A Universal Language for Finding Mass Spectrometry Data Patterns

Alan K. Jarmusch[1*], Allegra T. Aron[2*], Daniel Petras[3], Vanessa V. Phelan[4], Wout Bittremieux[2], Deepa D. Acharya[5], Mohammed M. A. Ahmed[6,7], Anelize Bauermeister[2], Matthew J. Bertin[8], Paul D. Boudreau[9], Ricardo M. Borges[10], Benjamin P. Bowen[11,12], Christopher J. Brown[13], Fernanda O. Chagas[10], Kenneth D. Clevenger[14], Mario S. P. Correia[15], William J. Crandall[16], Max Crüsemann[17], Tito Damiani[18], Oliver Fiehn[19], Neha Garg[20], William H Gerwick[21,2], Jeffrey R. Gilbert[13], Daniel Globisch[15], Paulo Wender P. Gomes[2], Steffen Heuckeroth[22], C. Andrew James[23], Scott A. Jarmusch[24], Sarvar A. Kakhkhorov[25], Kyo Bin Kang[26], Roland D Kersten[27], Hyunwoo Kim[28], Riley D. Kirk[29], Oliver Kohlbacher[30], Eftychia E. Kontou[31], Ken Liu[16], Itzel Lizama-Chamu[32], Gordon T. Luu[32], Tal Luzzatto Knaan[33], Michael T. Marty[34], Andrew C. McAvoy[35], Laura-Isobel McCall[36], Osama G. Mohamed[37,38], Omri Nahor[33], Timo H.J. Niedermeyer[39], Trent R. Northen[40,41], Kirsten E. Overdahl[42], Tomáš Pluskal[18], Johannes Rainer[43], Raphael Reher[39], Elys Rodriguez[19], Timo T. Sachsenberg[44], Laura M. Sanchez[32], Robin Schmid[2], Cole Stevens[45], Zhenyu Tian[46], Ashootosh Tripathi[38,47], Hiroshi Tsugawa[48,49,50], Kozo Nishida[48], Yuki Matsuzawa[48], Justin J.J. van der Hooft[51,52], Andrea Vicini[43], Axel Walter[44], Tilmann Weber[53], Quanbo Xiong[54], Tao Xu[55], Haoqi Nina Zhao[2], Pieter C. Dorrestein[2], Mingxun Wang[56]

[1]Immunity, Inflammation, and Disease Laboratory, Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, United States, [2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, 92093, United States, [3]Functional Metabolomics Lab, CMFI Cluster of Excellence, University of Tuebingen, University of Tuebingen, Tuebingen, Germany, [4]Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, Aurora, United States, [5]Biologicals and Natural Products Discovery, Crop Protection R & D, Corteva Agriscences, 9330 Zionsville Rd, Indianapolis, IN, 46268, United States, [6]BioMolecular Sciences, School of Pharmacy, University of Mississippi, 408 Faser Hall, University, Mississippi, 38677-1848, United States, [7]Pharmacognosy, Faculty of Pharmacy, Al-Azhar University, 1 El Mokhayam El Daem St., Nasr City, Cairo, 11371, Egypt, [8]Department of Biomedical and Pharmaceutical Sciences, College of Pharmacy, University of Rhode Island, Kingston, Rhode Island, 02881, United States, [9]BioMolecular Sciences, School of Pharmacy, University of Mississippi, 405 Faser Hall, University, Mississippi, 38677-1848, United States, [10]Walter Mors Institute of Research on Natural Products, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [11]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Lab, Berkeley, California, United States, [12]The Joint Genome Institute, Lawrence Berkeley National Lab. One Cyclotron Road. Berkeley, CA, 94720, Berkeley, California, United States, [13]Mass Spectrometry Center of Expertise, Regulatory and Stewardship, Mass Spectrometry Center of Expertise, Regulatory and Stewardship, Corteva Agriscences, 9330 Zionsville Road, Indianapolis, Indiana, 46268, United States, [14]Biologicals and Natural Products, Crop Protection R & D, Corteva Agriscences, Indianapolis, United States, [15]Department of Chemistry - BMC, Science for Life Laboratory, Uppsala University, Uppsala, Sweden, [16]Clinical Biomarkers Laboratory, School of Medicine, Emory University, Atlanta, GA, 30332, United States, [17]Institute of Pharmaceutical Biology,

University of Bonn, Nussallee 6, Bonn, 53115, Germany, [18]Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo nám. 2, Praha 6, 160 00, Czech Republic, [19]Department of Chemistry, University of California Davis, Davis, United States, [20]School of Chemistry and Biochemistry, Center for Microbial Dynamics and Infection, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta, GA, 30332, United States, [21]Scripps Institution of Oceanography and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, 92093, United States, [22]Institute of Inorganic and Analytical Chemistry, University of Münster, Corrensstraße 48, Münster, 48149, Germany, [23]Center for Urban Waters, University of Washington, Tacoma, United States, [24]Department of Biotechnology and Biomedicine, Technical University of Denmark, Søltofts Plads 221, Kongens Lyngby, Denmark, [25]Laboratory of Physical and Chemical Methods of Research, Center for Advanced Technologies, Tashkent, 100174, Uzbekistan, [26]College of Pharmacy, Sookmyung Women's University, Seoul, Republic of Korea, [27]Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, Ann Arbor, 428 Church Street, Ann Arbor, MI, 48104, United States, [28]College of Pharmacy, Dongguk University, 32 Dongguk-ro, Goyang, 10326, Republic of Korea, [29]College of Pharmacy, University of Rhode Island, Kingston, RI, 02881, United States, [30]Applied Bioinformatics, Department of Computer Science, University of Tuebingen, University of Tuebingen; Institute for Bioinformatics and Medical Informatics, University of Tuebingen; Institute for Translational Bioinformatics, University Hospital Tuebingen, Tübingen, 72076, Germany, [31]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet 220, building 220, Kongens Lyngby, 2800, Denmark, [32]Department of Chemistry and Biochemistry, UC Santa Cruz, 1156 High St, Santa Cruz, CA, 95064, United States, [33]Department of Marine Biology, The Leon H. Cherney School of Marine Sciences, University of Haifa, 199 Aba Koushy Ave, Haifa, 3498838, Israel, [34]Department of Chemstry and Biochemistry, University of Arizona, 1306 E. University Blvd., Tucson, AZ, 85721, United States, [35]School of Chemistry and Biochemistry,, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta, GA, 30332, United States, [36]Department of Chemistry and Biochemistry, Department of Microbiology and Plant Biology, Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, OK, 73019, United States, [37]Pharmacognosy Department, Faculty of Pharmacy, Cairo University, Kasr el-Aini St., Cairo, 11562, Egypt, [38]Natural Products Discovery Core, Life Sciences Institute, University of Michigan, Ann Arbor, MI, 48109, United States, [39]Institute of Pharmacy, Martin-Luther-University Halle-Wittenberg, Hoher Weg 8, Halle (Saale), 06114, Germany, [40]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Lab, Berkeley, United States, [41]The Joint Genome Institute, Lawrence Berkeley National Lab. One Cyclotron Road. Berkeley, CA, 94720, Berkeley, United States, [42]Immunity, Inflammation, and Disease Laboratory, Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, [43]Institute for Biomedicine (Affiliated to the University of Lübeck), Eurac Research, Bolzano, 39100, Italy, [44]Applied Bioinformatics, Department of Computer Science, University of Tuebingen, University of Tuebingen, Tübingen, Germany, [45]Department of BioMolecular Sciences, School of Pharmacy, University of Mississippi, University, MS, 38677-1848, United States, [46]Chemistry and Chemical Biology, Northeastern University, Boston, MA, 02115, United States, [47]Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, Ann Arbor, MI, 48109, United States, [48]Department of Biotechnology and Life Science, Tokyo

University of Agriculture and Technology, 2-24-16 Nakamachi, Koganei, Tokyo, 184-8588, Japan, [49]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan, [50]RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan, [51]Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, Wageningen, 6708 PB, the Netherlands, [52]Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg, 2006, South Africa, [53]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, building 220, Kongens Lyngby, 2800, Denmark, [54]Crop Protection R & D, Corteva Agriscences, 9330 Zionsville Road, Indianapolis, IN, 46268, United States, [55]Data Science and Bioinformatics, Corteva Agriscences, Dublin, United States, [56]Department of Computer Science, University of California Riverside, Riverside, CA, 92521, United States

* These authors contributed equally to the work

Abstract

Even though raw mass spectrometry data is information rich, the vast majority of the data is underutilized. The ability to interrogate these rich datasets is handicapped by the limited capability and flexibility of existing software. We introduce the Mass Spec Query Language (MassQL) that addresses these issues by enabling an expressive set of mass spectrometry patterns to be queried directly from raw data. MassQL is an open-source mass spectrometry query language for flexible and mass spectrometer manufacturer-independent mining of MS data. We envision the flexibility, scalability, and ease of use of MassQL will empower the mass spectrometry community to take fuller advantage of their mass spectrometry data and accelerate discoveries.

Main Text

Despite the widespread use of mass spectrometry (MS) in science to characterize proteins, peptides, polymers, small molecules, and nucleic acids, it remains difficult for scientists to search for known patterns of chemical classes within and across MS data sets. The variety of applications of MS and the diversity of class-specific chemical patterns makes automation difficult. Searches for specific chemicals or specific chemical classes within MS data are performed manually or using specialized software tools. These tools are generally designed to search for a specific pattern, to search data within a single research domain[1], or were created by computational scientists for their own use that present a high barrier for reuse[2]. This inability of the wider MS userbase to mine MS data rapidly across different MS datasets has left potential discoveries hidden in the data. To address the need for universal searching of MS data, we created MassQL, an open-source MS query language for flexible and mass spectrometer manufacturer-independent mining of MS data.

Based upon the concept that MS data captures unique characteristics of chemical structures, such as isotopic patterns (e.g., bromination), diagnostic fragmentation (e.g., product ion of sulfur trioxide), and neutral loss (e.g., loss of sugar moieties), MassQL implements common MS terminology to build a consensus vocabulary to search for MS patterns in a single mass spectrometry run up to entire data repositories (**Fig. 1a, 1b**). The MassQL language encompasses formal definition of common MS terms, including MS1 patterns, such as precursor ion *m/z* or isotopic patterns, and MS/MS fragmentation patterns (including support for data-dependent acquisition and data-independent acquisition, e.g., SWATH and MS^e), as well as terms for separation methods, including retention time and ion mobility drift time. Since this terminology is agnostic to the type of mass spectrometry data acquisition used, the MassQL querying language is compatible with all MS data. Additionally, MassQL query options include parameters for setting user-defined tolerances, such as ion intensities and mass accuracies, and boolean conjunctions, such as AND/OR, can be used to create more complex pattern queries marking inclusion or exclusion criteria (**Fig. 1b**). To facilitate adoption of MassQL, we made use of community input to establish commonly used terms required for a succinct language that could be readily shared and reused and new terms can be defined, which enables grammar and syntax evolution of MassQL to maintain compatibility of queries to advancing MS technologies. The resulting MassQL language provides users the flexibility and expressiveness

to query simple and complex MS patterns within their data and across public data regardless of their expertise in computational MS. Community members have written and applied MassQL queries to their own research (**SI Notes 1.1 - 5.2**) using MS/MS spectral information (**SI Notes 1.1 - 2.8**), precursor isotopic patterns (**SI Notes 3.1 - 3.9**), drift time (**SI Notes 4.1 - 4.2**), and other parameters (**SI Notes 5.1 - 5.2**) for querying MS data sets for chemically and biologically relevant molecules, such as identification of iron-binding compounds (**SI Note 3.3.1 and 3.3.2)** and distinguishing glycoconjugates (**SI Note 2.8**), among others.

MassQL is supported natively in a variety of community supported MS software, including MZmine[3], pyOpenMS[4], MS-DIAL[5], UniDec[6], GNPS[7], and the GNPS Dashboard[8] (**Fig. 1c**). To spur more widespread integration of MassQL into other platforms, MassQL is available as Python and R libraries and as a web API (for integration into software tools using a programming language without official libraries, including Java, Scala, C#). Further, a standalone command line tool and portable scalable computational workflow[9] are available to the community to run on their own compute clusters or in the cloud at GNPS[7]. To help users learn how to write and perform MassQL queries, we have created documentation (Link), instructional videos (Link), and an interactive MassQL sandbox (https://msql.ucsd.edu/). The MassQL sandbox enables users to interactively write and apply queries on demonstration data, including public MS/MS spectral libraries. As the research community is global, each MassQL query within the sandbox includes an automated translation into English, Portuguese, Spanish, German, French, Chinese, Japanese, Korean, and Russian, which can be included in manuscripts and grants to ensure reproducibility. From the MassQL sandbox, users can implement their desired search parameters and with a single click, apply the query to their own data in GNPS.

Community members have written and applied MassQL and contributed to a wiki-like community compendium using 35 applications of MassQL (Link, **Fig 1d**). The MassQL query compendium will function as an app store to provide a centralized location for MassQL query deposition. The compendium provides an opportunity for users to reuse successful MassQL queries to search for the same or similar classes of compounds in other MS datasets. Uniquely, MassQL is scalable and able to query individual files and across hundreds of thousands of data files from thousands of public projects in multiple repositories, including MassIVE/GNPS[7], Metabolomics Workbench[10], and MetaboLights[11]. This capability has thus far aided in the discovery of uncharacterized analogs of different chemical classes in a wide variety of fields, including environmental chemistry, (**SI Note 1.1.1 - Unexpected Organophosphate Compounds in the Environment**), bioinorganic chemistry (**SI Note 3.3.1 - Discovering Iron Binding Molecules from Fungi**), and natural product discovery (**SI Note 3.5.1 - Pentabrominated Natural Products**). Due to the unique scalability of MassQL, it is possible to focus queries towards a specific structural class across an entire repository of data. To enable users to efficiently explore the potentially large chemical diversity present in all public data, MassQL's output is interoperable with existing software tools such as molecular networking. For the first time, it is possible to use a straightforward language as a filter in conjunction with molecular networking to focus on a structural class across an entire data set or repository and visualize that class' full chemical diversity (**SI Notes 1.1.2, 1.5, 1.7, 1.9, 1.12, 3.8)**.

MassQL derives strength from the users in the community as an open-source, flexible, shareable, instrument agnostic, and scalable data analysis tool. We envision the flexibility of

MassQL and further development of the language and ecosystem over time to meet the scientific community's growing needs in mining MS data, a nearly-indispensable chemical analysis solution for science.

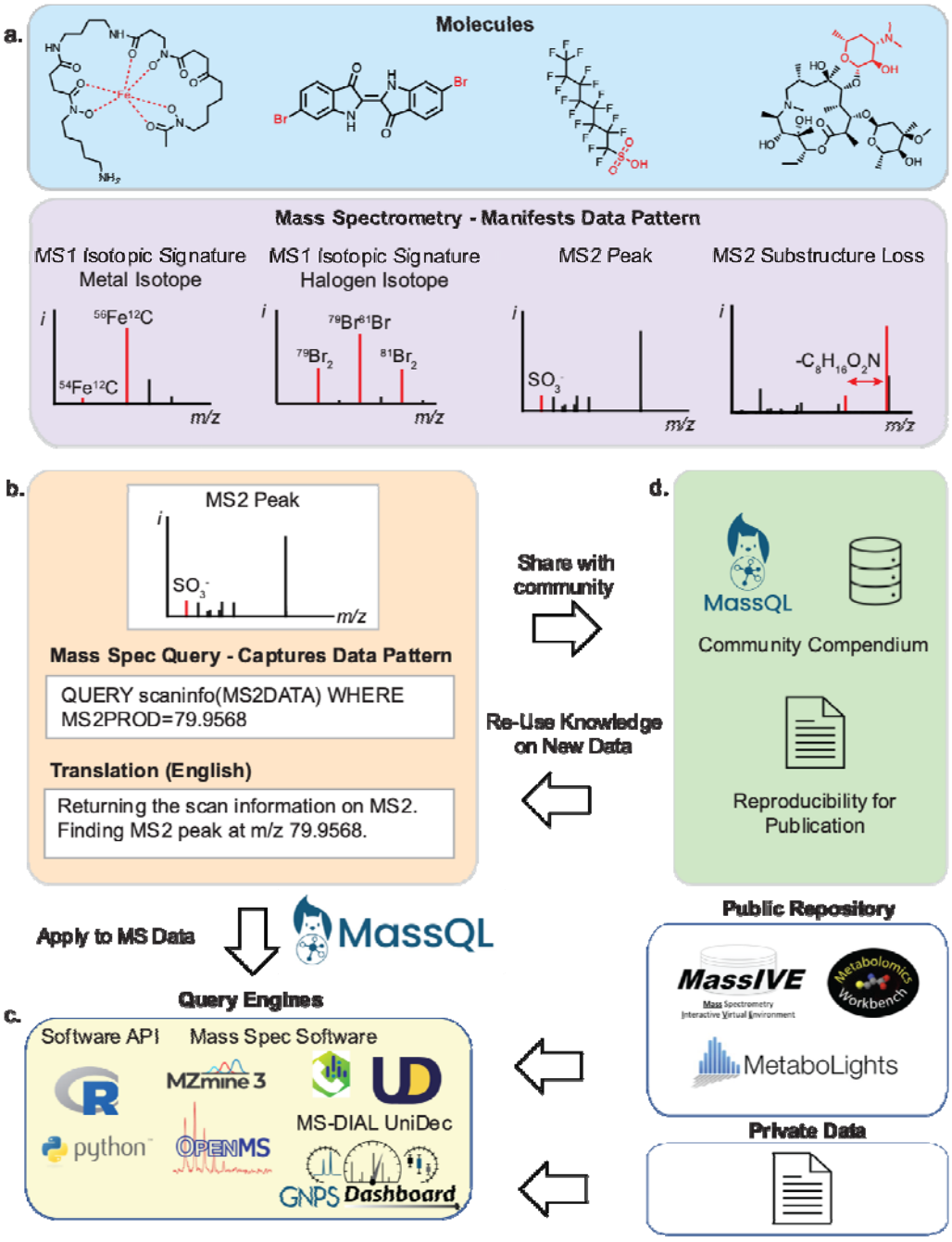

Figure 1. a) Examples of molecules that produce distinctive data patterns when measured by mass spectrometry. b) MassQL query representing MS/MS fragmentation patterns that encapsulates a characteristic mass loss. The query can be translated to 9 languages for enhanced accessibility. c) MassQL is a universal tool to query MS data. MassQL enables data searching in a single file to entire mass spectrometry repositories. MassQL has also been

198 incorporated into a wide range of mass spectrometry software. d) MassQL queries are shared
199 and reused via the Community Compendium, which increases reproducibility and knowledge
200 dissemination.

Code Availability

Reference Engine Implementation (Python), language formal grammar, GNPS Workflow, NextFlow Workflow, and interactive web interface can be found here:
https://github.com/mwang87/MassQueryLanguage

Also available in Pypi:
https://pypi.org/project/massql/

R API can be found here:
https://github.com/rformassspectrometry/SpectraQL

MZmine:
https://github.com/mzmine/mzmine3

OpenMS:
https://pyopenms.readthedocs.io/en/latest/massql.html

MS-DIAL 5:
http://prime.psc.riken.jp/compms/index.html

UniDec:
https://github.com/michaelmarty/UniDec

Language Documentation:
https://mwang87.github.io/MassQueryLanguage_Documentation/

Acknowledgements:

We thank Alan Leung for initial discussions and guidance on language design. This research was supported in part by the BBSRC-NSF award 2152526, the Intramural Research Program of National Institute of Environmental Health Sciences of the NIH (ES103363-01, Jarmusch), (ES030158, Fiehn), the National Institute of General Medical Sciences of the NIH (R01 GM125943, Sanchez; R01 GM107550, Gerwick/Dorrestein; R35 GM128690, Phelan), the National Institute of Allergy and Infectious Diseases of the NIH (R21AI156669, McCall), (R15AI137996, Stevens), National Science Foundation (2128044, Sanchez), (CHE-1845230, Marty), NSF CAREER Award (2047235 Garg), the Burroughs Wellcome Fund (1021280, McCall), Fundação de Amparo à Pesquisa do Estado de São Paulo (2018/24865-4), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (E-26/201.260/2021, Borges; E-26/211.314/2019, Chagas), Biological Sciences Scholars Program at the University of Michigan (Kersten), the National Research Foundation of Korea (NRF-2020R1C1C1004046, Kang), the German Research Foundation (EXC 2124, Petras), the Swedish Research Council (VR 2020-04707, Globisch), Fund for Financing Science and Innovation Support under the Ministry of Innovative Development of the Republic of Uzbekistan (Kakhkhorov), the German Research Foundation (DFG) TRR 261 (project 398967434, Walter), FOR2372 (project 290827466,

Crüsemann) the German Ministry for Education and Research (de.NBI, BMBF FKZ031 A 534A) (and EPIC-XS, project number 823839, funded by the Horizon 2020 programme of the European Union (Kohlbacher, Sachsenberg), the Czech Science Foundation (21-11563M, Pluskal), U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337; a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231 (Northen and Bowen) and Subcontract NO. 7601660, Wang. JSPS KAKENHI (21K18216, H.T.), the National Cancer Center Research and Development Fund (2020-A-9, H.T.), JST ERATO Grant (JPMJER2101, H.T.), AMED Japan Program for Infectious Diseases Research and Infrastructure (21wm0325036h0001, H.T.), JST National Bioscience Database Center (NBDC, H.T.), the Novo Nordisk Foundation, Denmark (NNF20CC0035580, NNF16OC0021746, Weber), University of Michigan Biological Science Initiative (UM-BSI, A.T.), Betty and Gordon Moore Foundation (ATA)

Author Contributions

MW conceived the project. MW and AKJ designed the language. MW, RS, JR, AV, HT, MTM, and TTS developed the software. MW, PCD supervised the development of the project. DP, ATA, AKJ, AB, MTM, WB, NG, VVP and PCD provided feedback. AB, JJJvdH, KBK, LIM, RS, MTM, SAK, TP, WB, RMB, FOC, QX tested the software. AB, JJJvdH, KBK, KL, LIM, WC, TD, TP, MSPC, DG, ACM, NG, MC, MJB, RMB, FOC, OGM, AT, EEK, TW, MMAA, PDB, SH, QX, ATA contributed a use case. MW, ATA, DP wrote the documentation. MW, ATA, AKJ, VVP, PCD wrote the manuscript. All authors edited and approved of the manuscript.

Competing interest statement

PCD is an advisor to Cybele and a Co-founder and scientific advisor to Ometa and Enveda with prior approval by UC San Diego. MW is a co-founder of Ometa Labs LLC. TRN. is an advisor of Brightseed Bio. JJJvdH is a member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy.

Bibliography (Journal limit 10, currently 10)

1. Herzog, R. *et al.* A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. *Genome Biol.* **12**, R8 (2011).
2. Matsuda, F. Regular expressions of MS/MS spectra for partial annotation of metabolite features. *Metabolomics* **12**, 113 (2016).
3. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
4. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry

data analysis. *Nat. Methods* **13**, 741–748 (2016).

5. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
6. Marty, M. T. *et al.* Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
7. 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).
8. Petras, D. *et al.* GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nat. Methods* 1–3 (2021) doi:10.1038/s41592-021-01339-5.
9. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
10. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
11. Haug, K. *et al.* MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440–D444 (2020).