El-MAVEN Documentation

Release v0.4.1

Elucidata Corporation

Oct 18, 2018

Contents

An intuitive, opensource LC-MS data processing engine from

Travis Build

Digital Object Identifier

Contents

1.1 El-MAVEN Features

MAVEN and El-MAVEN share the following features:

- Multi-file chromatographic aligner
- Peak-feature detector
- Isotope and adduct calculator
- Formula predictor
- Pathway visualizer
- Isotopic flux animator

El-MAVEN is faster, more robust and has more user friendly features as compared to MAVEN.

1.2 Download

El-MAVEN installers are available for Windows (7, 8, 10) and MacOS (1.5 and above). Download El-MAVEN's latest version or daily build for your preferred environment.

1.3 Build

Contributors can build El-MAVEN on Windows, Ubuntu or Mac systems by following these instructions. Users are recommended to download the installers provided here.

1.3.1 Windows

- Download MSYS2 installer and follow the installation instructions provided on their website.

- Open MSYS2 and give the following commands to set up libraries and tool chains for El-MAVEN. Reopen MSYS2 when required:

    – For 64 bit

    ```
    $ pacman --force -Sy
    $ pacman --force -Syu
    $ pacman --force -Su
    $ pacman --force -Sy base-devel msys2-devel mingw-w64-x86_64-toolchain␣
    ↪mingw-w64-x86_64-qt5 mingw64/mingw-w64-x86_64-hdf5 mingw64/mingw-w64-
    ↪x86_64-netcdf mingw64/mingw-w64-x86_64-boost msys/git
    ```

    – For 32 bit

    ```
    $ pacman --force -Sy
    $ pacman --force -Syu
    $ pacman --force -Su
    $ pacman --force -Sy base-devel msys2-devel mingw-i686-toolchain mingw-
    ↪i686-qt5 mingw32/mingw-i686-hdf5 mingw32/mingw-i686-netcdf mingw32/
    ↪mingw-i686-boost msys/git
    ```

- Open mingw64.exe from the MSYS2 folder and give the following commands:

    ```
    $ cd <PathToInstallationFolder> #for example: cd /c/User/Admin/Desktop
    $ git clone https://github.com/ElucidataInc/ElMaven.git
    $ ./run.sh
    $ ./bin/El_Maven_0.x #for example: ./bin/El_Maven_0.2
    ```

El-MAVEN loads with two windows: one for logging the application status and another for data analysis.

1.3.2 Ubuntu

- Open the terminal and give the following commands to set up the libraries and tool chains for El-MAVEN:

    ```
    $ sudo apt-get update
    $ sudo apt-get install g++
    $ sudo apt-get install qt5-qmake qtbase5-dev qtscript5-dev qtdeclarative5-
    ↪dev libqt5webkit5-dev libsqlite3-dev libboost-all-dev lcov libnetcdf-dev
    $ cd <PathToInstallationFolder> #for example: user@pc:~$ cd Desktop/
    $ git clone https://github.com/ElucidataInc/ElMaven.git
    $ ./run.sh
    $ ./bin/El_Maven_0.x #for example: ./bin/El_Maven_0.2
    ```

El-MAVEN loads with two windows: one for logging the application status and another for data analysis.

1.3.3 MacOS

- Install Xcode from App store
- Download and Install Qt5.6

This will give you the Qt5.6.2 dmg file. Using the dmg file install Qt under the directory /Users/Your_User_Name/

- Using the terminal execute the following commands:

```
$ sudo xcodebuild -license accept
$ xcode-select --install
$ /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/
→Homebrew/install/master/install)"
$ brew install boost
$ brew install llvm@3.7
$ brew install netcdf
$ cd ~
$ touch .profile
$ echo "PATH=/Users/$USER/Qt5.6.2/5.6/clang_64/bin/:$PATH" > .profile
$ source .profile
$ mkdir ~/maven_repo
$ cd ~/maven_repo
$ git clone https://github.com/ElucidataInc/ElMaven.git
$ cd ElMaven
$ source ~/.profile
$ qmake CONFIG+=debug -o Makefile build.pro
$ make -j4
```

1.3.4 Switching Versions

Users can switch between versions once they have compiled El-MAVEN successfully on their system. Follow these steps to pull a specific release:

- Choose the version you wish to install from the list of releases. (We recommend the version tagged "Current Release". Past releases are not stable and should be avoided)

- Find the version tag (v0.2.x, 0.1.x, etc) on the left of release notes.

- Open your terminal and move to the installation folder

- Give the following commands in the terminal:

```
$ cd ElMaven
$ ./uninstall.sh
$ git checkout develop
$ git pull
$ git checkout v0.x.y (Example: v0.4.1)
```

  - Build the new version using the following commands:

    * For Windows and Ubuntu

```
$ ./run.sh
```

    * For MacOS

```
$ source ~/.profile
$ qmake CONFIG+=debug -o Makefile build.pro
$ make -j4
```

1.4 User Documentation

Welcome to the El-MAVEN user documentation!

El-MAVEN is an open source LC-MS data processing engine that is optimal for isotopomer labeling and untargeted metabolomic profiling experiments. Currently El-MAVEN exists as a desktop application that runs on Windows, Ubuntu and MacOS. The software can be used to view the mass spectra, align chromatograms, perform peak-feature detection and alignment for labeled and unlabeled mass spectrometry data. The aim of this software package is to reduce complexity of metabolomics analysis by using a highly intuitive interface for exploring and validating metabolomics data.

Find out more about the software through the links below.

1.4.1 Introduction

Git Repository

El-MAVEN's repository can be found on GitHub here.

1.4.2 Getting Started

El-MAVEN User Interface

Following is the general workflow involved in El-MAVEN:
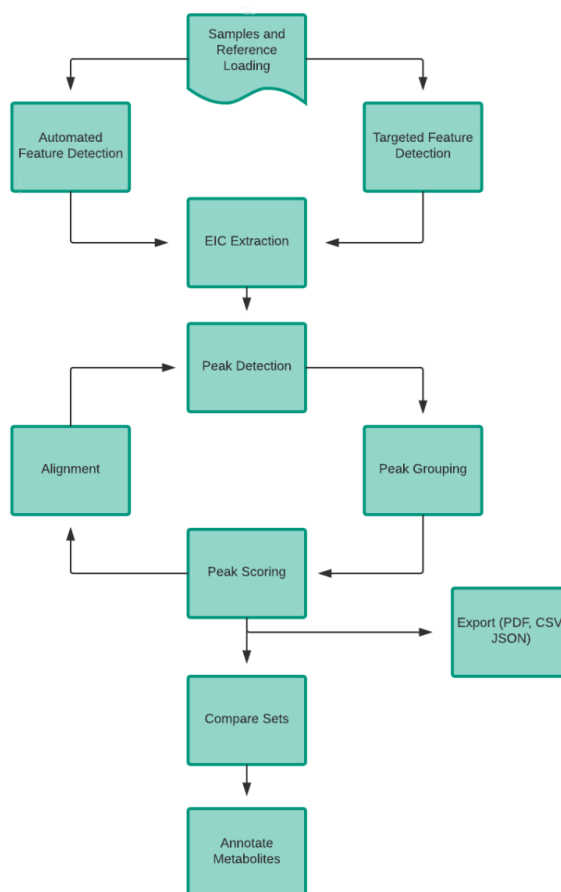
Peak detection, alignment, grouping and scoring are done multiple times for best results in the El-MAVEN workflow. Data from different cohorts can be compared using visualisation tools and easily exported to other formats.

## EI-MAVEN User Interface
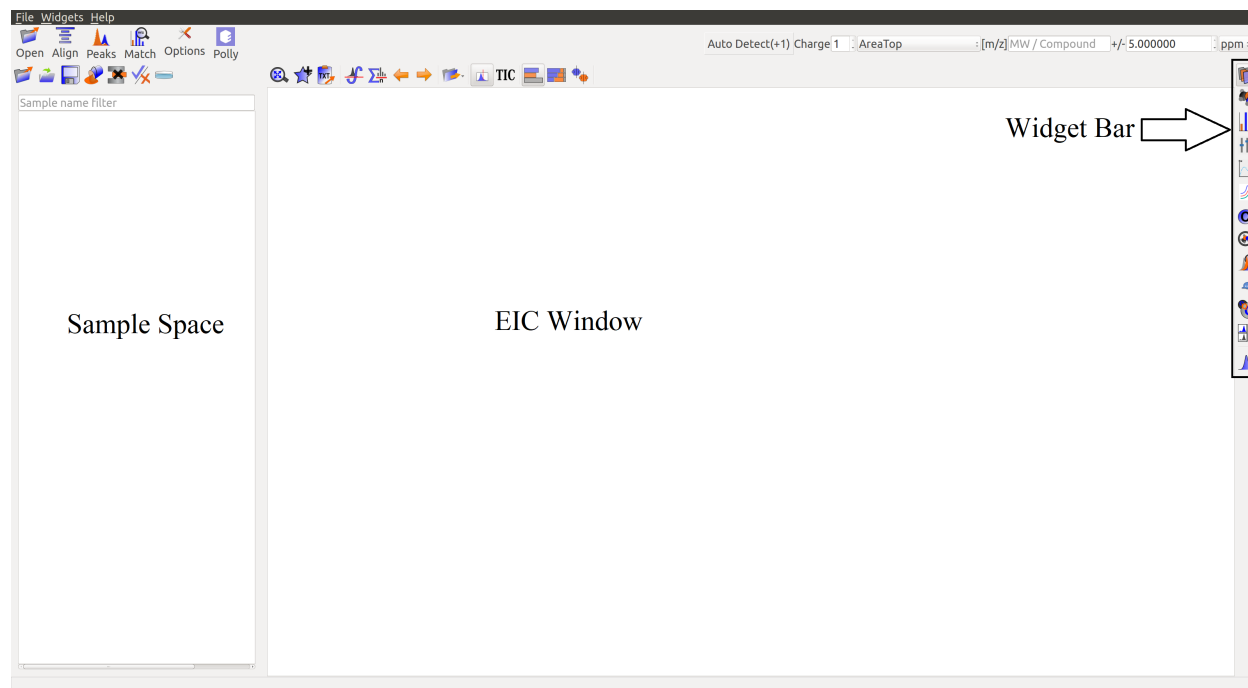
Global Settings

Global Settings can be changed from the Options dialog .

Instrumentation
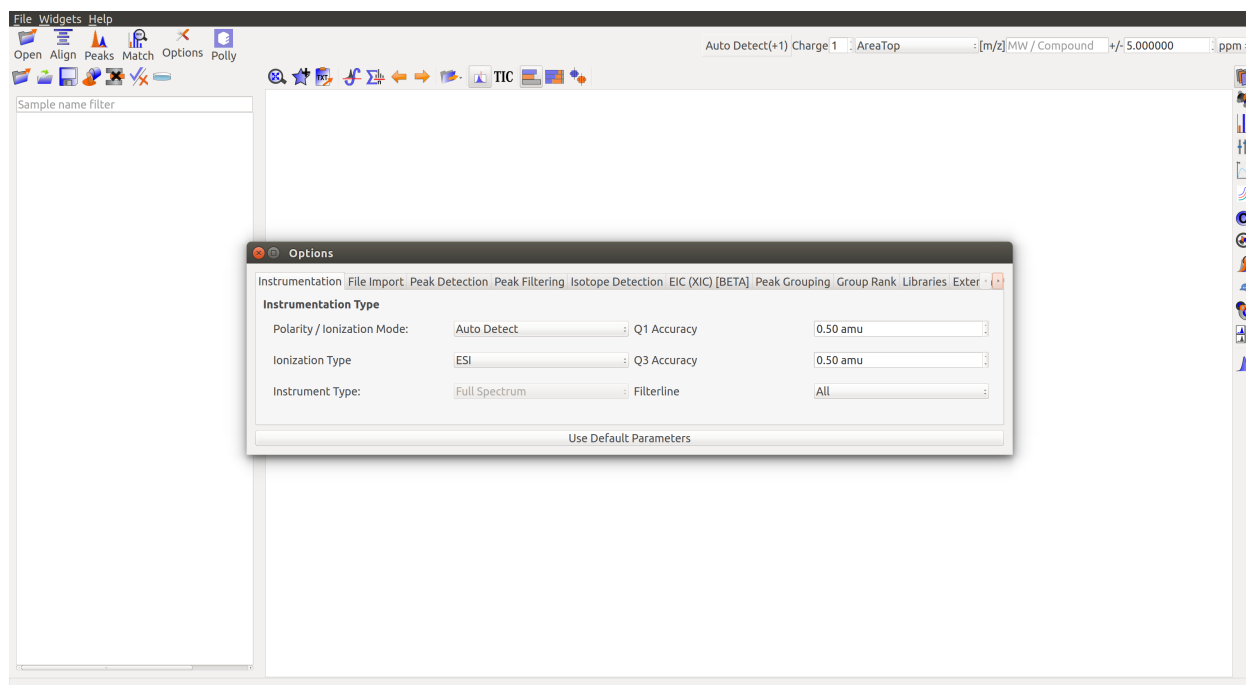
- *Polarity/Ionization mode*: Polarity information is required for m/z calculation. Users can set the polarity of the metabolites in their experiment from the drop-down list or set it to Auto-detect.

- *Ionization type*: Ionization methods can affect m/z calculation. Drop-down provides a list of the most popular ionization types.

- *Q1 accuracy*: This is the mass resolution in amu of the first quadrapole.

- *Q3 accuracy*: This is the mass resolution in amu of the third quadrapole.

- *Filterline*: The drop-down lists different mass ranges and allows the user to process the data in these ranges separately with different parameters. Primarily used for polarity-switching experiments.

File Import
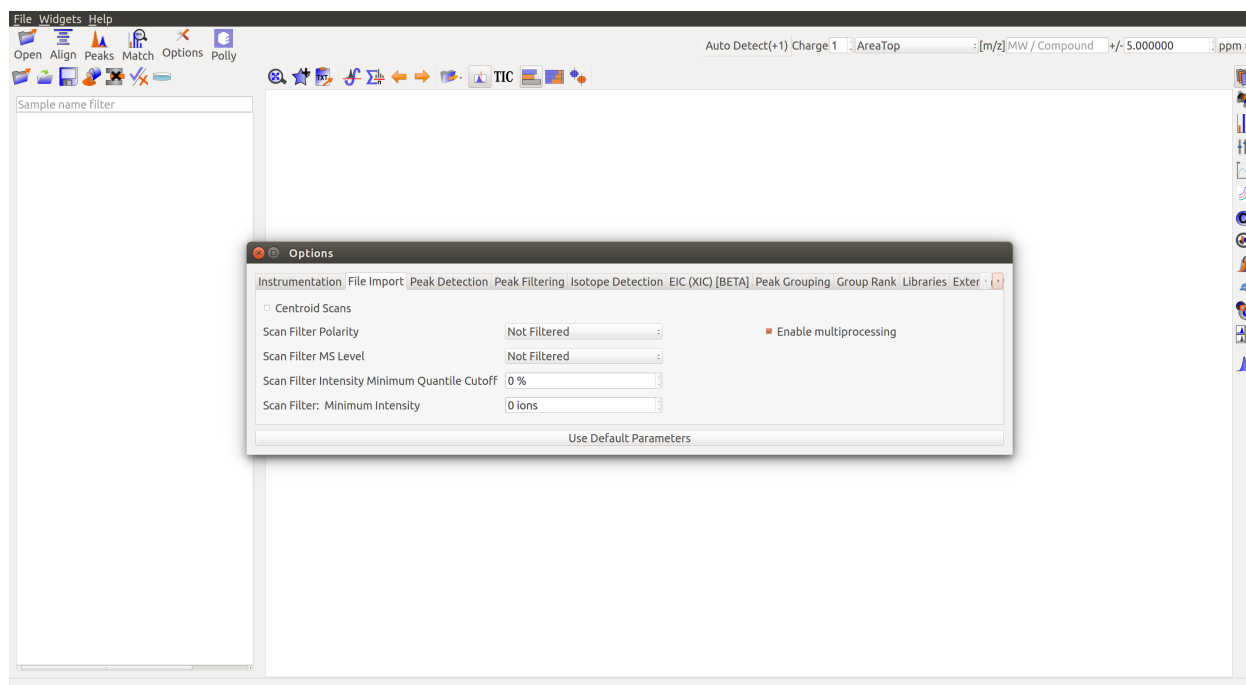
- *Centroid Scans*: Centroid acquisition is an acquisition method where only centroid m/z and intensity are stored. Centroid m/z is calculated based on the average m/z value weighted by the intensity and m/z values are assigned based on a calibration file. Users may leave the box unchecked if they have the centroid data or check the box if centroiding has to be done in El-MAVEN.

- *Scan Filter Polarity*: Users may choose to import scans based on the polarity of ions in the scan. Especially helpful in polarity-switching experiments.

- *Scan Filter MS Level*: Users may choose to import only MS1 or MS2 scans. This feature can be used with MS/MS data.

- *Scan Filter Minimum Intensity*: Sets a minimum threshold for reading in intensity values.

- *Scan Filter Intensity Minimum Quantile Cutoff*: Scans with x% of their intensity values below the threshold will be filtered out during import.

- *Enable Multiprocessing*: In order to reduce the sample load time, El-MAVEN uses multiprocessing. This behavior can be changed by the user.

Peak Detection
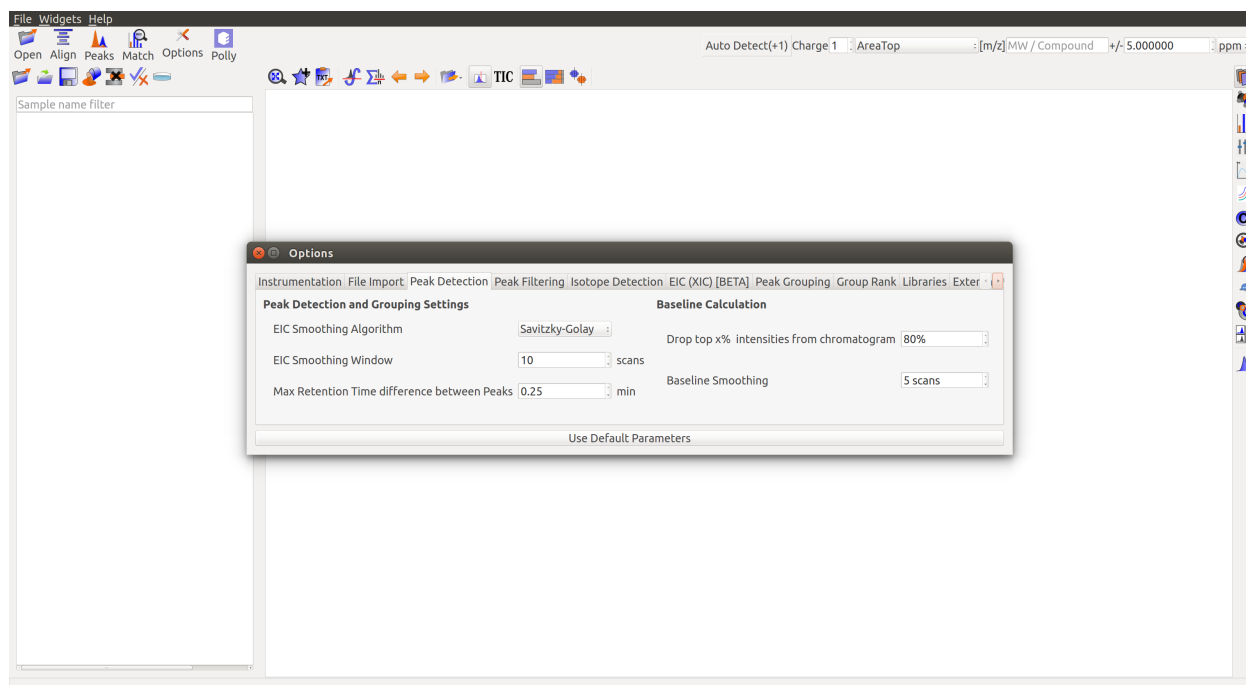
Peak Grouping and Grouping Settings

- *EIC Smoothing Algorithm*: Smoothing of data points helps in increasing the signal/noise ratio. There are three algorithms provided for EIC smoothing:

  1. *Savitzky-Golay*: It preserves the original shape and features of the signal better than most other filters (Learn more).

  2. *Gaussian*: It reduces noise by averaging over the neighborhood with the central pixel having higher weight but successfully preserves sharp edges. (Learn more).

  3. *Moving Average*: It takes the simple average of all points over time. Signal behavior is not natural. Least preferred method for smoothing (Learn more).

- *EIC Smoothing Window*: Number of scans used for fitting in the smoothing algorithm can be adjusted here.

- *Maximum Retention Time Difference Between Peaks*: Set a limit to retention time (RT) difference between peaks in a group. Increase the value if alignment fails to center peaks satisfactorily.

Baseline Calculation

- *Drop top x% intensities from chromatogram*: Set the baseline for every peak. Baseline is obtained once x% of the highest intensities in a peak are removed from consideration. Baseline should be set high when there is more noise in the data.

- *Baseline Smoothing*: Number of scans used for fitting in the smoothing algorithm can be adjusted here.

Peak Filtering
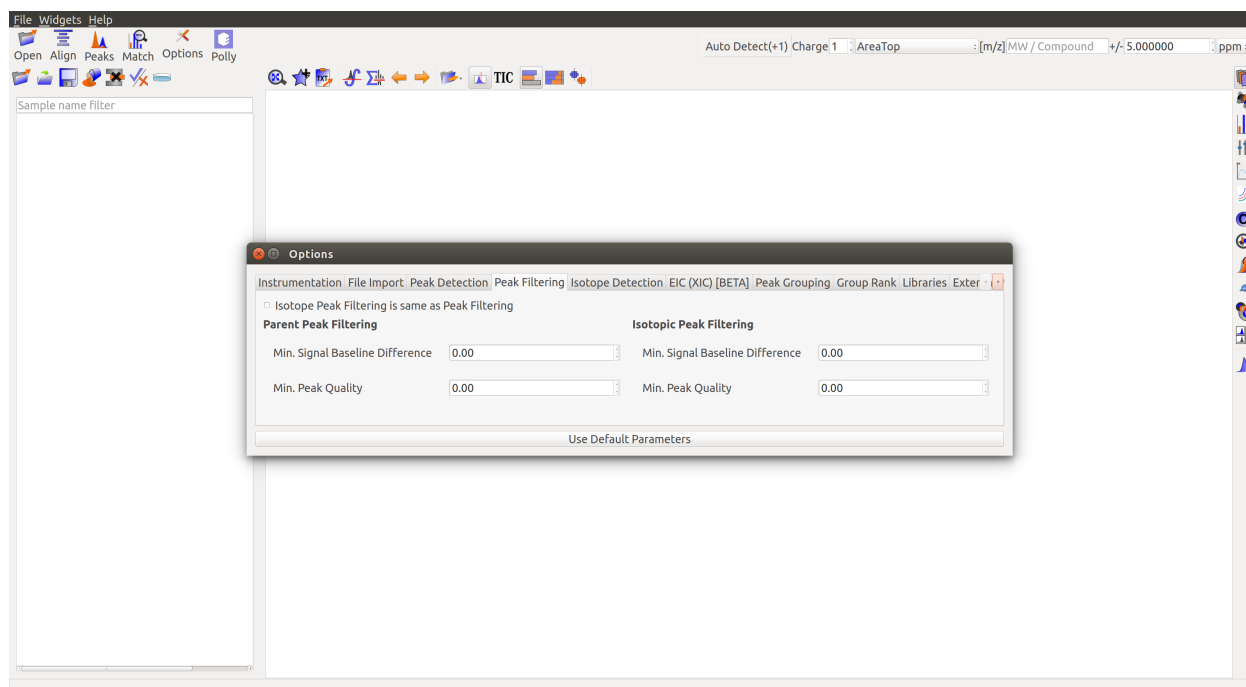
- *Minimum Signal Baseline Difference*: Set the minimum difference between intensity and baseline to detect any signal as a valid peak.

## Isotope Detection
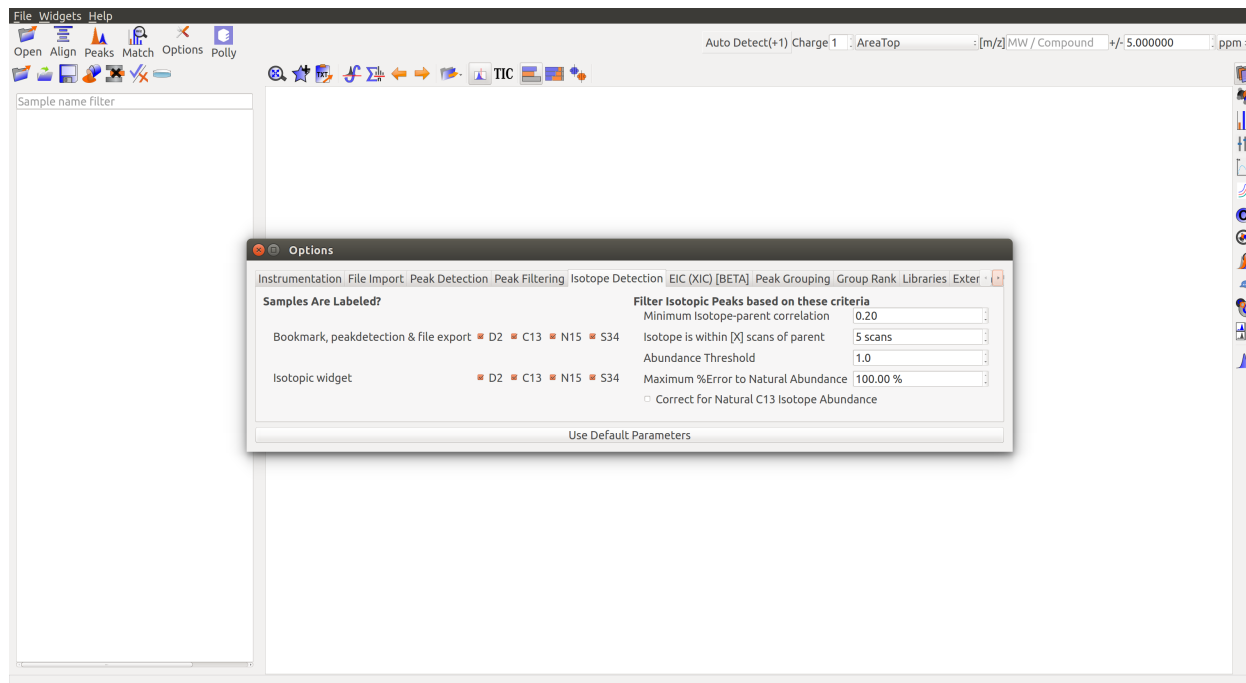
Are Samples Labeled?

- *Bookmarks, peak detection, file export*: Select the labeled atoms that should be used in bookmarking, peak detection and export. D2: Deuterium, C13: Labeled carbon, N15: Labeled nitrogen, S34: Labeled sulphur.

- *Isotopic widget*: Select the labeled atoms that should be displayed in the isotopic widget. D2: Deuterium, C13: Labeled carbon, N15: Labeled nitrogen, S34: Labeled sulphur.

- *Number of M+n isotopes*: Set the maximum number of labeled atoms per ion in the experiment.

- *Abundance Threshold*: Set the minimum threshold for isotopic abundance. Isotopic abundance is the ratio of intensity of isotopic peak over the parent peak.

Filter Isotopic Peaks based on these criteria

- *Minimum Isotope-Parent Correlation*: Set the minimum threshold for isotope-parent peak correlation. This correlation is a measure of how often they appear together.

- *Isotope is within [X] scans of parent*: Set the maximum scan difference between isotopic and parent peaks. This is a measure of how closely they appear together on the retention time scale.

- *Maximum % Error to Natural Abundance*: Set the maximum natural abundance error expected. Natural abundance of an isotope is the expected ratio of amount of isotope over the amount of parent molecule in nature. Error is the difference between observed and natural abundance as a fraction of natural abundance.

- *Correct for Natural C13 Isotope Abundance*: Check the box to correct for natural C13 abundance.

EIC (XIC) [BETA]
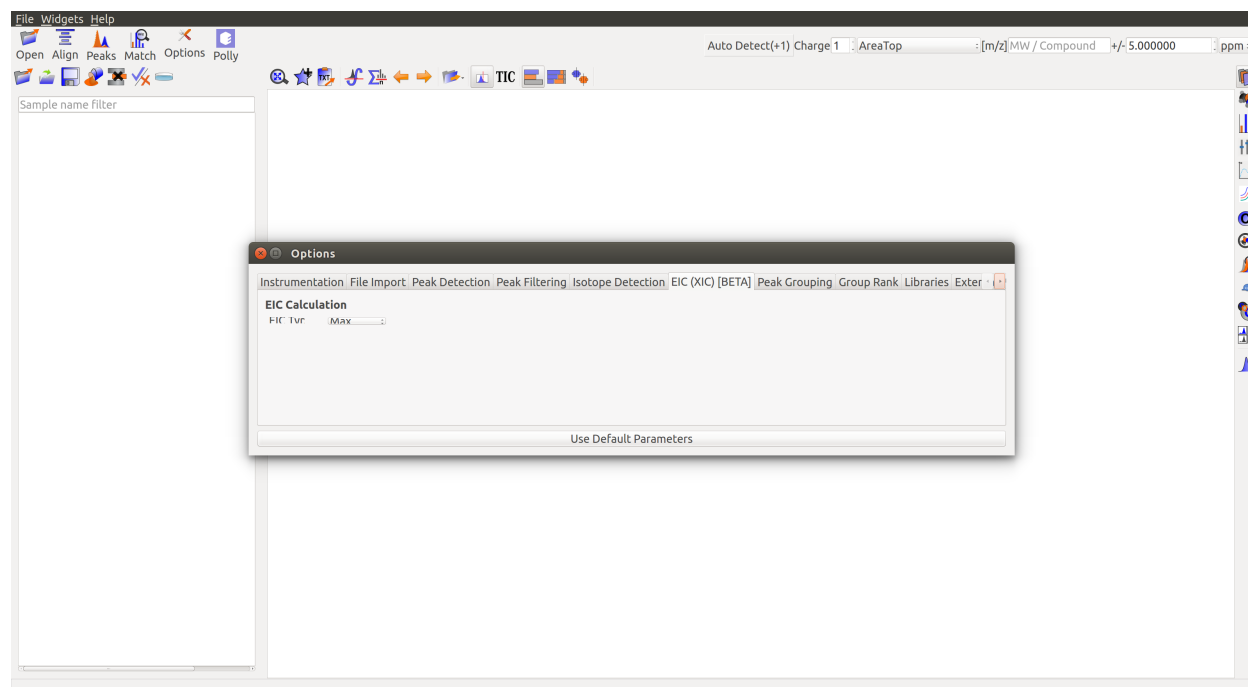
- *EIC Type*: Select a method to merge EICs over m/z. There are two options:

    1. *MAX*: Merged EIC is created by taking the maximum intensity across the m/z window at a particular scan.

    2. *SUM*: Merged EIC is created by taking the sum average of intensities across the m/z window at a particular scan.

Peak Grouping
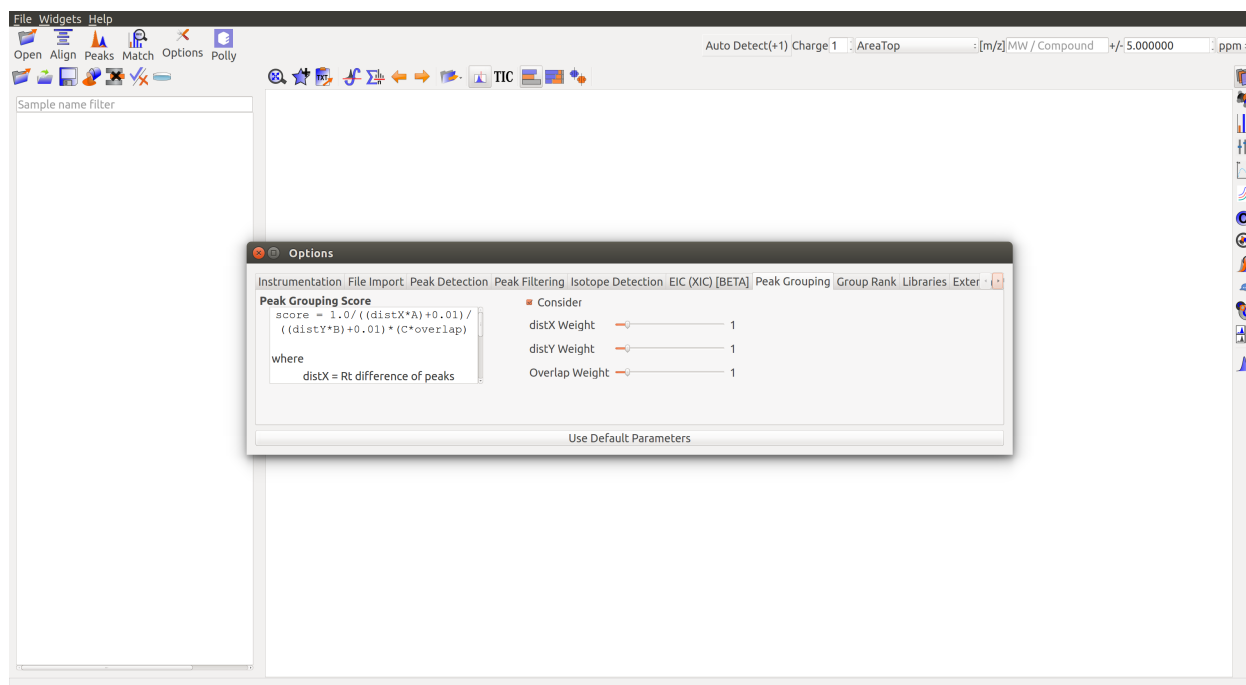
- *Peak Grouping Score*: Peaks are assigned a grouping score to determine whether they should be grouped together. There are two formulas for grouping score calculation:

  1. score = 1.0/((distX * A) + 0.01)/((distY * B) + 0.01) * (C * overlap)

  2. score = 1.0/((distX * A) + 0.01)/((distY * B) + 0.01)

The score depends on the following 3 parameters and their weights:

- *RT difference or DistX*: Difference in retention time between the peaks under comparison. Closer peaks are assigned a higher score.

- *Intensity difference or DistY*: Difference in intensity between peaks under comparison. Smaller difference accounts for a higher score.

- *Overlap*: Fraction of retention time overlap between the peaks under comparison. Greater overlap accounts for a higher score.

  – Uncheck *Consider Overlap* to calculate grouping score without overlap.

  – Sliders are provided to adjust the weights attached to each of the three parameters.

Group Rank
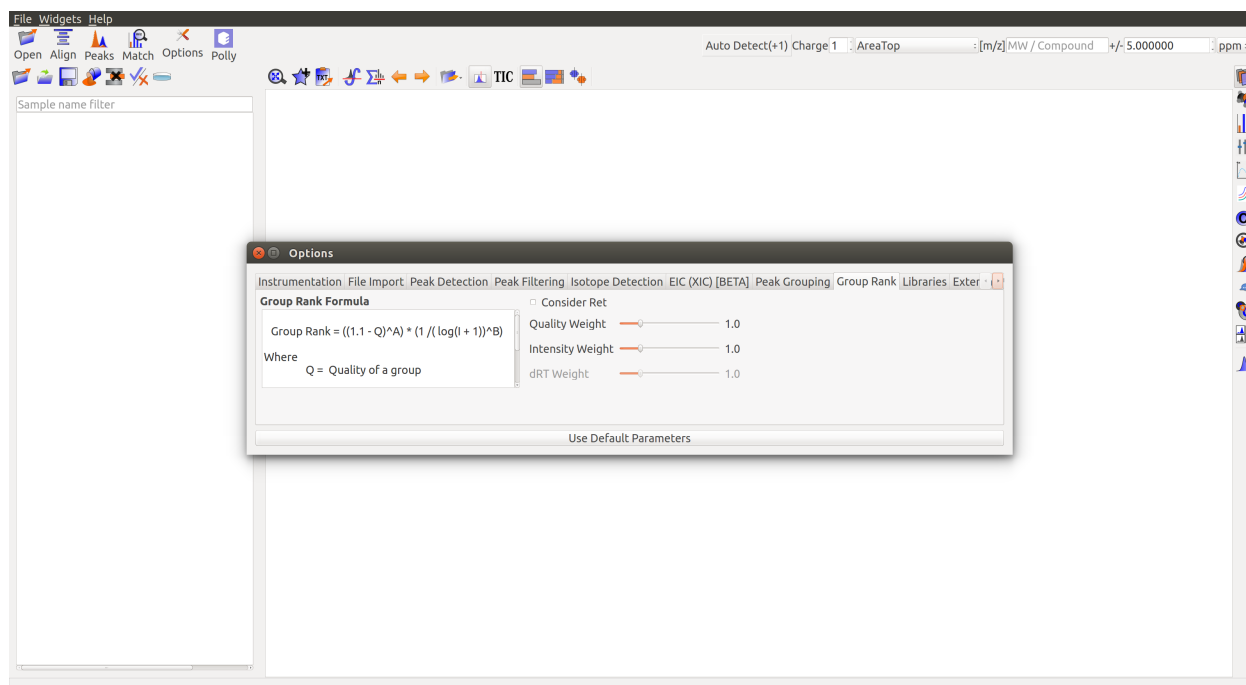
- *Group Rank Formula*: Group rank is one of the parameters for group filtering. There are two formulas below for group rank calculation:

  1. Group Rank = ((1.1 - Q) ^ A) * (1/(log(I + 1)) ^ B)

  2. Group Rank = ((1.1 - Q) ^ A) * (1/(log(I + 1)) ^ B) * (dRT) ^ (2 * C)

The score depends on the following 3 parameters and their respective weights A, B and C:

- *Q or Group Quality*: Maximum peak quality of a group. Peaks are assigned a quality score by a machine learning algorithm in El-MAVEN. Better quality leads to a higher rank.

- *I or Group Intensity*: Maximum intensity of a group. Better intensity leads to a higher rank.

- *dRT or RT difference*: Difference between expected retention time and group mean RT.

  – *Consider Retention Time*: Check the box to use formula (b) for group rank calculation. Formula (a) is used by default.

  – *Quality Weight*: Adjust slider to set weight for group quality in group rank calculation.

  – *Intensity Weight*: Adjust slider to set weight for group intensity in group rank calculation.

  – *dRT Weight*: Adjust slider to set weight for retention time difference in group rank calculation. The slider is disabled if Consider Retention Time is unchecked.

Sample Upload

Load Sample Files

Load sample files into El-MAVEN and click on *Show Samples Widget* on the widget bar to show/hide the project space. Blanks will not show up in the sample list if the file names start with 'blan' or 'blank'.

Load sample files into El-MAVEN and click on *Show Samples Widget* on the widget bar to show/hide the project space. Blanks will not show up in the sample list if the file names start with 'blan' or 'blank'.

There are three columns in the project space:

- *Sample*: This column has the sample name and the random color assigned to the sample. Double-click the sample name to change the color.

- *Set*: The set column holds the cohort name for every sample. For example: subjects and controls.

- *Scaling*: This column holds the normalization constant for every sample. For example, all intensities in a sample will be halved if the constant is two. This is done to normalize data if sample volumes are different.

Sample Space Menu

- *Load Project*: Sample files can be loaded here.

- *Load Meta*: Users may upload a meta file with sample and set names in a comma separated file (.csv) or double-click to enter text. Meta file template is shown below:

	A	B
1	**Sample**	**Set**
2	091215_5M	Test
3	091215_5i	Control
4	091215_45M	Test
5	091215_45i	Control
6	091215_3M	Test
7	091215_1M	Test
8		

- *Save Project as*: Current state of El-MAVEN can be saved in a .mzroll file for future use. All the settings, EICs and peak tables are stored in the file and may be reloaded at any point in the future.

- *Change Sample Color*: Sample colors can be changed by either clicking on this menu button or double-clicking the sample name. Users can pick a color of their choice to represent their samples.
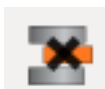
- *Remove Samples*: Apart from deselecting samples, users also have the option to remove samples from the project space. The sample files will not be deleted, only removed from El-MAVEN's project space.

- *Show/Hide Selected Samples*: Samples can be selected/deselected in batches. This is especially helpful when dealing with large datasets as the EIC window gets increasingly noisy with more samples.

- *Mark Sample as Blank*: Users can select sample files and set them as blanks as depicted below. Clicking the button again will reverse the move.
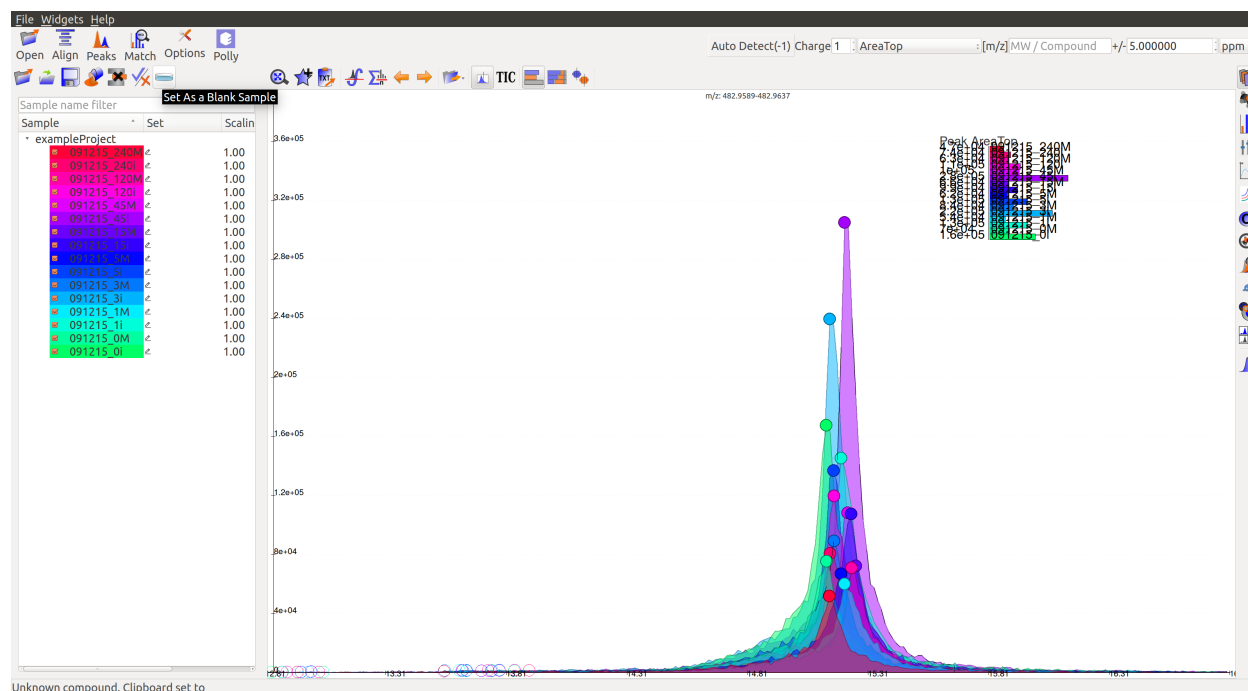
Compound Database

Load Reference File

	A	B	C	D	E
1	compound	formula	id	rt	category
2	cholesteryl sulfate	C27H46O4S	HMDB00653	17.25	
3	Deoxycholic acid	C26H43NO5	C05464	16.79	
4	trans_trans-farnesyl diphosphate	C15H28O7P2	C00448	16.74	
5	Cholic acid	C24H40O5	HMDB00619	16.69	
6	Geranyl-PP	C10H20O7P2	C00341	16.46	
7	Taurodeoxycholic acid	C26H45NO6S	C05463	16.23	
8	lipoate	C8H14O2S2	C00725	15.97	
9	3-hydroxy-3-methylglutaryl-CoA-nega	C27H44N7O20P3S	C00356	15.72	
10	butyryl-CoA	C25H42N7O17P3S	C00630	15.72	

Reference file contains a list of metabolites and their properties that are used for peak detection. This is a comma separated (.csv) or tab separated (.tab) file with compound name, id, formula, mass, expected retention time and category. It is preferable but not necessary to have retention time information in the reference file but either mass or formula is required. In case both mass and formula are provided, formula will be used to calculate the m/z. Click on the *Show Compounds Widget* on the widget toolbar to view the compounds panel. Users may upload a new reference file or use any of the default files loaded on start-up.
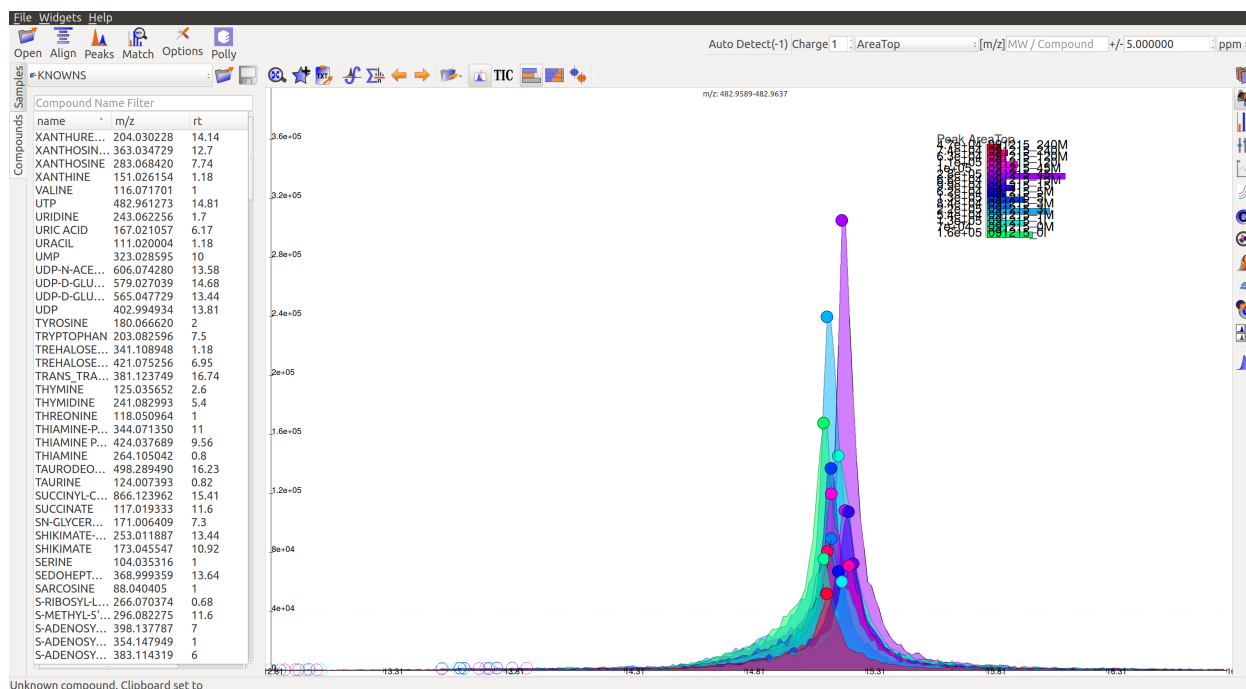
EIC
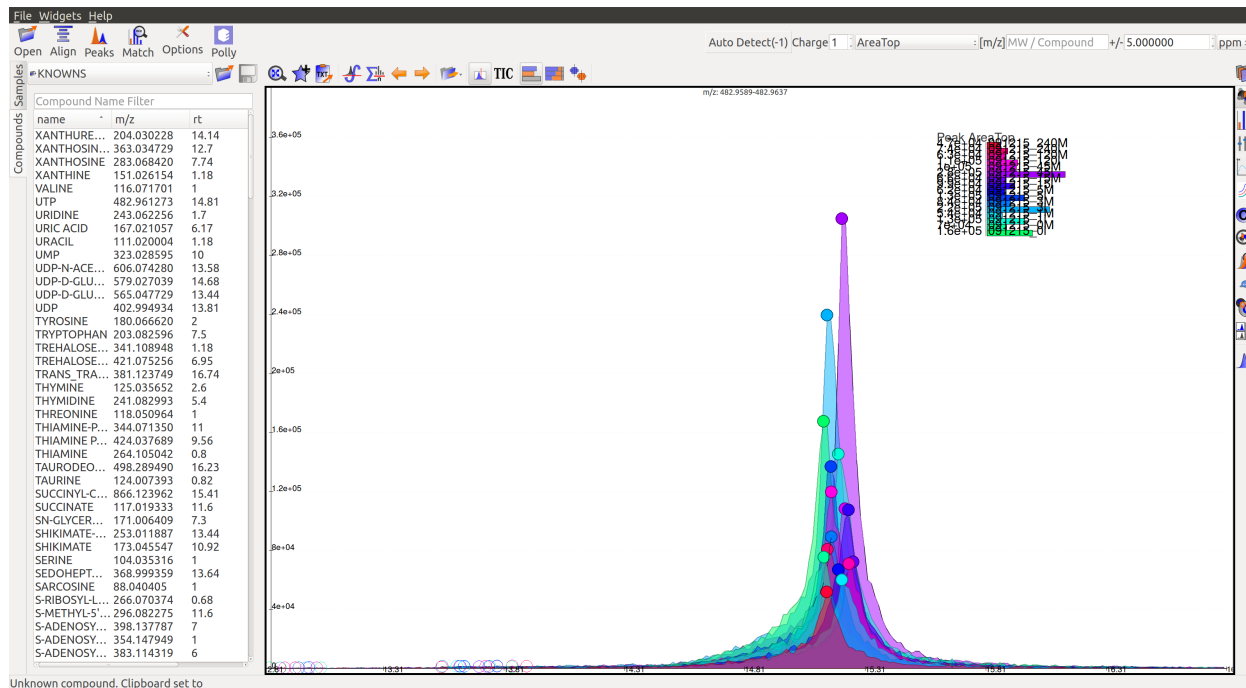
An Extracted Ion Chromatogram is a graph of Intensity vs. retention time for a certain m/z range. EIC window displays the EIC for every group/compound selected or m/z range provided. The group name and/or the m/z range is displayed at the top. Following are the different menu options on top of the EIC window:

-   *Zoom Out*: The EIC is initially zoomed-in to display the region near the expected retention time of a group. This button will zoom out and display the whole retention time range for the selected m/z range. Users may zoom in to a region by right dragging the mouse over it. Left-dragging will zoom out.

-   *Copy Group Information to Clipboard*: On clicking this button, group information is copied to the clipboard with every row representing a different sample.

-   *Bookmark as Good Group*: Users can manually curate a group as 'good' and store it in the bookmark table using this button. (Manual curation of groups has been covered here)
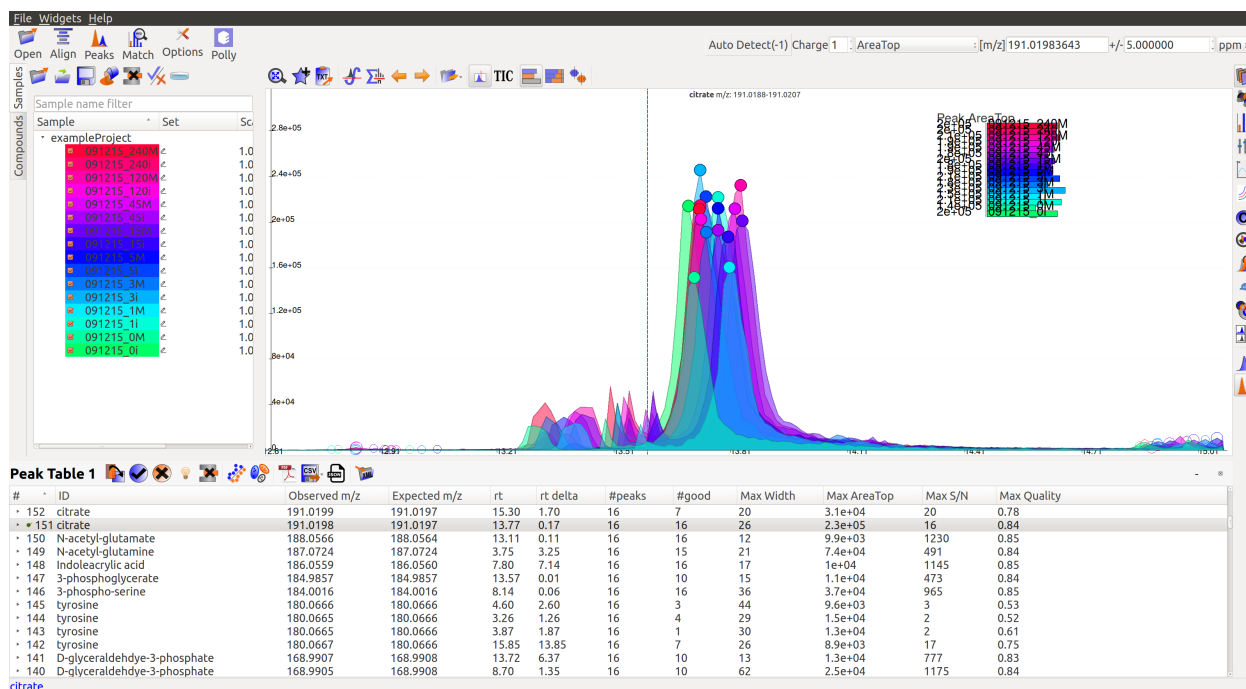
-   *Bookmark as Bad Group*: User can manually curate a group as 'bad' and store it in the bookmark table using this button. (Manual curation of groups has been covered here)
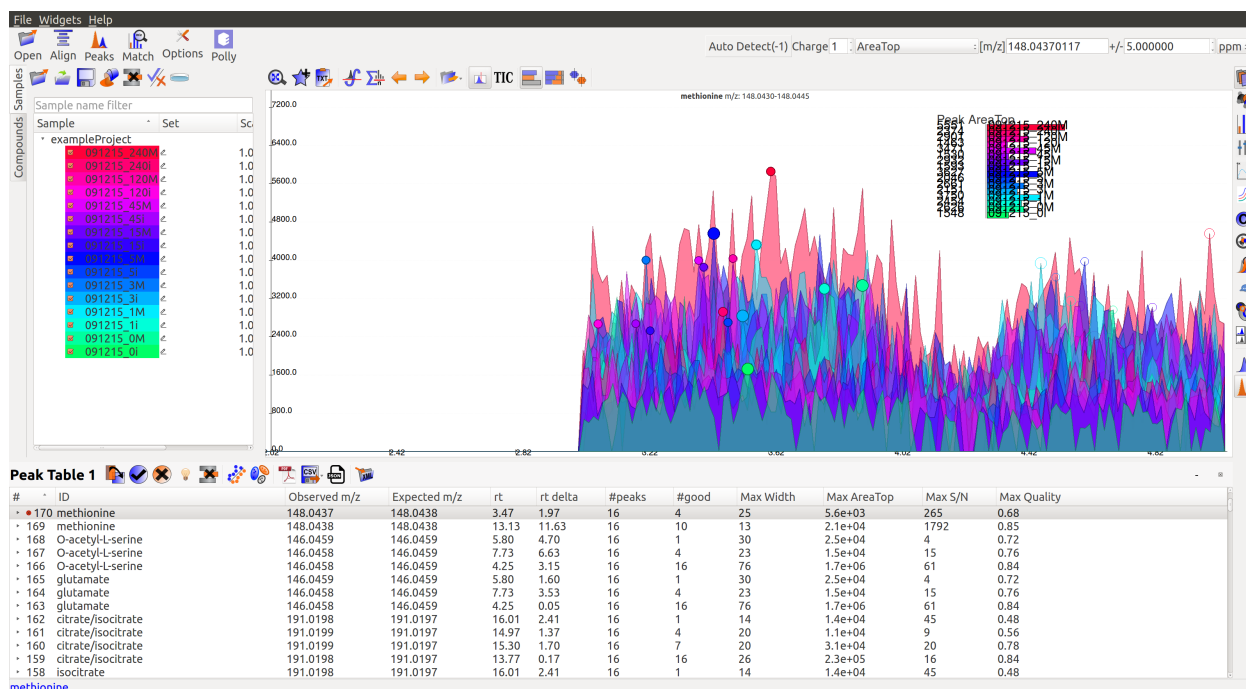
- *History Back*: EIC window display history is recorded. Clicking this button will display the previous state of the window.

- *History Forward*: EIC window display history is recorded. Clicking this button will display the next state of the window, if available.

- *Save EIC Image to PDF File*: Saves the current EIC window display in a PDF file.

- *Copy EIC Image to Clipboard*: Current EIC window display is copied to clipboard.

- *Print EIC*: Current EIC window display can be directly printed out.

- *Auto Zoom*: Auto Zoom is selected by default. It zooms-in and centers the EIC to the expected RT. The expected retention time is depicted as a dashed red line.
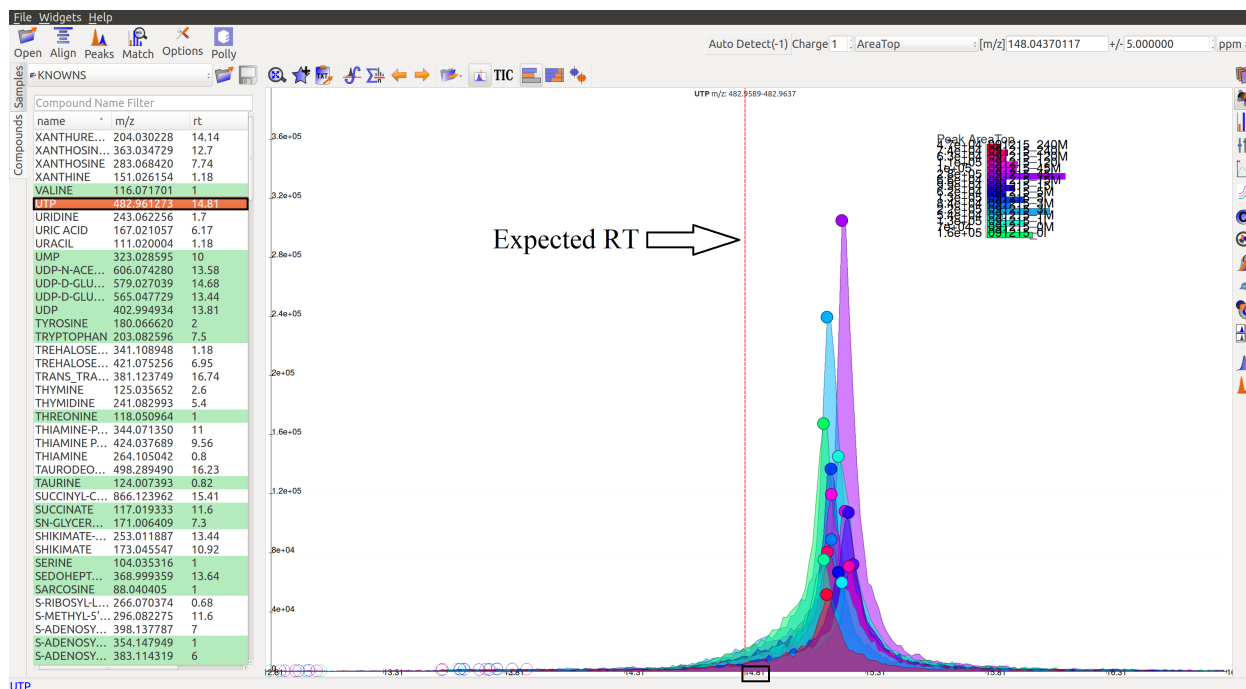
- *Show TICs*: Displays the Total Ion Current. TIC is the sum of all intensities in a scan.

- *Show Bar Plot*: Displays the peak intensity for a group in every sample. Intensity can be calculated by various methods known as quantitation types in El-MAVEN. Users can change the quantitation type from the drop-down list on the top right or choose to display other parameters like retention time and peak quality.
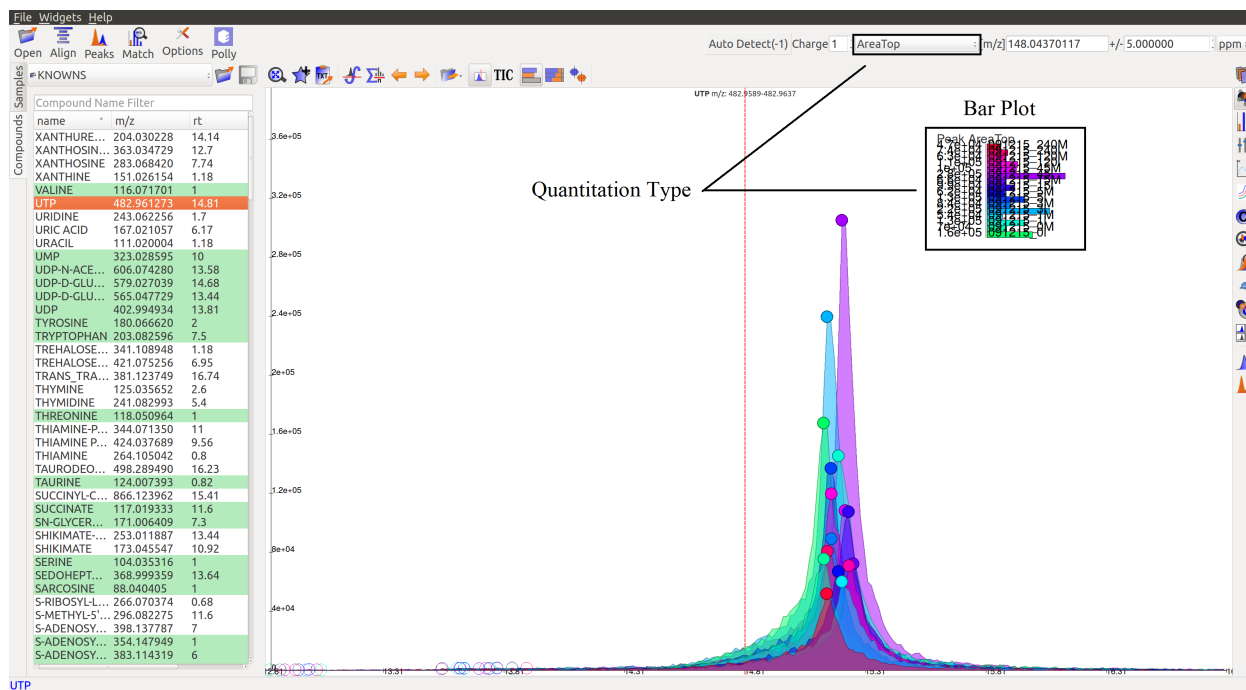
- *Show Isotope Plot*: Displays the isotope plot for a group. Each bar in the plot represents the relative percentage of different isotopic species for the selected group in a sample.
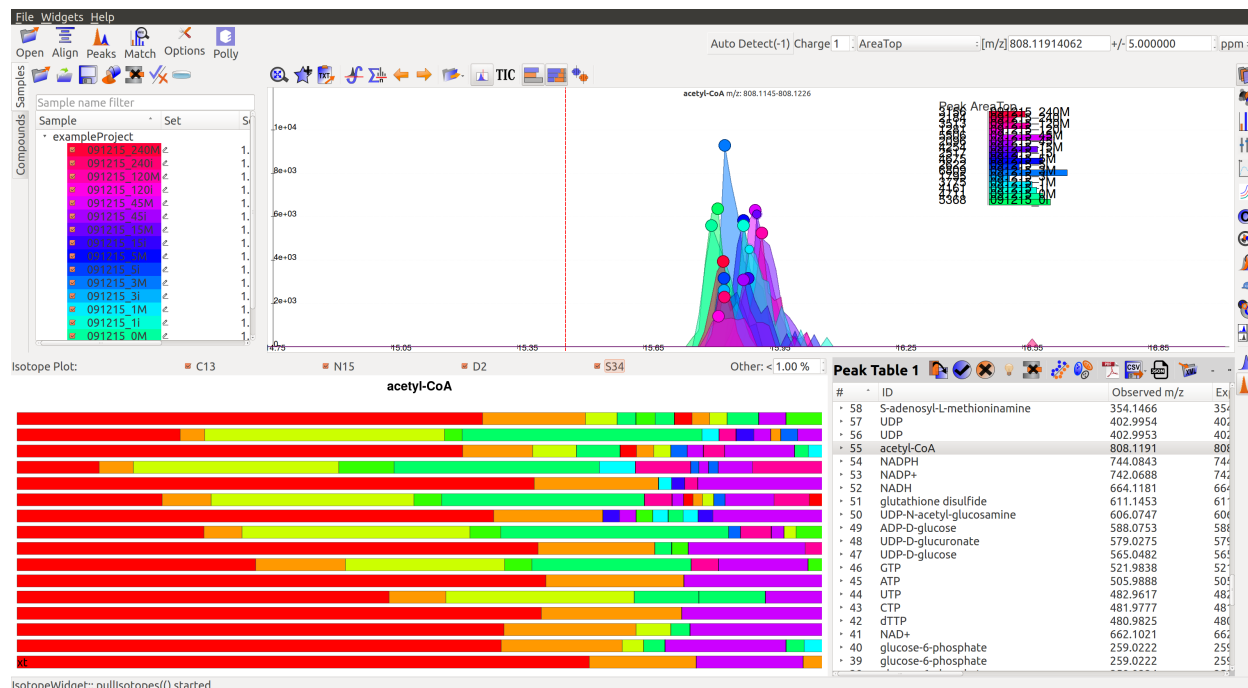
- *Show Box Plot*: Displays the boxplot for a group. The box plot shows the spread of intensities in the group and where each peak lies in relation to the median. Median of the intensities is the vertical line between the boxes.

Apart from the top menu, there are other features in the EIC window. Right-click anywhere in the window and go to Options.
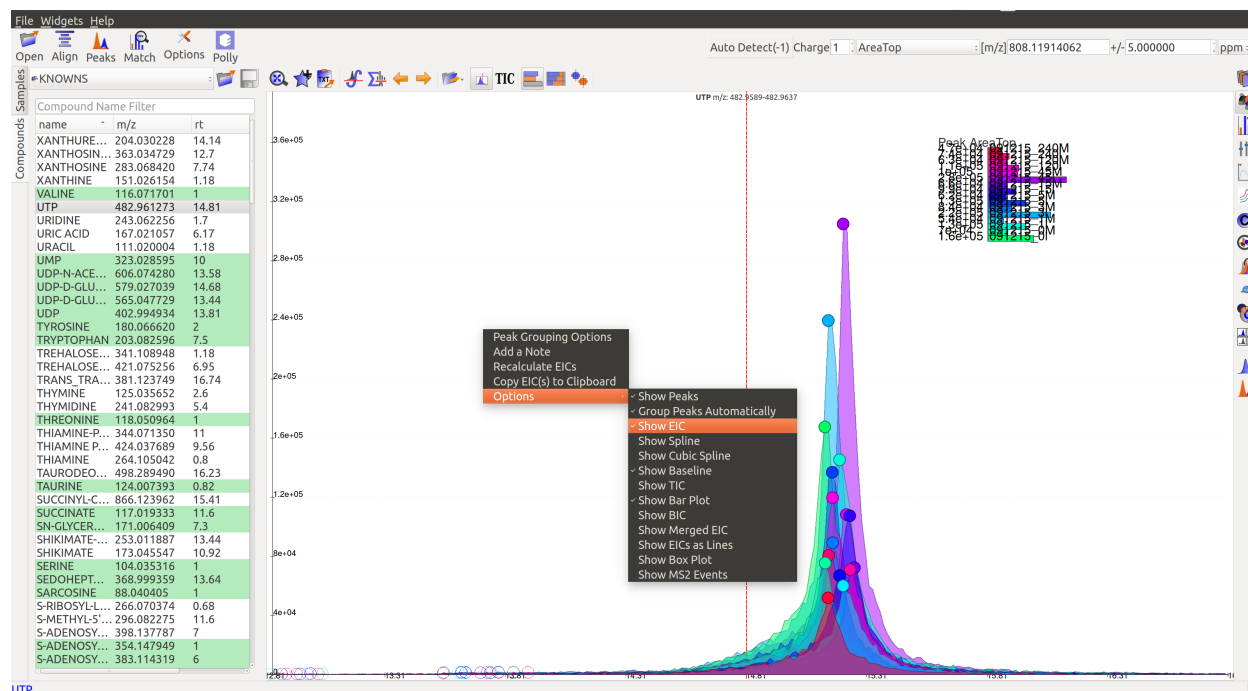
Some of the important options are:

- *Show Peaks*: Peaks are marked by the colored circles that represent the quality score of the peak. Bigger the circle, better the peak quality. This option allows the user to show/hide the peak quality score.

- *Group Peaks Automatically*: Peak grouping happens automatically when grouping parameters are changed. To prevent automatic grouping, user can uncheck this option.

- *Show Baseline*: Hide/Show the baseline for every peak. (Read more about baseline here).

- *Show Merged EIC*: Merged EIC is the sum average of EICs across samples. It smoothens the data and helps in grouping peaks.

- *Show EIC as Lines*: In case of large number of samples, it can get difficult to look at short individual peaks as they are obscured by larger peaks. Showing EIC as lines cleans up the display window and allows the user to look at small peaks.

Mass Spectra

Mass Spectra Widget displays each peak, its mass, and intensity for a scan. As the widget shows all detected masses in a scan, the ppm window for the EIC and consequently grouping can be adjusted accordingly. This feature is especially useful for MS/MS data and isotopic detection.
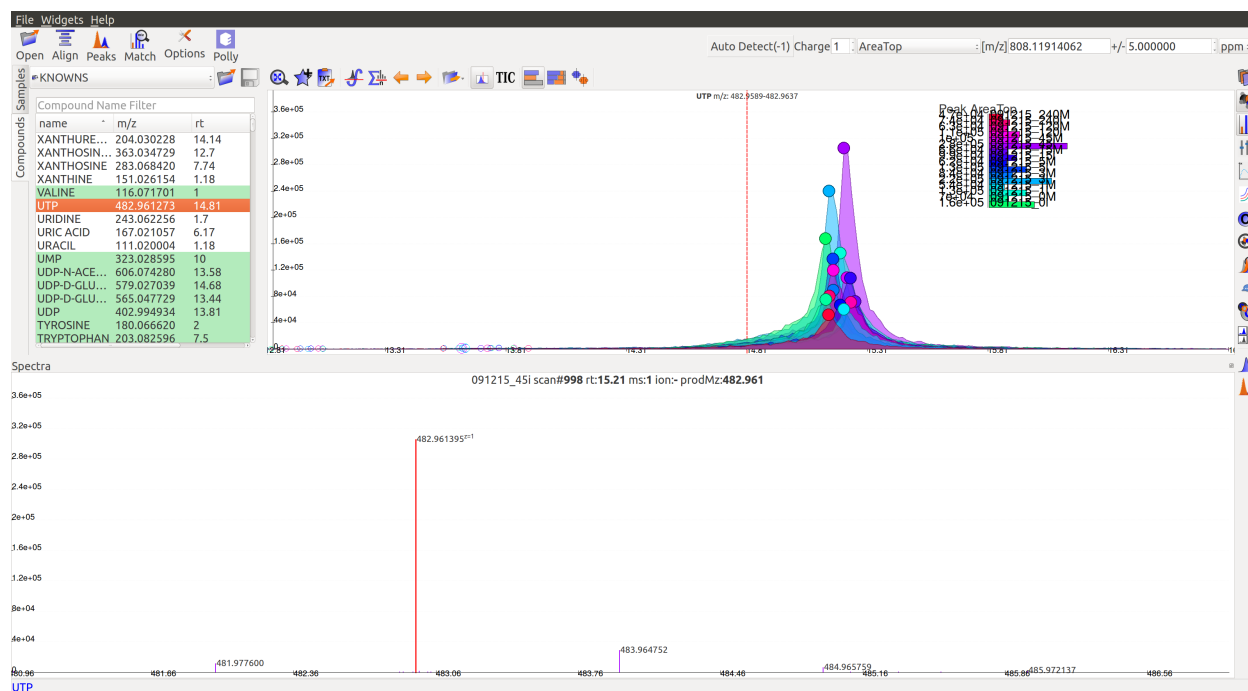
Alignment

Prolonged use of the LC column can lead to a drift in retention time across samples. Alignment shifts the peak RTs in every sample to correct for this drift and brings the peaks closer to median retention time of the group.

Click on the *Align* button [button icon] and adjust the settings.
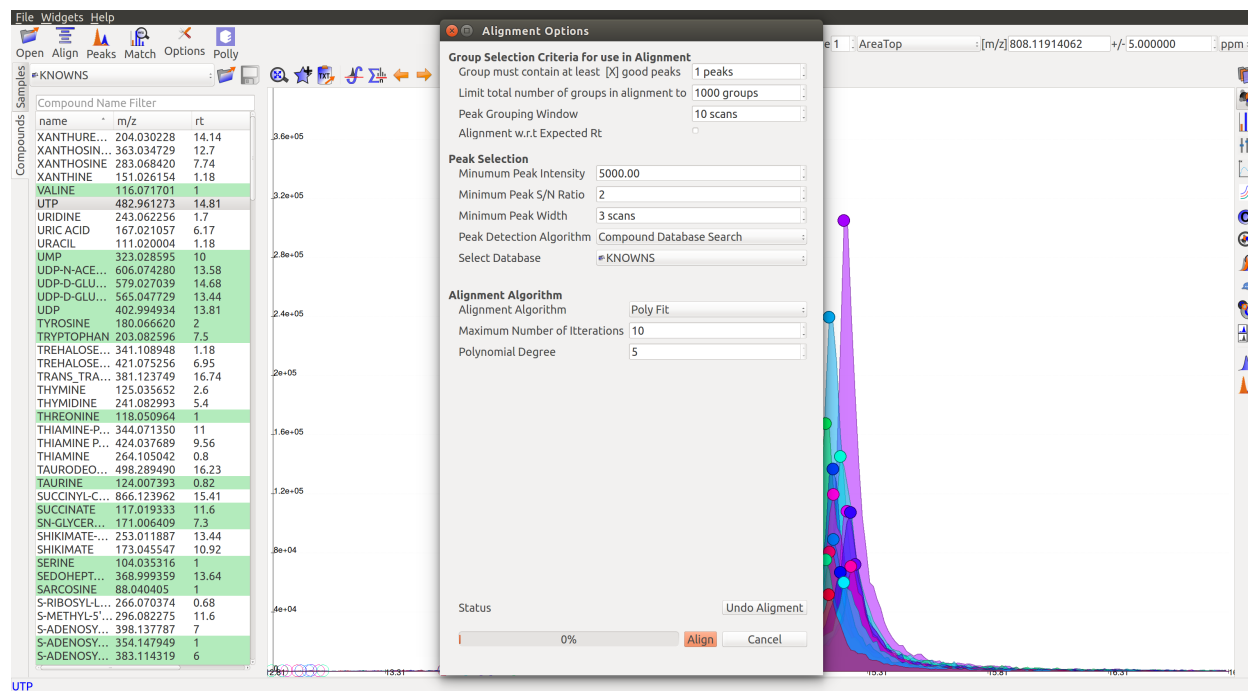
The first panel in Alignment options is for Group Selection criteria. 'Group' here refers to a set of peaks across samples that is annotated as a particular ion.

- *Group must contain at least [X] good peaks*: The value of x is set to filter out groups that do not have at least x good peaks from the alignment process. As there is only one peak per sample for a group, this value should not exceed the number of samples in your project. This option allows the users to discard groups with very few good peaks under the assumption that those could be stray peaks.

- *Limit total number of groups in alignment to*: Users can change the number of groups being used for alignment in case there are too many groups detected after the peak detection process.

- *Peak Grouping Window*: This value controls the number of scans required to get the most accurate peaks. Enter a high number if the reproducibility is low to ensure best results.

The next panel is for *Peak Selection* settings:

- *Minimum Peak Intensity*: The intensity value can be adjusted to only look at high or low intensity peaks in case you have prior information about the concentration of metabolite users are looking for.

- *Minimum peak S/N ratio*: This is the minimum signal to noise ratio of your experiment. Increase the value if you see too much noise in the data.

- *Minimum Peak Width*: This is the least number of scans to be considered to evaluate the width of any peak.

- *Peak Detection Algorithm*: Select the *Compound Database Search* algorithm and then choose an appropriate database from the next drop-down menu.

The *Alignment Algorithm* panel provides the following options:

- *Alignment Algorithm*: There are three alignment algorithms available in El-MAVEN: Obi-Warp, Poly fit and Loess fit. Loess fit has been released as a beta feature for now.

- *Maximum number of Iterations*: This parameter is only required for Poly fit algorithm. Enter the number of times El-MAVEN should fit a model to the data in order to align it.

- *Polynomial Degree*: This is the degree of the non-linear model we are trying to fit. Recommended settings are entered by default.

Click on *Align* at the bottom.

Alignment Visualizations

El-MAVEN provides three visualizations for alignment analysis.

- *Show Alignment Visualization*: Click on in the widget bar to open this visualization. Click on any grouped peak to look at its delta RT vs RT graph as shown.
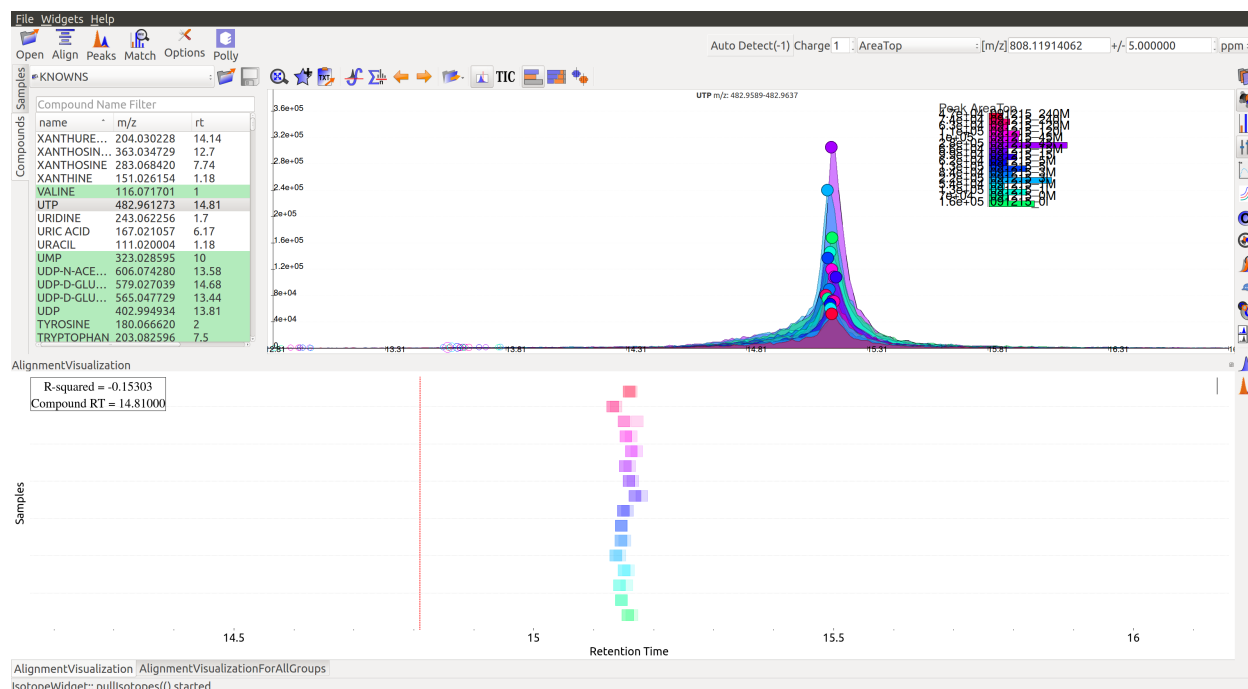
- *Show Alignment Visualization (For All Groups)*: Click on in the widget bar for this visualization.
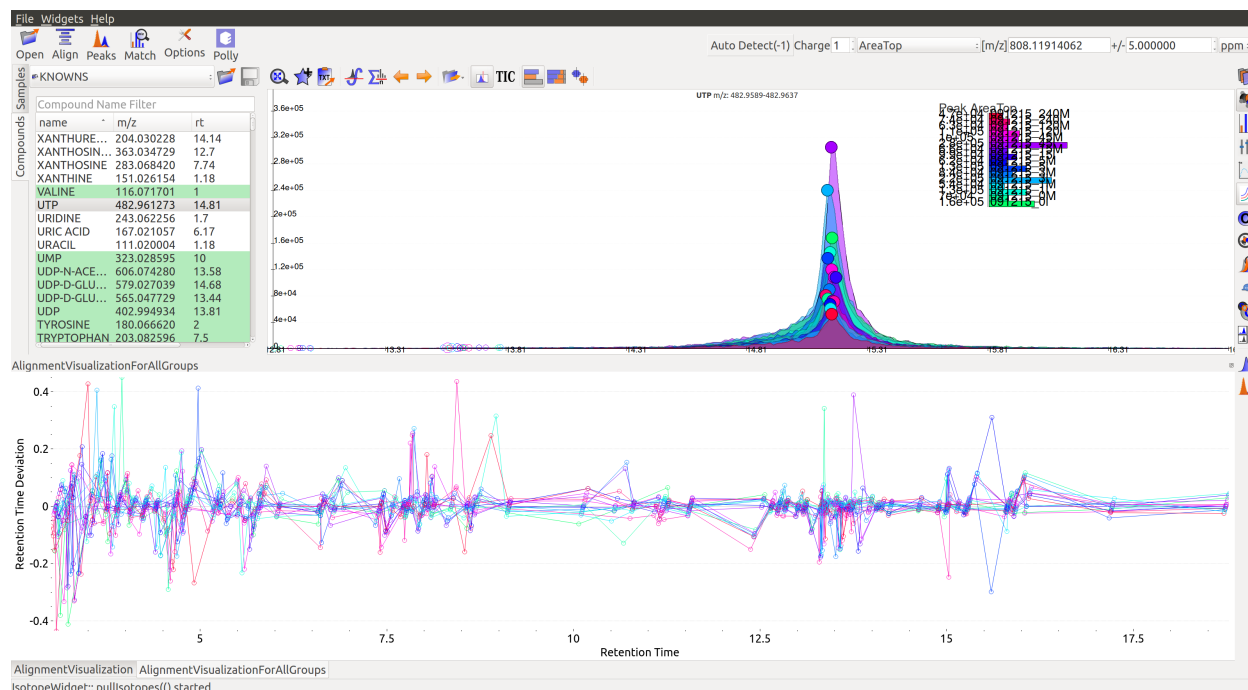
- *Show Alignment Polynomial Fit*: Click on in the widget bar for Poly fit alignment.
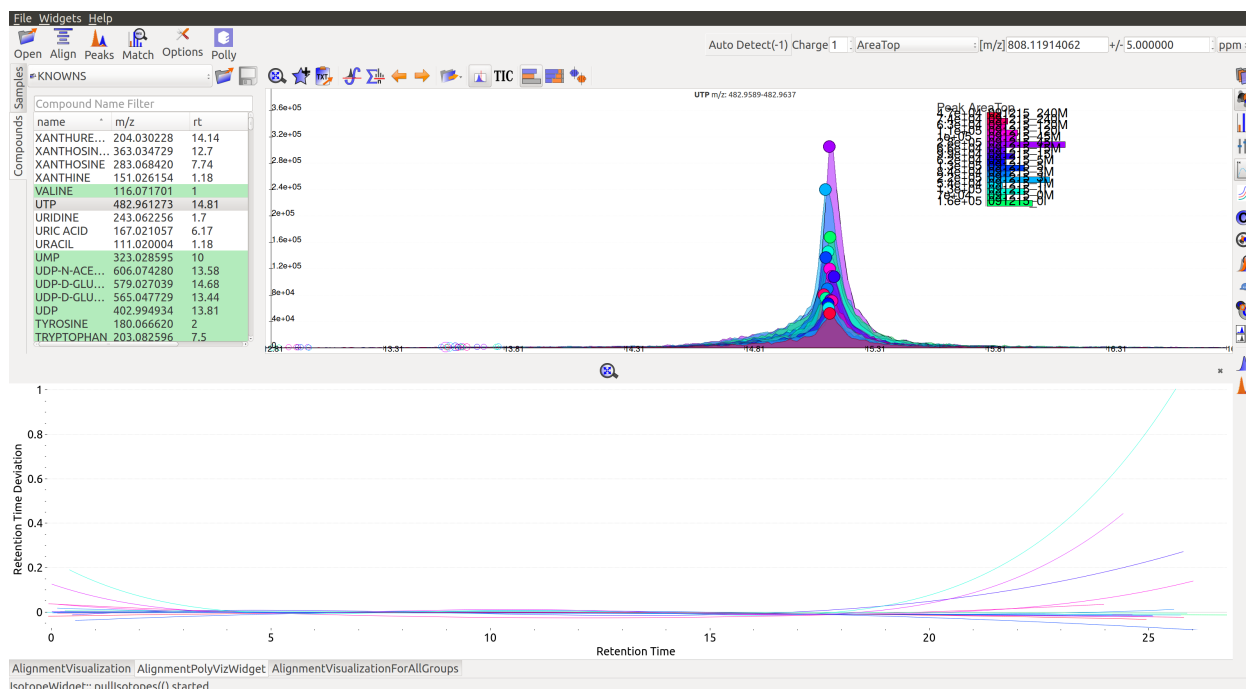
The above graphs give a clear indication of how aligned/misaligned the peaks are. Users may run alignment again with different parameters if required (or with a different algorithm).
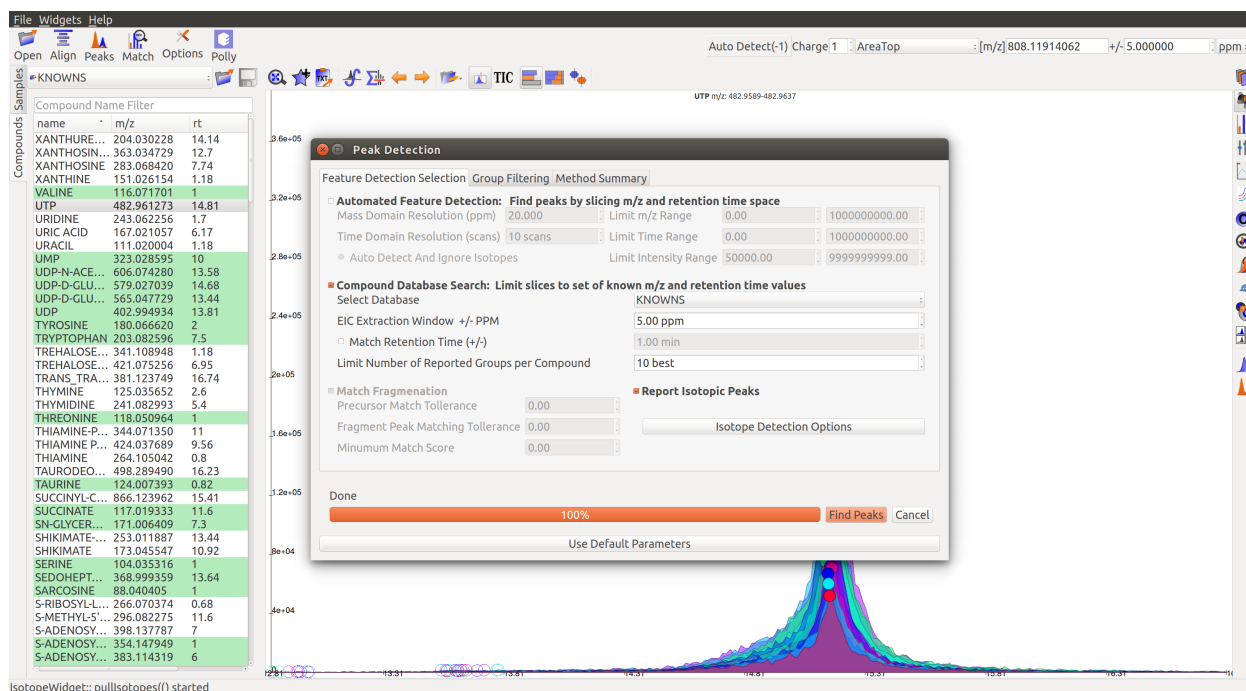
Peak Detection

Peak detection algorithm pulls the EICs, detects peaks and performs grouping and filtering based on parameters controlled by the users. The algorithm groups identical peaks across samples and calculates the quality score by a machine

learning algorithm. Click on the *Peaks* icon on the top to open the settings dialog.

There are 3 tabs for setting Peak Detection parameters:

1. Feature Detection Selection

The Feature Detection Selection panel has the following parameters:

- *Automated Feature Detection*: This is one of the two strategies for finding peaks. Automated search creates thousands of mass slices across the whole m/z and retention time space to find all peaks present in the sample. This strategy is used when looking for new/unknown metabolites in the samples.

    - *Mass Domain Resolution*: This value defines the m/z range of every mass slice in parts per million

    - *Time Domain Resolution*: This value defines the scan range (or retention time range) of every mass slice

    - *Limit Mass Range*: User can limit the automated search to a range of m/z according to their requirements

    - *Limit Time Range*: User can limit the automated search to a retention time range according to their requirements

- *Compound Database Search*: Database search is used to search for compounds listed in the reference file using their m/z information. For better accuracy, retention time information can also be used for the search.

    - *Select Database*: Select a desired reference file for the search from the drop-down list

    - *EIC Extraction Window*: Provide a ppm buffer range to all compound masses. A larger window is useful for processing low resolution data. The window should be smaller for high resolution data to reduce noise.

    - *Match Retention Time*: Enable/disable use of retention time information along with m/z to perform database search. Compounds can have different retention times in every experiment, therefore this option should only be checked if the reference file is specific to the experiment and the sampled used. Enter the time buffer in the accompanying box.

    - *Limit Number of Reported Groups Per Compound*: Multiple groups can be annotated as the same compound, especially when retention time is not taken into consideration for the search.

Users can set the value to only report X best groups according to their rank. The group rank formula will be discussed later in the tutorial.

- *Match Fragmentation*: This panel is activated for MS/MS data.

- *Report Isotopic Peaks*: Check this box to find and report isotopic peaks for labeled data.

To perform peak detection with reference, check the box next to *Compound Database Search* and choose the appropriate database. The *EIC Extraction Window* should be set according to the instrument's resolving power. Select the *Match Retention Time* option if you wish to search for compounds using both the m/z ratio and retention time value given in the database. In case of a generic database, searching by retention time is not recommended.

2. Group Filtering
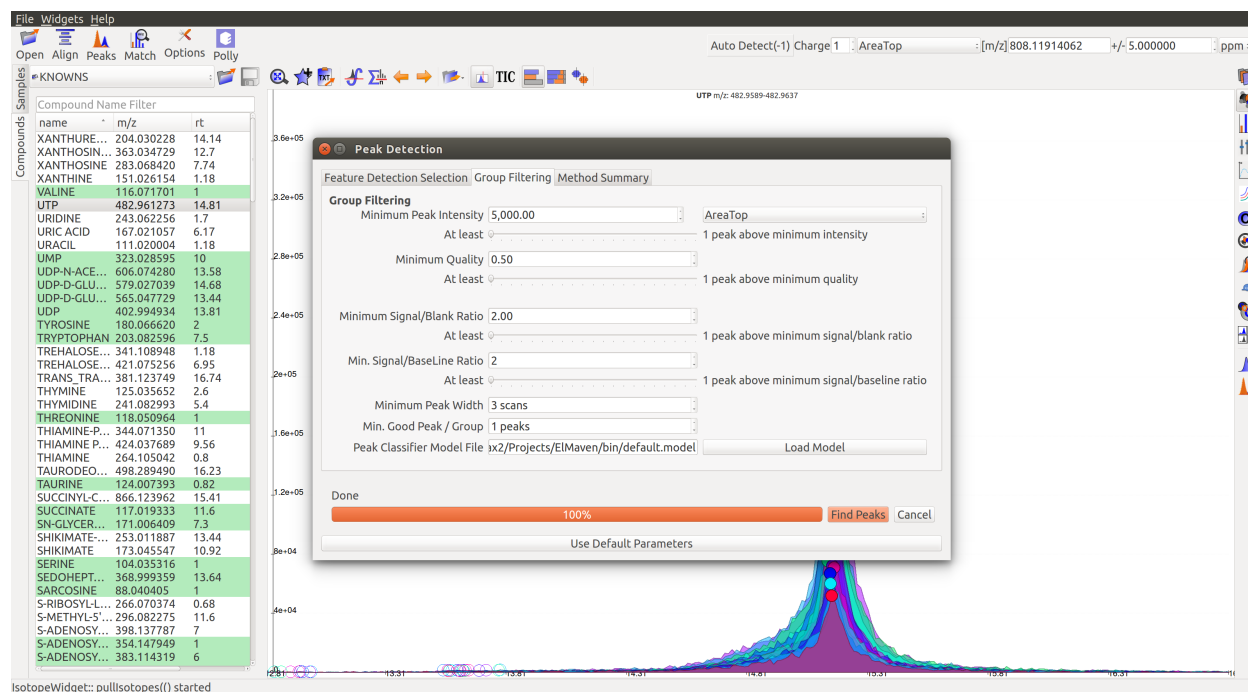
After grouping is done, groups that do not fulfill the criteria shown above are filtered out.

- *Minimum Peak Intensity*: Groups with no peak intensities above this threshold are filtered out. The drop-down list beside the input box defines how intensity is calculated. Different methods of intensity calculation are known as quantitation types. The slider below can be adjusted to change the minimum percentage of peaks per group that must pass the threshold (minimum number of peaks is 1).

- *Minimum Quality*: Quality of peaks is calculated using a machine learning algorithm. Groups with no peak qualities above this threshold are filtered out. The slider below can be adjusted to change the minimum percentage of peaks per group that must pass the threshold (minimum number of peaks is 1).

- *Minimum Signal/Blank Ratio*: Signal/Blank ratio is the ratio of peak intensity over maximum intensity observed in blanks. Groups with no peaks above this threshold are filtered out. The slider can be adjusted to change the minimum percentage of peaks per group that must pass the threshold (minimum number of peaks is 1). This helps in filtering out peaks that are also present in blanks.

- *Minimum Signal/Baseline Ratio*: Signal/Baseline ratio is the ratio of peak intensity over baseline value for that peak. Baseline calculation is used to filter out noise in the signal and will be discussed later in the tutorial. The slider can be adjusted to change the minimum percentage of peaks per group that must pass the threshold (minimum number of peaks is 1).

- *Minimum Peak Width*: Peak width is equal to the number of scans that a peak is spread over. Groups with no peak widths above this threshold are filtered out. Spurious signals can be filtered out using this option.

- *Peak Classifier Model File*: This is the default model that is used by the machine learning algorithm for classifying peaks according to their quality.

Change the settings according to the data and click on *Find Peaks* to run peak detection. For beginners, performing peak detection with default values at first is recommended. Users may then adjust the settings depending on their results.

3. Method Summary
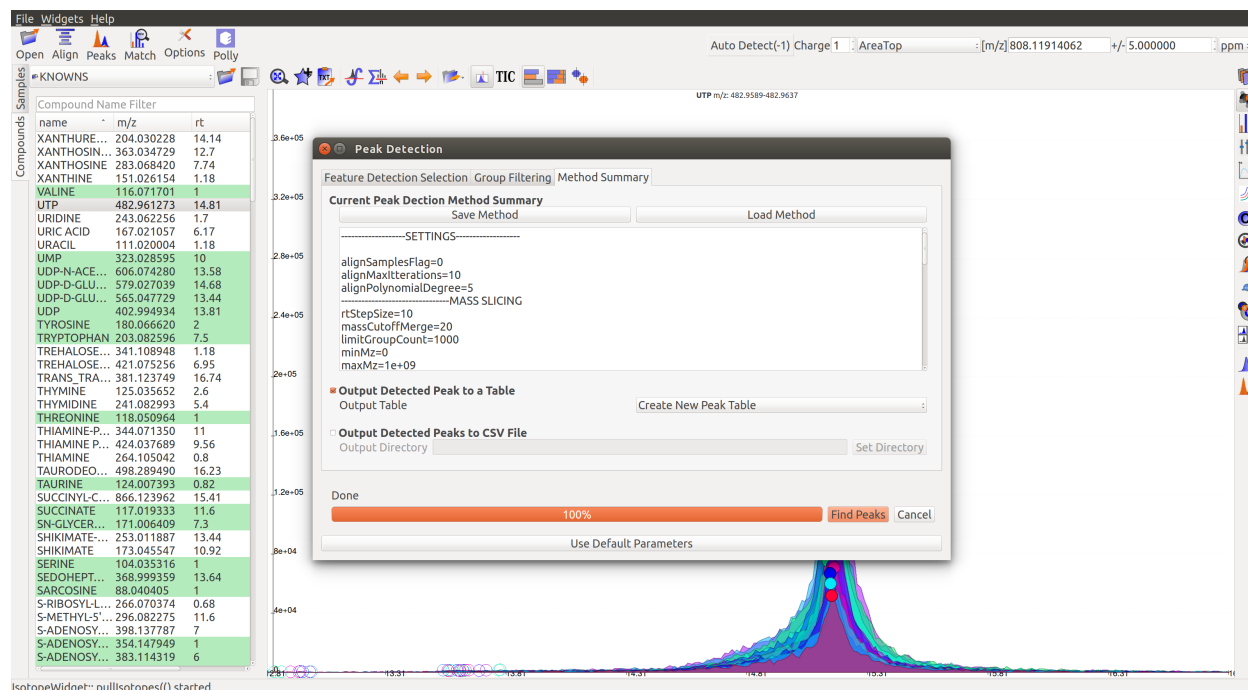
Peak Table
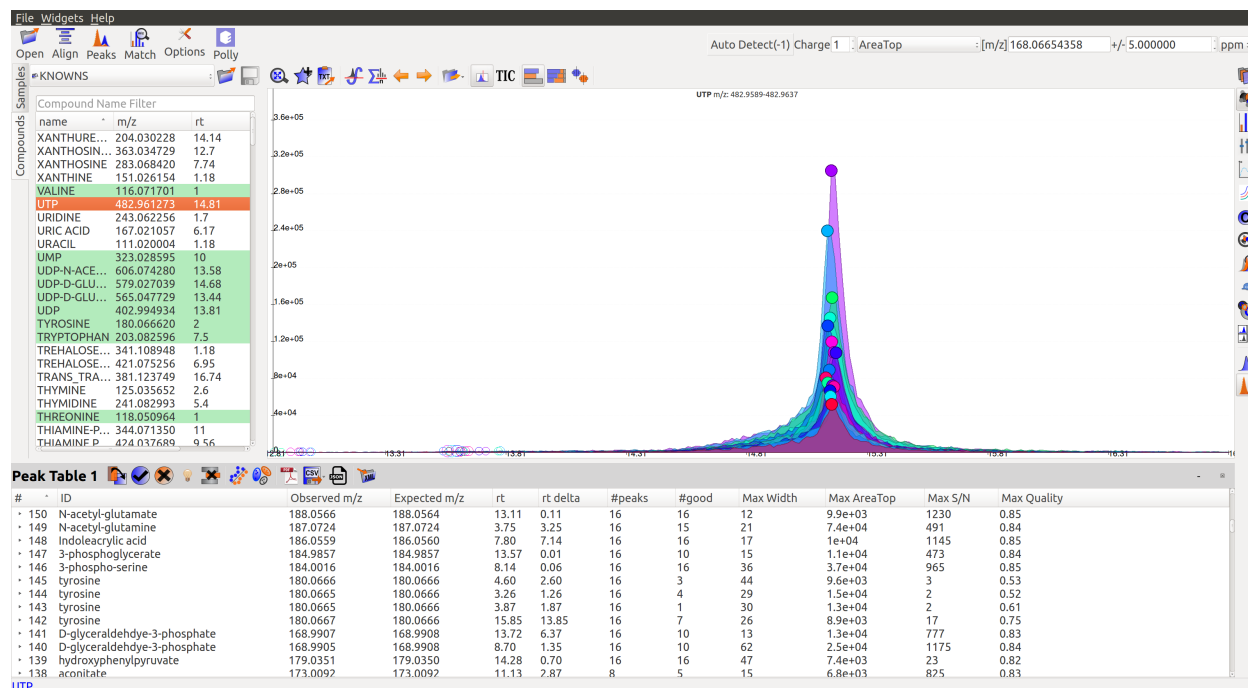
Groups information obtained after Peak Detection is stored and displayed in the form of a Peak Table with a row representing a group and its corresponding features in columns. Users can show/hide the peak table by clicking on

the widget bar.

Peak Table Features

Following are the different features/columns in a peak table:

- *#*: is the serial number for a group

- *ID*: Group ID is assigned according to the search mode used during peak detection. In case of Automated search, groups are named by their m/z and retention time values separated by '@' sign. For example, ID for a group with 230.2 m/z and 1.89 RT will be given as '230.2@1.89'. In case of Database search, groups are annotated as a compound from the reference file. For example, 'malate'.

- *Observed m/z*: is the median m/z of the group.

- *Expected m/z*: is the m/z value provided in the reference file for the compound represented by the group. This field is populated only in case of Database search.

- *rt*: is the median retention time of the group.

- *rt delta*: is the difference between expected retention time from the reference file and the observed RT. This field is set to -1 in case of Automated Search.

- *#peaks*: is the number of peaks in the group.

- *#good*: is the number of good peaks in a group. A good peak is defined as one with its quality score above the defined threshold in Peak Detection dialog.

- *Max Width*: is the maximum peak width in a group. Peak width is defined as the number of scans over which a peak is spread.

- *Max AreaTop*: is the maximum peak AreaTop intensity in a group. AreaTop is one of the quantitation types used to represent peak intensity in El-MAVEN. Read more about the different quantitation types here.

- *Max S/N*: is the maximum peak signal/noise ratio in a group.

- *Max Quality*: is the maximum peak quality score in a group.

- *Rank*: is the group rank. The formula and parameters involved have been explained here.

Peak Table Menu Bar
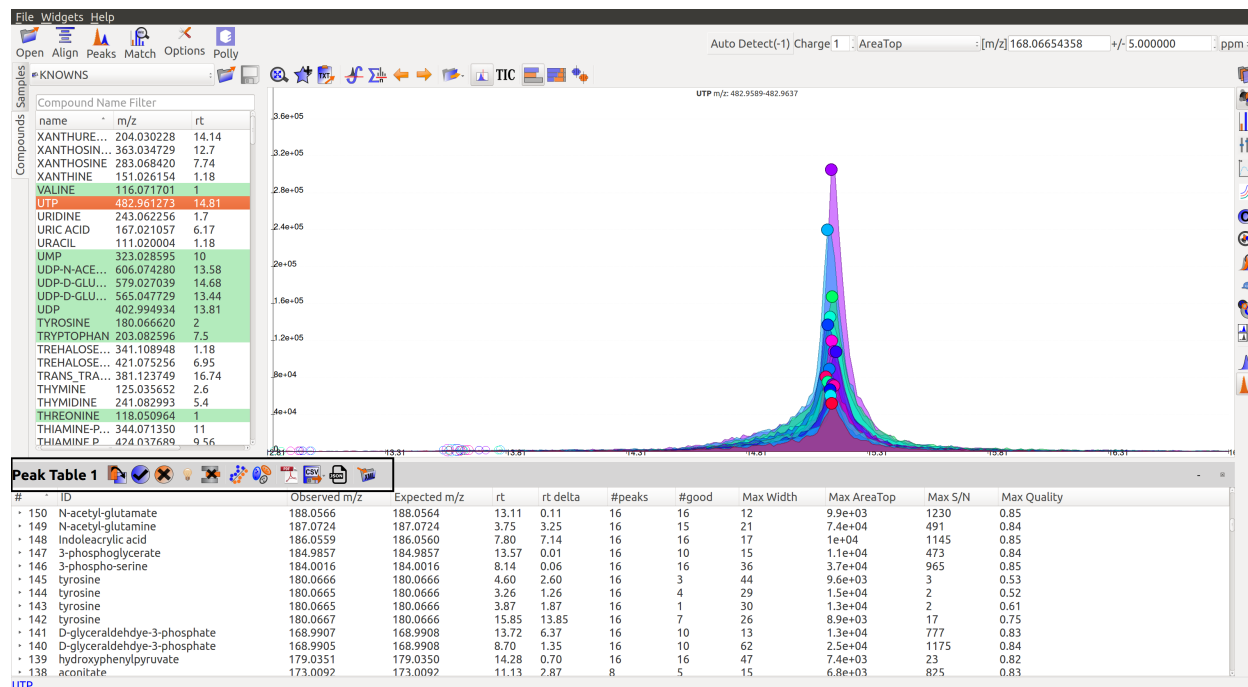
Multiple groups can be annotated as the same compound especially when retention time information is not used during Database search. The peak table provides options for filtering, comparing or exporting data from the table. Following are the different menu options available in the peak table:

- *Switch between group and peak views*: Switching to Peak view displays only Peak information. This includes group number, group ID, Expected m/z, Observed m/z, retention time and intensity of all peaks in the group with sample names as the respective column headers. Peak intensity cells are colored based on their relative values in a group. Highest intensity value has the lightest color and vice-versa.
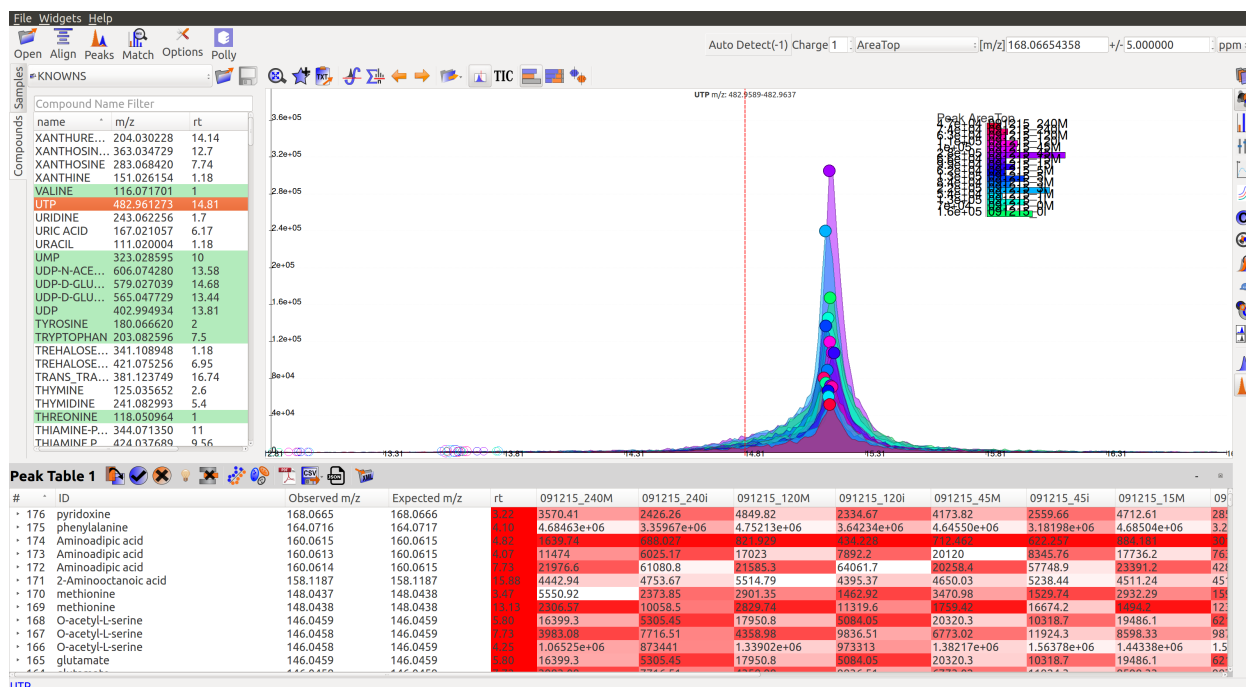
- *Mark Group as Good*: Used to manually curate selected peaks as 'good'. User can also press 'G' on their keyboard for the same. Manual curation has been described here.
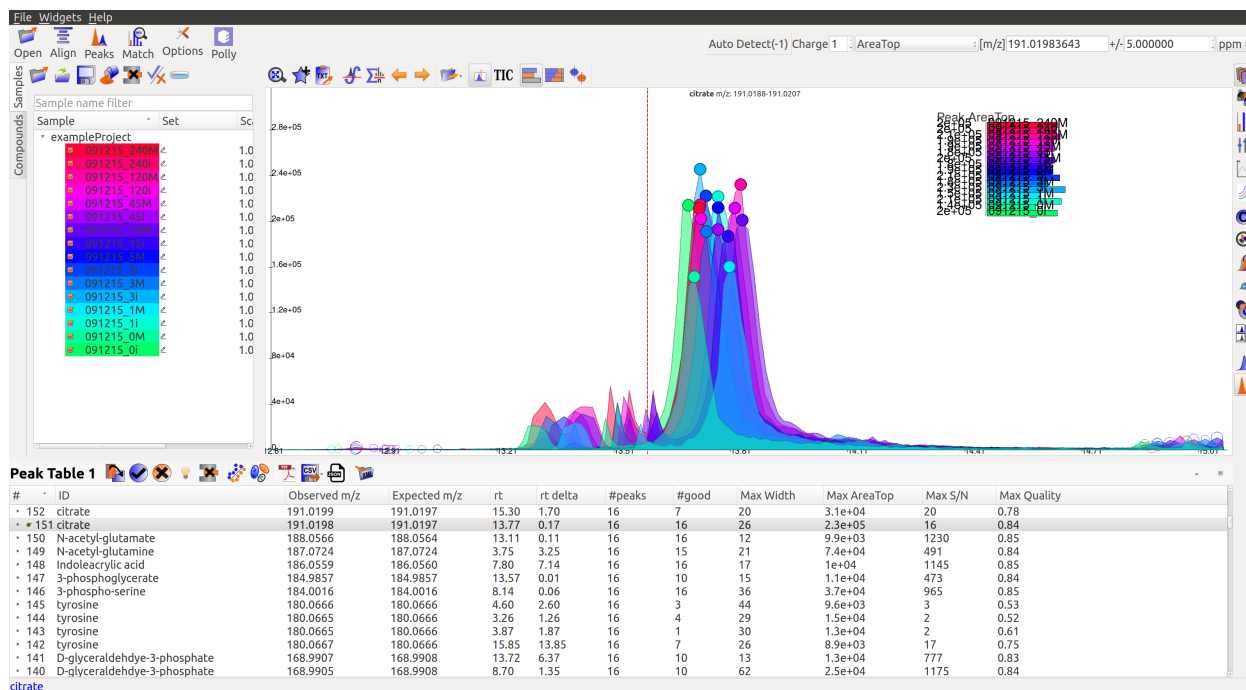
- *Mark Group as Bad*: Used to manually reject peaks by marking them as 'bad'. User can also press 'B' on their keyboard for the same. Manual curation has been described here.
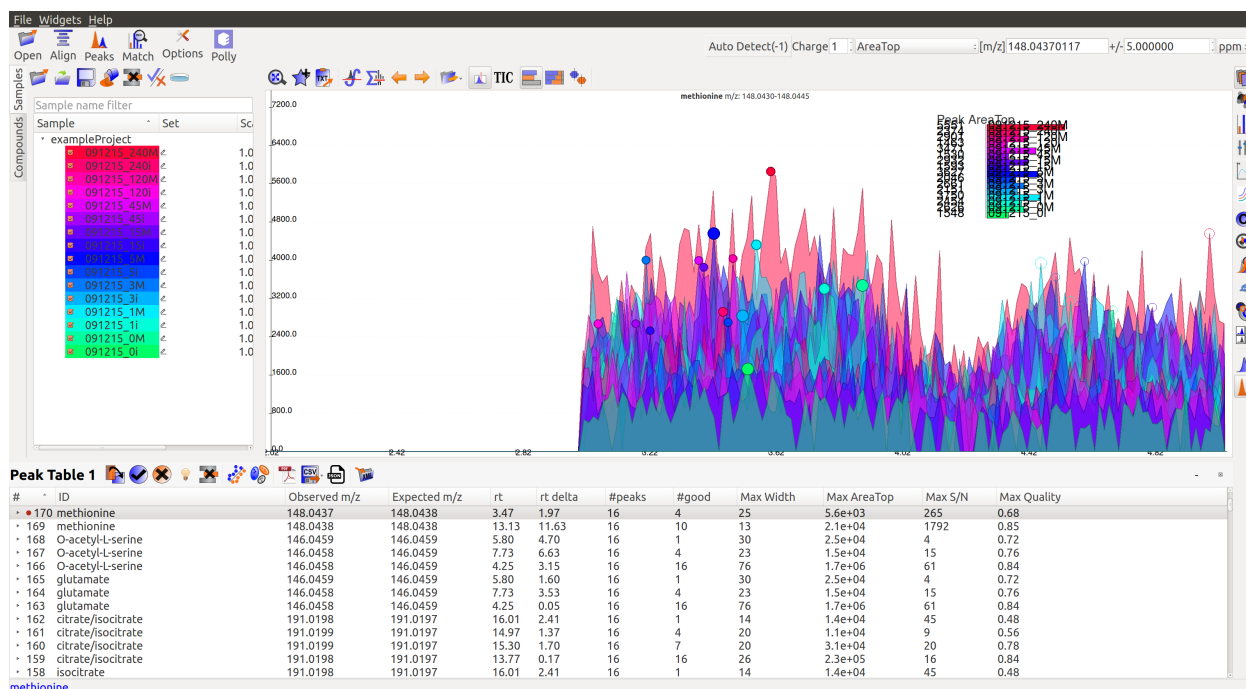
- *Train Neural Net*: Used to retrain the neural net algorithm to recognize good/bad peaks. User manually curates 100 peaks to train the algorithm.

- *Delete Group*: Deletes the selected group(s) from the peak table.

- *Show Scatter Plot*: Opens the Scatter plot widget used to compare different cohorts via Scatter plot and Volcano plot.

The remaining are export options and will be detailed in the Export section.

Statistics

El-MAVEN comes equipped with a statistics module for comparing data across different cohorts. Users can set the sample cohorts either by editing the Set column in the Sample space, or upload a meta file with sample and cohort names as detailed above under the Sample Space Menu section.

The statistics module can be accessed through the Peak Table menu.
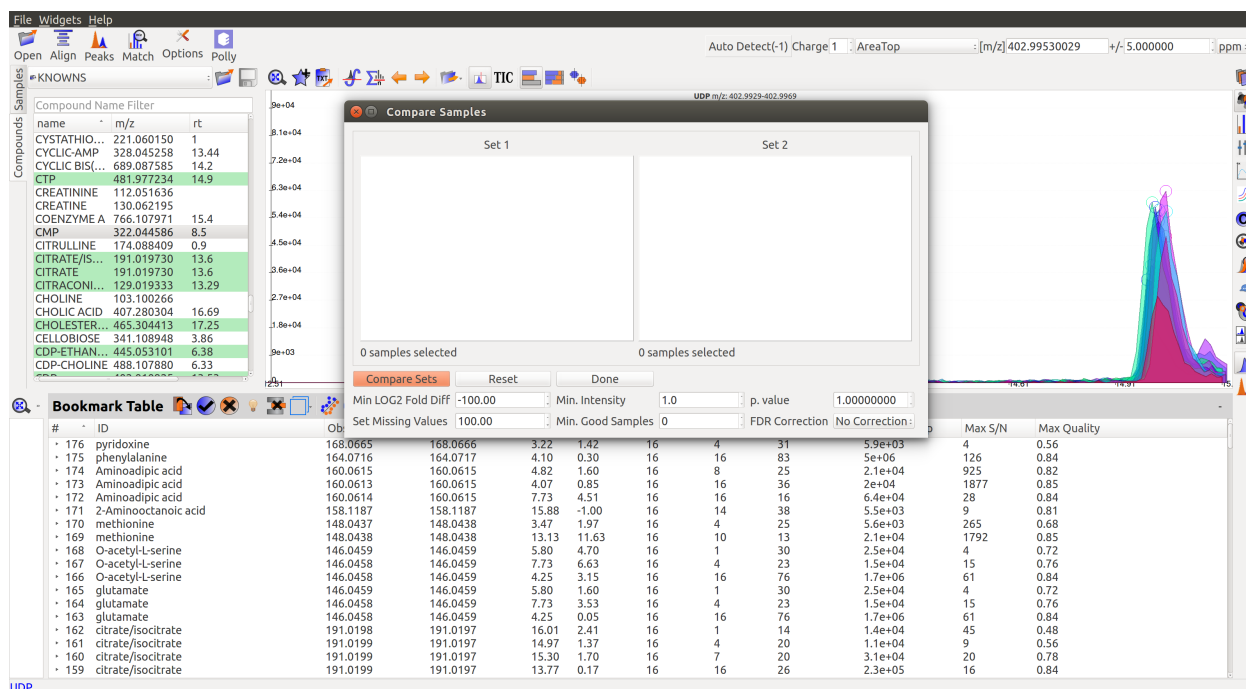
- *Set1/Set2*: Select two cohorts to be compared

- *Min Log2 Fold Difference*: Fold difference is a measure of how much the intensity of a group changes from one cohort to another. User can set the minimum threshold for this value in $\log_2$ format.

- *Min Intensity*: Groups with all peak intensities less than this value will be filtered out from the comparison process.

- *p value*: A t-test is performed to find if the intensity distributions of the two selected cohorts are significantly different from each other. This test returns a p-value indicating how significantly different a group behaves between the two cohorts. A lower p-value shows higher significance.

- *Set Missing Values*: User can set the default intensity value to be used in case the group is missing from a sample.

- *Min. Good Sample*: Groups should have a minimum number of good peaks (based on peak quality score) to be considered for comparison.

- *FDR Correction*: False discovery rate is the expected proportion of false positives in a test. There are a number of ways to correct for false positives. (Read More)

- *Compare Sets*: Click to get comparison results.

Compare Sets opens the scatter plot by default
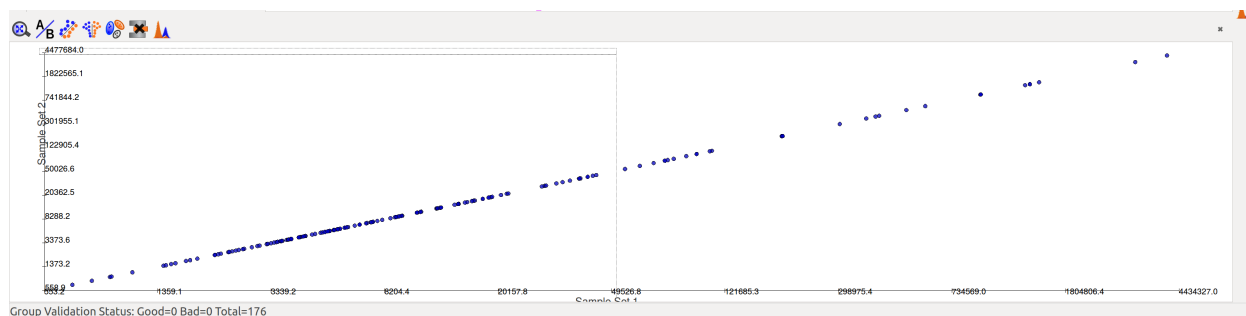
- *Zoom Out*: Zooms out of the plot.
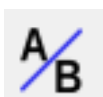
- *Compare Samples*: Opens the compare samples dialog again to adjust settings.

- *Scatter Plot*: The axes represent the average peak intensity (Peak Height) for sample 1 and 2 respectively. Each bubble is a group. The bubble size represents fold change between the samples. The significance (or p-value) of the fold change is represented by the bubble color. Red and blue signify higher intensity in sample 1 and 2 respectively. Opaqueness of the bubble represents the significance (or inverse of p-value) of the fold change between cohorts.

- *Volcano Plot*: The axes represent fold change and significance of fold change respectively. Red and blue bubbles represent positive and negative fold change respectively.

- *Delete*: Deletes a data point from the graph.

- *Scatter Plot Table*: A separate Peak Table is created with all filtered groups being used for statistical analysis. User can also export these in a CSV or JSON.

Export

Users can either save the state of the project or export only relevant data from the peak table. These are the different export options available in El-MAVEN:
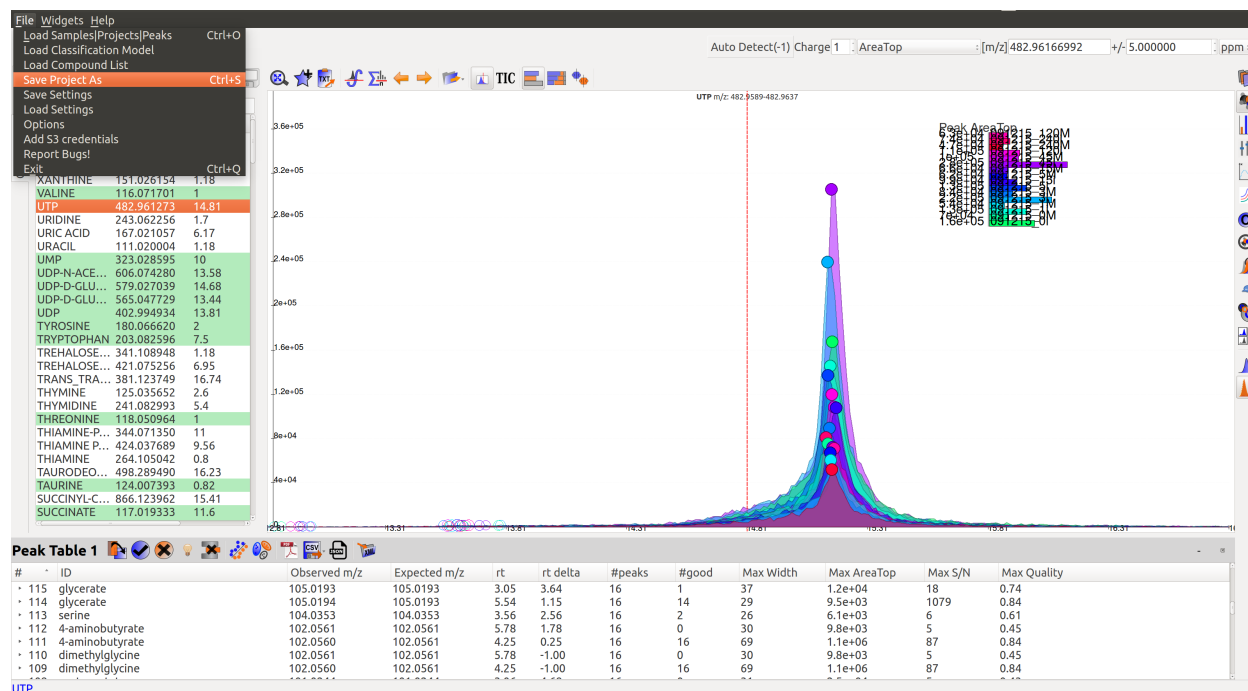
- *Save Project as*: This option is available in the File menu. It saves all peak tables and current settings in a .mzroll file. On loading the .mzroll file, all sample files are uploaded and the peak tables and EIC are available.

If the user wishes to save only certain Peak Tables, they can click on at the top of the Peak Table(s). This will only store that specific peak table instead of all.

- *Generate PDF Report*: This option is available on at the top of the Peak Table. It saves all EICs with their corresponding bar plots in a PDF file.

- *Export Groups to SpreadSheet (.csv)*: This option is available on top of the Peak Table . You can choose to export the whole table or a subset of the data. There are 4 possible selections: export only selected groups, export all groups, export only good groups or export only bad groups. The data is stored in a comma separated file.
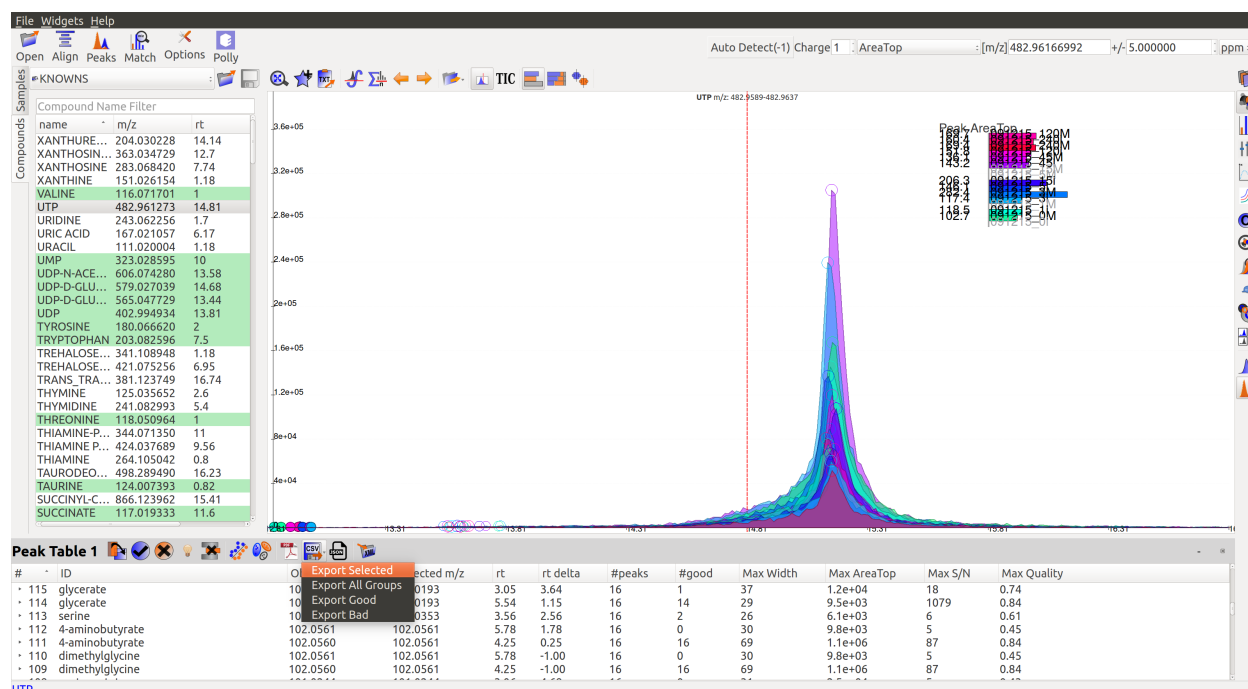
- *Export EICs to Json*: This option is available on top of the Peak Table . It exports all EICs to a Json file.

El-MAVEN Command Line Interface

The peakdetector executable in El-MAVEN can load sample files, perform peak detection and save the resulting peaks in a CSV or mzroll file. It reads a configuration file with user-set parameters and paths to input and output files.

This executable is not part of installer versions of El-MAVEN. Please clone the repository and build El-MAVEN to use this feature.

Create a default configuration file

Users need an XML file to specify their configuration and specify directory paths. The XML template can be generated using a simple command and modified according to user requirements.

- Open the terminal
- Move to your local El-MAVEN folder:

```
$ cd user@pc:~/ElMaven/ElMaven
```

- Execute the following command
  - For Windows and Ubuntu:

```
$ ./bin/peakdetector.exe --defaultxml
```

  - For MacOS:

```
$ ./bin/peakdetector.app/Contents/MacOS/peakdetector --xml config.
↪xml
```

- A default XML will be created in your current folder

Peak detector parameters

CLI provides a limited number of parameters for user-modification. Default values are used for all other parameters used in El-MAVEN GUI.

The template XML contains three sets of parameters:

1. OptionalDialogArguments: Global parameters

- Users need to know the ionization mode and accepted values are 0 (neutral), -1 (negative) and +1 (positive): `ionizationMode type="int" value="-1"`

- Ionization charge. This is the number of positive or negative charge added to every ion. Accepted values are positive integers: `charge type="int" value="1"`

- Mass accuracy. Set the +/- ppm range for m/z values: `compoundPPMWindow type="float" value="5"`

2. PeaksDialogArguments: Peak detection parameters

- Peak Detection Algorithm. Accepted values: "0" for Compound Database Search, "1" (or higher) for Automated Feature Detection: `processAllSlices type="int" value="0"`

Following are the arguments used in Automated Peak Detection and their GUI equivalents. These can be found in the 'Feature Detection Selection' tab of Peaks Dialog:

- 'Limit m/z Range' minimum: `minScanMz type="float" value="0"`

- 'Limit m/z Range' maximum: `maxScanMz type="float" value="1e+009"`

- 'Limit Time Range' minimum: `minScanRt type="float" value="0"`

- 'Limit Time Range' maximum: `maxScanRt type="float" value="1e+009"`

- 'Limit Intensity Range' minimum: `minScanIntensity type="float" value="50000"`

- 'Limit Intensity Range' maximum: `maxScanIntensity type="float" value="1e+010"`

- 'Mass Domain Resolution': `ppmMerge type="float" value="20"`

Following are the arguments used in Compound database search and their GUI equivalents. These can be found in the 'Feature Detection Selection' tab of Peaks Dialog:

- 'Select Database'. Enter full path to the database file: `Db type="string" value="0"`

- 'Match Retention Time (+/-)'. Value of '0' will set the flag as false. Positive integer value will set the allowed deviation in minutes: `matchRtFlag type="int" value="0"`

- 'Limit Number of Reported Groups per Compound': `eicMaxGroups type="int" value="0"`

Isotope detection. Following are the settings used to pull isotopes in case of database search and their GUI equivalents. These can be found in the 'Isotope Detection Options' in the Options Dialog:

- Choose labels present in your data. Enter a four digit value of '1's and '0's each representing a label: D2, N15, S34 and C13. For eg. 0000 signifies no labels, 0001 signifies your data has C13 label and 1111 signifies your data has all four labels: `pullIsotopes type="int" value="0"`

EIC processing. Following are the parameters used for EIC processing and filtering and their GUI equivalents. These can be found under the 'Peak Detection' tab of Options Dialog:

- 'Maximum Retention Time Difference between Peaks': `grouping _maxRtWindow type="float" value="0.5"`

- 'EIC smoothing window': `eicSmoothingWindow type="int" value="10"`

Group Filtering. Following are the parameters used for filtering detected groups and their GUI equivalents. These can be found under the 'Group Filtering' tab of Peak Dialog:

- 'Minimum Peak Intensity': `minGroupIntensity type="float" value="5000"`

- Quantitation type of the intensity set in 'Minimum Peak Intensity'. Accepted values are- "0":AreaTop, "1":Area, "2":Height, "3":AreaNotCorrected: `quantitationType type="int" value="0"`

- 'At least x% peaks above minimum intensity' slider. Enter value in percentage. 0% will default to 1 peak: `quantileIntensity type="float" value="0"`

- 'Minimum Quality': `minQuality type="float" value="0.5"`

- 'At least x% peaks above minimum quality' slider: `quantileQuality type="float" value="0"`

- 'Minimum peak width'. Enter number of scans: `minPeakWidth type="int" value="1"`

- 'Minimum Signal/Baseline Ratio': `minSignalBaseLineRatio type="float" value="2"`

- 'Minimum Good Peak/Group'. Value should be less or equal to the number of samples: `minGoodGroupCount type="int" value="1"`

- 'Peak Classifier Model File'. The default model file can be found in the bin folder of El-MAVEN installation. Enter full path to the 'default.model' file: `model type="string" value="0"`

- Not used by CLI: `rtStepSize type="float" value="20"`

3. GeneralArguments: General parameters

- Alignment. Positive integer value will run an alignment algorithm: `alignSamples type="int" value="0"`

- Export EIC in JSON. Positive integer value will save EICs JSON in user-specifed output folder: `saveEicJson type="int" value="0"`

- Output directory. Enter full path to desired output folder: `outputdir type="string" value="0"`

- Save Project. Positive integer value will save the project as a .mzroll file in user-specified output folder. This file can be loaded in El-MAVEN GUI for further processing or visualization: `savemzroll type="int" value="0"`

- Sample Path. Enter full path to a sample file in each row: `samples type="string" value="0"`

Run

Once the parameters and directory paths have been set in the configuration file, run peak detection from the terminal using the following command from the El-MAVEN root directory:

- Windows and Ubuntu:

```
$ ./bin/peakdetector.exe --xml config.xml
```

- MacOS:

```
$ ./bin/peakdetector.app/Contents/MacOS/peakdetector --xml config.xml
```

The resulting CSV file (and other files depending on the configuration) can be found in the specified output directory.

Help

To print the help commands, execute the following:

```
$ Peakdetector.exe -h
```

1.4.3 Tutorials

Unlabeled LC-MS Workflow

This is a tutorial for processing Unlabeled LC/MS data files through El-MAVEN.

Preprocessing

msConvert is a command-line/GUI tool that is used to convert between various mass spectroscopy data formats, developed and maintained by proteoWizard. Raw data files obtained from mass spectrometers need to be converted to certain acceptable formats before processing in El-MAVEN.

Input

msConvert supports the following formats:

- .mzXML
- .mzML
- .RAW ThermoFisher
- .RAW Walters
- .d Agilent
- .wiff ABSciex

The settings used for msConvert as a GUI tool are captured in the following screenshots:

Note: zlib compression is enabled by default in msConvert. El-MAVEN in its current form does not support zlib compression. It is important to uncheck "Use zlib compression" box.

Output

msConvert can convert to an array of different formats but El-MAVEN primarily uses .mzXML and .mzML formats.

Launch El-MAVEN

Once sample files are ready for processing, launch El-MAVEN.

Adjust Global Settings

Global Settings can be changed from the *Options* dialog . There are 9 tabs in the dialog. Each of these tabs has parameters related to a different module in El-MAVEN. For example, a tab for Instrumentation details, a tab for File Import settings etc.

To know more about the functionality of different tabs and their settings, users can see the Widgets page. Please be sure to set the desired settings before processing an input file.

Load Samples

Users can go to File in the Menu and click on *Load Samples|Projects|Peaks* option. Then navigate to the folder containing the sample data, select all .mzXML or .mzML files and click Open. A loading bar displays the progress at the bottom.

When the samples have loaded, users should see a sample panel on the left side. If it is not displayed automatically, click on the *Show Samples Widget* button on the toolbar. El-MAVEN automatically assigns a color to every sample. Users can select/deselect any sample by clicking the checkbox on the left of the sample name.

Load Compound Database

Users can click on *Compounds* option in the leftmost menu, navigate to the folder containing the standard database file, select the appropriate .csv file and click *Open*. Alternatively, users may use any of the default files loaded on start-up.

This is a sample Compound Database:

It lists all metabolite names, chemical formula, HMDB ID, and the class/category of compounds they belong to (if known).

Mark Blanks

Users can mark the blanks by selecting the blank samples from the list on screen, and clicking on the *Set as a Blank*

Sample icon in Samples menu.

Multiple blanks can be marked together. The blanks will appear black as shown in the image below:

Alignment

Prolonged use of the LC column can lead to a drift in retention time across samples. Alignment shifts the peak retention time in every sample to correct for this drift and brings the peaks closer to median retention time of the group.

In the above image, EIC for a UTP group is displayed. If the samples were aligned, all peaks would lie at the same retention time. Since this is not the case, the samples need to be aligned.

Alignment visualization can be used to judge the extent of deviation from median retention time.

In the above visualization, each box represents a peak from the selected group at its current retention time. Samples are said to be perfectly aligned when all peak boxes lie on the same vertical axis. The peaks are considerably scattered in the above image and therefore the samples should be aligned for better grouping of peaks.

Perform Alignment

Alignment settings can be adjusted using the Align button . This example was aligned with *Poly fit* algorithm with default parameters.

Post-alignment the peaks in the group should appear closer to the median retention time of the group.

Pre-alignment, the peaks were considerably scattered while the aligned peaks lie nearly on the same axis. Users can run alignment again with different parameters if required (or with a different algorithm). Further details on Alignment settings are available on the Widgets page.

Peak Grouping

Peak grouping is an integral part of the El-MAVEN workflow that categorizes all detected peaks into groups on the basis of certain user-controlled parameters. A group score is calculated for every peak during the process. The formula for this score takes into account the difference in retention time, intensities between peaks (smaller difference leads to a better score) and any existing overlap between them (higher extent of overlap leads to better score). All three parameters have certain weights attached to them that can be controlled by the users. The formula for the score is shown in the image. More details on it can be found on the Widgets page.

The above image shows two groups in the EIC window. The highlighted (solid circles) peaks belong to group A, the peaks to its left with empty circles belong to another group B. The short peaks in group A that are close to the baseline and peaks in group B come from the same samples. Additionally, the high intensity peaks of group A have a similar peak shape to group B peaks. These peaks might have been wrongly classified into separate groups because of the difference in retention time range of the two sets of peaks. The weights attached to difference in retention time and intensities, and extent of overlap can be adjusted for better grouping.

Grouping parameters can be changed from the Options dialog .

Giving less priority to difference in retention time and intensities results in the two groups being merged into a single group while the peaks that lay close to the baseline are no longer classified as valid peaks.

Baseline

When measuring a number of peaks, it is often more effective to subtract an estimated baseline from the data. This baseline should be set where ideally no peaks occur. Although sometimes the program sets a particular baseline such that one or more peaks occur below that baseline value. In the following image, the dashed line represents each baseline:

The corresponding peaks are indicated with solid circles:

The baseline correction can be done in the *Peak Detection* tab by clicking on *Options* button:

Further details on settings can be accessed here.

The *m/z* option scans the groups to find any specific m/z value and plot its corresponding EIC. The *+/-* option to its right is to specify the expected mass resolution error in parts per million (ppm).
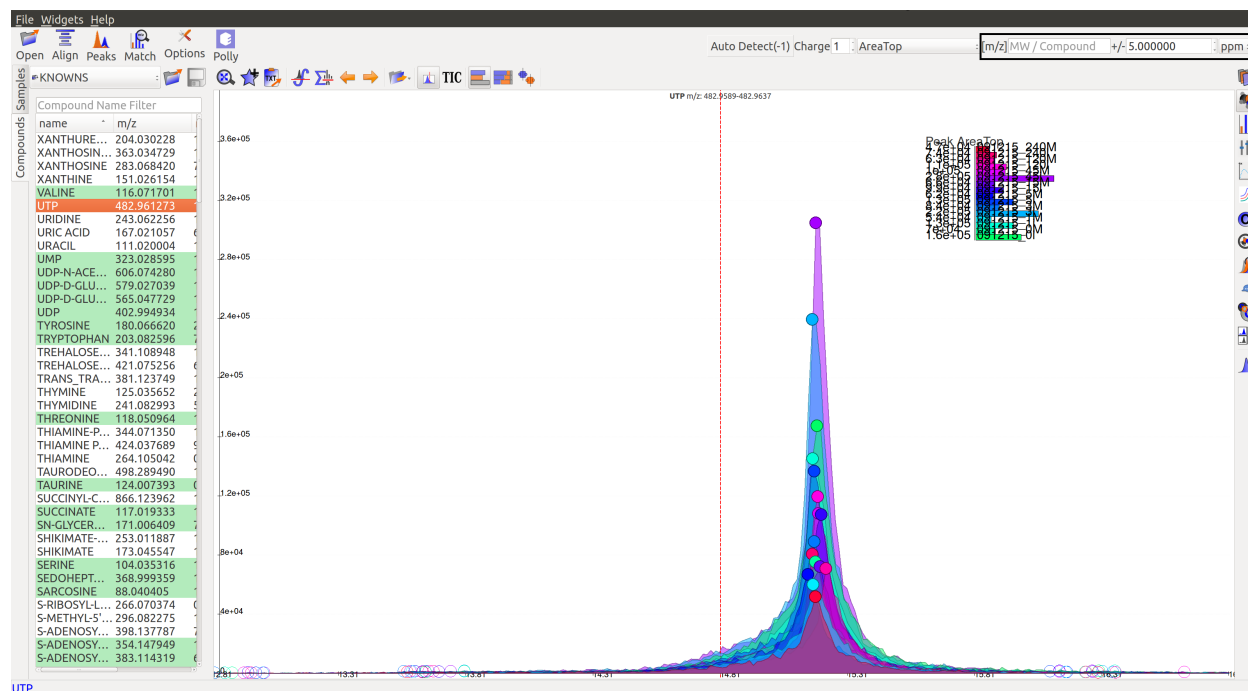
Mass Spectra

Mass Spectra Widget displays each peak, its mass, and intensity for a scan. As the widget shows all detected masses in a scan, the ppm window for the EIC and consequently grouping can be adjusted accordingly. This feature is especially useful for MS/MS data and isotopic detection.
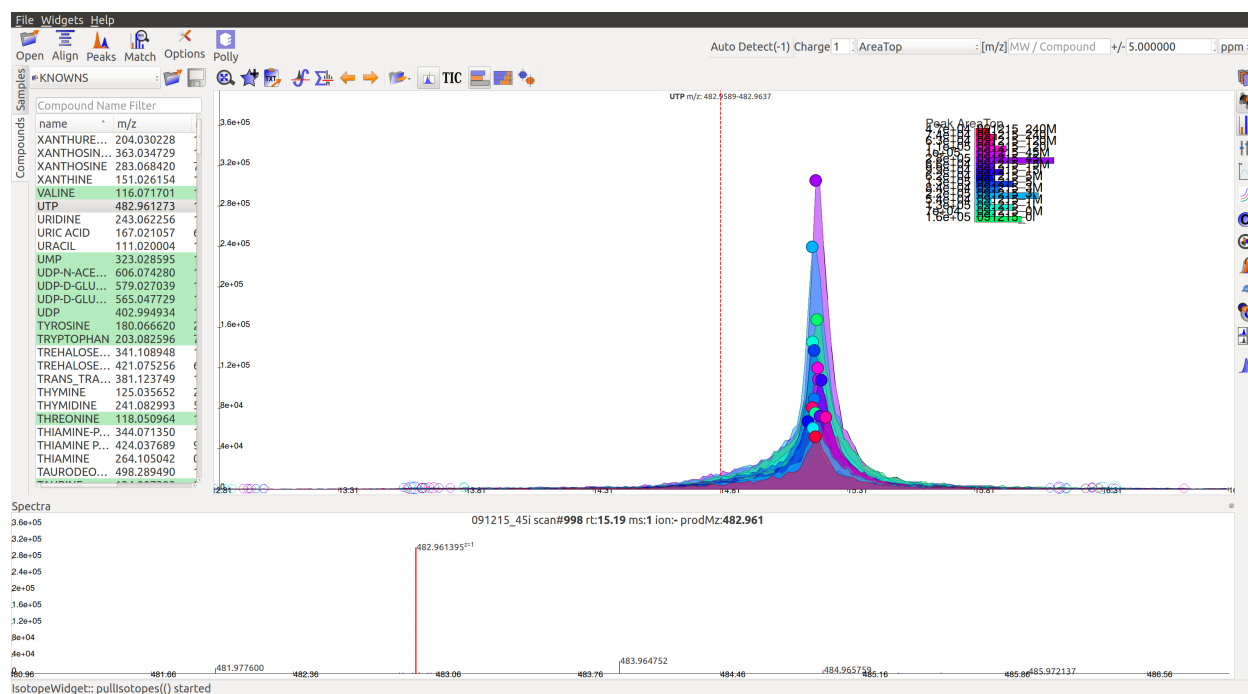
Peak Curation

There are multiple ways to curate peaks in El-MAVEN, though following are the two broad workflows:

1. Manual Peak Curation using Compound DB widget

To use manual curation using the compound DB widget, users have to iterate over all the compounds in the compound DB on the extreme left of the window, as highlighted in the image below.
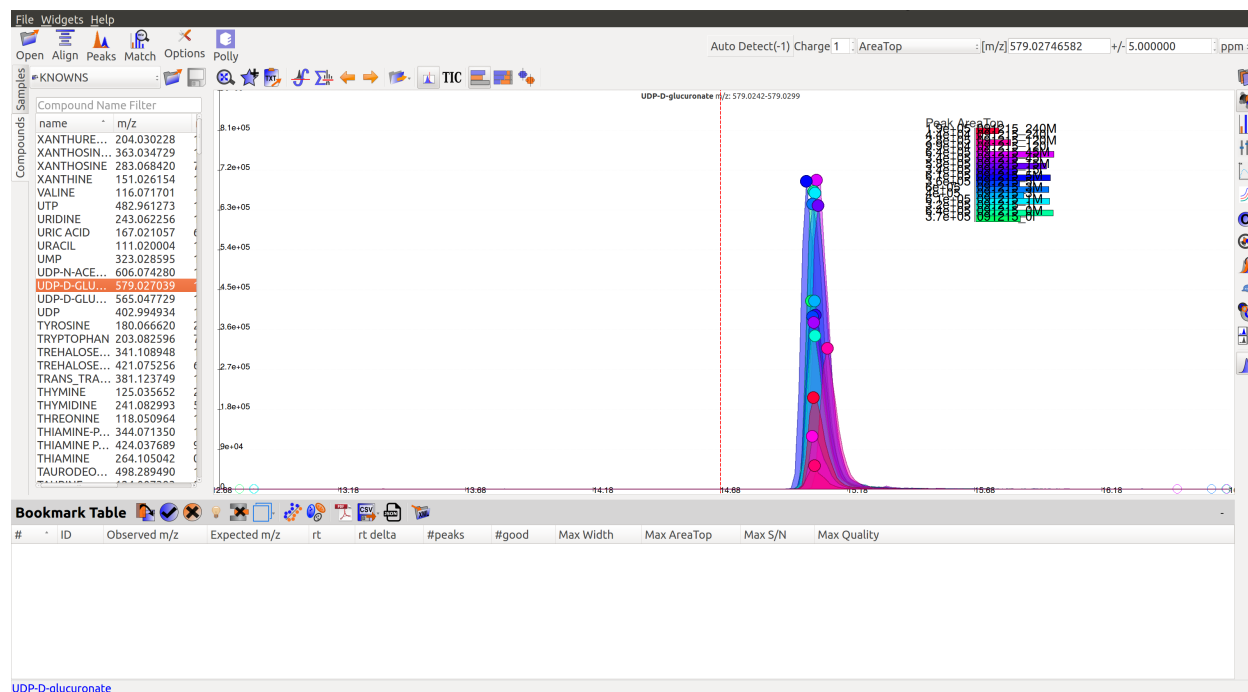
Once on a compound, El-MAVEN shows the highest ranked group for that m/z. Users can now choose a group or reject it.

First, users need to double click on the peak group of their choice. This will get the retention time line to the median of the group and also add the metabolite to the bookmarks table (as shown in the image below). Users can read more about the bookmarks table here.
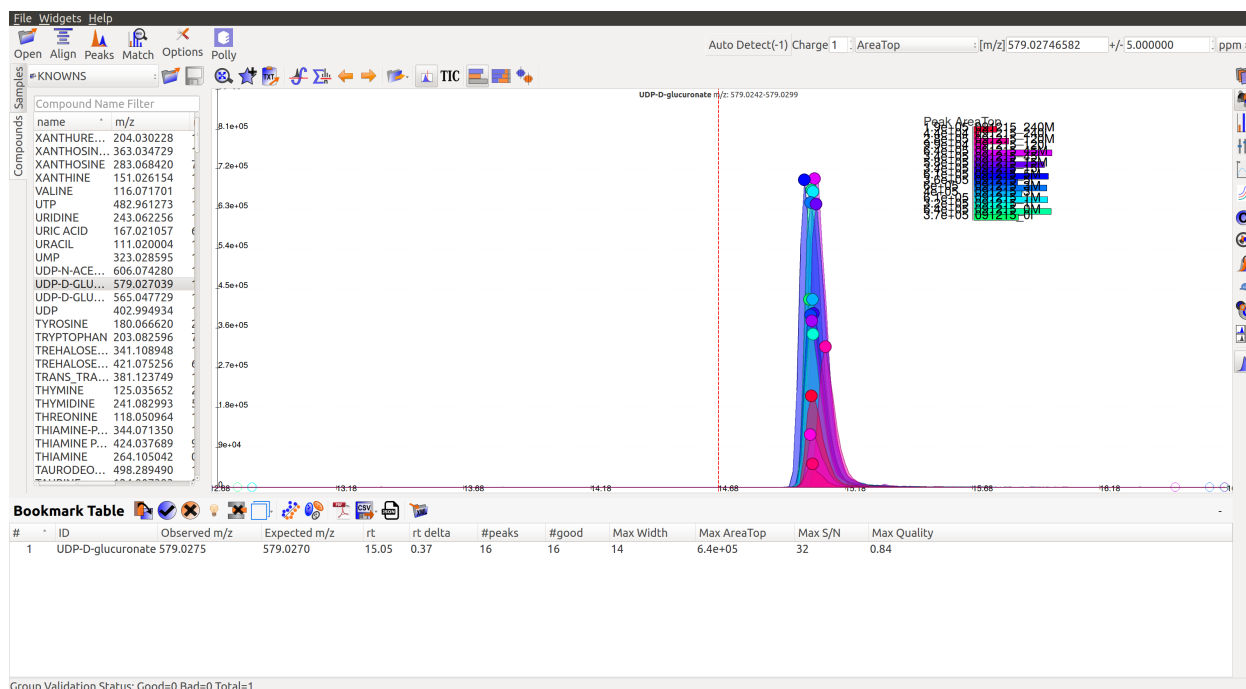
When the users select the first group, they would be asked if they would like to auto-save the state of the application. This feature allows the users to go back to their curated peaks if they so wish in future.
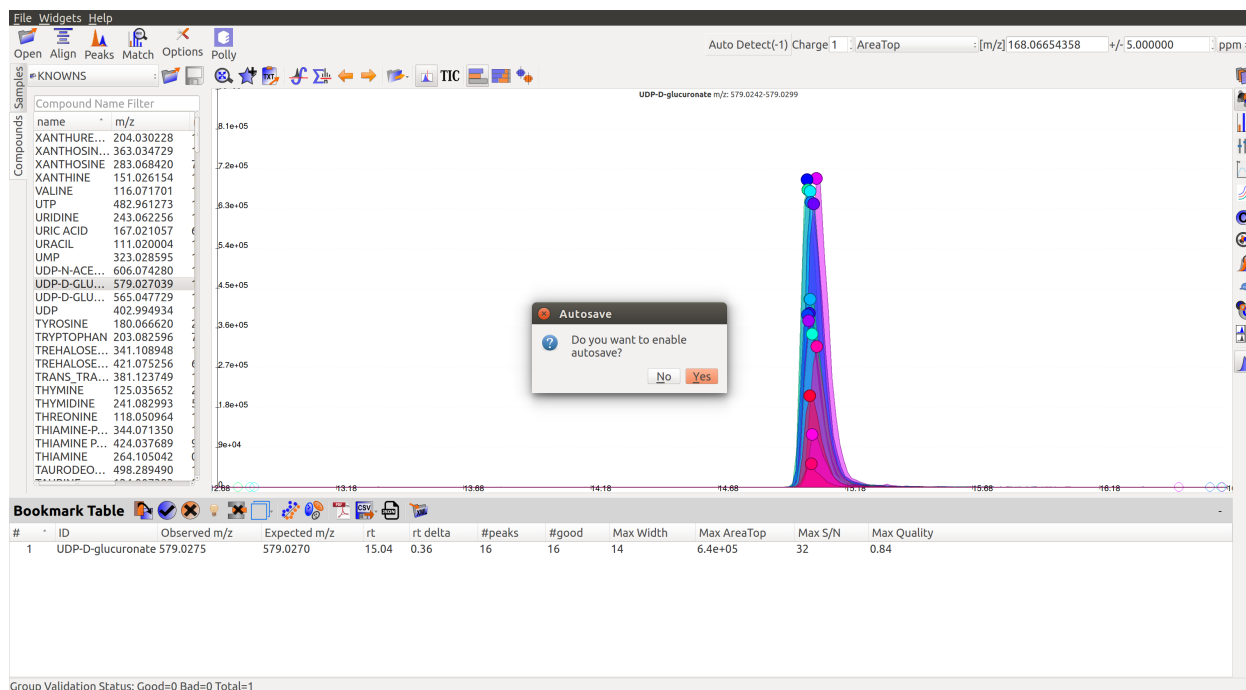
Qualifying peaks as good or bad is explained in the next few sections.

2. Automated Peak Curation

El-MAVEN can automatically select high intensity and high quality groups. This workflow is called automatic peak curation. To enable this workflow users have to click on the peak detection widget present in the top left of the window.

Upon clicking the peak detection widget the following dialog box will open.
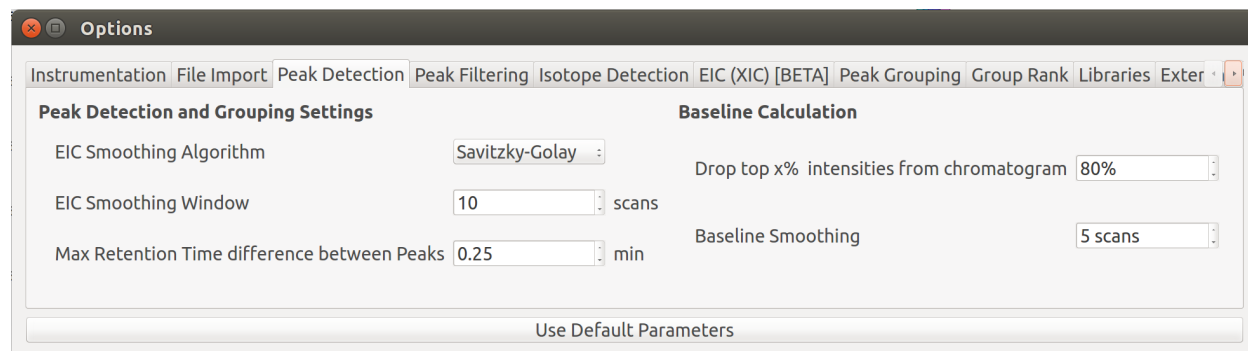
Users can read more about the peak detection widget here.

Upon selecting the default parameters, users can click on *Find Peaks* to select the most important peaks. Once the peak detection is completed, a peak table shows up at the bottom of the window.
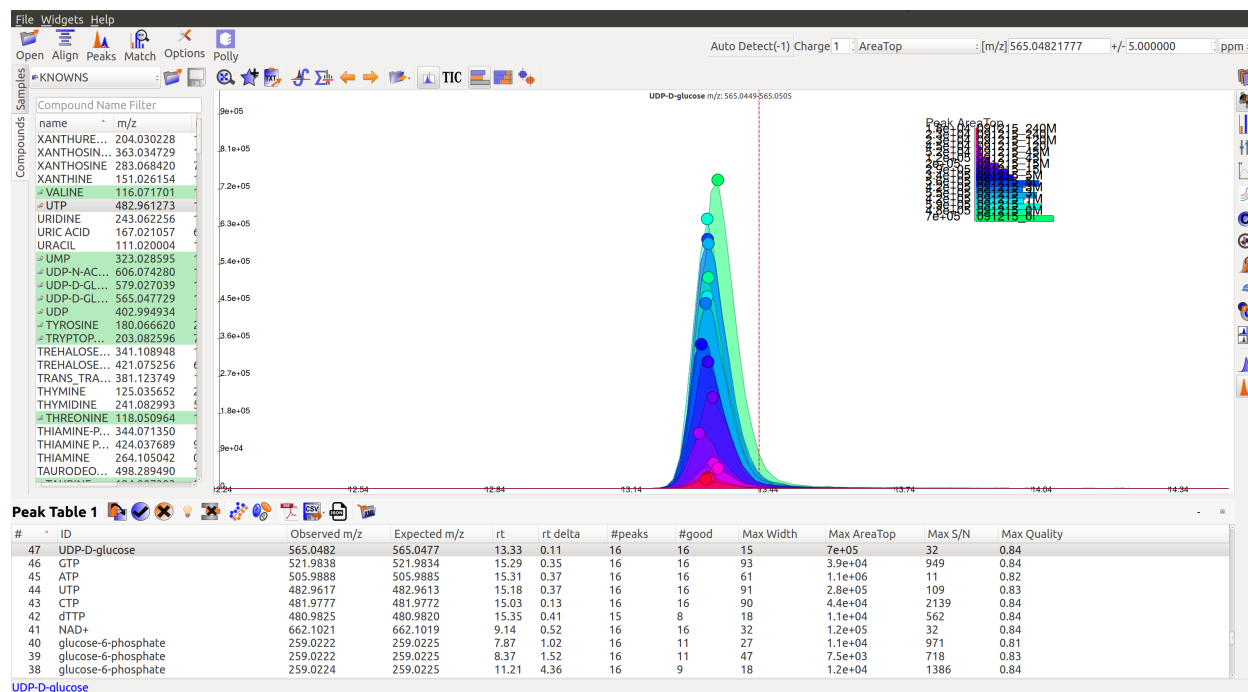
Users can now iterate over these peaks by marking them as good or bad by clicking on the good or bad buttons present in the peak table as shown below.
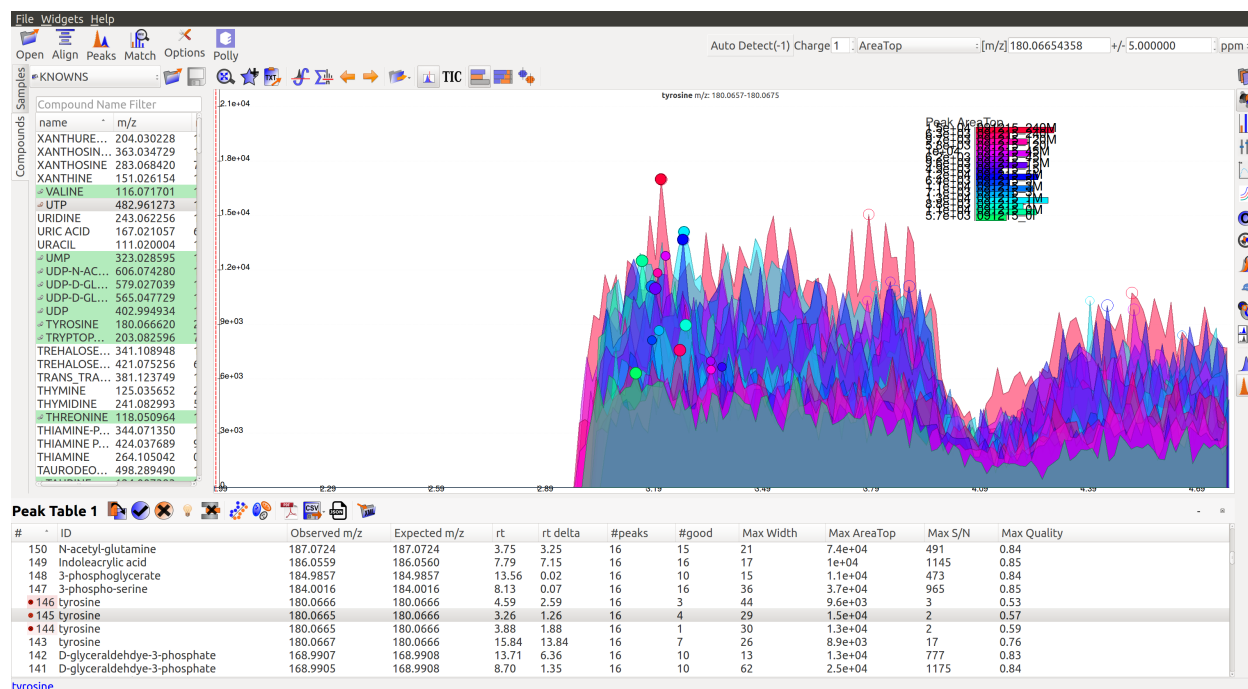
Guidelines for Peak Picking

- A peak's width and shape are two very crucial things to look at while classifying a peak as good or bad. A peak's shape should have a Gaussian distribution and width should not be spread across a wide range of retention time.
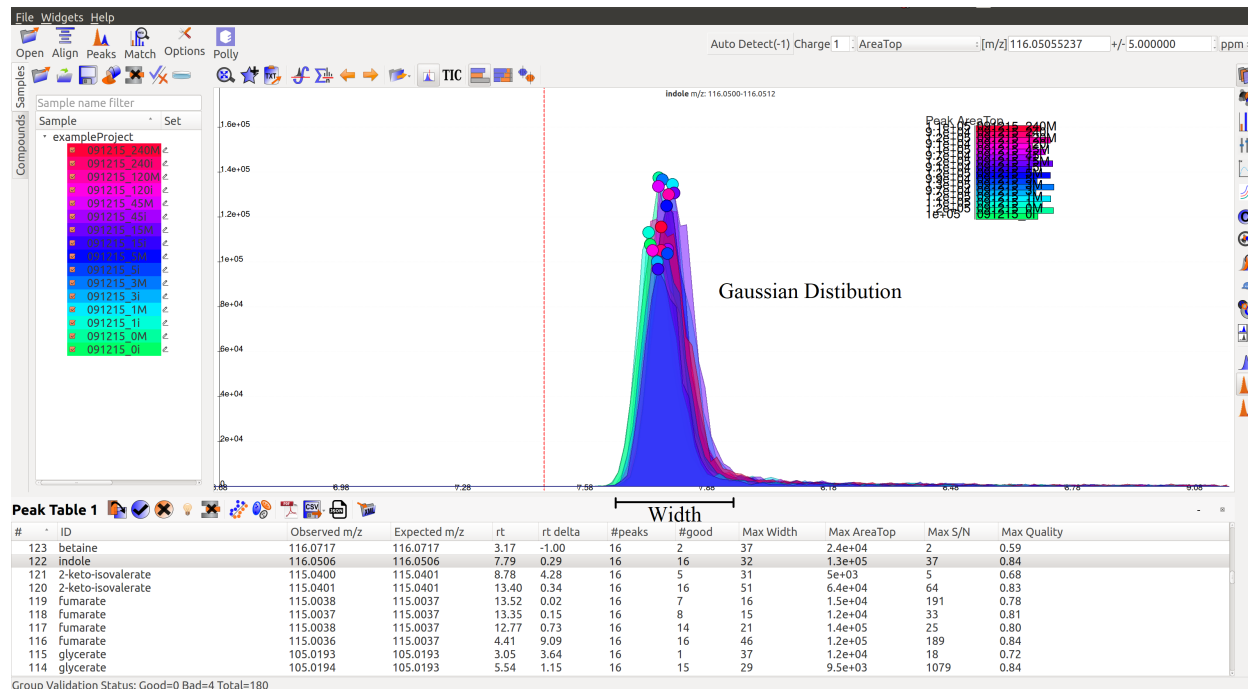
- Peak Intensities for a group are plotted as bar plots for all the samples. These bar plots have heights relative to the other samples.Thus, for a good peak the intensities should be high.
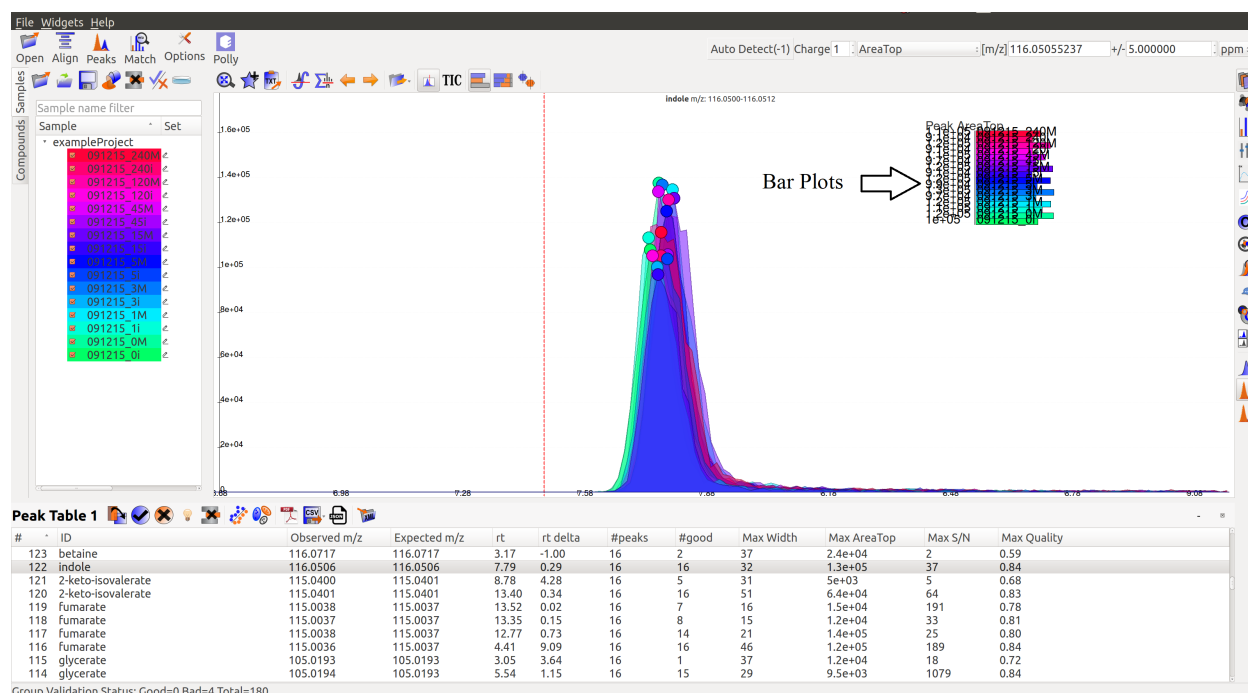
- Intensity Barplot heights should be higher for all the samples than Blank samples. We use intensities of Blank samples to set our group baseline. Blank intensities are used to calibrate intensity values across zero concentration.

- A good peak should have standards with varying intensity in a particular fashion (increasing or decreasing).

- Quality Control (QC) samples give us information about the quality of the data, i.e., it assesses reproducibility and software performance. Samples whose intensities and concentrations are already known are used as QCs to determine if the instrument is working as expected. Values (and scales) can be calibrated using QCs.

- If peak groups of a particular metabolite are separated apart (not aligned well) then we should use stringent alignment parameters to overcome this problem.

- For a particular metabolite, let's say if it has n number of groups, then the group which is much closer to the above guidelines should be selected as a good peak. Multiple groups can also be selected in case of ambiguity (if retention time information is not provided).

A good peak would look similar to the following peaks:
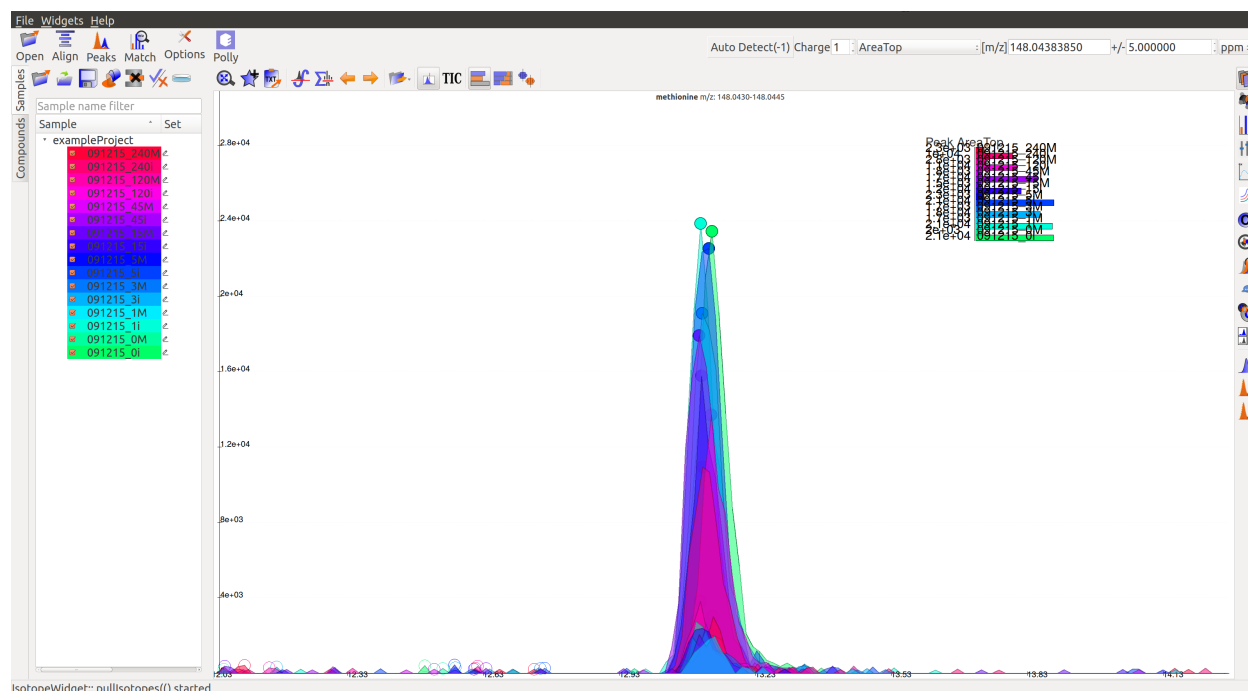
- Gaussian shape

- Perfect grouping, narrow retention time

- Good sample intensities

- Low blank intensities

- QCs look good

- An observable trend in intensity bars of standards, as well as samples.

Some examples of bad peaks are given below:

- The peaks do not have a Gaussian shape. Low intensity peaks are not grouping well. QC intensities (10^4) are too high with respect to the low sample intensities (10^2), which are very close to the noise level.
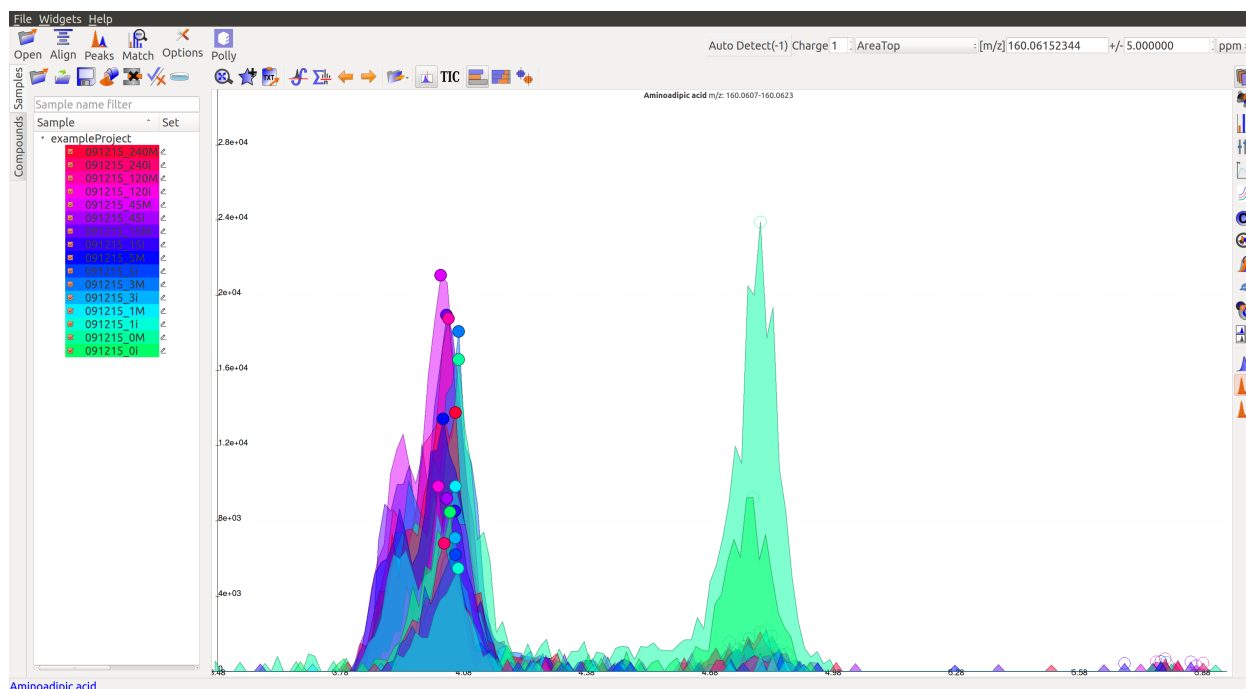
- The peaks have a good Gaussian shape. But the blank intensity bars are high. All the sample intensity bars are shorter or roughly equal to the blank intensities, implying that the peaks are noisy. This should be marked bad if better groups of the same metabolite are available.
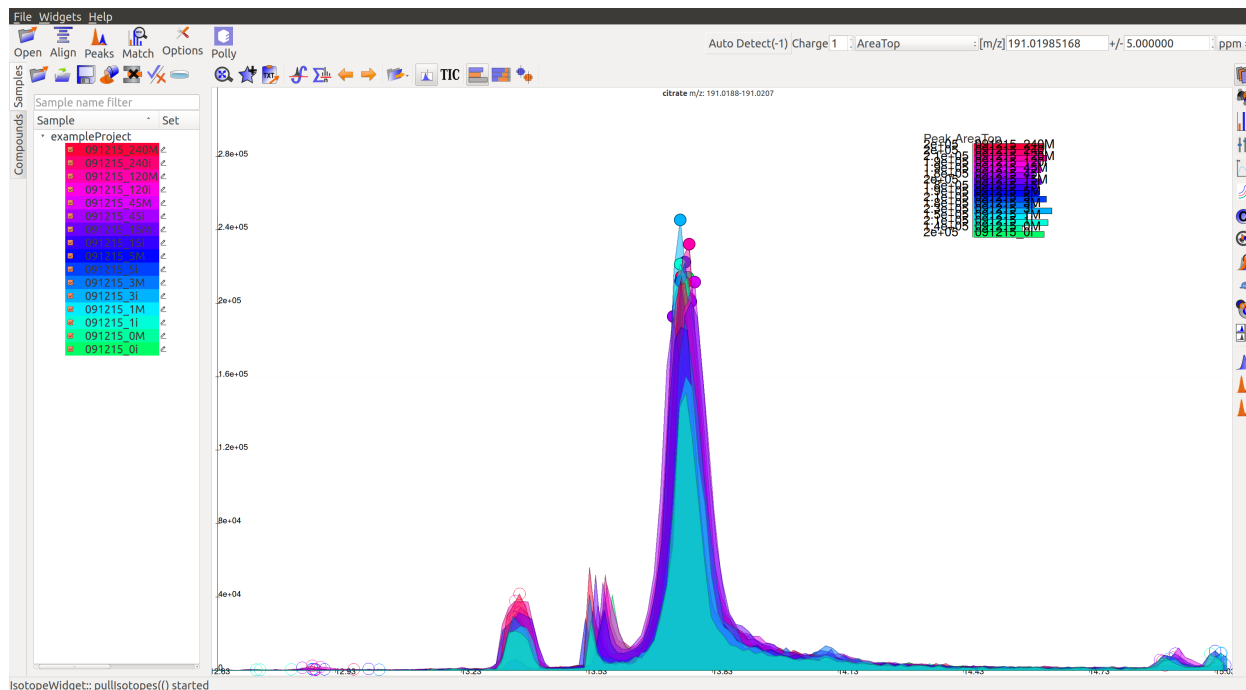
- The intensity levels are high. The blank intensities are lower. However, the peaks are spread over a long range of retention time, have poor grouping, and have forward trailing peaks. If the signal to noise ratio was improved, this peak would probably not be detected.
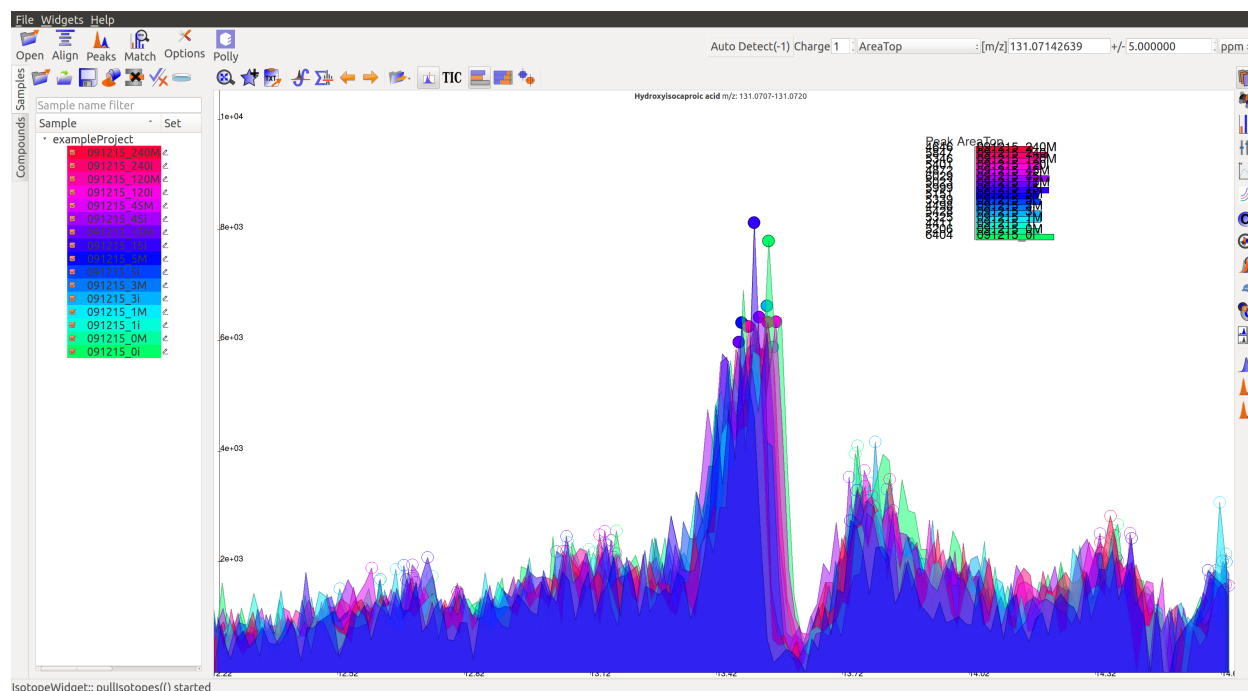
- In the following image, many sample intensities are missing from the intensities bar plots. Peaks do not have a Gaussian shape, nor good grouping. These peaks are probably noise which have been wrongly annotated. The blank intensities are high as well.
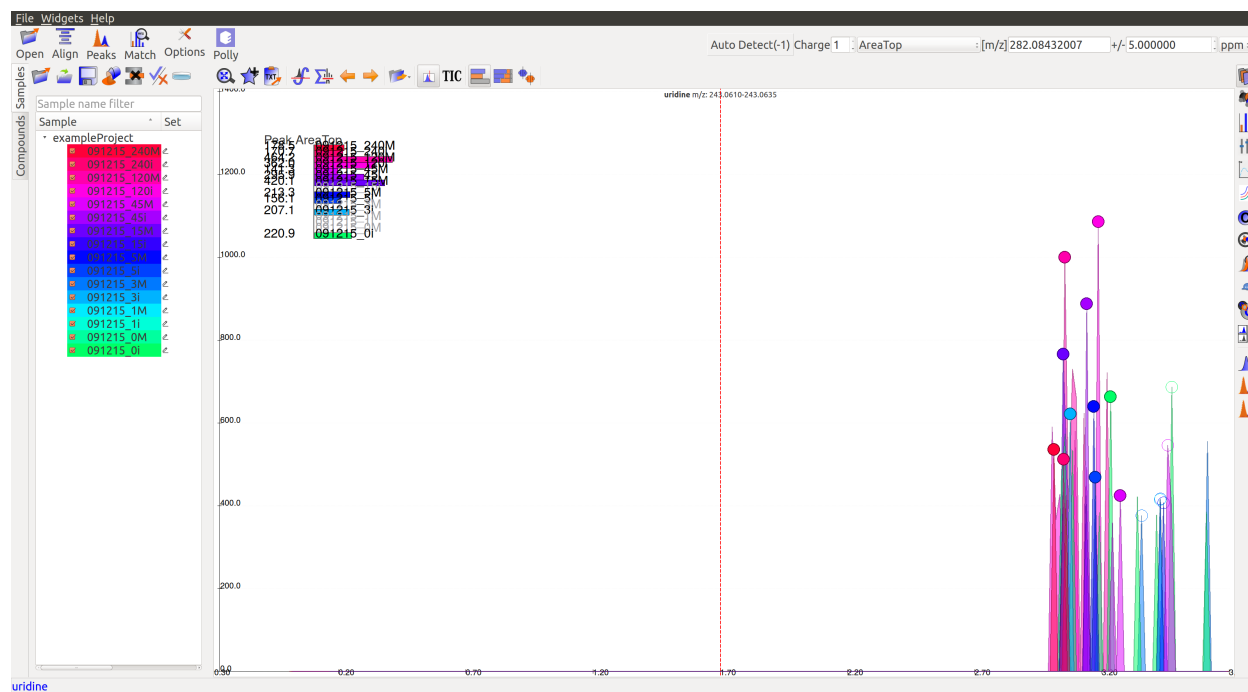
- This is a noisy group. There are no discrete peaks visible in the image. The X-axis is crowded with noise. The peak shape is sharp, triangular, or line-like; not Gaussian. The intensity levels are high, but so are the noise levels.
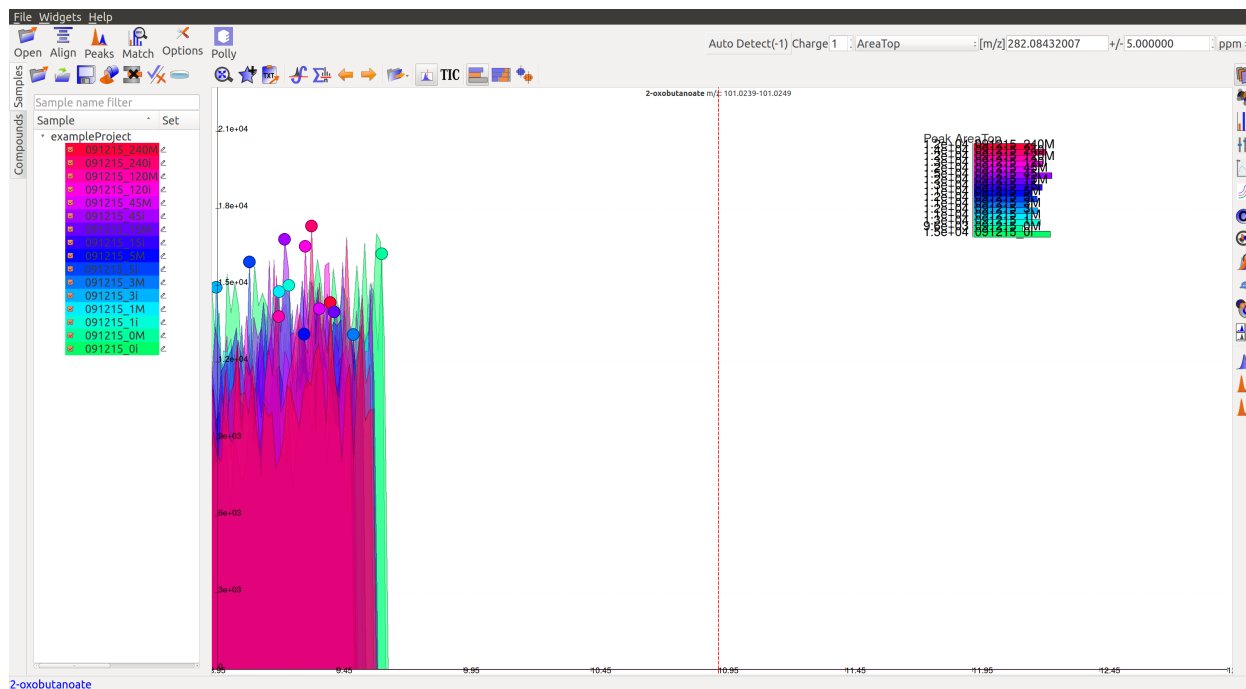
- The peaks don't have a Gaussian shape, and are also noisy. The intensity values are very low.
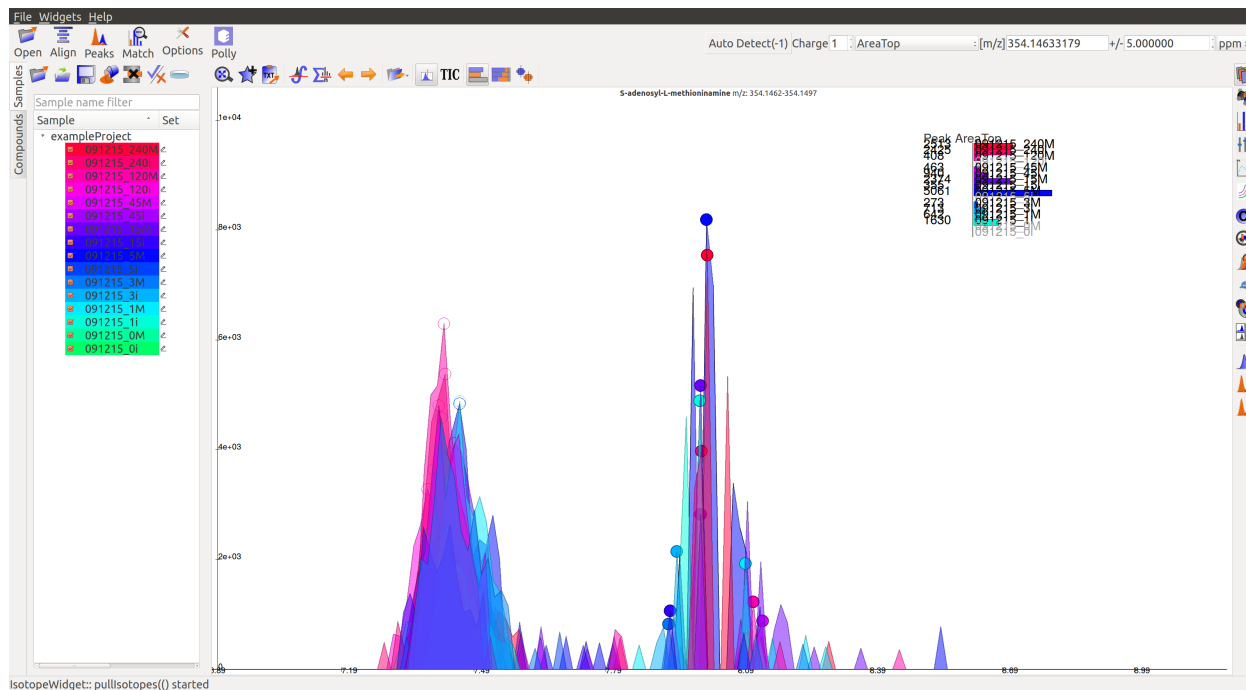
- For low intensity groups like this, the peak characteristics can be determined by zooming in.
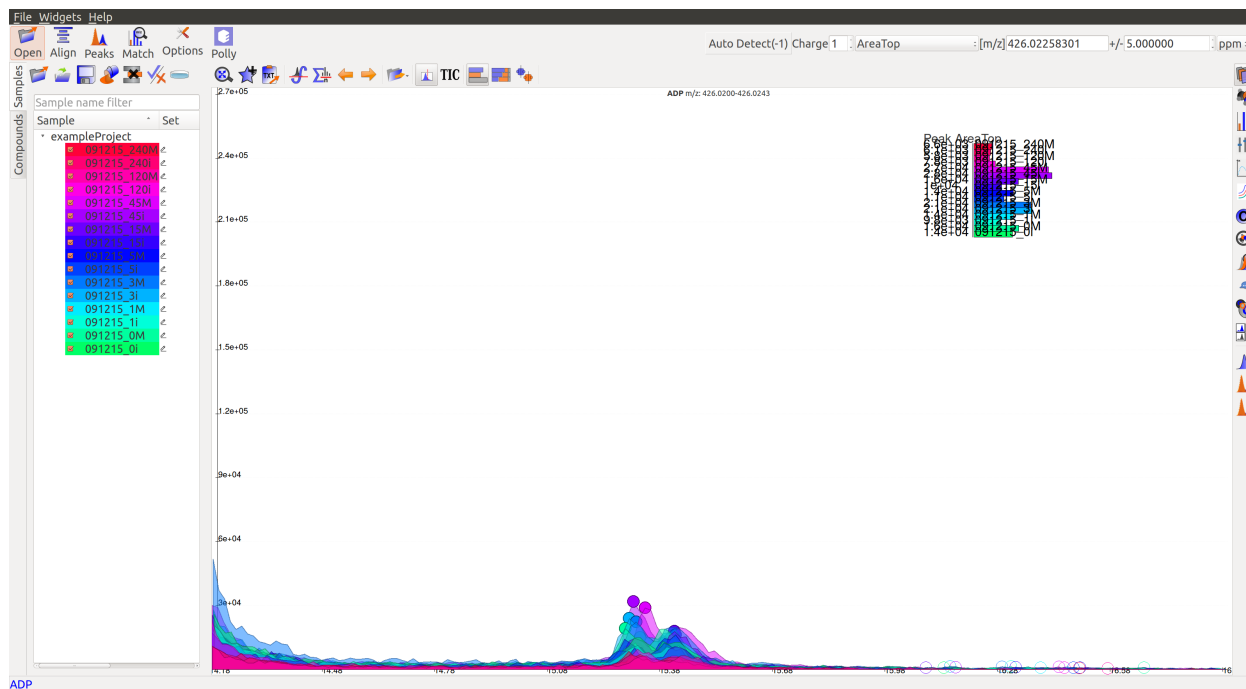
The mouse can be used to select the area of the peak as shown below
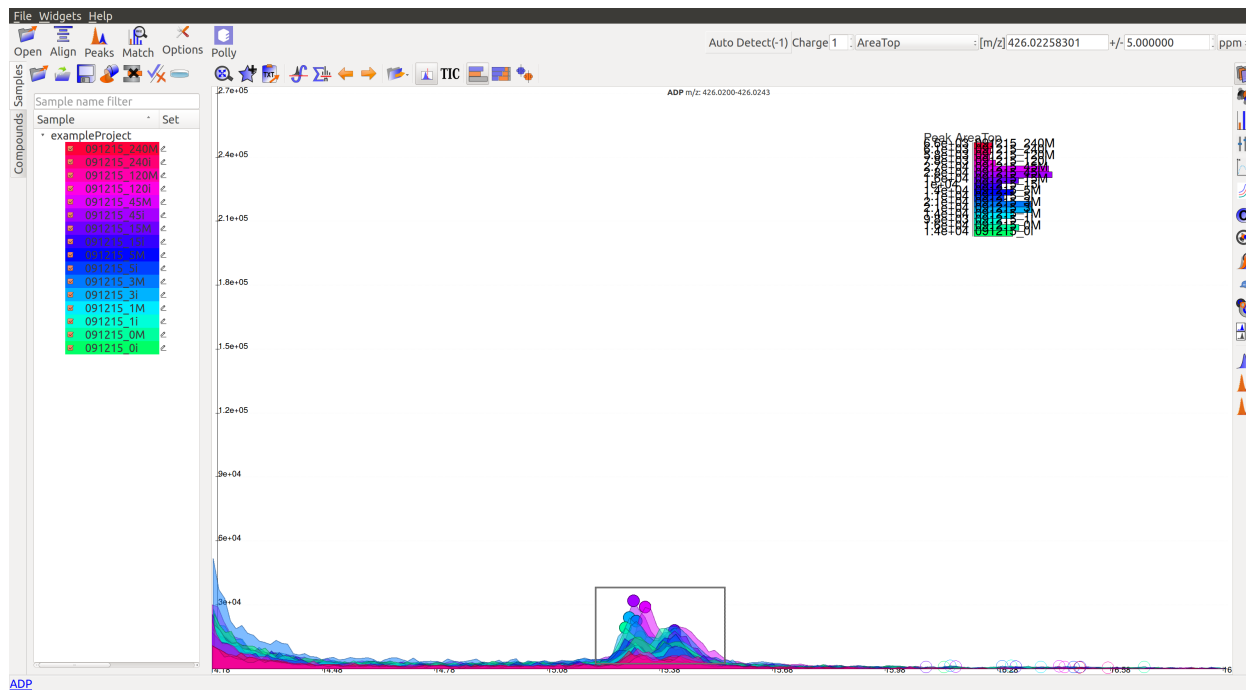
On zooming, it will be easy to make a decision on peak quality
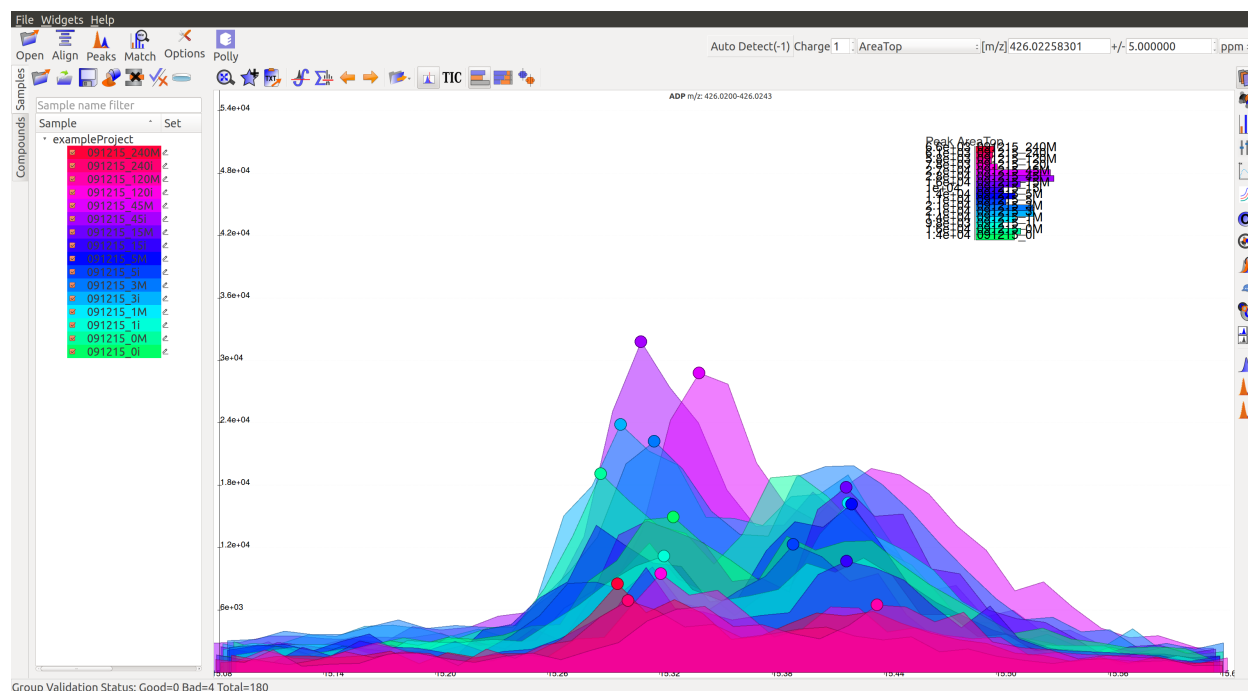
The user can mark any ambiguous peaks as good, and can review all such peaks later in the process.

Export

There are multiple export options available for storing marked peak data. Users can either generate a PDF report to save the EIC for every metabolite, export data for a particular group in .csv format, or export the EICs to a Json file as shown below.

Users can select *All, Good, Bad or Selected* peaks to export.

The *Export Groups to CSV* option lets the users save the 'good'/'bad' labels along with the peak table. Users also have the option to filter out rows that have a certain label while exporting the table.

Generate PDF Report option saves all EICs with their corresponding bar plots in a PDF file.

Export EICs to Json option exports all EICs to a Json file.

Another option is to export the peak data in .mzroll format that can be directly loaded into El-MAVEN by clicking on the Load *Samples|Projects|Peaks* option in the File menu. For this, go to the File option in the menu bar, and click on '*Save Project*'.

Labeled LC-MS Workflow

This is a tutorial for processing Labeled LC/MS data files through El-MAVEN.

Preprocessing

msConvert is a command-line/GUI tool that is used to convert between various mass spectroscopy data formats, developed and maintained by proteoWizard. Raw data files obtained from mass spectrometers need to be converted to certain acceptable formats before processing in El-MAVEN.

Input

msConvert supports the following formats:

- .mzXML
- .mzML
- .RAW ThermoFisher
- .RAW Walters
- .d Agilent
- .wiff ABSciex

The settings used for msConvert as a GUI tool are captured in the following screenshots:

Note: zlib compression is enabled by default in msConvert. El-MAVEN in its current form does not support zlib compression. It is important to uncheck "Use zlib compression" box.

Output

msConvert can convert to an array of different formats but El-MAVEN primarily uses .mzXML and .mzML formats.

Launch El-MAVEN

Once sample files are ready for processing, launch El-MAVEN.

Adjust Global Settings

Global Settings can be changed from the *Options* dialog . There are 9 tabs in the dialog. Each of these tabs has parameters related to a different module in El-MAVEN. For example, a tab for Instrumentation details, a tab for File Import settings etc.

The *m/z* option scans the groups to find any specific m/z value and plot its corresponding EIC. The +/- option to its right is to specify the expected mass resolution error in parts per million (ppm).

To know more about the functionality of different tabs and their settings, users can see the Widgets page. Please be sure to set the desired settings before processing an input file.

Load Samples

Users can go to File in the Menu and click on *Load Samples|Projects|Peaks* option. Then navigate to the folder containing the sample data, select all .mzXML or .mzML files and click Open. A loading bar displays the progress at the bottom.

When the samples have loaded, users should see a sample panel on the left side. If it is not displayed automatically, click on the *Show Samples Widget* button on the toolbar. El-MAVEN automatically assigns a color to every sample. Users can select/deselect any sample by clicking the checkbox on the left of the sample name.

Load Compound Database

Users can click on *Compounds* option in the leftmost menu, navigate to the folder containing the standard database file, select the appropriate .csv file and click *Open*. Alternatively, users may use any of the default files loaded on start-up.

This is a sample Compound Database:

It lists all metabolite names, chemical formula, HMDB ID, and the class/category of compounds they belong to (if known).
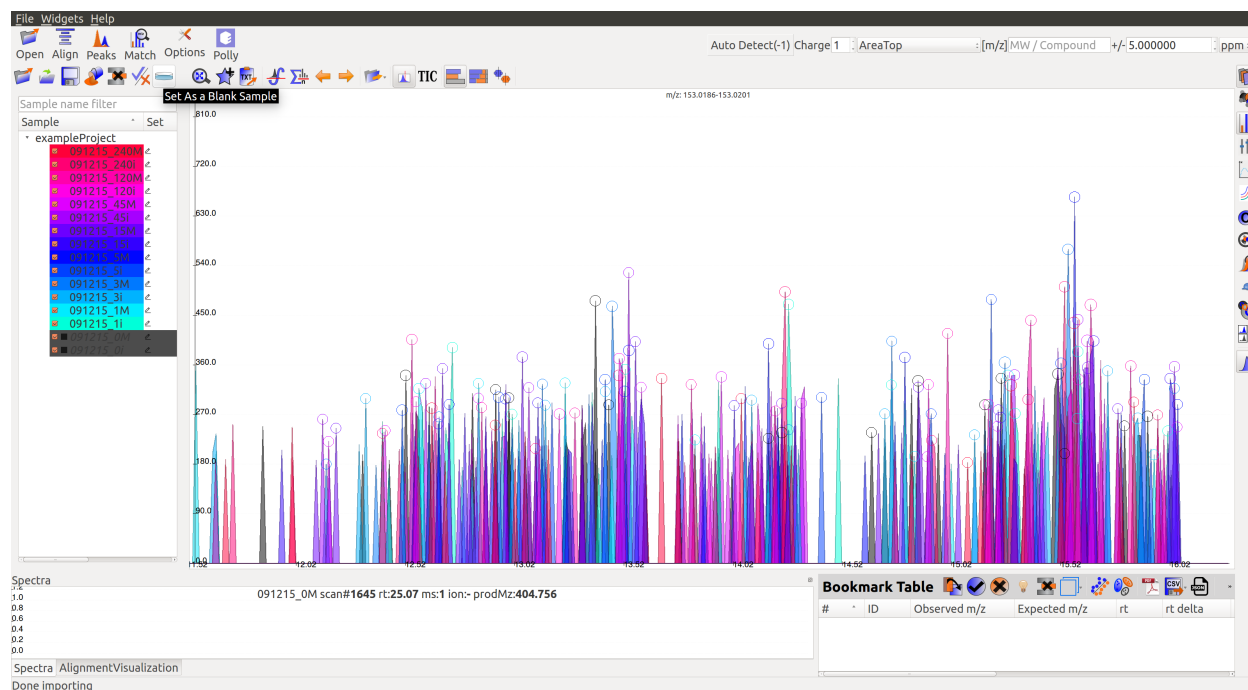
Mark Blanks

Users can mark the blanks by selecting the blank samples from the list on screen, and clicking on the *Set as a Blank*
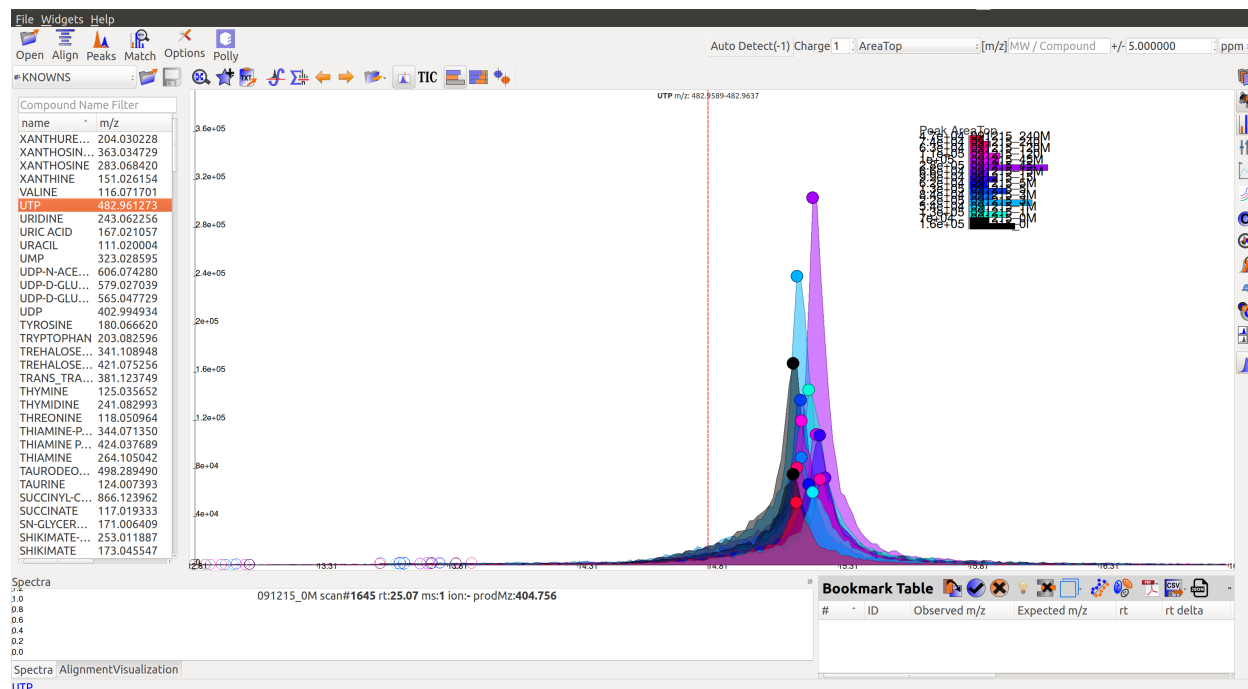
Sample icon in Samples menu.

Multiple blanks can be marked together. The blanks will appear black as shown in the image below:
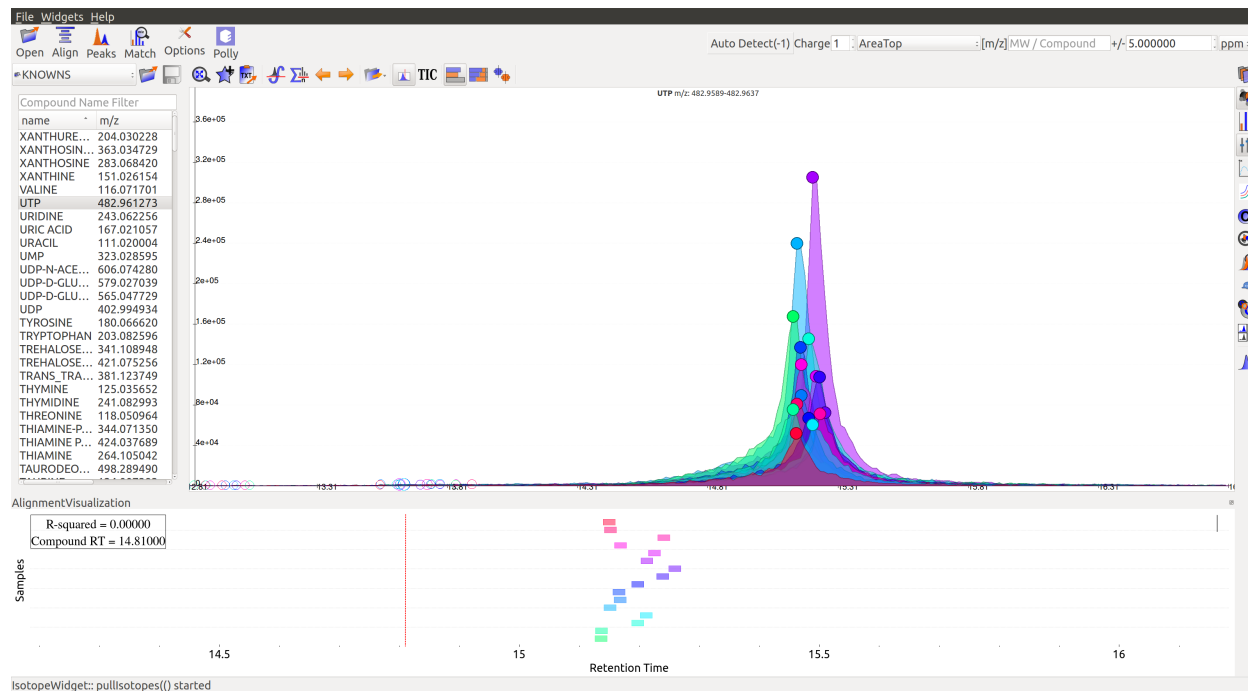

Alignment

Prolonged use of the LC column can lead to a drift in retention time across samples. Alignment shifts the peak retention time in every sample to correct for this drift and brings the peaks closer to median retention time of the group.

In the above image, EIC for a UTP group is displayed. If the samples were aligned, all peaks would lie at the same retention time. Since this is not the case, the samples need to be aligned.
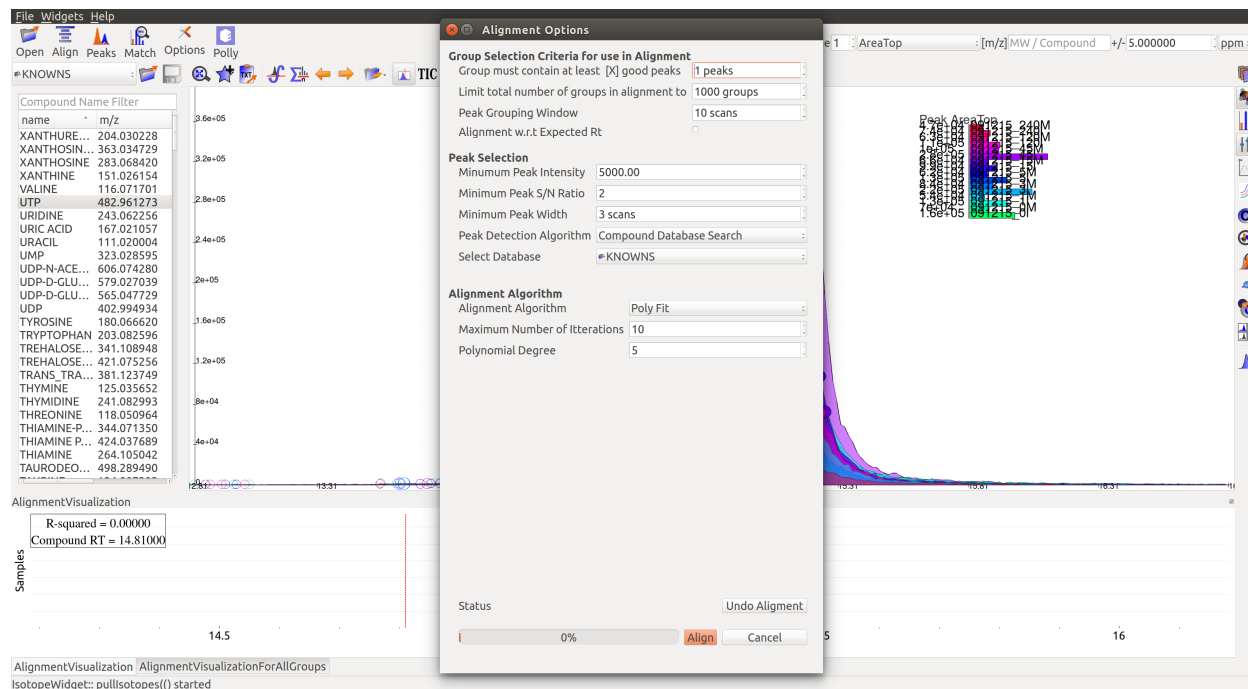
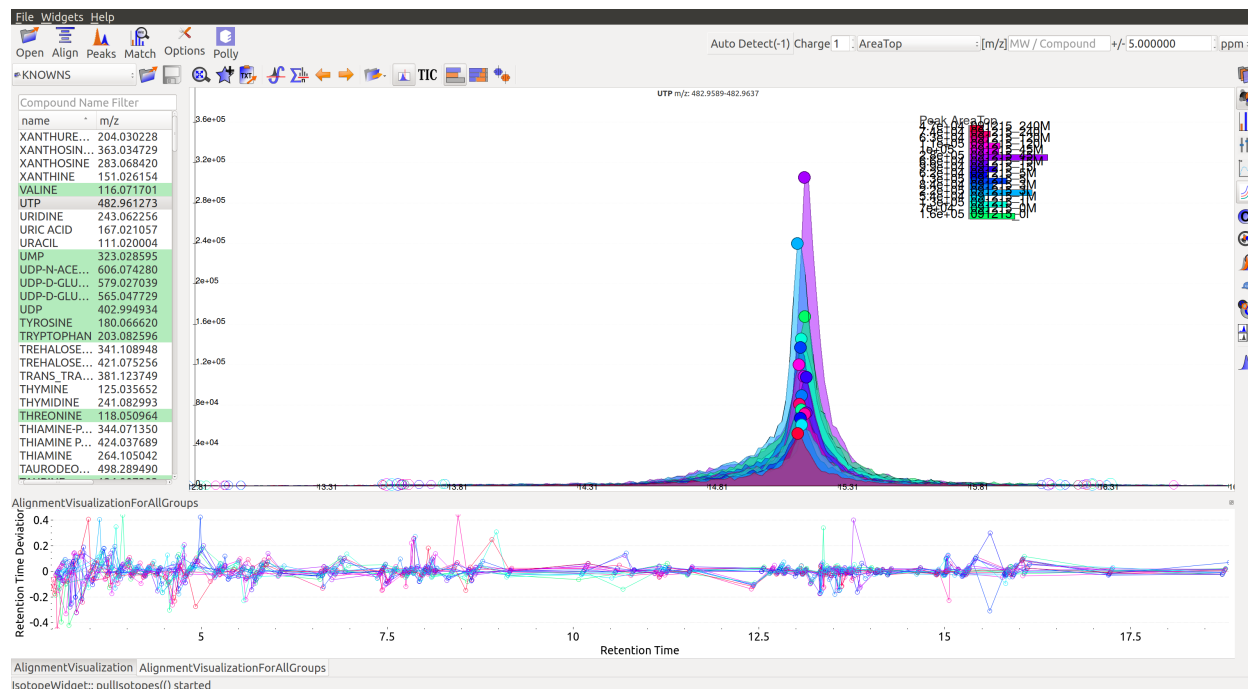*Alignment visualization* can be used to judge the extent of deviation from median retention time.

In the above visualization, each box represents a peak from the selected group at its current retention time. Samples are said to be perfectly aligned when all peak boxes lie on the same vertical axis. The peaks are considerably scattered in the above image and therefore the samples should be aligned for better grouping of peaks.
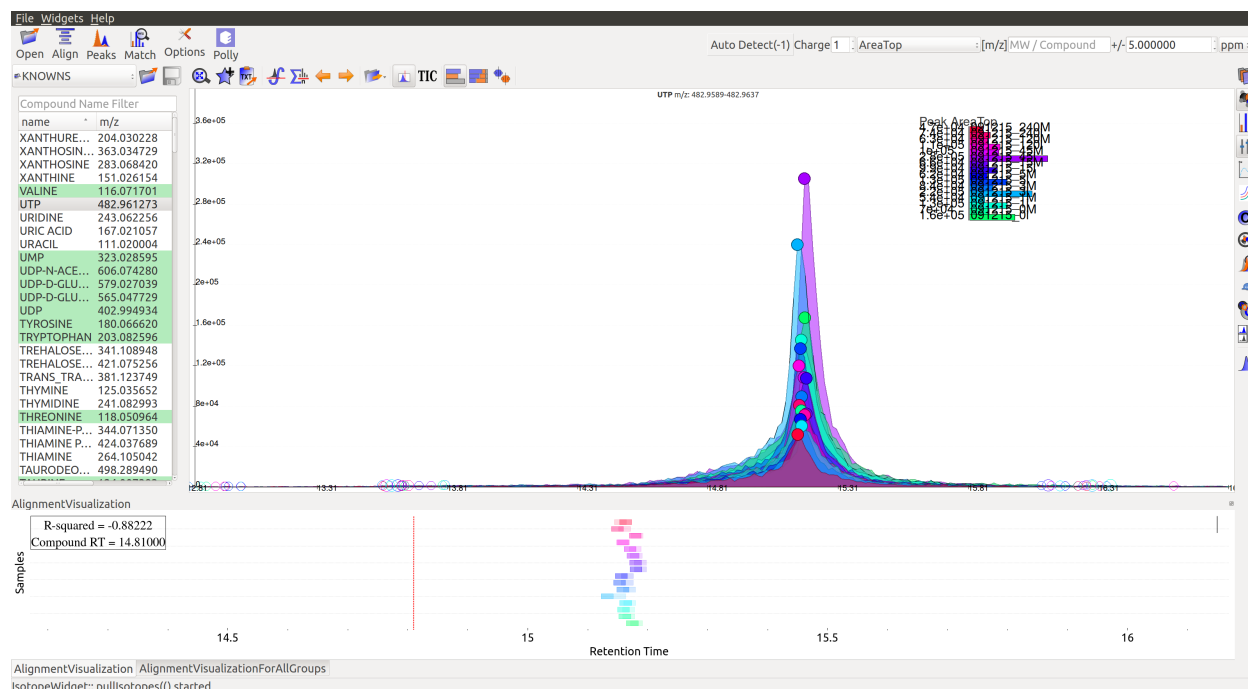
Perform Alignment

Alignment settings can be adjusted using the Align button [image]. This example was aligned with *Poly fit* algorithm with default parameters.

Post-alignment the peaks in the group should appear closer to the median retention time of the group.
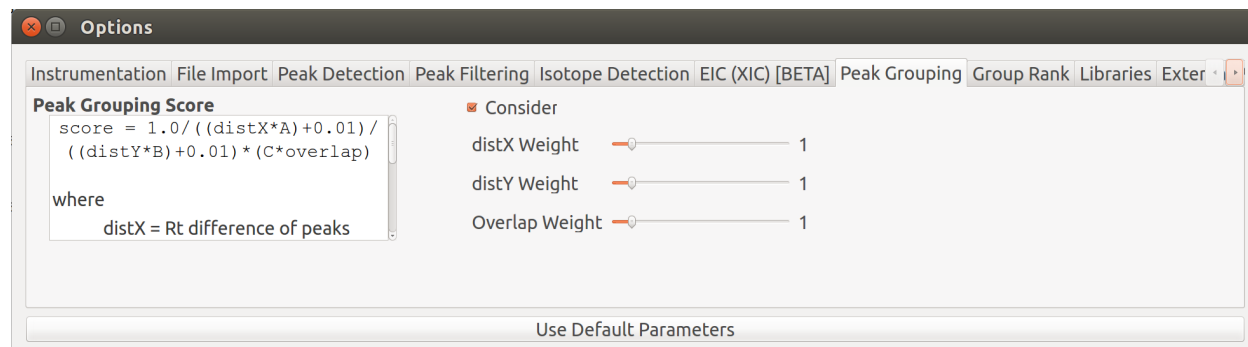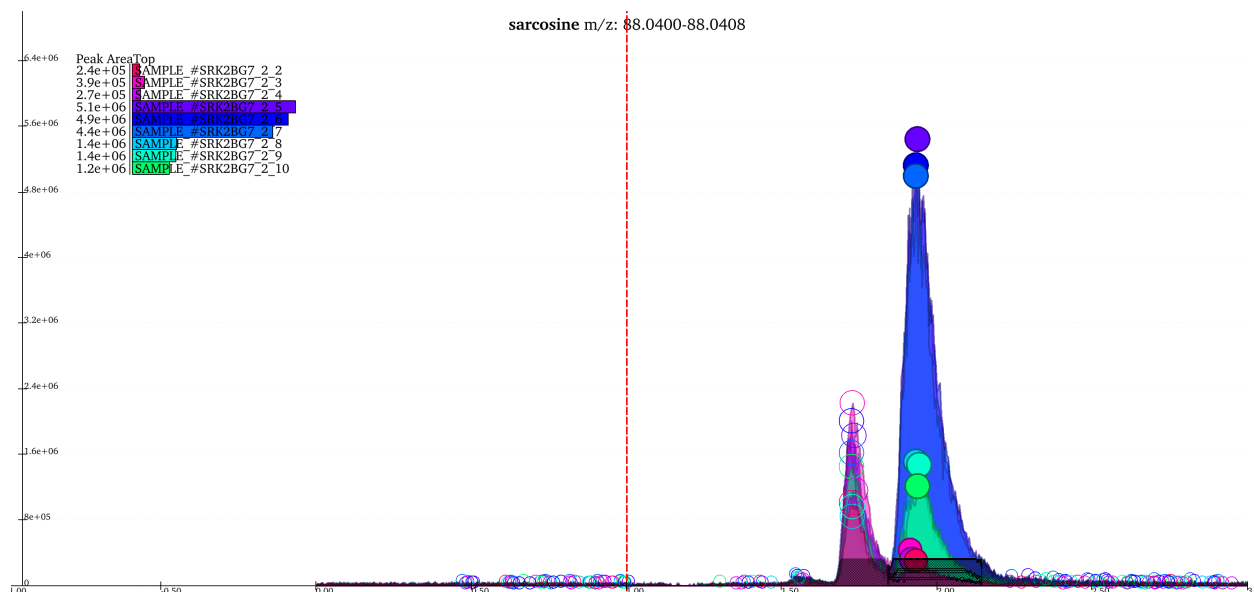
Pre-alignment, the peaks were considerably scattered while the aligned peaks lie nearly on the same axis. Users can run alignment again with different parameters if required (or with a different algorithm). Further details on Alignment settings are available on the Widgets page.

Peak Grouping

Peak grouping is an integral part of the El-MAVEN workflow that categorizes all detected peaks into groups on the basis of certain user-controlled parameters. A group score is calculated for every peak during the process. The formula for this score takes into account the difference in retention time, intensities between peaks (smaller difference leads to a better score) and any existing overlap between them (higher extent of overlap leads to better score). All three parameters have certain weights attached to them that can be controlled by the users. The formula for the score is shown in the image. More details on it can be found on the Widgets page.

The above image shows two groups in the EIC window. The highlighted (solid circles) peaks belong to group A, the peaks to its left with empty circles belong to another group B. The short peaks in group A that are close to the baseline and peaks in group B come from the same samples. Additionally, the high intensity peaks of group A have a similar peak shape to group B peaks. These peaks might have been wrongly classified into separate groups because of the difference in retention time range of the two sets of peaks. The weights attached to difference in retention time and intensities, and extent of overlap can be adjusted for better grouping.

Grouping parameters can be changed from the Options dialog .

Giving less priority to difference in retention time and intensities results in the two groups being merged into a single group while the peaks that lay close to the baseline are no longer classified as valid peaks.

Baseline

When measuring a number of peaks, it is often more effective to subtract an estimated baseline from the data. This baseline should be set where ideally no peaks occur. Although sometimes the program sets a particular baseline such that one or more peaks occur below that baseline value. In the following image, the dashed line represents each baseline:

The corresponding peaks are indicated with solid circles:

The baseline correction can be done in the *Peak Detection* tab by clicking on *Options* button:
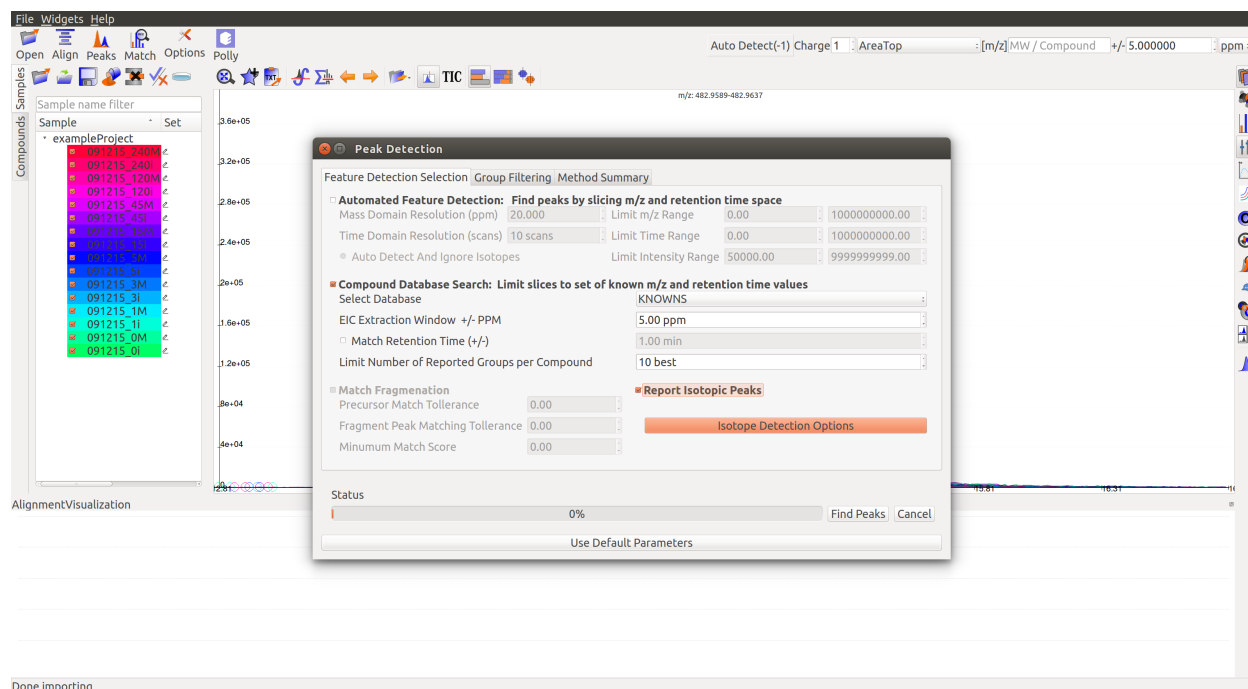

Further details on settings can be accessed here.
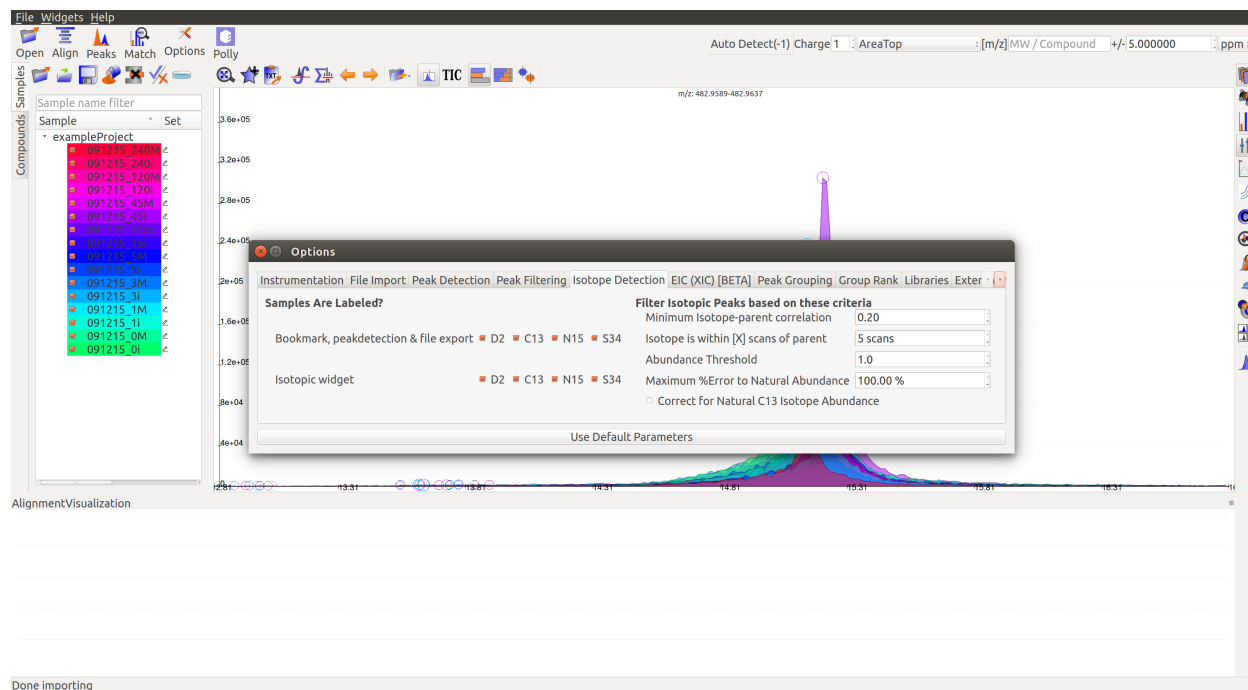
Isotope Detection

Samples are labeled?

The *Peaks* dialog can be used to detect labeled peaks along with the unlabeled ones in the Peaks Table.

On opening the *Feature Detection Selection* tab, the *Report Isotopic Peaks* box must be checked. Clicking on the *Isotope Detection Options* gives the following window. Alternately, these settings can also be accessed from the *Options* dialog.


- *Bookmarks, peak detection, file export*: To select the labeled atoms that should be used in bookmarking, peak detection and export. D2: Deuterium, C13: Labeled carbon, N15: Labeled nitrogen, S34: Labeled sulphur.

- *Isotopic widget*: To select the labeled atoms that should be displayed in the isotopic widget. D2: Deuterium, C13: Labeled carbon, N15: Labeled nitrogen, S34: Labeled sulphur.

- *Number of M+n isotopes*: To set the maximum number of labeled atoms per ion in the experiment.

- *Abundance Threshold*: To set the minimum threshold for isotopic abundance. Isotopic abundance is the ratio of intensity of isotopic peak over the parent peak.

Filter Isotopic Peaks based on these criteria

- *Minimum Isotope-Parent Correlation*: To set the minimum threshold for isotope-parent peak correlation. This correlation is a measure of how often they appear together.

- *Isotope is within [X] scans of parent*: To set the maximum scan difference between isotopic and parent peaks. This is a measure of how closely they appear together on the retention time scale.

- *Maximum % Error to Natural Abundance*: To set the maximum natural abundance error expected. Natural abundance of an isotope is the expected ratio of amount of isotope over the amount of parent molecule in nature. Error is the difference between observed and natural abundance as a fraction of natural abundance.

- *Correct for Natural C13 Isotope Abundance*: The box should be checked to correct for natural C13 abundance.

In the image below, Peak Table 3 has a drop down button with metabolites that shows all labeled isotopologues of that particular metabolite.
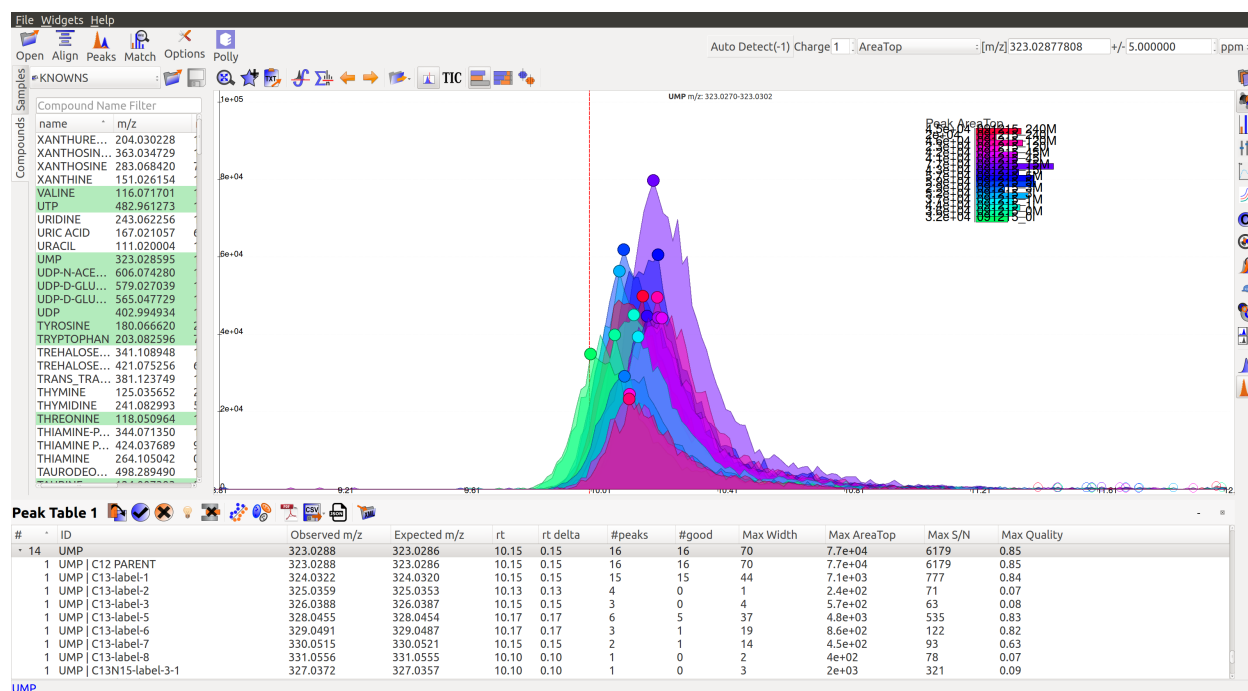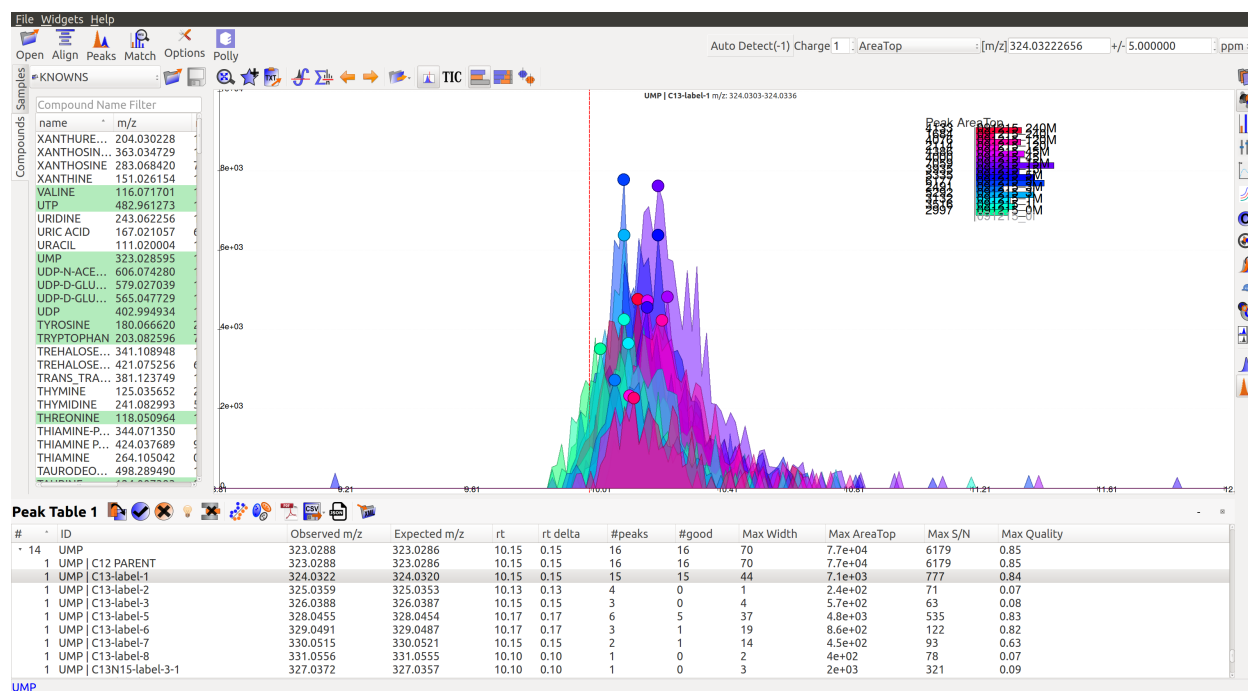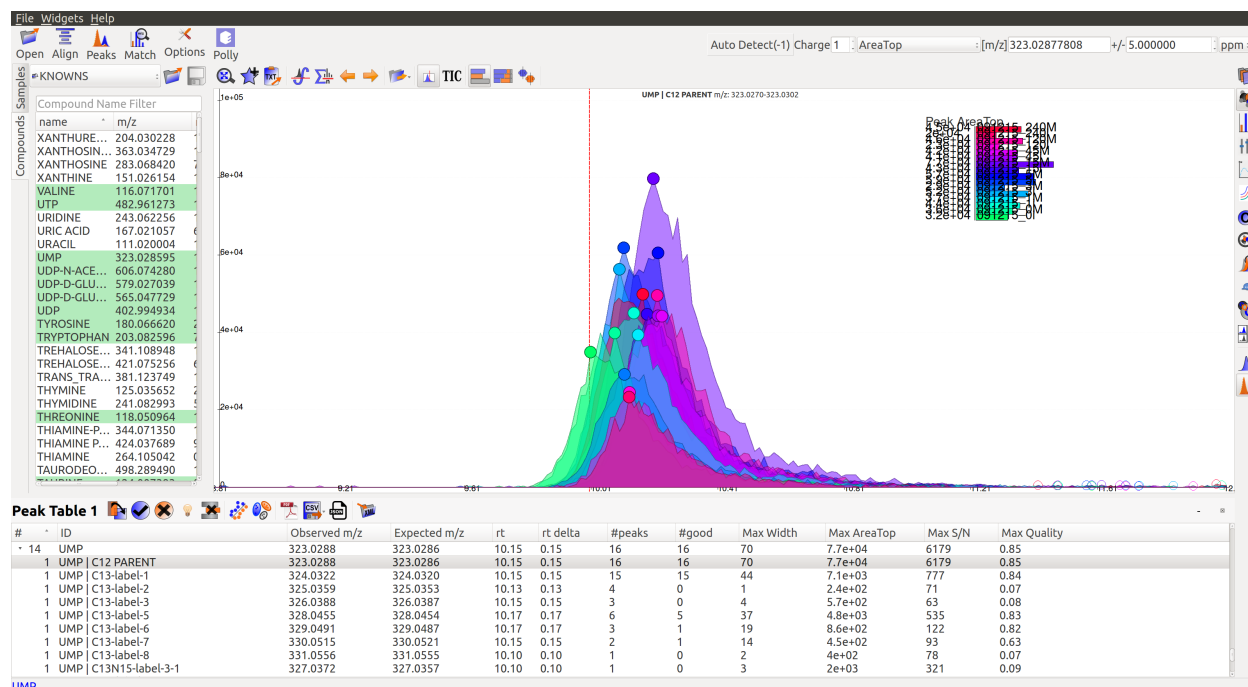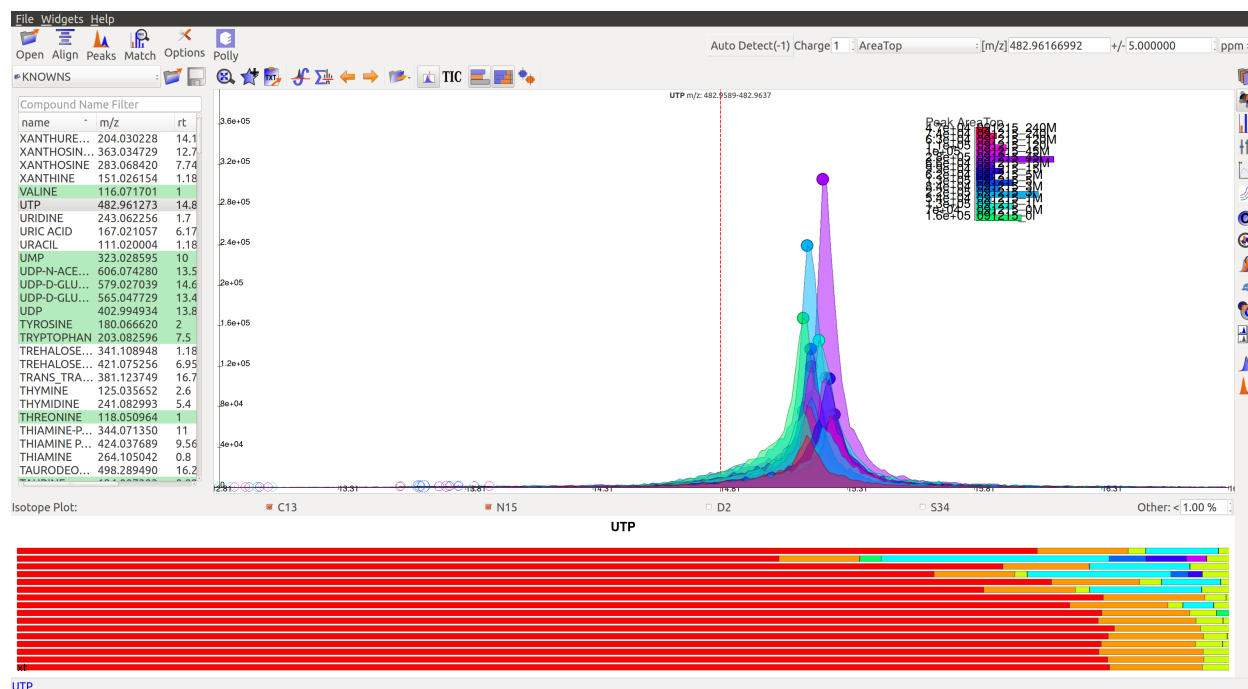
Show Isotope Plots

This icon on top displays the isotope plots for a group. The red colored bar plot for UTP group is shown below. Each bar in the plot represents the relative percentage of different isotopic species for the selected group in a sample.

Mass Spectra

Mass Spectra Widget displays each peak, its mass, and intensity for a scan. As the widget shows all detected masses in a scan, the ppm window for the EIC and consequently grouping can be adjusted accordingly. This feature is especially useful for MS/MS data and isotopic detection.
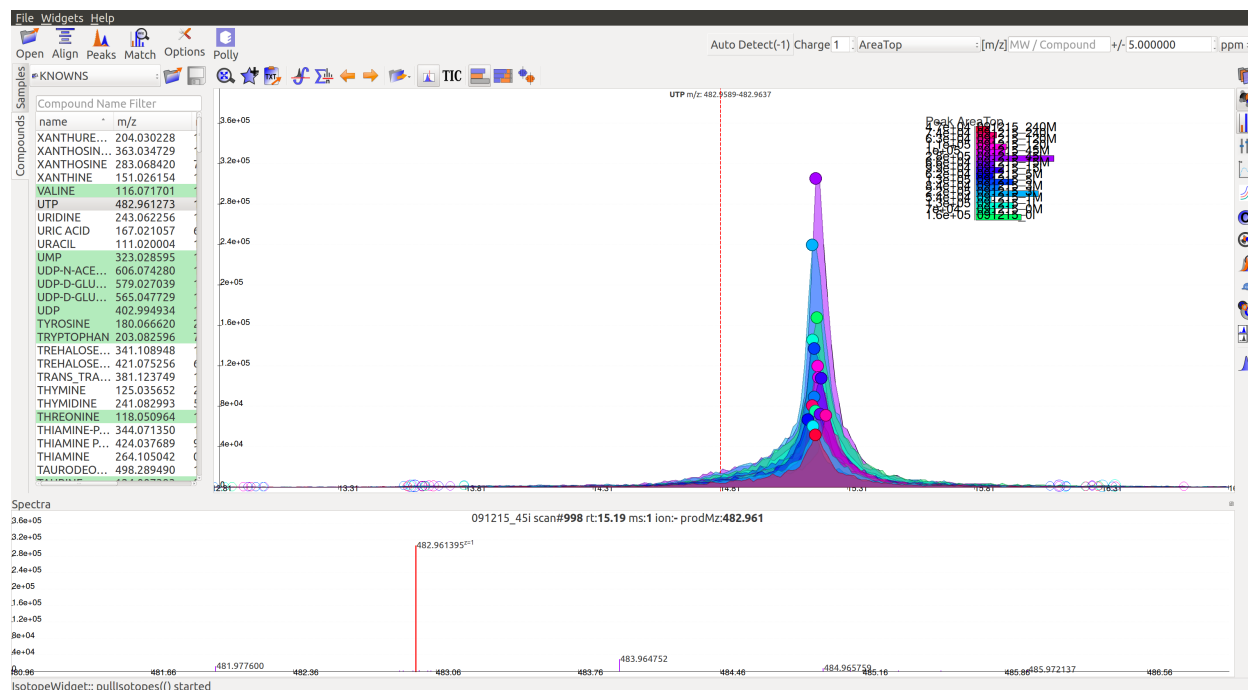
Peak Curation

Generally there are two broad workflows to curate peaks in El-MAVEN:

- Manual Peak Curation using Compound DB widget
- Automated Peak Curation

Although, for labeled data Automatic Peak Curation is not meaningful because it will not curate any labeled groups.

Manual Peak Curation using Compound DB widget

Clicking the *Peaks* icon on the top opens the settings dialog.
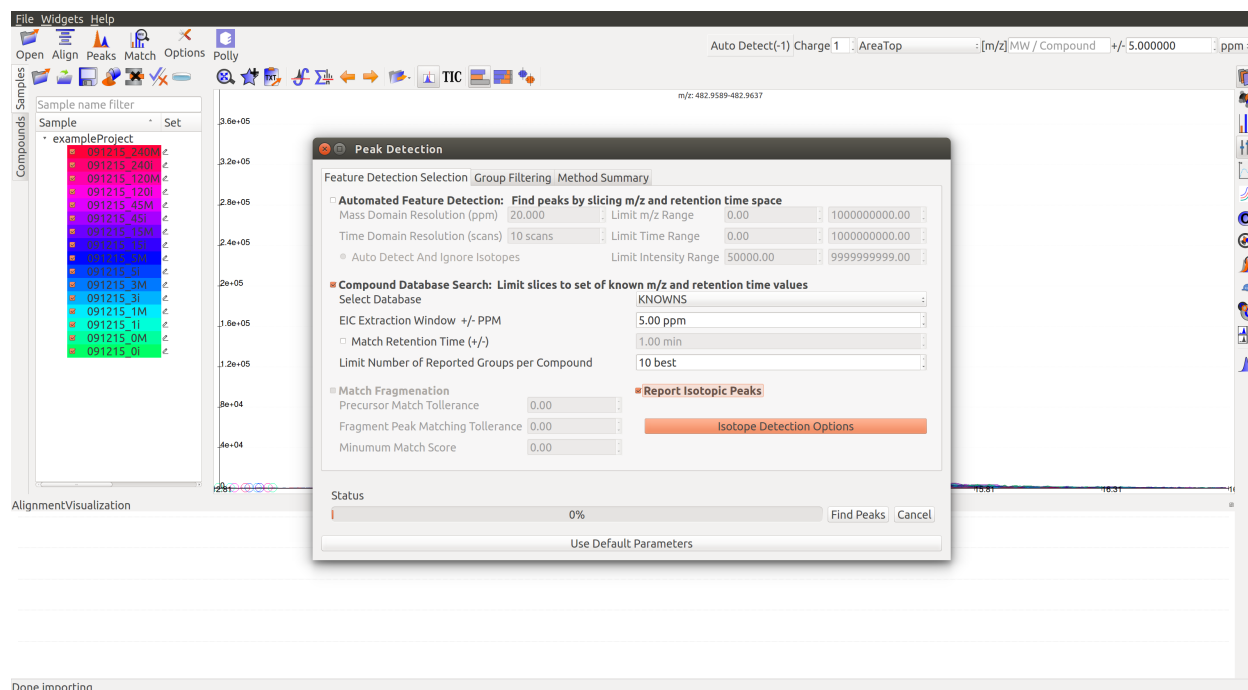
Users must check the box for *Report Isotopic Peaks* in the *Group Filtering* tab.

Note: Users should not click on *Find Peaks* after checking the box for manual curation. Clicking on that option would

start Automatic Peak Detection. For adjusting other settings, users can access them through the Options dialog .

For more details on how to access Peak Detection settings, read this Widgets page.

To use manual curation using the compound DB widget, users have to iterate over all the compounds in the compound DB on the extreme left of the window, as highlighted in the images below.
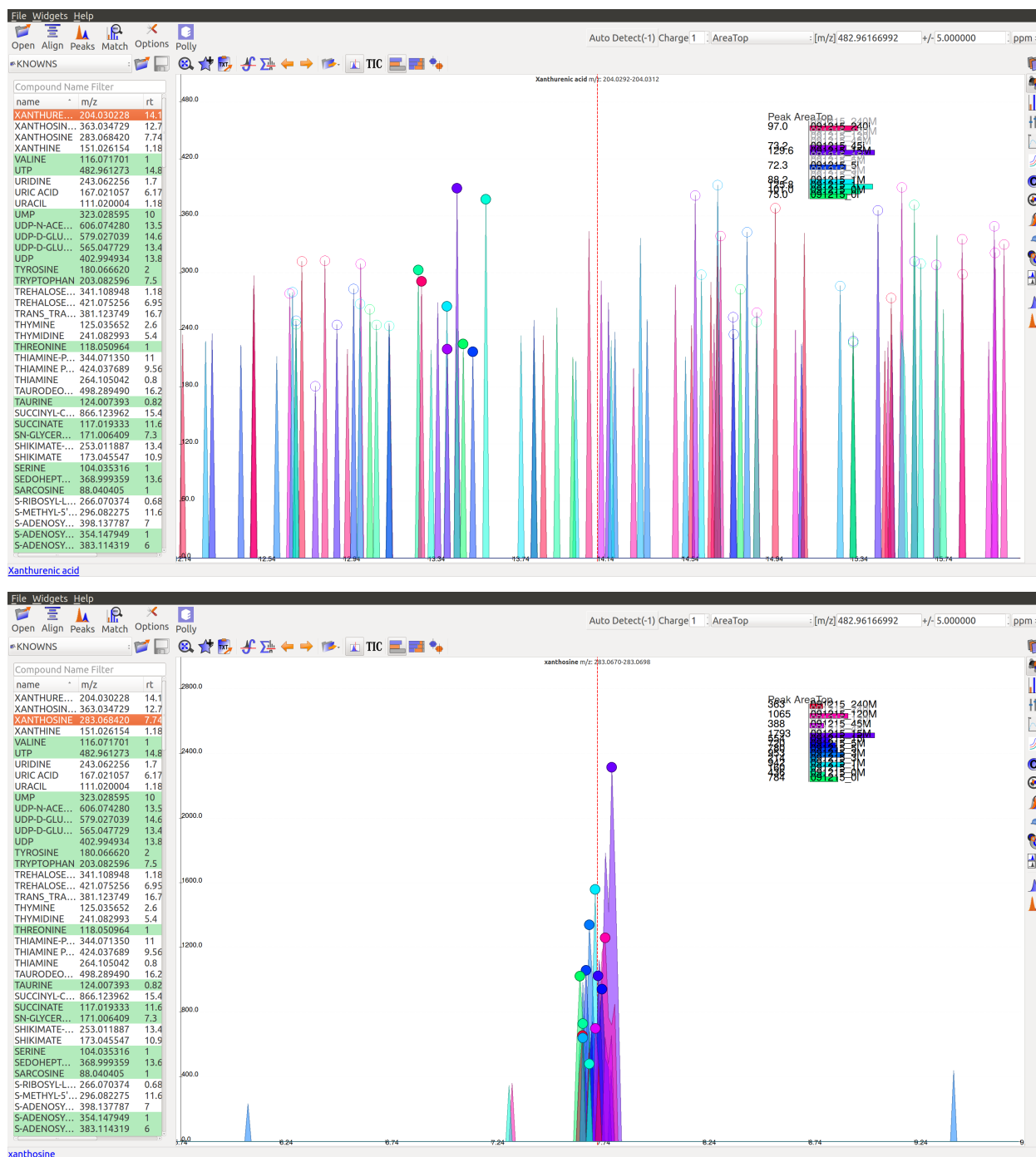
Once on a compound, El-MAVEN shows the highest ranked group for that m/z. Users can now choose a group or reject it. There are two ways to do this.

In the first workflow, users need to double click on the peak group of his choice. This will get the retention time line to the median of the group and also add the metabolite to the bookmarks table (as shown in the image below). Users can read more about the bookmarks table here.
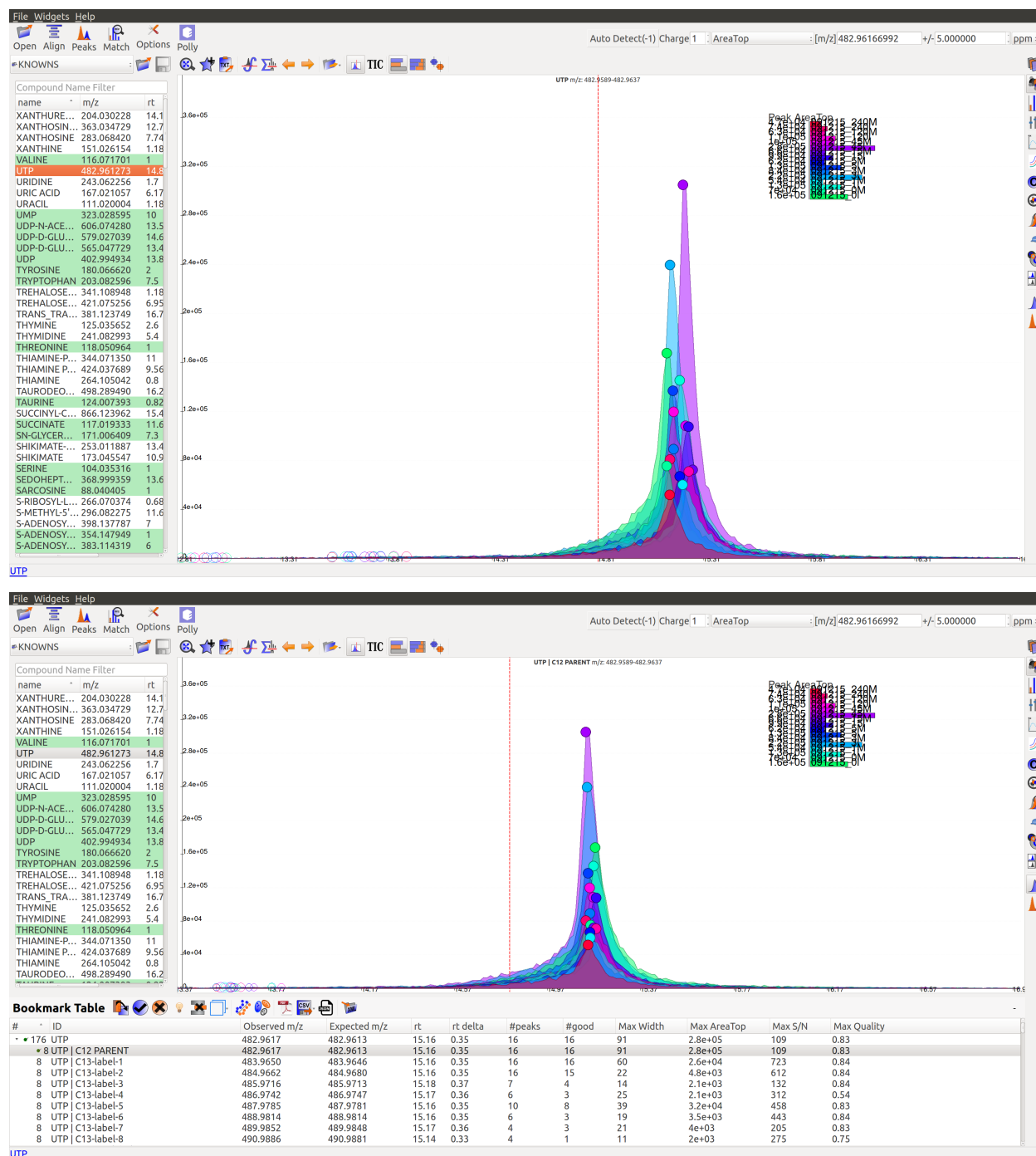
When the users select the first group they would be asked if they would like to auto-save the state of the application. This feature allows the users to go back to their curated peaks if they so wishes in future.
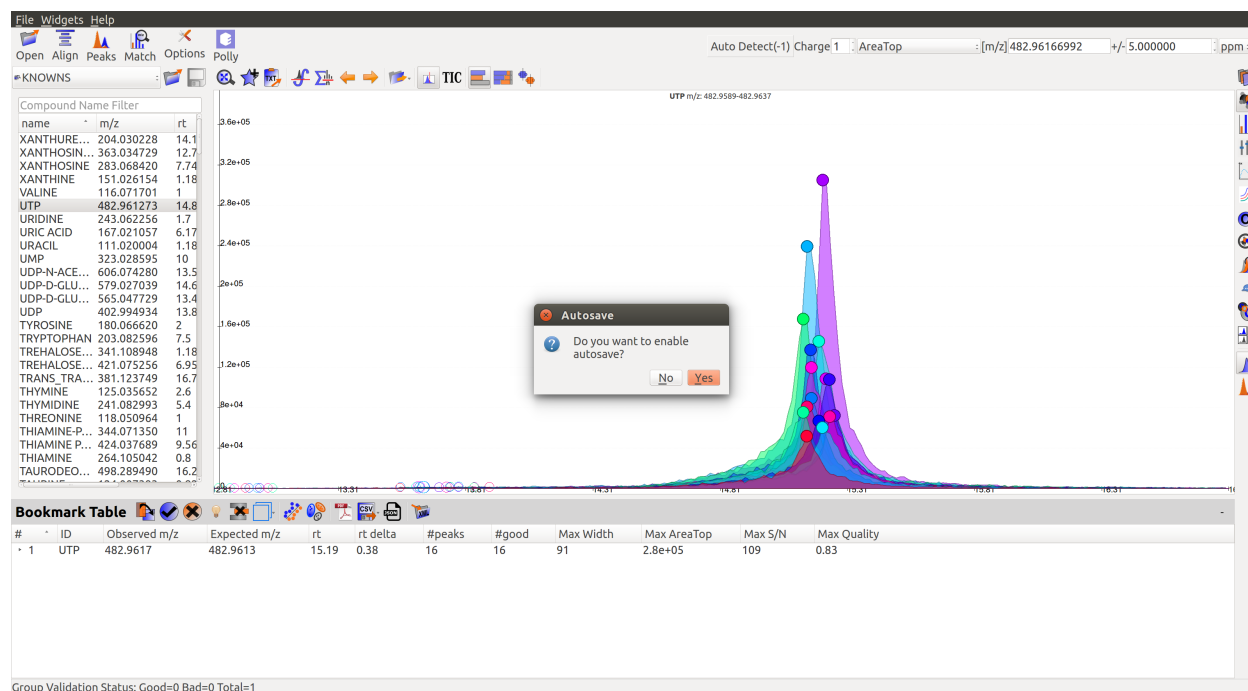
Users can then use dropdown arrow on bookmarked group to mark all its isotopologues as good or bad.
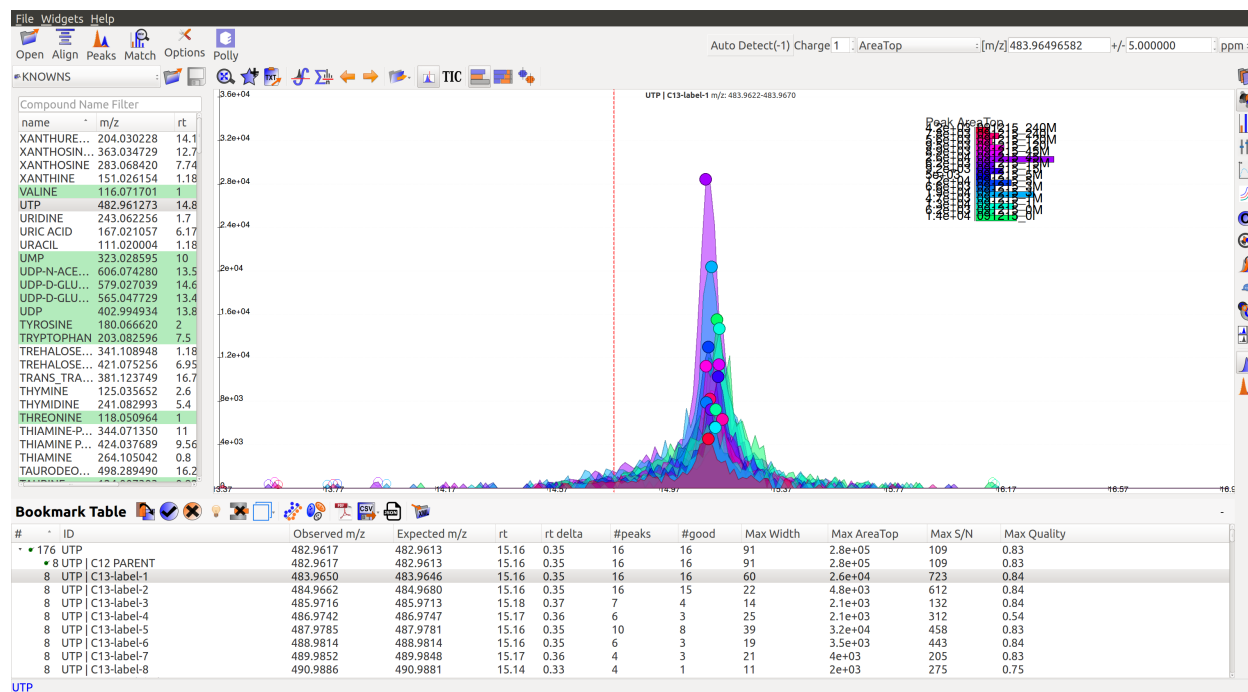
After marking all the groups for a compound, users can scroll down to the next compound and decide on the basis of shown EIC if the group should be marked for curation.
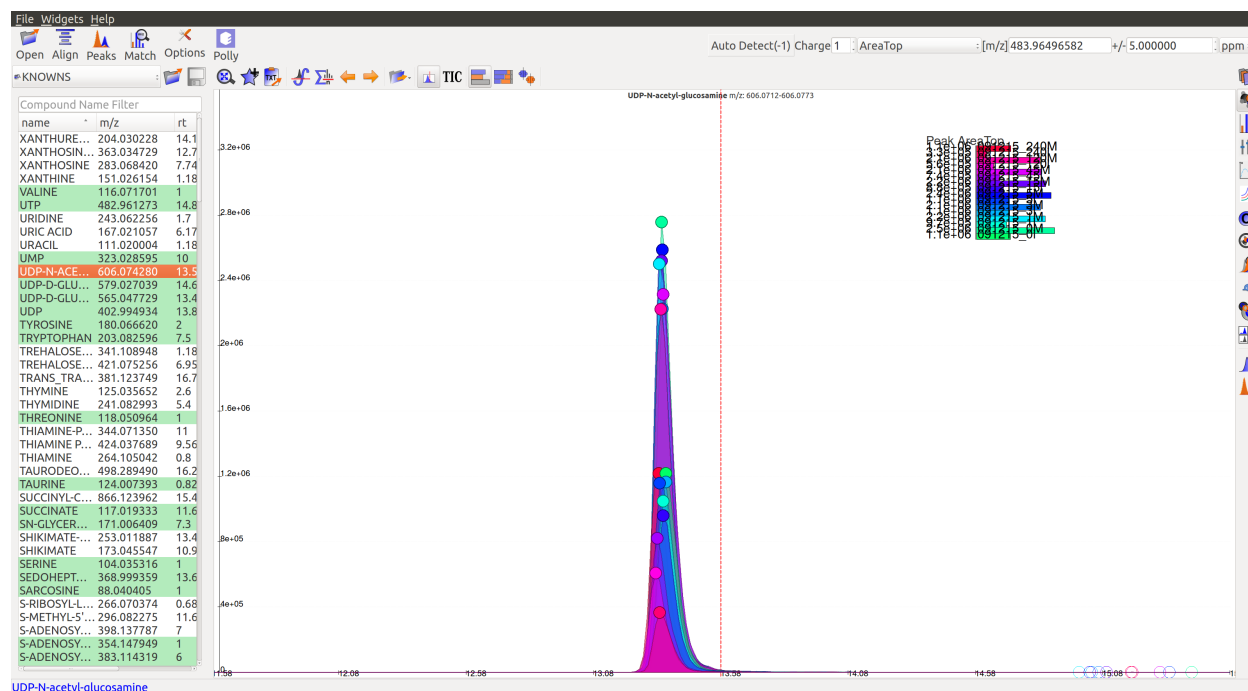
Double clicking on any peak (solid coloured circle) moves the retention time line along the group. And the group moves to the bookmark table.
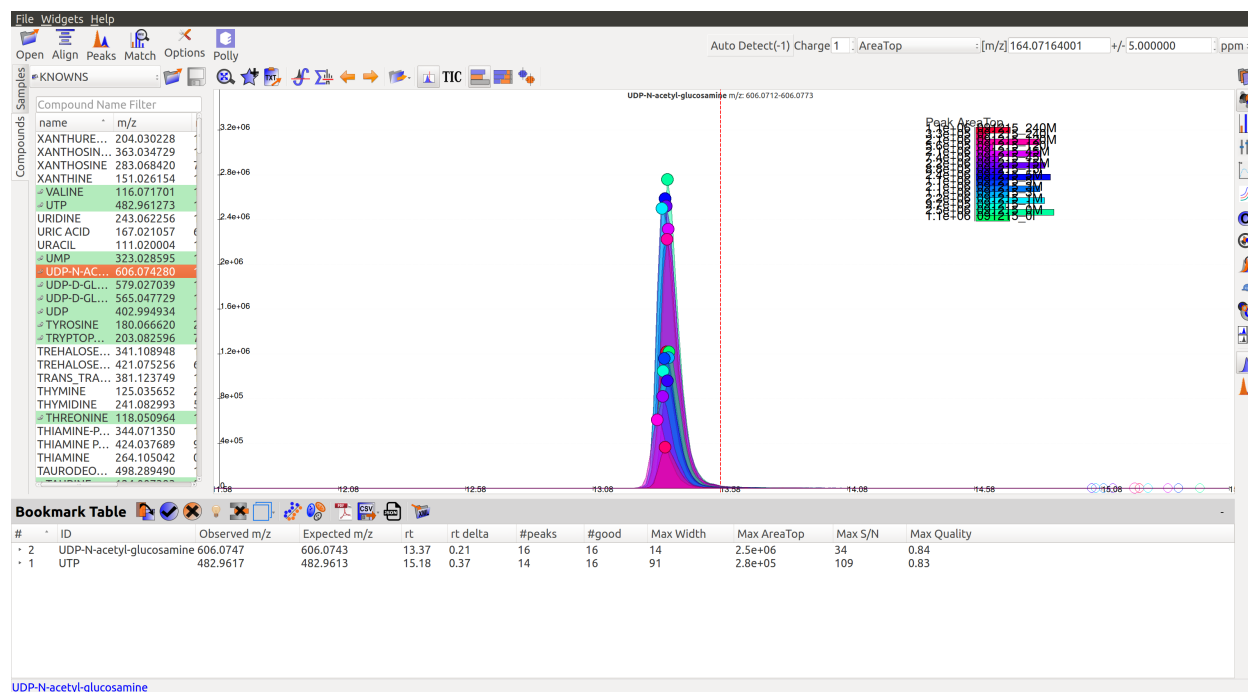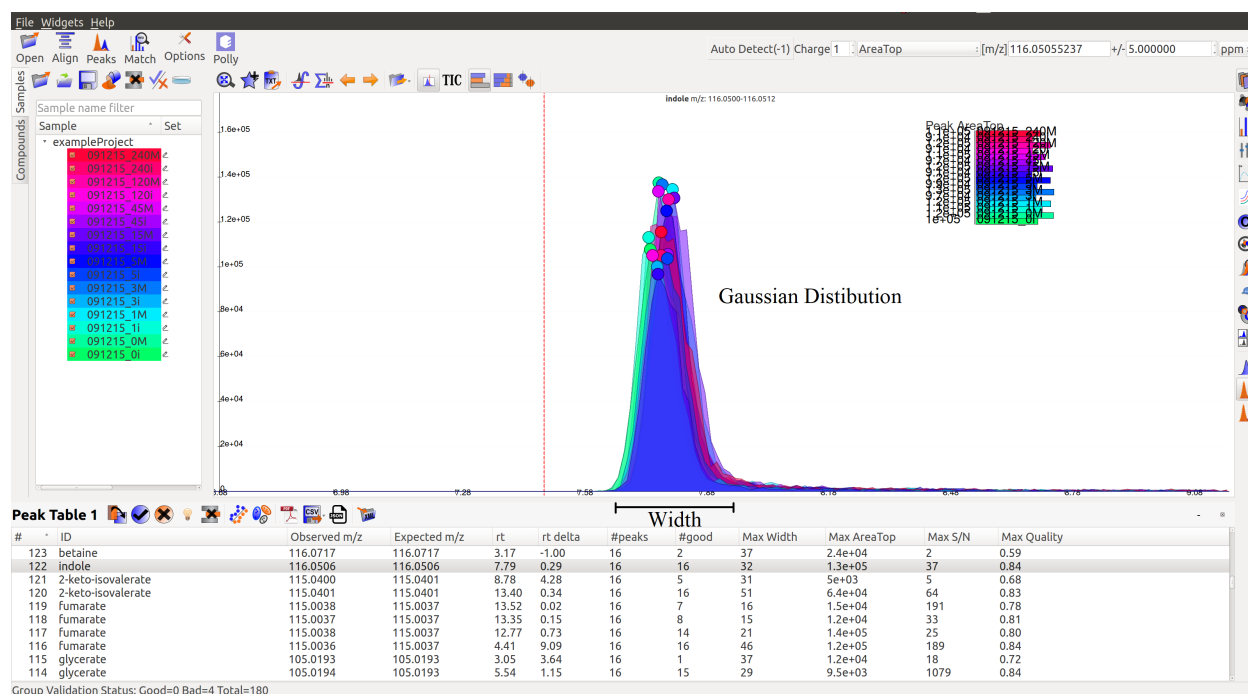

Qualifying peaks as good or bad is explained in the next section.

Guidelines for Peak Picking

- A peak's width and shape are two very crucial things to look at while classifying a peak as good or bad. The peak's shape should have a Gaussian distribution and width should not be spread across a wide range of retention

time.

- Peak Intensities for a group are plotted as bar plots for all the samples. These bar plots have heights relative to the other samples.Thus, for a good peak the intensities should be high.
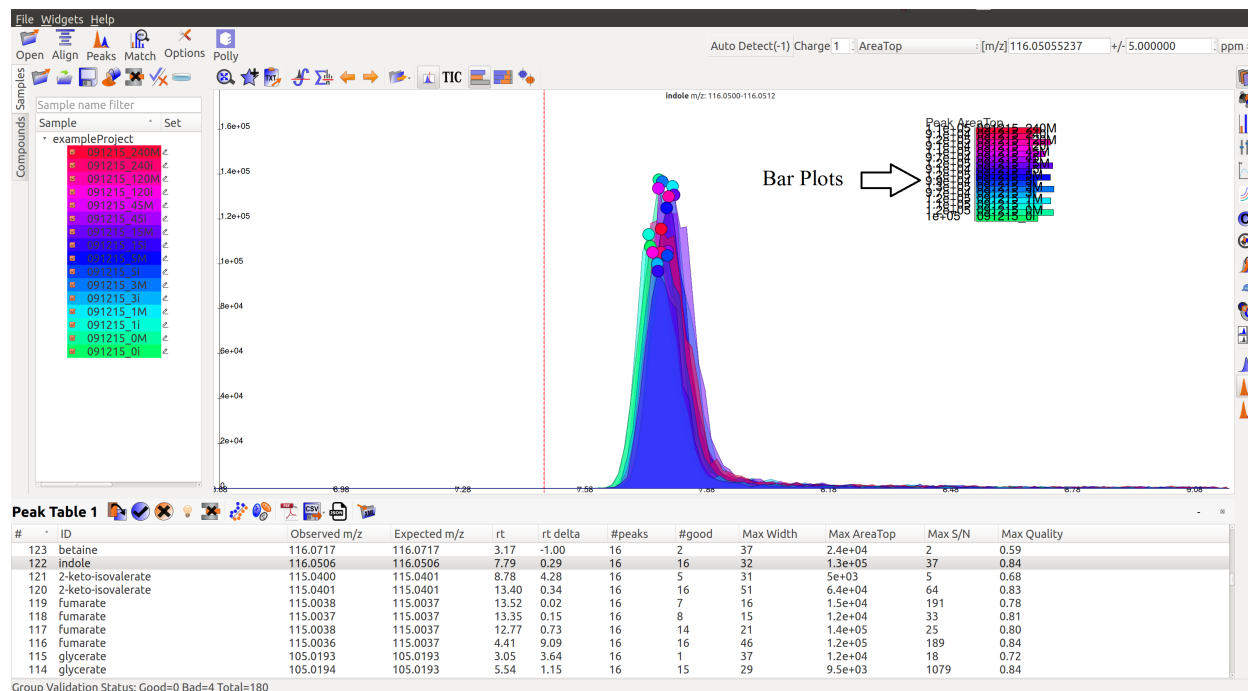

- Intensity Barplot heights should be higher for all the samples than Blank samples, as shown above. We use intensities of Blank samples to set our group baseline. Blank intensities are used to calibrate intensity values across zero concentration.

- A good peak should have standards with varying intensity in a particular fashion (increasing or decreasing).

- Quality Control (QC) samples give us information about the quality of the data, i.e., it assesses reproducibility

and software performance. Samples whose intensities and concentrations are already known are used as QCs to determine if the instrument are working as expected. Values (and scales) can be calibrated using QCs.

- If peak groups of a particular metabolite are separated apart (not aligned well) then we should use stringent alignment parameters to overcome this problem.

- For a particular metabolite, let's say if it has n number of groups, then the group which is much closer to the above guidelines should be selected as good peak. Multiple groups can also be selected in case of ambiguity (if retention time information is not provided).

**A good peak would look similar to the following peaks:**
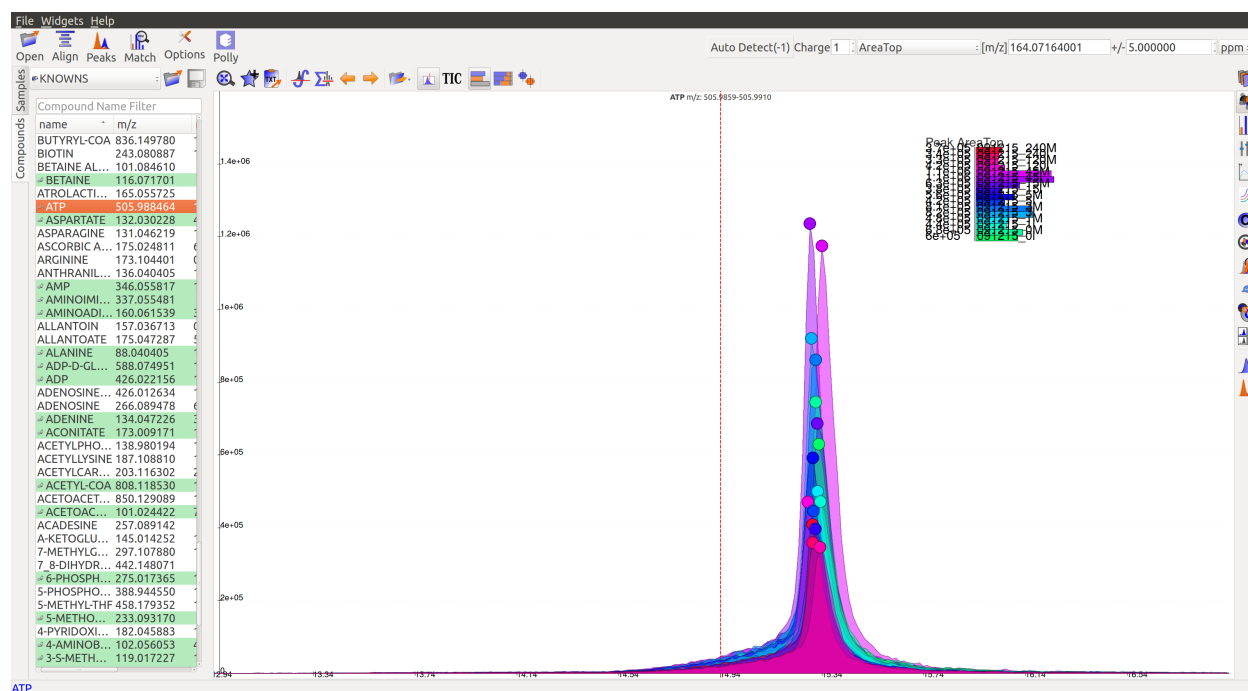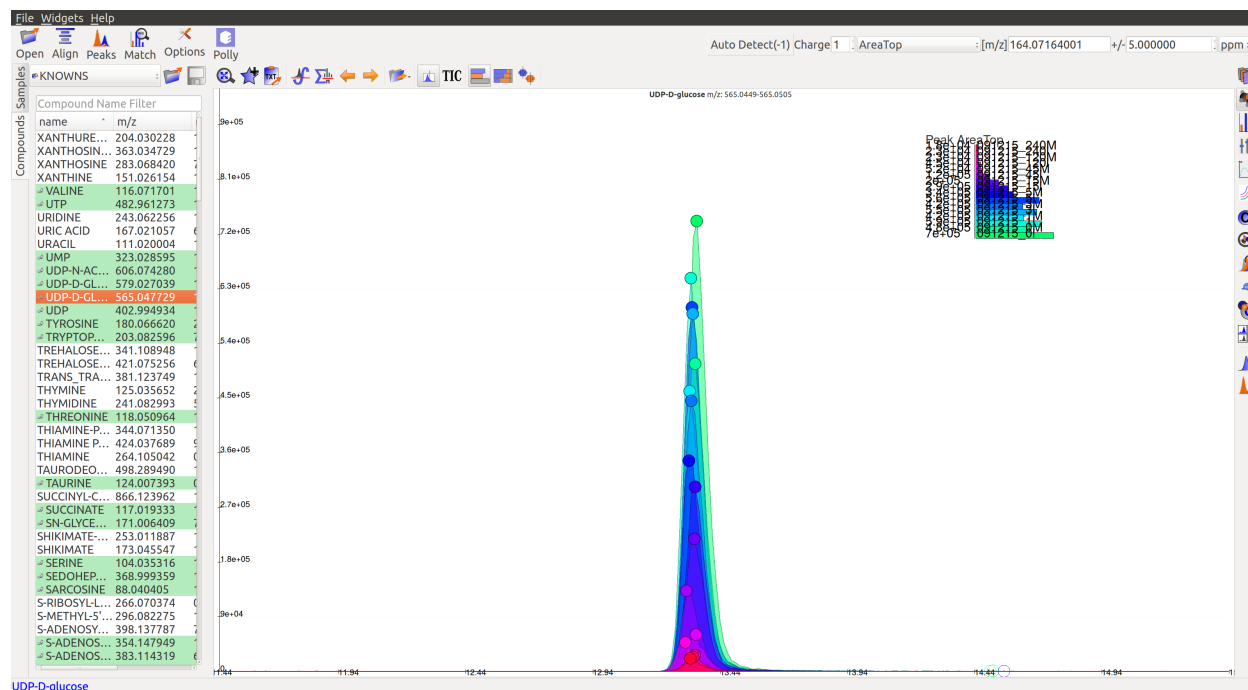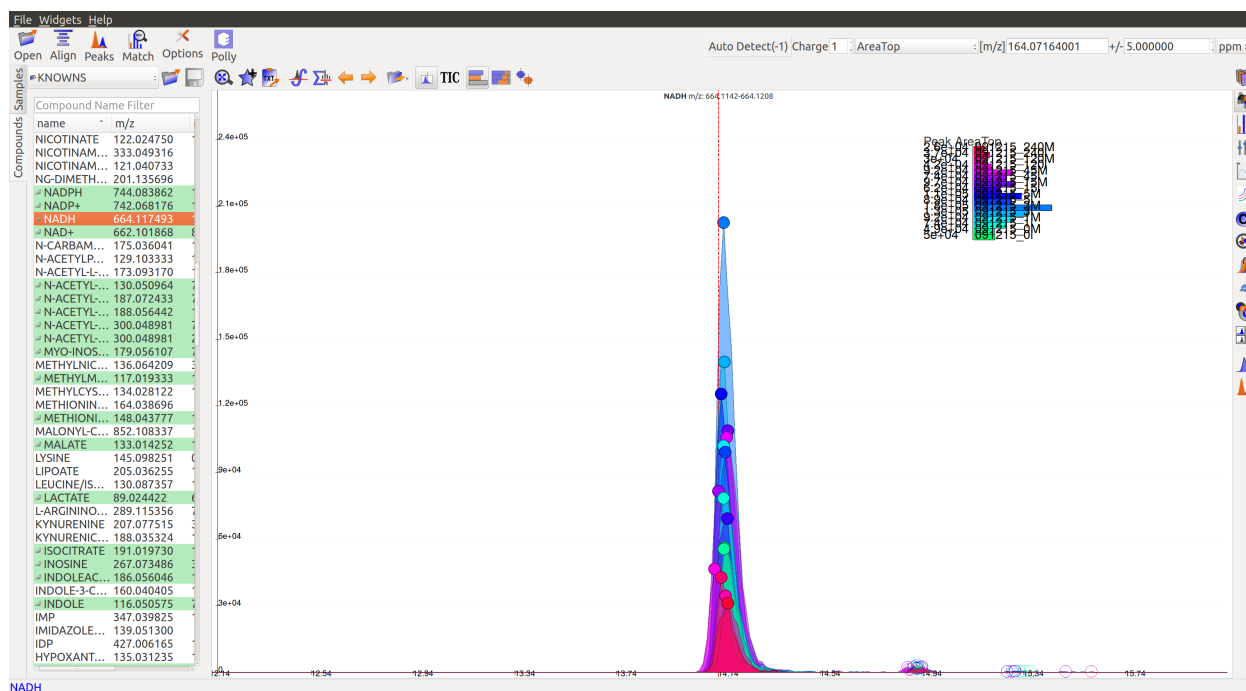
- Gaussian shape

- Perfect grouping, narrow retention time

- Good sample intensities

- Low blank intensities

- QCs look good

- An observable trend in intensity bars of standards, as well as samples.

**Some examples of bad peaks are given below:**

- The peaks have a good Gaussian shape. But the blank intensity bars are high. All the sample intensity bars are shorter or roughly equal to the blank intensities, implying the peaks are most likely noise.
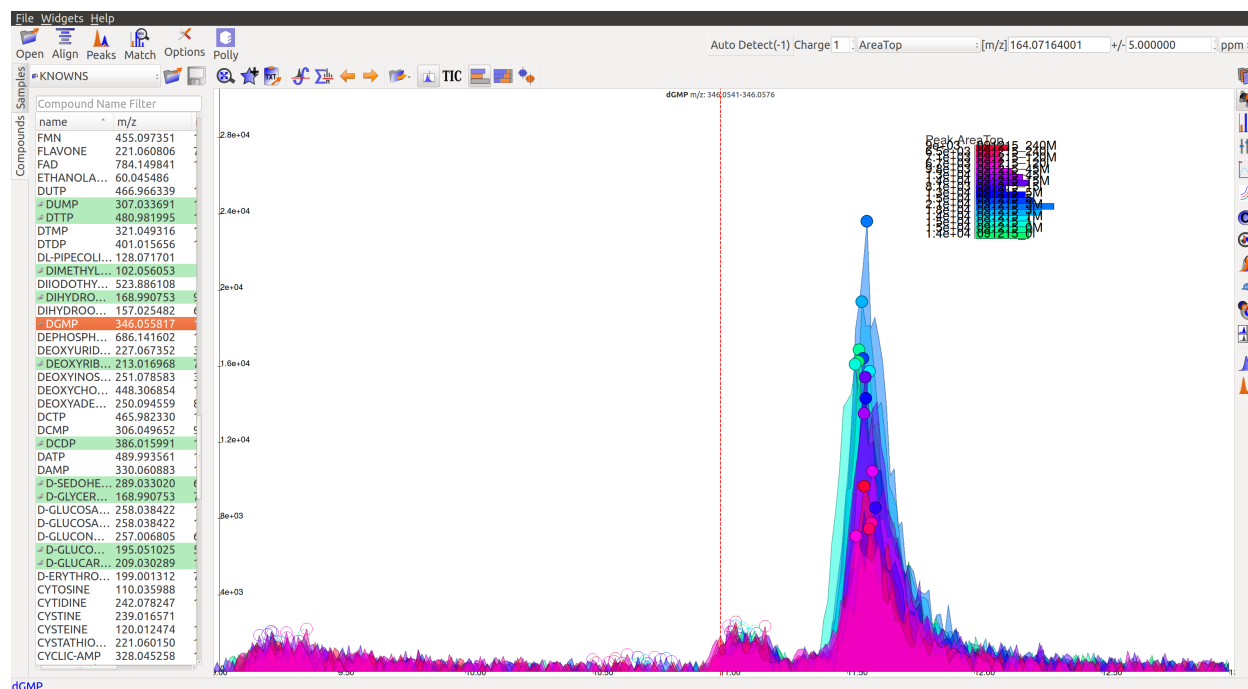
- The intensity levels are low relatively. The peaks are spread over a long range of retention time. They have poor shape, poor grouping and lie close to noise. If the signal to noise ratio was improved, this peak would probably not be detected.
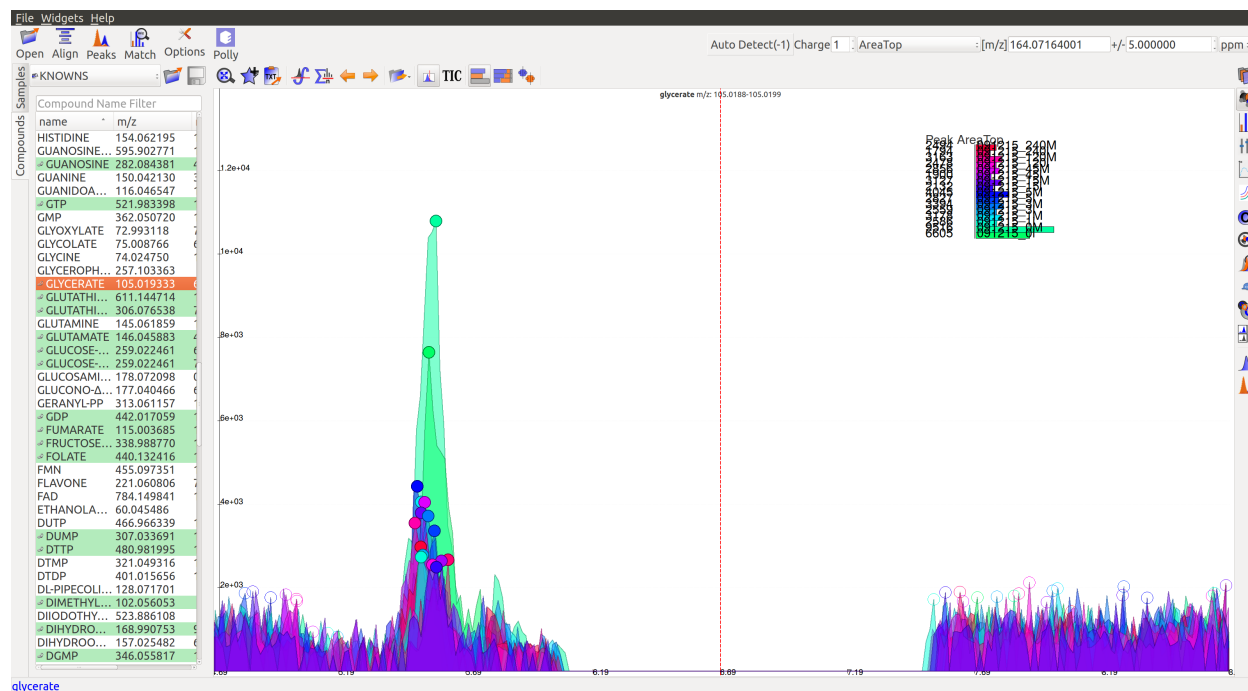

- In the following image, many sample intensities are missing from the intensities bar plots. Peaks do not have a Gaussian shape, nor good grouping. These peaks are probably noise which have been wrongly annotated.
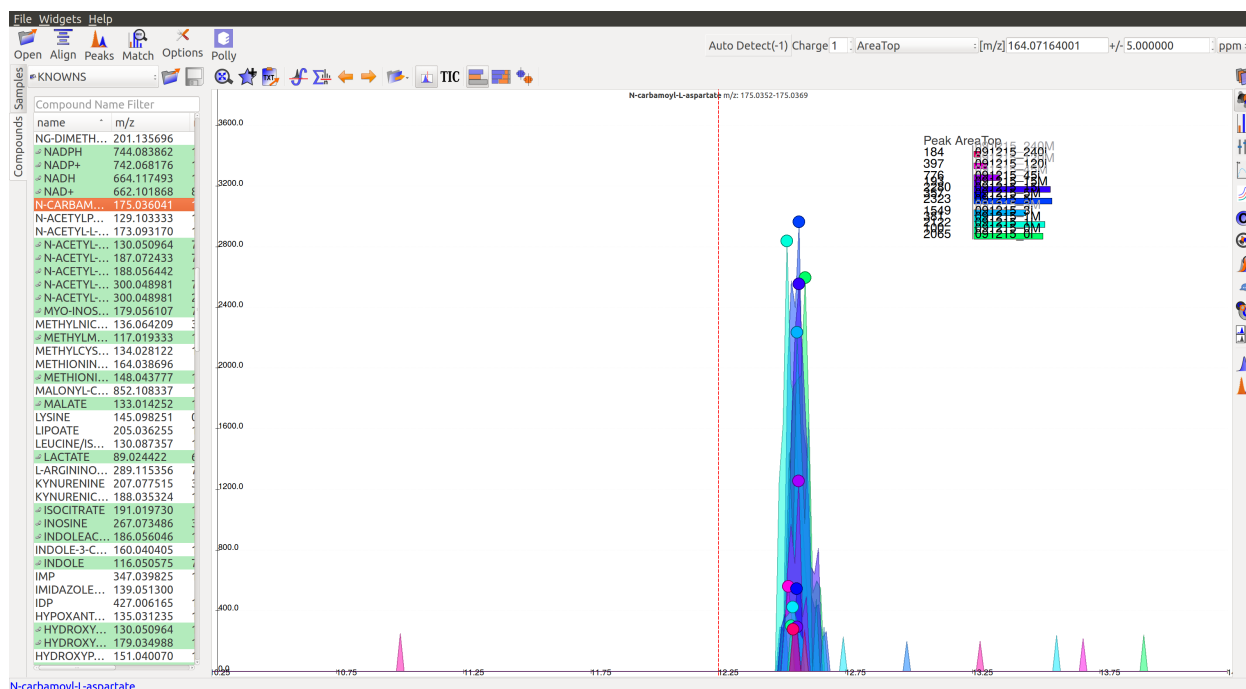
- This is a noisy group. There are no discrete peaks visible in the image. The X-axis is crowded with noise. The peak shape is sharp, triangular, or line-like; not Gaussian. The intensity levels are high, but so are noise levels.
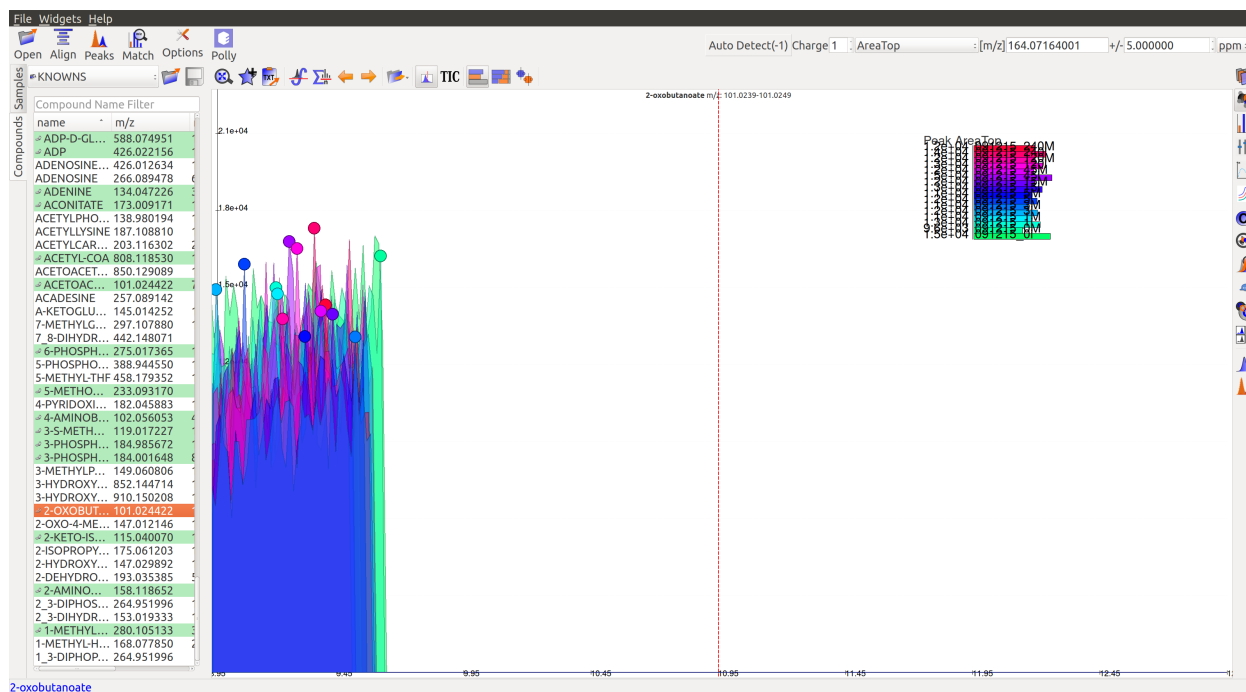
More examples of noisy peaks:
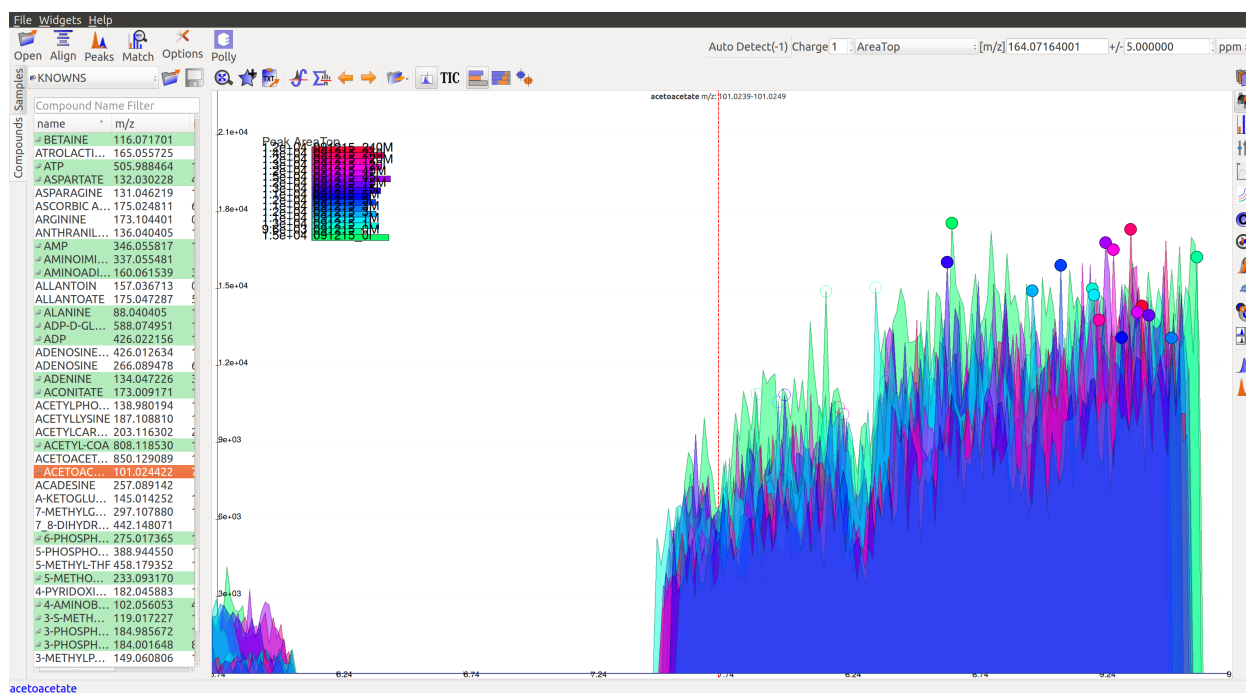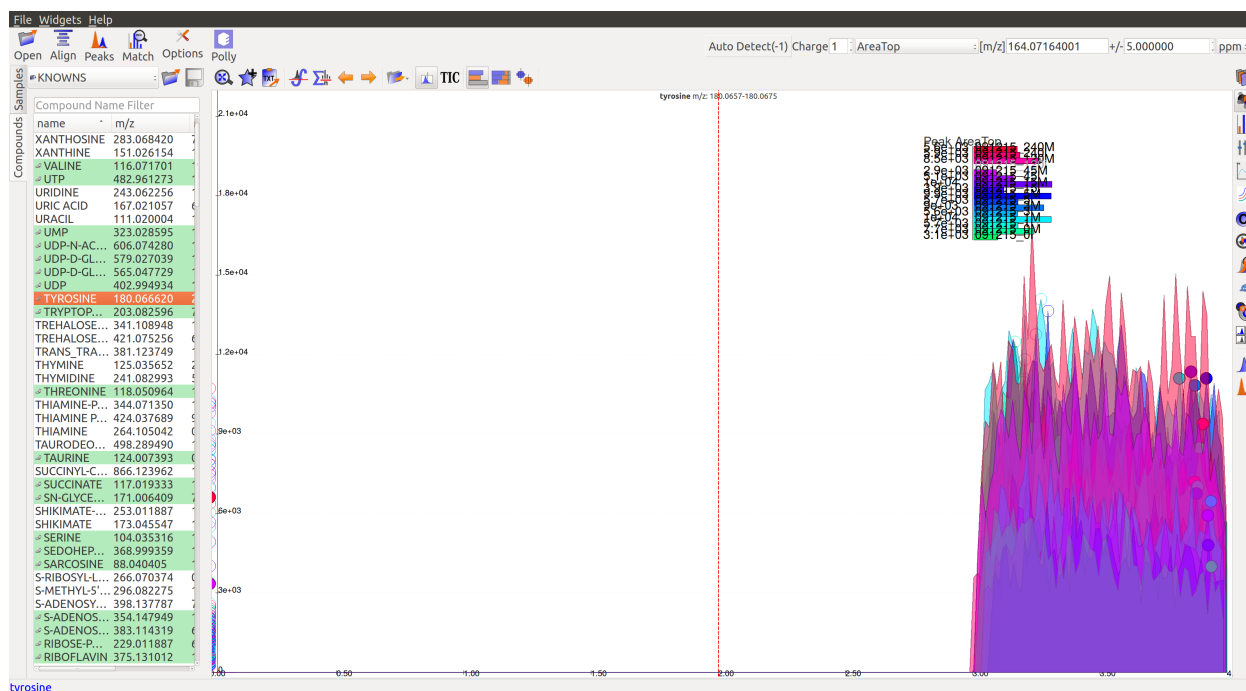
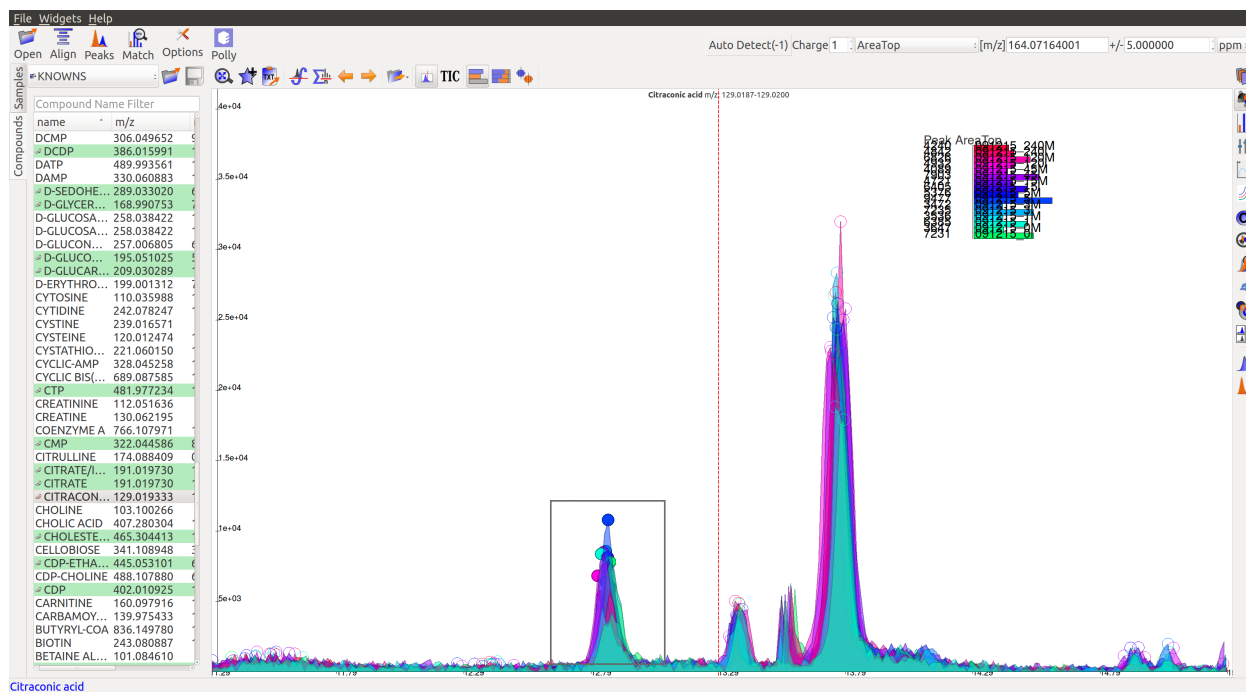- The peaks don't have a Gaussian shape, and are also noisy. The intensity values are very low.

- For low intensity groups like this, the peak characteristics can be determined by zooming in.
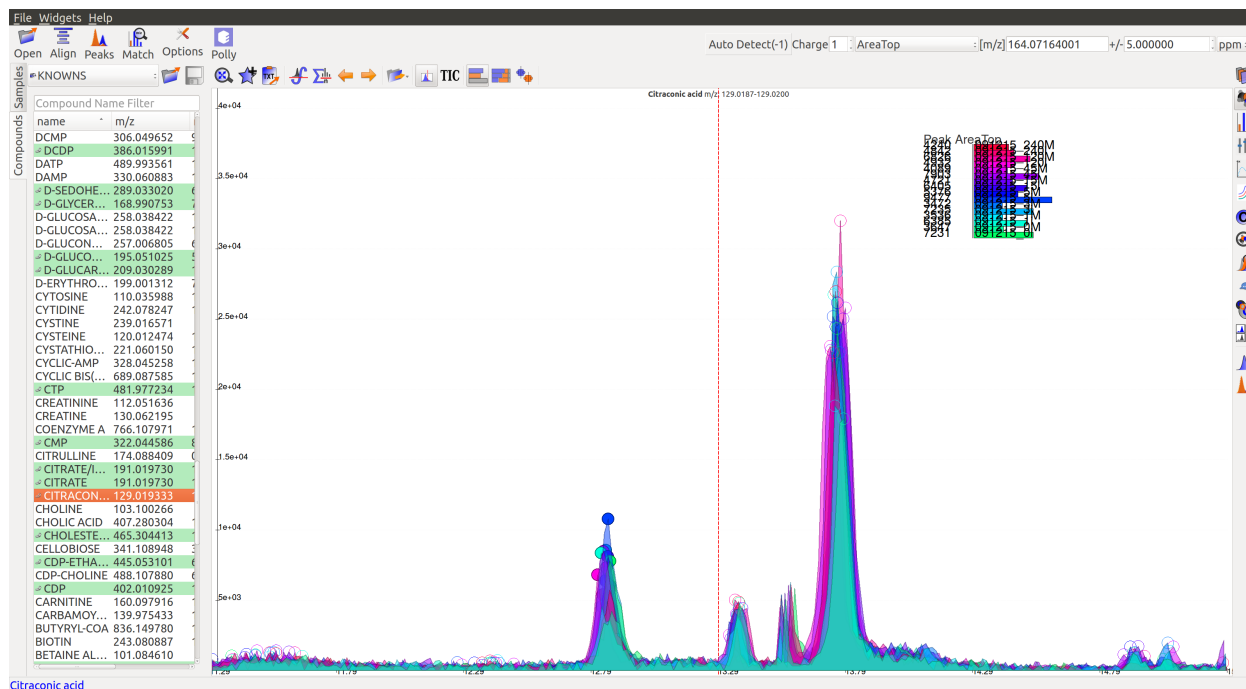
The mouse can be used to select the area of the peak as shown below

On zooming, it will be easy to make a decision on peak quality
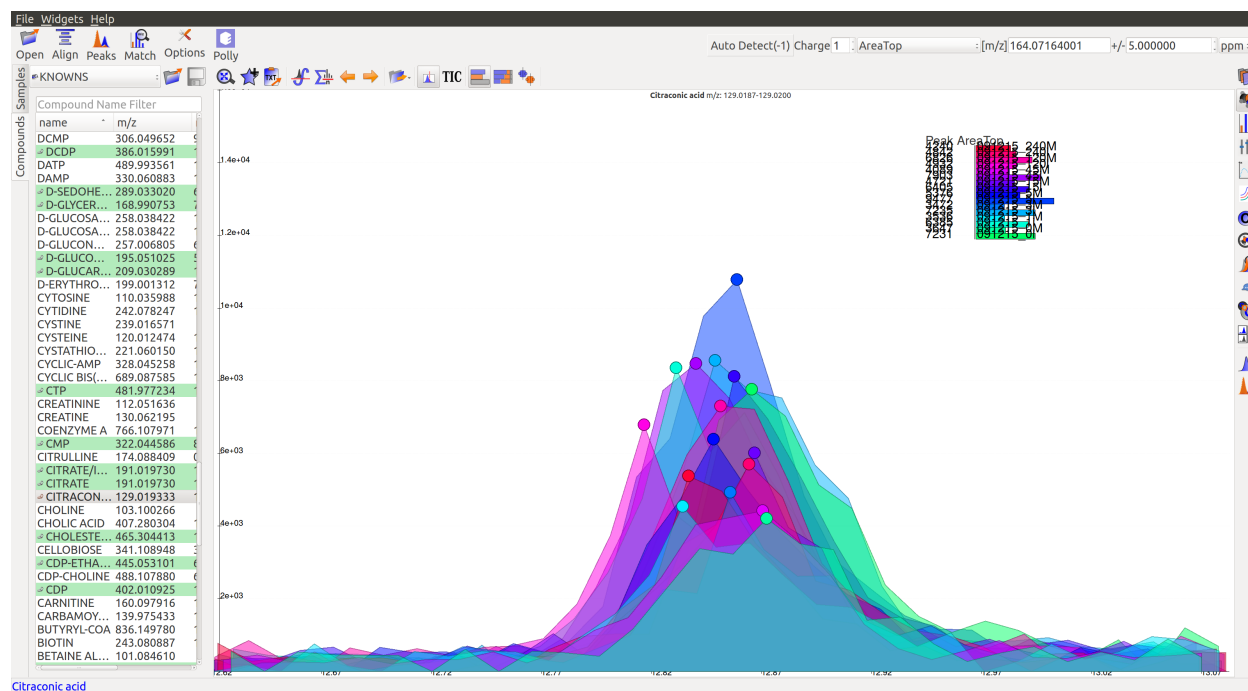
Note: Users can mark any ambiguous peaks as good, and can review all such peaks later in the process.

Export

There are multiple export options available for storing marked peak data. Users can either generate a PDF report to save the EIC for every metabolite, export data for a particular group in .csv format, or export the EICs to a Json file as shown below.

Users can select *All, Good, Bad or Selected* peaks to export.

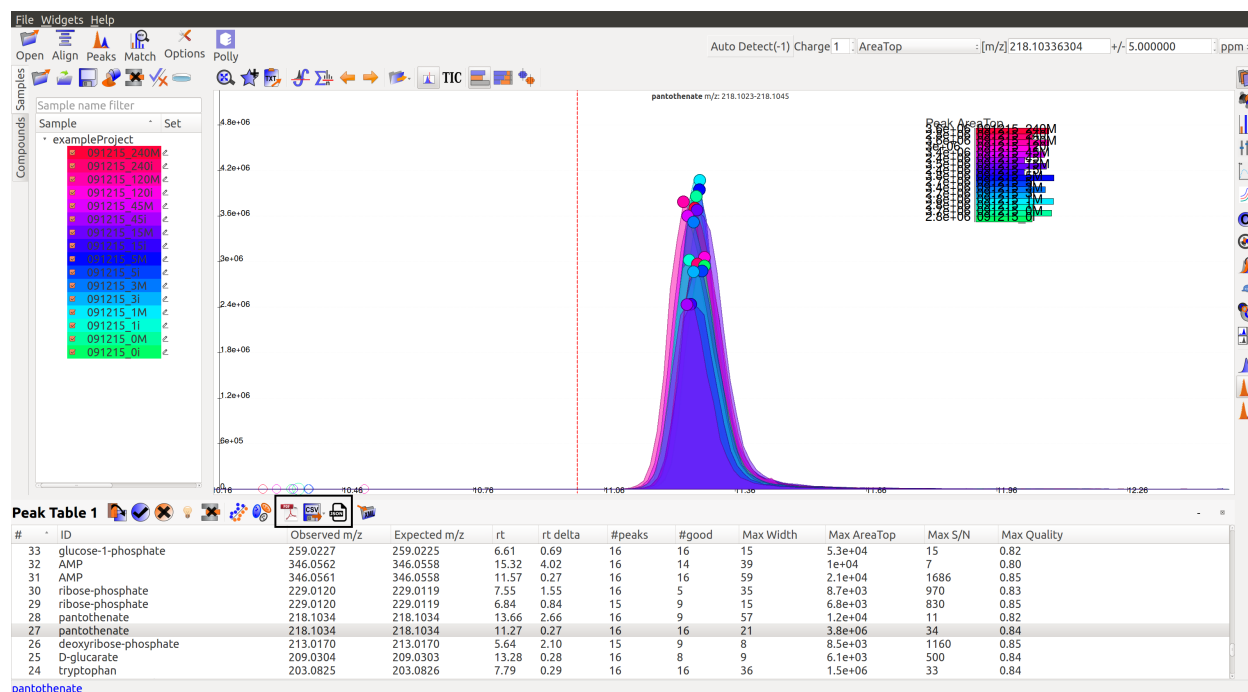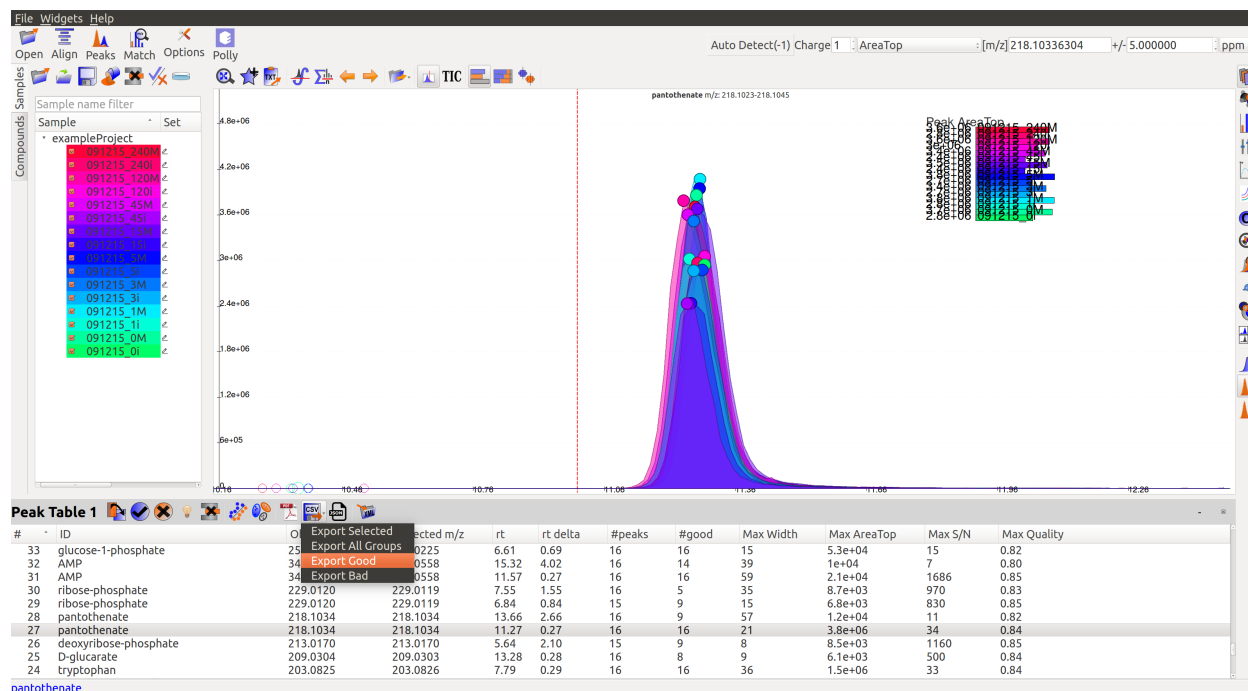
The *Export Groups to CSV* option lets the users save the 'good'/'bad' labels along with the peak table. Users also have the option to filter out rows that have a certain label while exporting the table.

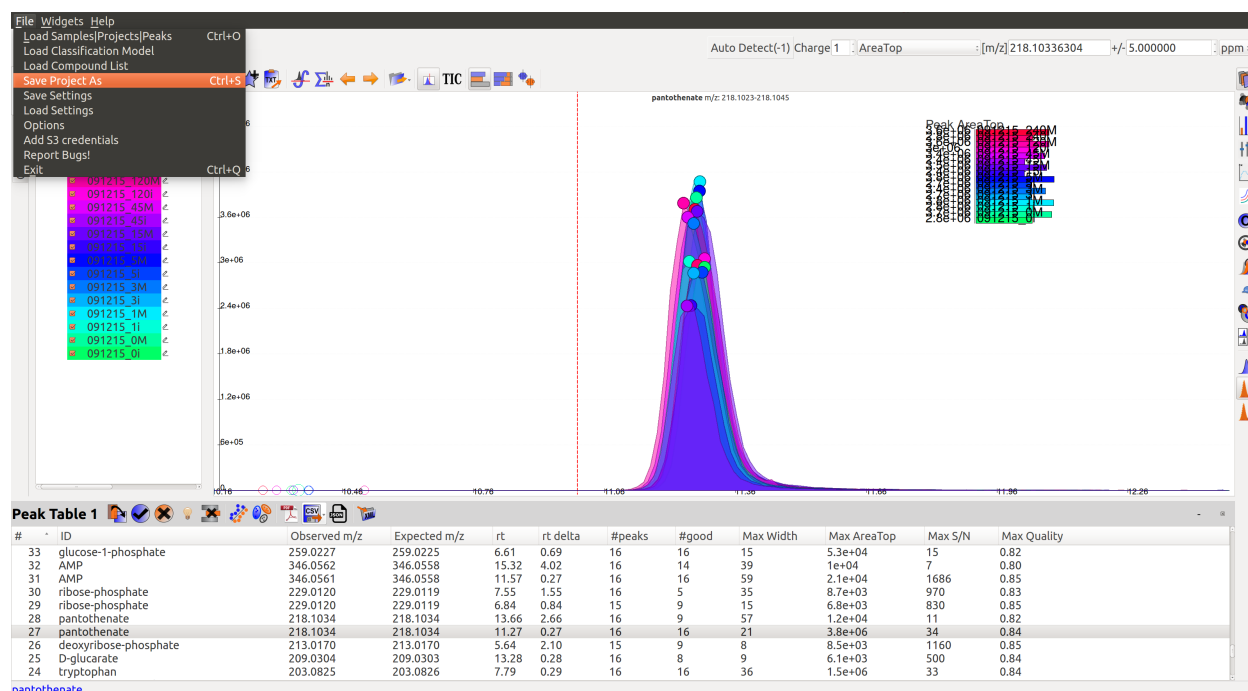*Generate PDF Report* option saves all EICs with their corresponding bar plots in a PDF file.

Export EICs to Json option exports all EICs to a Json file.

Another option is to export the peak data in .mzroll format that can be directly loaded into El-MAVEN by clicking on the Load *Samples|Projects|Peaks* option in the File menu. For this, go to the File option in the menu bar, and click on '*Save Project*'.

Labeled LC-MS/MS Workflow

This is a tutorial for processing LC/MSMS data files through El-MAVEN.

Preprocessing

msConvert is a command-line/GUI tool that is used to convert between various mass spectroscopy data formats, developed and maintained by proteoWizard. Raw data files obtained from mass spectrometers need to be converted to certain acceptable formats before processing in El-MAVEN.

Input

msConvert supports the following formats:

- .mzXML

- .mzML

- .RAW ThermoFisher

- .RAW Walters

- .d Agilent

- .wiff ABSciex

The settings used for msConvert as a GUI tool are captured in the following screenshot:
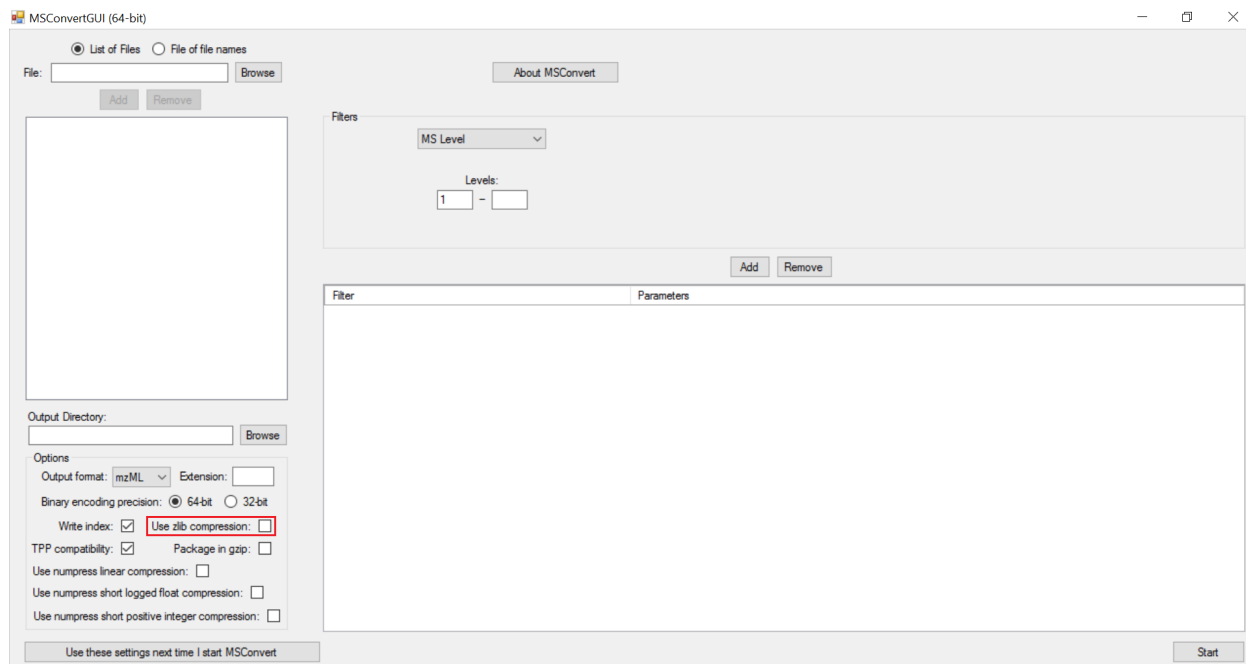
Note: It is important that zlib compression is enabled by default in msConvert. El-MAVEN in its current form does not support zlib compression. Make sure to uncheck "Use zlib compression" box.

Output

msConvert can convert to an array of different formats but El-MAVEN primarily uses .mzXML and .mzML formats.

Launch El-MAVEN

Once sample files are ready for processing, launch El-MAVEN.

Adjust Global Settings

Global Settings can be changed from the *Options* dialog . There are 9 tabs in the dialog. Each of these tabs has parameters related to a different module in El-MAVEN. For example, a tab for Instrumentation details, a tab for file import settings, etc.

To know more about the functionality of different tabs and their settings, users can see the Widgets page. Please be sure to set the desired settings before processing an input file.
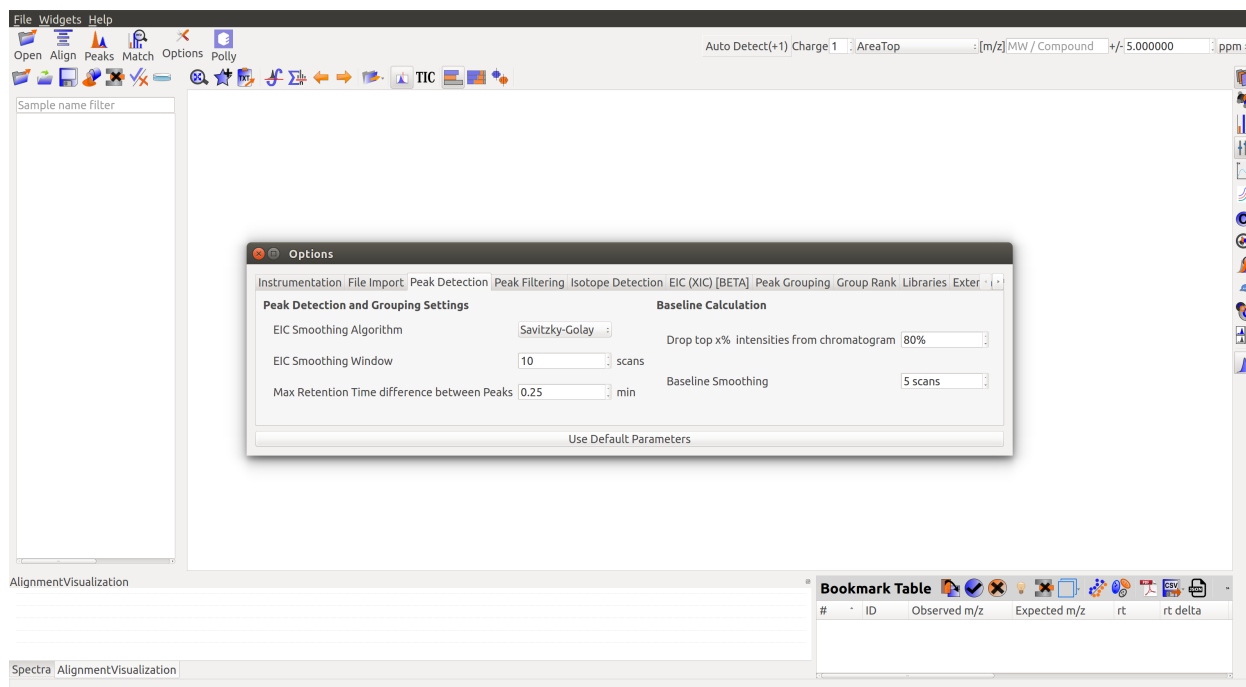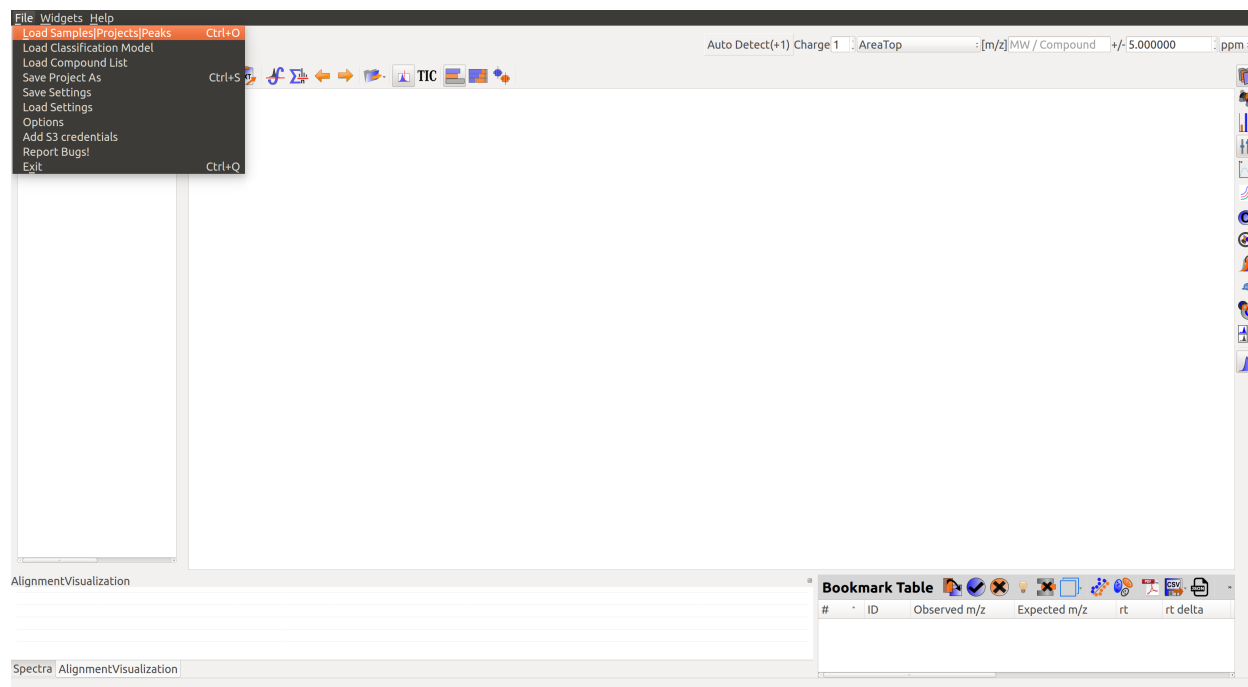
Load Samples

Users can go to File in the Menu, click on *Load Samples|Projects|Peaks* option in the File menu. Then navigate to the folder containing the sample data, select all .mzXML or .mzml files and click Open. A loading bar displays the progress at the bottom.

When the samples have loaded, users should see a sample panel on the left side. If it is not displayed automatically, click on the *Show Samples Widget* button on the toolbar. El-MAVEN automatically assigns a color to every sample. Users can select/deselect any sample by clicking the checkbox on the left of the sample name.

Load Compound Database

Users can click on *Compounds* option in the leftmost menu, navigate to the folder containing the standard database file, select the appropriate .csv file and click *Open.* Alternatively, users may use any of the default files loaded on start-up.

This is a sample Compound Database:

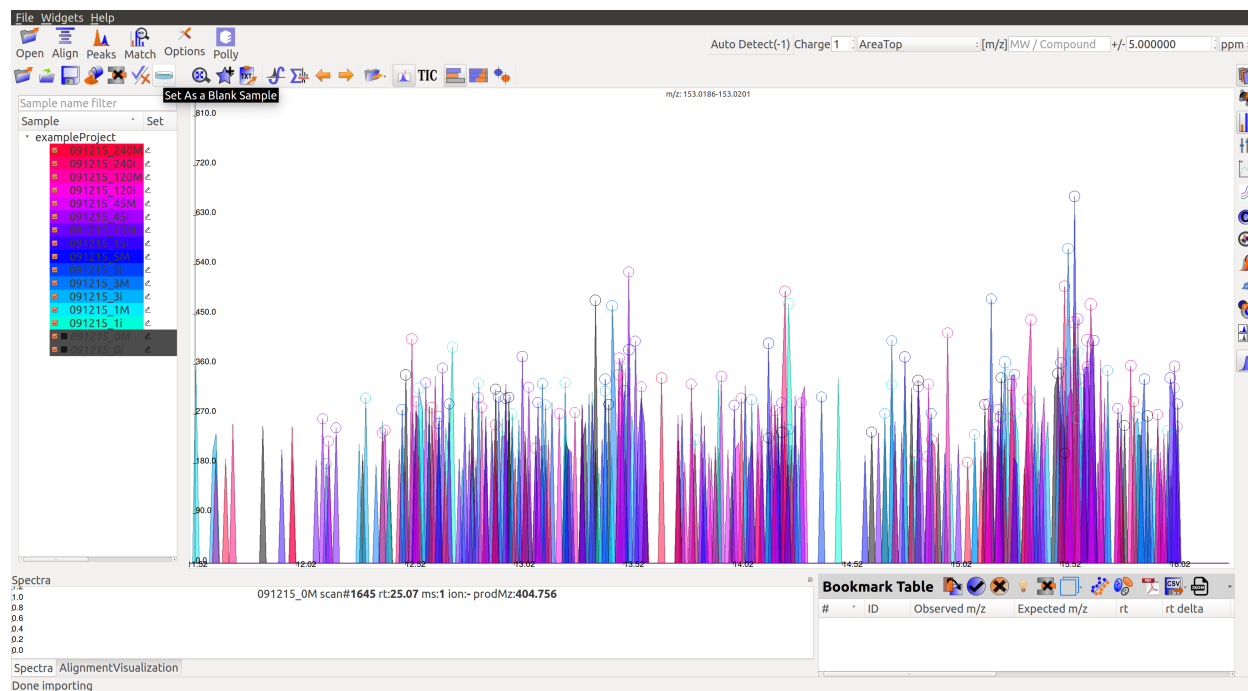|	A	B	C	D	E	F	G	H	I
1	polarity	compound	precursorMz	productMz	expectedRt				
2	-	Citrate 191/67	191	67	4.55				
3	-	Citrate 192/67	192	67	4.55				
4	-	Citrate 192/68	192	68	4.55				
5	-	Citrate 193/67	193	67	4.55				
6	-	Citrate 193/68	193	68	4.55				
7	-	Citrate 193/69	193	69	4.55				
8	-	Citrate 194/68	194	68	4.55				
9	-	Citrate 194/69	194	69	4.55				
10	-	Citrate 194/70	194	70	4.55				
11	-	Citrate 195/69	195	69	4.55				
12	-	Citrate 195/70	195	70	4.55				
13	-	Citrate 195/71	195	71	4.55				
14	-	Citrate 196/70	196	70	4.55				
15	-	Citrate 196/71	196	71	4.55				
16	-	Citrate 197/71	197	71	4.55				
17	-	Citrate 191/111	191	111	4.55				
18	-	Citrate 192/111	192	111	4.55				
19	-	Citrate 192/112	192	112	4.55				
20	-	Citrate 193/112	193	112	4.55				

Compound DB

It lists all metabolite names, chemical formula, HMDB ID, and the class/category of compounds they belong to (if known).

Mark Blanks

Users can mark the blanks by selecting the blank samples from the list on screen, and clicking on the *Set as a Blank Sample* icon in Samples menu.

Multiple blanks can be marked together. The blanks will appear black as shown in the image below:
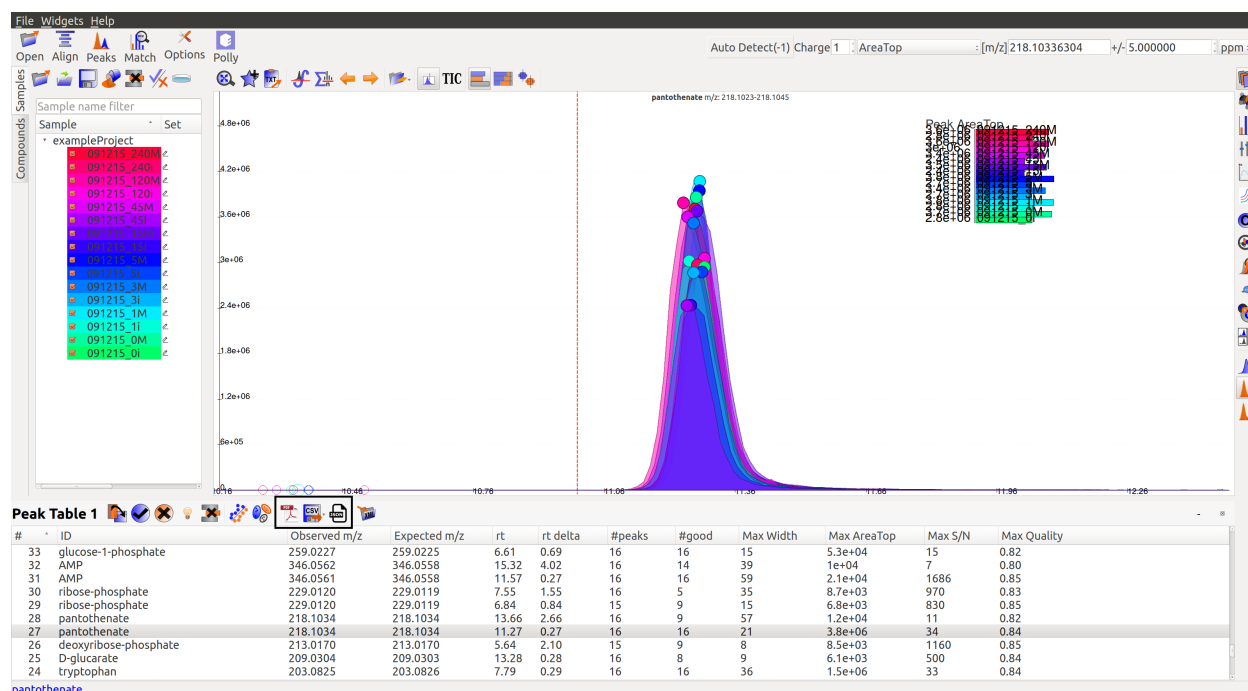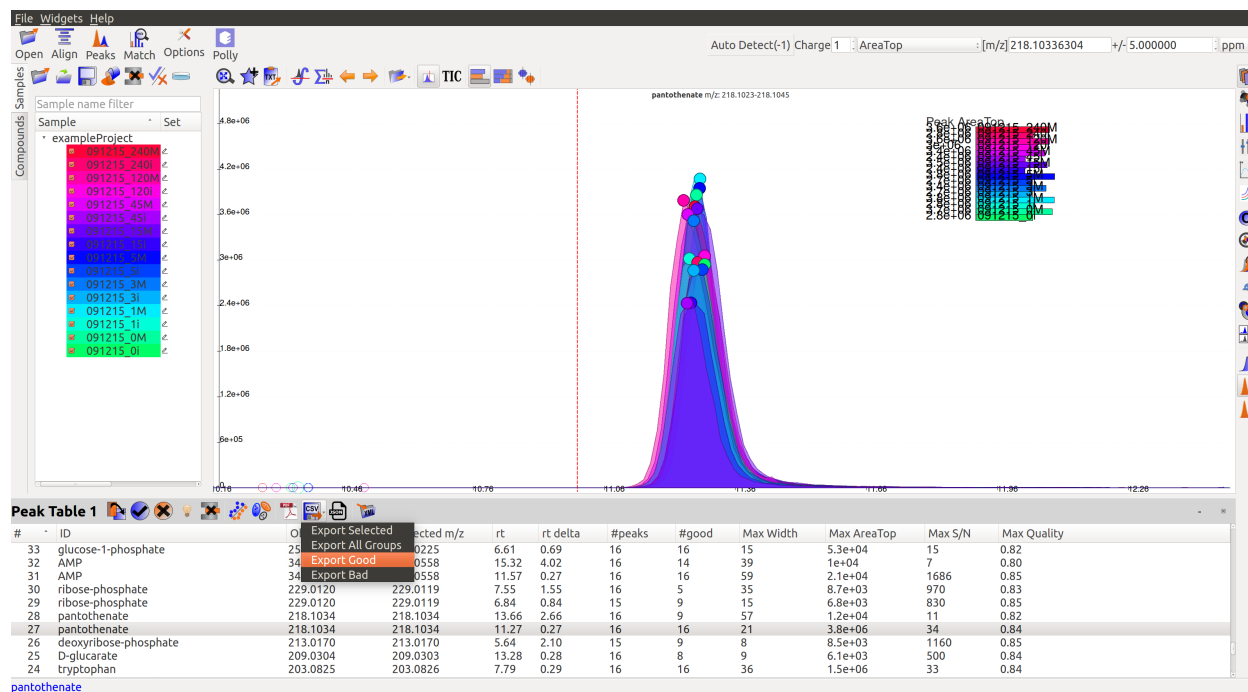

Alignment

(missing)

Export

There are multiple export options available for storing marked peak data. Users can either generate a PDF report to save the EIC for every metabolite, export data for a particular group in .csv format, or export the EICs to a Json file as shown below.

Users can select *All, Good, Bad or Selected* peaks to export.

The *Export Groups to CSV* option lets the users save the 'good'/'bad' labels along with the peak table. Users also have the option to filter out rows that have a certain label while exporting the table.
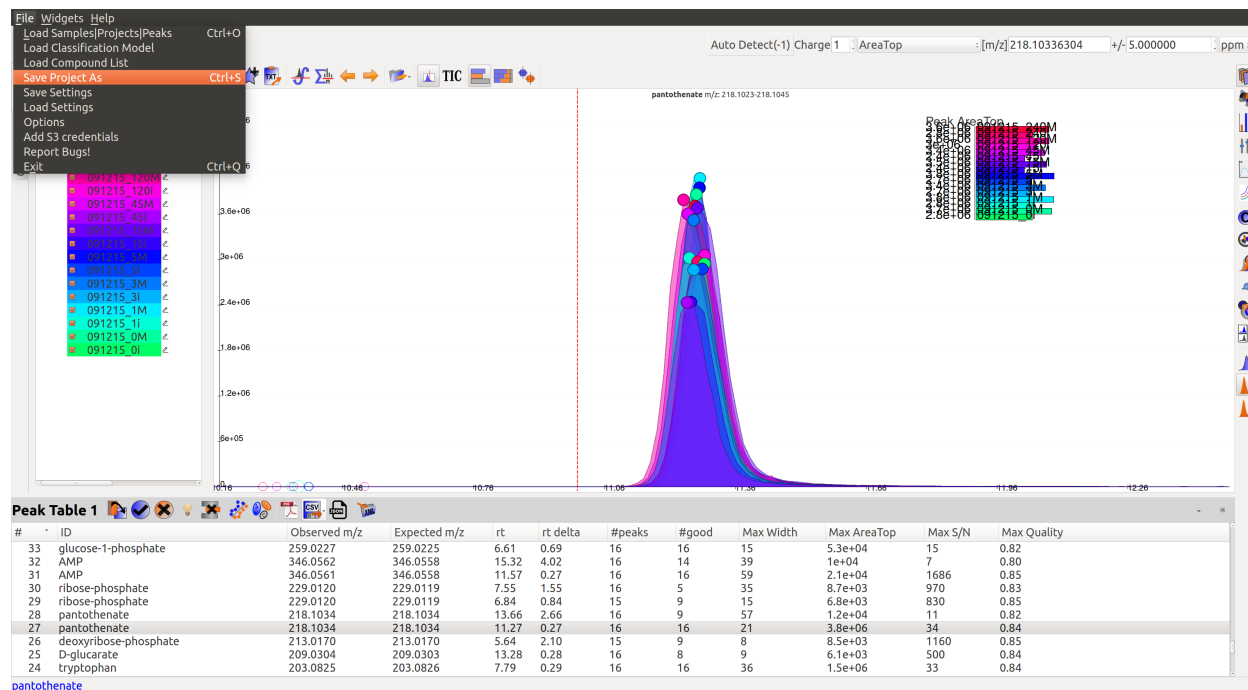
Generate PDF Report option saves all EICs with their corresponding bar plots in a PDF file.

Export EICs to Json option exports all EICs to a Json file.

Another option is to export the peak data in .mzroll format that can be directly loaded into El-MAVEN by clicking on the Load *Samples|Projects|Peaks* option in the File menu. For this, go to the File option in the menu bar, and click on '*Save Project*'.

1.4.4 Validation

Labeled MS/MS Validation

MS/MS Support in El-MAVEN

Tandem mass spectrometry or MS/MS is an important technique in analytical Mass spectrometry. Processing MS/MS data for a large batch of samples can be a time-consuming task. We have added some support for processing such data in El-MAVEN.

Data acquired using the following methods are currently supported in El-MAVEN: Multiple Reaction Monitoring, Parallel Reaction Monitoring (or Full-Scan MS2) and Data Dependant MS2.

Feature Additions

- Text search option using precursor and product m/z

- Targeted Peak Picking for MS/MS

- Manual annotation of SRM transition

Validation

In order to ensure correct reading and processing of MS2 sample files in El-MAVEN, MS spectra and chromatograms from El-MAVEN were validated using Skyline by ProteoWizard.
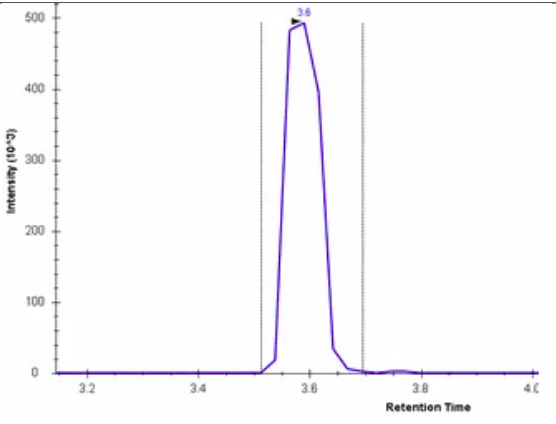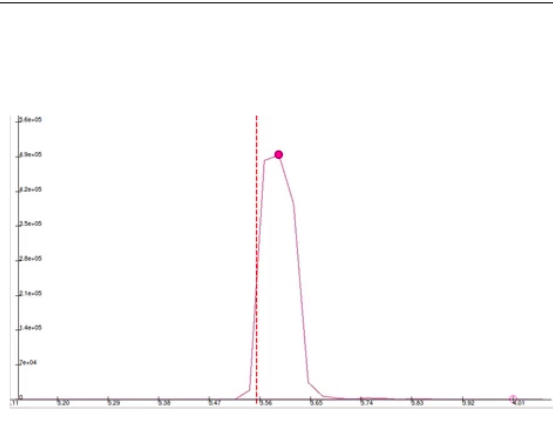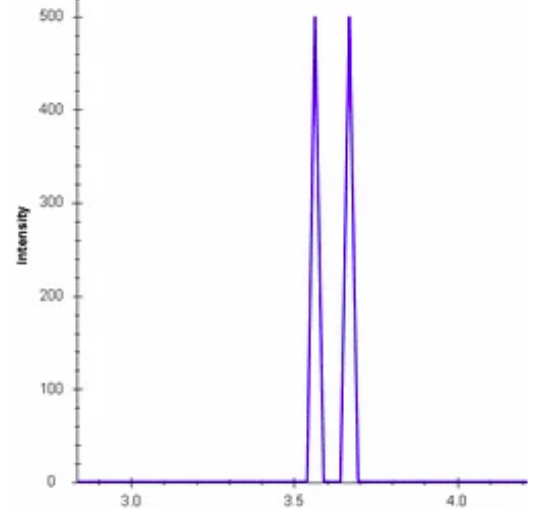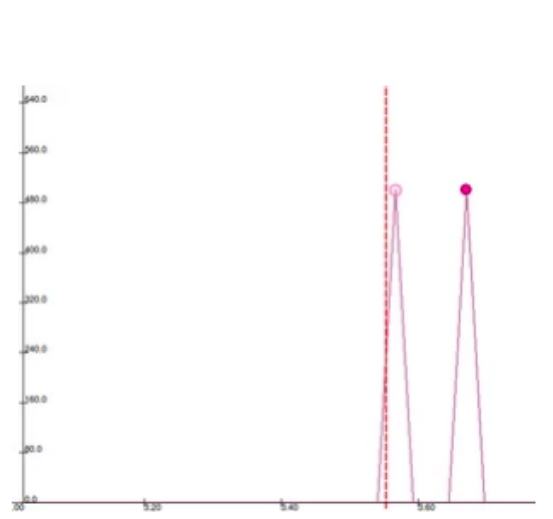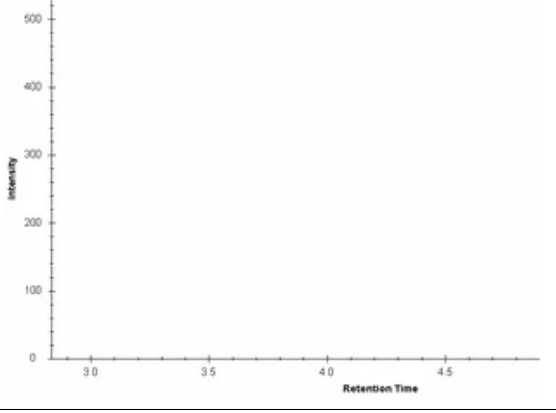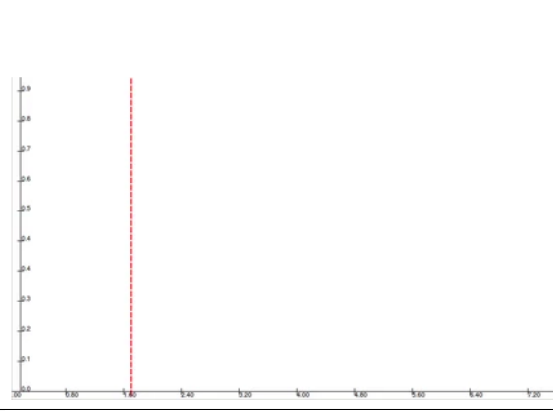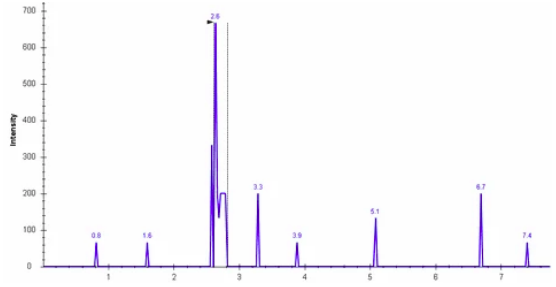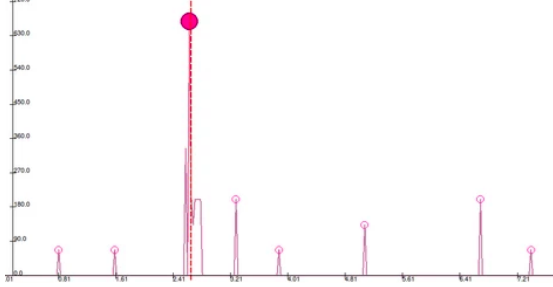
Mass Spectra Validation

Dataset: Full-scan MS2 sample file

Frag-ment	Skyline	EI-MAVEN
#1		
#2		
#3		
#4		
#5		
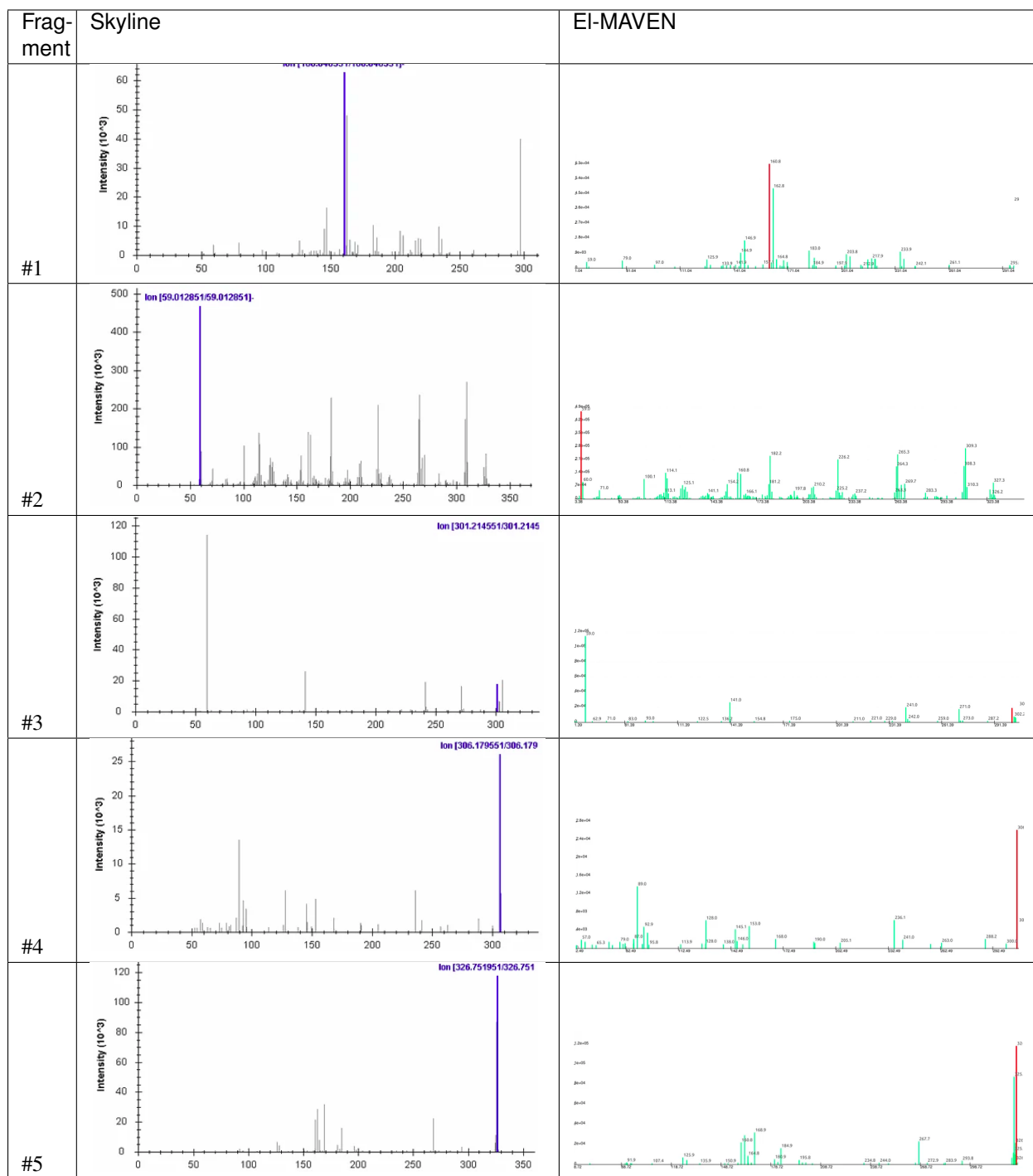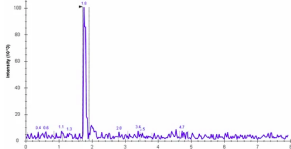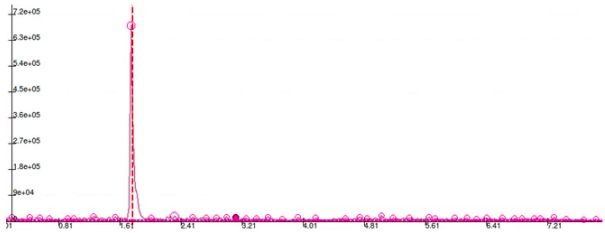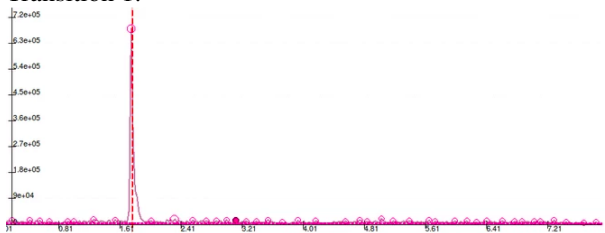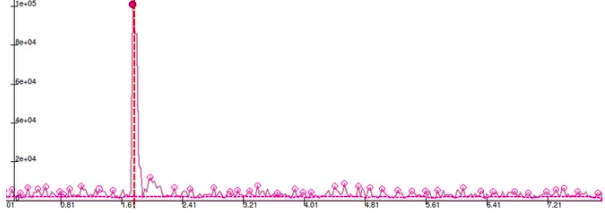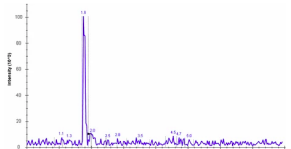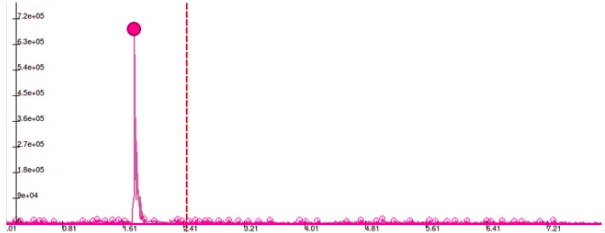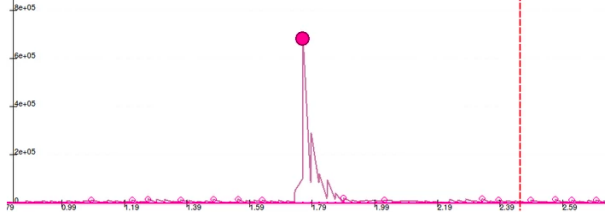
Chromatogram Validation

Dataset: Multiple Reaction Monitoring

Frag-ment	Skyline	EI-MAVEN
#1		
#2		
#3		
#4		

Multiple Transition cases: The same precursor/product pairs can be tracked in multiple runs, for example, in cases where two metabolites have common fragments like Pyruvate 89/89 and Lactate 89/89. El-MAVEN lists these runs separately in the SRM List widget. Following are some examples:
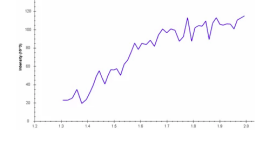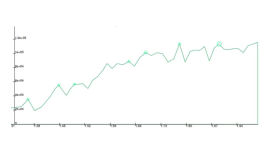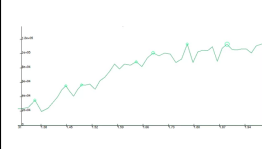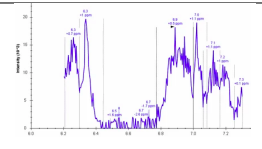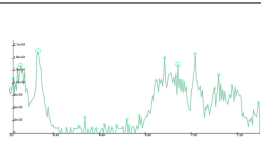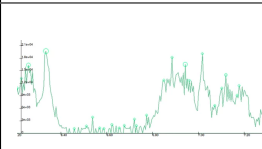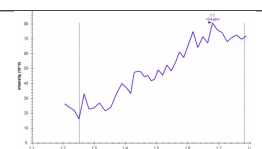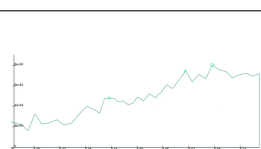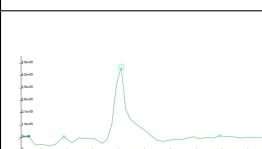
Frag- ment	Skyline	EI-MAVEN	Comments
Lac- tate Frag- ment #1		Compound widget: Both runs for this fragment have been annotated as Lac- tate in the SRM widget Transition 1: Transition 2: 	Skyline EIC matches Transi- tion 2 from the SRM widget in El-MAVEN
Pyru- vate Frag- ment #1	EIC is same as that of Lac- tate for this fragment. Dif- ferent peak is selected ac- cording to the expected re- tention time 	Compound widget: Zoomed in: SRM widget: No transitions mapped to Pyruvate	Unable to find record of another transition in Sky- line. The choppy peaks in EI- MAVEN are due to the merging of two transitions annotated as Lactate

Text search/Compound widget mismatch: Using the text search feature for MRM data can result in choppy peaks at

times as compared to those from the SRM/compound widget. Following are some examples:

Fragment	Skyline	EI-MAVEN Text Search	EI-MAVEN Compound Widget
#1			
#2			
#3			
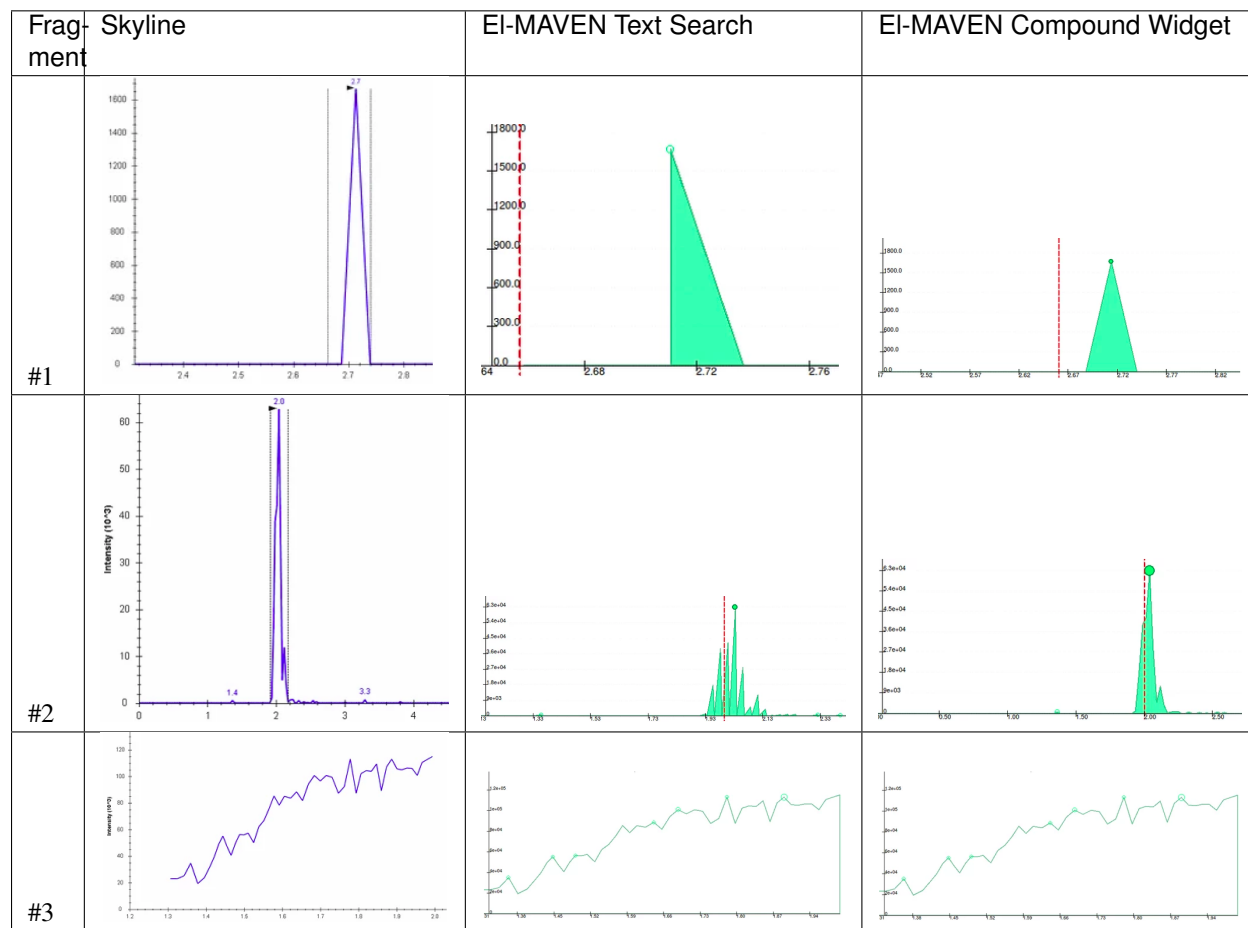
Dataset: Full-scan MS2

Frag-ment	Skyline	EI-MAVEN Text Search	EI-MAVEN Compund Widget	Comments
#1				All three EIC's match exactly
#2				All three EIC's match exactly
#3				EIC from the compound widget is different as one of the runs has been annotated as this fragment which pulls up the EIC for the whole run

Issues Observed

- Text search and compound/SRM widget EIC mismatch for MRM data: EIC obtained from the compound widget and Skyline are the same. The data points on the plot are same for all three. The error in text search is due to different processing. This has been filed as issue number #487; Status: Unresolved

- Text search and compound widget EIC mismatch for PRM data: EIC from Text search and Skyline are the same. The error in compound widget happens due to linking of SRM ID to the compound widget. This pulls up the whole run instead of the particular fragment. This is a known issue #442; Status: Unresolved

- Merged peaks in case of multiple transitions: This is being handled as part of issue number #405; Status: In progress

- Mass spectra displays wrong product M/z value in some cases: The prodMz field in the title of spectra widget displays the base peak m/z value which may or may not be the same as the fragment m/z. This will be resolved as part of #396; Status: Unresolved

- Peak table displays NA in Expected m/z column for MRM data: This has been filed as issue number #493

Conclusion

Mass spectra validation against Skyline passed. This indicates correct parsing of MS/MS data in EI-MAVEN. EIC validation against Skyline highlighted some known and unknown issues. We recommend using Text search for PRM data and Compound/SRM widget for MRM data processing until the reported issues have been fixed.

1.5 Contributors

- MAVEN team at Princeton University

- Eugene Melamud

- Victor Chubukov

- George Sabu

- Sahil

- Raghav Sehgal

- Shubhra Agrawal

- Raghuram Reddy

- Rishabh Gupta

- Pankaj Kumar

- Rahul Kumar

- Giridhari

- Kiran Varghese

- Naman Gupta

- Francis Lee

- Avijit Zutshi

1.6 Contributing

You are welcome to contribute. Please go through our contributing guidelines and code of conduct. These guidelines include directions for coding standards, filing issues and development guidelines.

1.7 References

To understand MAVEN's and El-MAVEN's workflows and features, please refer to the following literature on Maven:

- Metabolomic Analysis and Visualization Engine for LC-MS Data, Eugene Melamud, Livia Vastag, and Joshua D. Rabinowitz, Analytical Chemistry 2010 82 (23), 9818-9826

- LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine, Clasquin, M. F., Melamud, E. and Rabinowitz, J. D. 2012, Current Protocols in Bioinformatics. 37:14.11.1-14.11.23.

1.8 Acknowledgement

El-MAVEN would not have been possible without the unwavering support, constant feedback and financial support from Agios Pharmaceuticals, Inc. El-MAVEN thanks the metabolomics community for its immense contribution in taking the tool forward and making it a great success.

1.9 Release History

1.9.1 Current Release

Version 0.4.1

Publication Date: August 4th, 2018

- Optimisations
  - Faster sample upload
  - Checking/unchecking a sample in the presence of a large peak table is faster (#723)
  - No lag in switching between two large peak tables
- UI Changes
  - Added a close button for Scatter Plot widget
  - Removed 'Load Peaks' from Peak tables
  - 'Gallery' widget is only available on the side bar panel
  - Removed 'Ratio' and 'p-value' columns from Peak tables (available in Scatter Plot Peak Table)
  - Removed 'Good' and 'Bad' marking options from EIC widget
  - Polly upload button is not clickable during the upload process (#805)
  - Single button for exporting EIC on EIC widget
  - Removed 'Scatter Plot' button from side bar panel
- Bugs
  - Isotopic peak filtering based on peak quality has been fixed (#772)
  - Crash in bookmark button has been fixed (#768)
  - Empty tables will no longer be uploaded to Polly (#777)
  - Only a single compound database file (.csv) will be uploaded to Polly

1.9.2 Past Releases

Version 0.4.0-beta.1

Publication Date: July 16th, 2018

- Fixed Isotopic detection in shift-drag integration (#781)
- Groups with zero peaks will not be bookmarked on shift-drag integration
- Fixed a bug in changing compound database from the drop-down
- Getting Started window will not open behind the main application (#775)

Version 0.4.0-beta

Publication Date: July 5th, 2018

Version 0.3.2

Publication Date: May 17th, 2018

Version 0.3.1

Publication Date: April 27th, 2018

- Bug Fixes:
    - Fixed compound name in CSV export [(#635)](#635)
    - Fixed automated detection of high ranking groups
    - Fix installer issue on Mac (Installer not working due to netcdf addition)
- UI:
    - Change default tab for El-MAVEN - Polly Integration dialog

Version 0.3.1

Publication Date: April 26th, 2018

Version 0.3.0

Publication Date: April 24th, 2018

Version 0.3.0

Publication Date: April 23rd, 2018

Version 0.3.0-beta

Publication Date: April 11th, 2018

- Features
    - Export Scatter plot groups into a peak table
    - Introduced Exception handling in sample upload process
    - Send logs to sentry in the event of a crash
    - Upload data to Polly through CLI and GUI
    - Download project settings from Polly
    - New Alignment algorithm: Obi-warp
    - Support for cdf files on Windows
    - Highlight detected compounds in the database
- Enhancement
    - Consistent sample ordering across widgets
    - Append Sample number to Sample Name for mzML files
- Refactor
    - Isotope Widget
    - Isotope detection
- Bug fixes
    - Fixed Isotopes expected m/z in mzroll
    - Blank samples will be visible in sample widget
    - Fix trailing delimiter in peak detailed format
    - Isotopic intensities are consistent between isotope widget and peaks table

Version 0.2.4

Publication Date: February 7th, 2018

Version 0.2.4

Publication Date: February 5th, 2018

Version 0.2.4 Beta

Publication Date: January 24th, 2018

Version 0.2.3

Publication Date: January 23rd, 2018

Version 0.2.2

Publication Date: December 28th, 2017

1.10 Bugs and Feature Requests

Existing bugs and feature requests can be found on El-MAVEN's GitHub issue page. Please search the existing bugs and feature requests before you file one yourself.

1.11 Copyright and License

Code and documentation copyright 2017 Elucidata Inc. Code released under the GPL v2.0. Documentation is released under MIT license.