Hochschule Aalen

AriumMS Manual

Version 1.0

About AriumMS

Copyright (c) 2022, Adrian Haun
AriumMS is an open-source MATLAB application for processing, augmentation, visualization and feature extraction of metabolomic mass spectrometry data. Data is processed in groups. Groups can be different analytical techniques (e.g. liquid chromatography (LC) and capillary electrophoresis (CE)), samples (e.g. different growth conditions), or both. It does not matter which source these files come from, separation technology and mass spectrometer can be freely combined. When this manual refers to chromatograms, it also refers to electropherograms.

Reporting Bugs We appreciate any bug reports, encountered problems and feature requests from users of AriumMS.
Please report any issues to https://github.com/AdrianHaun/AriumMS/issues or contact the developer adrian.haun@hs-aalen.de

License This MATLAB application is free software under BSD 3-Clause "New" or "Revised" License. You can redistribute with or without modifications under the terms of this license. It is distributed "as is" and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed.

Contents

About AriumMS I

Table of contents II

Acronyms IV

1 Installation and starting 1
- 1.1 Hardware requirements . . . 1
- 1.2 System requirements . . . 1
- 1.3 Installation of AriumMS . . . 2
- 1.4 Starting AriumMS . . . 2

2 General workflow 4

3 MS Files 5
- 3.1 Supported file format . . . 5

4 AriumMS Main App 5
- 4.1 Main Options . . . 6
- 4.2 Data Selection . . . 7
- 4.3 ROI Parameters . . . 9
- 4.4 Blank Subtraction . . . 11
- 4.5 Alignment Parameters . . . 11
- 4.6 Baseline Correction Parameters . . . 16
- 4.7 Smoothing Parameters . . . 19
- 4.8 Internal Standard Normalization Selection . . . 21
- 4.9 Isotope Filter . . . 22
- 4.10 Adduct & Neutral Loss Filter . . . 22
- 4.11 Contaminant Filter . . . 24
- 4.12 Additional Options Panel . . . 25
- 4.13 Preview Parameters . . . 27

4.14 Batch Processing 27

5 AriumMSEval 28
5.1 Group Assignment 29
5.2 Internal standard (ISTD) Processing Options 30
5.3 Data Pre-Treatment 31
5.4 Feature Sorting Parameters 32
5.5 Sample Specific Scaling Factors 33
5.6 Feature Extraction 34
5.7 Data Visualization 36
5.7.1 Plot TICs 37
5.7.2 Entropy and S/N distribution 38
5.7.3 Rejected Features 39
5.7.4 Scatter Plot 40
5.7.5 Vulcano Plot 41
5.7.6 Cloud Plot 42
5.8 Multivariate Data Analysis 43
5.8.1 Heatmap 43
5.8.2 PCA 44
5.9 Export Results 46

Acronyms

BPC	Base peak chromatogram
CE	capillary electrophoresis
EIC	Extracted ion chromatogram
ISTD	Internal standard
LC	liquid chromatography
MS	mass spectrometry
OS	Operating System
PC	Principle component
PCA	Principle component analysis
ROI	Region of interest
RT	Retention time
TIC	Total ion chromatogram

1 Installation and starting

1.1 Hardware requirements

Minimum requirements: Any Intel or AMD x86–64 processor and 4 GB RAM

Recommended requirements: Any Intel or AMD x86–64 processor with four logical cores and AVX2 instruction set support and 4 GB RAM per core.
A hardware accelerated graphics card supporting OpenGL 3.3 with 1GB GPU memory.
An SSD is strongly recommended.

1.2 System requirements

Operating System (OS)

Windows: 11, 10 (version 20H2 or higher), Server 2019 or Server 2022

macOS: Monterey (12) or Big Sur (11.6)

Linux: Ubuntu 22.04 LTS, Ubuntu 20.04 LTS, Ubuntu 18.04 LTS, Debian 11, Debian 10, Red Hat Enterprise Linux 9, Red Hat Enterprise Linux 8 (minimum 8.4), Red Hat Enterprise Linux 7 (minimum 7.9), SUSE Linux Enterprise Desktop 15, SUSE Linux Enterprise Server 12 (minimum SP2), SUSE Linux Enterprise Server 15

Software prerequisites An installation of MATLAB R2022a (or newer) with Parallel Computing Toolbox, Bioinformatics Toolbox, Statistics and Machine Learning Toolbox, Signal Processing Toolbox, Wavelet Toolbox and Image Processing Toolbox

1.3 Installation of AriumMS

1. Download AriumMS here: https://github.com/AdrianHaun/AriumMS
2. Start MATLAB, in the Apps tab click "Install App" and open the file AriumMS.mlappinstall or open AriumMS.mlappinstall directly, this automatically opens MATLAB and installs the app.

1.4 Starting AriumMS

In MATLAB go to the Apps panel and click on AriumMS or type `AriumMS` into the command window and press enter.

Note that AriumMS tries to process as many files as possible at the same time through parallelization. Due to the size of mass spectrometry (MS) files, large amounts of memory can be required, causing Out of Memory errors when available memory is exceeded. To avoid this problem, the number of workers (cores) can be reduced in MATLAB under Preferences → Parallel Computing Toolbox.

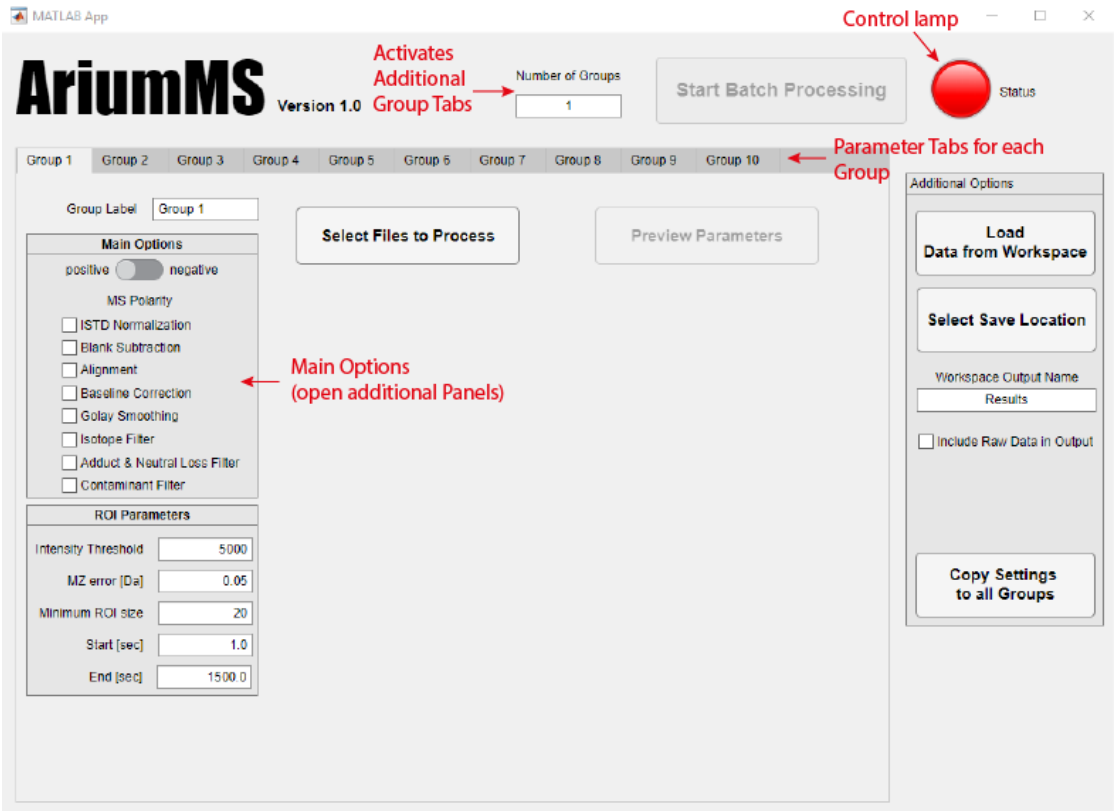


Figure 1: Initial User interface. Number of Groups controls the active tabs, the check boxes in the main option panel activate additional options. The control lamp shows if all options are set correctly for each active group.

2 General workflow

Generally, a workflow runs as follows:

1. Specify the number of groups
2. Select files for group 1
3. Preview the first file
4. Activate options and filters in the main options panel
5. Specify each parameter for each activated option
6. Check selected parameters with the Preview Parameters button (Note only ROI parameters, Baseline Correction, Golay Smoothing and contaminant filters are applied here)
7. Copy settings to all groups or set each group individually
8. Select files for remaining groups
9. Specify Save location and Workspace output
10. Start batch processing
11. Set feature extraction and scaling parameters
12. Specify reference groups and group augmentation
13. Extract features
14. Generate plots
15. Export results

3 MS Files

3.1 Supported file format

AriumMS can read .mzXML files in both profile or centroided modes. Files containing fragmentation (MS^n) scans can be used, but only MS^1 scans are evaluated. Vendor file formats can be converted to .mzXML using MSconvert (http://proteowizard.sourceforge.net/download.html)
Files should be converted using Write Index and TPP compatibility but **without zlib compression**. 32-bit encoding precision is recommended due to smaller data size and faster processing.

4 AriumMS Main App

The main app is used to select the files to be Region of interest (ROI) searched and to define the processing parameters for all groups. The ***Start Batch Processing button*** is deactivated as long as the necessary entries have not yet been made for activated actions. These include: Blank Subtraction (no blank files selected), ISTD normalization (information on internal standards incomplete) and Adduct & Neutral Loss filter (species filter not yet selected).
Control lamps in the corresponding panels indicate whether entries are still missing.
By pressing the ***Start Batch Processing button***, the files are automatically processed in groups with the corresponding parameters. Once processing is complete, the ROIs are transferred to AriumMSEval and output to the MATLAB workspace.

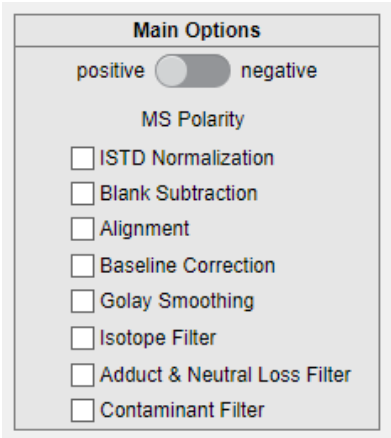
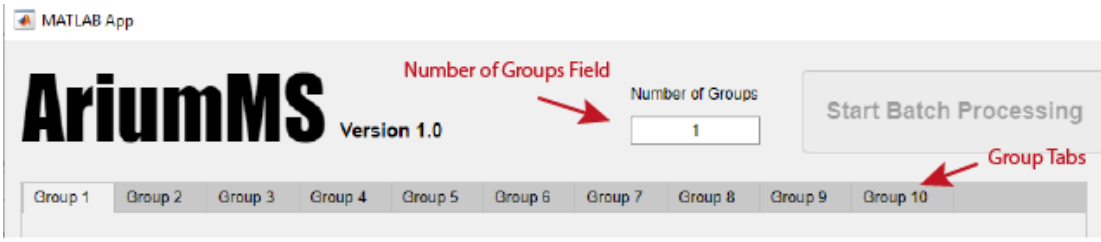


Figure 2: Main Options panel. Set the MS polarity and which preprocessing steps / filters should be used for the selected group.

4.1 Main Options

By activating the checkboxes, additional panels with the corresponding parameters of this processing step become visible.



The ***Number of Groups*** field is a numeric input field for specifying the number of groups. Allowed values 1-10. According to this number additional group tabs will be activated. The name of the group can be changed in each tab with the ***Group Label*** field.

Attention: Each group must be given an individual alphanumeric name. If duplicate group names are detected, the new name will not be applied.

4.2 Data Selection

Press the button ***Select files to Process***. A file explorer window opens in which all files for this group can be selected.
If blank subtraction is activated for this group, an identical window for the selection of the blanks opens after the selection of the measurement files.
The number of scans, start time and end time of the selected files are checked. If one of the values deviates by 30%, a warning is displayed.
Attention: By selecting the data, the start and end time in the ROI parameter panel is set to the limits of the shortest selected file. If these parameters were changed before selecting the files, the entries will be overwritten.
If groups are to be augmented, they must contain the same number of sample files.

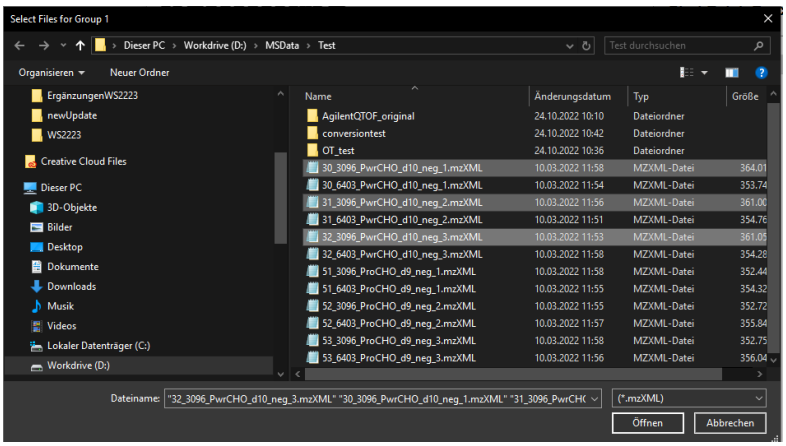


Figure 3: File Explorer for selection of data and blank files. Design depends on the OS.

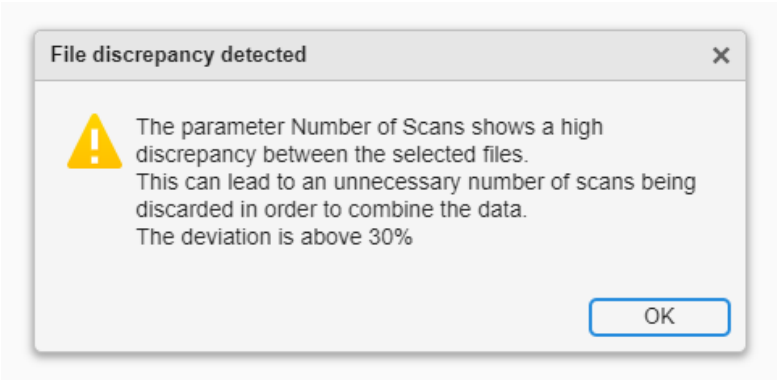


Figure 4: Warning if a file discrepancy is detected. Often caused by incorrect sorting or by MS^1 and MS^2 measurements being accidentally selected at the same time.

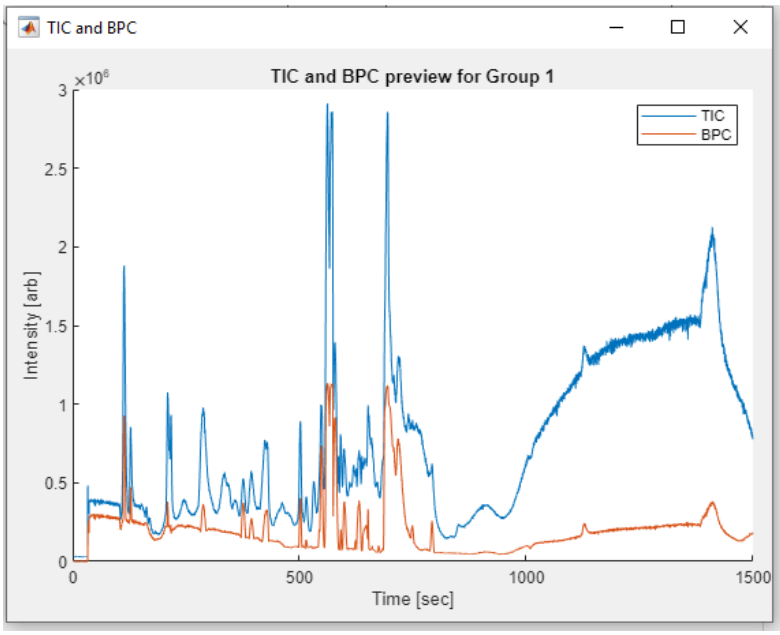


Figure 5: File Preview window showing TIC and BPC for the first selected file.

It is then possible to display a preview for the selected files. The preview is the Total ion chromatogram (TIC) and Base peak chromatogram (BPC) of the first file. MS files within a group should be measured with the same measurement conditions, usually biological replicates or measurement replicates.

4.3 ROI Parameters

The first step of the processing is the ROI search. ROI is a data compression without the loss of relevant information. All intensities below a limit and with to few occurrences are removed.
ROIs can be considered equivalent to Extracted ion chromatogram (EIC)s. By pressing the ***Preview Parameters button***, the currently selected parameters for the selected group can be tested.

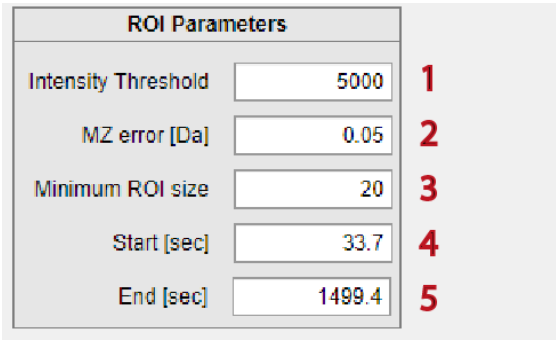


Figure 6: ROI Parameters panel

Method parameters

1. Intensity Threshold
   Minimum intensity of a m/z to be accepted as ROI. Great influence on the number of features found. A value that is too small leads to a massive number of false positive ROIs and thus to very long processing times. A value that is too high filters out too many real features. Guideline: Set this to 15-30% of minimum baseline intensity.

2. MZ error
   m/z range in Da in which ROI masses are considered to be inside the same ROI. A value that is too small leads to false positive ROIs, a value that is too large leads to a decrease in the MS mass resolution. Guideline: Set this to the mass error of the used mass analyzer.

3. Minimum ROI size
   Minimum number of scans a m/z must be over the intensity threshold to be considered as ROI. A value that is too small leads to a high number of false positive ROIs. A value that is too high filters out too many real features. Guideline: Set this to the number of scans of the smallest observable peak.

4. Start
   First scan to be processed. Previous scans will be discarded.

5. End
   Last scan to be processed. Subsequent scans will be discarded.

The limits for Start and End are automatically adjusted after file selection.

4.4 Blank Subtraction

The blank goes through the same processing steps as the sample files. the Blank ROI matrix obtained is subtracted from the sample ROI matrices. Thus, interferences are subtracted in a time-resolved manner. If several blank files are selected, an average is calculated.

4.5 Alignment Parameters

By activating the ***Alignment Checkbox***, the ***Alignment Parameters panel*** becomes active. Two different alignments can be applied here. Spectral alignment and Peak alignment.

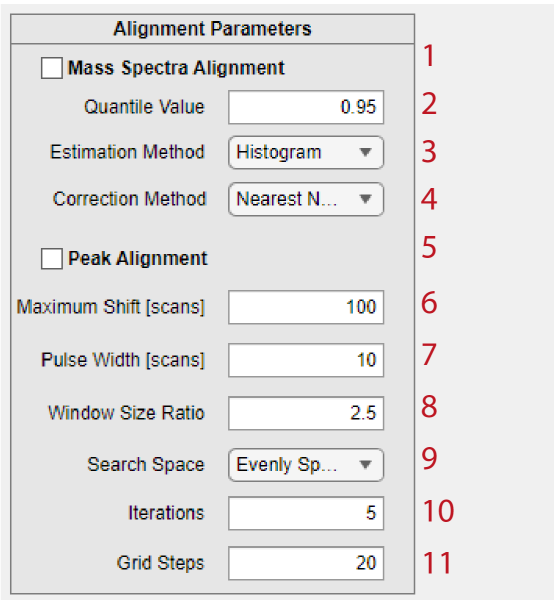
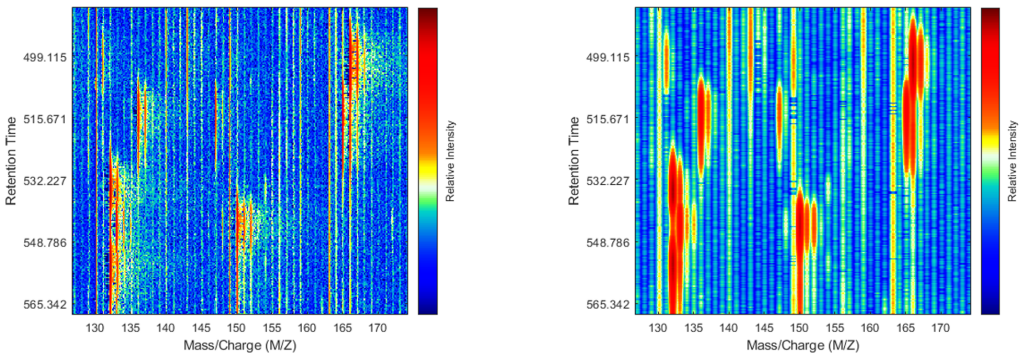


Figure 7: Alignment Parameters panel

With spectral alignment, all scans within one MS file are examined for common m/z values and then all scans are aligned to these m/z values.



(a) Before alignment

(b) After alignment

Figure 8: Comparison of m/z traces before and after spectral alignment with default values

Method parameters

1. Mass Spectra Alignment checkbox
   Applies the Mass spectra alignment processing step to this group. Recommended only for low resolution MS files due to high computational requirements.

2. Quantile Value
   Value that determines which peaks are selected by the estimation method to create the list of common m/z values. Choices are any value ≥ 0 and ≤ 1. Default value is 0.95.

3. Estimation Method

   Method for the estimation of the common m/z list, choices are:

   Histogram: Peak locations are clustered using a kernel density estimation approach. The peak ion intensity is used as a weighting factor. The center of all the clusters conform to the common m/z list. This option is less computational intensive, but causes a sharp reduction in mass resolution for data from high-resolution mass spectrometers.

   This is the default option.

   Regression: Takes a sample of the distances between observed significant peaks and regresses the inter-peak distance to create the common m/z list with similar inter element distances. This option has a higher computation time but causes no or significantly less reduction in mass resolution.

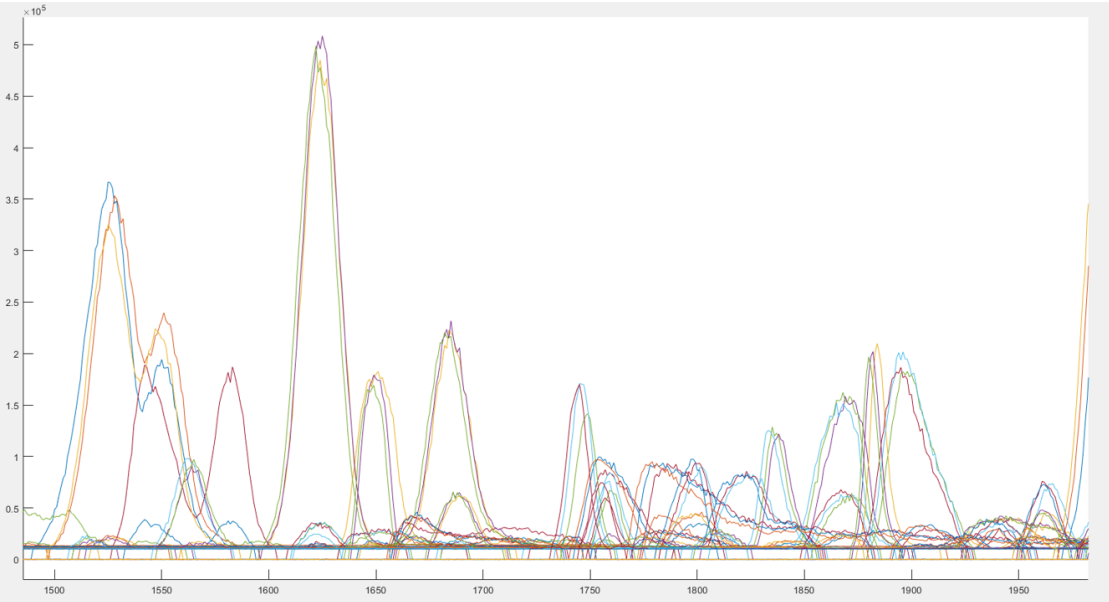
4. Correction Method

   Method to align each peak list to the common m/z list, choices are:

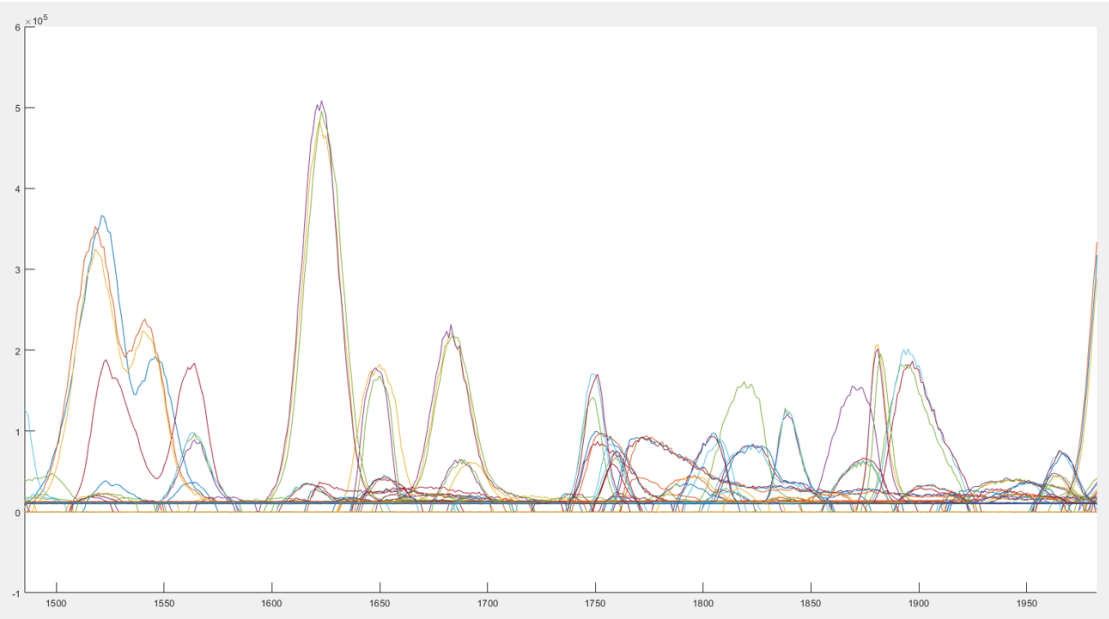
   Nearest neighbor: For each common peak in the common m/z list, its counterpart in each peak list is the peak that is closest to the common peak's m/z value. This is the default option.

   Shortest path: For each common peak in the common m/z list, its counterpart in each peak list is selected using the shortest path algorithm.

   In peak alignment, ROIs within a group are shifted by shifting the time axis so that they correlate maximally with a synthetic Gaussian peak created at the maximum intensity of all ROIs considered.



(a) Before alignment



(b) After alignment

Figure 9: ROIs before and after peak alignment with default values

5. Peak Alignment checkbox
   Applies the Peak alignment processing step to this group.

6. Maximum Shift
   Specifies the limits of the range ($\pm$x/2), in separation units (scans), relative to each peak. No peak shifts beyond these limits. Ideally, you should keep the range within the maximum expected shift. If you try to correct larger shifts by increasing the limits, you increase the possibility of picking incorrect peaks to align to the reference. Default value is 100.

7. Pulse Width
   Width of the synthetic Gaussian peak in separation units (scans). Tuning the spread of the Gaussian pulses controls a trade off between robustness (wider pulses) and precision (narrower pulses). However, the spread of the pulses is unrelated to the shape of the observed peaks in the signal. The purpose of the pulse spread is to drive the optimization algorithm. Default value is 10.

8. Window Size Ratio
   Specifies a scaling factor that determines the size of the window around every alignment peak. The synthetic signal is compared to the sample signal only within these regions, which saves computation time. Default value is 2.5.

9. Search Space
   Specifies the type of search space. Choices are:
   Evenly spaced lattice, this is the default option.
   Random Latin Hypercube with Grid Steps^2 samples.

10. Iterations
    Positive integer that specifies the number of refining iterations. At every iteration, the search grid is scaled down to improve the estimates. Default value is 5.

11. Grid Steps
    Positive integer that specifies the number of steps for the search grid. At every iteration, the search area is divided by Grid Steps^2. Default value is 20.

4.6 Baseline Correction Parameters

The baseline correction is performed as follows. Possible baseline points are determined in several shifted windows with fixed width, then a curve is regressed through all determined baseline points. This determined baseline is then removed while preserving peak heights.

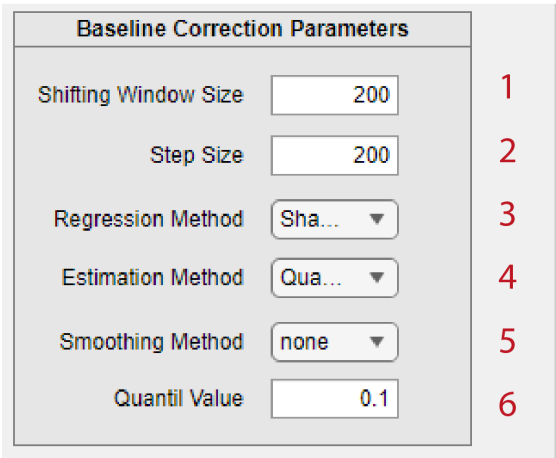
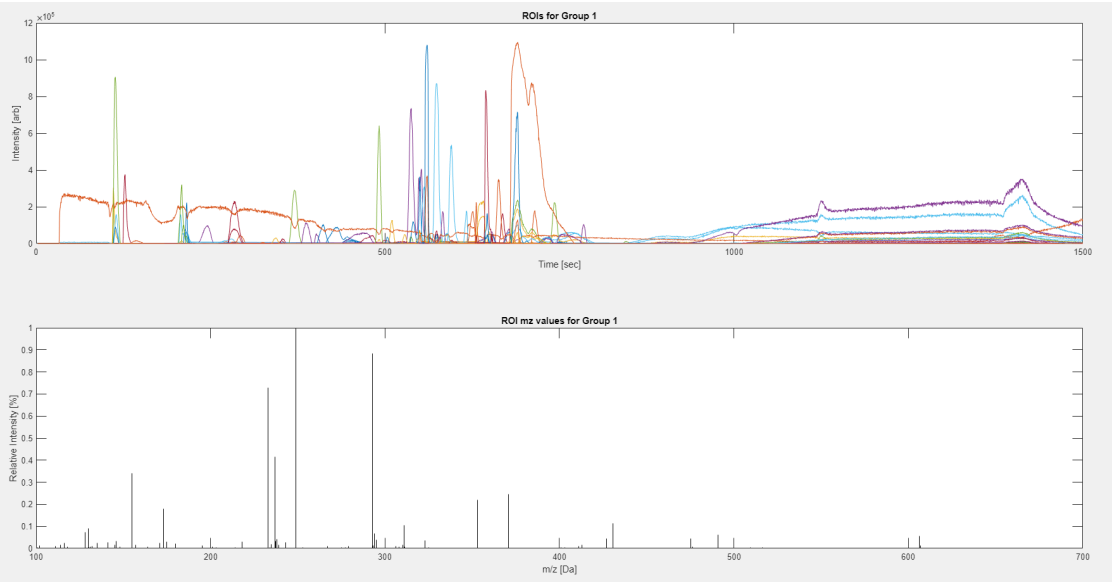
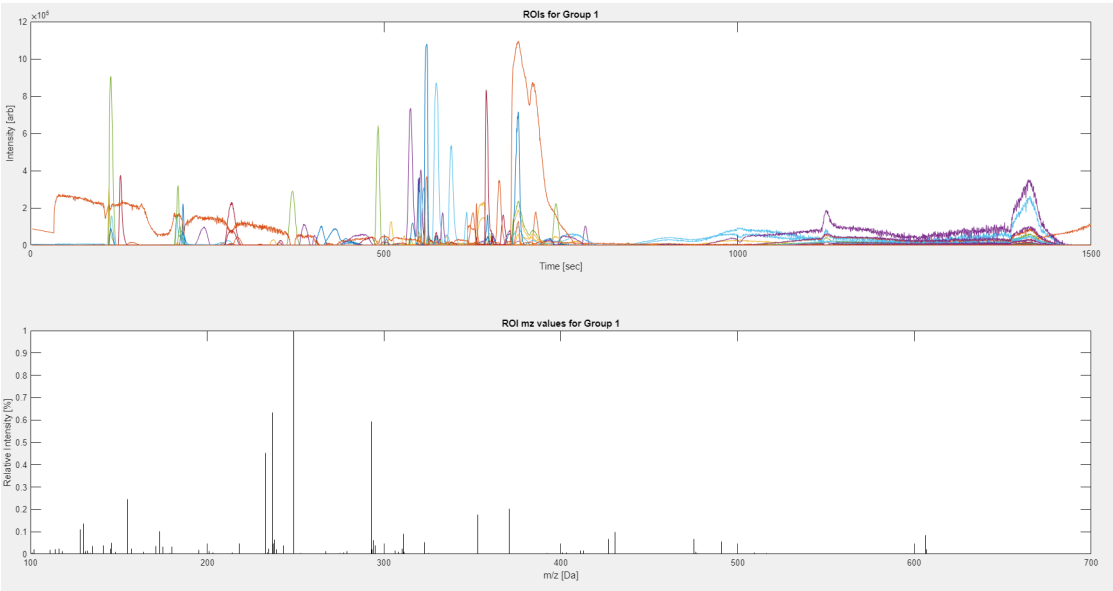


Figure 10: Baseline Correction Parameters panel



(a) Before baseline correction



(b) After baseline correction

Figure 11: ROIs before and after baseline correction with default values

Method parameters

1. Shifting Window Size
   Width of the shifting window in separation units (scans). Define the parameters based on the width of your peaks in the signal and the presence of possible drifts. Default value is 200.

2. Step Size
   Step size for the shifted window in separation units (scans). Default value is 200.

3. Regression Method
   Method of regression through the estimated baseline points. Choices are: **Shape-preserving piece wise cubic interpolation**
   The interpolated value at a query point is based on a shape-preserving piece wise cubic interpolation of the values at neighboring grid points. This is the default option.
   Linear interpolation
   The interpolated value at a query point is based on linear interpolation of the values at neighboring grid points in each respective dimension.
   Spline interpolation
   The interpolated value at a query point is based on a cubic interpolation of the values at neighboring grid points in each respective dimension.

4. Estimation Method
   Method for determining likely baseline points, choices are: **Quantile**
   Baseline points are all values that lie on the quantile that is defined by the parameter quantile value. This is the default option. **Estimation-Maximization**
   Every sample is the independent and identically distributed (i.i.d) draw of any of two normal distributed classes (background or peaks). Because the class label is hidden, the distributions are estimated with an Expectation-Maximization algorithm. The ultimate baseline value is the mean of the background class.

5. Smoothing Method
   Method to smooth the curve of estimated points, choices are: none, linear fit, quadratic fit, robust linear fit and robust quadratic fit.

6. Quantile Value
   Determines which values are used as baseline points in the quantile estimation. Choices are any value ≥ 0 and ≤ 1. Default value is 0.1.

4.7 Smoothing Parameters

ROIs are smoothed according to the Savitzky-Golay method. Here, the point under consideration is adjusted by regression of a polynomial of the n-th degree over a moving window.

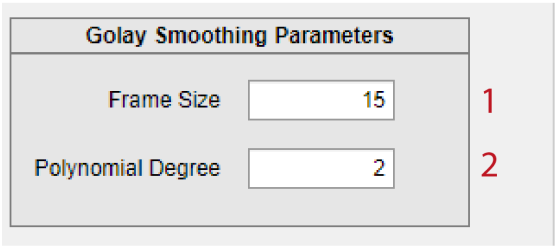
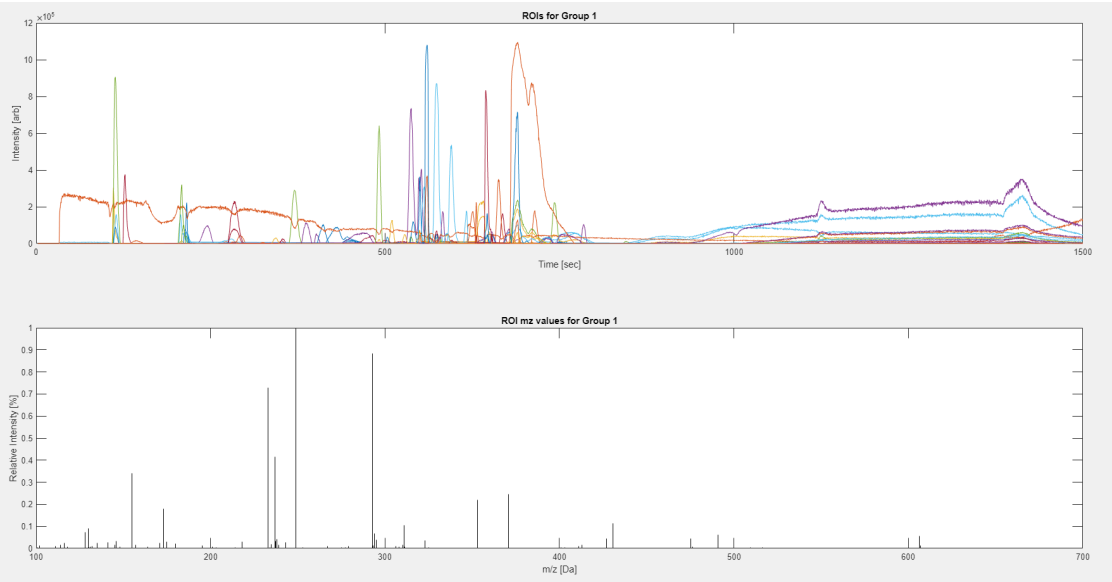
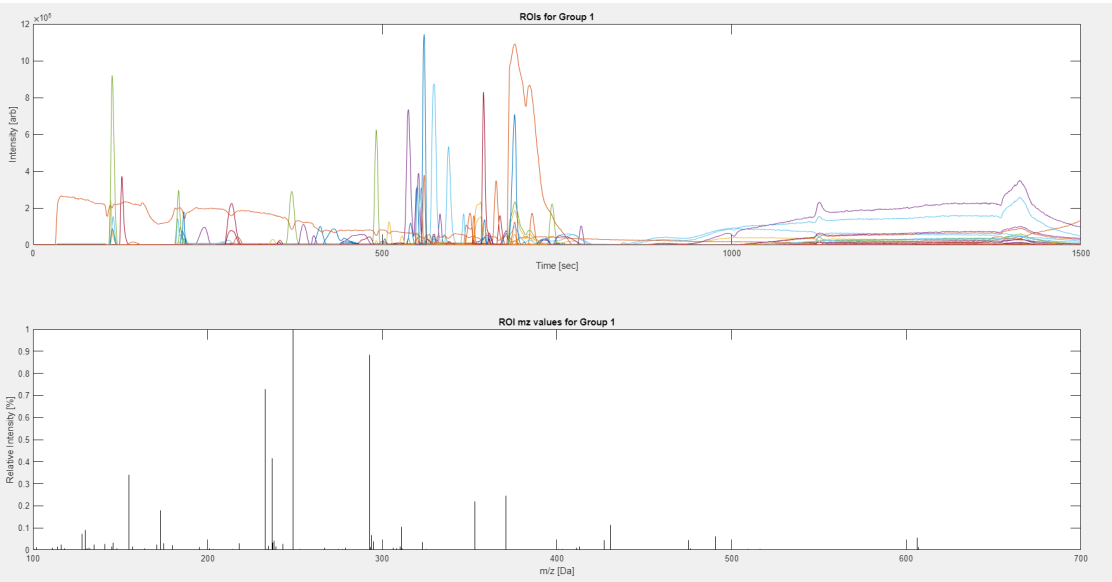


Figure 12: Golay Smoothing Parameters panel



(a) Before smoothing



(b) After smoothing

Figure 13: ROIs before and after Savitzky-Golay smoothing with default values

Method parameters

1. Frame size
   Moving frame size (scans), over which the polynomial is fitted. Higher values smooth the signal more with an increase in computation time. The default value is 15.

2. Polynomial Degree
   Specifies the degree of the polynomial fitted to the points in the moving frame. This value must be smaller than Frame size. The default value is 2.

4.8 Internal Standard Normalization Selection

The ISTD is used to normalize peak intensities and calibrate the mass dimension. Only activate if all defined ISTDs are present in the specified retention time window in all selected Data files of the selected group. If an ISTD is not found in one of the group files, it is not used. To select how the ISTD is used see 5.2 on page 30.

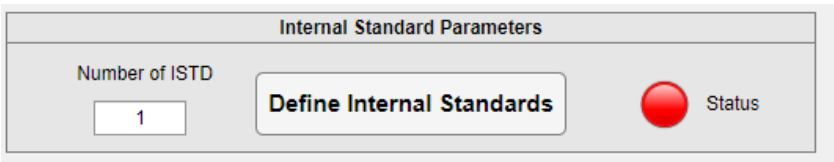


Figure 14: ISTD Parameters panel

In the ***internal standard parameters panel*** select the number of internal standards for this group. The status lamp indicates whether the internal standards have been defined for this group.
By clicking the ***Define Internal Standards button*** a window opens to define the internal standards. Required are the expected m/z, the expected Retention time (RT) and the acceptable RT drift for all ISTD

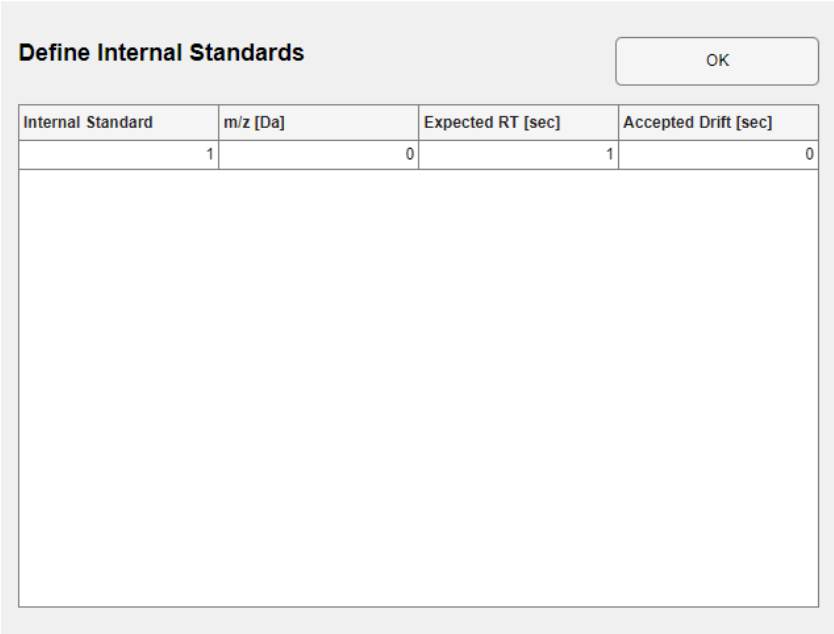


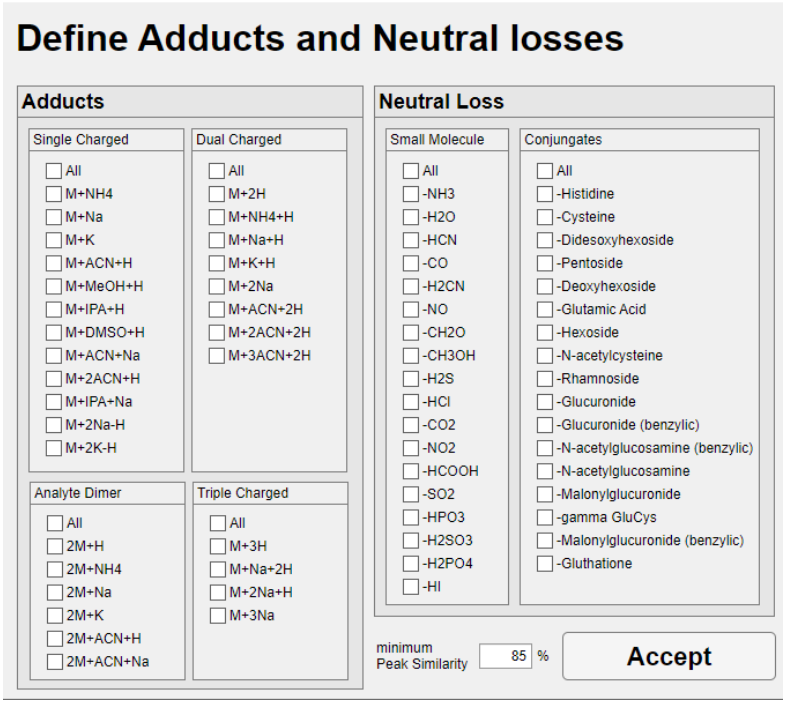
Figure 15: Define Internal Standards window

4.9 Isotope Filter

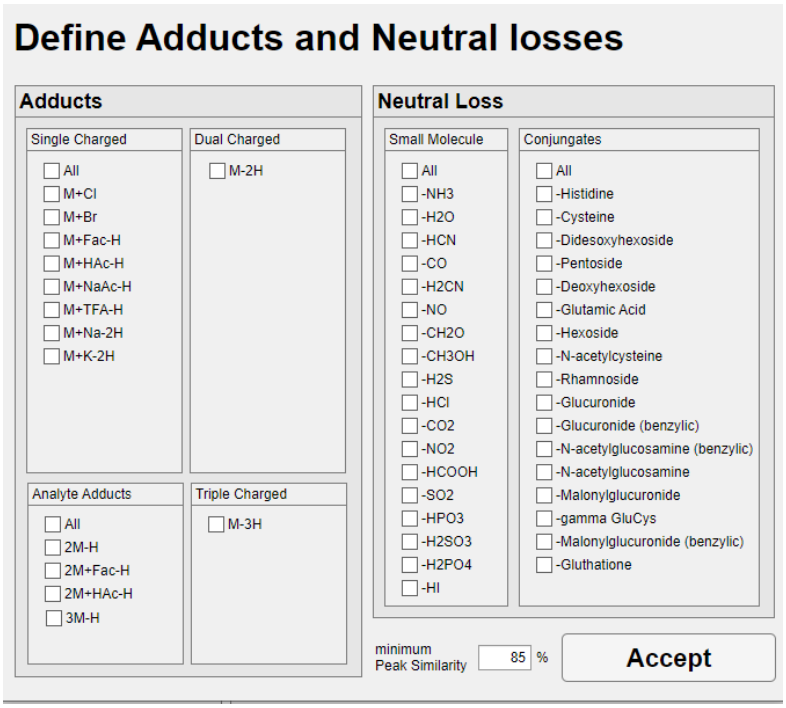
Selecting this filter removes possible isotope peaks. Candidates must have matching m/z within a certain tolerance (m/z tolerance defined in AriumMSEval), the isotope peak must have a cosine similarity of $\geq$ 85 % with the suspected non isotope peak and the intensity of the isotope peak must be within the expected intensity based on the natural abundance of the non isotope peak, within 10 %. Considered isotopes are: ^{13}C, ^{15}N, ^{18}O, ^{33}S and ^{34}S.

4.10 Adduct & Neutral Loss Filter

Selecting this filter removes possible adduct and or neutral loss peaks from the final peak list. The kind of adducts or neutral loss fragments to consider can be specified. Candidates must have matching m/z within a certain tolerance (m/z tolerance defined in AriumMSEval) and the adduct or neutral loss peak must have a cosine similarity grater than the specified peak similarity, with the suspected non adduct peak.



(a) Available positive Adducts & Neutral loss species



(b) Available negative Adducts & Neutral loss species

Figure 16: Adduct & Neural Loss windows

By ticking ***All***, all species in the respective panel will be toggled. The required peak similarity (calculated as cosine similarity) can be specified here. The default value is 85 %.
The status lamp in the ***Adduct & Neutral Loss Parameters panel*** indicates whether the species to filter have been defined for this group.

4.11 Contaminant Filter

Found ROIs are compared to a list of common mass spectrometry contaminants. ROIs with m/z within ±MZ error/2 are completely removed. Considered contaminants are from the UWPR Common Mass Spec Contaminants spreadsheet: https://proteomicsresource.washington.edu/protocols05/esi_background_ions_repeat_units.php

4.12 Additional Options Panel

This panel contains quality of life options.

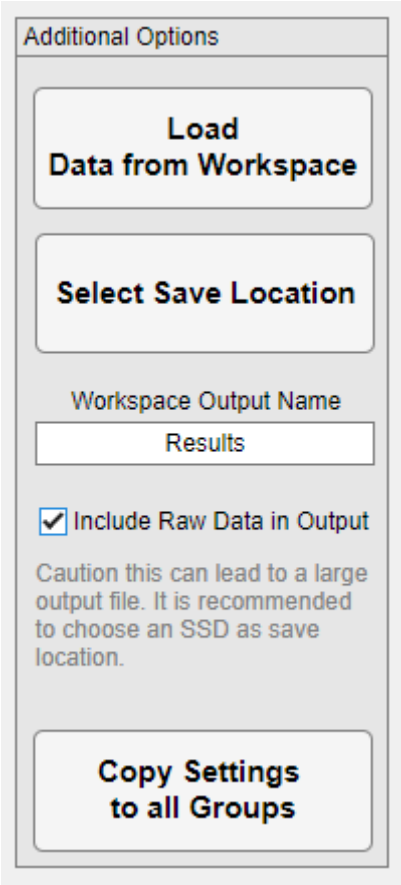


Figure 17: Additional Options panel

Load Data from Workspace allows you to load settings from previous sessions. **Caution**, the selected files are also selected. If the corresponding file is no longer in this path or has been renamed, an error occurs during batch processing.

Select Save Location allows you to select the file path and file name under which the result of batch processing is saved as a .mat file. In addition, the result is also output to the MATLAB workspace.
The name can be changed with the field ***Workspace Output Name***.
Caution: If the workspace already contains a variable with the same name, it will be overwritten. In addition, the output name must be compatible with the MATLAB variable naming scheme. (alphanumeric characters and _ allowed, starting with a letter, maximum 64 characters)

It is possible to save the raw data with the output, so that the most time-consuming step, importing the MS files into MATLAB, can be skipped for repeated processing with different parameter settings. Use the ***Include Raw Data in Output checkbox*** to activate this option.
Caution: The resulting Output files can take up several GB of disk space. Make sure you have enough space, and store the output on an SSD if possible.

Copy Settings to all Groups applies all parameters and options of the current selected group to all other groups. The already selected files are not affected by this.

4.13 Preview Parameters

By pressing the ***Preview Parameters button***, the preprocessing steps are applied to the first file of the selected group and the results are plotted.

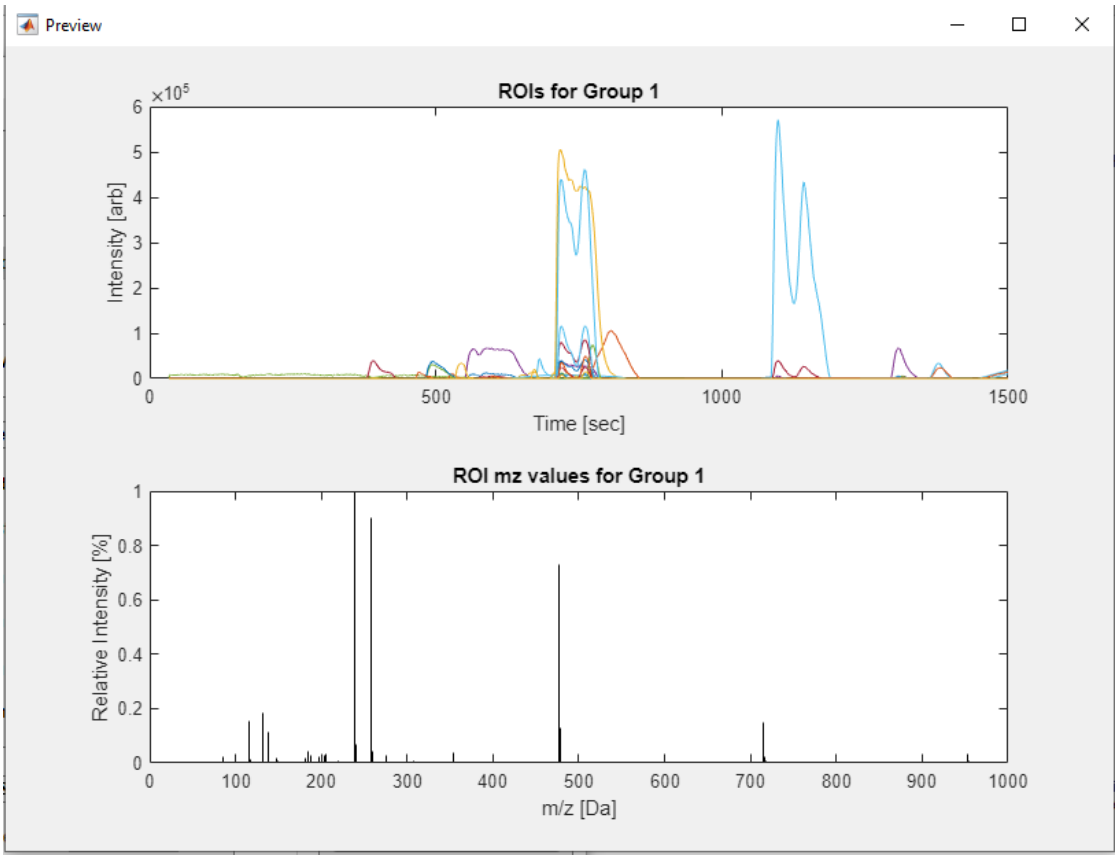


Figure 18: Parameter preview window. In the upper plot, the obtained ROIs are displayed. The lower plot contains the found m/z of the ROIs, the intensity is normalized to the most intensive one.

4.14 Batch Processing

As soon as all parameters are set correctly, the ***Start Batch Processing button*** becomes active. All files are processed in groups. Afterwards the output will be sent to the MATLAB workspace, and additionally saved as .mat file if the corresponding option is selected. AriumMSEval will open automatically.

5 AriumMSEval

After processing in AriumMS, the results are still stored as a ROI matrices. In this matrix, peaks must now be identified and integrated. This means transferring the data from the ROI level to the feature level. This is the main task of AriumMSEval. In addition, normalization, scaling, transformation, augmentation and statistics are performed here.

AriumMSEval opens automatically after AriumMS processing is finished. Alternatively, AriumMSEval can be started from the command line, this requires results from AriumMS as input argument. Load saved results from AriumMS into MATLAB workspace by opening the corresponding .mat file. Then open AriumMSEval by the command `AriumMSEval(Results)`. The name in the parenthesis must match the name of the result variable currently in the MATLAB workspace.

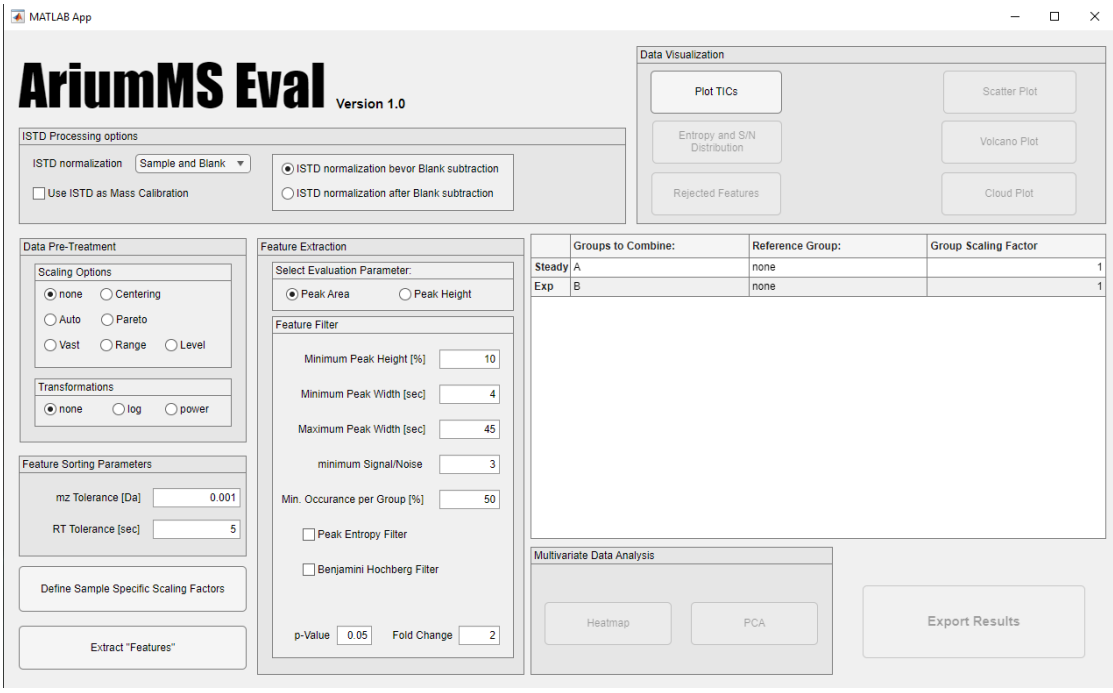


Figure 19: AriumMSEval initial window

All settings of AriumMSEval are applied to all groups in the same way.

5.1 Group Assignment

The table defines how the individual groups are related to each other. Which groups should be augmented (e.g. same sample, different measurement techniques) or which group should be used as normalization (e.g. effect cell line should be normalized to MOCK cell line).

	Groups to Combine:	Reference Group:	Group Scaling Factor
SteadyCE	A	none	1
ExpCE	B	A	2
SteadyLC	A	none	1
ExpLC	B	A	2

Figure 20: Group Assignment Table with example assignment. Groups SteadyCE and SteadyLC are assigned to A and will be augmented. ExpCE and ExpLC are assigned to B and will be augmented. In addition, the augmented group B is normalized to A. Both Exp groups are scaled by a factor of 2.

1. Groups to Combine:
   Groups with the same assignment in this column are augmented to one group.
2. Reference Group:
   Here you can select which group should be used as reference for this group. The assignment can be changed using the dropdown menu in the corresponding cell.
3. Group Scaling Factor:
   Scaling factor that is applied to all features of this group. E.g. different sample quantities

Caution: When selecting the groups for augmentation, they must contain the same number of measurements. In addition, the same number of groups must always be combined.

5.2 ISTD Processing Options

Here you can set the type of ISTD normalization. Use the dropdown menu to select between sample and blank, sample only or no normalization. In addition, whether the normalization is to take place before or after the blank value is subtracted. Likewise, the masses can be calibrated by the internal standard. Only applied to groups for which ISTD normalization has been activated in AriumMS.

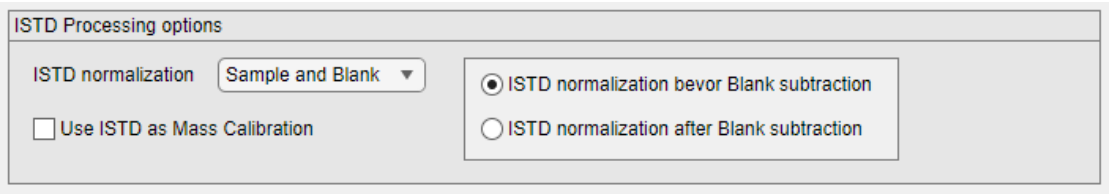


Figure 21: ISTD Processing options

If several ISTDs are defined for a group, the intensities are normalized to the internal standard that is closer in time. If the masses are to be calibrated by the internal standard, the same applies to the mass dimension. If an ISTD is not found in a sample (either no suitable m/z or not within the retention time window), this internal standard is not used for calibration within the group.

5.3 Data Pre-Treatment

Several preprocessing steps are available to scale or transform feature intensities.

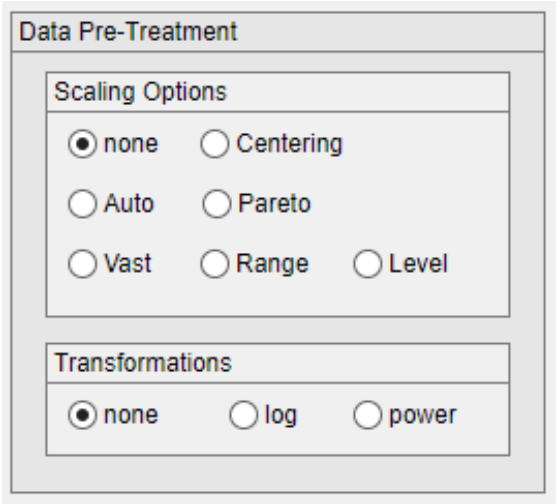


Figure 22: Data Pre-Treatment panel

Scaling Options Features are independently scaled within a group. Choices are:

1. none
2. Centering: Set feature mean to 0
3. Auto: Center and scale by standard deviation
4. Pareto: Center and scaly by square root of the standard deviation
5. Vast: Auto scale multiplied by mean divided by standard deviation
6. Range: Center and scale by the range of values
7. Level: Center and scale by mean

Transformation Options Features are independently transformed within a group. Choices are:

1. none: No transformation is performed
2. log: The $\log_{10}$ is formed and then centered
3. power: The square root is formed and then centered

5.4 Feature Sorting Parameters

These settings determine the tolerance with which features are sorted after integration. If the limits are too narrow, the same feature can be separated into different features (different mass or retention time) in repeated measurements. If the tolerances are set too high, different features can be incorrectly assigned to one feature.

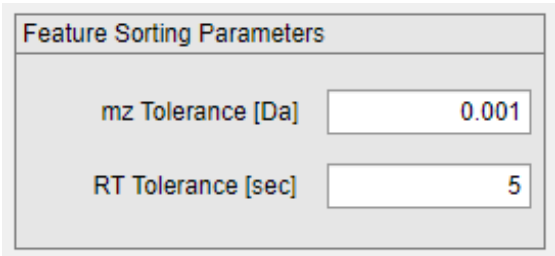
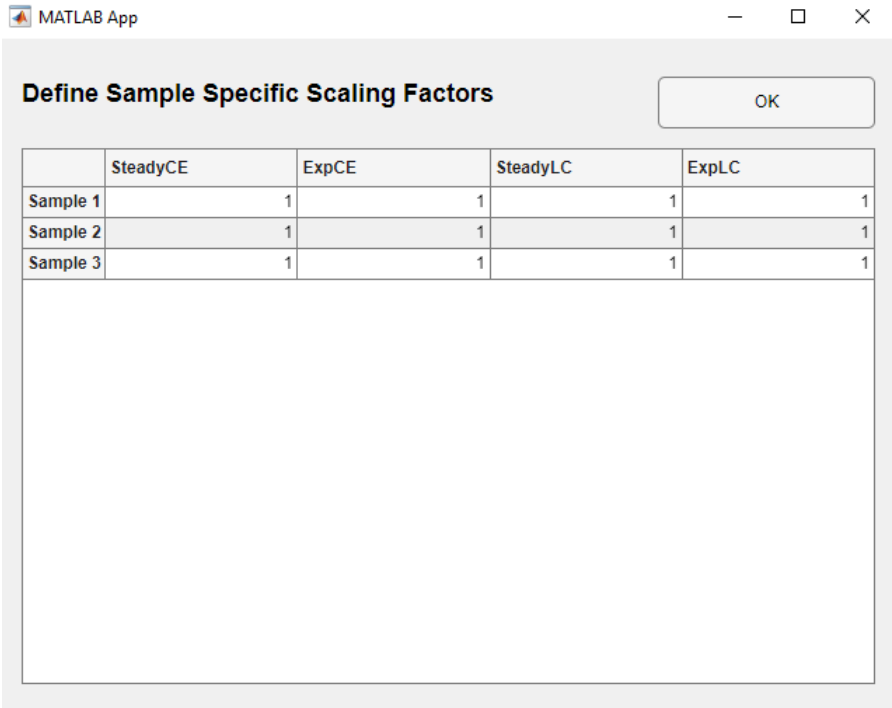


Figure 23: Feature sorting parameters

The m/z tolerance should correspond approximately to the mass accuracy of the MS used, and the retention time tolerance to the expected retention time drift.

5.5 Sample Specific Scaling Factors

By pressing the ***Define Sample Specific Scaling Factors button*** a new window opens. Specific scaling factors can be entered here for each sample. For example, the cell number for biological replicates to normalize biological replicates between each other.



	SteadyCE	ExpCE	SteadyLC	ExpLC
Sample 1	1	1	1	1
Sample 2	1	1	1	1
Sample 3	1	1	1	1

Figure 24: Sample specific scaling factor window

5.6 Feature Extraction

Within the ROIs found, peaks must be found, the retention time determined and integrated. Only then can we use the term "features".

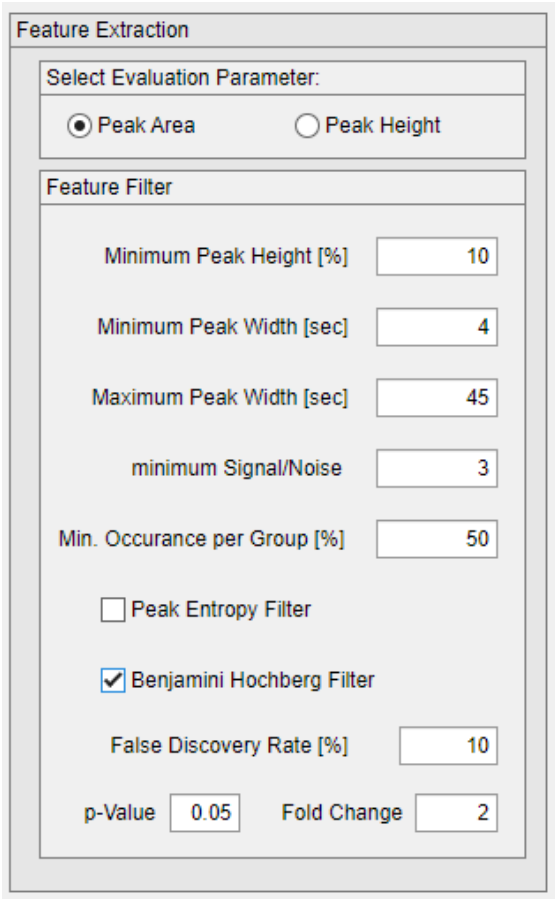


Figure 25: Feature extraction parameters

Evaluation Parameter Here you can select whether peak areas or peak heights are evaluated.

Feature Filter The number of false positive features can be reduced by using the following filters:

1. Minimum Peak Height
   The largest peak within the same m/z is set as 100 %.
2. Minimum Peak Width
   Removes mainly high frequency noise
3. Maximum Peak Width
   Removes mainly background signal
4. minimum Signal/Noise
   The noise is estimated for each ROI, by the standard deviation of the delta between the smoothed ROI and the original ROI.
5. Min. Occurrence per Group
   Removes features that occur in too few repeated measurements.
6. Peak Entropy Filter
   Determines whether the peak found is a real peak or noise by determining the entropy according to Shannon (information theory). Peaks with a larger entropy than the median entropy of the group are removed.
7. Benjamini Hochberg Filter
   Adjusts the detected p-values under the assumption that a part of the p-values are declared false positive as significant (false positive rate). Once the filter is selected, the false discovery rate can be adjusted.
8. p-Value
   The p value of a feature is determined by a two-sample t-test between groups. Features with a larger p-value are not considered significant. Also plotted as boundary line in the vulcano plot.
9. Fold Change
   The fold change filter determines the boundary that is displayed in the vulcano plot.

To start feature extraction press the ***Extract Features button***. All files are searched for features, identified peaks are integrated, the selected filter scalings are applied and groups are augmented and statistics are calculated.

5.7 Data Visualization

When starting AriumMSEval only the button ***Plot TICs*** is active. All others are available after feature extraction.

Figure 26: Data Visualization panel.

5.7.1 Plot TICs

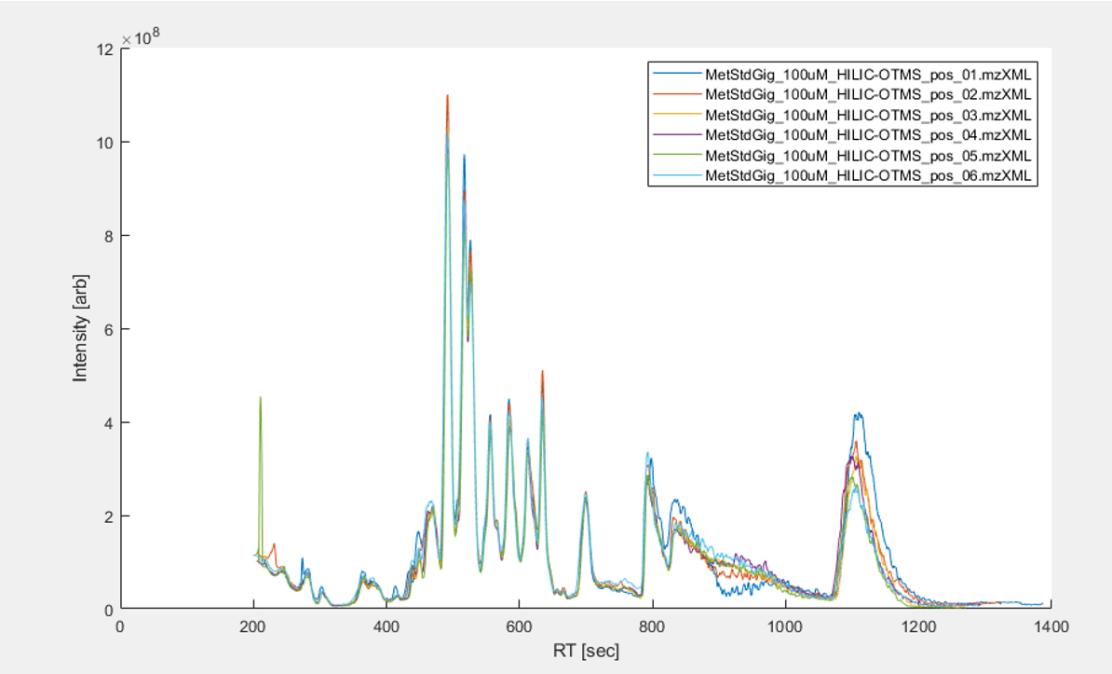


Figure 27: TIC for each sample in a group

Plot TICs plots all TICs and the blank or average blank, if selected, for each group.

5.7.2 Entropy and S/N distribution

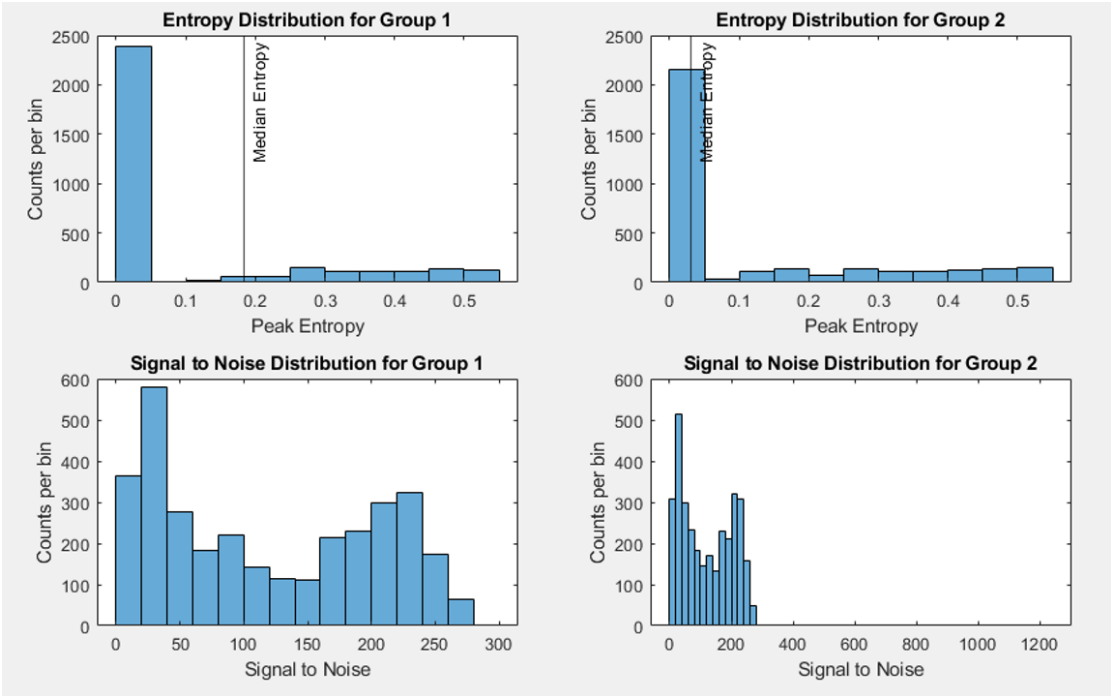


Figure 28: Entropy and Signal to Noise distribution

Plots a histogram with the distribution of the signal to noise ratios and the peak entropy for each group.

5.7.3 Rejected Features

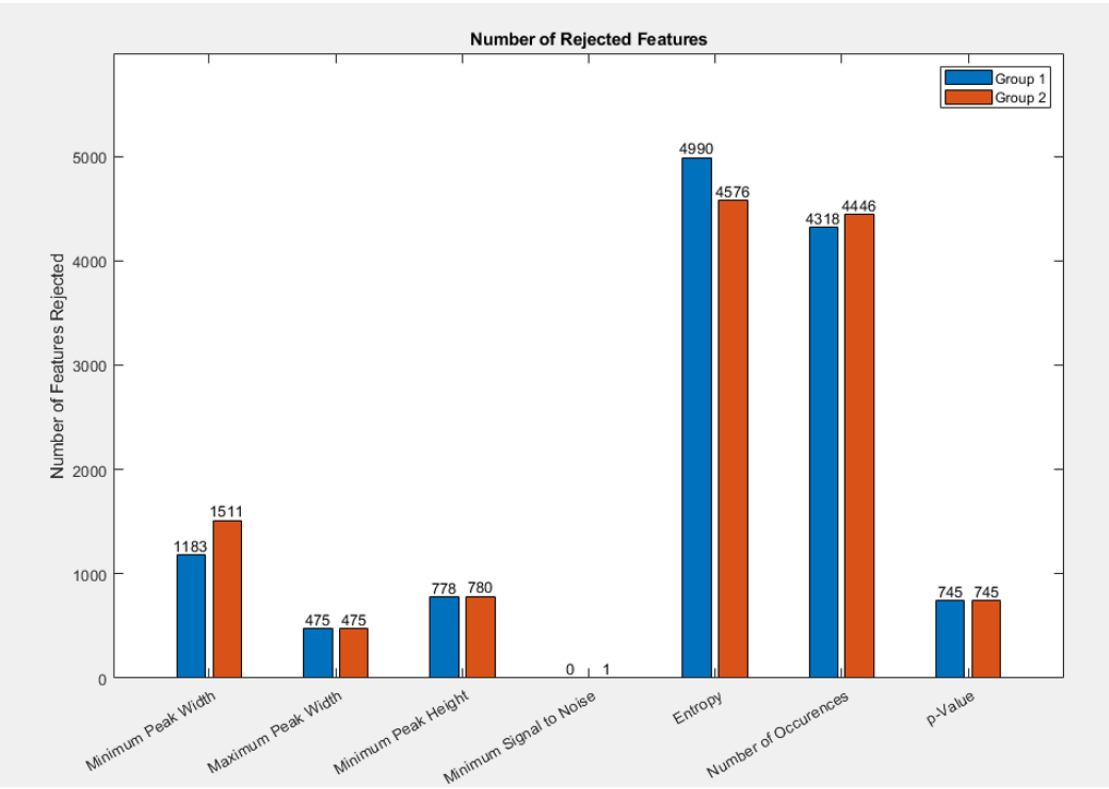


Figure 29: Number of rejected features sorted by filter option and group

Plots the number of features for each group removed from the total list by each filter.

5.7.4 Scatter Plot

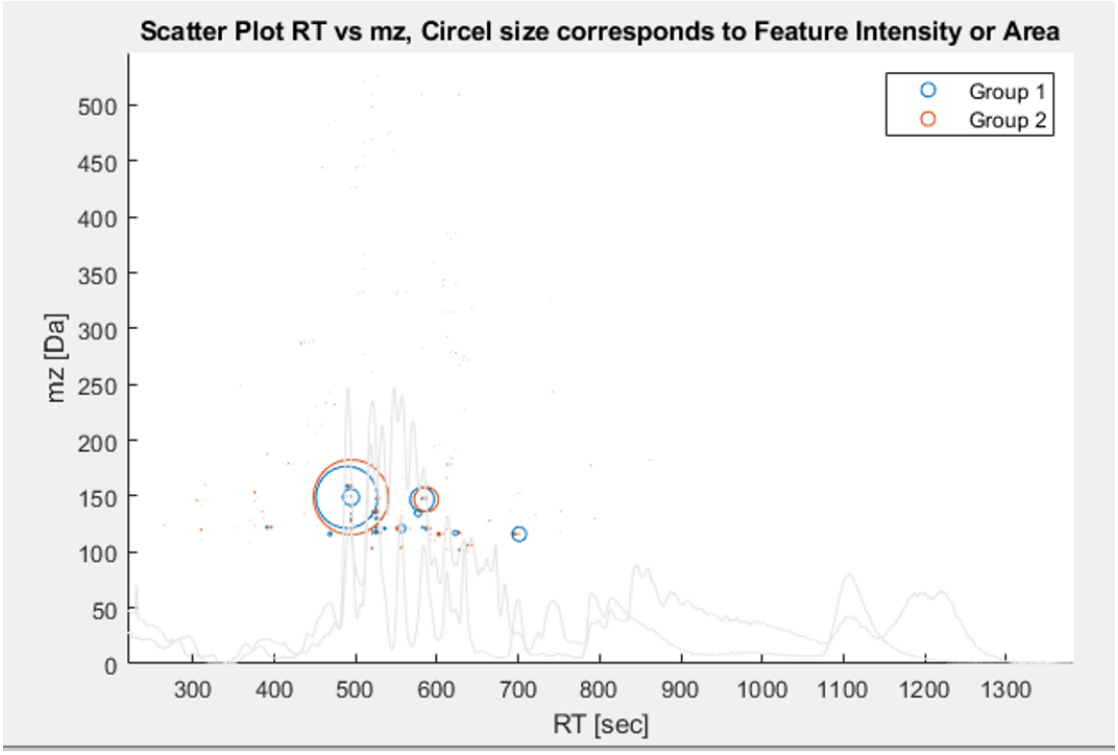


Figure 30: Scatter plot of found features

Scatter plot of the retention time against the m/z of the found features of all groups. The size of the circles corresponds to the peak height or area. For better orientation, an average TIC is shown in gray for each group.

5.7.5 Vulcano Plot

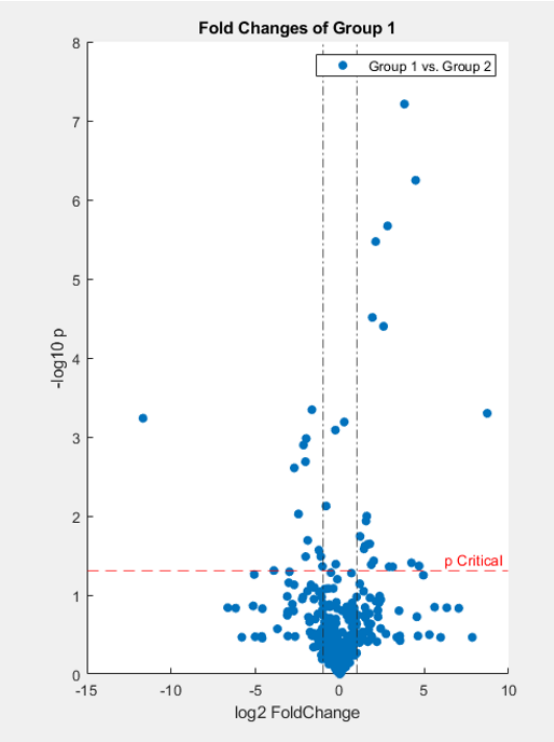


Figure 31: Volcano plots

Plots the $\log_2$ of the fold change against the -$\log_{10}$ of the p value of all features. Each group gets its own plot containing all possible group combinations.

5.7.6 Cloud Plot

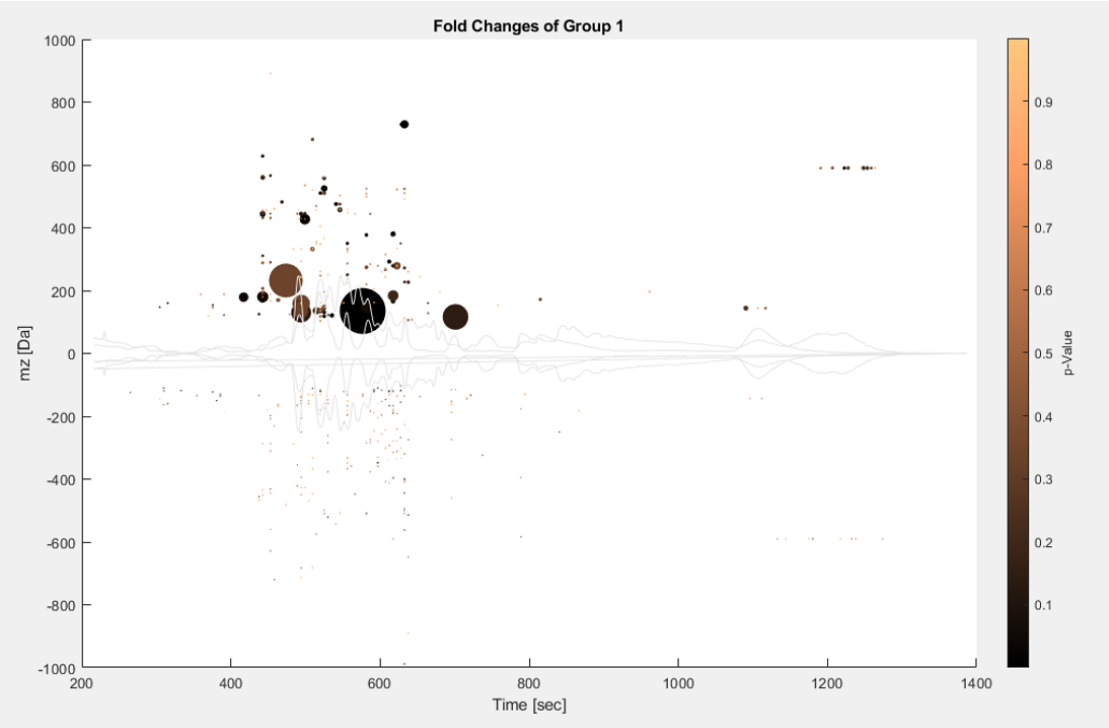


Figure 32: Cloud plot color corresponds to p value, size to fold change. Points above the x axis are up regulated, below down regulated

The cloud plot is an extension of the scatter plot. The position of the points is determined by the retention time and m/z of the feature. The size of the points correlates with the fold change of the feature. The points that lie below the x-axis (negative m/z) are regulated down. The color of the points indicates the p-value. Hovering the mouse pointer over a point displays the exact values. One plot for each group.

5.8 Multivariate Data Analysis

5.8.1 Heatmap

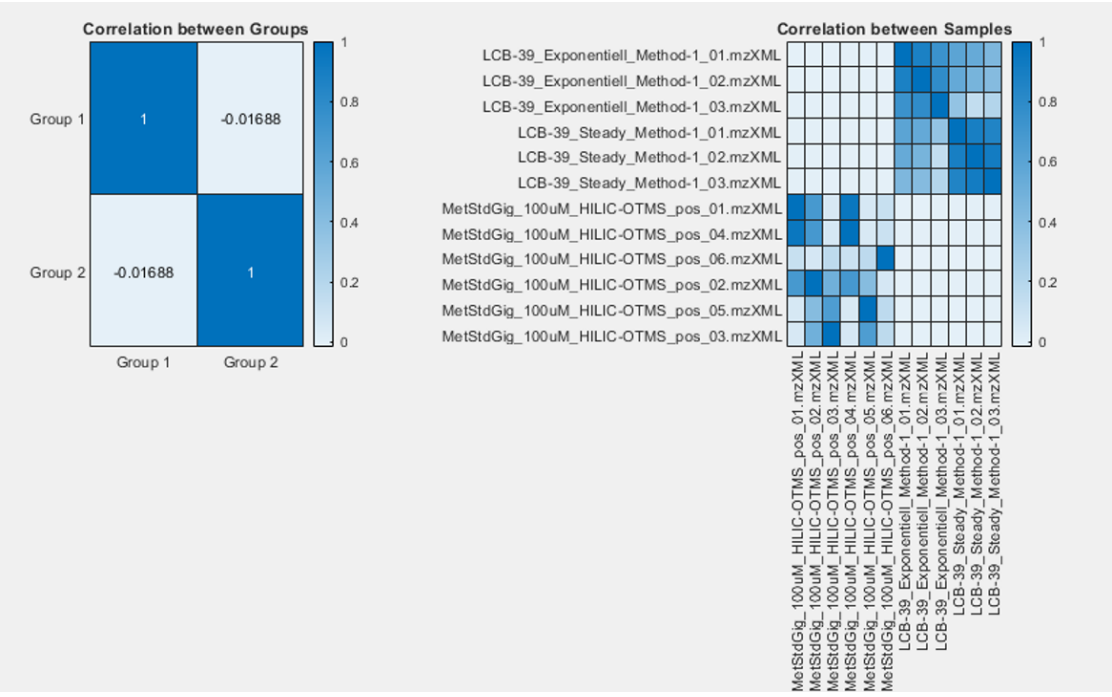


Figure 33: Correlation heatmaps between groups and between samples.

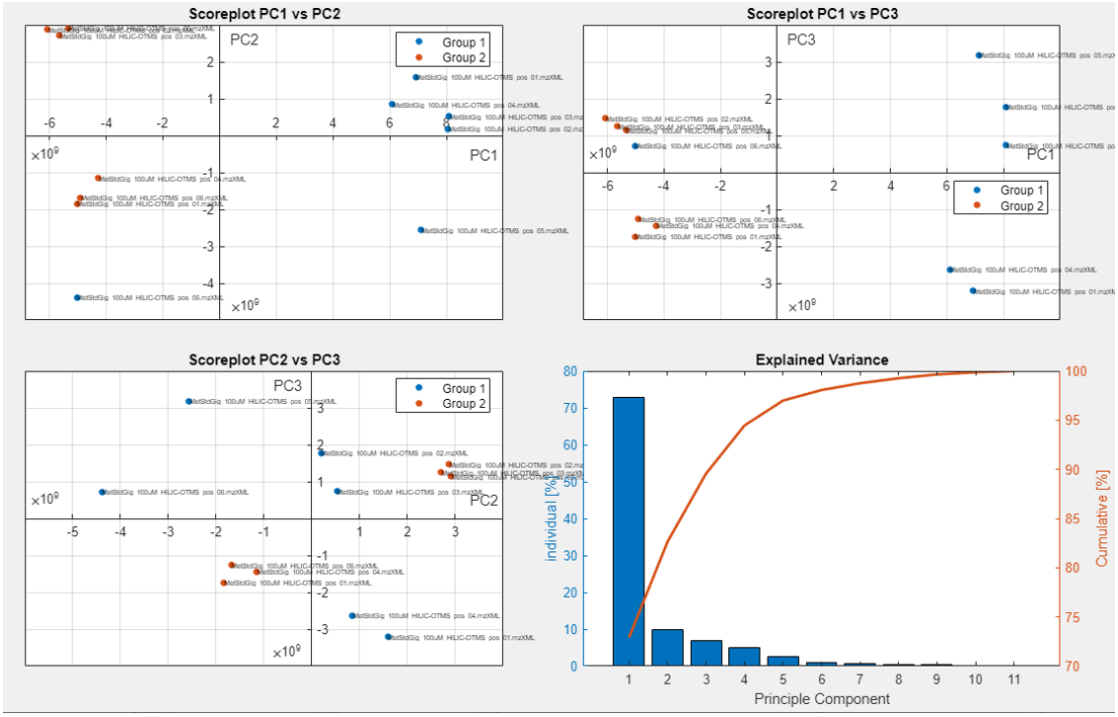
The heatmap plot shows the correlation between the groups and between the samples.

5.8.2 PCA

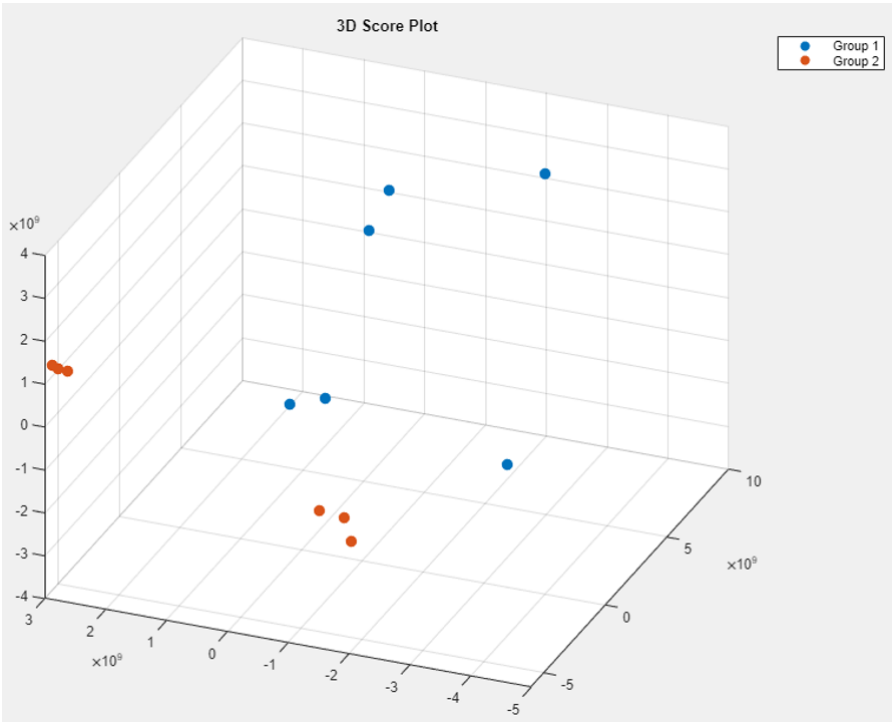
Principle component analysis (PCA) transforms high-dimensional data into a few Principle component (PC) by linear combination so that the new dimensions represent the greatest variance of the data in descending order. Multiple plots are generated.
In the first window, three 2D score plots are displayed. PC1 vs PC2, PC1 vs PC3 and PC2 vs PC3 as well as the variance explained individually for each principle component and cumulative. Points are labeled with the sample names and colored by group.

Plot 2 is a 3D plot of PC1 vs PC2 vs PC3.

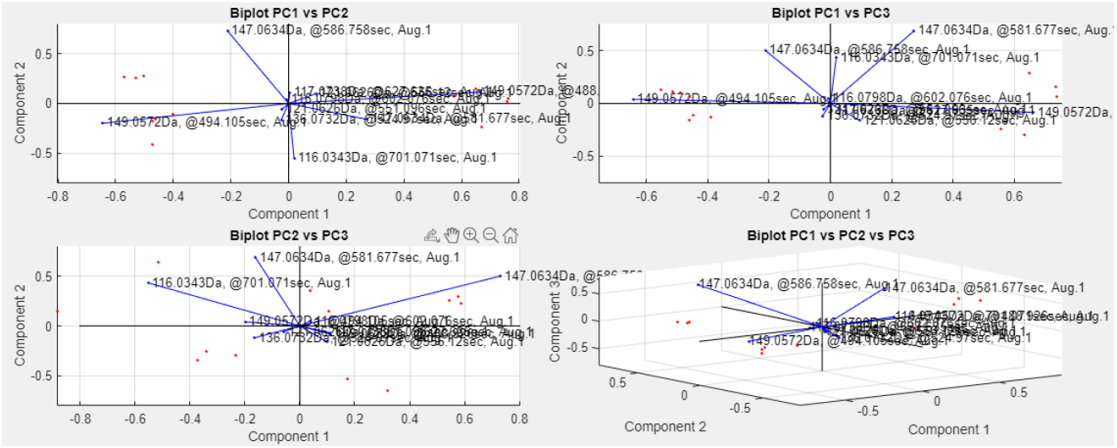
Window 3 contains the 2D biplots with the 20 largest loadings of individual features. The loadings are labeled with the corresponding feature. Additionally, a 3D biplot between PC1, PC2 and PC3 is displayed.



(a) Window 1 Score plots and explained variance



(b) Window 2 3D Score plot



(c) Window 3 Biplots

Figure 33: PCA result plots

5.9 Export Results

The ***Export Results button*** exports a table for each group to the MATLAB workspace. The naming scheme is: "FeatureTable_Groupname",
if groups were augmented: "FeatureTable_Groupname1_Groupname2". These can then be saved as a .mat file.
Caution:: If the workspace already contains a variable with the same name, it will be overwritten without warning.

The tables contain the following columns:

1. Name of the feature in the form m/z, @retention time, group.
2. Peak areas or heights for each measurement of the group.
3. Average peak area or height.
4. Standard deviation of the peak area or height.
5. Fold change against each other group.
6. p-value against each other group.
7. Whether the feature is significant against each other group (If significant 1, if not 0).
8. Whether the feature is unique to this group (If unique to this group 1, if not 0).
9. Whether the feature was also found in the reference group.
10. How often the feature was found in this group.
11. How often the feature was found in all groups.
12. The m/z of the feature.
13. the retention time of the feature.

If only one group was processed, all comparison columns are omitted.

List of Figures

1 Initial User Interface . . . 3
2 Main Options Panel . . . 6
3 Data selection window . . . 7
4 File discrepancy warning . . . 8
5 File preview window . . . 8
6 ROI Parameters panel . . . 9
7 Alignment Parameters panel . . . 11
8 Comparison of m/z traces before and after spectral alignment . . . 12
9 ROIs before and after Peak Alignment . . . 14
10 Baseline Correction Parameters panel . . . 16
11 ROIs before and after baseline correction . . . 17
12 Golay Smoothing Parameters panel . . . 19
13 ROIs before and after Savitzky-Golay smoothing . . . 20
14 ISTD Parameters panel . . . 21
15 Define Internal Standards window . . . 22
16 Adduct & Neural Loss windows . . . 23
17 Additional Options panel . . . 25
18 Parameter preview window . . . 27
19 AriumMSEval initial window . . . 28
20 Group Assignment Table . . . 29
21 ISTD Processing options . . . 30
22 Data Pre-Treatment panel . . . 31
23 Feature sorting parameters . . . 32
24 Sample specific scaling factor window . . . 33
25 Feature extraction . . . 34
26 Data Visualization panel . . . 36
27 Plotted TIC . . . 37
28 Entropy and Signal to Noise distribution . . . 38
29 Number of rejected features . . . 39
30 Scatter plot of found features . . . 40
31 Volcano plots . . . 41
32 Cloud plot . . . 42
33 Correlation Heatmap . . . 43

33 PCA result plots 46