

Introduction to Preprocessing of Untargeted Metabolomics Data

Xiuxia Du, Ph.D.

Department of Bioinformatics and Genomics
University of North Carolina at Charlotte

Outline

- Raw untargeted LC/MS and GC/MS metabolomics data
 - Profile and centroid data
 - Mass vs. retention time map
 - TIC
 - EIC
- Principles of LC/MS and GC/MS data preprocessing
- Feature identification
 - Identification of known compounds
 - Identification of unknown compounds

**Raw Untargeted
LC/MS and GC/MS
Metabolomics data**

List of mass spectra

list of scans in raw files

- MS scans in blue
- MS/MS scans in red
- # sequential number
- @ retention time
- MS level
- type of spectrum
 - p = profile
 - c = centroid
 - t = thresholded
- polarity of ionization
 - + = positive
 - - = negative
 - ? = unknown

Raw data files

PosMode_IR3.mzXML

- #1 @0.00 MS1 p +
- #2 @0.01 MS2 (61.0100) c +
- #3 @0.01 MS2 (81.5211) c +
- #4 @0.01 MS2 (92.0250) c +
- #5 @0.01 MS2 (105.9350) c +
- #6 @0.01 MS2 (125.9850) c +
- #7 @0.01 MS2 (132.0021) c +
- #8 @0.01 MS2 (141.9575) p +
- #9 @0.02 MS2 (146.9940) c +
- #10 @0.02 MS2 (158.9614) p +
- #11 @0.02 MS2 (162.9674) p +
- #12 @0.02 MS2 (182.9840) p +
- #13 @0.02 MS2 (188.0182) c +
- #14 @0.02 MS2 (202.9867) p +
- #15 @0.02 MS2 (236.9384) c +
- #16 @0.03 MS2 (371.0991) p +
- #17 @0.03 MS1 p +
- #18 @0.04 MS2 (61.0086) c +
- #19 @0.04 MS2 (105.9352) c +
- #20 @0.04 MS2 (112.9546) p +
- #21 @0.04 MS2 (125.9859) c +
- #22 @0.04 MS2 (132.0027) c +
- #23 @0.04 MS2 (141.9574) c +
- #24 @0.04 MS2 (146.9953) p +
- #25 @0.04 MS2 (153.0123) c +

Peak lists

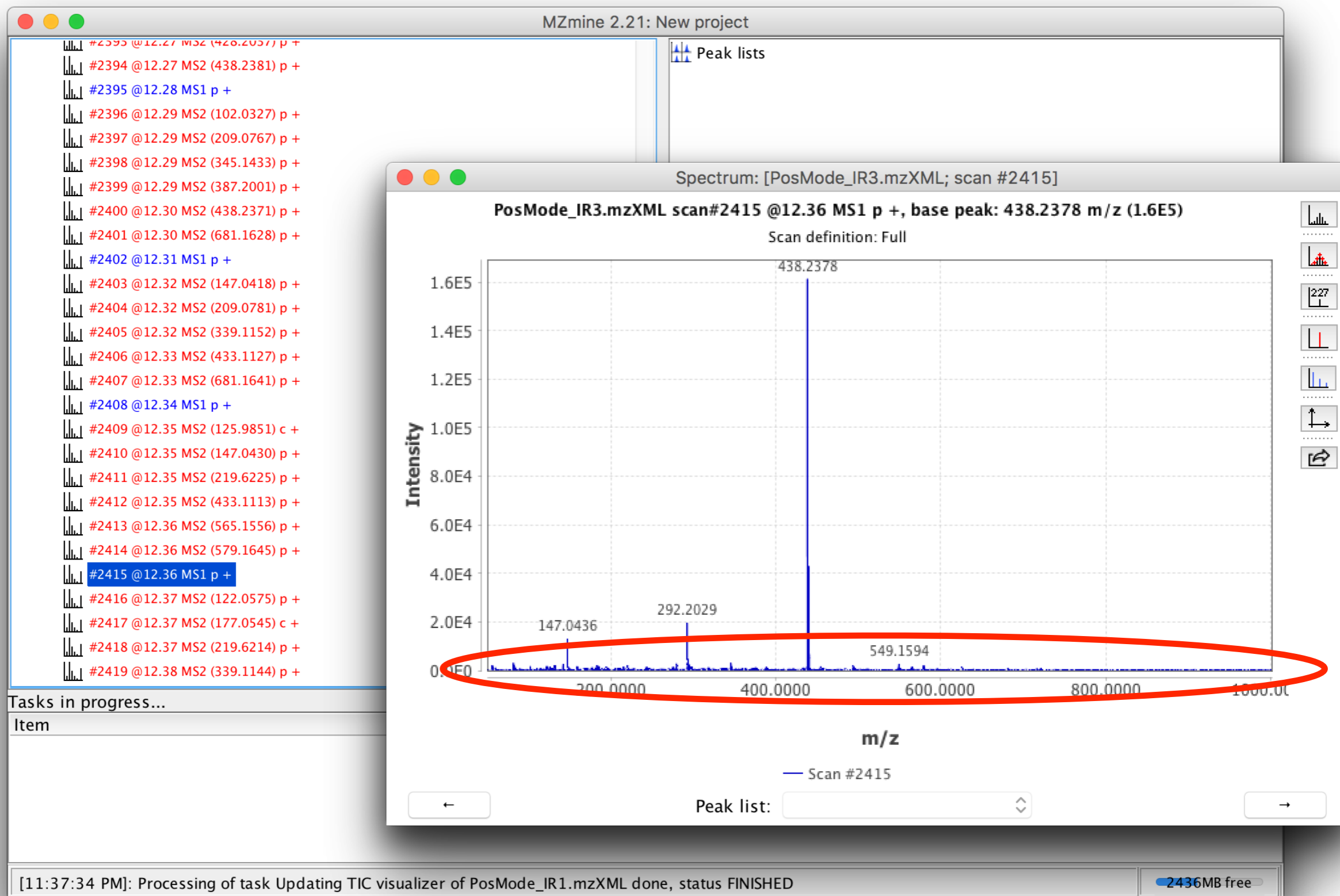
Tasks in progress...

Item	Priority	Status	% done
------	----------	--------	--------

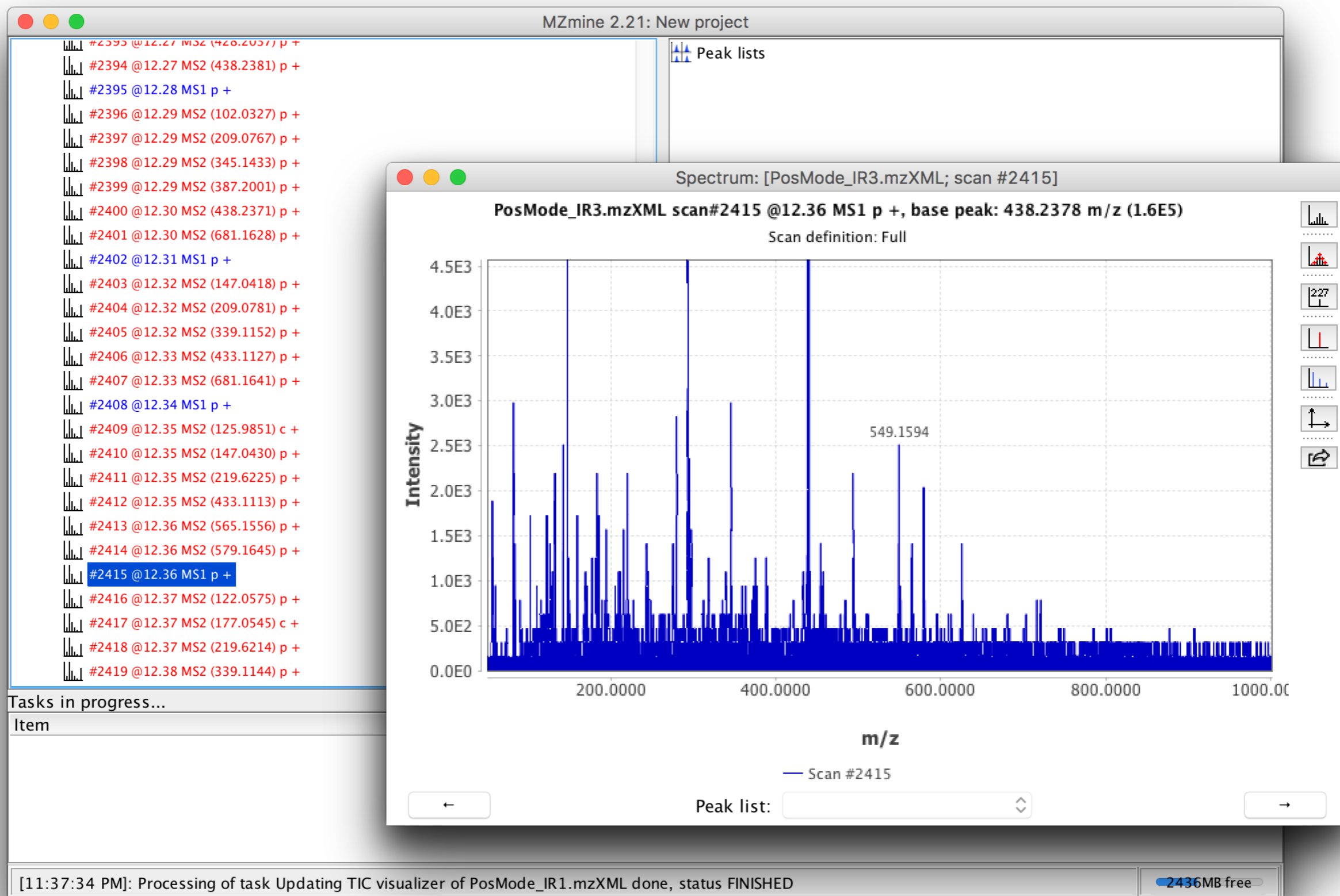
[11:37:34 PM]: Processing of task Updating TIC visualizer of PosMode_IR1.mzXML done, status FINISHED

3248MB free

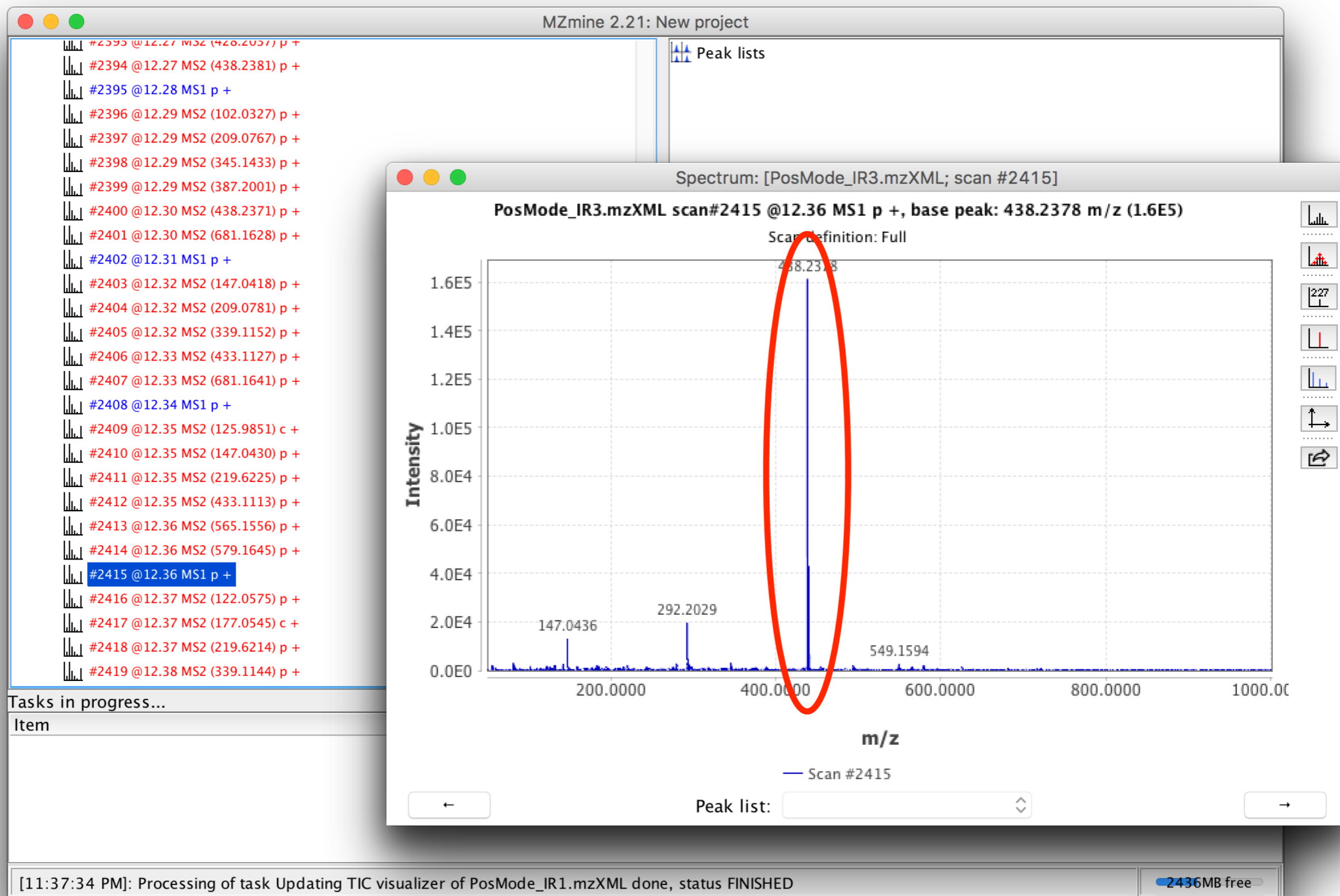
One mass spectrum



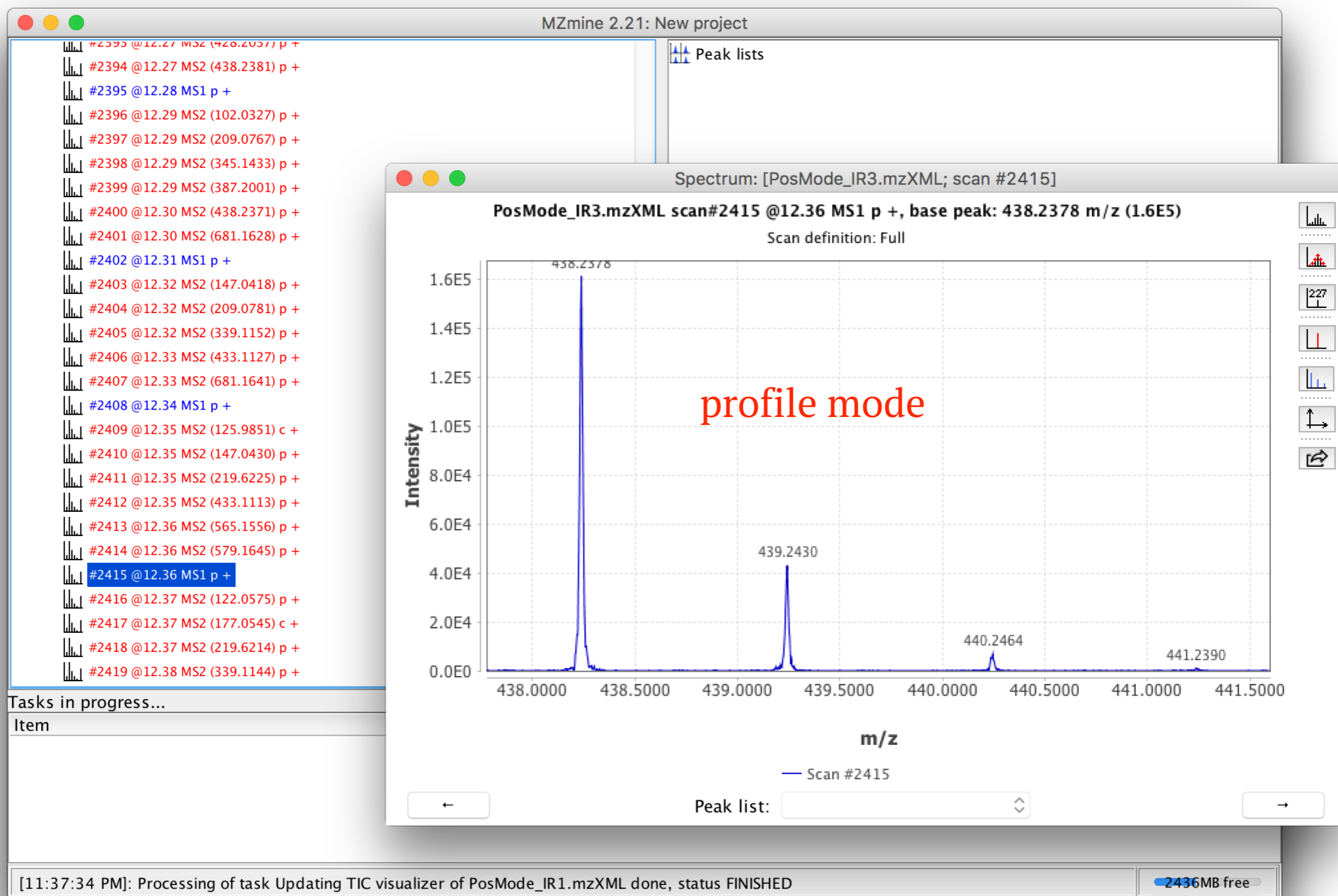
One mass spectrum



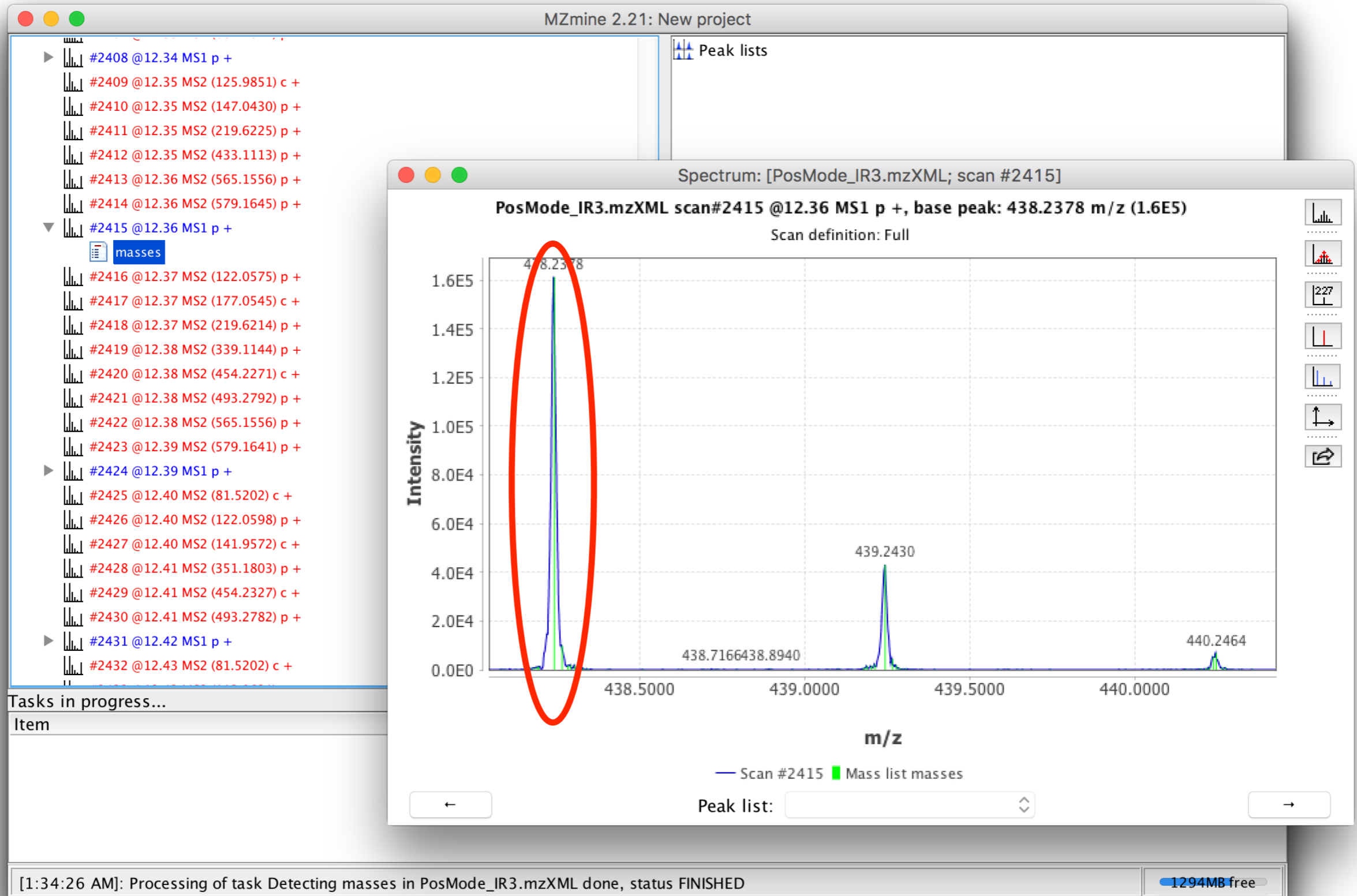
One mass spectrum



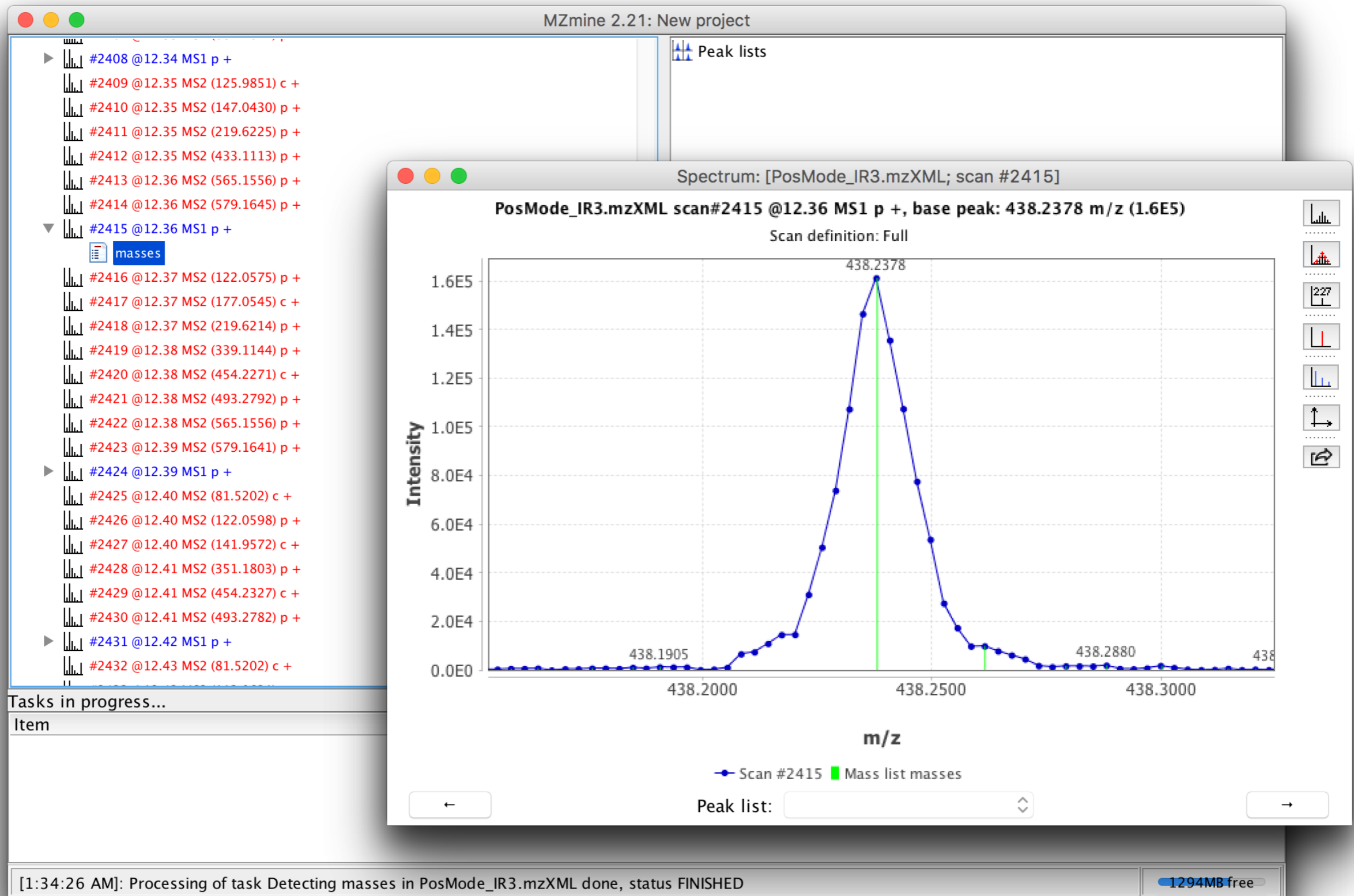
Zoom in one mass spectrum



Mass spectra in centroid mode



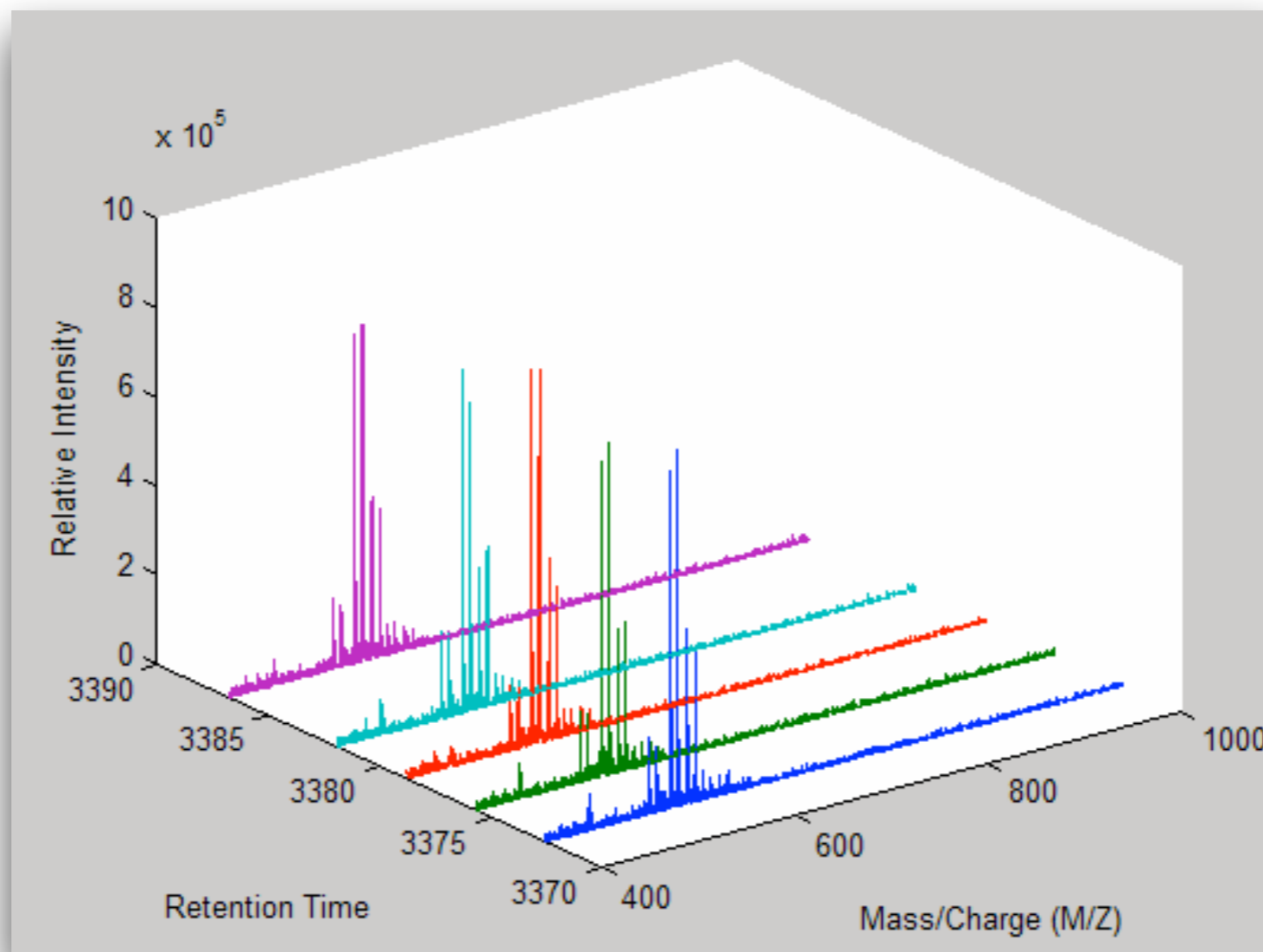
Mass spectra in centroid mode



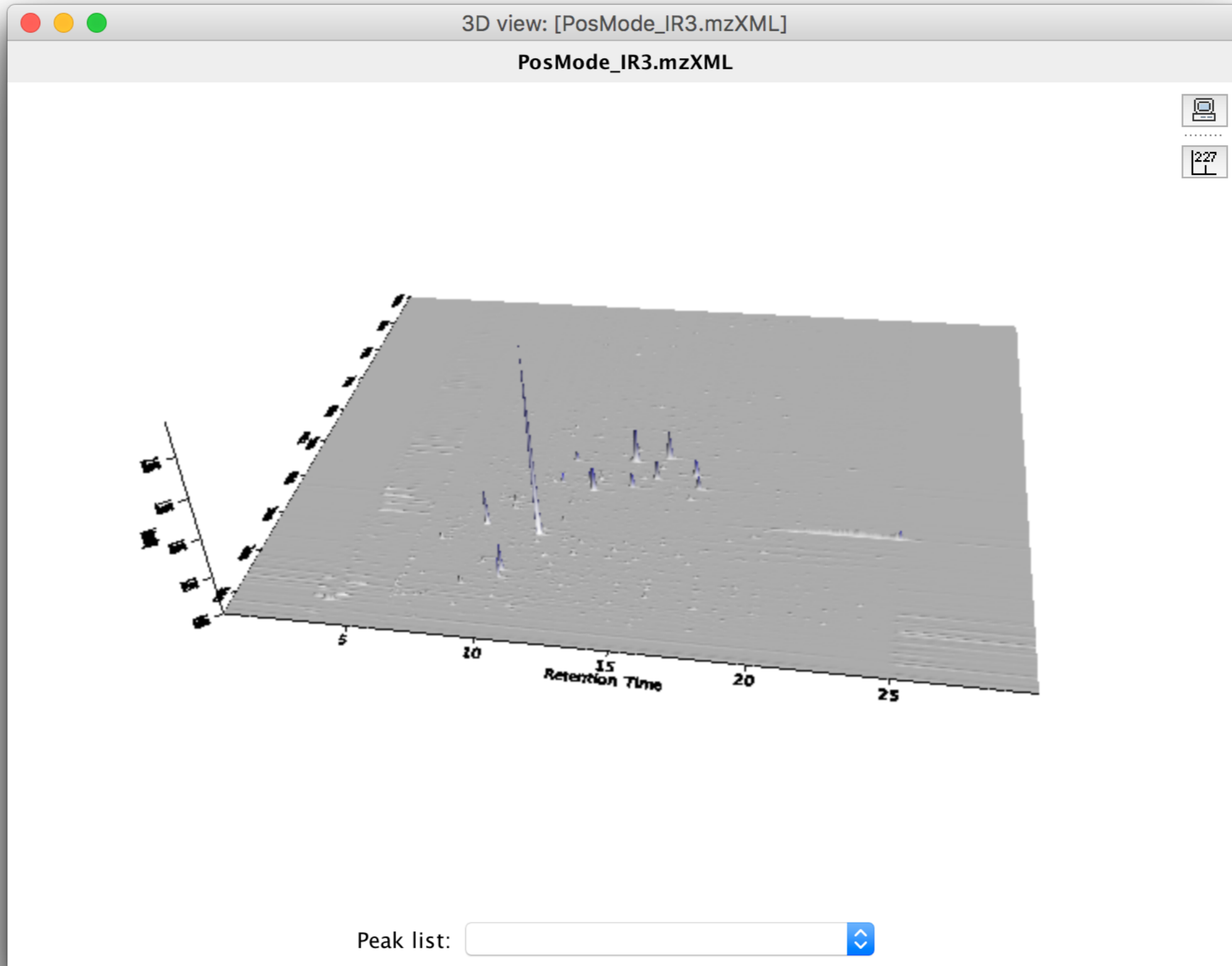
Spectrum in centroid mode

- Data files are much smaller than files in profile mode.
- We will use the centroid data for practicing data pre-processing using XCMS and MZmine 2.

LC-MS raw data in 3D



Raw data in 3D

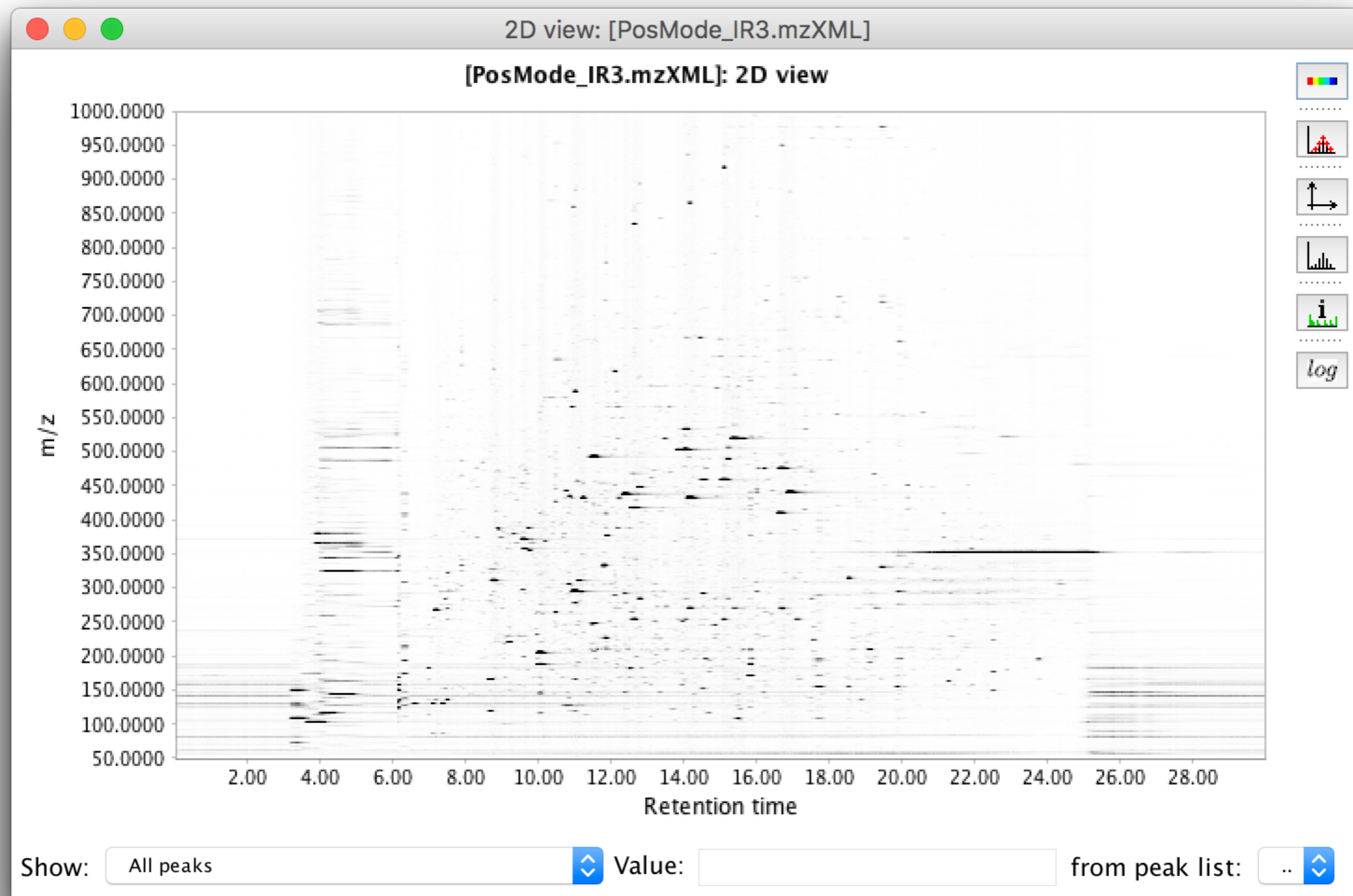


3D to 2D

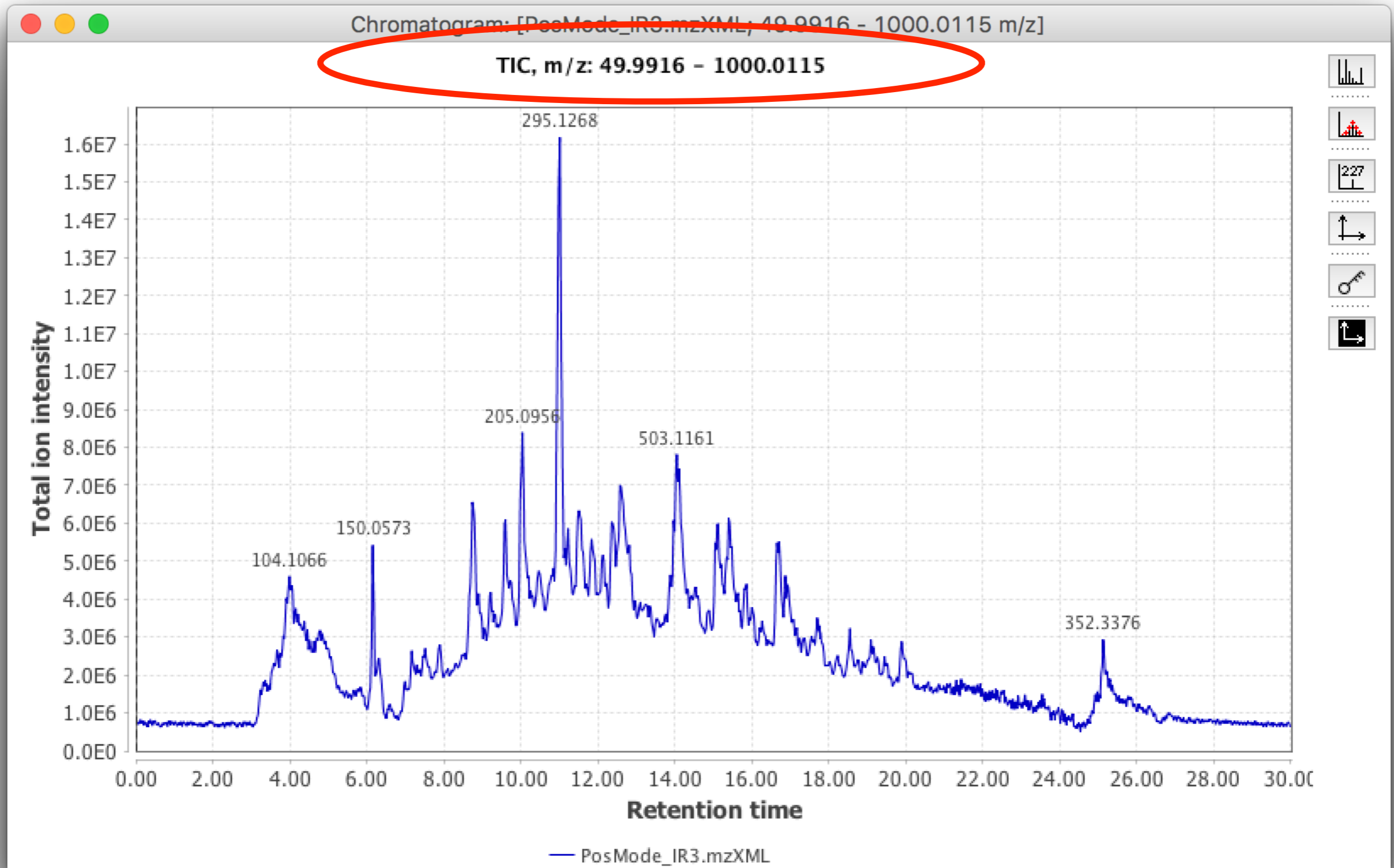
- Direct processing of the 3D data is NOT trivial

- Instead, we examine 2D
 - Mass vs. retention time
 - Total ion current vs. retention time: **TIC**
 - Ion current vs. retention time for a particular mass: **EIC** (Extracted Ion Chromatogram)

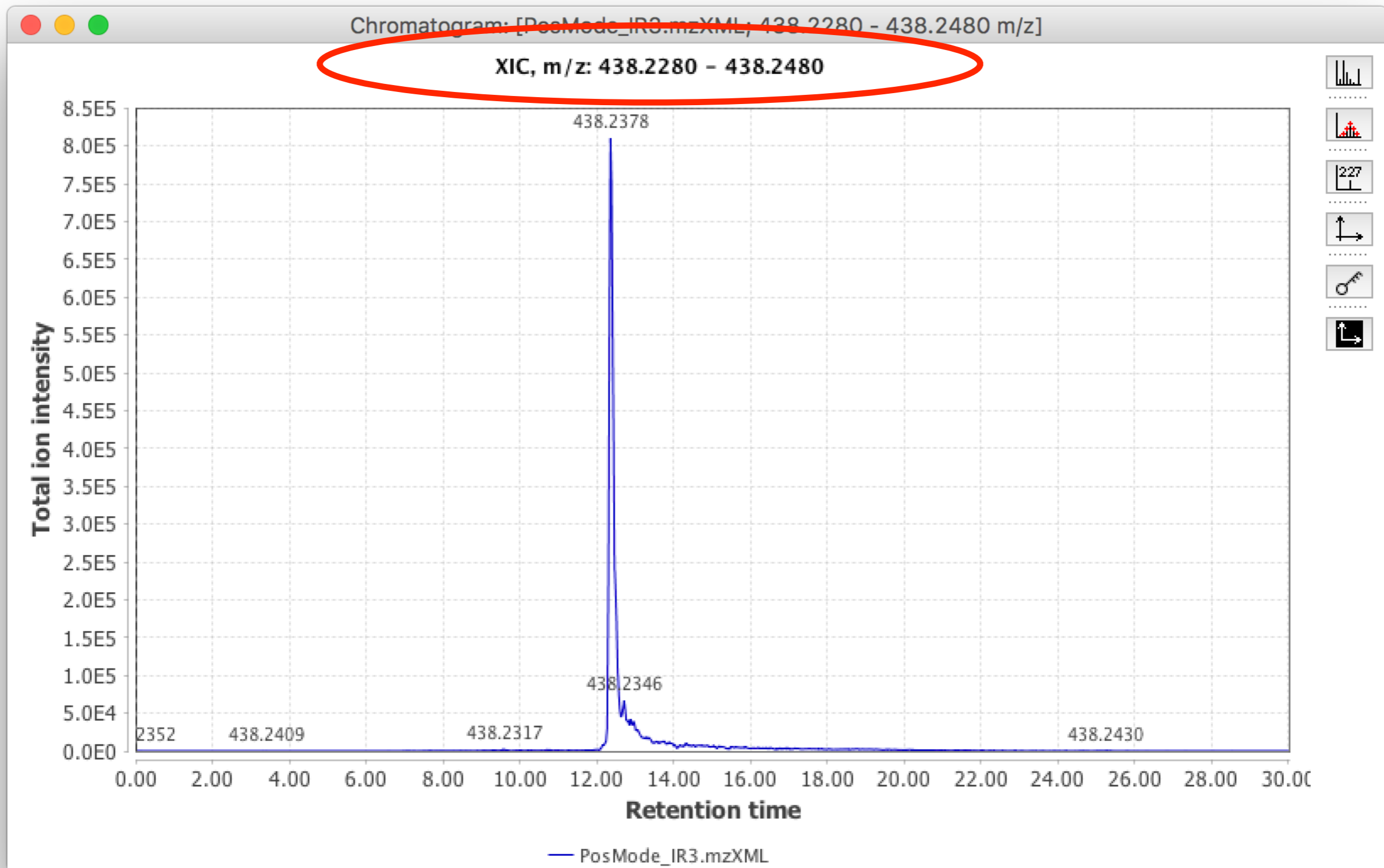
Mass vs. retention time map



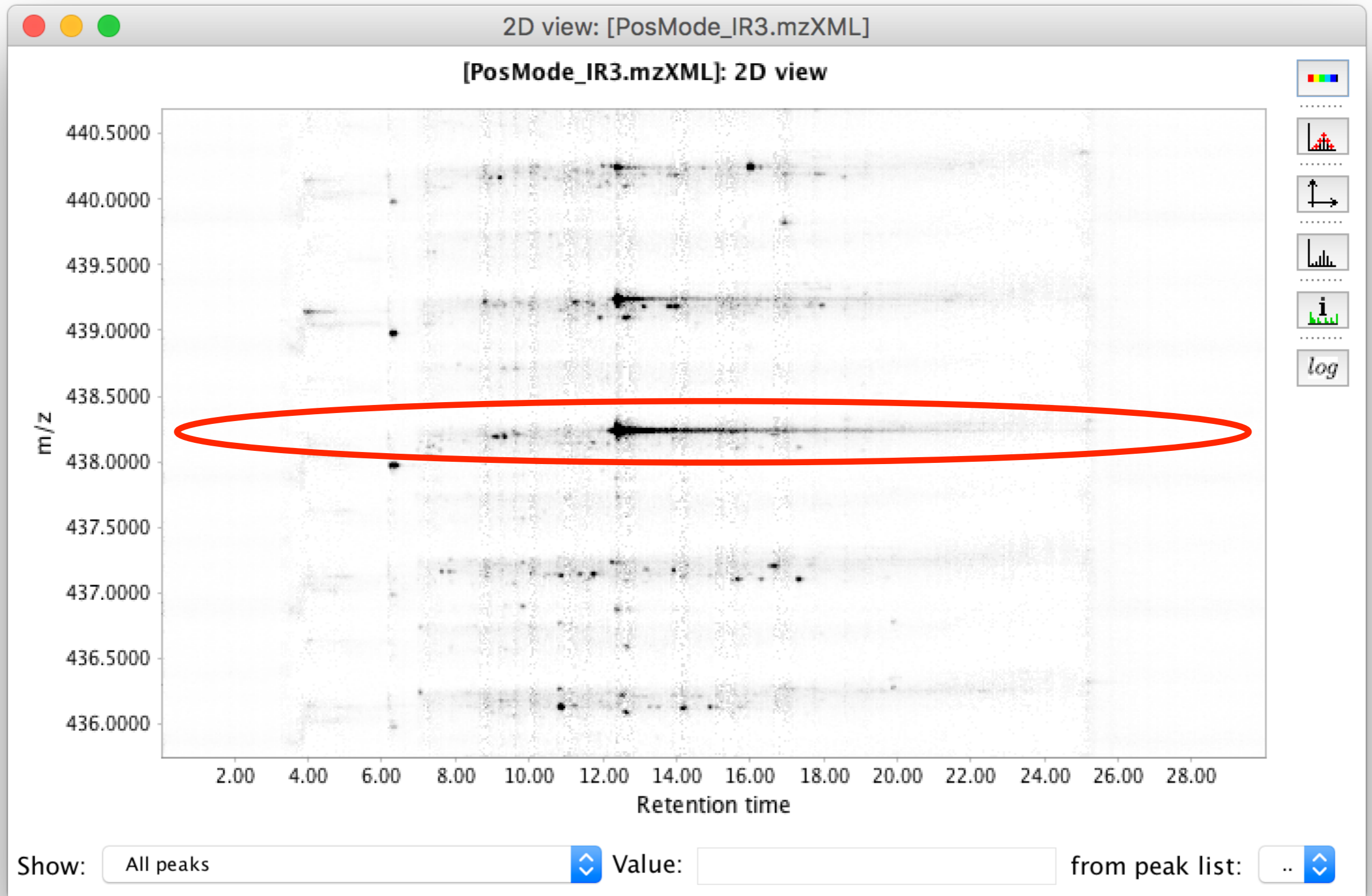
TIC



EIC



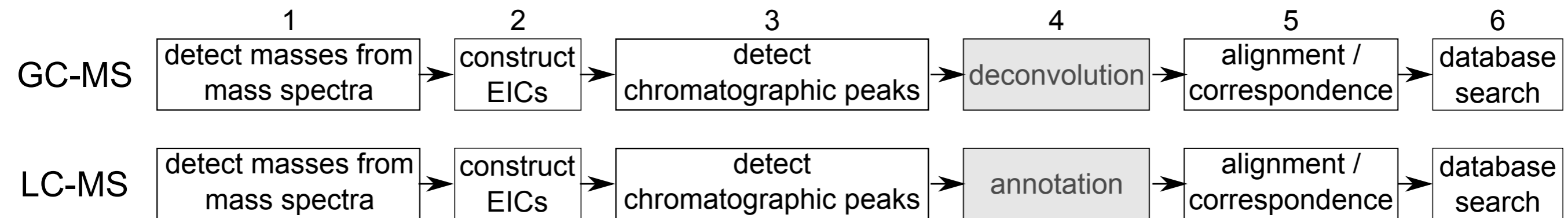
EIC



Principles of LC/MS and GC/MS

Data Preprocessing

Data preprocessing workflow



Construct EICs

MZmine 2.28: New project

Raw data files

- Neg_1e.mzXML

Peak lists

Neg_1e.mzXML chromatograms

- #1 61.9929 m/z @26.02
- #2 68.9982 m/z @26.25
- #3 112.9875 m/z @26.18
- #4 146.0607 m/z @14.32
- #5 154.9736 m/z @26.44
- #6 160.0405 m/z @9.50
- #7 162.0560 m/z @12.48
- #8 165.0561 m/z @15.75
- #9 171.1034 m/z @17.53
- #10 173.0832 m/z @15.02
- #11 174.0857 m/z @15.02
- #12 174.9560 m/z @26.58
- #13 186.1135 m/z @14.27
- #14 187.0973 m/z @16.99
- #15 188.1008 m/z @16.99
- #16 190.0511 m/z @14.30
- #17 191.0550 m/z @14.32

Tasks in progress...

Item	Priority	Status	% do...
------	----------	--------	---------

[11:30:02 PM]: Processing of task Updating TIC visualizer of Neg_1e.mzXML done, status FINISHED

9187MB free

Select one EIC

MZmine 2.28: New project

Raw data files

- Neg_1e.mzXML

#92	305.0686 m/z	@13.34
#93	305.1249 m/z	@14.73
#94	305.1594 m/z	@15.95
#95	307.0858 m/z	@15.43
#96	307.1382 m/z	@14.13
#97	309.1698 m/z	@23.27
#98	311.1689 m/z	@25.38
#99	312.1708 m/z	@25.35
#100	315.1432 m/z	@16.59
#101	317.1240 m/z	@14.13
#102	317.1592 m/z	@16.46
#103	317.1964 m/z	@18.20
#104	318.1253 m/z	@14.13
#105	319.1384 m/z	@13.34
#106	321.1183 m/z	@15.52
#107	321.1533 m/z	@15.92
#108	323.1226 m/z	@13.23
#109	324.0011 m/z	@15.05

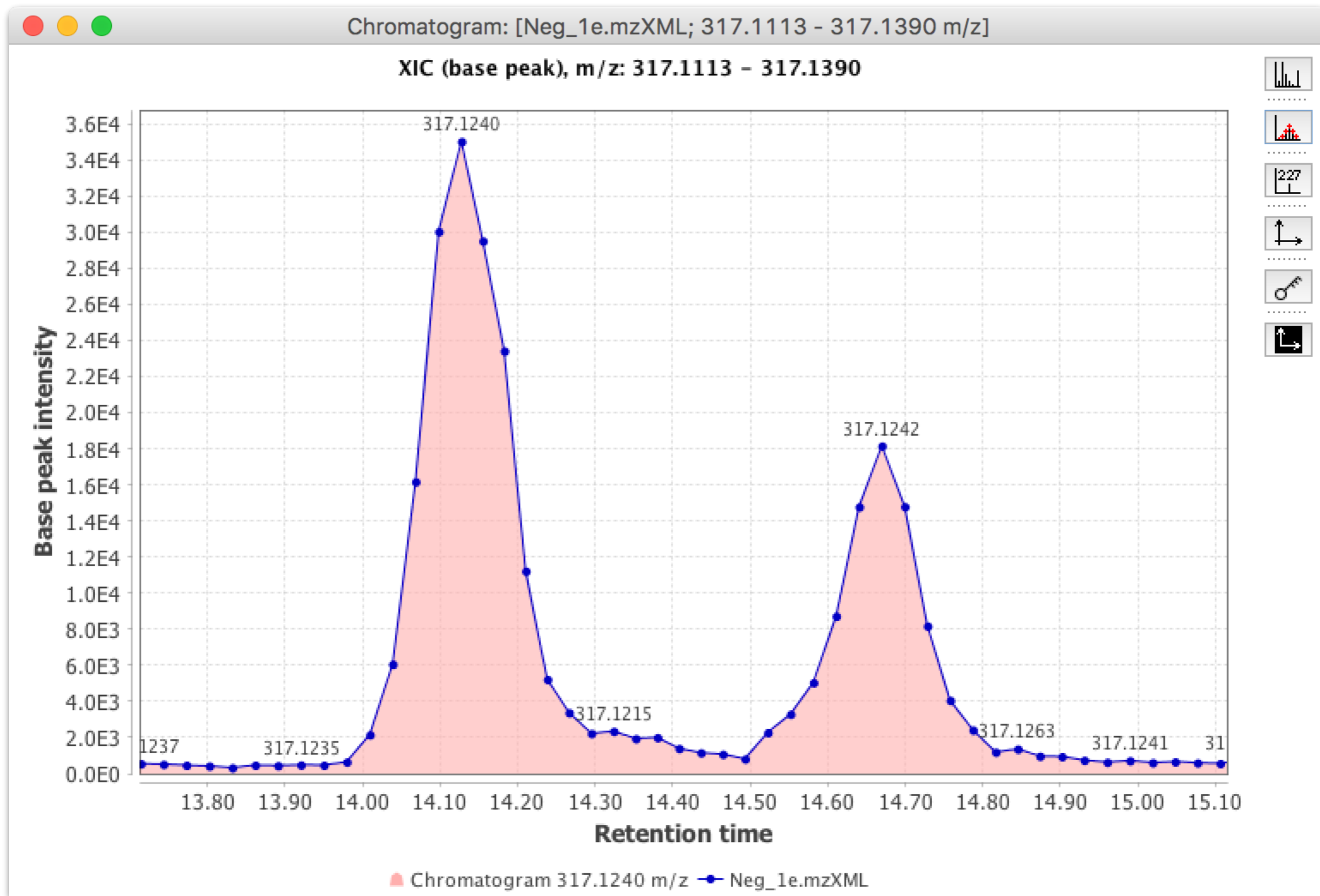
Tasks in progress...

Item	Priority	Status	% do...
------	----------	--------	---------

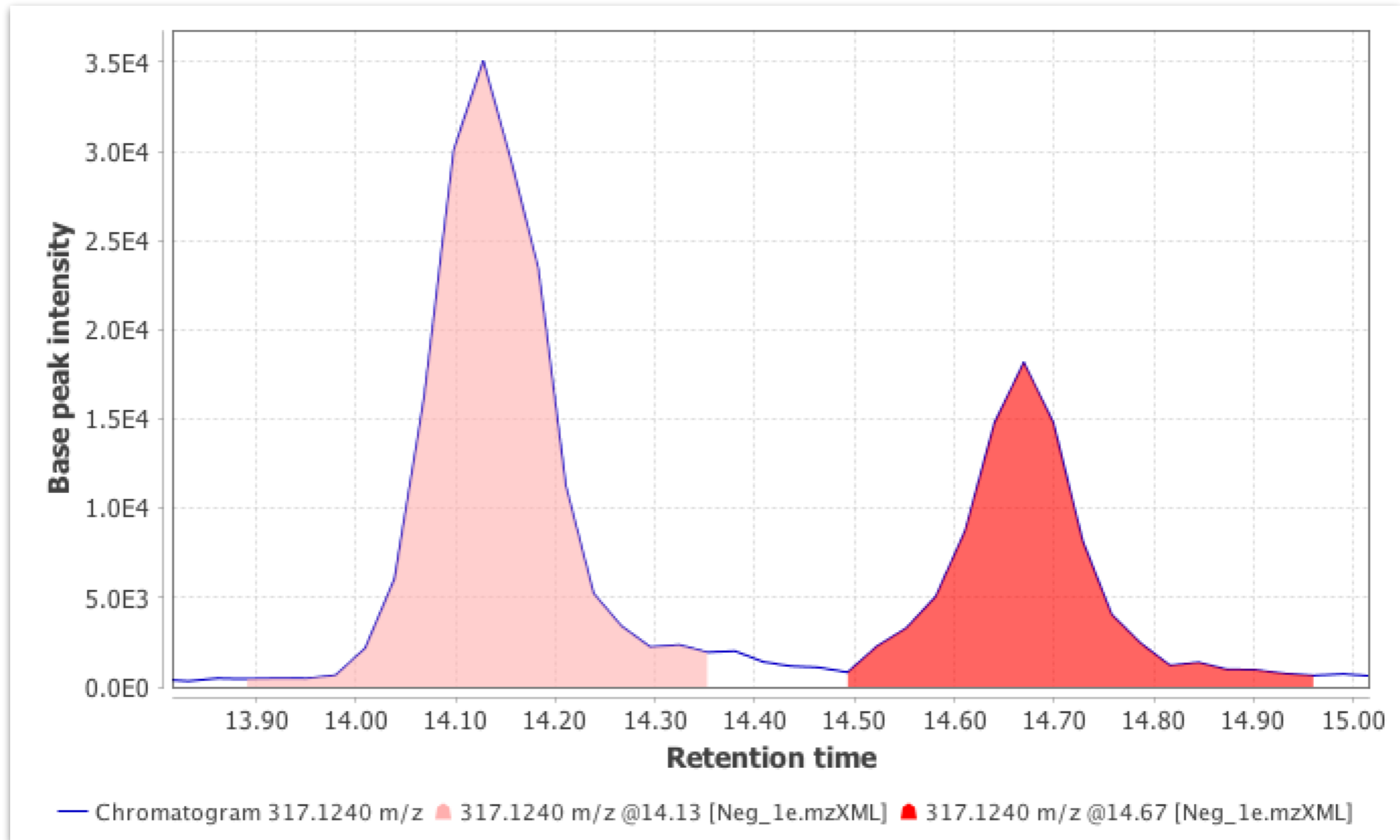
[11:41:31 PM]: Processing of task Updating TIC visualizer of Neg_1e.mzXML done, status FINISHED

6987MB free

One EIC

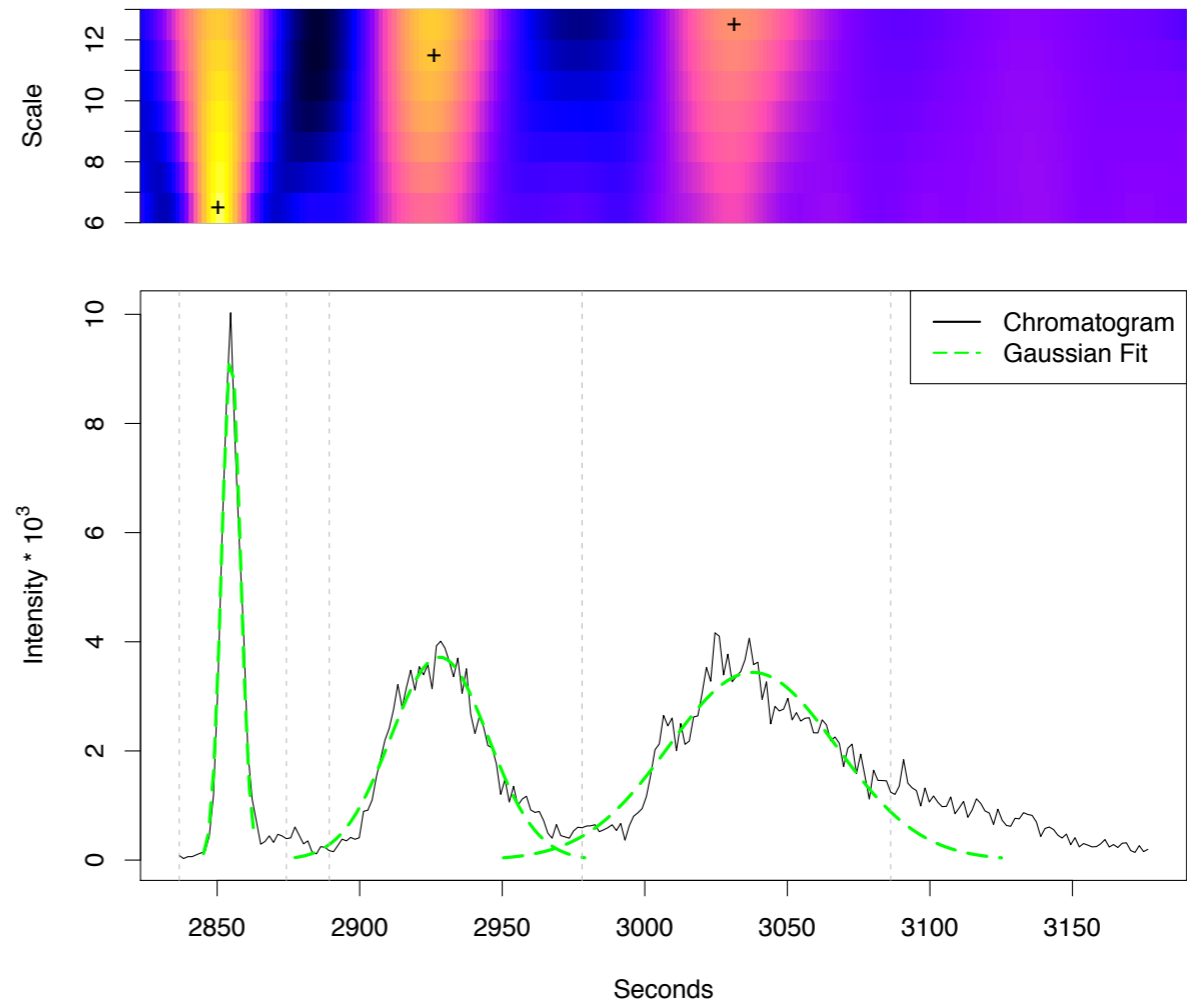
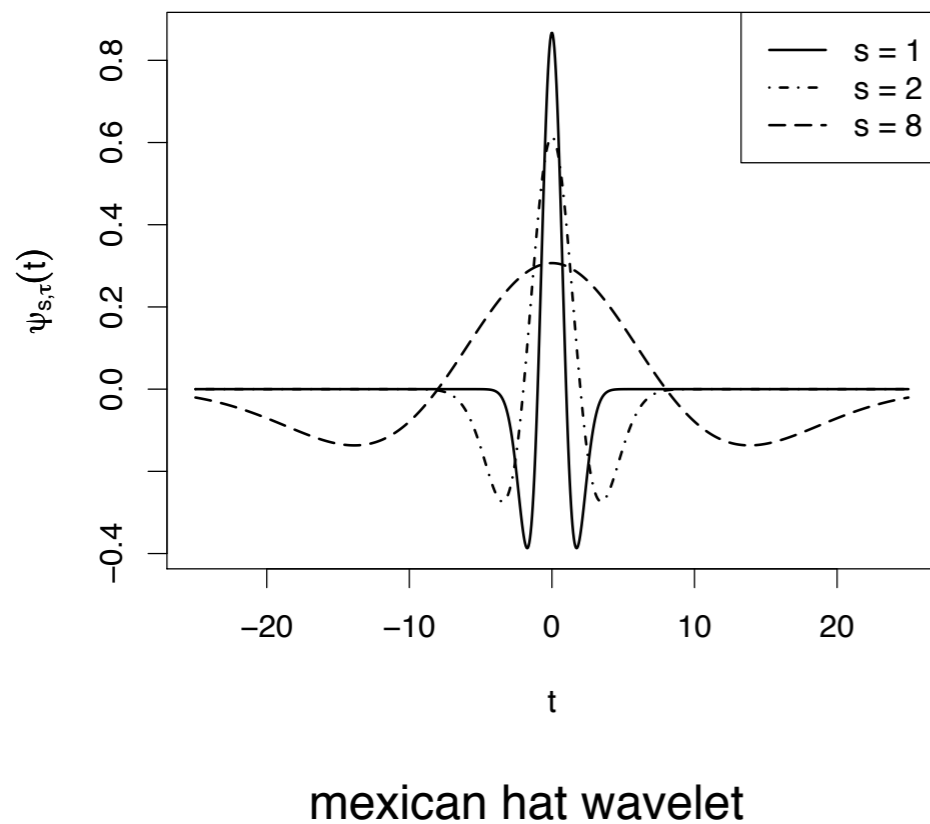


Detect EIC peaks



Detect EIC peaks

- Use wavelet transform



- Implemented in XCMS as the centWave method

Detected EIC peaks

MZmine 2.28: New project

Raw data files

- Neg_1e.mzXML

Peak lists

- Neg_1e.mzXML chromatograms
- Neg_1e.mzXML chromatograms deconvoluted
 - #1 61.9929 m/z @26.05
 - #2 146.0607 m/z @14.32
 - #3 162.0560 m/z @12.48
 - #4 173.0832 m/z @15.02
 - #5 173.0832 m/z @16.57
 - #6 174.0857 m/z @15.02
 - #7 186.1135 m/z @14.27
 - #8 187.0973 m/z @16.99
 - #9 188.1008 m/z @16.99
 - #10 190.0511 m/z @14.30
 - #11 191.0550 m/z @14.32
 - #12 192.0670 m/z @15.52
 - #13 193.0501 m/z @15.78
 - #14 195.0661 m/z @15.22
 - #15 197.0464 m/z @13.63
 - #16 199.0970 m/z @16.57

Tasks in progress...

Item	Priority	Status	% do...
------	----------	--------	---------

[11:41:31 PM]: Processing of task Updating TIC visualizer of Neg_1e.mzXML done, status FINISHED

6529MB free

LC/MS-specific Data Preprocessing

Find isotopes

Peak list: Neg_1e.mzXML chromatograms deconvoluted

ID	Average		Identity	Comment	Peak shape	Neg_1e.mzXML		
	m/z ▲	RT				Status	Height	Area
71	303.0177	12.39				●	2.7E3	1.8E4
72	303.0889	15.57				●	2.0E3	1.2E4
73	305.0686	13.34				●	4.4E3	3.3E4
74	307.0858	15.43				●	2.4E3	1.8E4
75	309.1698	23.27				●	3.9E3	3.0E4
76	309.1698	22.66				●	2.9E3	1.5E4
77	317.1240	14.13				●	3.5E4	2.9E5
78	317.1240	14.67				●	1.8E4	1.5E5
79	317.1964	18.20				●	3.8E3	2.9E4
80	318.1253	14.13				●	5.5E3	4.8E4
81	323.1226	13.23				●	3.6E3	2.0E4
82	324.0011	15.05				●	3.5E3	2.2E4
83	324.2170	21.52				●	2.7E3	1.6E4
84	329.2315	21.06				●	5.8E3	4.8E4
85	331.1395	13.31				●	3.3E3	2.6E4
86	331.2477	20.86				●	8.8E3	5.1E4
87	333.0069	12.91				●	1.4E4	1.2E5
88	333.1177	14.90				●	5.8E3	4.5E4

Find isotopes

MZmine 2.28: New project

Raw data files

- Neg_1e.mzXML

Peak lists

- Neg_1e.mzXML chromatograms
- Neg_1e.mzXML chromatograms deconvoluted
- Neg_1e.mzXML chromatograms deconvoluted deisotoped**
 - #8 187.0973 m/z @16.99
 - #175 514.2834 m/z @18.81
 - #111 352.9955 m/z @14.79
 - #77 317.1240 m/z @14.13
 - #95 343.1768 m/z @18.06
 - #10 190.0511 m/z @14.30
 - #159 443.1898 m/z @11.44
 - #100 345.2276 m/z @20.52
 - #46 243.1238 m/z @16.88
 - #149 431.1912 m/z @13.23
 - #4 173.0832 m/z @15.02
 - #98 345.1537 m/z @15.28
 - #34 225.1130 m/z @14.35
 - #1 61.9929 m/z @26.05
 - #40 239.0925 m/z @14.30

Tasks in progress...

Item	Priority	Status	% do...
------	----------	--------	---------

[12:00:07 AM]: Processing of task Updating TIC visualizer of Neg_1e.mzXML done, status FINISHED

7098MB free

Find isotopes

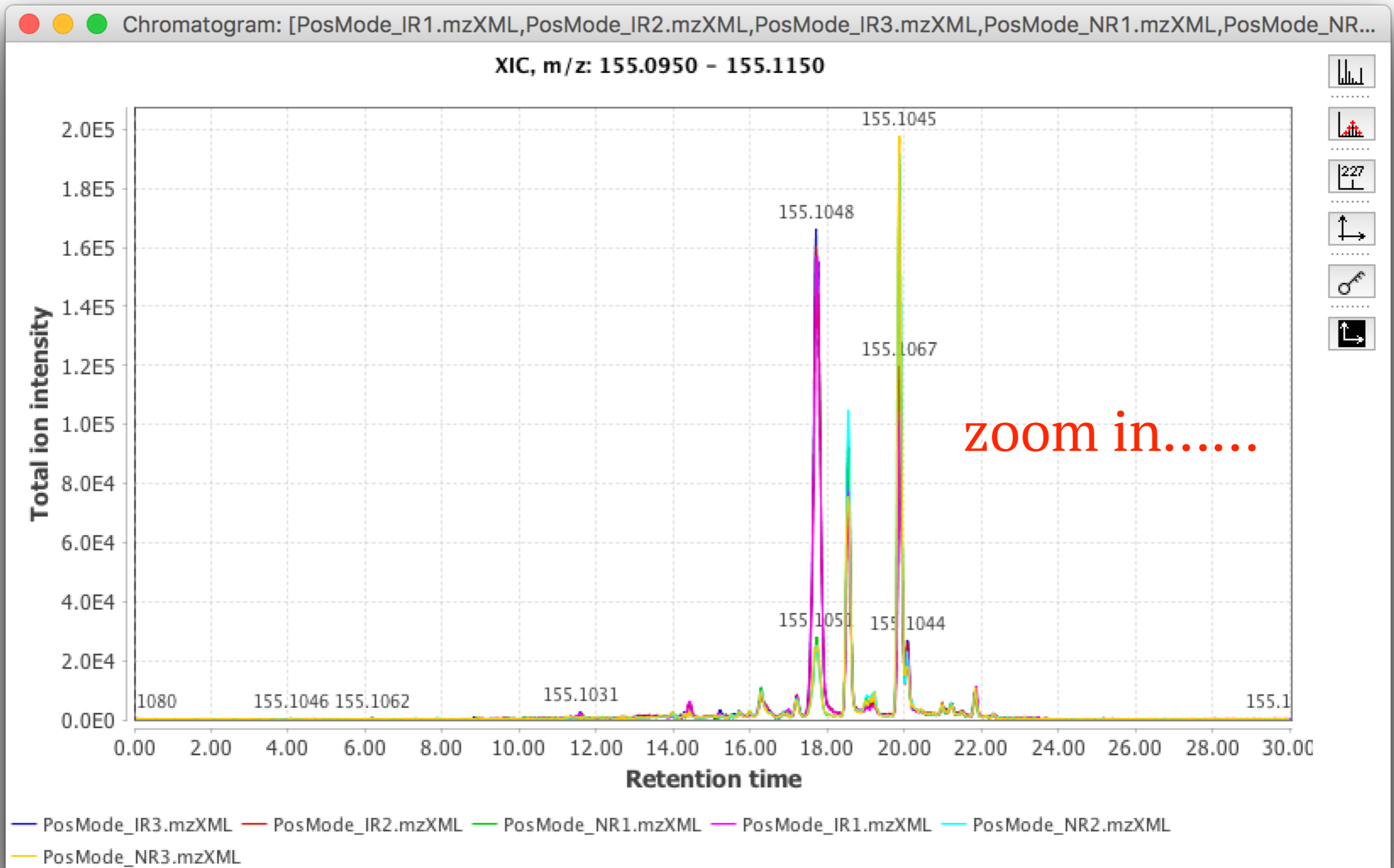
Peak list: Neg_1e.mzXML chromatograms deconvoluted deisotoped

ID	Average		Identity	Comment	Peak shape	Neg_1e.mzXML		
	m/z ▲	RT				Status	Height	Area
72	303.0889	15.57				●	2.0E3	1.2E4
73	305.0686	13.34				●	4.4E3	3.3E4
74	307.0858	15.43				●	2.4E3	1.8E4
75	309.1698	23.27				●	3.9E3	3.0E4
76	309.1698	22.66				●	2.9E3	1.5E4
77	317.1240	14.13				●	3.5E4	2.9E5
78	317.1240	14.67				●	1.8E4	1.5E5
79	317.1964	18.20						
81	323.1226	13.23						
82	324.0011	15.05						
83	324.2170	21.52						
84	329.2315	21.06						
85	331.1395	13.31				●	3.3E3	2.6E4
86	331.2477	20.86				●	8.8E3	5.1E4
87	333.0069	12.91				●	1.4E4	1.2E5
88	333.1177	14.90				●	5.8E3	4.5E4
89	336.1471	16.57				●	5.4E3	4.4E4
90	336.1471	13.98				●	3.4E3	2.5E4

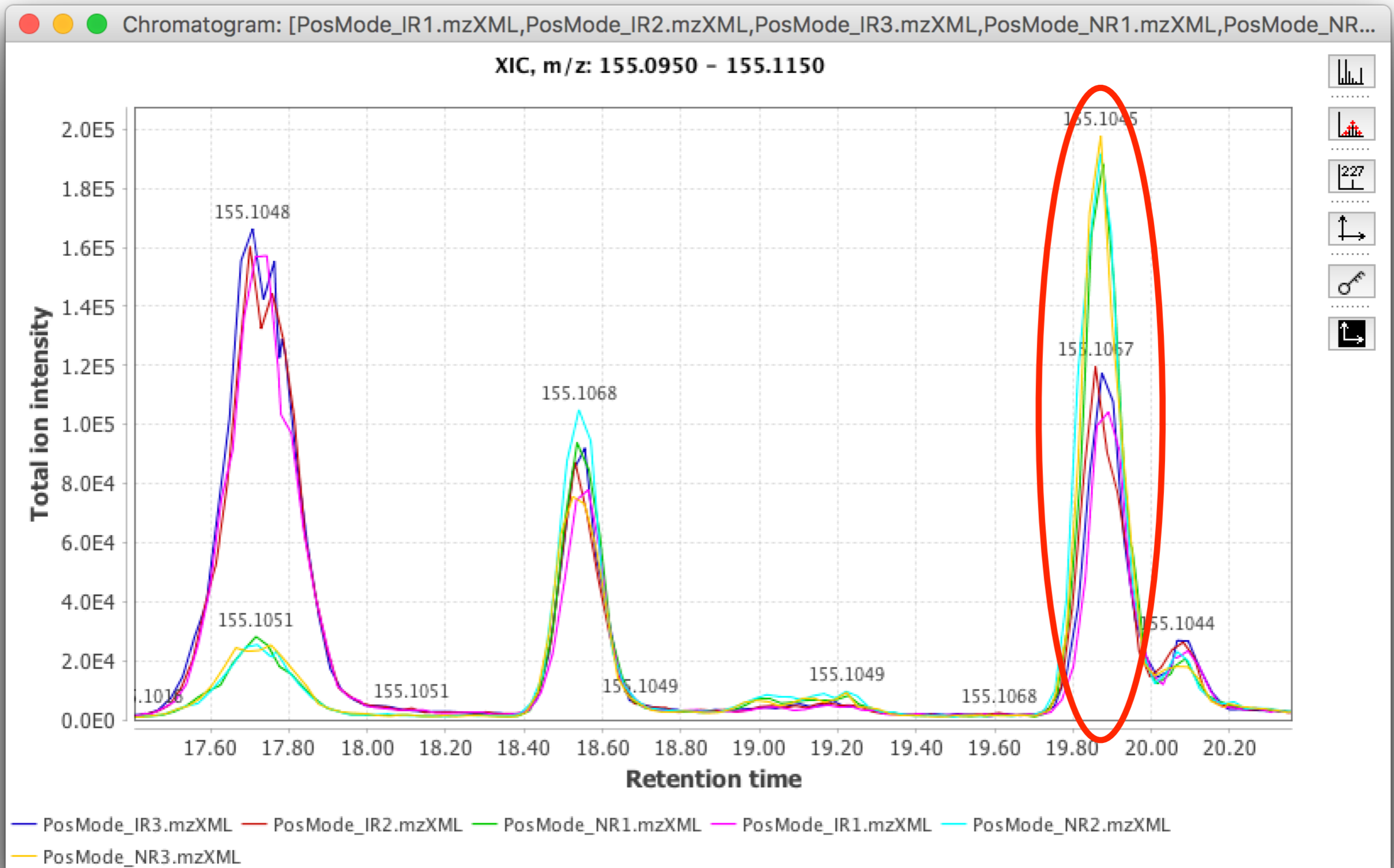
one_example_of_isotopic_group_in_text_format.txt

```
317.1239929199219 35057.0  
318.1253356933594 5473.0
```

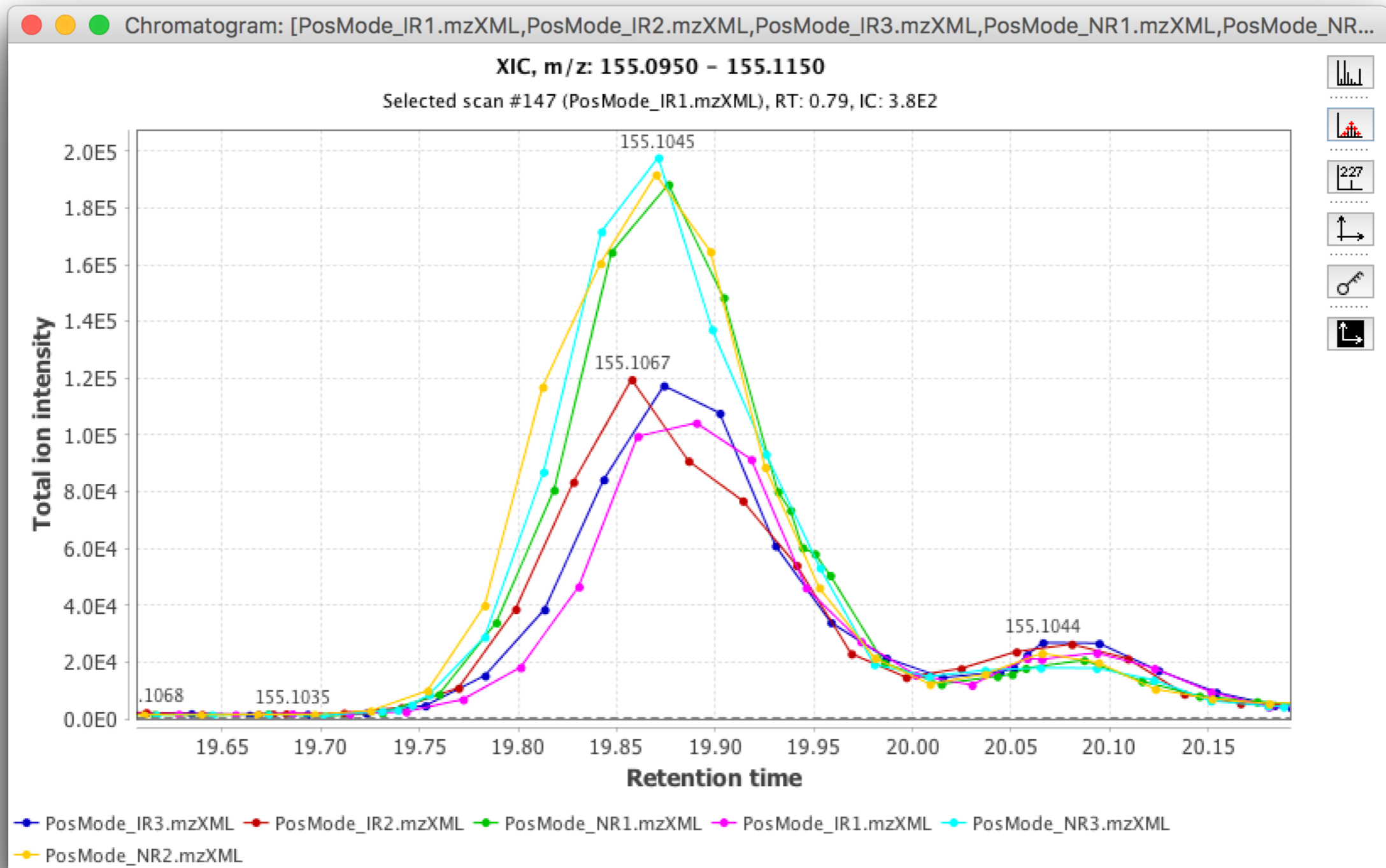
Alignment



Alignment



Alignment

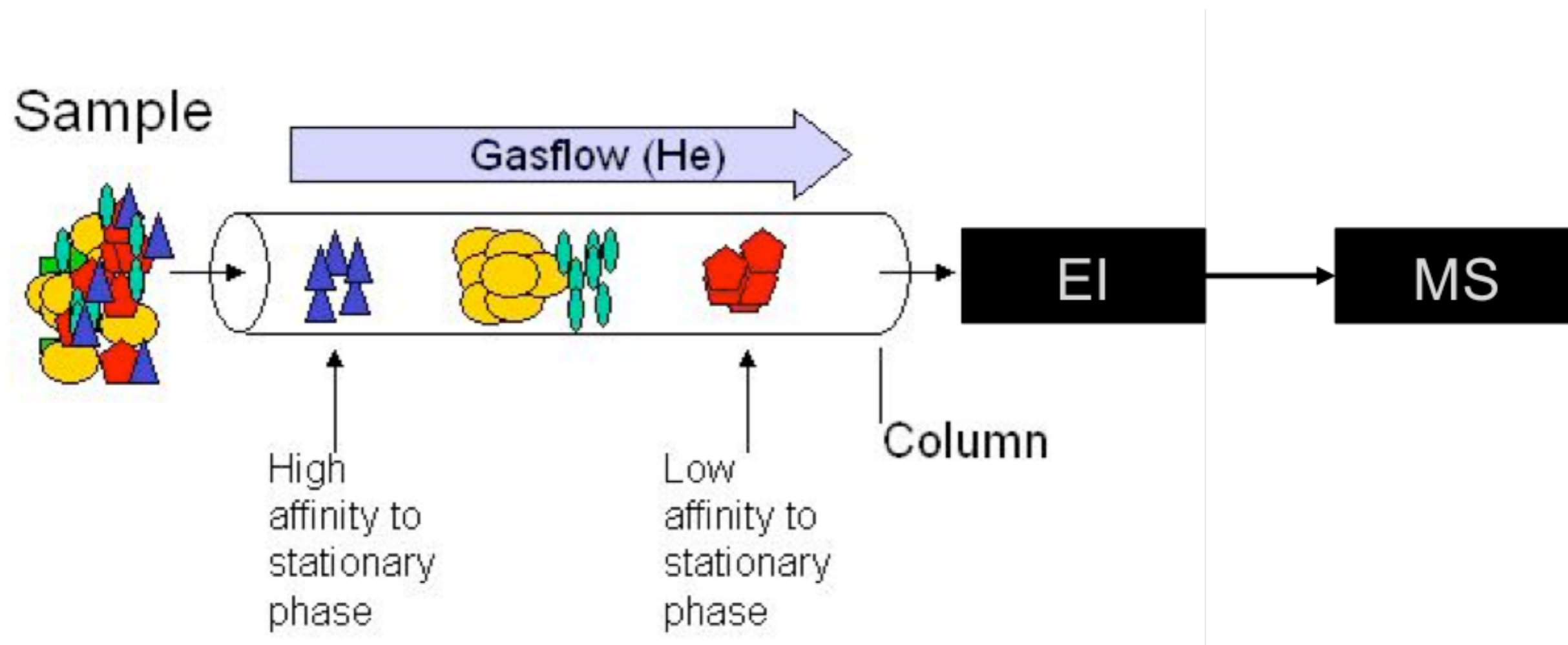


Peaks table after alignment

A	B	C	D	E	F	G	H	I	J	K
row ID	row m/z	row retentio	row identity	row commer	row number	Neg_2e.mzX	Neg_2e.mzX	Neg_2e.mzX	Neg_2e.mzX	Neg_2e.mzX
1	443.190292	11.4494583			2	DETECTED	443.190765	11.4633167	11.3171167	11.8174333
2	273.007701	12.9576056			3	DETECTED	273.007996	12.9542167	12.8374833	13.3464
3	187.097656	16.9978333			3	DETECTED	187.09787	17.0085	16.75	17.3798333
4	345.227895	20.4976667			3	DETECTED	345.22818	20.5048333	20.3618333	20.7863333
5	343.175323	18.043			1	DETECTED	343.175323	18.043	17.7836667	18.3851667
6	317.124016	14.1371542			4	DETECTED	317.124481	14.1317333	13.86585	14.33435
7	190.051648	14.3116067			5	DETECTED	190.051834	14.33435	14.18985	14.5621833
8	112.986198	26.3228333			3	DETECTED	112.986519	26.3391667	26.3126667	26.3668333
9	431.191948	13.2436333			4	DETECTED	431.1922	13.2616333	13.118	13.54485
10	514.282939	18.8240556			3	DETECTED	514.284119	18.8443333	18.4996667	19.1305
11	243.124359	16.8828889			3	DETECTED	243.12468	16.8938333	16.6649833	17.0368333
12	225.113566	14.3711944			3	DETECTED	225.113586	14.3897333	14.0723667	14.6212
13	206.046432	12.5004889			3	DETECTED	206.047104	12.5219333	12.3823	12.8968833
14	305.068522	13.3252444			3	DETECTED	305.069183	13.3181833	13.0626833	13.5740167
15	517.154953	12.0159			4	DETECTED	517.156433	12.0131333	11.9021333	12.3528833
16	239.093052	14.3080167			4	DETECTED	239.092896	14.3049667	14.0427167	14.6212
17	345.155792	15.2841278			3	DETECTED	345.154785	15.2977167	15.15055	15.4443667
18	303.018661	12.3884417			2	DETECTED	303.01886	12.3823	12.2933	12.7488833
19	173.083302	15.0110667			4	DETECTED	173.083511	15.0329333	14.79785	15.1800667
20	415.197021	19.6466667			1	DETECTED	415.197021	19.6466667	19.3603333	19.9328333
21	387.163747	13.54635			3	DETECTED	387.164093	13.54485	13.2616333	13.8954167
22	352.997826	14.7999			4	DETECTED	352.998566	14.82705	14.6507167	15.12105

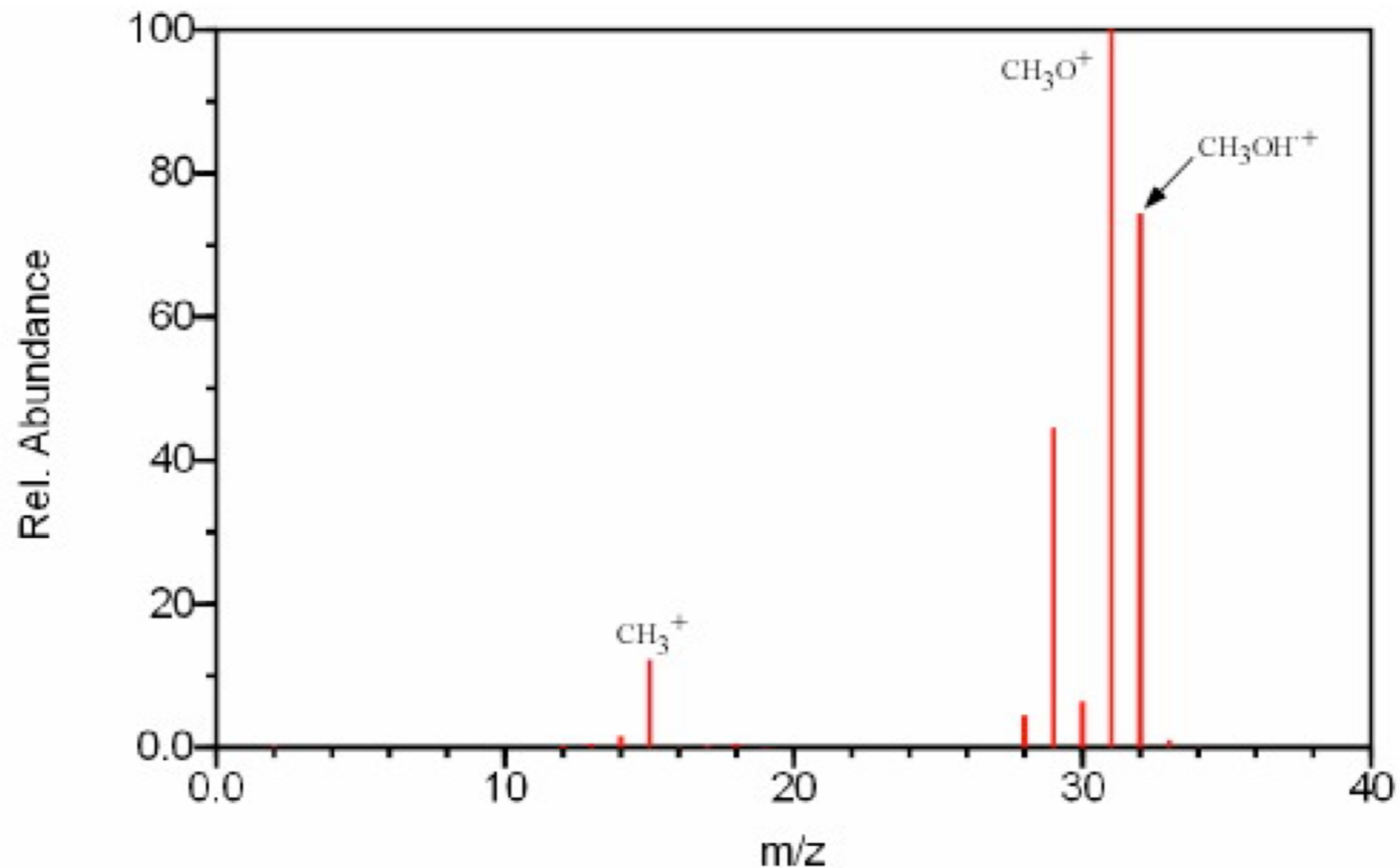
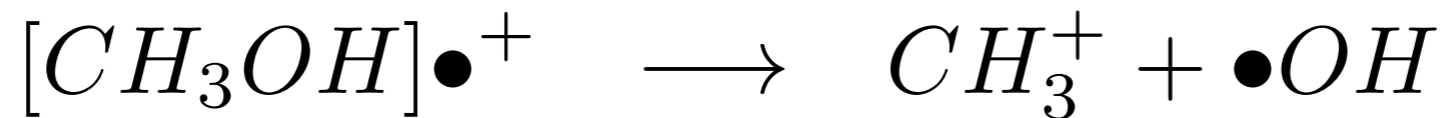
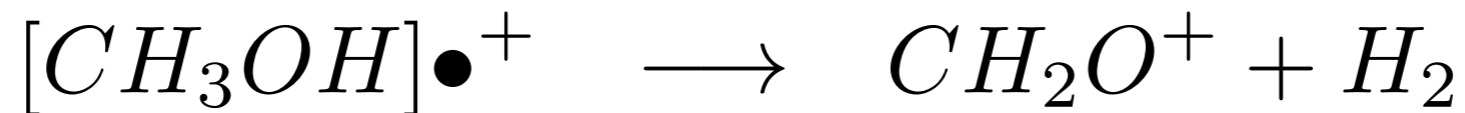
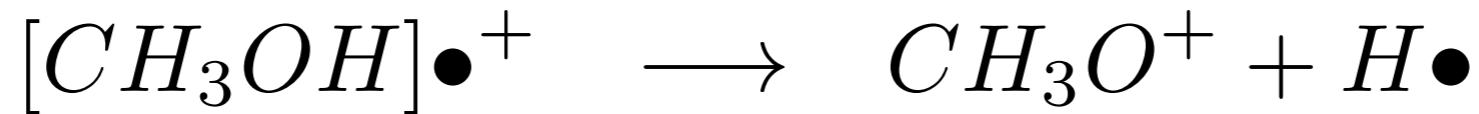
GC/MS-specific Data Preprocessing

GC-EI-MS

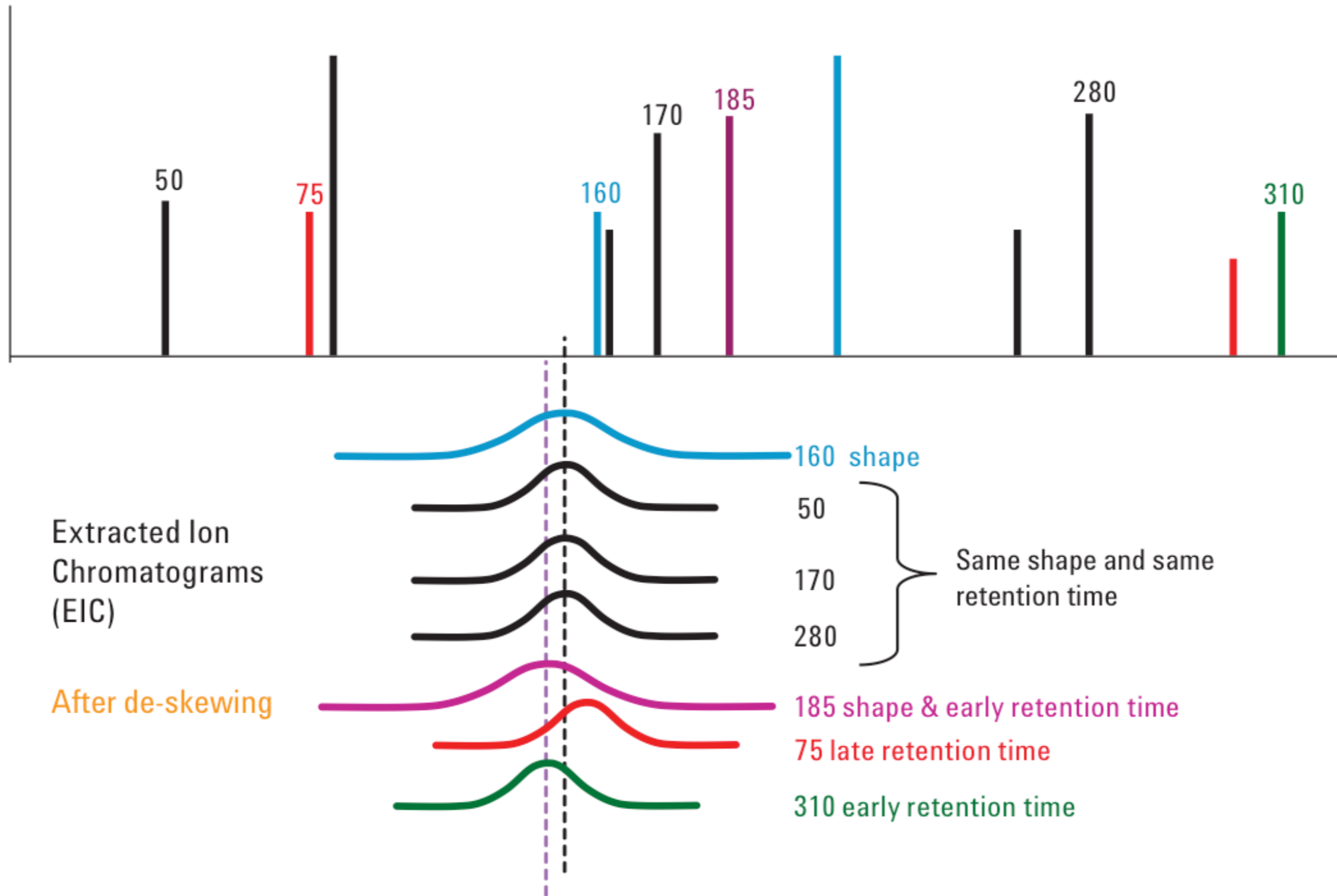


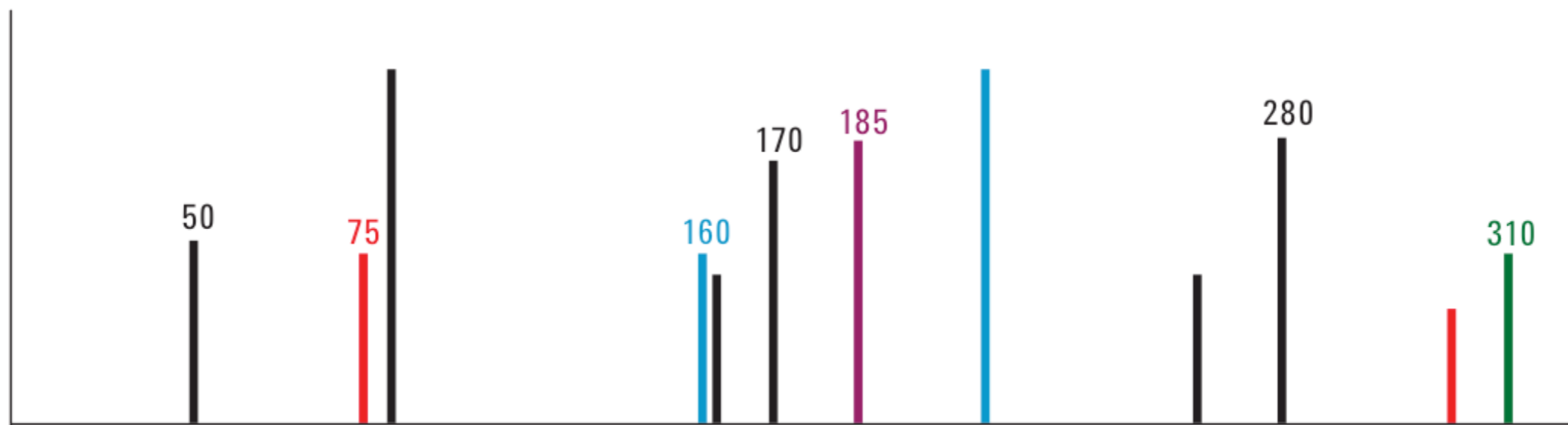
EI fragmentation

- **Example:** EI fragmentation of methanol

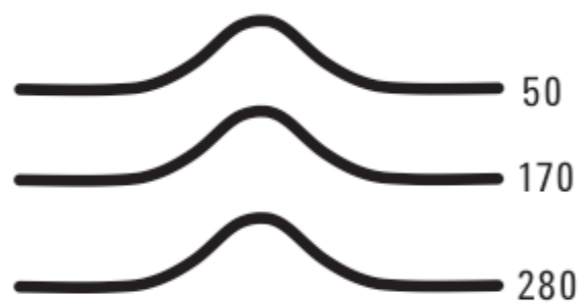


Deconvolution

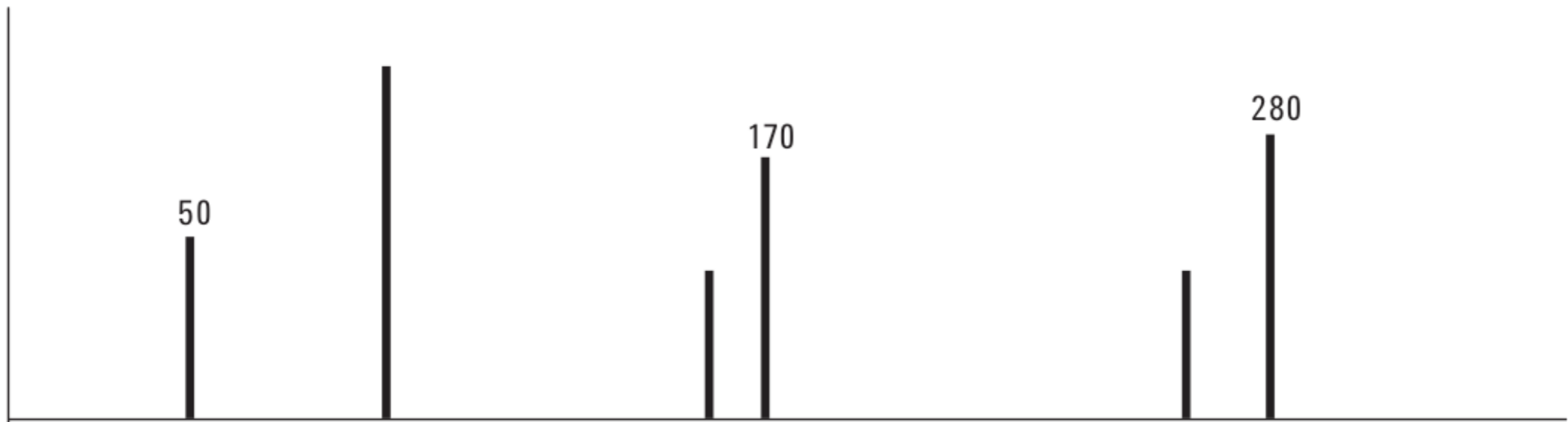




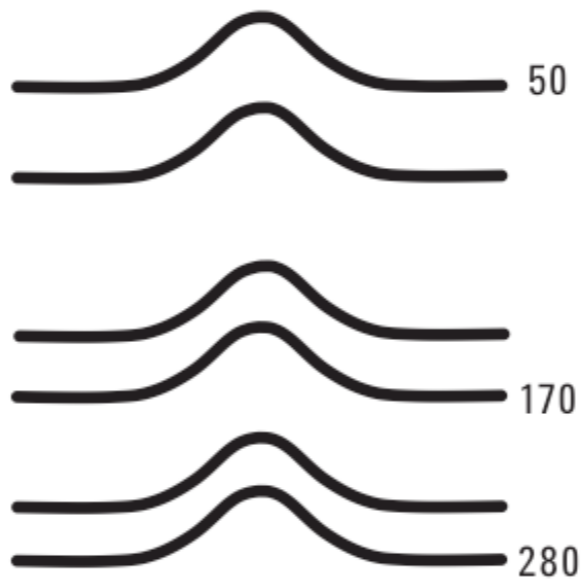
Extracted Ion Chromatograms (EIC)



Only the ions in black have the same shape and retention time as shown by 50, 170, 280- plus others

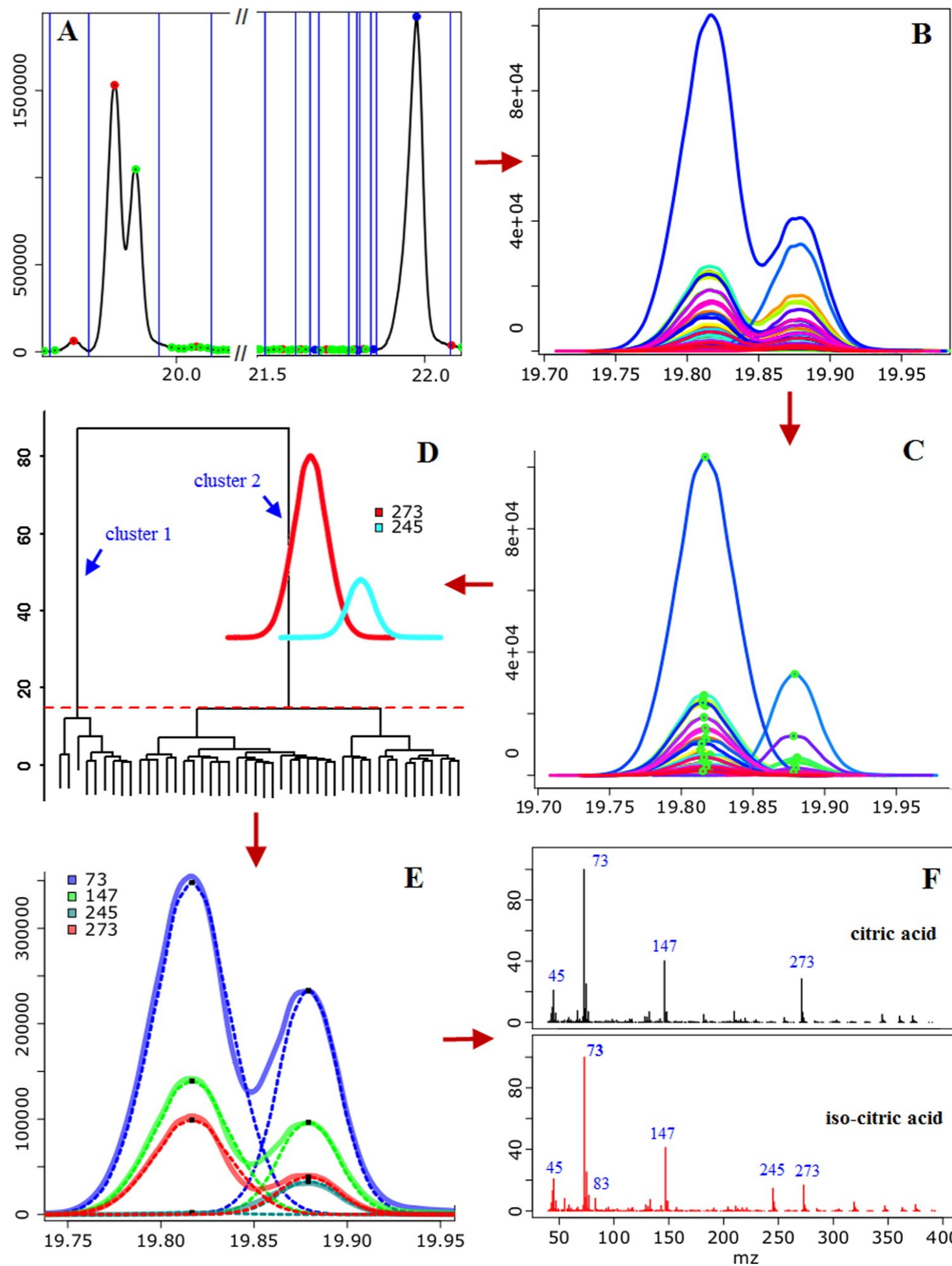


Extracted Ion
Chromatograms
(EIC)



These
deconvoluted ions
are grouped
together as a
component

ADAP-GC 2.0



ADAP-GC 2.0: Deconvolution of Coeluting Metabolites from GC/TOF-MS Data for Metabolomics Studies. *Analytical chemistry* 2012, 84 (15), 6619-29.

Feature identification

Identification of known compounds

- Screening search for compound ID based on LC-MS data
 - Searching monoisotopic mass and isotopic distribution against compound databases

- Library match for compound identification from both LC-MS/MS and GC-MS spectra

MS Search

MS/MS Search

GC/MS Search

1D NMR Search

2D NMR Search

Query Masses (Da)

147.11

Enter one mass per line (maximum 700 query masses per request)



Ionization

Ion Mode

Positive

Adduct Type

M+ACN+2H
M+2Na
M+2ACN+2H
M+3ACN+2H
M+H
M+NH4
M+Na
M+CH3OH+H

Hold Ctrl () or Command () to select multiple adducts

Molecular Weight
Tolerance ±

Da

Search

Load Example

Search Results

[Download Results As CSV](#)

MS search for 147.11 m/z

i Delta = abs(query mass - adduct mass)

Show entries

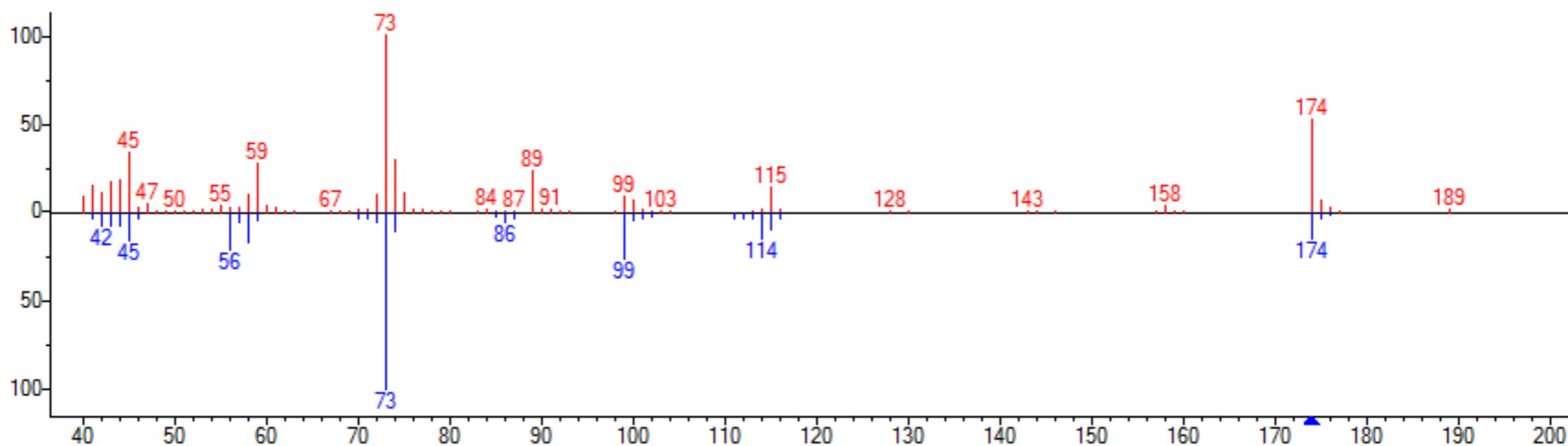
Compound	Name	Adduct	Adduct MW (Da)	Compound MW (Da)	Delta
HMDB12115	(3S,5S)-3,5-Diaminohexanoate	M+H	147.112804	146.105527702	0.002804
HMDB12114	(3S)-3,6-Diaminohexanoate	M+H	147.112804	146.105527702	0.002804
HMDB00182	L-Lysine	M+H	147.112804	146.105527702	0.002804
HMDB03405	D-Lysine	M+H	147.112804	146.105527702	0.002804
HMDB61808	(3-Methyl-2-butenyl)-benzene	M+H	147.116826	146.109550448	0.006826
HMDB39407	Methyl (±)-3-hydroxyhexanoate	M+H	147.10157	146.094294314	0.00843
HMDB61653	3-hydroxyheptanoic acid	M+H	147.10157	146.094294314	0.00843
HMDB36231	Methyl DL-Leucate	M+H	147.10157	146.094294314	0.00843
HMDB02207	3-Hydroxyisoheptanoic acid	M+H	147.10157	146.094294314	0.00843
HMDB32269	(+/-)-Ethyl 2-hydroxy-2-methylbutyrate	M+H	147.10157	146.094294314	0.00843

Showing 1 to 10 of 62 entries

[Previous](#)
[1](#)
[2](#)
[3](#)
[4](#)
[5](#)
[6](#)
[7](#)
[Next](#)

MS/MS or GC-MS spectra matching

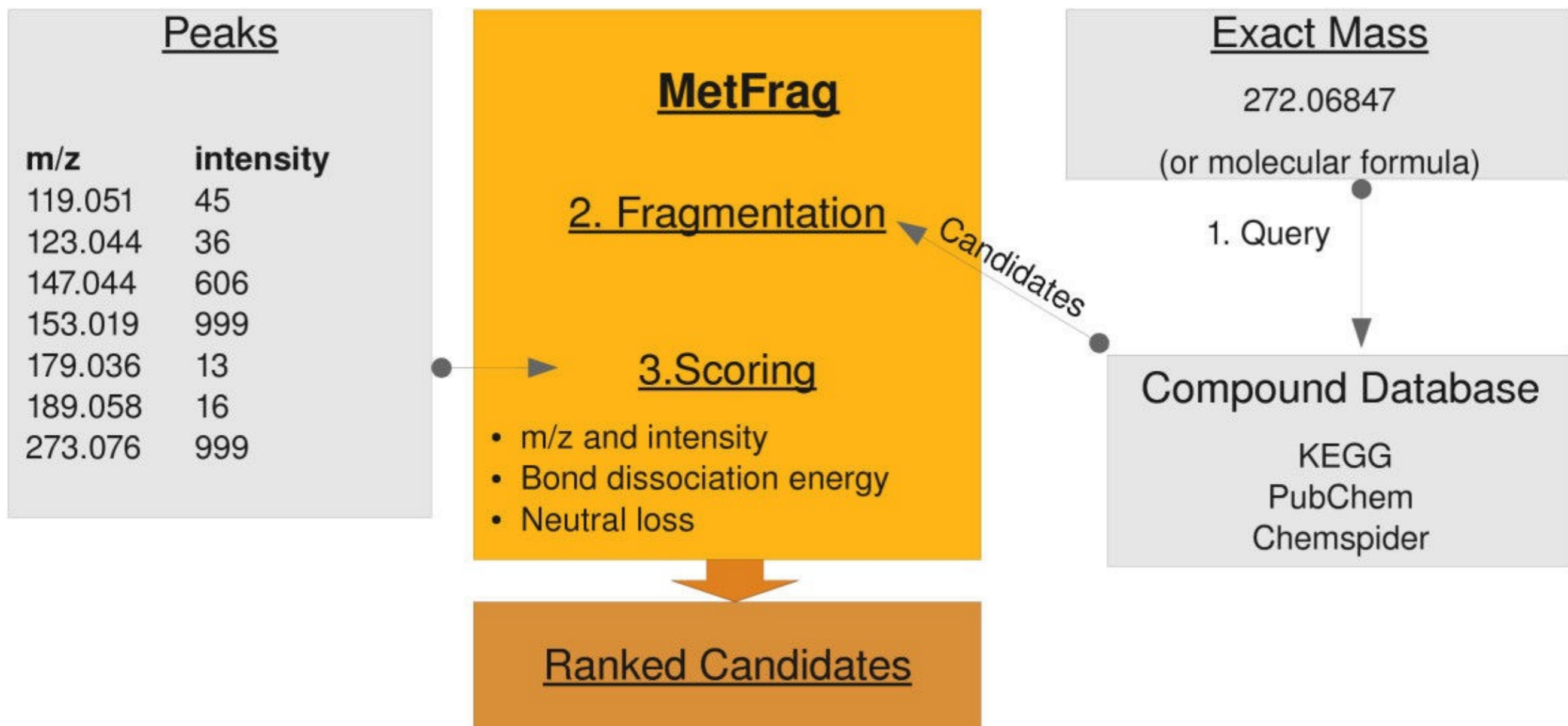
- Library match for compound identification from both LC-MS/MS and GC-MS spectra



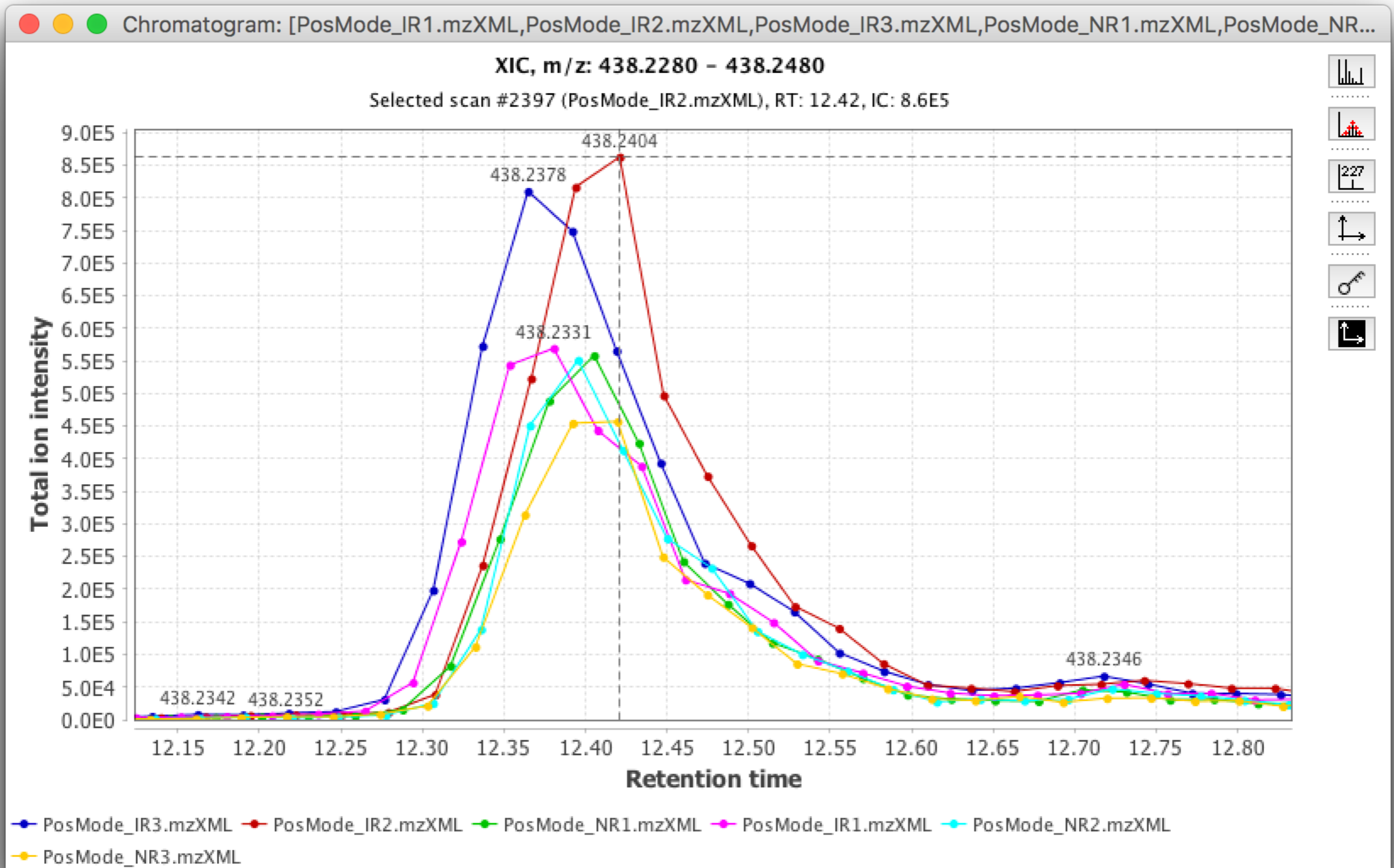
Identification of unknown compounds

- MS-FINDER
- CSI:FingerID
- CFM-ID
- MetFrag
- MIDAS
- MAGMA

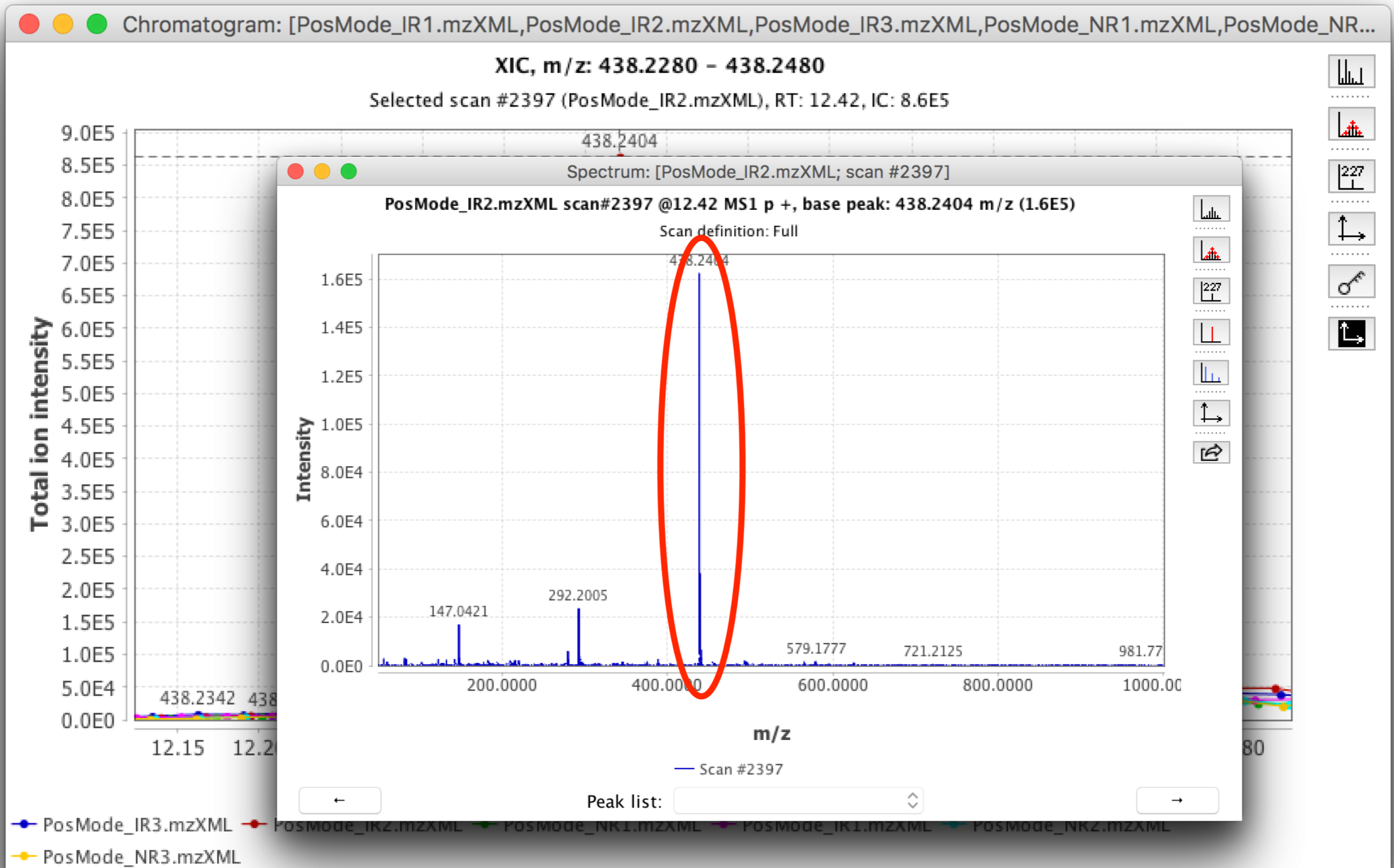
MetFrag



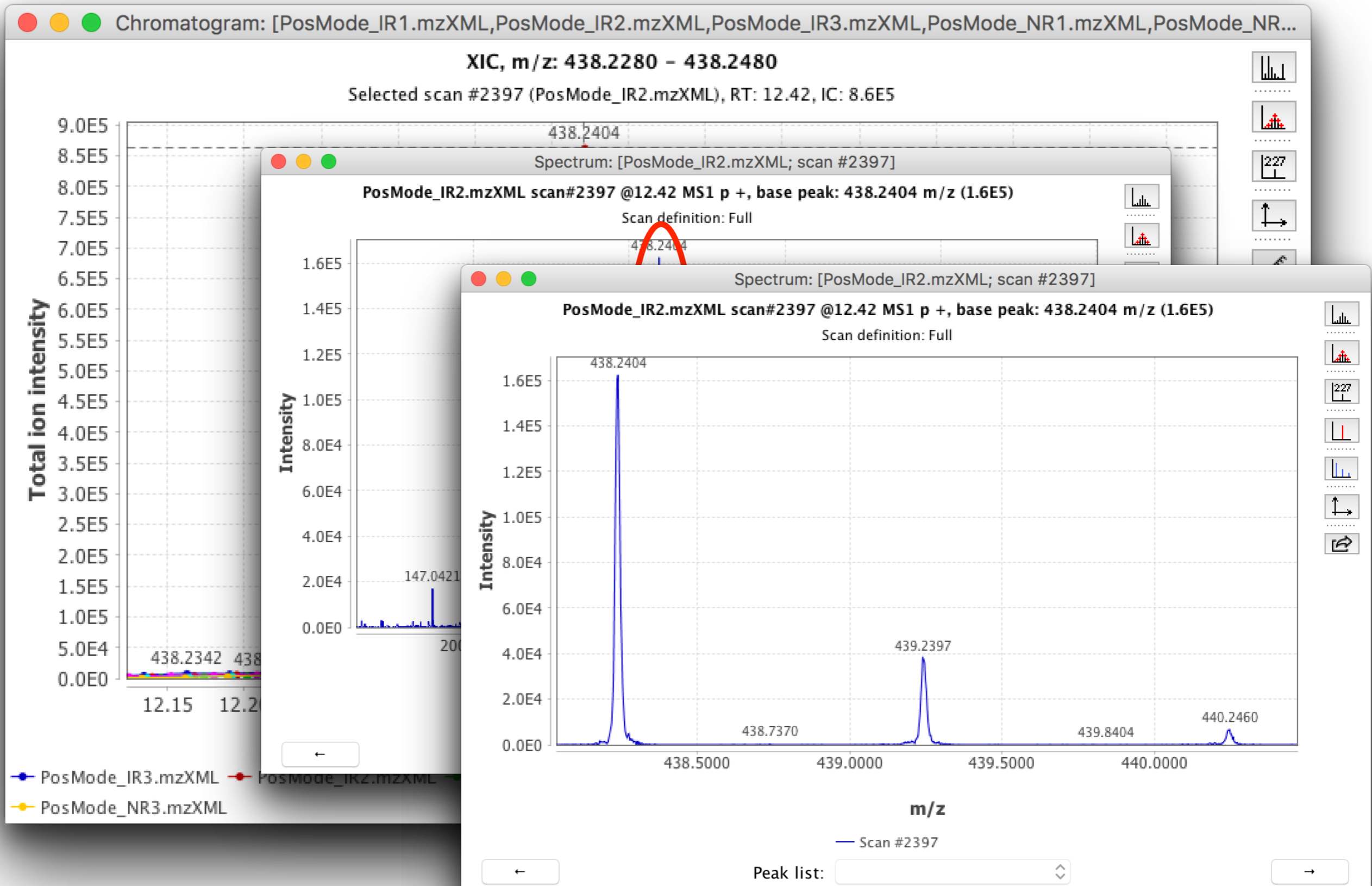
More on identification



More on identification



More on identification



More on identification

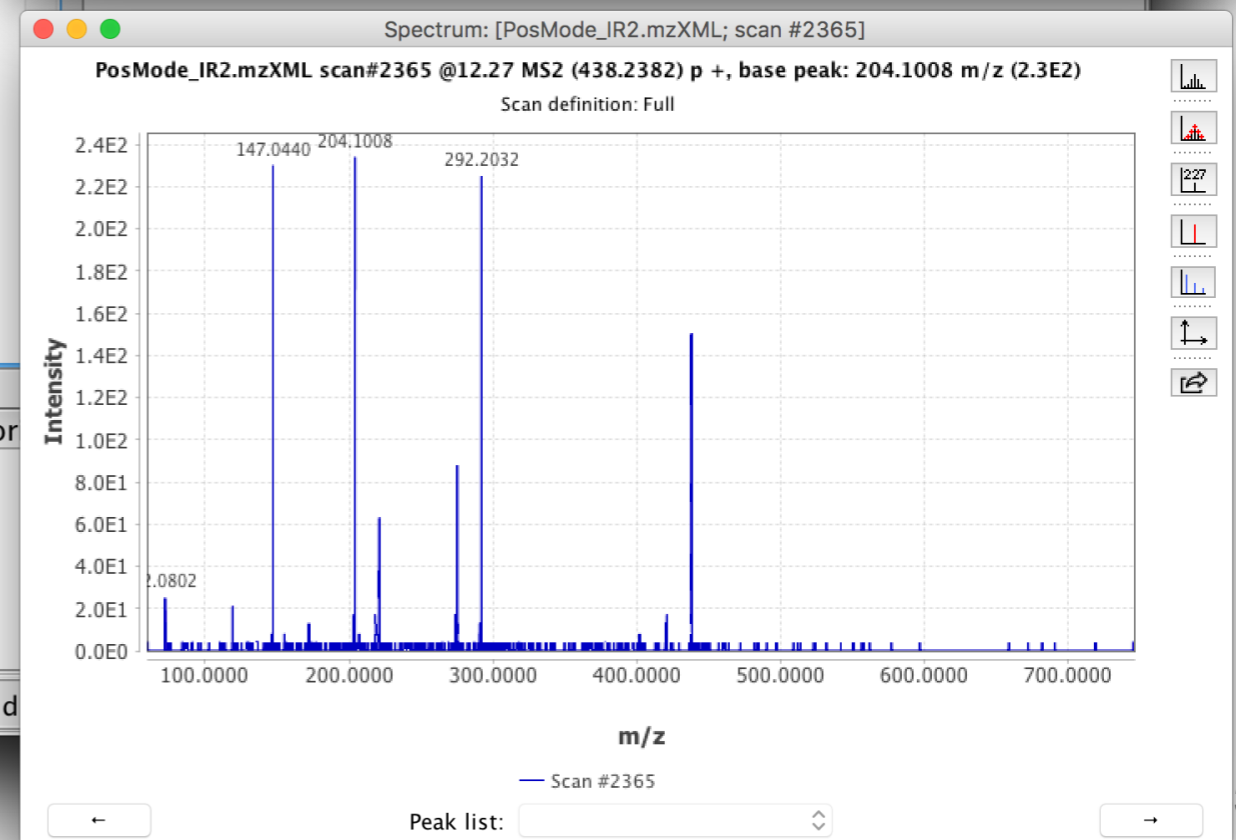
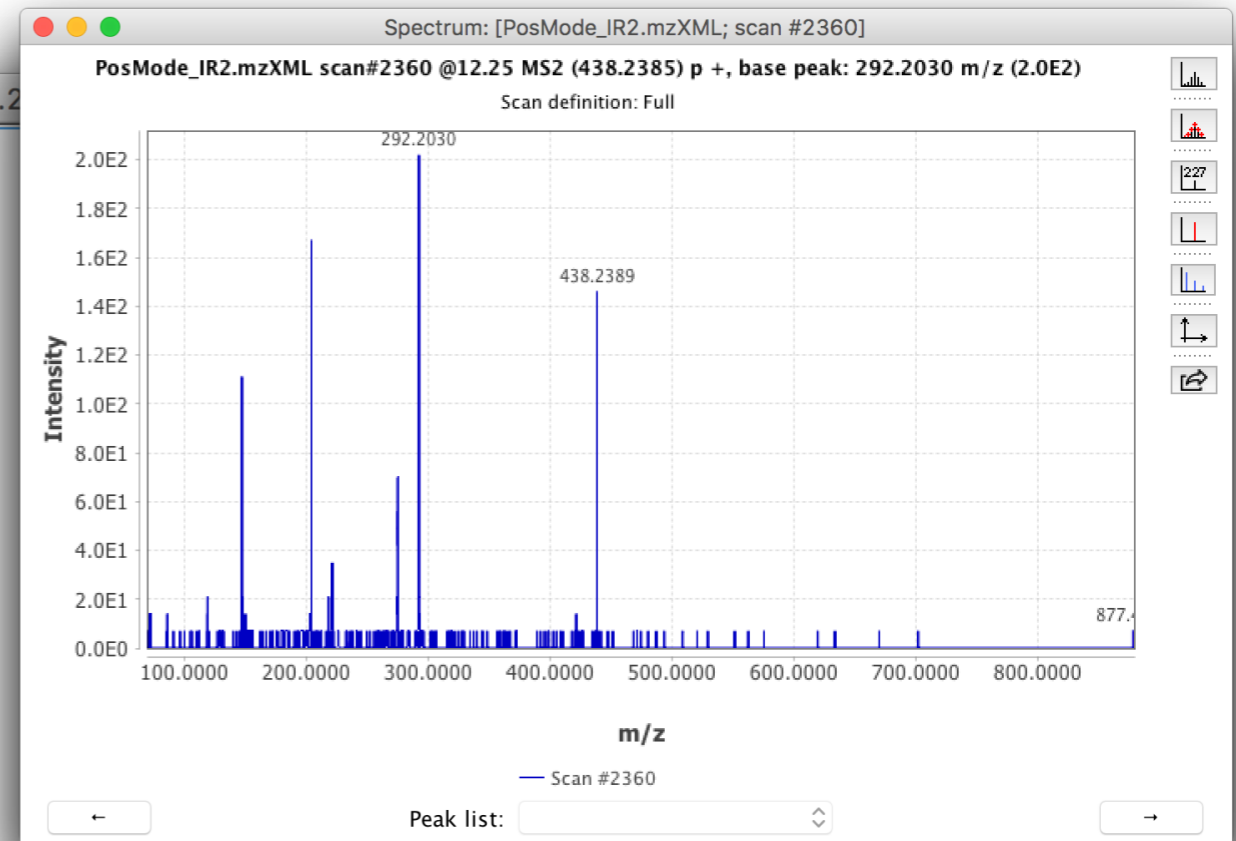
MZmine 2.2

- #2350 @12.21 MS2 (283.1066) p +
- #2351 @12.21 MS2 (541.2248) p +
- #2352 @12.22 MS2 (703.1956) p +
- #2353 @12.22 MS1 p +
- #2354 @12.23 MS2 (147.0314) p +
- #2355 @12.23 MS2 (169.0855) p +
- #2356 @12.24 MS2 (278.1413) p +
- #2357 @12.24 MS2 (279.0987) c +
- #2358 @12.24 MS2 (283.1070) p +
- #2359 @12.24 MS2 (428.2022) p +
- #2360 @12.25 MS2 (438.2385) p +
- #2361 @12.25 MS1 p +
- #2362 @12.26 MS2 (147.0296) p +
- #2363 @12.27 MS2 (387.2014) p +
- #2364 @12.27 MS2 (428.2038) p +
- #2365 @12.27 MS2 (438.2382) p +
- #2366 @12.28 MS1 p +
- #2367 @12.29 MS2 (209.0778) p +
- #2368 @12.30 MS2 (265.1169) p +
- #2369 @12.30 MS2 (387.2012) p +
- #2370 @12.31 MS1 p +
- #2371 @12.32 MS2 (209.0780) p +
- #2372 @12.32 MS2 (345.1436) p +
- #2373 @12.33 MS2 (373.1389) p +
- #2374 @12.33 MS2 (681.1649) p +
- #2375 @12.34 MS1 p +
- #2376 @12.35 MS2 (337.7160) p +

Tasks in progress...

Item	Prior
------	-------

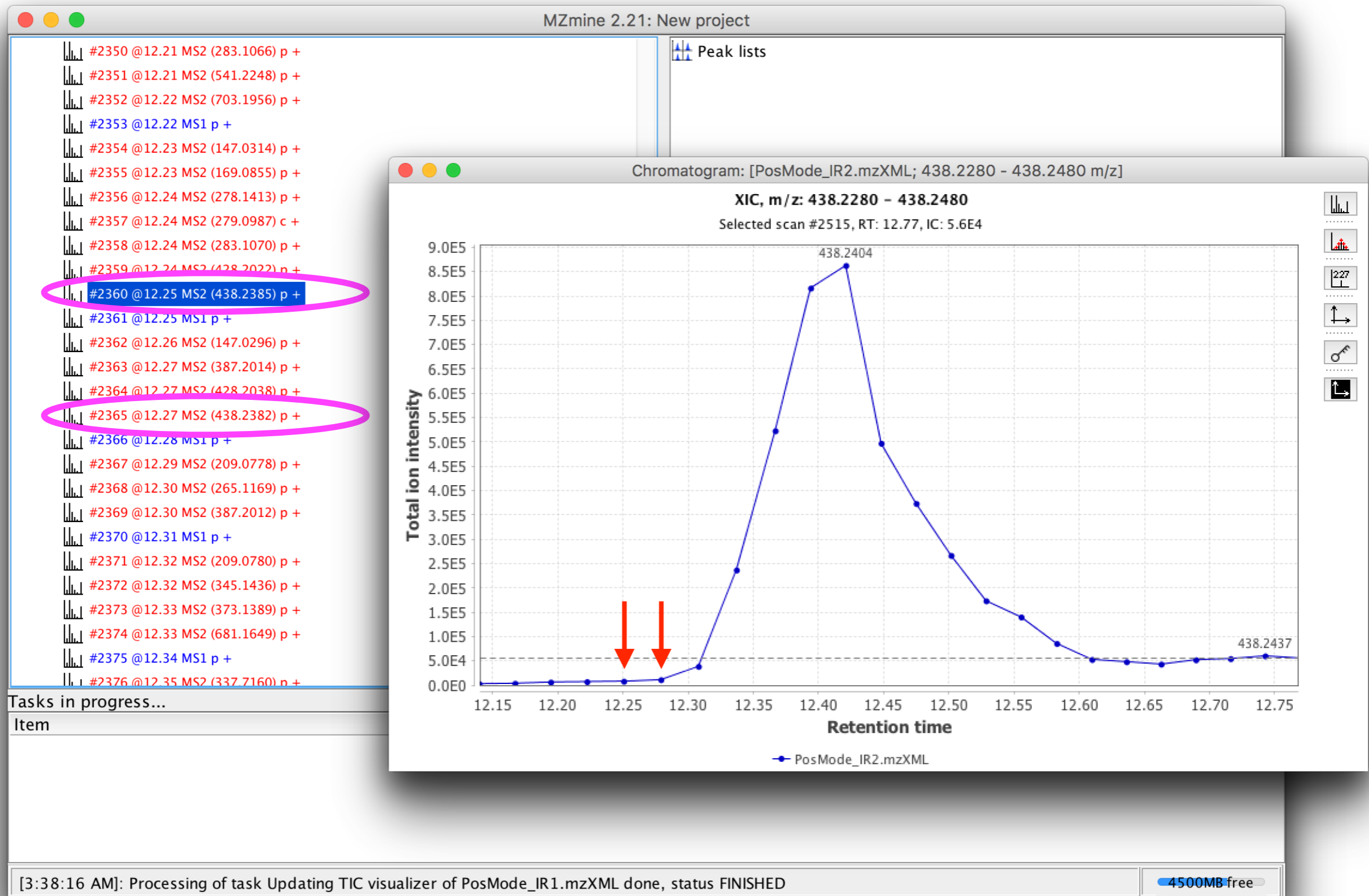
[3:38:16 AM]: Processing of task Updating TIC visualizer of PosMode_IR1.mzXML d



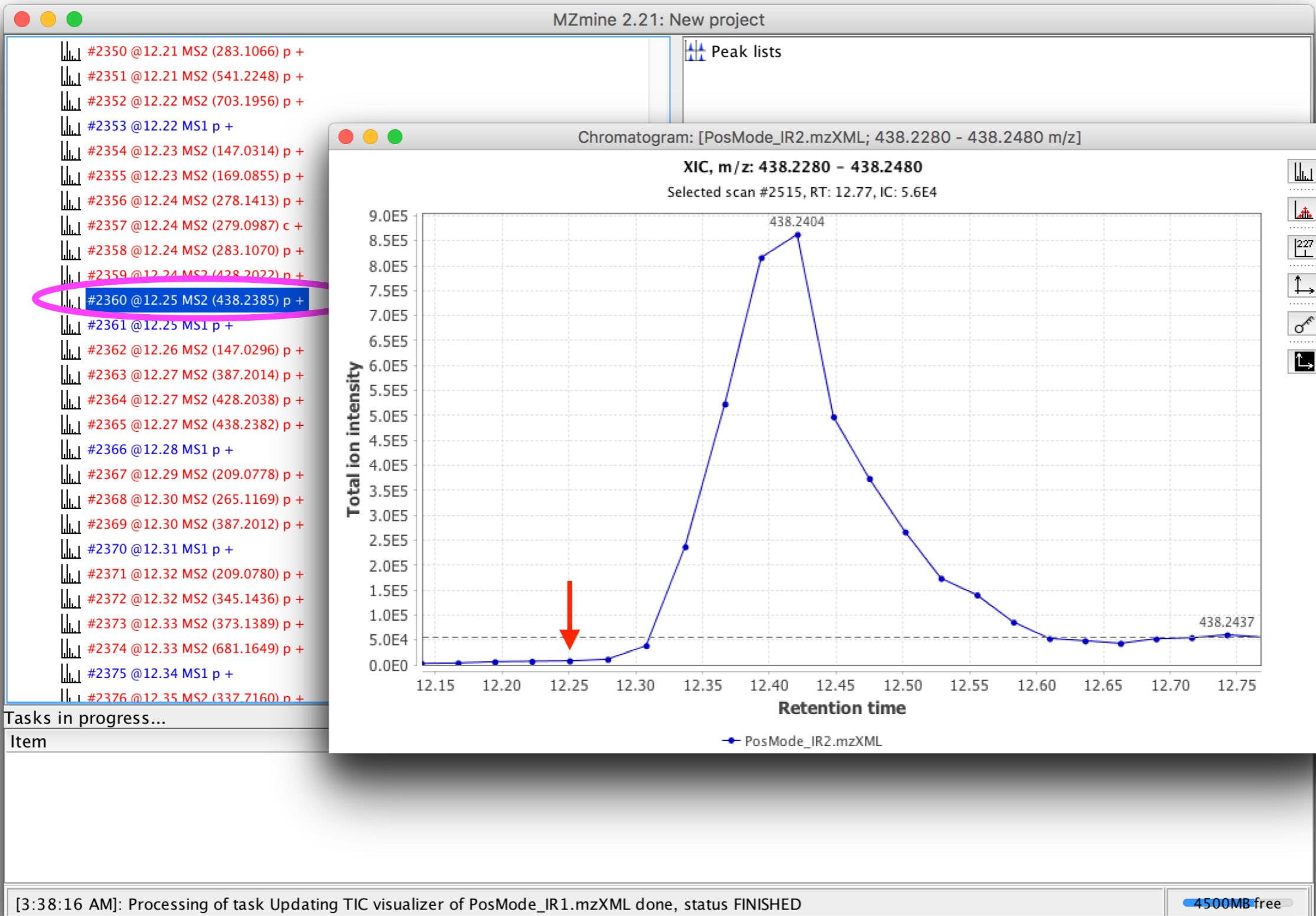
More on identification

- Information we have for identification of compounds based on MS/MS
 - M+H
 - Experimental isotopic identification
 - MS/MS

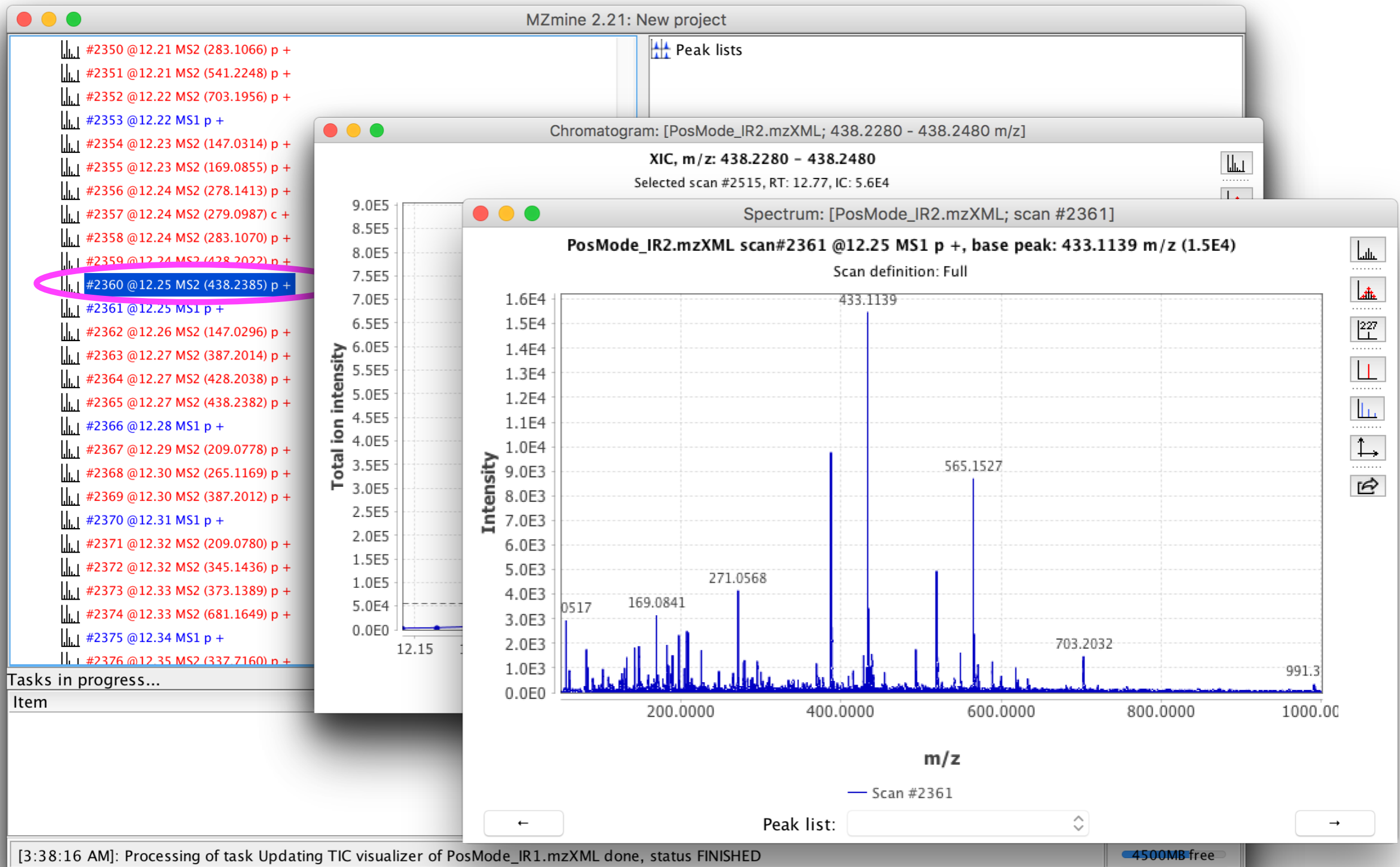
More on identification



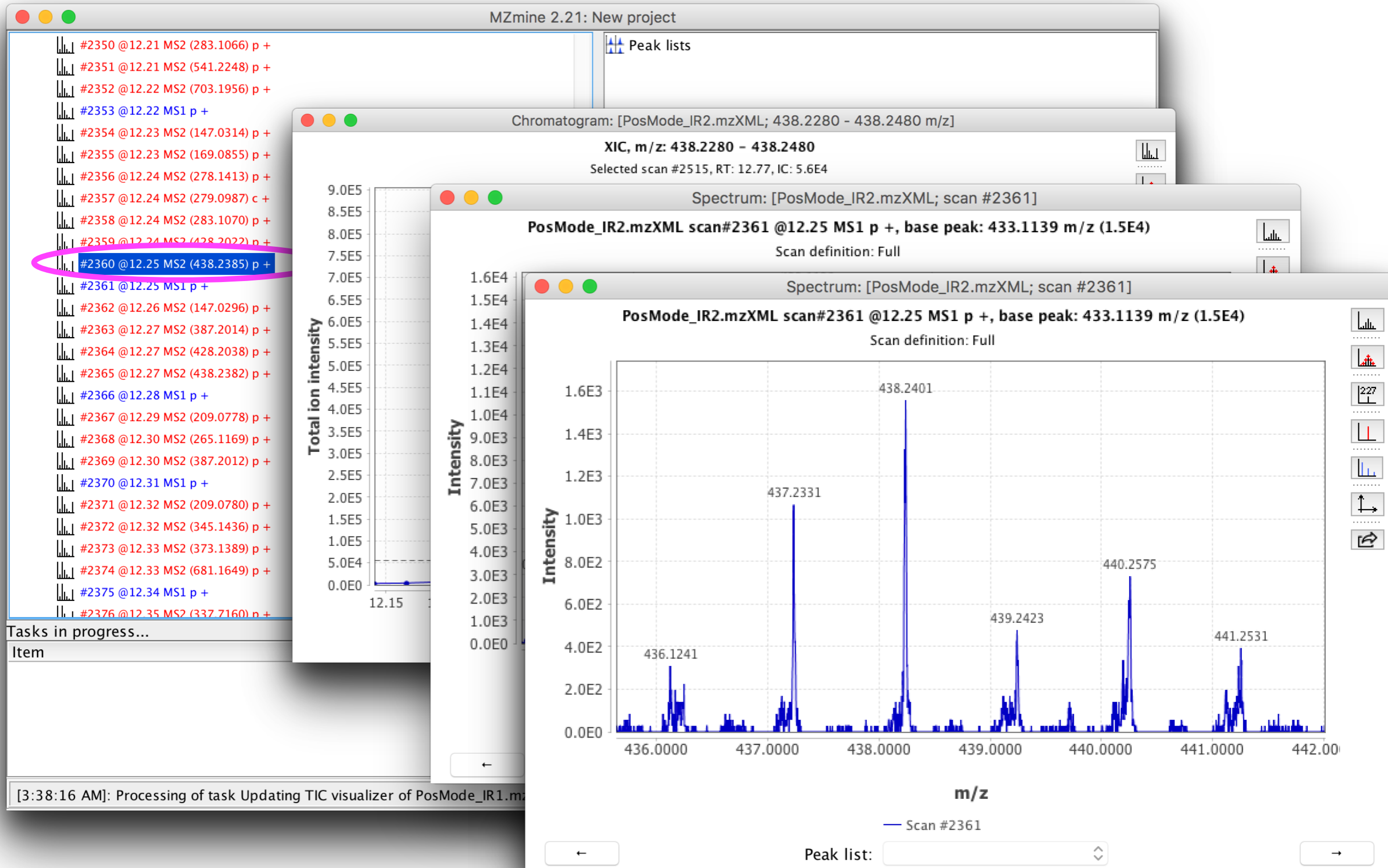
More on identification



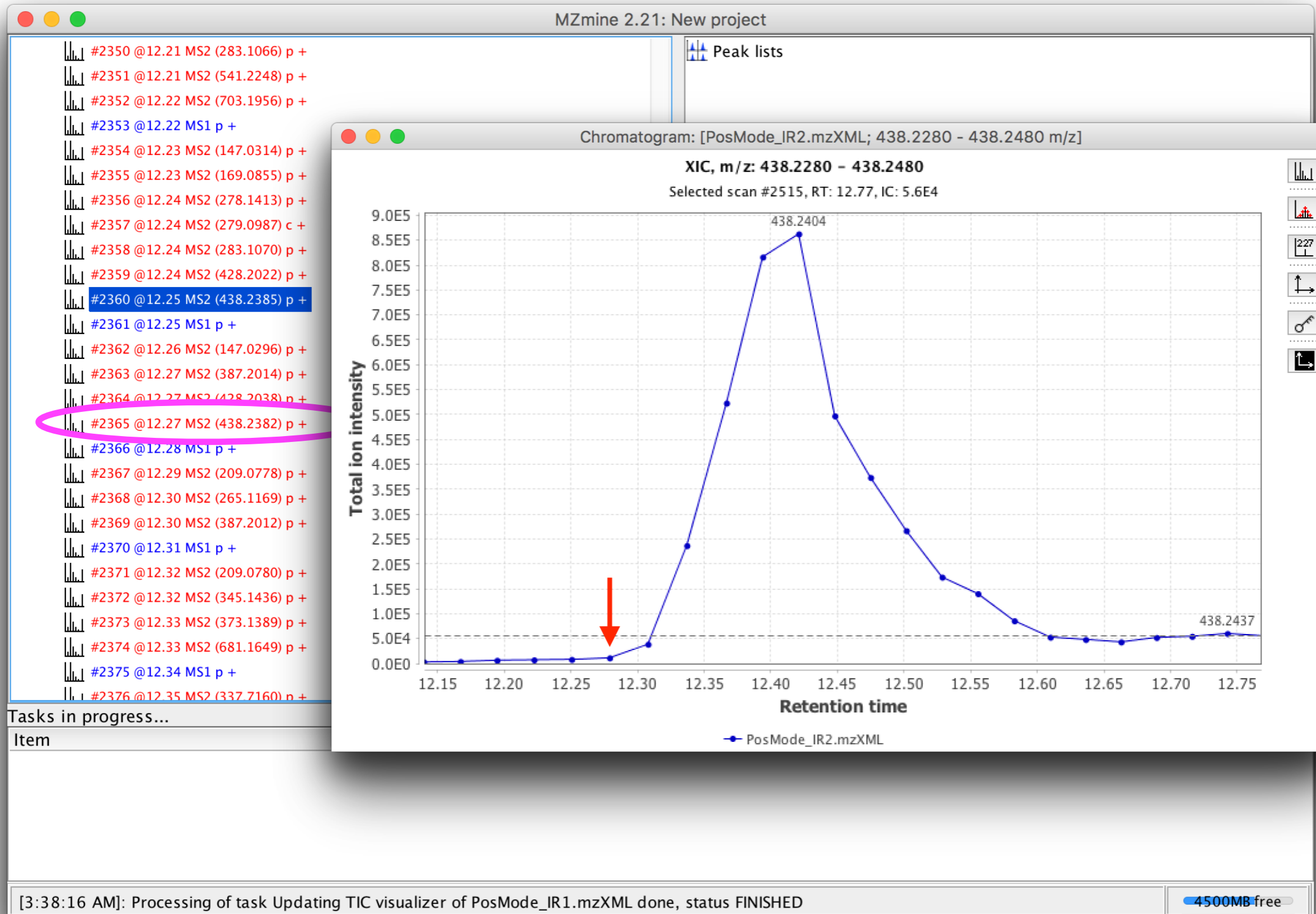
More on identification



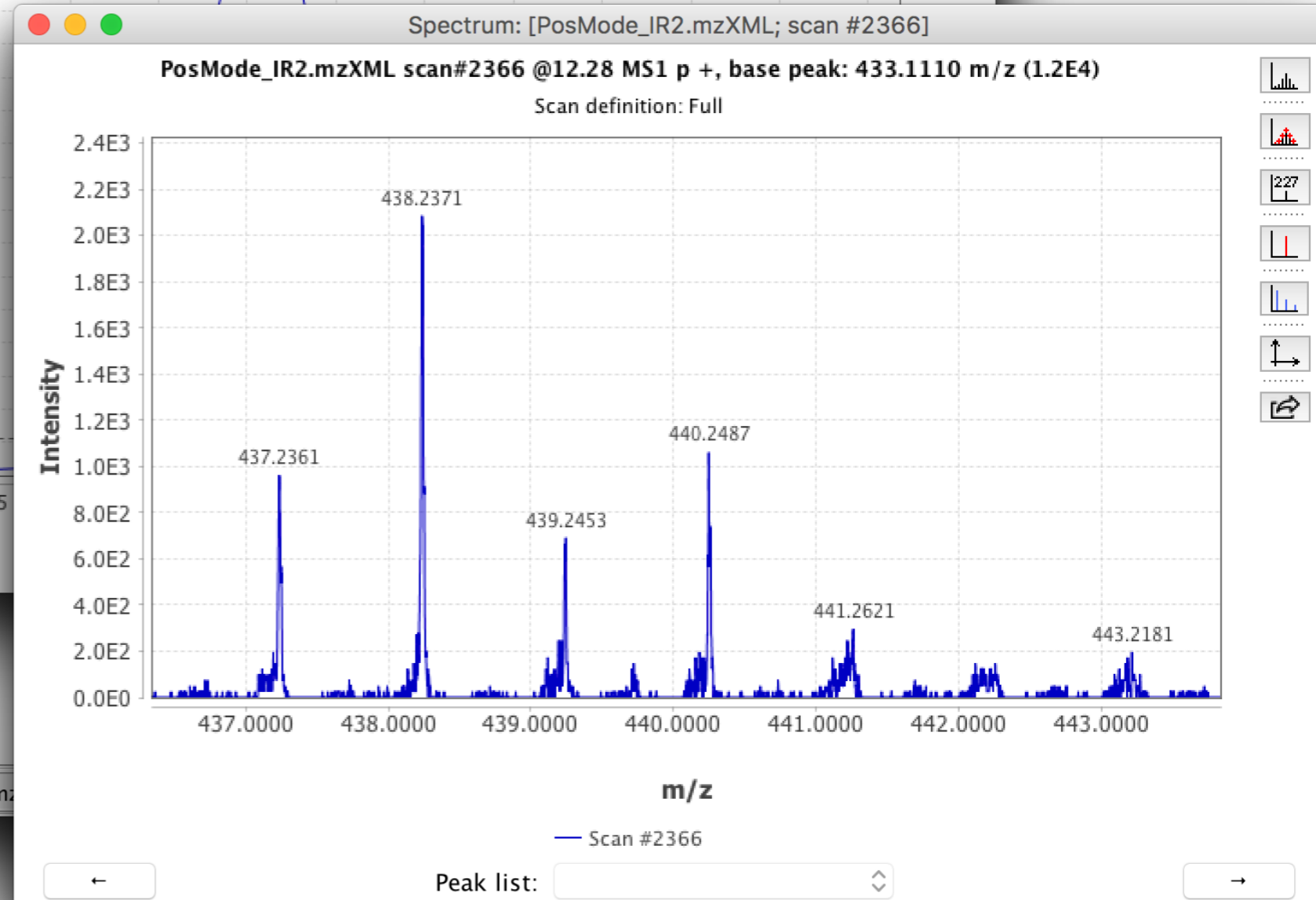
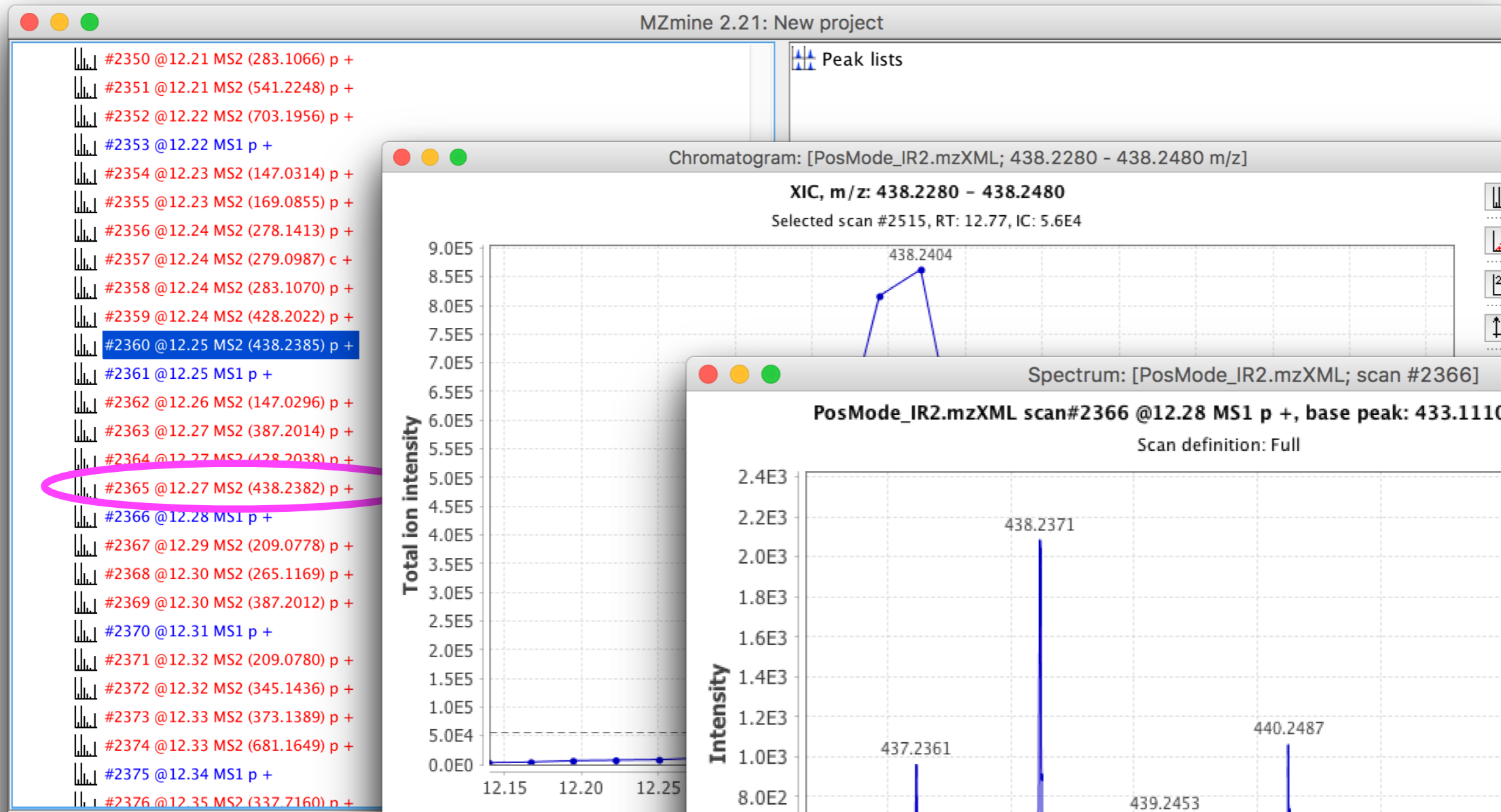
More on identification



More on identification



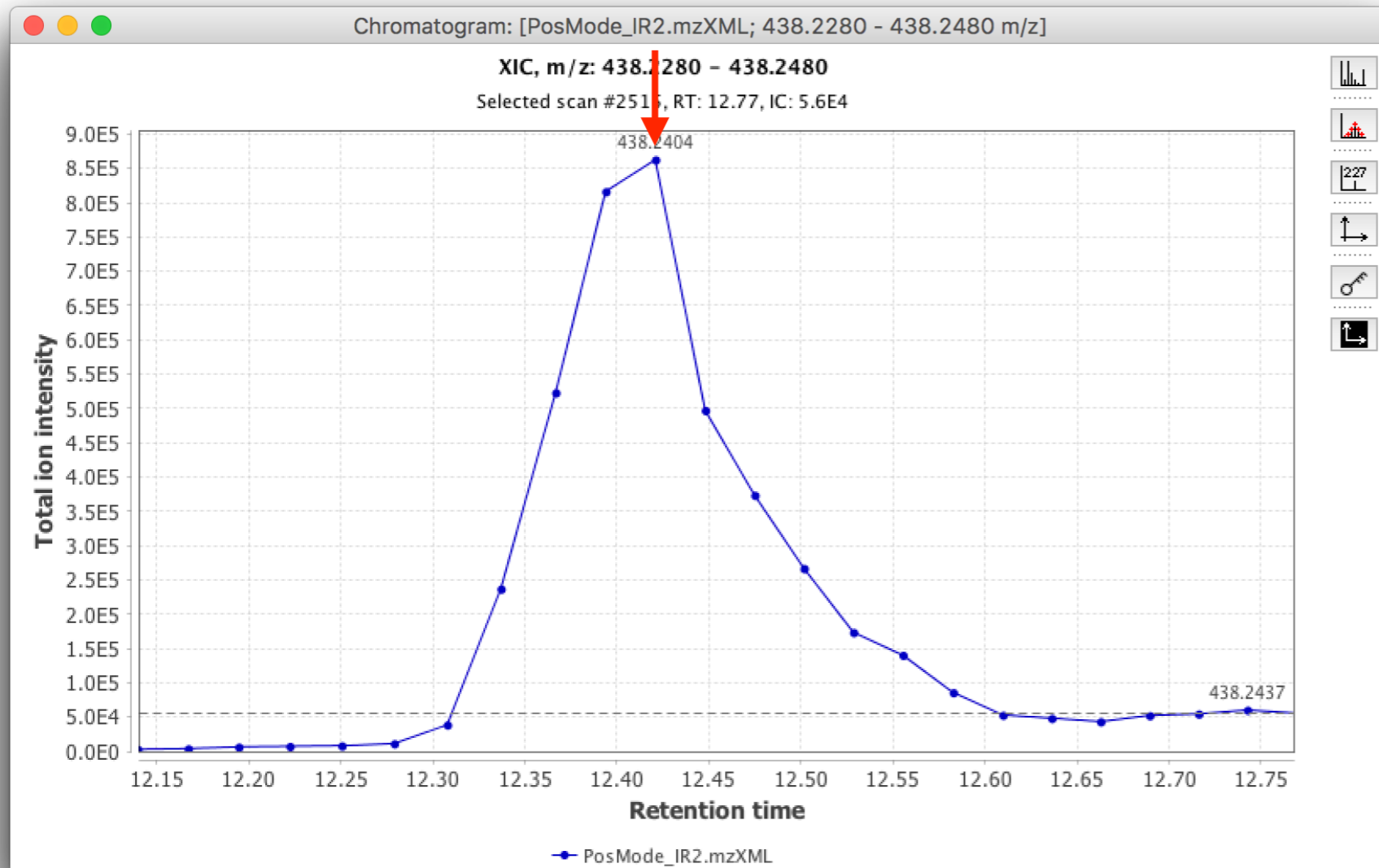
More on identification



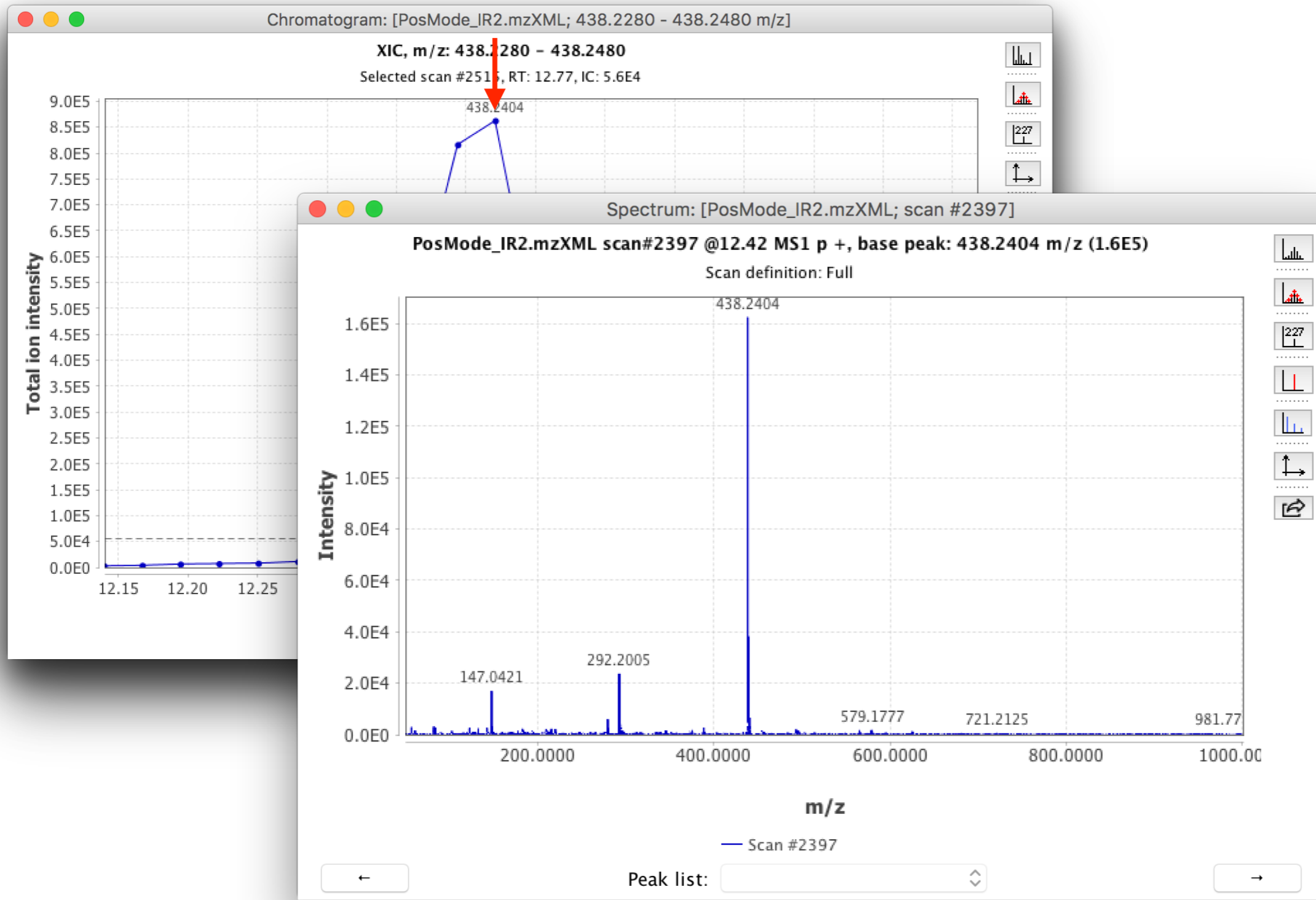
Tasks in progress...

Item
[3:38:16 AM]: Processing of task Updating TIC visualizer of PosMode_IR1.mz

More on identification



More on identification




More on identification



More on identification

← → ↻ https://metlin.scripps.edu/metabo_search_alt2.php



Scripps Center for Metabolomics

[MS HOME](#) [METLIN](#) [XCMS Online](#) [XCMS Institute](#) [XCM](#)

METLIN: Metabolite Search

Simple

[Simple \(Saved Searches\)](#) | [Advanced](#) | [Batch](#) | [Fragment](#) | [Neutral Loss](#)

Mass:

Tolerance (±):

Charge:

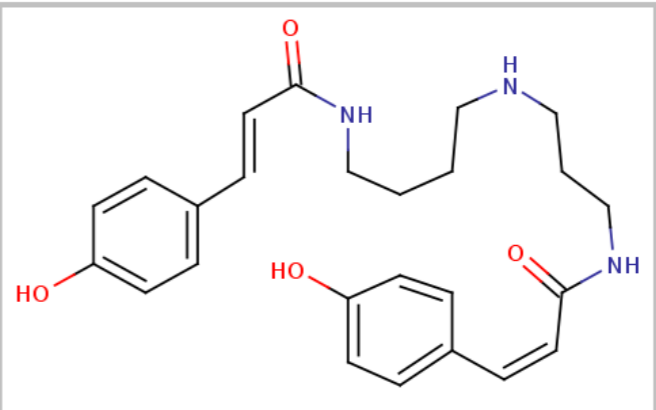
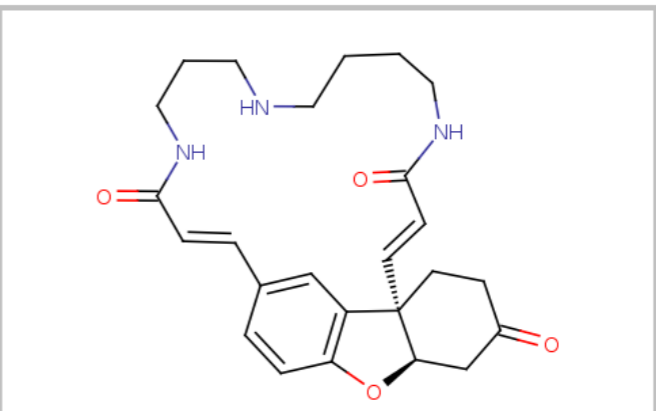
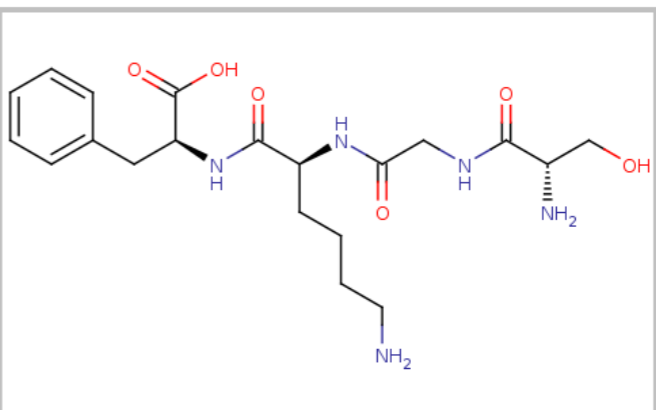
Neutral	M+H
Positive	M+NH4
Negative	M+Na
	M+H-2H2O
	M+H-H2O
	M+K
	M+ACN+H
	M+ACN+Na
	M+2Na-H
	M+2H
	M+3H
	M+H+Na
	M+2H+Na
	M+2Na
	M+2Na+H
	M+Li
	M+CH3OH+H

•To select multiple Adducts:
- Hit Ctrl + Adducts
- Hit Command + Adducts
Select: **all** | **none**

Remove peptides from search:

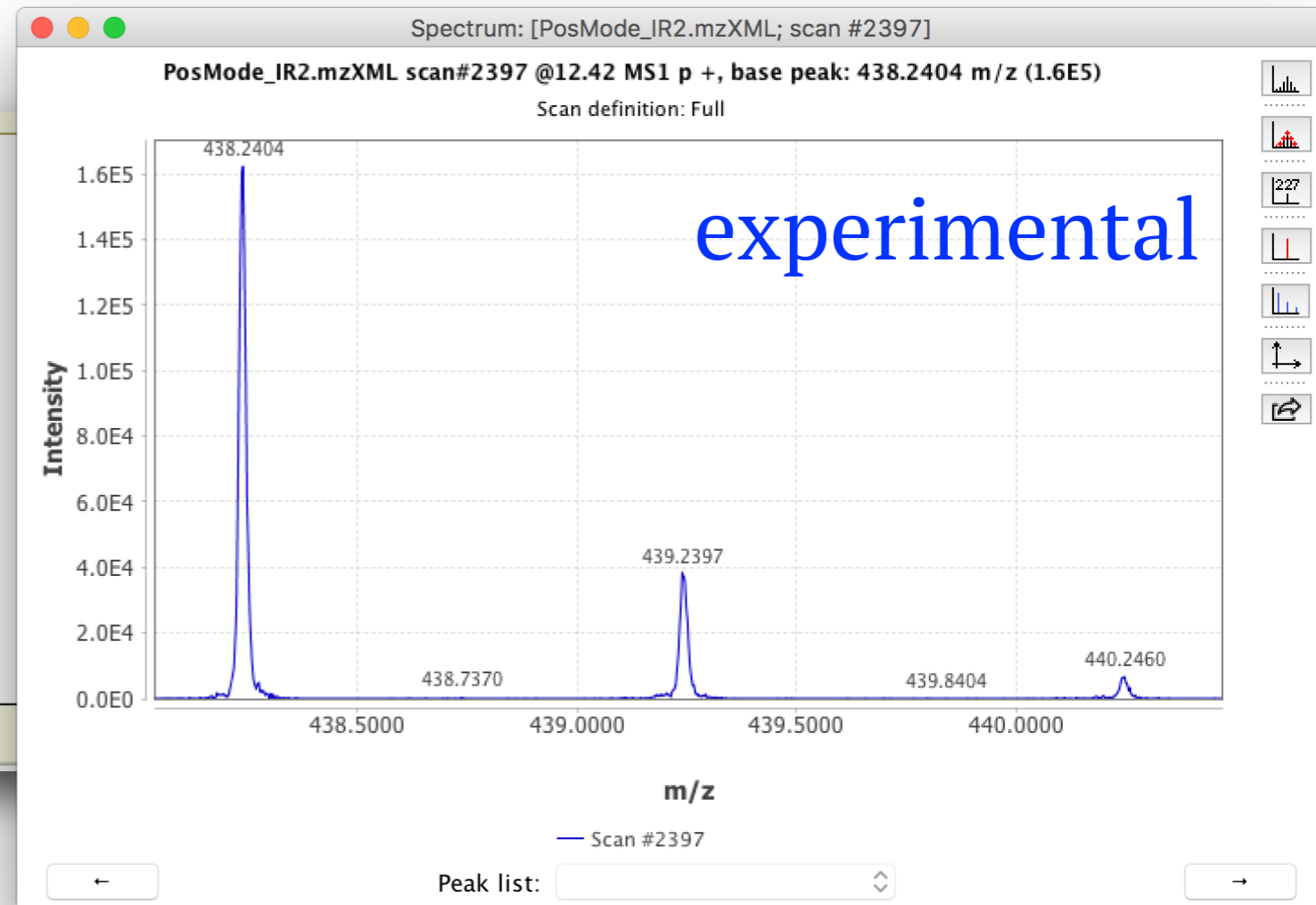
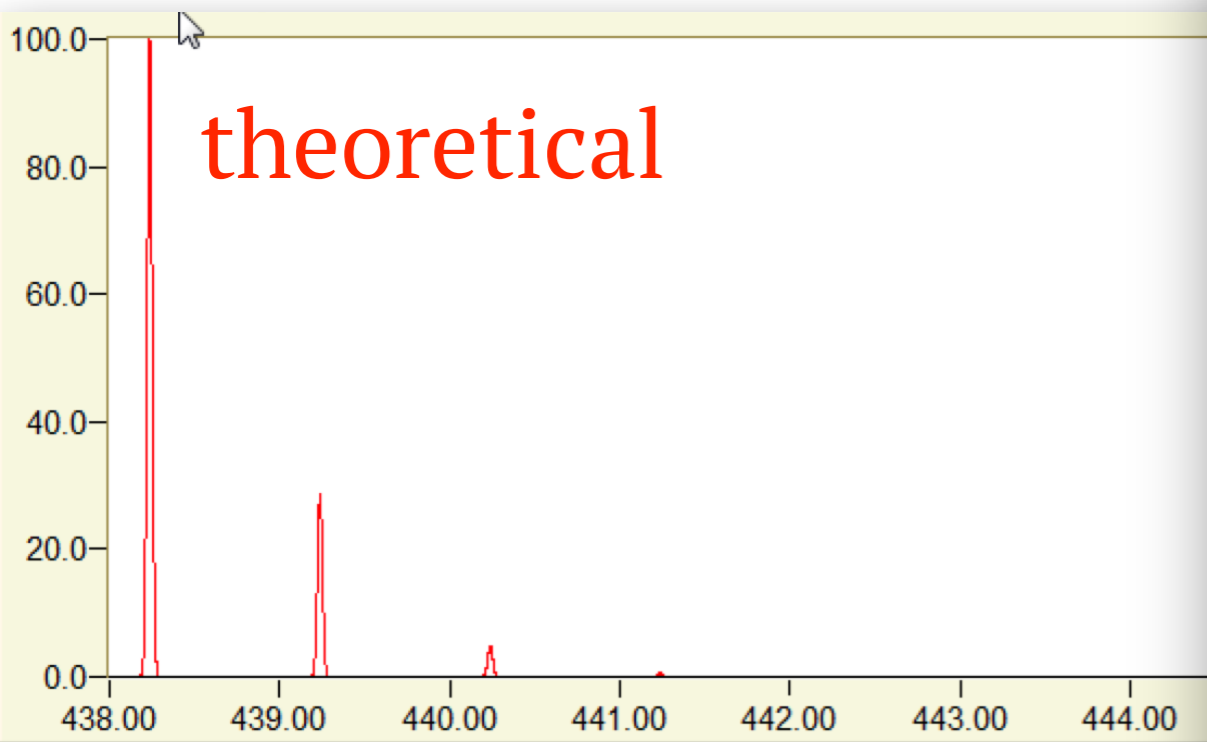
More on identification

Total: 87 Metabolites

METLIN ID	MASS	Δ ppm	NAME	MS/MS	STRUCTURE
89296 <input type="checkbox"/>	[M+H] ⁺ <u>m/z</u> 438.2387 M 437.2315	1	N1,N10- Dicoumaroylspermidine <i>Formula: C₂₅H₃₁N₃O₄</i> <i>CAS: 65715-79-9</i>	NO	 The structure shows a spermidine chain (a 9-membered ring with three nitrogen atoms) substituted with two coumaroyl groups. Each coumaroyl group consists of a coumarin ring system attached to a propenoic acid chain.
43760 <input type="checkbox"/>	[M+H] ⁺ <u>m/z</u> 438.2387 M 437.2315	1	Lunarine <i>Formula: C₂₅H₃₁N₃O₄</i> <i>CAS: 24185-51-1</i>	View	 The structure shows a complex polycyclic molecule with a central benzene ring fused to a five-membered ring containing an oxygen atom. It is substituted with two coumaroyl groups and a long chain containing two nitrogen atoms.
225643 <input type="checkbox"/>	[M+H] ⁺ <u>m/z</u> 438.2347 M 437.2274	7	Ser Gly Lys Phe <i>Formula: C₂₀H₃₁N₅O₆</i> <i>CAS:</i>	NO	 The structure shows a linear peptide chain consisting of four amino acids: Serine, Glycine, Lysine, and Phenylalanine. The Serine is at the C-terminus, followed by Glycine, Lysine, and Phenylalanine at the N-terminus.
225567 <input type="checkbox"/>	[M+H] ⁺	7	Ser Gly Phe Lys	NO	

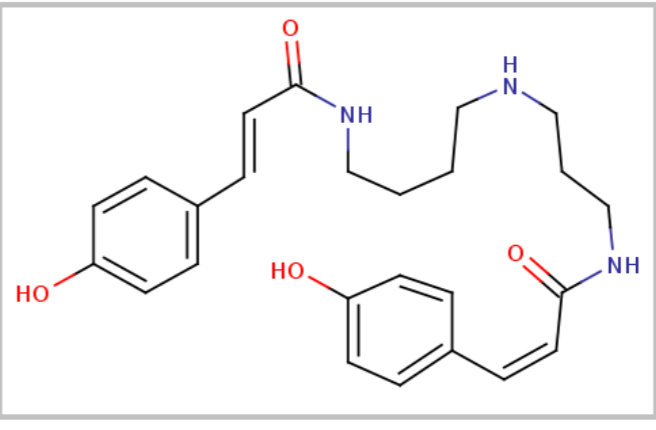
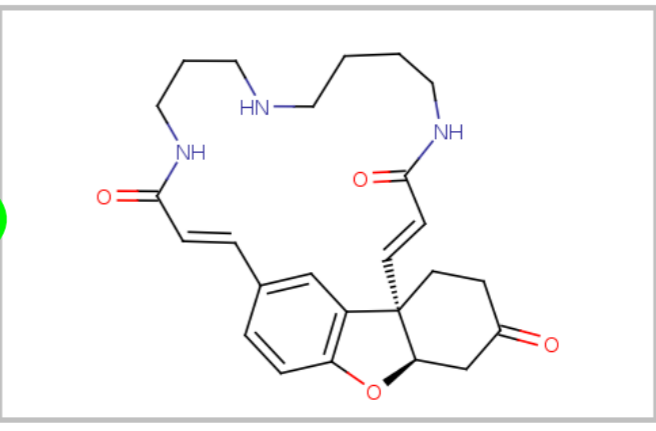
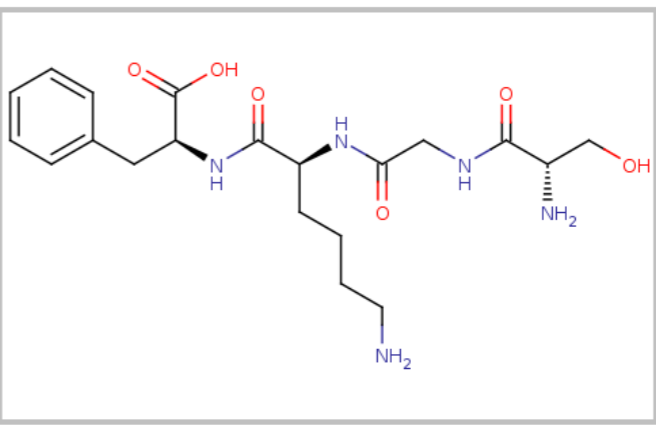
More on identification

- Compare isotopic distributions



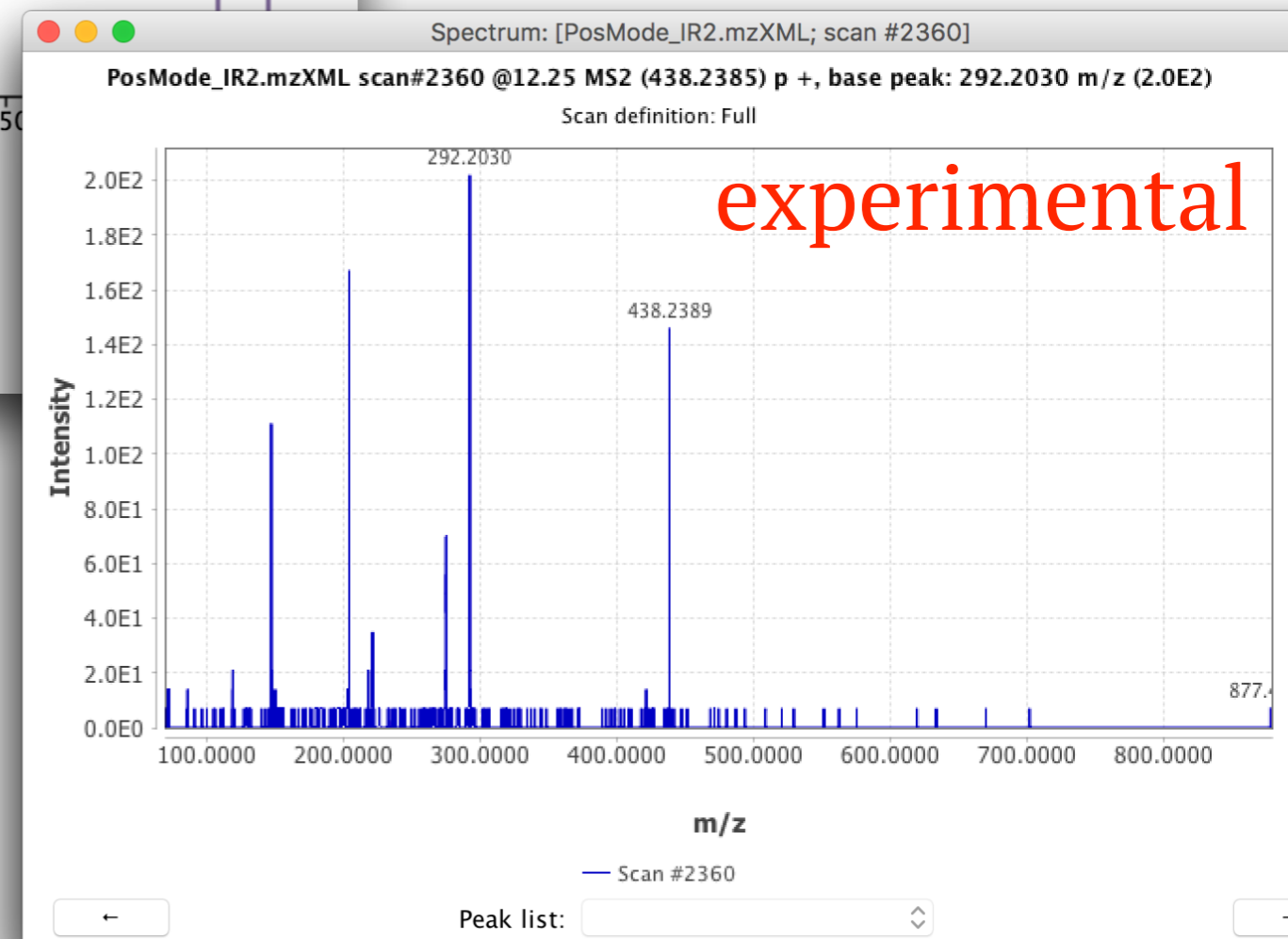
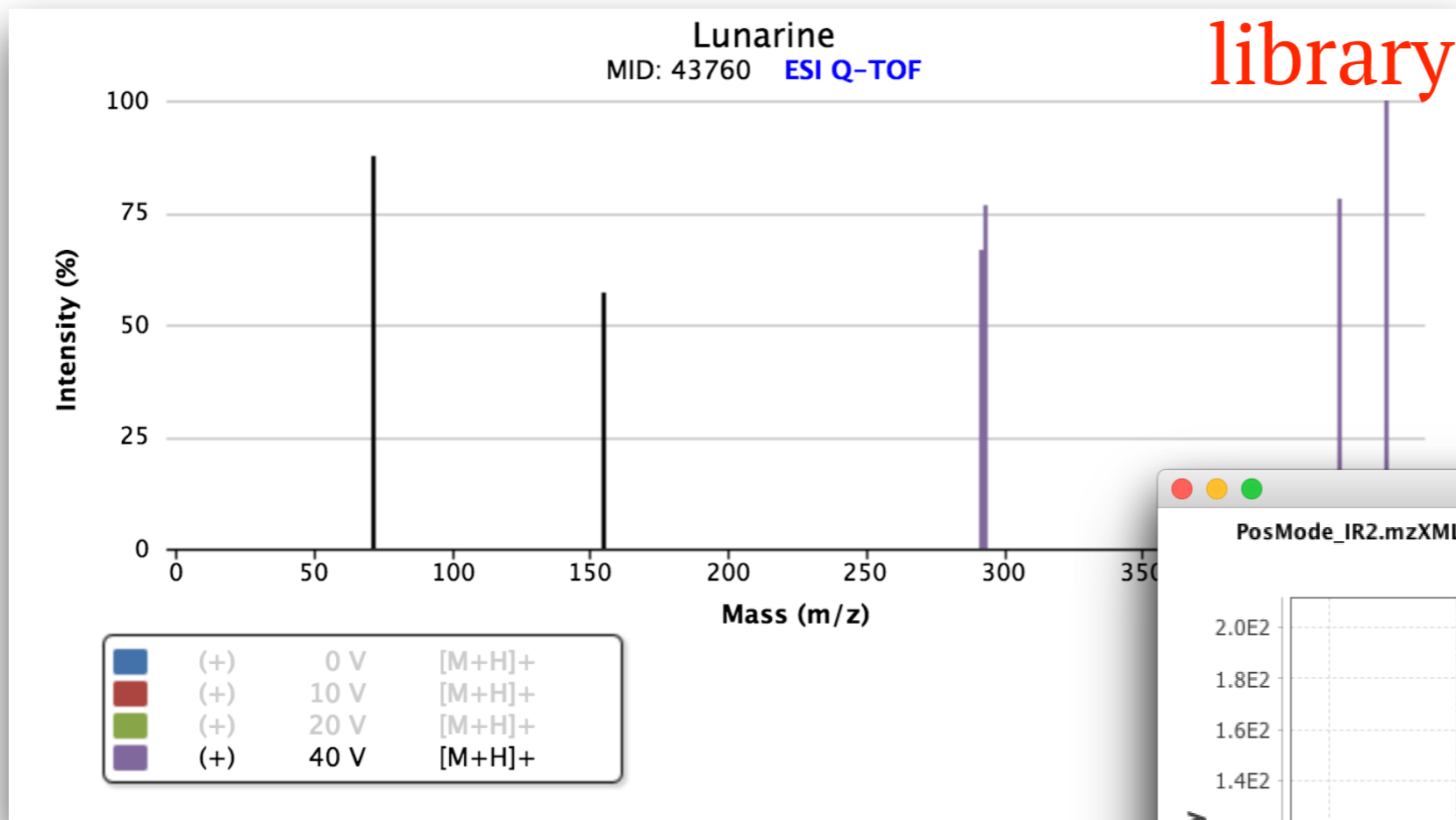
More on identification

Total: 87 Metabolites

METLIN ID	MASS	Δ ppm	NAME	MS/MS	STRUCTURE
89296 <input type="checkbox"/>	[M+H] ⁺ <u>m/z</u> 438.2387 M 437.2315	1	N1,N10- Dicoumaroylspermidine <i>Formula: C₂₅H₃₁N₃O₄</i> <i>CAS: 65715-79-9</i>	NO	
43760 <input type="checkbox"/>	[M+H] ⁺ <u>m/z</u> 438.2387 M 437.2315	1	Lunarine <i>Formula: C₂₅H₃₁N₃O₄</i> <i>CAS: 24185-51-1</i>	View	
225643 <input type="checkbox"/>	[M+H] ⁺ <u>m/z</u> 438.2347 M 437.2274	7	Ser Gly Lys Phe <i>Formula: C₂₀H₃₁N₅O₆</i> <i>CAS:</i>	NO	
225567 <input type="checkbox"/>	[M+H] ⁺	7	Ser Gly Phe Lys	NO	

More on identification

- CompareMS/MS



Thank you!