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

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

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

Introduction

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

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

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

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

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

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

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

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

Experimental Section

To showcase the jHRMS toolbox, two datasets have been used acquired on different instruments, in different labs covering both environmental and biological matrices. One dataset comprises of surface water samples measured on an Orbitrap instrument in centroided mode[34] while the other dataset dealt with two biological matrices measured on a ZenoTOF in profile model. Brief details on the datasets are provided below as the aim of the paper is to showcase the jHRMS toolbox without going too much in depth of what is found in the data. These two different data sets have specifically been chosen to show the compatibility of the toolbox with different types of measured data and applications' purposes. Furthermore, this section contains the processing details and parameters used for showcasing the toolbox capabilities and the code availability of the jHRMS toolbox.

Centroided Dataset

The centroided dataset is comprised of surface water samples coming from different rivers collected at 4 week intervals.[34] This data has been made publicly available and can be obtained from the MassIVE repository: `https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000087190`. In brief, the collected samples were extracted using solid phase extraction and analyzed with DDA LC-HRMS with a reversed phase column selectivity.[34] For the mobile phase a mixture of water with 0.1% formic acid (A) and methanol (B) was used. The gradient started with 90/10 A/B for 2 minutes, 0/100 at 15 minutes, 0/100 at 20 minutes, 90/10 from 21 to 30 minutes, using a flow rate of 0.2 mL/min. As for the DDA MS part, the samples measured with positive electrospray ionization were used. Here a scan range of 60-900 m/z, a MS1 resolution of 120,000 at 200 m/z, a maximum injection time of 70 ms, and an automatic gain control target of 1.0 $\times$ 106. For the top 5 data dependent analysis scans a resolution of 30,000 at 200 m/z, maximum injection time of 70 ms, 1.0 Da isolation window, and 30 (N)CE were used. For further details on instrumental settings and sample preparation see the citation.[34] For the showcase of workflow I, only the measurements acquired in positive mode of the 4 most frequently measured locations were used, which were location 1 to 4. Location 1 came from south west Luxembourg, location 2 and 3 from the middle, and location 4 from the east of Luxembourg. Each location was sampled 10 to 11 times in a time span of April 2019 till September 2020.

Profile Dataset

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

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

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

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

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

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

Showcase Workflows Settings

Workflow I

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

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

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

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

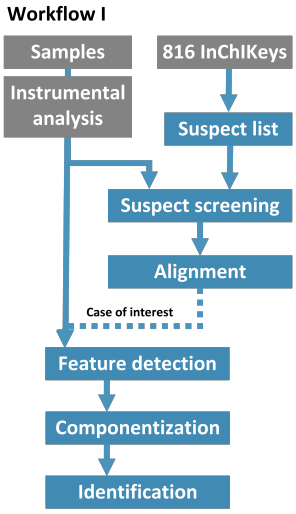

Figure 1: Overview of workflow I and the processing steps taken.

Workflow II

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

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

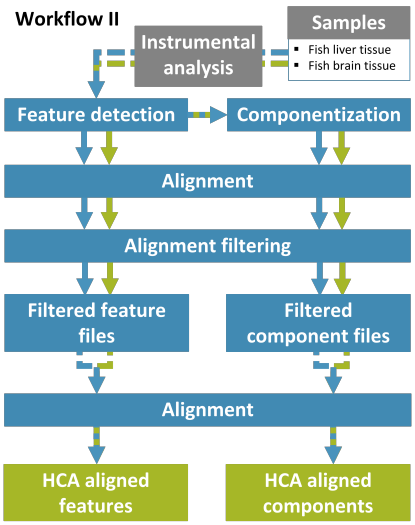

Figure 2: Overview of workflow II and the processing steps taken.

Calculations and Code Availability

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

Results and discussion

Modular Workflows

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

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

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

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

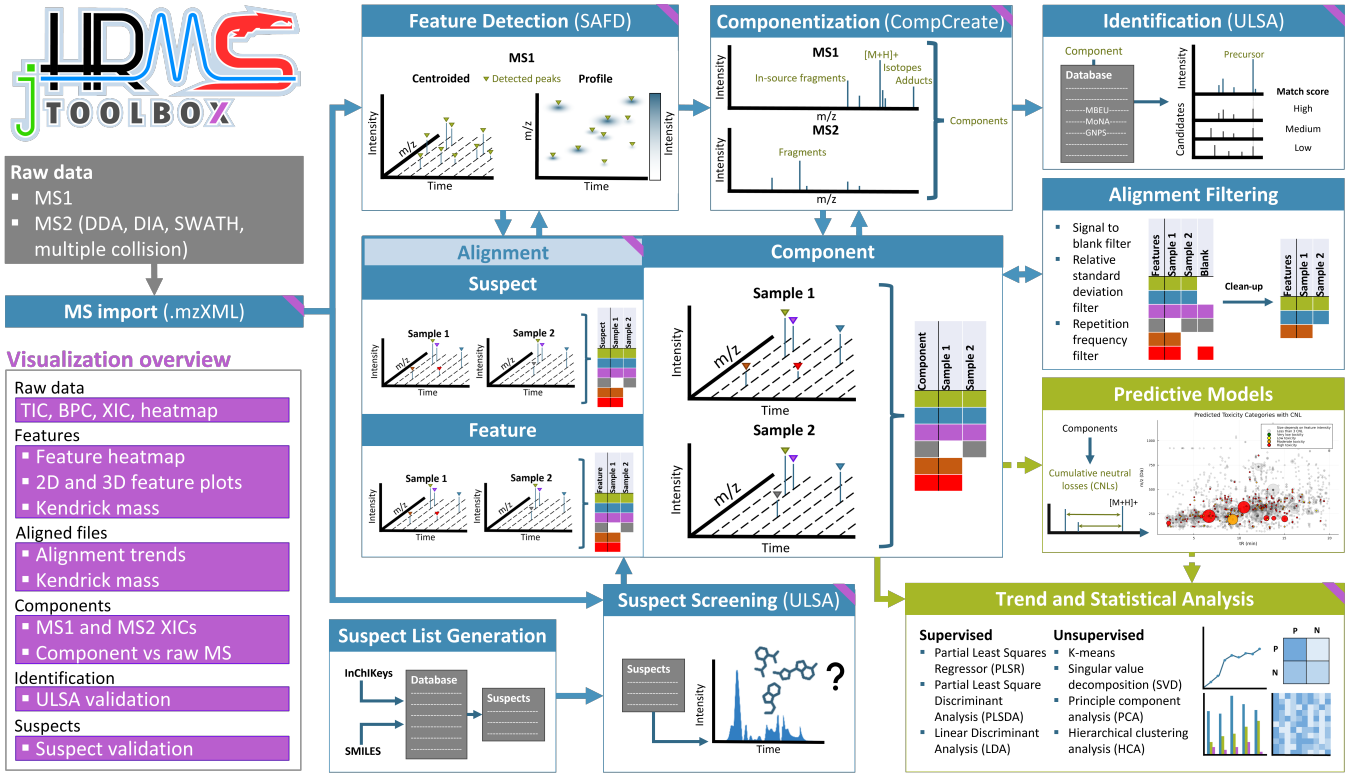

Figure 3: Overview of the jHRMS toolbox modular workflow. In blue are the algorithms and functions related to the HRMS data processing steps, in purple is the overview of the post-processing visualization where the purple ribbons refer to the post-processing visualization, and in green the possible predictive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

Data Processing Algorithms

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

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

for the componentization step, CompCreate (Section S1.3) is implemented, which is able to componentize both DDA and DIA data, including PRM, SWATH, and multi-collision analysis. Moreover, the algorithm reports all grouped information (i.e., (in-source) fragments, isotopes, and adducts) in the output, providing full traceability of grouped MS1 features instead of removing this information from the feature list. Moreover, CompCreate uses the parameter free naive Bayes isotope detection model[13] and cumulative neutral loss model for fragment deconvolution.[36] Since a wide variety of data types can be analyzed with these algorithms, consistent outputs can be obtained for further analysis.

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

Visualization and Trend analysis

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

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

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

Technical advantages

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

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

Showcase Workflow I

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

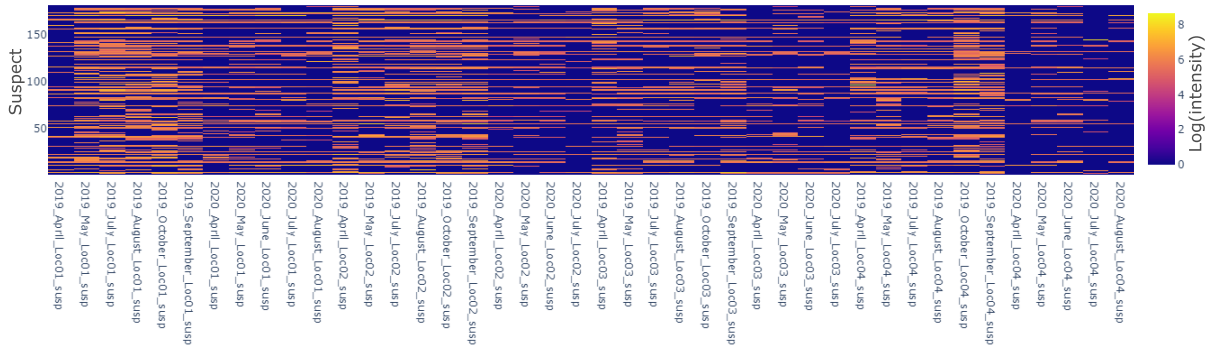

Figure 4: The intensity of the aligned suspect screening results, showing the suspect number on the y-axis,the samples on the x-axis, and the intensity on the z-axis.

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

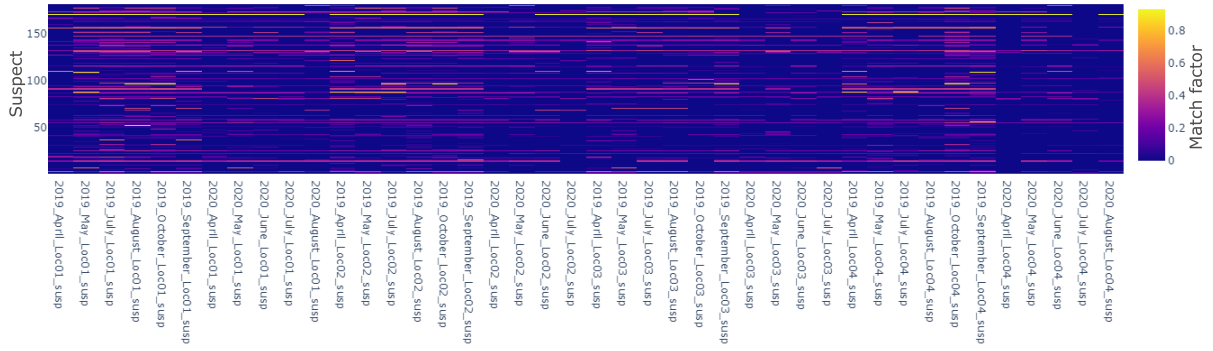

Figure 5: The match factor of the aligned suspect screening results, showing the suspect number on the y-axis, the samples on the x-axis, and the match factor on the z-axis.

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

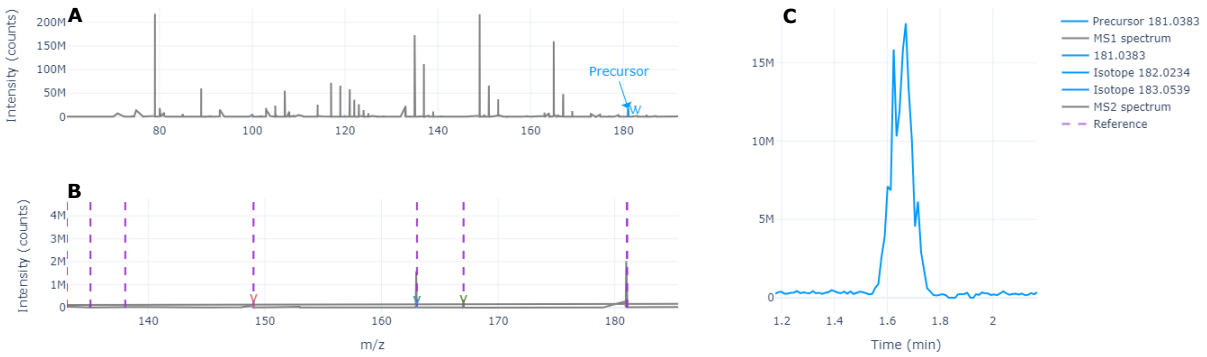

Figure 6: Visualized suspect screening result of the suspect aspirin at 1.67 minutes with an precursor m/z of 181.04. **A**) shows the MS1 signal with the detected precursor peak, **B**) shows the MS2 spectrum with the library reference signals in purple and the detected fragments colored and highlighted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

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

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

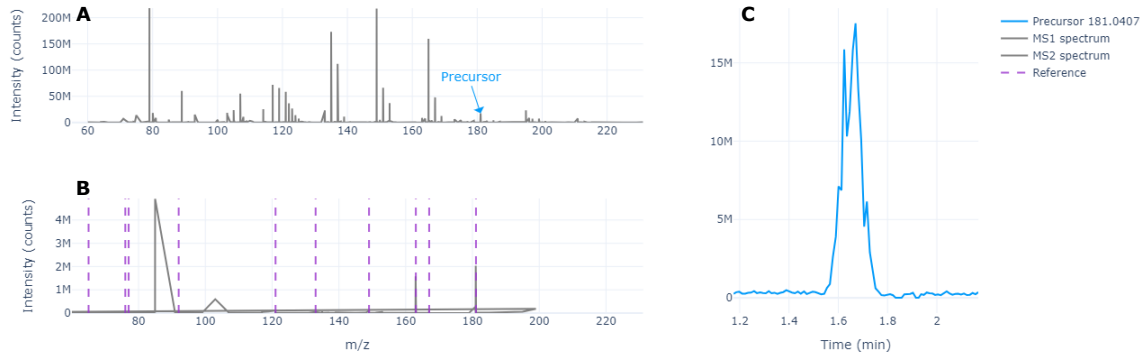

Figure 7: Visualized component identification of the component at 1.67 minutes with a precursor m/z of 181.04, matching the information of the aspirin suspect found previously. **A**) shows the MS1 signal with the detected precursor peak, **B**) shows the MS2 spectrum with the library reference signals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

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

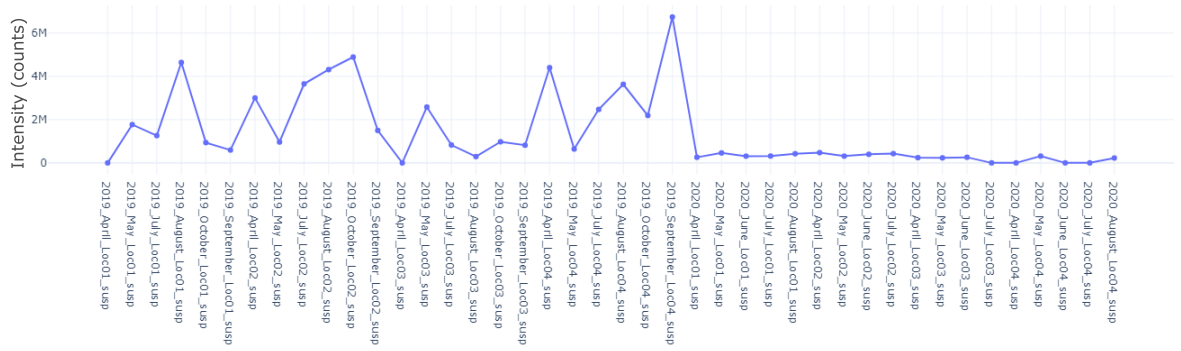

Figure 8: Intensity trend for the screened suspect sotalol for four different river sample locations obtained in 2019 and 2020.

Showcase Workflow II

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

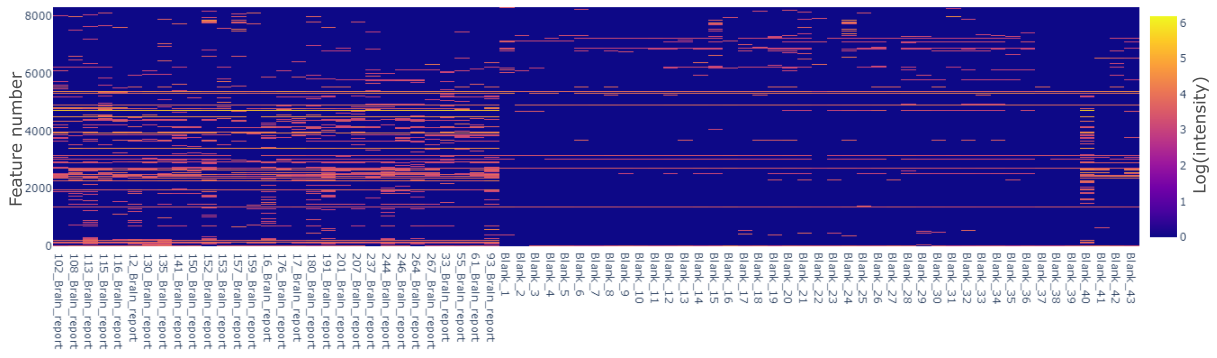

Figure 9: The aligned feature list of the brain samples and blanks, showing the feature number on the y-axis and the samples on the x-axis.

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

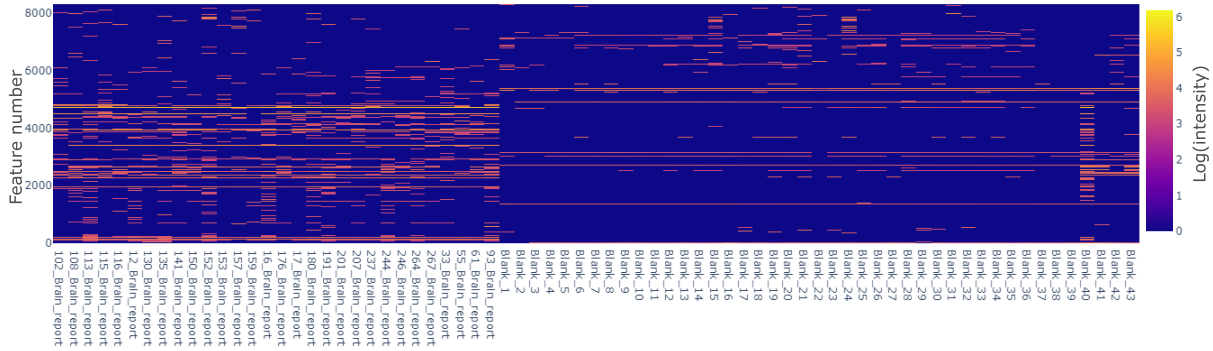

Figure 10: The filtered aligned feature list of the brain samples and blanks, showing the feature number on the y-axis and the samples on the x-axis. For the filtering a median blank intensity with a signal to blank ratio of 5 was used.

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

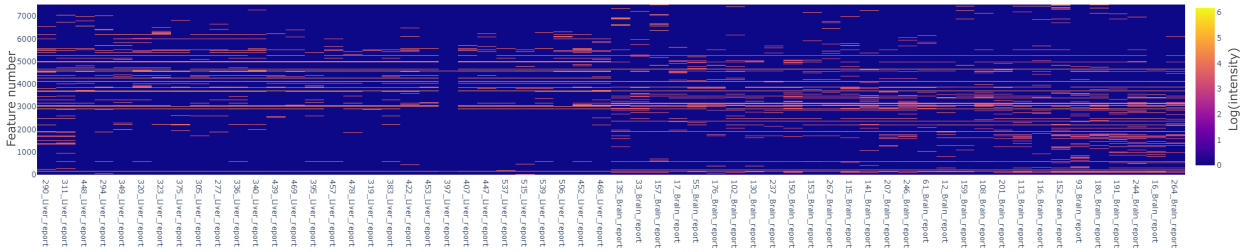

Figure 11: Aligned feature list of the individually blank filtered brain and liver samples, showing the feature number on the y-axis and the samples on the x-axis.

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

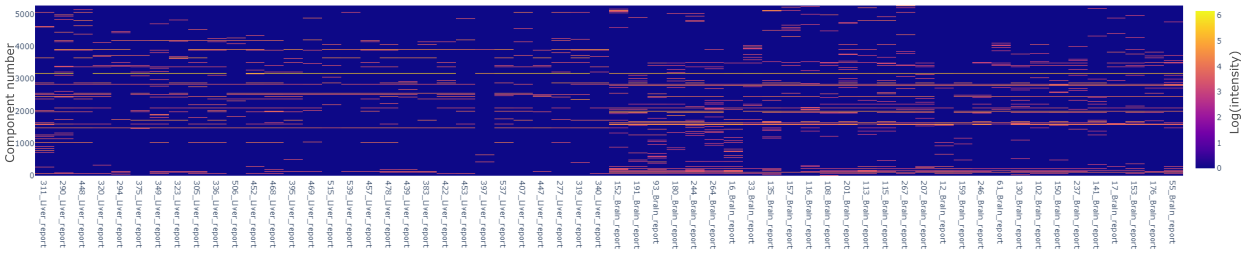

Figure 12: Aligned component list of the individually blank filtered brain and liver samples, showing the feature number on the y-axis and the samples on the x-axis. For the filtering a median blank intensity with a signal to blank ratio of 5 was used.

Conclusions

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

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

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

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

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

Acknowledgement

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

Supporting Information Available

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

Author Information

Corresponding Author:

Saer Samanipour

Van 't hoff institute for molecular sciences (HIMS),

University of Amsterdam,

the Netherlands

Email: s.samanipour@uva.nl

Denice van Herwerden

Van 't hoff institute for molecular sciences (HIMS),

University of Amsterdam,

the Netherlands

Email: d.vanherwerden@uva.nl

References

(1) Reymond, J. L. The Chemical Space Project. *Accounts of Chemical Research* **2015**, *48*, 722–730.

(2) Black, G. et al. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. *Analytical and Bioanalytical Chemistry* **2023**, *415*, 35–44.

(3) Hulleman, T.; Turkina, V.; O'Brien, J. W.; Chojnacka, A.; Thomas, K. V.; Samanipour, S. Critical Assessment of the Chemical Space Covered by LC-HRMS Non-Targeted Analysis. *Environmental Science and Technology* **2023**, *57*, 14101–14112.

(4) Samanipour, S.; Barron, L.; van Herwerden, D.; Praetorius, A.; Thomas, K.; O'Brien, J. Exploring the Chemical Space of the Exposome: How Far Have We Gone? *ChemRxiv* **2024**,

(5) van Herwerden, D.; Nikolopoulos, A.; Barron, L.; O'Brien, J.; Pirok, B.; Thomas, K.; Samanipour, S. Exploring the Chemical Subspace of RPLC: a Data Driven Approach. *ChemRxiv* **2024**,

(6) Schulze, B.; Jeon, Y.; Kaserzon, S.; Heffernan, A. L.; Dewapriya, P.; O'Brien, J.; Gomez Ramos, M. J.; Ghorbani Gorji, S.; Mueller, J. F.; Thomas, K. V.; Samanipour, S. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. *Trends in Analytical Chemistry* **2020**, *133*, 116063.

(7) Schymanski, E. L. et al. Non-target screening with high-resolution mass spectrometry: Critical review using a collaborative trial on water analysis. *Anal. Bioanal.Chem.* **2015**, *407*, 6237–6255.

(8) Werner, E.; Heilier, J.-F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.-C. Mass spec-

trometry for the identification of the discriminating signals from metabolomics: Current status and future trends. *J. Chromatogr. B* **2008**, *871*, 143–163.

(9) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. *Environ. Sci. Technol.* **2018**, *52*, 4694–4701.

(10) Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: A proof of concept. *Talanta* **2019**, *195*, 426–432.

(11) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.; Slobodnik, J.; Krauss, M. High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. *Environ. Sci. Eur.* **2019**, *31*, 62.

(12) Minkus, S.; Bieber, S.; Letzel, T. Spotlight on mass spectrometric non-target screening analysis: Advanced data processing methods recently communicated for extracting, prioritizing and quantifying features. *Analytical Science Advances* **2022**, *3*, 103–112.

(13) van Herwerden, D.; O'Brien, J. W.; Choi, P. M.; Thomas, K. V.; Schoenmakers, P. J.; Samanipour, S. Naive Bayes classification model for isotopologue detection in LC-HRMS data. *Chemometrics and Intelligent Laboratory Systems* **2022**, *223*, 104515.

(14) Schulze, B.; Heffernan, A. L.; Samanipour, S.; Ramos, M. J. G.; Veal, C.; Thomas, K. V.; Kaserzon, S. L. Is Nontarget Analysis Ready for Regulatory Application? Influence of Peak-Picking Algorithms on Data Analysis. *Analytical Chemistry* **2023**, *95*, 18361–18369.

(15) Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Comparison of Software Tools for Liquid Chromatography–High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Analytical Chemistry* **2020**, *92*, 1898–1907, PMID: 31840499.

(16) Höcker, O.; Flottmann, D.; Schmidt, T. C.; Neusüß, C. Non-targeted LC-MS and CE-MS for biomarker discovery in bioreactors: Influence of separation, mass spectrometry and data processing tools. *Science of The Total Environment* **2021**, *798*, 149012.

(17) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.

(18) Alygizakis, N. A.; Oswald, P.; Thomaidis, N. S.; Schymanski, E. L.; Aalizadeh, R.; Schulze, T.; Oswaldova, M.; Slobodnik, J. NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in "digitally frozen" environmental samples. 2019.

(19) FOR-IDENT-Platform: International hunt for unknown molecules combining international workflows and software tools. `https://water.for-ident.org/#!home`.

(20) Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**, *34*, 828–837.

(21) Feraud, M.; O'Brien, J. W.; Samanipour, S.; Dewapriya, P.; van Herwerden, D.; Kaserzon, S.; Wood, I.; Rauert, C.; Thomas, K. V. InSpectra – A platform for identifying emerging chemical threats. *Journal of Hazardous Materials* **2023**, *455*.

(22) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.;

Vandergheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12*, 523–526.

(23) 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 Bioinf.* **2010**, *11*, 395.

(24) Röst, H. L. et al. OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nature Methods* **2016**, *13*, 741–748.

(25) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. patRoon: open source software platform for environmental mass spectrometry based non-target screening. *Journal of Cheminformatics* **2021**, *13*.

(26) Helmus, R.; van de Velde, B.; Brunner, A. M.; ter Laak, T. L.; van Wezel, A. P.; Schymanski, E. L. patRoon 2.0: Improved non-target analysis workflowsincluding automated transformation product screening. *Journal of Open Source Software* **2022**, *7*, 4029.

(27) Peters, K. et al. PhenoMeNal: Processing and analysis of Metabolomics data in the Cloud. *GigaScience* **2018**, *8*.

(28) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **2019**, *16*, 299–302.

(29) Shen, X.; Yan, H.; Wang, C.; Gao, P.; Johnson, C. H.; Snyder, M. P. TidyMass an object-oriented reproducible analysis framework for LC–MS data. *Nature Communications* **2022**, *13*.

(30) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Process-

ing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, *78*, 779–787.

(31) Vosough, M.; Schmidt, T. C.; Renner, G. Non-target screening in water analysis: recent trends of data evaluation, quality assurance, and their future perspectives. *Analytical and Bioanalytical Chemistry* **2024**, *416*, 2125–2136.

(32) Liu, H.; Wang, R.; Zhao, B.; Xie, D. Assessment for the data processing performance of non-target screening analysis based on high-resolution mass spectrometry. *Science of the Total Environment* **2024**, *908*.

(33) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V. Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. *Anal. Chem.* **2019**, *91*, 10800–10807.

(34) Singh, R. R.; Lai, A.; Krier, J.; Kondić, T.; Diderich, P.; Schymanski, E. L. Occurrence and Distribution of Pharmaceuticals and Their Transformation Products in Luxembourgish Surface Waters. *ACS Environmental Au* **2021**, *1*, 58–70.

(35) Bade, R. et al. Workflow to facilitate the detection of new psychoactive substances and drugs of abuse in influent urban wastewater. *Journal of Hazardous Materials* **2024**, *469*.

(36) van Herwerden, D.; O'Brien, J. W.; Lege, S.; Pirok, B. W.; Thomas, K. V.; Samanipour, S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray Ionization High-Resolution Mass Spectrometry Data. *Analytical Chemistry* **2023**, *95*, 12247–12255.

(37) Samanipour, S.; Choi, P.; O'Brien, J. W.; Pirok, B. W. J.; Reid, M. J.; Thomas, K. V. From Centroided to Profile Mode: Machine Learning for Prediction of Peak Width in HRMS Data. *Analytical Chemistry* **2021**, *93*, 16562–16570.

(38) MassBank EU. https://massbank.eu/MassBank/Index.

(39) MassBank of North America. https://mona.fiehnlab.ucdavis.edu/downloads.

(40) NIST 20. https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries.

(41) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V.; Praetorius, A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. *Environmental Science and Technology* **2023**, *57*, 17950–17958.

(42) Boelrijk, J.; van Herwerden, D.; Ensing, B.; Forré, P.; Samanipour, S. Predicting RP-LC retention indices of structurally unknown chemicals from mass spectrometry data. *Journal of Cheminformatics* **2023**, *15*, 1–12.

(43) Peets, P.; Wang, W. C.; Macleod, M.; Breitholtz, M.; Martin, J. W.; Kruve, A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. *Environmental Science and Technology* **2022**, *56*, 15508–15517.

(44) Sepman, H.; Malm, L.; Peets, P.; MacLeod, M.; Martin, J.; Breitholtz, M.; Kruve, A. Bypassing the Identification: MS2Quant for Concentration Estimations of Chemicals Detected with Nontarget LC-HRMS from MS2 Data. *Analytical Chemistry* **2023**, *95*, 12329–12338.

(45) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. 2014.

TOC Graphic