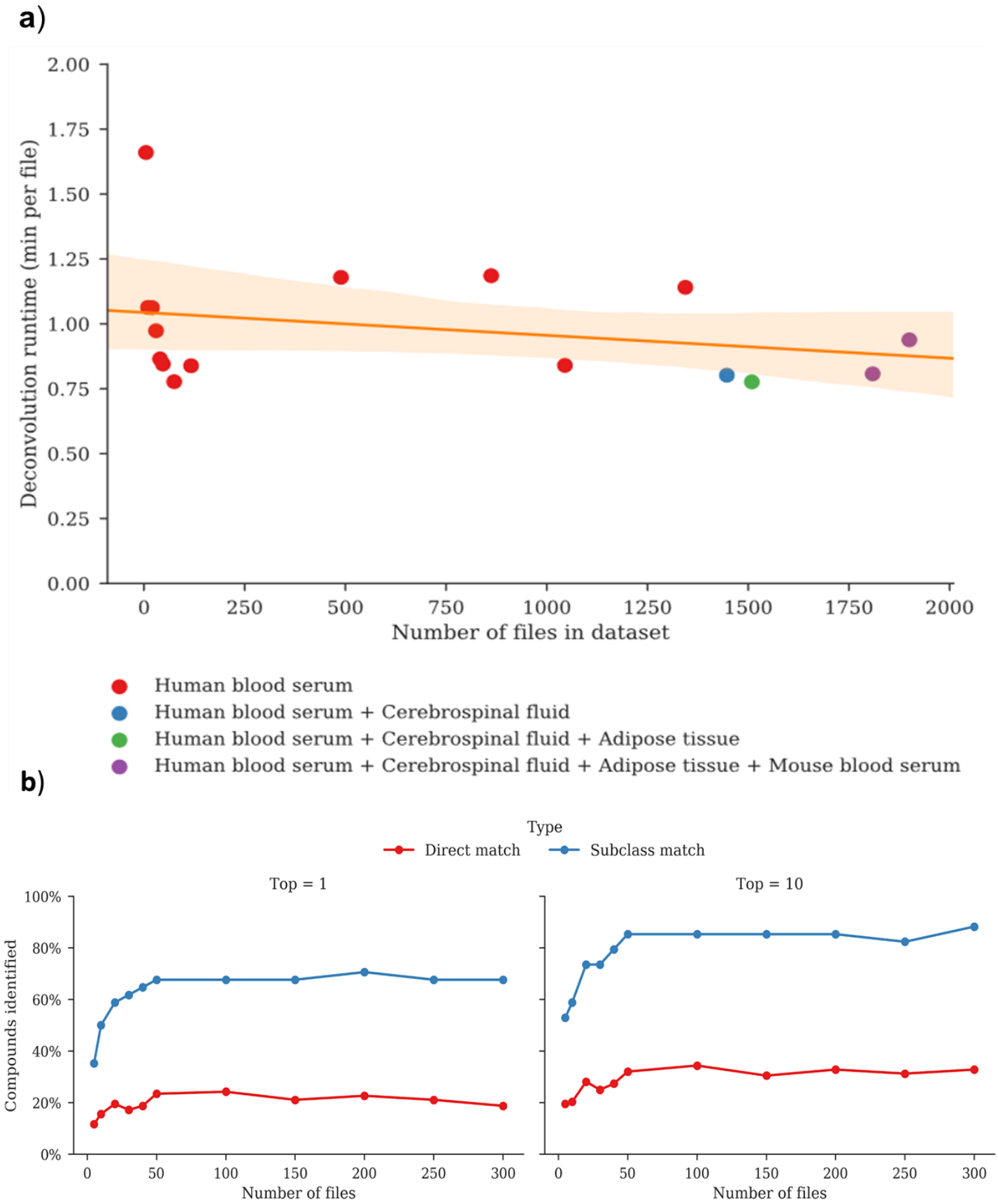
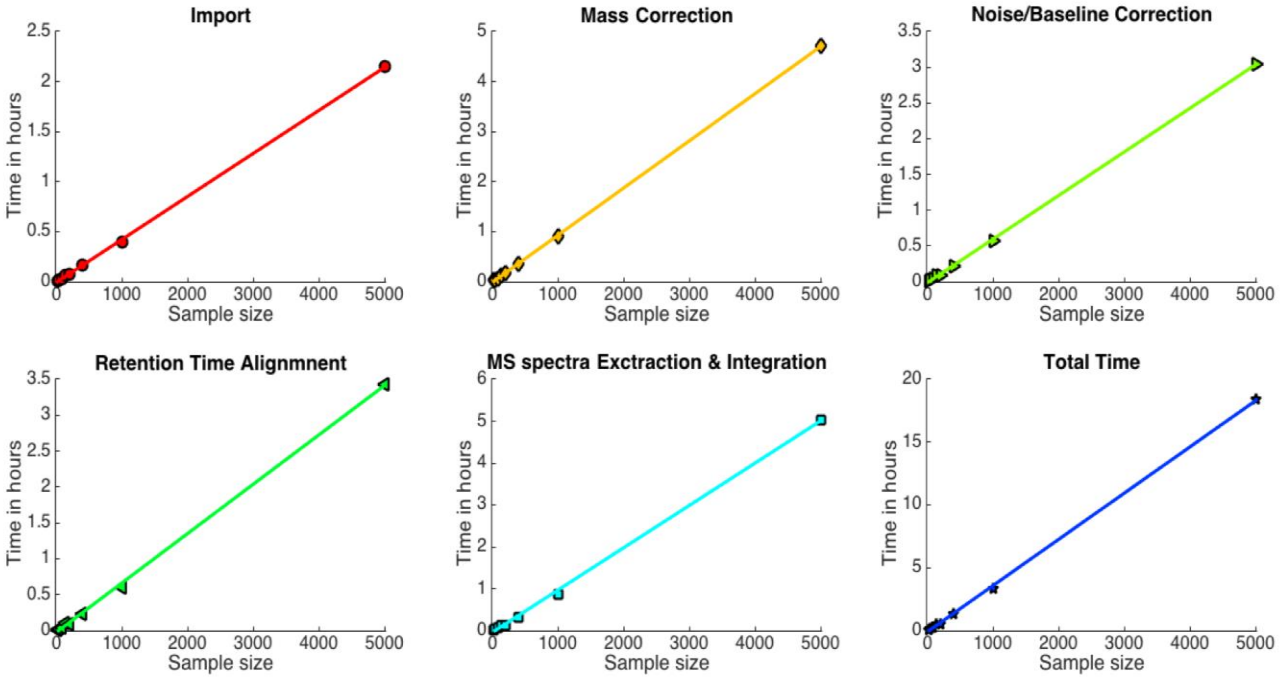


Figure S1. MSHub performance evaluation. a) MSHub analysis times per file in the dataset. The data collected using identical protocols have been processed by the GNPS MSHub workflow with the increasing number of files included in the dataset (UCD1 - UCD16, **Supplemental Table S1**). The human blood serum files only have been included initially followed by adding cerebrospinal fluid, adipose tissue and mouse blood serum samples. The graph is generated as described in the "Methods" section. b) The annotation percentage of the "known" (curated) compounds for subsets (**Table S8**) of the MSV000084039 dataset (Metabolomics Workbench ST001154, derivatized mouse blood serum) (**Table S6**). The ways data are subsetted can impact the results as different files contain different amounts of information (the randomness is mitigated when considering top 10 matches instead of top 1). For this dataset, it appears that MSHub is able to extract additional information for up to 50-100 files. The magnitude of increase is similar for top 1 and top 10 and is ~ 15% for direct match and ~33% for compound subclass.



The workflow memory load is constant (*i.e.* one sample data set is processed at a time). ~200Gb of raw data for 5000 samples

Figure S2. MSHub scalability testing on a single core desktop PC. The dependency between the number of samples processed and the processing time for each of the MSHub modules performed on a single core desktop PC.

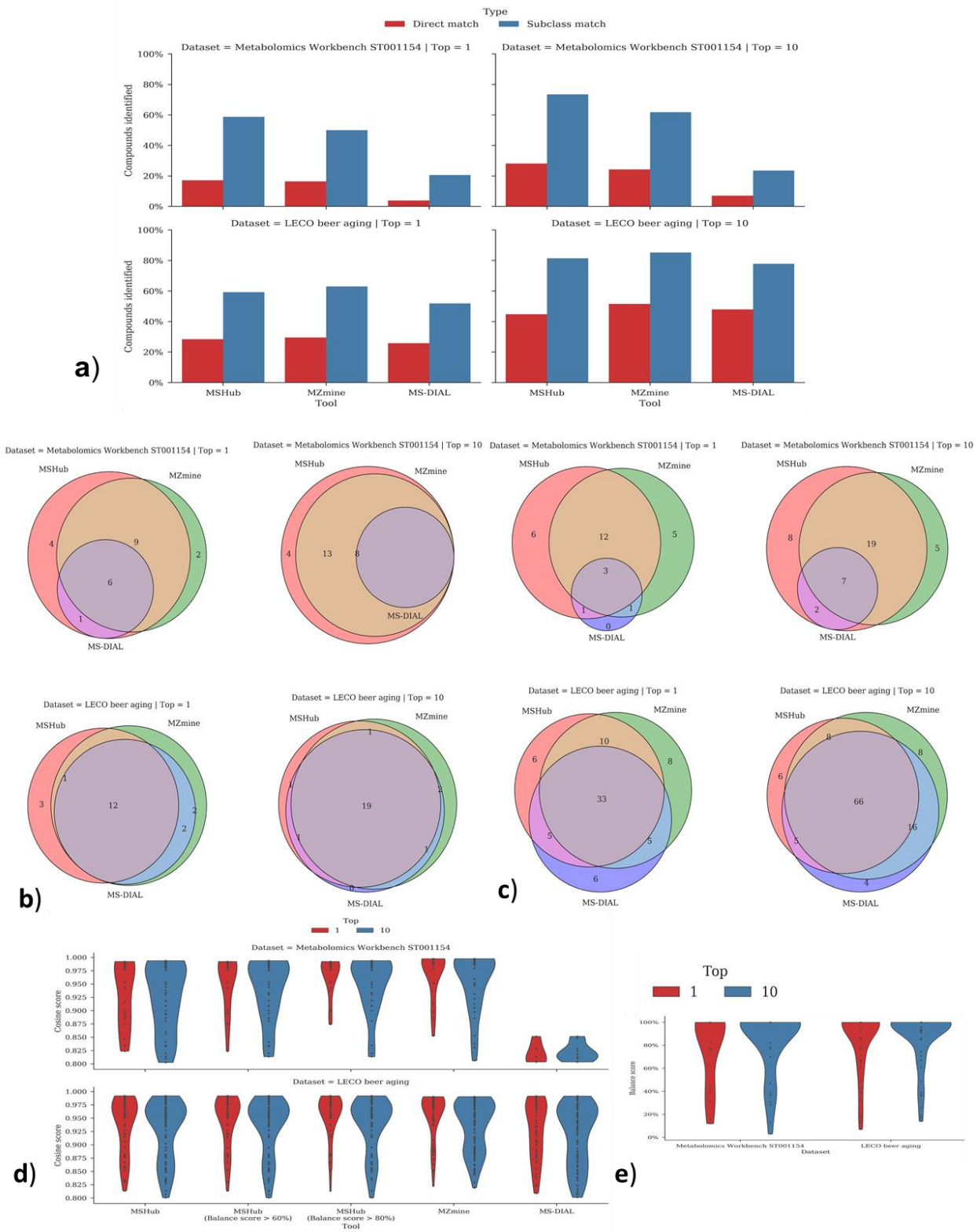


Figure S3. Comparison of deconvolution accuracy judged by proxy of annotation success rate (percentage of "known" compounds in the test samples that were automatically annotated) of MSHub, MZmine 2/ADAP and MS-DIAL deconvolution tools for test datasets: MSV000084039 (Metabolomics Workbench ST001154, derivatized mouse blood serum) and MSV000084349 (LECO beer aging reference data). The settings of MZmine 2/ADAP and MS-DIAL were optimized as described in the "Online Methods" by the ADAP algorithm developers; MSHub was run with automatically determined settings. For the dataset MSV000084039, the settings for MS-DIAL had to be adjusted by increasing the noise threshold tenfold to enable the processing without tool failing. The links to library search jobs are summarized in **Table S3**. **a**) Percentage of "known" (curated) compounds correctly annotated directly or by its subclass. Annotations at the chemical family level could serve as a good proxy benchmark of the ability to resolve spectral patterns, even if they are not annotated to the exact correct compound. For the MSV000084349 dataset, the sub-class level annotation in top 10 matches reaches close to 100% for MSHub and MZmine 2/ADAP, which is indicative of essentially all of the spectral patterns for the known/curated compounds being sufficiently well recovered by the deconvolution tool to be attributable to correct chemical family. **b**) Venn diagram of correctly directly annotated compounds from deconvolution by each tool. **c**) Venn diagram of correctly annotated compound classes. **d**) Violin plot of the cosine match scores for MSHub (directly and with 60% and 80% balance score filtering applied), MZmine 2/ADAP and MS-DIAL. Application of balance score filter for MSHub removes poor quality spectra **e**) Violin plot of match score values as the function of balance score for the datasets MSV000084039 (Metabolomics Workbench ST001154, derivatized mouse blood serum, **Table S6**) and MSV000084349 (LECO beer aging reference data, **Table S7**) for the "known" (curated) compounds correctly identified in the top 1 and top 10 matches.

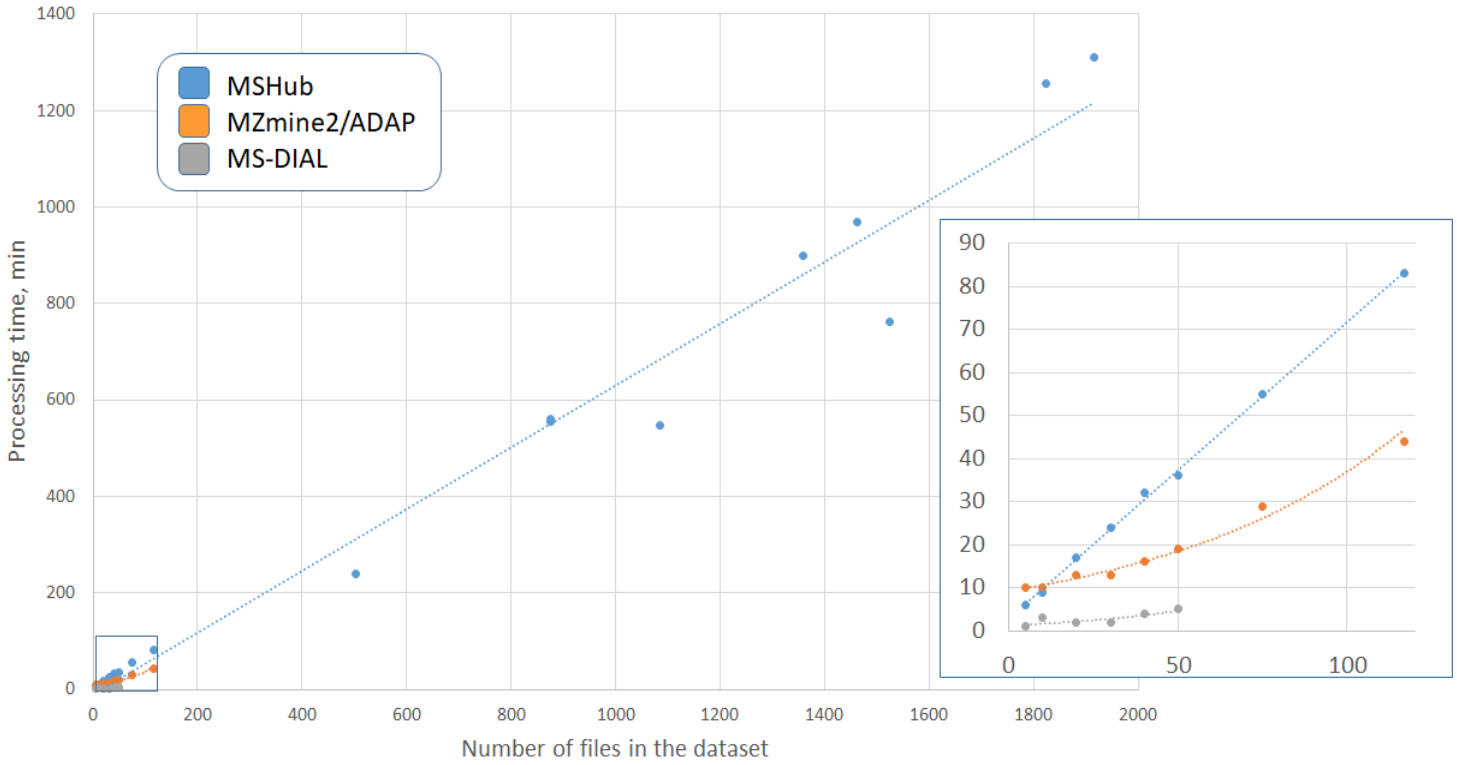


Figure S4. Comparison of processing times of MSHub with MZmine 2/ADAP and MS-DIAL on a 12 CPUs and 248GB RAM system of the test data (**Supplemental Table S4**). The times are plotted for successfully processed subsets; MSHub successfully processed all of the tested data (largest tested set is 1914 files), MZmine 2/ADAP fails with insufficient memory error at <530 files, MS-DIAL fails at <75 files. The inset panel shows the portion of the plot within the box, where all three tools can operate successfully; the scaling of deconvolution time with the number of files is exponential for MZmine 2/ADAP and MS-DIAL and linear for MSHub.

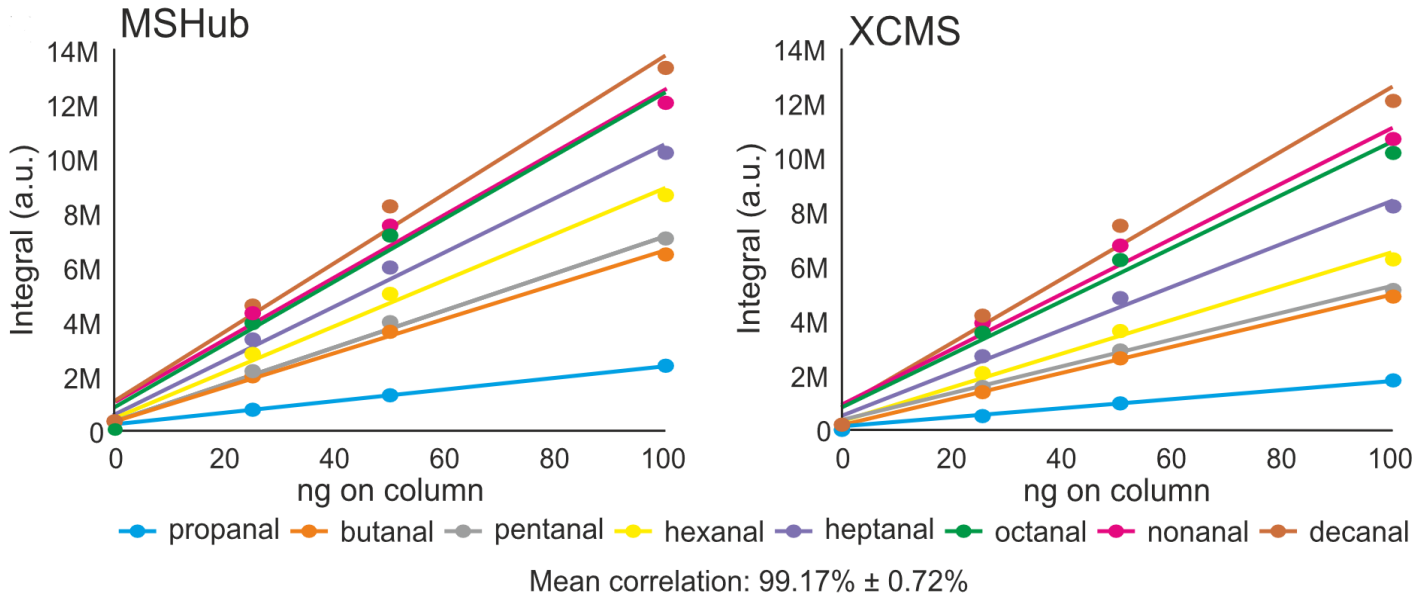


Figure S5. Comparison of quantitation accuracy of MSHub with XCMS of the dataset MSV000084622. The performance of both tools is nearly identical (the calibration curve is within 99.17% correlation with 0.72% STD.

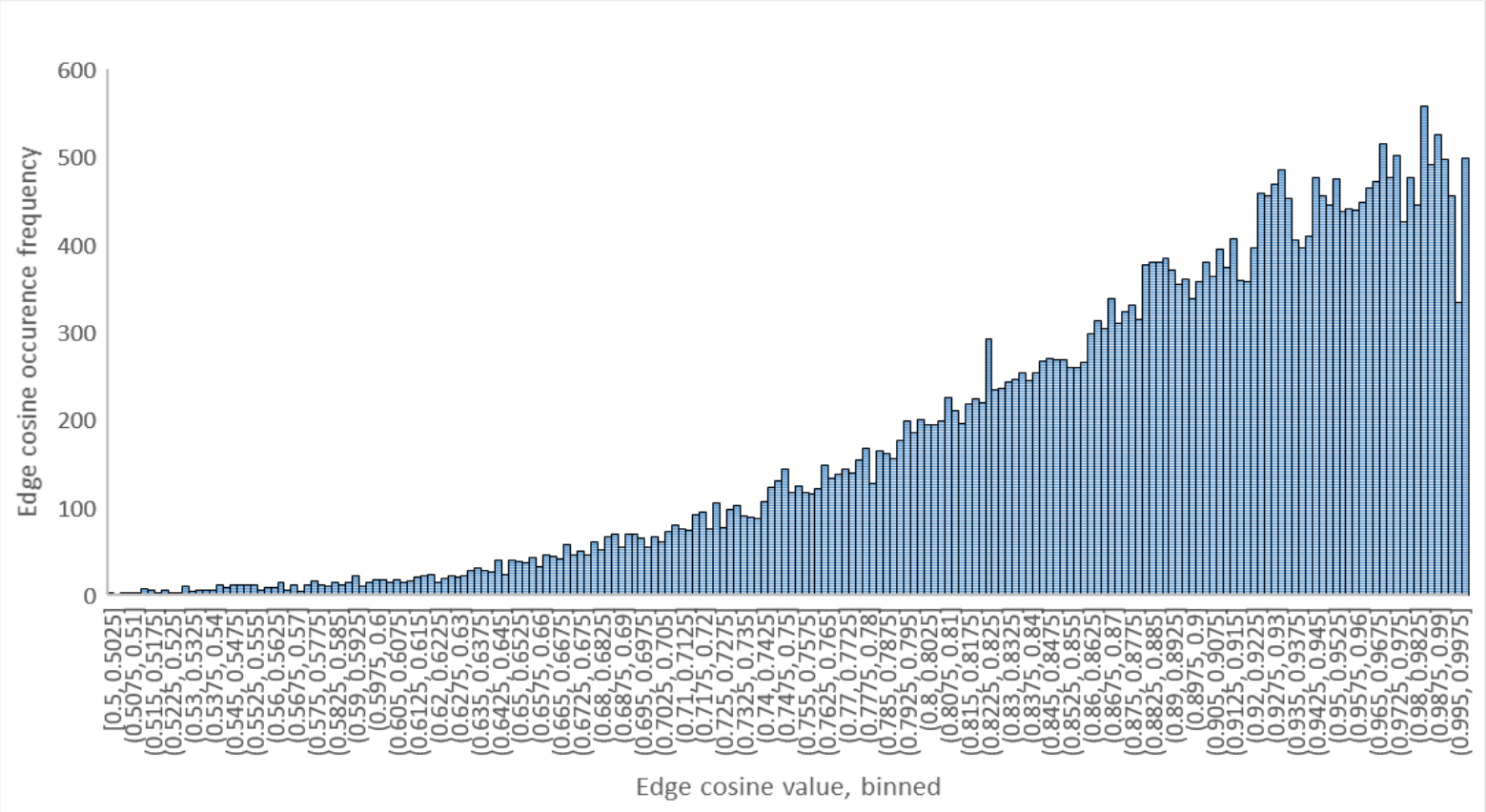


Figure S6. Global network overview (Figure 2, the datasets across GNPS (#1-38 in Table S1). Distribution of the cosine for all edges in the global network.

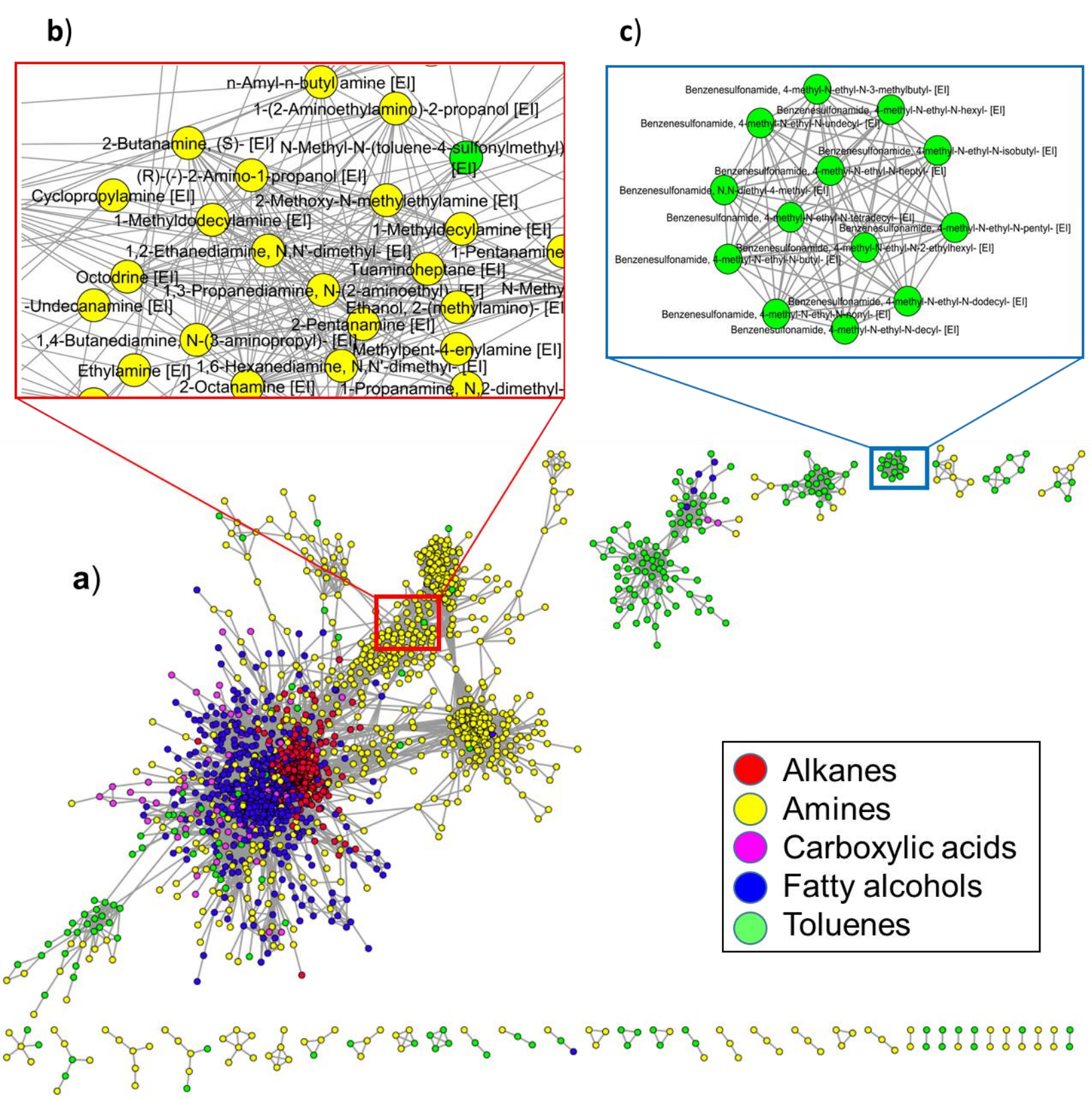


Figure S7. The basis for the EI-based molecular networking. Molecular network for the reference spectra of the NIST 14 reference library; only nodes with connections are displayed. All library compounds were classified using ClassyFire (*J. Cheminform.* **8**, 61, 2016). and examples of commonly encountered molecular subclasses were selected based on the following ClassyFire chemical subclass ontology terms of the top annotation for each node: alkanes, amines, carboxylic acids, fatty alcohols and toluenes. a) Network appearance at the cosine threshold of 0.85 for all of the compounds across five subclasses shows clustering largely follows chemical similarity: aliphatic compounds create a large cluster with separation by functional group, while aryls form distinct clusters. b) The part of the network within the red inset box on Figure 2a that shows portion of the amine cluster with high interconnectivity indicating broad spectral similarities of the EI spectra for these molecules. c) The part of the network within the blue inset box on Figure 2c showing the cluster of Benzenesulfonamides: all members of the clusters are interconnected, but not connected to any other nodes in the network indicating unique appearance of EI spectra of these compounds.

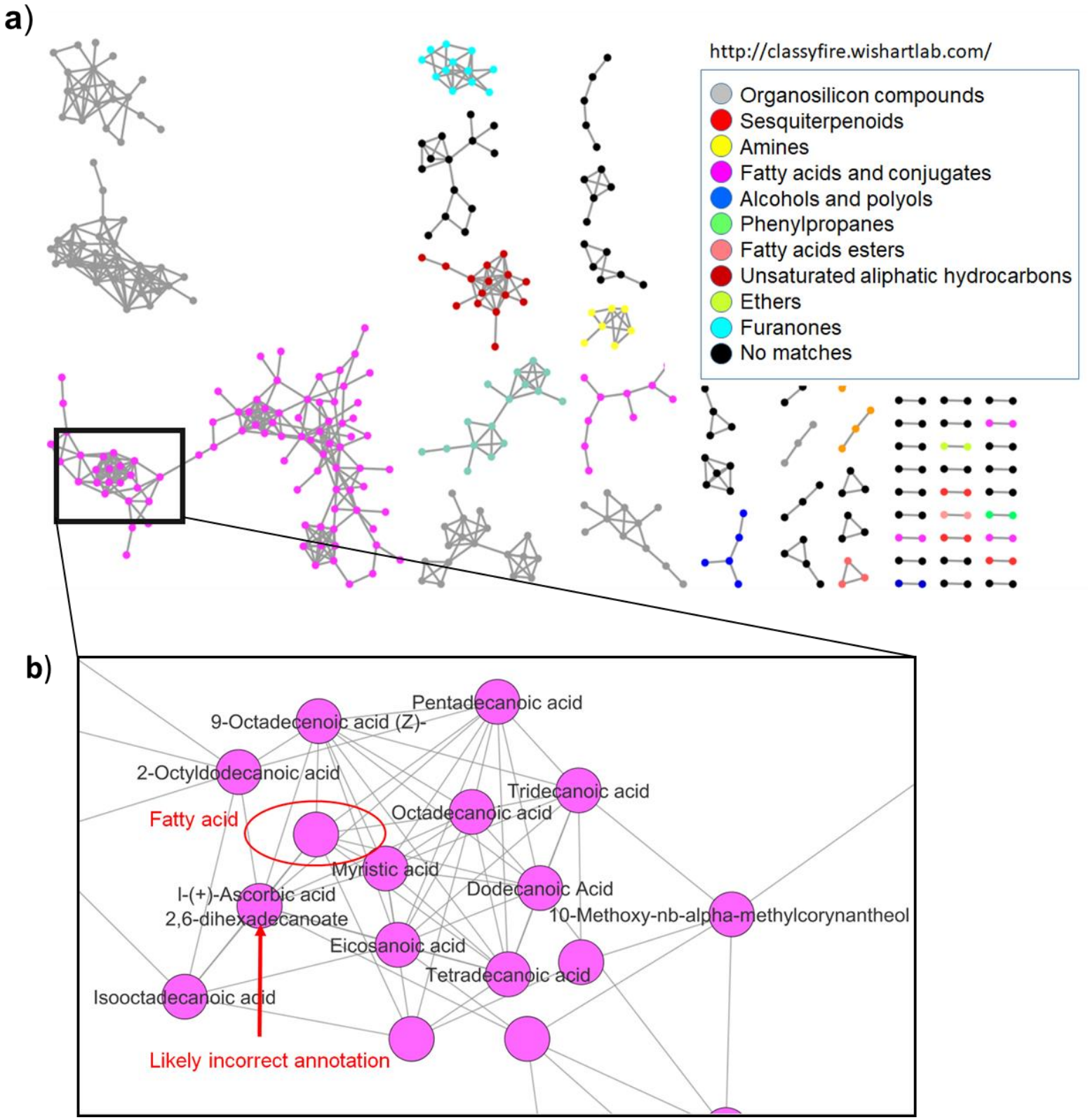


Figure S8. Example of molecular network application for "annotation guidance". a) a network of skin volatiles (dataset #19 in Table S1) is shown, only connected nodes are displayed. The MolNetEnhancer (*Metabolites* **9**, 2019). workflow was adapted to work with GC-MS data. As a result, nodes are colored by the chemical subclass term mostly occurring in the molecular family. For cases where no chemical structures (annotation) could be retrieved for any of the nodes within a molecular family using a cosine threshold of 0.5, the cluster is labelled as "No matches". b) Demonstration of the molecular network annotation guidance. The portion of the network corresponding to the "fatty acids and conjugates" cluster (box on the panel (a) is shown; the circled unannotated node is presumed to be the fatty acid according to the annotations of the nearby nodes. It was manually verified to be a fatty acid with 14 carbons or greater. Also shown is an example of a node (indicated by arrow) with an annotation (top hit) that does not fall within expected chemical subclass and is therefore suspected to be incorrect. It has been manually verified that the most likely correct annotation for this node is hexadecanoic acid which is more in line with the other surrounding annotations connected to it.

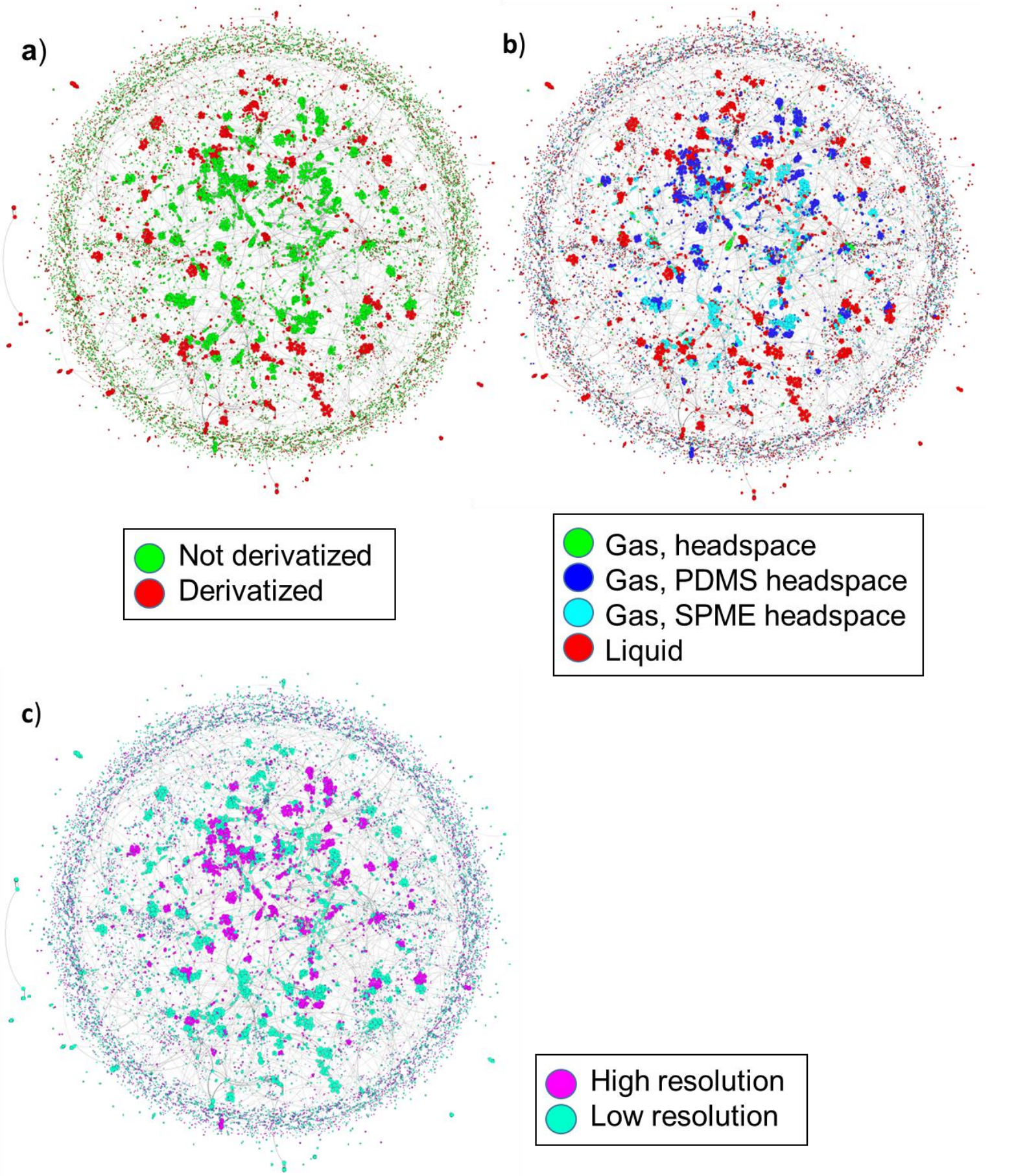


Figure S9. Networking in GNPS at repository scale. Global network containing 35,544 nodes from 8,489 files in 38 GNPS datasets for different types of samples. The size of the node is proportional to its connectivity (*Computer Networks and ISDN Systems* vol. 30 107–117, 1998), the edge thickness is proportional to the cosine score. Global network for the GNPS-wide data colored by a) derivatization status b) sample introduction mode c) input data resolution. In all cases distinct clustering can be observed, driven by distinct chemistries in the individual datasets.

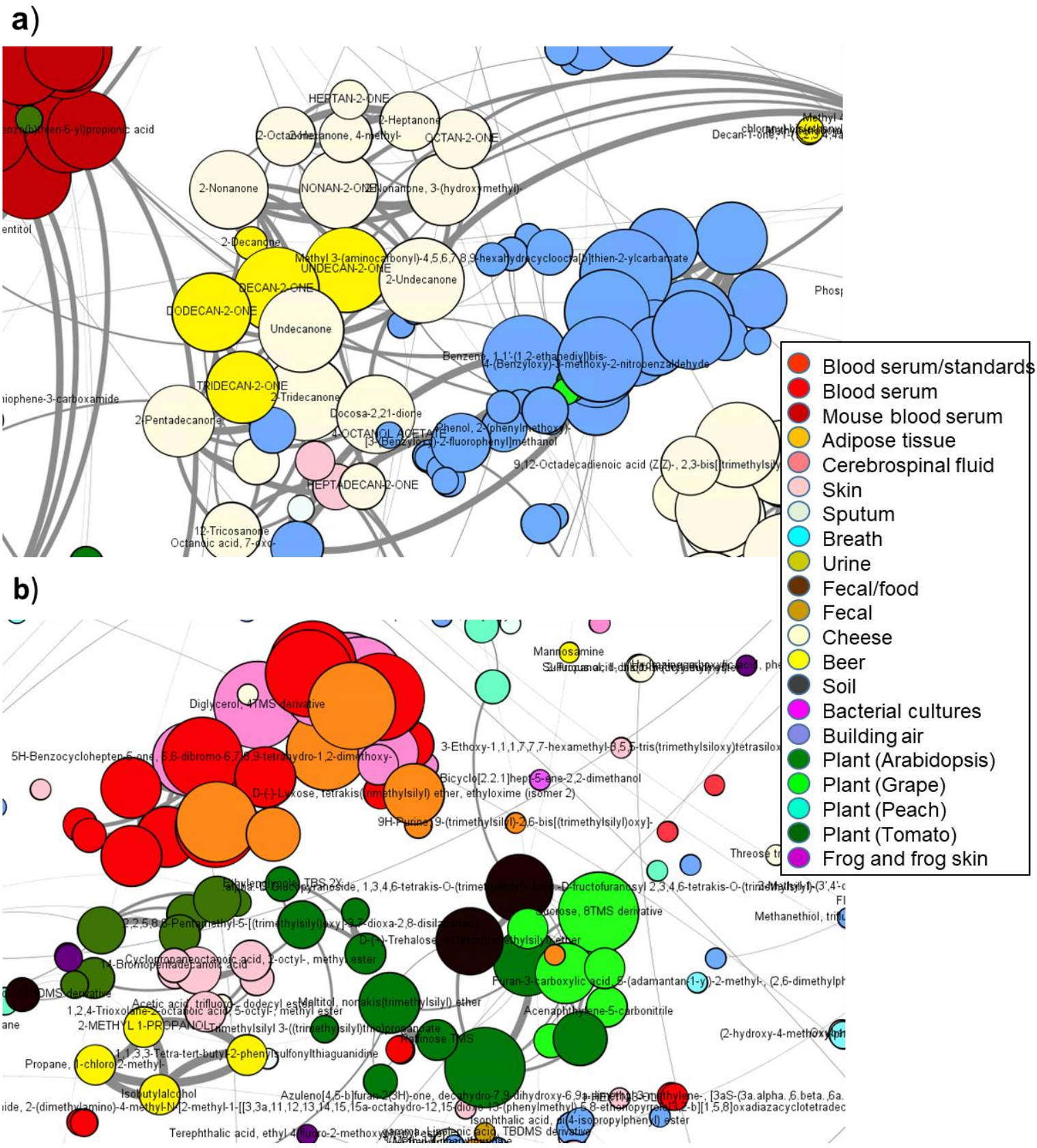


Figure S10. Networking in GNPS at repository scale. For the global network shown on Figure 3c colored by different types of samples. a) Close-up of a cluster of long and medium chain ketones which occur in cheese and beer and contribute to their flavors. b) Sugars are found across different types of plants and in soil.

Table S1. Pipeline testing

Dataset description	Dataset	MassIVE ID	Reference	Sample i	Derivatized?	Uberon/sample type	Instrument	Data type	Data file	Numbe	proteomics2 deconvolution link	GNPS deconvolution link	GNPS deconvolution link latest workflow	proteomics2 library search	GNPS library search link	GNPS library search link latest workflow
Metabolomics Workbench_ST000048	#1	MSV000084033		Liquid	Yes	Human blood serum	ToF	Low res	cdf	117	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f495c7318d6e4d45847ae979231c55a7
Metabolomics Workbench_ST000385	#2	MSV000085136		Liquid	Yes	Human blood serum	ToF	Low res	cdf	372	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=916a273dfd784949ae557a07594df421
Metabolomics Workbench_ST000386	#3	MSV000084034		Liquid	Yes	Human blood serum	ToF	Low res	cdf	373	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1da5ecf3f66e485bb03388c5025b3142
Metabolomics Workbench_ST000392	#4	MSV000084036		Liquid	Yes	Human blood serum	ToF	Low res	cdf	183	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=bc506e97d8b34869bf3ea93c2217a059
Metabolomics Workbench_ST000396	#5	MSV000084032		Liquid	Yes	Human blood serum	ToF	Low res	cdf	300	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2fb0f17e6ea74e34882d4a21812042d2
Metabolomics Workbench_ST000065	#6	MSV000084038		Liquid	Yes	Adipose tissue	ToF	Low res	cdf	62	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0c7bbfb0cdca476fa129407665a42334
Metabolomics Workbench_ST000061	#7	MSV000084042		Liquid	Yes	Adipose tissue	ToF	Low res	cdf	128	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0c7bbfb0cdca476fa129407665a42334
Metabolomics Workbench_ST001154	#8	MSV000084039		Liquid	Yes	Mouse blood serum	ToF	Low res	cdf	300	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=809a2a665cea466bba4fb62a7d1e8e9e
Metabolomics Workbench_ST000910	#9	MSV000084040		Liquid	Yes	Cerebrospinal fluid	ToF	Low res	cdf	91	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=75dfd2fbd1614d68aa8c09d1a549ae31
Metabolomics Workbench_ST000063	#10	MSV000084037		Liquid	Yes	Cerebrospinal fluid	ToF	Low res	cdf	103	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1ade4b468ea049de91394f1a45dd3aee
VOCs from Dendrobatids frogs from Colombia	#11	MSV000084211		Gas, SPM	No	Frog+frog skin	Linear trap	Low res	cdf	97	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e9a42cdb293242f59bae12639288aaa5
Air samples from HomeChem project built home	#12	MSV000083598		Gas, PDM	No	Home air	QToF	High res	mzML	150	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3505ddc9489e407480d6a5875b249c7a
Data from cheese Stilton and Bonde cheese sar	#13	MSV000080892		Gas, SPM	No	Cheese	QqQ	Low res	mzML	297	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=56e3686734f647b1867ad64ac5d40529
Data from cheese Stilton only	#14	MSV000080892		Gas, SPM	No	Cheese	QqQ	Low res	mzML	141	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e5f711d9b56b48e59935ec2c2770fad0
Data from cheese Bonde only	#15	MSV000080892		Gas, SPM	No	Cheese	QqQ	Low res	mzML	156	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0575066891164587872334ebfcbf25e7
PDMS armpit samples pre-_ post bacterial spray	#16	MSV000084337		Gas, PDM	No	Skin	QToF	High res	mzML	133	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=5b7c22d6feb94da9b580171b9d239c8e
Test data of standards spiked into TMS derivatiz	#17	MSV000083658		Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML	156	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7857bf0861c243159feea42f7d2bd5b9
GC data for the EMP project from PNNL	#18	MSV000083743		Liquid	Yes	Soil	Linear trap	Low res	cdf	520	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=925a12b45fe84558b35faa049d2dcc50
PDMS 3D skin map pilot from 3 subjects	#19	MSV000084226		Gas, PDM	No	Skin	QToF	High res	mzML	80	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b6998ae352534290bb7d869ffcbab9ff
Ben-Gurion University data fecal from safari ani	#20	MSV000083859		Liquid	Yes	Fecal+food	Linear trap	Low res	cdf	322	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d182726dc5284a2f9980405641178619
Bacillus fungal biocontrol study, mixed cultures	#21	MSV000083294		Gas, SPM	No	Bacterial and fungal cu	QToF	High res	mzML	34	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c814f3220317405ba7854ea0cb11f3ea
LECO beer aging reference data	#22	MSV000084349		Gas, SPM	No	Beer	ToF	Low res	cdf	42	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16af80c30d704d9ca17a13f7751f1f3b
Sputum VOC data from: Niche partitioning of a p	#23	MSV000081	PMID: 307	Gas, SPM	No	Sputum	QqQ	Low res	mzML	109	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=92c3676f4f9a4d3a8c91b79c228b3ede
Longitudinal sputum VOC data from: Neutrophili	#24	MSV000084	PMID: 302	Gas, SPM	No	Sputum	QqQ	Low res	mzML	230	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e484f2f8bc4240a8bccaf03b1282c462
Skin swabs extracted with ethanol, liquid injectio	#25	MSV000084	PMID: 278	Liquid	No	Skin	QqQ	Low res	mzML	186	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=195c08ed29d942279f048e6f1f27e18b
3D skin mapping longitudinal and high spatial re	#26	MSV000084338		Gas, PDM	No	Skin	QToF	High res	mzML	125	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a51d446f8c7b4c5681c3575308bacdee
3D skin mapping high spatial resolution data, liq	#27	MSV000084339		Liquid	No	Skin	QToF	High res	mzML	111	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b458565aa85840eca69c9c119fadacfa
Data from cheese with bacterial cultures	#28	MSV000081161		Gas, SPM	No	Cheese	QqQ	Low res	mzML	71	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1b957b599fba430c82fe4b32622e203f
Arabidopsis samples, wild types and other muta	#29	MSV000084	PMID: 270	Liquid	Yes	Plant (Arabidopsis)	ToF	Low res	cdf	671	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=5a6cf5cd5e2b494d896c677b3736b192
University of Avignon Prunus persica after aphic	#30	MSV000084377		Liquid	Yes	Plant (Peach)	ToF	Low res	cdf	15	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6482f02c003843ae8dd4d8c157e2fec7
Ben-Gurion University Fait lab grapes	#31	MSV000084145		Liquid	Yes	Plant (Grape)	Quadrupole	Low res	mzML	1187	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8747240e2ab6457990f48f9059f481ca
Ben-Gurion University Fait lab tomatoes	#32	MSV000084144		Liquid	Yes	Plant (Tomato)	Quadrupole	Low res	mzML	620	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=bdffe135e9554f0babd6c93c3e537845
Sterols in urine - ONIRIS data	#33	MSV000084146		Liquid	No	Urine	Quadrupole	Low res	cdf	44	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=edc67728dd7a4dc7abb72938f9ffc54f
ICL breath analysis cancer Study1	#34	MSV000084379		Gas	No	Breath	ToF	High res	cdf	387	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=aa29c69b734d4615b94fbdab4e8c70ab
ICL breath analysis cancer Study2	#35	MSV000084380		Gas	No	Breath	ToF	High res	cdf	275	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=58f957d6239946e698358637696daa5e
NYU fecal samples study of SCFAs and MCFAs	#36	MSV000084	PMID: 317	Liquid	Yes	Fecal	QToF	High res	mzML	158	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c4ff381901f6425eac2178fbf1f0df44
NYU fecal samples study of SCFAs and MCFAs	#37	MSV000084	PMID: 317	Gas, SPM	No	Fecal	QToF	High res	mzML	69	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=017cb37baf5544078fa8b84244a754ef
Methanolic extracts from Dendrobatids from Col	#38	MSV000084212		Liquid	No	Frog+frog skin	Linear trap	Low res	cdf	74	https://proteomics2.ucsd.edu/ProteoSA	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://gnps.ucsd.edu/ProteoSAFe/status.js	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=80c61c764cac4fdda82e38b737e55596
Global network for datasets 1-38	#1-38	N/A		All	All	All	All	All	All	8489	N/A	N/A	N/A	https://proteomics2.ucsd.edu	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=90019623856142648c3470862b7596f4	
Test data of standards spiked into TMS derivatized b		MSV000083658														
5 files from Test data of standards spiked into T	Test1			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=37ae4d69cb66469f9ece048bc4f9d1d0			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=b18c2127f6a24732898a12374acad882		
10 files from Test data of standards spiked into	Test2			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=114f1442a3b848e19668b17fda6bcfa0			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5ca73b82c2f34f1f802371c306d03ce1		
20 files from Test data of standards spiked into	Test3			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=ff25efe5dbc8499eafcd0456104917c4			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=dd2e18fb540747979e65272c6fb5fc06		
30 files from Test data of standards spiked into	Test4			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=48740e99772e472981ad81f924f15dde			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d05ad1d5836f4dd083033d7de2dd9cd1		
40 files from Test data of standards spiked into	Test5			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=ce84321db132472f82e67ecfc39e0983			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=7a5d042d0af2489a83aafd1ba1b1b50f		
50 files from Test data of standards spiked into	Test6			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=094ade46e2524a1390f389d1640a6482			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=bc0bdd9861684a50b78fd2a721445f3c		
60 files from Test data of standards spiked into	Test7			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5e559d11f21b416cbf65ac75f976ded2			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e9cc9a462027460fa365a611be1f4310		
70 files from Test data of standards spiked into	Test8			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=aa2d4b077fbf4718ad3e2ee9450d1450			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=44f02503074a4705a7bd369057ad5d03		
79 files from Test data of standards spiked into	Test9			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3a1cf196f1bf45908e10b3113c4c6a3c			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=df34c3e720a94b5fa6786a4b55997f46		
88 files from Test data of standards spiked into	Test10			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=65d646937a154b619d805970b2b55073			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=0e9a73c125a542a2abd7019682d59797		
Various standards Test data of standards spiked	Test11			Liquid	Yes	Blood serum+standard	Orbitrap	High res	mzML		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=59bd420a944148b6b79ee8ea35cd6b2e			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5e38ef43e706458b8c075fc38258fca2		
Testing with ICL breath analysis cancer Study1 data (non-derivatized)																
5 files from ICL breath analysis cancer Study1	ICL1			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=b787fc37d48343c6937d9de6542cd368			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5ec18d5817c340c29b02d4c4f296ecf0		
10 files from ICL breath analysis cancer Study1	ICL2			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=0cd954f560c844b5957f5061b1d5591b			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=de18da7db57f42a2bb8abbea1017bb61		
20 files from ICL breath analysis cancer Study1	ICL3			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=79377c40209f42a39732d816b9dc8cf7			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5ae92eff97524d72ae080a5692ba81ac		
30 files from ICL breath analysis cancer Study1	ICL4			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=b0caab2e60004710a85899305ab2e326			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=fb5d6b7ad7b04694b8c36c35d2efcc28		
40 files from ICL breath analysis cancer Study1	ICL5			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=ff2e28050f304e57a6fa8f09c9b427a7			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=bb8f95bd4ce1493ba37c96907d5fa407		
50 files from ICL breath analysis cancer Study1	ICL6			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e45216c9cd1744fb8c14e59022fdc088			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=28ed7022f8b34f9aa49291a199b263f6		
75 files from ICL breath analysis cancer Study1	ICL7			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3373d446cbb84e4cb1f9df4a3c922f70			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=fdbd3050c10d4210b47a70ff865e0acc		
100 files from ICL breath analysis cancer Study	ICL8			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=13565f2694ca4d199ac1b03b240c701f			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d0028d7c17274254b1696f0d0bf8c79b		
200 files from ICL breath analysis cancer Study	ICL9			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=f7cd2fcc1de84b9b9331e3760e8c69dc			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e6c95625a5c6483c8d2974d42285e90f		
300 files from ICL breath analysis cancer Study	ICL10			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=20188d53112f4c41af5b589485c621bb			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8df5bcda67144abd98b2b193ecc917b2		
ICL breath analysis cancer Study1 (full, 387 files	ICL11			Gas	No	Breath		High res	cdf		https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=c5c026993f3f4c06b8cfa7140cf375cb			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=020bcc95de624decb3eae751ac631abd		
Testing with UCDavis public data (derivatized)																
5 files from ST000048	UCD1			Liquid	Yes	Human blood serum	ToF	Low res	cdf	5	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=c569b57876774235a35b2b0744e9cd1f			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=732d02cb7aa14993b35ff73ad3c736eb		
10 files from ST000048	UCD2			Liquid	Yes	Human blood serum	ToF	Low res	cdf	10	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=283998fc51bb47608f58d6357a9e7ab3			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=dfb9d2465cbc4fd58f14646984a1fdd7		
20 files from ST000048	UCD3			Liquid	Yes	Human blood serum	ToF	Low res	cdf	20	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=21173413abcd4a96be9b7fc0345acc11			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=6d7288e89797429e8cd04b4aecbff272		
30 files from ST000048	UCD4			Liquid	Yes	Human blood serum	ToF	Low res	cdf	30	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5ae544f07cc2446b93c90ca84b7637a1			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=450424df0d1d4e83947ed83265ae08b8		
40 files from ST000048	UCD5			Liquid	Yes	Human blood serum	ToF	Low res	cdf	40	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d939e6b892b34bad8f98ae1d81558fd2			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=98f17975d98e44a5a17e3705b14e1cc0		
50 files from ST000048	UCD6			Liquid	Yes	Human blood serum	ToF	Low res	cdf	50	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=aea1ad1b44004146b10defe069405d7f			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5b3f80d513e64732a1a9802321876be4		
75 files from ST000048	UCD7			Liquid	Yes	Human blood serum	ToF	Low res	cdf	75	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d42d2392cd194580ab09026b6ddc2b97			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=a0ee3734d3d447b3b8bdb42c4a7b59d1		
ST000048	UCD8			Liquid	Yes	Human blood serum	ToF	Low res	cdf	117	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=746fe76afffc4487849746ec86316d33			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=6432df9649674a01bdc1edf0eccfdb81		
ST000048_ ST000385	UCD9			Liquid	Yes	Human blood serum	ToF	Low res	cdf	502	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8900b97a89264826a2ffce2557db3005			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=07a7d191377444038f765cd40443208a		
ST000048_ ST000385_ ST000386	UCD10			Liquid	Yes	Human blood serum	ToF	Low res	cdf	875	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=76f1f5dc56db48529ed8aee3548e323a			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3584a2f4242f4a91b29d7ba2a38f577a		
ST000048_ ST000385_ ST000386_ ST000392	UCD11			Liquid	Yes	Human blood serum	ToF	Low res	cdf	1085	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=47e1b0932a37440d9eb9a9226887f60d			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3522fa101c894109ac94bf5c61e0d8ca		
ST000048_ ST000385_ ST000386_ ST000392_	UCD12			Liquid	Yes	Human blood serum	ToF	Low res	cdf	1358	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=92e7b71151c642308e9d436a56411dff			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3e6588e42af5458d909dc6f126bf4206		
ST000048_ ST000385_ ST000386_ ST000392_	UCD13			Liquid	Yes	Human blood serum +	ToF	Low res	cdf	1461	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=204250593d4142e29bfe0dbaf4548827			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=127bae62562b419f8191cfa6f95737e7		
ST000048_ ST000385_ ST000386_ ST000392_	UCD14			Liquid	Yes	Human blood serum +	ToF	Low res	cdf	1523	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=7590e6dd68b14e8cbc0bad92dd24dad4			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8fe2d687ce444fb1953f257c54cd7d47		
ST000048_ ST000385_ ST000386_ ST000392_	UCD15			Liquid	Yes	Human blood serum +	ToF	Low res	cdf	1823	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=84bbb09ea7e74c0487f9f5c06b10611b			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=84a55ae4a89b4e218c4e5a806a48e893		
ST000048_ ST000385_ ST000386_ ST000392_	UCD16			Liquid	Yes	Human blood serum +	ToF	Low res	cdf	1914	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=bcf6493e56f64b6c873c016d65911c25			https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=288f4e30c00d474aad55a74f1ca166e7		

Table S2. Comparison of deconvolution tools for GC-MS data used in metabolomics research

	Platform and interface	Spectrum generation	Settings	Scalability	Reproducibility	Accessibility	Reference
MSHub (introduced 2019)	Cross-platform (python). Command line (API) interface. Web-based interface available as GNPS workflow.	Spectral decomposition	None (automatically determined).	Limit is currently unknown, >10,000 files tested.	Analysis settings and processing steps automatically stored in the HDF5 file.	Easily accessible to users with any level of expertise if used as the GNPS workflow.	The present manuscript
MS-DIAL (introduced 2015)	Windows (.NET). Graphical user interface and can be partially automated.	Spectral deconvolution using chromatogram decomposition	Manual (<20 settable parameters).	Limited by the RAM memory consumption (typically <100 files).	Settings can be entered manually or can be imported from a method. Parameters are automatically saved into an output file.	Easily accessible, but requires expertise for parameters selection.	https://www.nature.com/articles/nmeth.3393
MZmine/ADAP (introduced 2010)	Cross-platform (Java). Graphical user interface and can be partially automated.	Spectral deconvolution using chromatogram decomposition	Manual (>30 settable parameters).	Limited by the RAM memory consumption (typically few hundreds of files).	Settings can be entered manually or imported from a batch file.	Easily accessible, but requires expertise for parameters selection.	https://pubs.acs.org/doi/abs/10.1021/acs.jproteome.7b00633 https://pubs.acs.org/doi/abs/10.1021/acs.analchem.6b02222 https://pubs.acs.org/doi/abs/10.1021/acs.analchem.9b01424
XCMS (introduced 2005)	Cross-platform (R) and text-based interface.	Peak grouping	Manual or semi-automated with the IPO package (>30 settable parameters).	Scalablility unknown, but can process several hundred of files on a cluster.	Commandline tool. Settings can be entered manually, or imported from script.	Accessible for bioinformaticians, but requires training for other users.	https://pubs.acs.org/doi/abs/10.1021/ac051437y
XCMS online (introduced 2012)	Web-based interface.	Peak grouping	Automatically provided for instrument types and can be modified by the user.	Scalablility unknown, but can process several hundred of files.	Settings are saved automatically.	Easily accessible to users with moderate level of expertise at the XCMS online website.	https://pubs.acs.org/doi/abs/10.1021/ac300698c
AMDIS (introduced 1998)	Windows. Graphical user interface, cannot be automated.	Spectral deconvolution using chromatogram decomposition	Manual (~10 settable parameters).	One file at a time.	Settings need to be entered manually each time.	Easily accessible, but requires expertise for parameters selection.	https://doi.org/10.1002/(SICI)1097-0231(19990228)13:4<279::AID-RCM478>3.0.CO;2-I

Table S3. Testing of deconvolution accuracy of MSHub and existing deconvolution tools		
Dataset description	Metabolomics Workbench_ST001154	LECO beer aging reference data
MassIVE ID	MSV000084039	MSV000084349
ftp link	ftp://massive.ucsd.edu/MSV000084039/	ftp://massive.ucsd.edu/MSV000084349/
Reference	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6571919/	NA
Sample introduction mode	Liquid	Gas, SPME
Derivatized?	Yes	No
Uberon/sample type	Mouse blood serum	Beer
Instrument	ToF	ToF
Data type	Low res	Low res
Data file format	cdf	cdf
Number of files	300	42
GNPS deconvolution link for MSHub	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9f305f5b6db94e71a5775c77b124d8f7	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f1ee1e1d480448b197a08d804414ddf6
GNPS library search link, MSHub deconvolution	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4f2de61ac9b949ddb1dd68b0d2fcd76f	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3544543809e04e30b7848432182d0dc3
GNPS library search link, MZmine2/ADAP deconvolution	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=021f48182d4c4f9aaa759bd40c5ea14d	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=37971f05c990478596a463a94d201361
GNPS library search link, MS-DIAL deconvolution	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16d202bd23784024ba85b381ab9aa16d	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=dd21a47e2089475ab48a71a5bf163f3e
GNPS library search link, MSHub deconvolution (public librari	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=116e7bfc25074900a3da3a1817b9e9f9	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e9c642b2ae224862b4d7875cefdf20a6
GNPS library search link, MSHub deconvolution (NIST and Wil	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a693ad96f9b24b39967e7bfd519ebe3f	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=69d756fa78054e169da184cca90efd0e

Table S4. Testing of deconvolution time of MSHub and existing deconvolution tools

Data subset	# of files	Job link
5 files from ST000048	5	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=c569b57876774235a35b2b0744e9cd1f
10 files from ST000048	10	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=283998fc51bb47608f58d6357a9e7ab3
20 files from ST000048	20	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=21173413abcd4a96be9b7fc0345acc11
30 files from ST000048	30	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5ae544f07cc2446b93c90ca84b7637a1
40 files from ST000048	40	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d939e6b892b34bad8f98ae1d81558fd2
50 files from ST000048	50	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=aea1ad1b44004146b10defe069405d7f
75 files from ST000048	75	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d42d2392cd194580ab09026b6ddc2b97
ST000048	117	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=746fe76afffc4487849746ec86316d33
ST000048_ ST000385	502	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8900b97a89264826a2ffce2557db3005
ST000048_ ST000385_ ST000386	875	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=76f1f5dc56db48529ed8aee3548e323a
ST000048_ ST000385_ ST000386_ ST00039	1085	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=47e1b0932a37440d9eb9a9226887f60d
ST000048_ ST000385_ ST000386_ ST00039	1358	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=92e7b71151c642308e9d436a56411dff
ST000048_ ST000385_ ST000386_ ST00039	1461	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=204250593d4142e29bfe0dbaf4548827
ST000048_ ST000385_ ST000386_ ST00039	1523	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=7590e6dd68b14e8cbc0bad92dd24dad4
ST000048_ ST000385_ ST000386_ ST00039	1823	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=84bbb09ea7e74c0487f9f5c06b10611b
ST000048_ ST000385_ ST000386_ ST00039	1914	https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=bcf6493e56f64b6c873c016d65911c25

Dataset	MassIVE ID	ftp link
Metabolomics Workbench_ST000048	MSV000084033	ftp://massive.ucsd.edu/MSV000084033/
Metabolomics Workbench_ST000385	MSV000085136	ftp://massive.ucsd.edu/MSV000085136/
Metabolomics Workbench_ST000386	MSV000084034	ftp://massive.ucsd.edu/MSV000084034/
Metabolomics Workbench_ST000392	MSV000084036	ftp://massive.ucsd.edu/MSV000084036/
Metabolomics Workbench_ST000396	MSV000084032	ftp://massive.ucsd.edu/MSV000084032/
Metabolomics Workbench_ST000065	MSV000084038	ftp://massive.ucsd.edu/MSV000084038/
Metabolomics Workbench_ST000061	MSV000084042	ftp://massive.ucsd.edu/MSV000084042/
Metabolomics Workbench_ST001154	MSV000084039	ftp://massive.ucsd.edu/MSV000084039/
Metabolomics Workbench_ST000910	MSV000084040	ftp://massive.ucsd.edu/MSV000084040/
Metabolomics Workbench_ST000063	MSV000084037	ftp://massive.ucsd.edu/MSV000084037/

				MSHub					MZmine						MSDial
Test	Samples	Start	Finish	Minutes	Test	Samples	Start	Finish	Minutes		Test	Samples	Start	Finish	Minutes
1	5	7:06	7:12 AM	6	1	5	1:55 PM	2:05 PM	10		1	5	11:56	11:57 AM	1
2	10	7:19	7:28	9	2	10	1:41 PM	1:51 PM	10		2	10	11:01	11:04	3
3	20	7:35	7:52	17	3	20	1:25 PM	1:38 PM	13		3	20	11:16	11:18	2
4	30	7:56	8:20	24	4	30	2:12 PM	2:25 PM	13		4	30	11:22	11:24	2
5	40	8:38	9:10	32	5	40	2:30 PM	2:46 PM	16		5	40	11:27	11:31	4
6	50	9:15	9:51	36	6	50	3:28 PM	3:47 PM	19		6	50	11:33	11:38	5
7	75	9:56	10:51	55	7	75	4:06 PM	4:37 PM	29		7	75	11:40	-	FAIL during Alignment
8	117	11:10	12:33	83	8	117	4:45 PM	5:29 PM	44		8	117	11:51	-	FAIL during Alignment
9	502	17:25	21:25	240	9	502	5:45 PM	-	FAIL						
10	875	6:23	15:37	554											
10	875	10:51	20:12	561											
11	1085	16:29	1:36	547											
12	1358	5:54	20:53	899											
13	1461	5:55	22:05	970											
14	1523	16:30	5:12	762											
15	1823	5:55	2:51	1256											
16	1914	5:55	3:44	1309											

Table S5. Spiked compounds in the dataset MSV000083658			
compound name	**SMILES**	**InChIKey**	**subclass**
cis-13,16-Docosadienoic acid methyl ester	CCCCCC=CCC=CCCCCCCCCCCCC(=O)OC	UPIRHFZFPVEDCF-UHFFFAOYSA-N	CHEMONTID:0000324
cis-4,7,10,13,16,19-Docosahexaenoic acid methyl ester	CCC=CCC=CCC=CCC=CCC=CCC=CCCC(=O)OC	VCDLWFYODNTQOT-UHFFFAOYSA-N	CHEMONTID:0000324
cis-11,14-Eicosadienoic acid methyl ester	CCCCCC=CCC=CCCCCCCCCCC(=O)OC	GWJCFAOQCNNFAM-UHFFFAOYSA-N	CHEMONTID:0000324
cis-5,8,11,14,17-Eicosapentaenoic acid methyl ester	CCC=CCC=CCC=CCC=CCC=CCCCC(=O)OC	QWDCYFDDFPWISL-UHFFFAOYSA-N	CHEMONTID:0000324
cis-8,11,14-Eicosatrienoic acid methyl ester	CCCCCC=CCC=CCC=CCCCCCCC(=O)OC	QHATYOWJCAQINT-UHFFFAOYSA-N	CHEMONTID:0000324
cis-11,14,17-Eicosatrienoic acid methyl ester	CCC=CCC=CCC=CCCCCCCCCCC(=O)OC	XQAVRBUXEPJVRC-UHFFFAOYSA-N	CHEMONTID:0000324
cis-11-Eicosenoic acid methyl ester	CCCCCCCCC=CCCCCCCCCCC(=O)OC	RBKMRGOHCLRTLZ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl cis-10-heptadecenoate	CCCCCCC=CCCCCCCCCC(=O)OC	JNSUZRHLHDQGPN-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl hexanoate	CCCCCC(=O)OC	NUKZAGXMHTUAFE-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl ��-linolenate	CCC=CCC=CCC=CCCCCCCCC(=O)OC	DVWSXZIHSUZZKJ-UHFFFAOYSA-N	CHEMONTID:0000504
Methyl arachidate	CCCCCCCCCCCCCCCCCCCC(=O)OC	QGBRLVONZXHAKJ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl arachidonate	CCCCCC=CCC=CCC=CCC=CCCCC(=O)OC	OFIDNKMQBYGNIW-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl behenate	CCCCCCCCCCCCCCCCCCCCCC(=O)OC	QSQLTHHMFHEFIY-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl butyrate	CCCC(=O)OC	UUIQMZJEGPQKFD-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl decanoate	CCCCCCCCCC(=O)OC	YRHYCMZPEVDGFQ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl dodecanoate	CCCCCCCCCCCC(=O)OC	UQDUPQYQJKYHQI-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl elaidate	CCCCCCCCC=CCCCCCCCC(=O)OC	QYDYPVFESGNLHU-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl erucate	CCCCCCCCC=CCCCCCCCCCCCC(=O)OC	ZYNDJIBBPLNPOW-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl heneicosanoate	CCCCCCCCCCCCCCCCCCCCC(=O)OC	AJRICDSAJQHDSD-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl heptadecanoate	CCCCCCCCCCCCCCCCC(=O)OC	HUEBIMLTDXKIPR-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl linoleate	CCCCCC=CCC=CCCCCCCCC(=O)OC	WTTJVINHCBCLGX-UHFFFAOYSA-N	CHEMONTID:0000504
Methyl linolelaidate	CCCCCC=CCC=CCCCCCCCC(=O)OC	WTTJVINHCBCLGX-UHFFFAOYSA-N	CHEMONTID:0000504
Methyl linolenate	CCC=CCC=CCC=CCCCCCCCC(=O)OC	DVWSXZIHSUZZKJ-UHFFFAOYSA-N	CHEMONTID:0000504
Methyl myristate	CCCCCCCCCCCCCC(=O)OC	ZAZKJZBWRNNLDS-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl myristoleate	CCCCC=CCCCCCCCC(=O)OC	RWIPSJUSVXDVPB-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl oleate	CCCCCCCCC=CCCCCCCCC(=O)OC	QYDYPVFESGNLHU-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl octanoate	CCCCCCCC(=O)OC	JGHZJRVDZXSNKQ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl palmitate	CCCCCCCCCCCCCCCC(=O)OC	FLIACVVOZYBSBS-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl palmitoleate	CCCCCCC=CCCCCCCCC(=O)OC	IZFGRAGOVZCUFB-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl pentadecanoate	CCCCCCCCCCCCCCC(=O)OC	XIUXKAZJZFLLDQ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl cis-10-pentadecenoate	CCCCC=CCCCCCCCCC(=O)OC	JEDIPLFNJDRCFD-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl stearate	CCCCCCCCCCCCCCCCCC(=O)OC	HPEUJPJOZXNMSJ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl tricosanoate	CCCCCCCCCCCCCCCCCCCCCCC(=O)OC	VORKGRIRMPBCCZ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl tetracosanoate	CCCCCCCCCCCCCCCCCCCCCCCC(=O)OC	XUDJZDNUVZHSKZ-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl tridecanoate	CCCCCCCCCCCCC(=O)OC	JNDDPBOKWCBQSM-UHFFFAOYSA-N	CHEMONTID:0000324

Methyl undecanoate	CCCCCCCCCCC(=O)OC	XPQPWPZFBULGKT-UHFFFAOYSA-N	CHEMONTID:0000324
Methyl cis-15-tetracosenoate	CCCCCCCCC=CCCCCCCCCCCCCCC(=O)OC	AINIZSBLAFHZCP-UHFFFAOYSA-N	CHEMONTID:0000324
n-Octane	CCCCCCCC	TVMXDCGIABBOFY-UHFFFAOYSA-N	CHEMONTID:0002500
n-Decane	CCCCCCCCCC	DIOQZVSQGTUSAI-UHFFFAOYSA-N	CHEMONTID:0002500
n-Dodecane	CCCCCCCCCCCC	SNRUBQQJIBEYMU-UHFFFAOYSA-N	CHEMONTID:0002500
n-Tetradecane	CCCCCCCCCCCCCC	BGHCVCJVXZWKCC-UHFFFAOYSA-N	CHEMONTID:0002500
n-Hexadecane	CCCCCCCCCCCCCCCC	DCAYPVUWAIABOU-UHFFFAOYSA-N	CHEMONTID:0002500
n-Octadecane	CCCCCCCCCCCCCCCCCC	RZJRJXONCZWCBN-UHFFFAOYSA-N	CHEMONTID:0002500
n-Eicosane	CCCCCCCCCCCCCCCCCCCC	CBFCDTFDPHXCNY-UHFFFAOYSA-N	CHEMONTID:0002500
n-Docosane	CCCCCCCCCCCCCCCCCCCCCC	HOWGUJZVBDQJKV-UHFFFAOYSA-N	CHEMONTID:0002500
n-Tetracosane	CCCCCCCCCCCCCCCCCCCCCCCC	POOSGDOYLQNASK-UHFFFAOYSA-N	CHEMONTID:0002500
n-Hexacosane	CCCCCCCCCCCCCCCCCCCCCCCCCC	HMSWAIKSFDFLKN-UHFFFAOYSA-N	CHEMONTID:0002500
n-Octacosane	CCCCCCCCCCCCCCCCCCCCCCCCCCCC	ZYURHZPYMFLWSH-UHFFFAOYSA-N	CHEMONTID:0002500
n-Triacontane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	JXTPJDDICSTXJX-UHFFFAOYSA-N	CHEMONTID:0002500
n-Dotriacontane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	QHMGJGNTMQDRQA-UHFFFAOYSA-N	CHEMONTID:0002500
n-Tetratriacontane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	GWVDBZWVFGFBCN-UHFFFAOYSA-N	CHEMONTID:0002500
n-Hexatriacontane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	YDLYQMBWCWFRAI-UHFFFAOYSA-N	CHEMONTID:0002500
n-Octatriacontane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	BVKCQBBZBGYNOP-UHFFFAOYSA-N	CHEMONTID:0002500
n-Tetracontane	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	KUPLEGDPSCCPJI-UHFFFAOYSA-N	CHEMONTID:0002500
Ribitol	C(C(C(C(CO)O)O)O)O	HEBKCHPVOIAQTA-UHFFFAOYSA-N	CHEMONTID:0000011
Adonitol	C(C(C(C(CO)O)O)O)O	HEBKCHPVOIAQTA-UHFFFAOYSA-N	CHEMONTID:0000011

Table S6. Curated annotation in the dataset MSV000084039						
DataBaseID	**CompoundName**	**InChIKey**	**MZ**	**KEGGID**	**PubChemID**	**subclass**
31632	xylulose NIST	LQXVFWRQNMEDEE-PYHARJCCSA-N	173	C00312	439205	CHEMONTID:0000011
169	xylose	SRBFZHDQGSBBOR-IOVATXLUSA-N	103	C00181	135191	CHEMONTID:0000011
5857	xylitol	HEBKCHPVOIAQTA-NGQZWQHPSA-N	217	C00379	6912	CHEMONTID:0000011
3	valine	KZSNJWFQEVHDMF-BYPYZUCNSA-N	144	C00183	6287	CHEMONTID:0000013
14815	uridine	DRTQHJPVMGBUCF-XVFCMESISA-N	217	C00299	6029	
33210	uric acid	LEHOTFFKMJEONL-UHFFFAOYSA-N	441	C00366	1175	CHEMONTID:0000245
145496	urea	XSQUKJJJFZCRTK-UHFFFAOYSA-N	189	C00086	1176	CHEMONTID:0000517
1664	uracil	ISAKRJDGNUQOIC-UHFFFAOYSA-N	241	C00106	1174	CHEMONTID:0000075
16	tyrosine	OUYCCCASQSFEME-QMMMGPOBSA-N	218	C00082	6057	CHEMONTID:0000013
14	tryptophan	QIVBCDIJIAJPQS-VIFPVBQESA-N	202	C00078	6305	CHEMONTID:0001290
97	trans-4-hydroxyproline	PMMYEEVYMWASQN-DMTCNVIQSA-N	140	C01157	5810	CHEMONTID:0000013
1692	thymine	RWQNBRDOKXIBIV-UHFFFAOYSA-N	255	C00178	1135	CHEMONTID:0000075
87703	thymidine	IQFYYKKMVGJFEH-XLPZGREQSA-N	170	C00214	5789	CHEMONTID:0002180
26	threonine	AYFVYJQAPQTCCC-GBXIJSLDSA-N	218	C00188	6288	CHEMONTID:0000013
172	threonic acid	JPIJQSOTBSSVTP-STHAYSLISA-N	292	C01620	5460407	CHEMONTID:0000011
411	taurine	XOAAWQZATWQOTB-UHFFFAOYSA-N	326	C00245	1123	CHEMONTID:0000270
12266	sulfuric acid	QAOWNCQODCNURD-UHFFFAOYSA-N	227	C00059	1118	CHEMONTID:0001077
173	sucrose	CZMRCDWAGMRECN-UGDNZRGBSA-N	271	C00089	5988	CHEMONTID:0000011
161	succinic acid	KDYFGRWQOYBRFD-UHFFFAOYSA-N	247	C00042	1110	CHEMONTID:0000346
13	stearic acid	QIQXTHQIDYTFRH-UHFFFAOYSA-N	117	C01530	5281	CHEMONTID:0000262
25	serine	MTCFGRXMJLQNBG-REOHCLBHSA-N	218	C00065	5951	CHEMONTID:0000013
11214	saccharic acid	DSLZVSRJTYRBFB-LLEIAEIESA-N	333	C00818	33037	CHEMONTID:0000011
1662	ribose	HMFHBZSHGGEWLO-SOOFDHNKSA-N	217	C00121	5779	CHEMONTID:0000011
1683	ribonic acid	QXKAIJAYHKCRRA-BXXZVTAOSA-N	292	C01685	5460677	CHEMONTID:0000011
7362	ribitol	HEBKCHPVOIAQTA-ZXFHETKHSA-N	217	C00474	827	CHEMONTID:0000011
3190	raffinose	MUPFEKGTMRGPLJ-ZQSKZDJDSA-N	361	C00492	439242	CHEMONTID:0000011
66	pyruvic acid	LCTONWCANYUPML-UHFFFAOYSA-N	174	C00022	1060	CHEMONTID:0001113
1688	pseudo uridine	PTJWIQPHWPFNBW-GBNDHIKLSA-N	217	C02067	15047	
171970	proline	ONIBWKKTOPOVIA-BYPYZUCNSA-N	142	C00148	145742	CHEMONTID:0000013

16544	pinitol	DSCFFEYYQKSRSV-FEPQRWDDSA-N	260	C03844	164619	CHEMONTID:0000129
33429	pimelic acid	WLJVNTCWHIRURA-UHFFFAOYSA-N	155	C02656	385	CHEMONTID:0000262
4	phosphate	NBIIXXVUZAFLBC-UHFFFAOYSA-N	314	C00009	1004	CHEMONTID:0001073
2005	phenylethylamine	BHHGXPLMPWCGHP-UHFFFAOYSA-N	174	C05332	1001	CHEMONTID:0000186
33	phenylalanine	COLNVLDHVKWLRT-QMMMGPOBSA-N	218	C00079	6140	CHEMONTID:0000013
31889	phenol	ISWSIDIOOBJBQZ-UHFFFAOYSA-N	151	C00146	996	CHEMONTID:0004647
50	pelargonic acid	FBUKVWPVBMHYJY-UHFFFAOYSA-N	117	C01601	8158	CHEMONTID:0000262
31356	pantothenic acid	GHOKWGTUZJEAQD-ZETCQYMHSA-N	291	C12276	6613	CHEMONTID:0000129
96	palmitoleic acid	SECPZKHBENQXJG-FPLPWBNLSA-N	311	C08362	445638	CHEMONTID:0000262
11	palmitic acid	IPCSVZSSVZVIGE-UHFFFAOYSA-N	313	C00249	985	CHEMONTID:0000262
10	oxoproline	ODHCTXKNWHHXJC-VKHMYHEASA-N	156	C01879	7405	CHEMONTID:0000013
4923	oxalic acid	MUBZPKHOEPUJKR-UHFFFAOYSA-N	190	C00209	971	CHEMONTID:0000346
1821	ornithine	AHLPHDHHMVZTML-BYPYZUCNSA-N	142	C00077	6262	CHEMONTID:0000013
43	oleic acid	ZQPPMHVWECSIRJ-KTKRTIGZSA-N	339	C00712	445639	CHEMONTID:0000262
997	octadecanol	GLDOVTGHNKAZLK-UHFFFAOYSA-N	327		8221	CHEMONTID:0001334
127	myristic acid	TUNFSRHWOTWDNC-UHFFFAOYSA-N	285	C06424	11005	CHEMONTID:0000262
1741	myo-inositol	CDAISMWEOUEBRE-UHFFFAOYSA-N	305	C00137	892	CHEMONTID:0000129
1678	methionine sulfoxide	QEFRNWWLZKMPFJ-YGVKFDHGSA-N	128	C02989	158980	CHEMONTID:0000013
45	methionine	FFEARJCKVFRZRR-BYPYZUCNSA-N	176	C00073	6137	CHEMONTID:0000013
70	mannose	WQZGKKKJIJFFOK-QTVWNMPRSA-N	205	C00159	18950	CHEMONTID:0000011
1979	maltose	GUBGYTABKSRVRQ-PICCSMPSSA-N	204	C00208	439186	CHEMONTID:0000011
1391	malic acid	BJEPYKJPYRNKOW-UHFFFAOYSA-N	233	C00711	525	CHEMONTID:0001713
85112	lyxose	SRBFZHDQGSBBOR-AGQMPKSLSA-N	217	C00476	439240	CHEMONTID:0000011
233	lyxitol	HEBKCHPVOIAQTA-IMJSIDKUSA-N	217	C00532	439255	CHEMONTID:0000011
12	lysine	KDXKERNSBIXSRK-YFKPBYRVSA-N	156	C00047	5962	CHEMONTID:0000013
165	linoleic acid	OYHQOLUKZRVURQ-HZJYTTRNSA-N	150	C01595	5280450	CHEMONTID:0000504
9	leucine	ROHFNLRQFUQHCH-YFKPBYRVSA-N	158	C00123	6106	CHEMONTID:0000013
49	lauric acid	POULHZVOKOAJMA-UHFFFAOYSA-N	117	C02679	3893	CHEMONTID:0000262
80	lactic acid	JVTAAEKCZFNVCJ-UHFFFAOYSA-N	191	C01432	612	CHEMONTID:0001359
415	lactamide	SXQFCVDSOLSHOQ-UHFFFAOYSA-N	232		94220	CHEMONTID:0000129
1679	isothreonic acid	JPIJQSOTBSSVTP-GBXIJSLDSA-N	292	C00639	151152	CHEMONTID:0000011

15	isoleucine	AGPKZVBTJJNPAG-WHFBIAKZSA-N	158	C00407	6306	CHEMONTID:0000013
100866	isohexonic acid	RGHNJXZEOKUKBD-UHFFFAOYSA-N	333		604	CHEMONTID:0000299
32122	isocitric acid	ODBLHEXUDAPZAU-ZAFYKAAXSA-N	245	C00451	5318532	CHEMONTID:0001986
84524	inosine	UGQMRVRMYYASKQ-KQYNXXCUSA-N	230	C00294	6021	
31543	indoxyl sulfate	BXFFHSIDQOFMLE-UHFFFAOYSA-N	277		10258	CHEMONTID:0004258
112556	indole-3-propionic acid	GOLXRNDWAUTYKT-UHFFFAOYSA-N	202		3744	CHEMONTID:0001290
724	indole-3-lactate	XGILAAMKEQUXLS-UHFFFAOYSA-N	202	C02043	92904	CHEMONTID:0001290
69	indole-3-acetate	SEOVTRFCIGRIMH-UHFFFAOYSA-N	202	C00954	802	CHEMONTID:0001290
139	hydroxylamine	AVXURJPOCDRRFD-UHFFFAOYSA-N	146	C00192	787	
130396	hydroxycarbamate NIST	DRAJWRKLRBNJRQ-UHFFFAOYSA-M	278		16639161	
114270	hydrocinnamic acid	XMIIGOLPHOKFCH-UHFFFAOYSA-N	104	C05629	107	
33211	hippuric acid	QIAFMBKCNZACKA-UHFFFAOYSA-N	206	C01586	464	CHEMONTID:0000176
4554	hexuronic acid	IAJILQKETJEXLJ-UHFFFAOYSA-N	160		152867	CHEMONTID:0000011
20287	hexose	WQZGKKKJIJFFOK-UHFFFAOYSA-N	319	C00738	206	CHEMONTID:0000011
255	hexitol	FBPFZTCFMRRESA-UHFFFAOYSA-N	217	C00392	453	CHEMONTID:0000011
120765	hexadecane	DCAYPVUWAIABOU-UHFFFAOYSA-N	85		11006	CHEMONTID:0002500
727	heptadecanoic acid	KEMQGTRYUADPNZ-UHFFFAOYSA-N	117		10465	CHEMONTID:0000262
1971	glycolic acid	AEMRFAOFKBGASW-UHFFFAOYSA-N	177	C00160	757	CHEMONTID:0001359
6	glycine	DHMQDGOQFOQNFH-UHFFFAOYSA-N	248	C00037	750	CHEMONTID:0000013
1687	glycerol-alpha-phosphate	AWUCVROLDVIAJX-UHFFFAOYSA-N	357	C03189	754	CHEMONTID:0002214
1693	glycerol-3-galactoside	NHJUPBDCSOGIKX-NTXXKDEISA-N	204	C05401	16048618	CHEMONTID:0000637
30	glycerol	PEDCQBHIVMGVHV-UHFFFAOYSA-N	205	C00116	753	CHEMONTID:0000011
48	glyceric acid	RBNPOMFGQQGHHO-UWTATZPHSA-N	189	C00258	439194	CHEMONTID:0000011
171972	glutamine	ZDXPYRJPNDTMRX-VKHMYHEASA-N	156	C00064	5961	CHEMONTID:0000013
16576	glutamic acid	WHUUTDBJXJRKMK-VKHMYHEASA-N	246	C00025	33032	CHEMONTID:0000013
3167	glucose-1-phosphate	HXXFSFRBOHSIMQ-VFUOTHLCSA-N	217	C00103	65533	CHEMONTID:0000011
71	glucose	WQZGKKKJIJFFOK-VFUOTHLCSA-N	217	C00221	64689	CHEMONTID:0000011
14755	fumaric acid	VZCYOOQTPOCHFL-OWOJBTEDSA-N	245	C00122	444972	CHEMONTID:0000346
3009	fucose	SHZGCJCMOBCMKK-FPRJBGLDSA-N	160	C02095	439650	CHEMONTID:0000011
21	fructose	RFSUNEUAIZKAJO-ARQDHWQXSA-N	307	C02336	439709	CHEMONTID:0000011
3212	ethanolamine	HZAXFHJVJLSVMW-UHFFFAOYSA-N	174	C00189	700	CHEMONTID:0002449

92	erythritol	UNXHWFMMPAWVPI-ZXZARUISSA-N	217	C00503	222285	CHEMONTID:0000011
12092	dodecanol	LQZZUXJYWNFBMV-UHFFFAOYSA-N	243	C02277	8193	CHEMONTID:0001334
126903	D-erythro-sphingosine	WWUZIQQURGPMPG-KRWOKUGFSA-N	204	C00319	5280335	CHEMONTID:0002449
94	cystine	LEVWYRKDKASIDU-UHFFFAOYSA-N	218	C01420	595	CHEMONTID:0000013
65	cysteine	XUJNEKJLAYXESH-REOHCLBHSA-N	220	C00097	5862	CHEMONTID:0000013
31	creatinine	DDRJAANPRJIHGJ-UHFFFAOYSA-N	115	C00791	588	CHEMONTID:0000013
2670	conduritol-beta-epoxide	ZHMWOVGZCINIHW-SPHYCDKFSA-N	318		9989541	
1712	citrulline	RHGKLRLOHDJJDR-BYPYZUCNSA-N	157	C00327	9750	CHEMONTID:0000013
288	citric acid	KRKNYBCHXYNGOX-UHFFFAOYSA-N	273	C00158	311	CHEMONTID:0001986
31654	catechol	YCIMNLLNPGFGHC-UHFFFAOYSA-N	254	C00090	289	CHEMONTID:0001286
6866	caprylic acid	WWZKQHOCKIZLMA-UHFFFAOYSA-N	201	C06423	379	CHEMONTID:0000262
50422	capric acid	GHVNFZFCNZKVNT-UHFFFAOYSA-N	229	C01571	2969	CHEMONTID:0000262
148	beta-alanine	UCMIRNVEIXFBKS-UHFFFAOYSA-N	248	C00099	239	CHEMONTID:0000013
36	benzoic acid	WPYMKLBDIGXBTP-UHFFFAOYSA-N	179	C00180	243	CHEMONTID:0000176
79	aspartic acid	CKLJMWTZIZZHCS-REOHCLBHSA-N	232	C00049	5960	CHEMONTID:0000013
146	asparagine	DCXYFEDJOCDNAF-REOHCLBHSA-N	188	C00152	6267	CHEMONTID:0000013
291	arachidic acid	VKOBVWXKNCXXDE-UHFFFAOYSA-N	117	C06425	10467	CHEMONTID:0000262
413	aminomalonate	JINBYESILADKFW-UHFFFAOYSA-N	218	C00872	100714	CHEMONTID:0000013
294	alpha-ketoglutarate	KPGXRSRHYNQIFN-UHFFFAOYSA-N	198	C00026	51	CHEMONTID:0001115
125502	alpha-aminoadipic acid	OYIFNHCXNCRBQI-BYPYZUCNSA-N	260	C00956	92136	CHEMONTID:0000013
33396	allantoic acid	NUCLJNSWZCHRKL-UHFFFAOYSA-N	259	C00499	203	CHEMONTID:0000013
34178	alanine	QNAYBMKLOCPYGJ-REOHCLBHSA-N	116	C00041	5950	CHEMONTID:0000013
290	adenosine	OIRDTQYFTABQOQ-KQYNXXCUSA-N	236	C00212	60961	
92567	6-deoxyglucitol	SKCKOFZKJLZSFA-JGWLITMVSA-N	117		151266	CHEMONTID:0000011
284	5-methoxytryptamine	JTEJPPKMYBDEMY-UHFFFAOYSA-N	174	C05659	1833	CHEMONTID:0000183
85123	4-hydroxybutyric acid	SJZRECIVHVDYJC-UHFFFAOYSA-N	233	C00989	10413	CHEMONTID:0000262
1977	3-hydroxypropionic acid	ALRHLSYJTWAHJZ-UHFFFAOYSA-N	177	C01013	68152	CHEMONTID:0001713
46	3-hydroxybutyric acid	WHBMMWSBFZVSSR-GSVOUGTGSA-N	191	C01089	92135	CHEMONTID:0001713
12265	3-aminoisobutyric acid	QCHPKSFMDHPSNR-UHFFFAOYSA-N	248	C05145	64956	CHEMONTID:0000013
102714	3,6-anhydro-D-galactose	WZYRMLAWNVOIEX-BGPJRJDNSA-N	231	C06474	16069996	
49497	3-(3-hydroxyphenyl)propionic acid	QVWAEZJXDYOKEH-UHFFFAOYSA-N	192	C11457	91	

208	2-ketoisocaproic acid	BKAJNAXTPSGJCU-UHFFFAOYSA-N	89	C00233	70	CHEMONTID:0001416
2000	2-hydroxyglutaric acid	HWXBTNAVRSUOJR-UHFFFAOYSA-N	247	C02630	43	CHEMONTID:0000266
40	2-hydroxybutanoic acid	AFENDNXGAFYKQO-VKHMYHEASA-N	131	C05984	440864	CHEMONTID:0001359
1208	2-deoxytetronic acid	DZAIOXUZHHTJKN-UHFFFAOYSA-N	189		150929	CHEMONTID:0001713
160842	2-aminobutyric acid	QWCKQJZIFLGMSD-UHFFFAOYSA-N	130	C02721	6657	CHEMONTID:0000013
125664	1,3,5-trimethylcyanuric acid	AHWDQDMGFXRVFB-UHFFFAOYSA-N	171			CHEMONTID:0001920

Additional annotations: FAMEs			InChiKey			
BinID	CompoundName	Name (CAS)		RI		
14391	z C08 FAME internal standard	C8:0 Methyl caprylate (111-11-5), 4%	JGH	262320		
14356	z C09 FAME internal standard	Methyl nonanoate (1731-84-6)	IJXH	323120		
14348	z C10 FAME internal standard	C10:0 Methyl decanoate (110-42-9), 4%	YRH	381020		
15538	z C12 FAME internal standard	C12:0 Methyl dodecanoate (111-82-0), 4%	UQD	487220		
14330	z C14 FAME internal standard	C14:0 Methyl myristate (124-10-7), 4%	ZAZ	582620		
14328	z C16 FAME internal standard	C16:0 Methyl palmitate (112-39-0), 6%	FLIA	668720		
14344	z C18 FAME internal standard	C18:0 Methyl stearate (112-61-8), 4%	HPE	747420		
14338	z C20 FAME internal standard	C20:0 Methyl arachidate (1120-28-1), 4%	QGE	819620		
14350	z C22 FAME internal standard	C22:0 Methyl behenate (929-77-1), 4%	QSC	886620		
14373	z C24 FAME internal standard	C24:0 Methyl lignocerate (2442-49-1), 4%	XUD	948820		
14367	z C26 FAME internal standard	C24:0 Methyl Hexacosanoate	VHU	1006900		
14378	z C28 FAME internal standard	C24:0 Methyl Octacosanoate	ZKH	1061700		
14441	z C30 FAME internal standard	C24:0 Methyl Triacontanoate	BIRU	1113100		

Table S7. Curated annotation in the dataset MSV000084349							
Nist ID	**Name**	**Formula**	**CAS**	**R.T. (s)**	**XIC(Masses)**	**InChIKey**	**subclass**
34064	Dimethyl sulfide	C2H6S		115.704	62.018		
35900	Furan	C4H4O		130.244	68.026	YLQBMQCUIZJEEH-UHFFFAOYSA-N	
6707	Isobutanal	C4H8O		135.469	72.057	AMIMRNSIRUDHCM-UHFFFAOYSA-N	CHEMONTID:0001831
8607	Acetone	C3H6O		137.381	58.041	CSCPPACGZOOCGX-UHFFFAOYSA-N	CHEMONTID:0001831
1688	Ethyl formate	C3H6O2		142.582	74.036		
23642	2-Propenal	C3H4O		153.217	56.026	HGINCPLSRVDWNT-UHFFFAOYSA-N	CHEMONTID:0001831
4858	Tetrahydrofuran	C4H8O		158.808	72.057	WYURNTSHIVDZCO-UHFFFAOYSA-N	
55654	3-Methylfuran	C5H6O		166.158	53.039	KJRRQXYWFQKJIP-UHFFFAOYSA-N	
3818	Methacrolein	C4H6O		172.695	70.041	STNJBCKSHOAVAJ-UHFFFAOYSA-N	CHEMONTID:0001831
8863	Ethyl acetate	C4H8O2		177.359	61.028		
55653	2-Methylfuran	C5H6O		183.673	53.039	VQKFNUFAXTZWDK-UHFFFAOYSA-N	
9850	Butanone	C4H8O		186.91	72.057	ZWEHNKRNPOVVGH-UHFFFAOYSA-N	CHEMONTID:0001831
632	Methyl propanoate	C4H8O2		192.749	57.033		
2323	2-Methylbutanal	C5H10O		196.761	58.041	BYGQBDHUGHBGMD-UHFFFAOYSA-N	CHEMONTID:0001831
16524	3-Methylbutanal	C5H10O		201.574	58.041	YGHRJJRRZDOVPD-UHFFFAOYSA-N	CHEMONTID:0001831
9383	Methyl isobutanoate	C5H10O2		207.95	102.068		
6785	3-Methyl-2-butanone	C5H10O		214.162	86.073	SYBYTAAJFKOIEJ-UHFFFAOYSA-N	CHEMONTID:0001831
76298	2,5-Dimethylfuran	C6H8O		243.478	96.057	GSNUFIFRDBKVIE-UHFFFAOYSA-N	
616	Ethyl propanoate	C5H10O2		249.59	57.033		
9330	Ethyl isobutanoate	C6H12O2		259.629	89.06		
8875	n-Propyl acetate	C5H10O2		271.867	61.028		
53132	(E,E)-1,3,5-Heptatriene	C7H10		271.98	94.078	USKZHEQYENVSMH-YDFGWWAZSA-N	CHEMONTID:0002838
10763	2-Pentanone	C5H10O		272.792	86.073	XNLICIUVMPYHGG-UHFFFAOYSA-N	CHEMONTID:0001831
83720	Isobutanal diethyl acetal	C8H18O2		273.913	47.013		
10770	2,3-Butanedione	C4H6O2		281.881	86.036	QSJXEFYPDANLFS-UHFFFAOYSA-N	CHEMONTID:0001831
25801	2-Methyl-3-pentanone	C6H12O		297.994	100.088	HYTRYEXINDDXJK-UHFFFAOYSA-N	CHEMONTID:0001831
8623	Methyl isobutyl ketone	C6H12O		314.495	100.088		
28046	Methyl 2-methylbutanoic acid	C6H12O2		320.071	88.052		
72826	alpha-Pinene	C10H16		325.871	136.125	GRWFGVWFFZKLTI-UHFFFAOYSA-N	CHEMONTID:0001549
6013	3-Methyl-2-pentanone	C6H12O		327.084	100.088	UIHCLUNTQKBZGK-UHFFFAOYSA-N	CHEMONTID:0001831
8100	Isobutyl acetate	C6H12O2		327.884	56.062		
48778	Methyl 3-methylbutanoic acid	C6H12O2		336.36	74.036		

72814	alpha-Thujene	C10H16		338.622	136.125	KQAZVFVOEIRWHN-UHFFFAOYSA-N	CHEMONTID:0001549
75229	2-Ethyl-5-methylfuran	C7H10O		356.362	110.073	NBXLPPVOZWYADY-UHFFFAOYSA-N	
65916	Toluene	C7H8		363.862	92.062	YXFVVABEGXRONW-UHFFFAOYSA-N	CHEMONTID:0001091
41373	Ethyl butanoate	C6H12O2		365.312	88.052		
1686	1-Propanol	C3H8O		371.25	60.057	BDERNNFJNOPAEC-UHFFFAOYSA-N	CHEMONTID:0000129
24922	Propyl propanoate	C6H12O2		378.926	75.044		
10976	Methyl thioacetate	C3H6OS		383.426	90.013		
24934	Ethyl 2-methylbutanoate	C7H14O2		390.589	102.068		
9323	Propyl isobutanoate	C7H14O2		392.765	89.06		
6011	2,3-Pentanedione	C5H8O2		411.828	100.052	TZMFJUDUGYTVRY-UHFFFAOYSA-N	CHEMONTID:0001831
62925	Ethyl 3-methylbutanoate	C7H14O2		417.842	88.052		
73825	Dimethyl disulfide	C2H6S2		419.217	93.991		
73726	2-Vinylfuran	C6H6O		430.48	94.041	QQBUHYQVKJQAOB-UHFFFAOYSA-N	
83623	Isopentanal diethyl acetal	C9H20O2		433.63	103.075		
24900	Isobutyl propanoate	C7H14O2		437.868	57.033		
77201	2-Methylthiophene	C5H6S		450.119	98.018	XQQBUAPQHNYYRS-UHFFFAOYSA-N	
41541	Isobutyl isobutanoate	C8H16O2		454.794	89.06		
72372	beta-Pinene	C10H16		455.832	136.125	WTARULDDTDQWMU-UHFFFAOYSA-N	CHEMONTID:0001549
53146	1,3-(Z),5-(Z)-Octatriene	C8H12		469.133	108.093		
53571	1,3-(E),5-(Z)-Octatriene	C8H12		472.608	108.093		
137440	Pyranoid linalool	C10H18O		473.933	139.112		
6347	Isobutanol	C4H10O		473.896	74.073	ZXEKIIBDNHEJCQ-UHFFFAOYSA-N	CHEMONTID:0000129
72809	Sabinene	C10H16		483.534	136.125	NDVASEGYNIMXJL-UHFFFAOYSA-N	CHEMONTID:0001549
77202	3-Methylthiophene	C5H6S		500.898	98.018	QENGPZGAWFQWCZ-UHFFFAOYSA-N	
9248	1-Butanol, 2-methyl-, acetate	C7H14O2		509.999	101.06		
9247	1-Butanol, 3-methyl-, acetate	C7H14O2		513.236	87.044		
66555	1,4-Xylene	C8H10		522.162	106.078	URLKBWYHVLBVBO-UHFFFAOYSA-N	CHEMONTID:0004208
9378	Methyl thioisobutanoate	C5H10OS		525.737	118.045		
54349	2-Butylfuran	C8H12O		527.712	124.088	NWZIYQNUCXUJJJ-UHFFFAOYSA-N	
56990	4-Methyl-3-pentene-2-one	C6H10O		527.337	98.073	SHOJXDKTYKFBRD-UHFFFAOYSA-N	CHEMONTID:0001831
66554	1,3-Xylene	C8H10		533.663	106.078	IVSZLXZYQVIEFR-UHFFFAOYSA-N	CHEMONTID:0004208
719	Ethyl pentanoate	C7H14O2		534.938	88.052		
26352	Propyl 2-methylbutanoate	C8H16O2		540.288	116.083		
73174	2-Carene	C10H16		542.689	136.125	IBVJWOMJGCHRRW-UHFFFAOYSA-N	CHEMONTID:0001549
24903	Butyl propanoate	C7H14O2		545.902	75.044		

8608	5-Methyl-2-hexanone	C7H14O		546.314	114.104	FFWSICBKRCICMR-UHFFFAOYSA-N	CHEMONTID:0001831
73177	4-Carene	C10H16		550.677	136.125	LGNSZMLHOYDATP-UHFFFAOYSA-N	CHEMONTID:0001549
48779	Methyl isohexanoate	C7H14O2		550.589	74.036		
9620	Butyl isobutanoate	C8H16O2		556.49	89.06		
27879	2,5-Dimethyl-3-hexanone	C8H16O		558.165	128.12	TUIWMHDSXJWXOH-UHFFFAOYSA-N	CHEMONTID:0001831
1764	1-Butanol	C4H10O		562.7	56.062	LRHPLDYGYMQRHN-UHFFFAOYSA-N	CHEMONTID:0000129
72819	alpha-Phellandrene	C10H16		570.816	136.125	OGLDWXZKYODSOB-UHFFFAOYSA-N	CHEMONTID:0001549
60194	Propyl 3-methylbutanoate	C8H16O2		573.916	60.021		
4103	beta-Myrcene	C10H16		588.305	136.125	UAHWPYUMFXYFJY-UHFFFAOYSA-N	CHEMONTID:0001549
112331	alpha-Terpinene	C10H16		600.831	136.125	YHQGMYUVUMAZJR-UHFFFAOYSA-N	CHEMONTID:0001549
9239	n-Pentyl acetate	C7H14O2		606.482	61.028		
27927	Isobutyl 2-methylbutanoate	C9H18O2		608.769	130.099		
66558	1,2-Xylene	C8H10		613.994	106.078	CTQNGGLPUBDAKN-UHFFFAOYSA-N	CHEMONTID:0004208
52928	Pyridine	C5H5N		613.357	52.031	JUJWROOIHBZHMG-UHFFFAOYSA-N	
8609	2-Heptanone	C7H14O		619.558	114.104	CATSNJVOTSVZJV-UHFFFAOYSA-N	CHEMONTID:0001831
93779	2,3-Dehydroeucalyptol	C10H16O		621.42	109.065		
40010	Heptanal	C7H14O		624.708	70.078	FXHGMKSSBGDXIY-UHFFFAOYSA-N	CHEMONTID:0001831
24888	2-Methylbutyl propanoate	C8H16O2		631.871	114.104		
27102	3-Methylbutyl propanoate	C8H16O2		634.54	101.06		
36348	Limonene	C10H16		634.033	136.125	XMGQYMWWDOXHJM-UHFFFAOYSA-N	CHEMONTID:0001549
62832	Ethyl isohexanoate	C8H16O2		637.234	99.08		
59772	Isobutyl isovalerate	C9H18O2		637.721	60.021		
9573	3-Methylbutyl isobutanoate	C9H18O2		643.234	115.075		
41671	2-Methylbutyl isobutanoate	C9H18O2		646.822	129.091		
72661	beta-Phellandrene	C10H16		649.56	136.125	LFJQCDVYDGGFCH-UHFFFAOYSA-N	CHEMONTID:0001549
35299	(2R,5S)-2-Methyl-5-isopropenyl-2-vinylTHF	C10H16O		666.324	68.062		
25402	2-Methyl-1-Butanol	C5H12O		669.774	59.049	QPRQEDXDYOZYLA-UHFFFAOYSA-N	CHEMONTID:0000129
21983	3-Methyl-1-Butanol	C5H12O		672.949	46.041	PHTQWCKDNZKARW-UHFFFAOYSA-N	CHEMONTID:0000129
27851	Methyl thioisopentanoate	C6H12OS		692.926	132.06		
54840	2-Pentylfuran	C9H14O		708.015	138.104	YVBAUDVGOFCUSG-UHFFFAOYSA-N	
28366	Butyl 2-methylbutanoate	C9H18O2		708.84	85.065		
62769	Ethyl hexanoate	C8H16O2		715.052	88.052		
56662	Ethyl tiglate	C7H12O2		718.391	113.06		
72839	gamma-Terpinene	C10H16		722.591	136.125	YKFLAYDHMOASIY-UHFFFAOYSA-N	CHEMONTID:0004622
35300	(2R,5R)-2-Methyl-5-isopropenyl-2-vinylTHF	C10H16O		724.916	68.062		

24821	Pentyl propanoate	C8H16O2		725.503	75.044		
9622	Pentyl isobutanoate	C9H18O2		731.692	89.06		
59849	Thiazole	C3H3NS		738.679	84.998	FZWLAAWBMGSTSO-UHFFFAOYSA-N	CHEMONTID:0000095
59773	Butyl 3-methylbutanoate	C9H18O2		740.604	60.021		
72375	(3E)-3,7-Dimethyl-1,3,6-octatriene	C10H16		743.042	136.125	IHPKGUQCSIINRJ-CSKARUKUSA-N	CHEMONTID:0001549
5084	1-Pentanol	C5H12O		745.33	70.078	AMQJEAYHLZJPGS-UHFFFAOYSA-N	CHEMONTID:0000129
8527	3-Octanone	C8H16O		747.518	128.12	RHLVCLIPMVJYKS-UHFFFAOYSA-N	CHEMONTID:0001831
84492	Styrene	C8H8		750.93	104.062	PPBRXRYQALVLMV-UHFFFAOYSA-N	CHEMONTID:0000037
73751	Methylpyrazine	C5H6N2		763.394	94.053	CAWHJQAVHZEVTJ-UHFFFAOYSA-N	CHEMONTID:0000067
9895	2-Methyltetrahydrofuran-3-one	C5H8O2		766.094	100.052	FCWYQRVIQDNGBI-UHFFFAOYSA-N	CHEMONTID:0001982
109426	p-Cymene	C10H14		768.857	134.109	HFPZCAJZSCWRBC-UHFFFAOYSA-N	CHEMONTID:0001549
41664	Isoamyl butanoate	C9H18O2		769.145	71.049		
8084	n-Hexyl acetate	C8H16O2		782.933	61.028		
73151	alpha-Terpinolene	C10H16		787.721	136.125	MOYAFQVGZZPNRA-UHFFFAOYSA-N	CHEMONTID:0001549
40441	3-Methylbutyl 2-methylbutanoate	C10H20O2		789.358	129.091		
40466	2-Methylbutyl 2-methylbutanoate	C10H20O2		793.334	143.107		
8621	2-Octanone	C8H16O		801.622	128.12	ZPVFWPFBNIEHGJ-UHFFFAOYSA-N	CHEMONTID:0001831
40655	3-Methylbutyl 3-methylbutanoate	C10H20O2		821.174	129.091		
40657	2-Methylbutyl 3-methylbutanoate	C10H20O2		822.761	143.107		
55723	2,2,6-Trimethylcyclohexanone	C9H16O		844.538	140.12	ZPVOLGVTNLDBFI-UHFFFAOYSA-N	CHEMONTID:0001831
5489	2,5-Dimethylpyrazine	C6H8N2		857.414	108.068	LCZUOKDVTBMCMX-UHFFFAOYSA-N	CHEMONTID:0000067
11466	Propyl hexanoate	C9H18O2		860.489	99.08		
92497	2,6-Dimethylpyrazine	C6H8N2		867.928	108.068	HJFZAYHYIWGLNL-UHFFFAOYSA-N	CHEMONTID:0000067
83674	Pentyl 2-methylbutanoate	C10H20O2		875.928	103.075		
88496	1,2,3-Trimethylbenzene	C9H12		881.041	120.093	FYGHSUNMUKGBRK-UHFFFAOYSA-N	
62796	Ethyl heptanoate	C9H18O2		886.666	88.052		
6655	Sulcatone	C8H14O		892.18	126.104	UHEPJGULSIKKTP-UHFFFAOYSA-N	CHEMONTID:0001831
26732	Hexyl propanoate	C9H18O2		895.392	75.044		
10956	Hexyl isobutanoate	C10H20O2		900.505	89.06		
18170	Ethyl lactate	C5H10O3		904.118	45.033		
10666	Pentyl 3-methylbutanoate	C10H20O2		907.356	103.075		
137583	Rose oxide	C10H18O		910.331	139.112		
55542	2-Cyclopenten-1-one	C5H6O		915.306	82.041	BZKFMUIJRXWWQK-UHFFFAOYSA-N	CHEMONTID:0001831
79058	Isobutyl hexanoate	C10H20O2		917.894	99.08		
23911	1-Hexanol	C6H14O		918.907	56.062	ZSIAUFGUXNUGDI-UHFFFAOYSA-N	CHEMONTID:0001334

137588	trans-Rose oxide	C10H18O		932.72	139.112		
120606	Dimethyl trisulfide	C2H6S3		952.584	125.963		
9259	n-Heptyl acetate	C9H18O2		952.671	61.028		
8622	2-Nonanone	C9H18O		976.123	142.135	VKCYHJWLYTUGCC-UHFFFAOYSA-N	CHEMONTID:0001831
53225	(E,E)-1,3,5-Undecatriene	C11H18		983.174	150.14	JQQDKNVOSLONRS-JEGFTUTRSA-N	CHEMONTID:0002838
25252	Nonanal	C9H18O		983.437	98.109	GYHFUZHODSMOHU-UHFFFAOYSA-N	CHEMONTID:0001831
32467	3-Octanol	C8H18O		986.649	59.049	NMRPBPVERJPACX-UHFFFAOYSA-N	CHEMONTID:0001334
5576	Trimethylpyrazine	C7H10N2		991.037	122.084	IAEGWXHKWJGQAZ-UHFFFAOYSA-N	CHEMONTID:0000067
53227	(3E,5Z)-1,3,5-Undecatriene	C11H18		1001.905	150.14	JQQDKNVOSLONRS-STRRHFTISA-N	CHEMONTID:0002838
53222	(3Z,5E)-1,3,5-Undecatriene	C11H18		1009.355	150.14	JQQDKNVOSLONRS-BABZSUFTSA-N	CHEMONTID:0002838
38190	Perillene	C10H14O		1026.33	150.104	XNGKCOFXDHYSGR-UHFFFAOYSA-N	
109433	1,2,3,5-Tetramethylbenzene	C10H14		1027.08	134.109	BFIMMTCNYPIMRN-UHFFFAOYSA-N	
18456	2-Octanol	C8H18O		1028.305	45.033	SJWFXCIHNDVPSH-UHFFFAOYSA-N	CHEMONTID:0001334
83671	Hexyl 2-methylbutanoate	C11H22O2		1038.105	103.075		
109435	1,2,4,5-Tetramethylbenzene	C10H14		1042.83	134.109	SQNZJJAZBFDUTD-UHFFFAOYSA-N	
106316	p-Cymenene	C10H12		1049.943	132.093	MMSLOZQEMPDGPI-UHFFFAOYSA-N	CHEMONTID:0000045
63199	Ethyl octanoate	C10H20O2		1057.993	88.052		
27623	Heptyl propanoate	C10H20O2		1060.968	75.044		
10958	Heptyl isobutanoate	C11H22O2		1063.155	89.06		
60199	Hexyl 3-methylbutanoate	C11H22O2		1067.155	103.075		
67180	Cosmene	C10H14		1070.868	134.109	HPZWSJQQCJZBBG-LQPGMRSMSA-N	CHEMONTID:0002838
26298	1-Octen-3-ol	C8H16O		1076.218	72.057	VSMOENVRRABVKN-UHFFFAOYSA-N	CHEMONTID:0001334
23984	Isobutyl heptanoate	C11H22O2		1077.418	113.096		
40406	1-Heptanol	C7H16O		1081.568	70.078	BBMCTIGTTCKYKF-UHFFFAOYSA-N	CHEMONTID:0001334
40217	3-Methylbutyl hexanoate	C11H22O2		1087.543	143.107		
10837	2-Nonyl acetate	C11H22O2		1092.718	87.044		
7393	Acetic acid	C2H4O2		1097.068	60.021		
76294	Furfural	C5H4O2		1099.93	96.021	HYBBIBNJHNGZAN-UHFFFAOYSA-N	CHEMONTID:0001831
20404	Tetramethylpyrazine	C8H12N2		1101.255	136.099	FINHMKGKINIASC-UHFFFAOYSA-N	CHEMONTID:0000067
36123	Nerol oxide	C10H16O		1101.48	152.12		
40372	Isoamyl tiglate	C10H18O2		1102.18	101.06		
9257	n-Octyl acetate	C10H20O2		1113.543	61.028		
109428	1,2,3,4-Tetramethylbenzene	C10H14		1125.378	134.109	UOHMMEJUHBCKEE-UHFFFAOYSA-N	
165417	Copaene	C15H24		1125.953	204.187	VLXDPFLIRFYIME-BTFPBAQTSA-N	CHEMONTID:0001550
30326	2-Decanone	C10H20O		1139.465	156.151	ZAJNGDIORYACQU-UHFFFAOYSA-N	CHEMONTID:0001831

106758	Benzofuran	C8H6O		1156.628	118.041	IANQTJSKSUMEQM-UHFFFAOYSA-N	
75219	2-Acetylfuran	C6H6O2		1158.953	110.036	IEMMBWWQXVXBEU-UHFFFAOYSA-N	CHEMONTID:0001831
151708	Ethyl-trimethylpyrazine	C9H14N2		1162.003	149.107		
13358	Propyl octanoate	C11H22O2		1179.14	145.122		
18659	2-Nonanol	C9H20O		1180.39	69.07	NGDNVOAEIVQRFH-UHFFFAOYSA-N	CHEMONTID:0001334
51446	Benzaldehyde	C7H6O		1182.978	106.041	HUMNYLRZRPPJDN-UHFFFAOYSA-N	CHEMONTID:0000321
33622	2-Methylthiolan-3-one	C5H8OS		1188.24	116.029	YMZZPMVKABUEBL-UHFFFAOYSA-N	
33887	2-Methylthioethanol	C3H8OS		1191.903	61.011	WBBPRCNXBQTYLF-UHFFFAOYSA-N	CHEMONTID:0003862
63202	Ethyl nonanoate	C11H22O2		1205.078	88.052		
10959	Octyl isobutanoate	C12H24O2		1214.053	89.06		
163632	Ionene	C13H18		1221.19	174.14	LTMQZVLXCLQPCT-UHFFFAOYSA-N	
42410	Linalool	C10H18O		1224.615	71.049	CDOSHBSSFJOMGT-UHFFFAOYSA-N	CHEMONTID:0001549
26870	Isobutyl octanoate	C12H24O2		1229.403	145.122		
24040	1-Octanol	C8H18O		1235.665	84.093	KBPLFHHGFOOTCA-UHFFFAOYSA-N	CHEMONTID:0001334
40186	3-Methylbutyl heptanoate	C12H24O2		1237.14	113.096		
34200	Dimethyl sulfoxide	C2H6OS		1244.74	62.99		
36883	Methyl citronellate	C11H20O2		1245.59	152.12		
48881	Ethyl 3-(methylthio)propanoate	C6H12O2S		1249.075	148.055		
95439	5-Methylfurfural	C6H6O2		1260.663	110.036	OUDFNZMQXZILJD-UHFFFAOYSA-N	CHEMONTID:0001831
75361	2-Propanoylfuran	C7H8O2		1262.875	124.052	HCPORNAVHSWTOJ-UHFFFAOYSA-N	CHEMONTID:0001831
6713	Isobutanoic acid	C4H8O2		1263.913	73.028		
54756	Fenchol	C10H18O		1270.713	80.062	IAIHUHQCLTYTSF-UHFFFAOYSA-N	CHEMONTID:0001549
76037	4-Cyclopentene-1,3-dione	C5H4O2		1278.125	96.021	MCFZBCCYOPSZLG-UHFFFAOYSA-N	CHEMONTID:0001831
73222	Caryophyllene	C15H24		1281.025	204.187	NPNUFJAVOOONJE-GFUGXAQUSA-N	CHEMONTID:0001550
126716	2-Methylbenzofuran	C9H8O		1286.6	132.057	GBGPVUAOTCNZPT-UHFFFAOYSA-N	
30303	2-Undecanone	C11H22O		1295.1	170.167	KYWIYKKSMDLRDC-UHFFFAOYSA-N	CHEMONTID:0001831
94270	2-Acetyl-5-methylfuran	C7H8O2		1300.338	124.052	KEFJLCGVTHRGAH-UHFFFAOYSA-N	CHEMONTID:0001831
76974	5,5-Dimethylfuran-2-one	C6H8O2		1307.975	97.028		
32290	Myrcenol	C10H18O		1315.4	59.049	DUNCVNHORHNONW-UHFFFAOYSA-N	CHEMONTID:0000129
94074	N-Methyl-2-formylpyrrole	C6H7NO		1324.525	108.044	OUKQTRFCDKSEPL-UHFFFAOYSA-N	CHEMONTID:0001831
18283	2-Decanol	C10H22O		1325.438	69.07	ACUZDYFTRHEKOS-UHFFFAOYSA-N	CHEMONTID:0001334
63144	Ethyl decanoate	C12H24O2		1355.813	88.052		
66028	Benzeneacetaldehyde	C8H8O		1358.175	120.057	DTUQWGWMVIHBKE-UHFFFAOYSA-N	CHEMONTID:0001257
40694	Pinocarveol	C10H16O		1369.175	92.062	LCYXQUJDODZYIJ-UHFFFAOYSA-N	CHEMONTID:0001549
33774	2,3,5-Trithiahexane	C3H8S3		1377.71	139.978	MYIOBINSHMEDEY-UHFFFAOYSA-N	CHEMONTID:0004089

40875	3-Methylbutyl octanoate	C13H26O2		1379.773	171.138		
24075	1-Nonanol	C9H20O		1381.135	98.109	ZWRUINPWMLAQRD-UHFFFAOYSA-N	CHEMONTID:0001334
72752	Humulene	C15H24		1384.21	204.187	FAMPSKZZVDUYOS-HRGUGZIWSA-N	CHEMONTID:0001550
77694	2-Hydroxymethylfuran	C5H6O2		1386.36	98.036	XPFVYQJUAUNWIW-UHFFFAOYSA-N	
86753	Ethyl benzoate	C9H10O2		1387.973	150.068		
63265	Ethyl (4E)-4-decenoate	C12H22O2		1391.16	88.052		
3748	(6E)-beta-Farnesene	C15H24		1391.16	120.093	JSNRRGGBADWTMC-NTCAYCPXSA-N	CHEMONTID:0001550
33424	3-Methylbutanoic acid	C5H10O2		1405.023	60.021		
48854	2-Methylbutanoic acid	C5H10O2		1405.485	74.036		
80925	Diethyl succinate	C8H14O4		1405.823	101.023		
516	Ethyl (4Z)-4-decenoate	C12H22O2		1415.998	88.052		
62894	Ethyl 9-decenoate	C12H22O2		1425.373	88.052		
37225	Methyl trans-Geranate	C11H18O2		1430.673	182.13		
32871	alpha-Terpineol	C10H18O		1431.948	136.125	WUOACPNHFRMFPN-UHFFFAOYSA-N	CHEMONTID:0001549
59355	gamma-Hexalactone	C6H10O2		1435.898	85.028	JBFHTYHTHYHCDJ-UHFFFAOYSA-N	CHEMONTID:0001245
30304	2-Dodecanone	C12H24O		1440.298	184.182	LSKONYYRONEBKA-UHFFFAOYSA-N	CHEMONTID:0001831
90113	3-(Methylthio)-1-propanol	C4H10OS		1457.248	106.045	CZUGFKJYCPYHHV-UHFFFAOYSA-N	CHEMONTID:0003862
18837	2-Undecanol	C11H24O		1463.773	69.07	XMUJIPOFTAHSOK-UHFFFAOYSA-N	CHEMONTID:0001334
53701	Dimethyl tetrasulfide	C2H6S4		1486.243	157.935		
161112	Dehydroionene	C13H16		1489.585	172.125	ACFCPKXNISJZKZ-RMKNXTFCSA-N	CHEMONTID:0004622
63168	Ethyl undecanoate	C13H26O2		1490.498	88.052		
33540	Pentanoic acid	C5H10O2		1497.098	60.021		
165549	delta-Cadinene	C15H24		1505.595	204.187	FUCYIEXQVQJBKY-ZFWWWQNUSA-N	CHEMONTID:0001550
24197	Isobutyl decanoate	C14H28O2		1509.508	173.154		
40559	3-Methylbutyl nonanoate	C14H28O2		1515.758	141.127		
68292	Methyl 2-phenylacetate	C9H10O2		1516.258	150.068		
40357	1-Decanol	C10H22O		1518.608	112.125	MWKFXSUHUHTGQN-UHFFFAOYSA-N	CHEMONTID:0001334
36951	Citronellol	C10H20O		1524.283	156.151	QMVPMAAFGQKVCJ-UHFFFAOYSA-N	CHEMONTID:0001549
109391	Curcumene	C15H22		1530.683	202.172	VMYXUZSZMNBRCN-UHFFFAOYSA-N	CHEMONTID:0001550
110823	Methyl salicylate	C8H8O3		1532.908	152.047		
109177	(1S,4S,4aR)-Cubenene	C15H24		1536.595	204.187		
65175	Ethyl 2-phenylacetate	C10H12O2		1549.908	164.083		
109157	Cubenene diastereomer	C15H24		1549.97	204.187		
38458	Neryl propanoate	C13H22O2		1552.258	154.135		
53368	Myrtenol	C10H16O		1554.958	108.057	RXBQNMWIQKOSCS-UHFFFAOYSA-N	CHEMONTID:0001549

59357	gamma-Heptalactone	C7H12O2		1562.908	85.028	VLSVVMPLPMNWBH-UHFFFAOYSA-N	CHEMONTID:0001245
37118	Nerol	C10H18O		1567.645	154.135	GLZPCOQZEFWAFX-YFHOEESVSA-N	CHEMONTID:0001549
30310	2-Tridecanone	C13H26O		1579.095	198.198	CYIFVRUOHKNECG-UHFFFAOYSA-N	CHEMONTID:0001831
84130	2-Phenylethyl acetate	C10H12O2		1587.67	104.062		
38999	Damascenone	C13H18O		1590.745	190.135	POIARNZEYGURDG-FNORWQNLSA-N	CHEMONTID:0001831
37833	Geranyl propanoate	C13H22O2		1592.258	136.125		
18221	2-Dodecanol	C12H26O		1594.345	45.033	XSWSEQPWKOWORN-UHFFFAOYSA-N	CHEMONTID:0001334
161108	(1E)-1-(2,3,6-Trimethylphenyl)buta-1,3-diene	C13H16		1600.92	172.125		
163395	trans-Calamenene	C15H22		1601.498	159.117	PGTJIOWQJWHTJJ-QWHCGFSZSA-N	CHEMONTID:0001550
163396	cis-Calamenene	C15H22		1603.733	159.117	PGTJIOWQJWHTJJ-STQMWFEESA-N	CHEMONTID:0001550
63153	Ethyl dodecanoate	C14H28O2		1623.755	88.052		
36965	Geraniol	C10H18O		1628.968	154.135	GLZPCOQZEFWAFX-JXMROGBWSA-N	CHEMONTID:0001549
33536	Hexanoic acid	C6H12O2		1634.18	60.021		
145358	3-Phenylfuran	C10H8O		1636.818	144.057	BNANPEQZOWHZKY-UHFFFAOYSA-N	
94281	2-Methoxyphenol	C7H8O2		1646.13	124.052	LHGVFZTZFXWLCP-UHFFFAOYSA-N	CHEMONTID:0000190
40558	3-Methylbutyl decanoate	C15H30O2		1646.855	199.169		
21717	1-Undecanol	C11H24O		1649.955	126.14	KJIOQYGWTQBHNH-UHFFFAOYSA-N	CHEMONTID:0001334
166520	Safrole	C10H10O2		1660.83	162.068	ZMQAAUBTXCXRIC-UHFFFAOYSA-N	
53588	Benzyl Alcohol	C7H8O		1665.543	108.057		
177873	Epoxycalamenene	C15H20O		1668.38	173.096		
84137	2-Phenylethyl isobutanoate	C12H16O2		1668.505	104.062		
84437	Ethyl 3-phenylpropanoate	C11H14O2		1673.93	178.099		
161391	Dehydrohimachalene	C15H20		1692.918	200.156		
161103	alpha-Calacorene	C15H20		1707.193	200.156	CUUMXRBKJIDIAY-ZDUSSCGKSA-N	CHEMONTID:0001550
65943	Phenylethyl alcohol	C8H10O		1708.993	122.073		
30307	2-Tetradecanone	C14H28O		1711.443	212.213	POQLVOYRGNFGRM-UHFFFAOYSA-N	CHEMONTID:0001831
59356	gamma-Octalactone	C8H14O2		1712.343	85.028	IPBFYZQJXZJBFQ-UHFFFAOYSA-N	CHEMONTID:0001245
18319	2-Tridecanol	C13H28O		1719.605	45.033	HKOLRKVMHVYNGG-UHFFFAOYSA-N	CHEMONTID:0001334
42203	4-Methylhexanoic acid	C7H14O2		1732.893	60.021		
181204	trans-beta-Ionone	C13H20O		1736.13	177.127	PSQYTAPXSHCGMF-BQYQJAHWSA-N	CHEMONTID:0001550
99474	2-Hydroxymethylthiophene	C5H6OS		1748.778	114.013	ZPHGMBGIFODUMF-UHFFFAOYSA-N	
62801	Ethyl tridecanoate	C15H30O2		1749.39	88.052		
161104	beta-Calacorene	C15H20		1759.753	200.156	KFYISTOZYAKAPV-UHFFFAOYSA-N	CHEMONTID:0001550
45535	2-Ethylhexanoic acid	C8H16O2		1759.69	88.052		
33537	Heptanoic acid	C7H14O2		1764.803	60.021		

84193	2-Phenylethyl isopentanoate	C13H18O2		1769.553	104.062		
120738	Maltol	C6H6O3		1769.803	126.031	XPCTZQVDEJYUGT-UHFFFAOYSA-N	CHEMONTID:0000481
21317	1-Dodecanol	C12H26O		1774.978	140.156	LQZZUXJYWNFBMV-UHFFFAOYSA-N	CHEMONTID:0001334
74075	2-Acetylpyrrole	C6H7NO		1780.778	109.052	IGJQUJNPMOYEJY-UHFFFAOYSA-N	CHEMONTID:0001831
6735	Caryophyllene oxide	C15H24O		1784.03	187.148		
67176	2-Phenyl-1-butanol	C10H14O		1794.303	150.104	DNHNBMQCHKKDNI-UHFFFAOYSA-N	CHEMONTID:0002811
36815	Z-Nerolidol	C15H26O		1806.878	136.125		
160431	1,6-Dimethylnaphthalene	C12H12		1807.203	156.093	CBMXCNPQDUJNHT-UHFFFAOYSA-N	
36255	Perillol	C10H16O		1817.74	152.12	NDTYTMIUWGWIMO-UHFFFAOYSA-N	CHEMONTID:0001549
141556	Guajen	C12H12		1818.09	156.093	WWGUMAYGTYQSGA-UHFFFAOYSA-N	
72763	Humulene epoxide I	C15H24O		1823.328	220.182		
73738	Phenol	C6H6O		1827.74	94.041	ISWSIDIOOBJBQZ-UHFFFAOYSA-N	CHEMONTID:0004647
30617	2-Pentadecanone	C15H30O		1836.978	226.229	CJPNOLIZCWDHJK-UHFFFAOYSA-N	CHEMONTID:0001831
18328	2-Tetradecanol	C14H30O		1840.315	45.033	BRGJIIMZXMWMCC-UHFFFAOYSA-N	CHEMONTID:0001334
75061	2-Formylpyrrole	C5H5NO		1844.915	95.037	ZSKGQVFRTSEPJT-UHFFFAOYSA-N	CHEMONTID:0001831
59349	gamma-Nonalactone	C9H16O2		1846.29	85.028	OALYTRUKMRCXNH-UHFFFAOYSA-N	CHEMONTID:0001245
8547	Furaneol	C6H8O3		1854.328	128.047	INAXVXBDKKUCGI-UHFFFAOYSA-N	CHEMONTID:0001982
37123	E-Nerolidol	C15H26O		1862.553	136.125		
63166	Ethyl tetradecanoate	C16H32O2		1871.328	88.052		
109644	Epicubenol	C15H26O		1875.688	161.132	COGPRPSWSKLKTF-QPSCCSFWSA-N	CHEMONTID:0001550
189958	alpha-Corocalene	C15H20		1881.613	200.156	VTUZIFHLLUSULC-UHFFFAOYSA-N	CHEMONTID:0001550
109639	Diepicubenol	C15H26O		1882.375	161.132		
33546	Octanoic acid	C8H16O2		1890.588	60.021		
40221	Isoamyl laurate	C17H34O2		1892.875	183.174		
132467	Cuminol	C10H14O		1931.163	150.104	OIGWAXDAPKFNCQ-UHFFFAOYSA-N	CHEMONTID:0001549
30616	2-Hexadecanone	C16H32O		1957.288	240.245	XCXKZBWAKKPFCJ-UHFFFAOYSA-N	CHEMONTID:0001831
126603	Ethyl cinnamate	C11H12O2		1962.613	176.083		
54282	Neointermedeol	C15H26O		1966.363	222.198	DPQYOKVMVCQHMY-TUVASFSCSA-N	CHEMONTID:0001550
59347	gamma-Decalactone	C10H18O2		1977.838	85.028	IFYYFLINQYPWGJ-UHFFFAOYSA-N	CHEMONTID:0001245
63165	Ethyl pentadecanoate	C17H34O2		1986.713	88.052		
165009	tau-Cadinol	C15H26O		2003.523	161.132	LHYHMMRYTDARSZ-XQLPTFJDSA-N	CHEMONTID:0001550
84567	2-Phenylethyl hexanoate	C14H20O2		2004.485	104.062		
33554	Nonanoic acid	C9H18O2		2008.635	60.021		
7822	1-Tetradecanol	C14H28O		2009.735	168.187	HLZKNKRTKFSKGZ-UHFFFAOYSA-N	CHEMONTID:0001334
75317	tau-Muurolol	C15H26O		2019.673	204.187		

165024	delta-Cadinol	C15H26O		2032.023	161.132	LHYHMMRYTDARSZ-ZQDZILKHSA-N	CHEMONTID:0001550
132478	4-Vinyl-2-methoxyphenol	C9H10O2		2036.023	150.068	YOMSJEATGXXYPX-UHFFFAOYSA-N	CHEMONTID:0000190
111126	2-Acetylaniline	C8H9NO		2054.673	120.044	GTDQGKWDWVUKTI-UHFFFAOYSA-N	CHEMONTID:0001831
187944	Cadalene	C15H18		2055.073	198.14	VMOJIHDTVZTGDO-UHFFFAOYSA-N	CHEMONTID:0001550
75316	alpha-Cadinol	C15H26O		2062.548	222.198	LHYHMMRYTDARSZ-BYNSBNAKSA-N	CHEMONTID:0001550
30615	2-Heptadecanone	C17H34O		2065.248	254.26	TVTCXPXLRKTHAU-UHFFFAOYSA-N	CHEMONTID:0001831
63170	Ethyl hexadecanoate	C18H36O2		2086.123	88.052		
59352	gamma-Undecalactone	C11H20O2		2090.323	85.028	PHXATPHONSXBIL-UHFFFAOYSA-N	CHEMONTID:0001245
196879	Myristicin	C11H12O3		2095.26	192.078	BNWJOHGLIBDBOB-UHFFFAOYSA-N	
7302	Pyranone	C6H8O4		2098.798	144.042	ZPSJGADGUYYRKE-UHFFFAOYSA-N	CHEMONTID:0000481
33538	n-Decanoic acid	C10H20O2		2105.948	60.021		
191449	Diphenyl sulfide	C12H10S		2124.658	186.05		
9532	(2Z)-8-Hydroxylinalool	C10H18O2		2133.945	71.049		
3725	(2Z,6E)-Farnesol	C15H26O		2134.895	69.07	CRDAMVZIKSXKFV-PVMFERMNSA-N	CHEMONTID:0001550
37173	Geranic acid	C10H16O2		2157.558	100.052		
3724	trans-Farnesol	C15H26O		2165.208	69.07	CRDAMVZIKSXKFV-YFVJMOTDSA-N	CHEMONTID:0001550
63160	Ethyl heptadecanoate	C19H38O2		2166.633	88.052		
59346	gamma-Dodecalactone	C12H22O2		2180.733	85.028	WGPCZPLRVAWXPW-UHFFFAOYSA-N	CHEMONTID:0001245
21892	1-Hexadecanol	C16H32O		2181.808	196.219	BXWNKGSJHAJOGX-UHFFFAOYSA-N	CHEMONTID:0001334
84607	2-Phenylethyl octanoate	C16H24O2		2184.358	104.062		
33547	Undecanoic acid	C11H22O2		2186.358	60.021		
110639	Dihydrobenzofuran	C8H8O		2197.908	120.057	HBEDSQVIWPRPAY-UHFFFAOYSA-N	
79196	delta-Dodecalactone	C12H22O2		2215.445	99.044	QRPLZGZHJABGRS-UHFFFAOYSA-N	CHEMONTID:0001244
24154	Butyl palmitate	C20H40O2		2218.958	257.248		
105976	Indole	C8H7N		2231.145	117.057	SIKJAQJRHWYJAI-UHFFFAOYSA-N	CHEMONTID:0002497
63171	Ethyl octadecanoate	C20H40O2		2236.295	88.052		
160684	Triethyl citrate	C12H20O7		2241.97	157.05		
78766	Succinimide	C4H5NO2		2247.27	99.031	KZNICNPSHKQLFF-UHFFFAOYSA-N	CHEMONTID:0001158
7814	Ethyl oleate	C20H38O2		2249.48	264.245		
87049	Benzophenone	C13H10O		2252.343	182.073	RWCCWEUUXYIKHB-UHFFFAOYSA-N	CHEMONTID:0000120
84608	2-Phenylethyl nonanoate	C17H26O2		2257.455	104.062		
44889	Dodecanoic acid	C12H24O2		2257.643	60.021		
125785	3-Methylindole	C9H9N		2263.318	130.065	ZFRKQXVRDFCRJG-UHFFFAOYSA-N	CHEMONTID:0002497
77412	5-Hydroxymethylfurfural	C6H6O3		2269.98	126.031	NOEGNKMFWQHSLB-UHFFFAOYSA-N	CHEMONTID:0001831
35402	Ethyl linoleate	C20H36O2		2283.655	308.271		

155194	Vanillin	C8H8O3		2318.43	152.047	MWOOGOJBHIARFG-UHFFFAOYSA-N	CHEMONTID:0000190
76629	2-Furanylethanediol	C6H8O3		2325.53	128.047		
84563	2-Phenylethyl decanoate	C18H28O2		2339.755	104.062		
87937	Benzyl benzoate	C14H12O2		2370.405	212.083		
44887	Tetradecanoic acid	C14H28O2		2436.628	60.021		

Supplemental TableS8_MSHub annotation accuracy testing								
Metabolomics Workben	**Number of files**	**GNPS deconvolution link for MS**	**GNPS library search link, MSHub deconvolution**					
Subset 5 files	5	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=66af7b6ed1894b89a40ee1559f68f03f					
Subset 10 files	10	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f40db6d96c1e4ce6bbb4fe9ab88e5a1d					
Subset 20 files	20	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=af8a292b4c2e4226833e8d4f37a6141d					
Subset 30 files	30	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7fcb253c5d3743b4a0715bbe2216963a					
Subset 40 files	40	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8a126cec39704c8fa148220630200041					
Subset 50 files	50	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=562d5808353042da998ed2e561ea4833					
Subset 100 files	100	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ad5156516b924f06b5e78f34c793fce4					
Subset 150 files	150	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c757c34fd4b24bd8890ca9e41f2ccbe6					
Subset 200 files	200	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=307bcc2a80ab432d87028141b07e8483					
Subset 250 files	250	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d132c3e31f604136babb67d092e52410					
Full set 300 files	300	https://gnps.ucsd.edu/ProteoSAFe	https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4f2de61ac9b949ddb1dd68b0d2fcd76f					

Supplementary Notes

MSHub description:

The MSHub platform includes a set of modules for optimized analytical pre-processing workflow for raw gas chromatography-mass spectrometry (GC-MS) data. It accounts for common bio-analytical complexities (including data volume burden, platform-specific biases and noise structure) inherent to high-throughput GC-MS datasets of biological samples. The proposed pipeline includes a series of designated modules from raw data import, noise/baseline removal and peak alignment strategies to machine-learning driven fragmentation spectra extraction and quantity integral calculation of molecules. All processing modules have been designed to operate in an iterative fashion (i.e. processing a single sample dataset at a time) and are thus scalable for processing of hundreds to thousands of samples.

i) Raw data import: The module builds an HDF5 database file and writes raw GC-MS data from multiple data files into it. It supports reading of the data from multiple open-source data formats including ANDI-MS/.NetCDF and .mzML. The output database file contains a series of mass spectral scans taken along the retention time axis for each of the imported files. Each scan is a matrix of pairs of raw *m/z* values and peak intensities. *ii) m/z peak matching:* Since raw GC-MS data are generally acquired at nominal mass resolution, this module performs *m/z peak matching* by integrating the raw data to fixed *m/z* intervals with a nominal mass resolution. By default, the data are integrated at nominal mass with the boundaries set at -0.3 and +0.7 to reduce the shift of fragments between different *m/z* intervals[1]. For high-resolution mass spectrometry data, MSHub offers a kernel-based clustering approach to match m/z values within and between samples at a mass spectrometer's native resolution. This option is currently available as a stand-alone code. Each sample is then represented by 2D data matrix with each row corresponding to a nominal mass spectrum at a specific time point and each column corresponding to an ion chromatogram *iii) Noise filtering and baseline correction:* GC-MS data both contain high-frequency electronic noise and low frequency baseline distortions emerging due to various instrumental and experimental imperfections[2]. The module performs adjustments of high-frequency noise by applying a digital polynomial (Savitzky-Golay[3]) filter. It is particularly suited to filter out noise without distorting the chromatographic signal tendency[4]. In addition, the morphological filter is applied to correct baseline distortions[3]. The frame-lengths for these filters are automatically estimated with regard to the width of the average chromatographic peak. This procedure acts on single ion chromatograms of sample data matrix. *iv) Retention time alignment:* The inherent drifts of peaks in ion chromatograms between samples are adjusted by means of recursive segment-wise peak alignment strategy[5]. In brief, the strategy refines the alignment of a given ion chromatogram with regard to the reference one in a top-down fashion. It starts by shifting of all ion chromatographic peaks and then progresses to smaller segments when further alignment is required. The optimal shift position is found at the maximum of the cross correlation function:

$$corr(r,s) = \sum_{i=1}^{n} r(i) \cdot s(i+d) \quad [S.1]$$

$$d_{opt} = argmax(corr(r,s)_d) \quad [S.2]$$

where d_{opt} is an optimal shift between reference (*r*) and test (*s*) segments and n is a segment length. The calculation of this cross correlation function is accelerated by the Fast Fourier transform.

$$corr(r,s)_d \leftrightarrow R(j) \cdot S^*(j) \quad [S.3]$$

where R and S are the Fourier transform of functions of r and s, respectively, superscript star (*) denotes a complex conjugation and ↔ denotes fast Fourier transform (FFT). The cross correlation function is calculated by applying a fast Fourier transform to both r and s functions, multiplying the transformed functions, *R* and *S**, and finally performing the inverse Fourier transform of this product. The use of FFT reduces the alignment complexity from $O(N^2)$ *to* $O(N*logN)$ making it possible to perform alignment at a full chromatographic resolution. The reference profile is selected as a mean chromatographic ion profile across an entire dataset. The maximum peak shift and minimum segment width parameters are automatically defined with respect to the average peak width but can be overwritten by the user if known a priori. With this strategy, MSHub algorithm is capable of correcting both linear and considerable non-linear shifts of ion chromatograms, and doesn't rely on a simple linear shift of ion chromatograms. *v*) *Spectral deconvolution*: Following peak alignment, the individual chromatographic peaks are detected by means of the smoothed Savitzky-Golay first derivative and integrated using the trapezium rule. Each chromatographic peak with a given retention time index is arranged into the data matrix of nominal mass spectra with rows and columns representing samples and m/z channels, respectively. This data matrix may contain multiple co-eluting molecules with distinct but potentially correlated and overlapping fragmentation spectra. Mathematically, the deconvolution process of co-eluting molecules can then be formulated as:

$$X \approx \sum_{i=1}^{k} W_i \cdot H_i + E$$

where *X* is a mass spectral data matrix at a given retention time index; W_i and H_i are a fragmentation MS spectrum and level of the deconvoled i-th molecular component and E is the residual data matrix. We have recently shown that the solution to the above problem can be effectively found via unsupervised machine learning strategy using non-negative matrix factorization tools[6]. It allows the deconvolution of potentially multiple co-eluting molecules taking advantage of increased number of samples. The fragmentation pattern and levels of a molecule are recovered from component loadings and scores, respectively. To assess the reproducibility of fragmentation patterns between samples, we defined the balance score as the percentage of variance explained by the component as

$$B_{score} = 100 \cdot (1 - ||X - W_i \cdot H_i||^2 / ||X||^2)$$

$||X||^2$ denotes the sum-of-squares of the elements in X. If a given component explains the majority of variation (>95%), this means that the fragmentation patterns are consistent between samples.

The number of components for each chromatographic peak is determined automatically, so that each component has to explain at least 5% of the total variation between samples. In this regard, spectral patterns that are consistent across samples, even if only a few samples in the data contain them, would result in a high value of the balance score. The patterns with balance scores below 1% are removed as noise. At the results stage, users then can apply their own filtering to retain the spectral library matches that they deem of high quality using match score, balance score, number of shared peaks in the spectrum, total ion chromatogram (TIC), Kovats RI (if enabled by user) and used library(ies). The guidelines for the balance score filtering strategies are given in the tutorial.

We have designed the MSHub to follow the key requirements for large scale MS studies such as analytical transparency and reproducibility, workflow versatility, and scalability [7]. The unique features of the platform are:
(A) modular, customizable and readily extendable workflow with enhanced data pre-processing capabilities, covering all prerequisite steps for chromatography mass spectrometry data processing (**Figure 1**)
(B) a scalable "out of core" data pre-processing pipeline; out of core (or "external memory") processing is a technique used to process data that is too exhaustive to fit in a computer's main memory (RAM). All MSHub algorithms have been designed to operate iteratively for enhanced scalability by using high performance HDF5 technologies. A design feature incorporated into MSHub allows individual sample data to be uploaded one at a time into the specific module, where after data are processed, deleted from memory and deposited back into the data repository, with this procedure then repeated iteratively. **Figure S2a-f** illustrates linear dependency between the number of samples processed and the processing time, with less than 20 hours of total processing time for ~5000 GC-MS profiles on a single core desktop PC. Since only one sample is stored in memory at any given time, the workflow memory load is constant.
(C) Machine-learning enhanced fragmentation pattern deconvolution: in contrast to other solutions, MSHub extracts fragmentation spectra of individual components and their relative levels using unsupervised learning via non-negative matrix factorization tools. This allows the extraction of reproducible fragmentation patterns between spectra, as well as robust separation of co-eluting components.
(D) workflow reproducibility – all workflow steps and generated pre-processing metadata (e.g. common *m/z* or retention time feature vector, choice of baseline correction strategy, user defined parameter settings) are stored as part of the database file. This feature allows consistent and reliable comparison between newly collected and archived data.
The output of the pre-processing workflow includes fragmentation spectra and quantitative integrals of molecules. The putative annotation of the extracted fragmentation spectra from GC-MS data can be achieved with the use of library search workflow. The workflow utilizes any of the user-provided libraries (in all the examples given in the present manuscript the mass spectral databases of NIST 17 (http://www.nist.gov/srd/nist1a.cfm) and Wiley (https://www.wiley.com/en-us/Wiley+Registry%3A+Mass+Spectral+Library%2C+11th+Edition+NIST+2017-p-9781119412236) were used, as well as public libraries available on GNPS such as the Golm Metabolome Database (GMD, http://gmd.mpimp-golm.mpg.de) or user's own in-house spectral databases. The MSHub platform has been benchmarked with regard to the extraction of molecular fragmentation spectra and their quantitative integrals in small scale studies (up-to 100 breath samples) against existing state-of-the-art open-source ("XCMS")[8] and proprietary (MassHunter, Agilent and ChromaTOF, LECO) processing pipelines. In case of small scale studies of standard mixtures where the data size is sufficient for "in-core" processing (i.e. can fit into the main memory) and minor retention time drifts are expected, the high similarity (>99%)

between the extracted molecular fragmentation spectra and their quantity integrals by XCMS and MSHub processing pipelines was achieved (see **Figure 2t,u**).
We have benchmarked the MSHub against other tools, both in terms of deconvolution accuracy and processing times. Deconvolution accuracy of MSHub is similar or better than the other tools operated under optimal settings (**Figure S3, S5**). Alternative deconvolution/alignment tools include: MZmine[9], OpenChrom[8,10], AMDIS[11], MZmine/ADAP[12], MS-DIAL[13], BinBase[14], XCMS[15]/XCMS Online[8], MetAlign[16], SpecAlign[17], SpectConnect[18], PARAFAC2[19], MeltDB[20], eRah[21]).
The processing times of other tools scale exponentially with the number of files and thus they fail for datasets of a certain number of files, depending on the computing power of the server (**Figure S4**). As for all deconvolution algorithms, peak overloading reduces data quality. MSHub finds less accurate spectral patterns across samples for saturated peaks, as the frequency domain information is no longer consistent for those ions across all files, lowering the balance score. Overloading does not affect the signals - and the resulting balance scores - that are not saturated. The optimal solution to mitigate possible deconvolution errors is experimentally reducing the amount of sample that is loaded onto the GC-MS column.

GC-MS analysis for validation studies

Prior to analysis, all of the data were converted from vendor's proprietary formats to .NetCDF or .mzML, uploaded to MassIVE (https://massive.ucsd.edu; the corresponding dataset IDs are listed in the **Table S1**) and processed using GNPS GC-MS data analysis workflow as described in the tutorial. The compounds were annotated via matching spectra to the public reference libraries available on GNPS (currently GNPS has Fiehn[22], HMDB[23], MoNA[24], VocBinBase[14]) as well as NIST 17 and Wiley (the latter two combined make up 1,051,748 of spectra). All of the corresponding analysis links are listed in **Table S1**.

Testing and validation dataset of derivatized human blood serum spiked with standards (datasets Test1-Tes11, #17 in Table S1)
Acetonitrile, isopropanol, and pyridine were of LC-MS grade, and methoxyamine hydrochloride (MeOX), 1% TMCS in *N*-methyl-*N*-trimethylsilyl-trifluoroacetamide (MSTFA), and adonitol were obtained from Sigma-Aldrich, St. Louis, USA. Human serum (Cat. No. H6914) from a US origin individual male clotted whole blood, AB positive group, sterile-filtered) was obtained from Sigma-Aldrich, St. Louis, USA. Fatty acid methyl ester (FAMEs) mixtures (Supelco® 37 Component FAME Mix - 10 mg/mL in methylene chloride (varied concentration, 2-4% wt. basis) (Cat. No. 47885-U) was obtained from Sigma-Aldrich, St. Louis, USA. An n-hydrocarbon mixture of C8 - C40 (even C numbers), 17 components mixture (Cat. No. 95394) (concentration, 2000 µg/mL) was obtained from Absolute Standards Inc., CT, USA.
Post-thawing, human serum samples (30 µL) were sequentially extracted using solvent extraction, once with 1 mL of acetonitrile: isopropanol: water (3:3:2, v/v) ratio, followed by another extraction with 500 µL of acetonitrile: water (1:1, v/v) ratio mixture at 4 °C. Adonitol (Cat. No. A5502-5G, Sigma-Aldrich; 5 µL from a 10 mg/ml stock), an internal standard was added to all samples prior to solvent extraction. The pooled extracts (~ 1500 µL) resulting from the two steps were dried *in vacuo* at 4 °C prior to chemical derivatization. Dummy extractions were performed

in blank tubes that served as extraction blanks for background subtraction of noise resulting from extraction solvents, derivatization reagents, unwanted sources, and other sources of contamination from the analytical platform, i.e., septa, liner, column, vials, handling etc. All samples (i.e., except FAMEs and n-alkanes) were sequentially derivatized with methoxyamine hydrochloride (MeOX) and 1% TMCS in *N*-methyl-*N*-trimethylsilyl-trifluoroacetamide (MSTFA) as follows. The steps involved the addition of 10 µL of MeOX (20 mg/mL in pyridine) incubated at 55 °C for 60 min followed by trimethylsilylation at 60 °C for 60 min after adding 90 µL MSTFA.
The acquisition sequence started with blank solvent (pyridine) injections, followed by randomized lists of extraction blanks (B), reagent blanks (R), solvent (pyridine-P), and samples (S). FAMEs and n-hydrocarbon mixtures were spiked into the derivatized human serum samples as described in **Supplemental Table 1**. A robotic arm TriPlus™ RSH autosampler (Thermo Scientific™, Bremen, Germany) was used to inject 1 µL of derivatized sample into a split/splitless (SSL) injector set at 250 °C using splitless injection onto a TRACE™ 1310 gas chromatograph (Thermo Scientific™, Austin, TX). Helium was used as a carrier gas at a flow rate of 1 mL/min for separation on a TG-5SILMS™ 30 m length × 0.25 mm i.d. × 0.25 µm film thickness column (Thermo Scientific™, P/N, 26096-1420). The initial oven temperature held at 70 °C for 4 min was ramped through an initial gradient of 20 °C/min ramp rate. The final temperature was set to 320 °C and was held for 8 min. Eluting peaks moved through an auxiliary transfer line temperature of 250 °C into a QExactive™-GC mass spectrometer (Thermo Scientific™, Bremen, Germany). Electron ionisation (EI) at 70 eV energy, an emission current of 50 µA and an ion source temperature of 230 °C was used in all experiments. A filament delay of 5.3 min was selected to prevent excess reagents from being ionized. High resolution EI spectra were acquired using 60,000 resolution (FWHM at *m/z* 200) with a mass range of *m/z* 50-650.

Cheese volatiles (datasets #13-15,Table S1) and cheese bacteria grown on cheese-agar media (datasets #13-15, 28 in Table S1)

Two cheese types, Bonde and Stilton were purchased at Whole Foods store. Small cubes of approximately 1.25 cm^3 were cut from pieces of cheese, both rind and bulk, to collect samples across the entire piece, placed into Eppendorf tubes and stored in -80°C freezer until analysis. For the bacterial cultures samples, *Brevibacterium* and *Brachybacterium* isolates were grown on various cheeses (Langres, Cornish Yarg, Robiola della Spazza Camino, Ascutney Mountain, Sharfe Maxx, Gruyere, Challerhocker, Bayley Hazen Blue). Prior to analysis, the frozen samples were transferred into 2 mL borosilicate vials, capped with a screw cap with silicone septum and allowed to thaw. The GC-MS analysis was carried out on a Thermo Scientific TRACE 1310 GC [TG-5MS column; length, 30 m; ID (inner diameter), 0.25 mm; film thickness column, 0.25 µm] and a TSQ 8000 EVO mass spectrometer (Thermo Fisher Scientific), equipped with electron ionization (EI) source and a robotic sampler system. The PDMS/DVB SPME (Supelco) tip was inserted by the sampling robot into the vial, and the sample was heated to 160 °C and agitated for 10 min. Upon sample extraction, the SPME was inserted into the GC inlet maintained at 250 °C, and the extracted compounds were desorbed for 1 min. The GC protocol was as follows: cryofocusing on the head of the column at -10 °C for 1.25 min; 100 °C/min oven ramp to 40 °C (hold of 0.1 min), 15 °C/min oven ramp to 280 °C (hold of 0.1 min), and a 3 min hold period to purge the column. The helium carrier gas was set to constant 2 mL/min flow, splitless injection mode was used throughout. The scanned *m/z* range in a single quadrupole was 35-350. The

empty vial blanks were interspersed with the samples. Quality controls of natural mint oil extract were run along with samples throughout the analysis to monitor instrument performance and SPME wear.

Skin volatilome analysis

All of the sampling has been conducted under UC San Diego IRB #171662. Cleaned and degassed polydimethylsiloxane (PDMS, Goodfellow, Coraopolis, PA) 1 x 2 cm patches have been placed on 52 different points of the skin of a volunteer and secured in place with waterproof bandage. The sampling was conducted for 6 h during which volunteer performed normal daily activities.

Headspace sample introduction (datasets #19, 26 in Table S1)

After sampling, the bandage was removed, and patches placed into 2 mL borosilicate vials using clean tweezers and capped with a crimp camp with silicone septum. The GC-MS analysis was carried out using the Agilent 7200 GC/QToF (Agilent Technologies, Santa Clara, CA) equipped with a robotic sampler system. The separation was conducted on a HP-5MS column (30 m x 0.25 mm x 0.25 µm). The patch within the vial was heated for 20 min. at 200 °C to desorb volatiles from the patch and 0.5 mL of headspace injected (injector temperature set at 250 °C) into the instrument with a headspace syringe heated to 145 °C. The GC protocol analysis included: starting temperature 45 °C; 50 °C/min oven ramp to 100 °C (hold of 0.1 min), 15 °C/min oven ramp to 300 °C (hold of 0.1 min), and 50 °C/min oven ramp to 320 °C purge the column. The helium carrier gas was set to constant 1.2 mL/min flow and a splitless injection mode was applied. The scanned *m/z* range was 35-400 with the acquisition rate of 10 spectra/s. The empty vial blanks and blank PDMS patches were interspersed with the samples to assess background signal. Quality controls of natural mint oil extract were run along with samples before and after the analysis as described above.

Liquid sample extraction (datasets #27 in Table S1)

After sampling, the bandage was removed, patches placed into 2 mL borosilicate vials using clean tweezers and 150 µL of HPLC grade methanol added for extraction. The extraction was carried out overnight at room temperature. After extraction, the methanol extract was transferred into a 200 µL insert in a 2 mL vial and the vial capped with a cap with silicone septum. The GC-MS analysis was carried out using the Agilent 7200 GC-QTOF equipped with a robotic sampler system. The separation was conducted on an HP-5MS column (30 m x 0.25 mm x 0.25 µm). Sample (1 µL) was injected at split 2:1. The injector temperature was set at 250 °C. The GC protocol analysis was as follows: starting temperature 40 °C for 1 min; 20 °C/min oven ramp to 110 °C, 10 °C/min oven ramp to 300 °C and 50 °C/min oven ramp to 320 °C purge the column. The helium carrier gas was set to constant 1.2 mL/min flow and a splitless injection mode was applied. The scanned *m/z* range was 35-400 Th with the acquisition rate of 20 spectra/s. The methanol solvent blanks and blank PDMS patches extracts were interspersed with the samples to assess background signal. Quality controls of natural mint oil extract were run along with samples before and after the analysis as described above.

Oesophageal and gastric cancer (OGC) detection using exhaled breath analysis

Participant recruitment:

Ethical approval was obtained (REC 14/LO/1136) and all participants provided written informed consent. Participants for this study were recruited from Imperial College Healthcare NHS Trust into two possible groups – either those with known oesophago-gastric cancer (OGC) or a non-cancer control group. The participants who entered into the non-cancer control group were recruited from the endoscopy suite at Imperial College Healthcare NHS Trust and were undergoing an upper gastrointestinal endoscopy on the day of providing an exhaled breath sample (and were found not to have oesophageal or gastric cancer). If patients were also undergoing a colonoscopy at the same time (and had received bowel preparation medications) they were not eligible to participate in the study. The participants who entered into the OGC group were patients who had a biopsy proven invasive gastric or oesophageal adenocarcinoma. These patients were recruited from three possible locations – either on the day of undergoing a diagnostic upper gastrointestinal endoscopy, on the day of having a staging laparoscopy and upper gastrointestinal endoscopy under general anesthetic (part of the routine staging investigations for OGC), or on the day of review in the outpatient clinic. Patients who had already received surgical or oncological (chemotherapy or radiotherapy) treatment for oesophageal or gastric cancer were not eligible to participate in the study. Participants were between the ages of 18 and 90, and were excluded if they were known to have liver disease (including oesophageal varices and known portal hypertension, an acute infection, another type of cancer presently or within the past 5 yrs, or known inflammatory conditions of the small or large bowel.

Breath sample collection:

Exhaled breath samples were collected using a standardized breath-sampling device, 'Respiration Collector for In Vitro Analysis' (ReCIVA^{TM}) (Owlstone Medical, Cambridge, UK) in combination with a dedicated clean air supply 'Clean Air Supply Pump for ReCIVA' (CASPER) (Owlstone Medical, Cambridge, UK)[25]. All participants were fasted for a minimum of 4 h and rested for 20 min. prior to exhaled breath sample collection.

Two studies were performed. For study one, four sampling thermal desorption (TD) tubes (Tenax/Carbograph-5TD, Markes International Ltd, Llantrisant, UK) were used per participant. For study two, one sampling TD tube (Tenax/Carbograph-5TD, Markes International Ltd, Llantrisant, UK) was used per participant. Prior to sample collection all TD tubes were conditioned for 40 min. at 330 °C using a TC-20 tube conditioner (Markes International Ltd, Llantrisant, UK). The TD tubes were stored in an airtight polypropylene container at room temperature and used for sample collection within 24 h of conditioning.

Exhaled breath collection was performed using a standardized protocol with the participant performing normal tidal respiration whilst seated[25]. For study one exhaled breath sample collection with the ReCIVA device was performed using a sample volume 250 mL per TD tube, and a sample flow rate of 400 mL/min. For study two exhaled breath sample collection with the ReCIVA device was performed using a sample volume 500 mL per TD tube, and a sample flow rate of 200 mL/min. Prior to analysis TD tubes were stored in an airtight polypropylene container at room temperature and all TD tubes were analyzed within 12 h of breath sample collection. TD tubes that had been conditioned in preparation for exhaled breath sample collection and subsequently not used (due to a lower than expected number of participants

being recruited were) analyzed concurrently with the exhaled breath samples as 'blank' TD tubes.

Analysis with TD-GC-MS:

The exhaled breath and blank TD tubes samples were analyzed using TD-GC-MS. The TD tubes were desorbed using a Markes TD-100 thermal desorption unit (Markes International Ltd, Llantrisant, UK) using a two stage desorption program, applying a constant flow of helium at 50 mL/min. In the primary desorption stage, TD tubes were dry-purged for 3 min and heated at 280 °C for 10 min. In the secondary desorption stage, the cold trap (U-T12ME-2S, Markes International Ltd, Llantrisant, UK) was rapidly (99 °C/min) heated to from 10 °C to 290 °C. VOCs were transferred from the TD unit to the GC by means of a capillary line heated at 140 °C. GC-MS analysis was performed using an Agilent 7890B GC with 5977A MSD (Agilent Technologies Ltd, Santa Clara, USA) equipped with a ZB-642 capillary column (60 m x 0.25 mm ID x 1.40 µm df; Phenomenex Inc, Torrance, USA) with helium used as the carrier gas (1 mL/ min flow rate). The GC column temperature program was set as follows: 4 min at 40°C, ramp to 100 °C at 5 °C/min with a 1 min hold, ramp to 110 °C at 5 °C/min with a 1 min hold, ramp to 200°C at 5°C/min with a 1 min hold and finally ramp to 240 °C at 10 °C/min with a 4 min hold. The MS transfer line temperature was 240 °C and EI source conditions were 70 eV at 230 °C. Mass spectral acquisition was carried out in the range 20-250 *m/z* with a rate of approximately 6 scans/s.

Determination of OGC biomarkers

The feature tables were generated by the GNPS deconvolution workflow of the breath analysis study data (Study 1: dataset #34, Study 2: dataset #35 in **Table S1**). The discriminating features between the OGC and control groups were selected with maximum margin criterion[26]**.** For Study 1, the area under the curve for the ROC curve was 0.90227; for Study 2 the area under the curve for the ROC curve was 0.74000, leaving patients out cross validation was carried out in both cases. This indicates that cancer detection is possible in both studies, and with very high accuracy of Study 1. The lists of features with corresponding p and q values are given in the Supplemental **Tables S2** and **S3**. The corresponding feature annotations can be related from the GNPS library search jobs listed for the dataset #34 and dataset #35, respectively.

Skin volatome analysis for hedonic value estimation pre- and post-bacterial spray

All of the sampling has been conducted under Ghent University Ethical Approval #B670201835548, Belgium, and ClinicalTrials.gov identifier NCT03967470, with study called "Armpit bacteriotherapy". A total of 63 volunteers were recruited and selected with a stronger than normal underarm malodor, as determined by a trained odor panel (DOI: 10.1111/exd.13259). The volunteers were using bacterial sprays and placebo sprays in the underarm to verify the effect on the underarm odor. Odor analysis was done using hedonic value measurement by a trained odor panel, as well as GC-MS analysis. Cleaned and degassed PDMS patches were placed on the underarm skin of the volunteer for 4h and secured in place with a large cotton bandage. After sampling, the bandage was removed, patches placed into 2 mL borosilicate vials using clean tweezers and capped with a crimp camp with silicone septum.

The GC-MS analysis was carried out using the Agilent 7200 GC-QToF equipped with robotic sampler system. The separation was conducted on an HP-5MS column (30 m x 0.25 mm x 0.25 µm). The patch within the vial was heated for 20 min. at 200 °C to desorb volatiles from the patch and 0.5 mL of headspace injected into the instrument with a headspace syringe heated to 145°C. The GC protocol analysis included: The GC protocol analysis included: starting temperature 45 °C; 50 °C/min oven ramp to 100 °C (hold of 0.1 min), 15 °C/min oven ramp to 300 °C (hold of 0.1 min), and 50 °C/min oven ramp to 320 °C purge the column. The helium carrier gas was set to constant 1.2 mL/min flow and a splitless injection mode was applied. The scanned m/z range was 35-400 with the acquisition rate of 10 spectra/s. The empty vial blanks were interspersed with the samples to assess background signal. Quality controls of natural mint oil extract were run along with samples before and after the analysis.

3D cartography of skin volatiles

Skin sampling has been conducted as described in the [27] by swabbing areas of skin with a cotton swab and extracting the swabs using ethanol. The samples were stored in -80 °C freezer until analysis. The GC-MS analysis was carried out on a Thermo Scientific TRACE 1310 GC-MS [TG-5MS column; 30 m (length) x0.25 mm ID (inner diameter) x 0.25 µm(film thickness column)] and a TSQ 8000 EVO mass spectrometer (Thermo Fisher Scientific), equipped with electron ionization (EI) source, equipped with robotic sampler system. Aliquot of sample (1 µL) was injected into the GC inlet maintained at 250 °C. The GC protocol was as follows: starting temperature 40 °C (hold of 0.1 min), 15 °C/min oven ramp to 280 °C (hold of 0.1 min), and a 3 min hold period to purge the column. The helium carrier gas was set to constant 2 mL/ min flow, splitless injection mode was used throughout. The scanned *m/z* range in a single quadrupole was 35-350 with solvent delay of 4 min..

Volatilome of human-built environment/HomeChem project (dataset #12, Table S1)

Cleaned and degassed PDMS (Goodfellow, Coraopolis, PA) 1 x 2 cm patches have been placed on different locations across the test house (the detailed description of the project can be found at: https://indoorchem.org/projects/homechem/) and exposed to air for 3 h. After sampling, the patches were placed into 2 mL borosilicate vials using clean tweezers and capped with a crimp camp with silicone septum. The GC-MS analysis was carried out using the Agilent 7200 GC/QTOF equipped with a robotic sampler system. The separation was conducted on an HP-5MS column (30 m x 0.25 mm x 0.25 µm). The patch within the vial was heated for 20 min. at 200°C to desorb volatiles from the patch and 0.5 mL of headspace injected (injector temperature set at 250 °C) into the instrument with a headspace syringe heated to 145 °C. The GC protocol analysis included: starting temperature 45 °C; 50 °C/min oven ramp to 100 °C (hold of 0.1 min), 15 °C/min oven ramp to 300 °C (hold of 0.1 min), and 50 °C/min oven ramp to 320 °C purge the column. The helium carrier gas was set to constant 1.2 mL/min flow and a splitless injection mode was applied. The scanned *m/z* range was 35-400 with the acquisition rate of 10 spectra/s. The empty vial blanks and blank PDMS patches were interspersed with the samples to assess background signal. Quality controls of natural mint oil extract were run along with samples before and after the analysis.

VOCs from Dendrobatids frogs from Colombia (dataset # 11, Table S1)

Collection of specimens

Between 2 and 16 specimens were captured in the Departments of Chocó, Cundinamarca, and Leticia between 2016 and 2018. A framework permit to conduct this study was provided by Autoridad Nacional de Licencias Ambientales (ANLA) in the resolution 1177, granted to La Universidad de los Andes for the Collection of Specimens of Wild Species of Biological Diversity for Non-Commercial Scientific Research Purposes. The animals were collected by visual encounter protocols. All animals were collected using a plastic cup to avoid hand manipulation and kept in plastic bags with a small portion of water to avoid dehydration.

Afterward, the animals were transported to the Universidad de los Andes where three types of sampling methods were performed. First, an in vivo sampling of the VOCs released by each frog. Secondly, a VOCs sampling of the frog's skin after euthanization. Both procedures were performed following the protocol designed by Brunetti et al. (2015). Third, a solid-liquid extraction of the compounds in 1 mL of MeOH 98%. These analyses were performed for tacking and comparing the VOCs released by different species of Dendrobatids (Gonzalez et al., unpublished data). This study was carried out according to the regulations specified by the Institutional Animal Care and Use Committee of the Facultad de Ciencias de la Universidad de los Andes, (FUA_19-015).

In vivo VOCs sampling

Live frogs were sampled using HS-SPME/GC-MS analysis. Each specimen was freed from urine by applying a gentle pressure to the bladder. It was then placed inside a glass chamber designed especially for sampling frogs' VOCs by opening the bottom side of the chamber and immediately closed. Then, the upper side was sealed with a SPME compatible caps of 20 mm with PTFE septa (JG Finneran, Vineland, NJ, USA), and the chamber was introduced in a water bath set at 35 °C. It is worth mentioning that during field and laboratory handling, the characteristic smells of the frogs were perceived when the individuals were incidentally stressed. So, in order to promote the release of these compounds, we performed the *in vivo* sampling by applying a mild electrical stimulation. Two platinum electrodes were inserted through the small holes on the lateral sides of the glass chamber. After exposing a DVB/CAR/PDMS SPME fiber (SUPELCO, PA, USA) for 30 min, it was immediately inserted into the injection port of the gas chromatograph. Five min after being released from the device, the frog exhibited a normal behavior. Before each experiment, the *in vivo* glass chamber was washed with a non-ionic detergent, rinsed and heated to dryness. After cleaning, a blank run of this device was performed using the same extraction conditions in order to trace contaminants from the chamber or from the cleaning procedure.

VOCs sampling from skin

Immersion in liquid nitrogen was used as a method of euthanasia. Afterward, each specimen was left at room temperature until thawed. Then the complete skin of the specimen was dissected

carefully, placed in a ceramic mortar and homogenized with liquid nitrogen. Finally, the homogenized skin was placed in an empty glass vial of 22 mL. The weight of the skin was registered. For sampling the VOCs of the frog skin, HS-SPME procedure was performed introducing the glass vial in a water bath set at 45 °C, and exposing a DVB/CAR/PDMS SPME fiber (SUPELCO, PA, USA) for 40 min.

VOCs GC-MS analysis

The desorption process for *in vivo* VOCs sampling or for the skin's VOCs was carried out in the GC HP 6890 Series equipped with an Agilent Mass Selective Detector 5973 (Agilent Technologies, Palo Alto, CA, USA) at 250 °C using splitless injection. The separation was performed on a BP-5 capillary GC column (30 m × 0.25 mm × 0.25 µm, SGE, Austin, TX, USA) using helium as a carrier gas at a flow rate of 1 mL/ min. The temperature gradient program started at 40 °C for 3 min, followed by an increase to 100 °C at a rate of 6 °C/min, then the temperature was raised to 200 °C at 4 °C/min, and finally to 300 °C at 20 °C/min, and this temperature was maintained for 3 min. The GC-MS filament source and the quadrupole temperature were set at 230 °C and 150 °C, respectively. The electron ionization (EI) source was set at 70 eV, and the mass spectrometer was operated in full scan mode over a mass range from *m/z* 40 to 300 at a scan rate of 2.0 scan/s. All samples, including linear alkanes, were run under the same chromatographic conditions. Linear alkanes of the series C8–C20 were used for the determination of retention indices (RIs) and later for the tentative assignment of compounds. For validation, 4 biological replicates of the experimental procedure were performed.

To trace contaminants from the vial, a blank run was performed before placing the skins. This blank trial was carried out under identical conditions to those used during the extraction of frog skin samples. Blank runs of the fiber were also performed to trace compounds released by the polymers contained in the fiber.

Methanolic extracts from Dendrobatids frogs from Colombia (dataset # 38, Table S1)

Following the removal of the SPME-fiber from the glass vial, the skin was immediately placed in another glass vial of 4 mL were 1 mL of MeOH 98% was added to extract all the soluble compounds. Vortex (20 s) was applied to each vial and the methanolic extracts were kept at -80°C to avoid degradation until the chromatographic analysis was performed. For the GC-MS analysis, 100 µL of the methanolic extract were combined with 10 µL of Decahydroquinoline at 10 ppm as an internal standard (Sigma, St Louis, MO, USA).

The separation was performed on the same instrument and column used for the skins' VOCs analysis, but using a different temperature program for separation and characterization of amphibian alkaloids from Dendrobatids and facilitate the putative annotation of the alkaloids without having RI standardized for these types of compounds. A split injection was used with the objective of not saturate the column.

The flow rate was set to 1.1 mL/min. The temperature gradient program started at 100 °C for 3 min, followed by an increase to 190 °C at a rate of 10 °C/min and maintaining it by 1 min, then the temperature was raised to 200 °C at 2 °C/min and maintaining it by 1 min, and finally to 280 °C at 10 °C/min, and this temperature was maintained for 3 min. Conditions of the MS were equivalent at the method employed for the VOCs analysis from the skin. All samples, including linear alkanes, were run also under these chromatographic conditions.

Study of volatiles related to beer aging (dataset # 22, Table S1)

Gas Chromatography coupled to time-of-flight mass spectrometry (GC-TOFMS) was used to investigate the aging process of beer to determine time course trends for individual analytes within the aroma profile. A commercially-available IPA beer was purchased and artificially aged by storage at elevated temperatures. A previously reported protocol calibrated that a single day of storage at 40°C was comparable to one month of storage under appropriate conditions[28]. The protocol was used to generate a time course set of samples by storing samples at 40 °C for 0, 1, 2, 3, 4, 5, 6, 8, 12, or 20 d. For each sample, 5 mL of beer were pipetted into a 10 mL glass vial with septum cap. The samples were incubated for 10 min at 60°C. SPME extraction was performed with a DVB/CAR/PDMS fiber (Supelco) for 20 min at 60°C prior to analysis by GC-TOFMS. The analysis was conducted on Gas Chromatograph Agilent 7890B with LECO L-PAL3 autosampler coupled to LECO Pegasus® BT TOFMS. Chromatographic conditions were as follows: Injection: 3 min desorption in 250 ˚C inlet. Carrier Gas: He at 1 mL/min at Constant Flow. Column: Stabilwax, 30 m x 0.25 mm i.d. x 0.25 µm coating (Restek). Oven program: hold 4 min at 35 °C, ramp at 5 °C/min to 180 °C, ramp at 10 °C/min to 220 °C and hold 5 min. Transfer line temperature: 250 °C. Mass spectrometer conditions: Ion Source Temperature 250 °C, electron energy 70 eV, acquisition mass range 35-650 *m/z*, acquisition rate 10 spectra/s.

Investigation of polar compounds produced by Prunus persica after aphid infestation (dataset #30, Table S1)

The metabolic response of peach tree (*Prunus persica)* to green aphid (*Myzus persicae*) has been investigated by GC-MS profiling of the apex polar extracts. Six plants of two cultivars, GF305 (sensitive to *M. persicae*) and Rubira (resistant), were obtained from the germination of seeds after 3 months stratification at 4°C and grown for eight weeks in a green house before transfer in a growth chamber where the experiment took place. Three plants of each genotype were infested with 10 synchronized green aphid female adults (clones MP06). Aphids were removed after 48 h and the plant apex (about 100 mg) were collected in liquid nitrogen and thoroughly ground with a ball mill. Polar compounds were extracted as follow: 10 mg of fresh apex powder were extracted for 15 min. at 70 °C in 1.5 mL of 80% methanol containing 100 µM adonitol as internal standard. The samples were centrifuged, the supernatant was collected and passed through 0.22 µm filters. Then 50 µL of the polar extract was dried overnight under vacuum in glass vials and stored at -20°C before derivatization. GC-MS analysis was performed with an Agilent 7890B gas chromatograph coupled with a LECO Pegasus BT TOFMS. The samples were derivatized online by a Gerstel MPS XT Dual Head robot, according to the following procedure: 50 µL of Pyridine containing 20 mg/mL methoxyamine hydrochloride was added to the dried sample and mixed 1

min at 2000 rpm with a vortex shaker, before incubation 30 min at 80°C under constant shaking at 400 rpm. Then 80 µL of N,O-Bis(trimethylsilyl)trifluoroacetamide (BSTFA) containing a mixture of 9 n-alkanes (C10 to C36) was added and mixed 1 min at 2000 rpm with a vortex shaker before incubation 30 min at 80°C under constant shaking at 400 rpm. Each sample (1 µL) was injected immediately after derivatization was introduced in the injector at 270°C with a 50:1 split-ratio and 1 mL/min helium vector gas flux. The separation was achieved on a Phenomenex Zebron ZB-5MS column (30 m x 0.25 mm x 0.25 µm) with a temperature gradient: 70°C held 1 min., increase up to 220 °C at a rate of 9°C/min, then up to 330 °C at 15 °C/min, held 5 min before cooling. Both the transfer line and the source were set at 250 °C and the electronic impact source filament at 70 eV. Data were acquired after a 275 s solvent delay, at a rate of 10 spectra/s and in a *m/z* range 50 to 630.

Earth microbiome project (dataset #18, Table S1)

The samples were collected during the Earth Microbiome Project (http://earthmicrobiomeproject.org). Two protocols were used: one for the 105 soil samples and another for the 356 fecal samples. 105 soil samples were received at the Pacific Northwest National Laboratory and were processed as follows. Each soil sample (1 g) was weighed into microcentrifuge tubes (Eppendorf Biopur Safe-Lock, 2.0 mL). 1 mL of ddH_2O and one scoop of garnet/SS beads and 1 stainless steel (SS) bead was added to each tube. Samples were homogenized (Qiagen,Tissue Lyser) for 3 min. at 30 Hz. Samples were then transferred into 15 mL tubes (Olympus, polypropylene). Ice-cold water (1 mL) was used to rinse the smaller tube and combined into the 15 mL tube. 10 mL of 2:1 (v/v) chloroform:methanol added to each sample using a glass serological pipette and rotated at 4 °C for 10 min. The samples were then placed inside the -70 °C freezer for 10 min.. The samples were centrifuged at 4,000 rpm for 10 min. to separate phases. The top and bottom layers were combined into 40 mL glass vials and dried down until almost dry and frozen at -20 °C, the protein interlayer and debris were discarded. The following day 1 mL of 2:1 chloroform:methanol was added to each large glass vial and the sample was transferred into 1.5 mL Sorenston tubes, spun at 12,000 g to remove debris. The supernatant was transferred into smaller flat bottom glass MS vials and dried all the way down for analysis.

The remaining 356 samples that were received at UCSD which included fecal and sediment samples were processed as follows. Transferred 100 µL of each pellet to a 2 mL Sorenson microcentrifuge tube using a 100 µL scoop (Next Advance, MSP01). Brought final volume of sample to 1.5 mL ensuring the solvent ratio is 3:8:4 H_2O:$CHCl_3$:MeOH by adding the appropriate volumes of H_2O, MeOH, and $CHCl_3$. After transfer, we added 100 µL of 3 mm stainless steel beads (chloroform washed) to each tube. On the day of extraction, we added 400 µL of chilled methanol and kept the tubes on ice to prevent the sample from refreezing after transfer. Added 300 µL chilled nanopure water and vortexed to mix for 30 s. Added 800 µL chilled chloroform and vortexed to mix for 30 s. Transferred out of BSC wiping with a bleach wipe. Transferred any remaining, unused sample from the hood (wiping with a bleach wipe) directly into liquid nitrogen and stored at -70 °C. The samples were then derivatized for GC-MS analysis as follows. A methoxyamine solution with a pyridine concentration of 30 mg/mL was made. Aqueous metabolites dried briefly (>30 min.) SpeedVac if the samples were stored in a freezer. Performed

methoximation by adding 20 µL of methoxyamine solution to the sample vial and vortexed for 30 s at setting 5 on vortexer. A bath sonicator was used to ensure the sample is completely dissolved. Incubated the sample in the oven maintained at 37 °C for 1 h and 30 min with 1000 rpm shaking. Inverted the vial to mix the samples with condensed drops at the cap surface. Spun the sample down for 1 min. at 1000 g at room temperature in a centrifuge. Performed silylation by adding 80 µL of 1% TMCS in (MSTFA) solution using a syringe and vortex for 10 s. Incubated the sample in the oven maintained at 37 °C for 30 min with 1000 rpm shaking. Inverted the vial to mix the samples with condensed drops at the cap surface. Spun the sample down for 5 min at 2000 g at room temperature in a centrifuge. Transferred the reacted solution into two vials with inserts.

An Agilent 7890A gas chromatograph coupled with a single quadrupole 5975C mass spectrometer (Agilent Technologies, Inc.) was used for all analyses. A HP-5MS column (30 m × 0.25 mm × 0.25 µm; Agilent Technologies) was used for untargeted analysis. Samples (1 µL) were injected in splitless mode, and the helium gas flow rate was determined by the Agilent Retention Time Locking function based on analysis of deuterated myristic acid (Agilent Technologies, Santa Clara, CA). The injection port temperature was held at 250 °C throughout the analysis. The GC oven was held at 60 °C for 1 min after injection, and the temperature was then increased to 325 °C by 10 °C/min, followed by a 10 min hold at 325 °C. Data were collected over the mass range of *m/z* 50–600. A mixture of FAMEs (C8–C28) was analyzed each day with the samples for retention index alignment purposes during subsequent data analysis.

Detection of sterols in urine

All samples were taken from control experiments made on cattle urine to detect the misuse of anabolic agents for livestock fattening. For details see EU Directive (88/146/EEC and 96/22/EC) prohibiting the use, in livestock farming, of certain substances having a hormonal action on animals bred within the European Union. Urine samples were processed with the method described in [29].

GC-MS analysis was performed on a HP 6890 gas chromatograph coupled to a HP 5973 quadrupole mass analyzer (Hewlett-Packard, Palo Alto, CA, USA). A DB-5MS (30 m×0.25 mm i.d., 0.25 µm, Agilent, Palo Alto, CA, USA) column was used. For steroid separation, the following conditions were applied: injector 250 °C, splitless injection (1 min), source temperature (230 °C). The column flow rate was 1.5 mL/min (constant flow). The oven temperature was increased from 60 °C (1.5 min) to 220 °C at 40 °C/min, then to 240 °C (1 min) at 1 °C/min, then to 300 °C (1 min) at 20 °C/min. Mass spectra were recorded in the scan mode (*m/z* 50–550 mass range). The temperature of the transfer line was 280 °C.

Bacillus fungal biocontrol study, mixed cultures of Bacillus subtilis and Setophoma terrestris (dataset #21, **Table S1***)*

Bacterial and fungal strains used in this study are: *Bacillus subtilis* ALBA01 (isolate from onion rhizosphere), variants of *B. subtilis* ALBA01 obtained after co-culture with *S. terrestris*, *Setophoma terrestris* PH06 (fungal strain isolated from onion rhizosphere). Bacterial strains were routinely grown in LB medium (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) at 30ºC.

Biological replicates of *Bacillus subtilis* pre and post-ST variants and of the fungus *S. terrestris* were cultivated individually and in co-cultures in 10 mL-borosilicate vials with a screw cap with

silicon septum containing 3 mL of PDA medium for seven days at 30°C and then analyzed with GC-MS to determine the emitted volatile metabolites. After incubation, the capped vials were stored at -80 °C and thawed immediately prior to analysis. The GC-MS analysis was carried out using the Agilent 7200 GC/QTOF equipped with a robotic sampler system. The separation was conducted on an HP-5MS column (30 m x 0.25 mm x 0.25 µm). The volatiles from the sample were extracted from headspace using Polydimethylsiloxane/Divinylbenzene (PDMS/DVB) df 65 µm Solid Phase Microextraction Fiber (SPME) for 30 min. at 50 °C. The fiber was then inserted into the injector equipped with Merlin septum heated to 250 °C and the adsorbed compounds were desorbed for 1 min. The GC protocol analysis included: cryofocusing on the head of the column at -20 °C for 1 min; 115 °C/min oven ramp to 40 °C (hold of 0.1 min), 20 °C/min oven ramp to 300 °C (hold of 0.1 min), ramp to 320 °C to purge the column. The helium carrier gas was set to constant 2 mL/min flow and a splitless injection mode was applied. The scanned *m*/*z* range was 35-400 with the acquisition rate of 20 spectra/s. The empty vial blanks were interspersed with the samples to assess background signal. Quality controls of natural mint oil extract were run along with samples before and after the analysis.

Measurement of short chain fatty acids in fecal samples from patients with spondyloarthritis before and after treatment with biologic therapy (TNF inhibitor or IL-17A inhibitor; dataset #36, **Table S1***)*
Approximately 50 mg of sample were placed in 2 mL borosilicate vials capped with a screw cap with silicone septum. The capped vials were stored at -20 °C and thawed immediately prior to analysis. The GC-MS analysis was carried out using the Agilent 7200 GC/QTOF equipped with a robotic sampler system. The volatiles from the sample were extracted from headspace using Polydimethylsiloxane/Divinylbenzene (PDMS/DVB) df 65 µm Solid Phase Microextraction Fiber (SPME) for 30 min. at 50 °C. The fiber was then inserted into the injector equipped with Merlin septum heated to 250 °C and the adsorbed compounds were desorbed for 1 min. The separation was conducted on an HP-5MS column (30 m x 0.25 mm x 0.25 µm). The GC protocol analysis included: cryofocusing on the head of the column at -20 °C for 1 min; 115 °C/min oven ramp to 40 °C (hold of 0.1 min), 20 °C/min oven ramp to 300 °C (hold of 0.1 min), and 50 °C/min oven ramp to 320 °C purge the column. The helium carrier gas was set to constant 2 mL/min flow and a splitless injection mode was applied. The scanned *m/z* range was 35-400 with the acquisition rate of 10 spectra/s. The empty vial blanks were interspersed with the samples to assess background signal. Quality controls of natural mint oil extract were run along with samples before the analysis. Samples spiked with a 2.5 µL aliquot of standard mix of short chain fatty acids was analyzed along with the samples. The samples were randomized during the analysis.

Measurement of medium chain fatty acids in fecal samples from patients with spondyloarthritis before and after treatment with biologic therapy (TNFi inhibitor or IL-17A inhibitor; dataset # 37, **Table S1***).*
HPLC grade methanol, MTBSTFA (*N*-*tert*-Butyldimethylsilyl-*N*-methyltrifluoroacetamide with 1% *tert*-Butyldimethylchlorosilane (tBDMCS)]; short chain fatty acids were obtained from Sigma-Aldrich, St. Louis, USA. Post-thawing, human fecal (50 mg) were sequentially extracted using solvent extraction with 100 µL 100 % methanol (HPLC grade) followed by sonication for 10 min. and centrifugation (12000 rpm, 5 min.). Supernatant was then transferred to a new 1.7 mL tube and the samples were lyophilized followed by resuspension in 50 µL MTBSTFA (with 1%

tBDMCS). After sonication (5 min.) and centrifugation (12000 rpm, 10 min.) the prepared samples were placed in a 200 µL insert in 2 mL borosilicate vials and capped with a screw cap with silicone septum. The samples were stored at -20 °C and thawed immediately prior to analysis. The GC-MS analysis was carried out using the Agilent 7200 GC/QTOF equipped with a robotic sampler system. The separation was conducted on an HP-5MS column (30 m x 0.25 mm x 0.25 µm). A 1 µL aliquot of sample was injected with split 1:10, injector set at 250 °C. The GC protocol analysis included: solvent delay of 3.5 min; starting temperature of 50 °C; 30 °C/min oven ramp to 170 °C (hold 1 min), 15 °C/min oven ramp to 300 °C and 15 °C/min oven ramp to 310 °C with 3 min hold to purge the column. The helium carrier gas was set to constant 1.2 mL/min flow. The scan *m/z* range was 35-400 with the acquisition rate of 10 spectra/s. The derivatizing agent and empty vial blanks were interspersed with the samples to assess background signal. Samples spiked with a 2.5 µL aliquot of standard mix of short chain fatty acids was analyzed along with the samples. The samples were randomized during the analysis.
30

Other datasets

The data that have been previously published with methods described in the corresponding manuscript are listed in the **Table S1**: The methods for the datasets #1-10 are described in [3132]. The data are published in: datasets #1(MSV000084033)[30], #2(MSV000084033), #3(MSV000084034)[33], #4(MSV000084036)[34], #5(MSV000084032)[35], #6(MSV000084038)[36], #7(MSV000084042)[36], #8(MSV000084039)[32], #9(MSV000084040)[37], #23(MSV000081340)[38], #24(MSV000084348)[39], #25(MSV000084378)[40], #29(MSV000084350)[41].

Generation of molecular networks

Repository-wide network

The data were collected across multiple studies as described in the **Supplementary Notes**. The datasets #1-38 (**Table S1**) were processed on GNPS MSHub deconvolution workflow as described in the tutorial. The .mgf output file for each dataset was downloaded individually (**Table S1**) and all of the files were combined into a single .mgf using the script available at https://github.com/bittremieux/GNPS_GC/blob/master/src/merge_mgf.py. The combined file was then used as an input for the library search/molecular networking workflow. The graphml network file for the global network was downloaded and imported into Cytoscape software as described in the tutorial. The metadata regarding the datasets that include: Sample introduction mode, Derivatization status, Uberon/sample type, Instrument type, Data type (high resolution vs. low resolution), Data file format and Number of files in the dataset (**Table S1**) were imported into Cytoscape Version: 3.7.1. The **Figures 2, S9, S10** were prepared using Gephi software Version 0.9.2 (tutorial can be found at: http://www.martingrandjean.ch/gephi-introduction/). The graphml with metadata was exported from Cytoscape and imported into Gephi. The network was visualized using the “ForceAtlas2” layout with the following settings: Tolerance 1.0; Approximate Repul off; Scaling 1.0; Stronger Gravity off; Gravity 0.5; Dissuade Hubs on; LinLog mode off; Prevent Overlap on; Edge weight influence 1.0. The size of the node is proportional to the number of nodes that connected.

Library search for merged network:
https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8de1f720fd93476db728caa353f0fe50

Reference data network

Figure S7 was created from the reference spectra as follows. Spectra from the NIST 17 reference library were classified by ClassyFire to create the full chemical ontology for all reference compounds. The pairwise cosine was then calculated for all of the spectra in the library. The reference spectra were filtered by down-selecting classes of interest and imported into the Cytoscape software for visualization.

Skin volatilome reference network and MolNetEnhancer for GC-MS

The **Figure S8** was created from the dataset #19 in **Table S1** as follows. The graphml output of the GNPS library search job was downloaded from :
https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5d7f6a60f1a14126bfc8873231f4e7ca

The MolNetEnhancer workflow was created to provide a high-level chemical overview based on chemical class annotations of molecular families discovered in electrospray ionization-mass spectrometry fragmentation (ESI-MS/MS) data by counting ClassyFire chemical ontology terms of collected candidate structures from compound databases searched by the parent masses of each node in a molecular family. Here, the MolNetEnhancer workflow was adapted to work with GC-MS data as follows: i) the input structures are recognised from InChIs rather than SMILES, and, most importantly ii) a maximum of k candidate structures for each node in a molecular family were retrieved from library matches that were based on the following criteria: the top-ranked (best scoring) matches based on cosine scoring with a minimum of *u* that also passed user-defined Kovats filtering criteria (if available), and finally iii) molecular family information was retrieved from the graphml network file to calculate the most predominant chemical ontology terms per family (i.e., superclass, class, subclass, etc.) before adding the chemical class annotations and ClassyFire[42] scores to the network file. Both *k* (Top Hits Per Spectrum) and *u* (Score Threshold) are defined by the user upon launching the GNPS-GC-MS Library Search and Molecular Networking job. The links to GNPS-GC-MS and MolNetEnhancer scripts are provided under the “Code availability” section.

Comparison of deconvolution tools

Deconvolution accuracy/annotation success

The deconvolution was conducted on the GNPS workflow for the MSHub; MZmine2/ADAP v2.53 and MS-DIAL v4.18 were deployed on a desktop computer with 12 CPUs and 248GB RAM. The deconvolution settings were optimized for MZmine2/ADAP v2.53 and MS-DIAL v4.18 and are given below. MSHub was run “as is”, with no parameters optimization. The deconvolution results from each tool were used as an input for GNPS library search workflow, which were launched with identical search settings and the same set of libraries including public

GNPS libraries and commercial NIST14 and Wiley. The links to workflows are listed in the **Supplemental Table S3**.

Deconvolution settings:

MS-DIAL

Data type	
Data type	Centroid
Ion mode	Positive
Accuracy type	Nominal
Data collection parameters	
Retention time begin	0
Retention time end	100
Mass range begin	0
Mass range end	1000
Data processing	
Number of threads	2
Peak detection parameters	

Smoothing method	LinearWeightedMovingAverage
Smoothing level	3
Average peak width	20
Minimum peak height	1000
Mass slice width	0.5
Mass accuracy	0.5
MS1Dec parameters	
Sigma window value	0.5
Amplitude cut off	1
Alignment parameters setting	
Retention time	RT
Retention index tolerance	20
Retention time tolerance	0.1
EI similarity tolerance	60
Retention time factor	0.5

EI similarity factor	0.5
Peak count filter	0
QC at least filter	FALSE
Remove feature based on peak height fold-change	FALSE
Sample max / blank average	5
Sample average / blank average	5
Keep identified and annotated metabolites	TRUE
Keep removable features and assign the tag for checking	TRUE
Replace true zero values with 1/2 of minimum peak height over all samples	FALSE

MZmine2/ADAP

Mass Detection	
Mass detector	Centroid
Noise level	100
ADAP Chromatogram builder	
Min group size in # of scans	5

Group intensity threshold	1000
Min highest intensity	1000
M/z tolerance	0.001 Da, 0 ppm
Chromatogram deconvolution	
M/z center calculation	Median
Algorithm	Wavelets (ADAP)
S/N threshold	10
S/N estimator	Intensity window SN
Min feature height	1000
Coefficient/Area threshold	0
Peak duration range	0.0 - 0.5
RT wavelet range	0.001 - 0.05
Spectral Deconvolution / Multivariate Curve Resolution	
Deconvolution window width	0.2
Retention time tolerance	0.05

Minimum Number of Peaks	1
Adjust Apex Ret Time	False
ADAP Aligner (GC)	
Min confidence (between 0 and 1)	0.1
Retention time tolerance	0.5 (absolute)
M/z tolerance	0.01 Da, 0 ppm
Score threshold (between 0 and 1)	0.75
Score weight	0.1
Retention time similarity	Retention Time Difference (fast)

Both derivatized and non-derivatized samples were selected for testing (**Supplemental Table S3**, the “GC-MS analysis for validation studies” section above). Considering that a mixture of spiked compounds would be limited and an inadequate representation of true biological complexity, we also selected datasets that are real biological mixtures with curated annotations. One such dataset (MSV000084039) is a derivatized mouse blood serum data described in[43]. These data were curated by the authors of the manuscript as described in[43] to generate a list of 128 compounds, as well as 13 spiked FAME standards. The list of curated annotations for this dataset is given in the **Supplemental Table S6**.

Another real biological dataset is MSV000084349 of beer aging (this data set is used by LECO in the development of the various data processing software tools). The compounds in the dataset were annotated if following requirements were met:

- Good hit in NIST library search, either as high forward similarity score or as high probability in cases of less abundant compounds with compromised spectra that are nevertheless characteristic.

- Retention index matches either the NIST 2017 RI within ±40 units, or in the case of homologues that don't have an RI in NIST 2017, RI is within ±10 units of expectation based on bracketing homologues.

- Mass accuracy of main spectral signals within ±15 mDa of expectation, following mass recalibration of the acquired sample files, using siloxane and ubiquitous background ion masses. ±15 mDa value was chosen as it excludes O vs. CH_4 exchange, and is possible for the Pegasus BT system (GC-TOFMS with resolving power about 1500) on which the data were collected. For $m/z > 300$ this requirement is excessive for the Pegasus BT system, but is mitigated by a small number of analytes with such high mass in the beer headspace.

- Reasonable expectation of presence in beer, based on previous reports of the compound, compound class, or a chemical precursor or product in any kind of foodstuff.

The resultant curated list of 352 analytes generated using the above criteria is given in the **Supplemental Table S7**.

The results of GNPS library search jobs for each tool were downloaded (the links are given in **Supplemental Table S3**). Sparse patterns containing fewer than 10 peaks and poor matches with cosine below 0.7 were removed. In order to avoid ambiguities due to different possible names for the same compound, InChIKey was generated and matched for each annotation. The annotation was considered "correct" if it matched a compound on the corresponding curated list (**Supplemental Tables S6** and **S7**). It is important to point out that in the cases of actual biological mixtures, that analytes in the sample cannot be called "confirmed". To call an analyte "confirmed" or "identified" (or anything similarly unequivocal) requires coelution with an independently authenticated pure reference material and that the normalized spectrum remains identical on standard addition. Even then, the "identity" can be questioned, especially if enantiomers, diastereomers, or structural isomers with indistinguishable physicochemical descriptors are possible. Failing the above requirements, including where necessary a demonstration that other possible isomers could be separated, analytes are "assigned" or "provisionally identified". In this regard, although the analyte assignments are as confident as possible for the state of the art GC-MS experiment, there certainly exists a possibility of erroneous annotation, especially where multiple isomers are reasonable candidate assignments. However, it is reasonable to be certain that for every analyte on the list, there is an actual, at least related compound that possesses the listed quantitative mass. To reflect this ambiguity, we also demonstrate annotation accuracy at the level of chemical subclass, as reported by ClassyFire[42], instead of direct match (**Supplemental Figure S3**).

Processing time

The MSHub, MZmine2/ADAP and MS-DIAL tools were tested with respect to the number of files that can be processed and processing time (deconvolution and alignment of features). For the test, the data were split into subsets of an increasing number of files (**Supplemental Table S4**). The settings of MZmine2/ADAP and MS-DIAL were optimized as above and kept the same for

all subsets. The three tools were deployed and run on the same system with 12 CPUs and 248 GB RAM and the time from launching the analysis till completion was recorded; the resulting times are plots are shown on **Figure S4**. MS-DIAL and MZMine failed once the dataset reaches a certain size; for the system used in the test this failure occurred at 502 files for MZmine and 75 files for MS-DIAL. MSHub successfully completed processing of all tested data.

Generation of plots

Figure 1 k,l: Library search results and GNPS task metadata for the datasets Test1-Test11 (**Table S1**) were retrieved from GNPS (job links are listed in **Table S1**). Compound identifications in the InChi format were converted to canonical SMILES strings using RDKit and compared to a list of spiked-in compounds represented as SMILES strings. Cosine score densities for all identified compounds (**Figure 1k**) or for the identified compounds matching the spiked-in compounds (**Figure 1l**) for an increasing number of input files included in the analysis were plotted using the highest ranked identification for each scan in each run.

Figure 2 m,n: Library search results for the dataset Test11 (**Table S1**) were retrieved from GNPS (job links are listed in **Table S1**). The FDR was computed by considering the scans for which any of its top 10 identifications (based on cosine score) that matched the sub-classes of the spiked-in compounds, i.e. fatty acid esters or alkanes, as defined by ClassyFire as true positives, whereas scans for which this was not the case were considered false positives. The FDR was calculated at different cosine score thresholds ranging from 0.5 to 1 and balance score thresholds ranging from 0 to 100.

Figure 2o: Library search results for the dataset Test11 (**Table S1**) were retrieved from GNPS (job links are listed in **Table S1**). Compound identifications in the InChi format were converted to canonical SMILES strings using RDKit and compared to a list of spiked-in compounds represented as SMILES strings. The highest ranked identification for each scan in each run was considered, and the number of identifications matching the spiked-in compounds was calculated at different cosine score thresholds ranging from 0.5 to 1 and balance score thresholds ranging from 0 to 100.

Figure 2p,q: Library search results and GNPS task metadata for the datasets ICL1-ICL11

(**Figure 2p**) and the UCD1-UCD16 (**Figure 2q**) were retrieved from GNPS (job links are listed in **Table S1**). Compound identifications in the InChi format were converted to canonical SMILES strings using RDKit. Cosine score densities for an increasing number of files included in the analysis were plotted using the highest ranked identification for each scan in each run.

Figure 1r,s: Library search results and GNPS task metadata for the datasets ICL1-ICL11 (**Figure 2r**) and UCD1-UCD16 (**Figure 2s**) were retrieved from GNPS (job links are listed in **Table S1**). Compound identifications in the InChi format were converted to canonical SMILES strings using RDKit. The highest ranked identification for each scan in each run was considered,

and the number of unique compounds at different cosine score thresholds ranging from 0.5 to 1 was calculated for an increasing number of files included in the analysis.

Supplemental Figure S1: GNPS task metadata for the UCD1-UCD16 were retrieved from GNPS (job links are listed in **Table S1**). The deconvolution runtime (a) and the deconvolution runtime per file (b) were plotted for an increasing number of the files in the analyzed dataset.

Supplemental Figure S3: GNPS task metadata were retrieved from GNPS (job links are listed in the **Table S3**).

Supplemental Figure S4: GNPS task metadata were retrieved from GNPS (job links are listed in the **Table S4**).

Supplementary Notes References

1. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* vol. 26 51–78 (2007).
2. Likić, V. A. Extraction of pure components from overlapped signals in gas chromatography-mass spectrometry (GC-MS). *BioData Mining* vol. 2 (2009).
3. Li, Z. *et al.* Morphological weighted penalized least squares for background correction. *The Analyst* vol. 138 4483 (2013).
4. Zhang, J., Gonzalez, E., Hestilow, T., Haskins, W. & Huang, Y. Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry. *Current Genomics* vol. 10 388–401 (2009).
5. Veselkov, K. A. *et al.* Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Anal. Chem.* **81**, 56–66 (2009).
6. Gu, Q. & Veselkov, K. Bi-clustering of metabolic data using matrix factorization tools. *Methods* vol. 151 12–20 (2018).
7. Veselkov, K. *et al.* BASIS: High-performance bioinformatics platform for processing of large-scale mass spectrometry imaging data in chemically augmented histology. *Scientific Reports* vol. 8 (2018).
8. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry* vol. 84 5035–5039 (2012).
9. 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* vol. 11 (2010).
10. Wenig, P. & Odermatt, J. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. *BMC Bioinformatics* **11**, 405 (2010).

11. Stein, S. E. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* vol. 10 770–781 (1999).
12. Smirnov, A. *et al.* ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve Resolution to Spectral Deconvolution of Gas Chromatography–Mass Spectrometry Metabolomics Data. *Analytical Chemistry* vol. 91 9069–9077 (2019).
13. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
14. Skogerson, K., Wohlgemuth, G., Barupal, D. K. & Fiehn, O. The volatile compound BinBase mass spectral database. *BMC Bioinformatics* **12**, 321 (2011).
15. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**, 504 (2008).
16. Lommen, A. & Kools, H. J. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics* **8**, 719–726 (2012).
17. He, Q. P., Wang, J., Mobley, J. A., Richman, J. & Grizzle, W. E. Self-calibrated warping for mass spectra alignment. *Cancer Inform.* **10**, 65–82 (2011).
18. Styczynski, M. P. *et al.* Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. *Anal. Chem.* **79**, 966–973 (2007).
19. Amigo, J. M., Skov, T., Bro, R., Coello, J. & Maspoch, S. Solving GC-MS problems with PARAFAC2. *TrAC Trends in Analytical Chemistry* vol. 27 714–725 (2008).
20. Kessler, N. *et al.* MeltDB 2.0-advances of the metabolomics software system. *Bioinformatics* **29**, 2452–2459 (2013).
21. Domingo-Almenara, X. *et al.* eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. *Analytical Chemistry* vol. 88 9821–9829 (2016).
22. Kind, T. *et al.* FiehnLib: mass spectral and retention index libraries for metabolomics based

on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **81**, 10038–10048 (2009).

23. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
24. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
25. Doran, S. L. F., Romano, A. & Hanna, G. B. Optimisation of sampling parameters for standardised exhaled breath sampling. *Journal of Breath Research* vol. 12 016007 (2017).
26. Li, H., Jiang, T. & Zhang, K. Efficient and Robust Feature Extraction by Maximum Margin Criterion. *IEEE Transactions on Neural Networks* vol. 17 157–165 (2006).
27. Bouslimani, A. *et al.* Molecular cartography of the human skin surface in 3D. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E2120–9 (2015).
28. Marques, L., Espinosa, M. H., Andrews, W. & Foster, R. T. Advancing Flavor Stability Improvements in Different Beer Types Using Novel Electron Paramagnetic Resonance Area and Forced Beer Aging Methods. *Journal of the American Society of Brewing Chemists* vol. 75 35–40 (2017).
29. Buisson, C. *et al.* Application of stable carbon isotope analysis to the detection of 17β-estradiol administration to cattle. *Journal of Chromatography A* vol. 1093 69–80 (2005).
30. Olmstead, K. I. *et al.* Insulin induces a shift in lipid and primary carbon metabolites in a model of fasting-induced insulin resistance. *Metabolomics* vol. 13 (2017).
31. Fiehn, O. Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Current Protocols in Molecular Biology* 30.4.1–30.4.32 (2016) doi:10.1002/0471142727.mb3004s114.
32. Barupal, D. K. *et al.* A Comprehensive Plasma Metabolomics Dataset for a Cohort of Mouse Knockouts within the International Mouse Phenotyping Consortium. *Metabolites* **9**, (2019).

33. Fahrmann, J. F. *et al.* Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer. *Cancer Epidemiol. Biomarkers Prev.* **24**, 1716–1723 (2015).
34. Miyamoto, S. *et al.* Systemic Metabolomic Changes in Blood Samples of Lung Cancer Patients Identified by Gas Chromatography Time-of-Flight Mass Spectrometry. *Metabolites* **5**, 192–210 (2015).
35. Wikoff, W. R. *et al.* Diacetylspermine Is a Novel Prediagnostic Serum Biomarker for Non-Small-Cell Lung Cancer and Has Additive Performance With Pro-Surfactant Protein B. *J. Clin. Oncol.* **33**, 3880–3886 (2015).
36. Liesenfeld, D. B. *et al.* Metabolomics and transcriptomics identify pathway differences between visceral and subcutaneous adipose tissue in colorectal cancer patients: the ColoCare study. *Am. J. Clin. Nutr.* **102**, 433–443 (2015).
37. Nagy-Szakal, D. *et al.* Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. *Sci. Rep.* **8**, 10056 (2018).
38. Quinn, R. A. *et al.* Neutrophilic proteolysis in the cystic fibrosis lung correlates with a pathogenic microbiome. *Microbiome* **7**, 23 (2019).
39. Quinn, R. A. *et al.* Niche partitioning of a pathogenic microbiome driven by chemical gradients. *Sci Adv* **4**, eaau1908 (2018).
40. Bouslimani, A. *et al.* Lifestyle chemistries from phones for individual profiling. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7645–E7654 (2016).
41. Obata, T., Florian, A., Timm, S., Bauwe, H. & Fernie, A. R. On the metabolic interactions of (photo)respiration. *J. Exp. Bot.* **67**, 3003–3014 (2016).
42. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
43. Barupal, D. K. *et al.* A Comprehensive Plasma Metabolomics Dataset for a Cohort of Mouse Knockouts within the International Mouse Phenotyping Consortium. *Metabolites* **9**, (2019).