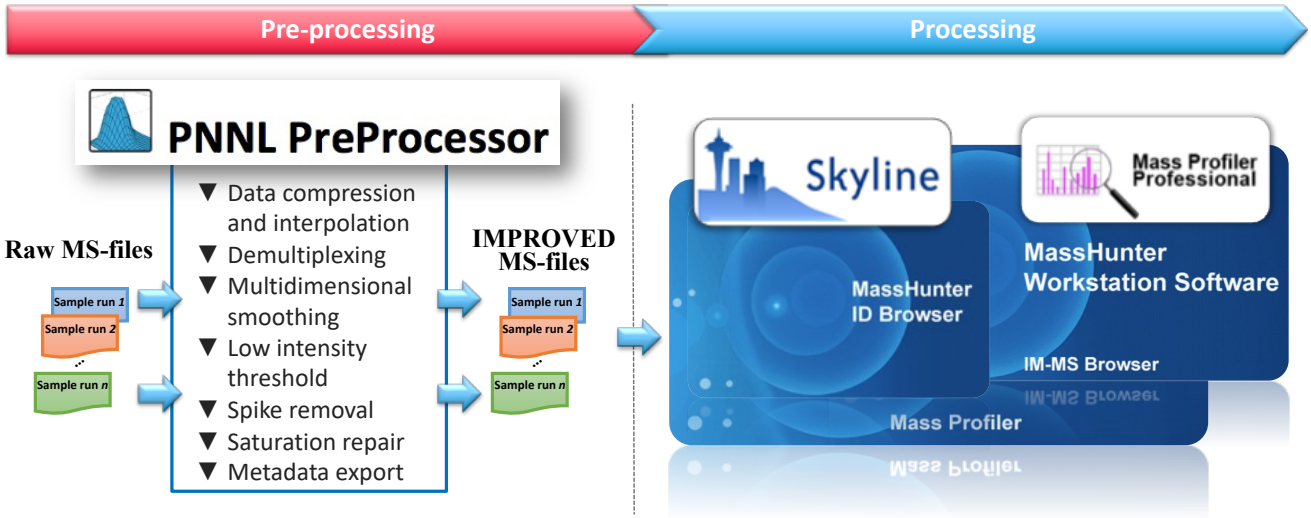
PNNL PreProcessor | Ion mobility MS

USER GUIDE – version 4.0 (2021.10.27)

In close collaboration between Pacific Northwest National Laboratory and Agilent Technologies, we have developed this user-friendly tool for Agilent MassHunter (.d) and UIMF mass spectrometry data files (MS-files) and generates new MS-files in the same instrument data format with enhanced signal quality.



Available algorithms and utilities in the PreProcessor include: data compression and interpolation, ion mobility demultiplexing, multidimensional smoothing, noise filtering by low intensity threshold and spike removal, saturation repair and metadata export.

If you use the PreProcessor, please cite: Bilbao et al. *A Preprocessing Tool for Enhanced Ion Mobility-Mass Spectrometry-Based Omics Workflows*. Journal of Proteome Research 2021 https://doi.org/10.1021/acs.jproteome.1c00425.

Contacts:

- aivett.bilbao@pnnl.gov
- john_fjeldsted@agilent.com

Do you have a problem to report? Please let us know any feedback!

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Agilent Technologies

1. What are the benefits?

- Remove artifacts in jagged peaks (i.e. low ion statistics) and enhance low-abundance real signals.
- Improve signal-to-noise and reproducibility with more features consistently detected across your multiple runs and reduce variations in abundance and collision-cross section (CCS).
- Reduce the file size and time for downstream processing.
- Recover the intensity, isotopic ratios and mass accuracy from saturated high-abundance analytes of different molecular ions (i.e., peptides, metabolites, lipids, glycans, and inorganic polymers).
- Computationally extend the dynamic range of your measurements and enable faster and memory-efficient downstream processing!

2. Requirements

- A computer running Windows 7 (64-bit) or later with at least 8GB of RAM (32GB recommended).
- The .NET Framework 4.7.2 (or later). Download available at https://dotnet.microsoft.com/download/dotnet-framework/thank-you/net472-web-installer.
- MS-files in UIMF or Agilent MassHunter (.d) format.

3. Software features

- Command-line and graphical (workflow style) user interfaces.
- Single-click batch processing of multiple raw MS-files.
- Data compression (by frame and mobility) and filtering by retention time range.
- Data interpolation of the ion mobility dimension to improve the results of the HRdm demultiplexing and peak deconvolution strategy.
- Multidimensional smoothing of data and repair of saturated peaks.
- PNNL demultiplexing and artifact removal algorithm integrated. A new selectable pulse coverage percentage to increase sensitivity for low level signals.
- An algorithm to remove noise in form of 'spikes'.
- Ion mobility MS with/without any separation: LC-IM-MS, solid phase extraction (e.g., RapidFire) IM-MS and direct infusion IM-MS.
- All Ions MS-files (data-independent acquisition) with alternating high/low collision energy fragmentation.
- Exporting metadata information of ion mobility frames (e.g., field, pressure, temperature) and MS actuals to text files.
- Multidimensional smoothing for non-ion mobility TOF-MS MS-files (e.g., produced by 6530, 6540, 6545 and 6550 Agilent instruments).

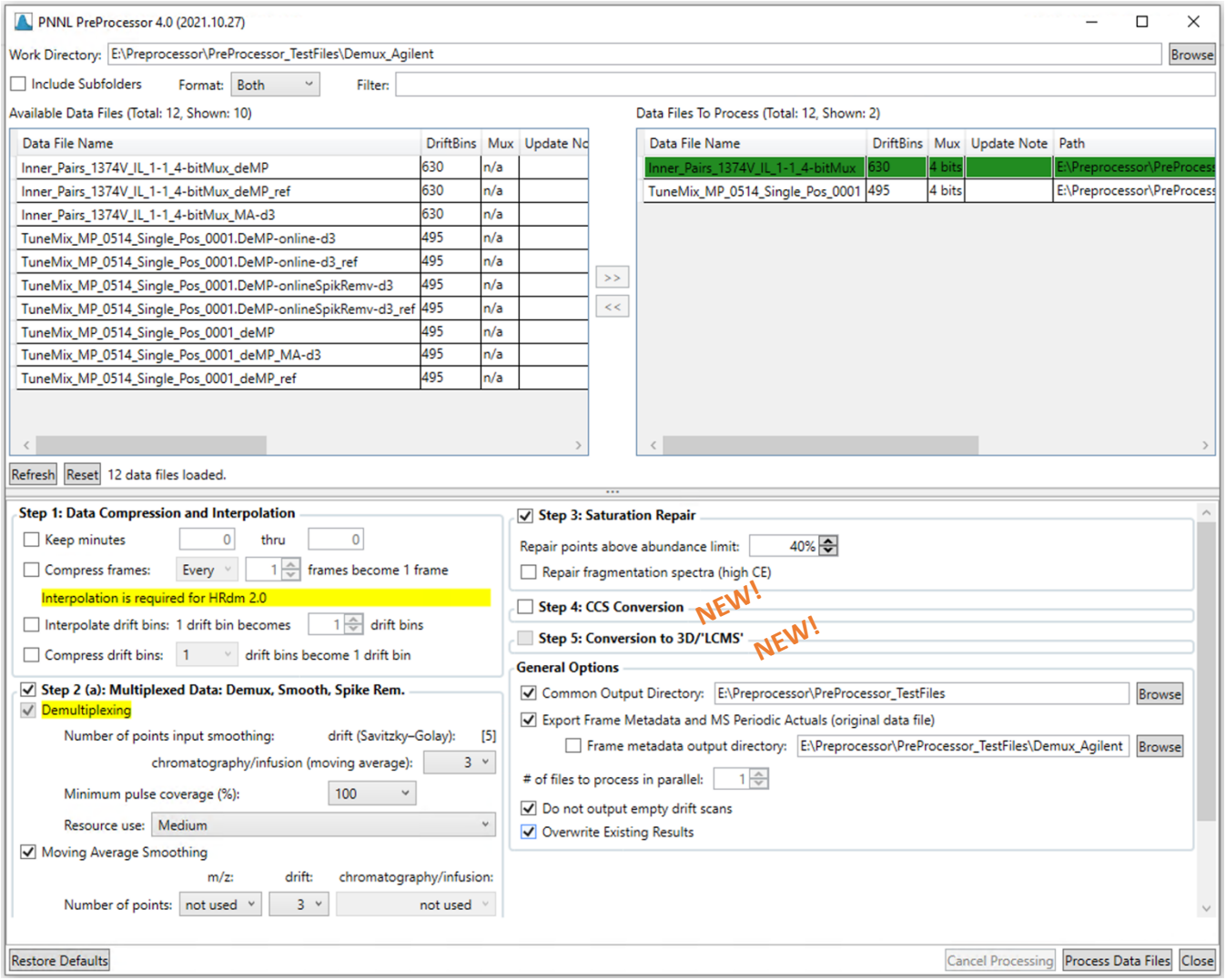
4. New features

- Conversion of arrival time to CCS in the raw data for SLIM (Step 4).
- Conversion of arrival time to retention time: single frame "IMMS" data to "LCMS" format (Step 5).

5. Usage

Installation is not required (users having issues with anti-virus software can alternatively use the installer). Download and unzip the file in your computer.

- Double click the 'PNNL-PreProcessor.exe' file to start it.
- Click the text box to search and select a single MS-file (.d folder) or a directory which contains several MS-files (.d folders).
- Select files and click '>>' to move to the right the files to be processed.
- Adjust the parameters and click the 'Process Data Files' button.



For each MS-file, the selected algorithms will be sequentially applied in the same order as they appear in the GUI from top to bottom. A progress bar will indicate the processing status and a new MS-file will be generated within the same folder (or selected Output directory if checked).

Suffix in the MS-file names indicates the preprocessing algorithms applied, for example:

- **_1.00-3.00.d** time range (minutes) kept.
- **_FC3.d** frame compression with every 3 frames summed into 1.
- **_DC3.d** drift bin (TOF transient or MS scan) compression with every 3 drift bins summed into 1.
- **_DI3.d** drift bin interpolation with addition of 3 interpolated drift bins per every original drift bin.
- **_MA-d3-c3.d** moving average smoothing with 3 points in drift and 3 points in chromatography.
- **_Min20.d** signal intensity lower threshold of 20.
- **_Spk.d** spike removal.
- **_SR.d** saturation repair.
- **_MA-d3-c3-Min20-Spk_SR.d** moving average smoothing with 3 points in drift and 3 points in chromatography, signal intensity lower threshold of 20, spike removal and saturation repair.
- **.DeMP.d** ion mobility demultiplexed combined or not with other algorithms. Only the final result file will be preserved and intermediate result files will be automatically deleted. To avoid file overwriting or loosing results, manually rename result files (after preprocessing is complete) to compare multiple combinations of preprocessing parameters with demultiplexing. Check the log file inside each result to verify which preprocessing algorithms were applied.

PreProcessor workflows with specific algorithms and parameters for different experimental approaches. Empty cells indicate that the module is not used. The output MS-files generated are named as the corresponding input MS-file plus a suffix that indicates the modules applied (due to compatibility with the HRdm software, the suffix "DeMP" is used for the demultiplexed MS-files regardless of additional post-demultiplexing modules applied).

			EXAMPLES OF EXPERIMENTS AND PARAMETERS				
ALGORITHM	**SLIM IM-MS**	**SLIM IM-MS**	**LC-DT IM-MS multiplexed**	**LC-DT IM-MS multiplexed + HRdm**	**LC-DT IM-MS default**	**LC-DT IM-MS, saturated ions**	**DT IM-MS (direct infusion)**
Data compression	15 drift bins become 1	All frames become 1					All frames become 1
Data interpolation				1 drift bin becomes 5			
Ion mobility demultiplexing			5 points LC	5 points LC			
Multidimensional smoothing			3 points IM	3 points IM	3 points LC, 3 points IM	3 points LC, 3 points IM	3 points IM
Low intensity threshold	1 count		20 counts	20 counts	20 counts	20 counts	50 counts
Spike removal					1 adjacent point per dimension		
Saturation repair						40% (MS1 only)	
CCS conversion		Selected					
Conversion to 'LCMS'		Selected					
Suffix in output MS-file name	_DC15-Min1	FCsum_ CCSz_3D	_DeMP	_DI5.d.DeMP	_MA-c3-d3-Min20-Spk	_MA-c3-d3-min20-Spk-SR	_FCsum-d3-min50

CAUTION with temporary files: temporary files with suffix such as '_temp1.d' will be generated with intermediate results. Do not delete the temp files while any preprocessing is in progress and verify that all temp files were deleted after all preprocessing is finished. For example, an error has occurred if saturation repair was performed and only the file 'SR_temp1.d' was generated, please report it.

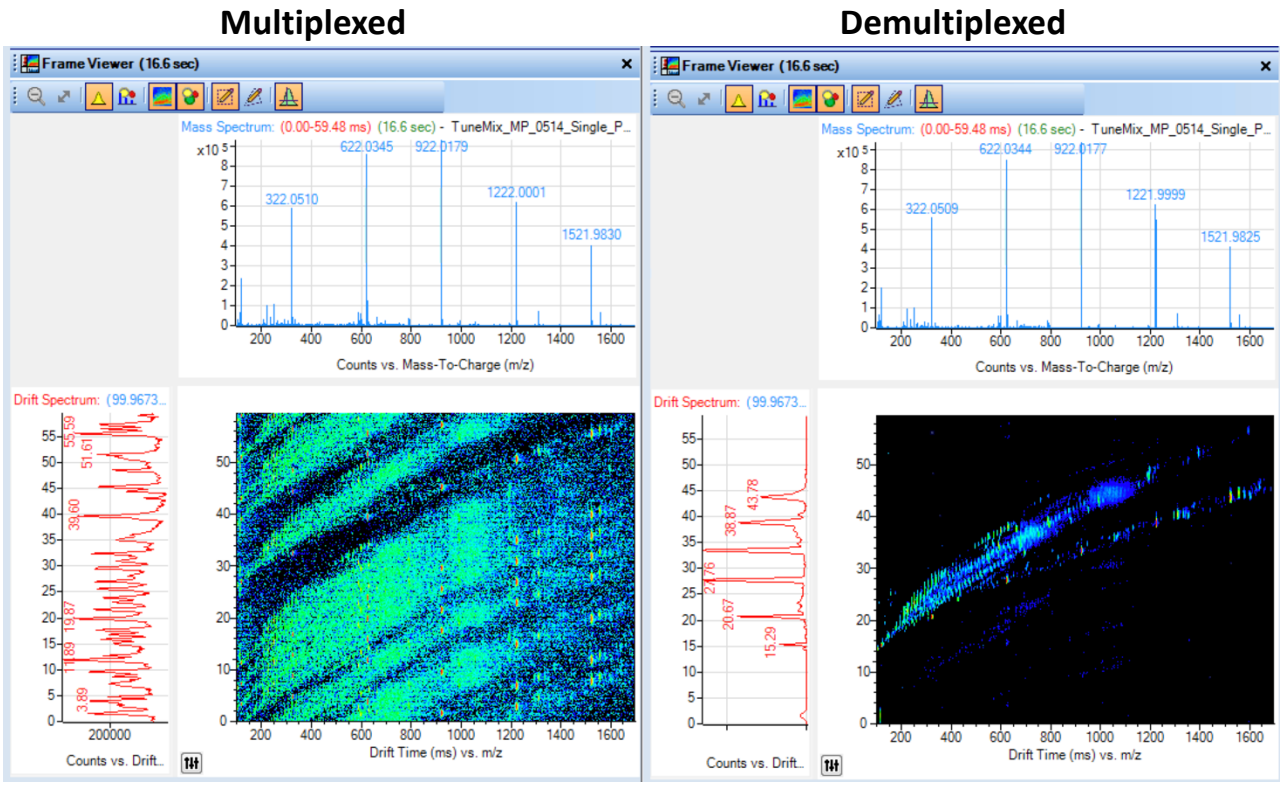
6. Data Compression and Interpolation

New to version 3 are options to select and alter the sampling rate of the data that will be treated.

- It is now possible to select a time range from the original input MS-file for treatment. The final output MS-file will only contain data from the frames within the time range selected.
- When configuring the acquisition of IM data, the user selects either the time interval or the specific number of frames acquired and summed before being written to disk. For long separations which may have wide chromatographic peak widths this number could be low resulting in many points acquired per LC peak. By using the data compression post-acquisition, additional in-frame sensitivity can be realized by further combining (i.e., summing or "compressing") adjacent frames. For data that has been acquired as part of an infusion experiment, options are also available to sum all frames into one.
- To increase the effective sampling rate of the ion mobility separation the user can now add additional TOF Transients by means of drift bin interpolation. Specifying 1 bin to become 3 (a good default) results in an effective ion mobility sampling period of 40 μsec as opposed to 120 μsec without interpolation when operating in the 1700 *m/z* range mode. This operation should always be used for multiplexed data that will be subsequently processed with HRdm 2.0. It can also be used for single pulse data.
- To better support SLIM data it is now possible to effectively decrease the TOF sampling rate by combining (i.e., summing or "compressing") the data in adjacent TOF transients. This is very helpful because SLIM separations occur over a much longer time scale (approximately 1 second) resulting in TOF transient oversampling. Typical drift bin compression values range from 3 to 10.

7. Demultiplexing

The previously developed [PNNL demultiplexing and artifact removal algorithm] was integrated into the PNNL PreProcessor. Multiplexing is a data acquisition technique applied during ion mobility separation which can significantly improve the duty cycle and signal-to-noise as well as extend the dynamic range.



For ion mobility multiplexing, the drift time dimension is encoded with a pseudo random sequence by allowing multiple ion packets to be concurrently resident in the drift tube. As the multiple packets traverse the drift cell, fast moving ions from one packet overlap with slow moving ions from a previous packet and the information for all packets is recorded by the detector in one MS-file. The multiplexed raw data is demultiplexed post-acquisition using a Hadamard transform-based decoding algorithm which is based on the gating of the ions and the pseudo random sequence used for multiplexing during acquisition and a second algorithm developed at PNNL to detect and remove artifacts.

Version 3.0 provides the following new capabilities:

- The user can now select the Minimum pulse coverage (%). Following demultiplexing the software performs a noise filtering or artifact removal by applying a reverse confirmation for all output results to check that corresponding signal is present. This is one important contributor to why the PNNL PreProcessor produces results with a very low level of artifacts, compared to other demultiplexing methods. Reducing this value from 100% allows preserving signals with relatively low ion statistics while still providing very low artifact presence.

- Three options for resource use are available for demultiplexing workflows. As demultiplexing is a computationally intensive operation, the user can now select either low, medium or high. The selection is not absolute, but rather relative to the computer's number of cores.

CAUTION: The demultiplexing algorithm has been optimized to use multiple threads and supports all post-demultiplexing operations selected, except chromatographic smoothing which occurs as a single-thread operation in a second step. To avoid very slow processing *do not use the parameter 'Moving algorithm smoothing: chromatography/infusion'*.

8. Multidimensional smoothing

Raw intensity values in each frame (an IM separation) are smoothed first in the drift dimension followed by smoothing of the chromatographic dimension considering neighboring frames. The parameters are described in the following subsection.

8.1. Number of points

The number of points used to compute the average or new smoothed value should be specified for each dimension. For instance, 5 points for drift mean that each new point is computed as the average considering also the 2 drift spectra before and the 2 drift spectra after (same *m/z* and within the same frame).

- As a general rule, the number of points should be less than the number of sampling points covering the peak width at half height. Smoothing with a large number of points cause a detrimental lost in resolution (see section below on Over Smoothing).
- Smoothing can be applied to all *m/z*, drift and chromatographic dimensions or any subset combination.
- Smoothing across the *m/z* dimension is typically not necessary for Agilent IM-MS systems.
- Larger chromatographic number of points requires more memory and processing time.

8.2. Over Smoothing

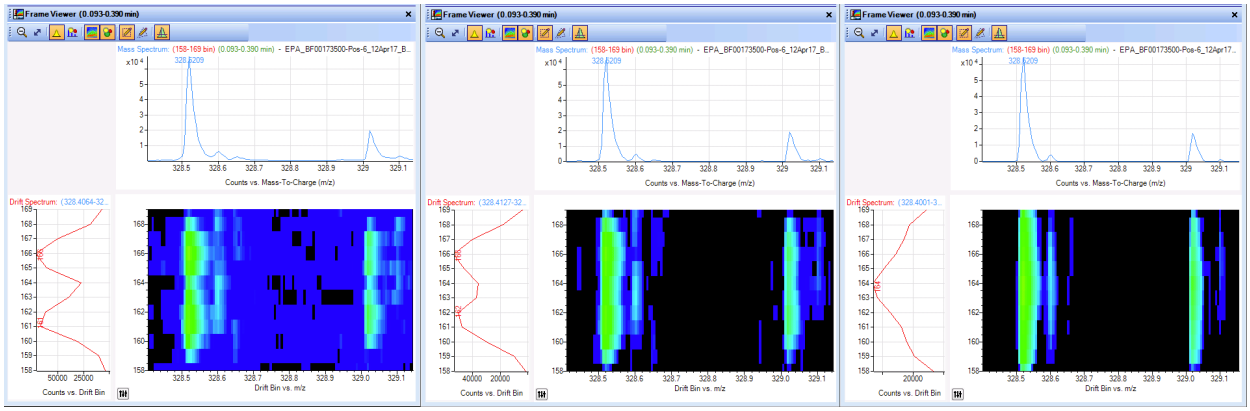
When first using the software or processing data from a new sample type, you should check your raw data to inspect the results. One of the possible problems that you may run into is over smoothing or blurring two adjacent peaks into one. This happens when smoothing is applied with a number of points that is too large (i.e. smooth too many points, relative to the width of a peak).

To avoid over smoothing, please do the following: check one of your files and see how many points you have across the early drift peaks that you are interested in, the low *m/z* range has usually narrower drift peaks and earlier drift times.

In the IM-MS Browser, select: Configuration -> Set Drift Spectrum units to Drift Bins.

In the following example, the file in the left is the original or unsmoothed one, those drift peaks have less than ~6 points per peak (the first one from 158-164 Drift Bin), those are likely isomers. To preserve them, use only 3 points for smoothing (as in the middle file), since using 5 or 7 points will merge those drift peaks (as in the right file).

Unsmoothed **3 points drift** **5 points drift**

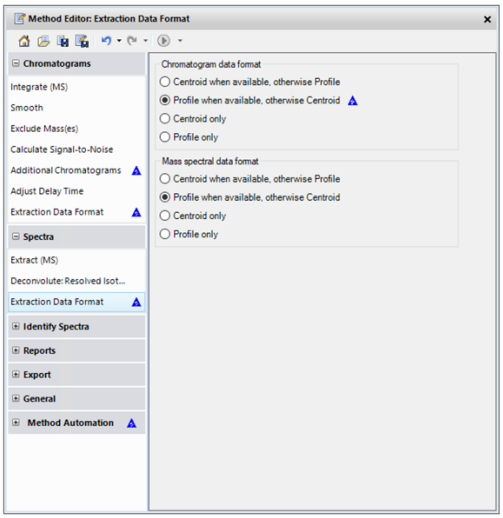


8.3. Smoothing for non-ion mobility TOF-MS MS-files and using them in QUAL software

Smoothing is also available for files produced by 6530, 6540, 6545 and 6550 Agilent TOF instruments. However, the PNNL PreProcessor does not generate centroid peak data. *Use the Agilent 'IM-MS Reprocessor' to recentroid the result file prior to opening it in Agilent QUAL software*.

Alternatively, change the settings in QUAL to show the raw data instead of the centroid peak data:

- Go to 'Method -> Method Editor'…
- Select 'Spectra -> Extraction Data Format'
- Select 'Chromatogram data format -> Profile when available, otherwise Centroid'
- Select 'Mass spectral data format -> Profile when available, otherwise Centroid'



9. Signal intensity lower threshold

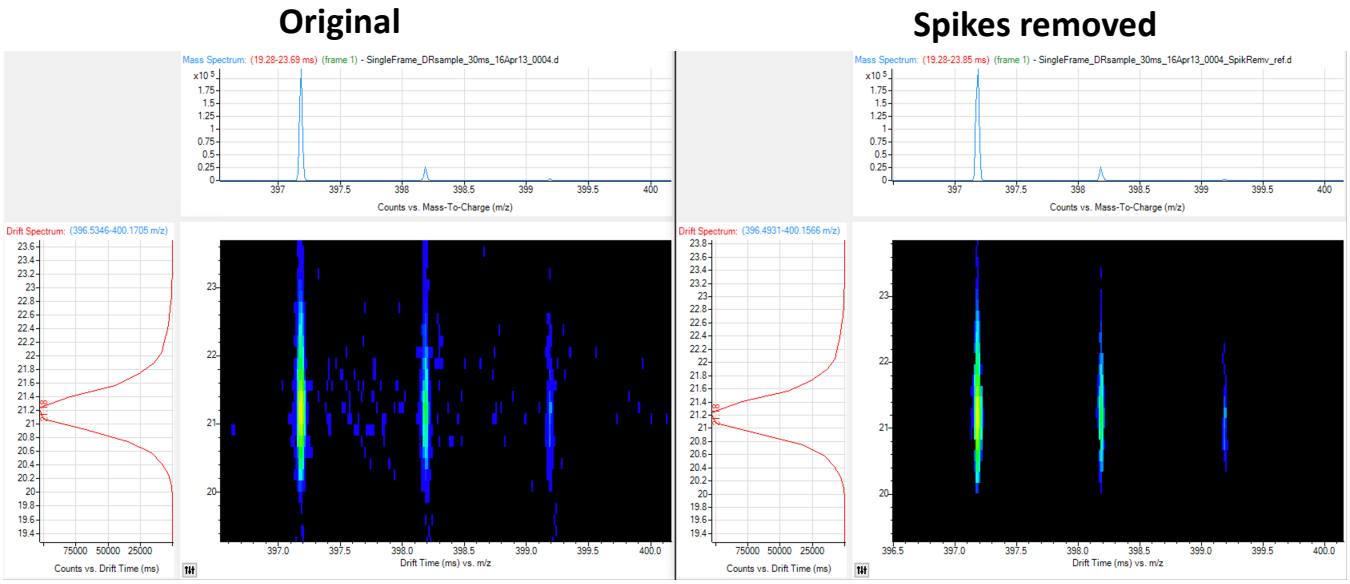
After applying the smoothing, intensity values that do not exceed the threshold will be removed.

CAUTION: a threshold higher than the recommended default value (20) in the default settings may cause non-detection or decreased quality of features. However, a higher value could be useful to reduce required memory and increase processing speed for very complex data files.

10. Spike removal

We refer to as spikes to the noisy peaks that have inconsistent and very few raw data points. Since the spikes may have intensity values above the lower intensity threshold, we detect them by comparing each point against the neighboring or adjacent points in drift and *m/z*. Each point must have either the 1 or 2 user-specified neighbors in each dimension in order to pass the filter. Points that do not pass the filter will be removed.

Note: the spike removal algorithm is not recommended for All Ions MS-files because some relevant low-mass and low-abundance fragment ions could be removed from the fragmentation spectra.



11. Saturation repair

Complete raw MS-files from untargeted LC-IM-MS analyses are automatically preprocessed to find ions exceeding a defined intensity level (a percentage of the abundance limit given the MS instrumentation).

The undistorted isotopic distribution is recovered from spectra at unsaturated time from the peak tail (intensity below the selected saturation level). In contrast to our previous algorithm, this new method does not require theoretical isotopic distributions, works at the raw or profile level (i.e., non-centroid spectra) and produces MS-files in instrument format (instead of text tables) as output.

The most intense unsaturated isotopic peak is utilized as reference to successively reconstruct the time profile of the saturated ones.

11.1. IMPORTANT NOTES FOR CURRENT VERSION

- Chromatographic separation required: saturated analytes must have a well-defined chromatographic peak (even if several isotopic peaks have a 'flat top', peaks should have a clear start and end time points and at least one isotopic peak below the saturated threshold defined).
- Interferences: the software ignores saturated ions if interferences are detected.
- Ion charge states from 1-5 supported.
- All Ions MS-files supported (high fragmentation energy frames are ignored by default).

11.2. Output MS-files

The tool works in a serial way, two files are generated if both smoothing and saturation repair options are selected: the first file is only smoothed, and the second file is both smoothed and saturation repaired. For example, with default settings, the suffix in the MS-file names indicates the preprocessing applied:

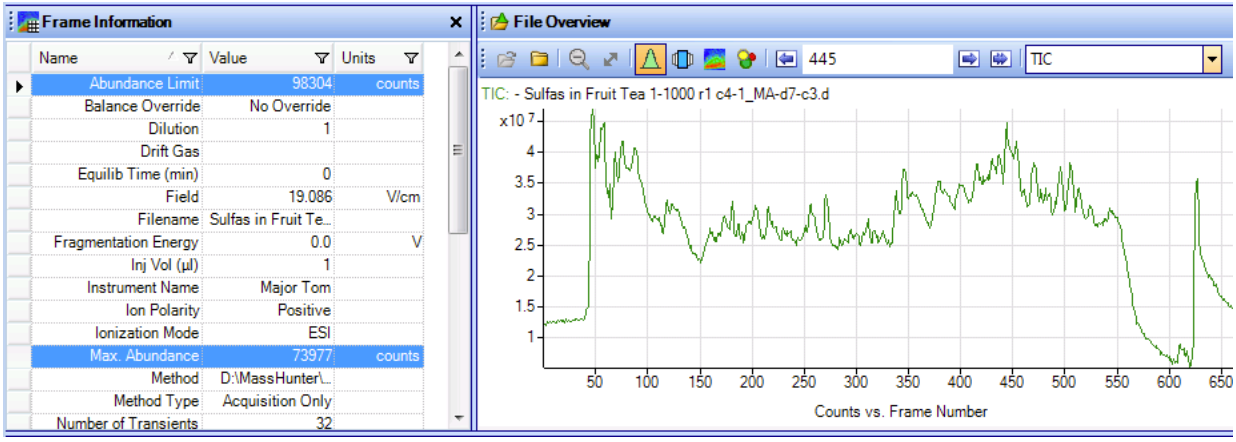
- _MA-d3-c3.d: smoothed only
- _MA-d3-c3_SR.d: both smoothed and saturation repaired

Why both '_MA-d3-c3.d' and '_MA-d3-c3_SR.d' files are kept? Because the saturation repair results may not be comparable across multiple runs, it is recommended to verify the results and delete the unused files after. For example, when the sample is very complex and there are interferences for a saturated peak in some runs but not in the others, the peak will not be repaired across all runs and therefore quantitation results will not be comparable across runs. Unused files can be manually deleted by filtering them either typing 'SR.d' or similar in the Explorer windows.

11.3. Abundance limit

The percent saturation limit default is 40% and is the approximate onset of non-linearity, peak broadening and an increasing mass error due to saturation of the detector. The following is an example of a threshold at 70% and shows where the associated Abundance Limit can be found. It should be noted that saturation effects worsen as a peak's percentage abundance increases and that 100% is a theoretical and not realized maximum.

You can see the Abundance values in your file using the IM-MS Browser by selecting 'View -> Frame Information':

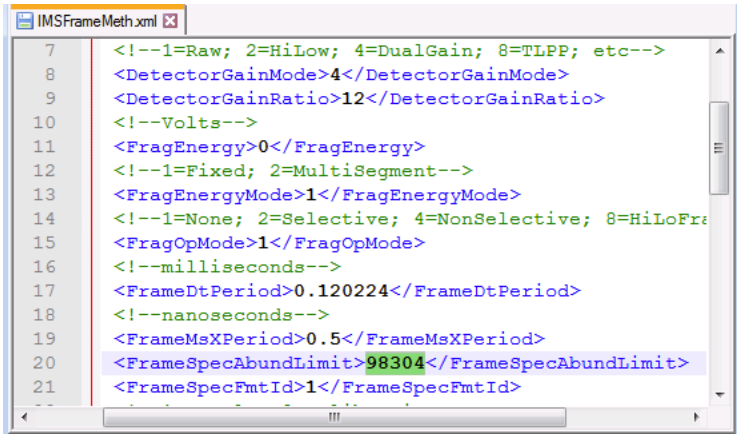


The Abundance Limit is 98304 (given by the detection system and acquisition settings) and

the Max. Abundance of the detected ions (for the actual signals recorded in that frame) is 73977.

Therefore, the software will attempt to repair all ions in that frame that have raw intensity between 68812 (70% of the Abundance Limit 98304) and 73977.

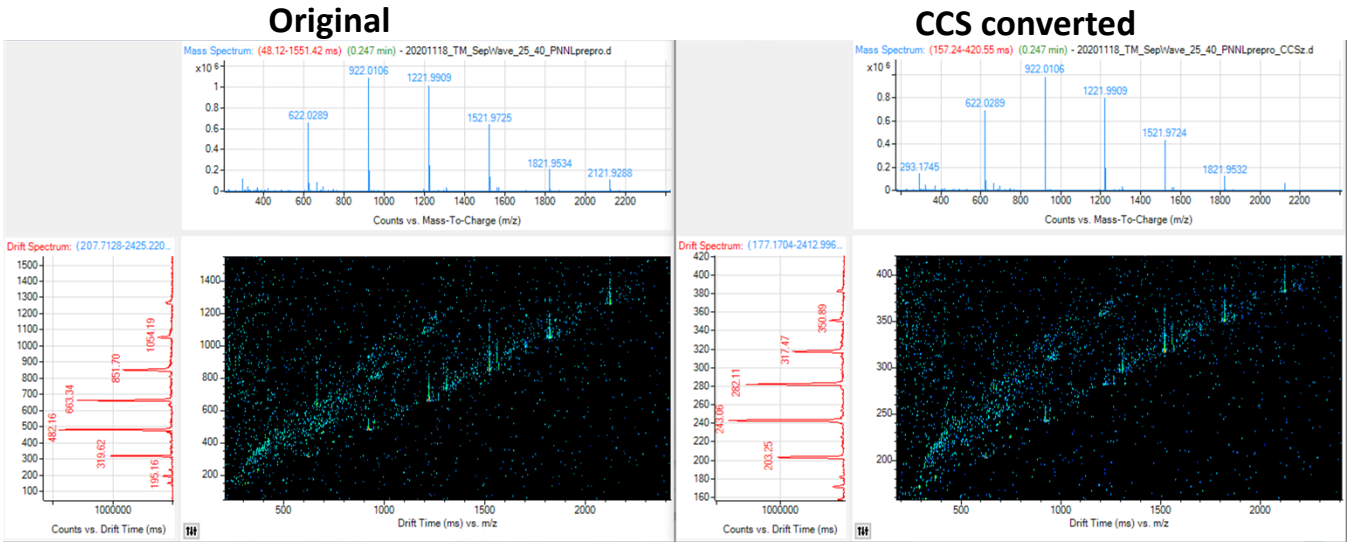
The software reads the actual value FrameSpecAbundLimit from the IMSFrameMeth.xml file in the AcqData subfolder.



12. CCS Conversion

Conversion of arrival time to CCS in the raw data for SLIM. The time axis with the arrival time in the raw data file is parsed to CCS using the polynomial function and calibration coefficients (found in file SLIMImsCal.xml).

The figure below shows an example for the Agilent tune mix ions. In the CCS converted file on the right, the Drift Time axis (red trace) represents the ion's CCS directly in the raw data.

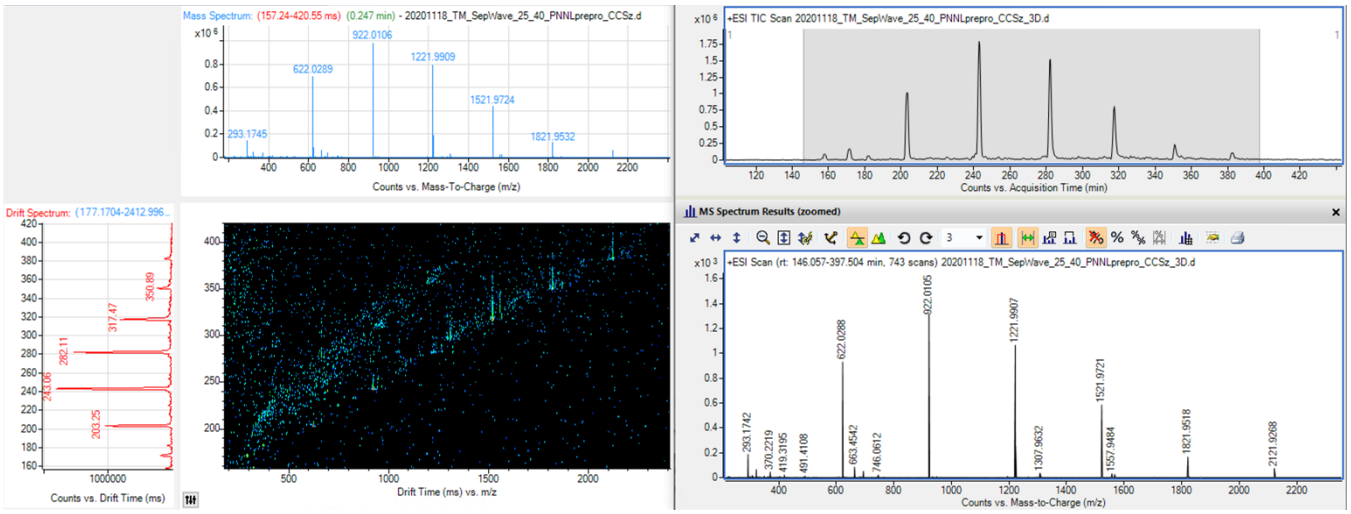


13. Conversion to 3D/'LCMS'

Conversion of arrival time to retention time: single frame "IMMS" data to "LCMS" format. The time axis with arrival times (or converted CCS) are parsed to retention time (RT) in a new file in the MassHunter QTOF/3D format which can be processed by existing LC-MS software tools (e.g., Agilent Qualitative Analysis or Skyline). In the case of CCS converted data files, 1 sq Angstrom is equivalent to 1 min RT in the LCMS data file. For unconverted CCS data files, several options are available to scale the units from milliseconds in arrival time to seconds or minutes in RT.

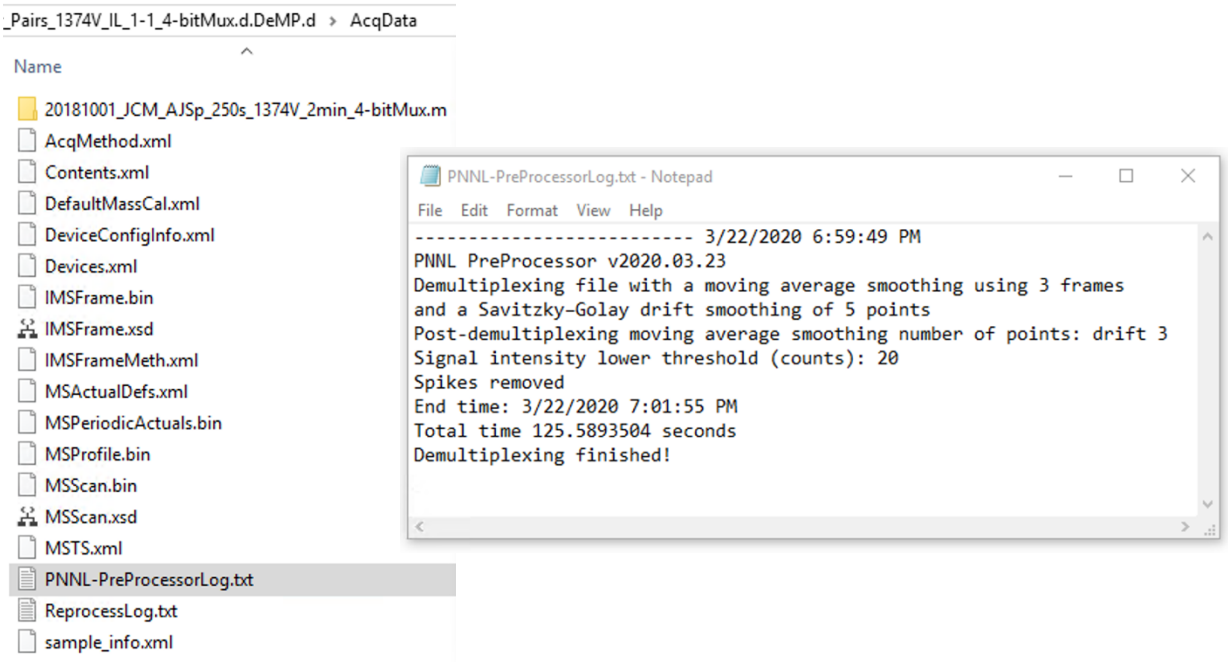
The figure below shows an example for the Agilent tune mix ions. After conversion to 'LCMS', the data file with suffix "_3D" can be visualized using the Agilent Qualitative Analysis software (IM-MS Browser will not work for this format). While the *m/z* dimension remains the same, the Drift Time axis (red trace on the left) becomes the Acquisition Time or RT axis on the right (top trace).

CCS converted **'LCMS' converted**



14. Log files

The PNNL-PreProcessorLog.txt file in the AcqData subfolder contains information about the software version, preprocessing algorithms applied, parameters used and execution time, e.g.,



Note: Information about the saturated ions found, repaired points and specific messages for ions that were not repaired are in the text file PNNL-PreProcessorSaturation.csv.