Supporting Information

Mass Spectral Feature List Optimizer (MS-FLO): a tool to minimize false positive peak reports in untargeted LC-MS data processing

Brian C. DeFelice[1], Sajjan Singh Mehta[1], Stephanie Samra[1], Tomáš Čajka[1], Benjamin Wancewicz[1], Johannes F. Fahrmann[1, 2], Oliver Fiehn1, 3

[1] *University of California, Davis, West Coast Metabolomics Center, 451 E. Health Sciences Drive Rm 1300, Davis, California, 95616, United States*

[2]*Department of Clinical Cancer Prevention, University of Texas MD Anderson Cancer Center, 6767 Bertner Avenue Houston, TX 77030-2603*

[3] *Department of Biochemistry, Faculty of Sciences, King Abdulaziz University, Abdullah Sulayman, Jeddah 21589, Saudi Arabia*

Oliver Fiehn, E-mail: ofiehn@ucdavis.edu

Contents

S1

In order to tailor the data processing parameters to specific sample matrices and the respective LC-MS methods used for analysis small modifications were made to the following general data processing parameters. Studies processed with MZmine2 used the following parameters or slightly modified variations of the following parameters: 1) Files were centroided and peak heights below 100 intensity counts were considered noise and removed when converting to the .mzXML data format. 2) Mass detection threshold was set to a peak height of 1000. 3) Chromatogram builder m/z tolerance, 0.005 Dal; RT tolerance, 0-60s. 4) Deconvolution algorithm used was 'local minimum search' with the following settings: chromatographic threshold, 40%; minimum retention time range, 1.0 min; absolute peak height, 1000; minimum ratio of top or edge, 1.8; peak duration, 0-1.5 min. 5) Samples were ordered as such: study samples, quality control samples, and blanks and aligned using the join aligner. Join aligner settings: m/z tolerance, 0.005 or 50ppm; weight for m/z, 30; RT tolerance, 0.2 min; weight for RT, 35. 6) Identification was performed by accurate mass-RT database matching, from an in-house database created from MS/MS annotation and injection of authentic standards.

MS-DIAL studies were processed using the following parameters or slight modifications of those parameters depending on LC-MS and matrix conditions: 1) Data files were converted to the .abf format. 2) Smoothing method, linear weighted moving average smoothing; smoothing level, 1 scan; minimum peak width, 5 scans; minimum peak height, 3000 amplitude; mass slice width, 0.1 Da. 3) Deconvolution parameters: band width, 5 scans; segment number, 1; peak consideration, both; sigma window value, 0.001 min. 4) Identification was performed by accurate mass-RT database matching, and adducts were selected based on LC-MS system used. 5) Alignment parameter tab settings: RT tolerance, 0.1 min; MS1 tolerance, 0.025 Da; RT factor, 0.5; MS1 factor, 0.5; peak count filter, 30%.

S2

MZmine .csv Export Format. Data processed with MZmine should be exported using a comma as a field separator. All common elements (row ID, row m/z, row retention time, row comment and row number of detected peaks) and identity elements should be checked and exported. Additionally, 'Export peak height' should be checked. .csv files exported in this fashion can be directly submitted to MS-FLO

Submitting Non-MZmine/MS-DIAL/XCMS Datasets to MS-FLO. There are three column headers required for any .csv file to be recognized by MS-FLO's MZmine configuration: "row *m/z*", "row retention time" and "peak height". The columns titled "row *m/z*" and "row retention time" must contain mass-to-charge ratio and retention time information respectively. All columns containing sample height information must contain the phrase "peak height". For example, a column containing peak height information for 'Sample 1' should be renamed, 'Sample 1 peak height'. All of the above mentioned .csv file modifications can be performed in any cell based program such as MS Excel or Open Office.

Table S1

study description	dataset number	LC method	data processing program
Allantoin differences in Synechococcus cells grown in high versus low lightgrowth	1	Lipids (+)	MZmine 2
Analysis of various points along the canine gastrointestinal tract	2	HILIC (+)	MZmine 2
Changes in the metabalome and lipidome in response to exercise training	3	Lipids (+)	MS-DIAL
Changes in the metabalome and lipidome in response to exercise training	4	Lipids (-)	MS-DIAL
Crude Algae Oil	5	Lipids (+)	MZmine 2
Effects of dietary supplement on hamster metabolism	6	Lipids (+)	MS-DIAL
Effects of dietary supplement on hamster metabolism	7	Lipids (-)	MS-DIAL
Effects of the probiotic "LGG" on gut metabolism of alcoholics	8	Lipids (-)	MZmine 2
Effects of the probiotic "LGG" on gut metabolism of alcoholics	9	Lipids (+)	MZmine 2
Identification and validation of interstitial cystitis/painful bladder syndrome metabolites	10	Reverse Ph	MZmine 2
Metabolite changes associated with methionine stress sensitivity of cancer	11	Lipids (-)	MZmine 2
Metabolite changes associated with methionine stress sensitivity of cancer	12	Lipids (+)	MZmine 2
Metabolite comparison of mouse gastric tissue and glands	13	Lipids (-)	MS-DIAL
Metabolite comparison of mouse gastric tissue and glands	14	Lipids (+)	MS-DIAL
Metabolites detected from human bronchoalveolar lavage of varying asthma severities	15	Lipids (-)	MS-DIAL
Metabolites detected from human bronchoalveolar lavage of varying asthma severities	16	Lipids (+)	MS-DIAL
Metformin effects on liver and kidney tissue	17	HILIC (+)	MZmine 2
NOD diabetic mice progressors vs nonprogressors	18	Lipids (-)	MZmine 2
NOD diabetic mice progressors vs nonprogressors	19	Lipids (+)	MZmine 2
Progesterone level effects on primary metabolites in uterus, blood, and ovaries (Follicle)	20	Lipids (+)	MZmine 2
Progesterone level effects on primary metabolites in uterus, blood, and ovaries (Plasma)	21	Lipids (+)	MZmine 2
Renal metabolic pathways indicating ischemic or inflammatory changes	22	Lipids (-)	MS-DIAL
Renal metabolic pathways indicating ischemic or inflammatory changes	23	Lipids (+)	MS-DIAL
Role of medium in bacterial growth	24	Lipids (+)	MZmine 2
Role of medium in bacterial growth	25	Reverse Ph	MZmine 2
Role of medium in bacterial growth	26	HILIC (+)	MZmine 2
Single treatment gene impact on Arabidopsis metabolites	27	Lipids (-)	MZmine 2
Single treatment gene impact on Arabidopsis metabolites	28	Lipids (+)	MZmine 2
Average			
Standard Deviation			

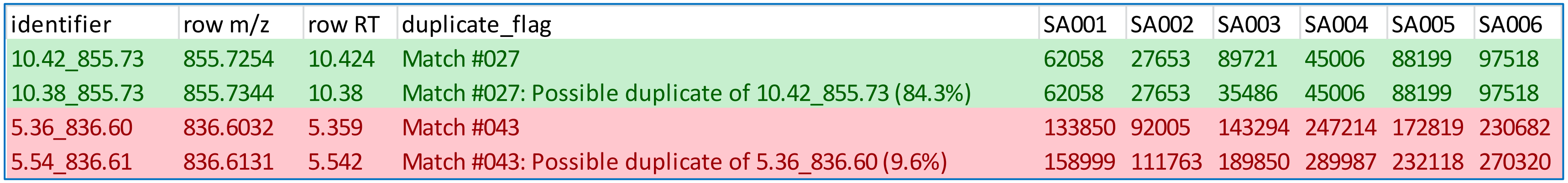
Table S1. Details about studies, samples mass spectrometry parameters, LC parameters, and metadata.

matrix	workbench ID	sample count	initial row count	final row count	rows removed	duplicates flagged	duplicates removed	adducts pairs flagged
synechococcus cells	ST000318	8	1263	1145	118	1	1	138
canine colon, duodenum, ileum, rectum	ST000327	24	2282	2260	22	78	0	6
human blood plasma	ST000387	302	841	720	121	17	2	21
human blood plasma	ST000387	302	1266	1170	96	119	42	36
algae	ST000319	3	2806	2509	297	21	11	79
spent media from DG44 chinese hamster ovary cells	ST000344	18	354	306	48	4	0	20
spent media from DG44 chinese hamster ovary cells	ST000344	17	459	425	34	21	0	10
mouse stool	ST000321	36	1471	1335	136	5	4	47
mouse stool	ST000321	36	1441	1364	77	5	4	8
human urine	ST000382	100	1364	1284	80	19	34	63
human cancer cells	ST000077	35	1068	922	146	6	0	35
human cancer cells	ST000077	35	756	728	28	2	0	6
mouse stomach tissues	ST000354	22	2171	1896	275	34	0	175
mouse stomach tissues	ST000354	26	1595	1505	90	31	1	25
human bronchoalveolar lavage fluid	ST000346	20	412	367	45	8	1	22
human bronchoalveolar lavage fluid	ST000346	17	181	173	8	4	0	2
Mouse Liver and Kidney	ST000340	23	744	725	19	1	0	24
blood plasma	ST000075	85	663	609	54	6	1	23
blood plasma	ST000075	85	128	123	5	1	0	0
cow preovulatory follicle fluid	ST000324	12	1155	1091	64	9	8	24
cow blood plasma	ST000322	88	1125	1070	55	8	13	9
human renal tissue	ST000342	37	939	824	115	17	0	57
human renal tissue	ST000342	37	356	333	23	12	0	1
bacterial culture medium	ST000317	21	695	639	56	0	0	29
bacterial culture medium		22	1050	973	77	4	5	26
bacterial culture medium	ST000326	24	546	501	45	7	3	13
genetically modified arabidopsis	ST000320	22	498	463	35	1	2	6
genetically modified arabidopsis	ST000320	24	246	245	1	1	0	0
		52.9	995.54	918.0	77.5	15.8	4.7	32.3
		74.6	657.0	605.2	71	25.7	10.1	40.5

adduct rows removed by joining	possible isotope sets	possible isotope w/ mean R^2 > 0.8	total rows auto-removed	contaminant ions removed
117	41	33	118	0
0	7	2	22	22
119	33	29	121	0
54	161	127	96	0
285	76	76	297	1
48	9	7	48	0
34	3	3	34	0
132	73	73	136	0
73	26	26	77	0
45	72	57	80	1
146	17	17	146	0
28	11	10	28	0
275	158	108	275	0
89	38	27	90	0
44	12	10	45	0
8	2	1	8	0
19	37	8	19	0
53	25	25	54	0
5	1	1	5	0
56	71	70	64	0
42	66	73	55	0
115	31	24	115	0
23	1	1	23	0
50	2	1	56	6
70	14	12	77	2
29	19	16	45	13
33	28	28	35	0
1	1	1	1	0
71.2	37.0	30.9	77.5	1.6
71.3	42.3	34.4	71.2	4.8

Figure S1

A)



identifier	row m/z	row RT	duplicate_flag	SA001	SA002	SA003	SA004	SA005	SA006
10.42_855.73	855.7254	10.424	Match #027	62058	27653	89721	45006	88199	97518
10.38_855.73	855.7344	10.38	Match #027: Possible duplicate of 10.42_855.73 (84.3%)	62058	27653	35486	45006	88199	97518
5.36_836.60	836.6032	5.359	Match #043	133850	92005	143294	247214	172819	230682
5.54_836.61	836.6131	5.542	Match #043: Possible duplicate of 5.36_836.60 (9.6%)	158999	111763	189850	289987	232118	270320

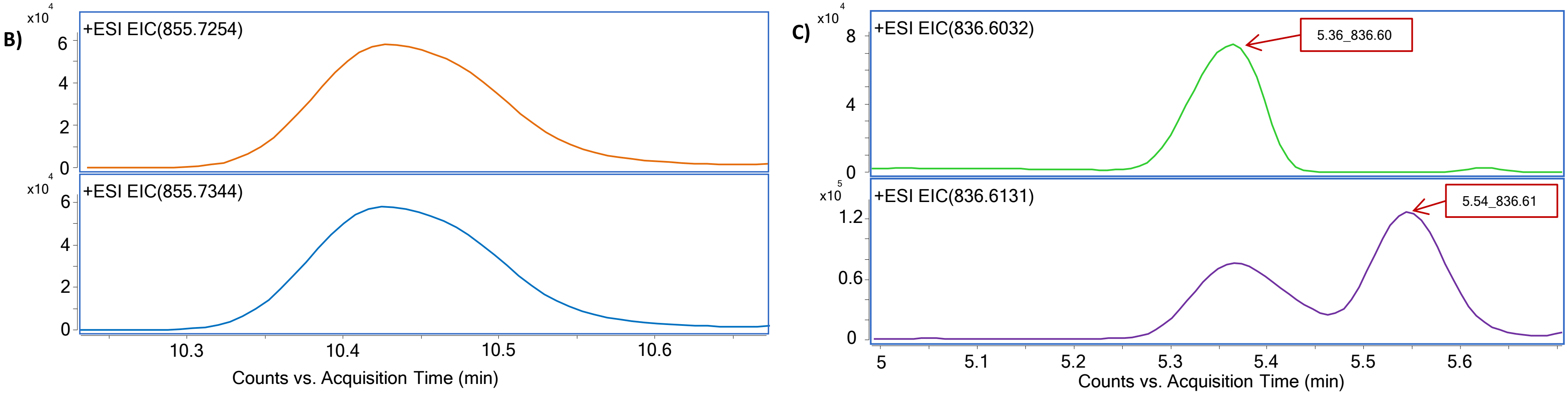
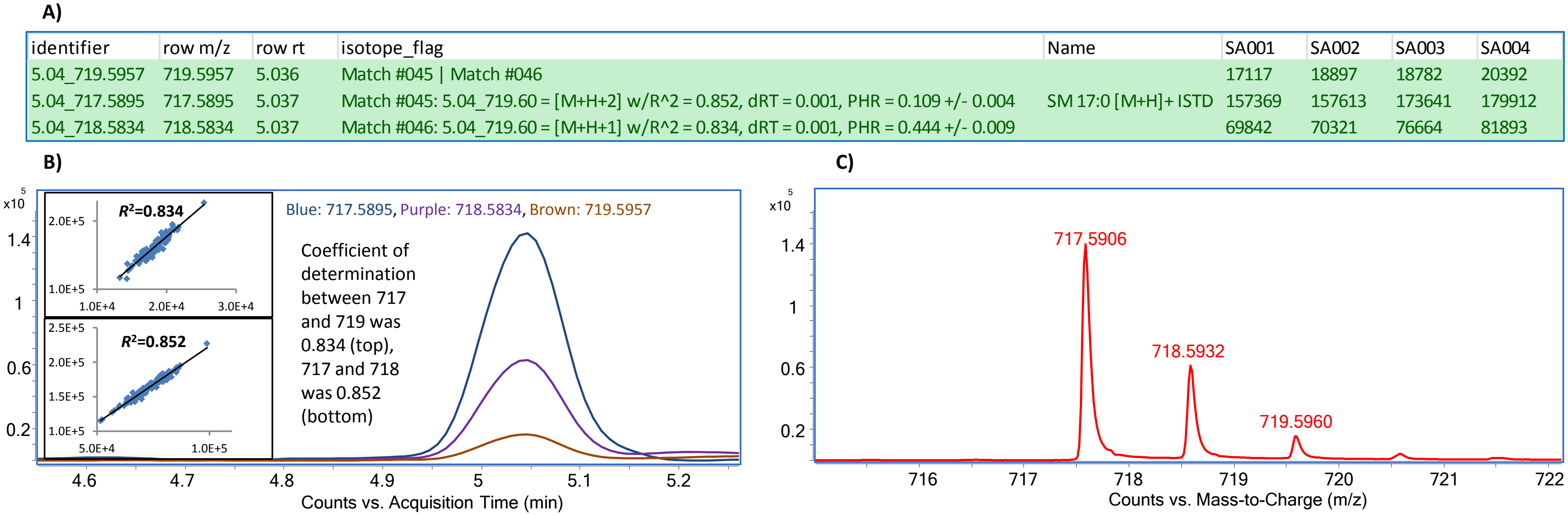


Figure S1. Examples of potential duplicate peaks detected by MS-FLO. **A)** Section of a .csv file exported from MS-FLO, showing flagged potential duplicate features, green features have been determined to be duplicates, red features have been determined to be unique features. **B)** Extracted ion chromatograms (*10mDa window*) representing two duplicate features that were misaligned, as separate features, in data processing; blue: 855.7344 and gold: 855.7254. **C)** Features 5.54_836.61 and 5.36_836.60 were flagged as potential duplicates due to their close RT and *m/z* proximity however once visually inspected it is clear they are separate features.

Figure S2

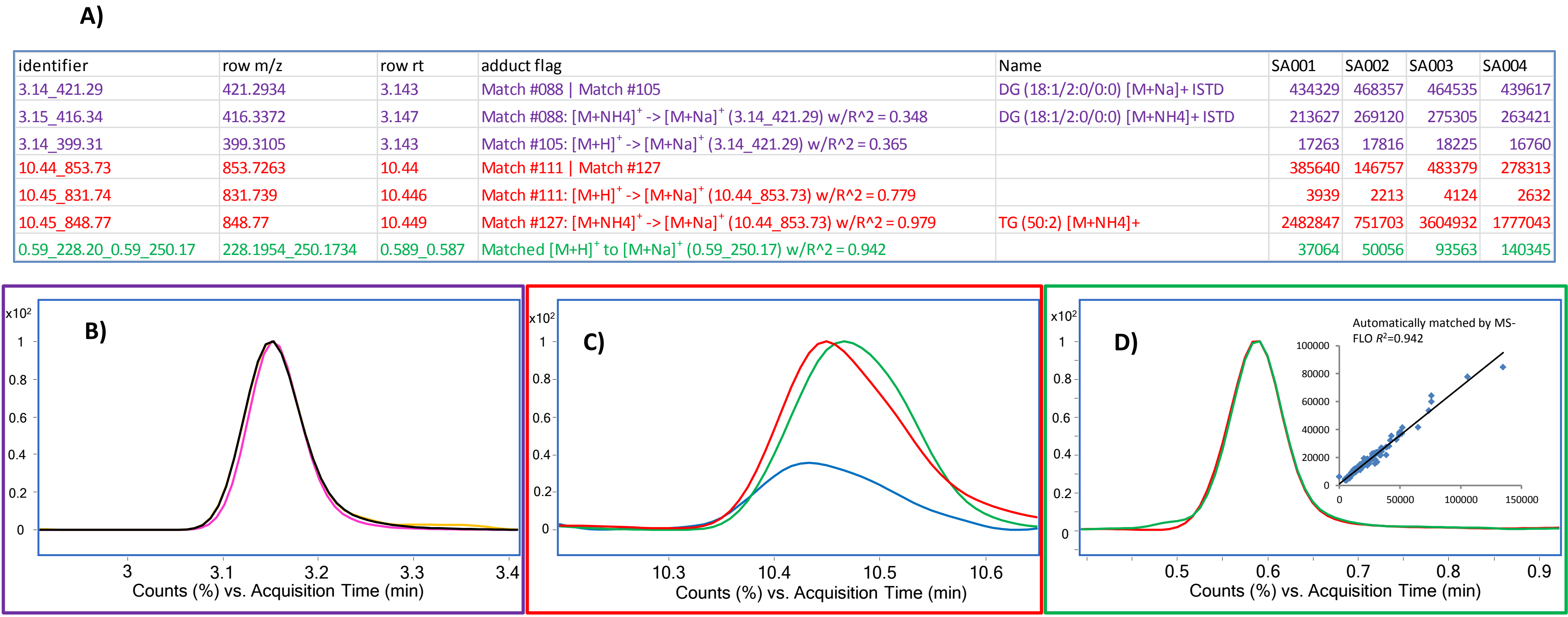


A)

identifier	row m/z	row rt	isotope_flag	Name	SA001	SA002	SA003	SA004
5.04_719.5957	719.5957	5.036	Match #045 | Match #046		17117	18897	18782	20392
5.04_717.5895	717.5895	5.037	Match #045: 5.04_719.60 = [M+H+2] w/R^2 = 0.852, dRT = 0.001, PHR = 0.109 +/- 0.004	SM 17:0 [M+H]+ ISTD	157369	157613	173641	179912
5.04_718.5834	718.5834	5.037	Match #046: 5.04_719.60 = [M+H+1] w/R^2 = 0.834, dRT = 0.001, PHR = 0.444 +/- 0.009		69842	70321	76664	81893

Figure S2. Example of potential isotopic pairs detected by MS-FLO **A)** Portion of a .csv file exported from MS-FLO, showing isotopic relationships of 3 features, including close RT and common isotopic peak height ratios (PHR). **B)** Extracted ion chromatograms for the three flagged masses Blue: 717.5895, Purple: 718.5834, Brown: 719.5957 (right). Graphical representation of the coefficient of determination between ions 717 and 719, R^2= 0.834 (left, top), 717 and 718 R^2=0.852 (left, bottom). **C)** Mass spectral data extracted from the apex of the three potentially isotopic peaks. Features identified as 5.04_718.58 and 5.04_719.60 were determined to be isotopic peaks and removed from the final feature list.

Figure S3



A)

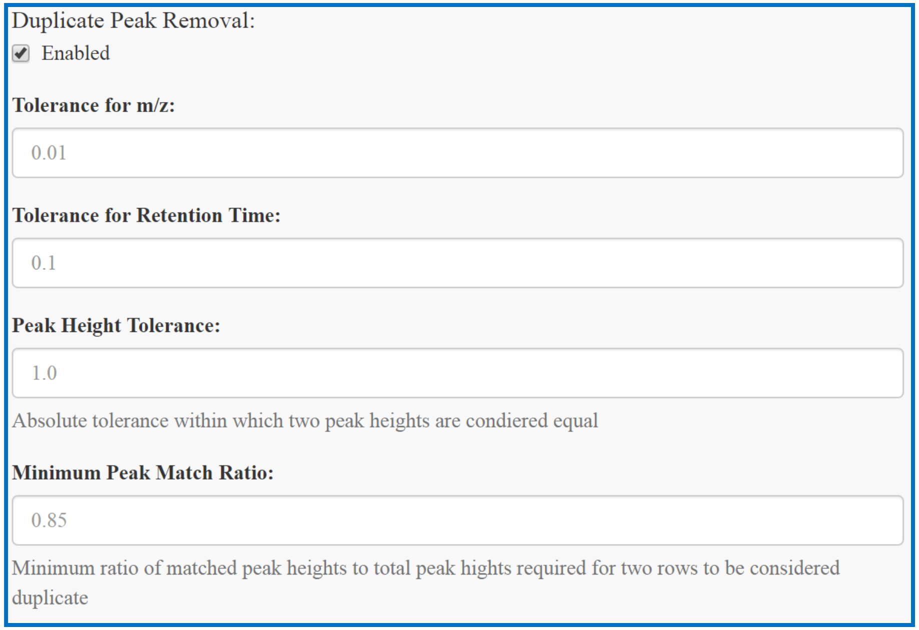
identifier	row m/z	row rt	adduct flag	Name	SA001	SA002	SA003	SA004
3.14_421.29	421.2934	3.143	Match #088 | Match #105	DG (18:1/2:0/0:0) [M+Na]+ ISTD	434329	468357	464535	439617
3.15_416.34	416.3372	3.147	Match #088: [M+NH4]$^+$ -> [M+Na]$^+$ (3.14_421.29) w/R^2 = 0.348	DG (18:1/2:0/0:0) [M+NH4]+ ISTD	213627	269120	275305	263421
3.14_399.31	399.3105	3.143	Match #105: [M+H]$^+$ -> [M+Na]$^+$ (3.14_421.29) w/R^2 = 0.365		17263	17816	18225	16760
10.44_853.73	853.7263	10.44	Match #111 | Match #127		385640	146757	483379	278313
10.45_831.74	831.739	10.446	Match #111: [M+H]$^+$ -> [M+Na]$^+$ (10.44_853.73) w/R^2 = 0.779		3939	2213	4124	2632
10.45_848.77	848.77	10.449	Match #127: [M+NH4]$^+$ -> [M+Na]$^+$ (10.44_853.73) w/R^2 = 0.979	TG (50:2) [M+NH4]+	2482847	751703	3604932	1777043
0.59_228.20_0.59_250.17	228.1954_250.1734	0.589_0.587	Matched [M+H]$^+$ to [M+Na]$^+$ (0.59_250.17) w/R^2 = 0.942		37064	50056	93563	140345

B)
x10²
1
0.8
0.6
0.4
0.2
0
3
3.1
3.2
3.3
3.4
Counts (%) vs. Acquisition Time (min)
C)
x10²
1
0.8
0.6
0.4
0.2
0
10.3
10.4
10.5
10.6
Counts (%) vs. Acquisition Time (min)
D)
x10²
1
0.8
0.6
0.4
0.2
0
0.5
0.6
0.7
0.8
0.9
Counts (%) vs. Acquisition Time (min)
Automatically matched by MS-FLO R²=0.942
100000
80000
60000
40000
20000
0
0
50000
100000
150000

Figure S3. Examples of Adduct Joining/Flagging **A)** Excerpt of .csv file exported from MS-FLO showing adducts/molecular ions automatically matched (green text) and potential adducts flagged for manual review **B)** Extracted ion chromatograms (EICs) of multiple features Orange: 399.3105, Pink: 416.3372, Black: 421.2934. Despite the low R^2 value (~0.35) these peaks are all adducts from the same molecule. **C)** EICs of 3 features that have been flagged as potential adducts; Blue: 831.7390, Green: 848.7700, Red: 853.7263 These flagged features have a mass error of 10 mDa, and when the EICs are overlaid it is clear there is a RT shift. It is doubtful that these ions are all generated from the same molecule. **D)** EICs of two ions (Red: 228.1953 Green: 250.1774) that were automatically joined by MS-FLO based on the user settings reveal the two features overlay. Additionally, the accurate mass difference between the theoretical (21.9787 Da) and experimental (21.9821 Da) masses corresponding to a [M+H]$^+$ → [M+Na]$^+$ ion transition is small (3.4 mDa) and the R^2 value supports a very strong correlation.

Figure S4

A)



Duplicate Peak Removal:

☑ Enabled

Tolerance for m/z:

0.01

Tolerance for Retention Time:

0.1

Peak Height Tolerance:

1.0

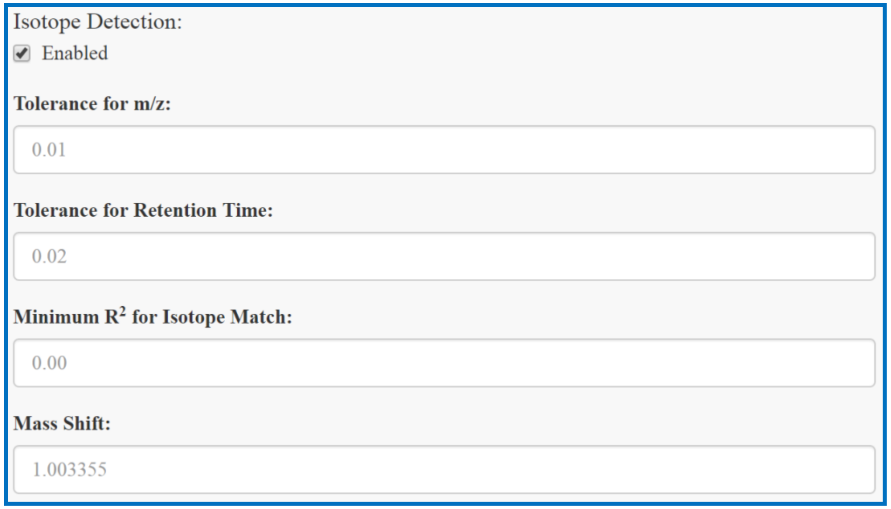
Absolute tolerance within which two peak heights are condiered equal

Minimum Peak Match Ratio:

0.85

Minimum ratio of matched peak heights to total peak hights required for two rows to be considered duplicate

B)



Isotope Detection:

☑ Enabled

Tolerance for m/z:

0.01

Tolerance for Retention Time:

0.02

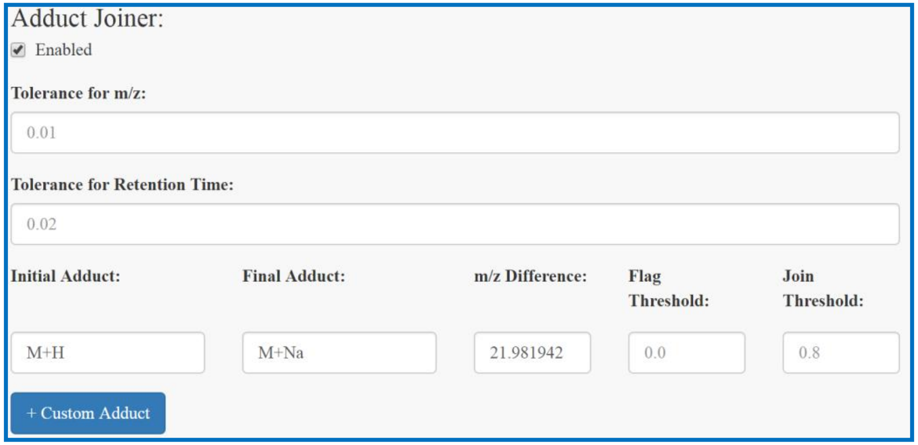
Minimum R^2 for Isotope Match:

0.00

Mass Shift:

1.003355

C)



Adduct Joiner:

☑ Enabled

Tolerance for m/z:

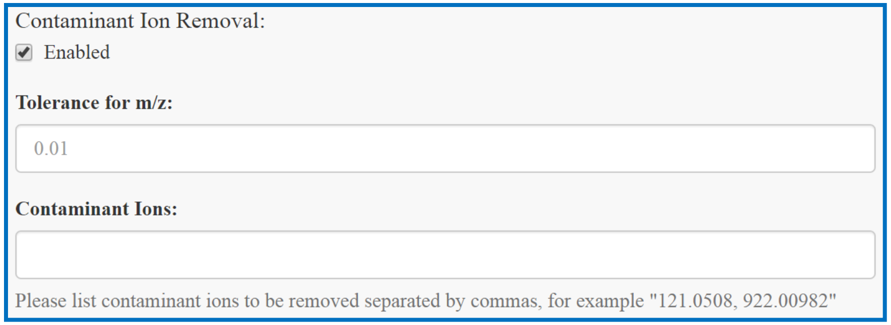
0.01

Tolerance for Retention Time:

0.02

Initial Adduct:	**Final Adduct:**	**m/z Difference:**	**Flag Threshold:**	**Join Threshold:**
M+H	M+Na	21.981942	0.0	0.8

D)



Contaminant Ion Removal:

☑ Enabled

Tolerance for m/z:

0.01

Contaminant Ions:

Please list contaminant ions to be removed separated by commas, for example "121.0508, 922.00982"

Figure S4. The user interface of MS-FLOs four primary modules and the recommended default settings. **A)** Duplicate peak removal **B)** Isotope detection **C)** Adduct joiner **D)** Contaminant ion removal

Figure S5

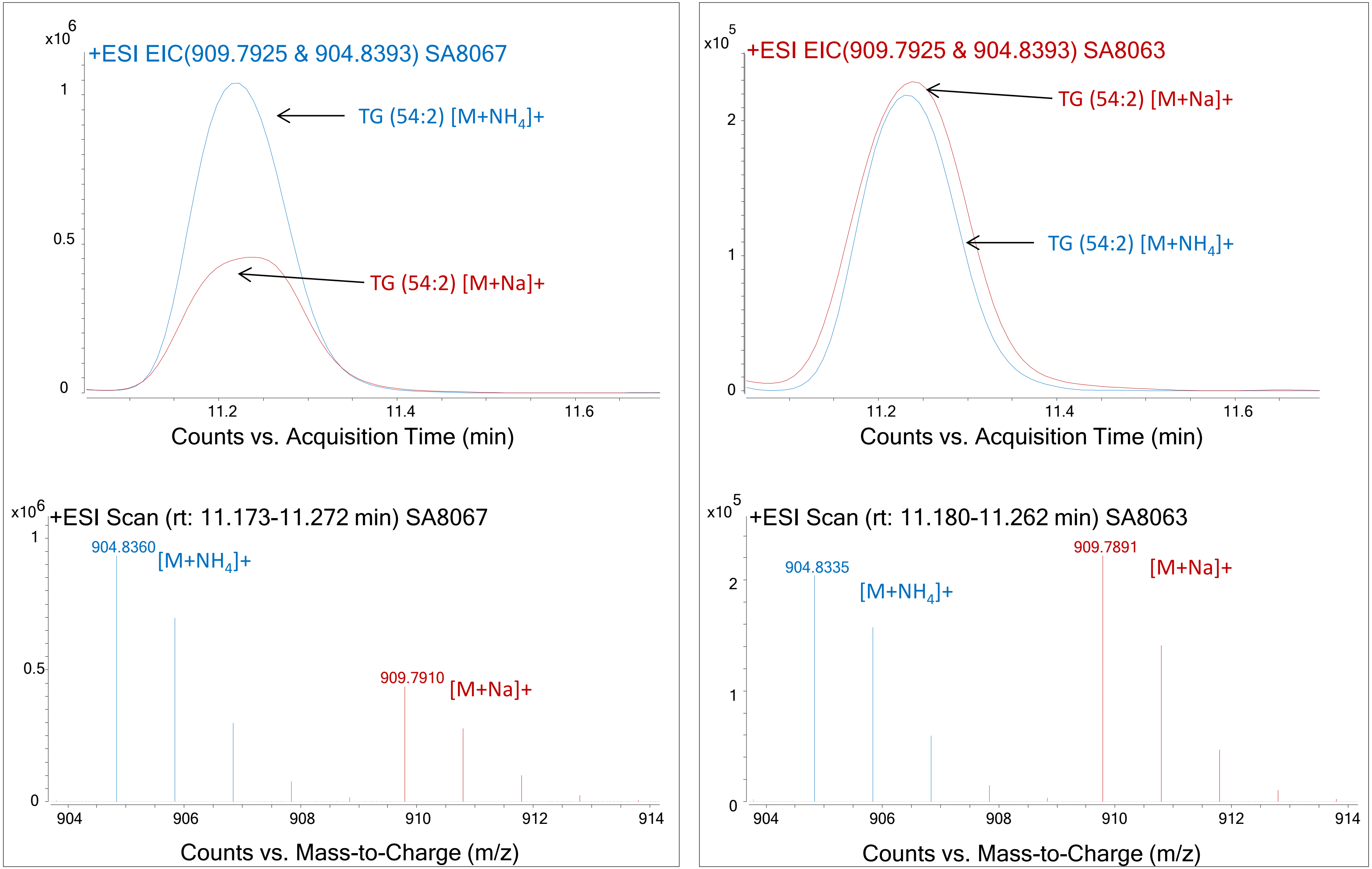


Figure S5. Variation in ratio of triacylglycerol ion adducts from sample to sample. Depending on metabolite abundance the peak height ratio of sodiated adduct to ammoniated adduct can vary greatly. Left: [M+NH4]+ to [M+Na]+ ratio is 1:0.44. Right: [M+NH4]+ to [M+Na]+ ratio is 0.97:1