

Feature-based Molecular Networking in the GNPS Analysis Environment

Louis Felix Nothias,^{1,2,#} Daniel Petras,^{1,2,3,#} Robin Schmid,⁴ Kai Dührkop,⁵ Johannes Rainer,⁶
Abinesh Sarvepalli,^{1,2} Ivan Protsyuk,⁷ Madeleine Ernst,^{1,2,8} Hiroshi Tsugawa,^{9,10} Markus
Fleischauer,⁵ Fabian Aicheler,^{11,12} Alexander Aksenov,^{1,2} Oliver Alka,^{11,12} Pierre-Marie Allard,¹³
Aiko Barsch,¹⁴ Xavier Cachet,¹⁵ Mauricio Caraballo,^{1,2} Ricardo R. Da Silva,^{2,16} Tam Dang,^{2,17}
Neha Garg,¹⁸ Julia M. Gauglitz,^{1,2} Alexey Gurevich,¹⁹ Giorgis Isaac,²⁰ Alan K. Jarmusch,^{1,2}
Zdeněk Kameník,²¹ Kyo Bin Kang,^{1,2,22} Nikolas Kessler,¹⁴ Irina Koester,^{1,2,3} Ansgar Korf,⁴
Audrey Le Gouellec,²³ Marcus Ludwig,⁵ Christian Martin H.,²⁴ Laura-Isobel McCall,²⁵ Jonathan
McSayles,²⁶ Sven W. Meyer,¹⁴ Hosein Mohimani,²⁷ Mustafa Morsy,²⁸ Oriane Moyne,^{23,29} Steffen
Neumann,^{30,31} Heiko Neuweiger,¹⁴ Ngoc Hung Nguyen,^{1,2} Melissa Nothias-Esposito,^{1,2} Julien
Paolini,³² Vanessa V. Phelan,³³ Tomáš Pluskal,³⁴ Robert A. Quinn,³⁵ Simon Rogers,³⁶ Bindesh
Shrestha,¹⁹ Anupriya Tripathi,^{1,29,37} Justin J.J. van der Hoof,^{1,2,38} Fernando Vargas,^{1,2} Kelly C.
Weldon,^{1,2,39} Michael Witting,⁴⁰ Heejung Yang,⁴¹ Zheng Zhang,^{1,2} Florian Zubeil,¹⁴ Oliver
Kohlbacher,^{11,12,42,43} Sebastian Böcker,⁵ Theodore Alexandrov,^{1,2,7} Nuno Bandeira,^{1,2,44} Mingxun
Wang,^{1,2,44,*} and Pieter C. Dorrestein^{1,2,29,39,*}

1. Skaggs of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, San Diego, CA, USA
2. Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, San Diego, CA, USA
3. Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA
4. Institute of Inorganic and Analytical Chemistry, University of Münster, Münster, Germany
5. Chair for Bioinformatics, Friedrich-Schiller-University Jena, Jena, Germany
6. Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy
7. Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
8. Center for Newborn Screening, Department of Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark
9. RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan
10. RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan
11. Applied Bioinformatics, Department of Computer Science, University of Tübingen, Tübingen, Germany
12. Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen, Germany
13. Department of Phytochemistry and Bioactive Natural Products, University of Geneva, Geneva, Switzerland
14. Bruker Daltonics, Bremen, Germany

- 1 15. Equipe PNAS, UMR 8038 CiTCoM CNRS, Faculté de Pharmacie de Paris, Université
2 Paris Descartes, Paris, France
- 3 16. Department of Physics and Chemistry, School of Pharmaceutical Sciences of Ribeirão
4 Preto, University of São Paulo, Ribeirão Preto, Brazil
- 5 17. Technische Universität Berlin, Faculty II Mathematics and Natural Sciences, Institute of
6 Chemistry, Berlin, Germany
- 7 18. School of Chemistry and Biochemistry, Center for Microbial Dynamics and Infection,
8 Georgia Institute of Technology, Atlanta, GA, USA
- 9 19. Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St.
10 Petersburg State University, St. Petersburg, Russia
- 11 20. Waters Corporation, Milford, MA, USA
- 12 21. Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic
- 13 22. College of Pharmacy, Sookmyung Women's University, Seoul, Republic of Korea
- 14 23. Univ. Grenoble Alpes, CNRS, Grenoble INP, CHU Grenoble Alpes, TIMC-IMAG,
15 Grenoble, France
- 16 24. Centro de Biodiversidad y Descubrimiento de Drogas, INDICASAT AIP, Panama,
17 Republic of Panama
- 18 25. Department of Chemistry and Biochemistry, Department of Microbiology and Plant
19 Biology and Laboratories of Molecular Anthropology and Microbiome Research,
20 University of Oklahoma, USA
- 21 26. Nonlinear Dynamics, Milford, MA, USA
- 22 27. Computational Biology Department, School of Computer Sciences, Carnegie Mellon
23 University, Pittsburgh, Pennsylvania, USA
- 24 28. Department of Biological and Environmental Sciences, University of West Alabama,
25 Livingston, USA
- 26 29. Department of Pediatrics, University of California San Diego, La Jolla, San Diego, CA,
27 USA
- 28 30. Bioinformatics and Scientific Data, Leibniz Institute of Plant Biochemistry, Halle,
29 Germany
- 30 31. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,
31 Germany
- 32 32. Laboratoire de Chimie des Produits Naturels, UMR CNRS SPE, Université de Corse
33 Pascal Paoli, France
- 34 33. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado,
35 Denver, Aurora, CO, USA
- 36 34. Whitehead Institute for Biomedical Research, Cambridge, MA, USA
- 37 35. Department of Biochemistry and Molecular Biology, Michigan State University, East
38 Lansing, 48823, MI, USA
- 39 36. School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK
- 40 37. Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA
- 41 38. Bioinformatics Group, Wageningen University, Wageningen, the Netherlands
- 42 39. Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA,
43 USA
- 44 40. Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München
- 45 41. College of Pharmacy, Kangwon National University, Republic of Korea

- 1 42. Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen,
- 2 Germany
- 3 43. Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen,
- 4 Germany
- 5 44. Department of Computer Science and Engineering, University of California San Diego,
- 6 CA, USA
- 7
- 8

1 Abstract

2 Molecular networking has become a key method used to visualize and annotate the chemical space in
3 non-targeted mass spectrometry-based experiments. However, distinguishing isomeric compounds and
4 quantitative interpretation are currently limited. Therefore, we created Feature-based Molecular
5 Networking (FBMN) as a new analysis method in the Global Natural Products Social Molecular
6 Networking (GNPS) infrastructure. FBMN leverages feature detection and alignment tools to enhance
7 quantitative analyses and isomer distinction, including from ion-mobility spectrometry experiments, in
8 molecular networks.

9

10 Main text

11 Introduced in 2012¹, molecular networking has become an essential bioinformatics tool to visualize and
12 annotate non-targeted mass spectrometry data^{2,3}. The first application of molecular networking was
13 described by Traxler and Kolter⁴ as holding “*great promise in providing the next quantum leap in*
14 *understanding the fascinating world of microbial chemical ecology*”. Molecular networking goes beyond
15 spectral matching against reference spectra, by aligning experimental spectra against one another and
16 connecting related molecules by their spectral similarity⁵. In a molecular network, related molecules are
17 referred to as a “molecular family”, differing by simple transformations such as glycosylation, alkylation,
18 and oxidation/reduction. Molecular networking became publicly accessible in 2013 through the initial
19 release of the Global Natural Product Social Molecular Networking (GNPS), a web-enabled mass
20 spectrometry knowledge capture and analysis platform (<http://gnps.ucsd.edu>)⁶, and has been widely
21 applied in mass spectrometry-based metabolomics to aid in the annotation of molecular families from
22 their fragmentation spectra (MS²).

23 Powered by 3,000+ CPU cores at the Center for Computational Mass Spectrometry at the
24 University of California San Diego and the MassIVE data repository, GNPS has provided researchers
25 from more than 150 countries with the ability to perform molecular networking. To build upon the
26 success of the first molecular networking method (referred to as “classical” molecular networking,
27 classical MN) which is based on the MS-Cluster algorithm⁷, we introduce a complementary tool named
28 Feature-based Molecular Networking (FBMN). FBMN accepts the outputs of well-established mass
29 spectrometry processing software and improves upon classical MN by incorporating MS¹ information,
30 such as isotope patterns and retention time, but also ion-mobility separation when performed. As a result,
31 molecular networks obtained with FBMN can distinguish isomers that may have remained hidden,
32 facilitates spectral annotation, and incorporates relative quantitative information which enables robust
33 downstream metabolomics statistical analysis. Whereas users of the classical molecular networking would
34 have had to perform molecular networking and MS¹ analysis separately before performing a cumbersome
35 linking of the outputs, a key advantage of FBMN is that the analysis is fully integrated throughout the
36 analysis pipeline.

37 To fully utilize the MS¹ and MS² content collected during a non-targeted liquid chromatography
38 coupled to tandem mass spectrometry data (LC-MS²) metabolomics experiment in a streamlined fashion,
39 we have created an online infrastructure to support the outputs of feature detection and alignment tools for
40 FBMN analysis (<https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking>),
41 including the standard output format for small molecules analysis (mzTab-M)⁸ (Fig. 1a). The diversity of
42 supported software, each offering different functionalities/modules, serves experimentalists,
43 bioinformaticians, and software developers. FBMN is already the second most utilized analysis tool
44 within the GNPS environment (Fig. 1b) with more than 600 jobs performed in August 2019 and has
45 already been used in more than 80 publications using FBMN during its development since Nov 2017.
46 The molecular networks generated with FBMN enable the efficient visualization and annotation of
47 isomers in LC-MS² datasets, as demonstrated below with LC-MS² data from a drug discovery project
48 from *Euphorbia* plant extract⁹ (Fig. 2a-b), and the detection of human microbiome-derived lipids,
49 belonging to the commendamide family¹⁰, detected in fecal samples from the American Gut Project¹¹ (a

1 crowd-sourced citizen science microbiome project) (Fig. 2c-d). In both cases, FBMN resolved positional
2 isomers in the molecular networks that have similar MS² spectra but distinct retention times, that would
3 not have been resolved with classical MN. In the study of metabolites produced by *Euphorbia* plants, the
4 annotation of isomers in the FBMN facilitated the subsequent isolation of antiviral compounds¹². With
5 samples from the American Gut Project, FBMN enabled the annotation of commensal isomers,
6 including a putative novel derivative¹¹.

7 By incorporating chromatographic information, FBMN reduced the complexity of molecular
8 networking analysis. In non-targeted LC-MS² data acquisition, the same precursor ion can be fragmented
9 multiple times during chromatographic elution, which ultimately leads to multiple nodes representing the
10 same compound in classical MN. Additionally, coeluting isobaric ions are often isolated and fragmented
11 together leading to the generation of chimeric spectra which result in different MS² spectra for one
12 precursor ion. With FBMN, the integration of feature detection with the alignment of the mass
13 spectrometry signal discerns that these different MS² spectra are related to the same precursor molecule
14 and selects a singular representative consensus spectrum for the feature. The benefit of using FBMN in
15 such a case can be illustrated with the metal chelating agent ethylenediaminetetraacetic acid (EDTA)
16 observed in the LC-MS² analysis of plasma samples (Fig 2e), in which it is used as an anticoagulant
17 agent. Classical MN resulted in 13 duplicated nodes with identical precursor *m/z* values in one molecular
18 family, ten of which have spectral library matches to EDTA reference MS² data (Fig. 2e and f). On the
19 contrary, FBMN displays a unique representative MS² spectrum that matches EDTA spectra in the
20 library, because the multiple MS² spectra detected were part of a single feature in the chromatographic
21 dimension. The reduction of redundancy within the resulting molecular network simplifies the discovery
22 of structurally related compounds.

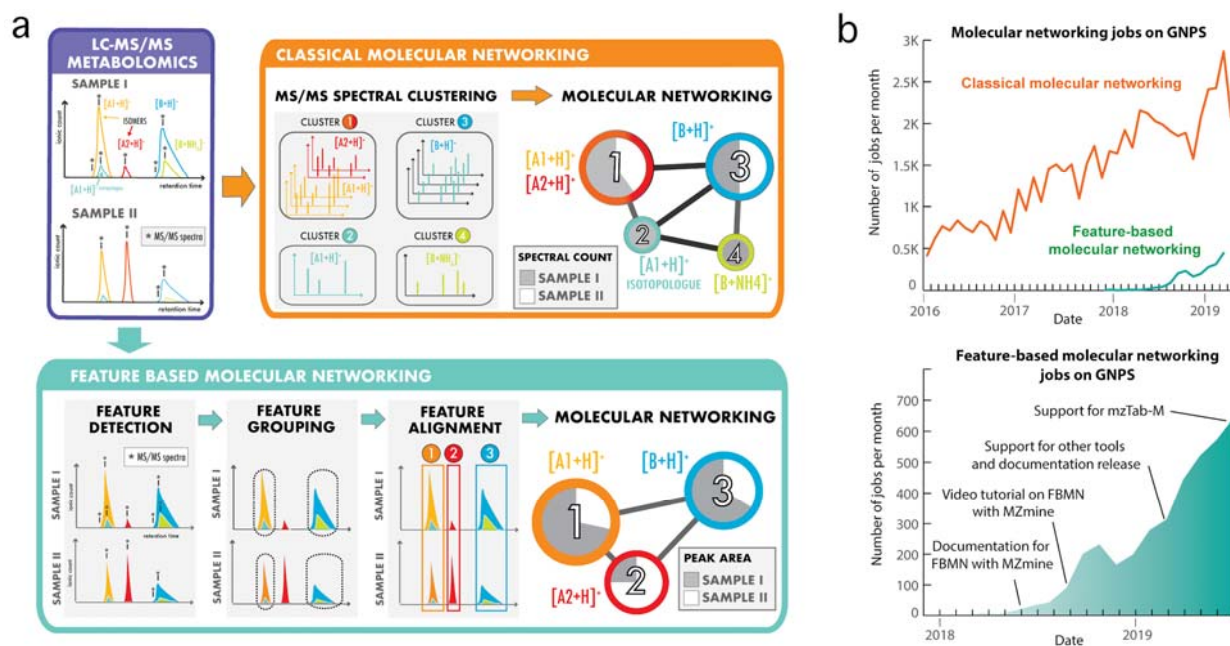
23 While classical MN uses the spectral count or the sum precursor ion count, FBMN uses the LC-
24 MS feature abundance (peak area or peak height), resulting in a more accurate estimation of the relative
25 ion intensity. The method of FBMN simplifies, organizes the data, and adds relative quantitative
26 information and precursor isotope patterns. FBMN enables robust statistical analysis by providing relative
27 ion intensities across a dataset. This capacity is demonstrated with a serial dilution series dataset of the
28 NIST1950 serum reference standard¹³, containing 150 spiked standards. Here, the LC-MS² were
29 processed with MZmine¹⁴ or OpenMS¹⁵ for FBMN (Fig 2g-h). A linear regression analysis was used to
30 evaluate the relative quantification between classical MN and FBMN. Figure 2h shows that for FBMN,
31 relative quantification has a correlation coefficient (*r*) value distribution mostly above 0.7, while this was
32 not the case when the precursor ion abundance was obtained from classical MN via spectral counts (Fig
33 2g). The improved distribution of correlation coefficients towards 1 indicates a more linear response
34 between concentration and ion abundance, which improves the accuracy and precision of quantification
35 results. FBMN facilitates the direct application of existing statistical, visualization, and annotation tools,
36 such as QIIME2¹⁶, MetaboAnalyst¹⁷, ili¹⁸, NAP¹⁹, MS2LDA²⁰, MolNetEnhancer²¹, and SIRIUS²².

37 FBMN further enables the creation of molecular networks from ion mobility spectrometry
38 experiments coupled with LC-MS² analysis. As an orthogonal separation method, the use of ion mobility
39 offers additional resolving power to differentiate isomeric ions in the molecular network. The integration
40 of ion mobility with FBMN on GNPS can currently be performed with MetaboScape, MS-DIAL²³, and
41 Progenesis QI. An example of such isomer separation using trapped ion mobility spectrometry (TIMS)
42 coupled to LC-MS² is shown in Supplementary Fig. 1.

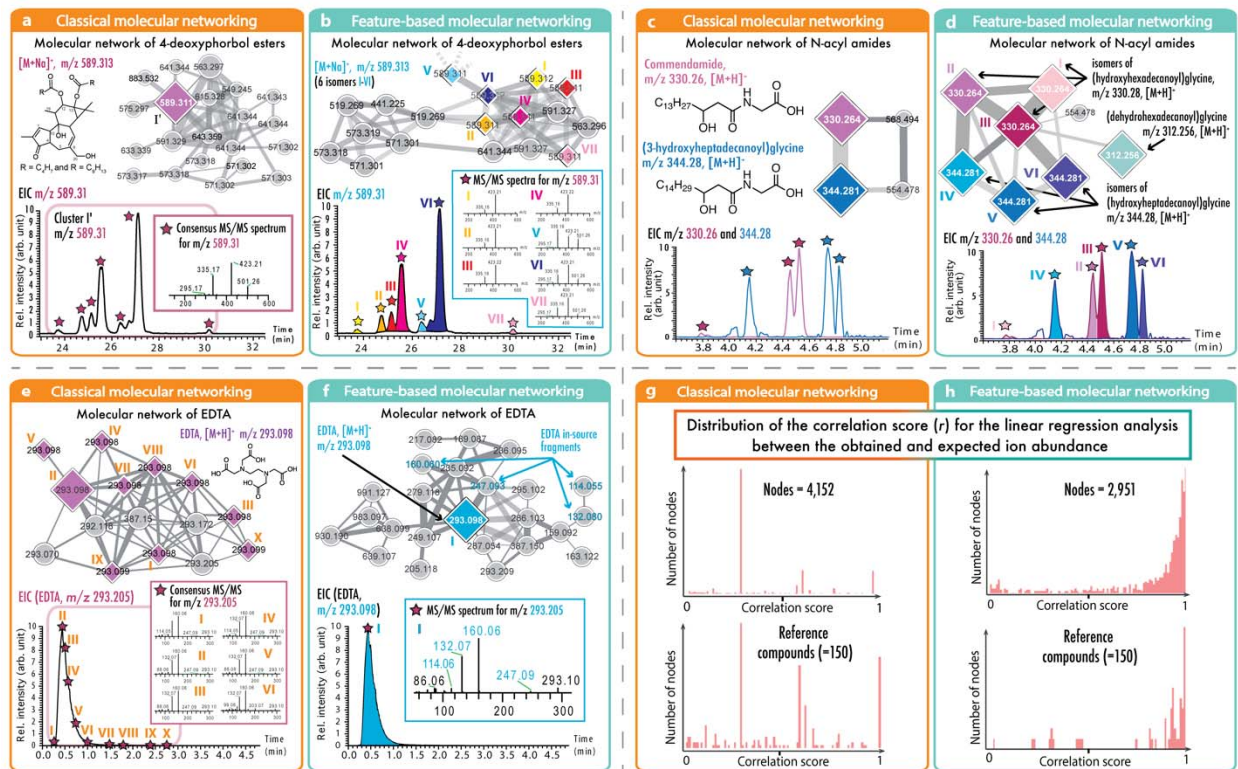
43 Available on the GNPS web platform at <https://gnps.ucsd.edu>, FBMN is ideally suited for
44 advanced molecular networking analysis, enabling the characterization of isomers, the incorporation of
45 relative quantification, and the integration of ion mobility data. FBMN is the recommended way to
46 analyse a single LC-MS² metabolomics study, but care must be taken when applied across multiple
47 studies due to different experimental conditions and possible batch effects. Moreover, the use of FBMN
48 for the analysis of very large datasets (containing several thousand samples) is limited by the scalability
49 of most feature detection and alignment software tools. Thus, while FBMN offers an improvement upon
50 many aspects of molecular networking analysis, classical MN remains essential for repository-scale meta-
51 analysis large dataset processing, and is convenient for rapid analysis of LC-MS² data with less user

1 defined parameters: one important aspect of molecular networks obtained with FBMN is the use of
2 adequate processing steps and parameters, which otherwise could negatively impact the resulting
3 molecular networks. To facilitate dissemination, education of the FBMN method, and the supported
4 processing software, we created detailed tutorials and step-by-step instructions, available at [https://ccms-ucsd-](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking)
5 [github.io/GNPSDocumentation/featurebasedmolecularnetworking](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking).

6 The FBMN workflow offers not only automated spectral library search and spectral entry
7 curation, but is also integrated with other annotation tools available on GNPS environment, such as
8 MASST²⁴, while promoting data analysis reproducibility by saving the FBMN jobs on the user's private
9 online workspace. The GNPS environment conveniently enables the user to evaluate different parameters
10 and enables the sharing of the results via a web URL for publication.
11
12



13
14
15 **Fig. 1: Methods for the generation of molecular networks from non-targeted mass spectrometry**
16 **data with the GNPS web platform. a)** Two methods exist for the generation of molecular networks on
17 the GNPS web platform: classical MN and feature-based molecular networking (FBMN). For both
18 methods, mass spectrometry data files have to be converted to the mzML format using tools such as
19 Proteowizard MSConvert²⁵. The classical MN method runs entirely on the GNPS platform. In that
20 method, MS² spectra are clustered with MS-Cluster and the consensus MS² spectra obtained are used for
21 classical MN generation. In the case of FBMN, a feature detection and alignment tool is used to first
22 process the LC-MS² data (such as MZmine, MS-DIAL, XCMS, OpenMS, Progenesis QI, or
23 MetaboScape) instead of using MS-Cluster (classical MN). Results are then exported and uploaded to the
24 GNPS web platform for molecular networking analysis. **b)** Graphs showing the number of molecular
25 networking jobs performed on GNPS. The upper graph shows the number of classical MN and FBMN
26 jobs since 2016. The lower graph shows the number of FBMN jobs since its introduction in 2017 and key
27 events accelerating its use.
28



1
2 **Fig. 2: Comparisons of classical MN and feature-based molecular networking.** In these examples, the
3 node size corresponds to the spectral count in classical MN (orange boxes, left) or to the sum of LC-MS²
4 peak area in FBMN (blue boxes, right). Panel (a) displays the results from classical MN with the LC-MS²
5 data of *Euphorbia dendroides* plant samples; classical MN resulted in one node for the ion at m/z
6 589.313, while (b) FBMN performed after MZmine processing was able to detect seven isomers of this
7 ion. Classical MN in the data of a cohort from the American Gut Project (c) showed two different *N*-acyl
8 amides while the use of FBMN (d) processed with MZmine allowed the annotation of three different
9 isomers per *N*-acyl amides. Classical MN (e) and FBMN (f) performed with OpenMS, were used to
10 analyse the network of EDTA in plasma (373 samples) by compressing MS² spectra of EDTA eluting
11 over 2.5 min into one best-quality MS² spectrum. FBMN recovered the molecular similarity of in-source
12 fragments observed for EDTA, which were not displayed with classical MN, due to the fixed top-K rank
13 for connected nodes (typically set to 10) of MS² spectral similarity. Evaluation of quantitative
14 performance using a serial dilution of serum reference sample (NIST1950SRM) analyzed with (g)
15 classical MN and (h) FBMN. The plots are showing the correlation score (r) distribution between the
16 observed and expected relative ion abundance. The upper charts present the distribution of the correlation
17 score for all the nodes (features) generated, and the bottom charts show the distribution for 150 reference
18 compounds. While classical MN uses the spectral count or the sum precursor ion count to estimate the
19 molecular network node abundance, FBMN uses the LC-MS feature abundance (peak area or peak
20 height), resulting in a more accurate estimation of the relative ion intensity.

21
22 **Methods**

23
24 **Development of Feature-based Molecular Networking (FBMN)**

25 The FBMN method consists of two main steps: 1) LC-MS feature detection and alignment, then 2) a
26 dedicated molecular networking workflow on GNPS. Our first prototype for FBMN was developed with
27 the Optimus workflow^{12,18} that uses OpenMS tools¹⁵ and the KNIME Analytics²⁶ platform. Following step
28 1 (feature detection and alignment), two files are exported: a *feature quantification table* (.TXT format)

1 and a *MS² spectral summary* (.MGF format). The feature quantification table contains information about
2 LC-MS features across all considered samples including a unique identifier (feature ID) for each feature,
3 *m/z* value, retention time, and intensity. The *MS² spectral summary* contains a list of *MS²* spectra, with
4 one representative *MS²* spectrum per feature. The mapping information between the feature quantification
5 table and the *MS² spectral summary* is stored in these files using the feature ID and scan number,
6 respectively. This simple mapping enables to relate LC-MS feature information or statistically derived
7 results to the molecular network nodes. This approach was also used for the integration of other tools with
8 FBMN, and does not require third party software like it was proposed in the past^{27,28}. Finally, the FBMN
9 workflow also supports the mzTab-M format⁸, a standardized output format designed for the report of
10 metabolomics *MS*-data processing results. In this case, the mzTab-M file is used instead of *feature*
11 *quantification table* and requires the input of the mzML files instead of the *MS² spectral summary* file.
12 Support for the mzTab-M format enables the possibility to perform FBMN with any existing and future
13 processing tools that support this standardized format.

14 The FBMN workflow has been integrated into the GNPS ecosystem and thus benefits from the
15 connection with other GNPS features, e.g. the possibility to perform automatic *MS²* spectral library
16 search, the direct addition and curation of library entries, the search of a spectrum against public datasets
17 with MASST²⁴, and the visualization of molecular networks directly in the web browser²⁹ or with
18 Cytoscape³⁰. The FBMN workflow is available on the GNPS platform (<https://gnps.ucsd.edu/>) via a web
19 interface (See Supplementary Fig. 2). Jobs are computed and stored on the computational infrastructure of
20 the Center for Computational Mass Spectrometry at the University of California San Diego. Each finished
21 job is saved in the private user space for future examination and has a permanent static link that enables
22 data sharing and collaborative analyses. We strongly recommend the sharing of this static link along with
23 publications using GNPS workflows to facilitate results accessibility and data analysis reproducibility.
24 Instructions to perform FBMN with the supported tools are provided in the GNPS documentation
25 (<https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking> and Supplementary
26 Fig. 3).

27 **FBMN with MZmine**

29 MZmine¹⁴ is a popular open-source cross-platform software for mass spectrometry data processing with
30 an advanced Graphical User Interface (GUI) that enables the users to visually optimize parameters and
31 examine the results of each processing step. Moreover, MZmine allows for the export of a batch file
32 containing all the steps and parameters used in the processing, thus enabling its reproducibility. To
33 support FBMN in MZmine, the feature detection step (peak “Deconvolution module”) was modified to
34 provide the ability to pair a feature with its *MS²* scans using an *m/z* and retention time range defined by
35 the user (Supplementary Fig. 4). Due to a new data structure and to support older projects (created with
36 release < 2.38), an additional specific filtering module (*Group *MS²* scans with features*) was developed to
37 assign all *MS²* scans to the features of existing peak list (see this video for instructions:
38 <https://www.youtube.com/watch?v=EL5pmFvpTFE>). Moreover, a GNPS export and direct submission
39 module was created (Supplementary Fig. 5) which offers two modes: 1) Export of the *feature*
40 *quantification table* and the *MS² spectral summary* file and 2) Direct FBMN analysis on the GNPS web
41 platform (release 2.37+). The direct GNPS job submission generates all the files and uploads them
42 together with an optional metadata table and default parameters (Supplementary Fig. 6) to the FBMN
43 workflow on GNPS. By providing the user’s GNPS login credentials (optional), a new job can be created
44 in the personal user space
45 (https://www.youtube.com/watch?v=vFcGG7T_44E&list=PL4L2Xw5k8ITzd9hx5XIP94vFPxj1sSafB&index=4&t=0s). Otherwise, the user can be notified by email or directly redirected to the job webpage
46 after the submission. With the option “*most intense*”, the GNPS Export uses the most intense *MS²*
47 spectrum as a representative spectrum for each LC-*MS²* feature. When using the “merge *MS/MS*” spectra
48 option (release 2.40+), a representative high quality *MS²* spectrum is instead generated from all spectra
49 and exported as a representative spectrum (Supplementary Note 1). The detailed documentation is

1 available at <https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-mzmine2/>.

4 **FBMN with OpenMS**

5 OpenMS is an open-source cross-platform software specifically designed for the flexible and reproducible
6 analysis of high-throughput MS data analysis, including more than 200 tools for common mass
7 spectrometric data processing tasks¹⁵. Building on our experience with the Optimus development, the
8 integration of OpenMS and FBMN was achieved by creating a *GNPSExport* tool (TOPP tool) as a part of
9 the OpenMS tool collection (<https://github.com/Bioinformatic-squad-DorresteinLab/OpenMS>). A detailed
10 description of the *GNPSExport* module and how to use it for FBMN is available at the following webpage
11 <https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-openms/>. In
12 brief, after running an OpenMS non-targeted metabolomics pipeline, the *GNPSExport* TOPP tool can be
13 applied to the consensusXML file resulting from *FeatureLinkerUnlabeledKD* or
14 *FeatureLinkerUnlabeledQT* tools (alignment step), and the corresponding mzML files. For each
15 consensusElement (LC-MS² feature) in the consensusXML file, the *GNPSExport* generates one
16 representative consensus MS² spectrum that will be exported in the *MS² spectral summary* file (using
17 either the option “most intense” or “merged spectra”, see Supplementary Note 1). The *TextExport* tool is
18 applied to the same consensusXML file to generate the *feature quantification table*. Note that the
19 *GNPSExport* requires the use of the *IDMapper* tool on the featureXML files (from the feature detection
20 step) prior to feature linking, in order to associate MS² scans [peptide annotation in OpenMS
21 terminology] with each feature. These MS² scans are used by the *GNPSExport* for the generation of the
22 representative MS² spectrum. Additionally, the *FileFilter* has to be run on the consensusXML file, prior to
23 the *GNPSExport*, in order to remove consensus Elements without associated MS² scans. The two files
24 exported (*feature quantification table* and *MS² spectral summary*) can be directly used for FBMN analysis
25 on GNPS. The OpenMS-GNPS workflow for metabolomics data processing was implemented as a
26 python wrapper around OpenMS TOPP tools ([https://github.com/Bioinformatic-squad-](https://github.com/Bioinformatic-squad-DorresteinLab/openms-gnps-tools)
27 [DorresteinLab/openms-gnps-tools](https://github.com/Bioinformatic-squad-DorresteinLab/openms-gnps-tools)), and released as a workflow ([https://github.com/Bioinformatic-squad-](https://github.com/Bioinformatic-squad-DorresteinLab/openms-gnps-workflow)
28 [DorresteinLab/openms-gnps-workflow](https://github.com/Bioinformatic-squad-DorresteinLab/openms-gnps-workflow)) on the GNPS/MassIVE web platform³¹. OpenMS version 2.4.0
29 was used¹⁵. The OpenMS + GNPS workflow can be accessed and run here:
30 <https://proteomics2.ucsd.edu/ProteoSAFe/>.

32 **FBMN with XCMS**

33 The XCMS package (<https://github.com/sneumann/xcms> for the most recent version) is one of the most
34 widely used software for processing of mass spectrometry-based metabolomics data³². The integration of
35 XCMS and FBMN is currently possible using a custom utility function “formatSpectraForGNPS”
36 creating the *MS² spectral summary*. This function is available on the following GitHub repository
37 <https://github.com/jorainer/xcms-gnps-tools> and is compatible with the CAMERA algorithm for isotopes
38 and adduct annotation³³. Representative XCMS R scripts in markdown and Jupyter notebook formats are
39 available in the following GitHub repository
40 https://github.com/DorresteinLaboratory/XCMS3_FeatureBasedMN. The two exported files (*feature*
41 *quantification table* and *MS² spectral summary*) can be directly used for FBMN analysis on GNPS. The
42 detailed documentation is available at [https://ccms-](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-xcms3/)
43 [ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-xcms3/](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-xcms3/).

45 **FBMN with MS-DIAL**

46 MS-DIAL is an open-source mass spectrometry data processing software²³ (available for Windows only,
47 http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/). The integration of MS-DIAL and FBMN
48 was made possible since ver. 2.68 by exporting the “Alignment results” using the “GNPS export” option.
49 In addition to LC-MS² data processing, MS-DIAL can process data from SWATH-MS² (data-independent
50 LC-MS² acquisition), and ion mobility spectrometry coupled to LC-MS². The two files exported (*feature*
51 *quantification table* and *MS² spectral summary*) can be directly used for FBMN analysis on GNPS. A

1 video tutorial on the use of MS-DIAL for FBMN is available at
2 <https://www.youtube.com/watch?v=hxk40jwAkcc&t=7s>. The detailed documentation is available at
3 <https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-ms-dial/>.

4 5 **FBMN with MetaboScope**

6 MetaboScope is a proprietary mass spectrometry metabolomics data processing software commercialized
7 by Bruker and available on Windows. MetaboScope can perform feature detection, alignment and
8 annotation of non-targeted LC-MS² data acquired on Bruker mass spectrometers. Support for the
9 processing of trapped ion mobility spectrometry (TIMS) coupled to non-targeted LC-MS² (LC-TIMS-
10 MS²) was added in MetaboScope 4.0, which results in LC-TIMS-MS features. Feature-based molecular
11 networking can be performed on LC-MS² or LC-TIMS-MS² data by exporting the *feature quantification*
12 *table* and *MS² spectral summary* from the “bucket table” using the “Export to GNPS format” function.
13 These files can be uploaded to GNPS for FBMN analysis. Information from MetaboScope, such as the
14 Collision Cross Section values, or other spectral annotations can be mapped into the molecular networks
15 using Cytoscape³⁰. The detailed documentation is available at [https://ccms-
16 ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-metaboscape/](https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-metaboscape/)
17

18 **FBMN with Progenesis QI**

19 Progenesis QI is a proprietary feature detection and alignment software developed by Nonlinear
20 Dynamics (Waters) that is compatible with various proprietary and open mass spectrometry data formats.
21 Progenesis QI can perform feature detection, alignment and annotation of non-targeted LC-MS² data
22 acquired either in data-dependent acquisition (DDA) or data independent analysis (DIA, such as MS^E),
23 and can also utilize the ion mobility spectrometry (IMS) dimension. FBMN can be performed on any of
24 these data types processed with Progenesis QI (ver 4.0), by exporting the *feature quantification table*
25 (.CSV format) and the *MS² spectral summary* (.MSP format). These two files can be exported from the
26 “Identify Compounds” submenu by using the function “Export compound measurement” and “Export
27 fragment database”, respectively. These files can be uploaded to GNPS for FBMN analysis. Information
28 from Progenesis QI, such as the Collision Cross Section values, or other spectral annotations can be
29 mapped into the molecular networks using Cytoscape³⁰. The detailed documentation is available at
30 <https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-progenesisqi/>.

31 32 **FBMN makes it possible to resolve isomers in a drug lead discovery effort**

33 The examination of the LC-MS² data ([MSV000080502](#)) from the *Euphorbia dendroides* plant extract
34 showed the presence of numerous chromatographic peaks for ions in the range *m/z* 500-900,
35 corresponding to diterpene ester derivatives. These specialized metabolites consist of a polyhydroxylated
36 diterpene core acylated with various acidic moieties, that are typically found as positional isomers based
37 on their acylation pattern³⁴. The extracted ion chromatogram (EIC) for the ion *m/z* 589.31 in the
38 *Euphorbia dendroides* extract data (Supplementary Fig. 7) shows the presence of at least seven distinct
39 LC-MS peaks between 24.5 and 27.3 min, including five peaks with an associated MS² spectra. The
40 analysis of the extract and the fractions where these molecules were originally isolated (fractions 13 and
41 14) with [classical MN](#) resulted in a [molecular network](#) with two nodes for the *m/z* 589.31 ions (Fig. 2a
42 and Supplementary Fig. 8). These MS² spectra (cluster index [5352](#) and [5354](#)) resulted from merging 96
43 fragmentation spectra spanning from 23.6 to 26.5 min by MS-Cluster (Fig. 2b and Supplementary Fig. 9).
44 Close examination of the clustered spectra revealed that while all MS² spectra for the precursor *m/z*
45 589.31 present fragment ions *m/z* 501.26, 423.21, 335.16, and 295.17, three distinct spectral types could
46 be established based on the ions relative intensities (Supplementary Fig. 10). FBMN of the dataset with
47 MZmine processing ([see the GNPS job](#)) enabled the differentiation of the MS² spectra of seven isomers
48 (Figure 2b and Supplementary Fig. 11 for the [molecular network view](#)). A detailed discussion on the
49 differences observed between the two methods can be found in the Supporting Information
50 (Supplementary Note 2 and Supplementary Table 1). Interestingly, in the original study¹² OpenMS was
51 used for FBMN and resulted in the observation of three different positional isomers instead of seven,

1 which shows that different processing methods can lead to different results with FBMN. These three
2 isomers were subsequently isolated and differed by the position of one double bond on the C-12 acyl
3 chain, or from carbon C-4 configuration¹². Because FBMN connects the accurate relative abundance of
4 the ions across the fractions and the molecular networks, it allowed to create bioactivity-based molecular
5 networks¹², which were used to predict and target potentially antiviral compounds. For detailed
6 description of the extraction, mass spectrometry analysis, and structural elucidation, see the original
7 manuscript¹². The MZmine project and parameters used can be accessed on the MassIVE submission
8 ([MSV000080502](https://massive.ucsd.edu/MSV000080502)).

9 **FBMN resolves isomers in large scale metabolomics studies**

11 FBMN was applied on a cohort of the American Gut Project (AGP), a citizen-scientist research project
12 that enabled the observation of the commendamide in humans, along with other new *N*-acyl amide
13 derivatives using molecular networking¹¹. Commendamide is a recently discovered bacterial *N*-acyl
14 amide that was shown to modulate host metabolism via G-protein-coupled receptors (GPCRs) in the
15 murine intestinal tract³⁵.

16 The use of FBMN for the AGP data (Figure 2d) made possible to observe the presence of two additional
17 commendamide isomers (*m/z* 330.26) and of an analogue (hydroxyheptadecanoyl)glycine (*m/z* 344.28),
18 while classical MN resulted in the observation of one single consensus spectrum for each compound
19 (Figure 2c). In addition, FBMN allowed to observe a putative commendamide derivative
20 (dehydrohexadecanoyl)glycine ([CCMSLIB00005436498](https://massive.ucsd.edu/CCMSLIB00005436498) and Supplementary Fig. 12) in the
21 commendamide molecular network. The sample collection and mass spectrometry acquisition methods
22 are described in the original manuscript¹¹. The data were downloaded from MassIVE ([MSV000080186](https://massive.ucsd.edu/MSV000080186))
23 and processed with MZmine (2.37). The MZmine project along with parameters and export files were
24 deposited to the MassIVE repository ([MSV000084095](https://massive.ucsd.edu/MSV000084095)). The chromatograms for *m/z* 330.26 and *m/z*
25 344.28 displayed in Figure 2c-d are from samples 43076_P3_RB9_01_314.mzML and
26 38131_P5_RA4_01_538.mzML, respectively. Chromatograms were exported with MZmine. The results
27 were exported with the “Export for/Submit to GNPS” module for FBMN analysis on GNPS. The
28 corresponding job can be accessed here:

29 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0a8432b5891a48d7ad8459ba4a89969f> (only logged
30 users can see all the input files). The mzML files were used for the classical MN job can be accessed
31 here: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3c27e43d908c4044bace405cc394cd25>.

33 **FBMN reduces spectral redundancy and deobfuscates spectral similarity relationships: the case of 34 EDTA**

35 The benefit of using FBMN can be illustrated with the metal chelating agent ethylenediaminetetraacetic
36 acid (EDTA), widely used in beauty products, food, and scientific protocols. A search for its occurrences
37 in public spectral datasets with the mass spectrometry search tool (MASST)²⁴ showed that it [is frequently
38 observed](#) in plasma samples where it is used during the sample preparation. We took one of the public
39 human serum sample datasets ([MSV00008263](https://massive.ucsd.edu/MSV00008263)) where EDTA was observed. For a detailed description of
40 the protocol and mass spectrometry parameters, see Supplementary Note 3. The analysis of the data with
41 classical MN showed that the EDTA ions are found in two molecular networks. One network consists of
42 [\[M+H\]⁺](#) spectra and the other of [\[M+Na\]⁺](#) spectra. Interestingly, each of these networks have one node
43 with a large number of clustered spectra (node 91205 for 4655 spectra, and node 116470 for 571 spectra,
44 respectively), but yet EDTA ions are represented by multiple nodes although these nodes have the same
45 precursor ion mass and retention time. Detailed analysis showed that while the median pairwise cosine
46 values between EDTA spectra are high (median value of 0.93 and 0.94), the spectra are not clustering into
47 a single node. Examination of the multiple fragmentation spectra for EDTA ions showed that some 1) are
48 chimeric spectra “contaminated” by fragment ions produced by co-eluting isobaric ions, and 2) that other
49 spectra were dominated by low intensity fragment ions resulting from MS² spectra acquired at low
50 intensity. The method of FBMN was applied on that same dataset using the OpenMS-GNPS workflow
51 ([see the job](#)), and the results showed that it efficiently reduces the appearance of these redundant node

1 patterns from the same molecule (see the FBMN job, Figure 2f), both for the molecular networks
2 containing the $[M+H]^+$ and $[M+Na]^+$ spectra. FBMN recovers the molecular similarity of in-source
3 fragments observed for EDTA, which were not displayed with classical MN, as they now fall within the
4 top-K rank (typically set to 10) of MS^2 spectral similarity considered in the network topology. The
5 parameters used for OpenMS tools can be accessed in the OpenMS-GNPS job (see the job). OpenMS ver.
6 2.4.0 was used¹⁵.

7 **FBMN enables the use of relative quantification in the molecular networks**

8 While classical MN uses the spectral count or the sum of precursor ion intensity to estimate the ion
9 abundance, FBMN uses the accurate ion intensities obtained from LC-MS feature detection. The FBMN
10 method brings in ion abundance across all samples by using the value of the chromatographic peak area
11 or peak height as determined by the LC-MS feature detection and alignment software. Using a serial
12 dilution of the NIST 1950 serum reference metabolome sample¹³ analyzed on an Orbitrap mass
13 spectrometer (Q Exactive, ThermoFisher) and processed with OpenMS or MZmine, we show the linearity
14 of the relative quantification with FBMN and the improvement compared to classical MN (Figure 2h).
15 The sample preparation and mass spectrometry methods are described in Supplementary Note 4. The files
16 along with the parameters for MZmine are available on the following MassIVE repository
17 ([MSV000084092](https://massive.ucsd.edu/MSV000084092)). The linear regression analysis was performed with python 2 (ver. 2.7.15) with the
18 *LinearRegression* function of the *sklearn* package (ver. 0.20.1)³⁶. The parameters used for OpenMS can
19 be obtained from the following job link:

20 <https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=aae71e9b72cf431d9b2606170c3f7a7d>. The
21 molecular networking jobs can be accessed here: classical MN
22 (<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=daf3f0d7cec94104b2c9001739964c31>), MZmine
23 processing (<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f443cad083be4979aedd2af0f97b9fe9>) and
24 OpenMS processing
25 (<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2c48a477ec094123987cdf90db4be8e4>).

26 **FBMN enables molecular networking with ion mobility spectrometry**

27 The sample NIST 1950 serum¹³ was analyzed using a timsTOF Pro (Bruker Daltonics, Bremen) in data-
28 dependent acquisition mode using PASEF (Parallel Accumulation-Serial Fragmentation)³⁷. The data were
29 then processed with MetaboScape (ver. 5.0) and the results were exported for FBMN analysis on GNPS.
30 The mass spectrometry acquisition method, data, and parameters used for the processing were deposited
31 on MassIVE ([MSV000084402](https://massive.ucsd.edu/MSV000084402)). Classical MN were annotated with the GNPS⁶, NIST17 and LipidBlast³⁸
32 spectral libraries

33 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2adc2cf33c646548798d0e285197a96>). Lipid
34 annotation in MetaboScape was performed using SimLipid (ver. 6.04, Premier Biosoft, Palo Alto) and
35 mapped to the FBMN
36 (<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0d89db67b0974939a91cb7d5bfe87072>). The
37 molecular networks were visualized with Cytoscape ver. 3.7.1³⁰, and the results are presented in the
38 Supplementary Information (Fig. S1). Classical MN were annotated with the GNPS⁶, NIST17 and
39 LipidBlast³⁸ spectral libraries
40 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2adc2cf33c646548798d0e285197a96>). FBMN were
41 annotated by mapping annotations from the MetaboScape annotation engine
42 (<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0d89db67b0974939a91cb7d5bfe87072>).

43 **Integration with other computational mass spectrometry annotation tools**

44 The MGF file format is accepted by numerous computational mass spectrometry annotation tools. The
45 use of these tools with the MS^2 spectral summary file enables subsequent direct mapping of these
46 annotations to the molecular networks produced by the feature-based molecular networking method.
47

1 These tools include SIRIUS²², DEREPLICATOR,^{39,40} NAP,¹⁹ MS2LDA²⁰, MolNetEnhancer²¹ (see
2 Supplementary Note 5), as well as other software such as MetWork⁴¹, CFM-ID⁴², MetGem⁴³, MetFrag⁴⁴.

4 **Large dataset processing with OpenMS and XCMS**

5 The processing of large metabolomics datasets (more than a thousand samples) is limited by the
6 scalability of existing LC-MS feature detection tools, especially those based on a GUI (such as MZmine
7 and MS-DIAL). We showed that with specific peak picking parameters the use of XCMS or OpenMS
8 enables using FBMN for large metabolomics study ([MSV000080030](https://doi.org/10.1101/812404), approximately 2,000 samples). See
9 the Supplementary Note 6, Supplementary Table 2, and Supplementary Fig. 13.

11 **Code availability**

12 The FBMN workflow is available as web-interface on the GNPS web platform (<https://gnps-quickstart.ucsd.edu/featurebasednetworking>). The workflow code is open source and available on GitHub
13 (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/feature-based-molecular-networking).
14 It is released under the licence of The Regents of the University of California and free for non-profit
15 research (https://github.com/CCMS-UCSD/GNPS_Workflows/blob/master/LICENSE). The workflow
16 was written in Python (ver. 3.7) and deployed with the ProteoSAFE workflow manager employed by
17 GNPS (<http://proteomics.ucsd.edu/Software/ProteoSAFe/>). We also provide documentation, support,
18 example files, and additional information on the GNPS documentation website (<https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/>). The source code of the
19 GNPSExport module in MZmine is available at (<https://github.com/mzmine/mzmine2>) under the GNU
20 General Public License. The source code of the GNPSExport tool in OpenMS is available at
21 (<https://github.com/Bioinformatic-squad-DorresteinLab/OpenMS>) under the BSD licence. The source
22 code for the GNPSExport custom function for XCMS is available at [https://github.com/jorainer/xcms-
23 gnps-tools](https://github.com/jorainer/xcms-gnps-tools) under the GNU General Public License.

27 **Data availability**

28 The LC-MS² data for the *Euphorbia dendroides* dataset, along with the MZmine project and parameters
29 used can be accessed on the MassIVE submission ([MSV000080502](https://massive.ucsd.edu/MSV000080502), Creative Commons CC0 1.0
30 Universal license). The classical MN and FBMN jobs can be accessed via the GNPS website at
31 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=189e8bf16af145758b0a900f1c44ff4a> and
32 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=672d0a5372384cff8c47297c2048d789>, respectively.
33 The LC-MS² data for the American Gut Project (AGP) were downloaded from MassIVE
34 ([MSV000080186](https://massive.ucsd.edu/MSV000080186) Creative Commons CC0 1.0 Universal license) and processed with MZmine (2.37). The
35 MZmine project along with parameters and export files were deposited ([MSV000084095](https://massive.ucsd.edu/MSV000084095), Creative
36 Commons CC0 1.0 Universal license). The classical MN and FBMN jobs can be accessed at
37 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3c27e43d908c4044bace405cc394cd25> and
38 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0a8432b5891a48d7ad8459ba4a89969f>, respectively.
39 The LC-MS² data for the EDTA case are available on the MassIVE submission ([MSV00008263](https://massive.ucsd.edu/MSV00008263), Creative
40 Commons CC0 1.0 Universal license). The classical MN job can be accessed at
41 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fbac1a5061ba4ad683a284ef55d45df6>. The OpenMS
42 and the FBMN job at
43 <https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=83a0a417a49b4b76b61e9a8191a6ea2d> at
44 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8f40420c11694cf9ab06fdf7a5a4c53b>, respectively.
45 The mass spectrometry acquisition method, data, and parameters used for the processing of the serum
46 analysis with the timsTOF mass spectrometer were deposited ([MSV000084402](https://massive.ucsd.edu/MSV000084402)). Classical MN and
47 FBMN jobs can be accessed here:
48 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2adc2cf33c646548798d0e285197a96>, and
49 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0d89db67b0974939a91cb7d5bfe87072>, respectively.
50

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

References

1. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci.* **109**, E1743–52 (2012).
2. Quinn, R. A. *et al.* Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends Pharmacol. Sci.* **38**, 143–154 (2017).
3. Fox Ramos, A. E., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat. Prod. Rep.* **36**, 960–980 (2019).
4. 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).
5. Yang, J. Y. *et al.* Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76**, 1686–1699 (2013).
6. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
7. Frank, A. M. *et al.* Clustering Millions of Tandem Mass Spectra. *J. Proteome Res.* **7**, 113–122 (01/2008).
8. Hoffmann, N. *et al.* mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. *Anal. Chem.* **91**, 3302–3310 (2019).
9. Esposito, M. *et al.* Euphorbia dendroides Latex as a Source of Jatrophane Esters: Isolation, Structural Analysis, Conformational Study, and Anti-CHIKV Activity. *J. Nat. Prod.* (octobre 27, 2016). doi:10.1021/acs.jnatprod.6b00644
10. Cohen, L. J. *et al.* Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E4825–E4834 (2015).
11. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research.

- 1 *mSystems* **3**, (2018).
- 2 12. Nothias, L.-F. *et al.* Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in
3 Natural Product Bioassay-Guided Fractionation. *J. Nat. Prod.* **81**, 758–767 (2018).
- 4 13. Simón-Manso, Y. *et al.* Metabolite profiling of a NIST Standard Reference Material for human
5 plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-
6 based resources. *Anal. Chem.* **85**, 11725–11731 (2013).
- 7 14. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for
8 processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC*
9 *Bioinformatics* **11**, 395 (2010).
- 10 15. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data
11 analysis. *Nat. Methods* **13**, 741–748 (2016).
- 12 16. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using
13 QIIME 2. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0209-9
- 14 17. Xia, J., Sinelnikov, I. V., Han, B. & Wishart, D. S. MetaboAnalyst 3.0—making metabolomics more
15 meaningful. *Nucleic Acids Res.* **43**, W251–W257 (2015).
- 16 18. Protsyuk, I., Melnik, A. V., Nothias, L. F. & Rappetz, L. 3D molecular cartography using LC–MS
17 facilitated by Optimus and 'ili software. *Nat. Protoc.* (2018).
- 18 19. da Silva, R. R. *et al.* Propagating annotations of molecular networks using in silico fragmentation.
19 *PLoS Comput. Biol.* **14**, e1006089 (2018).
- 20 20. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for
21 untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 13738–
22 13743 (2016).
- 23 21. Ernst, M. *et al.* MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining
24 and Annotation Tools. *Metabolites* **9**, (2019).
- 25 22. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure
26 information. *Nat. Methods* **16**, 299–302 (2019).

- 1 23. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive
2 metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
- 3 24. Wang, M. *et al.* MASST: A Web-based Basic Mass Spectrometry Search Tool for Molecules to
4 Search Public Data. *bioRxiv* 591016 (2019). doi:10.1101/591016
- 5 25. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat.*
6 *Biotechnol.* **30**, 918–920 (2012).
- 7 26. Berthold, M. R. *et al.* KNIME: The Konstanz Information Miner. in *Data Analysis, Machine*
8 *Learning and Applications* 319–326 (Springer Berlin Heidelberg, 2008).
- 9 27. Winnikoff, J. R., Glukhov, E., Watrous, J., Dorrestein, P. C. & Gerwick, W. H. Quantitative
10 molecular networking to profile marine cyanobacterial metabolomes. *J. Antibiot.* **67**, 105–112
11 (2014).
- 12 28. Olivon, F., Grelier, G., Roussi, F., Litaudon, M. & Touboul, D. MZmine 2 Data-Preprocessing To
13 Enhance Molecular Networking Reliability. *Anal. Chem.* (2017). doi:10.1021/acs.analchem.7b01563
- 14 29. Ono, K., Demchak, B. & Ideker, T. Cytoscape tools for the web age: D3.js and Cytoscape.js
15 exporters. *F1000Res.* **3**, 143 (2014).
- 16 30. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular
17 interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
- 18 31. Wang, M. Spectral Library Construction and Matching of MS/MS Spectra at Repository Scales. (UC
19 San Diego, 2017).
- 20 32. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution
21 LC/MS. *BMC Bioinformatics* **9**, 504 (2008).
- 22 33. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated
23 Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass
24 Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289.
- 25 34. Vasas, A. & Hohmann, J. Euphorbia Diterpenes: Isolation, Structure, Biological Activity, and
26 Synthesis (2008–2012). *Chem. Rev.* **114**, 8579–8612.

- 1 35. Cohen, L. J. *et al.* Commensal bacteria make GPCR ligands that mimic human signalling molecules.
2 *Nature* **549**, 48–53 (2017).
- 3 36. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830
4 (2011).
- 5 37. Meier, F. *et al.* Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped
6 Ion Mobility Mass Spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).
- 7 38. Kind, T. *et al.* LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat.*
8 *Methods* **10**, 755–758 (2013).
- 9 39. Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass
10 spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
- 11 40. Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra.
12 *Nat. Commun.* **9**, 4035 (2018).
- 13 41. Beauxis, Y. & Genta-Jouve, G. Network: a web server for natural products anticipation.
14 *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty864
- 15 42. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for
16 putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).
- 17 43. Olivon, F. *et al.* MetGem Software for the Generation of Molecular Networks Based on the t-SNE
18 Algorithm. *Anal. Chem.* (2018). doi:10.1021/acs.analchem.8b03099
- 19 44. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched:
20 incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **8**, 1–16 (2016).

21 **Author Contributions**

22 L.F.N., D.P., M.W., and P.D. conceived the method and supervised its implementation and wrote the
23 manuscript.

24 L.F.N and D.P contributed equally to the work.

25 I.P., L.F.N., M.E., and T.A. created the FBMN prototype in Optimus.

26 M.W., L.F.N. D.P. and Z.Z. created the FBMN workflow on GNPS.

27 R.S., L.F.N, M.W., D.P., A.K., M.F, Z.Z., A.S., and T.P. developed the GNPS Export module in
28 MZmine.

29 K.D., A.K., M.L., and S.B. developed the spectral clustering algorithm and SIRIUS export in MZmine.

30 A.S., and L.F.N. created the GNPS Export tool in OpenMS, with the guidance from F.A, O.A., and O.K.

1 J.R. and M.Wit. created the XCMS export tool.
2 H.T, M.W. and L.F.N. made possible the integration with MS-DIAL.
3 L.F.N., A.B., H.N., F.Z. and T.D. made possible the integration with MetaboScape.
4 M.W., G.I., B.S., S.W.M. and J.M. made possible the integration with Progenesis QI.
5 F.V. performed the mass spectrometry for the plasma and NIST1950SRM samples.
6 A.A. performed the mass spectrometry for the American Gut Project samples.
7 A.K.J, L.F.N, and A.Tri. analyzed the results of the plasma samples.
8 J.R and L.F.N. performed the XCMS processing of the forensic dataset.
9 L.F.N. and M.W. created the general documentation.
10 D.P., L.F.N. and R.d.S. created the MZmine documentation.
11 K.B.K., H.Y. created the MS-DIAL documentation.
12 F.V., J.M.G. K.W., and A.K.J. prepared the MS-DIAL video tutorial.
13 M.W., R.S., D.P. prepared the MZmine video tutorials.
14 M.E., R.d.S., J.R., O.M., and S.N. created the XCMS documentation.
15 L.F.N and A.S. created the OpenMS documentation.
16 L.F.N, N.H.N., and T.D, created the MetaboScape documentation.
17 M.C., and L.-I. M. documented the FBMN interface workflow.
18 M.N.-E., I.K., and C.M created the Cytoscape documentation.
19 H.M, A.G., M.W. and L.F.N. made the integration with DEREPLICATOR.
20 M.W., J.J.J.v.d.H, M.E. and S.R. made the integration with MS2LDA.
21 R.d.S made the integration with NAP.
22 M.M., N.B., X.C., V.V.P., N.G. R.A.Q, A.A., Z.K., and S.N. tested and provided suggestions on how to
23 improve the methods.
24 J.J.J.v.d.H., V.P., T.A., A.K.J., T.P., A.L.G, L.-I.M., P.-M.A., S.B., and S.N. improved the manuscript.
25 All authors have contributed to the final manuscript.

26

27 **Ethics/COI declaration**

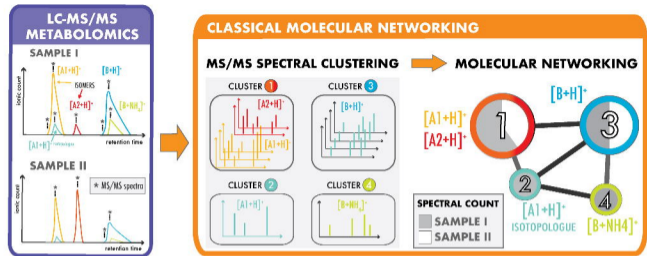
28 Pieter C. Dorrestein is a scientific advisor for Sirenas LLC.
29 Mingxun Wang is a consultant for Sirenas LLC and the founder of Ometa labs LLC.
30 Tomáš Pluskal is a consultant for Ginkgo Bioworks.
31 Alexander Aksenov is a consultant for Ometa labs LLC.
32 Theodore Alexandrov is on the Scientific Advisory Board of SCiLS, a Bruker company.
33 Kai Dührkop, Marcus Ludwig, Markus Fleischauer and Sebastian Böcker are founders of Bright Giant
34 GmbH.
35 Aiko Barsch, Sven W. Meyer, Heiko Neuweiger and Florian Zubeil are employes of Bruker Daltonics
36 GmbH.
37 Giorgis Isaac, Jonathan McSayles, and Bindesh Shrestha are employees of Waters Corporation.
38
39

40 **Acknowledgements**

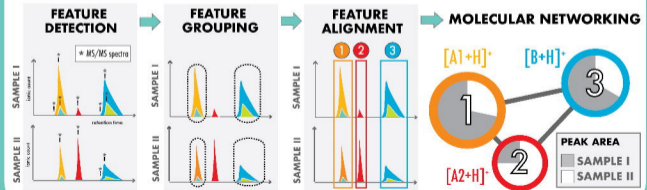
41
42 We gratefully acknowledge financial support by: the U.S. National Institutes of Health for the Center for
43 Computational Mass Spectrometry grant (P41 GM103484), the reuse of metabolomics data (R03
44 CA211211), and the tools for rapid and accurate structure elucidation of natural products (R01
45 GM107550) to P.D; the European Union's Horizon 2020 grants 704786 (MSCA-GF to L.F.N) 634402,
46 777222 (T.A. and I.P.) and the ERC Consolidator grant METACELL (T.A.). The German Research
47 Foundation (DFG) with grant number PE 2600/1 to D.P.; S.N. acknowledges funding from
48 Bundesministerium für Bildung und Forschung (FKZ 031L0107) and the European Commission
49 (EC654241); NSF grant IOS-1656481 to PCD and AMCR; O.A., acknowledge the funding from: the

1 Bundesministerium für Ernährung und Landwirtschaft (FKZ 2816501214), the Bundesministerium für
2 Wirtschaft und Energie (FKZ AiF18475N), the Bundesministerium für Bildung und Forschung (FKZ
3 031A430C), and the European Commission (823839) that also benefited to F.A. and O.K.; S.B.
4 acknowledges funding from Deutsche Forschungsgemeinschaft (BO 1910/20); J.J.J.v.d.H. was supported
5 by an Accelerating Scientific Discoveries Grant funded by the Netherlands eScience Center [NLeSC]
6 (No. ASDI.2017.030). Z.K. was supported by the project International Mobility of Researchers
7 (CZ.02.2.69/0.0/0.0/16_027/0007990). The authors would like to thank Nils Hoffman for maintaining the
8 mzTab-M format.

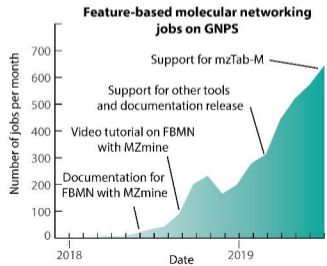
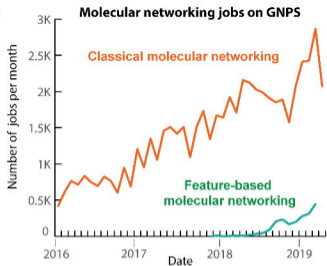
a



FEATURE BASED MOLECULAR NETWORKING

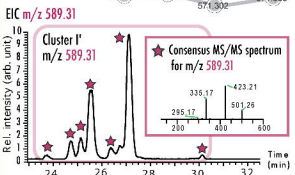
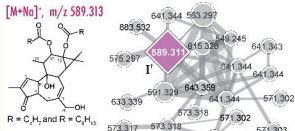


b



a Classical molecular networking

Molecular network of 4-deoxyphorbol esters



b Feature-based molecular networking

Molecular network of 4-deoxyphorbol esters

