Widely-Targeted Volatilomics (WTV) 2.0 Manual

1 Introduction

WTV2.0 is open-source software with a user-friendly interface, providing a one-stop solution for the entire process of gas chromatography-mass spectrometry (GC-MS) based on widely-targeted volatilomics. The *library builder* manages the integration and deduplication of the mass spectra libraries and retention time (RT) data. The *method generator* selects the characteristic qualitative ions of each compound from the library and constructs the optimized acquisition segments. The *data analyzer* is used for qualitative and semi-quantitative analysis of widely-targeted data. WTV 2.0 is also applicable to GC-MS-based primary metabolites profiling and species beyond plants.

1.1 Library Builder

The *library builder* features mass spectra libraries and retention time information integration and deduplication, retention index (RI) calibration and unknown signals integration. A comprehensive plant volatile (contains 2101 signals) library can be found on GitHub at:

https://github.com/yuanhonglun/WTV_2.0/blob/main/sample_data/library_builder_sample_data/export/Remove_Duplicates.msp

1.2 Method Generator

The *method generator* selects characteristic qualitative ions with the minimum ion number of each compound in the library and performs optimized segmentation. It can generate the comprehensive-Selective Ion Monitoring (cSIM) acquisition method with high sensitivity, coverage and annotation accuracy, or regular widely-targeted methods

based on Selective Ion Monitoring (SIM) mode. A cSIM acquisition method (XML-formatted) can be found on GitHub at:
https://github.com/yuanhonglun/WTV_2.0/blob/main/sample_data/method_generator_sample_data/export/qqqacqmethod.xml

1.3 Data Analyzer

The *data analyzer* performs the qualitative and semi-quantitative analysis of cSIM data, using algorithms similar to that of untargeted data analysis. Note that for regular (widely) targeted data analysis, we recommend using software provided by vendors.

1.4 Other Information

Download link: https://github.com/yuanhonglun/WTV_2.0
Operating system and dependencies: Windows 10 or Windows 11. RAM: 8.0 GB or more. It is recommended to run this software as an administrator.
Installation: No installation required.
Programming language: Python.
License of use: GNU General Public License, version 3 (GPL-3.0).

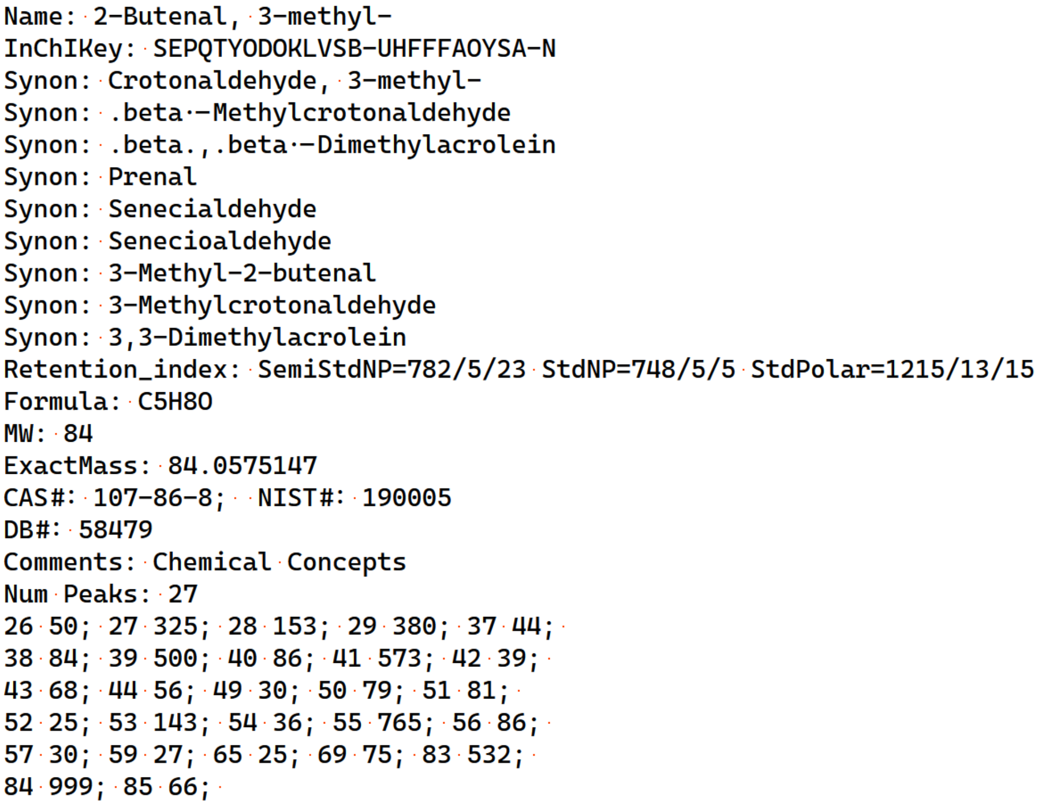
2 Instruction

2.1 Library Builder

2.1.1 Data Preparation

MSP file(s)

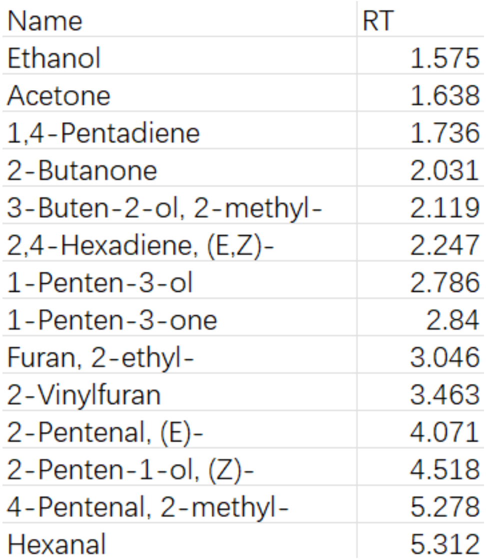
Prepare MSP files that contain mass spectrometry (MS) information and retention index. It is recommended to export MSP files from NIST library.



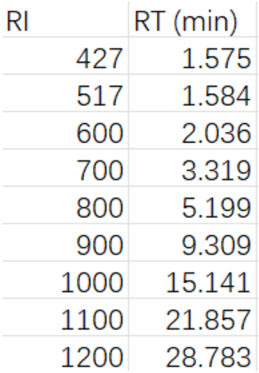
```
Name: 2-Butenal, 3-methyl-
InChIKey: SEPQTYODOKLVSB-UHFFFAOYSA-N
Synon: Crotonaldehyde, 3-methyl-
Synon: .beta.-Methylcrotonaldehyde
Synon: .beta.,.beta.-Dimethylacrolein
Synon: Prenal
Synon: Senecialdehyde
Synon: Senecioaldehyde
Synon: 3-Methyl-2-butenal
Synon: 3-Methylcrotonaldehyde
Synon: 3,3-Dimethylacrolein
Retention_index: SemiStdNP=782/5/23 StdNP=748/5/5 StdPolar=1215/13/15
Formula: C5H8O
MW: 84
ExactMass: 84.0575147
CAS#: 107-86-8;  NIST#: 190005
DB#: 58479
Comments: Chemical Concepts
Num Peaks: 27
26 50; 27 325; 28 153; 29 380; 37 44; 
38 84; 39 500; 40 86; 41 573; 42 39; 
43 68; 44 56; 49 30; 50 79; 51 81; 
52 25; 53 143; 54 36; 55 765; 56 86; 
57 30; 59 27; 65 25; 69 75; 83 532; 
84 999; 85 66; 
```

Retention Time Information

Prepare RT files that contain compound names and measured RT. Prepare retention index (RI) calibration data. Note that in addition to alkane standards, the *library builder* was compatible with using of RT and RI from other compounds in the RI calibration data.



Name	RT
Ethanol	1.575
Acetone	1.638
1,4-Pentadiene	1.736
2-Butanone	2.031
3-Buten-2-ol, 2-methyl-	2.119
2,4-Hexadiene, (E,Z)-	2.247
1-Penten-3-ol	2.786
1-Penten-3-one	2.84
Furan, 2-ethyl-	3.046
2-Vinylfuran	3.463
2-Pentenal, (E)-	4.071
2-Penten-1-ol, (Z)-	4.518
4-Pentenal, 2-methyl-	5.278
Hexanal	5.312



RI	RT (min)
427	1.575
517	1.584
600	2.036
700	3.319
800	5.199
900	9.309
1000	15.141
1100	21.857
1200	28.783

2.1.2 *Library Builder* Algorithm

The imported MSP-formatted library files and CSV-formatted retention information files were integrated respectively. The input of library files is mandatory while input of RT and RI files is optional. After formatting standardization of compound names (e.g., replace ".alpha." with "alpha") and removing the invalid mass spectrum, the redundant information was excluded based on compound names, synonyms, and CAS numbers. The *library builder* verifies whether compound names in the RT list match those in the library, removes compounds not found in the library and unifies the compound names. Duplicates in the RT list were removed according to compound names. Using RI calibration data, the *library builder* calculates the theoretical RT using the equation (equation 2.1-1) from AMDIS. The theoretical RT of the target compound (tc) is calculated by the previous reference compound (prc) and the next reference compound (nrc).

$$theoretical\ RT_{tc} = RT_{prc} + \frac{(RI_{tc}-RI_{prc})\times(RT_{nrc}-RT_{prc})}{RI_{nrc}-RI_{prc}} \quad \text{(equation 2.1-1)}$$

The *library builder* flags compounds with user-defined deviations between the library's RI value and the calculated RI value. Users have the option to replace the measured RT with the theoretical RT. Depending on the imported retention information, the *library builder* will export measured RT, theoretical RT, calibrated RT (including theoretical RT), or do not export RT list.

Finally, the non-redundant library and RT list were exported. The *library builder* supports the integration of unknown signals: the integrated library was used as background, and imported unknowns were compared with all adjacent compounds. Only the unknowns with spectral similarity scores lower than the user-defined threshold were incorporated into the library.

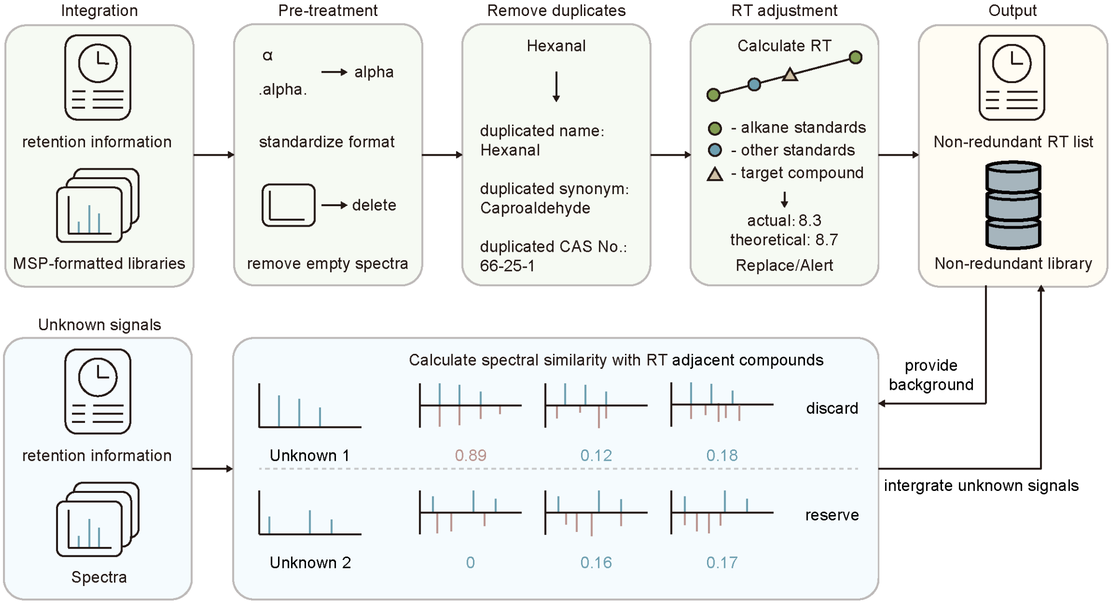
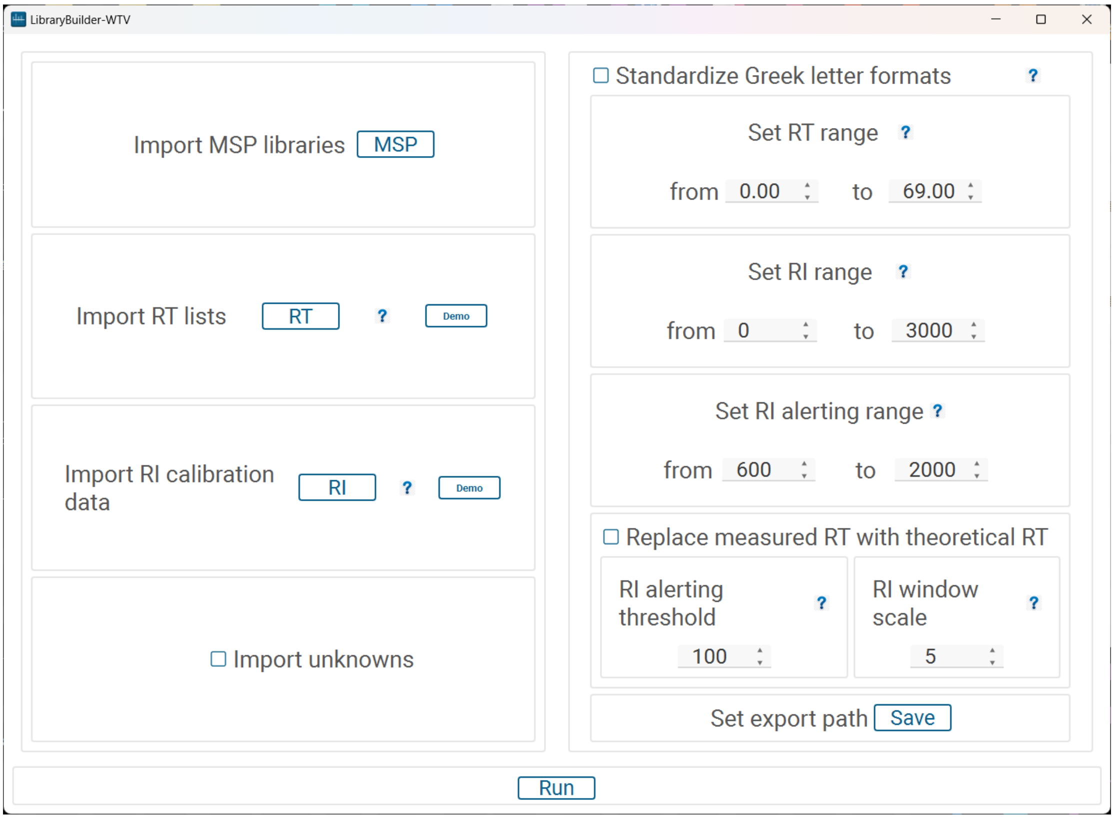


Fig. 2.1-1 The workflow of the *library builder* module of WTV 2.0.

2.1.3 User Interface of The *Library Builder*


LibraryBuilder-WTV

Import MSP libraries MSP

Import RT lists RT ? Demo

Import RI calibration data RI ? Demo

□ Import unknowns

□ Standardize Greek letter formats ?

Set RT range ?
from 0.00 to 69.00

Set RI range ?
from 0 to 3000

Set RI alerting range ?
from 600 to 2000

□ Replace measured RT with theoretical RT

RI alerting threshold ? 100

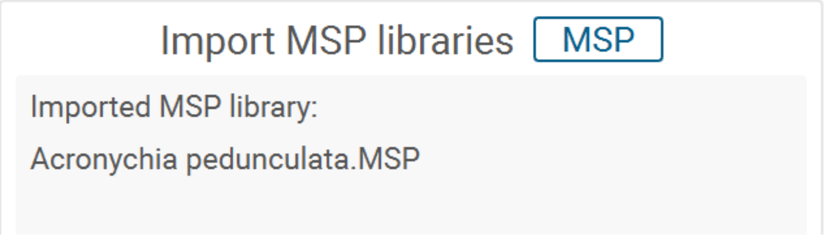
RI window scale ? 5

Set export path Save

Run

2.1.4 Parameter Configuration

Import MSP library: To import an MSP file, click the 'MSP' button and select the MSP file. Once selected, the interface will display all the selected files, confirming a successful input.


Import MSP libraries MSP

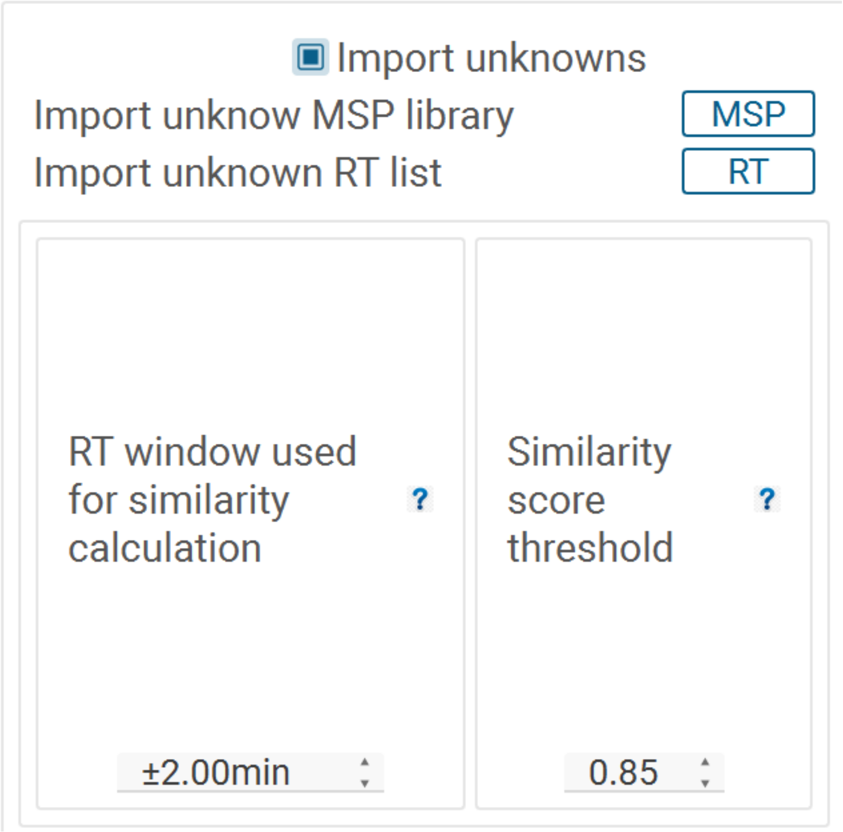
Imported MSP library:

Acronychia pedunculata.MSP

Import RT list: Import a list of compounds and their RT in CSV format.

Import RI calibration data: Import RI calibration data in CSV format.

Import unknowns: Import MS and RT information of unknown signals.



☑ Import unknowns

Import unknow MSP library

Import unknown RT list

RT window used for similarity calculation ±2.00min

Similarity score threshold 0.85

Similarity is calculated between unknowns and compounds within the user-defined RT window. Only when the similarity is below the similarity score threshold with all compounds in the RT window will the unknown be integrated into the library.

Standardize Greek letter formats: *e.g.*: ".alpha." -> "alpha"

Set RT range: The output retention times will be constrained within this range.

Set RI range: The output retention index will be constrained within this range.

Set RI alerting range: Compounds with the RI falling within this range will be flagged.

Replace measured RT with theoretical RT: Selecting this option will replace the measured RT of flagged compounds with calculated RT.

RI alerting threshold: When the deviation between the library's RI value and the calculated RI value exceeds this threshold, the compound will be flagged.

RI window scale: The larger this value, the larger the RI alerting threshold when the RI is larger. Setting it to 0 disables this feature. For more information, please refer to the AMDIS manual.

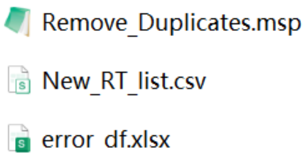
Set export path: Click the 'Save' button to set the export path.

Tips:

1. Any file directory entered can be deleted by double-clicking the entered entry.
2. Clicking the corresponding "demo" button allows viewing the required format.
3. Clicking the corresponding "?" button allows viewing the help tip.

2.1.5 Results and Reports

After configuring all parameters, click the 'Run' button. The program will automatically generate three result files.

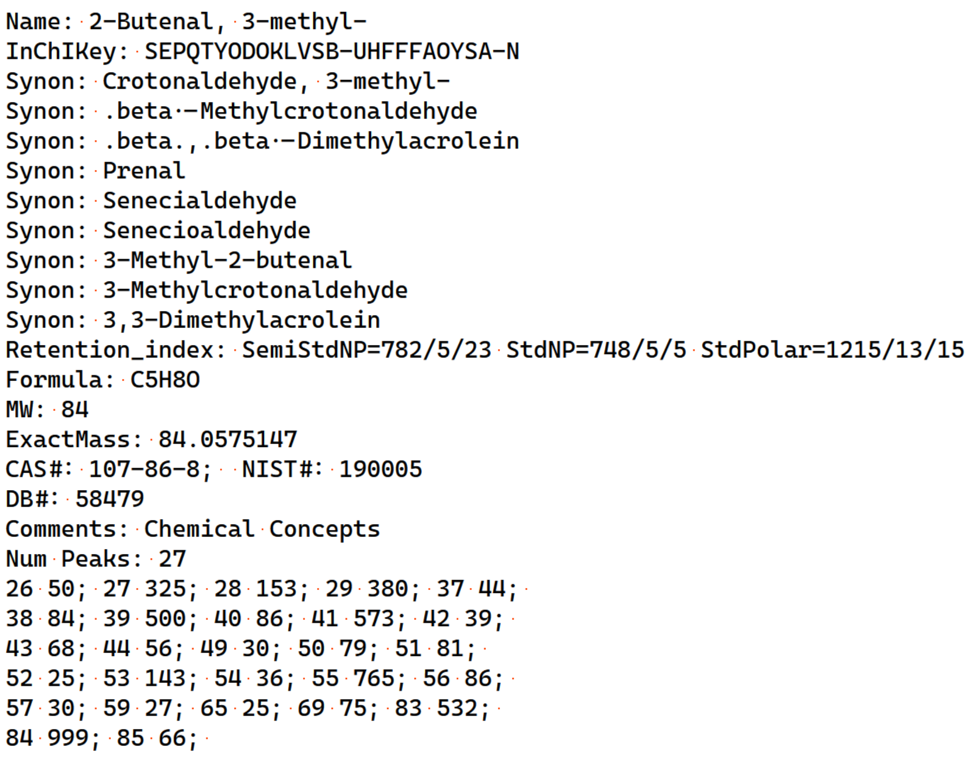


Remove_Duplicates.msp

New_RT_list.csv

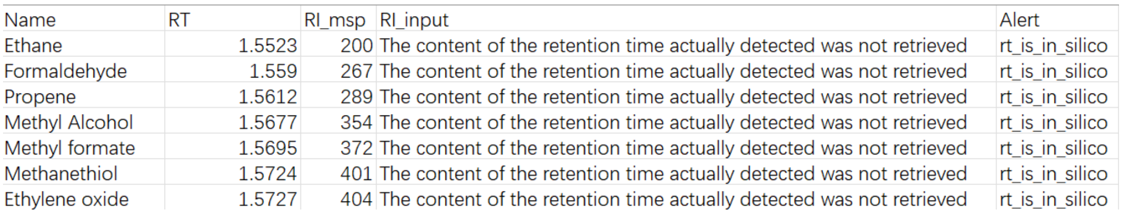
error df.xlsx

The MSP result file: Contains the non-redundant MS information



```
Name: 2-Butenal, 3-methyl-
InChIKey: SEPQTYODOKLVSB-UHFFFAOYSA-N
Synon: Crotonaldehyde, 3-methyl-
Synon:  .beta.-Methylcrotonaldehyde
Synon:  .beta.,.beta.-Dimethylacrolein
Synon: Prenal
Synon: Senecialdehyde
Synon: Senecioaldehyde
Synon: 3-Methyl-2-butenal
Synon: 3-Methylcrotonaldehyde
Synon: 3,3-Dimethylacrolein
Retention_index: SemiStdNP=782/5/23 StdNP=748/5/5 StdPolar=1215/13/15
Formula: C5H8O
MW: 84
ExactMass: 84.0575147
CAS#: 107-86-8;  NIST#: 190005
DB#: 58479
Comments: Chemical Concepts
Num Peaks: 27
26 50; 27 325; 28 153; 29 380; 37 44;
38 84; 39 500; 40 86; 41 573; 42 39;
43 68; 44 56; 49 30; 50 79; 51 81;
52 25; 53 143; 54 36; 55 765; 56 86;
57 30; 59 27; 65 25; 69 75; 83 532;
84 999; 85 66;
```

The RT result file: Contains the non-redundant RT information.



Name	RT	RI_msp	RI_input	Alert
Ethane	1.5523	200	The content of the retention time actually detected was not retrieved	rt_is_in_silico
Formaldehyde	1.559	267	The content of the retention time actually detected was not retrieved	rt_is_in_silico
Propene	1.5612	289	The content of the retention time actually detected was not retrieved	rt_is_in_silico
Methyl Alcohol	1.5677	354	The content of the retention time actually detected was not retrieved	rt_is_in_silico
Methyl formate	1.5695	372	The content of the retention time actually detected was not retrieved	rt_is_in_silico
Methanethiol	1.5724	401	The content of the retention time actually detected was not retrieved	rt_is_in_silico
Ethylene oxide	1.5727	404	The content of the retention time actually detected was not retrieved	rt_is_in_silico

The warning information file: Contains the warning information, including:

WARNING: The file format cannot be recognized

WARNING: This compound was not found in the provided MSP library.

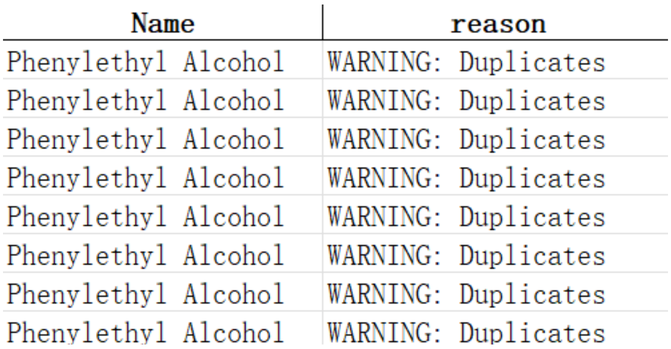
WARNING: The RI of this compound is 0.

WARNING: The RT value is out of the setting range.

WARNING: The synonym name has been changed to unified Name.

WARNING: The synonym name was not found in the library.

WARNING: Duplicates



Name	reason
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates
Phenylethyl Alcohol	WARNING: Duplicates

Tips: A demo results can be found at:

https://github.com/yuanhonglun/WTV_2.0/tree/main/sample_data/library_builder_sample_data/export

2.2 Method Generator

2.2.1 Data Preparation

Prepare library and RT files that contains MS and RT information for all targeted compounds. The results of *library builder* can be directly imported to the *method generator.*

2.2.2 *Method Generator* Algorithm

The *method generator* operates on the MSP-formatted library file and CSV-formatted RT and compound list. While the input of the library and RT list are mandatory, the input of the compound list is optional. The compound list contains part of the compound names in the RT list. If imported, the developed acquisition method will only contain the compounds in the list.

For qualitative ions selection, initially, the minimum ion number (m) was defined. For each target compound, the software calculates the number of available ions after screening by intensity and *m/Z* threshold, and the compound was excluded if the available ion number was less than 2. Then, compounds with adjacent RT were recorded according to the user-defined RT window. If the target compound has no adjacent compounds, ion combination that contained m ions were listed, and the weight score of each ion combination was calculated using equation (2.3-1), The ion combination with the highest weight score was selected as the qualitative ions. In particular, if the ion's *m/Z* was smaller than the user-defined preferred *m/Z* threshold, its value was set to 1 (instead of intensity$^{0.5}$×m/Z^3) in weight score calculation.

$$\text{weight score} = \sum_{i=1}^{n} \sqrt{intensity} \times m/Z^3 \qquad (2.3\text{-}1)$$

If adjacent compounds were present, the software first calculates similarity scores between the target compound and its adjacent compounds (equation 2.3-2~4). Users can define the minimum ion number to introduce the F_R term (F_R factor). When the ion number is less than the F_R factor, the similarity score uses the F_D value instead of the composite score. This feature lowers the overestimated similarity score when the ion number is limited (e.g., less than 3).

$$F_D = \sqrt{\frac{\left(\sum_{i=1}^{n}(x_i \times y_i)\right)^2}{\sum_{i=1}^{n}(x_i)^2 \times \sum_{i=1}^{n}(y_i)^2}} \quad (x_i or\ y_i = \sqrt{intensity} \times m/Z^2) \qquad (2.3\text{-}2)$$

$$F_R = \frac{1}{N_{L\&U}} \times \sum_{i}^{L\&U} \left(\frac{W_{Li}}{W_{Li-1}} \times \frac{W_{Ui-1}}{W_{Ui}}\right)^n (W = intensity) \qquad (2.3\text{-}3)$$

where n = 1 or -1 when the term in parentheses is less than or greater than unity, respectively

$$Composite\ score = \frac{N_U \times F_D + N_{L\&U} \times F_R}{N_U + N_{L\&U}} \qquad (2.3\text{-}4)$$

Next, adjacent compounds that exceeding the user-defined similarity score threshold were excluded. This is necessary because subsequent ion selection aims to achieve spectral separation of the target compound from all adjacent compounds with less ions. If an adjacent compound already exhibits high similarity to the target compound when the mass spectrum is complete, it is impossible to achieve spectral separation with less ions. In data analysis, these compounds are distinguished from the target compound based on RT or RI. If no adjacent compounds were present after exclusion, the qualitative ions were selected based on weight score.

The software selects the qualitative ions according to the workflow described below. Firstly, starting from one qualitative ion, it uses the modified cosine similarity score to calculate the separation score (see below), and selects the characteristic qualitative ions with the minimum ion number. For instance, to select qualitative ions for compound X (Fig. 2.3-1a), adjacent compounds (A, B, C, and D) were recorded according to the user-defined RT window (default 2.0 minutes), then the low-intensity and unwanted ions were excluded based on user-defined intensity and *m/Z* thresholds (default 5% of maximum intensity and 35-500, respectively), which resulted in the remaining of ions 57, 71, 96, and 128. Afterwards, the modified cosine similarity scores between target compound X and each adjacent compound were calculated, and adjacent compounds showing similarity scores above the user-defined threshold (default 0.85) were excluded (e.g., compound D), which enables subsequent ion selection. The software then selects qualitative ions by listing possible qualitative ion combinations for compound X (in Fig. 2.3-1, the ion number of the current round (n) starting from two for better illustration), which resulted in a total of six ion combinations that contained two ions, including 57-71, 57-96 and 96-128, etc. (Fig. 2.3-1b). The similarity score was calculated between each ion combination of X and each adjacent compound. The number of adjacent compounds with a similarity score lower than the threshold was recorded and referred to separate number. For instance, the separated number of 57-71, 57-96 and 96-128, was 1, 2 and 2, respectively. The separation score was then calculated by dividing the separated number by the total number of adjacent compounds. Accordingly, the separation score of 57-71, 57-96 and 96-128, was 0.33, 0.67 and 0.67, respectively.

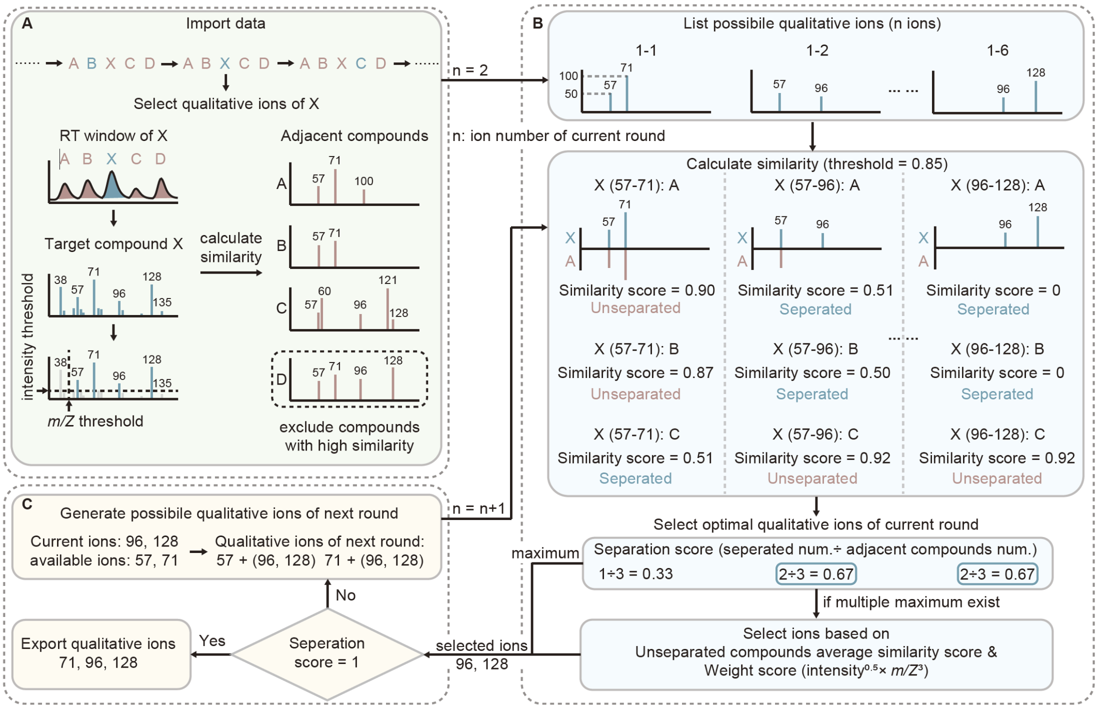


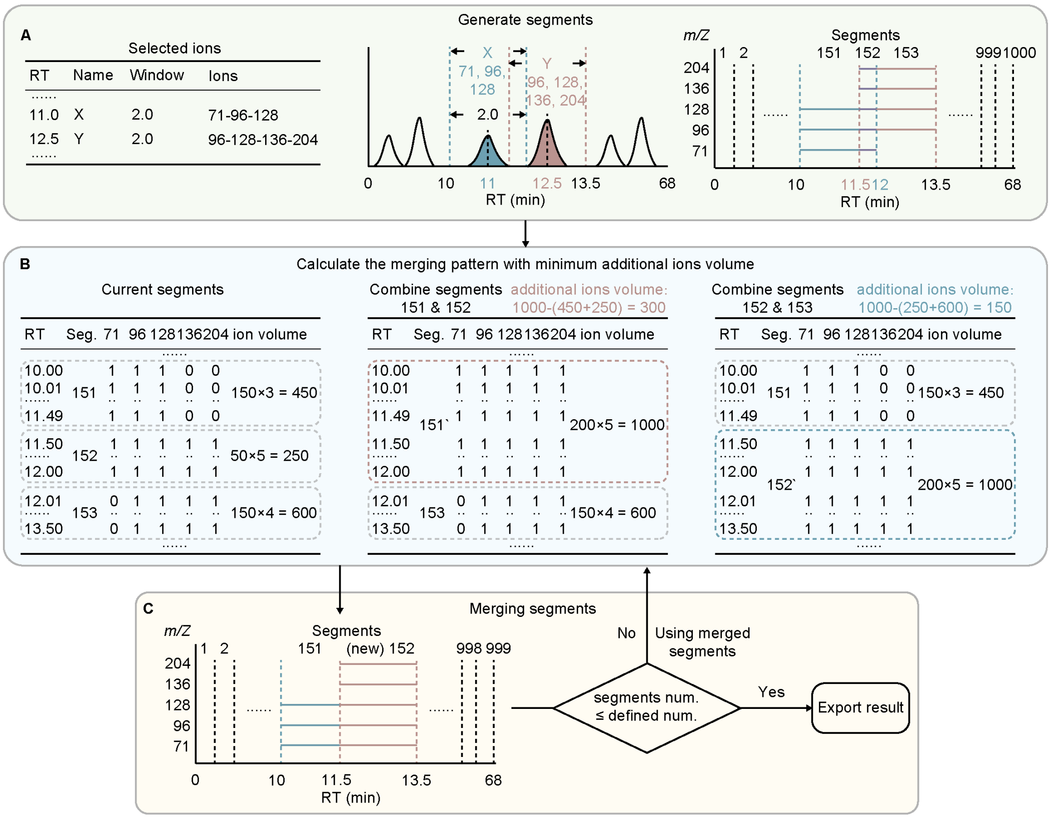
Fig. 2.3-1 The workflow of *method generator* module

The combination with the highest separation score was selected as the optimal ion combination of the current round. If multiple ion combinations have the same highest separation score, adjacent compounds with similarity scores above the threshold were collected, their average similarity score was calculated, the ion combination with lowest average similarity score was selected. If multiple ion combinations have the same lowest average similarity score, they were compared for weight score. The combination with the highest weight score was selected.

At the end of each round, the software assesses whether the separation score of the current optimal ion combination equals one, and ion number was greater than or equal to minimum ion number (m). If met, this optimal ion combination was designated as the qualitative ions for the target compound. Otherwise, the software assesses the remaining ion number. If remaining ion number equals 0 and the separation score was 1, the ion combination was still selected as the qualitative ions. If remaining ion number equals 0 and the separation score was less than 1, the target compound was discarded. If there are remaining ions, the possible ion combinations for the next round were generated by including one more remaining ion. This iterates until the qualitative ions was selected or the target substance was discarded. For compound X, 71-96-128 was selected as the final qualitative ions.

After the selection of qualitative ions for each compound, the software produces a table containing the compound RT, qualitative ions, and RT window. This table can be employed to create a SIM mode acquisition method using vendor-provided software.

The *method generator* can create SIM mode acquisition method based on selected qualitative ions. Based on the selected ions, the ions to be detected at each time point can be determined. Time points with the same ions to be detected can spontaneously generate a segment. For instance, compound X and Y have qualitative ions of 71-96-128 and 96-128-136-204, with RT of 11.0 and 12.5 min, respectively. Based on a 2.0-minutes RT window, the ions to be detected in each time point from 10.0 to 13.5 minutes were determined, and segments 151, 152, and 153 were generated accordingly (Fig. 2.3-2a). The ion volume was calculated by multiplying the number of time points by the number of ions to be detected, e.g., the ion volume of segment 151, 152 and 153 was 450, 250 and 600, respectively (Fig. 2.3-2b). At this point, the segment number and total ion volume reached its maximum and minimum, respectively. Merging segments 151 and 152, 152 and 153, resulted in an additional ion volume of 300 and 150, respectively. Hence, segments 152 and 153 were merged to minimize the additional ion volume. In each round, the additional ion volumes of all merging patterns were calculated, pattern with the minimum additional ion volume was selected for merging. This process was repeated until the number of segments was reduced to or below the maximum number allowed by the instrument (Fig. 2.3-2c).



RT	Name	Window	Ions
......			
11.0	X	2.0	71-96-128
12.5	Y	2.0	96-128-136-204
......			

Current segments

RT	Seg.	71	96	128	136	204	ion volume
10.00		1	1	1	0	0	
10.01	151	1	1	1	0	0	150×3 = 450
11.49		1	1	1	0	0	
11.50	152	1	1	1	1	1	50×5 = 250
12.00		1	1	1	1	1	
12.01	153	0	1	1	1	1	150×4 = 600
13.50		0	1	1	1	1	

Combine segments 151 & 152 additional ions volume: 1000-(450+250) = 300

RT	Seg.	71	96	128	136	204	ion volume
10.00		1	1	1	1	1	
10.01		1	1	1	1	1	
11.49	151`	1	1	1	1	1	200×5 = 1000
11.50		1	1	1	1	1	
12.00		1	1	1	1	1	
12.01	153	0	1	1	1	1	150×4 = 600
13.50		0	1	1	1	1	

Combine segments 152 & 153 additional ions volume: 1000-(250+600) = 150

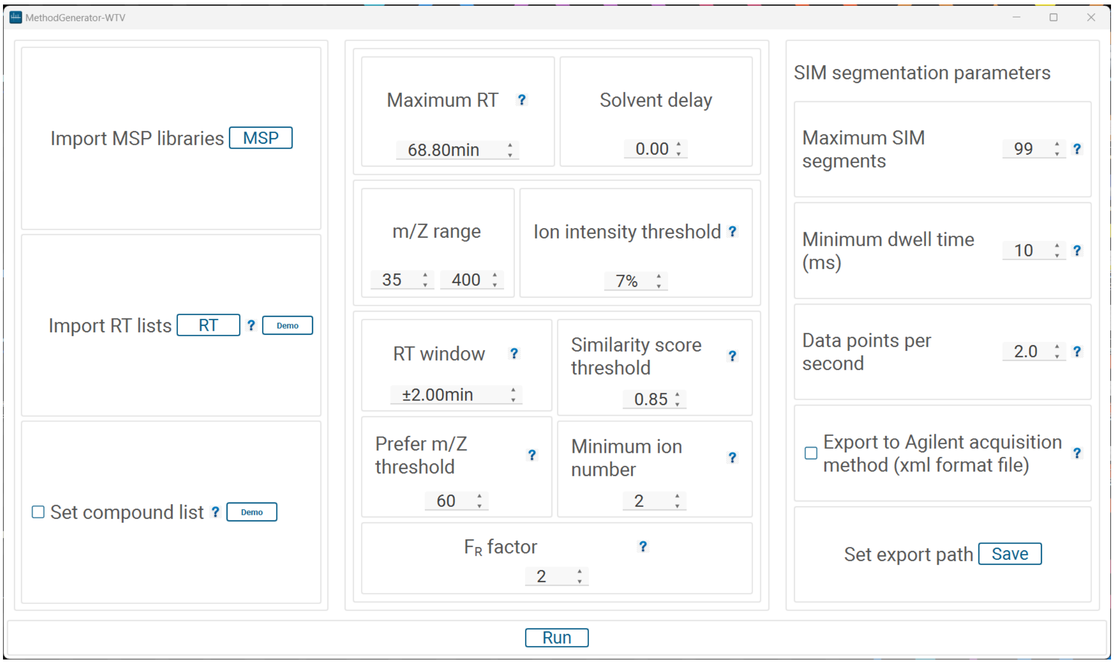
RT	Seg.	71	96	128	136	204	ion volume
10.00		1	1	1	0	0	
10.01	151	1	1	1	0	0	150×3 = 450
11.49		1	1	1	0	0	
11.50		1	1	1	1	1	
12.00	152`	1	1	1	1	1	200×5 = 1000
12.01		1	1	1	1	1	
13.50		1	1	1	1	1	

Fig. 2.3-2 The SIM method segmentation algorithms

Afterward, according to the user-defined number of data points per second, the software calculates the dwell time of ion in each SIM segment. If the calculated dwell time was greater than or equal to the user-defined minimum dwell time, then the dwell time was set to the calculated dwell time. Otherwise, the dwell time was adjusted by reducing the data points per second until it was greater than or equal to the minimum dwell time.

The software then exports the SIM segmentation results. For Agilent users, the *method generator* can export an XML-formatted file compatible with Agilent acquisition software. Users can replace the corresponding file in the original SIM acquisition method with the generated file and directly employ it.

2.2.3 User Interface of The *Method Generator*



Maximum RT 68.80min

Solvent delay 0.00

m/Z range 35 400

Ion intensity threshold 7%

RT window ±2.00min

Similarity score threshold 0.85

Prefer m/Z threshold 60

Minimum ion number 2

F_R factor 2

SIM segmentation parameters

Maximum SIM segments 99

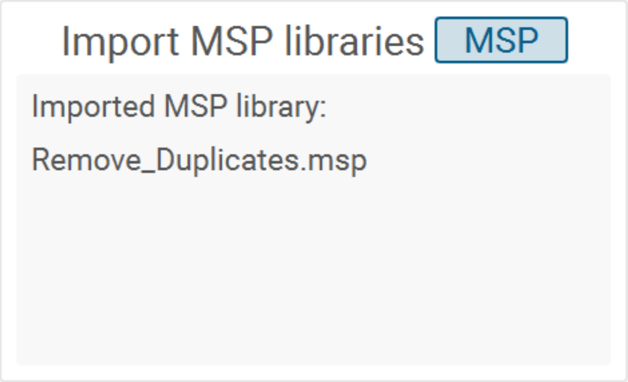
Minimum dwell time (ms) 10

Data points per second 2.0

□ Export to Agilent acquisition method (xml format file)

2.2.4 Parameter Configuration

Import MSP library: To import an MSP file, click the 'MSP' button and select the MSP file. Once selected, the interface will display all the selected files, confirming a successful input.



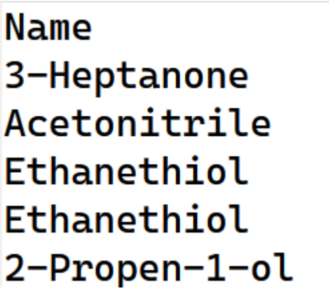
Imported MSP library:

Remove_Duplicates.msp

Import RT list: Import a list of compounds and their RT in CSV format.

Set compound list: When "set compound list" is selected and a compound list is imported, the generated collection method will only contain qualitative ions of compounds in the list.

The format of the input compound list file is shown below:



Name
3-Heptanone
Acetonitrile
Ethanethiol
Ethanethiol
2-Propen-1-ol

Maximum RT: Enter the end time of the temperature program.

Solvent delay: Enter the solvent delay time.

***m/Z* range:** The qualitative ions will be selected within this range.

Ion intensity threshold: Ions with abundances lower than the maximum ion abundance multiplied by this threshold value will be excluded in the selection of qualitative ions.

RT window: When selecting qualitative ions, target compounds are compared for similarity with compounds within the user-defined RT window.

Similarity score threshold: When the similarity score is below this threshold, two spectra are considered distinguishable. The default value is 0.85.

Prefer *m/Z* threshold: If the ion's *m/Z* value was smaller than this threshold, its value was set to 1 in weight score calculation.

Minimum ions number: The minimum number of ions of the selected qualitative ions.

F_R factor: The 'Ratio of Peak Pairs' term is considered in the similarity score calculation only if the number of ions is greater than this threshold. The F_R factor is typically set to be consistent with the minimum number of ions. For more information, please refer to the AMDIS manual.

SIM segmentation parameters:

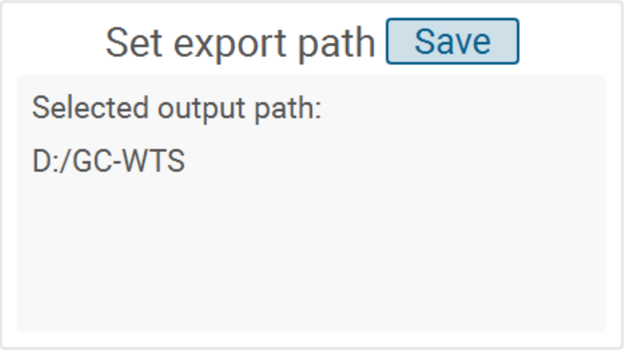
Maximum SIM segments: The maximum number of segments allowed in the SIM acquisition method. Default: 99.

Minimum dwell time (ms): The allowed minimum dwell time. Default: 10 ms.

Data points per second: Adjust the number of data points per second. If the calculated dwell time based on the specified data points per second falls below the minimum dwell time, the data points per second will be decreased to ensure that the dwell time remains greater than or equal to the minimum dwell time. The default setting is 2.

Export to Agilent data acquisition method (xml format file): Select this option to export the XML-formatted file used by Agilent data acquisition software.

Set export path: Click the 'Save' button to set the export path.

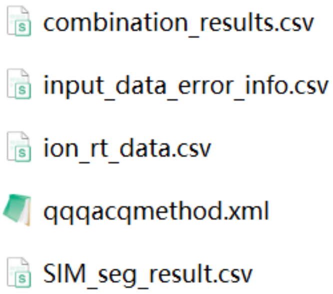


Selected output path:

D:/GC-WTS

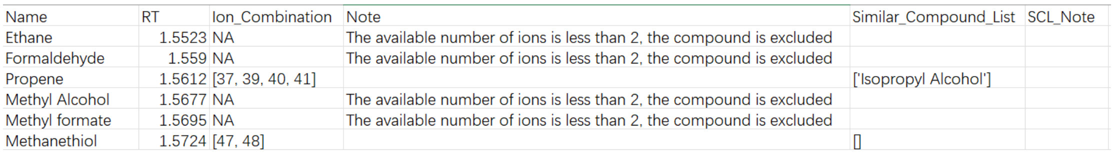
2.2.5 Generated Widely-Targeted SIM Method

After configuring all parameters, click the 'Run' button. The program will automatically generate five result files.



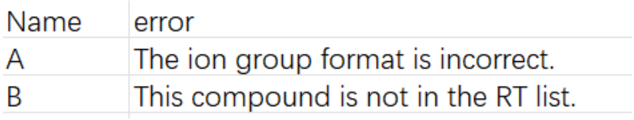
- combination_results.csv
- input_data_error_info.csv
- ion_rt_data.csv
- qqqacqmethod.xml
- SIM_seg_result.csv

combination_results: This is the qualitative ions selection results. SCL_Note: If the target compound has no adjacent compounds, the message 'No adjacent compounds.' will be displayed.



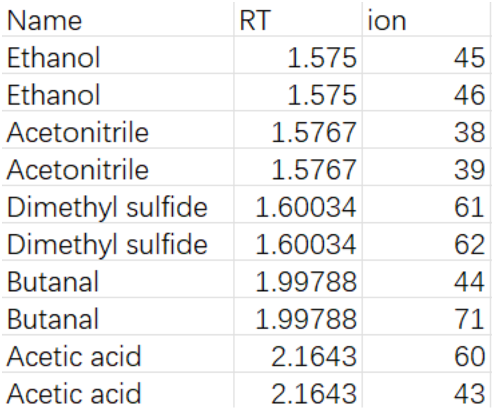
Name	RT	Ion_Combination	Note	Similar_Compound_List	SCL_Note
Ethane	1.5523	NA	The available number of ions is less than 2, the compound is excluded		
Formaldehyde	1.559	NA	The available number of ions is less than 2, the compound is excluded		
Propene	1.5612	[37, 39, 40, 41]		['Isopropyl Alcohol']	
Methyl Alcohol	1.5677	NA	The available number of ions is less than 2, the compound is excluded		
Methyl formate	1.5695	NA	The available number of ions is less than 2, the compound is excluded		
Methanethiol	1.5724	[47, 48]		[]	

input_data_error_info: This is the error information of imported data.



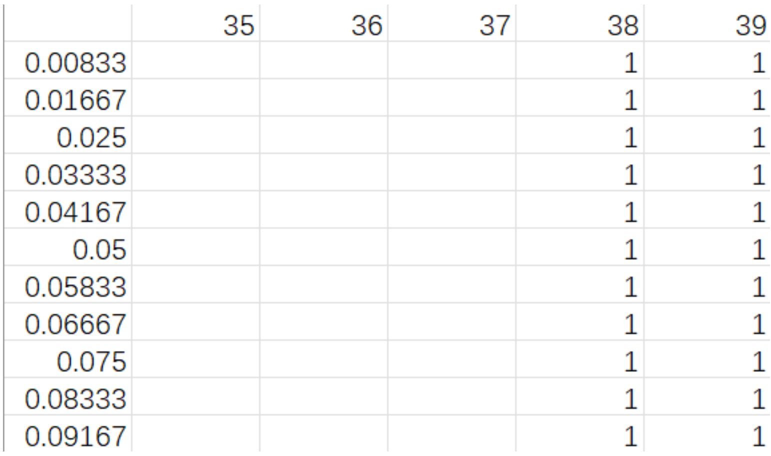
Name	error
A	The ion group format is incorrect.
B	This compound is not in the RT list.

ion_rt_data: This is the qualitative ions selection results in another format.



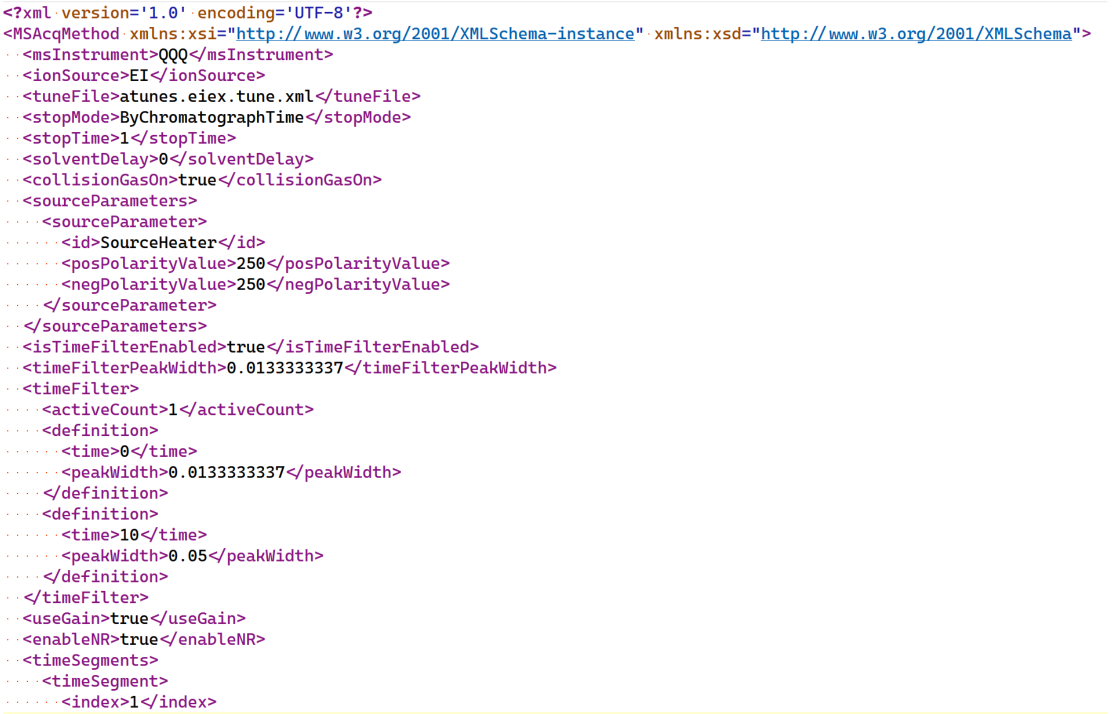
Name	RT	ion
Ethanol	1.575	45
Ethanol	1.575	46
Acetonitrile	1.5767	38
Acetonitrile	1.5767	39
Dimethyl sulfide	1.60034	61
Dimethyl sulfide	1.60034	62
Butanal	1.99788	44
Butanal	1.99788	71
Acetic acid	2.1643	60
Acetic acid	2.1643	43

SIM_seg_result: This is the segmentation result.



	35	36	37	38	39
0.00833				1	1
0.01667				1	1
0.025				1	1
0.03333				1	1
0.04167				1	1
0.05				1	1
0.05833				1	1
0.06667				1	1
0.075				1	1
0.08333				1	1
0.09167				1	1

qqqacqmethod.xml: The XML-formatted file used by Agilent data acquisition software.



```xml
<?xml version='1.0' encoding='UTF-8'?>
<MSAcqMethod xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <msInstrument>QQQ</msInstrument>
  <ionSource>EI</ionSource>
  <tuneFile>atunes.eiex.tune.xml</tuneFile>
  <stopMode>ByChromatographTime</stopMode>
  <stopTime>1</stopTime>
  <solventDelay>0</solventDelay>
  <collisionGasOn>true</collisionGasOn>
  <sourceParameters>
    <sourceParameter>
      <id>SourceHeater</id>
      <posPolarityValue>250</posPolarityValue>
      <negPolarityValue>250</negPolarityValue>
    </sourceParameter>
  </sourceParameters>
  <isTimeFilterEnabled>true</isTimeFilterEnabled>
  <timeFilterPeakWidth>0.0133333337</timeFilterPeakWidth>
  <timeFilter>
    <activeCount>1</activeCount>
    <definition>
      <time>0</time>
      <peakWidth>0.0133333337</peakWidth>
    </definition>
    <definition>
      <time>10</time>
      <peakWidth>0.05</peakWidth>
    </definition>
  </timeFilter>
  <useGain>true</useGain>
  <enableNR>true</enableNR>
  <timeSegments>
    <timeSegment>
      <index>1</index>
```

Tips: A demo results can be found at:

https://github.com/yuanhonglun/WTV_2.0/tree/main/sample_data/method_generator_sample_data/export

2.3 Data Analyzer

2.3.1 Data Preparation

2.3.1.1 Mass Spectrometry File Format Conversion

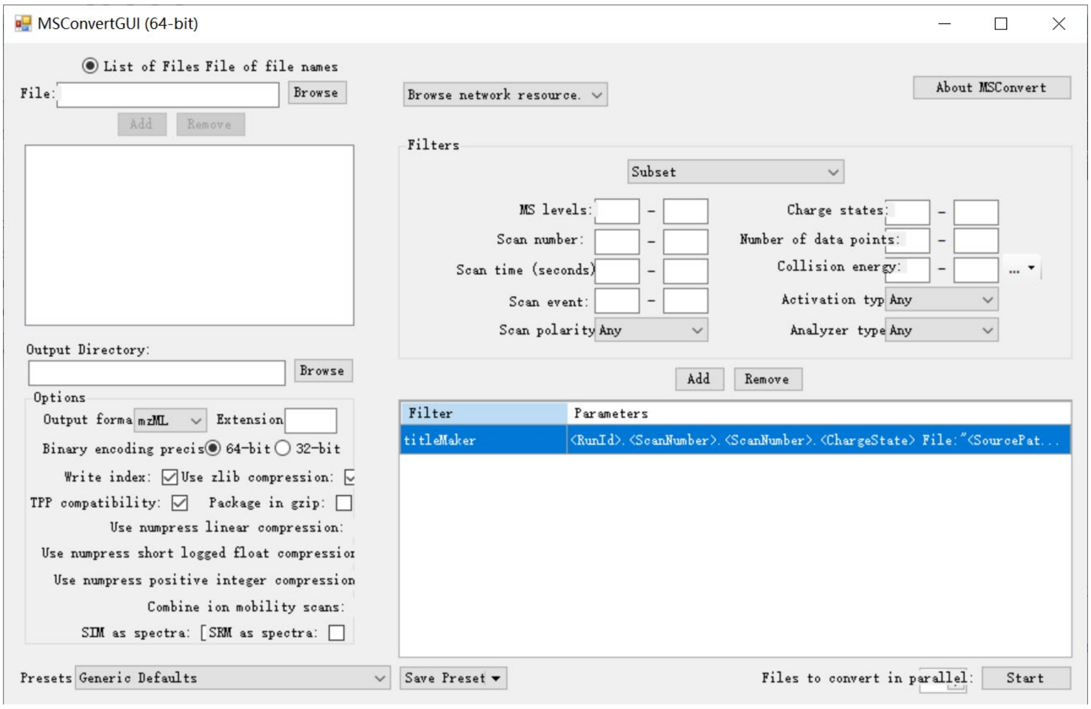
Before data analysis, user need to convert data format to mzML or cdf format.

tomato_fruit-spme-20231213.cdf

tomato_fruit-spme-20231213.mzml

Convert to mzML:

Download and install *ProteoWizard*: (http://proteowizard.sourceforge.net/downloads.shtml). After installation, open the software and click Browse and select file(s) for conversion. Then click Add to add them to the MSConvert workflow. Choose an Output Directory. Choose mzML (preferred) or mzXML for output format, 32-bit for binary encoding precision, SIM as spectra and uncheck Use zlib compression. Click Start. Check your folder for the converted .mzML files.



Output forma mzML Extension

Binary encoding precis ● 64-bit ○ 32-bit

Write index: ☑ Use zlib compression: ☑

TPP compatibility: ☑ Package in gzip: ☐

Use numpress linear compression:

Use numpress short logged float compressio

Use numpress positive integer compression

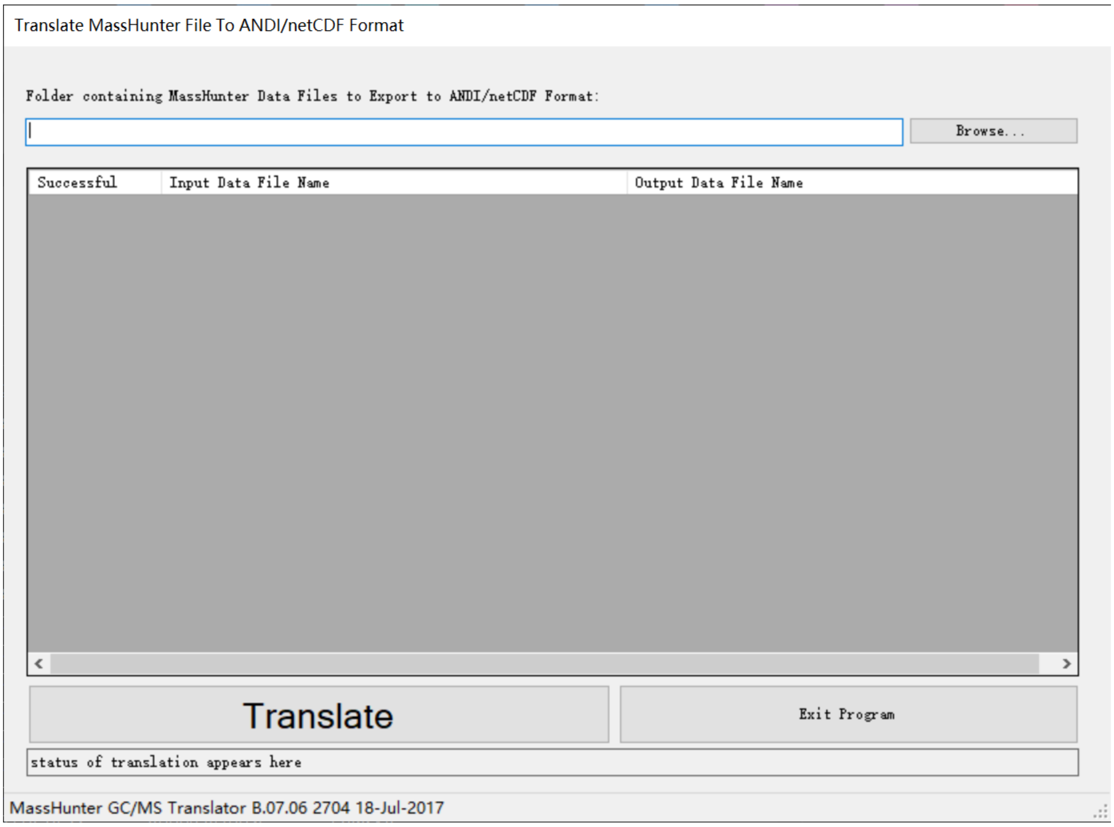
Combine ion mobility scans:

SIM as spectra: [SRM as spectra: ☐

Filter	Parameters
titleMaker	<RunId>.<ScanNumber>.<ScanNumber>.<ChargeState> File:"<SourcePat...

Convert to cdf:

For users of *Agilent GC-MS translator*: Click Browse and select the folder containing the file(s) for conversion. Click Translate. The CDF files are generated in the same folder as the raw data files.



Translate MassHunter File To ANDI/netCDF Format

Folder containing MassHunter Data Files to Export to ANDI/netCDF Format:

Successful	Input Data File Name	Output Data File Name

status of translation appears here

MassHunter GC/MS Translator B.07.06 2704 18-Jul-2017

For other vendor-provided software, please refer to the manual.

2.3.1.2 Compound Library (MSP file) for Identification

The library used for identification should be the same library that used to generate the acquisition method.

2.3.1.3 Retention Information

The RT list used for identification should be the same list that used to generate the acquisition method. Alternatively, user can prepare the RI calibration data for identification.

2.3.2 *Data Analyzer* Algorithm

The *data analyzer* is used to perform the automatic qualitative and semi-quantitative analysis of cSIM data (Fig. 2.3-1): After data import, the initial step involves missing value imputation using linear interpolation: If a data point displays a response of 0, while its preceding and subsequent data points exhibit responses greater than 0, then its response was replaced with the average response of its adjacent points. Subsequently, the *data analyzer* employs a linearly weighted smoothing average to perform data smoothing (equation 2.3-1). Users can define the smoothing factor (n in the equation), with a higher factor leading to a more pronounced smoothing effect, albeit potentially causing the loss of low-intensity peaks.

$$f(\mathrm{x})_{\mathrm{new}} = \frac{\sum_{i=-n}^{n}(n-i+1)\times f(\mathrm{x+i})}{n^2} \qquad (2.3\text{-}1)$$

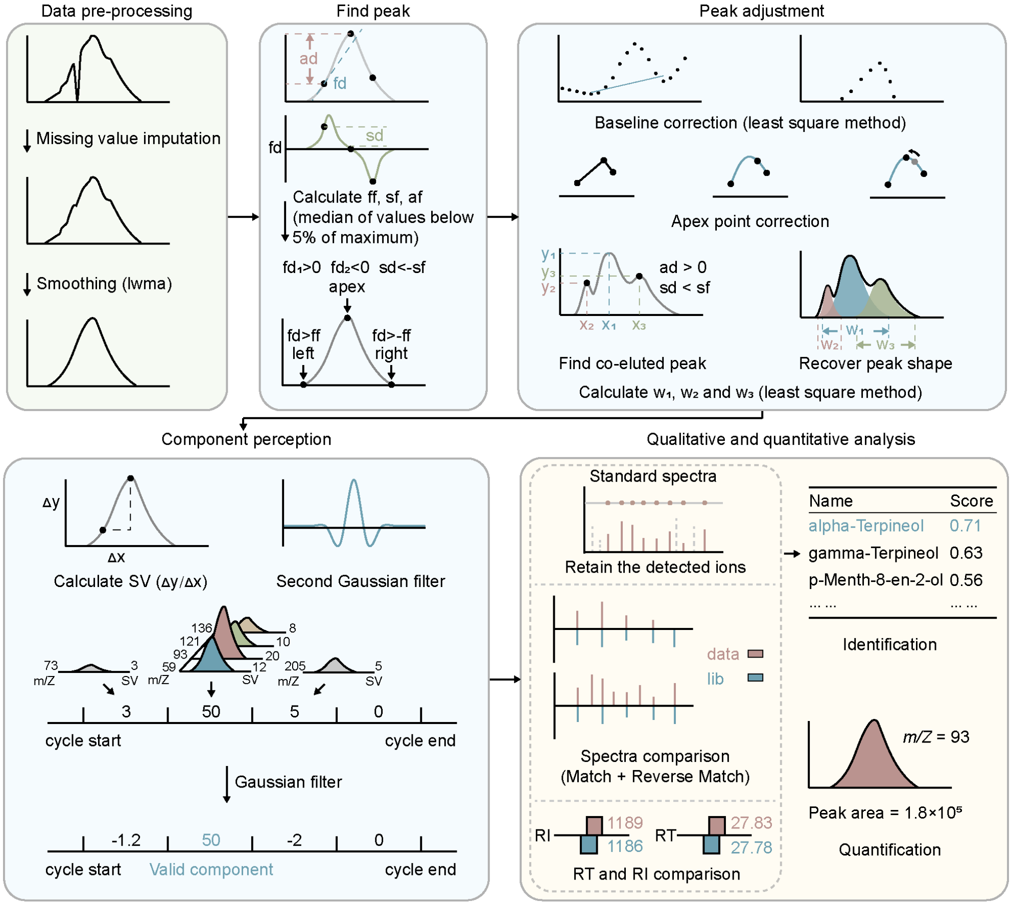


Fig. 2.3-1 The workflow of the *data analyzer* module.

For peak detection, the *data analyzer* employs the algorithm of MS-DIAL. For each extraction ion current (EIC), the first derivative (fd), second derivative (sd) and the abundance difference (ad) of each data point were calculated (equations 2.3-2~4). For the calculation of the first derivative filter (ff), all absolute values of fd below 5% of the maximum absolute value of fd are collected, and the median value from these fd absolute values is chosen as ff. Similarly, when calculating the second derivative filter (sf), all absolute values of sd that are smaller than 5% of the maximum negative value of sd are collected, and the median value from these sd absolute values is designated as sf. The calculation of the abundance difference filter (af) follows the same approach as that of ff.

The *data analyzer* utilizes these values and filters to perform peak detection: When two adjacent data points both exhibit fd greater than ff multiplied by a user-defined peak filtering factor, and both data points have intensity greater than 0, the first data point was chosen. A local minimum within the adjacent 5-point window was explored, and the data point with the lowest intensity was identified as the left boundary of the peak. The software continues to assess data points: If a data point's fd or the following data point's fd was less than 0, the preceding data point's fd was greater than 0, and the data point's sd was smaller than the negative value of sf, then the data point was selected. A local maximum within the adjacent 5-point window was explored, and the data point with the highest intensity was identified as the apex of the peak. The software continues to assess data points: If two adjacent data points exhibit fd greater than the negative value of ff multiplied by the peak filtering factor, the first data point is selected. Alternatively, if a data point's intensity was less than 5% of the apex point's intensity, it was selected. Based on the selected data point, a local minimum within the adjacent 5-point window was determined, and the data point with the lowest intensity was recognized as the right boundary of the peak.

$$fd = \frac{-2x_{-2} - x_{-1} + x_{+1} + 2x_{+2}}{10} \quad (2.3\text{-}2)$$

$$sd = \frac{2x_{-2} - x_{-1} - 2x_0 - x_{+1} + 2x_{+2}}{7} \quad (2.3\text{-}3)$$

$$ad = x_{+1} - x_0 \quad (2.3\text{-}4)$$

Subsequent to peak detection, the software removes peaks with incorrect RT order (e.g., the RT of the apex was smaller than the RT of the left boundary). The software identifies and removes the false peaks caused by SIM segmentation by evaluating whether the intensity of the left or right boundary point of a peak was zero in the unsmoothed data. Following this, the software calculates the noise factor (NF) performs baseline correction, removes redundant peaks, adjusts the RT and intensity of the apex point and calculates sharpness value (SV), employing the algorithms used in AMDIS.

The nf is calculated using equation 2.3-5. Each EIC is divided into segments of 13 scans. If any abundance in a segment is zero, the segment is rejected. For each accepted segment, a mean abundance is computed and the number of times that this mean value is "crossed" within the segment is counted (crossings occur for adjacent mass spectral scans where one abundance is above the mean and other abundance is below the mean). If the number of crossings is less than one-half the number scans in the segment (six or less), the segment is rejected. For each accepted segment, the median deviation from the mean abundance for that segment is found. This deviation is divided by the square root of the mean abundance for that segment to obtain a sample nf value, which is then saved.

$$nf = \frac{average\ random\ deviation}{\sqrt{intensity}} \quad (2.3\text{-}5)$$

Then, four steps for determining whether a peak is large enough to be used for peak perception. (1) A scan window is set using minima on each side of the peak; (2) a tentative baseline is drawn between the lowest points on each side (readjusted if a point between these end points falls below the line); (3) a least-squares line is drawn using the lowest one-half of points as measured from the baseline in step 2; (4) signal height between the maximum and least squares line is computed. Peaks must have heights larger than the peak perception threshold for use in peak perception (equation 2.3-6).

$$peak\ perception\ threshold = 4 \times \mathrm{nf} \times \sqrt{raw\ intensity} \quad (2.3\text{-}6)$$

Afterwards, a precise apex point is computed by fitting a parabola to the maximum and its two adjacent scans. The peak is then time shifted to center the scans at this computed apex point.

The software then performs the deconvolution: For each peak, points between the left edge and apex were examined. If the ad value of the adjacent three data points follows a pattern of positive, positive and negative, and the sd value of the second data point was smaller than sf, and its intensity was greater than 10% of the intensity of the apex point, then the second point was identified as the apex of a convoluted peak. Similarly, points between the apex and the left edge were examined as well. Utilizing the apex RT and intensity of convoluted peak(s), Gaussian peak(s) were fitted using the least squares method, which yields the fitted peak width and baseline height. The data points of convoluted peak(s) were replaced with that of the fitted Gaussian peak(s).

Subsequently, the *data analyzer* calculates the SV of each peak. SVs are the maximum rate of decline in abundance between the central scan and scans on either side (equations 2.3-7 and 2.3-8). Each scan was divided into an array of user-defined number of subintervals (bins). SVs were added to the bin corresponding to its RT, and a matched second derivative Gaussian filter was applied, each peak of the Gaussian filter represents a detected component (equation 2.3-9). All single ions that have a maximum within this range were assigned to this component.

$$SV_{left\ or\ right} = max_n \frac{A_{max} - A(n)}{n\sqrt{A_{max}}} \quad (2.3\text{-}7)$$

$$SV = \frac{SV_{left} + SV_{right}}{2} \quad (2.3\text{-}8)$$

$$second\ derivative\ Gaussian\ filter\text{:}\ \left\{1-\left(\frac{x}{\delta}\right)^2\right\}\times exp\left\{-\frac{1}{2}\left(\frac{x}{\delta}\right)^2\right\}\quad \text{(2.3-9)}$$

After component perception, the *data analyzer* calculates the similarity score between the spectra of the component and standard (equations 2.3-10~12), and the RT/RI penalty score (equation 2.3-13). The overall similarity score was calculated using the following equation (equation 2.3-14):

$$F_D=\sqrt{\frac{\left(\sum_{i=1}^{n}(x_i\times y_i)\right)^2}{\sum_{i=1}^{n}(x_i)^2\times\sum_{i=1}^{n}(y_i)^2}}\quad (x_i\ or\ y_i=\sqrt{intensity}\times m/Z^2)\quad \text{(2.3-10)}$$

$$F_R=\frac{1}{N_{L\&U}}\times\sum_{i}^{L\&U}\left(\frac{W_{Li}}{W_{Li-1}}\times\frac{W_{Ui-1}}{W_{Ui}}\right)^n\ (W=intensity)\quad \text{(2.3-11)}$$

where n = 1 or -1 when the term in parentheses is less than or greater than unity, respectively

$$Composite\ score=\frac{N_U\times F_D+N_{L\&U}\times F_R}{N_U+N_{L\&U}}\quad \text{(2.3-12)}$$

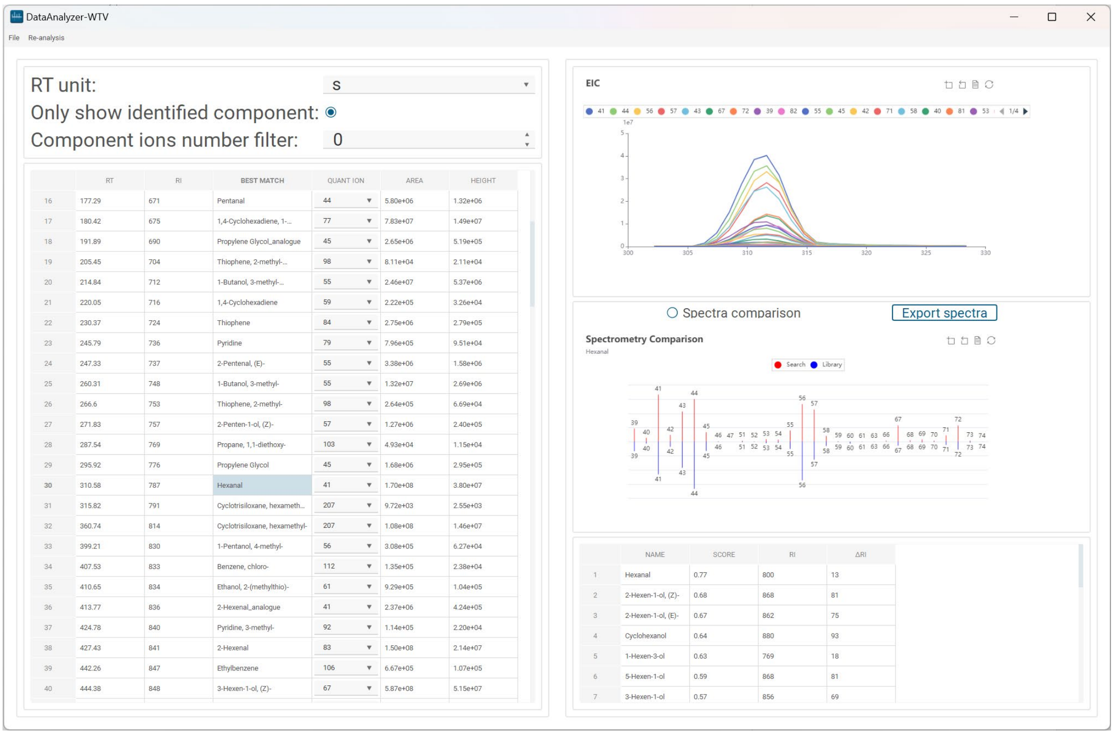
$$Retention\ penalty\ score=\left(\frac{abs(RI/RT_{measured}-RI/RT_{library})}{RI/RT\ window}-1\right)\times level\ factor\quad \text{(2.3-13)}$$

$$Overall\ similarity\ score=\frac{S_M\times W_M+S_R\times W_R}{W_M+W_R}-Retention\ penalty\ score\quad \text{(2.3-14)}$$

Note that in similarity score calculation, the spectra of the standard retain only the ions detected by the acquisition method. S_M (Match score) was the similarity score calculated between the spectra of the component and the spectra of the standard. S_R (Reverse Match Score) was obtained by comparing only the ions present in the standard spectra against those in the spectra of the component. W_M and W_R represent the weights of S_M and S_R, respectively. Users can also define parameters of the retention information penalty score, including level factor, max penalty, and no retention information penalty. In addition, the *data analyzer* also introduces an inaccurate RI threshold and inaccurate RI level factor: If the RI of a component is less than the inaccurate RI threshold, then it uses the inaccurate RI level factor to calculate the RI penalty score, which mitigates the impact of RI penalty score on overall similarity score when RI calculation was inaccurate.

Finally, the *data analyzer* selects the ion with the highest response among the non-coeluting ions in each component as the quantitative ion, and calculates its peak height and peak area to perform the semi-quantitative analysis. If all ions in a component were co-eluting ions, the ion with the highest response was selected as the quantitative ion.

2.3.3 User Interface of The *Data Analyzer*



RT unit: s

Only show identified component: ◉

Component ions number filter: 0

	RT	RI	BEST MATCH	QUANT ION	AREA	HEIGHT
16	177.29	671	Pentanal	44	5.80e+06	1.32e+06
17	180.42	675	1,4-Cyclohexadiene, 1-...	77	7.83e+07	1.49e+07
18	191.89	690	Propylene Glycol_analogue	45	2.65e+06	5.19e+05
19	205.45	704	Thiophene, 2-methyl-...	98	8.11e+04	2.11e+04
20	214.84	712	1-Butanol, 3-methyl-...	55	2.46e+07	5.37e+06
21	220.05	716	1,4-Cyclohexadiene	59	2.22e+05	3.26e+04
22	230.37	724	Thiophene	84	2.75e+06	2.79e+05
23	245.79	736	Pyridine	79	7.96e+05	9.51e+04
24	247.33	737	2-Pentenal, (E)-	55	3.38e+06	1.58e+06
25	260.31	748	1-Butanol, 3-methyl-	55	1.32e+07	2.69e+06
26	266.6	753	Thiophene, 2-methyl-	98	2.64e+05	6.69e+04
27	271.83	757	2-Penten-1-ol, (Z)-	57	1.27e+06	2.40e+05
28	287.54	769	Propane, 1,1-diethoxy-	103	4.93e+04	1.15e+04
29	295.92	776	Propylene Glycol	45	1.68e+06	2.95e+05
30	310.58	787	Hexanal	41	1.70e+08	3.80e+07
31	315.82	791	Cyclotrisiloxane, hexameth...	207	9.72e+03	2.55e+03
32	360.74	814	Cyclotrisiloxane, hexamethyl-	207	1.08e+08	1.46e+07
33	399.21	830	1-Pentanol, 4-methyl-	56	3.08e+05	6.27e+04
34	407.53	833	Benzene, chloro-	112	1.35e+05	2.38e+04
35	410.65	834	Ethanol, 2-(methylthio)-	61	9.29e+05	1.04e+05
36	413.77	836	2-Hexenal_analogue	41	2.37e+06	4.24e+05
37	424.78	840	Pyridine, 3-methyl-	92	1.14e+05	2.20e+04
38	427.43	841	2-Hexenal	83	1.50e+08	2.14e+07
39	442.26	847	Ethylbenzene	106	6.67e+05	1.07e+05
40	444.38	848	3-Hexen-1-ol, (Z)-	67	5.87e+08	5.15e+07

	NAME	SCORE	RI	ΔRI
1	Hexanal	0.77	800	13
2	2-Hexen-1-ol, (Z)-	0.68	868	81
3	2-Hexen-1-ol, (E)-	0.67	862	75
4	Cyclohexanol	0.64	880	93
5	1-Hexen-3-ol	0.63	769	18
6	5-Hexen-1-ol	0.59	868	81
7	3-Hexen-1-ol	0.57	856	69

RT unit: Switch the RT unit between minute and second.

Only show identified component: If selected, the results interface will only display identified components.

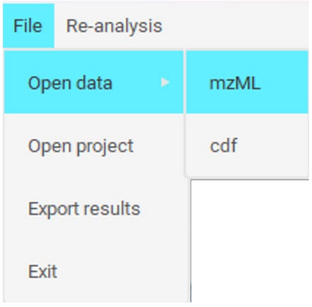
Component ions number filter: The results interface will only display components with ion number exceeding the threshold.

Results interface: The left panel displays the detected components, including their RT, RI, annotation results, quantitative ions, peak areas and heights. The right panel displays the extraction ion currents (EICs), spectra comparison, and possible qualitative results for the selected component. Users can export results as a CSV file.

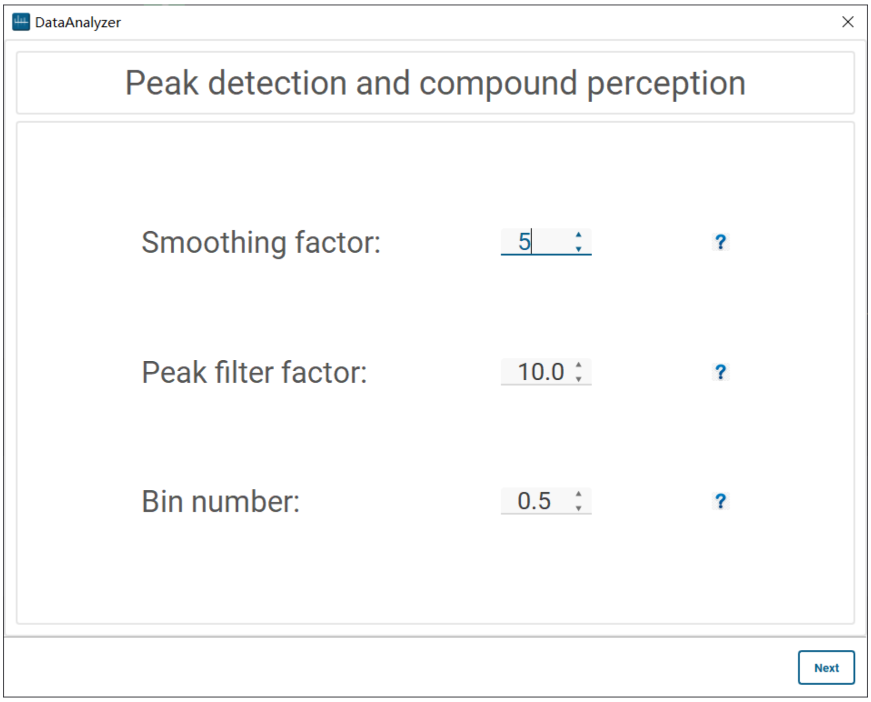
2.3.4 Parameter Configuration

2.3.4.1 Data Import

Select the imported file format.



2.3.4.2 Peak Detection



Smoothing factor: 5

Peak filter factor: 10.0

Bin number: 0.5

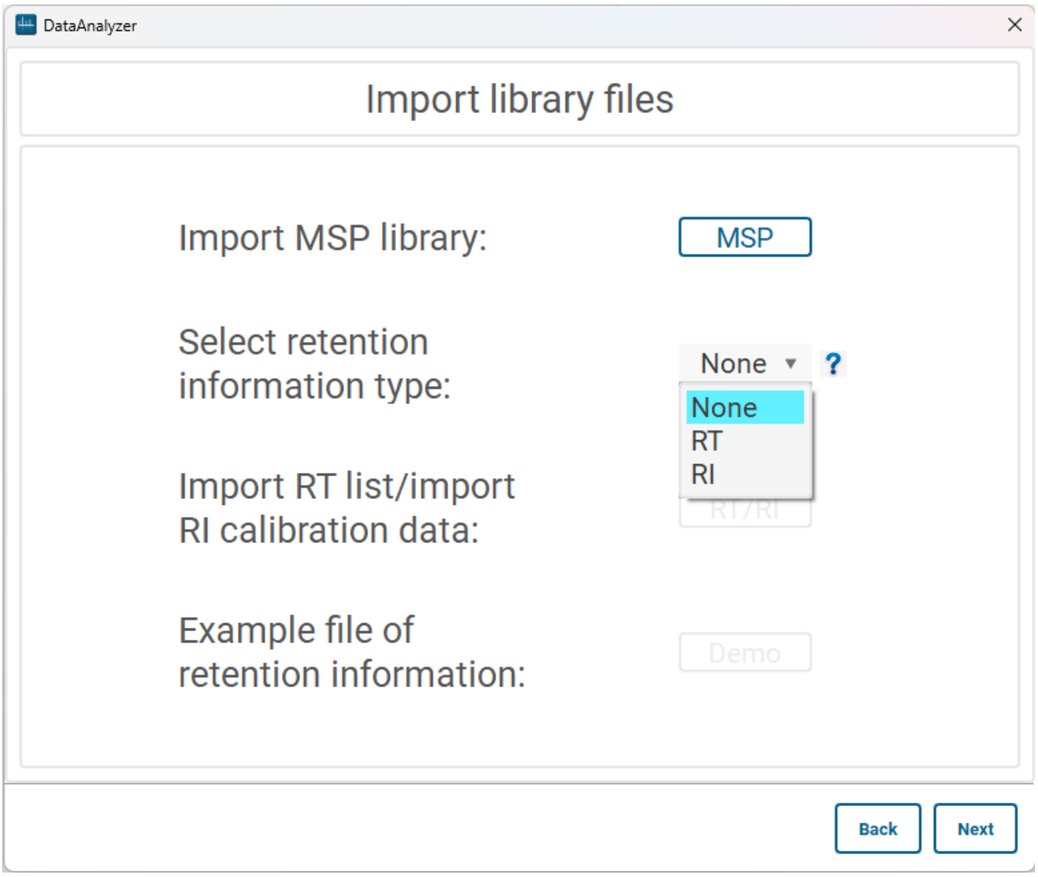
Smoothing factor: A high smooth factor can make the peaks smoother, but it may also lead to the loss of low-abundance peaks. The default value is set to 5.

Peak filter factor: A high peak filter factor can remove low-abundance peaks. If this value is lower than 10, it will significantly increase the program's runtime. The default setting is 10.

Bin number: Set a smaller bin number when using a smaller data points per second in acquisition method development. *e. g.*, set bin number as 0.5 when the data point per second is 2. For more information, please refer to the reference:

Stein, S.E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J Am Soc Mass Spectr 10:770-781. 10.1016/S1044-0305(99)00047-1.

2.3.4.3 Import library files



Import library files

Import MSP library:

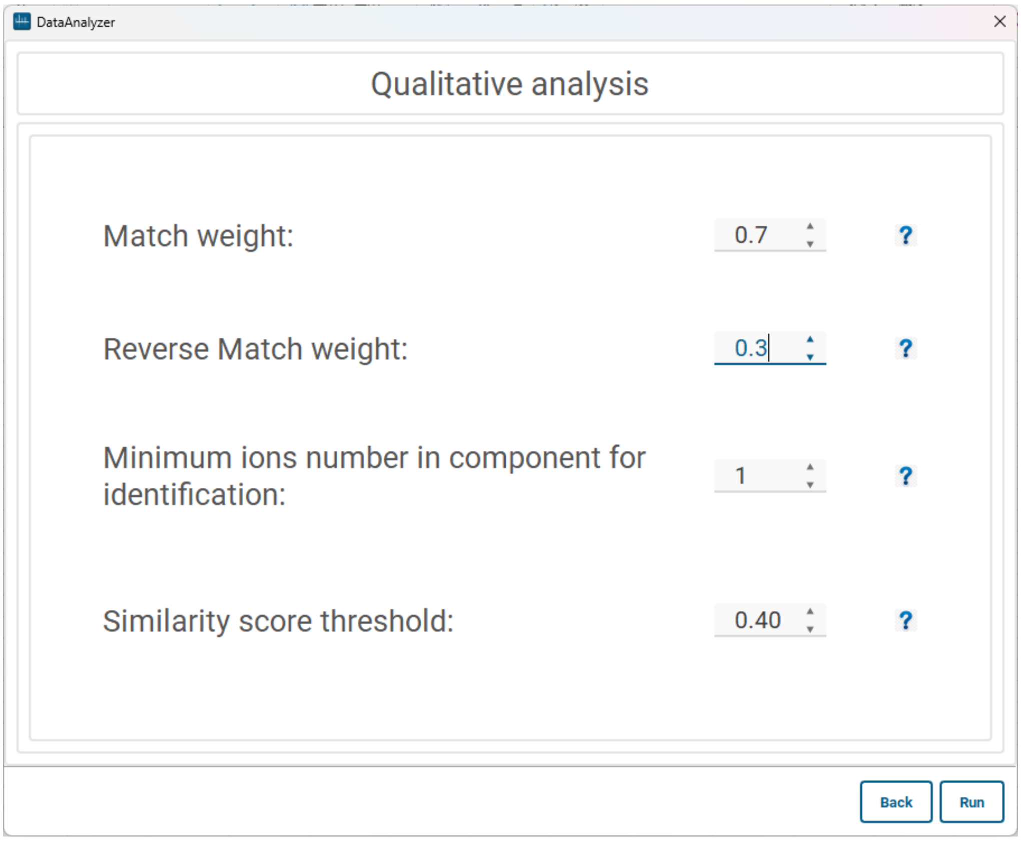
Select retention information type: None

Import RT list/import RI calibration data:

Example file of retention information:

Click the MSP button to import the library file.

In “None” mode



Qualitative analysis

Match weight: 0.7

Reverse Match weight: 0.3

Minimum ions number in component for identification: 1

Similarity score threshold: 0.40

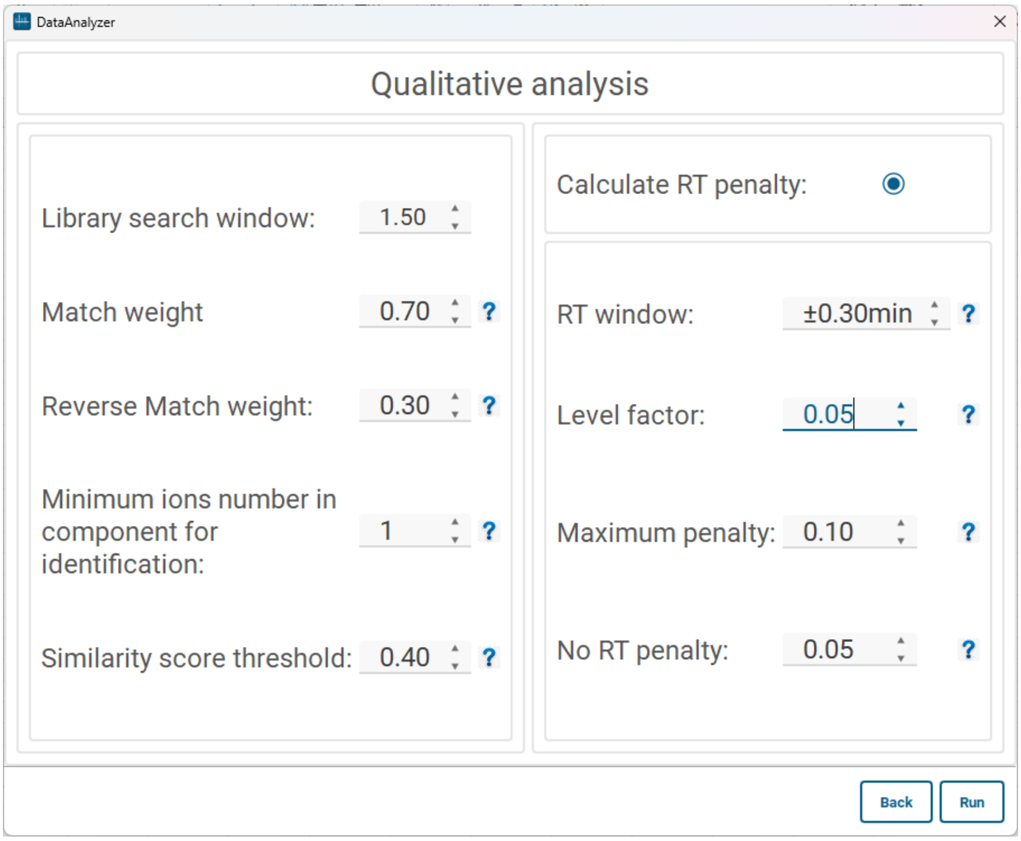
Match weight: Peak group forward retrieval matching score weight. Default: 0.7.

Reverse Match weight: Peak group reverse retrieval matching score weight. Default: 0.3.

Minimum ions number in component for identification: Peak group with ion number below this threshold will not be subjected to qualitative analysis.

Similarity score threshold: The integrated similarity score threshold, where only compounds with scores above this threshold will be considered candidate compounds. Default: 0.4.

In RT mode



Qualitative analysis

Library search window: 1.50

Match weight 0.70 ?

Reverse Match weight: 0.30 ?

Minimum ions number in component for identification: 1 ?

Similarity score threshold: 0.40 ?

Calculate RT penalty: ●

RT window: ±0.30min ?

Level factor: 0.05 ?

Maximum penalty: 0.10 ?

No RT penalty: 0.05 ?

Library search window: Configure the retrieval target compound window in RT mode. The program will compare all compounds within the specified matching window.

Match weight: Peak group forward retrieval matching score weight. Default: 0.7.

Reverse Match weight: Peak group reverse retrieval matching score weight. Default: 0.3.

Minimum ions number in component for identification: Component with ion number below this threshold will not be subjected to qualitative analysis.

Similarity score threshold: The integrated similarity score threshold, where only compounds with scores above this threshold will be considered candidate compounds. Default: 0.4.

Calculate RT penalty: Selecting this option will penalize candidate compounds with a significant difference in RT values.

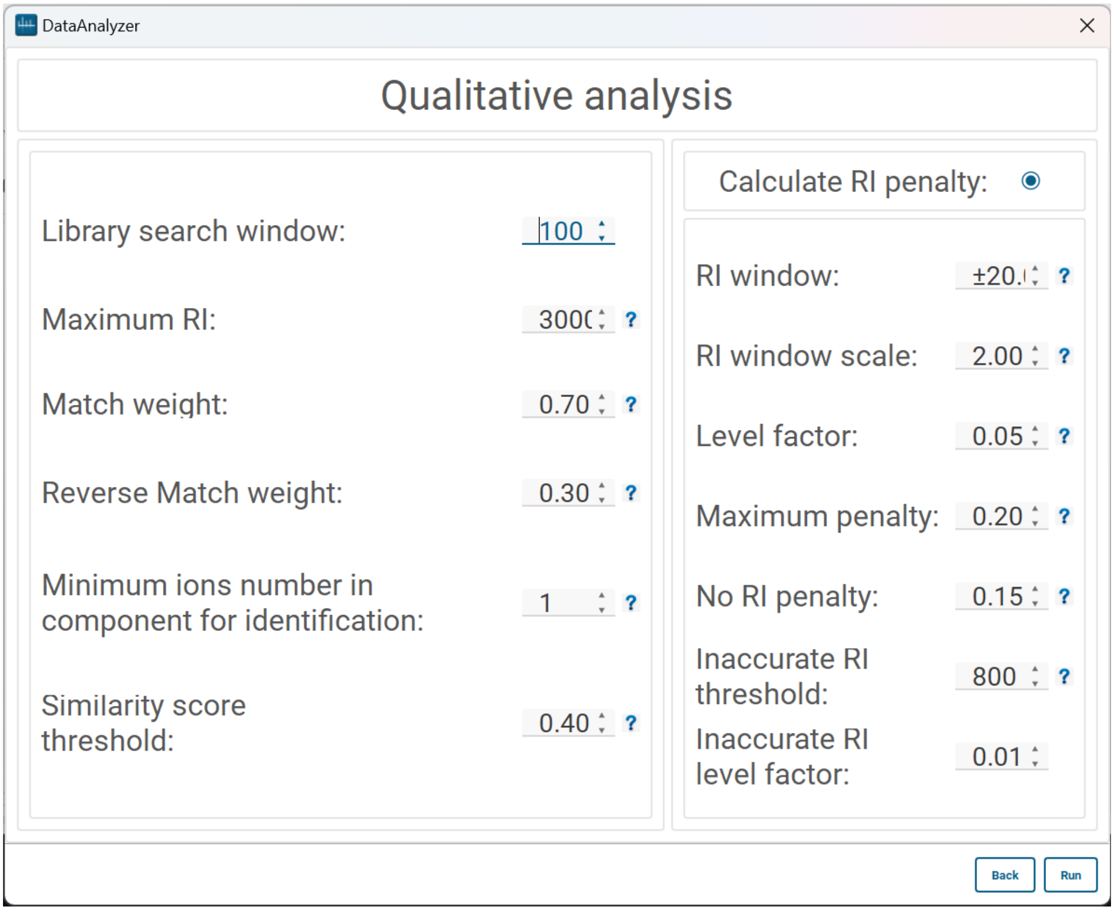
RT window: RT penalty will be applied only when RT difference exceeds this threshold.

Level factor: A higher number indicates a more severe penalty. Default: 0.05.

Maximum penalty: Set the maximum penalty score. Default: 0.1.

No RT penalty: Peak groups with RT lower than this value will not receive a penalty for mismatch. Default: 0.05.

In RI mode



Qualitative analysis

Library search window: 100

Maximum RI: 300(

Match weight: 0.70

Reverse Match weight: 0.30

Minimum ions number in component for identification: 1

Similarity score threshold: 0.40

Calculate RI penalty: ●

RI window: ±20.(

RI window scale: 2.00

Level factor: 0.05

Maximum penalty: 0.20

No RI penalty: 0.15

Inaccurate RI threshold: 800

Inaccurate RI level factor: 0.01

Library search window: Configure the retrieval target compound window in RI mode. The program will compare all compounds within the specified matching window.

Maximum RI: Enter the allowed maximum RI.

Match weight: Peak group forward retrieval matching score weight. Default: 0.7.

Reverse Match weight: Peak group reverse retrieval matching score weight. Default: 0.3.

Minimum ions number in component for identification: Component with ion number below this threshold will not be subjected to qualitative analysis.

Similarity score threshold: The integrated similarity score threshold, where only compounds with scores above this threshold will be considered candidate compounds. Default: 0.4.

Calculate RI penalty: Selecting this option will penalize candidate compounds with a significant difference in RI values.

RI window: RI penalty will be applied only when RI difference exceeds this threshold.

RI window scale: RI penalty window will be linearly scaled by this factor, setting it to 0 disables this feature. Default: 2.

Level factor: A higher number indicates a more severe penalty. Default: 0.05.

Maximum penalty: Set the maximum penalty score. Default: 0.2.

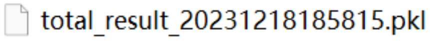
No RI penalty: Peak groups without RI will receive a penalty. Default: 0.15.

Inaccurate RI threshold: RI below this threshold will use an alternative penalty factor.

Inaccurate RI level factor: The specified range for the Inaccurate RI threshold mentioned above is used for the following purposes.

2.3.3 Results File

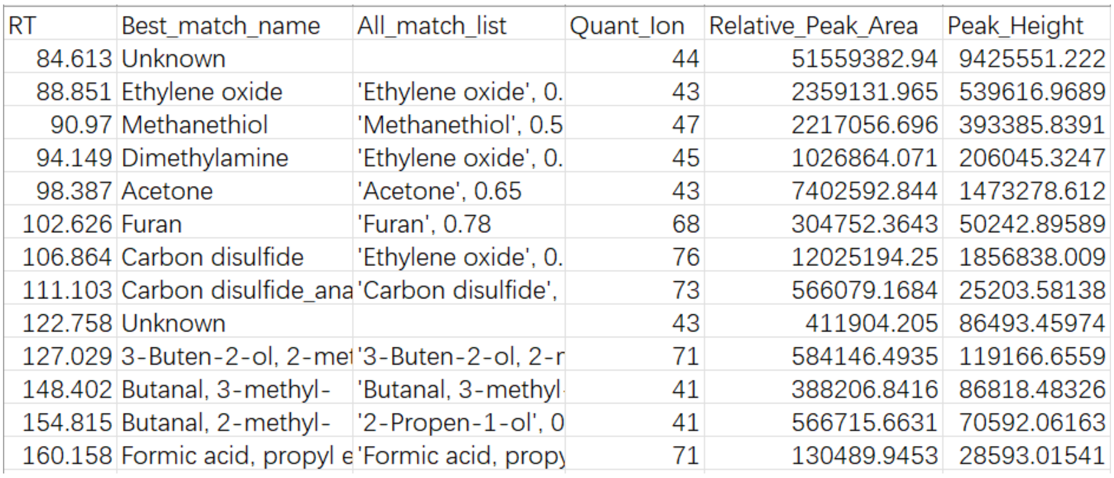
A pkl binary containing all the results information:



total_result_20231218185815.pkl

qualitative_and_quantitative_analysis_result.csv

The results including RT, best matching name, a list of all candidate names with their corresponding scores, the quantitative ion, peak area and peak height.



RT	Best_match_name	All_match_list	Quant_Ion	Relative_Peak_Area	Peak_Height
84.613	Unknown		44	51559382.94	9425551.222
88.851	Ethylene oxide	'Ethylene oxide', 0.	43	2359131.965	539616.9689
90.97	Methanethiol	'Methanethiol', 0.5	47	2217056.696	393385.8391
94.149	Dimethylamine	'Ethylene oxide', 0.	45	1026864.071	206045.3247
98.387	Acetone	'Acetone', 0.65	43	7402592.844	1473278.612
102.626	Furan	'Furan', 0.78	68	304752.3643	50242.89589
106.864	Carbon disulfide	'Ethylene oxide', 0.	76	12025194.25	1856838.009
111.103	Carbon disulfide_ana	'Carbon disulfide',	73	566079.1684	25203.58138
122.758	Unknown		43	411904.205	86493.45974
127.029	3-Buten-2-ol, 2-met	'3-Buten-2-ol, 2-r	71	584146.4935	119166.6559
148.402	Butanal, 3-methyl-	'Butanal, 3-methyl	41	388206.8416	86818.48326
154.815	Butanal, 2-methyl-	'2-Propen-1-ol', 0	41	566715.6631	70592.06163
160.158	Formic acid, propyl e	'Formic acid, propy	71	130489.9453	28593.01541

Tips: A demo results can be found at: https://github.com/yuanhonglun/WTV_2.0/tree/main/sample_data/data_analyzer_sample_data/export

3 Contact

Dr. Honglun Yuan
School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication)
Hainan University
Sanya, Hainan, 572025, China
E-mail: yuanhonglun@hotmail.com