'.
- 16) Two windows should open ('Compass AutomationEngine' and 'Compass DataAnalysis ProcessQueuer')
- 17) In the 'Compass AutomationEngine', click 'Method' and select the method that was saved in the steps above.

- 18) Click 'Select' and choose the raw .d files to be processed.
- 19) Once the desired files are in the analysis list, highlight the file names and click 'Process'.
- 20) The file names will move into the 'Compass DataAnalysis ProcessQueuer'. Large batch file conversion takes a significant amount of time and processing power. Therefore, it is recommended to run it overnight and avoid using DataAnalysis in the meantime.
- 21) Once complete, the converted .mzXML files will appear in the folder specified in the script above. Open the files in Insilicos to ensure that they were properly converted.

Supplementary Tables

SI Table 9: Metadata table format (Standard format)

Table from documentation <https://ccms-ucsd.github.io/GNPSDocumentation/networking/>

SI Table 10: Metadata table format (Format for ReDU-MS2)

Table from documentation <http://dorresteinappshub.ucsd.edu:5005/addmetadata>

SI Table 11: Speclibs libraries in GNPS used for examples (January 2019,

<https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp>)

speclibs/MIADB/MIADB.mgf
 speclibs/GNPS-NIH-NATURALPRODUCTSLIBRARY_ROUND2_POSITIVE/GNPS-NIH-NATURALPRODUCTSLIBRARY_ROUND2_POSITIVE.mgf
 speclibs/GNPS-COLLECTIONS-PESTICIDES-NEGATIVE/GNPS-COLLECTIONS-PESTICIDES-NEGATIVE.mgf
 speclibs/MASSBANK/MASSBANK.mgf
 speclibs/GNPS-NIH-CLINICALCOLLECTION1/GNPS-NIH-CLINICALCOLLECTION1.mgf
 speclibs/HMDB/HMDB.mgf
 speclibs/DEREPLICATOR_IDENTIFIED_LIBRARY/DEREPLICATOR_IDENTIFIED_LIBRARY.mgf
 speclibs/GNPS-SELLECKCHEM-FDA-PART1/GNPS-SELLECKCHEM-FDA-PART1.mgf
 speclibs/RESPECT/RESPECT.mgf
 speclibs/GNPS-FAULKNERLEGACY/GNPS-FAULKNERLEGACY.mgf
 speclibs/MMV_NEGATIVE/MMV_NEGATIVE.mgf
 speclibs/MASSBANKEU/MASSBANKEU.mgf
 speclibs/GNPS-LIBRARY/GNPS-LIBRARY.mgf
 speclibs/GNPS-PRESTWICKPHYTOCHEM/GNPS-PRESTWICKPHYTOCHEM.mgf
 speclibs/GNPS-EMBL-MCF/GNPS-EMBL-MCF.mgf
 speclibs/GNPS-NIH-CLINICALCOLLECTION2/GNPS-NIH-CLINICALCOLLECTION2.mgf
 speclibs/GNPS-NIH-SMALLMOLECULEPHARMACOLOGICALLYACTIVE/GNPS-NIH-SMALLMOLECULEPHARMACOLOGICALLYACTIVE.mgf
 speclibs/GNPS-SELLECKCHEM-FDA-PART2/GNPS-SELLECKCHEM-FDA-PART2.mgf
 speclibs/GNPS-COLLECTIONS-PESTICIDES-POSITIVE/GNPS-COLLECTIONS-PESTICIDES-POSITIVE.mgf
 speclibs/GNPS-NIH-NATURALPRODUCTSLIBRARY_ROUND2_NEGATIVE/GNPS-NIH-NATURALPRODUCTSLIBRARY_ROUND2_NEGATIVE.mgf

speclibs/PNNL-LIPIDS/PNNL-LIPIDS-
 POSITIVE.mgf
 speclibs/PNNL-LIPIDS/PNNL-LIPIDS-NEGATIVE.mgf
 speclibs/GNPS-COLLECTIONS-MISC/GNPS-COLLECTIONS-MISC.mgf
 speclibs/GNPS-NIH-NATURALPRODUCTSLIBRARY/GNPS-NIH-
 NATURALPRODUCTSLIBRARY.mgf
 speclibs/CASMI/CASMI.mgf
 speclibs/MMV_POSITIVE/MMV_POSITIVE.mgf
 speclibs/MONA/MONA.mgf
 speclibs/SUMNER/SUMNER.mgf
 speclibs/GNPS-NIST14-MATCHES/GNPS-NIST14-MATCHES.mgf

SI Table 12: Field requirements for batch annotation uploading (as referred to in Section 3.5)

Header	Desc.	Default Value To Put When Not Known	Possible Values	Required
FILENAME	Input Filename (mzXML)	REQUIRED	only letters numbers underscores and periods are allowed	Yes
SEQ	Peptide Sequence	*..*		No
COMPOUND_NAME	Compound Common Name	REQUIRED		Yes
MOLECULEMASS	Corrected Precursor MZ for Compound	0		No
INSTRUMENT	Mass Analyzer Instrument	REQUIRED	qTof, QQQ, Ion Trap, Hybrid FT, Orbitrap	Yes
IONSOURCE	Source of Ions	REQUIRED	LC-ESI, DI-ESI	Yes
EXTRACTSCAN	Spectrum Scan of Spectrum	REQUIRED		Yes
SMILES	Smiles Structure	N/A		No
INCHI	Inchi Structure	N/A		No
INCHIAUX	Inchi Auxiliary Structure	N/A		No
CHARGE	Charge of Precursor (0 to pull from spectrum file)	0		No
IONMODE	Ionization Mode	Positive	Positive, Negative	Yes

	(Positive/Negative)			
PUBMED	Pubmed ID where compound or data was published	N/A		No
ACQUISITION	Sample source of compound	Crude	Crude, Lysate, Commercial, Isolated, Other	Yes
EXACTMASS	Compound Exact Mass	0		No
DATACOLLECTOR	Individual collecting data	REQUIRED		Yes
ADDUCT	Adduct of Ion fragmented in MS2 (M+H, M+2H, etc.)	REQUIRED		Yes
INTEREST	N/A	N/A		No
LIBQUALITY	Quality of Library (1 for Gold, 2 for Silver, 3 for Bronze)	3 rd	1,2,3	Yes
GENUS	Genus	N/A		No
SPECIES	Species	N/A		No
STRAIN	Strain	N/A		No
CASNUMBER	Chemical Abstracts Service compound identification number	N/A		No
PI	PI overseeing compound acquisition and analysis	REQUIRED		Yes

SI_Document_20190806_submission.pdf (16.11 MiB)

[view on ChemRxiv](#) • [download file](#)
