nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-020-0933-6

Feature-based molecular networking in the GNPS analysis environment

Louis-Félix Nothias[1,2,45], Daniel Petras[1,2,3,45], Robin Schmid[4,45], 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 A. Aksenov[1,2], Oliver Alka[11,12], Pierre-Marie Allard[13], Aiko Barsch[14], Xavier Cachet[15], Andres Mauricio Caraballo-Rodriguez[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 Neuweger[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[20], Anupriya Tripathi[1,29,37], Justin J. J. van der Hooft[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] ✉

Molecular networking has become a key method to visualize and annotate the chemical space in non-targeted mass spectrometry data. We present feature-based molecular networking (FBMN) as an analysis method in the Global Natural Products Social Molecular Networking (GNPS) infrastructure that builds on chromatographic feature detection and alignment tools. FBMN enables quantitative analysis and resolution of isomers, including from ion mobility spectrometry.

Since its introduction in 2012 (ref. [1]), molecular networking has become an essential bioinformatics tool to visualize and annotate non-targeted mass spectrometry (MS) data[2,3]. Molecular networking, uniquely, goes beyond spectral matching against reference spectra, by aligning experimental spectra against one another and connecting related molecules by their spectral similarity. In a molecular network, related molecules are referred to as a 'molecular family', differing by simple transformations such as glycosylation, alkylation and oxidation/reduction. Molecular networking became publicly accessible in 2013 through the initial release of GNPS, a web-enabled MS knowledge capture and analysis platform (https://gnps.ucsd.edu/)[4], and has been widely applied in MS-based metabolomics to aid in the annotation of molecular families from their fragmentation spectra (MS^2).

Powered by more than 3,000 CPU cores at the University of California San Diego Center for Computational Mass Spectrometry and the MassIVE data repository, GNPS has provided researchers from more than 150 countries with the ability to perform molecular networking. To build upon the success of the first molecular networking method referred to as 'classical' molecular networking (classical MN), which is based on the MS-Cluster algorithm[5], we introduce a complementary tool named FBMN. FBMN leverages the capability of well-established MS processing software and improves upon classical MN by incorporating not only MS^1 information, such as isotope patterns and retention time, but also ion mobility separation when performed. By relying on processed spectral information, molecular networks obtained with FBMN can (1) distinguish isomers producing similar MS^2 spectra that are resolved by chromatographic or ion mobility separation, which may have remained hidden in classical MN, (2) facilitate spectral annotation, and (3) incorporate relative quantitative information that enables robust downstream metabolomics statistical analysis. Whereas users of the classical MN would have had to perform molecular networking and MS^1 analysis separately before performing a cumbersome linking of the outputs, the FBMN method accepts the output of feature detection and alignment tools, making them directly compatible with annotation tools and the entirety of the analysis pipeline.

To fully utilize the MS^1 and MS^2 data collected during a non-targeted metabolomics experiment in liquid chromatography coupled to tandem MS (LC–MS^2), we have created an online and streamlined workflow (Fig. 1a) infrastructure that supports the outputs of feature detection and alignment tools for FBMN analysis (https://ccms-ucsd.github.io/GNPSDocumentation/featurebased-molecularnetworking/), including the standard output format for analysis of small molecules (mzTab-M)[6]. The diversity of supported software, each offering different functionalities and modules, serves experimentalists, bioinformaticians, and software developers. FBMN is the second most commonly used analysis tool within the GNPS environment (Fig. 1b), with more than 6,767 jobs performed

A full list of affiliations appears at the end of the paper.

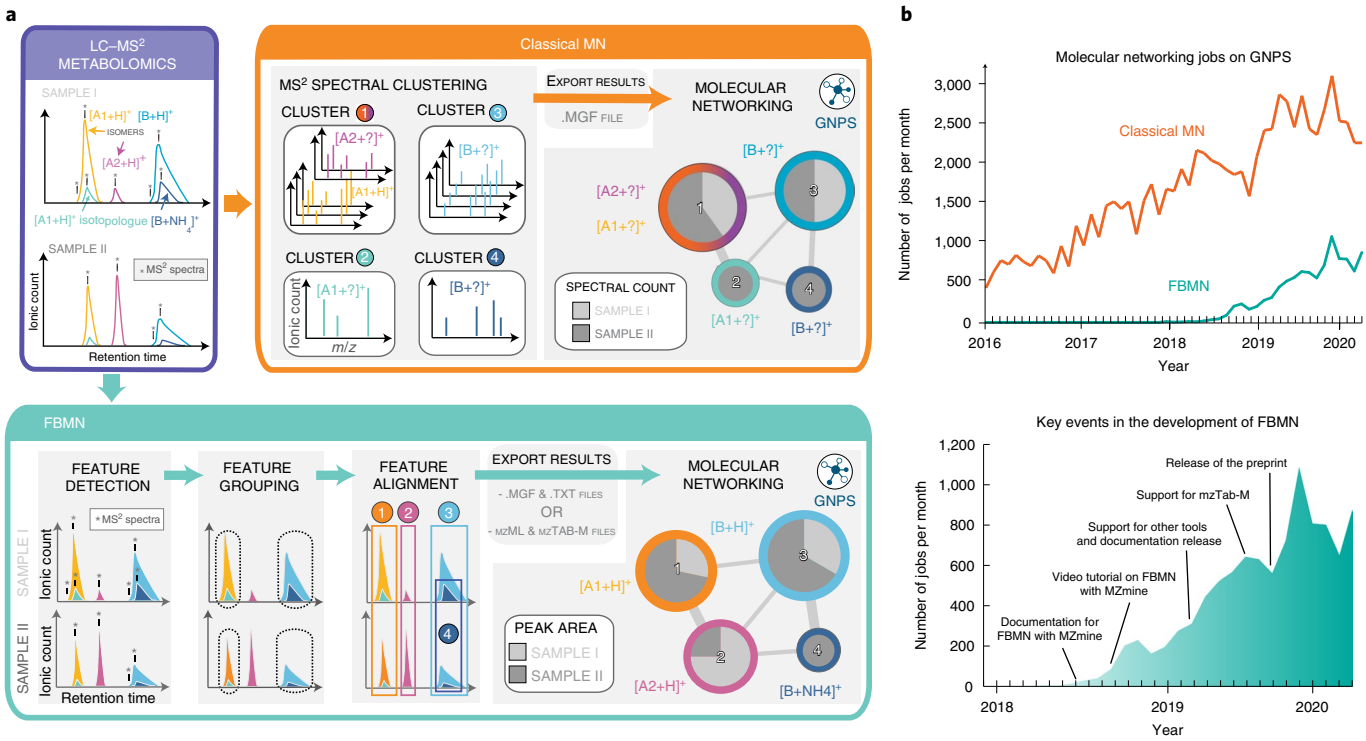

Fig. 1 | Methods for the generation of molecular networks from non-targeted MS data with the GNPS web platform. a, Two methods exist for the generation of molecular networks on the GNPS web platform: classical MN and FBMN. For both methods, the MS data files are converted to the mzML format using tools such as Proteowizard MSConvert[21]. The classical MN method runs entirely on the GNPS platform, in which MS^2 spectra are clustered with MS-Cluster and the consensus MS^2 spectra obtained are used for molecular network generation. For FBMN, the user first applies a feature detection and alignment tool to process the LC–MS^2 data (such as MZmine, MS-DIAL, XCMS, OpenMS, Progenesis QI or MetaboScape) instead of using MS-Cluster (classical MN) on GNPS. Results are then exported as a feature quantification table (.TXT format) and MS^2 spectral summary (.MGF format) or an mzTab-M file and uploaded to the GNPS web platform for molecular networking analysis with the FBMN workflow. **b**, Graphs show the number of molecular networking jobs performed on GNPS. Top: the number of classical MN and FBMN jobs since 2016. Bottom: the number of FBMN jobs since its inception and key events accelerating its use.

in 2019, and has already been used in more than 80 publications since its introduction in November 2017.

The molecular networks generated with FBMN enable the efficient visualization and annotation of isomers in LC–MS^2 datasets, as demonstrated below with LC–MS^2 data from a drug discovery project from *Euphorbia* plant extract[7] (Fig. 2a,b) and the detection of human microbiome-derived lipids belonging to the commendamide family[8], detected in fecal samples from the American Gut Project (AGP[9]; a crowd-sourced citizen-science microbiome project; Fig. 2c,d). In both cases, FBMN resolved positional isomers/stereoisomers in the molecular networks that have similar MS^2 spectra but distinct retention times, that would not have been resolved with classical MN. The uses of FBMN facilitated the discovery of antiviral compounds[7] (Fig. 2c), and the annotation of commendamide isomers[9] and of a putative new derivative, the *N*-(dehydrohexadecanoyl)glycine (Fig. 2d).

In non-targeted LC–MS^2 data acquisition, the same precursor ion is frequently fragmented multiple times during chromatographic elution. While MS-Cluster is often able to cluster these spectra into one single node in classical MN, there are cases where it will fail and produce multiple nodes representing the same compound. For example, this can happen for compounds producing mostly low-intensity fragment ions or for chimeric spectra resulting from coeluting isobaric ions isolated and fragmented together. With FBMN, a singular representative MS^2 spectrum is selected for the LC–MS feature (defined as the detected ion signal for an eluting molecule)[10]. The benefit of using FBMN in such instances can be illustrated with the metal chelating agent EDTA in the LC–MS^2 analysis of plasma samples (Fig. 2e), in which it was used as an anticoagulant agent. Classical MN resulted in 13 duplicated nodes with identical precursor *m/z* values in one molecular family, 10 of which had spectral library matches to EDTA reference MS^2 data (Fig. 2e,f). On the contrary, FBMN displayed a unique representative MS^2 spectrum that matched EDTA spectra in the library. The reduction of redundancy within the resulting molecular network simplifies the discovery of structurally related compounds.

While classical MN uses the spectral count or the summed precursor ion count, FBMN uses the LC–MS feature abundance (peak area or peak height), resulting in a more accurate estimation of the relative ion intensity. FBMN simplifies and aggregates data by including relative quantitative information and other MS^1 derived information (that is, precursor isotope patterns, adduct annotation). FBMN enables robust statistical analysis by providing accurate relative ion intensities across a dataset. This capacity is demonstrated with a serial dilution series dataset of the NIST 1950 serum reference standard, containing 150 spiked standards. Here the LC–MS^2 data were processed with MZmine[11] or OpenMS[10] for FBMN (Fig. 2g,h). A linear regression analysis was used to evaluate the relative quantification between classical MN and FBMN. Figure 2h shows that for FBMN, relative quantification had a coefficient of determination (R^2) value distribution mostly above 0.7, whereas this was not found when the precursor ion abundance was obtained from classical MN via spectral counts (Fig. 2g). The improved distribution of correlation coefficients toward 1 indicates a more linear response between molecular concentration and ion abundance, which improves the accuracy and precision of the quantification of results. In addition, FBMN facilitates

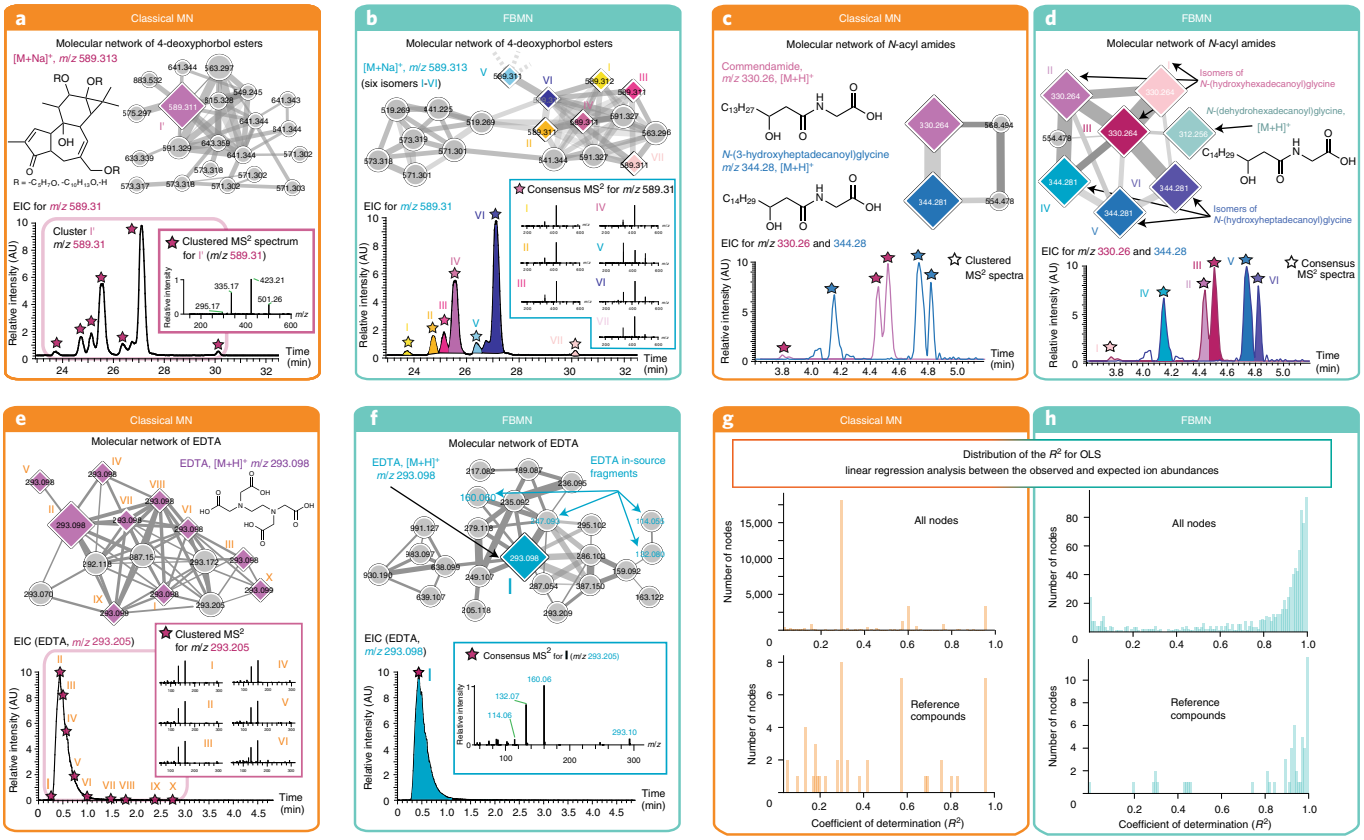

Fig. 2 | Comparisons of classical MN and FBMN. a–h, In these examples, the node size corresponds to the relative spectral count in classical MN (orange boxes, left) or to the sum of LC–MS peak area in FBMN (blue boxes, right); diamond shape nodes are spectra annotated by spectral library matching; the edge color gradient indicates the spectral similarity degree (lighter colors correspond to less similarity). **a**,**b**, Results from classical MN with LC–MS^2 data of *Euphorbia dendroides* plant samples (14 samples; $n=1$ LC–MS^2 experiment per sample); classical MN resulted in one node for the ion at *m/z* 589.313 (**a**), while FBMN was able to detect seven isomers (**b**). AU, arbitrary units; EIC, extracted ion chromatogram. **c**,**d**, Classical MN with data from the AGP (201 samples; $n=1$ LC–MS^2 experiment per sample) showed two different *N*-acyl amides (**c**), while the use of FBMN allowed the annotation of three different isomers per *N*-acyl amides (**d**). **e**,**f**, Classical MN (**e**) and FBMN (**f**) were used to analyze the network of EDTA in plasma (373 samples; $n=1$ LC–MS^2 experiment per sample). By merging MS^2 spectra of EDTA eluting over 2.5 min into one representative MS^2 spectrum, FBMN recovered the molecular similarity of in-source fragments observed for EDTA. **g**,**h**, Evaluation of quantitative performance using multiple dilutions of a reference serum sample (5 dilutions; $n=3$ LC–MS^2 experiments per sample). The plots show the distribution of the coefficient of determination (R^2) from the ordinary least squares (OLS) linear regression analysis between the observed and expected relative ion abundances for molecular network nodes in classical MN (**g**) or FBMN (**h**). The upper charts present the distribution of the R^2 for the network nodes with classical MN ($n=3{,}367$) and FBMN ($n=877$), and the lower charts show the R^2 distribution for the annotated reference compounds with classical MN ($n=49$) and FBMN ($n=54$).

the direct application of existing statistical, visualization and annotation tools, such as QIIME2 (ref. [12]), MetaboAnalyst[13], 'ili[14], SIRIUS[15], DEREPLICATOR[16], MS2LDA[17] and Qemistree[18].

FBMN further enables the creation of molecular networks from ion mobility spectrometry (IMS) experiments coupled within LC–MS^2 analysis. As an orthogonal separation method, the use of ion mobility offers additional resolving power to differentiate isomeric ions in the molecular network based on their collisional cross-section. The integration of ion mobility with FBMN on GNPS can currently be performed with MetaboScape, MS-DIAL[19] and Progenesis QI. An example of such isomer separation using trapped IMS (TIMS) coupled to LC–MS^2 is shown in Supplementary Fig. 1.

Available on the GNPS web platform at https://gnps.ucsd.edu/, FBMN is ideally suited for advanced molecular networking analysis, enabling the characterization of isomers, incorporation of relative quantification and integration of ion mobility data. FBMN analysis is recommended for a single LC–MS^2 metabolomics study, but its applicability is limited when applied across multiple studies due to different experimental conditions and possible batch effects. Moreover, the use of FBMN for the analysis of very large datasets (containing several thousand samples) is limited by the scalability of most feature detection and alignment software tools. Thus, while FBMN offers an improvement upon many aspects of molecular networking analysis, classical MN remains essential for meta-analysis of large-scale datasets and is convenient for rapid analysis of LC–MS^2 data with less user-defined parameters; one important aspect of molecular networks obtained with FBMN is the use of adequate processing steps and parameters, which otherwise could negatively affect the resulting molecular networks. To facilitate dissemination and education of the FBMN method and the supported processing software, we have created detailed tutorials and step-by-step instructions, available at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/.

The FBMN workflow not only offers automated spectral library search and spectral library entry curation, but is also integrated with other annotation tools available on the GNPS environment, such as MASST[20], while promoting data analysis reproducibility by saving the FBMN jobs on the user's private online workspace. The GNPS environment conveniently enables the user to evaluate different parameters and share the results via a URL for publication.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0933-6.

Received: 18 October 2019; Accepted: 22 July 2020;
Published online: 24 August 2020

References

1. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109**, E1743–E1752 (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. Traxler, M. F. & Kolter, R. A massively spectacular view of the chemical lives of microbes. *Proc. Natl Acad. Sci. USA* **109**, 10128–10129 (2012).
4. 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).
5. Frank, A. M. et al. Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).
6. Hoffmann, N. et al. mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal. Chem.* **91**, 3302–3310 (2019).
7. Nothias, L.-F. et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *J. Nat. Prod.* **81**, 758–767 (2018).
8. 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. USA.* **112**, E4825–E4834 (2015).
9. McDonald, D. et al. American Gut: an open platform for citizen-science microbiome research. *mSystems* **3**, e0031–18 (2018).
10. Röst, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
11. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
12. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
13. Xia, J., Sinelnikov, I. V., Han, B. & Wishart, D. S. MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Res.* **43**, W251–W257 (2015).
14. Protsyuk, I., Melnik, A. V., Nothias, L. F. & Rappez, L. 3D molecular cartography using LC–MS facilitated by Optimus and'ili software. *Nat. Protoc.* **13**, 134–154 (2018).
15. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
16. Mohimani, H. et al. Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
17. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl Acad. Sci. USA.* **113**, 13738–13743 (2016).
18. Tripathi, A. et al. Chemically-informed analyses of metabolomics mass spectrometry data with qemistree. Preprint at *bioRxiv* 2020.05.04.077636 (2020) https://doi.org/10.1101/2020.05.04.077636.
19. Tsugawa, H. et al. A lipidome atlas in MS-DIAL 4. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0531-2 (2020).
20. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
21. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [2]Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, 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, 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]Section for Clinical Mass Spectrometry, Department of Congenital Disorders, Danish Center for Neonatal Screening, Statens Serum Institut, Copenhagen, Denmark. [9]RIKEN Center for Sustainable Resource Science, Yokohama, Japan. [10]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [11]Applied Bioinformatics, Department of Computer Science, University of Tübingen, Tübingen, Germany. [12]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany. [13]Department of Phytochemistry and Bioactive Natural Products, University of Geneva, Geneva, Switzerland. [14]Bruker Daltonics, Bremen, Germany. [15]Equipe PNAS, UMR 8038 CiTCoM CNRS, Faculté de Pharmacie de Paris, Université Paris Descartes, Paris, France. [16]Department of Physics and Chemistry, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil. [17]Institute of Chemistry, Technische Universität Berlin, Berlin, Germany. [18]School of Chemistry and Biochemistry, Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, GA, USA. [19]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia. [20]Waters Corporation, Milford, MA, USA. [21]Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic. [22]College of Pharmacy, Sookmyung Women's University, Seoul, Republic of Korea. [23]Univ. Grenoble Alpes, CNRS, Grenoble INP, CHU Grenoble Alpes, TIMC-IMAG, Grenoble, France. [24]Centro de Biodiversidad y Descubrimiento de Drogas, Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), Panama, Republic of Panama. [25]Department of Chemistry and Biochemistry, Department of Microbiology and Plant Biology and Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, OK, USA. [26]Nonlinear Dynamics, Milford, MA, USA. [27]Computational Biology Department, School of Computer Sciences, Carnegie Mellon University, Pittsburgh, PA, USA. [28]Department of Biological and Environmental Sciences, University of West Alabama, Livingston, AL, USA. [29]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [30]Bioinformatics and Scientific Data, Leibniz Institute of Plant Biochemistry, Halle, Germany. [31]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. [32]Laboratoire de Chimie des Produits Naturels, UMR CNRS SPE, Université de Corse Pascal Paoli, Corte, France. [33]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado, Denver, Aurora, CO, USA. [34]Whitehead Institute for Biomedical Research, Cambridge, MA, USA. [35]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. [36]School of Computing Science, University of Glasgow, Glasgow, UK. [37]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. [38]Bioinformatics Group, Wageningen University, Wageningen, the Netherlands. [39]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. [40]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, München, Germany. [41]College of Pharmacy, Kangwon National University, Chuncheon-si, Republic of Korea. [42]Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen, Germany. [43]Biomolecular Interactions, Max Planck Institute for Developmental Biology, Tübingen, Germany. [44]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [45]These authors contributed equally: Louis-Félix Nothias, Daniel Petras, Robin Schmid. ✉e-mail: miw023@ucsd.edu; pdorrestein@ucsd.edu

Methods

Development of FBMN. The FBMN method consists of two main steps: (1) LC–MS feature detection and alignment and (2) a dedicated molecular networking workflow on GNPS. Our first prototype for FBMN was developed with the Optimus workflow[7,14] using OpenMS tools[10]. Following the first step, two files are exported: a feature quantification table (.TXT format) and a MS^2 spectral summary (.MGF format). The feature quantification table contains information about LC–MS features across all considered samples including a unique identifier (feature ID) for each feature, *m/z* value, retention time and intensity. The MS^2 spectral summary contains a list of MS^2 spectra, with one representative MS^2 spectrum per feature. The mapping of information between the feature quantification table and the MS^2 spectral summary is stored in these files using the feature ID and scan number, respectively. This simple mapping enabled us to relate LC–MS feature information or statistically derived results to the molecular network nodes. This approach was also used for the integration of other tools with FBMN and does not require third-party software, as proposed previously[22,23]. Finally, the FBMN workflow also supports the mzTab-M format[6], a standardized output format designed for the report of metabolomics MS data-processing results. In this case, the mzTab-M file is used instead of the feature quantification table and requires the input of the mzML files instead of the MS^2 spectral summary file. Support for the mzTab-M format enables the possibility to perform FBMN with any existing and future processing tools that support this standardized format.

The FBMN workflow has been integrated into the GNPS ecosystem and thus benefits from the connection with other GNPS features, for example, the possibility to perform automatic MS^2 spectral library searching, the direct addition and curation of library entries, the search of a spectrum against public datasets with MASST[20] and the visualization of molecular networks directly in the web browser[24] or with Cytoscape[25]. The FBMN workflow is available on the GNPS platform (https://gnps.ucsd.edu/) via a web interface (Supplementary Fig. 2). Jobs are computed and stored on the computational infrastructure of the University of California San Diego Center for Computational Mass Spectrometry. Each finished job is saved in the private user space for future examination and has a permanent static link that enables data sharing and collaborative analyses. We strongly recommend the sharing of this static link along with data publications using GNPS workflows to facilitate accessibility and reproducibility of results. Instructions to perform FBMN with the supported tools and input file format requirements are provided in the GNPS documentation (https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking; Supplementary Fig. 3).

Processing mass spectrometry data for FBMN. FBMN supports the output from several feature detection and alignment processing software programs. Depending on the type and size of MS data and the intended user (for example, bioinformatician, mass spectrometrists and biologists), different software might be more appropriate. In general, tools with a graphical user interface (GUI), for example, MZmine[11], MS-DIAL[26], MetaboScape and Progenesis QI, are convenient for data visualization and empirical parameters optimization, but often have limited scalability, which might prevent their usage for large datasets (>500 files). For these large datasets, tools that were designed to operate on a cluster/cloud computer are preferred (XCMS[27], OpenMS[10] and, to some extent, MZmine). Regardless of the software or application, the processing steps and parameters used should be determined according to the recommendations from tool developers and experienced users through community feedback. Finally, automated optimization modules can be used to finely tune parameters, which is particularly valuable when using command-line interface tools[28,29]. While we acknowledge that many tools and configurations are available to analyze MS data, we provide a summary of processing steps on the supported tools in the FBMN documentation (https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking). These steps constitute an aggregation of institutional knowledge from tool developers and experienced tool users that do not encompass all possible applications, but rather provide a starting point for new users.

Generation of a representative MS^2 spectrum from a LC–MS^2 file. The selection of the representative MS^2 spectrum for detected features in the MS^2 spectral summary file is performed using several methods. Available in all tools supported, the default method ('most intense') uses the MS^2 spectrum with the highest precursor ion intensity or total ion current, in the specified *m/z* and retention time range, as the representative MS^2 spectrum for a LC–MS feature. Experimental spectral 'clustering' methods for the creation of the representative MS^2 spectrum in FBMN are implemented in MZmine, OpenMS and XCMS. The spectral clustering method implemented in MZmine ('merge' option in the GNPS/SIRIUS export modules) and OpenMS ('merge spectra' option in the GNPSExport tool) works as follows: for each LC–MS^2 feature, the purity of each fragmentation spectra is calculated with a function inspired by msPurity[30]. Briefly, adjacent MS^1 scans are examined to determine if other isobaric ions were co-fragmented. In these MS^1 scans, the ratio between the precursor ion intensity and the other isobaric ions in the precursor ion isolation range is calculated. Then, the purest MS^2 spectrum (highest purity score) is selected as the reference spectrum and pairwise comparison (cosine score) is computed between the purest spectrum and all the other MS^2 spectra for the feature. All MS^2 spectra reaching a cosine score threshold with the reference MS^2 spectra are then merged into one representative MS^2 spectrum. The mass accuracy, isolation width for the ion filtering and cosine score are defined by the user.

FBMN after MZmine processing. MZmine[11] is an open-source cross-platform software for MS data processing with an advanced GUI that enables the users to visually optimize parameters and examine the results of each processing step. Moreover, MZmine allows for the export of a batch file containing all the steps and parameters used in the processing, thus enabling reproducibility. To support FBMN in MZmine, the feature detection step (peak 'deconvolution module') was modified to provide the ability to pair a feature with its MS^2 scans using an *m/z* and retention time range defined by the user (Supplementary Fig. 4). Due to a new data structure and to support older projects (before version 2.38), an additional specific filtering module (group MS^2 scans with features) was developed to assign all MS^2 scans to the features of the existing peak list (for instructions see https://www.youtube.com/watch?v=EL5pmFvpTFE). Moreover, a GNPS export and direct submission module was created (Supplementary Fig. 5), which offers two modes: (1) export of the feature quantification table and the MS^2 spectral summary file and (2) direct FBMN analysis on the GNPS web platform (version 2.37+). The direct GNPS job submission generates all the files and uploads them together with an optional metadata table and default parameters (Supplementary Fig. 6) to the FBMN workflow on GNPS. By providing the user's GNPS login credentials (optional), a new job can be created in the personal user space (https://www.youtube.com/watch?v=vFcGG7T_44E&list=PL4L2Xw5k8ITzd9hx5XIP94vFPxj1sSafB&index=4&t=0s/). Otherwise, the user can be notified by email or redirected to the job web page after the submission. With the option 'most intense', the GNPSExport module uses the most intense MS^2 spectrum as a representative spectrum for each LC–MS^2 feature. When using the 'merge MS/MS' spectra option (version 2.40+), a representative high quality MS^2 spectrum is instead generated from all spectra and exported as a representative MS^2 spectrum. For detailed documentation, see https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-mzmine2/.

FBMN after OpenMS processing. OpenMS is an open-source cross-platform software specifically designed for the flexible and reproducible analysis of high-throughput MS data analysis, including more than 200 tools for common mass spectrometric data-processing tasks[10]. Building on our experience with the Optimus development, the integration of OpenMS and FBMN was achieved by creating a GNPSExport tool (TOPP tool) as a part of the OpenMS tool collection (https://github.com/OpenMS/OpenMS). A detailed description of the GNPSExport module and instructions for use with FBMN is available at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-openms/. Briefly, after running an OpenMS non-targeted metabolomics pipeline, the GNPSExport TOPP tool can be applied to the consensusXML file resulting from FeatureLinkerUnlabeledKD or FeatureLinkerUnlabeledQT tools (alignment step) and the corresponding mzML files. For each consensusElement (LC–MS^2 feature) in the consensusXML file, the GNPSExport tool generates one representative MS^2 spectrum that will be exported in the MS^2 spectral summary file (using either the option 'most intense' or 'merged spectra'). The TextExport tool is applied to the same consensusXML file to generate the feature quantification table. Note that GNPSExport requires the use of the IDMapper tool on the featureXML files (from the feature detection step) before feature linking, to associate MS^2 scans (peptide annotation in OpenMS terminology) with each feature. These MS^2 scans are used by the GNPSExport tool for the generation of the representative MS^2 spectrum. Additionally, the FileFilter has to be run on the consensusXML file before the GNPSExport, to remove consensusElements without associated MS^2 scans. The two exported files (feature quantification table and MS^2 spectral summary) can be directly used for FBMN analysis on GNPS. The OpenMS-GNPS workflow for metabolomics data processing was implemented as a Python wrapper around OpenMS TOPP tools (https://github.com/Bioinformatic-squad-DorresteinLab/openms-gnps-tools/) and released as a workflow (https://github.com/Bioinformatic-squad-DorresteinLab/openms-gnps-workflow/) on the GNPS/MassIVE web platform and could be run in OpenMS TOPPAS workflow[31]. The OpenMS and GNPS workflow can be accessed and run at https://proteomics2.ucsd.edu/ProteoSAFe/.

FBMN after XCMS processing. XCMS (for the most recent version, see https://github.com/sneumann/xcms/) is one of the most widely used software packages for processing of MS-based metabolomics data[27]. The integration of XCMS and FBMN is currently possible using a custom utility function 'formatSpectraForGNPS' to create the MS^2 spectral summary. This function is available on GitHub (https://github.com/jorainer/xcms-gnps-tools/) and is compatible with the CAMERA algorithm for isotopes and adduct annotation[32]. Representative XCMS R scripts in Markdown and Jupyter notebook formats are available in GitHub at https://github.com/DorresteinLaboratory/XCMS3_FeatureBasedMN/. The two exported files (feature quantification table and MS^2 spectral summary) can be directly used for FBMN analysis on GNPS. The detailed documentation is available at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-xcms3/.

FBMN after MS-DIAL processing. MS-DIAL is an open-source MS data-processing software[26] (available for Windows only; http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/). The integration of MS-DIAL and FBMN

has been made possible since v2.68 by exporting the 'alignment results' using the 'GNPSExport' option. In addition to LC–MS^2 data processing, MS-DIAL can process data from SWATH-MS^2 (data-independent LC–MS^2 acquisition) and IMS coupled to LC–MS[19]. The two files exported (MS^2 spectral summary and feature quantification table) can be directly used for FBMN analysis on GNPS. A video tutorial on the use of MS-DIAL for FBMN is available at https://www.youtube.com/watch?v=hxk40jwAkcc&t=7s/ and detailed documentation is available at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-ms-dial/.

FBMN after MetaboScape processing. MetaboScape is a commercial MS metabolomics data-processing software commercialized by Bruker and available on Windows. MetaboScape can perform feature detection, alignment, and annotation of non-targeted LC–MS^2 data acquired on Bruker mass spectrometers. Support for the processing of TIMS coupled to non-targeted LC–MS^2 (LC-TIMS-MS^2) was added in MetaboScape 4.0, which results in LC-TIMS-MS features. FBMN can be performed on LC–MS^2 or LC-TIMS-MS^2 data by exporting the feature quantification table and MS^2 spectral summary from the 'bucket table' using the 'export to GNPS format' function. These files can be uploaded to GNPS for FBMN analysis. Information from MetaboScape, such as the collision cross-section values or other spectral annotations, can be mapped into the molecular networks using Cytoscape[25]. The detailed documentation is available at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-metaboscape/.

FBMN after Progenesis QI processing. Progenesis QI is a commercial feature detection and alignment software developed by Nonlinear Dynamics (Waters) that is compatible with various proprietary and open MS data formats. Progenesis QI can perform feature detection, alignment and annotation of non-targeted LC–MS^2 data acquired either in data-dependent acquisition or data-independent analysis and can also utilize the IMS dimension. FBMN can be performed on any of these data types processed with Progenesis QI (v4.0), by exporting the feature quantification table (.CSV format) and MS^2 spectral summary (.MSP format). These two files can be exported from the 'identify compounds' submenu by using the functions 'export compound measurement' and 'export fragment database', respectively. These files can be uploaded to GNPS for FBMN analysis. Information from Progenesis QI, such as the collision cross-section values or other spectral annotations, can be mapped into the molecular networks using Cytoscape[25]. The detailed documentation is available at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-progenesisQI/.

Running time and scalability of the FBMN method. While the molecular networking computation part of the FBMN method is performed online on the GNPS web server (runtime of 5 min to several hours depending on the number of features and job parameters), the data-processing part has to be performed with the computational resources available to the researcher (laptop or desktop computer, workstation and cluster/cloud infrastructure). The computational cost of the data-processing part depends on (1) the software employed, (2) the number of samples in the dataset and (3) the parameters set. For this reason, the computational cost of the method in all scenarios cannot be comprehensively established. Nevertheless, our experience and feedback from the FBMN community with open-source tools such as MZmine or MS-DIAL showed that small datasets (<50 samples) can be processed in 10–60 min with a computer equipped with 8–16 GB RAM. Medium-sized datasets (>100 samples) require the use of a workstation equipped with 16–32 GB RAM, and large datasets (>500 samples) require 32–64 GB RAM. For very large datasets (>1,000 samples), it is currently recommended to use OpenMS or XCMS on a cluster/cloud infrastructure.

Integration with other computational mass spectrometry annotation tools. The .MGF file format is accepted by numerous computational MS annotation tools. The use of these annotation tools with the MS^2 spectral summary file enables (1) a reduction in the computation time compared to when using the unprocessed MS file(s) and (2) the subsequent mapping of these annotations to the molecular networks produced by the FBMN method. Some of these tools are directly available in the GNPS environment, including SIRIUS[15], DEREPLICATOR[16], Network Annotation Propagation (NAP)[33], MS2LDA[17], MolNetEnhancer[34] and Qemistree[18] (see below for details), as well as other software such as MetWork[35], CFM-ID[36] and MetFrag[37].

SIRIUS. SIRIUS is an advanced software for the computational annotation of small molecules from LC–MS^2 data[15]. It is capable of identifying compounds at the molecular formula[38], and annotating substructural, class[39] and structural levels[40] from the compound MS^2 spectra. The MS^2 spectral summary file (.MGF format) generated for the FBMN is compatible with SIRIUS, either running locally or with the dedicated GNPS workflow (https://ccms-ucsd.github.io/GNPSDocumentation/sirius/). Results from SIRIUS can be mapped on the molecular networks, which is essential since spectral library matching usually results frequently in a 1–5% annotation rate. In addition, a dedicated SIRIUS export function compatible with FBMN was created in MZmine and MetaboScape that exports a modified MS^2 spectral summary file with representative MS^1 and MS^2 spectra for each feature. The MS^1 spectra contains information about the detected isotopic pattern and can be used for automated detection of adduct/rare elements in SIRIUS, which restricts the molecular formula search space to speed up computation and improve molecular formula identification rates.

DEREPLICATOR. DEREPLICATOR[16], along with DEREPLICATOR VarQuest[41], is a collection of computational MS tools specialized in the annotation of peptidic small molecules often produced by microorganisms endowed with various biological activities. DEREPLICATOR tools can be run directly through the FBMN workflow results on GNPS. Alternatively, and for advanced parameterizing, the DEREPLICATOR workflow on GNPS accepts the MS^2 spectral summary file (.MGF format) as input and can directly map into the FBMNs (https://ccms-ucsd.github.io/GNPSDocumentation/dereplicator/).

Network annotation propagation. NAP[33] uses MetFrag/MetFusion[37,42] for the prediction of putative structures and the network topology to rerank structure predictions by propagating the expected structural similarity. NAP is available on GNPS as a dedicated workflow and offers direct support to FBMN (https://ccms-ucsd.github.io/GNPSDocumentation/nap/).

Unsupervised substructure annotation with MS2LDA. MS2LDA uses the latent Dirichlet allocation algorithm to mine for motifs (Mass2Motifs) of co-occurring fragments and neutral losses in MS2 spectra[17,43]. MS2LDA accepts the outputs of FBMN, allowing the direct mapping between MS2LDA annotations and the molecular networks. MS2LDA can be run on the GNPS web platform (https://ccms-ucsd.github.io/GNPSDocumentation/ms2lda/) and/or in the MS2LDA web application[43].

MolNetEnhancer. MolNetEnhancer[34] combines the outputs from molecular networking, substructure annotation with MS2LDA and other structural annotation tools, including SIRIUS, NAP and DEREPLICATOR, together with automated chemical classification through ClassyFire[44] into a single molecular network[34]. MolNetEnhancer accepts input files from classical MN and FBMN. The MolNetEnhancer accepts the MS^2 spectral summary from FBMN and is available through the GNPS web platform (https://ccms-ucsd.github.io/GNPSDocumentation/molnetenhancer/).

Qemistree. Qemistree is an MS data exploration strategy based on hierarchical organization of structural fingerprints[18] predicted from fragmentation spectra. The fingerprints are predicted with SIRIUS/CSI:FingerID[15,40] and the tree-based structure obtained allows the application of ecological tools to study the chemical composition. Qemistree is available as a GNPS workflow and directly accepts the outputs of FBMN (https://ccms-ucsd.github.io/GNPSDocumentation/qemistree/).

FBMN applications. *FBMN makes it possible to resolve isomers in a drug lead discovery effort.* The examination of LC–MS^2 data (MSV000080502) from the *E. dendroides* plant extract showed the presence of numerous chromatographic peaks for ions in the range of m/z 500–900, corresponding to diterpene ester derivatives. These specialized metabolites consist of a polyhydroxylated diterpene core acylated with various acidic moieties, which are typically found as positional isomers based on their acylation pattern. The EIC for the ion m/z 589.31 in the *E. dendroides* extract data (Supplementary Fig. 7) shows the presence of at least seven distinct LC–MS peaks between 24.5 min and 27.3 min, including five peaks with associated MS^2 spectra. The analysis of the extract and the fractions where these molecules were originally isolated (fractions 13 and 14) with classical MN resulted in a molecular network with two nodes for the m/z 589.31 ions (Fig. 2a and Supplementary Fig. 8). These MS^2 spectra (cluster index of 5352 and 5354) resulted from the merging of 96 fragmentation spectra spanning 23.6 min to 26.5 min by MS-Cluster (Fig. 2b and Supplementary Fig. 9). Close examination of the clustered spectra revealed that, while all MS^2 spectra for the precursor m/z 589.31 present fragment ions m/z 501.26, 423.21, 335.16 and 295.17, three distinct spectral types could be established based on the relative intensities of the ions (Supplementary Fig. 10). FBMN of the dataset with MZmine processing (see the GNPS job) enabled the differentiation of the MS^2 spectra of seven isomers (Fig. 2b and Supplementary Fig. 11; see the molecular network view). A detailed discussion of the differences observed between the two methods can be found in the Supplementary Note 1 and Supplementary Table 1. Interestingly, in the original study[7], OpenMS was used for FBMN and resulted in the observation of three different positional isomers instead of seven, which shows that different processing methods and/or parameters can lead to different results with FBMN. These three isomers were subsequently isolated and differed by the position of one double bond on the C-12 acyl chain or from carbon C-4 configuration[7]. Because FBMN connects the accurate relative abundance of the ions across the fractions and the molecular networks, it allowed us to create bioactivity-based molecular networks[7], which were used to predict and target potentially antiviral compounds. For a detailed description of the extraction, MS analysis and structural elucidation, see the original paper[7]. The MZmine project and parameters used can be accessed on the MassIVE submission (MSV000080502).

FBMN resolves isomers in large-scale metabolomics studies. FBMN was applied on a cohort of the AGP, a citizen-science research project that enabled the observation of commendamide in humans, along with other new *N*-acyl amide derivatives using molecular networking[9]. Commendamide is a recently discovered bacterial *N*-acyl amide that has been shown to modulate host metabolism via G-protein-coupled receptors in the murine intestinal tract[45].

The use of FBMN for the AGP data (Fig. 2d) allowed the observation of two additional commendamide isomers (*m/z* 330.26) and of an analog, *N*-(hydroxyheptadecanoyl)glycine (*m/z* 344.28), while classical MN resulted in the observation of one single consensus spectrum for all the isomers (Fig. 2c). In addition, FBMN allowed the observation of a putative commendamide derivative *N*-(dehydrohexadecanoyl)glycine (CCMSLIB00005436498; Supplementary Fig. 12) in the commendamide molecular network. The sample collection and MS acquisition methods are described in the original manuscript[9]. The data were downloaded from MassIVE (MSV000080186) and processed with MZmine (v2.37). The MZmine project data, along with parameters and export files, were deposited to the MassIVE repository (MSV000084095). The chromatograms for *m/z* 330.26 and *m/z* 344.28 displayed in Fig. 2c,d are from samples 43076_P3_RB9_01_314.mzML and 38131_P5_RA4_01_538.mzML, respectively. Chromatograms were exported with MZmine. The results were exported with the 'Export for/submit to GNPS' module for FBMN analysis on GNPS. The corresponding job can be accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0a8432b5891a48d7ad8459ba4a89969f (only logged-in users can access all the input files). The mzML files used for the classical MN job are available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3c27e43d908c4044bace405cc394cd25.

FBMN reduces spectral redundancy and de-obfuscates spectral similarity relationships: the case of EDTA. The benefit of using FBMN can be illustrated with the metal chelating agent EDTA, widely used in beauty products, food and scientific protocols. A search for its occurrences in public spectral datasets with the MS search tool (MASST)[20] showed that it is frequently observed in plasma samples where it is used during the sample preparation. Using classical MN, we analyzed a public dataset of human plasma where EDTA was observed (MSV00008263; see Supplementary Note 2 for protocol and MS parameters); the results showed that the EDTA ions are found in two molecular networks: one network consisting of $[M+H]^+$ spectra and the other of $[M+Na]^+$ spectra. Interestingly, each of these networks have one node with a large number of clustered spectra (node 91,205 for 4,655 spectra and node 116,470 for 571 spectra), yet EDTA ions are represented by multiple nodes, although these nodes have the same precursor ion mass and retention time. Detailed analyses showed that while the median pairwise cosine values between EDTA spectra were high (median values of 0.93 and 0.94), the spectra were not clustering into a single node. Examination of the multiple fragmentation spectra for EDTA ions showed that (1) some are chimeric spectra that are 'contaminated' by fragment ions produced by coeluting isobaric ions and (2) other spectra were dominated by low-intensity fragment ions resulting from MS^2 spectra acquired at low intensity. The method of FBMN was applied on that same dataset using the OpenMS-GNPS workflow (see the job), and the results showed that it efficiently reduced the appearance of these redundant node patterns from the same molecule (see the FBMN job; Fig. 2f), both for the molecular networks containing the $[M+H]^+$ and $[M+Na]^+$ spectra. FBMN recovered the molecular similarity of in-source fragments observed for EDTA, which were not displayed with classical MN, as they now fall within the top-K rank (typically set to ten) of MS^2 spectral similarity considered in the network topology. The parameters used for OpenMS tools can be accessed in the OpenMS-GNPS job (see the job). OpenMS v2.4.0 was used[10].

FBMN enables the use of relative quantification in the molecular networks. While classical MN uses the spectral count or the sum of precursor ion intensity to estimate the ion abundance, FBMN uses the accurate ion intensities obtained from LC–MS feature detection. The FBMN method brings in ion abundance across all samples by using the value of the chromatographic peak area or peak height as determined by the LC–MS feature detection and alignment software. Multiple dilutions ($n=5$) of the NIST 1950 serum reference metabolome sample[46] were analyzed by LC–MS^2 (three independent experiments per sample) on an Orbitrap mass spectrometer (Q Exactive, Thermo Fisher) and processed with MZmine or OpenMS. OLS linear regression analysis between the feature intensity and expected relative abundance in samples of known dilution factor (serial dilution) showed improved linearity of the relative quantification with FBMN compared to classical MN (Fig. 2h, Supplementary Note 3 and Supplementary Figs. 13–19). The sample preparation and MS methods are described in Supplementary Note 3. The files along with the parameters for MZmine are available on the following MassIVE repository (MSV000084092). The OLS analysis was performed with Python 2 (v2.7.15) using the linear regression function of the sklearn package (v0.20.1)[33]. The analysis is available as a Jupyter notebook at https://github.com/lfnothias/FeatureBasedMolecularNetworking_RelativeQuantEval/. The molecular networking jobs and parameters can be accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=daf3f0d7cec94104b2c9001739964c31 for classical MN, https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f443cad083be4979aedd2af0f97b9fe9 for FBMN with MZmine, https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d6a430cc6da2458f8135ae76126eb763 for GNPS-OpenMS and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=53bcfa39fa674c749b4da0b613df1b8d for FBMN with OpenMS.

FBMN enables molecular networking with IMS. The sample NIST 1950 serum[46] was analyzed using a timsTOF Pro (Bruker Daltonics) in data-dependent acquisition mode using PASEF[47]. The data were then processed with MetaboScape (v5.0), and the results were exported for FBMN analysis on GNPS. The MS acquisition method, data and parameters used for the processing were deposited on MassIVE (MSV000084402). Classical MN data were annotated with the GNPS[4], NIST17 and LipidBlast[48] spectral libraries (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2adc2cf33c646548798d0e285197a96). Lipid annotation in MetaboScape was performed using SimLipid (v6.04; Premier Biosoft) and mapped to the FBMN (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0d89db67b0974939a91cb7d5bfe87072). The molecular networks were visualized with Cytoscape (v3.7.1; ref. [25]), and the results are presented in Supplementary Fig. 1.

Large dataset processing with OpenMS and XCMS. The processing of very large metabolomics datasets (>1,000 samples) is limited by the scalability of existing LC–MS feature detection tools, especially those based on a GUI (such as MZmine and MS-DIAL). We showed that, with specific peak-picking parameters, the use of XCMS or OpenMS enables the processing of large metabolomics studies for FBMN (MSV000080030; approximately 2,000 samples; Supplementary Note 4, Supplementary Table 2 and Supplementary Fig. 19).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The LC–MS^2 data for the *E. dendroides* dataset, along with the MZmine project and parameters used, can be accessed on the MassIVE submission (MSV000080502; Creative Commons CC0 1.0 Universal license). The classical MN and FBMN jobs can be accessed via the GNPS website at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=189e8bf16af145758b0a900f1c44ff4a and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=672d0a5372384cff8c47297c2048d789, respectively. LC–MS^2 data for the AGP were downloaded from MassIVE (MSV000080186; Creative Commons CC0 1.0 Universal license) and processed with MZmine (v2.37). The MZmine project along with parameters and export files were deposited (MSV000084095; Creative Commons CC0 1.0 Universal license). The classical MN and FBMN jobs can be accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3c27e43d908c4044bace405cc394cd25 and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0a8432b5891a48d7ad8459ba4a89969f, respectively. The LC–MS^2 data for the EDTA case are available on the MassIVE submission (MSV00008263; Creative Commons CC0 1.0 Universal license). The classical MN job can be accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fbac1a5061ba4ad683a284ef55d45df6. The OpenMS and FBMN jobs are available at https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=83a0a417a49b4b76b61e9a8191a6ea2d at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8f40420c11694cf9ab06fdf7a5a4c53b, respectively.
The MS acquisition method, data and parameters used for the processing of the serum analysis with the timsTOF mass spectrometer were deposited (MSV000084402). Classical MN and FBMN jobs can be accessed at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2adc2cf33c646548798d0e285197a96 and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0d89db67b0974939a91cb7d5bfe87072, respectively.

Code availability

The FBMN workflow is available as a web interface on the GNPS web platform (https://gnps-quickstart.ucsd.edu/featurebasednetworking/). The workflow code is open source and available on GitHub (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/feature-based-molecular-networking/). It is released under the license of The Regents of the University of California San Diego and free for non-profit research (https://github.com/CCMS-UCSD/GNPS_Workflows/blob/master/LICENSE/). The workflow was written in Python (v3.7) and deployed with the ProteoSAFE workflow manager used by GNPS (https://proteomics.ucsd.edu/Software/ProteoSAFe/). We also provide documentation, support, example files and additional information on the GNPS documentation website (https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/). The source code of the GNPSExport module in MZmine is available at https://github.com/mzmine/mzmine2/ under the GNU General Public License. The source code of the GNPSExport tool in OpenMS is available at https://github.com/Bioinformatic-squad-DorresteinLab/OpenMS/under the BSD license. The source code for the GNPSExport custom function for XCMS is available at https://github.com/jorainer/xcms-gnps-tools/ under the GNU General Public License.

References

22. Winnikoff, J. R., Glukhov, E., Watrous, J., Dorrestein, P. C. & Gerwick, W. H. Quantitative molecular networking to profile marine cyanobacterial metabolomes. *J. Antibiot.* **67**, 105–112 (2014).
23. Olivon, F., Grelier, G., Roussi, F., Litaudon, M. & Touboul, D. MZmine 2 data-preprocessing to enhance molecular networking reliability. *Anal. Chem.* **89**, 7836–7840 (2017).

24. Ono, K., Demchak, B. & Ideker, T. Cytoscape tools for the web age: D3.js and Cytoscape.js exporters. *F1000Res.* **3**, 143 (2014).
25. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
26. Tsugawa, H. et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
27. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).
28. Libiseller, G. et al. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16**, 118 (2015).
29. McLean, C. & Kujawinski, E. B. AutoTuner: high fidelity and robust parameter selection for metabolomics data processing. *Anal. Chem.* **92**, 5724–5732 (2020).
30. Lawson, T. N. et al. msPurity: automated evaluation of precursor ion purity for mass spectrometry-based fragmentation in metabolomics. *Anal. Chem.* **89**, 2432–2439 (2017).
31. Junker, J. et al. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *J. Proteome Res.* **11**, 3914–3920 (2012).
32. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography–mass spectrometry datasets. *Anal. Chem.* **84**, 283–289 (2012).
33. da Silva, R. R. et al. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **14**, e1006089 (2018).
34. Ernst, M. et al. MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* **9**, 144 (2019).
35. Beauxis, Y. & Genta-Jouve, G. Metwork: a web server for natural products anticipation. *Bioinformatics* **35**, 1795–1796 (2019).
36. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).
37. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **8**, 1–16 (2016).
38. Ludwig, M. et al. ZODIAC: database-independent molecular formula annotation using Gibbs sampling reveals unknown small molecules. Preprint at *bioRxiv* https://doi.org/10.1101/842740 (2019).
39. Dührkop, K. et al. Classes for the masses: systematic classification of unknowns using fragmentation spectra. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.17.046672 (2020).
40. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
41. Gurevich, A. et al. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* **3**, 319–327 (2018).
42. Gerlich, M. & Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **48**, 291–298 (2013).
43. Wandy, J. et al. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **34**, 317–318 (2017).
44. Feunang, Y. D. et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
45. Cohen, L. J. et al. Commensal bacteria make GPCR ligands that mimic human signalling molecules. *Nature* **549**, 48–53 (2017).
46. Simón-Manso, Y. et al. Metabolite profiling of a NIST standard reference material for human plasma (SRM 1950): GC–MS, LC–MS, NMR and clinical laboratory analyses, libraries and web-based resources. *Anal. Chem.* **85**, 11725–11731 (2013).
47. Meier, F. et al. Online parallel accumulation-serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).
48. Kind, T. et al. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods* **10**, 755–758 (2013).

Acknowledgements

We gratefully acknowledge financial support from the U.S. National Institutes of Health (NIH) for the Center for Computational Mass Spectrometry grant (P41 GM103484), the reuse of metabolomics data (R03 CA211211) and the tools for rapid and accurate structure elucidation of natural products (R01 GM107550 and U19 AG063744 01) to P.C.D.; the NIH grants R24GM127667 and 1R01LM013115 and a National Science Foundation (NSF) award (ABI 1759980) to N.B.; the European Union's Horizon 2020 grants 704786 (MSCA-GF to L.-F.N.), 634402 and 777222 (T.A. and I.P.) and a European Research Council Consolidator grant METACELL (T.A.). L.-F.N. was supported by the Center for Microbiome Innovation from the University of California San Diego (support program award). D.P. was supported by the German Research Foundation (DFG; grant no. PE 2600/1). S.N. acknowledges funding from Bundesministerium für Bildung und Forschung (FKZ 031L0107) and the European Commission (EC654241). R.S. acknowledges funding by the German Chemical Industry Fund (FCI) fellowship. H.T. was supported by KAKENHI (18H02432 and 18K19155). A.M.C.-R. was supported by an NSF grant (IOS-1656481) to P.C.D. O.A. acknowledges funding from the Bundesministerium für Ernährung und Landwirtschaft (FKZ 2816501214), the Bundesministerium für Wirtschaft und Energie (FKZ AiF18475N), the Bundesministerium für Bildung und Forschung (FKZ 031A430C) and the European Commission (823839), which also supported F.A. and O.K. S.B. acknowledges funding from Deutsche Forschungsgemeinschaft (BO 1910/20). M.L. was supported by the Deutsche Forschungsgemeinschaft (BO 1910/20-1). J.J.J.v.d.H. was supported by an Accelerating Scientific Discoveries Grant funded by the Netherlands eScience Center (NLeSC; no. ASDI.2017.030). S.N. acknowledges funding from BMBF (grant no. 031L0107) and the European Commission (PhenoMeNal grant EC654241). F.V. was funded by the Department of Navy, Office of Naval Research Multidisciplinary University Research Initiative (MURI) award (N00014-15-1-2809). V.V.P. acknowledges support from the ALSAM Foundation (Therapeutic Innovation Award and L.S. Skaggs Professorship) and the NIH (R35 GM128690). T.P. is a Simons Foundation Fellow of the Helen Hay Whitney Foundation. Z.K. was supported by the project International Mobility of Researchers (CZ.02.2.69/0.0/0.0/16_027/0007990). A.K.J. was supported by the American Society for Mass Spectrometry (Postdoctoral Career Development Award). K.B.K. was supported by a grant from the National Research Foundation (NRF) of Korea (MSIT; NRF-2019R1F1A1058068). H.Y. was supported by the Basic Science Research Program through the NRF grant (NRF- 2018R1C1B6002574). A.L.G. was supported by Vaincre la mucoviscidose and Association Grégory Lemarchal. The work of H.M. was supported by a research fellowship from the Alfred P. Sloan Foundation and an NIH New Innovator Award (DP2GM137413). The authors thank N. Hoffman for maintaining the mzTab-M format. Finally, we acknowledge the continuous feedback from the GNPS community and the contribution of all researchers and associated institutions who are committed to depositing their MS data in public repositories.

Author contributions

L.-F.N., D.P., M.W. and P.C.D. conceived the method and supervised its implementation and wrote the manuscript. I.P., L.-F.N., M.E. and T.A. created the FBMN prototype in Optimus. M.W., L.-F.N., D.P. and Z.Z. created the FBMN workflow on GNPS. R.S., L.-F.N, M.W., D.P., A.K., M.F., Z.Z., A.S. and T.P. developed the GNPSExport module in MZmine. K.D., A.K., M.L. and S.B. developed the spectral clustering algorithm and SIRIUS export in MZmine. A.S. and L.-F.N. created the GNPSExport tool in OpenMS, with guidance from F.A., O.A. and O.K. J.R. and M.W. created the XCMS export tool. H.T., M.W. and L.-F.N. enabled the integration with MS-DIAL. L.-F.N., A.B., H.N., F.Z. and T.D. enabled the integration with MetaboScape. M.W., G.I., B.S., S.W.M. and J.M. enabled the integration with Progenesis QI. F.V. performed the MS for the plasma and NIST1950SRM samples. A.A.A. performed the MS for the AGP samples. A.K.J., L.-F.N. and A.T. analyzed the results of the plasma samples. J.R. and L.-F.N. performed the XCMS processing of the forensic dataset. L.-F.N. and M.W. created the FBMN documentation. The serum sample analysis in PASEF mode and the data processing with MetaboScape were performed by F.Z., and the subsequent FBMN analysis was performed by L.-F.N. D.P., L.-F.N. and R.d.S. created the MZmine documentation. K.B.K. and H.Y. created the MS-DIAL documentation. F.V., J.M.G., K.W. and A.K.J. prepared the MS-DIAL video tutorial. M.W., R.S. and D.P. prepared the MZmine video tutorials. M.E., R.d.S., J.R., O.M. and S.N. created the XCMS documentation. L.-F.N. and A.S. created the OpenMS documentation. L.-F.N., N.H.N. and T.D. created the MetaboScape documentation. A.M.C.-R. and L.-I.M. documented the FBMN interface workflow. M.N.-E., I.K. and C.M. created the Cytoscape documentation. H.M., A.G., M.W. and L.-F.N. made the integration with DEREPLICATOR. M.W., J.J.J.v.d.H., M.E. and S.R. made the integration with MS2LDA. R.d.S made the integration with NAP. M.M., N.B., X.C., V.V.P., J.P., N.G., R.A.Q., A.A.A., Z.K. and S.N. tested and provided suggestions on how to improve the methods. J.J.J.v.d.H., T.A., A.K.J., T.P., V.V.P., A.L.G., L.-I.M., P.-M.A., S.B. and S.N. improved the manuscript. All authors contributed to the final manuscript.

Competing interests

P.C.D. is a scientific advisor for Sirenas, Galileio and Cybele and scientific advisor and founder of Ometa labs and Enveda. M.W. is a founder of Ometa Labs. T.P. is a consultant for Ginkgo Bioworks. A.A.A. is a consultant for Ometa Labs. T.A. is on the Scientific Advisory Board of SCiLS, a Bruker company. K.D., M.L., M.F. and S.B. are founders of Bright Giant. A.B., S.W.M., H.N. and F.Z. are employees of Bruker Daltonics. G.I., J.M. and B.S. are employees of Waters.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0933-6.

Correspondence and requests for materials should be addressed to M.W. or P.C.D.

Peer review information Arunima Singh was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Ming Wang and Pieter C. Dorrestein

Last updated by author(s): May 6, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The mass spectrometry data, acquisition method files, and results files used in this study were deposited to GNPS/MassIVE repository. The accession numbers are: MSV000080502, MSV000080186, MSV00008263, MSV000084092, MSV000084402, MSV000080030.
Data analysis	The results for the molecular networking jobs are accessible via their web link provided in the manuscript. The parameters used for the feature detection/tools are described in the manuscript and deposited to GNPS/MassIVE and ar All softwares used is this study are open source and free to use, except for MetaboScape (Bruker Daltonics GmbH) and Progenesis QI (Nonlinear dynamics, Waters Corporation). The FBMN workflow is available as a web-interface on the GNPS web platform (https://gnps-quickstart.ucsd.edu/featurebasednetworking). The workflow code is open source and available on GitHub (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/feature-based-molecular-networking). It is released under the licence of The Regents of the University of California and free for non-profit research (https://github.com/CCMS-UCSD/GNPS_Workflows/blob/master/LICENSE). The workflow was written in Python (ver. 3.7) and deployed with the ProteoSAFE workflow manager employed by GNPS (http://proteomics.ucsd.edu/Software/ProteoSAFe/). We also provide documentation, support, example files, and additional information on the GNPS documentation website (https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/). The source code of the GNPSExport module in MZmine is available at (https://github.com/mzmine/mzmine2) under the GNU General Public License. The source code of the GNPSExport tool in OpenMS is available at (https://github.com/Bioinformatic-squad-DorresteinLab/OpenMS) under the BSD licence. The source code for the GNPSExport custom function for XCMS is available at https://github.com/jorainer/xcms-gnps-tools under the GNU General Public License. The custom code used for the evaluation of of relative quantification between classical molecular networking and feature-based molecular networking (python version 2.7.15) is available as jupyter notebook at https://github.com/lfnothias/FeatureBasedMolecularNetworking_RelativeQuantEval

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the data were deposited on the GNPS/MassIVE public repository (https://gnps.ucsd.edu) under the following accession numbers: MSV000080502, MSV000080186, MSV00008263, MSV000084092, MSV000084402, MSV000080030.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the evaluation of the quantitative performance, 5 different dilutions were analyzed.
Data exclusions	No data were excluded from the study
Replication	For the evaluation of the quantitative performance, 3 independent LC-MS2 experiments were performed per sample.
Randomization	For the evaluation of the quantitative performance, the acquisition of LC-MS2 data was randomized
Blinding	Doesn't apply to the study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging