
Widely-Targeted Volatilomics (WTV) 2.0



1 General Introduction

WTV2.0 is an open-source software platform with an intuitive graphical interface which provides users with a convenient one-stop solution for the whole process of widely targeted GC-MS data. WTV2.0 provides a complete metabolomics workflow encompassing the integration of metabolite libraries, the generation of widely targeted acquisition methods, qualitative and semi-quantitative analysis of data.

1.1 *library builder* - Metabolite Library Integration

library builder is a module focusing on integrating mass spectra libraries and retention time information. Initially, the imported libraries, RT information and retention index (RI) calibration data were integrated. After format standardization of compound names (e.g., replace “.beta.” with “beta”) and removes of invalid mass spectrum, the redundant information was excluded based on compound names, synonyms, and CAS numbers. The RTs were calculated or calibrated using RI calibration data. A feature of *library builder* is in addition to alkane standards, it is compatible with the use of RT and RI from other compounds in the RI calibration dataset, thus broadening calibration coverage. For example, due to the inconvenience in obtaining C4 and C5 alkanes, ethanol (RI = 427) and ethanethiol (RI = 517) can be incorporated to the dataset for the RT calibration of compounds with RI < 500. Afterward, the non-redundant RT list and library were exported. An additional function of the *library builder* is the integration of unknown signals in untargeted or cSIM data to the library, thereby expanding the detection coverage of acquisition method. However, the imported unknown signals might be compounds already existing in the library. To avoid redundancy, the integrated and non-redundant library was used as

background, and imported unknowns are compared for spectral similarity with compounds having adjacent RTs. Only when the similarity score between an unknown signal and all its neighboring compounds falls below the user-defined threshold is the unknown integrated into the library. This ensures the accuracy and consistency of the metabolite library, laying a reliable foundation for subsequent processes, such as widely targeted acquisition method generation and data analysis.

1.2 *method generator* - Widely Targeted Acquisition Method Generation

method generator, another key component of WTV2.0, is responsible for generating widely targeted acquisition methods. The software first picks the minimum group of characteristic ions for each compound based on metabolite library, which are sufficient to characterize from all compounds within a user-set window to form the highest sensitivity detected ion segmentation. However, due to the instrument detection's certain limitation on the number of SIM segments, in order to generate an acquisition method file that the instrument can read, we need to adjust the segment information according to the requirements of each manufacturer. We do this by automatically merging the segments that have the least sensitivity reduction to meet the instrument requirements, that is, adding the minimum number of ions to be detected.

The scope of this program is very broad. For high-throughput detection of high-coverage libraries, the software generates comprehensive SIM (cSIM) detection methods through the above algorithm. While for low-throughput detection, whether it is a low-coverage library or a high-coverage library, the program can assist in generating high-sensitivity detection method and accurate qualitative analysis of targeted metabolites by comparing the mass-to-charge ratio and corresponding intensity of compounds in the mass spectrum and selecting an appropriate number of detection ions.

1.3 *data analyzer* - Data Qualitative and Semi-Quantitative Analysis

data analyzer is tailored for qualitative and semi-quantitative analysis of GC-MS cSIM data. Raw data was first performed the missing value imputation, followed by data smoothing using a linearly weighted smoothing average method. For each data point, its first derivative (fd), the second derivative (sd), and the abundance difference

(ad) were calculated, and used to calculate the first derivative filter (ff), second derivative filter (sf) and abundance difference filter (af). Utilizing these filters, the software identifies the left boundary, apex and right boundary of a peak. Afterward, the software conducts the baseline correction and adjusts the RT and intensity of the apex point. For each peak, the software utilizing the ad and sd value to identify the apex point of co-eluted peaks, and recovers the peak shape using least square method. The sharpness value (SV) of each peak was calculated and added to the subintervals of scans. A second derivative Gaussian filter was applied to identify the component. For identification, the standard mass spectrum retains only the ions detected by the acquisition method and compared with the spectra of component to calculates the match score and reverse match score. The RT or RI penalty score was also calculated. These scores are collectively used to calculate the final similarity score. Annotation results with similarity score greater than the user-defined threshold are listed. Finally, semi-quantitative analysis was carried out by calculation of the peak area and peak height of quantitative ion.

For more detailed information, read on in the documentation, checkout the paper. Source codes are also available on GitHub (https://github.com/yuanhonglun/WTV_2.0) under the GNU General Public License, version 3 (GPL-3.0) License.

1.4 Installation Guide

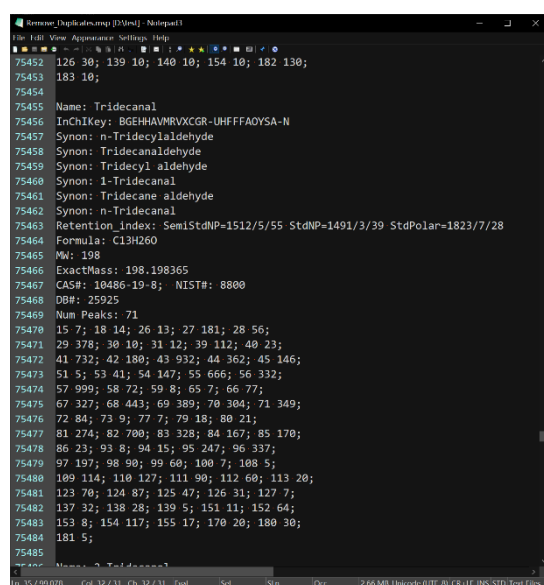
WTV2.0 is written in the Python language. It currently works in the Windows OS (Windows 10 or higher). The standalone software can be freely downloaded from GitHub (https://github.com/yuanhonglun/WTV_2.0). What we are releasing is the exe application, which is out of the box. Go to the WTV2.0 directory. Run .exe to start the graphical user interface by double-clicking but recommend to run with admin permissions. Since we do not pay Microsoft for certification you might have to confirm that you want to trust software from an unknown source on Windows.

2 Using the Software

2.1 Building a Compound Library (MSP) for Detection and Identification (the *library builder* module)

2.1.1 Data Preparation

MSP file(s)



```
Remove Duplicates.msp [UTF-8] - Notepad++
File Edit View Appearance Settings Help
75452 126 30; 139 10; 140 10; 154 10; 182 130;
75453 183 10;
75454
75455 Name: Tridecanal
75456 InChIKey: BGEHIAWVRVXGGR-UHFFFAOYSA-N
75457 Synon: n-Tridecylaldehyde
75458 Synon: Tridecanaldehyde
75459 Synon: Tridecyl aldehyde
75460 Synon: 1-Tridecanal
75461 Synon: Tridecane aldehyde
75462 Synon: n-Tridecanal
75463 Retention_index: SemiStdNP=1512/5/55 StdNP=1491/3/39 StdPolar=1823/7/28
75464 Formula: C13H26O
75465 MW: 198
75466 ExactMass: 198.198365
75467 CAS#: 10486-19-8; NIST#: 8800
75468 DB#: 25925
75469 Num Peaks: 71
75470 15 7; 18 14; 26 13; 27 181; 28 56;
75471 29 378; 30 10; 31 12; 39 112; 40 23;
75472 41 732; 42 180; 43 932; 44 362; 45 146;
75473 51 5; 53 41; 54 147; 55 666; 56 332;
75474 57 999; 58 72; 59 8; 65 7; 66 77;
75475 67 327; 68 443; 69 389; 70 304; 71 349;
75476 72 84; 73 9; 77 7; 79 18; 80 21;
75477 81 274; 82 700; 83 328; 84 167; 85 170;
75478 86 23; 93 8; 94 15; 95 247; 96 337;
75479 97 197; 98 90; 99 60; 100 7; 108 5;
75480 109 114; 110 127; 111 90; 112 60; 113 20;
75481 123 70; 124 87; 125 47; 126 31; 127 7;
75482 137 32; 138 28; 139 5; 151 11; 152 64;
75483 153 8; 154 117; 155 17; 170 20; 180 30;
75484 181 5;
75485
75486 Name: n-Tridecanal
75487
```

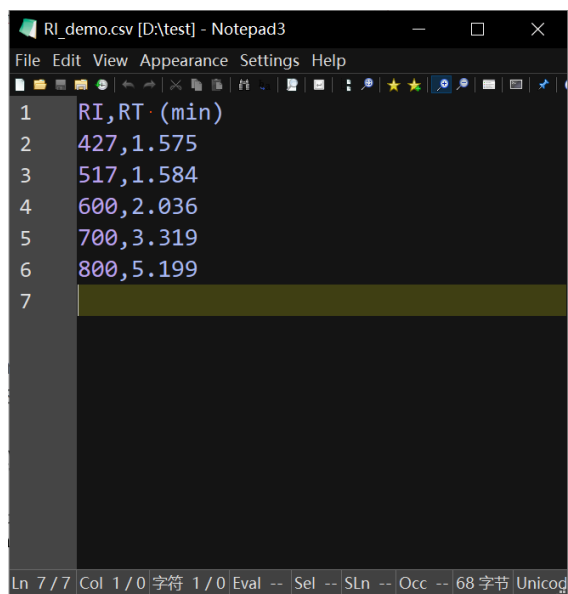
Prepare MSP files that contains mass spectrometry information and theoretical retention times for substances.

Retention Time Information

Using actual measured retention times (RT)

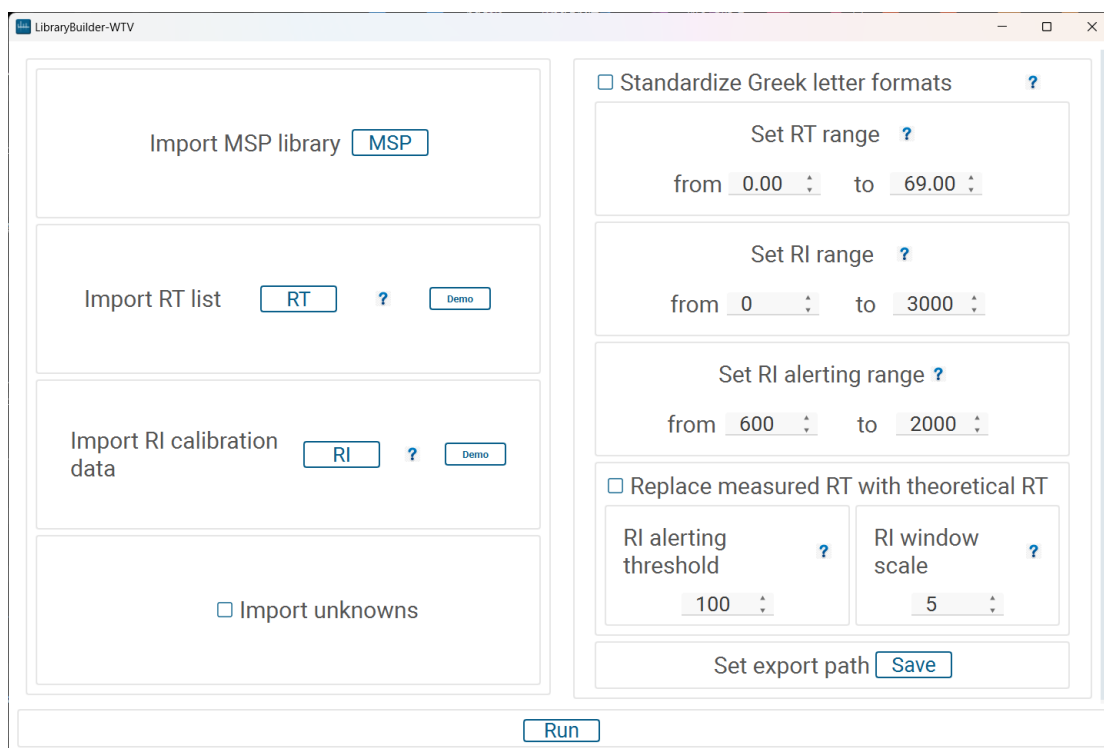
	A	B
1	Name	RT
2	Acetaldehyde	1.496
3	Methanethiol	1.526
4	Ethane	1.5523
5	Ethyl Chloride	1.555
6	Formaldehyde	1.559
7	Propene	1.5612
8	Methyl Alcohol	1.5677
9	Methyl formate	1.5695
10	Ethylene oxide	1.5727
11	Dimethylamine	1.5753
12	Ethylamine	1.5773
13	2-Propenal	1.5779
14	Ethyl formate	1.5791
15	Butane, 2-methyl-	1.5798
16	Ethyl ether	1.5808

Or / And RI calibration



```
RI_demo.csv [D:\test] - Notepad3
File Edit View Appearance Settings Help
Ln 1 Col 1 / 0 字符 1 / 0 Eval -- Sel -- SLn -- Occ -- 68 字节 Unicod
1 RI,RT:(min)
2 427,1.575
3 517,1.584
4 600,2.036
5 700,3.319
6 800,5.199
7
```

2.1.2 Parameter Configuration



LibraryBuilder-WTV

Import MSP library

Import RT list ?

Import RI calibration data ?

☐ 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

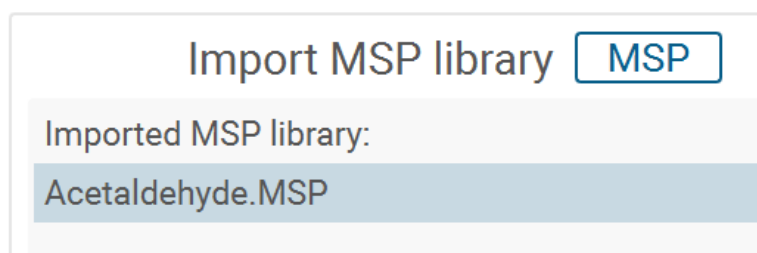
RI window scale ? 5

Set export path

Import MSP library

To input an MSP file, simply click the 'MSP' button and select the desired MSP file.

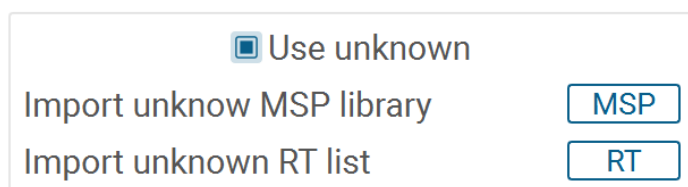
Once selected, the interface will display all the selected file directories, confirming a successful input.



Import RT list: Import a list of compounds and their RT.

Import RI calibration data: Import RI calibration data.

Use unknown: Users can use this feature when users need to enhance signals from compounds that cannot be qualitatively identified.



When users wish to use this feature, simply select the white box to highlight it, and the hidden menu will become visible. Within the hidden menu, input the corresponding information for the unknown item.

Similarity is calculated between unknowns and compounds within the user-defined retention time (RT) window. Only when the similarity is below the set threshold with all nearby compounds will the unknown be integrated into the database.

Standardize Greek letter formats: This option allows for a unified representation of Greek letters. Eg: ".alpha." -> "alpha"

Set RT range: Once configured, the output retention times will be constrained within this range.

Set RI range: Once configured, the output retention index will be constrained within this range.

Set RI alerting range: Only unreasonable RT (Retention Time) values outside the configured range will trigger a warning.

Replace measured RT with theoretical RT: Selecting this option will replace measured retention times outside the set theoretical retention time window with the theoretical retention time.

RI alerting threshold: Compound will be marked if the difference between the measured RI and the database RI exceeds this threshold.

RI window scale: The larger this value, the larger the RI error threshold when the RI is larger. Setting it to 0 disables this feature. For more specific and detailed information, please refer to the official documentation of the AMDIS software.

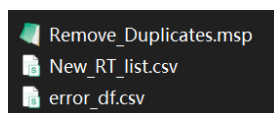
Set export path: Click the 'Save' button, choose a saving path, and the processed results will be saved to this directory.

Tips:

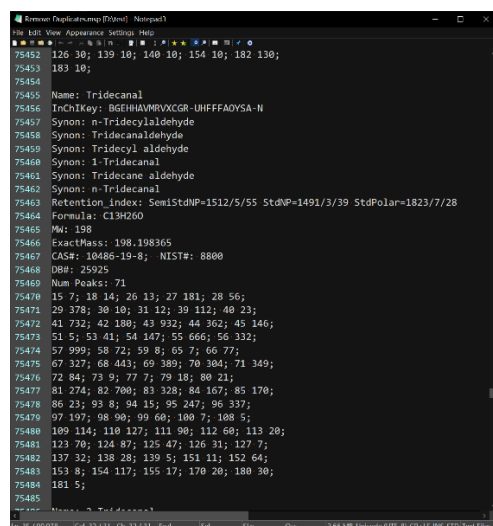
1. Any file directory entered can be deleted by double-clicking the entered entry.
2. Clicking the corresponding "demo" button allows viewing the required format.

2.1.3 Results and Reports

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



The MSP result file: Contains the MSP file information after deduplication.



The RT (Retention Time) result file: Contains the Retention Time information after deduplication.

	A	B	C	D	E
1	Name	RT	RI_msp	RI_input	Alert
842	5-Nonanol	21.528	1092		1095
843	2-Furanmethanol, propar	21.5884	1096	The content of the retention time actually detected was not retrieved	rt_is_in_silico
844	Fenchone	21.5884	1096	The content of the retention time actually detected was not retrieved	rt_is_in_silico
845	Pyrazine, 2-methoxy-3-(1	21.6555	1097	The content of the retention time actually detected was not retrieved	rt_is_in_silico
846	2-Undecene, (E)-	21.6555	1097	The content of the retention time actually detected was not retrieved	rt_is_in_silico
847	7-Octen-4-ol, 2-methyl-	21.6555	1097	The content of the retention time actually detected was not retrieved	rt_is_in_silico
848	o-Guaiacol	21.689			1097
849	Guaicol	21.7			1098
850	Pentanoic acid, butyl este	21.7227	1098	The content of the retention time actually detected was not retrieved	rt_is_in_silico
851	3-(Methylthio)propanoic	21.7227	1098	The content of the retention time actually detected was not retrieved	rt_is_in_silico
852	trans-Linalool oxide (fura	21.743	1086		1098
853	3-Nonanol	21.754	1095		1098
854	4-Nonanol	21.808	1088		1099
855	2H-Pyran-2-one, tetrahy	21.817	1092		1099
856	Methyl benzoate	21.87			1100
857	Benzoic acid, methyl ester	21.915	1094		1101
858	Butanoic acid, 2-methyl-	21.9263	1101	The content of the retention time actually detected was not retrieved	rt_is_in_silico
859	6-Nonenal, (Z)-	21.9263	1101	The content of the retention time actually detected was not retrieved	rt_is_in_silico
860	2-Cyclohexen-1-one, 4,4	21.9263	1101	The content of the retention time actually detected was not retrieved	rt_is_in_silico
861	Furan, 3-(4-methyl-3-per	21.9263	1101	The content of the retention time actually detected was not retrieved	rt_is_in_silico
862	2-Nonanone	21.94	1092		1101
863	Caryophyllenyl alcohol	21.95			1101
864	3-Hexen-1-ol, propanoat	21.962	1100	The content of the retention time actually detected was not retrieved	rt_is_in_silico
865	(Z,Z)-3,6-Nonadienal	21.962	1100	The content of the retention time actually detected was not retrieved	rt_is_in_silico
866	2-Undecene, (Z)-	21.9955	1102	The content of the retention time actually detected was not retrieved	rt_is_in_silico
867	Diallyl disulphide	21.999	1081		1102
868	Thujone	22.0648	1103	The content of the retention time actually detected was not retrieved	rt_is_in_silico
869	(E)-1-Allyl-2-(prop-1-en	22.0648	1103	The content of the retention time actually detected was not retrieved	rt_is_in_silico
870	1-Nonen-4-ol	22.0648	1103	The content of the retention time actually detected was not retrieved	rt_is_in_silico
871	Decane, 2,6,8-trimethyl-	22.134	1104	The content of the retention time actually detected was not retrieved	rt_is_in_silico
872	2,2,6-Trimethyl-3-keto-6	22.186			1105
873	Butanoic acid, 2-methyl-	22.2033	1105	The content of the retention time actually detected was not retrieved	rt_is_in_silico
874	Ethanone, 1-(4,5-dihydro	22.2726	1106	The content of the retention time actually detected was not retrieved	rt_is_in_silico
875	3-Cyclohexene-1-methar	22.2726	1106	The content of the retention time actually detected was not retrieved	rt_is_in_silico
876	Disulfide, dipropyl	22.3418	1107	The content of the retention time actually detected was not retrieved	rt_is_in_silico
877	1,3,8-p-Menthatriene	22.353	1119		1107
878	Unknown S	22.376			1107
879	Propanoic acid, hexyl este	22.4111	1108	The content of the retention time actually detected was not retrieved	rt_is_in_silico
880	2H-Pyran-3(4H)-one, 6-ε	22.4111	1108	The content of the retention time actually detected was not retrieved	rt_is_in_silico
881	2,6,6-Trimethylbicyclo[3,2	22.4111	1108	The content of the retention time actually detected was not retrieved	rt_is_in_silico
882	Phenol, 2,6-dimethyl-	22.4111	1108	The content of the retention time actually detected was not retrieved	rt_is_in_silico
883	Pentanoic acid, 3-methyl	22.4111	1108	The content of the retention time actually detected was not retrieved	rt_is_in_silico
884	Filifolone	22.42			1108
885	Hexanoic acid, propyl est	22.445	1094		1108

The warning information file: Contains the warning information after deduplication.

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

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.

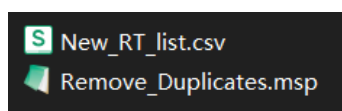
WARNING: Duplicates

Tips: Path to the resulting file for the sample data of the *library builder*:
https://github.com/yuanhonglun/WTV_2.0/tree/main/sample_data/library_builder_sample_data/export

2.2 Generating a Widely-Targeted SIM Method for Compound Detection (the *method generator module*)

2.2.1 Data Preparation

Before using this software, it is necessary to prepare two input files. The input files can directly utilize the corresponding result files generated by the *library builder* module.



These input files should include the compound names and their corresponding RT (Retention Time) information.

	A	B
1	Name	RT
2	Acetaldehyde	1.496
3	Methanethiol	1.526
4	Ethane	1.5523
5	Ethyl Chloride	1.555
6	Formaldehyde	1.559
7	Propene	1.5612
8	Methyl Alcohol	1.5677
9	Methyl formate	1.5695
10	Ethylene oxide	1.5727
11	Dimethylamine	1.5753
12	Ethylamine	1.5773

In addition, you should also have a Mass Spectrometry (MS) information file in the MSP format.

```
75452 126 30; 139 10; 140 10; 154 10; 182 130;  
75453 183 10;  
75454  
75455 Name: Tridecanal  
75456 InChIKey: BGEHHAVMRVXCGR-UHFFFAOYSA-N  
75457 Synon: n-Tridecylaldehyde  
75458 Synon: Tridecanaldehyde  
75459 Synon: Tridecyl aldehyde  
75460 Synon: 1-Tridecanal  
75461 Synon: Tridecane aldehyde  
75462 Synon: n-Tridecanal  
75463 Retention_index: SemiStdNP=1512/5/55 StdNP=1491/3/39 StdPolar=1823/7/28  
75464 Formula: C13H26O  
75465 MW: 198  
75466 ExactMass: 198.198365  
75467 CAS#: 10486-19-8; NIST#: 8800  
75468 DB#: 25925  
75469 Num Peaks: 71  
75470 15 7; 18 14; 26 13; 27 181; 28 56;  
75471 29 378; 30 18; 31 12; 39 112; 40 23;  
75472 41 732; 42 180; 43 932; 44 362; 45 146;  
75473 51 5; 53 41; 54 147; 55 666; 56 332;  
75474 57 999; 58 72; 59 8; 65 7; 66 77;  
75475 67 327; 68 443; 69 389; 70 304; 71 349;  
75476 72 84; 73 9; 77 7; 79 18; 80 21;  
75477 81 274; 82 700; 83 328; 84 167; 85 170;  
75478 86 23; 93 8; 94 15; 95 249; 96 339;  
75479 97 197; 98 90; 99 60; 100 7; 108 5;  
75480 109 114; 110 127; 111 90; 112 60; 113 20;  
75481 123 70; 124 87; 125 47; 126 31; 127 7;  
75482 137 32; 138 28; 139 5; 151 11; 152 64;  
75483 153 8; 154 117; 155 17; 170 20; 180 30;  
75484 181 5;  
75485  
75486 Name: 1-Tridecanal
```

2.2.2 Parameter Configuration

MethodGenerator-WTV

Import MSP library **MSP**

Import RT list **RT** ? **Demo**

☐ Set compound list ? **Demo**

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 ions 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) ?

Set export path **Save**

Run

Import MSP library

To input an MSP file, simply click the 'MSP' button and select the desired MSP file.

Import MSP library **MSP**

Imported MSP library:
Acetaldehyde.MSP

Once selected, the interface will display all the selected file directories, confirming a successful input.

Import RT list: Import a list of compounds and their RT.

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.

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

Solvent delay: Enter the start time for mass spectrometry detection, also known as the solvent delay time.

m/z range: The qualitative ions will be selected from within this range.

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

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

Similarity score threshold: When the similarity score is below the threshold of 0.85, two spectra are considered distinguishable. The default threshold is set to 0.85.

Prefer m/z threshold: M/Z values below this threshold will be assigned a lower weight in the calculation of the weighted score. The default threshold is set to 60.

Minimum ions number: This represents the minimum number of ions required for the selection of 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 the specified threshold. The F_R factor is typically set to be consistent with the minimum number of ions. For more specific and detailed information, please refer to the official documentation of the AMDIS software.

SIM segmentation parameters:

Maximum SIM segments: The maximum number of segments allowed in the SIM (Selected Ion Monitoring) 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

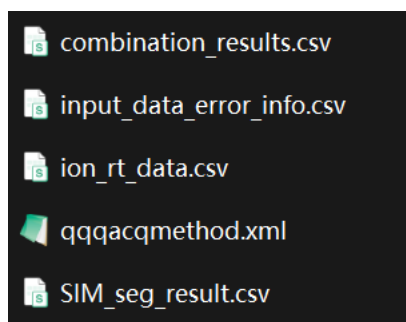
minimum dwell time, the data points per second will be decreased to ensure that the dwell time remains greater than or equal to the minimum dwell time. The default setting is 2.

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, choose a saving path, and the processed results will be saved to this directory.

2.2.3 Generated Widely-Targeted SIM Method

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



combination_results: This displays the combined information of target compounds after merging.

Name	RT	Ion_Combination	Note	Similar_Compound_List	SCL_Note
Ethanol	1.575	[45, 46]		☐	
Acetonitrile	1.5767	[38, 39]		☐	
Dimethyl sulfide	1.60034	[61, 62]		☐	
Butanal	1.99788	[44, 71]		☐	

input_data_error_info: This file highlights discrepancies in the input data that led to improper identification and analysis.

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

ion_rt_data: This file displays the results of the final compounds and their qualitative ions.

The content of this file is identical to the 'combination_results' file, but it is in a different format to make it more user-friendly for other purposes.

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 file presents the results of the SIM segments.

	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.

```

1  <?xml version="1.0" encoding="UTF-8" ?>
2  <MSAcqMethod xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xm
3  <msInstrument>QQQ</msInstrument>
4  <ionSource>EI</ionSource>
5  <tuneFile>atunes.elex.tune.xml</tuneFile>
6  <stopMode>ByChromatographTime</stopMode>
7  <stopTime>1</stopTime>
8  <solventDelay>0</solventDelay>
9  <collisionGasOn>true</collisionGasOn>
10 <sourceParameters>
11   <sourceParameter>
12     <id>SourceHeater</id>
13     <posPolarityValue>250</posPolarityValue>
14     <negPolarityValue>250</negPolarityValue>
15   </sourceParameter>
16 </sourceParameters>
17 <isTimeFilterEnabled>true</isTimeFilterEnabled>
18 <timeFilterPeakWidth>0.0133333337</timeFilterPeakWidth>
19 <timeFilter>
20   <activeCount>1</activeCount>
21   <definition>
22     <time>0</time>
23     <peakWidth>0.0133333337</peakWidth>
24   </definition>
25   <definition>
26     <time>10</time>
27     <peakWidth>0.05</peakWidth>
28   </definition>
29 </timeFilter>

```

Tips: Path to the resulting file for the sample data and demo data of the *method generator*:
https://github.com/yuanhonglun/WTV_2.0/tree/main/sample_data/method_generator_sample_data/export

2.3 Data Analysis (the *data analyzer* module)

2.3.1 Data Preparation

2.3.1.1 Mass Spectrometry File Format Conversion

Before data analysis, we need convert raw data to open formats (.mzML/.cdf) by open access application.

Data conversion to open format (.mzML)

Conversion with MSConvert

Download and install ProteoWizard from here (<http://proteowizard.sourceforge.net/downloads.shtml>) .

After installation, from the Start Menu, click the ProteoWizard folder and open MSConvert.

Click Browse and select file(s) for conversion. Then click Add to add them to the MSConvert workflow.

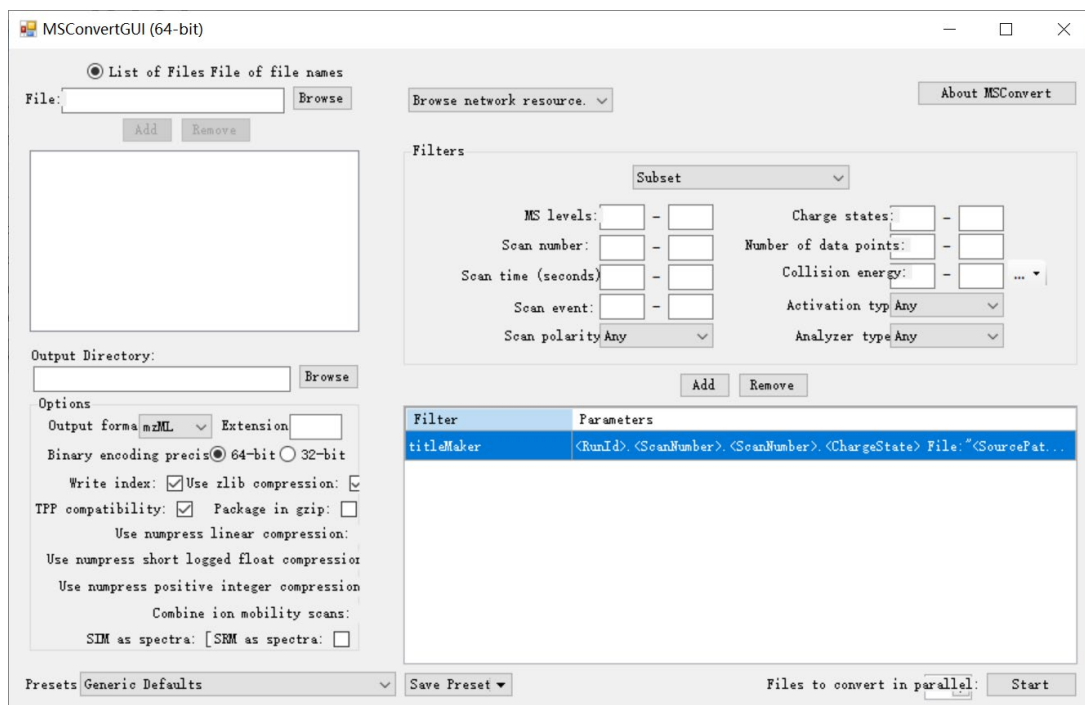
Choose an Output Directory.

Under Options, choose mzML (preferred) or mzXML for output format, 32-bit for binary encoding precision, SIM as spectra and uncheck Use zlib compression.

Click Add to add the filter.

Save the parameters for the next conversion. This will save you some time and prevent misconfiguration. In Presets (left bottom), click on Save Presets, and select "Save as default for the format".

Click on Start. Check your folder for the new .mzML files. Verify that these files open properly in Insilicos or TOPP View (OpenMS <http://www.openms.de/>).

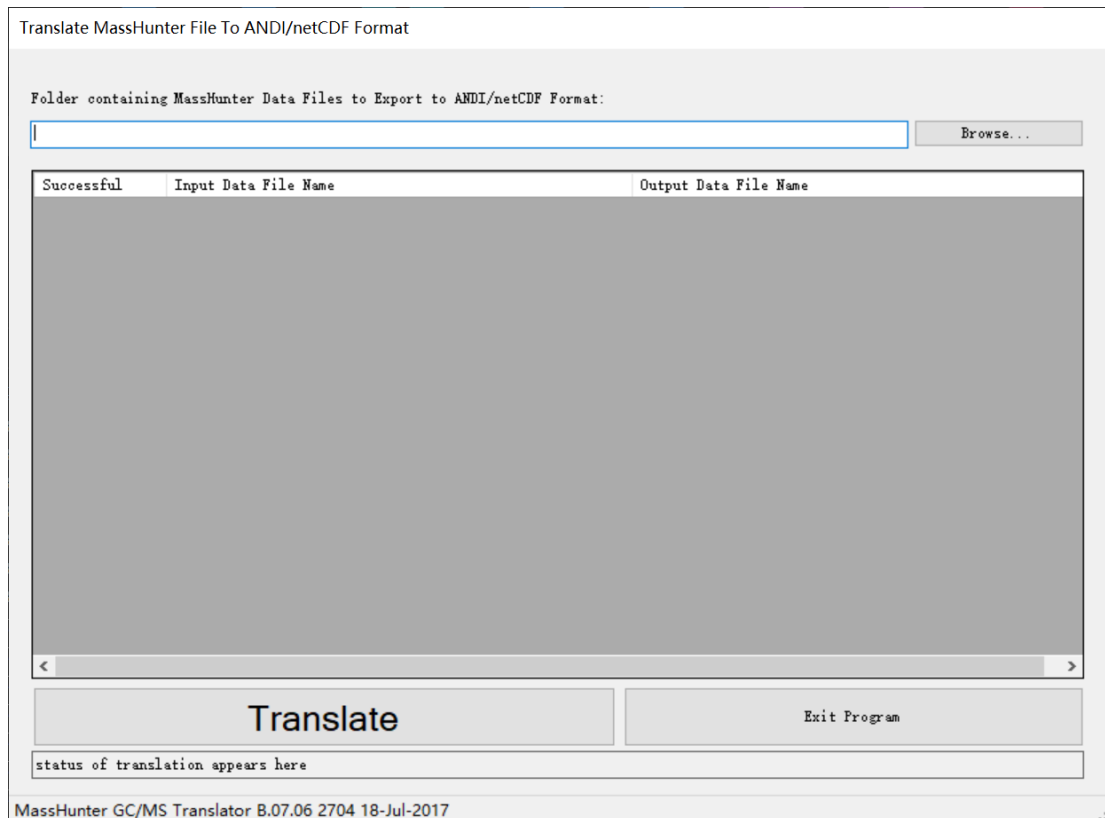


Data conversion to open format (.cdf)

Agilent: Click Browse and select folder containing file(s) for conversion.

Click Translate

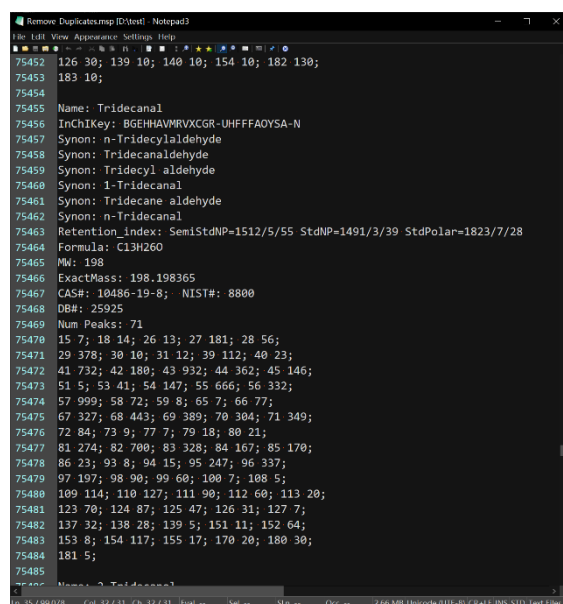
The CDF files are generated in the same folder as the raw data files.



For data processing methods from other vendors, please refer to the relevant data conversion instructions.

2.3.1.2 Compound Library (MSP) for identification

Input an MSP file that contains all the target metabolites. The input files can directly utilize the corresponding result files generated by the *library builder* module.



2.3.1.3 Optional: Retention Time Information

To identify compounds using actual measured retention times (RT), you will need the required files. The input files can directly utilize the corresponding result files generated by the *library builder* module.

	A	B
1	Name	RT
2	Acetaldehyde	1.496
3	Methanethiol	1.526
4	Ethane	1.5523
5	Ethyl Chloride	1.555
6	Formaldehyde	1.559
7	Propene	1.5612
8	Methyl Alcohol	1.5677
9	Methyl formate	1.5695
10	Ethylene oxide	1.5727
11	Dimethylamine	1.5753
12	Ethylamine	1.5773
13	2-Propenal	1.5779
14	Ethyl formate	1.5791
15	Butane, 2-methyl-	1.5798
16	Ethyl ether	1.5808

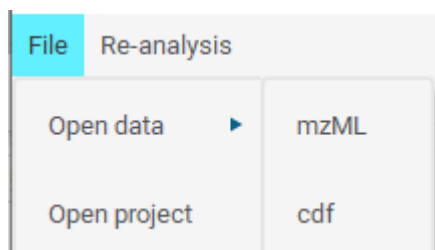
To identify compounds using Kovats Index calculation, you will need the necessary files.

	A	B
1	Num	RT (min)
2	8	5.219
3	9	9.34
4	10	15.177
5	11	21.877
6	12	28.764
7	13	35.514
8	14	41.988
9	15	48.144
10	16	53.937
11	17	59.401
12	18	63.787
13	19	64.851
14	20	65.484

2.3.2 Parameter Configuration

2.3.2.1 Data Import

Select the appropriate option based on the transformed file format.



2.3.2.2 Peak Detection

DataAnalyzer

×

Peak detection and compound perception

Smoothing factor:

5

↑

↓

?

Peak filter factor:

10.0

↑

↓

?

Bin number:

0.5

↑

↓

?

Next

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 exceeds 10, it will significantly extend the program's runtime. The default setting is 10.

Bin number: When using a low data point acquisition mode, such as two points per second, it is recommended to set a wide peak grouping width, for example, 0.5. For more specific and detailed information, please refer to the official documentation of the AMDIS software.

2.3.2.3 Qualitative Analysis

In RI mode

The screenshot shows the 'Qualitative analysis' window in the DataAnalyzer software. The window is divided into two main panels. The left panel contains the following settings:

- Library search window: 100
- Maximum RI: 300
- Peak group Match weight: 0.70
- Peak group Reverse Match weight: 0.30
- Minimum ions number in peak group for identification: 1
- Similarity score threshold: 0.40

The right panel is titled 'Calculate RI penalty:' and has a radio button selected. It contains the following settings:

- 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

At the bottom right of the window are 'Back' and 'Run' buttons.

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: The maximum RI (Retention Index) value.

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

Peak group Reverse Match weight: Peak group reverse retrieval matching score weight

Minimum ions number in peak group for identification: The minimum number of ions required for a peak group.

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

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

RI window: Set the RI window range in which no penalty will be applied.

RI window scale: The RI penalty window will be scaled linearly by this factor. Setting it to 0 disables this feature. The default value is 2.

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

Maximum penalty: Set the maximum penalty score.

No RI penalty: Peak groups without RI (Retention Index) will receive a penalty.

Inaccurate RI threshold: RI values below this threshold will be affected by the Inaccurate RI level factor.

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

In RT mode

The screenshot shows the 'DataAnalyzer' window with the 'Qualitative analysis' tab selected. The settings are organized into two columns. The left column contains: 'Library search window' (1.50), 'Peak group Match weight' (0.70), 'Peak group Reverse Match weight' (0.30), 'Minimum ions number in peak group for identification' (1), and 'Similarity score threshold' (0.40). The right column contains: 'Calculate RT penalty' (selected with a radio button), 'RT window' (±0.30mi), 'Level factor' (0.05), 'Maximum penalty' (0.10), and 'No RT penalty' (0.05). Each numerical input has a small up/down arrow and a question mark icon. At the bottom right, there are 'Back' and 'Run' buttons.

Qualitative analysis	
Library search window:	1.50
Peak group Match weight	0.70 ?
Peak group Reverse Match weight:	0.30 ?
Minimum ions number in peak group for identification:	1 ?
Similarity score threshold:	0.40 ?
Calculate RT penalty:	<input checked="" type="radio"/>
RT window:	±0.30mi ?
Level factor:	0.05 ?
Maximum penalty:	0.10 ?
No RT penalty:	0.05 ?

Back Run

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

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

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

Minimum ions number in peak group for identification: The minimum number of ions required for a peak group.

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

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

RT window: Set the RT window range in which no penalty will be applied.

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

Maximum penalty: Set the maximum penalty score.

No RT penalty: Peak groups with retention times (RT) lower than this value will not receive a penalty for mismatch.

Choose the mode for not entering retention time information.

The screenshot shows a window titled "DataAnalyzer" with a sub-header "Qualitative analysis". Inside the window, there are four configuration options, each with a text label, a numeric input field, and a question mark icon:

Parameter	Value	Icon
Peak group Match weight:	0.7	?
Peak group Reverse Match weight:	0.3	?
Minimum ions number in peak group for identification:	1	?
Similarity score threshold:	0.40	?

At the bottom right of the window, there are two buttons: "Back" and "Run".

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

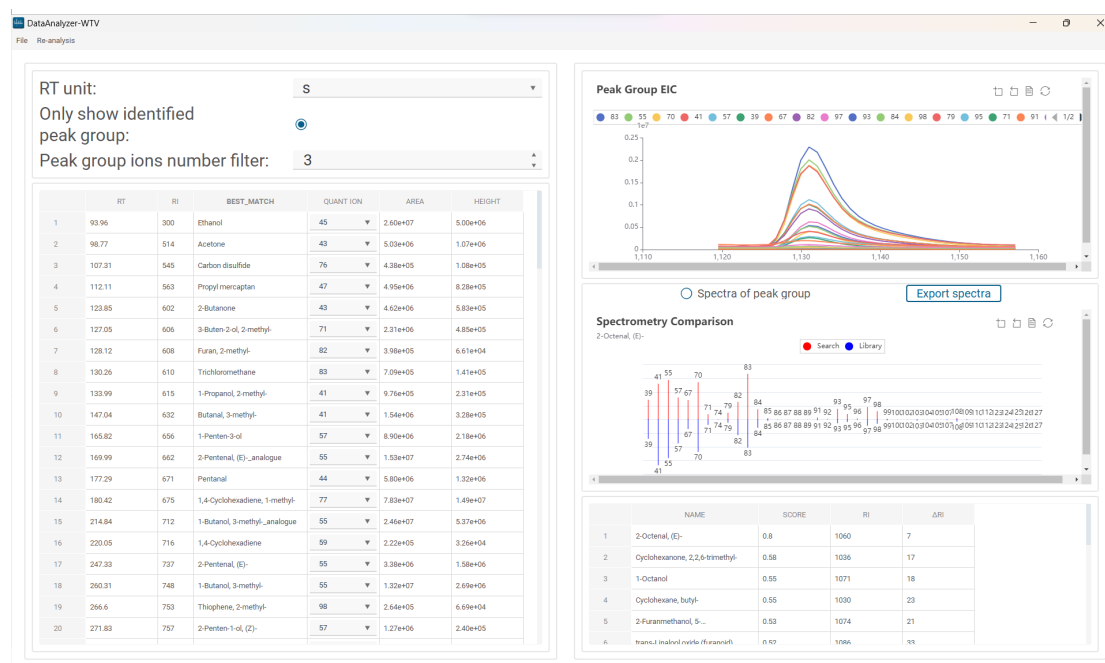
Peak group Reverse Match weight: Peak group reverse retrieval matching score weight.

Minimum ions number in peak group for identification:

The minimum number of ions required for a peak group.

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

2.3.3 Main Workspace Overview



RT unit: Users can use this option to switch the RT unit between min and s.

Only show identified peak group: After the user makes a selection, the result interface will only display peak groups that can match candidate compounds under the current threshold.

Peak group ions number filter: Users can use this feature to filter peak groups that contain a number of ions not less than the current setting.

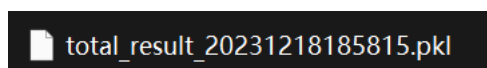
Results window:

The left panel displays the detected components, including their RT, RI, annotation results, quantitative ions, peak areas and heights. Users can also define the unit of

displayed RT, whether to show only qualified compounds, and the minimum ion number for showed components. The right panel displays the extraction ion currents (EICs), spectra comparison, and possible qualitative results for the selected component. Users can export qualitative results as a CSV file.

2.3.4 Results file

A pkl binary containing all the resulting information:



qualitative_and_quantitative_analysis_result.csv

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-mel	'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: Path to the resulting file for the sample data of the *data analyzer*:
https://github.com/yuanhonglun/WTV_2.0/tree/main/sample_data/data_analyzer_sample_data/export