MetMatch

User Guide

About MetMatch

MetMatch is a software tool for the semi-automated comparison of different datasets of untargeted metabolomics experiments. It provides methods to correct LC-HRMS data exhibiting *m/z* and retention time shifts relative to a reference chromatogram.

License

Table of contents

1 Installation

1.1 System Requirements

OS: Windows 7/8/10 64-bit

Minimum: CPU with 2 cores, 2 GB RAM

Recommended: CPU with 4 cores, 8 GB RAM

1.2 Installation

Note: Although MetMatch uses R for some of its internal functions, R does not need to be installed separately. The necessary files are included in MetMatch, and MetMatch doesn't interfere with any previous or future installations of R.

1.2.1 Requirements

MetMatch requires Java (at least version 8, 64bit) to be installed on your system. Java can be downloaded from https://java.com/download.

1.2.2 Setup

To install MetMatch, extract the contents of "MetMatch-Setup.exe" to your desired installation path. To do so, simply double click the file.

1.2.3 Troubleshooting

In case MetMatch doesn't properly launch, please double click "Initialize-Test.bat". Check for the following two error messages:

1):

```
ERROR: Access to the registry path is denied.
```

When starting MetMatch for the first time, a user account with administrative rights is needed, as some environment variables have to be set. If you don't have access to such an account, please contact your system administrator.

2):

```
'setx' is not recognized as an internal or external command, operable
program or batch file.
```

Please check if "C:\Windows\System32" is part of the PATH in your Environment Variables.

If “Initialize-Test.bat” finishes without any error messages, double click “MetMatch.bat” and check for the following error message:

```
Cannot find JRI native library! Please make sure that the JRI native
library is in a directory listed in java.library.path.
```

Please add the following entries (replace METMATCHPATH with the path of your MetMatch installation) to the PATH in your Environment Variables:

- METMATCHPATH\R\R-3.2.3\bin\x64
- METMATCHPATH\R\R-3.2.3\library\rJava\jri\x64

In case the problem persists, please report an issue at MetMatch’s GitHub repository (https://github.com/S-T-K/MetMatch).

2 Working with MetMatch

2.1 Introduction

MetMatch is as software tool for the semi-automatic comparison of different LC-HRMS metabolomics experiments. A typical use-case is the comparison of experiments acquired under different conditions (e.g. instruments) or separated by long stretches of time. Such comparisons would be hard to perform manually, since LC-HRMS datasets can be quite large (thousands of metabolites). But apart from the size, the main problem is the occurrence of shifts and drifts in both retention time (Rt) and mass to charge ratio (*m/z*). Using novel algorithms, MetMatch automatically detects and corrects them.

One experiment is defined as the reference. A list of metabolites is extracted from this experiment (consisting of one or multiple datafiles). This can be done manually or automatically, using a suitable software tool such as XCMS or MetExtract. The list has to be provided in the form of a tab separated file with column headers in the first row. Metabolites contained in this list are then matched to other experiments, provided as .mzXML files. Accurate results require a reference list that is compatible to the other experiments, meaning that a certain number of metabolites has to be present in both of them (the exact number varies from case to case). Matched metabolites can be investigated using the interactive interface, enabling the user to easily confirm or correct the match if necessary.

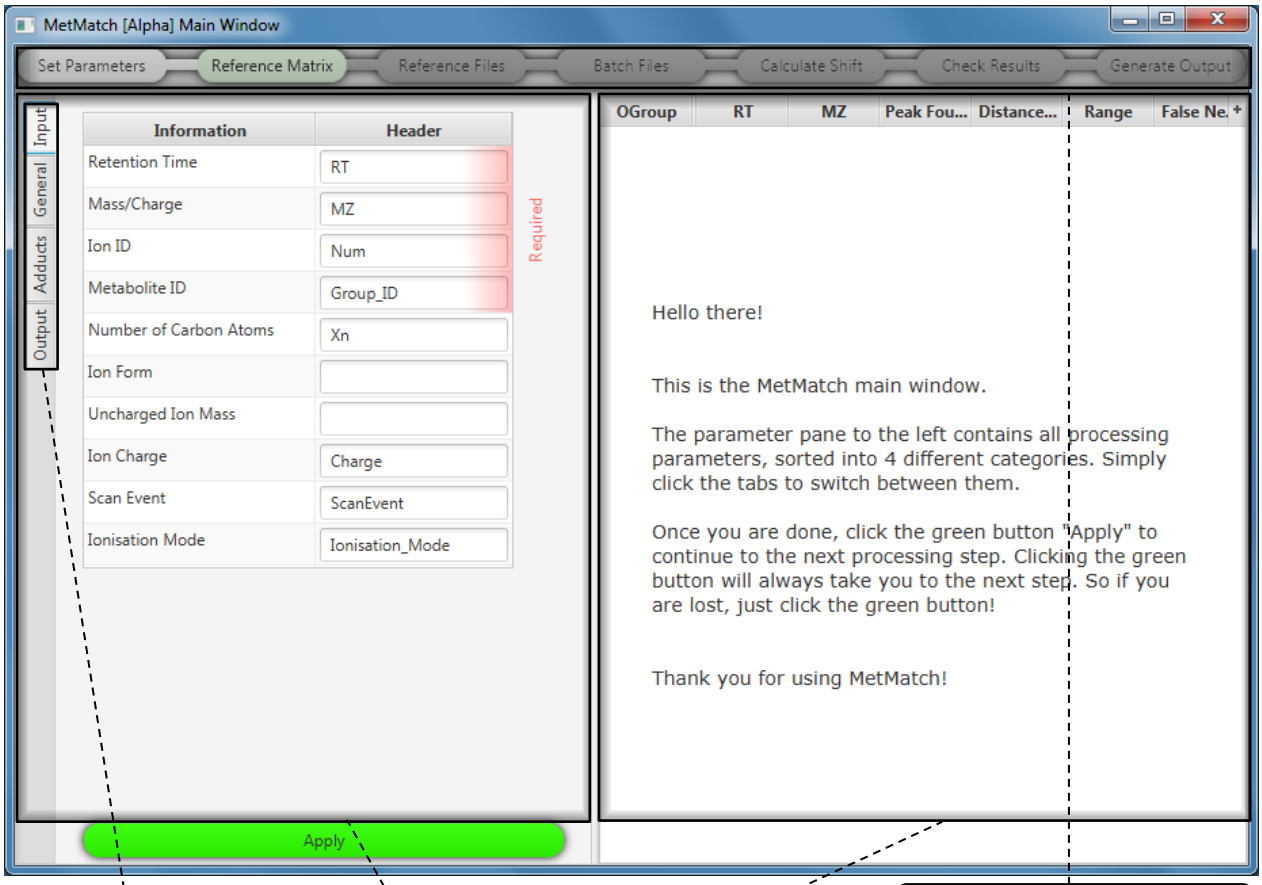
At the end, MetMatch outputs a list of metabolites closely resembling the input list. This list now also contains information derived from the other experiments, such as new metabolite adducts and most importantly the abundance of each of these metabolites in all processed files. This information could now be used for statistical analysis, for instance investigating the levels of newly identified metabolites in older experiments. Two example use cases can be found at http://www.mdpi.com/2218-1989/6/4/39.

2.2 Starting MetMatch

To start MetMatch, double click “MetMatch.bat” inside the main folder of your MetMatch installation path. A command line window (console) is opened immediately, displaying some routine checks. Don’t close this window, as this will terminate MetMatch. MetMatch’s main window should open up just a few seconds later.

2.3 Main Window

Figure 1 shows the main window of MetMatch. The main window displays the current processing step (green button) in the progress indication bar on top. This bar also serves as menu bar, which is used to start the respective task.



Parameter pane tabs:
Parameters are accessible by clicking on the respective category.

Parameter pane:
Contains various processing parameters, some of which may be disabled at certain processing steps (greyed out), and some may be changed throughout processing.

Adduct table:
Metabolite Data will be visible here once the reference matrix file has been loaded.

Progress indication bar:
Helps keep track of the current processing step

Figure 1: Main Window of MetMatch.

2.4 Data Processing

This chapter outlines the successive data processing steps and describes their function and the relevant parameters. Some of these steps are optional and are marked as such. Each step has its own item in the progress indication bar.

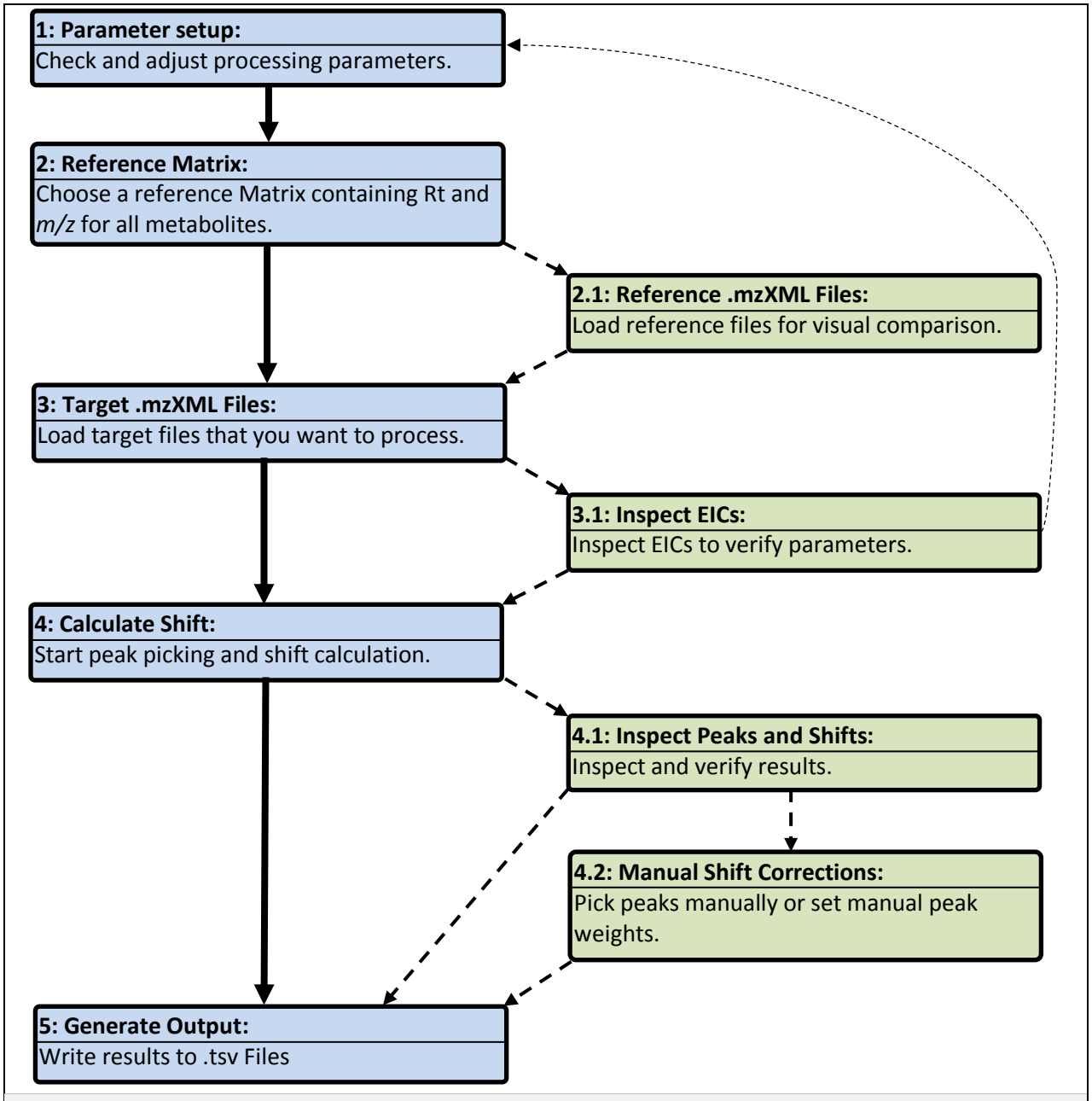


Figure 2: Mandatory (blue) and optional (green) processing steps.

2.4.1 Parameter Setup

Once you start MetMatch, the parameter pane is already displayed. Please adjust them to the data you are about to process. Some parameters may be changed during processing and are accessible by clicking "Set Parameters".

Once you have adjusted the parameters, proceed to the next step by clicking "Apply".

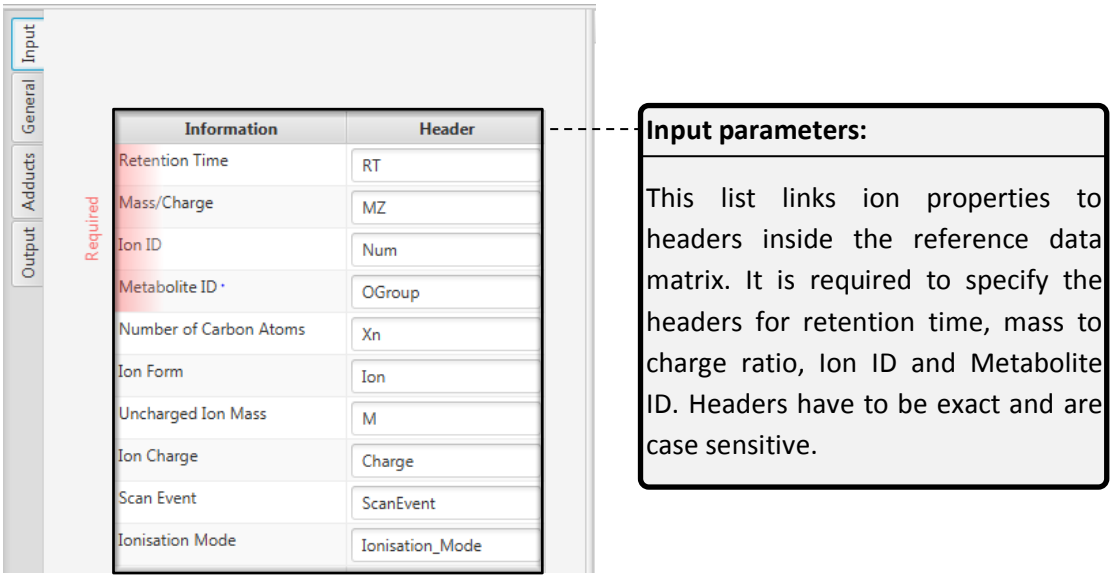


Figure 3: Input Parameters.

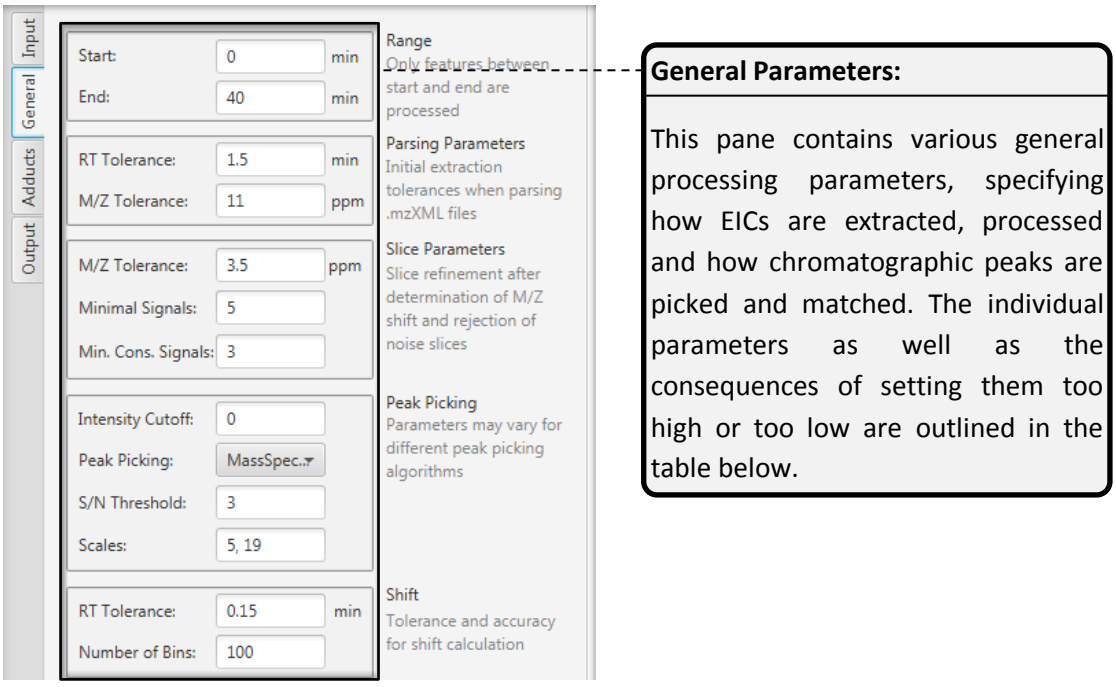
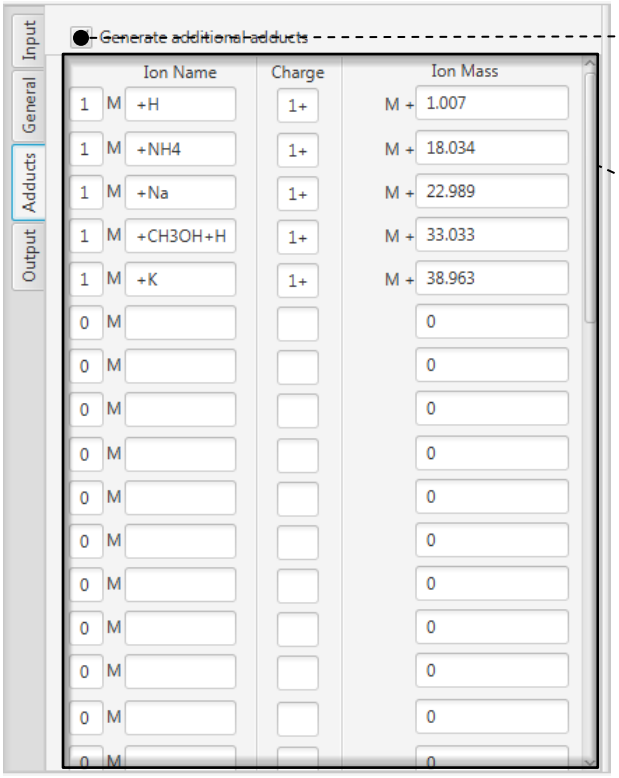


Figure 4: General Parameters.

Range:	
Start: Metabolites with a Rt lower than this value are not parsed from the input file	▲ Peaks with low Rt values are not included
	▼ Injection peaks are included
End: Metabolites with a Rt higher than this value are not parsed from the input file	▲ Peaks with high Rt values (possibly injection peaks) will be included
	▼ Peaks with high Rt values are not included
Parsing Parameters	
Rt Tolerance: Signals with a Rt deviation lower or equal to this value will be extracted. The deviation is relative to the Rt specified in the reference data matrix	▲ Setting this value too high will decrease performance because many unrelated signals are processed and could lead to wrong results.
	▼ Signals of interest are not extracted.
***m/z* Tolerance:** Signals with an *m/z* deviation lower or equal to this value will be extracted. The deviation is relative to the Rt specified in the reference data matrix	▲ Performance is decreased because many unrelated signals are extracted.
	▼ Signals of interest are not extracted.
Slice Parameters	
***m/z* Tolerance:** Any signal that deviates by more than this value from the calculated *m/z* shift will be deleted from the EIC	▲ Unrelated signals are part of the peak
	▼ Peaks appear spiky, because some of their signals are not included
Minimal Signals: All slices with less than this number of signals get rejected, as they hold no valuable information.	▲ Slices containing signals of interest are rejected.
	▼ Slices only containing noise signals are not rejected.
Min. Cons. Signals: All slices with less than this number of consecutive signals will be rejected.	▲ Slices containing signals of interest are rejected.
	▼ Slices only containing noise signals are not rejected.
Peak Picking	
Cutoff: Any peak with an abundance less than this value will be rejected	▲ Small peaks are rejected
	▼ Noise will be included
Peak Picking: for details see chapter 2.4.6	
NU (S/N) Threshold: Depending on the chosen peak picking algorithm, peaks with a S/N (signal-to-noise ratio) or NU (noise unit value) smaller than this value will be rejected	▲ Small peaks are rejected
	▼ Noise will be included
Scales: Scale parameters for the MassSpecWavelet algorithm, comma separated.	▲ Narrow peaks are rejected
	▼ Noise peaks will be included
Shift	
Rt Tolerance: When matching chromatographic peaks, only peaks with a deviation of less than this value from the calculated Rt shift are taken into consideration	▲ Unrelated peaks are matched
	▼ True peaks are not matched
Number of Bins: This parameter is used during the calculation of the retention time shifts	▲ Performance is decreased
	▼ Accuracy is decreased

Table 1: General Parameters and the consequences of setting them too high or too low.

▲/▼: too high/low values are not critical to the correctness of the results
▲/▼: too high/low values are critical to the correctness of the results
▲/▼: too high/low values are extremely critical to the correctness of the results

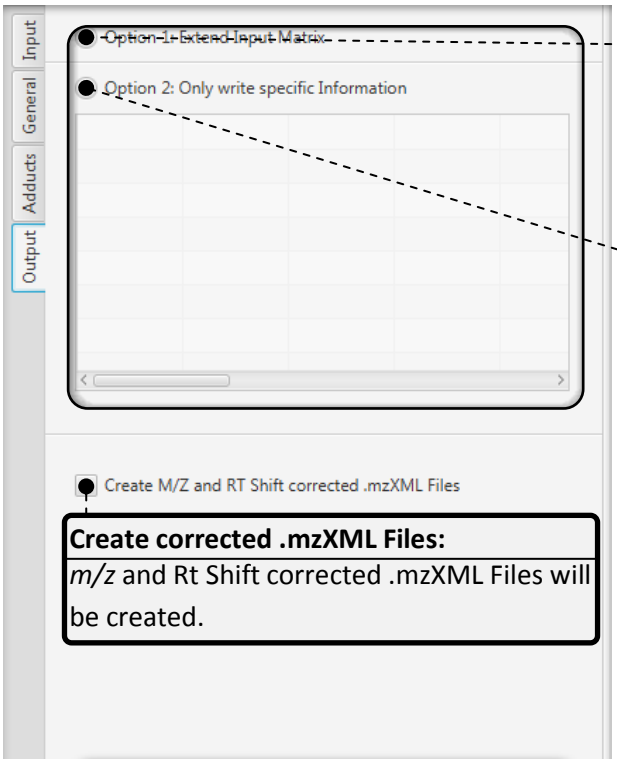


Generate additional adducts:
The adduct generation step can be skipped by unchecking this box.

Adduct Parameters:

Adducts that should be included in the adduct generation step are listed here, each with their respective number of metabolite molecules, ion name, charge, and ion mass. Changing any of these parameters after choosing a data matrix is not possible.

Figure 5: Adduct Parameters.



Option 1: Extended Input Matrix:
Extends the given input reference data matrix with the new processed information.

Option 2: Only write specific information:

To generate a more lightweight or completely customized output matrix, this option can be checked. The table represents the final output. To customize it, simply click on a column header (the first row) and start entering the desired information. MetMatch suggests matching possible options which can be chosen by clicking on them. The prefix "Files:" indicates that this information will be written for each file. At least one "Files:" information has to be included.

Create corrected .mzXML Files:
m/z and Rt Shift corrected .mzXML Files will be created.

Figure 6: Output Parameters.

All output parameters can be changed at any time. Thus several different output files containing information of the same processing run can be generated.

The option to create corrected .mzXML files can be used to check the quality of the alignment using any other tool that can parse .mzXML files. Further iterative refinements can be performed on the corrected file by reprocessing it with MetMatch.

2.4.2 Reference Matrix

MetMatch needs a matrix of reference ions in the form of a tab separated file (.tsv or .txt). These files can be generated automatically or manually. The headers need to be in the first row. Linking headers with specific metabolite information is done through parameters (Figure 3) inside the MetMatch user interface before actually choosing a matrix file. To choose a matrix, click “Reference Matrix”.

After processing, taking a few seconds up to minutes depending on the number of adducts specified (Figure 5), the main window should look like this:

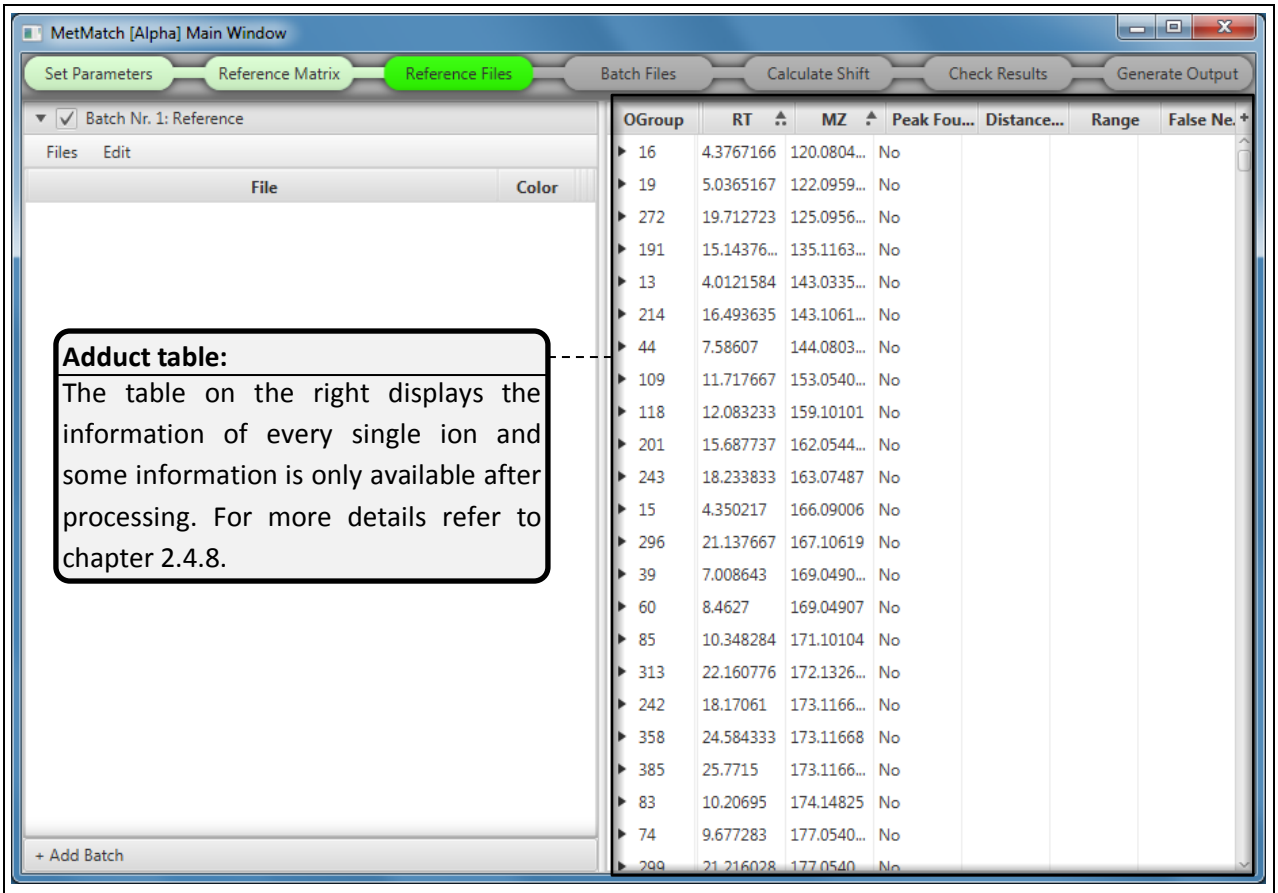


Figure 7: Main GUI once a reference Matrix has been loaded.

2.4.3 Reference .mzxml Files

MetMatch parses all information needed directly from .mzxml files. It is recommended to also provide the respective files from which the reference data matrix has been extracted. This helps estimating the quality of the alignment visually through comparison with the reference EIC. Because of this reason, the first batch has been named “Reference” by default.

If the original files are not available, or the reference matrix has been created manually, this step can be skipped by loading any .mzXML file and deleting it afterwards. Files can be deleted using the context menu (right click on file).

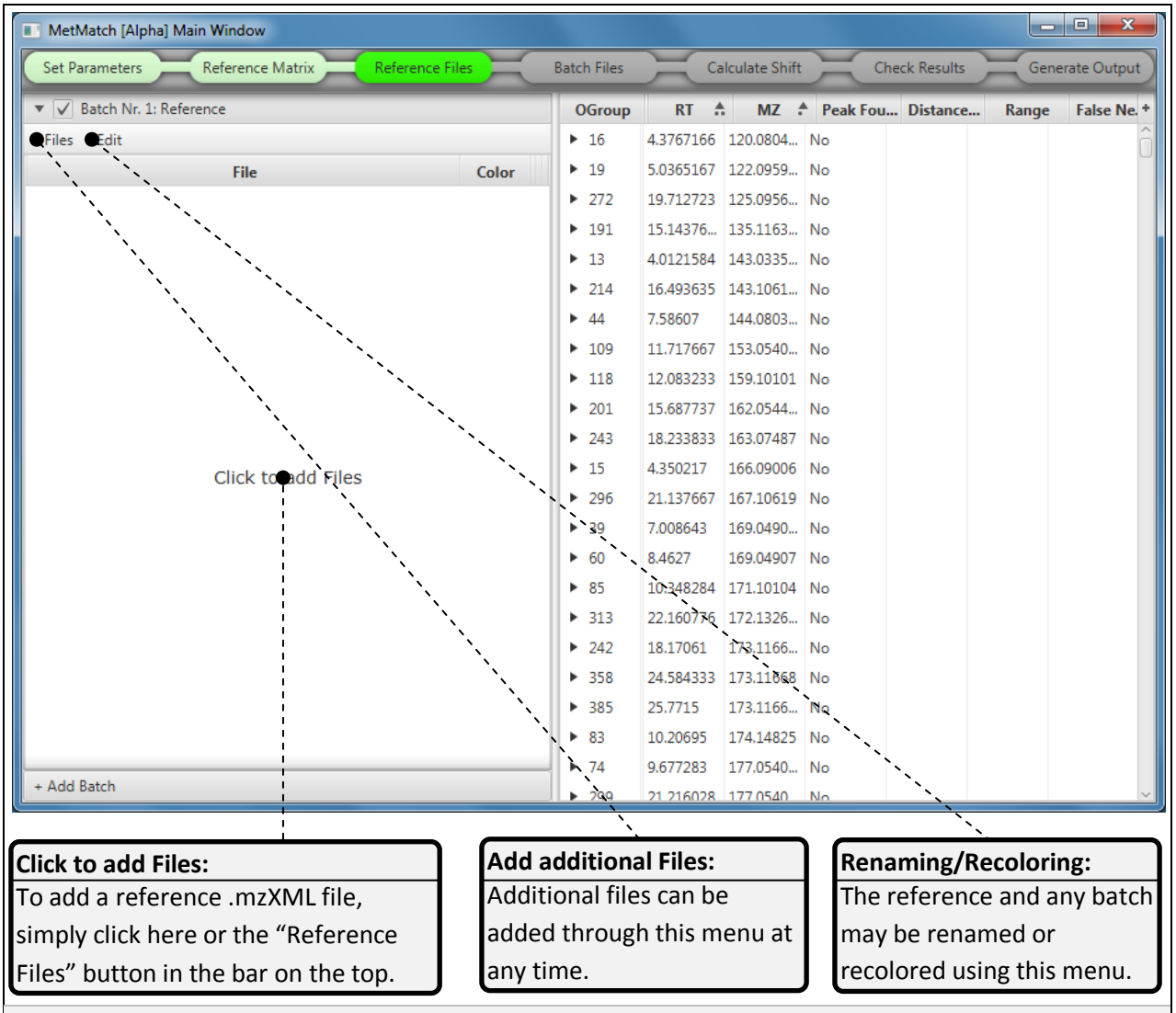


Figure 8: Reference file pane.

2.4.4 Adding new Batches

Files are organized into batches within MetMatch. Batches can be used to load and compare files from different experiments and can be visually customized to easily distinguish them.

To create a new batch, click “Add Batch” and then “Files”->”Add new .mzXML” to load files, or simply click “Target Files” in the progress indication bar, which automatically creates a new batch and opens up the file choice dialog.

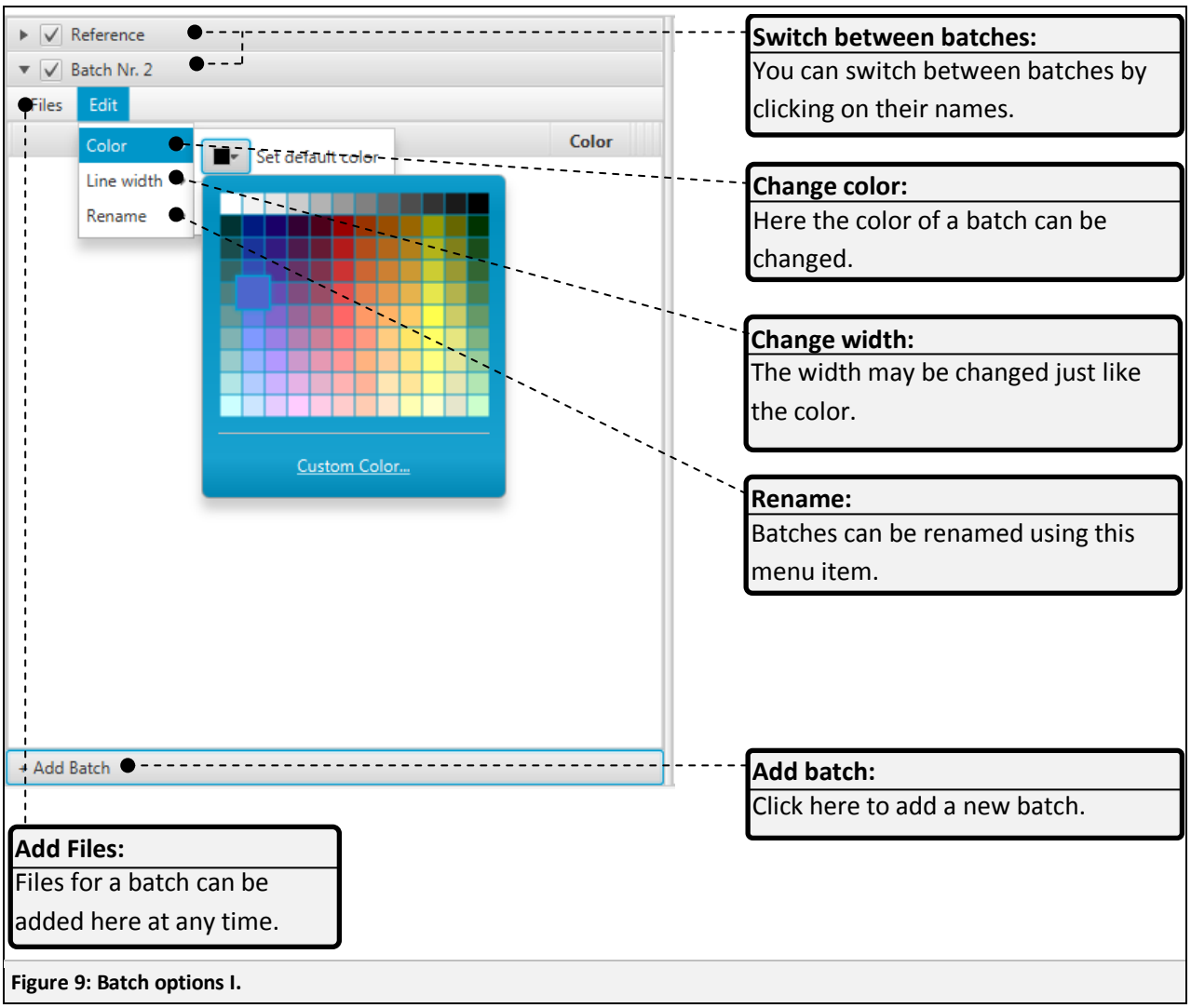


Figure 9: Batch options I.

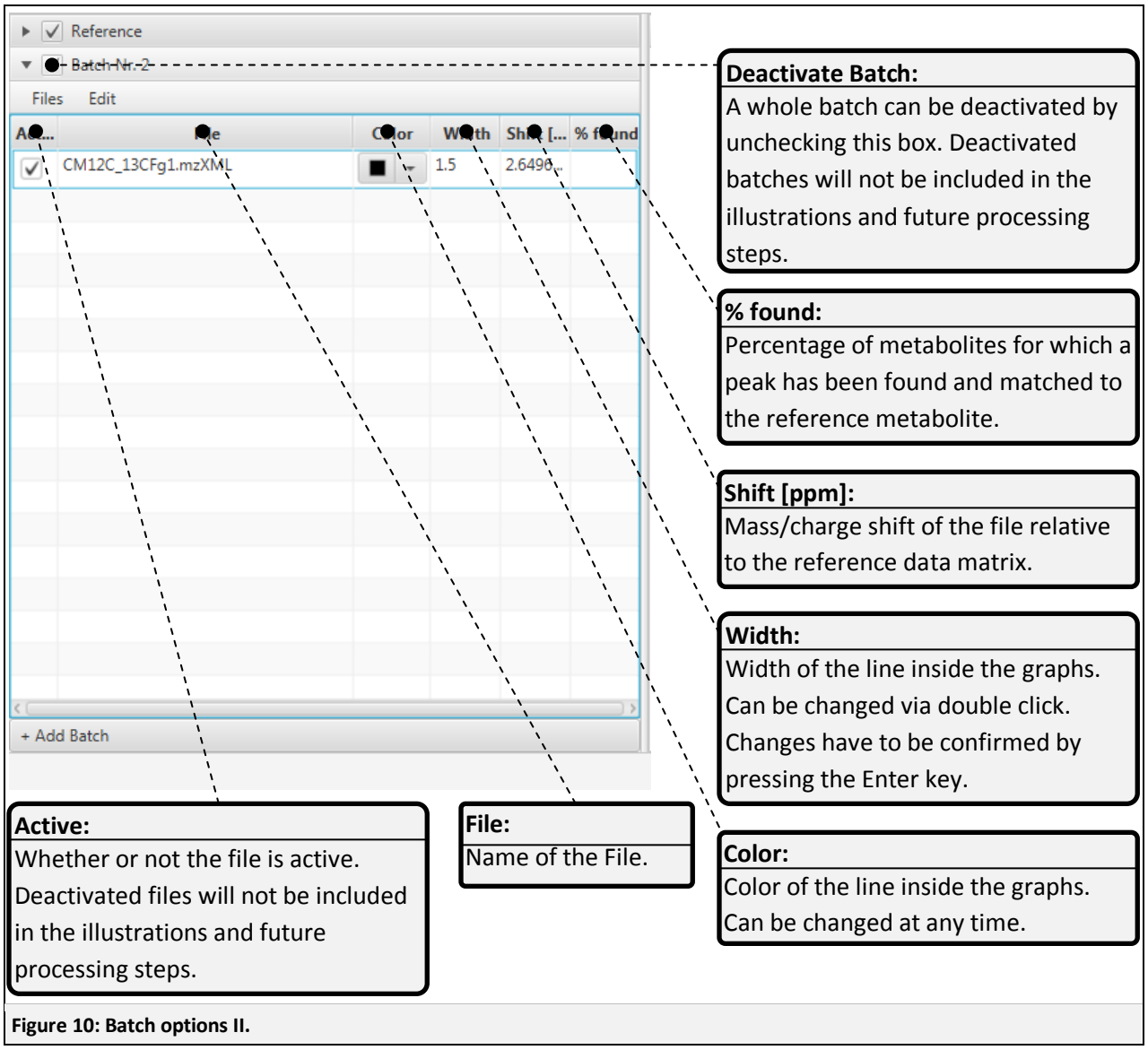


Figure 10: Batch options II.

2.4.5 Inspecting EICs

In this optional step, the EICs can be inspected to ensure the various parameters have been set correctly.

The available chart types (accessible in the menu “Show/Hide Charts”) are:

- **EIC:** this is a basic EIC chart.
- **Normalized EIC:** this chart normalizes the EIC of each file to improve comparing the shape of peaks with different abundances.
- ***m/z*:** this chart visualizes the ppm deviation for each individual signal.
- **Shift:** this is a slightly different version of the chart displayed in Figure 18. It shows the retention time shift function for all files and an indicator for the current retention time (once the function has been computed).

To see the EIC of specific metabolites ions, double click the respective row inside the adduct table (Figure 7). This should open up a new window resembling this one:

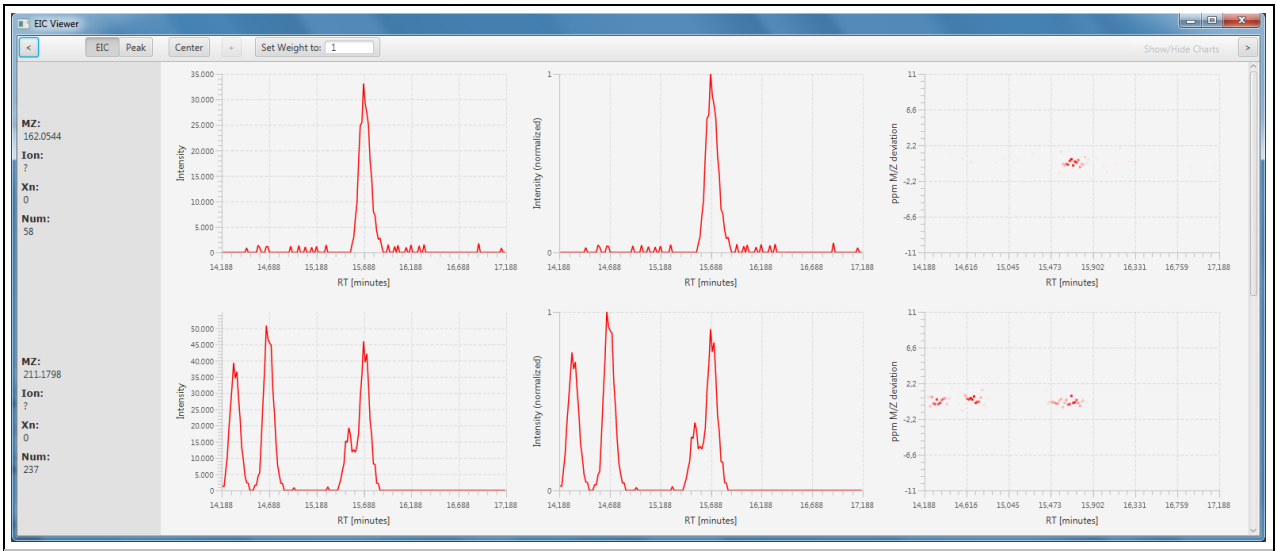
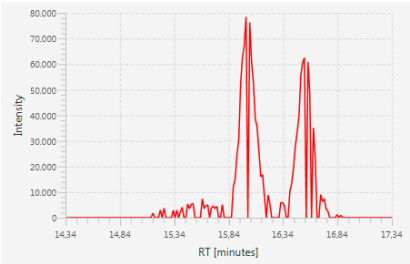
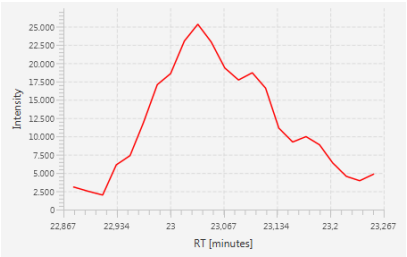


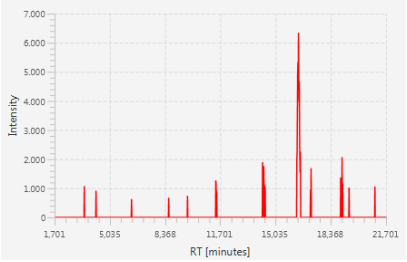
Figure 11: EIC window, for further details on this window, see Figure 20.



In this case the *m/z* tolerance (Figure 4) was set too narrow, leading to spiky EICs because of missing signals.



This shows an Rt tolerance (Figure 4) that was set too narrow, leading to EICs consisting only of a few signals (they appear jerky) and peaks that appear to be very broad. If the actual retention time shift is bigger than the Rt tolerance, the correct peak cannot be found. Limiting the Rt tolerance might shorten processing times, but it is not recommended to do so without verification.



Here on the other hand, the Rt tolerance (Figure 4) was set too high. This unnecessarily increases processing times and could potentially throw off peak picking and shift detection. To prevent this, the number of bins would have to be set accordingly (around 10 times the value of Rt tolerance).

Figure 12: Examples of parameters that need to be adjusted.

2.4.6 Peak Picking and Rt Shift Detection

MetMatch offers 3 different peak picking algorithms.

- **Gauss Peak Correlation:** This algorithm detects peaks by comparing the EIC to an idealized Gaussian peak. If the EIC has a high correlation to the Gaussian peak, a peak is detected.
- **Savitzky-Golay Filter:** Each EIC is smoothed using a Savitzky Golay filter, and the minima and maxima of the resulting function are used to detect the peaks. This algorithm is very fast.
- **MassSpecWavelet:** Peaks are detected using a wavelet calculation. This takes very long to compute (it uses an R script).

To change the currently active algorithm, click “Set Parameters”, then choose the tab “General” and pick the desired algorithm under “Peak Picking”. Apply the changes by clicking on “Apply”.

If your reference matrix has been calculated using one of these algorithms, it is recommended to use the same one to ensure comparability.

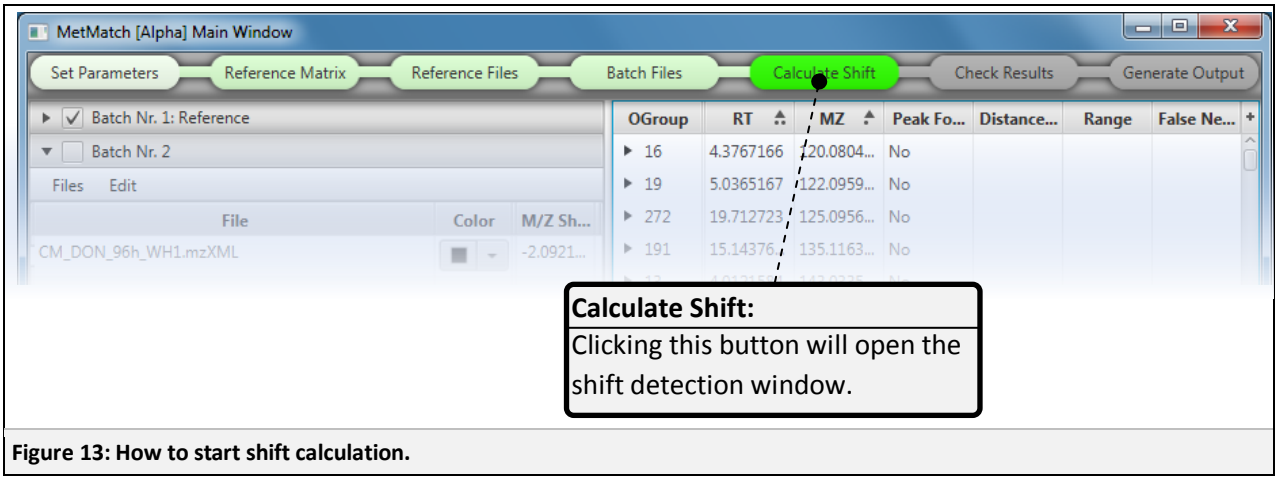


Figure 13: How to start shift calculation.

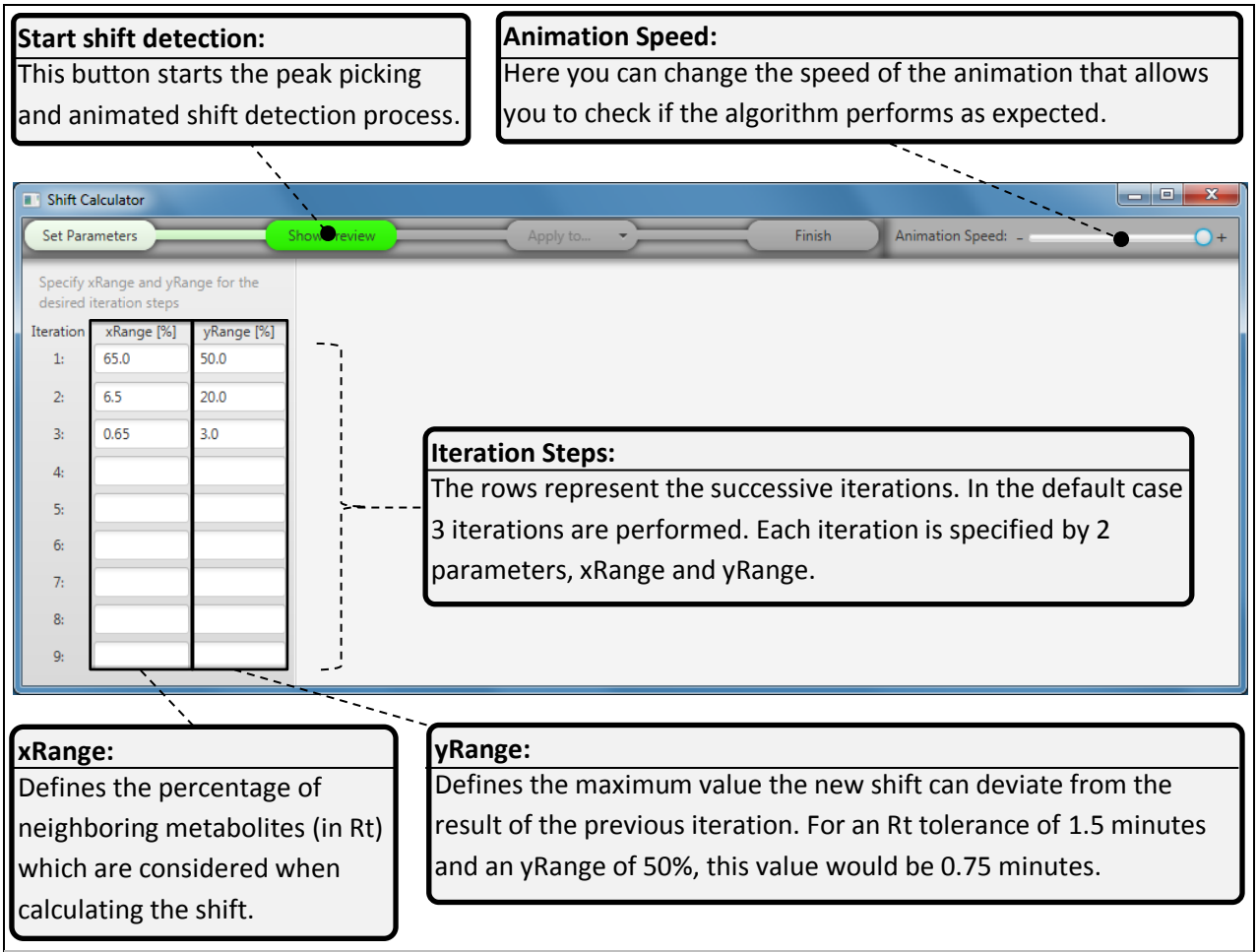
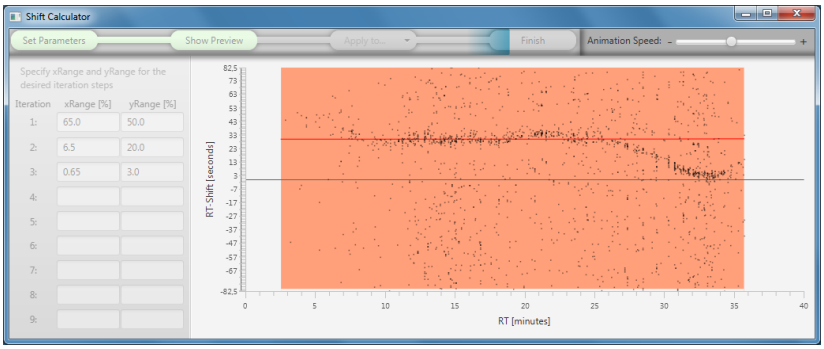


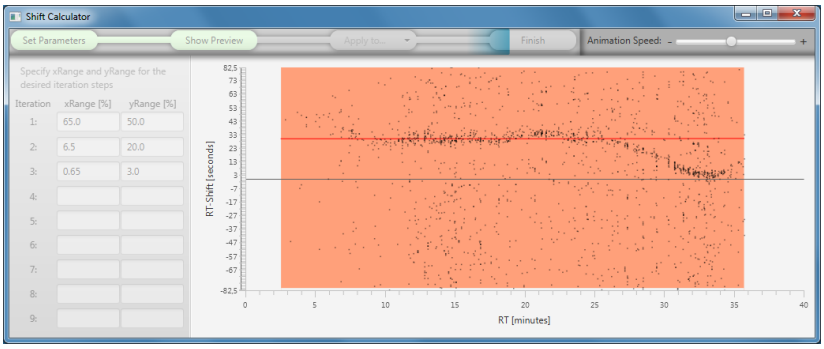
Figure 14: Shift calculation parameters.

For a detailed description of how the algorithm works, please refer to http://www.mdpi.com/2218-1989/6/4/39.

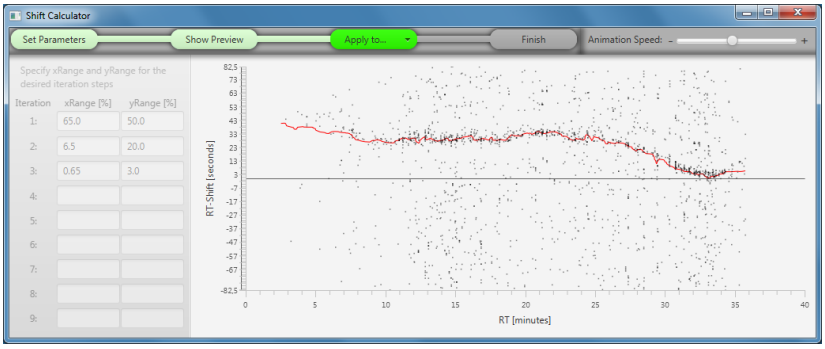
Note: If you wish to illustrate the algorithm on a specific file, select it before clicking on “Show Preview”. Otherwise the first target file in the list is chosen by default. Peak picking is performed, and all detected peaks are marked as dots on the graph. The first iteration is calculated and its yRange and resulting shift function (orange area and red line in Figure 15) are displayed. The consecutive iteration steps are animated, allowing you to estimate whether or not it works as expected, and to tweak the parameters if necessary.



The xRange of 65% yields an almost linear shift in the first iteration.

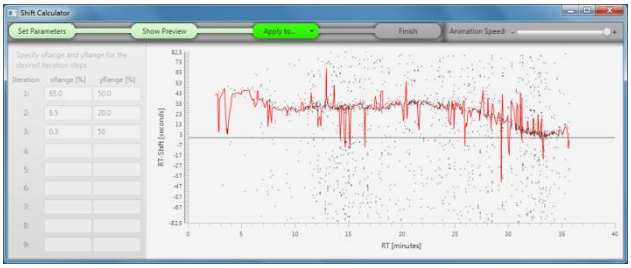


Local minima are taken into account using a smaller yRange of 6.5% in the second iteration.

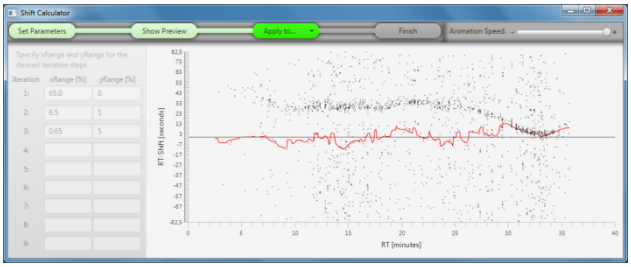


Even more local minima are found during the last iteration. A relatively small xRange (3%) ensures that they don't deviate too much from the results of the previous iteration.

Figure 15: Three iteration steps of the Rt shift detection algorithm.

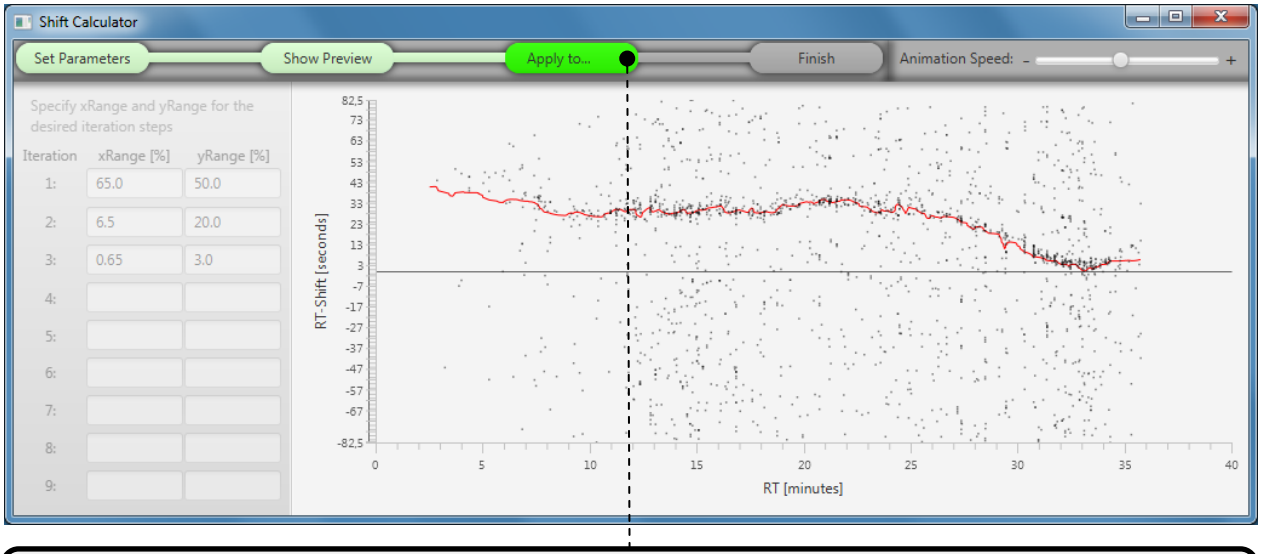


Very spiky shifts like this one are the result of an iteration step with a small xRange and big yRange. Because of this, the algorithm tried to find local maxima too far away from a better global maximum.



It is recommended to set the yRange to 50% for the first iteration. This ensures that the global maximum is found. In this case the global maximum was outside of the allowed range.

Figure 16: Examples of parameters that need to be adjusted.



Apply to…:

Once the parameters have been verified, you can use them on:

- All Files: Shift detection is started for all files using the parameters specified.
- This Batch: Only calculates the current batch (first batch with a selected file). This option is useful if you wish to fine tune the parameters for every single batch.

Figure 17: How to apply shift calculation to all files or one batch.

Figure 18: Shift calculation window after processing two different batches.

In Figure 18, the calculated shift for the target (green) and the reference files (blue) is displayed.

Double clicking on any of the lines opens up a new window displaying the EICs for the selected metabolite. It automatically selects the file represented by the line and highlights it in bright red.

In this case the reference file inside the red circle was chosen to inspect the relatively large shift when compared to the other reference files.

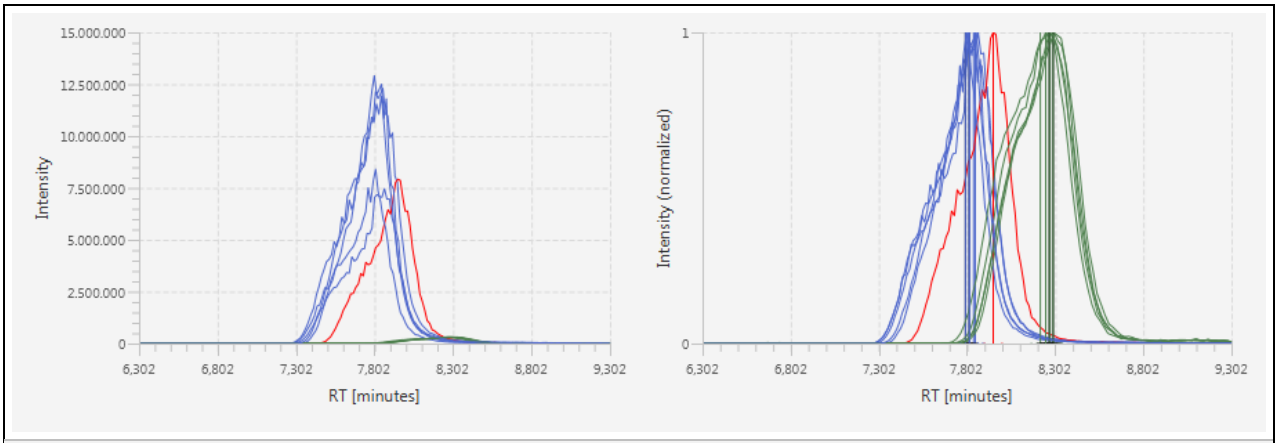


Figure 19: EICs showing shift within the reference files. Blue: reference files. Red: reference file with large shift.

The calculated shift is easily confirmed by looking at the EICs. The selected file indeed shows some shift relative to the other reference files (blue).

The region around 4 to 6 minutes shows no consensus within the batch files, indicating either only few peaks and/or inconsistent Rt shifts. To help the algorithm in calculating the shift, you can use some more advanced MetMatch features.

Here you can see the EIC window after shift detection.

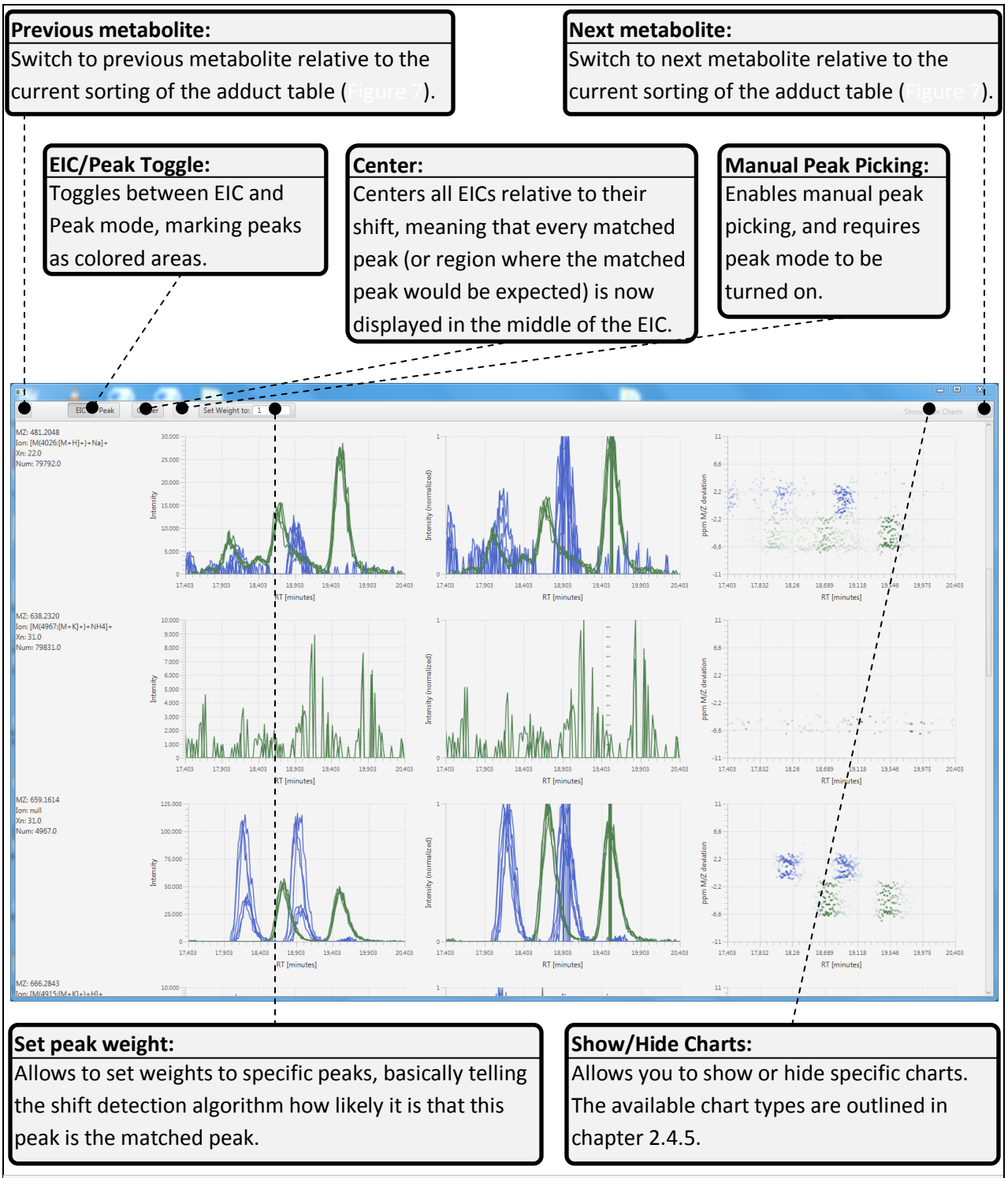


Figure 20: EIC window options.

2.4.7 Manual Shift Corrections

Manual peak picking:

For manual peak picking, simply select the desired files (highlighting them in red) and click the "+" button. Now locate the peak and draw a box around it via click and drag inside the normalized EIC chart. MetMatch will look for peaks inside this box. Note that it doesn't matter whether or not you draw the box high enough to include the peak height, only the Rt information is taken into account. A very basic peak picking algorithm is applied.

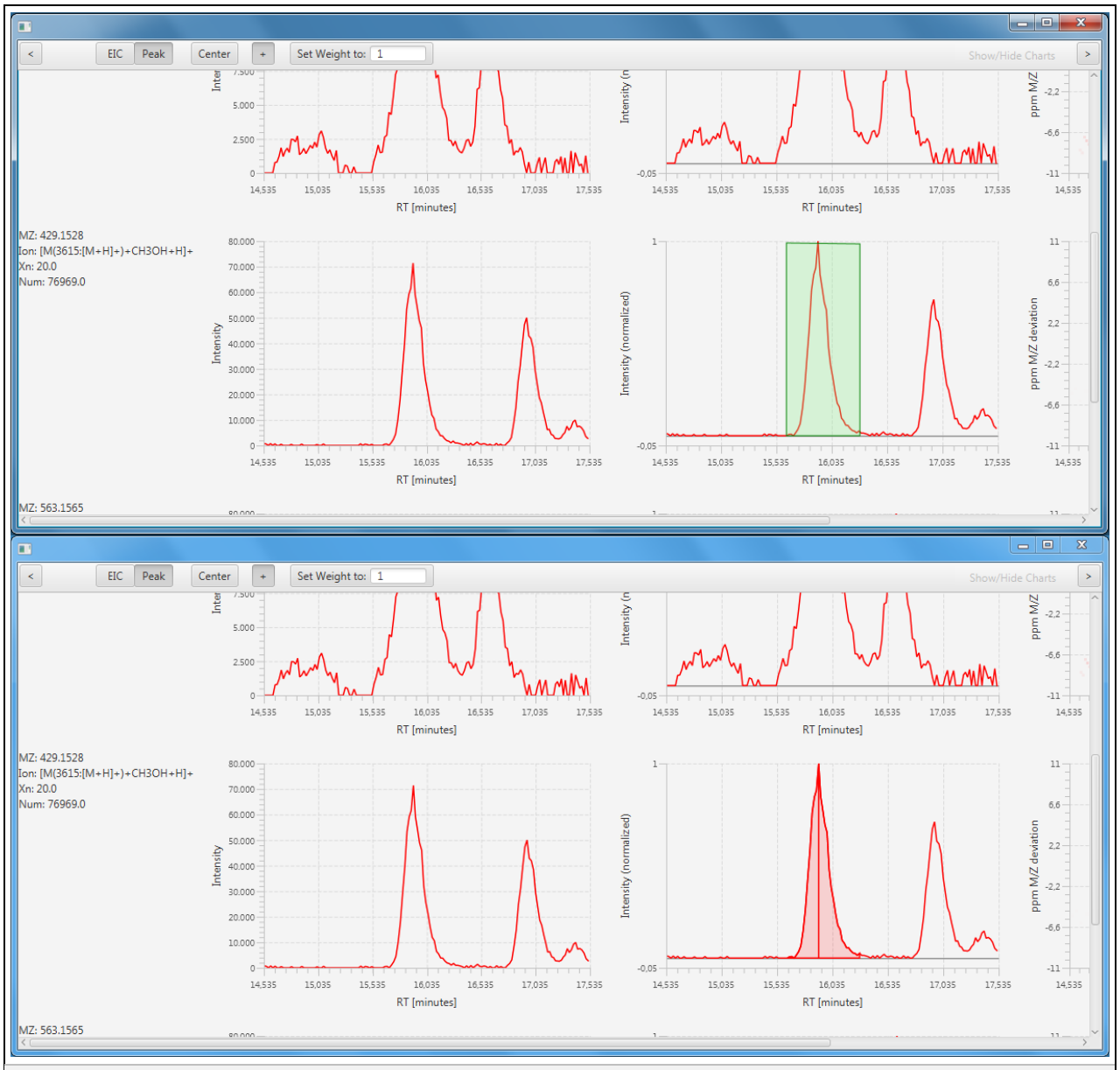


Figure 21: Manual peak picking. Top: User drawing a box around desired peak. Bottom: Successfully detected peak.

Peak weight:

All detected peaks have the same weight by default, meaning the algorithm treats all as equally likely matches. Higher weights indicate that a peak is more likely to be the matched peak.

To set a peak weight, select the desired files, click on "Set Weight to:", enter the weight (or drag the slider) and click and drag a box inside the normalized EIC chart, analogous to manual peak picking. Any peak whose maximum is inside this box will be assigned the new weight.

2.4.8 Inspecting Peaks and Shifts

In this optional step, peaks and shifts can be inspected.

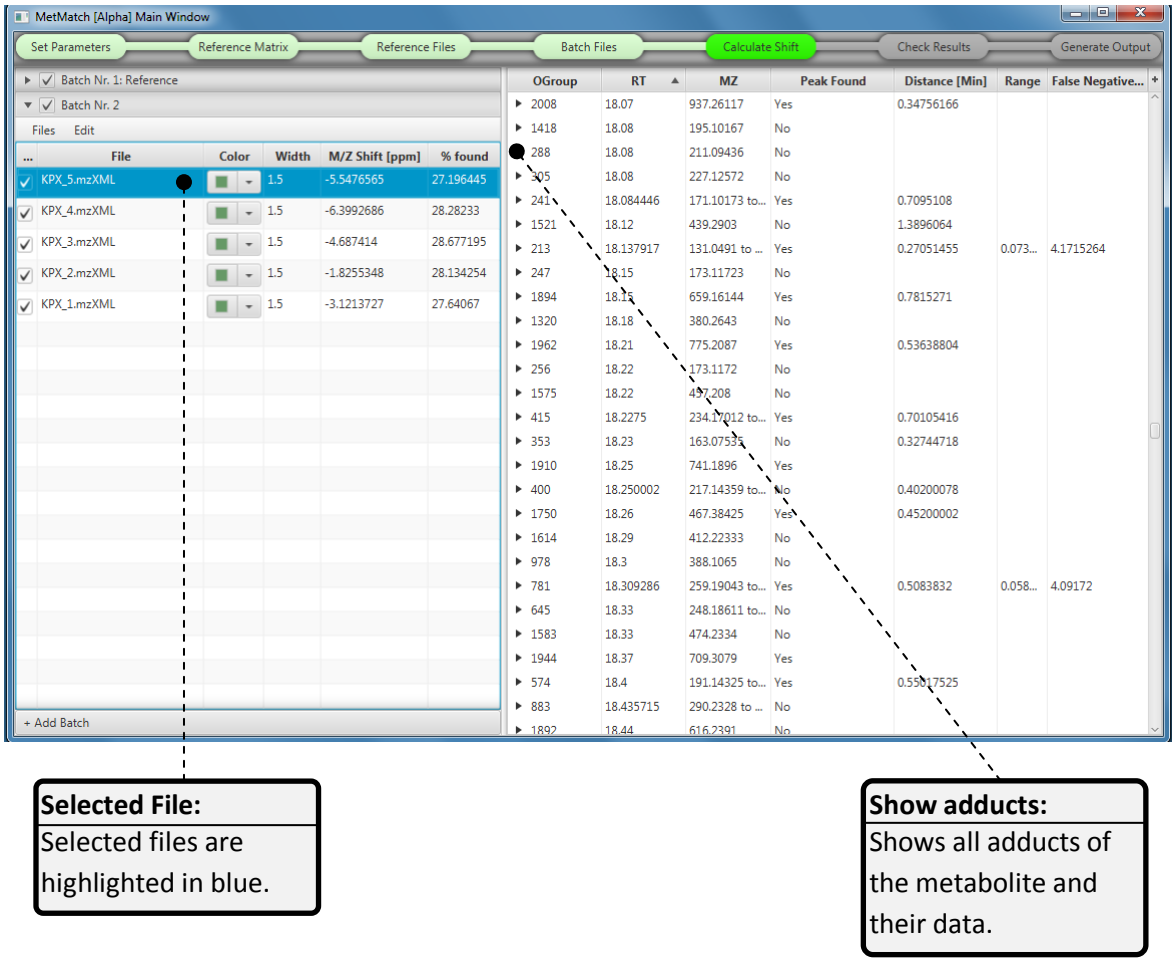


Figure 22: How to get detailed information about the adducts of one or multiple files.

The adduct table in the main GUI can help you finding specific EICs that could be worth investigating. The displayed information is either the lowest or the highest value of all selected files.

Information	Metabolite shows which value of all its adducts?	Table shows which value of all selected files?
PeakFound: This tells you whether or not a matched peak could be found. 0=not found, 1=found.	maximum	maximum
Distance: The distance of the peak closest to the calculated shift is shown. Peaks that have been marked as the matched peak are not considered.	minimum	minimum
Range: The range is the maximum distance between all matched peaks of a metabolite. If the matched peaks for all adducts have the exact same retention time, it is 0. This value cannot exceed 2 times the value of the shift Rt tolerance (Figure 4).	Range is only defined for metabolites	maximum
False Negative NU: If no matched peak has been found, the noise unit value of a hypothetical peak at the calculated shift is displayed. High values indicate that a potential peak was not detected during peak picking.	maximum	maximum

Table 2: Information displayed in the adduct table.

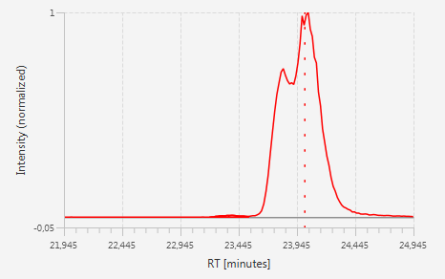
In the following, some use cases are outlined.

Finding Metabolites with/without matched peaks

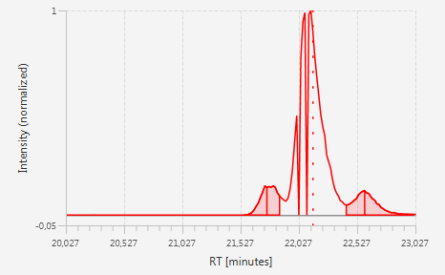
1) Select the desired files.
2) Click the “PeakFound” header once if you want to find metabolites without matched peak, or twice if you want matched peaks.
3) Metabolites are now sorted in the desired order and can be inspected.

Finding peaks not detected during peak picking

1) Select the desired files.
2) Click the “False Negative” header twice, sorting it in descending order.
3) Take a closer look at the metabolites with high values.



This not detected peak has a noise unit value of about 50, and was probably not detected due to some error in peak detection.



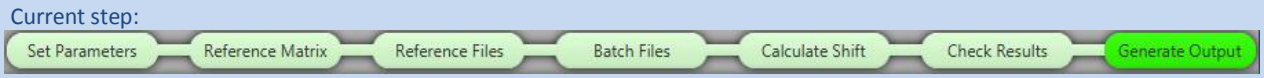
This peak was not detected because of a very spiky EIC. Take a closer look at the m/z chart to determine whether you should widen the slice *m/z* tolerance (Figure 4).

Finding peaks that were just barely not detected as matched peaks

1) Select the desired files.
2) Click the “PeakFound” header once, sorting it in ascending order.
3) On your keyboard hold “Shift” and click the “Distance” header once, sorting it in ascending order.
4) Metabolites that were not found, but have a peak very close to the calculated shift are now at the top of the list.

Finding peaks with a suspiciously high range

1) Select the desired files.
2) Click the “Range” header twice, sorting it in descending order.
3) Metabolites whose matched peaks have a high retention time range in between them are now at the top of the list.



2.4.9 Output generation

Once you are satisfied with the processing results, you can generate the output file. You have 2 options, outlined in Figure 6.

To generate the output file, simply click “Generate Output”.

After this process is completed (it should only take a few seconds) an explorer-window showing the file opens automatically. The name of the result file is the same as the reference file with the prefix “matched_”. Corrected .mzxml files are marked by the prefix “MetMatch-corrected_”.