

MET-COFEI User Manual

(Beta version - last updated: 04/16/2014)

The Samuel Roberts Noble Foundation, Inc.

MET-COFEI Description

MET-COFEI is a GC-MS Data Processing Platform for METabolite COMpound Feature Extraction and Identification, which is aiming to extract the pure spectrum that associated with metabolite compounds from the inputting GC-MS files, and then identify the compound by searching against an user specific GC-MS spectrum library. It mainly includes 3 sequential modules (Figure.1): compound feature extraction, compound identification and compound alignment. Compound feature extraction module include 3 sequential sub-modules: EIC extraction and Peak detection and peak filtering while compound identification module include 3 sequential sub-modules: peak grouping, pure spectrum reconstructing and library searching. EIC extraction aims to extract the meaningful mass trace slices from the start scan to the end scan. Peak detection aims to detect the local chromatograph peak for each EIC. Peak filtering aims to filter out some 'bad' quality peaks. Peak grouping is to cluster the detected peaks with the close retention time and peak shape similarity. Pure spectrum reconstructing is to build a compound related spectrum by combining all of the mz-intensity pair at apex position from all the peaks with the same group_id. Library searching is to search the constructed spectrum against an user specified GC-MS library (.msl file). Compound alignment is to align the same compound across different samples. The library searching and alignment are based on the calculated similarity score between two spectrums. The following figure.1 is the flow chart of MET-COFEI data processing.

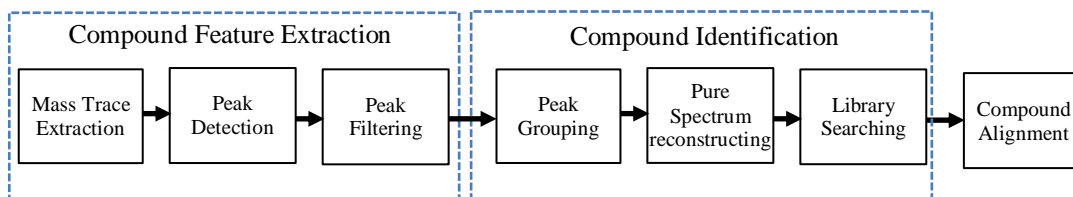


Fig.1 Flow chart of MET-COFEI Data Processing

In the latest version, all of the processing algorithms are coded in C++ and all of visualization parts are coded in C++/CLI. During the data processing, 3 peaklist files named as xxx_chromatograph_peaklist.csv, xxx_grouped_chromatograph_peaklist.csv, and xxx_identified_grouped_chromatograph_peaklis.csv will be produced for each sample (xxx_ means the sample name). If you choose some samples to align, another aligned peaklist file named as aligned_identified_grouped_chromatograph_peaklist.csv will also be generated in the end. Additionally, all of the final output peaklist and the intermediate extracted mass traces are stored in database by SQLite, each sample has a corresponding database file named as xxx_identified_grouped_chromatograph_peaklist.db, all of the files that chosen to align, will have a corresponding database file aligned_identified_grouped_chromatograph_peaklist.aligndb. The corresponding database file contains the peak shape information. Therefore, it realized the complete separation between data processing and result visualization. For the raw data (.CDF), you can view graphics such as TIC (Total Ion Chromatograph), spectrum data of each scan, 2D display of the raw data, and binning based EIC. For the output results visualization, you can open the database file only to view, the extracted EIC by mass trace method, the detected individual chromatograph peak, peaklist that have been grouped, and aligned.

The latest version of MET-COFEI support Batch mode and Parallel mode (MPI: Message Passing Interface) to run your multiple samples, depending the core number of your PC. Of course, the data processing time will saved and the required memory will increased, if you run at Parallel mode.

Additionally, considering the compatibility for 32bit and 64 bit CPU in MPI package, we are separated them into two packages. The users should download the corresponding package according to their own CPU hardware.

MET-COFEI Application

The following screenshot is METCOFEI software interface. All the application operation and parameters configuration can be finished by the software. There are 4 main parts (they are displayed as 4 item property page): **Data Process** for raw Data visualization and processing, **Parameter Setup** for processing, **Identification Result** for individual sample result visualization, and **Alignment Result** for multiple sample visualization after alignment.

Data Process

This property page let user to select the GC-MS data file from the loaded data file name list(.CDF) and visual the raw data, which include the TIC, spectrum of the specific scan determined by user mouse click position, 2D (spanned by mz-retention time) binarization visualization at the specific cutoff threshold. Additionally, in this property page, user can select the files to run and align from the loaded file name list. After select the parameter file (or change the processing parameters and click ‘Apply’ in the property page of Parameter Setting), the user can run the selected files. The following is the normal procedures for this property page:

1. Input the GC-MS data file(s): Click Browse button to select CDF file(s) and click “Load Data” button to display the list of the file name(s) to the table. See Fig.2.

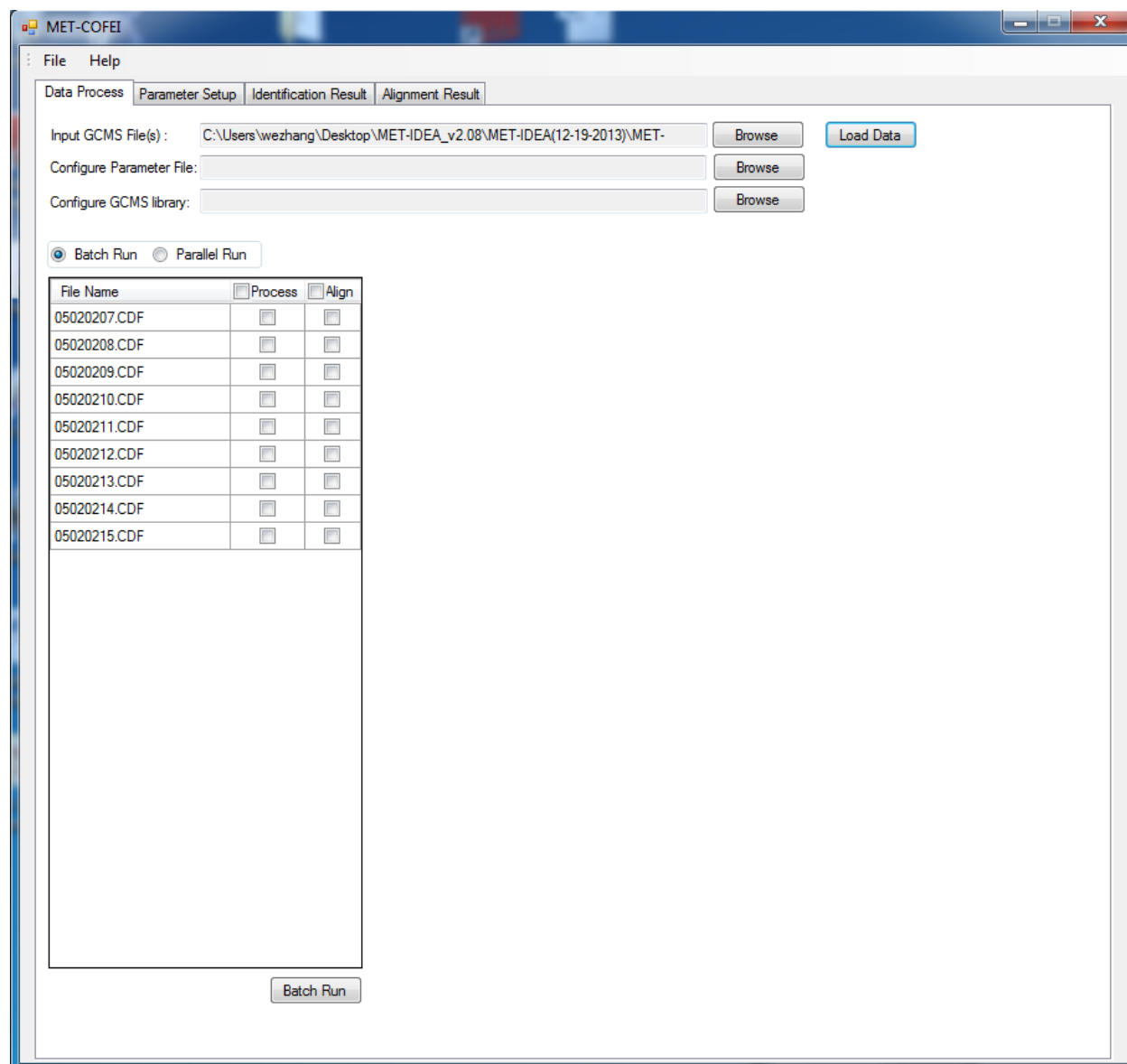


Fig.2 Load Data file names

2. Once data file names are loaded, select the file name on the table to visualize the TIC and the spectrum of each scan. TIC can be displayed as scan number mode(see Fig.3) or retention time mode(see Fig.4). Here, the retention time unit is second. Additionally, the raw data also can be displayed as 2D model, if user specific a cutoff threshold(see Fig.5).

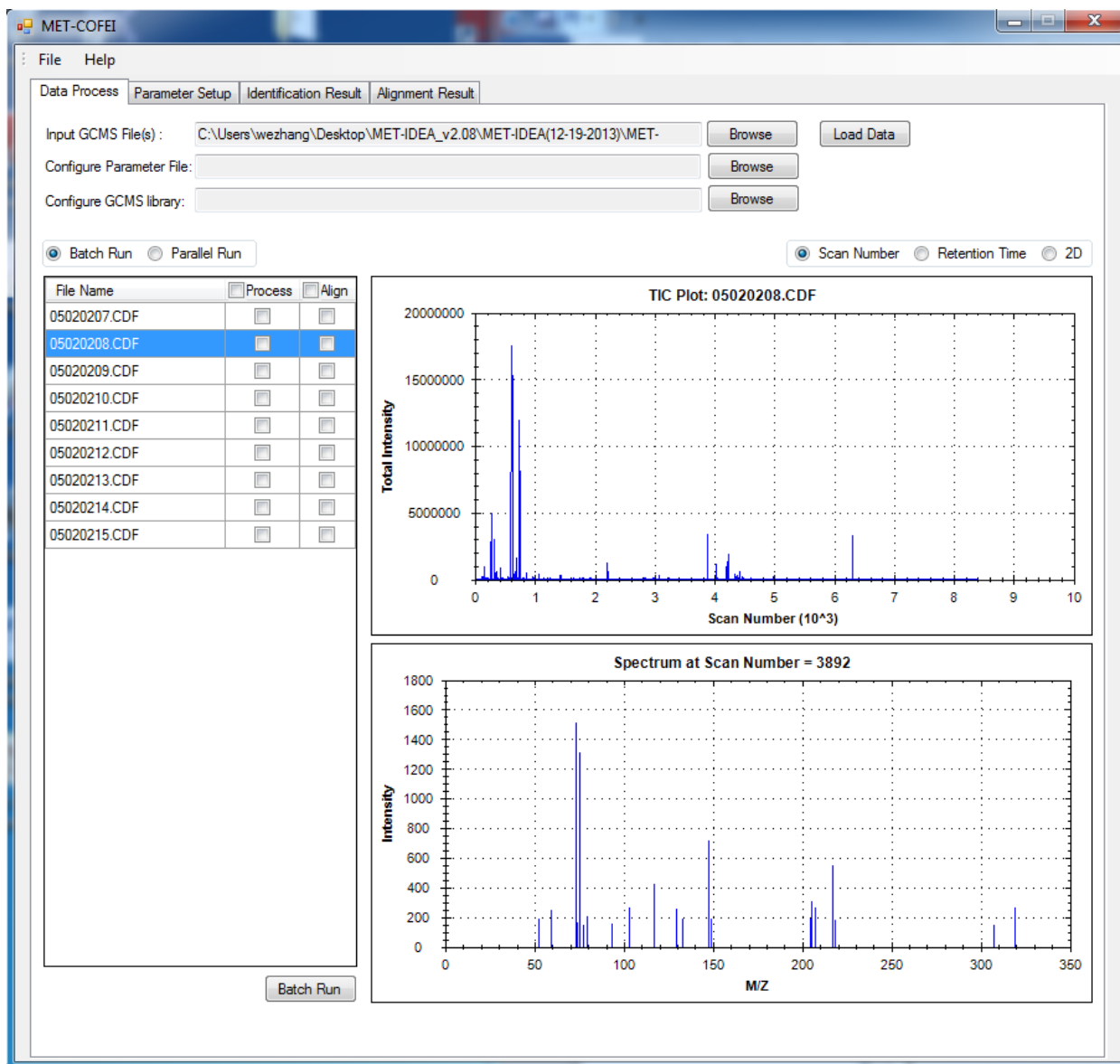


Fig.3 Visualization of Raw GC-MS Data and TIC are plotted at Scan mode

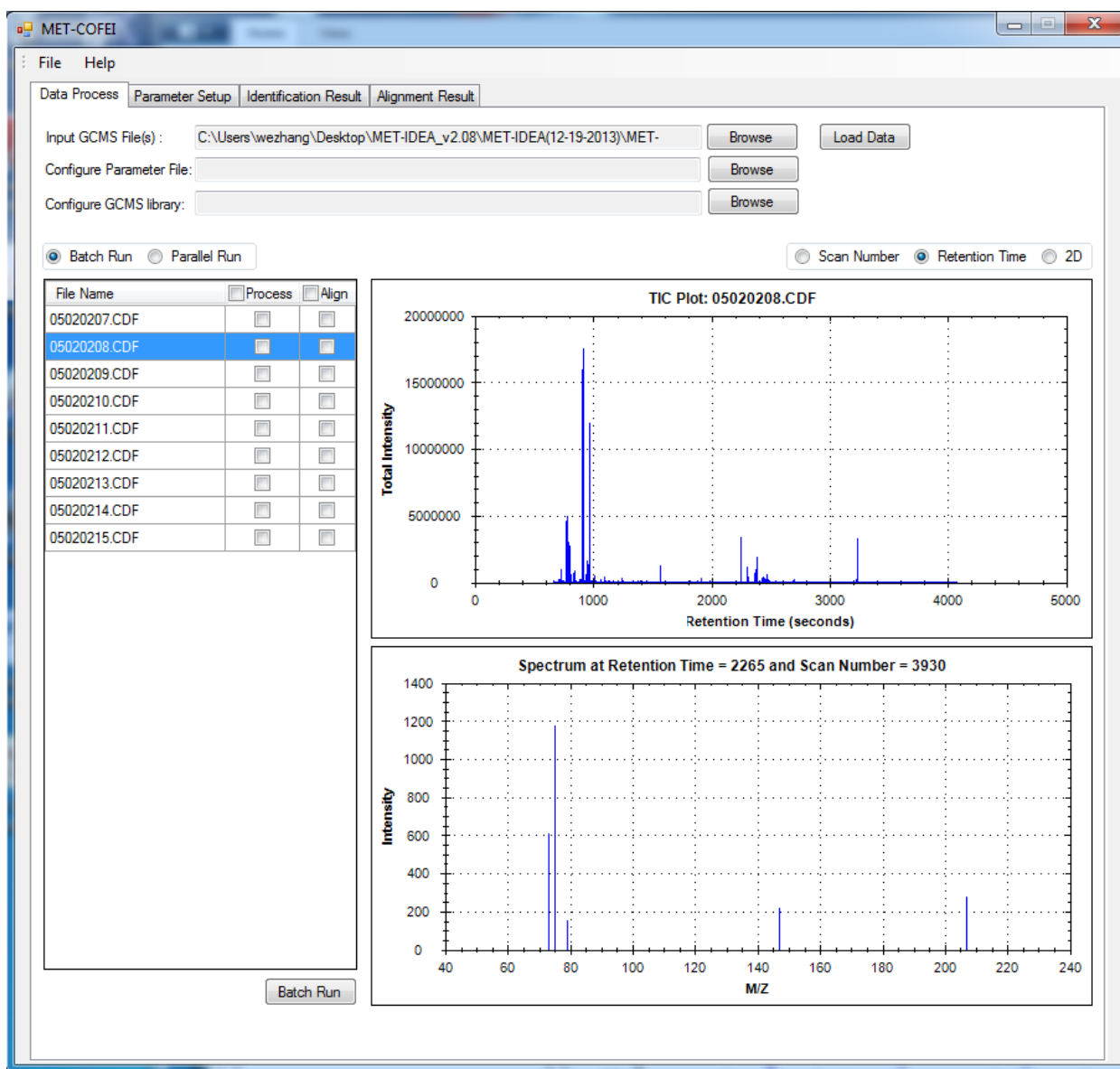


Fig.4 Visualization of Raw GC-MS Data and TIC are plotted at retention time mode

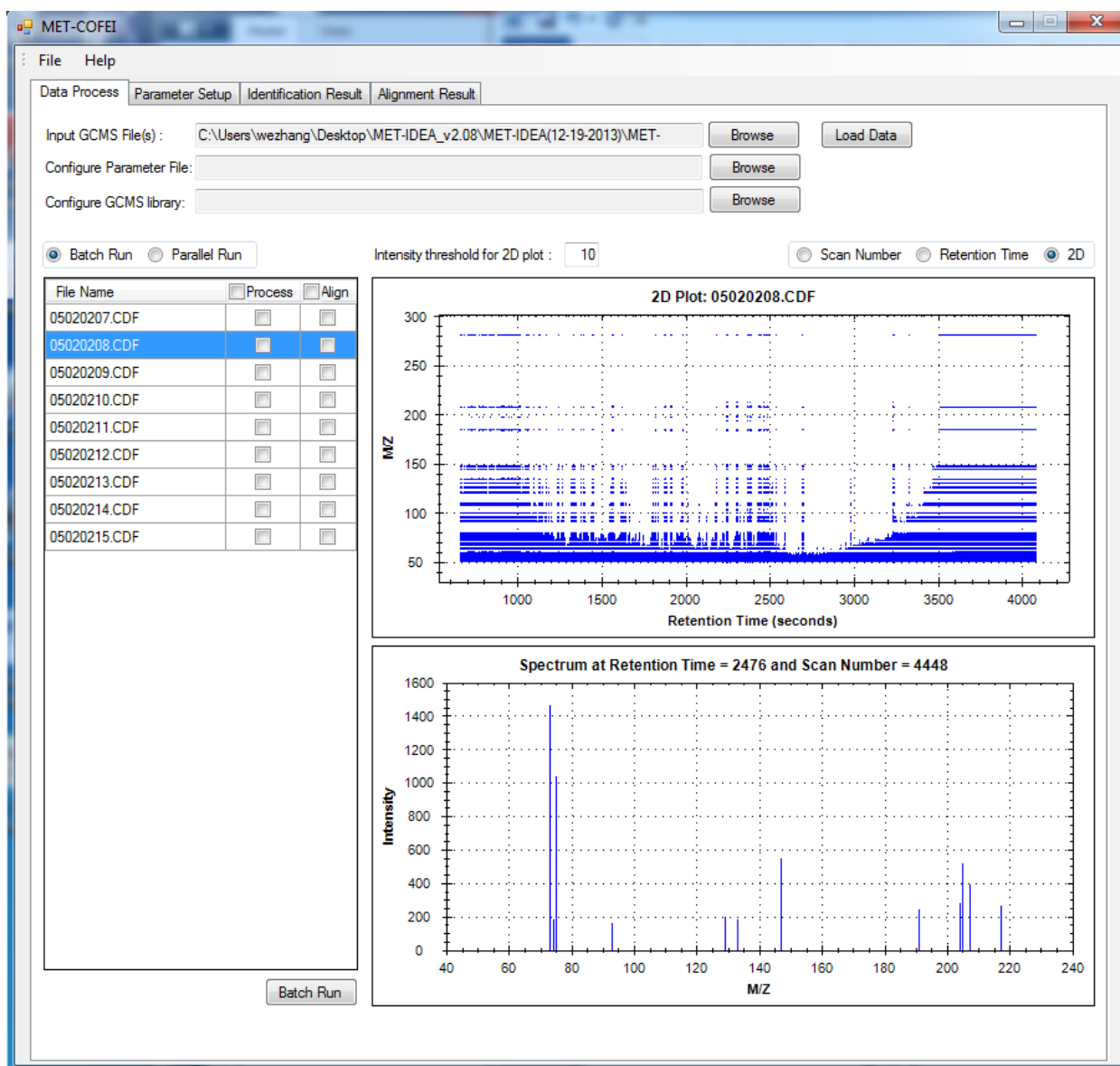


Fig.5. Visualization of Raw GC-MS Data in 2D mode.

3. Select the Configure Parameter File. A default configure parameter file named as “config_para.csv” is snapshot in Fig.6. (Parameter can be further configured by software property page of Parameter Setup).
4. Select the configured GC-MS library file (.msl), a default library file named as “MET_COFEI_Test_Lib.msl” is snapshot in fig.7.
5. Check process checkbox and (or) align checkbox for desired sample file(s) to run.
6. Click “Batch Run” or “Parallel Run” button to run. See Fig.8 and Fig.9.

Then a file named as “Job_Run_Config.csv” will automatically generated, which recorded the information for this job. See Fig.10.

- The output files will be created in the “Result” folder where the GC-MS CDF files are loaded.

	A	B	C	D
1	Para_Profile_Centroid	1		
2	Para_Cutoff_Mode	0		
3	Para_Cutoff_Intensity	70		
4	Para_Cutoff_TopNumber	500		
5	Para_Cutoff_LowPercent	0.1		
6	Para_Start_Scan	100		
7	Para_End_Scan	1960		
8	Para_PPM	80		
9	Para_Min_EIC	6		
10	Para_Misscount_Allow	3		
11	Para_InterTrace_Dist_Min	0.06		
12	Para_Min_PeakWidth	6		
13	Para_Max_PeakWidth	50		
14	Para_Branch_Gap_Allow	2		
15	Para_Min_Branch_Length	6		
16	Para_Min_Span	2		
17	Para_MarkerBranch_Dif_Th	8		
18	Para_PeakIntensity_Th	500		
19	Para_SNR_Th	2		
20	Para_Peak_Significance_Th	1		
21	Para_TPASR_Th	0.7		
22	Para_Zig_Zag_Index_Th	0.6		
23	Para_Group_Scan_Shift_Tol	3		
24	Para_Group_Shape_Angle_Th	20		
25	Para_Library_Fragment_Mass_Tol	0.1		
26	Para_Library_Matching_Score_Th	750		
27	Para_One_Two_Phase_Align	1		
28	Para_Align_Fragment_Mass_Tol	0.08		
29	Para_Align_Matching_Score_Th	700		
30	Para_Align_Window1	10		
31	Para_Align_Window2	5		
32				

Fig.6. One Configured Parameters file for MET-COFI

MET_COFI_Test_Lib.ms1 - Notepad	
File Edit Format View Help	
NAME:Ribitol TMS	
CASNO:0120-30070101-N1001	
RT:37.624	
RW:	
NUM PEAKS: 170	
(53 2) (54 2) (55 6) (56 1) (57 2) (58 6) (59 26) (60 2) (61 3) (66 1) (67 1) (69 4) (70 2) (71 3) (73 1000) (74 84) (75 60) (76 3) (77 2) (81 4) (83 2) (85 3) (86 1) (87 5) (88 4) (89 22) (90 2) (91 1) (97 1) (99 3) (101 24) (103 39) (104 38) (105 17) (106 1) (111 2) (113 1) (114 1) (115 6) (116 14) (117 123) (118 12) (119 11) (120 1) (121 1) (125 1) (127 3) (129 190) (130 23) (131 42) (132 6) (133 94) (134 13) (135 8) (136 1) (141 2) (142 2) (143 11) (144 2) (145 6) (147 501) (148 78) (149 48) (150 5) (151 2) (153 1) (155 5) (156 1) (157 21) (158 3) (159 4) (161 7) (162 1) (163 9) (164 1) (165 1) (169 1) (170 1) (171 3) (172 9) (173 2) (174 1) (175 15) (176 3) (177 9) (178 2) (179 1) (185 1) (187 1) (189 171) (190 35) (191 74) (192 12) (193 6) (194 1) (201 2) (202 1) (203 49) (204 164) (205 454) (206 92) (207 50) (208 7) (209 2) (215 3) (217 775) (218 200) (219 79) (220 14) (221 29) (222 6) (223 1) (224 1) (229 24) (230 7) (231 5) (232 1) (237 1) (242 2) (243 62) (244 14) (245 8) (246 1) (247 1) (249 1) (263 1) (265 3) (266 1) (277 73) (278 23) (279 13) (280 3) (281 1) (289 1) (291 27) (292 9) (293 5) (294 1) (303 1) (305 12) (306 21) (307 278) (308 79) (309 39) (310 8) (311 2) (317 35) (318 12) (319 405) (320 120) (321 58) (322 11) (323 3) (331 3) (332 47) (333 16) (334 8) (335 2) (351 1) (395 10) (396 4) (397 2) (398 1) (407 3) (408 1) (409 2) (422 22) (423 8) (424 5) (425 1)	
NAME:5-methoxy salicylic acid - 2TMS	
FORM:C14H24O4S12	
CASNO:5-methoxy salicylic acid - 2TMS	
RT:27.401	
COMMENT: 27.401 min MEDIA_LWS12DEC05100.FIN	
SOURCE:C:\DATABASES & LIBRARIES\MS_LIBRARIES\Noble_LIB5\methoxySalicyclic acid.ms1	
NUM PEAKS: 152	
(51 11) (52 3) (53 9) (54 2) (55 8) (56 1) (57 3) (58 9) (59 21) (60 3) (61 6) (62 2) (63 12) (64 3) (65 5) (66 3) (67 5) (68 1) (69 4) (70 3) (71 5) (73 790) (74 67) (75 55) (76 5) (77 19) (78 11) (79 26) (80 2) (81 1) (83 20) (84 3) (85 6) (87 2) (88 1) (89 7) (90 3) (91 21) (92 5) (93 6) (94 3) (95 5) (96 1) (97 3) (98 2) (99 3) (101 1) (103 8) (104 4) (105 12) (106 5) (107 27) (108 3) (109 6) (110 2) (111 4) (113 2) (115 6) (116 1) (117 4) (118 2) (119 11) (121 33) (122 7) (123 4) (124 1) (125 1) (127 1) (131 9) (133 33) (134 7) (135 21) (136 6) (137 27) (138 4) (139 2) (141 3) (143 2) (145 2) (147 40) (148 8) (149 52) (151 15) (152 3) (153 9) (154 1) (157 2) (159 1) (161 1) (163 4) (165 76) (166 13) (167 5) (168 1) (177 2) (179 17) (180 52) (181 25) (182 2) (183 2) (189 2) (191 3) (192 1) (193 5) (194 3) (195 4) (196 5) (197 1)	

Fig.7. One Configured GC-MS library file for MET-COFI

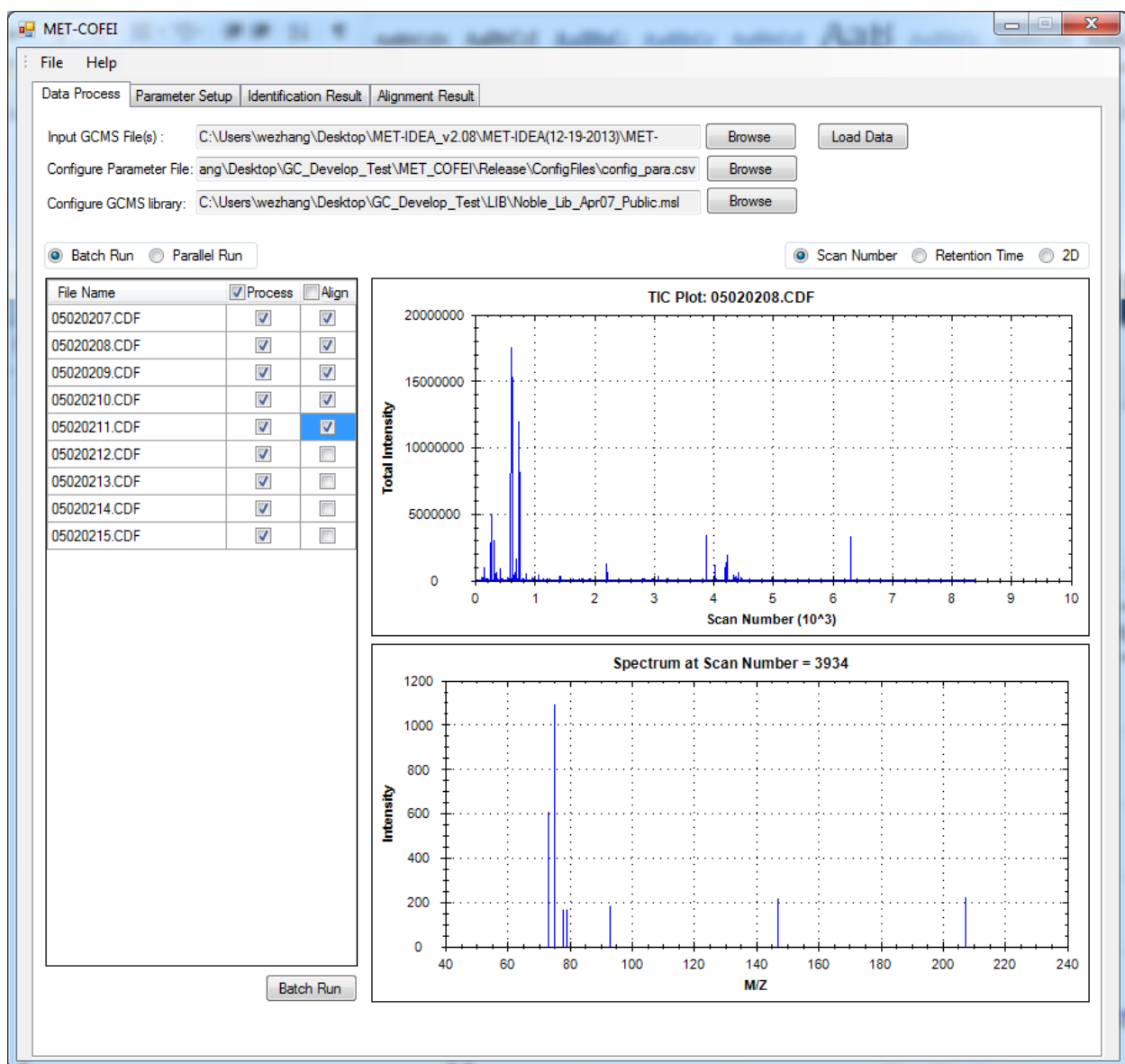


Fig.8 Select files to Batch Run

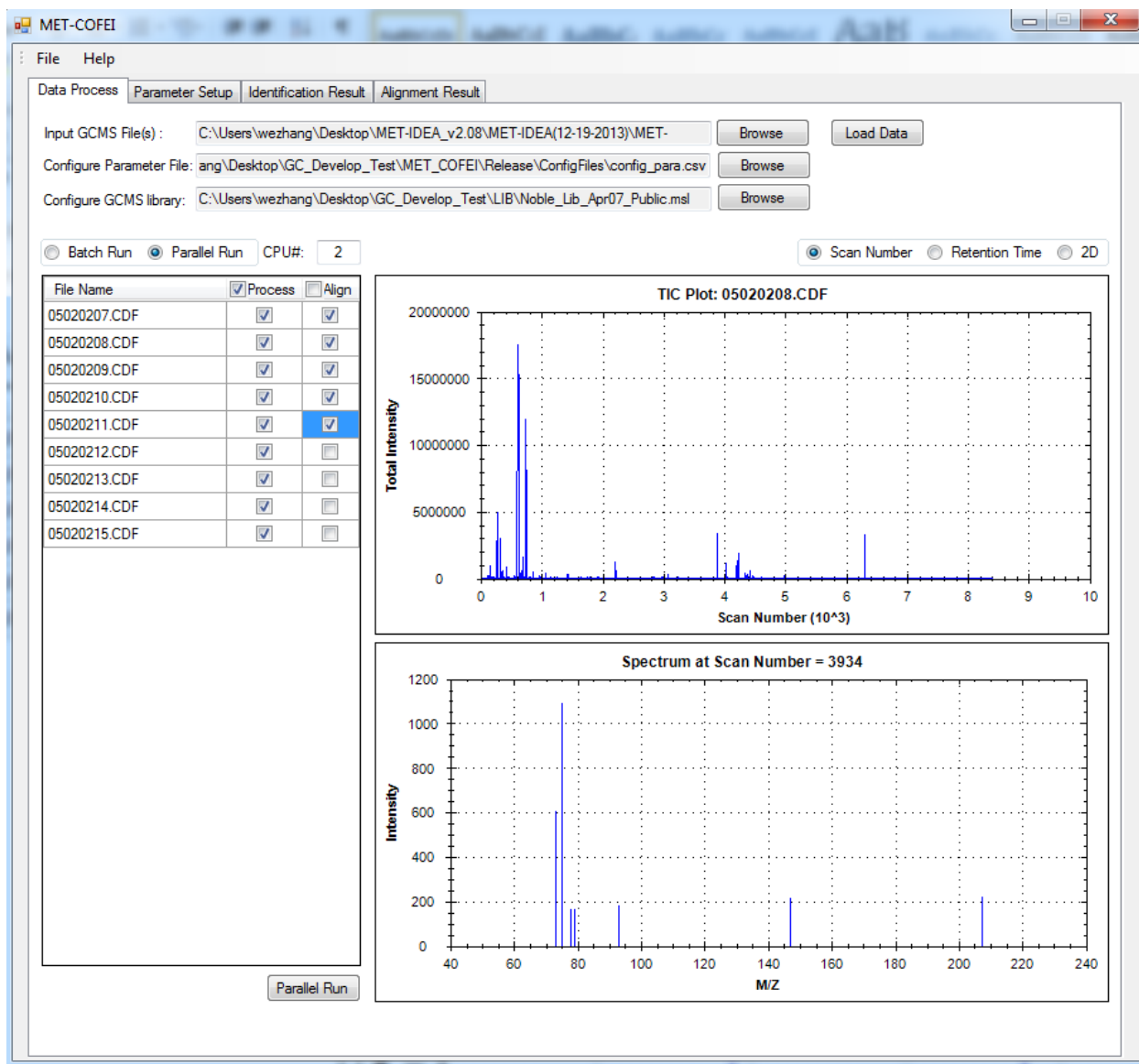


Fig.9 Select files to Parallel Run

Job_Run_Config.csv													
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Job Create Time:	Tue Apr 15 17:59:33 2014											
2	Config Parameter Path	C:\Users\wezhang\Desktop\GC_Develop_Test\MET_COFEI\Release\ConfigFiles											
3	Data File Path	C:\Users\wezhang\Desktop\MET-IDEA_v2.08\MET-IDEA(12-19-2013)\MET-IDEA_2.05_Mar_16_2011\Lloyd\050202001-30CDFs											
4	Output Path	C:\Users\wezhang\Desktop\MET-IDEA_v2.08\MET-IDEA(12-19-2013)\MET-IDEA_2.05_Mar_16_2011\Lloyd\050202001-30CDFs\Result											
5	Run Mode	Parallel M CPU Num											
6	Process File Name	Aligned											
7	05020207.CDF	1											
8	05020208.CDF	1											
9	05020209.CDF	1											
10	05020210.CDF	1											
11	05020211.CDF	1											
12	05020212.CDF	0											
13	05020213.CDF	0											
14	05020214.CDF	0											
15	05020215.CDF	0											
16													
17													
18													
19													

Fig.10 Content of Job Run Config File

Parameter Setup

In this property page, the user can configure the processing parameter (Fig.11). The different parameter setting will generate very different results. If the user wants to acquire the optimal results by setting the optimal parameters, please read the parameter explanation section and parameter optimization section first.

After parameter configuring, then Click “Apply” button, it will save the modified parameters to file user specific parameter file (default as “config_para.csv”) (see Data Process property page). Only the user click the button of ‘Apply’, the newly configured parameters can be loaded into parameter setup panel and can used in the following data processing. If the user wants to use the new configured parameter, go back to “Data Process” page, and click Batch Run or Parallel Run.

The parameters configuration includes 6 parts: Data type configuration, Parameter configuration for Mass traces (EIC) Extraction, Parameter configuration for CWT based Peak detection, Parameter configuration for peak quality filtering, Parameter configuration for peak grouping, and library searching, Parameter configuration for Compound alignment. Regarding Data type configuration, you need to select the inputting data as profile data or centroid data. For profile data, the centroid processing module will be called at first. These data configuration should be correct, otherwise, you can’t get the correct meaningful run result.

The same Job configure file(Job_Run_Config.csv) and the same processing parameter file (config_para.csv) can ensure the same result for the same data file set.

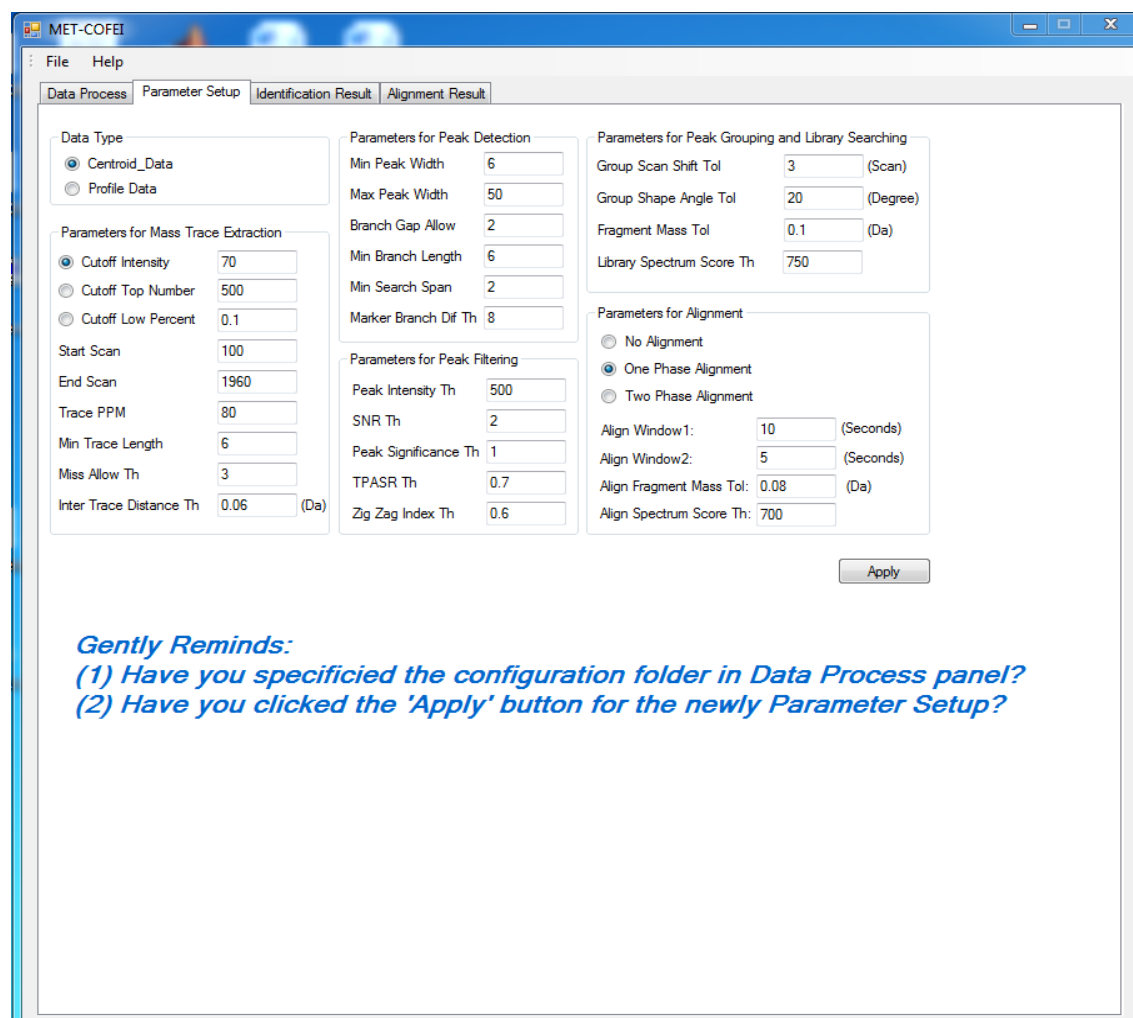


Fig.11 Parameter Configuring

Identification Result

In this property page, the user can view the chromatograph peak shape for each detected peak, peaklist for a group with the same Group_ID. We suppose the chromatograph peaks from the same metabolite should have the close retention time and peak shape, so we separated peaklist into different Group_ID, if the peaks' retention time (corresponds to peak apex) and peak shape meets some criteria. In this property, MET-COFEI visualization can clearly provide the relationship of an individual peak with the related peaklist with the same Group_ID, so, the user can double check the performance of peak grouping module(also called de-convolution in other GC-MS tools). The user only need to click the radio check button at "Peak Plot", or "Group Plot", the detailed visualization for the individual peak, peaklist with the same Group_ID will be plotted.

Additionally, MET-COFEI visualization also can clearly provide the relationship of the individual peak and the related whole extracted EIC by mass trace method or binning method (you need to specific the raw CDF file).

Regarding to the peaklist with the same Group_ID, the corresponding pure re-constructed spectrum can be generated and plotted. So, the user can click 'Export' button to export the pure re-constructed spectrum into a 'txt' file. Compared with the raw mixed spectrum, the spectrum constructed with the same Group_ID become more pure by separation according to the criteria of retention time and peak shape. The constructed spectrum is considered to a pure spectrum that related to a potential compound, which can be searched against a library. If the user also specific the library based identification file (xxx.IDEN), the matched degree between the constructed spectrum(upward) and library searched spectrum(downward) can be together plotted.

The following is the normal procedures for this property page:

1. Select the result file (xxx_identified_grouped_chromatograph_peaklist.db) from result directory. Usually, after the data processing, the processing result will be generated in the same folder of the original GC-MS data files located.
2. Click a cell in the table to visualize the individual peak (Check Radio of 'Peak Plot'), associated peaks with the same Group_ID(Check Radio of 'Group Plot')
3. Export the constructed spectrum into a txt file.
4. Select the library based identification file (xxx.IDEN) to check matching degree between the constructed spectrum and the library searching spectrum.
5. Select the raw CDF file and manually configure the m/z tolerance (ppm or Da) to view the binning based EIC.

If the user wants to view the meaningful peak or peaklist, you can click the head of the column with the Group_ID, or Apex_MZ, and then the whole table displayed in the left part will be sorted as ascending or descending order. The user can view the associated peaks in the lower right part at the normalized (focus on peak shape similarity) or standard, Scan Number or Retention time mode, see Fig.12, and Fig.13.

For Group Plot, a constructed-spectrum will be generated according to the m/z values and apex intensity values of the associated peaks, if the user click the button of 'Export'. One of the constructed spectrum using peaklist with same Group_ID is showed in Fig.14.

For more details about the matching between the constructed spectrum and the library spectra, the user can open the identification file(xxx.IDEN) to double check. One of a identification result file is showed in Fig.15.

From the extracted EIC, the user can know the shape information of the front and back of the specific peak. Fig.16 provide the visualization of the specific peak and the whole extracted EIC. Because the EIC is extracted based an object tracing method, and the following chromatograph peak is detected by a CWT based method, so the user can optimize the parameters for mass tracing module and CWT based peak detection module.

If the raw CDF file data is selected, the binning EIC and TIC also can be displayed. Compared the difference between the extracted EIC and binning EIC (specific m/z and tolerance), the user can optimize parameter for mass trace extraction. See Fig.17.

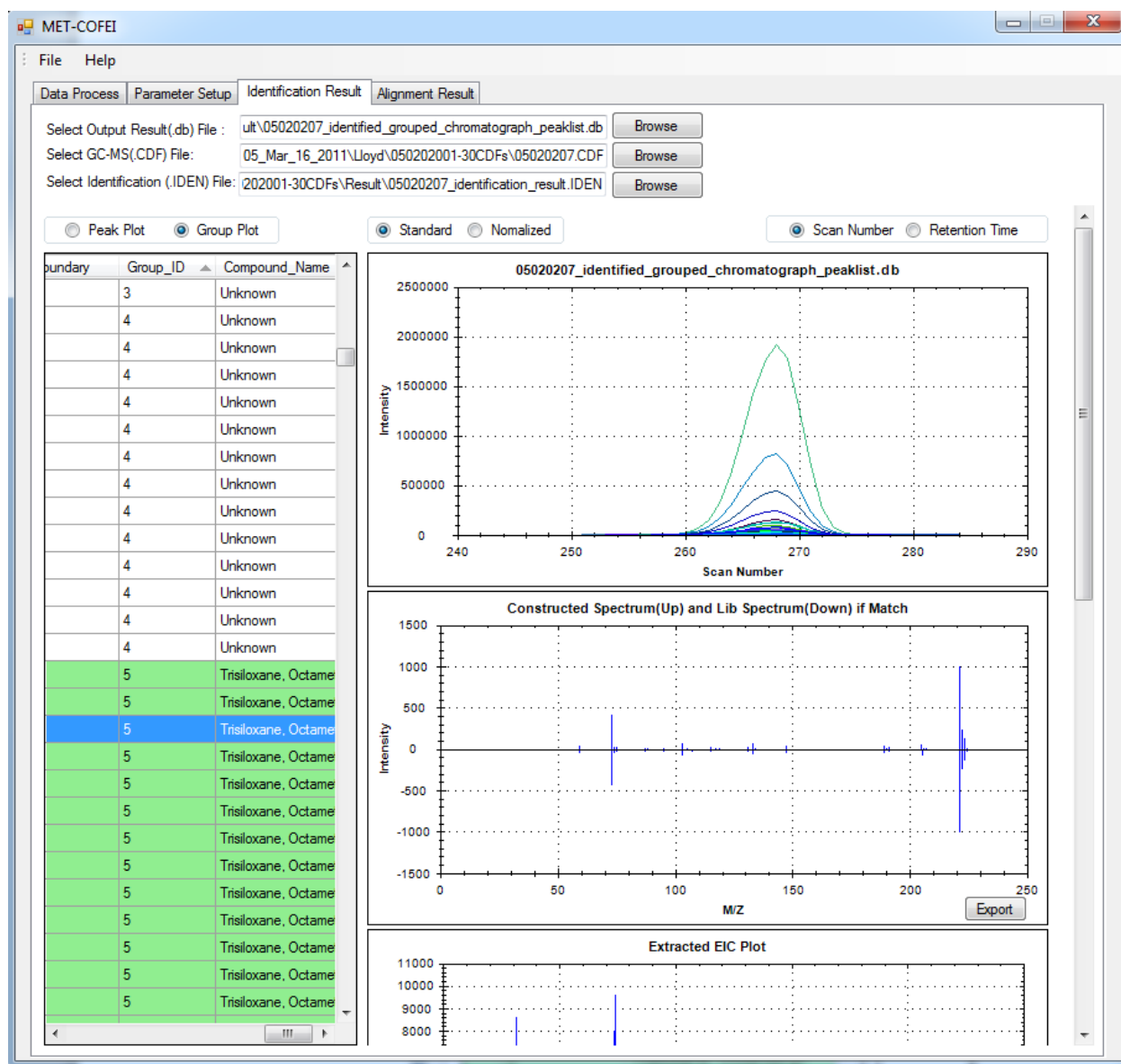


Fig.12 Visualization of the associated un-normalized peaks with the same Group_ID and the constructed and library matching spectrums.

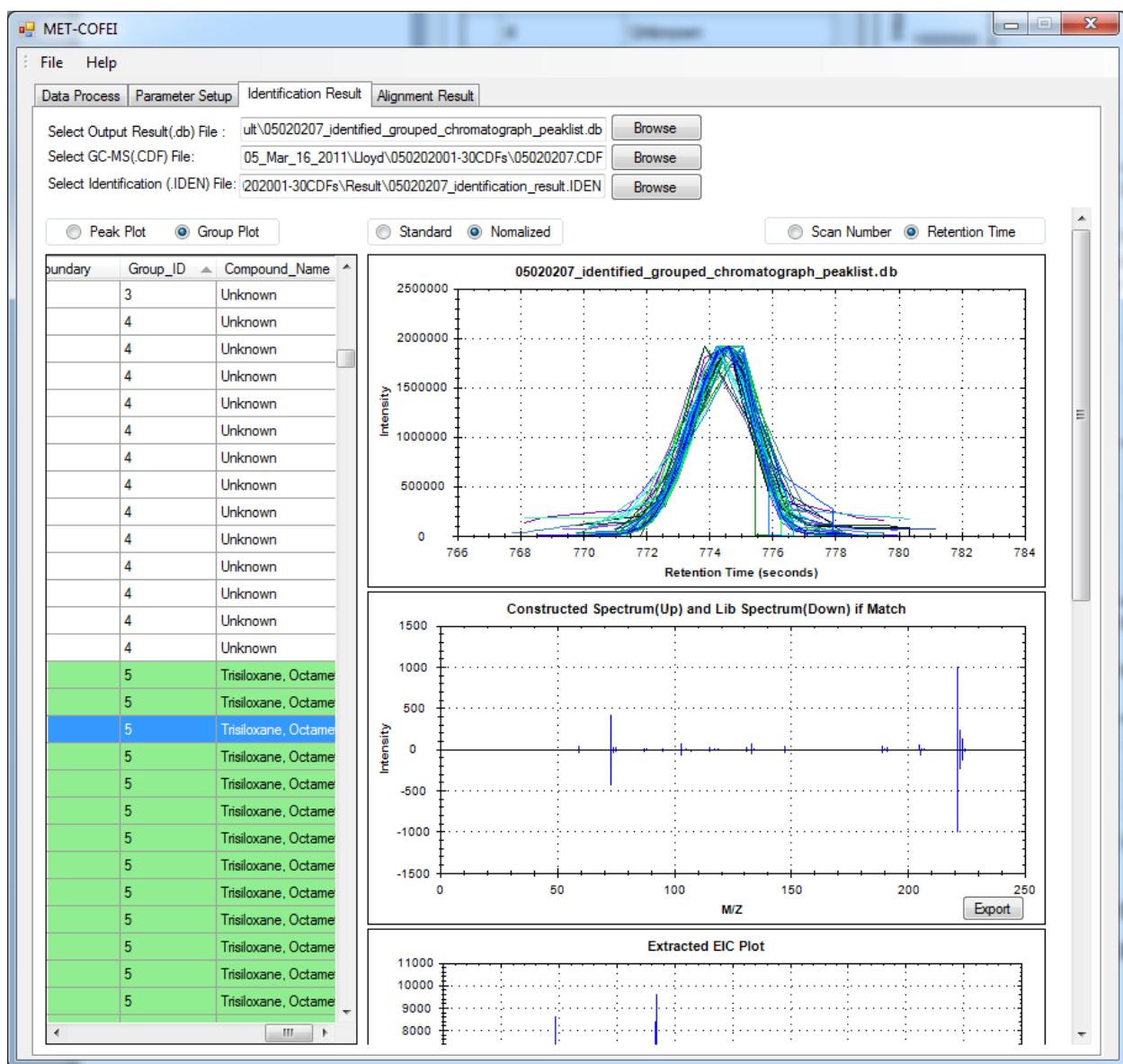


Fig.13 Visualization of the associated Normalized peaks with the same Group_ID and the constructed and library matching spectrums

pseudospectrum_group_id5.txt					
	A	B	C	D	E
1	M/Z	Intensity			
2	59	38.53861			
3	115.9	3.456452			
4	88	6.438059			
5	148	6.196463			
6	76	1.440188			
7	223	126.8702			
8	75	25.42246			
9	73	424.5926			
10	132.9	67.65128			
11	86.9	13.75067			
12	142.9	1.320694			
13	222	230.3299			
14	207	10.35266			
15	204.9	62.47495			
16	158.9	5.163806			
17	193.9	0.383007			
18	95	15.89634			
19	70.9	3.674568			
20	59.9	2.425882			
21	191.9	4.071663			
22	56.9	1.986521			
23	79.9	3.643781			
24	81.9	2.276646			
25	60	3.166849			
26	190.9	22.10376			
27	147	38.68054			
28	144.9	5.148673			
29	224	19.94974			
30	60.9	6.303955			
31	88.8	5.287474			
32	116.9	16.56008			
33	225	5.136672			
34	206.9	10.07714			
35	80.9	3.273819			
36	100.9	2.839676			
37	118.9	11.25434			
38	117.9	2.803671			
39	90.9	0.718529			
40	221	1000			
41	66.9	1.446972			
42	103	77.04382			
43	85	6.291953			
44	208	1.461582			
45	128.9	1.858156			
46	122.9	0.289171			

Fig.14 Content of constructed Spectrum exported by peaklist with Group_ID=5


```
05020207_identification_result.IDEN - Notepad
File Edit Format View Help
Group_ID:5
Extracted Spectrum:
(59 73856) (115.9 6624) (88 12338) (148 11875) (76 2760)
(223 243136) (75 48720) (73 813696) (132.9 129648) (86.9 26352)
(142.9 2531) (222 441408) (207 19840) (204.9 119728) (158.9 9896)
(193.9 734) (95 30464) (70.9 7042) (59.9 4649) (191.9 7803)
(56.9 3807) (79.9 6983) (81.9 4363) (60 6069) (190.9 42360)
(147 74128) (144.9 9867) (224 38232) (60.9 12081) (88.8 10133)
(116.9 31736) (225 9844) (206.9 19312) (80.9 6274) (100.9 5442)
(118.9 21568) (117.9 5373) (90.9 1377) (221 1.91642e+006) (66.9 2773)
(103 147648) (85 12058) (208 2801) (128.9 3561) (133.9 17800)
(189.9 18752) (101 5928) (101.9 3887) (188.9 92952) (192.9 4360)
(91 1129) (74.9 48680) (114.9 37192) (54.9 2483) (104.9 15358)
(75.9 2600) (55 1880) (80 9156) (176.9 9087) (87.9 13838)
(174.9 16672) (162.9 3375) (130.9 46616) (74 71336) (58.9 64456)
(106.9 1671) (71 6561) (134.9 10166) (89.9 1090) (148.9 6299)
(205.9 28256)
Library Match Spectrum:
Name:Trisiloxane, Octamethyl-
Match Score:753.722
(55 4) (57 2) (59 38) (60 3) (61 6)
(63 1) (65 2) (66 5) (67 7) (71 4)
(73 424) (74 37) (75 25) (80 6) (81 5)
(82 3) (85 6) (87 14) (88 7) (89 5)
(91 10) (92 2) (95 24) (101 3) (103 75)
(105 11) (107 21) (108 3) (110 4) (113 2)
(115 16) (117 14) (118 3) (119 11) (122 7)
(129 2) (131 23) (132 3) (133 53) (134 7)
(135 5) (145 6) (147 34) (148 6) (149 3)
(159 5) (161 4) (175 8) (177 6) (189 43)
(190 10) (191 20) (192 4) (205 65) (206 13)
(207 9) (208 2) (221 1000) (222 226) (223 123)
(224 19) (225 6)
Group_ID:7
Extracted Spectrum:
(86.9 25432) (103 8110) (114 8343) (101.9 9557) (82.9 1782)
(74.9 11773) (69.9 12716) (144.9 3341) (60.9 6681) (175 154752)
(118.9 3324) (72.9 328640) (114.9 6339) (163 8824) (55.9 8235)
(99.9 44360) (161 169664) (68.9 2952) (102 11925) (116 9358)
(160 986496) (177 12324) (64.5 8552) (131 6205) (88 20640)
(86 486656) (58 13600) (118 7773) (116.9 49680) (57.9 8341)
(87.9 21216) (52.9 2211) (83.9 3283) (117.9 7922) (112.9 4133)
(103.9 1882) (51 5383) (115.9 9178) (59.9 12084) (130.9 5647)
(162 78096) (113.9 8496) (128.9 812) (129.9 20440) (87 50320)
(54.9 4987) (102.9 8839) (75 13199) (59 186048) (70.9 7345)
(56.9 8569) (73.9 29272) (57 9314) (100.9 12070) (53 1920)
(56 8447) (104.9 670) (60 14871) (79.9 1289) (176 28064)
(117 51424) (80.9 856)
Library Match Spectrum:
Name:bis-trimethylsilyl amine
Match Score:949.058
(51 3) (53 1) (54 1) (55 4) (56 7)
(57 8) (58 11) (59 164) (60 13) (61 6)
(64 7) (69 3) (70 12) (71 7) (73 303)
(74 27) (75 10) (80 1) (81 1) (82 1)
(83 2) (84 3) (86 526) (87 52) (88 22)
(89 2) (98 1) (99 6) (100 43) (101 12)
(102 10) (103 9) (104 2) (105 1) (113 4)
(114 9) (115 6) (116 10) (117 53) (118 8)
(119 4) (128 1) (129 1) (130 20) (131 6)
(132 3) (133 1) (144 23) (145 4) (146 4)
(158 2) (160 1000) (161 172) (162 79) (163 9)
(164 2) (172 1) (175 156) (176 29) (177 12)
(178 1)
Group_ID:8
Extracted Spectrum:
(175 86864) (57 5095) (174 481344) (189 29456) (100.9 35312)
(75 8280) (59.9 7233) (86.9 7489) (176.9 3794) (55 2968)
(54.9 2976) (71 3737) (86 67040) (131 15226) (69.9 8801)
(115 4519) (61 2703) (88 3113) (130 24840) (130.9 11454)
(69 2385) (83.9 1506) (68.9 2294) (131.9 4696) (172 4296)
(116 8374) (73.9 16076) (158.9 1051) (71.9 7700) (190 6848)
```

Fig.15 Content of the identification result file

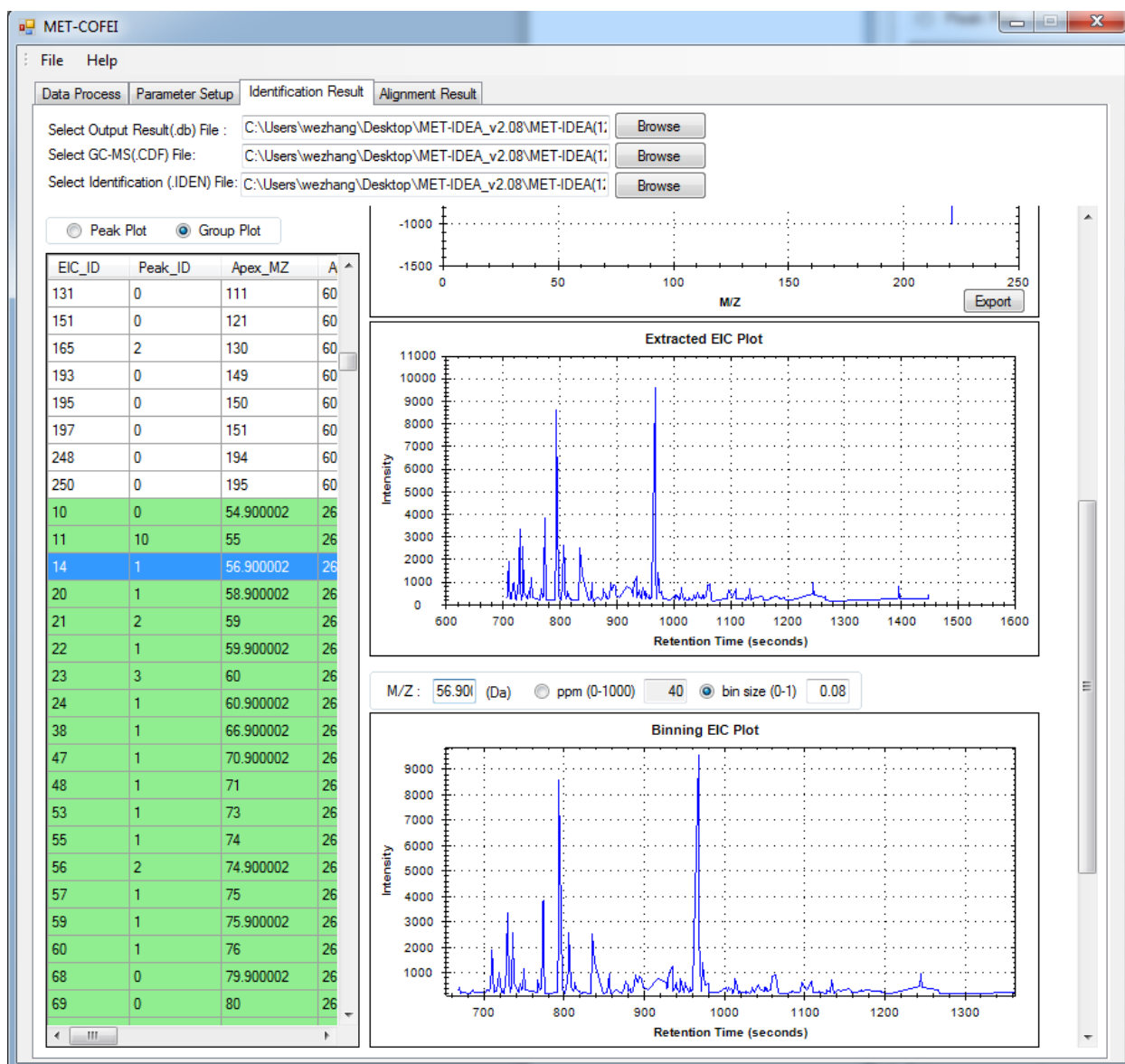


Fig.17 Visualization of the extracted EIC for the specific peak and the binning EIC from raw CDF file

Alignment Result

In this property page, the user can visualize the peaklist with the same Align_ID across different samples. From the visualization (see Fig.18, Fig.19), all of the peaks with the same Align_ID across different samples are plotted and the peaks from the same Sample are plotted with one specific color (Right top panel). Additionally, the peak associated with the mouse click can also be plotted individually, or with the associated peaklist with the same Group_ID (Right bottom panel). So, in this property page, the user can clearly know the relationship of the same compound associated peaks across different samples, and the relationship between the individual peak and the peaklist with the same Group_ID.

The alignment procedure is based on the retention time and the similarity score between each constructed spectrum pairs. If the retention time and constructed spectrum similarity score fall in the users' specific tolerance, the compound related peaks will be aligned together.

The following is the normal procedures for this property page:

1. Select the result file name “aligned_identified_grouped_chromatograph_peaklist.aligndb”.
(This database file will be generated only if you choose some files for alignment.(See the Property Page of Data Process)
2. Switch the retention time mode between RT_Original and RT_Corrected to check the alignment results.
3. Click a cell in the table to visualize the peaklist that have been aligned into the same Align_ID across different sample files.

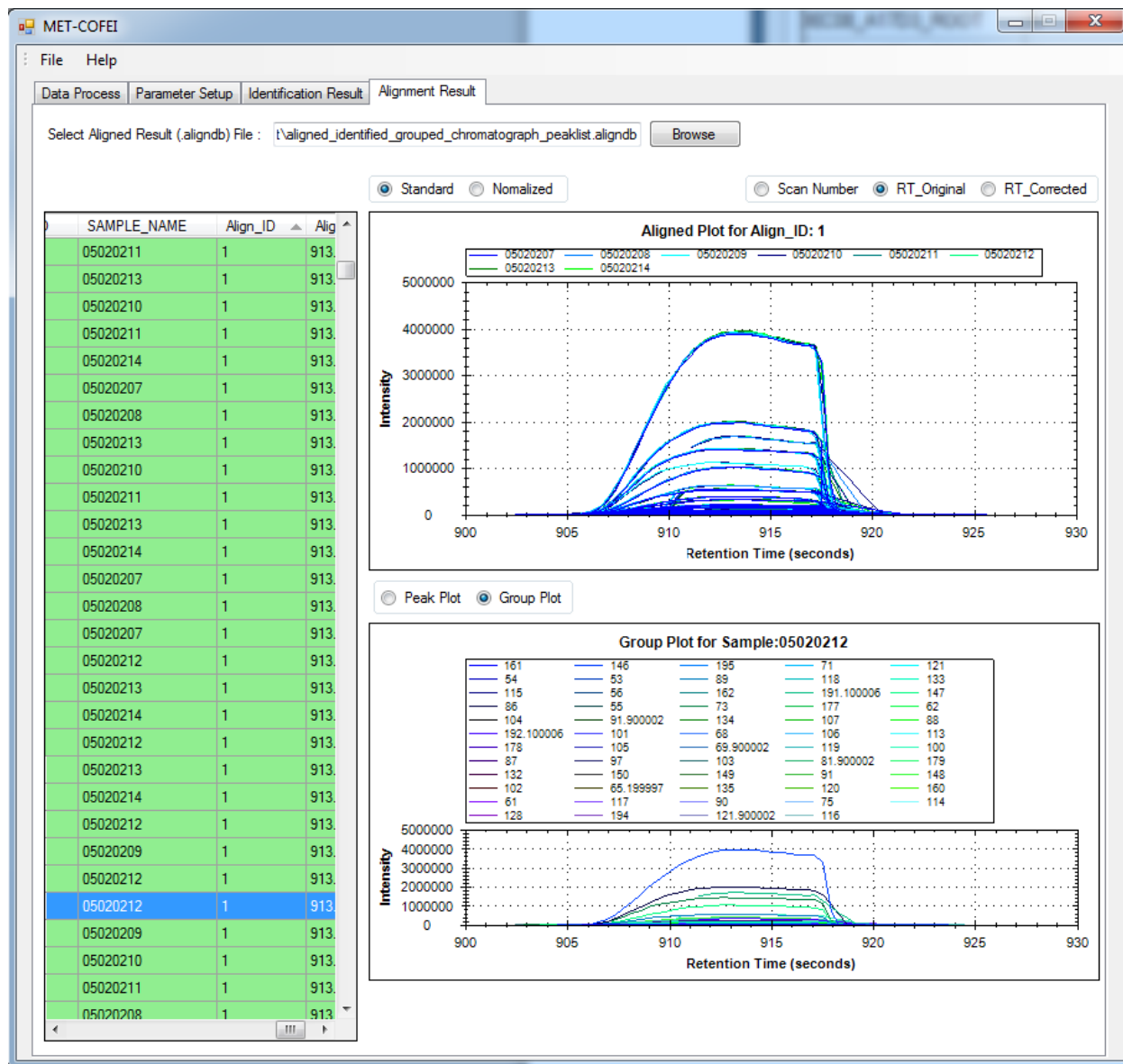


Fig.18 Visualization of peaklist with the same Align_ID across different samples displayed at RT_Original mode



Fig.19 Visualization of peaklist with the same Align_ID across different samples and displayed at RT_Corrected mode

Parameter Explanation

1. Cutoff Intensity: Intensity cutoff threshold for each scan. Only the data point with the intensity is larger than the threshold is used to do mass trace extraction, by which the low intensity noisy data point can be filtered
2. Cutoff Top Number: For each scan, only the Top Number intensity data points are considered to do mass trace extraction, by which the low intensity noisy data point can be filtered.
3. Cutoff Low Percent: For each scan, only the percentage of low intensity data points is filtered.
4. Start Scan: Specific the start scan for mass trace extraction.
5. End Scan: Specific the end scans for mass trace extraction.
6. Trace PPM: Specific the PPM threshold of m/z value variation for a valid mass trace.
7. Min Trace Length: Specific the minimum mass trace length.
8. Miss Allow Th: Specific the maximum allowed miss data point number during the mass trace extraction.
9. Inter Trace Distance Th: Specific the minimum distance between two neighboring mass trace

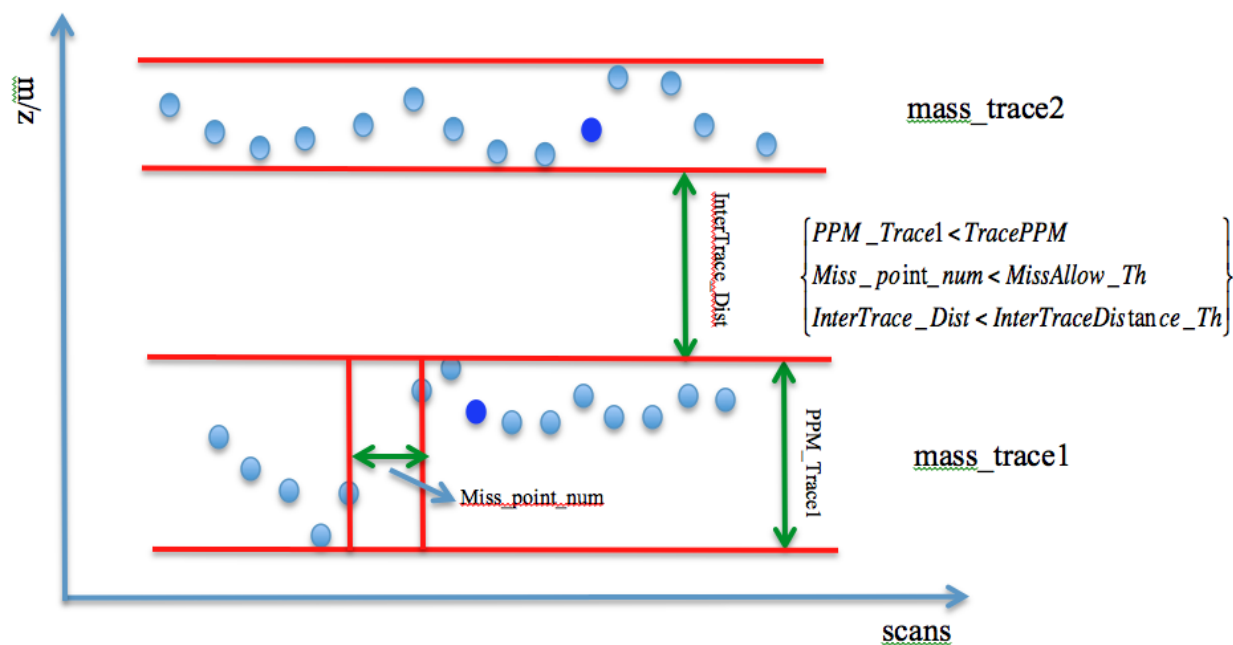


Fig.20 Illustration for some Parameter of Mass Trace

10. Min Peak Width: The minimum width of a valid peak, the very thin peaks with peak width smaller than MinPeakWidth will be filtered.
11. Max Peak Width: The maximum width of a valid peak, the very fat peaks with peak width larger than MaxPeakWidth will be filtered.
12. Peak intensity Th: The minimum intensity value of a valid peak, the very low intensity peaks with intensity smaller than Peakintensity_Th will be filtered.

13. SNR Th: The minimum S (Signal)/N (Noise) value of a valid peak, the peaks with SNR smaller than SNR Th will be filtered. SNR is defined in the wavelet domain by the ratio of the CWT coefficient at marker point to 95% quintile of the absolute CWT coefficient in scale 1.
14. Peak Significance Th: The minimum Peak Significant level of a valid peak, the peaks with the Peak Significant level smaller than Peak Significance Th will be filtered. Peak significant level is defined by the ratio between the mean intensity value of data points near the peak apex and the mean intensity value of data points near the two boundaries.
15. TPASR Th: The maximum Triangle Peak Area Similarity Ratio (TPASR) of a valid peak, the peaks with TPASR larger than TPASR Th will be filtered.

Triangle Peak Area Similarity Ratio (TPASR) is defined as the following formula,

$$\begin{cases} TPA = 0.5 * Peak_Width * Intensity(Peak_Apex) \\ RPA = \sum_{i=Left_Boundary}^{Right_Boundary} Intensity(i) \\ TPASR = \frac{|TPA - RPA|}{TPA} \end{cases}$$

Here, TPA is the Triangle peak area and RPA is the real peak area. TPASR provides an index for the closeness of the detected peak and triangle peak in area. The TPASR value is more close to 0, the better of the peak quality.

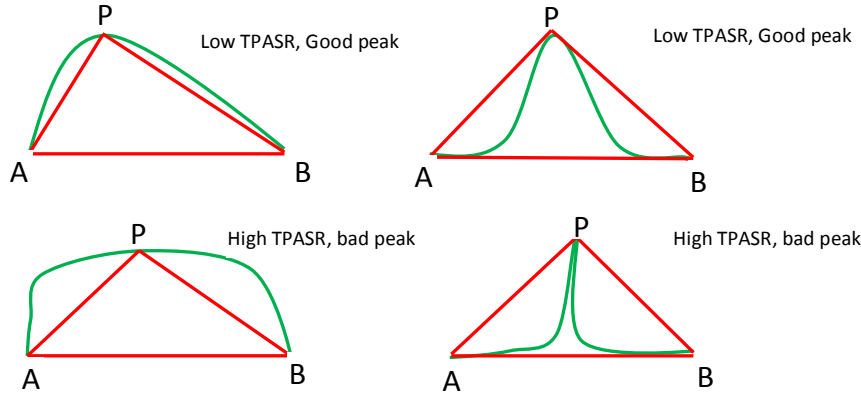


Fig. 21 Illustration of Parameter of TPASR for fat peak and thin peak

16. Zig Zag Index Th: The maximum Zig Zag Index of a valid peak, the peaks with Zig Zag Index larger than Zig Zag Index Th will be filtered. Zig Zag Index is adopted to evaluate the degree of zig-zag of a chromatograph peak. Zig Zag Index can be calculated by the following procedure.

Suppose the intensity array of a chromatograph peak is represented as

$$I_1, I_2, \dots, I_{n-1}, I_n, I_{n+1}, \dots, I_N.$$

- 1) Calculate the effective peak intensity by subtract the baseline at the peak apex.
 $EPI = \text{Max}(I_1, I_2, \dots, I_{n-1}, I_n, I_{n+1}, \dots, I_N) - \text{Baseline}(\text{Apex}).$
- 2) Calculate the first order derivative of the peak and acquire the increment for each data

point pair.

$$d_n = I_n - I_{n-1}, d_{n+1} = I_{n+1} - I_n \quad n = 2, 3 \dots N.$$

- 3) Calculate the variance of each two neighbor increment pair as:

$$v(d_n, d_{n+1}) = (d_n - d_{nmean})^2 + (d_{n+1} - d_{nmean})^2 \text{ and } d_{nmean} = \frac{(d_n + d_{n+1})}{2.0}$$

After some simple deducing, the variance can be represented as:

$$v(d_n, d_{n+1}) = 0.5 * (2I_n - I_{n-1} - I_{n+1})^2$$

Here $(2I_n - I_{n-1} - I_{n+1})^2$ indicate the local zig zag degree of data point I_{n-1}, I_n, I_{n+1} .

- 4) Sum all of the local zig zag, we get

$$Sum_zig_zag = \sum_{n=2}^{n=N-1} (2I_n - I_{n-1} - I_{n+1})^2$$

- 5) Calculate the average and normalized Sum_zig_zag then get Zig Zag Index as following:

$$Zig_Zag_Index = \frac{\sum_{n=2}^{n=N-1} (2I_n - I_{n-1} - I_{n+1})^2}{N * EPI^2}$$

Based on the real data's testing, the proposed Zig_Zag_Index can evaluate the zig zag degree of a chromatograph peak shape, and the lower the Zig_Zag_Index, the better the peak quality.

Parameter 17-20 is about peak branch pattern detection in CWT domain. In MET-COFEI, 1D mass trace (EIC) is firstly transformed into 2D CWT coefficients. Then local maximum detection is utilized for each scale. Several continuous meaningful local maximum points across the 2D scan-scale space are defined as a meaningful peak pattern branch; in general, a meaningful branch should be composed of the local maximum points across several continuous scales, and corresponds to one valid peak of the original EIC. A meaningful peak branch pattern should be larger than a specific length, and its searching span should be limited to a specific value, and all of its Branch pattern searching gap should be smaller than a specific value.

17. Min Branch Length: The minimum value for a meaningful branch. The branch with its final length is smaller than Min Branch Length will not be considered a valid peak branch pattern.
18. Min Search Span: The minimum search span for search another local maximum point in its neighboring scale.
19. Branch Gap Allow: The maximum gap across several continuous neighboring scales. Only the branches with its maximum gap is smaller than Branch Gap Allow are considered as a meaningful branch, and finally a meaningful profile peak.
20. Marker Branch Dif Th: For all data points of a branch, the point with its coefficient value larger than its neighboring scales is defined as a marker points. Usually, for the good quality peak shape, there is only one marker point for a branch. But at the case of peak overlapping or low peak shape quality, there maybe exists several Marker points for one branch. If the distances of two marker points across scans are larger than Marker Branch Dif Th, the branch should be split into two meaningful branches, and finally two peaks should be identified.

A: The original EIC signal. B: CWT coefficients at different scales. C: local maximum detection for each scale and 3 meaningful branches can be recognized, for the branch 2,

there exist two marker points and also the distance of the two marker points are larger than Marker Branch Dif Th. So, there are 4 meaningful peaks are detected. D: The peak's parameters such as peak's apex, left/right boundary, etc. are retrieved according to the detected marker points.

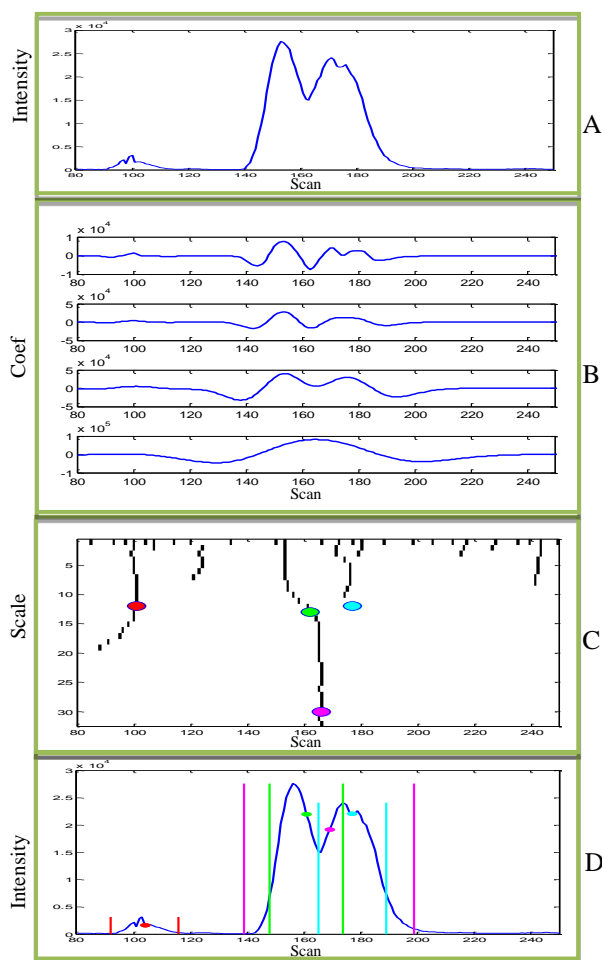


Fig. 22 CWT-based peak detection

21. Group Scan Shift Tol: The peaks with its peak apexs fall in the ranges of Group Scan Shift Tol can be considered a peak group.
22. Group Shape Angle Tol: The peaks with their peak shape similarities (defined by dot product and then \cos^{-1}) are smaller than Group Shape Angle Tol can be considered as a meaningful peak group. Here, HCA (Hierarchical Cluster Analysis) are adopted.
23. Fragment Mass Tol: When calculate the similarity score between the constructed spectrum and library spectrum, MET-COFEI carefully considers the cases of the integer and float type of the spectrums, so the fragment's m/z difference is allowed. One m/z pair(one from the constructed spectrum, the other from library spectrum) can be considered as a matched m/z pair, if their m/z tolerance smaller than Fragment Mass Tol. Only the matched m/z pair will contribute to the final spectrum similarity score.

24. Library Spectrum Score Th: Only the similarity score between the constructed spectrum and library spectrum is higher than Library Spectrum Score Th, it will be considered as a matched identification.
25. None/One/Two Phase alignment: If The user need alignment to align the compound associated peaklist across samples, please choose one or two phase alignment strategy. One phase alignment means a loosen alignment but a fast alignment, while two phase align means a stringent alignment but a time-consuming alignment.
26. Align Window1: The align window for the first phase alignment across retention time.
27. Align Window2: The align window for the second phase alignment across retention time. The second alignment means a more accurate align. So, AlignWindow2 < AlignWindow1.
28. Align Fragment Mass Tol: When calculate the similarity score between the two constructed spectra from two samples, MET-COFEI carefully considers the observed m/z difference between the two spectrums, so the fragment's m/z difference is allowed. One m/z pair(one m/z value from one constructed spectrum, the other m/z value from another spectrum) can be considered as a matched m/z pair, if their m/z tolerance smaller than Align Fragment Mass Tol. Only the matched m/z pair will contribute to the final spectrum similarity score
29. Align Spectrum Score Th: Only the similarity score between the two constructed spectra is higher than Align Spectrum Score Th, it will be considered as a matched alignment. The two components represented by the two constructed spectrum can be considered the same metabolite compound.

Parameter Setting and Optimization Procedure

Given one GC-MS Data sample, parameter configuration will affect the processing speed and performance greatly. Usually, the loosen parameter can get more peak information but with the high computation burden and longer computation time. The user needs to balance it in real application. If there are not too much dataset, the user always wants to acquire the optimal analysis results. Here, I provide a normal procedure for parameter setting and optimization.

1. Open a representative CDF file (refer to Data Process property page). Determine the Start Scan and End Scan from the TIC curve.
2. Move mouse to a representative scan, from the corresponding spectrum, determine the intensity cutoff threshold. Bear in mind, only the higher intensity point of a scan will be used for EIC extraction and peak detection.
3. Still from the opened representative spectrum, zoom out, you can easily know the data is centroid data or profile data. In profile data, the spectrum displayed as many spectrum peaks while centroid data displayed as sticks.
4. Parameter configuring for EIC extraction.

Bear in mind, MET-COFEI adopt a more advanced method to extract EIC, mass trace (not binning)based method, in principle, it is same to object tracing problem in video tracing. So, a meaning mass trace (EIC) should be a continuous point trace across several continuous scans and the m/z value varies (shift) in the specific tolerance.

4-a, from the mass spectrometry instrument facility personnel, you should know the mass spectrometry accuracy/tolerance, and then you can configure Trace PPM.

4-b, also, you can know the minimum meaningful chromatograph peak width from the mass spectrometry instrument facility personnel, and then you can configure the Min Trace Length.

4-c, Allow some point missing during EIC tracing, you should configure Miss Allow Th (default value=3.)

4-d, Allow the tolerance between two neighboring mass trace, you should configure Inter Trace Distance Th.

You can refer to fig.20 to know the real physical meaning at first and then configure the parameters for mass trace extraction.

5. Parameter configuring for Peak detection

Bear in mind, MET-COFEI adopt a more advanced method to detect meaningful profile peak. The peak detection is implemented in 2D CWT domain spanned by retention time-scan by detect the peak branch pattern. One meaningful peak corresponds to a meaningful branch pattern. A meaningful branch pattern should be across several scales and the continuity is not too bad.

5-a, Min Branch Length: the minimum branch length in 2D CWT domain for a peak. Default value =6;

5-b, Branch Gap Allow: allow the branch have a limited gap across neighboring scales. Default value =2-3

5-c, Min Search Span: specific the branch searching range in the neighboring scales. Default value =2-3.

5-d, Marker Branch Dif Th: for sharing peak issues, the branch pattern is a little challenge; you can consider it a very wide peak, or two sharing peaks. So, one peak branch pattern in the high scale will split into two at some scale. In this case, there will two marker points. When the distance between the two marker points is larger than Marker Branch Dif Th, MET-COFEI will consider it as two sharing peaks.

You can refer to fig.22 to know the real meaning of this method and then configure the parameters for peak detection.

6. Parameter configuring for peak quality filtering

Bear in mind, MET-COFEI is aiming to output the high-quality chromatograph peak by some criteria.

6-a: Min Peak Width (default =6), Max Peak Width (default =50), Peak intensity Th (default =50), can be known from mass spectrometry instrument facility personnel, or,

you can tentative configure this parameter as default values and run MET-COFEI, if found some peaks missing, try to configure a loosen parameters to find more peaks.

6-b: SNR Th (default =2.0), Peak Significance Th (default =1.0), TPASRTh (default =0.7), Zig Zag Index (default =0.6)

You can tentative configure this parameter as default values and run MET-COFEI, if found some peaks missing, try to configure a loosen parameters to find more peaks.

7. Parameter configuring for peak grouping, identification and alignment

7-a: Group Scan Shift_Tol (default=3), Group Shape Angle Tol (default=20), Fragment Mass Tol (default=0.1), Library Spectrum Score Th(default=750). The calculation method for spectrum similarity score is same to the NIST method.(The maximum score =1000)

7-b: One/Two Phase alignment, Align Window1, Align Window2. Usually, use One Phase alignment, if you find some compounds are misalign, you can use two phase alignment strategy to refine the alignment results, keep in mind, the AlignWindo2 for Two Phase alignment should be narrower than AlignWindow1 for One Phase alignment. Regarding the configured value for Align Window1/ Align Window2, you should know these parameters from the mass spectrometry instrument facility personnel, because they are related to the real dynamic range for retention time shift.

Additionally, Start Scan, End Scan, and Cutoff Intensity are mainly factor to affect the computation burden.

Potential problems in application

MET-COFEI development team tried to find any potential bugs or problems in application, if you find any problems, please contact us by pzhao@noble.org or wezhang@noble.org. We are very appreciated for your feedback and suggestion.

During our test, the potential problems that have been found include:

1. MET-COFEI may fail in parallel processing mode if the specific file folder name and file name have any space. This problem is generated by the MPI command line parsing module.
Solutions: remove any space in the folder name and file name.
2. In some Windows 7 machine, MPI based parallel processing may fail,
Solutions: Check Windows OS and update into the latest version, ensure it support MPI, .NET.