RMet Manual

Table of contents

S1- Installation

- Open https://github.com/SUTChemometricsGroup and select the RMet repository

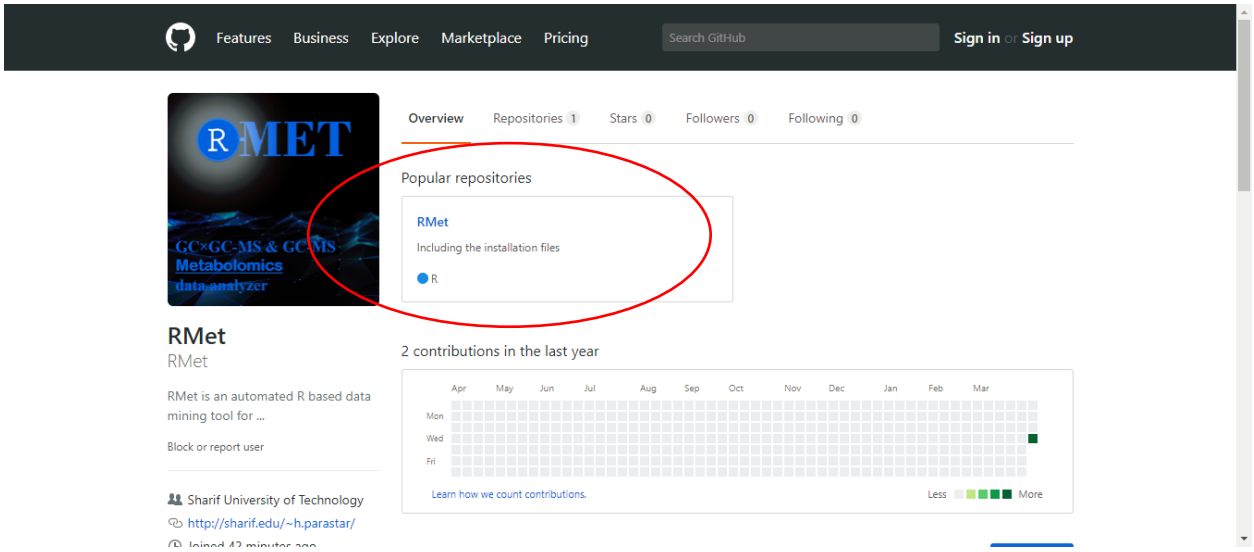

Figure S1. RMet on GitHub

- Download and run RMet.exe file using the Google drive link provided in Intallation_file folder

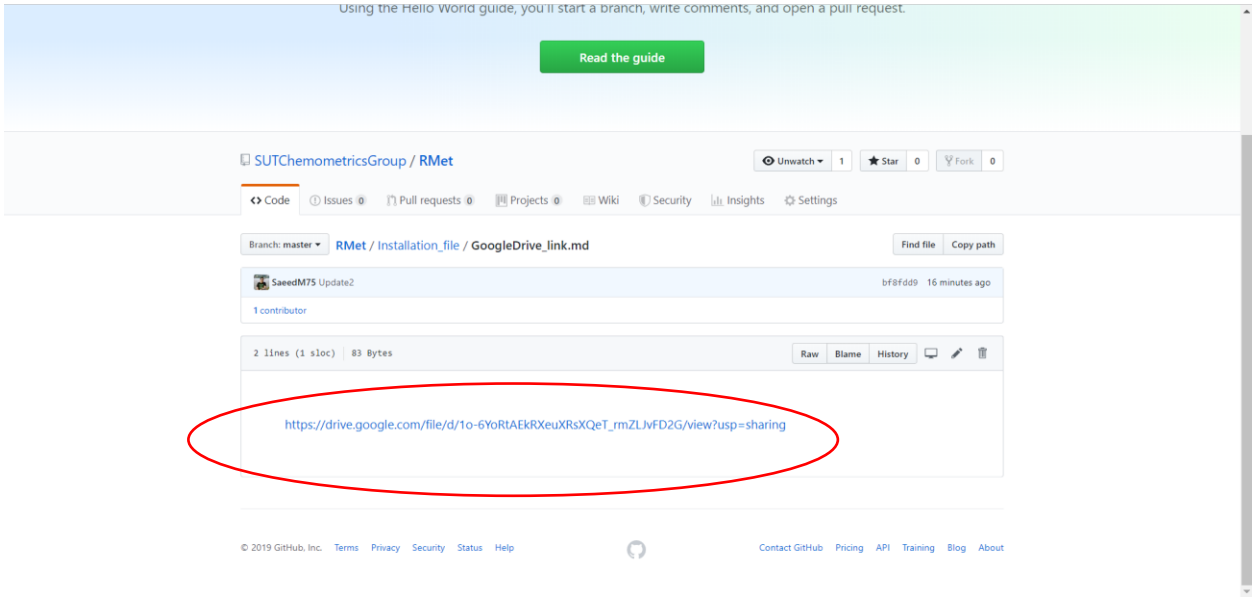

Figure S2. Downloading the installation file from the RMet repository

- RMet will be easily installed just by multiple clicking on Next button!

A)

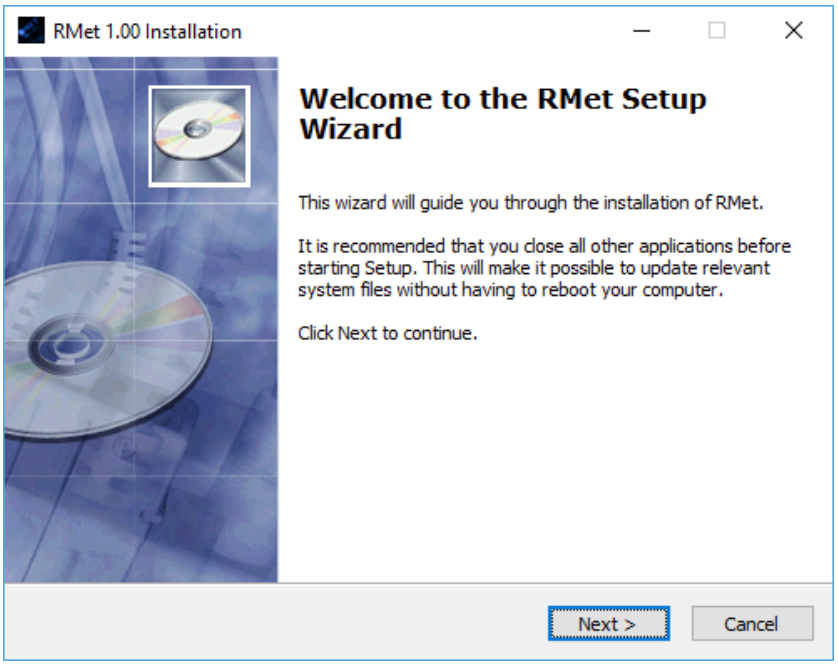

Figure S3. RMet installation step

B)

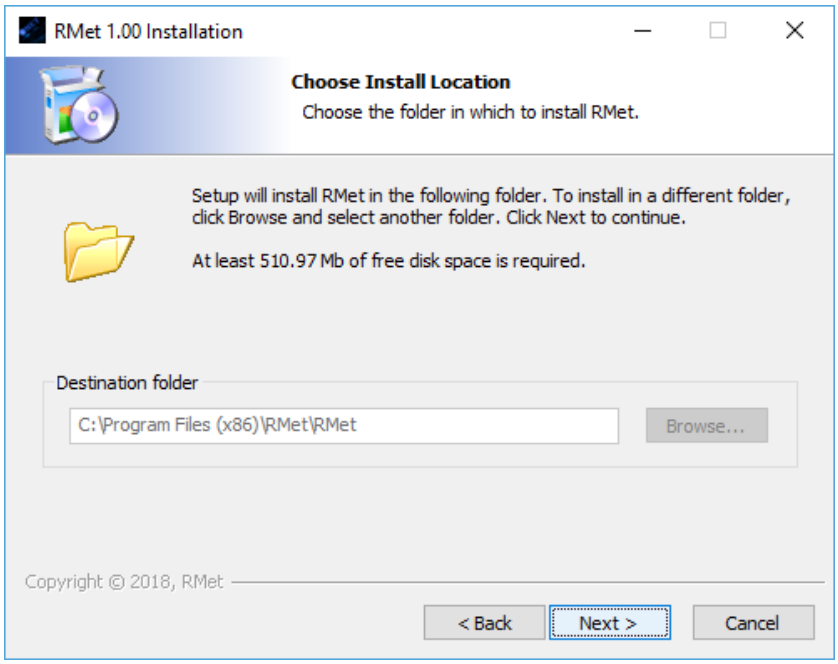

Figure S4. RMet installation step

C)

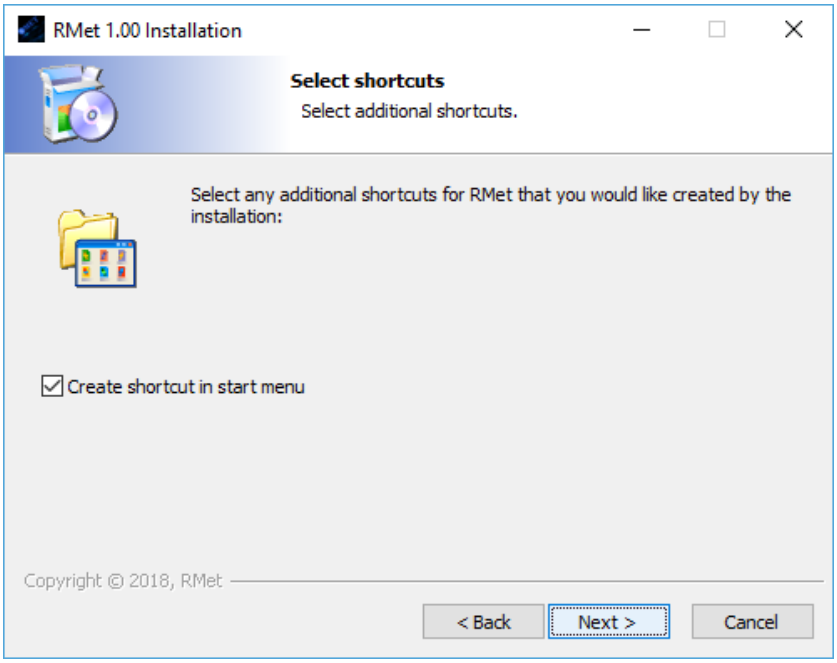

Figure S5. RMet installation step

D)

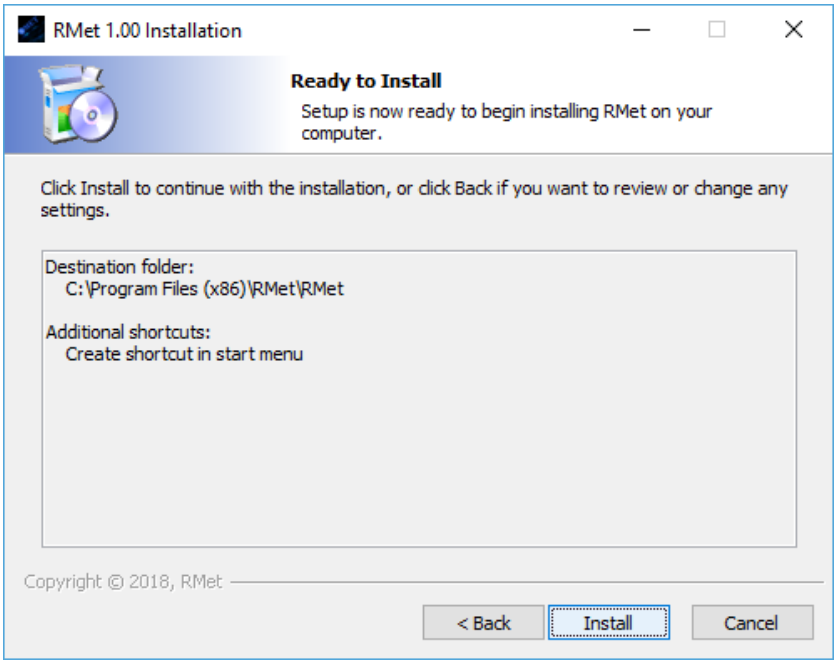

Figure S6. RMet installation step

E)

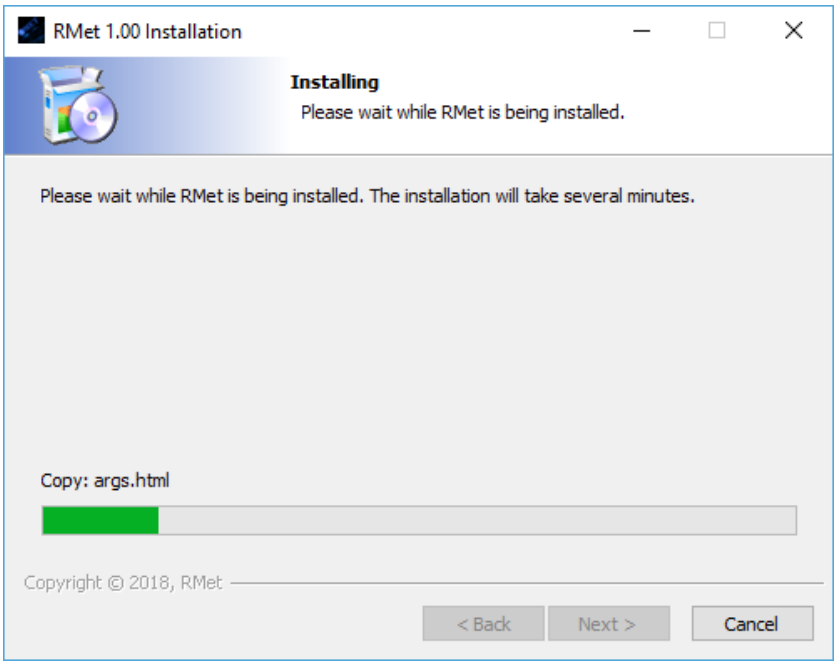

Figure S7. RMet installation step

F) After the installation is complete you can open RMet from the start menu or desktop

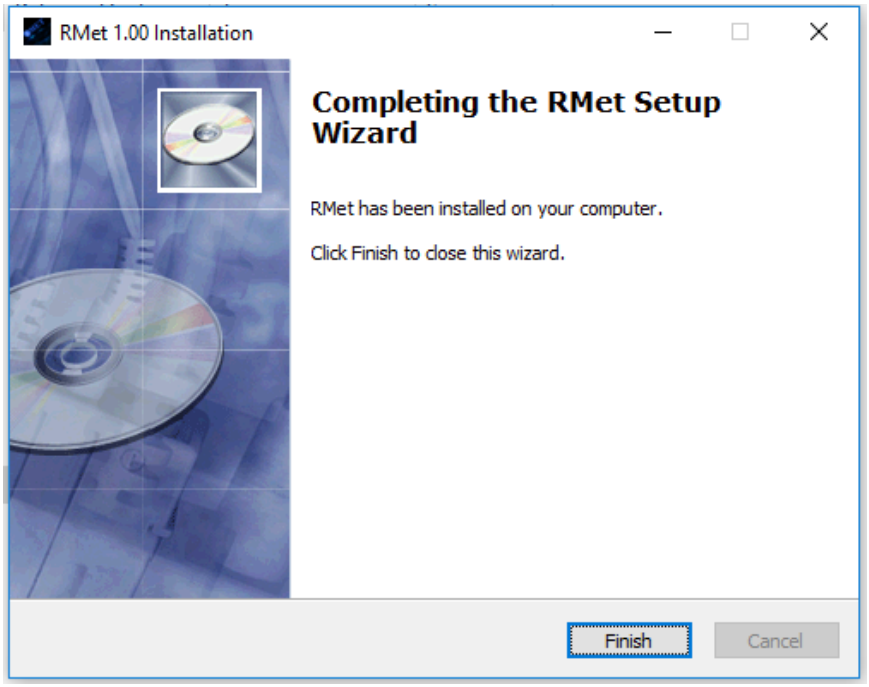

Figure S8. RMet installation step

S2. GC×GC-MS data analysis procedure

- Open RMet software and after the packages are load click on “Start Analysis” button; here you can also see more information about the RMet and its developers by clicking on “About” button.

Figure S9. RMet software initial page

After clicking on the “Start Analysis” button the analysis workflow will be shown; click on “Data import” button in order to begin analysis.

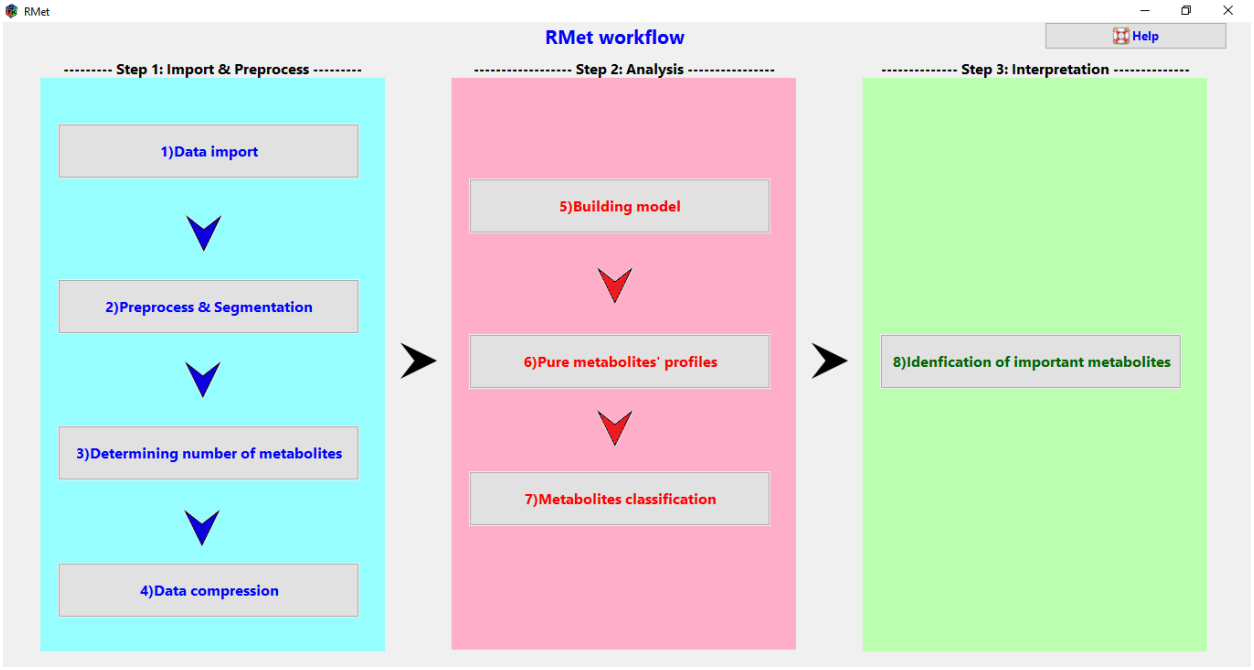

Figure S10. RMet workflow

S2.1. Data import

In this section, you can upload the raw GC×GC-MS data or results of an incomplete RMet analysis in three different formats including CDF, CSV, and RData. First, you should select the data category, since a complete analysis of big GC×GC-MS data may be time-consuming you can perform some steps and save the result in order to continue analysis in another time that's why RMet has a Data category combo button which includes Raw data, Segment, Augment, 2nd column profiles, Pure spectral profiles, and 2nd column peak areas. Select the raw data in order to analysis GC×GC-MS instrument data for the first time. Second, you should select the Data type which is GC×GC-MS; after selecting it a window will pop up and you should input the instrument Modulation period(time) in seconds and the detector frequency in Hz. Finally, select the data format and click on "Browse & import" button and select the data destination in order to import data.

When the uploading is done you can see each data matrix's information by double-clicking on it. If you have uploaded the data in any format other than Rdata you save the data matrices in Rdata format by clicking on "Save the as .Rdata" button, by doing so you can upload these matrices much more quickly. RMet also provides you with both 2D and 3D Total Ion Chromatograms (TIC). You can view these plots by selecting the dimensions and clicking on "Plot" button. Note that you can zoom in or out or turn the 3D-TIC plot.

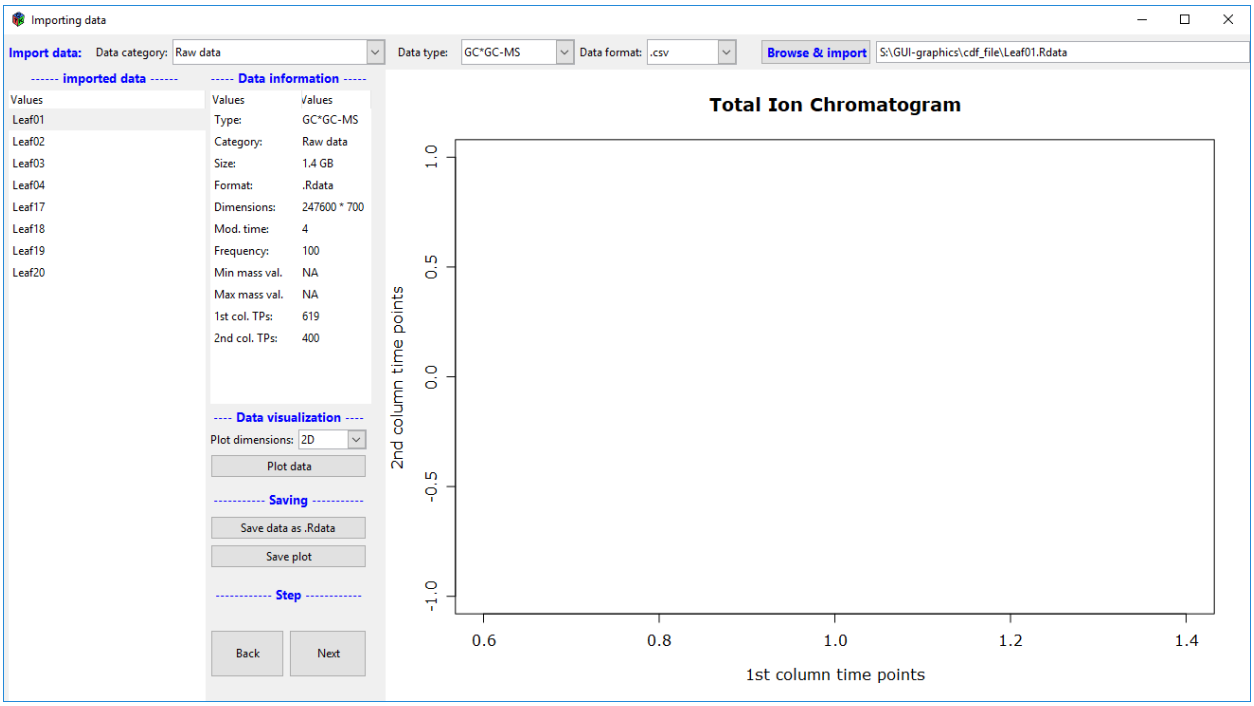

Figure S11. Importing data in RMet.

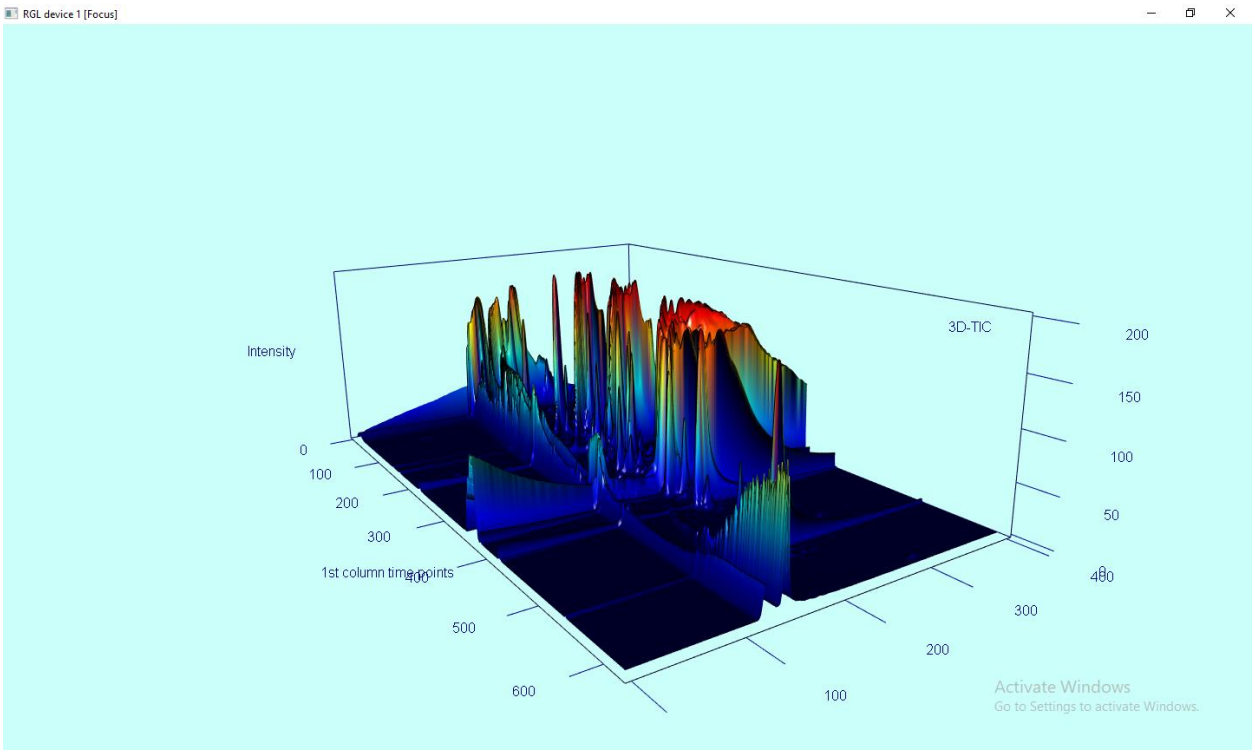

Figure S12. 3D-TIC of a sample GC×GC-MS data

S2.2. Preprocessing & Segmentation

In this section, you can segment your data or delete part of it. First, you should select a data matrix and an intensity ceiling in order to view the 2D TIC. The less you set intensity ceiling the more component will be shown in plot (Figure S13 vs Figure S14).

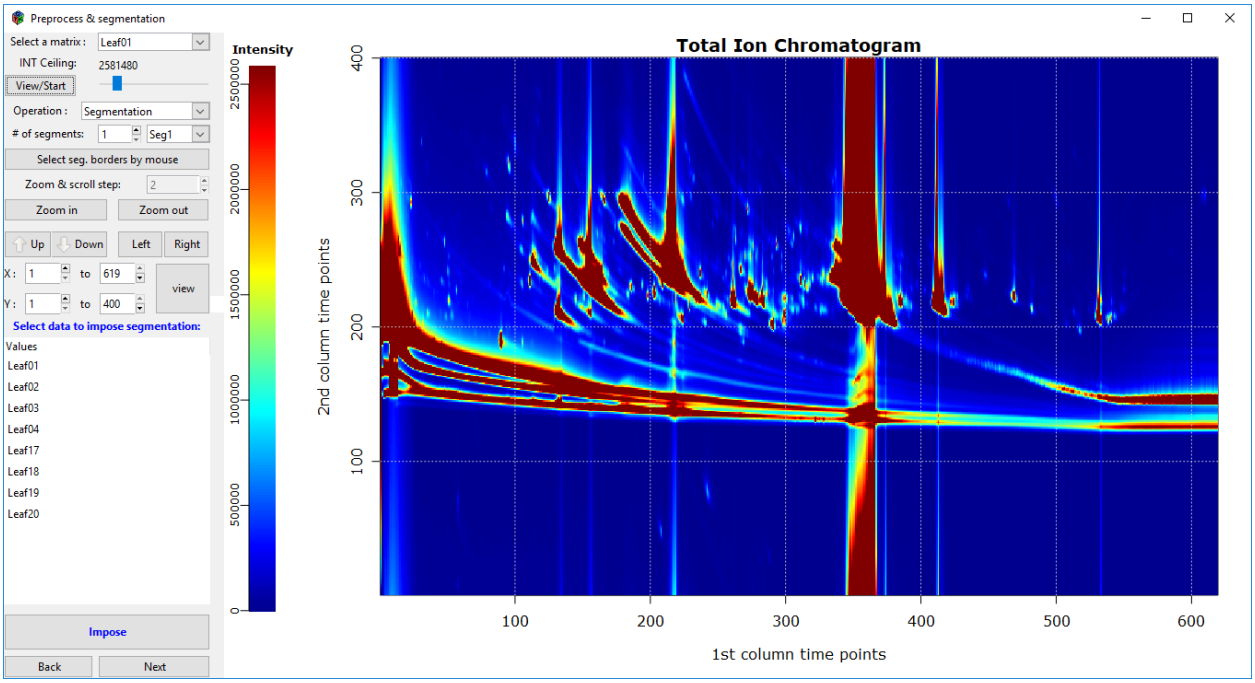

Figure S13. TIC for a sample data matrix (Leaf01) setting INT Ceiling on 2581480

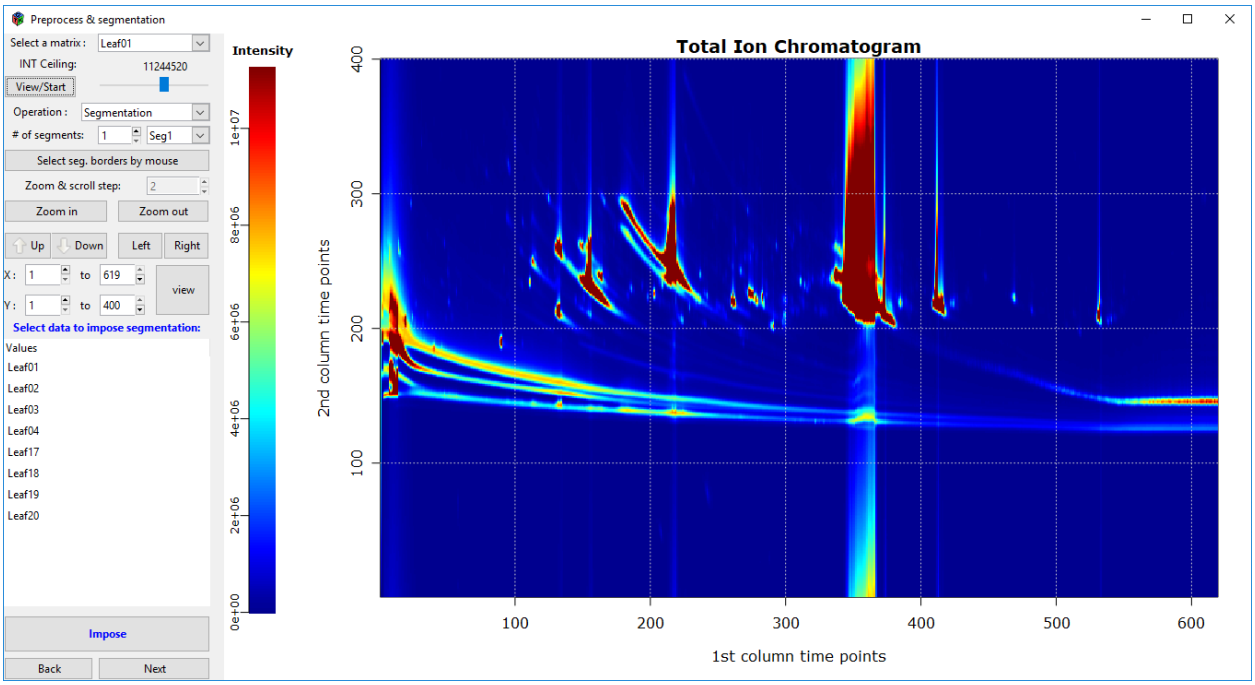

Figure S14. TIC for a sample data matrix (Leaf01) setting INT Ceiling on 11244520

RMet performs three types of operations for preprocessing data. Consider TIC of a sample matrix (Leaf01). Let's suppose you want to segment data in a way to remove the derivatizing agent, column bleeding, and overload areas. You can do it easily be performing a segmentation operation followed by a vertical remove operation (FigureS15).

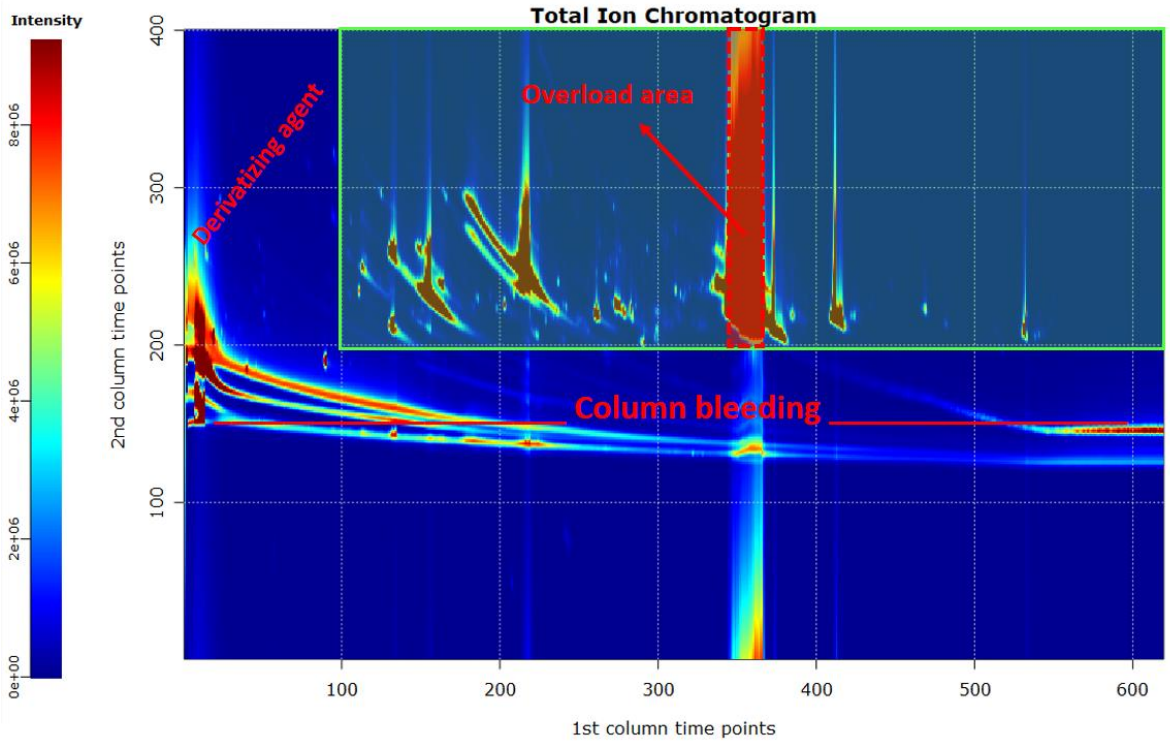

Figure S15. Derivatizing agent, Column bleeding, and Overload areas in TIC of a sample data matrix (Leaf01)

First, you should set the operation on “Segmentation” in order to segment data matrices to the green area in FigureS15. Click on “Select seg. borders by mouse” button. By doing so you can select two vertices of a rectangle which you are going to select. The selection mechanism is shown in Figure S16. When you click on point A the left edge and the top edge of a rectangle is drawn, then you should click on point B, by doing so the right edge and the button edge will be drawn too and the final selection will be automatically shown.

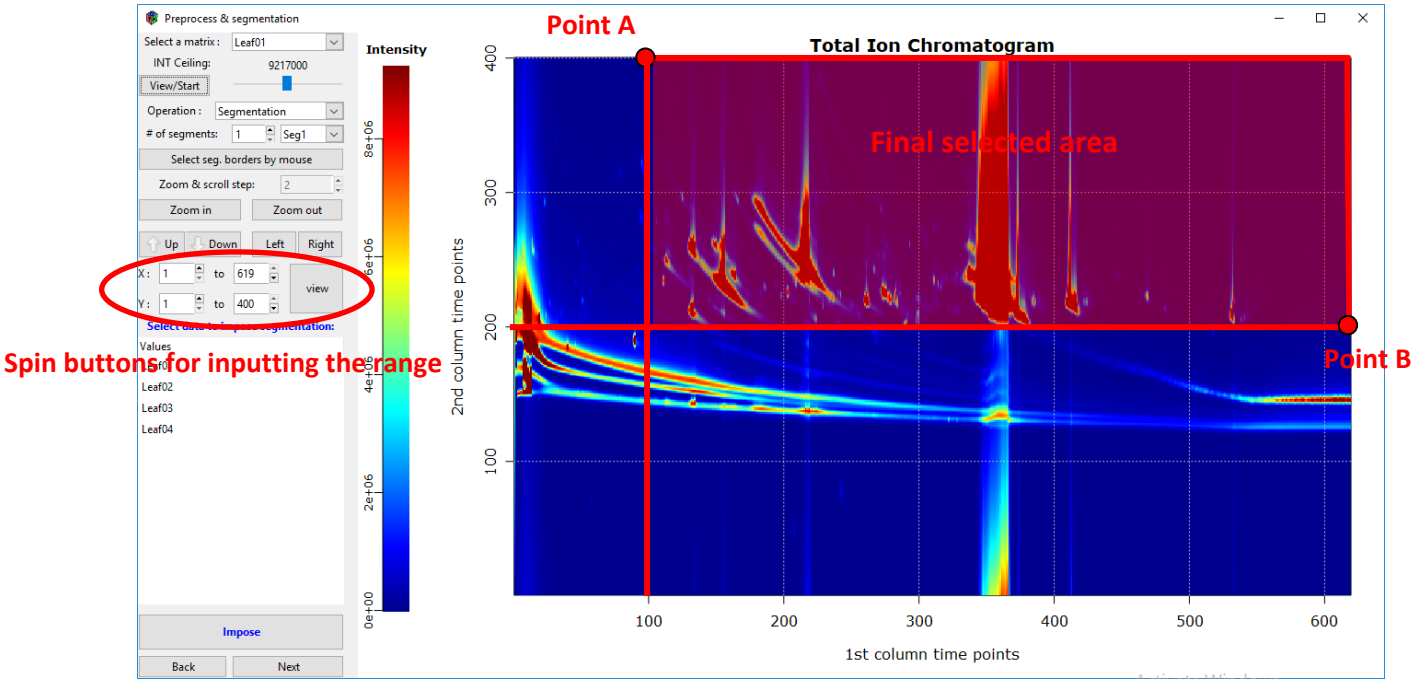

Figure S16. TIC area selection in RMet

Note that you can also input the 1st and 2nd column time points ranges (X and Y ranges) manually with the spin buttons. You can also Zoom in or out or move the selected area in every direction by clicking on a proper button. After reaching the desired area select the matrices which are supposed to be segmented and click on “Impose” button. The result is shown in Figure S17.

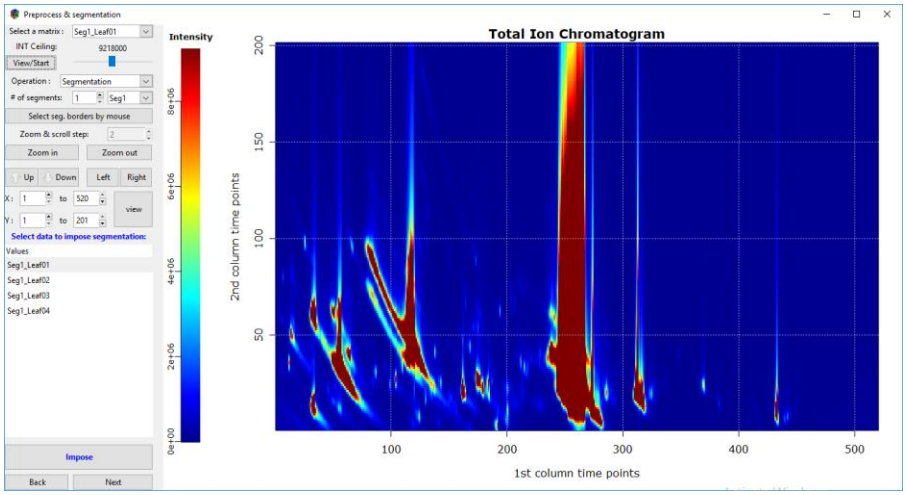

Figure S17. Segmentation of Leaf01 data matrix

Now you should select the “Vertical remove” operation to remove the overload area, just like the segmentation you can select the area by mouse click or you can manually input the range (Figure S18).

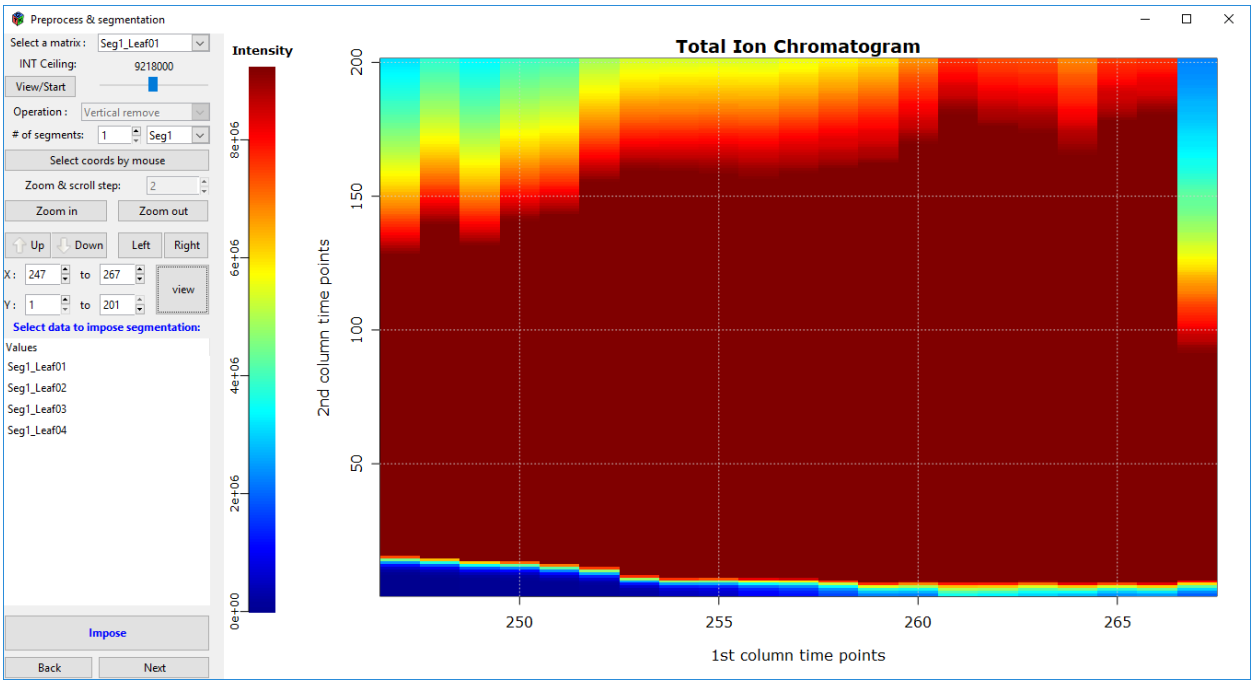

Figure S18. Select vertical area for removal.

In order to reach the final result select the segments and click on “Impose” button. The final resulted TIC which excludes the overload is shown in Figure S19. Click on next to proceed.

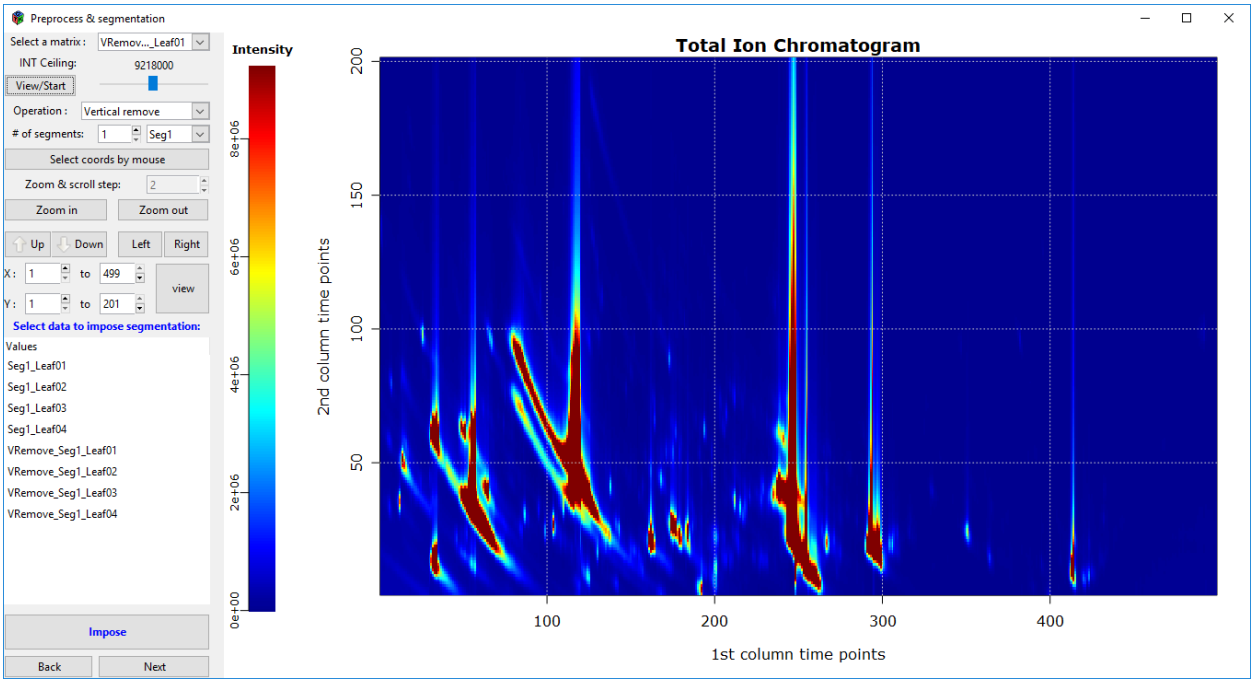

Figure S19. Final preprocessed TIC of Leaf01 data matrix

S2.3. Determining the number of metabolites

In this section, you should determine the number of metabolites in your samples. In order to do it, First, you should create an augmented matrix by selecting all desired segments, defining a name, and clicking on the “Create Aug.” button. Then you should input an over-estimated metabolites number of metabolites as the “Number of PCs”. In fact, RMet utilizes the Singular Value Decomposition (SVD) algorithm to calculate Eigenvalue of each metabolite and plots the logarithm of Eigenvalue ratios. In order to perform an SVD analysis, one should determine the number of Principal Components (PCs). In our case, you should have an estimation of the maximum number of metabolites in your sample; for example if you know that the maximum number of metabolites in your sample may not be over than 100 , set the “Number of PCs” on 120 in order to see the Eigenvalue plot of all 120 components and make sure that the actual number of metabolites is less than 100. After inputting “The number of PCs” click on the “Plot” button in order to view the Eigenvalue plot (Figure S20). Note that an SVD analysis make take several minutes for a big augmented matrix and a high number of metabolites, but it usually does not take more than 5 minutes for a 10 Gigabytes matrix with 100 metabolites on an Intel Core i7 desktop processor.

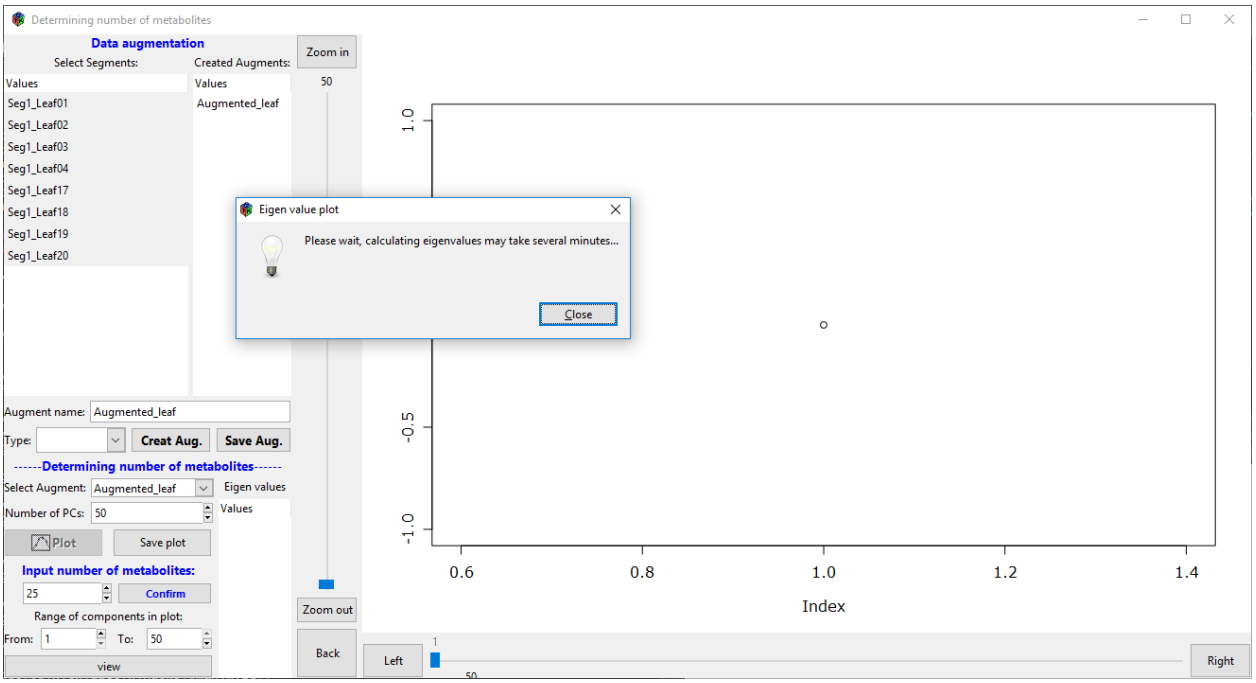

Figure S20. Data augmentation and SVD analysis in RMet.

After the Eigenvalue plot is shown, you should zoom and scroll the plot in order to find the numbers which their Eigenvalues are much more than the next ones, these are the possible number of metabolites. The difference in the amount of Eigenvalue is accompanied by the difference in color, it means that if Eigenvalue of a number is plotted in dark (black) color in and the next one is white it means that their values vary a lot and that might be a possible number of metabolites. After finding all possible number of metabolites should one of them based on your knowledge about the studied biological system. The possible number of metabolites for a sample augmented data matrix is shown in figure S21.

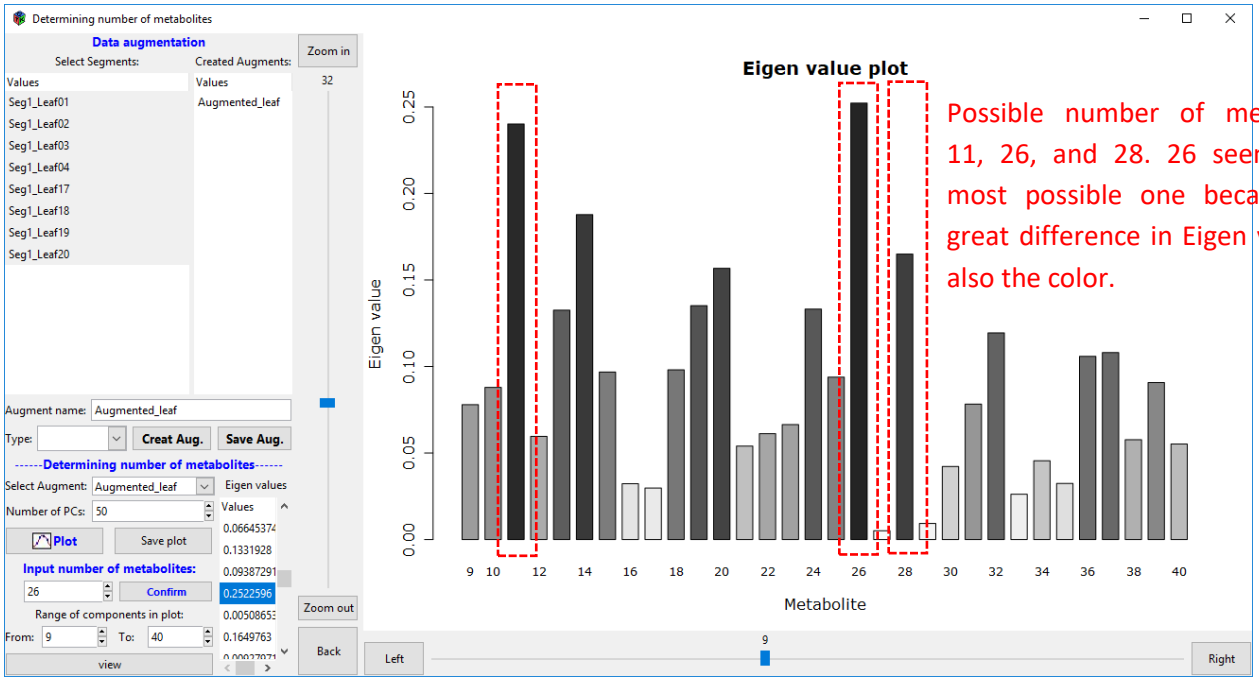

Figure S21. Determining the number of metabolites using the SVD Eigenvalue plot

S2.4. Data compression

If you do not have access to a high-end PC or a High-Performance Computer (HPC) analysis of big GC×GC-MS data will be time-consuming or even impossible. One solution to overcome this challenge is to perform a proper compression algorithm for reducing the data size. RMet uses the Discrete Wavelet Transform (DWT) as its compressing algorithm. Select an augment data matrix and select the level of compression and click on “Compress” button to run the compression by doing so the chromatogram of the compressed data matrix will be shown. In order to find the proper compression level, you should start from the level 1 and increase the compression level until the negative values (plotted in red) take a considerable portion of all values. In our example, negative values are acceptable until level 3, as you can see in figure S22 a great number of negative values are created on compression level 4 it so we cannot go further than 3. Note that data size is reduced by 2 ^ (compression level) times (8 times for level 3).

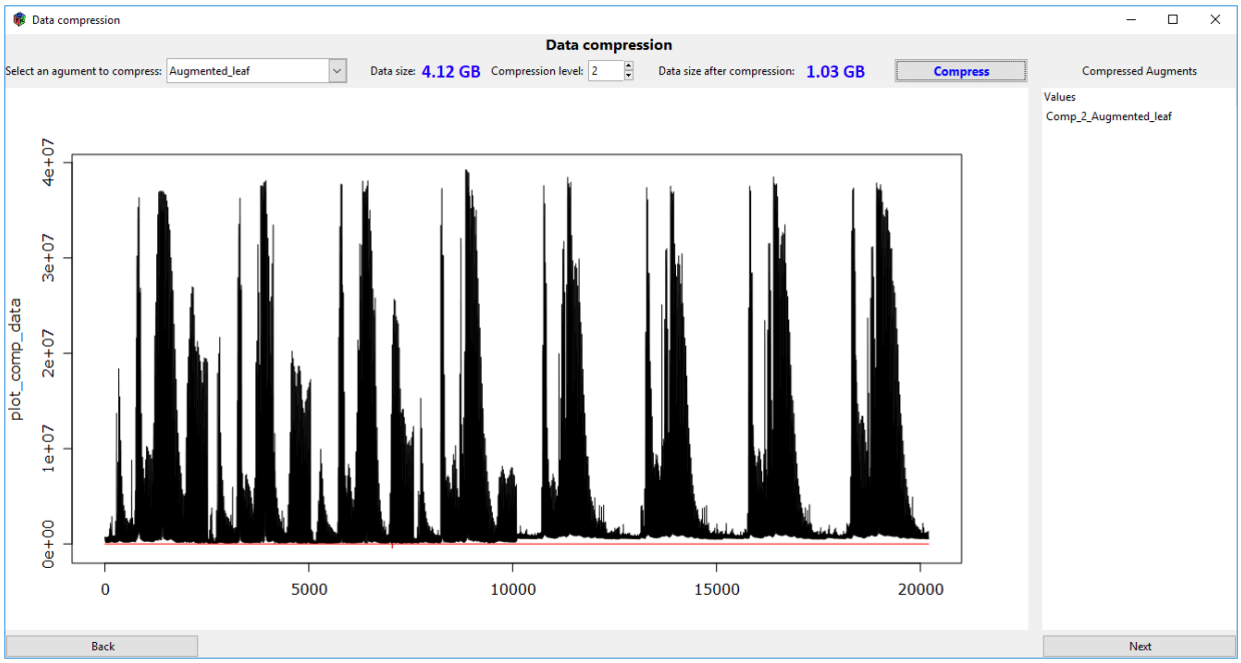

Figure S22. Chromatogram of compressed augmented matrix by 2 levels.

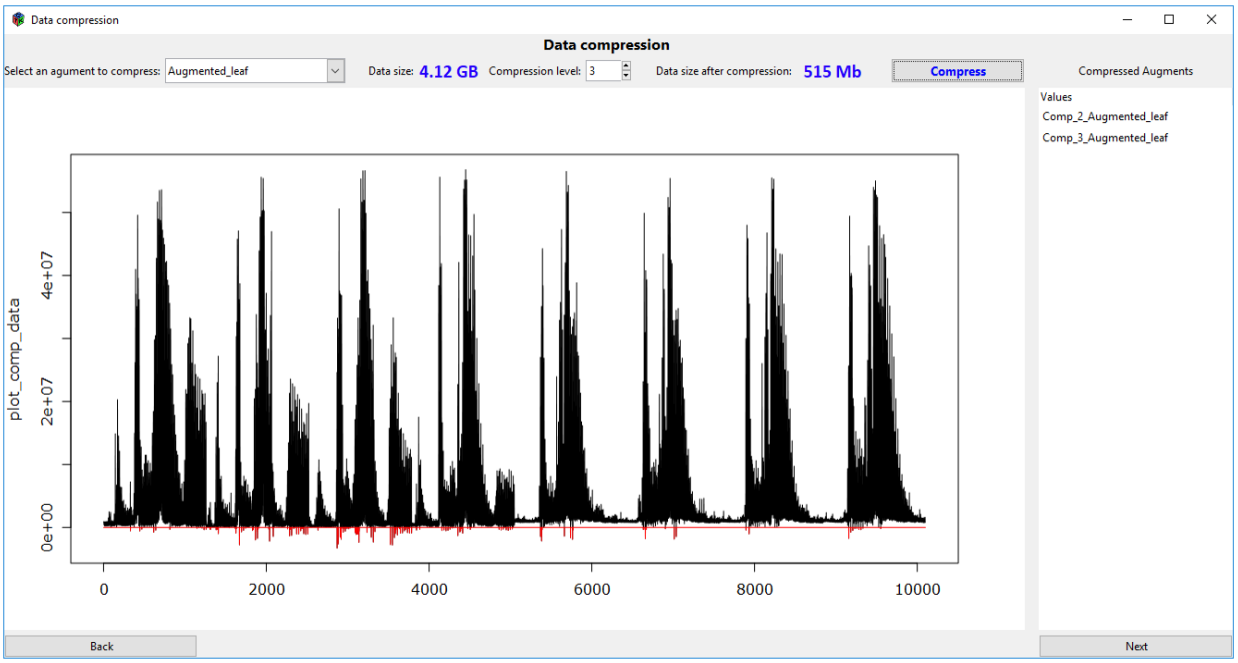

Figure S23. Chromatogram of the compressed augmented matrix by 3 levels.

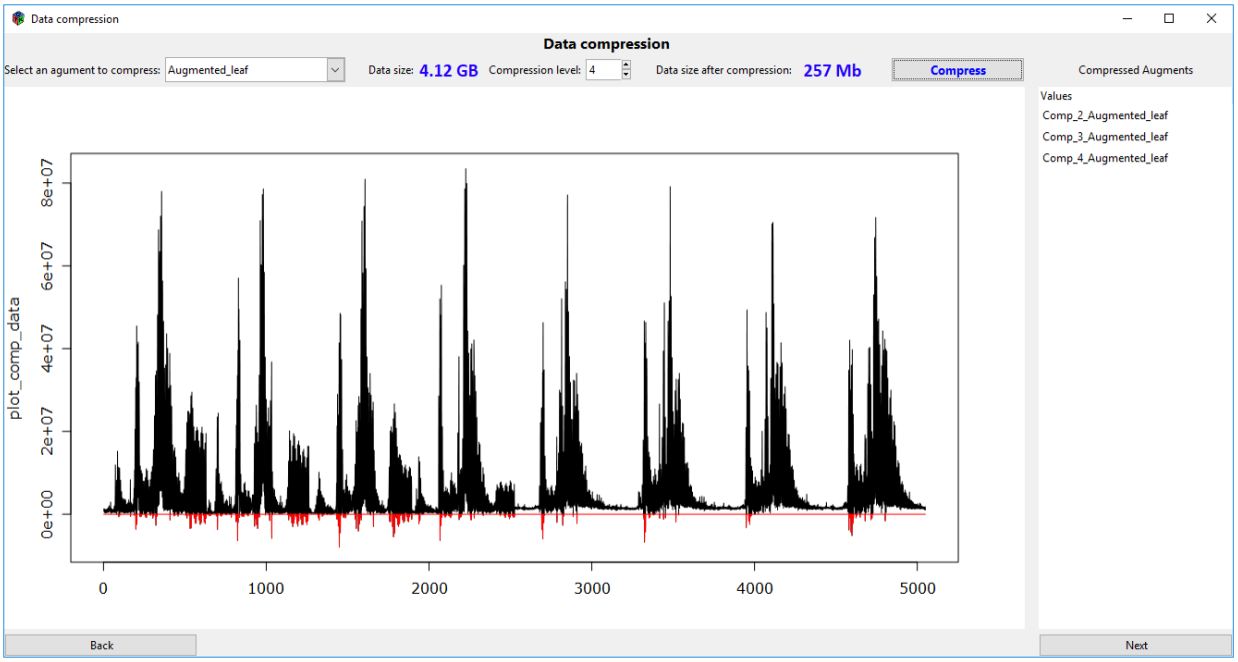

Figure S24. Chromatogram of the compressed augmented matrix by 3 levels.

S2.5. Building model

This step is the most time-consuming one. In this section compressed augmented data matrices will be analyzed using the Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) algorithm. In order to perform an MCR-analysis you should select a data and determine the number of metabolites and number of samples (matrices), it worth mentioning that these are automatically selected by the information provided from the previous steps but you can also change them here. You should also define a name for the result of analysis which are two matrices: Pure mass spectra and the pure elution profiles (which will be automatically transformed to the First column and second column elution profiles for each sample). There are three constraints which are applied in MCR: Non-negativity, closure, and unimodality however only None-negativity constrain is applied in most of the metabolomics study which means that the Intensity in both spectra and elution profile cannot be negative. MCR-ALS is actually an iterative algorithm; it continues until reaching a predefined convergence criterion. Thus you should input a convergence criterion which is a numeric value that defaults to 0.01; if ((old RSS - RSS) / old RSS) < convergence criterion then the optimization stops, where old RSS is the Residual Sum of Squares at iteration x-1 and RSS is the residual sum of squares at iteration x. Although, RMet provide you with the information about whether the optimization is converging in each iteration you should set a maximum for the number of iterations because some analysis is never going to converge. If an optimization does not converge after 500 iterations it is very unlikely to converge in next iterations, so you will avoid wasting time on a diverging optimization by setting a maximum on the number of iterations. Finally, you can start the analysis by clicking on "Run" button. During the optimization, you can see the RSS for each iteration and check whether the optimization is converging or not. Note that if an optimization diverges for 30 continuous iterations it is very unlikely to converge in 31st iteration so it is better to stop the analysis and try with a different number of metabolites or compression level. After the analysis is completed you can see the model statistics include noise level, lack of fit, initial RSS, and final RSS. Pay attention that noise level should be near to lack of fit; if they are very different from each other you may have selected an incorrect number of metabolites. RMet also provides users with a great option which is exporting the preprocessed augmented data matrices and an R code to a High Perform Computer (HPC) in order to perform MCR-ALS analysis. This option will be explained in following pages.

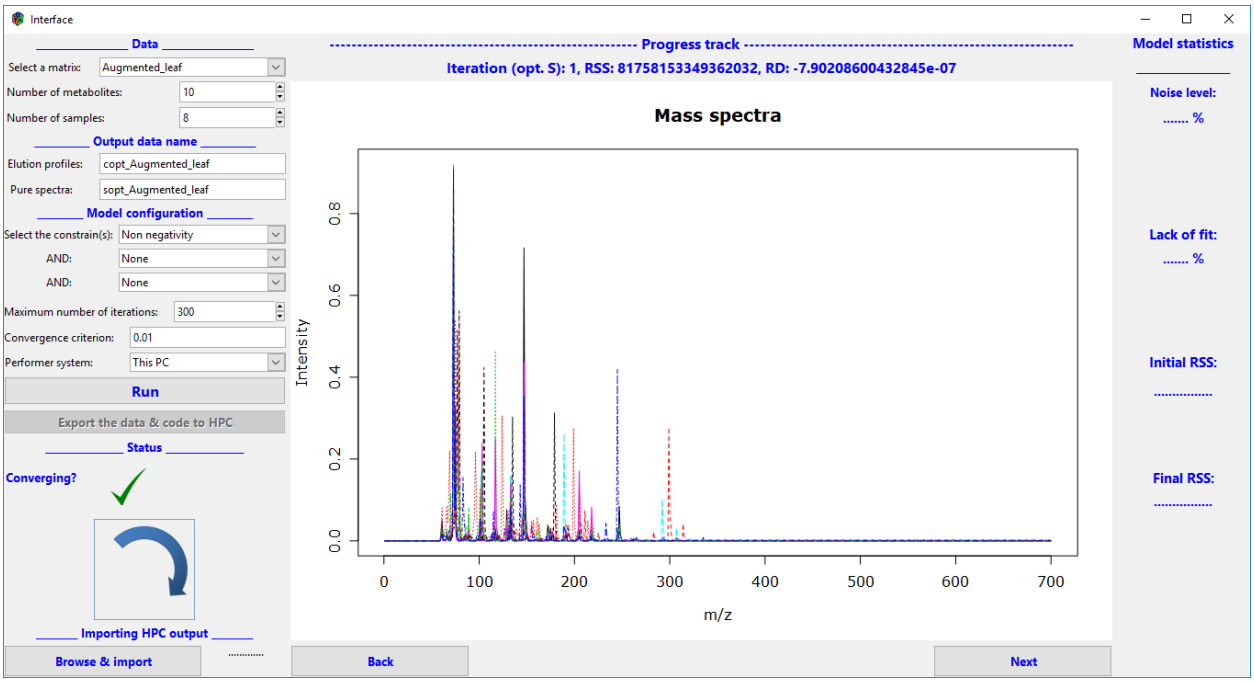

Figure S25. MCR-ALS analysis in RMet

S2.6. Pure profiles

In this section, you can view the resulted chromatograms and mass spectra (Figure S26) by selecting a data matrix and clicking on the "View data" button. You can also view the first and second column elution profile of a specified metabolite in one sample or in all samples (Figure S27).

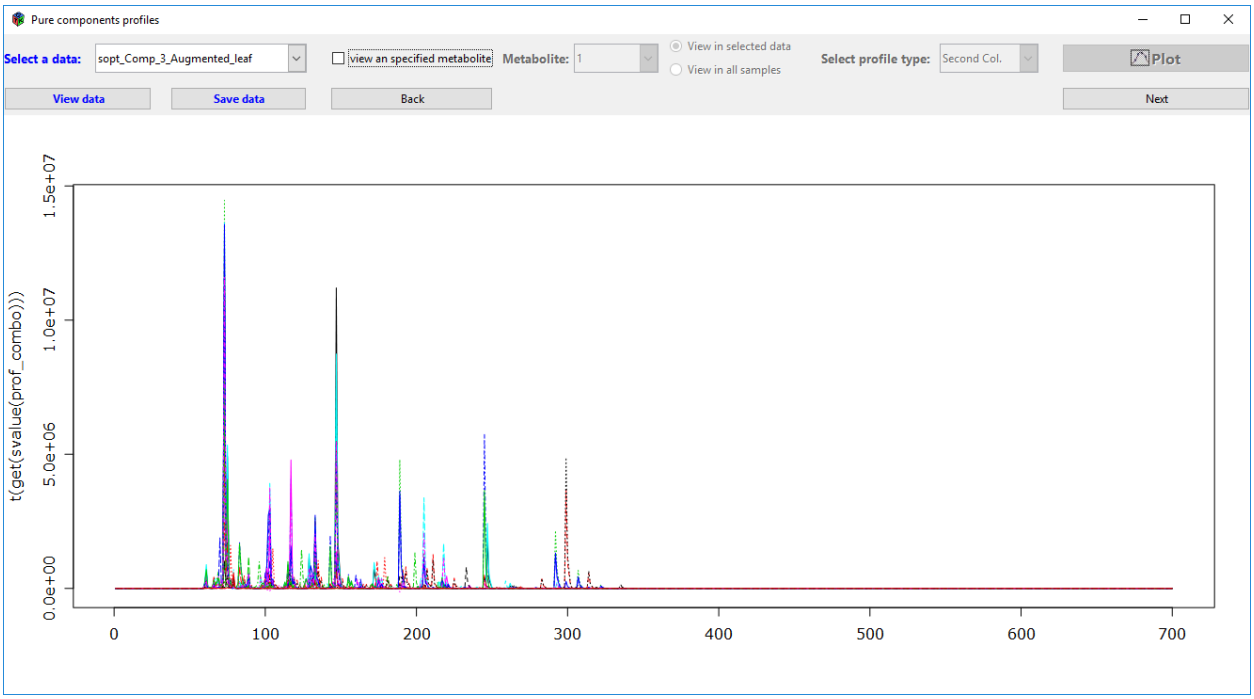

Figure S26. Plotting MCR results in RMet – Pure mass spectra

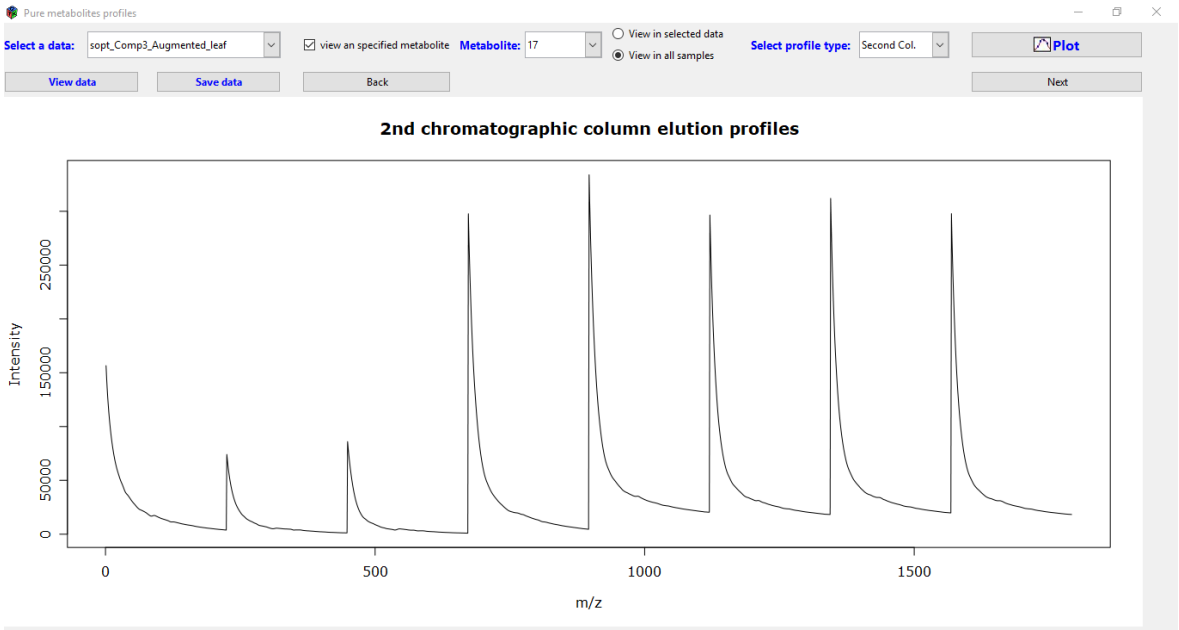

Figure S27. Plotting MCR results in RMet – 2nd Column elution profile for metabolite 13 in all samples

2.7- Metabolites classification

In this section, RMet utilizes the Partial Least Squares - Discriminant Analysis (PLS-DA) in order to identify the important (significantly affected) metabolites. First, you should arrange the 2nd column elution profiles matrices into the right classes (including the control, and exposed classes). Then you should select the preprocessing method for your classification among the auto scaling, mean centering, and scaling. The last parameter which should be determined is the number of Latent Variables (LVs). Its default value is equal to number of classes – 1; in order to find the proper number of LVs you can use the default number to perform the classification and you can check the X Cumulative variance plot and find out the suitable number of LVs which bring you an X cumulative variance over than 90% (Figure S28). You should test different preprocessing methods in order to find the best model in which the classed are highly distinguished from each other. In our example, auto-scaling (Figure S29) and scaling (Figure S30) seems to be good but using only mean-centering (Figure S31) does not seem to be suitable. RMet also provides all required statistical information of the PLS-DA model for the model validation including RMSE, X Residuals, Sensitivity, and so on (Figure S32). After you classified the metabolites click on the "Next button" in order to proceed to the last step.

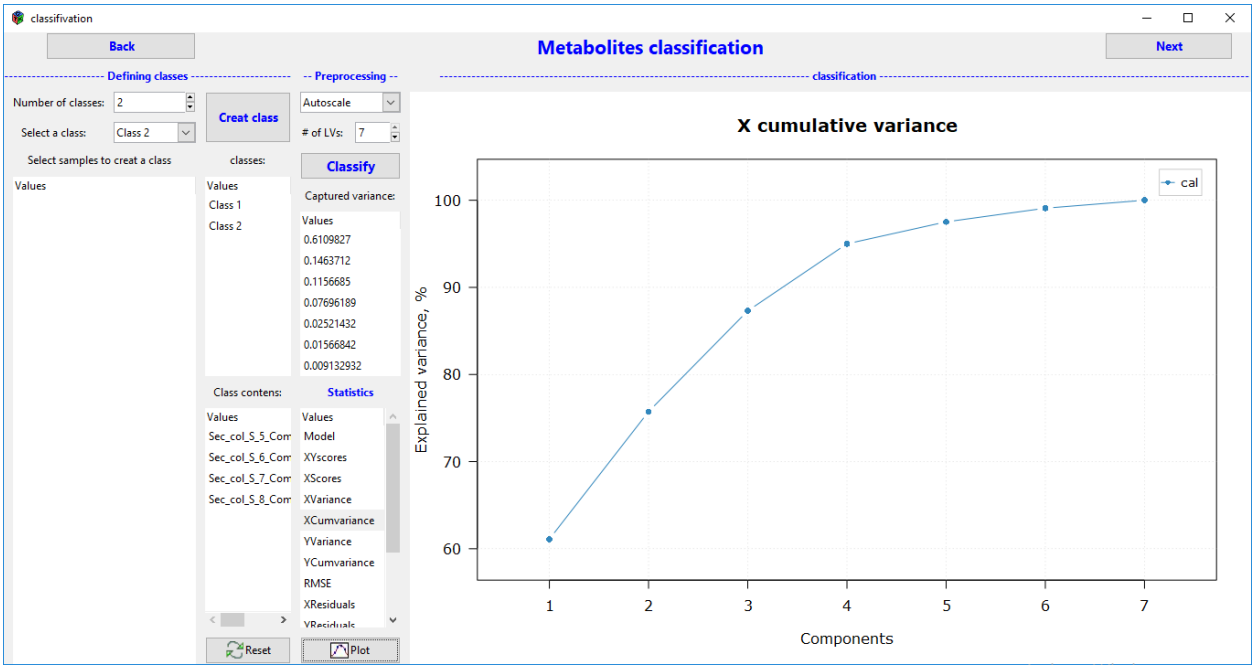

Figure S28. X cumulative variance plot considering 7 LVs, 4 LVs is enough since the cumulative variance will be about 95% with 4 LVs

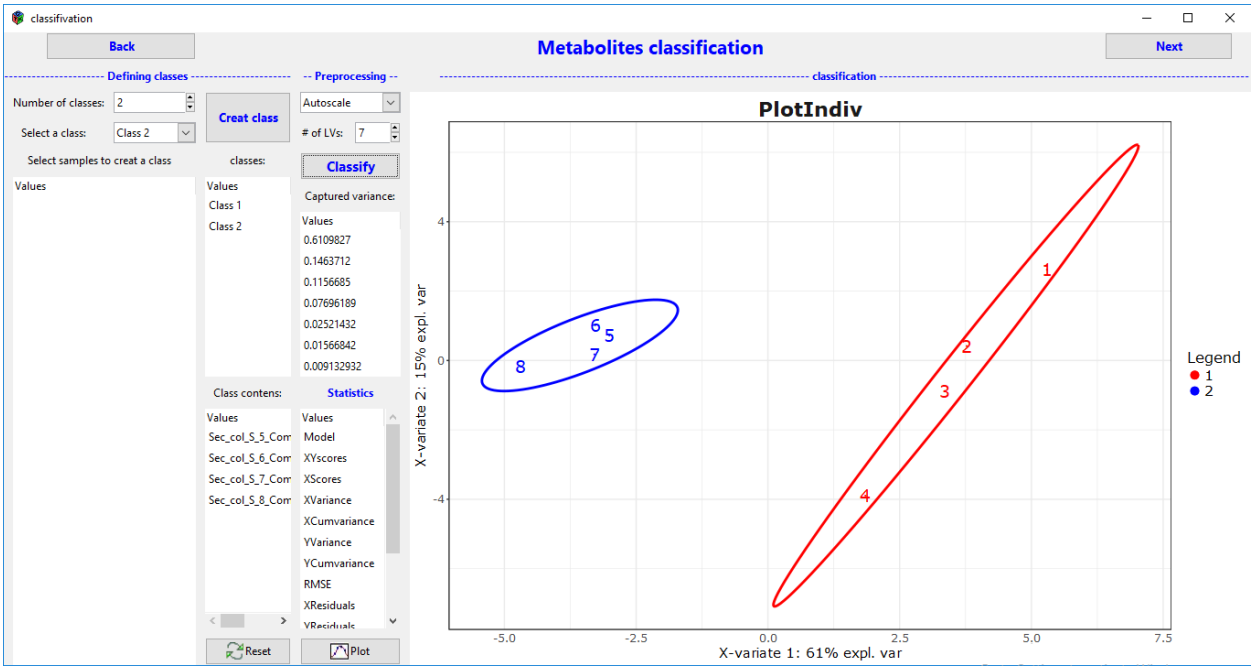

Figure S29. PLS-DA classification result for sample data matrices by auto-scaling preprocessing

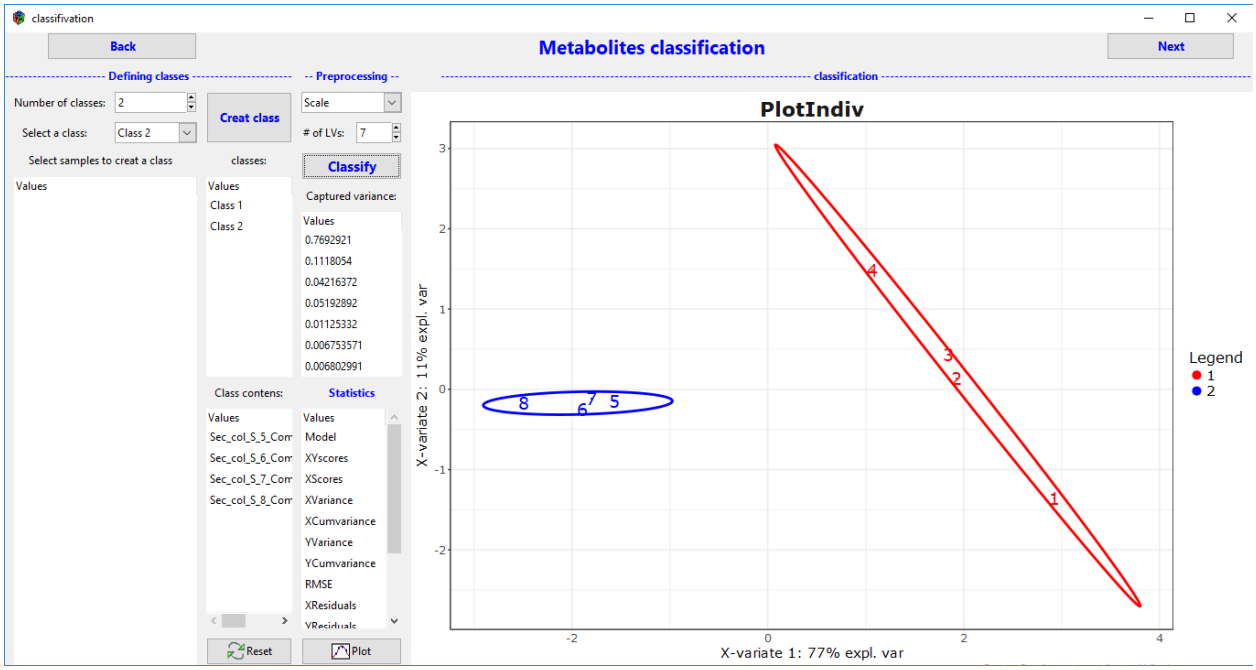

Figure S30. PLS-DA classification result for sample data matrices by Scaling preprocessing

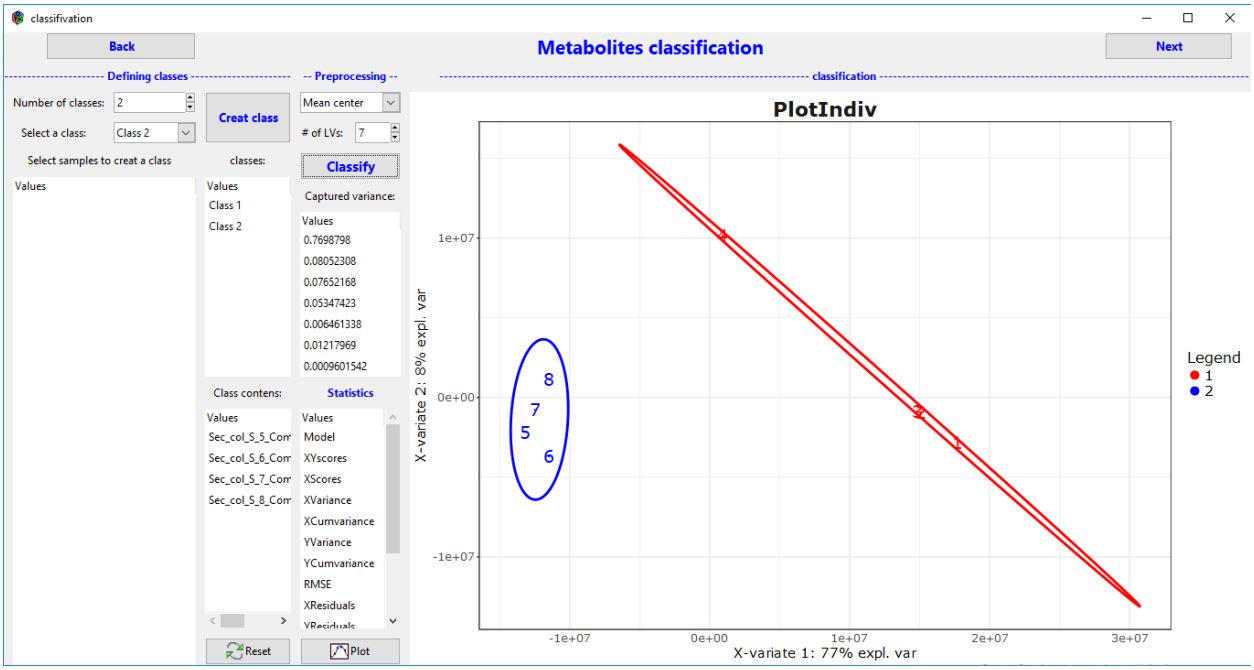

Figure S31. PLS-DA classification result for sample data matrices by Scaling preprocessing

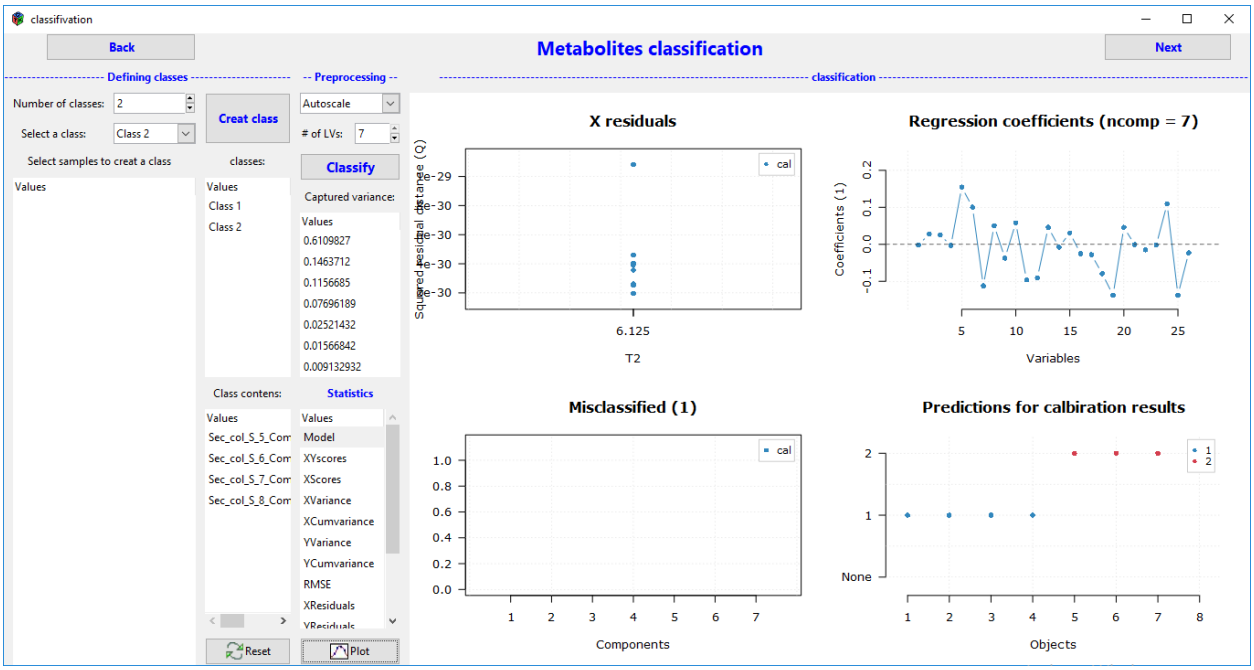

Figure S32. PLS-DA Model statistics for model validation

S2.8. Identification of important metabolites

In this last section, RMet introduces you the important metabolites based on their Variable Importance in Projection (VIP) scores. Click on “View” button in order to make RMet calculate and display the VIP score plot. In this plot, metabolites with a score more than one (the blue ones) are considered as an important metabolite (Figure S34). Finally, you should export the spectra to the NIST software in order to identify the unknown important metabolites based on a parameter called match factor which indicates the similarity level of your unknown metabolite with the NIST proposed compound. In order to export, input a peak selection criterion (the ratio of a peak intensity to the maximum peak intensity); note that although less selection criterion means more accuracy the less you set this parameter the less match factor you obtain in NIST, click on “Export to NIST” button and select a directory in which the NIST compatible text files of metabolites spectra are going to be created (Figure S35). Now you can upload these text file to NIST and identify the important metabolites (Figure S36).

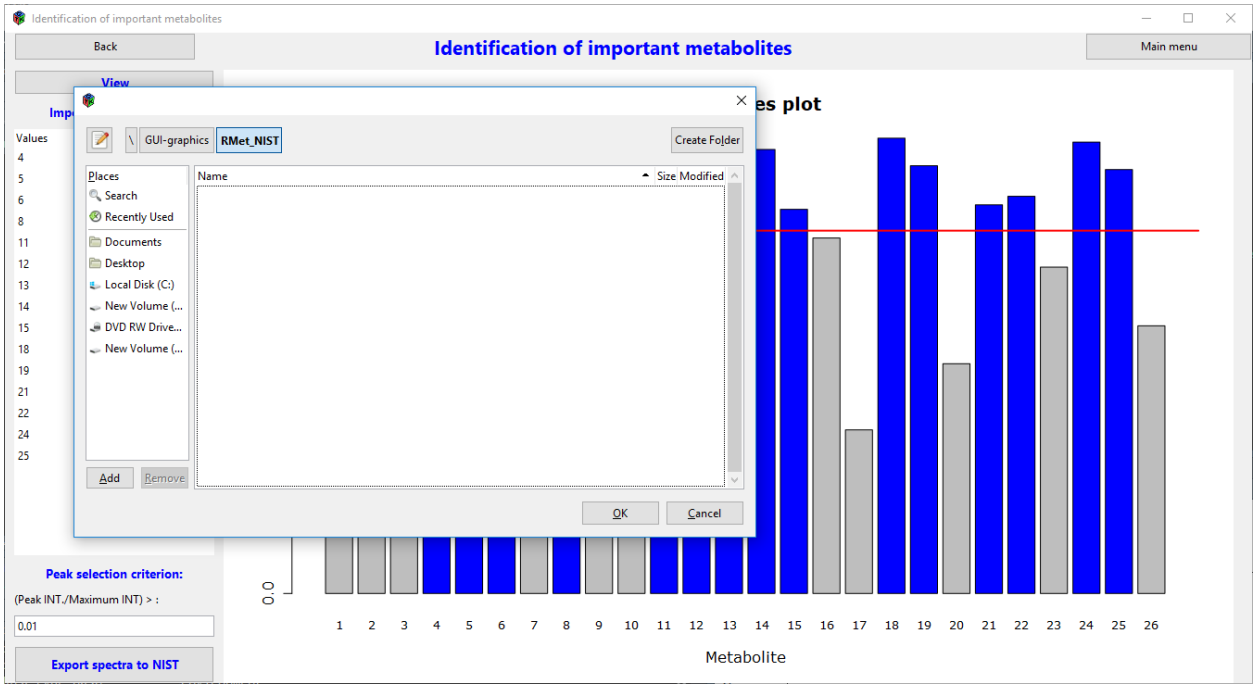

Figure S33. VIP scores for a sample PLS-DA Model and exporting the spectra

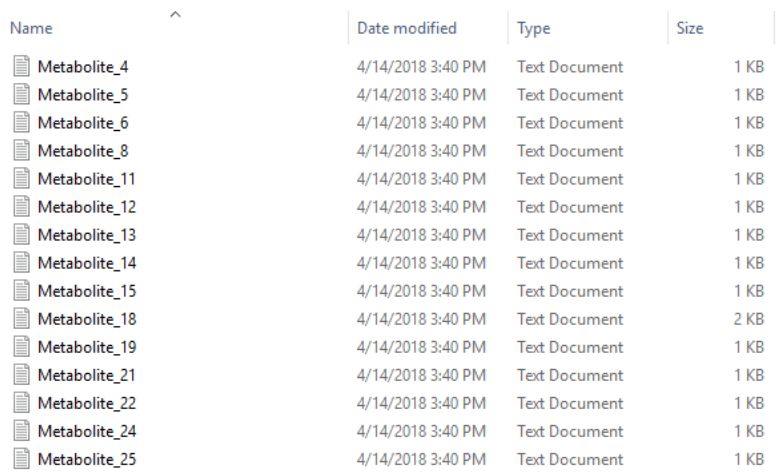

Figure S34. RMet generated NIST compatible text files of unknown metabolites’ spectra

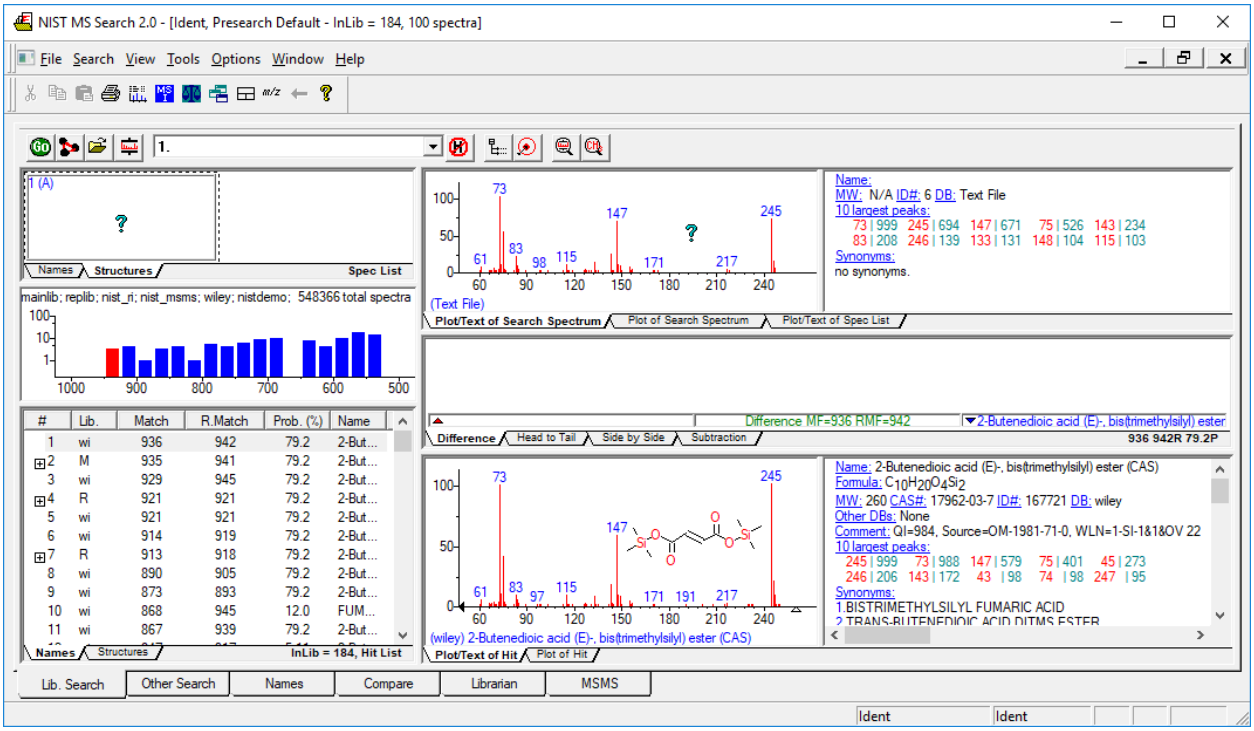

Figure S35. Final NIST search result for an important metabolite

S3. GC-MS data analysis procedure

GC-MS metabolomics is similar to the GC×GC-MS, the only different section will be the “Segmentation & preprocessing” section. For GC-MS data matrices you can only select a range of Retention times by inputting the desired range in spin buttons and clicking on the “Impose button” (Figure S38). All the following steps will be as same as the GC×GC-MS procedures but the compression is can be skipped since GC-MS data sets have usually low sizes.

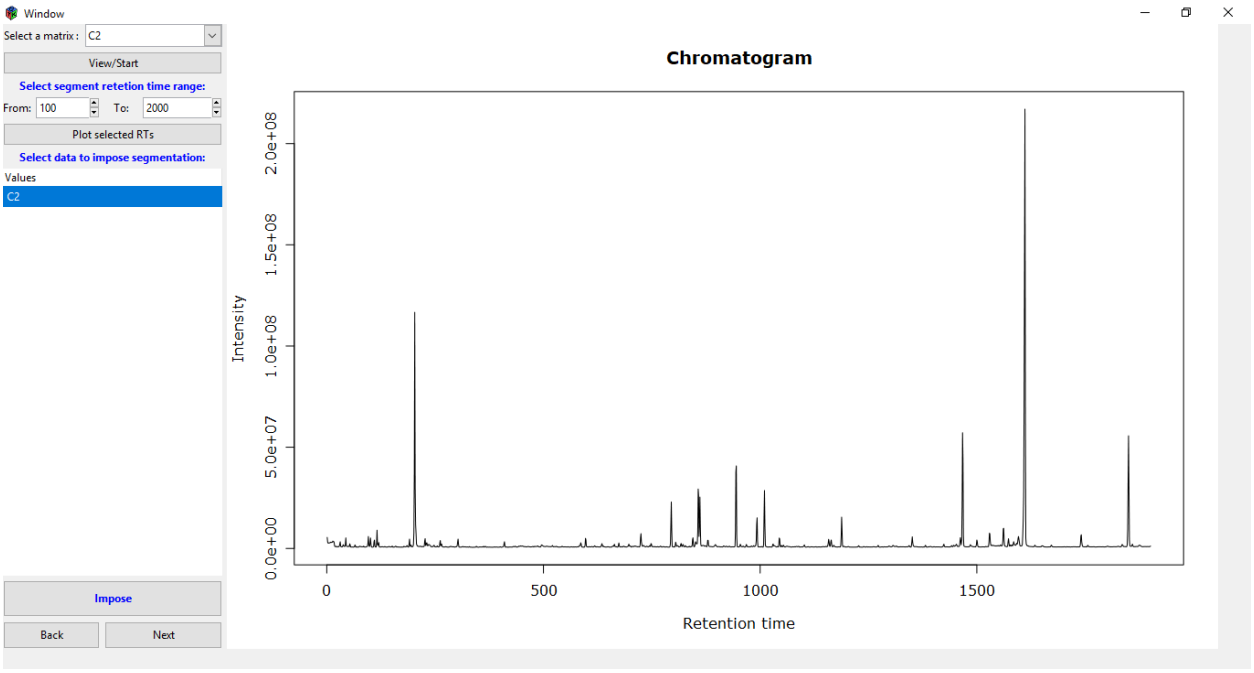

Figure S38. GC-MS data segmentation in RMet