

Modular Open-access and Open-source julia language Toolbox for Processing of HRMS Data: jHRMSToolBox

Denice van Herwerden,^{*,†} Etienne Kant,[†] Miranda Jackson,[‡] Chloe L. Fender,[‡]
Manuel Garcia-Jaramillo,[‡] Jake W. O'Brien,^{¶,†} Kevin V. Thomas,[¶] and Saer
Samanipour^{*,†,§}

[†]*Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam,
Amsterdam, 1098 XH, the Netherlands*

[‡]*Department of Environmental and Molecular Toxicology, Oregon State University,
Corvallis, OR, USA*

[¶]*Queensland Alliance for Environmental Health Sciences (QAEHS), The University of
Queensland, Brisbane, QLD 4102, Australia*

[§]*UvA Data Science Center, University of Amsterdam, Amsterdam, 1012 WP, the
Netherlands*

E-mail: d.vanherwerden@uva.nl; s.samanipour@uva.nl

Abstract

1
2 There is a growing need for understanding the exposome chemical space. Non-target
3 analysis is, generally, used for the analysis of the thousand of known and unknown
4 chemicals in environmentally and biologically relevant samples. However, algorithm
5 limitations arise with regard to flexibility and suitability for the processing of such

6 data. Hence, the modular open-access and open-source jHRMS toolbox was devel-
7 oped, providing both a user-interface and the freedom to modify and add workflows as
8 required. The default implemented algorithms have been developed for high-resolution
9 mass spectrometry data and can handle MS1 and various data-dependent and data-
10 independent analysis data types both in profile and centroided formats. Moreover, the
11 identification algorithm provides extensive match quality reporting. Besides the data
12 processing workflow, the toolbox comes with built in post processing (i.e., visualiza-
13 tion) for individual steps of the workflow and statistical analysis. Finally, the results
14 are reported step-by-step, parameters can be saved, and it is operating system agnostic.
15 To showcase the potential of the jHRMS toolbox, two datasets from different origins
16 environmental and biological were analyzed and reported. For the environmental case
17 study the trends of some pharmaceuticals in river waters were evaluated. While for
18 the biological samples it was possible to differentiate between liver and brain tissues
19 based on the extracted information.

20 Introduction

21 There is a growing importance of understanding chemicals (i.e., anthropogenic and naturally
22 produced) in environmental and biological samples, which can be referred to as the expo-
23 some chemical space.¹⁻⁵ Depending on the method used to analyze such samples, a different
24 subspace (i.e., region) of the exposome chemical space is covered.^{2,4,5} The method used are
25 impacted by various aspects including: sample preparation, experimental analysis setup, and
26 data processing, implying that the remainder of the exposome space is excluded that can
27 contain highly exposome relevant or toxic chemical.^{2,4,5} Hence, variety in analysis methods
28 is required to (un)cover as much as possible of the exposome chemical space, including the
29 data processing side of the workflow.

30
31 To obtain as much information as possible on the thousands of chemicals that may be

32 present in samples (e.g., biological or environmental), non-target analysis (NTA) utilizing
33 liquid or gas chromatography (LC or GC) coupled with a high-resolution mass spectrome-
34 ter (HRMS) is a commonly used technique.^{4,6-13} Here data-dependent analysis (DDA) and
35 data-independent analysis (DIA) are often used to obtain the related MS2 information (i.e.,
36 fragments) for LC-HRMS.⁶ DDA data generally has less overlap and cleaner MS2 spectra
37 but does not analyze all MS1 signals, as only ions of interest are further fragmented. On the
38 other hand, DIA aims to analyze and fragment all MS1 signals, obtaining generally complex
39 MS2 spectra that can come from overlapping compounds and could obscure low intensity
40 compounds. Through data processing the information is generally extracted by performing
41 feature detection, componentization where information from unique chemical constituents
42 is grouped (i.e., parent, isotopologue, adduct, and (in-source) fragment ions), and identi-
43 fication.⁶ Besides the measurement parameters that influence the covered chemical space,
44 adequate data processing techniques that can handle the complex data and do not further
45 limit the detectable chemical space are required.^{2,4} For example, the algorithms should not
46 be compound class specific as it is generally unknown what the sample is comprised of with
47 NTA experiments.

48

49 It is extremely difficult to objectively compare the available algorithms, as the ground
50 truth is almost impossible to obtain for most NTA data.¹⁴ For example, feature detection
51 algorithms are often compared by investigating the overlap of detected peaks regardless of
52 the peak quality or true positive peaks.¹⁴⁻¹⁶ While a feature detected by multiple algorithms
53 may be more likely to be true, it is still not certain that the signal is truly coming from a
54 chemical or the background. Moreover, this is often only the first step in the full workflow
55 that can influence outcomes further in the workflow.¹⁴ Therefore, increased flexibility and
56 freedom in data processing workflows and algorithms is needed to extract as much relevant
57 information of the subspace as possible. This is similar to varying the measuring methods
58 of NTA approaches to cover a larger region of the exposome chemical space.⁴

59

60 To process LC/GC-HRMS data, a variety of tools have been developed including CAM-
61 ERA,¹⁷ DSFP,¹⁸ FOR-IDENT,¹⁹ GNPS,²⁰ InSpectra,²¹ MS-Dial,²² MZmine,²³ OpenMS,²⁴
62 Patroon,^{25,26} Phenomenal,²⁷ SIRIUS,²⁸ TidyMass,²⁹ and XCMS.³⁰ However, the majority
63 have been developed with the focus on metabolomics applications.^{17–20,22,25–31} The difficulty
64 here is that in NTA a broad range of chemical classes can be found besides metabolites.
65 On the other hand, commercial software generally limits the user with the options provided
66 by the program and can only process vendor specific data formats.³¹ Moreover, the closed
67 source code makes it difficult to understand what happens with the data in case there is, for
68 example, loss of information (i.e., reduced subspace). This leaves only a few options non-
69 vendor, open-source, and open-access software options: InSpectra,²¹ MS-Dial,²² MZmine,²³
70 patRoön,²⁵ and OpenMS.²⁴

71

72 One of the data processing limitations in these software packages is the heavy focus on
73 DDA.³² Where only a part of the MS1 information is further fragmented in MS2. Often
74 focusing on either known precursor masses of interest or intense peaks. Meanwhile, for
75 NTA, DIA is a valuable approach as this focuses more on the unknown compounds. Here
76 all MS1 signals are further fragmented to obtain fragmentation information. While the self-
77 adjusting feature detection algorithm (SAFD)³³ and CompCreate for componentization are
78 implemented in InSpectra, this platform lacks visualization options, a front-end, and the
79 possibility to process highly confidential data as it needs to be uploaded. Meanwhile, SAFD
80 and CompCreate enable the possibility to perform feature detection on both centroided and
81 profile data and componentization on various MS1 and MS2 data types, of which the latter
82 is specifically valuable with DIA data types.

83

84 In this paper we introduce and showcase the jHRMS toolbox with its functionalities
85 and capabilities. The implemented algorithms have generally been optimized and tested for

86 small molecules (i.e., ≤ 1000 Da), providing a broad general application range. The tool-
87 box provides numerous NTA HRMS data processing workflows for a variety of data types,
88 including MS1 only data, DDA, parallel reaction monitoring (PRM), DIA, sequential win-
89 dowed acquisition of all theoretical fragment (SWATH), and multi-collision analysis. The
90 jHRMS toolbox is designed to be highly modular with the ease-of-use through a graphical
91 user interface, while maintaining the complete freedom to add processing workflows and
92 functionalities. To enable this, the toolbox is fully open-access, open-source, and functional
93 on windows, MacOS, and Linux systems. It has been written in the programming language
94 Julia, which is known for the balance between ease of use for data processing and its similar
95 computing performance of low-level languages like C, making it highly suitable for processing
96 HRMS data. Additionally, the toolbox comes with numerous post-processing visualization
97 options and built-in trend and statistical analysis.

98

99 **Experimental Section**

100 To showcase the jHRMS toolbox, two datasets have been used acquired on different instru-
101 ments, in different labs covering both environmental and biological matrices. One dataset
102 comprises of surface water samples measured on an Orbitrap instrument in centroided mode³⁴
103 while the other dataset dealt with two biological matrices measured on a ZenoTOF in profile
104 model. Brief details on the datasets are provided below as the aim of the paper is to showcase
105 the jHRMS toolbox without going too much in depth of what is found in the data. These
106 two different data sets have specifically been chosen to show the compatibility of the toolbox
107 with different types of measured data and applications' purposes. Furthermore, this section
108 contains the processing details and parameters used for showcasing the toolbox capabilities
109 and the code availability of the jHRMS toolbox.

110

111 **Centroided Dataset**

112 The centroided dataset is comprised of surface water samples coming from different rivers
113 collected at 4 week intervals.³⁴ This data has been made publicly available and can be
114 obtained from the MassIVE repository: <https://massive.ucsd.edu/ProteoSAFe/data>
115 [set.jsp?accession=MSV000087190](https://massive.ucsd.edu/ProteoSAFe/data). In brief, the collected samples were extracted using
116 solid phase extraction and analyzed with DDA LC-HRMS with a reversed phase column
117 selectivity.³⁴ For the mobile phase a mixture of water with 0.1% formic acid (A) and methanol
118 (B) was used. The gradient started with 90/10 A/B for 2 minutes, 0/100 at 15 minutes,
119 0/100 at 20 minutes, 90/10 from 21 to 30 minutes, using a flow rate of 0.2 mL/min. As for
120 the DDA MS part, the samples measured with positive electrospray ionization were used.
121 Here a scan range of 60-900 m/z, a MS1 resolution of 120,000 at 200 m/z, a maximum
122 injection time of 70 ms, and an automatic gain control target of 1.0×10^6 . For the top 5
123 data dependent analysis scans a resolution of 30,000 at 200 m/z, maximum injection time of
124 70 ms, 1.0 Da isolation window, and 30 (N)CE were used. For further details on instrumental
125 settings and sample preparation see the citation.³⁴ For the showcase of workflow I, only the
126 measurements acquired in positive mode of the 4 most frequently measured locations were
127 used, which were location 1 to 4. Location 1 came from south west Luxembourg, location
128 2 and 3 from the middle, and location 4 from the east of Luxembourg. Each location was
129 sampled 10 to 11 times in a time span of April 2019 till September 2020.

130 **Profile Dataset**

131 The profile dataset is comprised of liver and brain tissue samples of the salmonid species
132 Chinook (*Oncorhynchus tshawytscha*). This data has been made publicly available and can
133 be obtained from the Metabolomics Workbench repository: ST004904.

134

135 Fish husbandry and euthanasia were performed in accordance with the Standard Guide
136 for Conducting Acute Toxicity Tests on Test Materials with Fishes, Macroinvertebrates, and

137 Amphibians 1 (ASTM, 2014). These methods were approved by Oregon State University
138 Institutional Animal Care and Use Committee (IACUC-2022-0260). Fish were housed in
139 the Aquatic Animal Health Laboratory (AAHL) at Oregon State University (OSU). Fish
140 were acclimated to AAHL holding conditions in 100 L constant flow-through tanks contain-
141 ing ambient well water (16°C) and constant oxygen supply. Fish were fed a commercial
142 salmonid feed at a daily rate of 1% body weight daily during acclimation until the average
143 weight of approximately 3.5 g was achieved as a target experimental weight, including the
144 un-fed control group. Each tank included five fish replicates and each condition was repeated
145 to collect three biological replicates. Fish were sacrificed using MS-222 (Tricaine mesylate
146 powder) + 50 g/L bicarbonate. In this study, the measurements from the fed and un-fed
147 fish were used. The un-fed fish were no longer fed 24 hour prior to the euthanasia.

148

149 Fish liver and brain tissues were dissected, weighed, and flash frozen in aluminum foil
150 packets using liquid nitrogen and stored at - 80°C. Liver and brain tissue samples (10 mg)
151 were aliquoted in 2 mL screw cap vials prefilled with 40 μ L volume of 1.4 mm ceramic beads.
152 Chilled methanol:water, 80:20, was added to the vials (300 μ L). Samples were spiked prior
153 extraction with a mixture of isotope labeled metabolites (Mix 2 QReSS Kit, Cambridge Iso-
154 tope Labs, Tewksbury, MA) to account for extraction variability among samples. Samples
155 were extracted in a Precellys homogenizer (3x15s; 5,500 rpm). Homogenized tissue (200
156 μ L) was transferred to a clean Eppendorf vial. Samples were centrifuged at 10,000g for 3
157 min at 4°C. Samples were stored overnight at - 24°C to allow for precipitation of remaining
158 proteins. Samples were centrifuged again at 13,000g for 15 min at 4°C. Supernatant (150
159 μ L) was recovered in LCMS glass vials with 300 μ L insert, spiked with a mixture of isotope
160 labeled metabolites (Mix 1 QReSS Kit, Cambridge Isotope Labs) and stored at 4°C before
161 analysis. Mix 1 internal standard mixture was used to check for injection accuracy and
162 platform performance due to the large number of samples and extended batch run time.

163

164 Non-targeted UPLC–HRMS/MS analyses were performed using a previously published
165 method with minor modifications. Briefly, data-dependent acquisition in the positive ion
166 mode was conducted on a Sciex ZenoTOF 7600 mass spectrometer (AB SCIEX, Concord,
167 Canada) coupled to an ultra-high performance liquid chromatography system (Sciex Ex-
168 ionLC AD). Chromatographic separation was performed on an Inertsil Phenyl-3 column (2.1
169 x 150 mm, Intersil Ph-3 column, GL Sciences, Torrance, CA) held at 40°C. A gradient with
170 two mobile phases was used: (A) water (LC-MS grade) with 0.1% v/v formic acid; (B)
171 methanol (LC-MS grade) with 0.1% v/v formic acid, using a flow rate of 0.3 mL/min. The
172 injection volume was 2 μ L. Samples were analyzed in a fully randomized batch. The ion
173 spray voltage was set at 4,500 V and the source temperature was 500°C. Period cycle time
174 was 641 ms; accumulation time 80 ms; m/z scan range 50 - 1200 Da. The collision energy was
175 set at 35 V with a collision energy spread setting of 15 V, and a declustering potential of 80 V.

176

177 The mass calibration was automatically performed every 10 injections using a positive
178 calibration solution (AB SCIEX) via a calibration delivery system (CDS). Quality control
179 was assured by (i) randomization of the sequence, (ii) injection of QC pool samples at the
180 beginning and the end of the sequence and between each 10 actual samples, (iii) procedure
181 blank analysis, and (iv) checking the peak shape and the intensity of spiked extraction in-
182 ternal standards (Mix 2 QReSS Kit) and the internal standard added prior to injection (Mix
183 1 QReSS Kit).

184

185 **Showcase Workflows Settings**

186 **Workflow I**

187 The centroided data (i.e. river water samples measured via Orbitrap) was processed ac-
188 cording to the steps shown in figure 1. The data was first screened for potential suspects,
189 which were obtained from the publication of the dataset.³⁴ This list comprised of 816 phar-

190 maceutical compounds from which the InChIKeys were used to setup a suspect list through
191 the jHRMS toolbox (Section S1.5). This generated suspect list contained the positive and
192 negative merged spectra, when found in the database, for each of the 816 compounds. The
193 merged part entails that, if a compound has multiple spectra for an ionization mode, the MS2
194 information will be combined into a single entry. With this suspect list, suspect screening
195 from the Universal Library Search Algorithm (ULSA) was performed with a mass tolerance
196 of 0.05 Da, a minimum precursor intensity of 1000 counts, a retention width of 0.05 min-
197 utes, and a maximum isotopic tree depth of 5. These parameters were selected based on
198 previously processed data sets with suspect screening.^{35,36} Details on the algorithm can be
199 found in section S1.5. Subsequently, the suspect presence across samples was obtained by
200 performing suspect screening alignment (Section S1.6). Here, based on a set of criteria, the
201 best matches for each suspect are obtained across all the samples. The set criteria where a
202 minimum precursor intensity of 10,000 counts and a minimum of 1 detected fragment. For
203 this case, due to the lack of a retention time and a unknown/non-optimized match factor
204 tolerance, the retention time tolerance and minimum match factor criteria were disregarded.

205

206 On the other hand, to investigate and validate the suspect cases found during the suspect
207 screening part, a full identification workflow was executed. Therefore, feature detection
208 (Section S1.2) was first performed on the centroided dataset, using 10,000 iterations, a
209 maximum peak width of 100 scans, a resolution of 50,000, a minimum mass peak width of
210 0.01 m/z, a correlation threshold of 0.8, a minimum intensity of 100 counts, an increasing
211 signal threshold of 5%, a signal to background ratio of 2, a minimum peak width of 3 seconds,
212 and the m/z peak width was estimated using the best guess method (i.e., based on the m/z
213 and resolution). These parameters have been selected based on optimal parameters used
214 with previously processed data sets.^{13,33,36,37} Additionally, an in-depth explanation of the
215 parameters can be found in the supporting information (Section S1.2). As for the peak
216 width estimation, generally, the random forest model would be the best option to use for

217 the peak width. However, this model has only been trained on time of flight data, making it
218 unsuitable for orbitrap data. After feature detection, CompCreate (Section S1.3) was used
219 to perform componentization with a mass window percentage of 0.75%, a retention window
220 percentage of 0.5%, a correlation threshold of 0.8, and a minimum MS2 intensity of 50
221 counts. Finally, identification was performed through ULSA (Section S1.4) with an external
222 database comprised of MassBank EU,³⁸ MassBank of North America,³⁹ and the NITS20
223 database.⁴⁰ For this step, the used settings were to only use positive mode ESI spectra from
224 the database and the scoring was performed with equal weights (i.e., a weight of 1 for each
225 of the 7 scoring parameters).

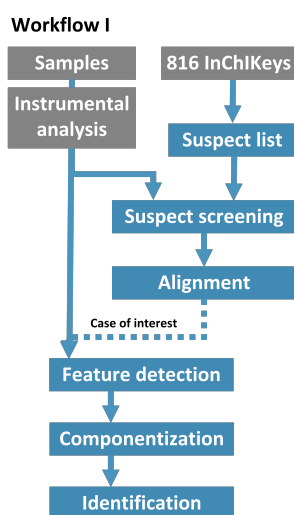


Figure 1: Overview of workflow I and the processing steps taken.

226 **Workflow II**

227 The second workflow was used for the profile dataset (i.e. biological tissues measured via a
228 ZenoTOF instrument). Here a combination of feature detection (i.e., SAFD), componentiza-
229 tion (i.e., CompCreate), alignment, clean-up, and hierarchical clustering analysis was used to
230 show the difference between liver and brain tissues from the Chinook fish (Figure 2). First,
231 feature detection with SAFD (Section S1.2) was performed on the liver, brain, and blank
232 samples, using 10000 iterations, 100 scans maximum peak width in the time domain, 20000

233 resolution, 0.02 minimum m/z window, 0.8 correlation threshold, 150 minimum intensity,
234 5% signal increment threshold, signal to background ratio of 2, and 1 second of minimum
235 peak width in the time domain.^{13,33,36,37} Second, CompCreate (Section S1.3) was performed,
236 using a 0.8 correlation threshold, 0.5 retention window percentage, 0.8 mass window per-
237 centage, and 50 minimum intensity. Next, both the feature and component lists were aligned
238 using the same principle (Section S1.6). In other words the liver and brain files were aligned
239 separately with their corresponding blank files, using a m/z tolerance of 0.005 Da and a time
240 tolerance of 0.1 minutes. This allowed to use the blank filtering functions (Section S1.7) to
241 filter the matrix specific blank features from the liver and brain samples, respectively. For
242 this, the mean blank signal was used with a signal to blank ratio of 5. Then, based on the
243 blank filtering information, the individual feature and component lists were filtered. These
244 filtered lists were then used for the second alignment where the liver and brain samples were
245 combined, using again a m/z tolerance of 0.005 Da and a time tolerance of 0.1 minutes. This
246 resulted in a aligned feature file and a aligned component file of all the samples (i.e., liver
247 and brain). Finally, hierarchical cluster analysis was performed on the two aligned lists.

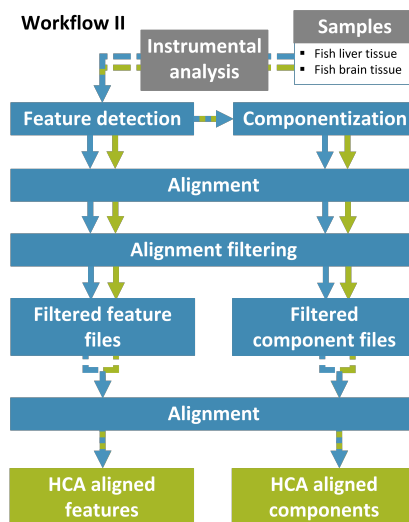


Figure 2: Overview of workflow II and the processing steps taken.

248 **Calculations and Code Availability**

249 The jHRMS toolbox has been developed and tested on a personal computer with 12 CPUs
250 and 32 GB of RAM, using Windows 10. The jHRMS toolbox was developed with the Julia
251 programming language (v1.6). The source code, installation manual, and basics on how to
252 use the toolbox is available at: https://bitbucket.org/Denice_van_Herwerden/jhrms
253 `toolbox.jl/src/main/`. This package contains the functions related to the graphical user
254 interface, visualization options, and statistical analysis. Whereas the function related to the
255 processing of the ‘raw’ data can be found in the description of their respective functions (SI
256 ‘Individual Algorithm Descriptions’).

257

258 **Results and discussion**

259 **Modular Workflows**

260 One of the main advantages of the jHRMS toolbox is the combination of the modular im-
261 plementation of the workflow steps and full access to all the implemented algorithms. The
262 latter will be further discussed after the advantages of the implemented algorithms. Figure
263 3 shows the currently implemented algorithms in the modular workflow. It should be noted
264 that various types of data can be analyzed, including MS1, DDA, PRM, DIA, SWATH, and
265 multi-collision analysis. These data types can also be combined during the alignment.

266

267 As a basis, the workflow contains both the generally used suspect screening and iden-
268 tification workflows. In addition to this, when working on larger data sets, it is possible
269 to perform alignment on the features, components, or suspects. The most straight forward
270 option that this enables is the possibility to perform trend and statistical analysis. However,
271 with the combination of the alignment clean-up functions, this allows for even more flexible

272 workflows. For example, if there is a dataset with 2 different sample types that each have
273 their own method blanks, feature detection followed by alignment of the feature files for each
274 of the sample type could be performed. The alignment clean-up then enables to filter the
275 features from the method blanks for each sample type. From this point, it is possible to
276 filter the individual feature list based on the aligned file and only maintain the features that
277 belong to each sample type. From this, filtered feature lists are obtained that provide the
278 possibility to continue with componentization using these reduced lists. Throughout this
279 process the changes are tracked and deleted features can be restored. Overall, this means
280 that the setup of the toolbox allows to only process information of interest further down
281 the pipeline. This functionality already provides 15+ workflows excluding all the alignment
282 clean-up options.

283

284 The workflow also shows a future option for the use of predictive models in between
285 the aligned components and trend and statistical analysis. As more models are developed
286 that use spectral information for the prediction of, for example, toxicity and ionization
287 efficiencies.⁴¹⁻⁴⁴ Additional functions that predict these values prior to trend analysis can be
288 implemented. The overall workflow and collection of functions can expand over time. The
289 current overview of the specific functions and their names in the toolbox can be found in
290 Figure S1.

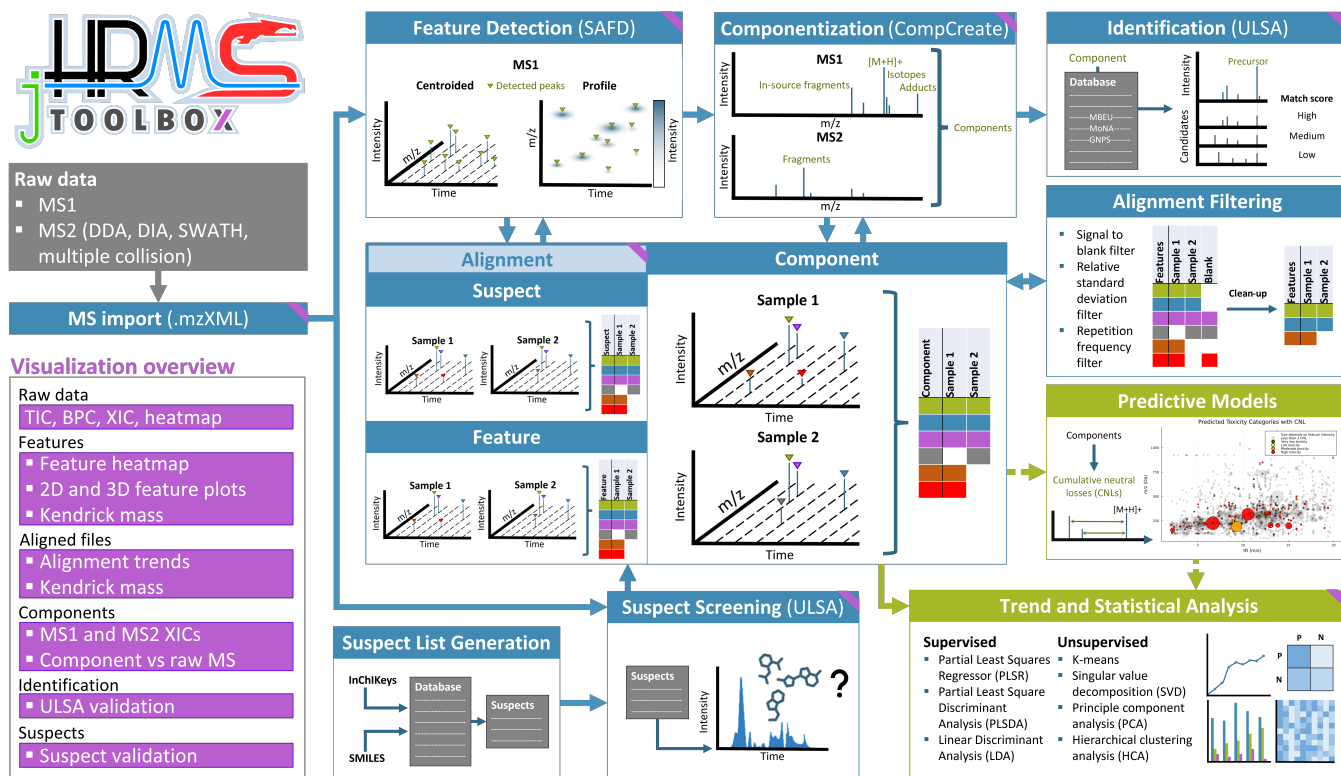


Figure 3: Overview of the jHRMS toolbox modular workflow. In blue are the algorithms and functions related to the HRMS data processing steps, in purple is the overview of the post-processing visualization where the purple ribbons refer to the post-processing visualization, and in green the possible predictive models and the trend and statistical analysis.

291 Data Processing Algorithms

292 As mentioned above there are several algorithms from simple data import to feature detec-
 293 tion and identification are incorporated into the toolbox. The algorithms relevant to the two
 294 discussed workflows are described below. In depth details on the data processing algorithms
 295 can be found in the supporting information section S1.

296
 297 The implemented feature detection algorithm SAFD (Section S1.2) has the main advan-
 298 tage that it can perform feature detection on both profile and centroided data.³³ Where the
 299 former, even though computationally more expensive, can avoid false peaks introduced by
 300 centroiding or other steps and work directly with the raw MS1 data. If preferred, the profile
 301 data can also be centroided using the SAFD package and then perform feature detection. As

302 for the componentization step, CompCreate (Section S1.3) is implemented, which is able to
303 componentize both DDA and DIA data, including PRM, SWATH, and multi-collision anal-
304 ysis. Moreover, the algorithm reports all grouped information (i.e., (in-source) fragments,
305 isotopes, and adducts) in the output, providing full traceability of grouped MS1 features
306 instead of removing this information from the feature list. Moreover, CompCreate uses the
307 parameter free naive Bayes isotope detection model¹³ and cumulative neutral loss model for
308 fragment deconvolution.³⁶ Since a wide variety of data types can be analyzed with these
309 algorithms, consistent outputs can be obtained for further analysis.

310

311 The library search algorithm implemented in the jHRMS toolbox is ULSA (Section S1.4).⁹
312 This algorithm provides extensive reporting on the match quality (Section S1.4). In total,
313 there are 7 parameters that are used to provide a final score, for which the weights of each
314 quality reporting parameter can be set. The algorithm can be used with either the provided
315 database, containing MassBank EU,³⁸ MassBank of North America,³⁹ and GNPS,²⁰ or a
316 local database in a specified format. Finally, the implemented suspect screening algorithm
317 (S1.5) allows for screening of MS/MS spectra. Initially it screens for the precursor ion and
318 records all instances the precursor ion were found, including which fragments from the ref-
319 erence spectrum were matched and a matching score. Additionally, a suspect list with MS2
320 information can be easily constructed with InChIKeys or SMILES from the internal database.

321

322 **Visualization and Trend analysis**

323 The visualization capabilities of the jHRMS toolbox are another unique feature and will also
324 be extensively showcased in workflow I and II. At almost every point of the workflow the
325 data can be visualized and inspected, ranging from the more standard raw data visualization
326 to post-processing extracted data (i.e., features, components, identifications, suspects, and
327 alignments). The detected features can be visualized on a heatmap of the chromatogram

328 with which the user can interact and pull up the corresponding plots of the MS1 and MS2
329 scans. On the other hand, the raw data from the time and mass domains behind specific
330 features or components can be plotted, allowing the user to inspect the quality of extracted
331 information. Additionally, suspect screening and identification matches can also be visual-
332 ized to perform post-processing quality control, showing both the matching performance in
333 the time and mass domain. More advanced features are the possibility to plot the Kendrick
334 masses for feature lists and aligned files and the possibility to plot the alignment of features,
335 components, and suspects.

336

337 As for the trend and statistical analysis, multiple supervised and un-supervised methods
338 have been implemented together with their visualization that can be used on the aligned
339 data. The implemented unsupervised methods are k-means, singular value decomposition
340 (SVD), principal component analysis (PCA), and hierarchical clustering analysis (HCA).
341 These methods can be used for exploratory analysis of the data and unsupervised ‘clus-
342 tering’. While for the supervised methods, partial least squares regressor (PLSR), partial
343 least square discriminant analysis (PLS-DA), and linear discriminant analysis (LDA) have
344 been implemented. These supervised methods require a list of values corresponding to each
345 file/column from the aligned data, enabling the analysis of correlating features to a certain
346 underlying trend.

347 **Technical advantages**

348 Calling the jHRMS toolbox in Julia loads all related packages and makes all the functions
349 accessible in the command line. This also provides the user with the option to use the
350 graphical user interface and/or the command line. Since the jHRMS toolbox is fully open-
351 source and open-access, the toolbox can be tailored to the users’ needs. This could include
352 developing new algorithms as part of a full workflow, implementing other algorithms of
353 interest, and adding visualization features. These changes can then also be provided to

354 other users in a team, enabling them to use more advanced and tailored features through the
355 user interface. The Julia programming language also allows calling functions and packages
356 written in other programming languages, including Python, C/C++, and R. This enables
357 the user to implement algorithms written in other languages into the jHRMS toolbox. Also,
358 the toolbox is operating system agnostic, thus can be used on Windows, macOS, and Linux
359 operating systems. The toolbox has also an option to save and load methods, allowing
360 the user to keep track of which dataset has been processed with what settings. Finally, the
361 jHRMS toolbox provides extensive and step-by-step reporting at each point of the workflow.
362 Results are generated for every processing step and saved in .csv files. When processing of
363 a workflow is interrupted, the toolbox picks up the workflow where it left off the next time
364 it is started. This also enables the user to investigate the data at each step and backtrack
365 information of cases of interest. Moreover, this allows the user to easily get results and use
366 them with other algorithms both inside and outside the jHRMS toolbox.

367 **Showcase Workflow I**

368 For the first workflow, suspect screening of 816 pharmaceuticals was performed followed by
369 filtering and alignment of the results. From the 816 pharmaceuticals, 280 chemicals were
370 found in the database comprised of MassBank EU, MassBank of North America, and NIST20.
371 For these 280 chemicals the merged suspect entries were obtained using the jHRMS toolbox.
372 From this suspect screening list, 181 unique compounds were found during screening of the
373 river samples with an intensity above 10000 counts and at least 1 detected fragment. Figure 4
374 shows the intensity trend of these compound across the samples. Overall, a higher frequency
375 of detection of the pharmaceuticals was found in 2019 compared to 2020. The latter trend
376 was also found in the results of the study by Singh et. al., for which the data was originally
377 measured.³⁴ In that study, the compounds were confirmed at level 1 and 2a, according to
378 the Schymanski scheme,⁴⁵ and in silico fragmentation was used to obtain spectra for the
379 compounds that were missing from the MassBank of North America database.

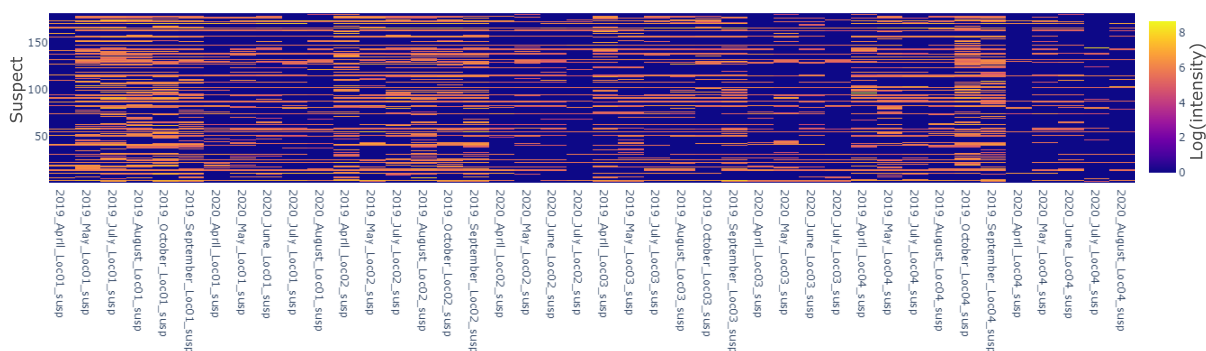


Figure 4: The intensity of the aligned suspect screening results, showing the suspect number on the y-axis, the samples on the x-axis, and the intensity on the z-axis.

380 Alternatively, the match factor between the suspect entry with the best matching can-
 381 didate signal can be visualized similarly. Even though no match factor filter was applied
 382 during trend analysis for the suspect screening, the occurrence frequency of likely suspects
 383 is still higher in 2019 compared to 2020. Some compounds are frequently found with high
 384 confidence for all locations and months while other are less frequent or have an overall less
 385 confident candidate. In this case the lower match factors do not completely correlate with
 386 being a more likely match, since the spectra of all database candidates for a give InChIKey
 387 were merged for setting up the suspect screening entries. This means that, for cases where
 388 many spectra were found with a variety of unique fragments, the match factor is inherently
 389 lower.

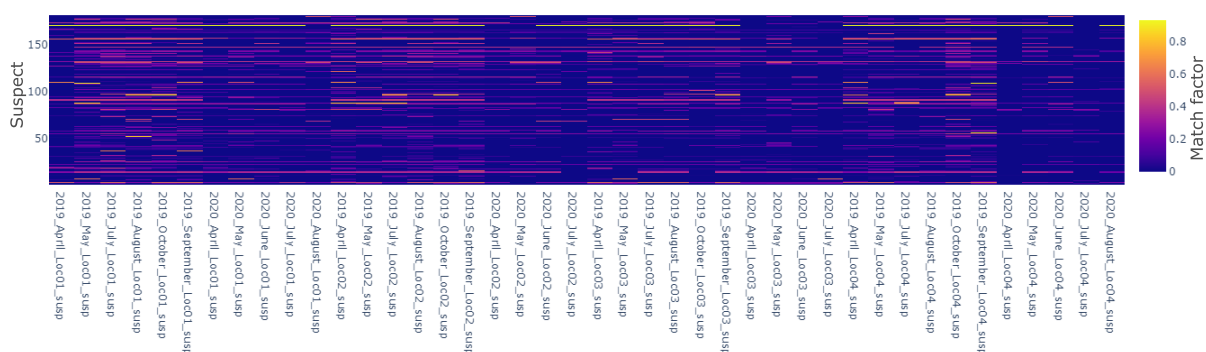


Figure 5: The match factor of the aligned suspect screening results, showing the suspect number on the y-axis, the samples on the x-axis, and the match factor on the z-axis.

390 To highlight the post-processing capabilities of the toolbox, one of the suspects was
391 further investigated. For this aspirin (BSYNRYMUTXBXSQ-UHFFFAOYSA-N) was chosen
392 and detected in 39 samples with match factors varying between 0.2 and 0.57 with 0.25 being
393 the median and the number of fragments matched between 1 and 4 with a median of 1
394 fragment. Figure 6 shows the matched aspirin suspect information from the sample take in
395 April 2019. Overall, a match at 1.67 minutes was found where two isotopes were detected
396 and three fragments were matched with the suspect entry, making this indeed a likely match.
397 On the other hand, aspirin was not found in three of the 42 samples. Figure S3 shows that
398 for the measurement from May 2020 indeed no signal was found around 1.67 minutes for the
399 precursor mass of aspirin.

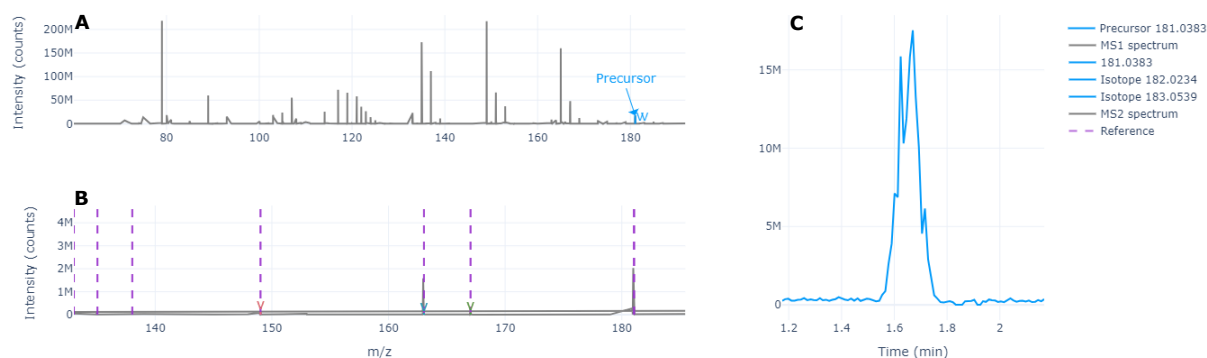


Figure 6: Visualized suspect screening result of the suspect aspirin at 1.67 minutes with an precursor m/z of 181.04. **A)** shows the MS1 signal with the detected precursor peak, **B)** shows the MS2 spectrum with the library reference signals in purple and the detected fragments colored and highlighted with downward arrows, and **C)** shows the XIC in the time domain of the precursor ion.

400 To complement the suspect screening, a full identification workflow using feature detec-
401 tion and componentization was also performed. It is known that depending on the workflow
402 used different compounds can be detected.²¹ Hence, the use of both suspect screening and
403 identification workflows can complement each other. Using the sample from April 2019, the
404 identification entry for aspirin was evaluated (Figure 7). It can be seen that the same pre-
405 cursor peak was detected but no fragments were matched with the library entry. Looking at

406 the component, the fragment of 167.004 Da was componentized to the aspirin precursor ion.
407 However, the mass tolerance based on the precursor peak width (i.e., $0.0334/2 = 0.0167$ Da)
408 was too small to match this fragment with the library entry (i.e., $167.034 - 167.004 \pm 0.0167$
409 m/z). Overall, this showed the post-processing capabilities of the toolbox with regard to
410 the suspect screening and identification, which are two workflows that can complement each
411 other in NTA.

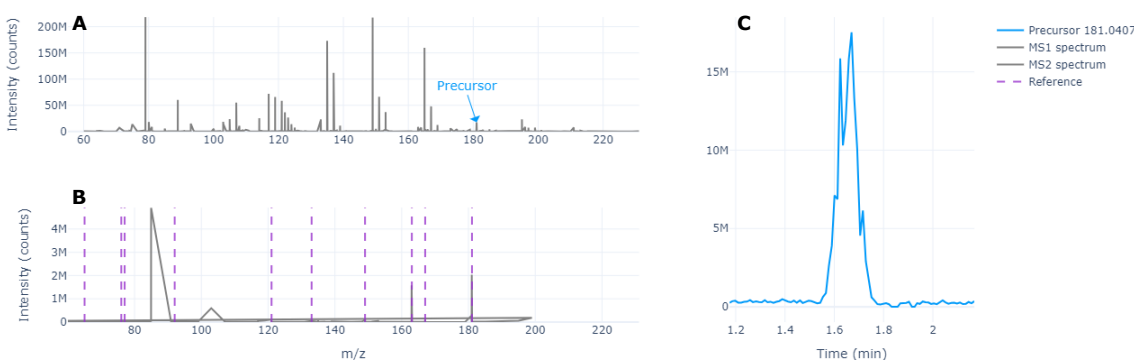


Figure 7: Visualized component identification of the component at 1.67 minutes with a precursor m/z of 181.04, matching the information of the aspirin suspect found previously. **A)** shows the MS1 signal with the detected precursor peak, **B)** shows the MS2 spectrum with the library reference signals in purple, and **C)** shows the XIC in the time domain of the precursor ion.

412 Finally, the study by Singh et. al. reported a few frequently detected chemicals. One of
413 those chemicals was the antihypertensive drugs sotalol (ZBMZVLHSJCTVON-UHFFFAOYSA-
414 N), for which a overall higher concentration was found in 2019 compared to 2020. Even
415 though no quantification has been performed with the showcase, an overall higher intensity
416 for sotalol in 2019 was found via our retrospective analysis of the data and can be seen in
417 figure 8.

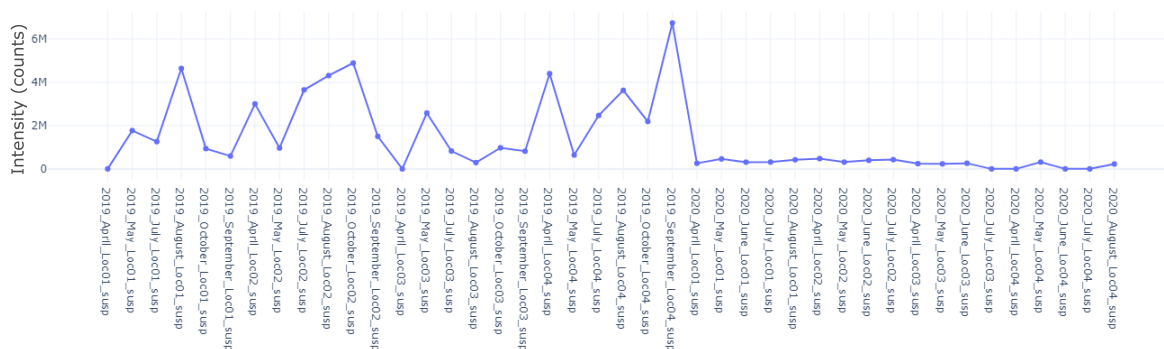


Figure 8: Intensity trend for the screened suspect sotalol for four different river sample locations obtained in 2019 and 2020.

418 Showcase Workflow II

419 For the second workflow, feature detection and componentization were performed on fish
 420 samples coming from brain and liver tissues. Alignment and blank filtering have been per-
 421 formed on both the feature lists and components, which were finally clustered to investigate
 422 the potential of separating the two tissues (Figure 2). First the brain and liver samples with
 423 their corresponding blanks were aligned separately. Figure 9 shows the aligned features for
 424 the brain samples. Here, the similarities and differences between the tissue sample features
 425 can be seen. Additionally, there are a few frequently occurring features found in both of the
 426 tissues and the blank samples, which are not likely to contain relevant information.

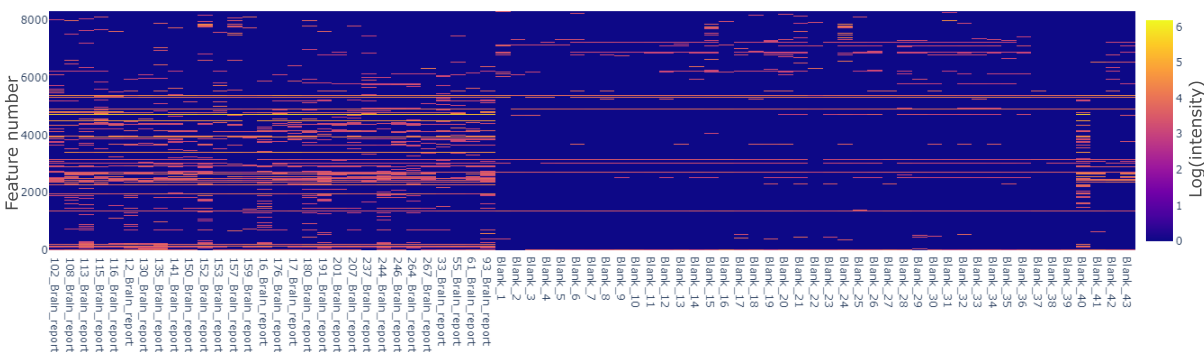


Figure 9: The aligned feature list of the brain samples and blanks, showing the feature number on the y-axis and the samples on the x-axis.

427 Figure 10 shows the remaining sample features after using a filter with a signal to median
 428 blank ratio of 5. It can be seen that frequently occurring blank features are removed (i.e.,
 429 set to 0 intensity) from the samples. Overall, reducing the total number of aligned features
 430 from 19,557 to 8,812 in the samples. The visualized filtered and unfiltered aligned feature
 431 list for the liver samples can be found in the SI (Figure S4 and S5), showing similar results.

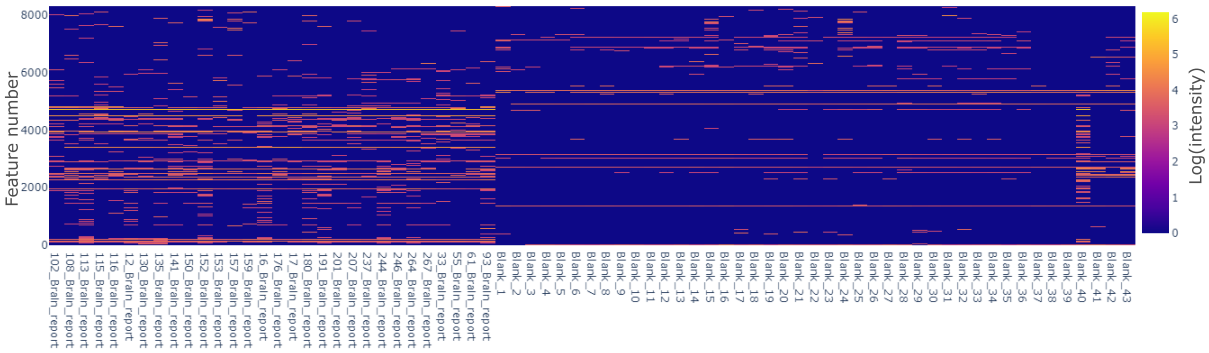


Figure 10: The filtered aligned feature list of the brain samples and blanks, showing the feature number on the y-axis and the samples on the x-axis. For the filtering a median blank intensity with a signal to blank ratio of 5 was used.

432 Subsequently, the individual feature lists of the brain and liver (Figure S4 and S5) samples
 433 were reduced based on the blank filtering performed above, obtaining only the tissue features
 434 of interest. These filtered feature lists from the brain and liver samples were then aligned
 435 with each other and hierarchical clustering was performed. Figure 11 shows that based on
 436 the features it is possible to cluster or differentiate between liver and brain tissue samples.
 437 Both similarities and difference between the samples can clearly be seen in these plots.

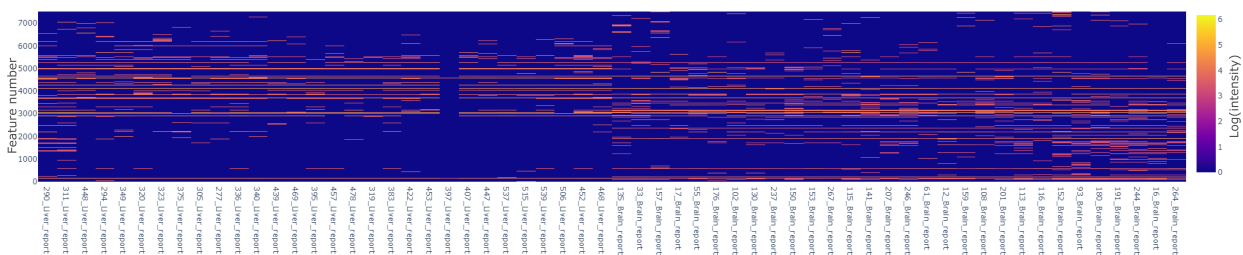


Figure 11: Aligned feature list of the individually blank filtered brain and liver samples, showing the feature number on the y-axis and the samples on the x-axis.

438 Finally, the same steps were taken for the component files of the brain and liver samples.
439 The component alignment and filtering for the samples can be found in the SI (Figure
440 S6, S7, S8, and S9), showing a similar trend in information removal as the brain samples
441 above. Overall, after clustering of the aligned liver and brain tissue samples, again a clear
442 separation between the two groups can be found. It can also be seen that a lower number of
443 unique components can be found compared to the number of unique features, which would
444 be expected as features from the same compound are grouped together. An advantage of this
445 is that compounds contribute equally to the clustering. Not all compounds have an equal
446 number of features present in the feature list (e.g., in-source fragments, isotopes, adducts,
447 and precursor ion), meaning that there can be an unequal contribution between compounds
448 to the result of the clustering. This was not an issue during the showcasing of this dataset,
449 but may need special attention, depending on the type and origin of the investigated data.
450 Overall, these plots have shown that it is possible to differentiate between the two tissues,
451 showing that the algorithms were able to extract important chemical information from the
452 data.

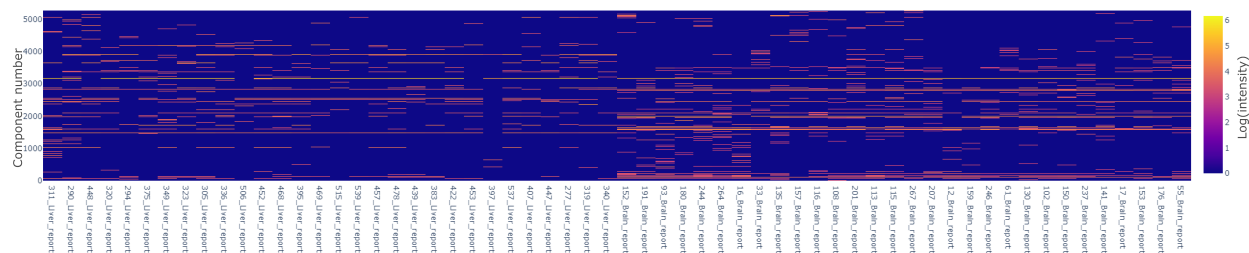


Figure 12: Aligned component list of the individually blank filtered brain and liver samples, showing the feature number on the y-axis and the samples on the x-axis. For the filtering a median blank intensity with a signal to blank ratio of 5 was used.

453 Conclusions

454 The jHRMS toolbox is an open-source and open-access modular toolbox that, on one hand,
455 provides a user interface for NTA HRMS data processing algorithms and, on the other hand,

456 allows full freedom to modify and add workflows. The first environmental dataset that was
457 processed with the jHRMS toolbox, showed similar trends in the data as the original study
458 by Singh et. al.³⁴ and the extensive visualization and reporting by the toolbox. Meanwhile
459 the second workflow showcased that the algorithms are able to extract important informa-
460 tion that can differentiate between biological tissues. Additionally, the toolbox comes with
461 built-in statistical analysis and visualization (i.e., post-processing) at almost every point in
462 the workflow. The specific algorithms that are implemented at the time of publication are:
463 SAFD for the feature detection of both profile and centroided data, CompCreate for the
464 componentization of both DDA and DIA measurements, ULSA that provides extensive re-
465 porting of match score and quality, and suspect screening from ULSA that also screens MS2
466 information for fragments. Moreover, the toolbox runs on all operating systems, allows for
467 saving parameters for specific methods, and reports algorithms results for each step in the
468 workflow in .csv files. The latter allows for easy transfer to and from other platforms to,
469 for example, use the feature detection results from another algorithms and proceed with the
470 workflow in the jHRMS toolbox.

471

472 As for the implemented algorithms themselves, further improvements can still be made.
473 For example, the component alignment algorithms currently uses the MS1 precursor infor-
474 mation to align features across multiple samples while the fragment information also contains
475 crucial knowledge on which features are the same. However, this requires further research
476 on how much information is required to confidently group components from different sam-
477 ples. Meanwhile work is being done to extend and test the algorithms for GC-HRMS data
478 processing. Finally, the predictive models that can use components (i.e., cumulative neutral
479 losses) need to be implemented to use their predicted information during trend analysis (e.g.,
480 toxicity) or even alignment (e.g., retention indices).

481

482 For transfer of results between the jHRMS toolbox and algorithms from other platforms,

483 application programming interfaces (APIs) might be needed to make the input/output com-
484 patible. Namely, information can be formatted or named differently or even missing. How-
485 ever, this does not mean that the algorithms are not compatible. Currently, such APIs have
486 not yet been developed and will expand as the need for it arises. Finally, the algorithms
487 implemented in the jHRMS toolbox can be expanded over time as interest for certain func-
488 tionalities or algorithms arises. Luckily, these algorithms are not limited by the programming
489 language used as packages from other programming languages can be called through Julia.

490 **Acknowledgement**

491 The authors thank the Environmental Monitoring and Computational Mass Spectrometry
492 (www.emcms.info) group for their insights and feedback and Lukas W.Y. Robbertsen for
493 the design of the jHRMS toolbox logo. SS is thankful to the UvA Data Science Center
494 and ChemistryNL for financial support. The Queensland Alliance for Environmental Health
495 Sciences gratefully acknowledges the financial support from the Queensland Department
496 of Health. J.W.O is the recipient of an NHMRC Emerging Leadership Fellowship (EL1
497 2009209).

498 **Supporting Information Available**

499 Overview of the implemented algorithms and their related information, which provides either
500 a detailed or a general description depending on whether the algorithm has been published.
501 An XIC of aspirin for Workflow I and alignment figures for workflow II of the filtered and
502 unfiltered liver feature lists, brain component lists, and liver components lists.

503 **Author Information**

504 Corresponding Author:

505 Saer Samanipour

506 Van 't hof institute for molecular sciences (HIMS),

507 University of Amsterdam,

508 the Netherlands

509 Email: s.samanipour@uva.nl

510

511 Denice van Herwerden

512 Van 't hof institute for molecular sciences (HIMS),

513 University of Amsterdam,

514 the Netherlands

515 Email: d.vanherwerden@uva.nl

516

References

- 517
- 518 (1) Reymond, J. L. The Chemical Space Project. *Accounts of Chemical Research* **2015**, *48*,
519 722–730.
- 520 (2) Black, G. et al. Exploring chemical space in non-targeted analysis: a proposed
521 ChemSpace tool. *Analytical and Bioanalytical Chemistry* **2023**, *415*, 35–44.
- 522 (3) Hulleman, T.; Turkina, V.; O'Brien, J. W.; Chojnacka, A.; Thomas, K. V.; Sama-
523 nipour, S. Critical Assessment of the Chemical Space Covered by LC-HRMS Non-
524 Targeted Analysis. *Environmental Science and Technology* **2023**, *57*, 14101–14112.
- 525 (4) Samanipour, S.; Barron, L.; van Herwerden, D.; Praetorius, A.; Thomas, K.; O'Brien, J.
526 Exploring the Chemical Space of the Exposome: How Far Have We Gone? *ChemRxiv*
527 **2024**,
- 528 (5) van Herwerden, D.; Nikolopoulos, A.; Barron, L.; O'Brien, J.; Pirok, B.; Thomas, K.;
529 Samanipour, S. Exploring the Chemical Subspace of RPLC: a Data Driven Approach.
530 *ChemRxiv* **2024**,
- 531 (6) Schulze, B.; Jeon, Y.; Kaserzon, S.; Heffernan, A. L.; Dewapriya, P.; O'Brien, J.; Gomez
532 Ramos, M. J.; Ghorbani Gorji, S.; Mueller, J. F.; Thomas, K. V.; Samanipour, S. An
533 assessment of quality assurance/quality control efforts in high resolution mass spectrom-
534 etry non-target workflows for analysis of environmental samples. *Trends in Analytical*
535 *Chemistry* **2020**, *133*, 116063.
- 536 (7) Schymanski, E. L. et al. Non-target screening with high-resolution mass spectrometry:
537 Critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.* **2015**,
538 *407*, 6237–6255.
- 539 (8) Werner, E.; Heilier, J.-F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.-C. Mass spec-

- 540 trometry for the identification of the discriminating signals from metabolomics: Current
541 status and future trends. *J. Chromatogr. B* **2008**, *871*, 143–163.
- 542 (9) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and
543 a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent
544 Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Re-
545 sults. *Environ. Sci. Technol.* **2018**, *52*, 4694–4701.
- 546 (10) Samanipour, S.; Kaserzon, S.; Vijayasathy, S.; Jiang, H.; Choi, P.; Reid, M. J.;
547 Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS
548 analysis for a risk warning system of chemical hazards in drinking water: A proof of
549 concept. *Talanta* **2019**, *195*, 426–432.
- 550 (11) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.;
551 Slobodnik, J.; Krauss, M. High-resolution mass spectrometry to complement monitor-
552 ing and track emerging chemicals and pollution trends in European water resources.
553 *Environ. Sci. Eur.* **2019**, *31*, 62.
- 554 (12) Minkus, S.; Bieber, S.; Letzel, T. Spotlight on mass spectrometric non-target screening
555 analysis: Advanced data processing methods recently communicated for extracting,
556 prioritizing and quantifying features. *Analytical Science Advances* **2022**, *3*, 103–112.
- 557 (13) van Herwerden, D.; O'Brien, J. W.; Choi, P. M.; Thomas, K. V.; Schoenmakers, P. J.;
558 Samanipour, S. Naive Bayes classification model for isotopologue detection in LC-
559 HRMS data. *Chemometrics and Intelligent Laboratory Systems* **2022**, *223*, 104515.
- 560 (14) Schulze, B.; Heffernan, A. L.; Samanipour, S.; Ramos, M. J. G.; Veal, C.;
561 Thomas, K. V.; Kaserzon, S. L. Is Nontarget Analysis Ready for Regulatory Appli-
562 cation? Influence of Peak-Picking Algorithms on Data Analysis. *Analytical Chemistry*
563 **2023**, *95*, 18361–18369.

- 564 (15) Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Compari-
565 son of Software Tools for Liquid Chromatography–High-Resolution Mass Spectrometry
566 Data Processing in Nontarget Screening of Environmental Samples. *Analytical Chem-*
567 *istry* **2020**, *92*, 1898–1907, PMID: 31840499.
- 568 (16) Höcker, O.; Flottmann, D.; Schmidt, T. C.; Neusüß, C. Non-targeted LC-MS and CE-
569 MS for biomarker discovery in bioreactors: Influence of separation, mass spectrometry
570 and data processing tools. *Science of The Total Environment* **2021**, *798*, 149012.
- 571 (17) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An
572 integrated strategy for compound spectra extraction and annotation of liquid chro-
573 matography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.
- 574 (18) Alygizakis, N. A.; Oswald, P.; Thomaidis, N. S.; Schymanski, E. L.; Aalizadeh, R.;
575 Schulze, T.; Oswaldova, M.; Slobodnik, J. NORMAN digital sample freezing platform:
576 A European virtual platform to exchange liquid chromatography high resolution-mass
577 spectrometry data and screen suspects in “digitally frozen” environmental samples.
578 2019.
- 579 (19) FOR-IDENT-Platform: International hunt for unknown molecules combining interna-
580 tional workflows and software tools. <https://water.for-ident.org/#!home>.
- 581 (20) Wang, M. et al. Sharing and community curation of mass spectrometry data with
582 Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**,
583 *34*, 828–837.
- 584 (21) Feraud, M.; O’Brien, J. W.; Samanipour, S.; Dewapriya, P.; van Herwerden, D.; Kaser-
585 zon, S.; Wood, I.; Rauert, C.; Thomas, K. V. InSpectra – A platform for identifying
586 emerging chemical threats. *Journal of Hazardous Materials* **2023**, *455*.
- 587 (22) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.;

- 588 Vandergheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-independent MS/MS decon-
589 volution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12*, 523–526.
- 590 (23) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework
591 for processing, visualizing, and analyzing mass spectrometry-based molecular profile
592 data. *BMC Bioinf.* **2010**, *11*, 395.
- 593 (24) Röst, H. L. et al. OpenMS: A flexible open-source software platform for mass spectrom-
594 etry data analysis. *Nature Methods* **2016**, *13*, 741–748.
- 595 (25) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. patRoon:
596 open source software platform for environmental mass spectrometry based non-target
597 screening. *Journal of Cheminformatics* **2021**, *13*.
- 598 (26) Helmus, R.; van de Velde, B.; Brunner, A. M.; ter Laak, T. L.; van Wezel, A. P.;
599 Schymanski, E. L. patRoon 2.0: Improved non-target analysis workflows including au-
600 tomated transformation product screening. *Journal of Open Source Software* **2022**, *7*,
601 4029.
- 602 (27) Peters, K. et al. PhenoMeNal: Processing and analysis of Metabolomics data in the
603 Cloud. *GigaScience* **2018**, *8*.
- 604 (28) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.;
605 Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass
606 spectra into metabolite structure information. *Nature Methods* **2019**, *16*, 299–302.
- 607 (29) Shen, X.; Yan, H.; Wang, C.; Gao, P.; Johnson, C. H.; Snyder, M. P. TidyMass an
608 object-oriented reproducible analysis framework for LC–MS data. *Nature Communica-*
609 *tions* **2022**, *13*.
- 610 (30) Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Process-

- 611 ing mass spectrometry data for metabolite profiling using nonlinear peak alignment,
612 matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.
- 613 (31) Vosough, M.; Schmidt, T. C.; Renner, G. Non-target screening in water analysis: recent
614 trends of data evaluation, quality assurance, and their future perspectives. *Analytical
615 and Bioanalytical Chemistry* **2024**, *416*, 2125–2136.
- 616 (32) Liu, H.; Wang, R.; Zhao, B.; Xie, D. Assessment for the data processing performance
617 of non-target screening analysis based on high-resolution mass spectrometry. *Science
618 of the Total Environment* **2024**, *908*.
- 619 (33) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V. Self adjusting algorithm
620 for the nontargeted feature detection of high resolution mass spectrometry coupled with
621 liquid chromatography profile data. *Anal. Chem.* **2019**, *91*, 10800–10807.
- 622 (34) Singh, R. R.; Lai, A.; Krier, J.; Kondić, T.; Diderich, P.; Schymanski, E. L. Occurrence
623 and Distribution of Pharmaceuticals and Their Transformation Products in Luxem-
624 bourgish Surface Waters. *ACS Environmental Au* **2021**, *1*, 58–70.
- 625 (35) Bade, R. et al. Workflow to facilitate the detection of new psychoactive substances and
626 drugs of abuse in influent urban wastewater. *Journal of Hazardous Materials* **2024**,
627 *469*.
- 628 (36) van Herwerden, D.; O'Brien, J. W.; Lege, S.; Pirok, B. W.; Thomas, K. V.; Sama-
629 nipour, S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray
630 Ionization High-Resolution Mass Spectrometry Data. *Analytical Chemistry* **2023**, *95*,
631 12247–12255.
- 632 (37) Samanipour, S.; Choi, P.; O'Brien, J. W.; Pirok, B. W. J.; Reid, M. J.; Thomas, K. V.
633 From Centroided to Profile Mode: Machine Learning for Prediction of Peak Width in
634 HRMS Data. *Analytical Chemistry* **2021**, *93*, 16562–16570.

- 635 (38) MassBank EU. <https://massbank.eu/MassBank/Index>.
- 636 (39) MassBank of North America. <https://mona.fiehnlab.ucdavis.edu/downloads>.
- 637 (40) NIST 20. <https://www.nist.gov/programs-projects/nist20-updates-nist-tan>
638 [dem-and-electron-ionization-spectral-libraries](https://www.nist.gov/programs-projects/nist20-updates-nist-tan).
- 639 (41) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V.; Praetorius, A. From
640 Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach
641 to Chemical Prioritization. *Environmental Science and Technology* **2023**, *57*, 17950–
642 17958.
- 643 (42) Boelrijk, J.; van Herwerden, D.; Ensing, B.; Forré, P.; Samanipour, S. Predicting RP-
644 LC retention indices of structurally unknown chemicals from mass spectrometry data.
645 *Journal of Cheminformatics* **2023**, *15*, 1–12.
- 646 (43) Peets, P.; Wang, W. C.; Macleod, M.; Breitholtz, M.; Martin, J. W.; Krueve, A. MS2Tox
647 Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Wa-
648 ter by Nontarget LC-HRMS. *Environmental Science and Technology* **2022**, *56*, 15508–
649 15517.
- 650 (44) Sepman, H.; Malm, L.; Peets, P.; MacLeod, M.; Martin, J.; Breitholtz, M.; Krueve, A.
651 Bypassing the Identification: MS2Quant for Concentration Estimations of Chemicals
652 Detected with Nontarget LC-HRMS from MS2 Data. *Analytical Chemistry* **2023**, *95*,
653 12329–12338.
- 654 (45) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J.
655 Identifying small molecules via high resolution mass spectrometry: Communicating
656 confidence. 2014.

657 **TOC Graphic**

658

