

1 **Title:** Algorithmic Learning for Auto-deconvolution of GC-MS Data to Enable Molecular
2 Networking within GNPS.

3
4 **Authors:** Alexander A. Aksenov^{1,2,#}, Ivan Laponogov^{3,#}, Zheng Zhang¹, Sophie LF Doran³,
5 Ilaria Belluomo³, Dennis Veselkov⁴, Wout Bittremieux^{1,2,35}, Louis Felix Nothias^{1,2}, Mélissa
6 Nothias-Esposito^{1,2}, Katherine N. Maloney^{1,27}, Biswapriya B. Misra⁵, Alexey V. Melnik¹,
7 Kenneth L. Jones II¹, Kathleen Dorrestein^{1,2}, Morgan Panitchpakdi¹, Madeleine Ernst^{1,33},
8 Justin J.J. van der Hoof^{1,38}, Mabel Gonzalez⁶, Chiara Carazzone⁶, Adolfo Amézquita⁷, Chris
9 Callewaert^{8,9}, James Morton⁹, Robert Quinn¹⁰, Amina Bouslimani^{1,2}, Andrea Albarracín
10 Orio¹¹, Daniel Petras^{1,2}, Andrea M. Smania^{31,32}, Sneha P. Couvillion¹², Meagan C. Burnet¹²,
11 Carrie D. Nicora¹², Erika Zink¹², Thomas O. Metz¹², Viatcheslav Artaev¹³, Elizabeth Humston-
12 Fulmer¹³, Rachel Gregor³⁷, Michael M. Meijler³⁷, Itzhak Mizrahi³⁶, Stav Eyal³⁶, Brooke
13 Anderson¹⁵, Rachel Dutton¹⁵, Raphaël Luga¹⁶, Pauline Le Boulch¹⁶, Yann Guitton¹⁷,
14 Stephanie Prevost¹⁷, Audrey Poirier¹⁷, Gaud Dervilly¹⁷, Bruno Le Bizec¹⁷, Aaron Fait¹⁴, Noga
15 Sikron Persi¹⁴, Chao Song¹⁴, Kelem Gashu¹⁴, Roxana Coras¹⁸, Monica Guma¹⁸, Julia
16 Manasson²¹, Jose U. Scher²¹, Dinesh Barupal¹⁹, Saleh Alseekh^{20,29}, Alisdair Fernie^{20,29}, Reza
17 Mirnezami²⁸, Vasilis Vasiliou²², Robin Schmid²³, Roman S. Borisov²⁴, Larisa N. Kulikova²⁵,
18 Rob Knight^{9,26,34,35}, Mingxun Wang^{1,2}, George B Hanna³, Pieter C. Dorrestein^{1,2,9,26,*} & Kirill
19 Veselkov^{3**}.

20
21 **Addresses:**

- 22 1. Skaggs of Pharmacy and Pharmaceutical Sciences, University of California San
23 Diego, La Jolla, San Diego, CA
- 24 2. Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and
25 Pharmaceutical Sciences, University of San Diego, California, 9500 Gilman Dr. La
26 Jolla CA 92093
- 27 3. Department of Surgery and Cancer, Imperial College London, South Kensington
28 Campus, London SW7 2AZ
- 29 4. Intelligify Limited, 160 Kemp House, City Road, London, United Kingdom, EC1V 2NX
- 30 5. Center for Precision Medicine, Department of Internal Medicine, Section of Molecular
31 Medicine, Wake Forest School of Medicine, Medical Center Boulevard, Winston-
32 Salem NC 27157
- 33 6. Department of Chemistry, Universidad de los Andes, Bogotá, Colombia
- 34 7. Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia
- 35 8. Center for Microbial Ecology and Technology, Coupure Links 653, 9000 Ghent,
36 Belgium
- 37 9. Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA
- 38 10. Department of Biochemistry and Molecular Biology, Michigan State University, 603
39 Wilson Rd, East Lansing, MI 48824
- 40 11. IRNASUS, Universidad Católica de Córdoba, CONICET, Facultad de Ciencias
41 Agropecuarias. Córdoba, Argentina
- 42 12. Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA
43 99352
- 44 13. LECO Corporation, 3000 Lakeview Avenue, St. Joseph, MI 49085

- 45 14. The French Associates Institute for Agriculture and Biotechnology of Dryland, The
46 Jacob Blaustein Institutes for Desert Research, Ben Gurion University of the Negev,
47 84990 Sede Boqer Campus, Israel
- 48 15. Division of Biological Sciences, University of San Diego, California, 9500 Gilman Dr.
49 La Jolla CA 92093
- 50 16. UMR Qualisud, Université d'Avignon et des Pays du Vaucluse, Agrosiences, 84000
51 Avignon, France
- 52 17. Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA),
53 Oniris, INRA, 44307 Nantes, France
- 54 18. Division of Rheumatology, Department of Medicine, University of California San
55 Diego, La Jolla, CA
- 56 19. NIH West-Coast Metabolomics Center for Compound Identification, University of
57 California Davis, Davis CA USA
- 58 20. Max-Planck Institute for Molecular Plant Physiology, 14476, Potsdam-Golm, Germany
- 59 21. Division of Rheumatology, Department of Medicine, New York University School of
60 Medicine, New York, NY
- 61 22. Department of Environmental Health Sciences, Yale School of Public Health, Yale
62 University, New Haven, CT, USA
- 63 23. Institute of Inorganic and Analytical Chemistry, Corrensstr. 28/30, 48149 Münster,
64 Germany
- 65 24. A.V.Topchiev Institute of Petrochemical Synthesis RAS, 29 Leninsky pr., Moscow,
66 119991 Russian Federation
- 67 25. Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St,
68 Moscow, 117198 Russian Federation
- 69 26. UCSD center for Microbiome Innovation, University of San Diego, California, 9500
70 Gilman Dr. La Jolla CA 92093
- 71 27. Department of Chemistry, Point Loma Nazarene University, 3900 Lomaland Drive,
72 San Diego, CA, 92106 USA
- 73 28. Department of Colorectal Surgery, Royal Free Hospital NHS Foundation Trust, Pond
74 Street, Hampstead, London NW3 2QG
- 75 29. Center of Plant Systems Biology and Biotechnology (CPSBB), Plovdiv, Bulgaria
- 76 30. University of Antwerp, Antwerp, Belgium
- 77 31. Universidad Nacional de Córdoba, Facultad de Ciencias Químicas, Departamento de
78 Química Biológica Ranwel Caputto, Córdoba, Argentina.
- 79 32. CONICET, Universidad Nacional de Córdoba, Centro de Investigaciones en Química
80 Biológica de Córdoba (CIQUIBIC), Córdoba, Argentina.
- 81 33. Center for Newborn Screening, Department of Congenital Disorders, Statens Serum
82 Institut, Copenhagen, Denmark
- 83 34. Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA
- 84 35. Department of Computer Science, University of California, San Diego, La Jolla, CA,
85 USA
- 86 36. Department of Life Sciences and the National Institute for Biotechnology in the
87 Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel

88 37. Department of Chemistry and the National Institute for Biotechnology in the Negev,
89 Ben-Gurion University of the Negev, Beer-Sheva, Israel

90 38. Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB,
91 Wageningen, the Netherlands

92

93 # Co-first author

94 * To whom correspondence should be addressed regarding mass spectrometry and
95 GNPS – pdorrestein@health.ucsd.edu

96 ** To whom correspondence should be addressed regarding MSHub.

97 kirill.veselkov04@imperial.ac.uk

98

99 **Author contributions:**

100 PCD, AAA, MW, LFN came up with the concept of GNPS for GC-MS data.

101 KV designed and supervised MSHub platform development

102 IL, DV, VV, KV developed the MSHub platform

103 MW, ZZ, AAA developed the workflows

104 AAA, ZZ, MW, BBM, RSB performed infrastructure testing and benchmarking

105 AAA, ZZ assessed EI-based molecular networking

106 WB generated plots for MSHub algorithm performance testing

107 ZZ, AA, ME generated molecular network plots

108 ME, JJJvdH adapted the MolNetEnhancer workflow for GC-MS Molecular Networks

109 AAA, AVM, MP, KJ, KD conducted 3D skin volatilome mapping studies

110 SD, IB, GH conducted oesophageal and gastric breath analysis cancers detection study

111 AAA, ZZ, MP, MW converted and added public libraries to GNPS

112 AAA, AVM, SD, BBM, MG, CC, AA, JM, RQ, AB, AAO, DP, AMS, SPC, TOM, MCB, CDN,

113 EZ, VA, EHF, RG, MMM, IM, SE, PLB, BA, RL, YG, SP, AP, GD, BLB, AF, NS, KG, CS, RC,

114 MG, JM, JUS, DB, SA, AF generated GC-MS data

115 RSB, LNK, AAA assembled the initial version of the public reference spectra library

116 RS created MZmine export module for GNPS GC-MS input files and RI markers file export

117 AAA, RS, IB, AAO, AMS, BA, MG, KNM, RSB produced training videos

118 AAA, MNE, MG, LFN wrote and compiled tutorials and documentation

119 PCD, AAA, WB, KV, RM, RK wrote the paper

120

121 **Ethics/COI declaration**

122 Pieter C. Dorrestein is a scientific advisor for Sirenas LLC.

123 Mingxun Wang is a consultant for Sirenas LLC and the founder of Ometa labs LLC.

124 Alexander A. Aksenov is a consultant for Ometa labs LLC.

125

126 **Online Tutorial:** <https://ccms-ucsd.github.io/GNPSDocumentation/gcanalysis/>

127

128 **Abstract:** Gas chromatography-mass spectrometry (GC-MS) represents an analytical
129 technique with significant practical societal impact. Spectral deconvolution is an essential
130 step for interpreting GC-MS data. No public GC-MS repositories that also enable repository-
131 scale analysis exist, in part because deconvolution requires significant user input. We

132 therefore engineered a scalable machine learning workflow for the Global Natural Product
133 Social Molecular Networking (GNPS) analysis platform to enable the mass spectrometry
134 community to store, process, share, annotate, compare, and perform molecular networking of
135 GC-MS data. The workflow performs auto-deconvolution of compound fragmentation patterns
136 *via* unsupervised non-negative matrix factorization, using a Fast Fourier Transform-based
137 strategy to overcome scalability limitations. We introduce a “balance score” that quantifies the
138 reproducibility of fragmentation patterns across all samples. We demonstrate the utility of the
139 platform with breathomics analysis applied to the early detection of oesophago-gastric
140 cancer, and by creating the first molecular spatial map of the human volatilome.

141
142

143 **Introduction:**

144 Electron ionization gas chromatography-mass spectrometry (GC-MS) is widely used
145 in numerous analytical applications with profound societal impact, including screening for
146 inborn errors of metabolism, toxicological profiling in humans and animals, basic science
147 investigations into biochemical pathways and metabolic flux, understanding of
148 chemoattraction, doping investigations, forensics, food science, chemical ecology, ocean and
149 air quality monitoring, and many routine laboratory tests including cholesterol¹, vitamin D²
150 and lipid levels³. GC-MS is widely adopted because of its key advantages, including low
151 operational cost, excellent chromatographic resolution, reproducibility and ease of use.

152 In GC-MS, the predominant ionization technique is electron ionization (EI), in which all
153 compounds that elute from the chromatography column are ionized by high energy (70eV)
154 electrons in a highly reproducible fashion to yield a combination of fragment ions. Because
155 fragmentation occurs simultaneously with ionization, an essential computational step in the
156 analysis of all GC-MS data is the “spectral deconvolution” - the process of separating
157 fragmentation ion patterns for each eluting molecule into a composite mass spectrum⁴. The
158 deconvolution is particularly computationally challenging for complex biological systems
159 where co-elution of compounds is inevitable as raw GC-MS data consist of mass spectra
160 originating from hundreds-to-thousands of molecules.

161 Annotation of GC-MS data is achieved by matching the deconvoluted fragmentation
162 spectra against reference spectral libraries of known molecules. The 70eV energy for ionizing
163 electrons in GC-MS was set as the standard early, making it possible to use decades-old EI
164 reference spectra for annotation^{5,6} and compare EI data across instruments. There are now
165 ~1.2 million reference spectra, accumulated and curated over a period of >50 years, that are
166 commercially or publicly available for the annotation of GC-MS data^{6,7}. To date, many
167 analytical tools and several repositories for GC-MS data have been introduced^{5,8-16}. Despite
168 these developments, much GC-MS data processing is restricted to vendor-specific formats
169 and software (e.g. VocBinBase¹⁵ uses Leco ChromaTOF data). Moreover, the deconvolution
170 requires multiple parameters to be set by the user or manual peak integration. Further, none
171 of the tools are integrated into a mass spectrometry/metabolomics public data repository that
172 retains every setting and result of an analysis job, features that are vital for reproducibility of
173 data processing. A public informatics resource that can not only be integrated with a public
174 repository, but also perform GC-MS deconvolution, alignment, and mass spectral library
175 matching for large numbers (>100) of data files is needed. Technical reasons, such as the
176 lack of a shared and uniform data format, often preclude GC-MS data comparison between

177 different laboratories and prevents taking advantage of repository-scale information and
178 community knowledge about the data. This, coupled to a lack of incentive to deposit data into
179 public domain, leads to GC-MS datasets being infrequently shared and rarely reused across
180 studies and/or biological systems^{15,17–21}.

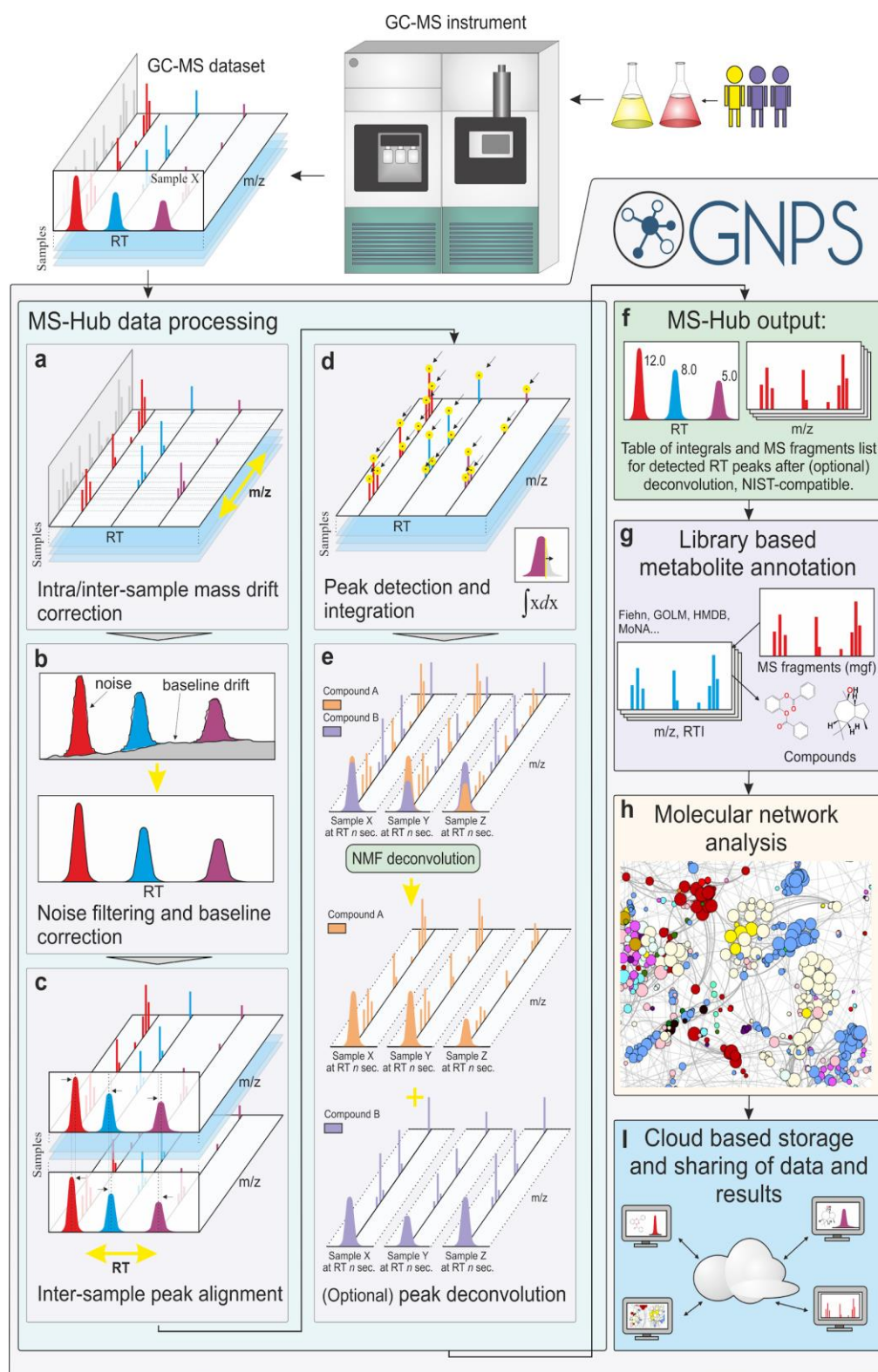
181 One of the developments that enabled finding additional structural relationships within
182 mass spectrometry data is spectral alignment, which forms the basis for molecular
183 networking^{22–26}. Here, we develop a repository-scale analysis infrastructure for GC-MS data
184 enable to create networks within the Global Natural Products Social (GNPS) Networking
185 platform. GNPS promotes Findable, Accessible, Interoperable, and Reusable (FAIR) use
186 practices for mass spectrometry data²⁷. The community infrastructure can be accessed at
187 <https://gnps.ucsd.edu> under the header “GC-MS EI Data Analysis”.

188

189 **Results: Creating a web-based scalable strategy for spectral deconvolution.** Current EI
190 spectral deconvolution strategies can save settings and apply them to the next analysis, but
191 require initial manual parameter setting (e.g. AMDIS⁵, MZmine/ADAP⁸, MS-DIAL⁹,
192 PARAFAC2¹²); some require extensive computational skills to run (e.g. XCMS²⁸, eRah¹⁴).
193 Although batch modes exist, they do not enhance deconvolution quality by utilizing
194 information from other files of the dataset. To use this across-file information, improve
195 scalability of spectral deconvolution, and eliminate manual parameter setting, we developed
196 an algorithmic learning strategy for deconvolution of entire datasets (**Figure 1a-f**). We
197 deployed this functionality within GNPS/MassIVE²⁹ (**Figure 1f-i**). To promote analysis
198 reproducibility, all GNPS jobs performed are retained in the “My User” space and can be
199 shared as hyperlinks in collaborations or publications.

200 Classically, when performing spectral deconvolution of GC-MS data, the user defines
201 parameters specific to their data to the best of their abilities. The user must therefore have a
202 thorough understanding of the characteristics (i.e., peak shape, peak width, resolution etc.) of
203 the particular GC-MS data set before spectral deconvolution. In our approach, the
204 parameters for spectral deconvolution (m/z drift of the ions, peak shape - slopes of raising
205 and trailing edges, peak shifts, and noise/intensity threshold) are auto-estimated. This user-
206 independent ‘automatic’ parameter optimization is accomplished via fast Fourier
207 transformation, multiplication, and inverse Fourier transformation for each ion across entire
208 data sets, followed by an unsupervised matrix factorization (one layer neural network):

209 **Figure 1a-e**. Then, the compositional consistency of spectral patterns, for each spectral
210 feature deconvoluted across the entire data set, can be summarized as a parameter that we
211 termed “balance score”. The balance score (definition is described in the Methods) gives
212 insight into how well the spectral feature is explained across the entire data set: when high,
213 the spectrum is consistent across different samples. Even when a compound is present in
214 only a few samples in the dataset, as long as the spectral patterns are highly conserved
215 across samples (e.g. not contaminated by spurious noise peaks), it would result in a high
216 balance score. Balance score thus allows discarding low-quality spectra that are more likely
217 to be noise, and provides an orthogonal metric to matching scores when searching spectral
218 libraries. We refer to the dataset-based spectral deconvolution tool within the GNPS
219 environment as “MSHub”. MSHub converts raw GC-MS data of any kind (e.g. **Table S1**) into
220 spectral patterns, enabling molecular networking within GNPS.



221
222
223
224
225

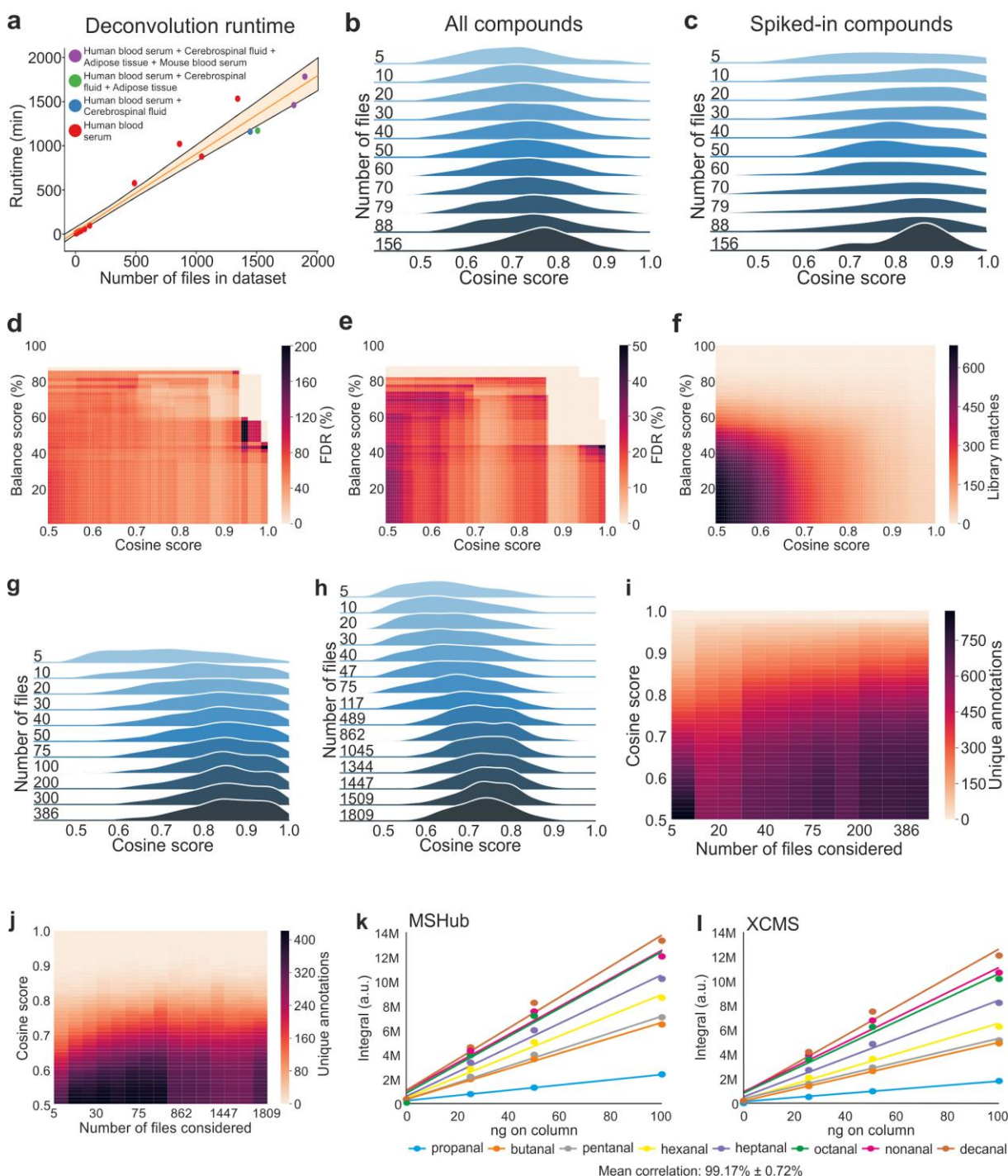
Figure 1. Schematic representation of the MSHub processing pipeline within GNPS. MSHub accepts netCDF, mzML formats of any EI GC-MS data for input. **a**) Spectra are aligned and binned in *m/z* dimension noise is filtered and **b**) baseline is corrected in each spectrum in RT dimension to address

226 issues such as baseline rise with thermal gradient due to bleeds. **c)** Common profile established across
227 entire dataset and peaks in RT dimension are aligned to it using FFT-accelerated correlation in full
228 resolution via iterative approach. **d)** Fast peak detection picks and integrates peaks to generate both
229 peak integrals for all samples and their common fragmentation patterns. For datasets with more than 5
230 samples peak deconvolution step **e)** is employed to separate overlapping peaks with different patterns
231 across samples using NMF approach. **f)** MSHub produces peak integrals for all samples and canonical
232 fragmentation patterns. **g)** GNPS employs either public or user-provided reference libraries to annotate
233 peaks. **h)** Molecular networks are built for further metabolite analysis. **i)** Data and results are shared
234 between users via GNPS's cloud architecture. NMF - Non-negative matrix factorization, FFT - Fast
235 Fourier Transform, RT - retention time, m/z - mass-to-charge ratio.

236

237 All MSHub algorithms operate iteratively for enhanced scalability, using high-
238 performance HDF5 technologies saving settings for each analysis step. The Fourier
239 transform step with multiplication dramatically improves MSHub's efficiency, resulting in
240 deconvolution times that scale linearly rather than exponentially with the number of files
241 (**Figure 2a, S2**). The GNPS GC-MS workflow can process thousands of files in hours (**Figure**
242 **2a**), which is faster than data acquisition, making data processing no longer a bottleneck. We
243 achieved this performance using out-of-core processing, a technique used to process data
244 that are too large to fit in a computer's main memory (RAM): MSHub uploads files one at a
245 time into the specific RAM module, data are then processed and deleted from memory,
246 iteratively. **Figure 2a** illustrates the linear dependency between the number of samples
247 processed and the processing time. Because only one sample is stored in memory at any
248 given time, the workflow memory load is constant. Spectral deconvolution scales linearly
249 because each step in the processing pipeline is linear with respect to time (**Figure S2a-f**),
250 taking ~1 min per file (**Figure S1**). The machine learning approaches gain improved
251 performance with increasing amounts of data, which means that increasing dataset size
252 would boost learning each spectral pattern. Indeed, larger volume of analyzed data leads to
253 better scores of spectral matches for the known compounds in derivatized blood serum
254 samples that were spiked with 37 fatty acid methyl esters (FAMES) and 17 long-chain
255 hydrocarbons (**Figure 2b, c**). Cosine and balance score can be jointly used as filters for
256 processing the final results (**Figure 2d-f**). In the analysis of biological samples, similar trends
257 are found as for the reference dataset: the spectral matching scores against the library
258 increase with increasing number of processed files while their distributions become narrower,
259 a reflection that more data leads to better quality of results (**Figure 2g, h**). When there are
260 more files deconvoluted, MSHub is leveraged to reduce chimeric spectra and discover more
261 real spectral features, which leads to higher quality spectra and a rise in the number of
262 unique annotations with greater match scores (**Figure 2i, j**). If the user only has a few files
263 (fewer than 10), spectral deconvolution and alignment should be performed using alternative
264 methods (e.g. MZmine³⁰, OpenChrom^{28,31}, AMDIS⁵, MZmine/ADAP⁸, MS-DIA⁹, BinBase¹⁵,
265 XCMS³²/XCMS Online²⁸, MetAlign¹⁰, SpecAlign³³, SpectConnect¹¹, PARAFAC¹², MeltDB¹³,
266 eRah¹⁴). Using those tools, molecular networking can be performed in the same fashion as
267 for MSHub, as the library search GNPS workflow accepts input from other tools into the
268 GNPS/MassIVE environment. We have further benchmarked the MSHub against XCMS²⁸
269 (MassIVE dataset MSV000084622) and the quantitative results were nearly identical (the
270 calibration curve was within 99.17% correlation with 0.72% STD, **Figure 2k,l**).

271 Spectral deconvolution using MSHub in GNPS generates an .mgf file that contains
 272 deconvoluted spectra with aligned retention times and a feature table of peak areas of
 273 features across all files. This generated .mgf spectral deconvolution summary file is used for
 274 searching against spectral libraries and for molecular networking. GNPS saves this
 275 information, so the deconvolution step does not need to be re-performed for any future
 276 analyses. The output results can be downloaded and explored using many external tools,
 277 e.g. MetExpert³⁴.
 278



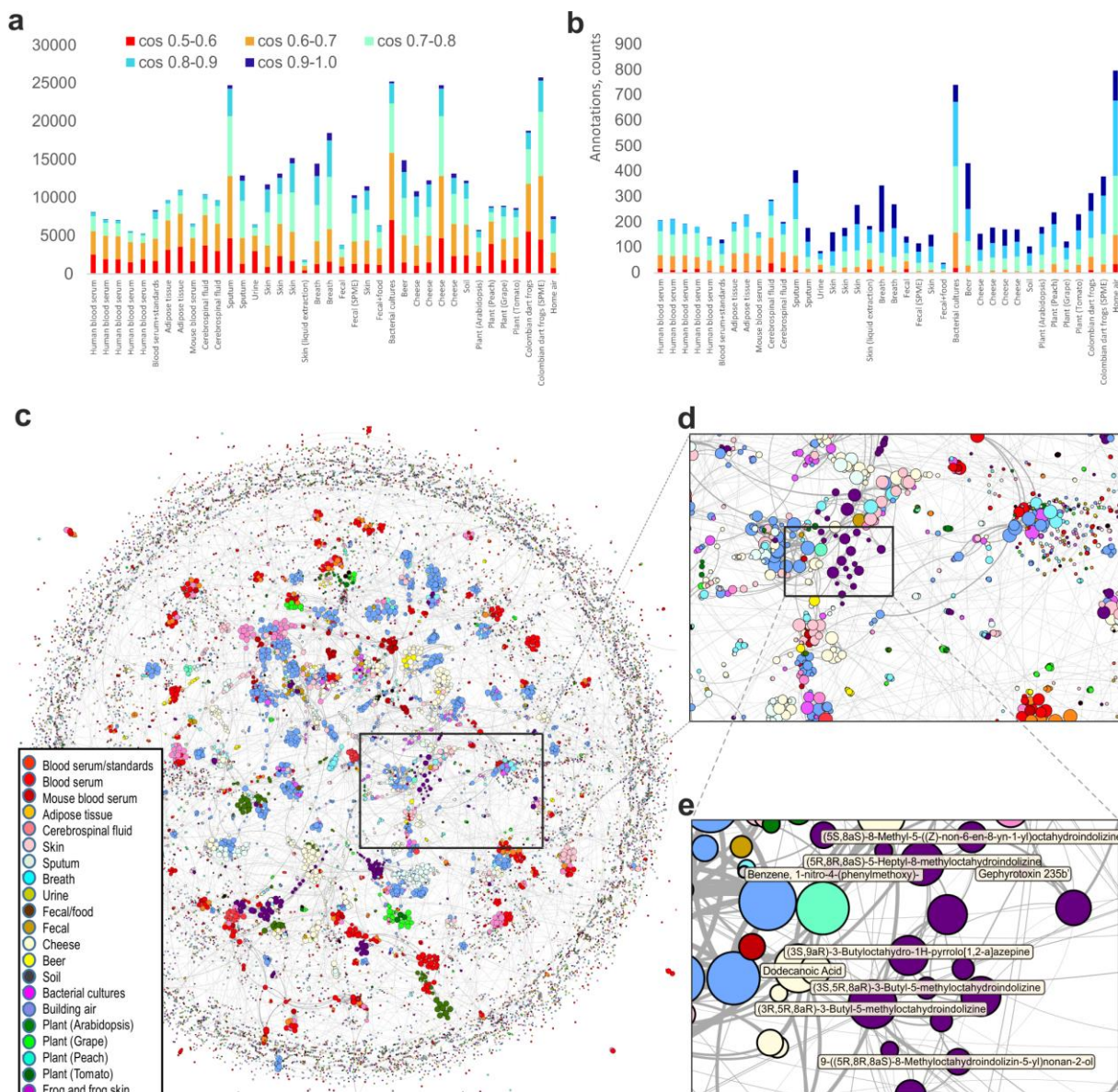
279

280 **Figure 2: Performance evaluation of dataset-wide deconvolution.** a) Linear dependence between
281 the number of samples and the processing time on a single compute node. b) Distributions of library
282 matching scores for the test dataset of reference compounds spiked in a complex blood serum matrix
283 with an increased volume of input data (**datasets Test1-Test11, Table S1**) for all matches and c) for
284 the reference compounds only. d) False discovery rate (FDR) for the sub-class³⁵ annotations (**dataset**
285 **Test11 in Table S1**) of the top match and (e) top ten matches. More restrictive thresholds minimize
286 misannotations. f) Heat map of the number of library matches for spiked compounds. g) Distributions
287 of library matching scores for the top match in a study of oesophageal and gastric cancer detection
288 using breath analysis (non-derivatized, **datasets ICL1-ICL11 in Table S1**) and h) studies of human
289 and mouse blood serum, adipose tissue and cerebrospinal fluid (silalated, **datasets UCD1-UCD16 in**
290 **Table S1**). i) Heat map of the number of unique annotations (top hit only) for the data across **datasets**
291 **ICL1-ICL11 in Table S1** and j) **datasets UCD1-UCD16 in Table S1**; no balance score filtering
292 applied. Spurious features corresponding to the low cosine tail of the distribution on panels (g) and (h)
293 are improved as higher volume of the data enhances the frequency domain for deconvolution quality.
294 k, l) Quantitative integrals of abundances quantitation for the mixture of standards (MassIVE dataset
295 MSV000084622) evaluated using XCMS (l) and MSHub (k).

296
297 **GNPS enables searches against public spectral reference libraries and molecular**
298 **networking at repository scale.** Once the .mgf file is generated by GNPS-MSHub or
299 imported from another deconvolution tool, the spectral features can be searched against
300 public libraries³⁶ (currently GNPS has Fiehn³⁷, HMDB³⁸, MoNA¹⁷, VocBinBase¹⁵) or the user's
301 own private or commercial libraries (such as NIST 2017³⁹ and Wiley). Matches are narrowed
302 down based on user-defined filtering criteria such as number of matched ions, Kovats
303 retention index (RI, calculated if hydrocarbon reference values are provided), balance score,
304 cosine score, and abundance. With this release, we also provide additional freely available
305 reference data compiled by co-authors of this manuscript of 19,808 spectra for 19,708
306 standards. Although the possible candidate annotations can be further narrowed by retention
307 index (RI), they should still be considered level 3, a molecular family, annotation according to
308 the 2007 metabolomics standards initiative (MSI)⁴⁰. Calculation of RIs is enabled and
309 encouraged but not enforced. When multiple annotations can be assigned, GNPS provides
310 all candidate matches within user's filtering criteria.

311 No matter how the spectral library is searched in GC-MS, due to the absence of a
312 parent mass, a list of spectral matches is more likely contain mis-annotations, both related
313 (isomers, isobars) or less frequent, entirely unrelated compounds⁵. However to spot
314 misassignments at the molecular family level, we propose to explore deconvoluted GC-MS
315 data via molecular networking, a strategy that has been effective for LC-MS/MS data. In the
316 case of EI, unlike in LC-MS/MS where the precursor ion mass is known, the molecular ion is
317 often absent. For this reason, the molecular networks are created through spectral similarity
318 of the deconvoluted fragmentation spectrum without considering the molecular ion. For GC-
319 MS data that do have a molecular ion or precursor ion mass, e.g. from chemical ionization
320 (CI) or with MS/MS spectra, the feature-based molecular networking workflows should be
321 used^{29,41}. We explored clustering patterns for the EI data (**Figure S4**) and observed that the
322 EI-based cosine similarity networks are predominantly driven by structural similarity (**Figure**
323 **S4a**)³⁵. These EI networks can be used to visualize chemical distributions and guide
324 annotations (**Figure S5**). Networking enables data co- and re-analysis, as it is agnostic to the
325 data origin once the features are deconvoluted. To demonstrate this, we have built a global

326 network of various public GC-MS datasets deposited on GNPS (38 datasets comprising
327 ~8,500 GC-MS files, **Figure 3c**). These data encompass various types of samples, modes of
328 sample introduction etc. and thus the global network is a snapshot of all chemistries
329 detectable by GC-MS (**Figure 3c-e, S6**). Prior to networking, we applied a balance score of
330 65%, which allowed us to remove a bulk of spurious low quality matches (**Figure 3 a,b**). The
331 balance score filter ensures that the best-explained deconvoluted features are matched
332 against the reference library. The annotation is usually done by ranking potential matches
333 according to a similarity measure (forward match, reverse match, and probability^{42,43}) and
334 when possible, filtering by retention index then reporting the top match. Molecular networking
335 can further guide the annotation at the family level by utilizing information from connected
336 nodes (**Figure S5**) rather than focusing on individual annotations⁴⁴. The global network can
337 be colored by metadata such as sample type (**Figure 3c**), derivatized vs. non-derivatized,
338 instrument type or other metadata (**Figure S6**) to reveal interpretable patterns. When coloring
339 the data by sample type, for example, a cluster of nitrogen-containing heterocyclic
340 compounds was observed to be unique to dart frogs from Dendrobatoidea superfamily
341 (**Figure 3e**), while the long-chain ketones occur in cheese and beer (**Figure S7**). To highlight
342 the broader utility of GNPS GC-MS and GC-MS based molecular networking, 6 supplemental
343 videos were created that carry the user through how to navigate and perform analysis with
344 the tools (**Supporting Videos 1-6**).



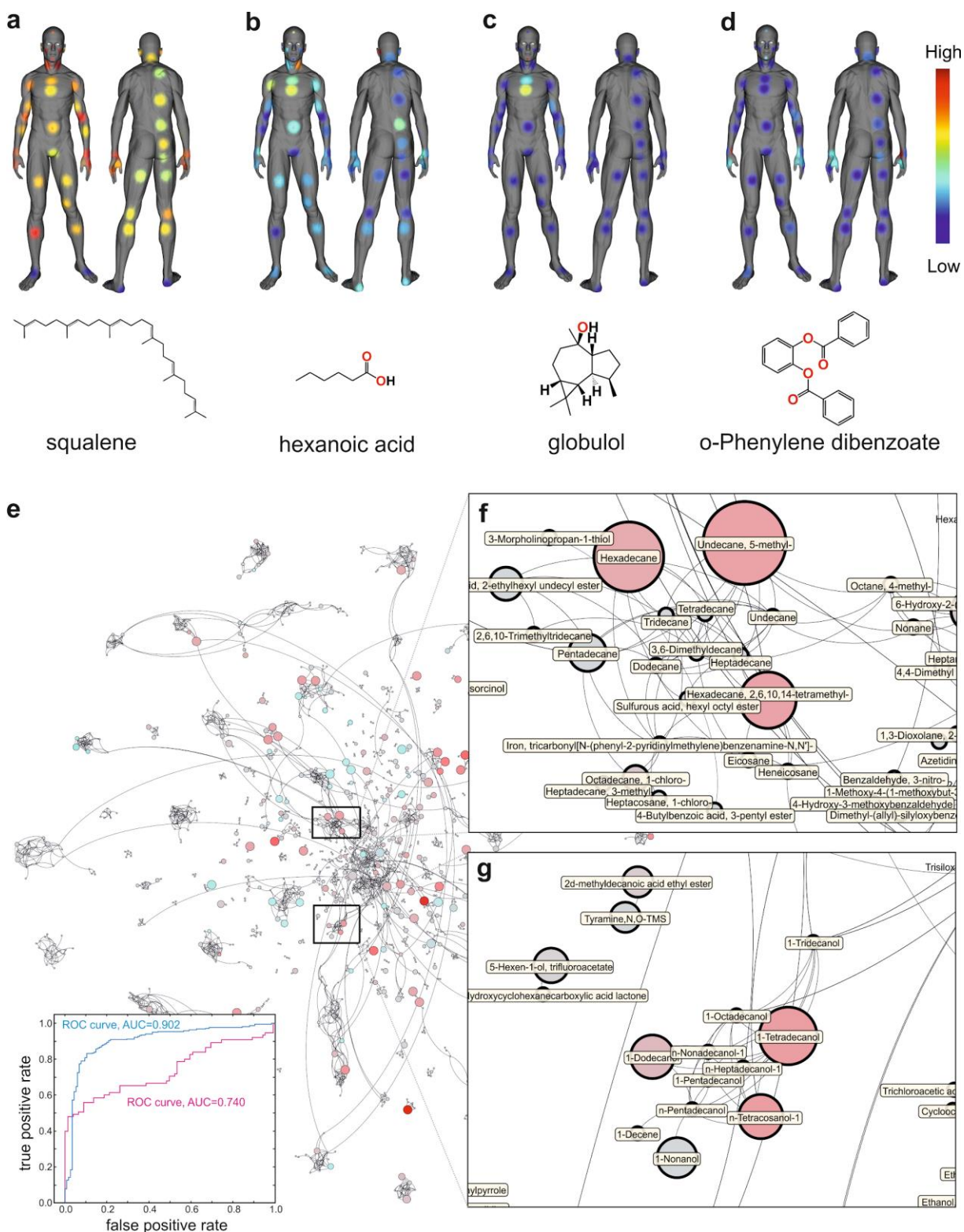
345
 346 **Figure 3: Molecular networking of GC-MS data in GNPS.** Features that are annotated (pass library
 347 match threshold of 0.5) are included **a)** without filtering and **b)** with a 65% balance score filtering. **c)**
 348 Global network containing 35,544 nodes from 8,489 files in 38 GNPS datasets for different types of
 349 samples, including human, derived from various animal, plant, microbial and environmental samples.
 350 Nodes are connected if cosine > 0.5. The size of the node is proportional to the number of nodes that
 351 connect to it ⁴⁵, the edge thickness is proportional to the cosine score (**Figure S3**), the annotation is
 352 the top match with cosine above 0.65. **d)** The inset shows the zoomed-in portion of the network. **e)**
 353 Close up of a cluster of compounds found in the dart frog skin samples with the top spectral library
 354 match shown - all nodes are nitrogen heterocyclic alkaloids such as gephyrotoxin ⁴⁶ that are unique to
 355 these frogs.

356
 357 The output from GNPS deconvolution, annotations, and molecular network analysis
 358 can be exported for use in a statistical analysis environment such as Qiime ^{47,48}, Qiita ⁴⁹, or
 359 MetaboAnalyst ^{50,51}, or for data visualization in tools such as Cytoscape ⁵² or Gephi ⁵³ (e.g.

360 **Supplementary Figures S4-S7, Figure 3, 4e-g**), or for molecular cartography in 'ili⁵⁴
361 (**Figure 4 a-d**). To demonstrate how to use GNPS GC-MS for the latter, we collected
362 samples from 52 body locations from one person using a sampling patch that absorbs
363 volatiles (**Figure 4 a-d**). These samples were subjected to headspace desorption followed by
364 GC-MS, deconvoluted and annotated using the GNPS GC-MS pipeline. The abundances
365 from the deconvoluted spectra are superimposed onto a 3D model of a human (**Figure 4 a-**
366 **d**). Using balance filters at 50% and >0.9 cosine, we arrived at annotations that, once
367 visualized, revealed the distributions of skin volatiles. For example, squalene was found on
368 all locations, but less on the feet. Hexanoic acid was most abundant on the chest and
369 armpits. Globulol, an ingredient of the personal care product this individual used on the chest,
370 was most intense on the chest, while phenylenedibenzoate, a skincare ingredient, was found
371 on the face and hands.

372 We also conducted two studies (Study 1: n=631 samples and Study 2: n=219
373 samples, respectively) on breath analysis associated with oesophageal and gastric cancers
374 (OGC, **Figure 4 e-g**). In breath, biological signatures are usually obscured by intra- and inter-
375 subject variability, experimental conditions, e.g. ambient air quality, different diets etc.
376 Biologically relevant compounds are often present at low abundances. Both studies predicted
377 OGC (inset in **Figure 4e**). The next important step is to consider features that are the most
378 discriminant between categories of interest (OGC vs. control) to investigate whether their
379 chemical identity can be linked to a plausible biological rationale. However, even though
380 OGC prediction was achieved in each study, the "OGC signatures" do not appear to overlap
381 between the two studies, which is very typical for breath analysis field in general⁵⁵⁻⁵⁸. As
382 molecular networking organizes chemically similar compounds into clusters, it facilitates
383 recognition of patterns at a chemical family level. Exploring the two studies as a single
384 network revealed an increase of related but not identical medium/long chain alcohols,
385 aldehydes and hydrocarbons (**Figure 4 e-g**). Only a handful of these compounds appeared
386 as top discriminating features in either study. Aldehydes are known to be found
387 endogenously, mostly due to lipid peroxidation, and have been proposed as potential
388 biomarkers in exhaled breath in several different types of cancer including lung⁵⁹⁻⁶², breast⁶³,
389 ovarian⁶⁴, colorectal⁶⁵, and, most notably, OGC⁶⁶⁻⁶⁸. The alkanes and methyl branched
390 alkanes have not been previously associated with oesophageal or gastric cancer, but have
391 been associated with lung and breast cancer in exhaled breath^{60,61,63,69,70}. Lipid peroxidation
392 of polyunsaturated fatty acids in cell membranes generates alkanes that can then be
393 excreted in the breath⁷¹, which makes their observation in relation to OGC biologically
394 plausible. Although few individual alkanes were found significantly increased in OGC cohort
395 in both studies, none of them overlapped, and without considering these data as a single
396 network, association of long-chain alkanes with OGC would be far more difficult to recognize.

397
398
399
400



401
402
403
404
405

Figure 4. Examples of results with GNPS processed GC-MS data. 3D visualization of human surface volatome visualized with ^{54}Li as described in the tutorial (<https://ccms-ucsd.github.io/GNPSDocumentation/gcanalysis/>). Molecular distributions on skin of a volunteer shown for: **a**) squalene, a key component of natural skin grease. Low abundance of squalene is aligned with

406 areas of dry skin ⁷² **b**) hexanoic acid, one of the malodour molecules responsible for the unpleasant
407 sour body odor ⁷³ **c**) globulol, naturally occurring in plant essential oils, likely introduced via use of skin
408 cosmetics **d**) phenylenedibenzoate, also introduced via use of a skin product. **e**) Chemical distributions
409 that relate to cancer status are visualized via molecular network that combines two studies. Each node
410 represents a unique mass spectrometry feature obtained from deconvolution. The top annotation is
411 given for matches with $\cos > 0.65$. The size of the node represents the importance of the feature for
412 discrimination by the maximum margin criterion ⁷⁴ with leave patient out cross validation of the OGC
413 group vs control volunteers; the color of the node represents average fold-change in abundance
414 between OGC vs. control groups (red - higher in OGC, teal - higher in control, gray - neither), the size
415 represents $-\log(p \text{ value})$, larger circle corresponds to greater values. The inset shows ROC for both
416 studies (Study 1 - blue, Study 2- red). **f**) Example of cluster of hydrocarbons and **g**) long-chain
417 alcohols. Both human studies are approved by the institutional review boards as described in the
418 Methods.

419

420 **Discussion:**

421 GNPS provides a platform for data sharing and accumulation of public knowledge.
422 Community adoption of GNPS has sharply increased the volume of MS data in the public
423 domain⁷. It has also spurred new tools development (MASST⁷⁵, FBMN⁴¹, ReDU⁷⁶) and
424 enabled many biological discoveries. Due to the fundamental differences between CID and
425 EI fragmentation, the GNPS infrastructure could not previously support the analysis of EI
426 data. Adopting existing solutions for deconvolution was not possible as all of them required
427 too much manual input from the user and could not operate at repository scale. Here we
428 used an unsupervised non-negative matrix factorization and a Fast Fourier Transform-based
429 approach to scale the deconvolution step. Such strategies are most effective when large-
430 scale datasets become available, as features can be extracted with increasing quality of
431 fragmentation patterns, as defined by the balance score. Currently, all 1D EI GC-MS data are
432 amenable and we will extend the same approach to 2D GC-MS data.

433 These features can then be subjected to EI-based molecular networking. The
434 algorithm for molecular networking within GNPS had to be modified to accommodate EI data
435 to function without molecular ion information and can reinforce candidate annotations to level
436 3 by assessing if the annotations are similar at the family level and if annotations share
437 chemical class terms. Such analysis can now be achieved with the data at repository scale,
438 enabling co- and re-analysis of GC-MS data. Here we show how the co-analysis could be
439 beneficial for two cancer breathomics data sets, but in the same fashion other breathomics
440 (or other volatilome) data can now be co-analyzed with these datasets as long as they are
441 publicly available in an open file format. Co-analyzing multiple disparate GC-MS studies
442 would be challenging otherwise. Further, when considering GC-MS data as networks, in
443 addition to conventional statistical approaches, strategies such as networks on graphs⁷⁷
444 could be deployed to investigate global biochemical patterns rather than differences in
445 individual compounds. The networks, in principle, are not limited to any one kind of data and
446 can be extended to any number or type of datasets as shown in **Figure 3c**.

447 Surprisingly, although GC-MS is the oldest and most established of MS-based
448 methods, and the sheer volume of existing EI reference data accumulated over decades (far
449 exceeding that for any other kind of MS), researchers still use decades-old data analysis
450 strategies. We anticipate that the new GNPS community infrastructure will incentivize moving

451 raw EI data into the public domain for data reuse, comparable to the trajectory for tandem
452 MS^{7,75,76}. GC-MS analysis within GNPS/MassIVE will lower the expertise threshold required
453 for analysis, encourage FAIR practices²⁷ through reusable deposition of the data in the public
454 domain, and promote data analysis reproducibility and “recycling” of GC-MS data. Finally,
455 this work is a piece of the puzzle to democratize scientific analysis from all over the world.
456 GC-MS the most widely used MS method, in part due to its competitive operational cost. It is
457 often the only mass spectrometry method available at smaller, e.g. undergraduate
458 institutions, non-metabolomics laboratories, or local testing facilities. The proposed
459 infrastructure will enable labs with fewer resources, including those from developing
460 countries, to have free access to data and reference data in a uniform format, and to free,
461 powerful computing infrastructures.

462

463 **Data and code availability**

464 All of the data used in preparation of this manuscript are publicly available at the MassIVE
465 repository at the UCSD Center for Computational Mass Spectrometry website
466 (<https://massive.ucsd.edu>). The dataset accession numbers are: #1 (MSV000084033), #2
467 (MSV000084033), #3 (MSV000084034), #4 (MSV000084036), #5 (MSV000084032), #6
468 (MSV000084038), #7 (MSV000084042), #8 (MSV000084039), #9 (MSV000084040), #10
469 (MSV000084037), #11 (MSV000084211), #12 (MSV000083598), #13 (MSV000080892), #14
470 (MSV000080892), #15 (MSV000080892), #16 (MSV000084337), #17 (MSV000083658), #18
471 (MSV000083743), #19 (MSV000084226), #20 (MSV000083859), #21 (MSV000083294), #22
472 (MSV000084349), #23 (MSV000081340), #24 (MSV000084348), #25 (MSV000084378), #26
473 (MSV000084338), #27 (MSV000084339), #28 (MSV000081161), #29 (MSV000084350), #30
474 (MSV000084377), #31 (MSV000084145), #32 (MSV000084144), #33 (MSV000084146), #34
475 (MSV000084379), #35 (MSV000084380), #36 (MSV000084276), #37 (MSV000084277), #38
476 (MSV000084212).

477 All of the GNPS analysis jobs for all of the studies are summarized in **Table S1**.

478 The source code of the MSHub software is available online at Github (version used in GNPS)
479 (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/mshub-gc/tools/mshub-gc/proc)
480 and at BitBucket (standalone version in MSHub developers’ repository:
481 https://bitbucket.org/iAnalytica/mshub_process/src/master/). Scripts used to parse, filter,
482 organize data and generate the plots in the manuscript are available online at Github
483 (https://github.com/bittremieux/GNPS_GC_fig). Script for merging individual .mgf files into a
484 single file for creating global network is available at Github:

485 https://github.com/bittremieux/GNPS_GC/blob/master/src/merge_mgf.py)

486 The 3D model, feature table with coordinates used for the mapping and snapshots shown on
487 the Figure 4a-d are available at: [https://github.com/aaksenov1/Human-volatilome-3D-](https://github.com/aaksenov1/Human-volatilome-3D-mapping-)
488 [mapping-](https://github.com/aaksenov1/Human-volatilome-3D-mapping-)

489

490 **Methods:** These are provided as supporting information. The tools are accessible through
491 gnps.ucsd.edu and the documentation on how to use the GNPS GC-MS Deconvolution
492 workflow and molecular networking workflows can be found here [https://ccms-](https://ccms-ucsd.github.io/GNPSDocumentation/gcanalysis/)
493 [ucsd.github.io/GNPSDocumentation/gcanalysis/](https://ccms-ucsd.github.io/GNPSDocumentation/gcanalysis/). Representative examples and short “how
494 to” video can be found here:

495 <https://www.youtube.com/watch?v=yrru-5nrsk&feature=youtu.be>
496 <https://www.youtube.com/watch?v=MblruOSgql&feature=youtu.be>
497 https://www.youtube.com/watch?v=iX03r_mGi2Q&feature=youtu.be
498 <https://www.youtube.com/watch?v=mv-fw2zSgss&feature=youtu.be>
499 <https://www.youtube.com/watch?v=nUhCZ9LwoM4&feature=youtu.be>
500 <https://www.youtube.com/watch?v=PehOiBqzzY&feature=youtu.be>

501

502 **Acknowledgments:** The conversion of the data from different repositories was supported by
503 R03 CA211211 on reuse of metabolomics data, to build enabling chemical analysis tools for
504 the ocean symbiosis program, the development of a user-friendly interface for GC-MS
505 analysis was supported by the Gordon and Betty Moore Foundation through Grant
506 GBMF7622. The UC San Diego Center for Microbiome Innovation supported the campus
507 wide SEED grant awards for data collection that enabled the development of some of this
508 infrastructure. PCD was supported by NSF grant IOS-1656475. KV and IL are very grateful
509 for the support of Vodafone Foundation as part of the project DRUGS/DreamLab. AB was
510 supported by the National Institute of Justice Award 2015-DN-BX-K047. Additional support
511 for data acquisition and data storage was provided by P41 GM103484 Center for
512 Computational Mass Spectrometry, the collection of data from the HomeChem project was
513 supported by the Sloan Foundation. GBH, SD, IL, KV and IB are grateful for the support of
514 the OG cancer breath analysis study by the NIHR London Invitro Diagnostic Co-operative
515 and Imperial Biomedical Research Centre, Rosetrees and Stonegate Trusts and Imperial
516 College Charity. IB acknowledges the contribution of Qing Wen and Dr Michelangelo Colavita
517 for the production of the training video. CC was supported by the Research Foundation
518 Flanders (FWO), with support from the industrial research fund of Ghent University. WB was
519 supported by the Research Foundation Flanders (FWO). AAO acknowledges the support of
520 Fulbright Commission and Consejo Nacional de Investigaciones Científicas y Técnicas
521 (CONICET-Argentina). The work of RL and PLB on the dataset 30 was supported by the
522 Metaboscope, part of the "Platform 3A" funded by the European Regional Development
523 Fund, the French Ministry of Research, Higher Education and Innovation, the region
524 Provence-Alpes-Côte d'Azur, the Departmental Council of Vaucluse and the Urban
525 Community of Avignon. SA and ARF acknowledge the PlantaSYST project by the European
526 Unions Horizon 2020 research and innovation programme (SGA-CSA No 664621 and No
527 739582 under FPA No. 664620). VV acknowledges the support by the National Institute On
528 Alcohol Abuse and Alcoholism award R24AA022057. MG and RC acknowledge the support
529 of the Krupp Endowed Fund grant. A portion of mass spectra in the public reference library
530 was produced within the framework of the State Task for the Topchiev Institute of
531 Petrochemical Synthesis RAS and with the support of the RUDN University Program 5-100.
532 RSB acknowledges support of the State Task for the Topchiev Institute of Petrochemical
533 Synthesis RAS. LNK acknowledges support of the RUDN University Program 5-100. IM
534 acknowledges support of the Israel Science Foundation project number 1947/19 and
535 European Research Council under the European Union's Horizon 2020 research and
536 innovation program (project number 640384). JS has been supported by NIH/NIAMS
537 R03AR072182, The Colton Center for Autoimmunity, Rheumatology Research Foundation,
538 The Riley Family Foundation and The Snyder Family Foundation. JM acknowledges support

539 from 2017 Group for Research and Assessment of Psoriasis and Psoriatic Arthritis
540 (GRAPPA) Pilot Research Grant and NIH/NIAMS T32AR069515. RG is grateful to the Azrieli
541 Foundation for the award of an Azrieli Fellowship. JJJvdH acknowledges support from an
542 ASDI eScience grant, ASDI.2017.030, from the Netherlands eScience Center-NLeSC. BA
543 was supported by the National Science Foundation (NSF) through the Graduate Research
544 Fellowship Program. GC-MS analyses for collection of the dataset MSV000083743 were
545 supported by the Pacific Northwest National Laboratory, Laboratory Directed Research and
546 Development Program, and were contributed by the Microbiomes in Transition Initiative; data
547 were collected in the Environmental Molecular Sciences Laboratory, a national scientific user
548 facility sponsored by the Department of Energy (DOE) Office of Biological and Environmental
549 Research and located at Pacific Northwest National Laboratory (PNNL). PNNL is operated by
550 Battelle Memorial Institute for the DOE under contract DEAC05-76RLO1830. Authors are
551 grateful to Drs. Marina Vance and Delphine Farmer who have organized the sampling for
552 HomeChem indoor chemistry project (<https://indoorchem.org/projects/homechem/>) that
553 allowed to collect samples for the dataset MSV000083598. Brandon Ross has assisted with
554 collecting data for the dataset MSV000084348. GC-MS analyses for collection of the
555 datasets MSV000084211 and MSV000084212 were supported by the announcement N757
556 Doctorados Nacionales and project EXT-2016-69-1713 from Departamento Administrativo de
557 Ciencia, Tecnología e Innovación (COLCIENCIAS), the seed project INV-2019-67-1747 and
558 FAPA project of Chiara Carazzone from the Faculty of Science at Universidad de los Andes,
559 and the grant No. FP80740-064-2016 of COLCIENCIAS. Authors are grateful to Lida M.
560 Garzón, Pablo Palacios, Marco Gonzalez and Jack Hernandez for their contributions
561 collecting the samples, and to Jhony Oswaldo Turizo for designing and manufacturing the
562 amphibian electrical stimulator.

563

564

565

566

567

568

569 **References**

570

571

- 572 1. Pizzoferrato, L., Nicoli, S. & Lintas, C. GC-MS characterization and quantification of
573 sterols and cholesterol oxidation products. *Chromatographia* vol. 35 269–274 (1993).
- 574 2. Coldwell, R. D., Porteous, C. E., Trafford, D. J. H. & Makin, H. L. J. Gas
575 chromatography—mass spectrometry and the measurement of vitamin D metabolites in
576 human serum or plasma. *Steroids* vol. 49 155–196 (1987).
- 577 3. Mr, J., Johnston, M. R. & Sobhi, H. F. Advances in Fatty Acid Analysis for Clinical
578 Investigation and Diagnosis using GC/MS Methodology. *Journal of Biochemistry and*
579 *Analytical studies* vol. 3 (2018).
- 580 4. Du, X. & Zeisel, S. H. Spectral deconvolution for gas chromatography mass
581 spectrometry-based metabolomics: current status and future perspectives. *Comput.*
582 *Struct. Biotechnol. J.* **4**, e201301013 (2013).
- 583 5. Stein, S. E. An integrated method for spectrum extraction and compound identification
584 from gas chromatography/mass spectrometry data. *Journal of the American Society for*
585 *Mass Spectrometry* vol. 10 770–781 (1999).
- 586 6. Stein, S. Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical
587 Identification. *Analytical Chemistry* vol. 84 7274–7282 (2012).
- 588 7. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical
589 analysis of biology by mass spectrometry. *Nature Reviews Chemistry* vol. 1 (2017).
- 590 8. Smirnov, A. *et al.* ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve
591 Resolution to Spectral Deconvolution of Gas Chromatography–Mass Spectrometry
592 Metabolomics Data. *Analytical Chemistry* vol. 91 9069–9077 (2019).
- 593 9. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive
594 metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).

- 595 10. Lommen, A. & Kools, H. J. MetAlign 3.0: performance enhancement by efficient use of
596 advances in computer hardware. *Metabolomics* **8**, 719–726 (2012).
- 597 11. Styczynski, M. P. *et al.* Systematic identification of conserved metabolites in GC/MS data
598 for metabolomics and biomarker discovery. *Anal. Chem.* **79**, 966–973 (2007).
- 599 12. Amigo, J. M., Skov, T., Bro, R., Coello, J. & MasPOCH, S. Solving GC-MS problems with
600 PARAFAC2. *TrAC Trends in Analytical Chemistry* vol. 27 714–725 (2008).
- 601 13. Kessler, N. *et al.* MeltDB 2.0—advances of the metabolomics software system.
602 *Bioinformatics* **29**, 2452–2459 (2013).
- 603 14. Domingo-Almenara, X. *et al.* eRah: A Computational Tool Integrating Spectral
604 Deconvolution and Alignment with Quantification and Identification of Metabolites in
605 GC/MS-Based Metabolomics. *Analytical Chemistry* vol. 88 9821–9829 (2016).
- 606 15. Skogerson, K., Wohlgemuth, G., Barupal, D. K. & Fiehn, O. The volatile compound
607 BinBase mass spectral database. *BMC Bioinformatics* **12**, 321 (2011).
- 608 16. Akiyama, K. *et al.* PRIME: a Web site that assembles tools for metabolomics and
609 transcriptomics. *In Silico Biol.* **8**, 339–345 (2008).
- 610 17. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life
611 sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
- 612 18. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics
613 data and metadata, metabolite standards, protocols, tutorials and training, and analysis
614 tools. *Nucleic Acids Res.* **44**, D463–70 (2016).
- 615 19. Carroll, A. J., Badger, M. R. & Harvey Millar, A. The MetabolomeExpress Project:
616 enabling web-based processing, analysis and transparent dissemination of GC/MS
617 metabolomics datasets. *BMC Bioinformatics* **11**, 376 (2010).
- 618 20. Haug, K. *et al.* MetaboLights—an open-access general-purpose repository for
619 metabolomics studies and associated meta-data. *Nucleic Acids Research* vol. 41 D781–
620 D786 (2013).

- 621 21. Hummel, J. *et al.* Mass Spectral Search and Analysis Using the Golm Metabolome
622 Database. *The Handbook of Plant Metabolomics* 321–343 (2013)
623 doi:10.1002/9783527669882.ch18.
- 624 22. Kim, S., Gupta, N., Bandeira, N. & Pevzner, P. A. Spectral dictionaries: Integrating de
625 novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell.*
626 *Proteomics* **8**, 53–69 (2009).
- 627 23. Guthals, A., Watrous, J. D., Dorrestein, P. C. & Bandeira, N. The spectral networks
628 paradigm in high throughput mass spectrometry. *Mol. Biosyst.* **8**, 2535–2544 (2012).
- 629 24. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc.*
630 *Natl. Acad. Sci. U. S. A.* **109**, E1743–52 (2012).
- 631 25. Varmuza, K., Karlovits, M. & Demuth, W. Spectral similarity versus structural similarity:
632 infrared spectroscopy. *Analytica Chimica Acta* vol. 490 313–324 (2003).
- 633 26. Elie, N., Santerre, C. & Touboul, D. Generation of a Molecular Network from Electron
634 Ionization Mass Spectrometry Data by Combining MZmine2 and MetGem Software.
635 *Anal. Chem.* **91**, 11489–11492 (2019).
- 636 27. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and
637 stewardship. *Sci Data* **3**, 160018 (2016).
- 638 28. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based
639 Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry* vol. 84 5035–
640 5039 (2012).
- 641 29. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global
642 Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
- 643 30. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for
644 processing, visualizing, and analyzing mass spectrometry-based molecular profile data.
645 *BMC Bioinformatics* vol. 11 (2010).
- 646 31. Wenig, P. & Odermatt, J. OpenChrom: a cross-platform open source software for the

- 647 mass spectrometric analysis of chromatographic data. *BMC Bioinformatics* **11**, 405
648 (2010).
- 649 32. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high
650 resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).
- 651 33. He, Q. P., Wang, J., Mobley, J. A., Richman, J. & Grizzle, W. E. Self-calibrated warping
652 for mass spectra alignment. *Cancer Inform.* **10**, 65–82 (2011).
- 653 34. Qiu, F., Lei, Z. & Sumner, L. W. MetExpert: An expert system to enhance gas
654 chromatography–mass spectrometry-based metabolite identifications. *Anal. Chim. Acta*
655 **1037**, 316–326 (2018).
- 656 35. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a
657 comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
- 658 36. AIST:Spectral Database for Organic Compounds,SDBS.
659 <https://sdfs.db.aist.go.jp/sdfs/cgi-bin/ENTRANCE.cgi>.
- 660 37. Kind, T. *et al.* FiehnLib: mass spectral and retention index libraries for metabolomics
661 based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal.*
662 *Chem.* **81**, 10038–10048 (2009).
- 663 38. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic*
664 *Acids Res.* **46**, D608–D617 (2018).
- 665 39. Remoroza, C. A., Mak, T. D., De Leoz, M. L. A., Mirokhin, Y. A. & Stein, S. E. Creating a
666 Mass Spectral Reference Library for Oligosaccharides in Human Milk. *Analytical*
667 *Chemistry* vol. 90 8977–8988 (2018).
- 668 40. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis.
669 *Metabolomics* vol. 3 211–221 (2007).
- 670 41. Nothias, L. F. *et al.* Feature-based Molecular Networking in the GNPS Analysis
671 Environment. *Bioinformatics* **143** (2019).
- 672 42. Stauffer, D. B., McLafferty, F. W., Ellis, R. D. & Peterson, D. W. Probability-based-

- 673 matching algorithm with forward searching capabilities for matching unknown mass
674 spectra of mixtures. *Analytical Chemistry* vol. 57 1056–1060 (1985).
- 675 43. Stauffer, D. B., McLafferty, F. W., Ellis, R. D. & Peterson, D. W. Adding forward
676 searching capabilities to a reverse search algorithm for unknown mass spectra.
677 *Analytical Chemistry* vol. 57 771–773 (1985).
- 678 44. Ernst, M. *et al.* MolNetEnhancer: enhanced molecular networks by integrating
679 metabolome mining and annotation tools. *bioRxiv* 654459 (2019) doi:10.1101/654459.
- 680 45. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine.
681 *Computer Networks and ISDN Systems* vol. 30 107–117 (1998).
- 682 46. Daly, J. W., Witkop, B., Tokuyama, T., Nishikawa, T. & Karle, I. L. Gephyrotoxins,
683 histrionicotoxins and pumiliotoxins from the neotropical frog *Dendrobates histrionicus*.
684 *Helv. Chim. Acta* **60**, 1128–1140 (1977).
- 685 47. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing
686 data. *Nat. Methods* **7**, 335–336 (2010).
- 687 48. Bolyen, E. *et al.* Author Correction: Reproducible, interactive, scalable and extensible
688 microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 1091 (2019).
- 689 49. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*
690 **15**, 796–798 (2018).
- 691 50. Chong, J., Wishart, D. S. & Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and
692 Integrative Metabolomics Data Analysis. *Current Protocols in Bioinformatics* vol. 68
693 (2019).
- 694 51. Xia, J. & Wishart, D. S. Metabolomic Data Processing, Analysis, and Interpretation Using
695 MetaboAnalyst. *Current Protocols in Bioinformatics* vol. 34 14.10.1–14.10.48 (2011).
- 696 52. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of
697 biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
- 698 53. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph

- 699 Layout Algorithm for Handy Network Visualization Designed for the Gephi Software.
700 *PLoS ONE* vol. 9 e98679 (2014).
- 701 54. Protsyuk, I. *et al.* 3D molecular cartography using LC-MS facilitated by Optimus and 'ili
702 software. *Nat. Protoc.* **13**, 134–154 (2018).
- 703 55. Einoch Amor, R., Nakhleh, M. K., Barash, O. & Haick, H. Breath analysis of cancer in the
704 present and the future. *Eur. Respir. Rev.* **28**, (2019).
- 705 56. Schmidt, K. & Podmore, I. Current Challenges in Volatile Organic Compounds Analysis
706 as Potential Biomarkers of Cancer. *J Biomark* **2015**, 981458 (2015).
- 707 57. Lawal, O., Ahmed, W. M., Nijsen, T. M. E., Goodacre, R. & Fowler, S. J. Exhaled breath
708 analysis: a review of 'breath-taking' methods for off-line analysis. *Metabolomics* **13**, 110
709 (2017).
- 710 58. Alkhalifah, Y. *et al.* VOCCluster: Untargeted Metabolomics Feature Clustering Approach
711 for Clinical Breath Gas Chromatography - Mass Spectrometry Data. *Anal. Chem.* (2019)
712 doi:10.1021/acs.analchem.9b03084.
- 713 59. Poli, D. *et al.* Determination of aldehydes in exhaled breath of patients with lung cancer
714 by means of on-fiber-derivatisation SPME–GC/MS. *Journal of Chromatography B* vol.
715 878 2643–2651 (2010).
- 716 60. Phillips, M. *et al.* Volatile organic compounds in breath as markers of lung cancer: a
717 cross-sectional study. *The Lancet* vol. 353 1930–1933 (1999).
- 718 61. Phillips, M. *et al.* Detection of lung cancer with volatile markers in the breath. *Chest* **123**,
719 2115–2123 (2003).
- 720 62. Fuchs, P., Loeseken, C., Schubert, J. K. & Miekisch, W. Breath gas aldehydes as
721 biomarkers of lung cancer. *International Journal of Cancer* NA–NA (2009)
722 doi:10.1002/ijc.24970.
- 723 63. Phillips, M. *et al.* Prediction of breast cancer using volatile biomarkers in the breath.
724 *Breast Cancer Research and Treatment* vol. 99 19–21 (2006).

- 725 64. Amal, H. *et al.* Assessment of ovarian cancer conditions from exhaled breath.
726 *International Journal of Cancer* vol. 136 E614–E622 (2015).
- 727 65. Altomare, D. F. *et al.* Exhaled volatile organic compounds identify patients with colorectal
728 cancer. *British Journal of Surgery* vol. 100 144–150 (2013).
- 729 66. Markar, S. R. *et al.* Assessment of a Noninvasive Exhaled Breath Test for the Diagnosis
730 of Oesophagogastric Cancer. *JAMA Oncology* vol. 4 970 (2018).
- 731 67. Amal, H. *et al.* Detection of precancerous gastric lesions and gastric cancer through
732 exhaled breath. *Gut* **65**, 400–407 (2016).
- 733 68. Kumar, S. *et al.* Mass Spectrometric Analysis of Exhaled Breath for the Identification of
734 Volatile Organic Compound Biomarkers in Esophageal and Gastric Adenocarcinoma.
735 *Ann. Surg.* **262**, 981–990 (2015).
- 736 69. Sihvo, E. I. T. *et al.* Oxidative stress has a role in malignant transformation in Barrett's
737 oesophagus. *International Journal of Cancer* vol. 102 551–555 (2002).
- 738 70. Phillips, M., Greenberg, J. & Sabas, M. Alveolar gradient of pentane in normal human
739 breath. *Free Radic. Res.* **20**, 333–337 (1994).
- 740 71. Risby, T. H. & Sehnert, S. S. Clinical application of breath biomarkers of oxidative stress
741 status. *Free Radical Biology and Medicine* vol. 27 1182–1192 (1999).
- 742 72. Pappas, A. Epidermal surface lipids. *Dermatoendocrinol.* **1**, 72–76 (2009).
- 743 73. Martin, A. *et al.* A functional ABCC11 allele is essential in the biochemical formation of
744 human axillary odor. *J. Invest. Dermatol.* **130**, 529–540 (2010).
- 745 74. Li, H., Jiang, T. & Zhang, K. Efficient and robust feature extraction by maximum margin
746 criterion. *IEEE Trans. Neural Netw.* **17**, 157–165 (2006).
- 747 75. Wang, M. *et al.* MASST: A Web-based Basic Mass Spectrometry Search Tool for
748 Molecules to Search Public Data. doi:10.1101/591016.
- 749 76. Jarmusch, A. K. *et al.* Repository-scale Co- and Re-analysis of Tandem Mass
750 Spectrometry Data. *bioRxiv* 750471 (2019) doi:10.1101/750471.

751 77. Masci, J., Rodolà, E., Boscaini, D., Bronstein, M. M. & Li, H. Geometric deep learning.

752 *SIGGRAPH ASIA 2016 Courses on - SA '16* (2016) doi:10.1145/2988458.2988485.

753