LC-MS Data Pre-Processing

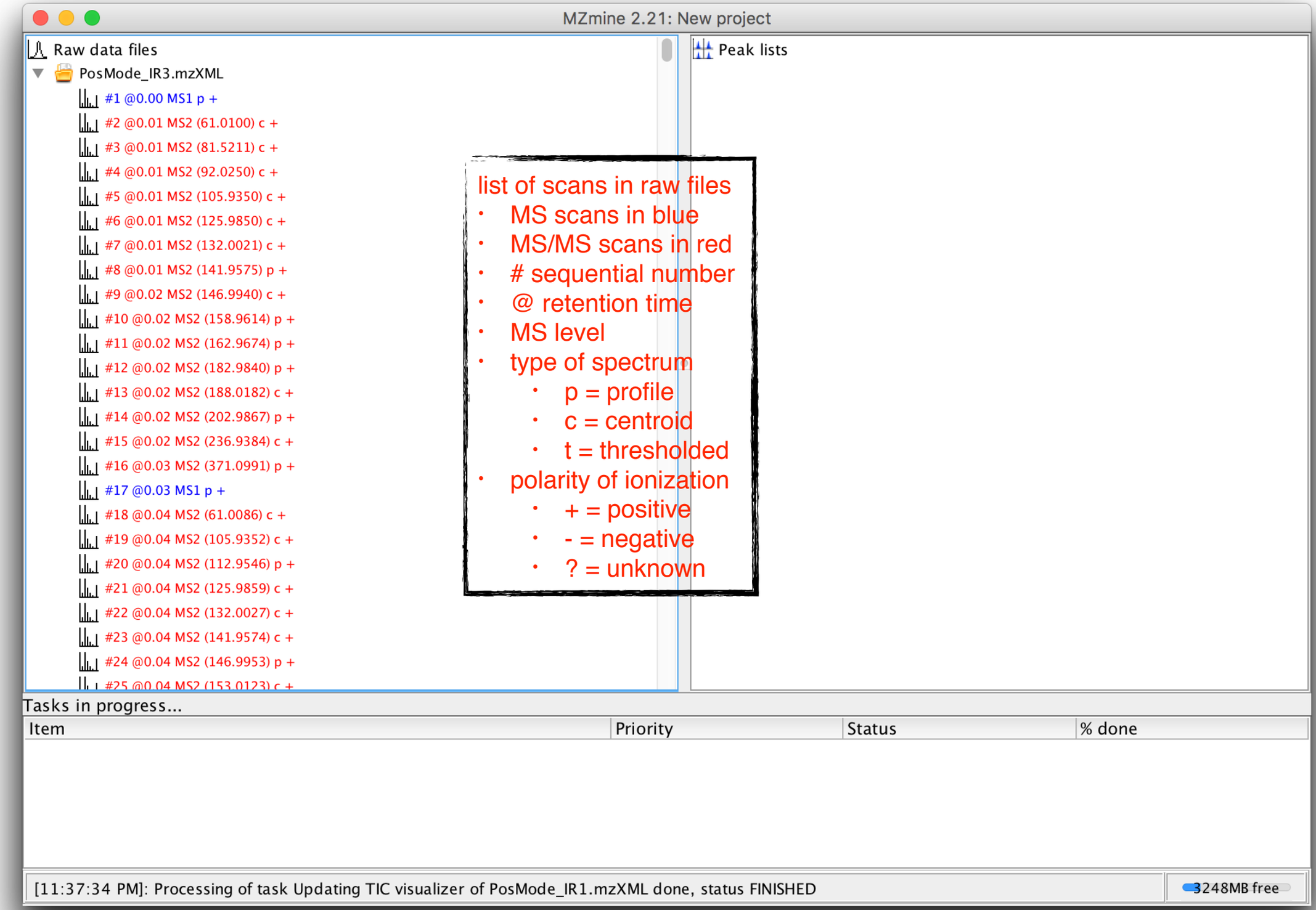
Xiuxia Du, Ph.D.
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte

Outline

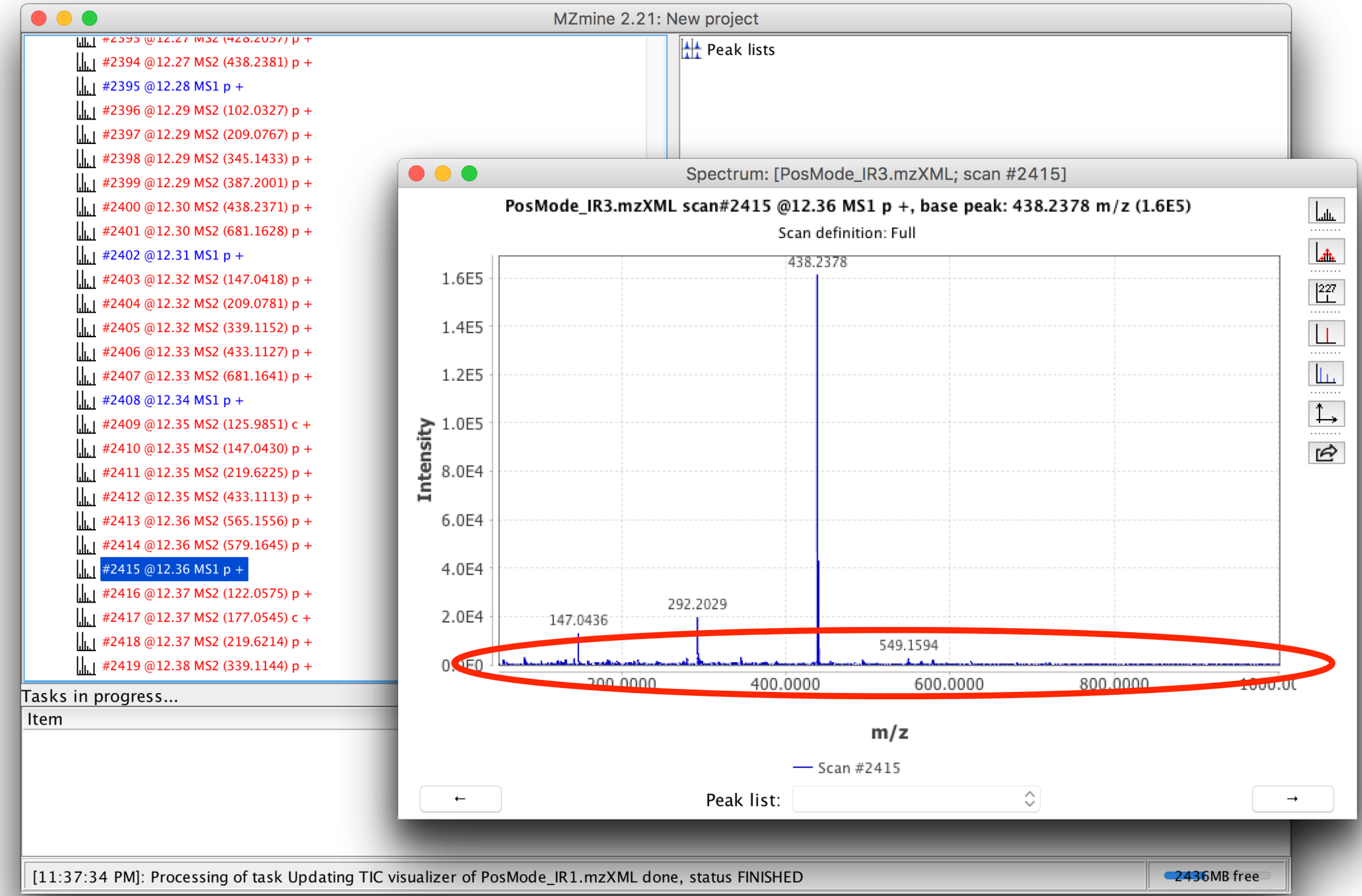
- Raw LC-MS data
  - Profile and centroid data
  - Mass vs. retention time map
  - TIC
  - EIC
  - Feature
- Data pre-processing
  - Feature detection
  - Feature grouping
  - Feature alignment
- Feature identification

Raw LC-MS data

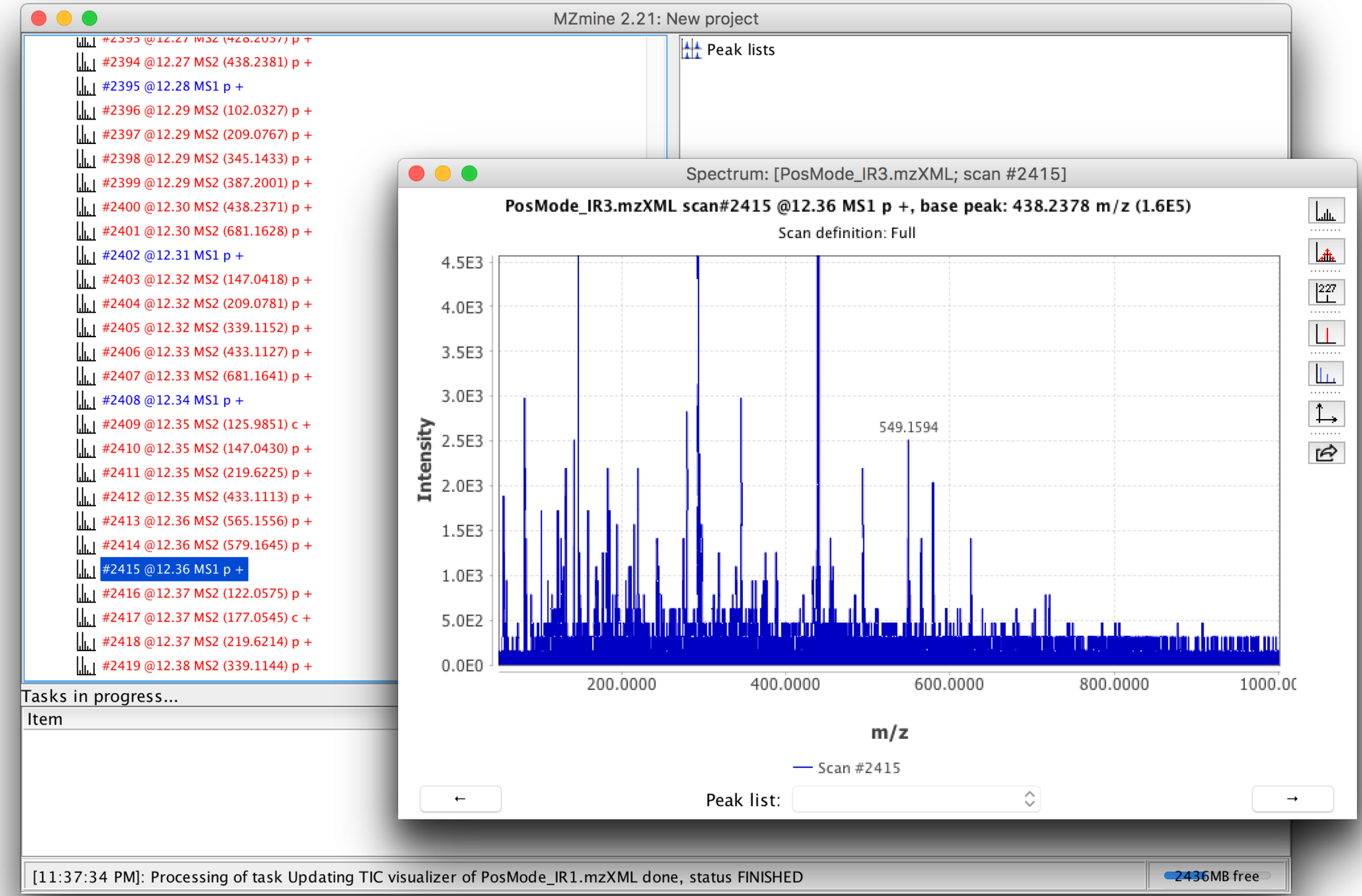
List of mass spectra



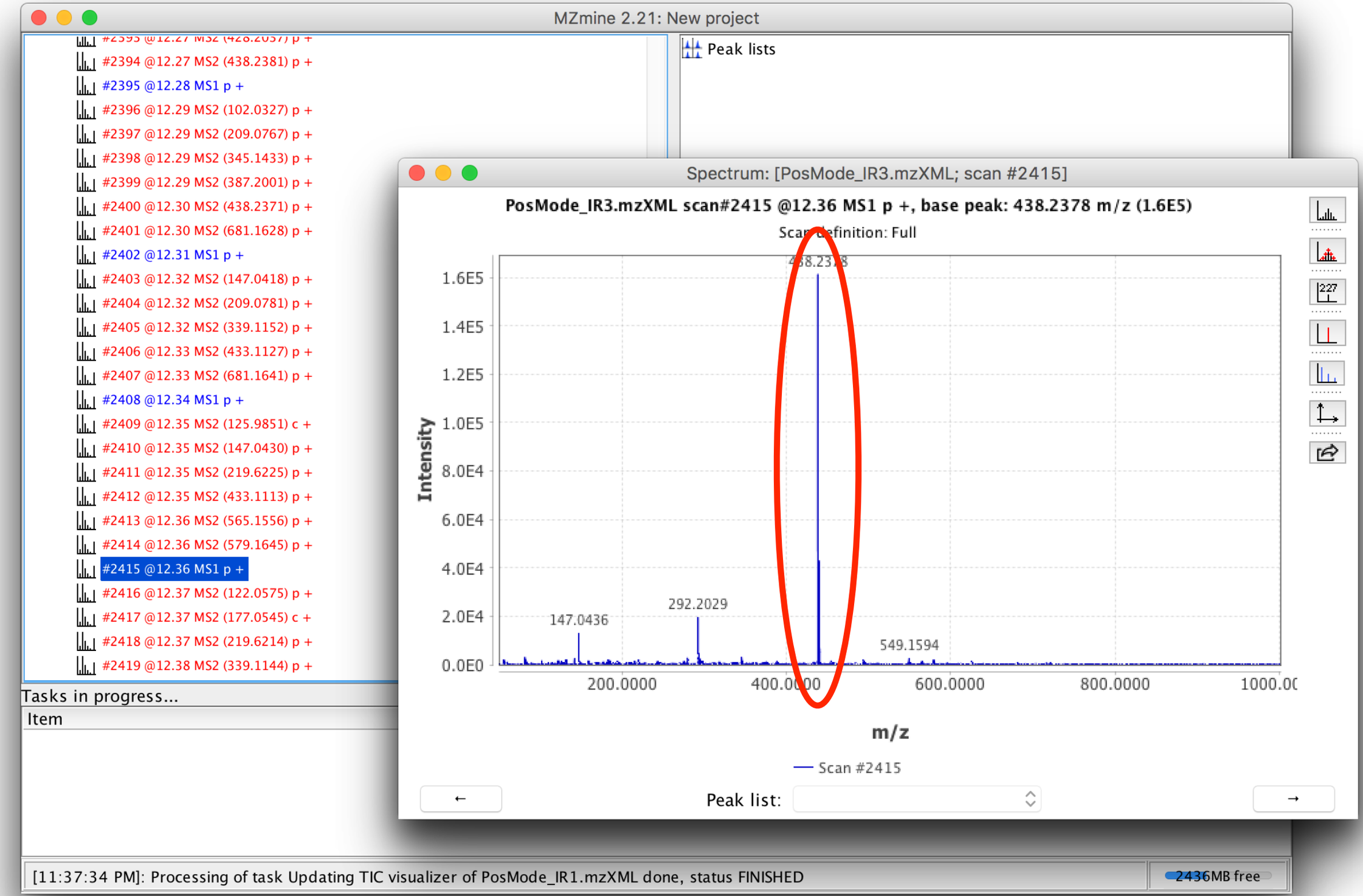
One mass spectrum



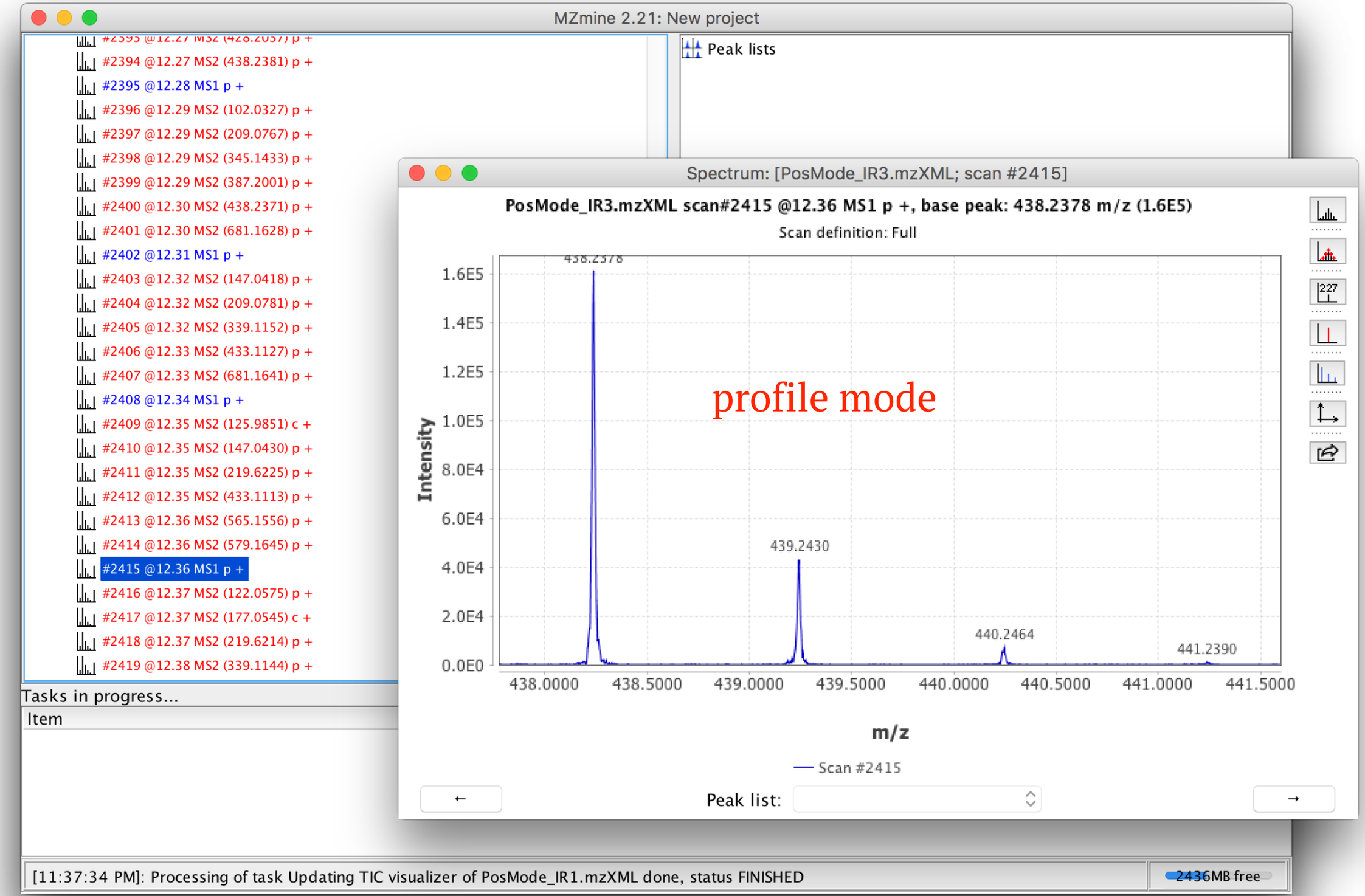
One mass spectrum



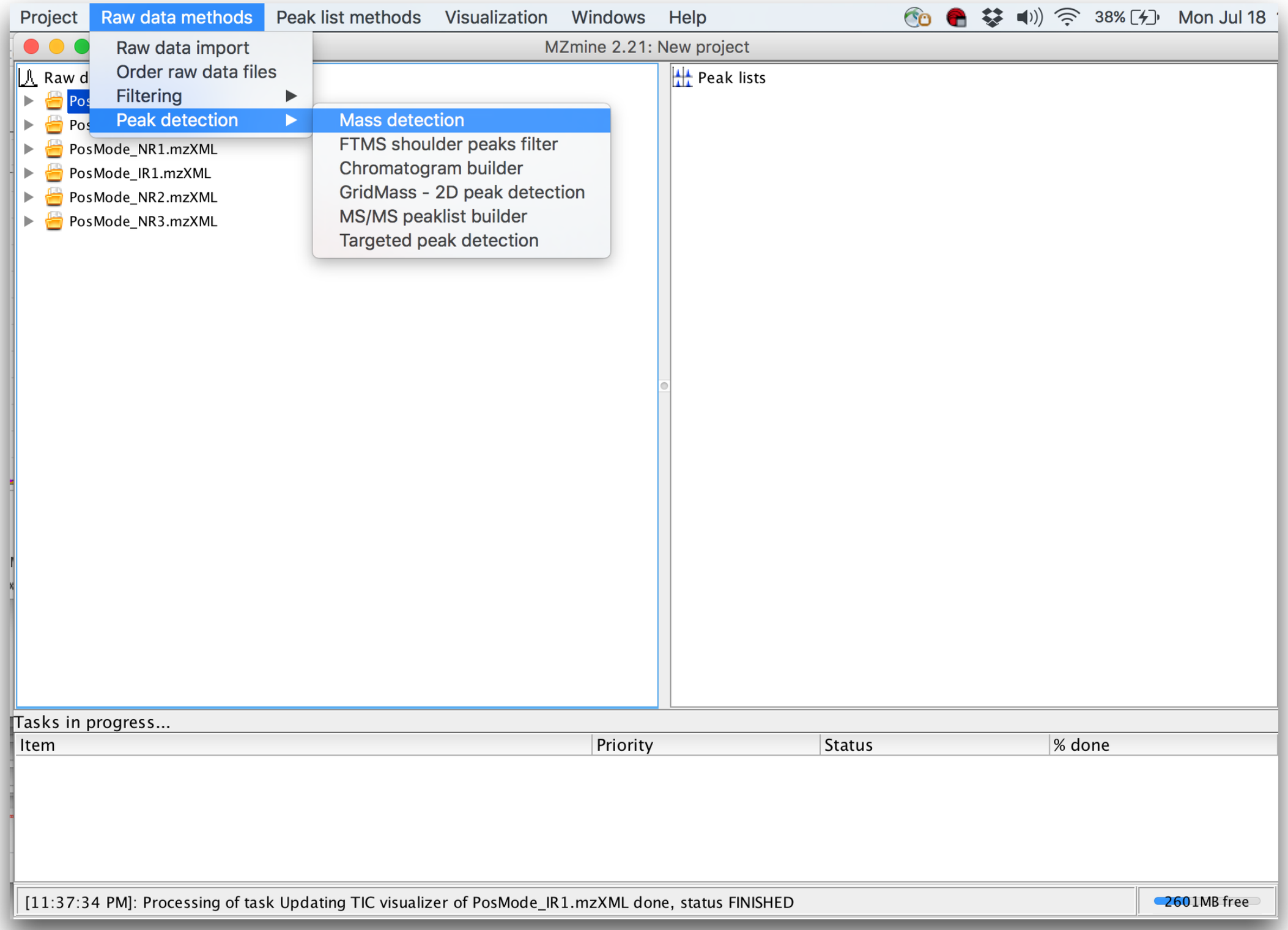
One mass spectrum



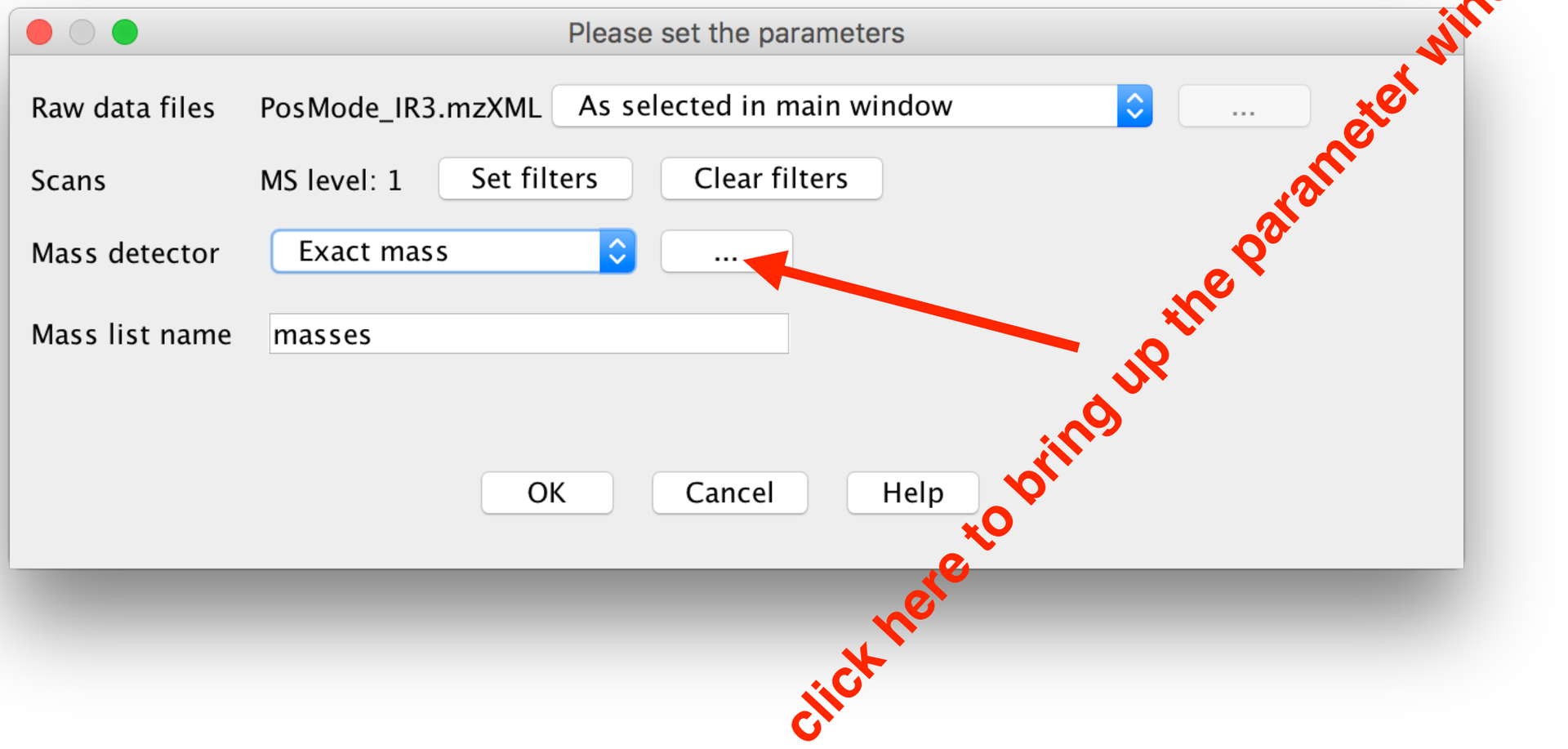
Zoom in one mass spectrum



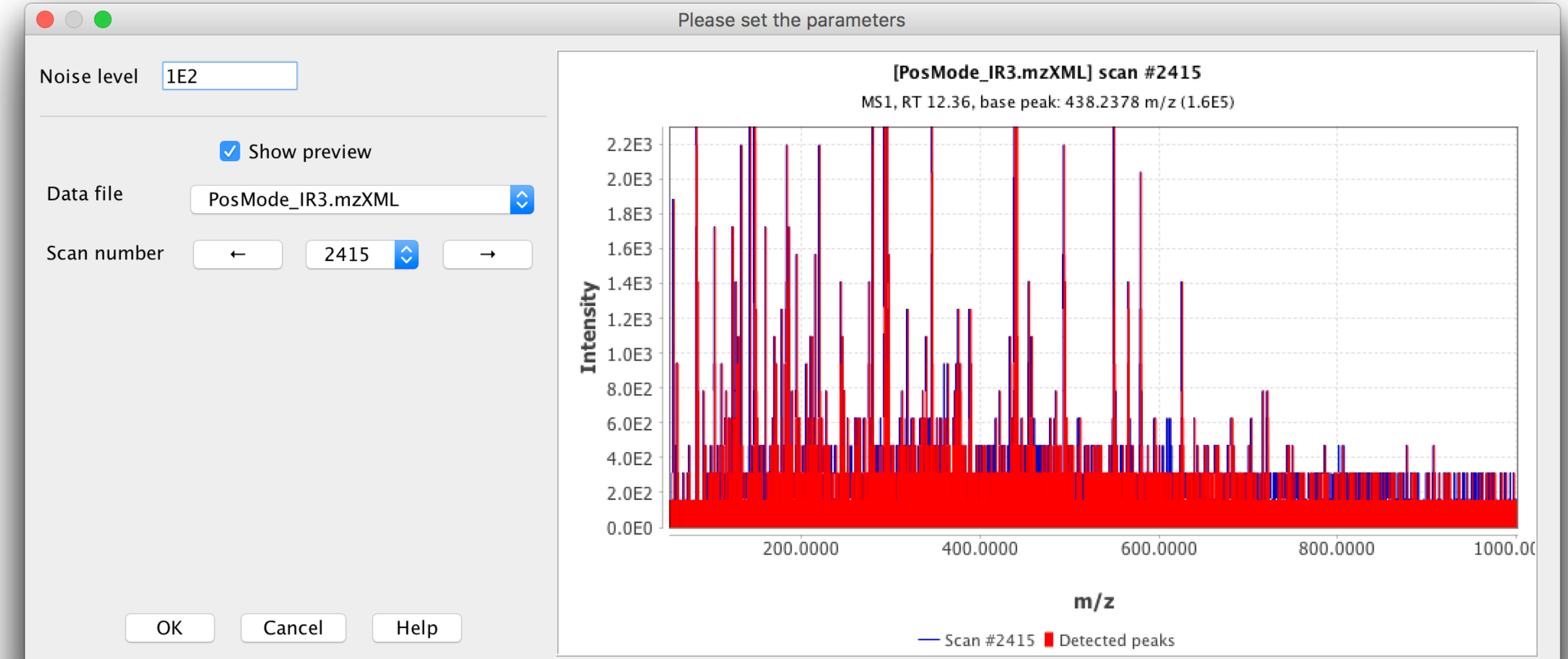
Conversion to centroid mode



Conversion to centroid mode



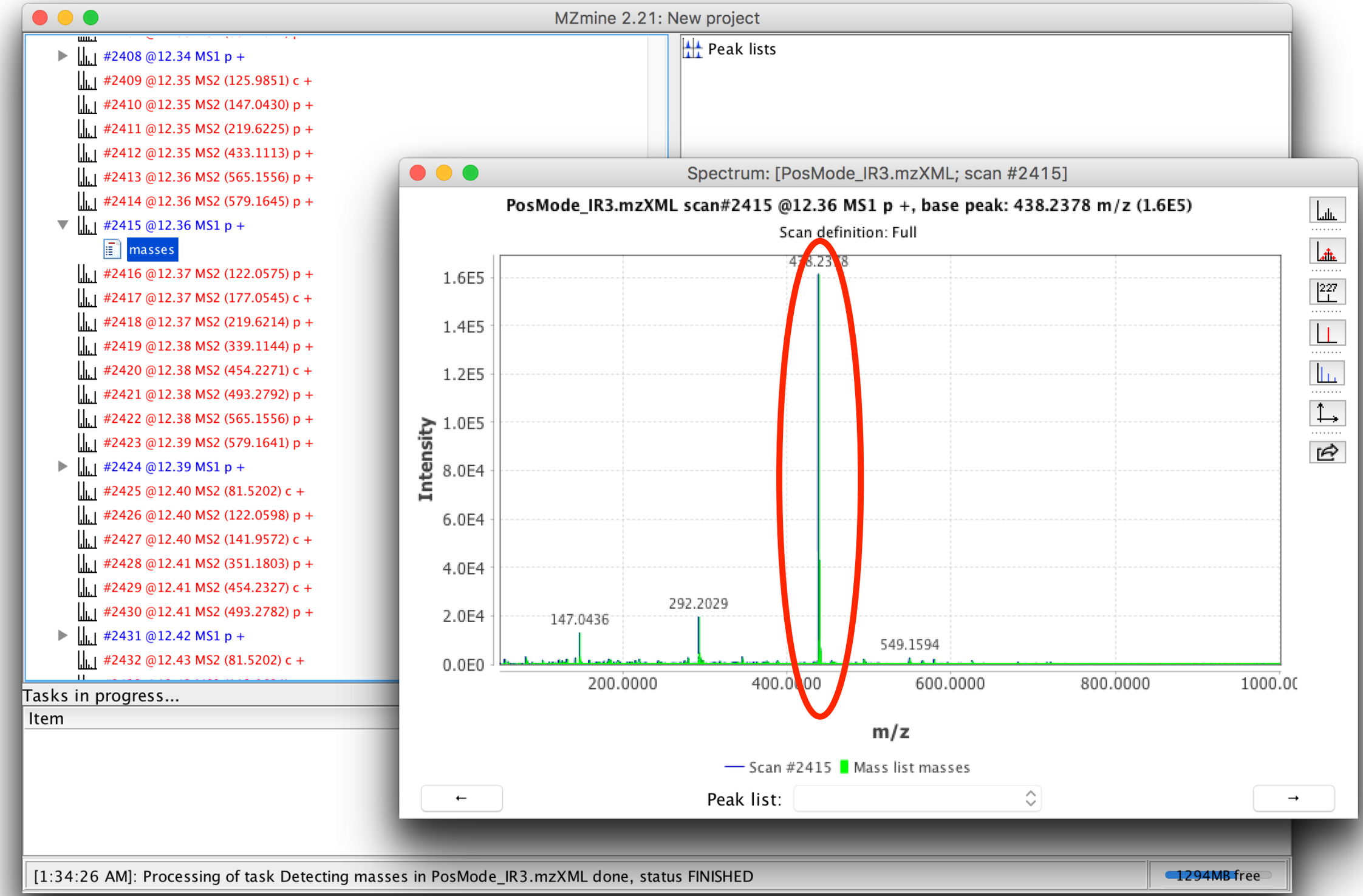
Conversion to centroid mode



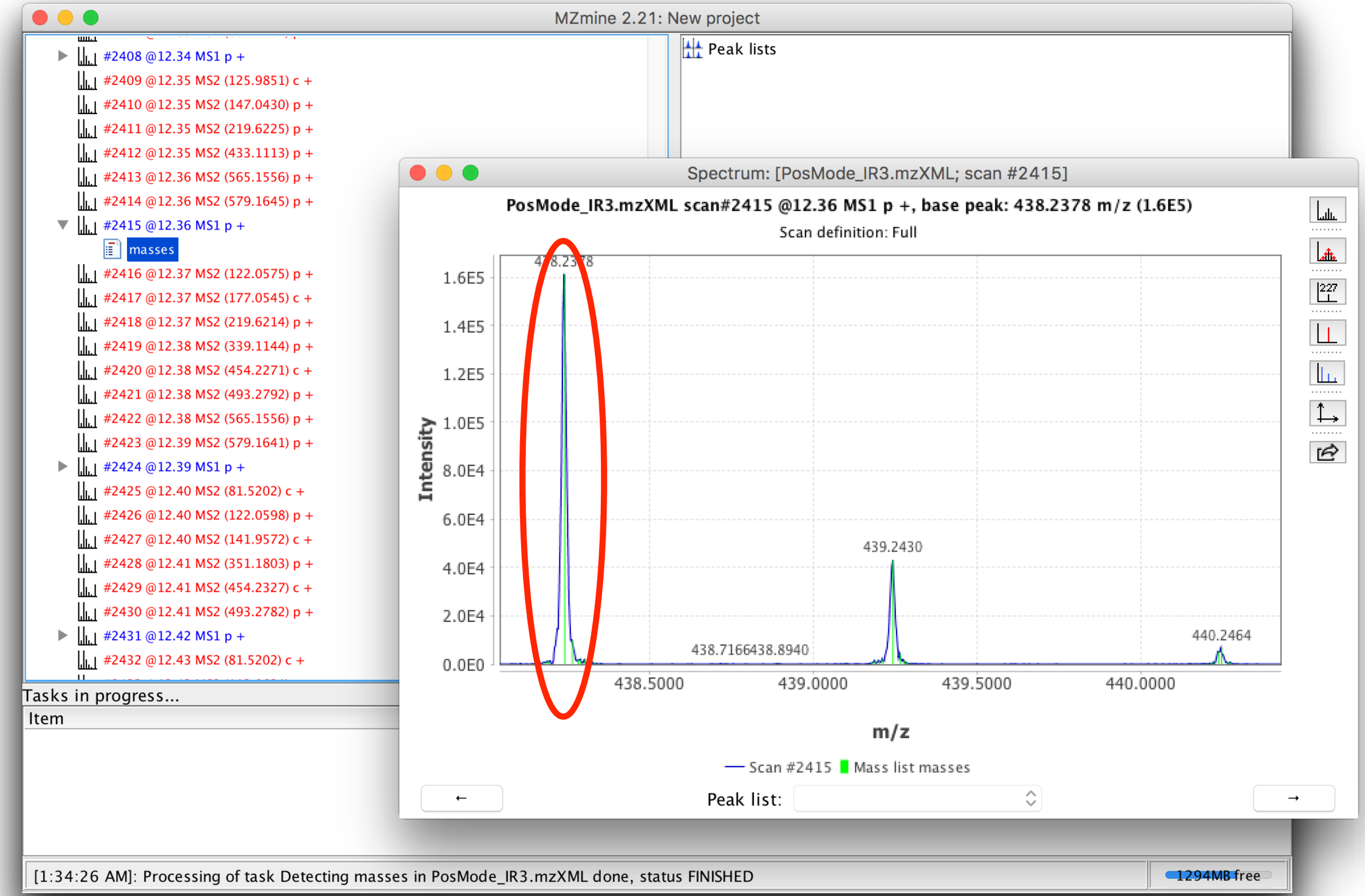
Conversion to centroid mode

Mass detection in progress ……

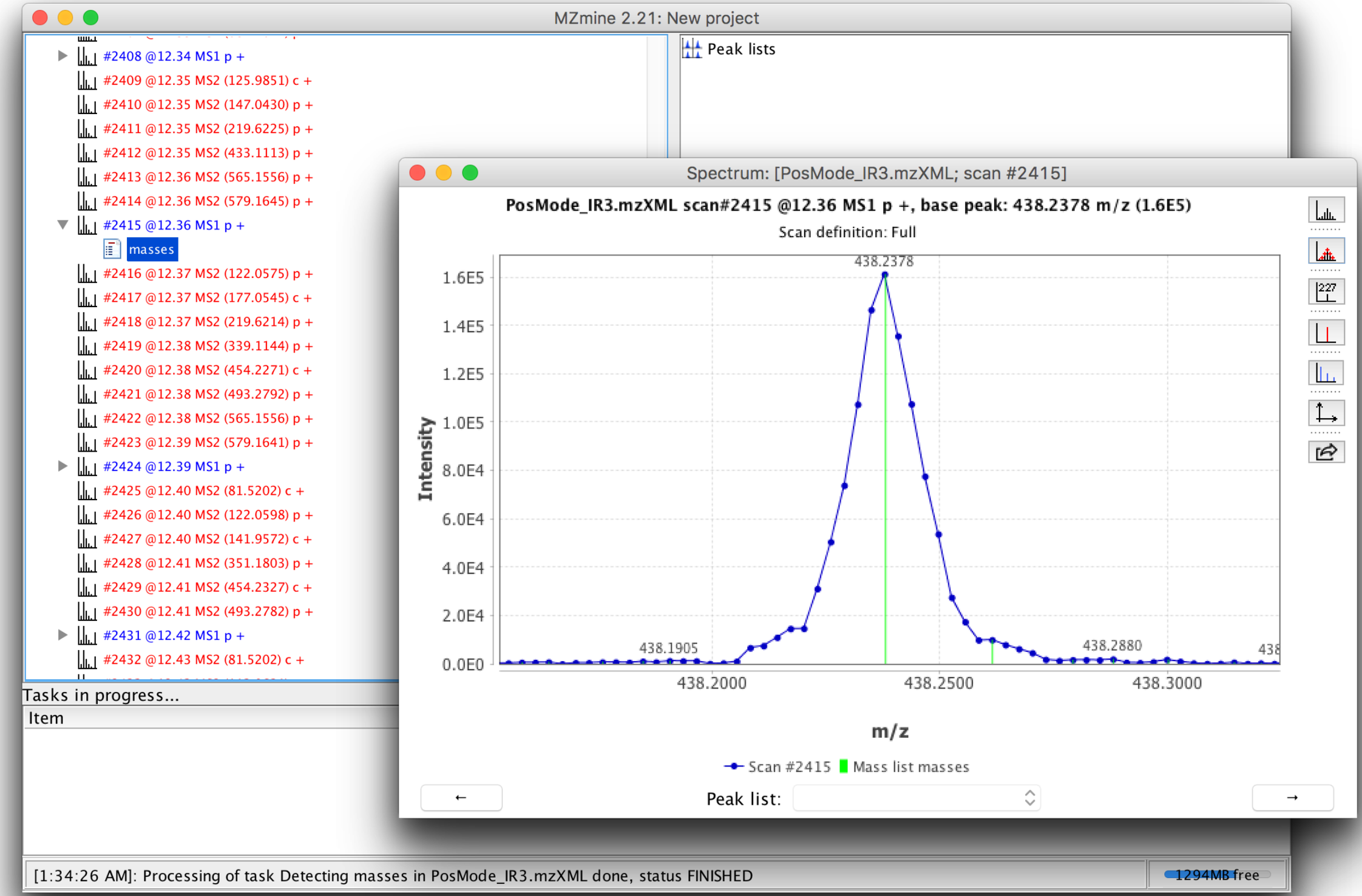
Mass spectra in centroid mode



Mass spectra in centroid mode



Mass spectra in centroid mode

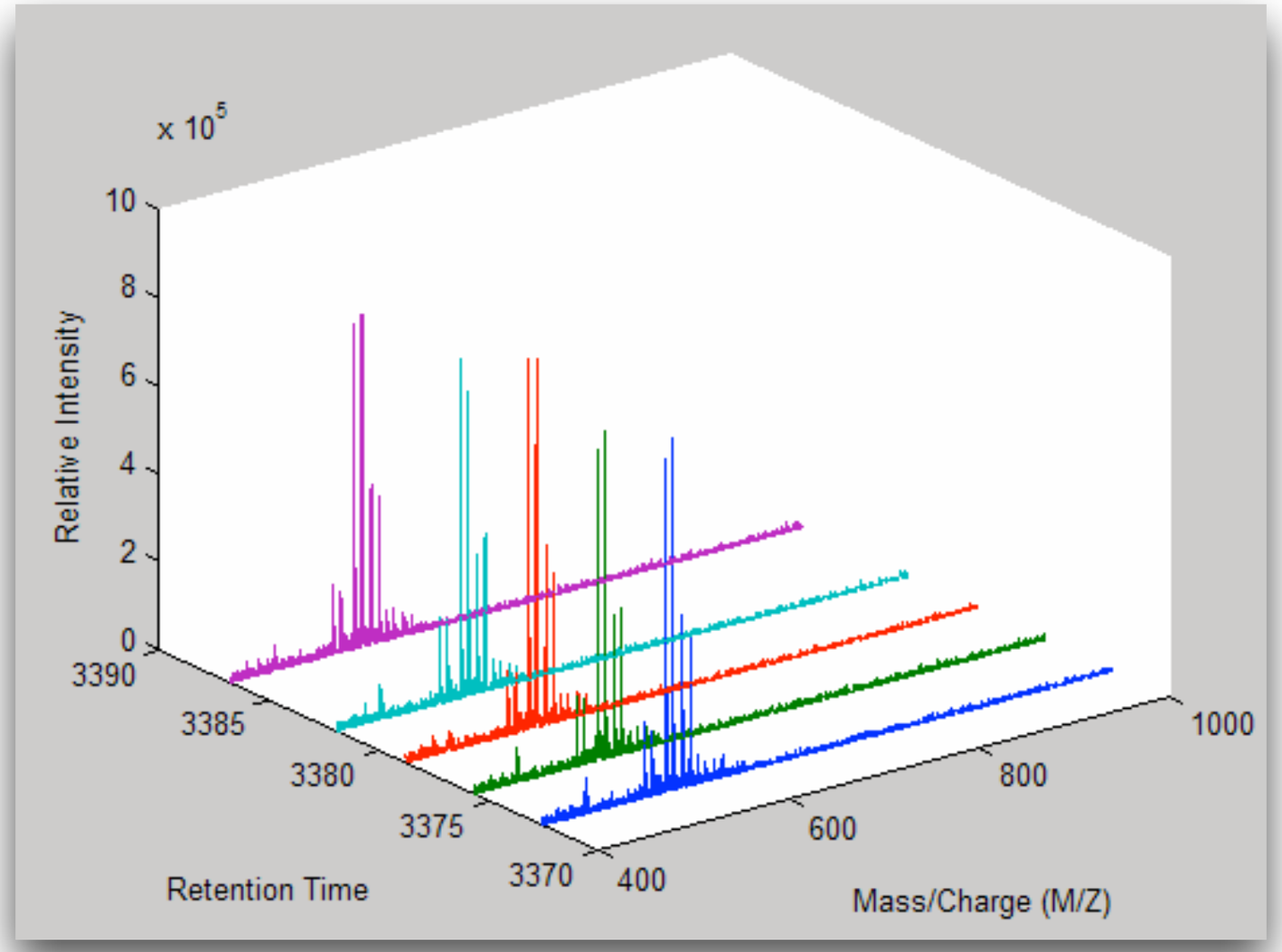


Spectrum in centroid mode

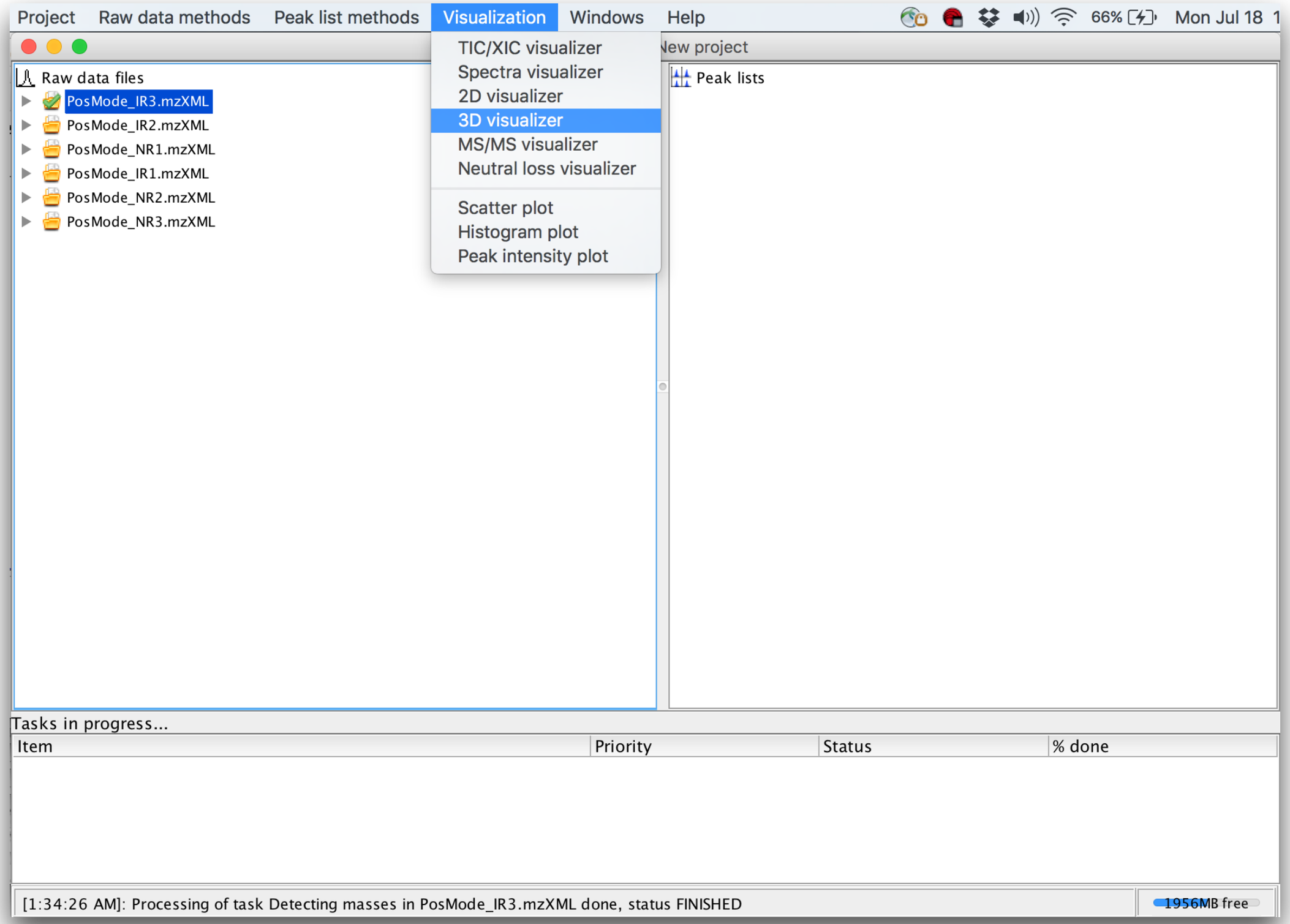
- Data files are much smaller than files in profile mode.

- We will use the centroid data for practicing data pre-processing using XCMS in R.

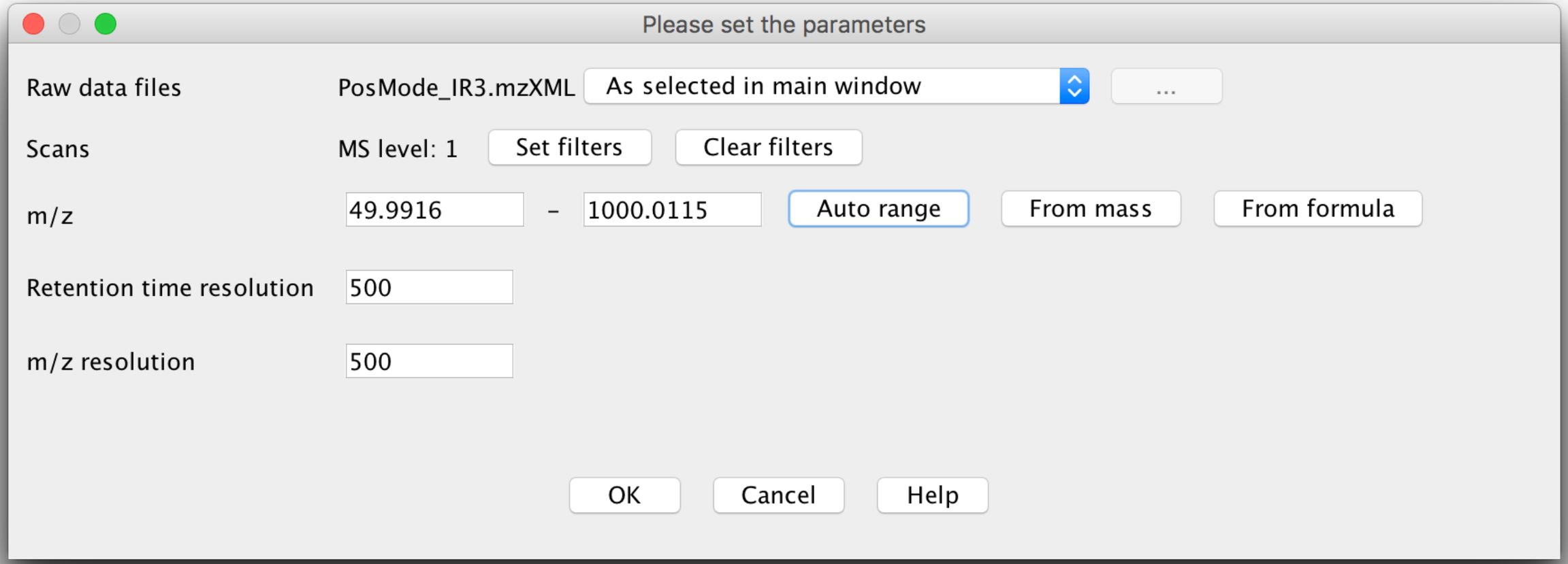
LC-MS raw data in 3D



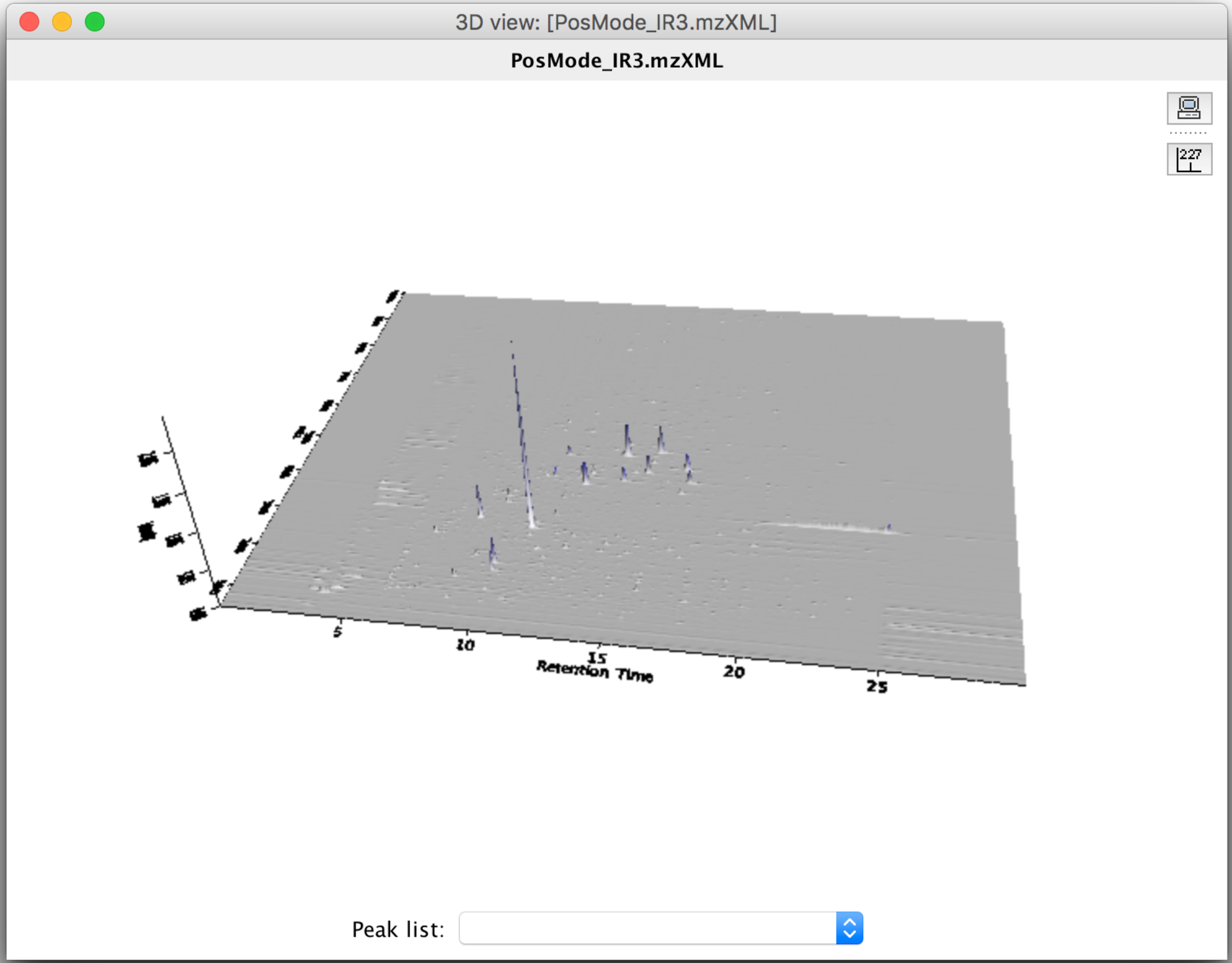
Raw data in 3D



Raw data in 3D



Raw data in 3D

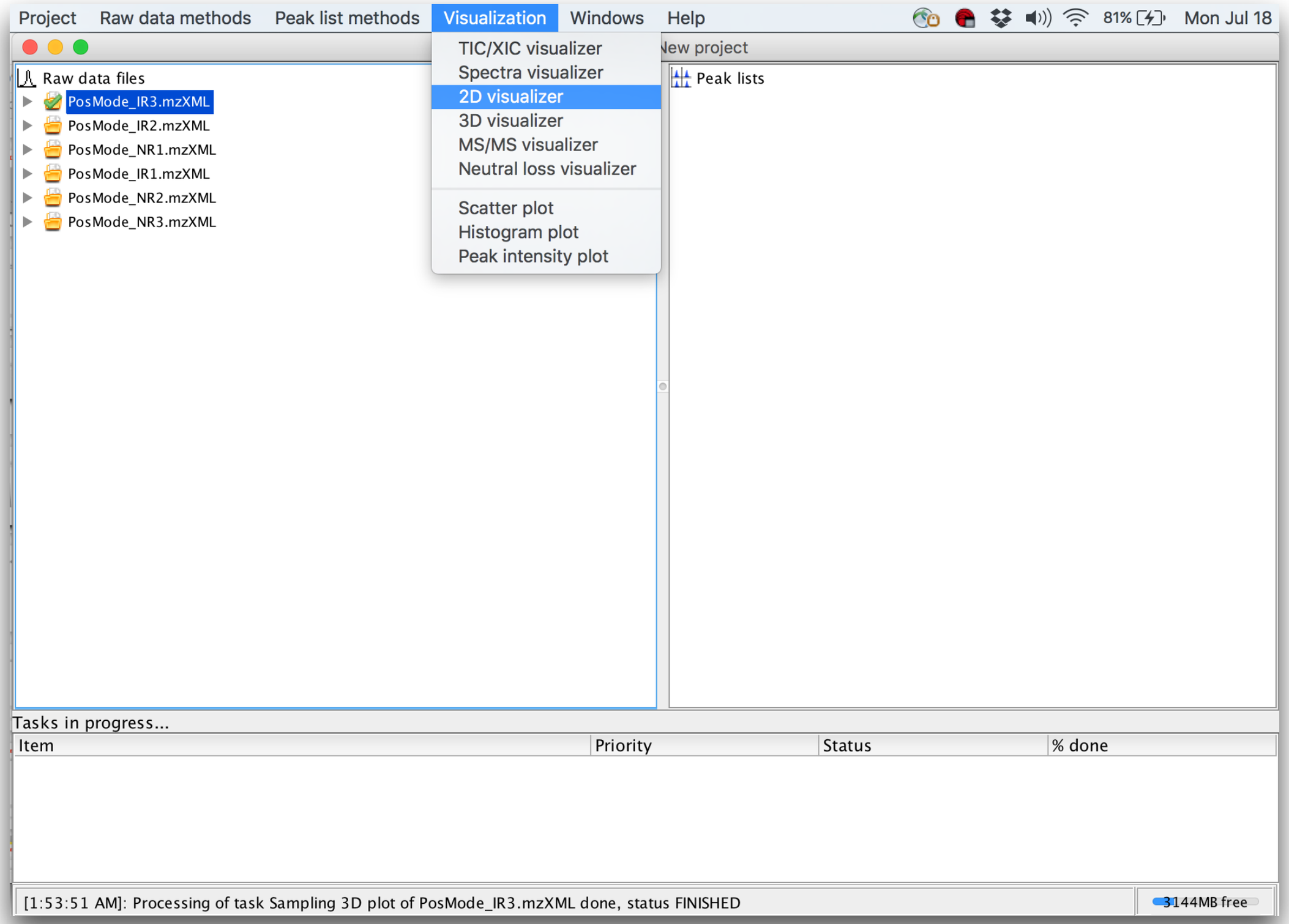


3D to 2D

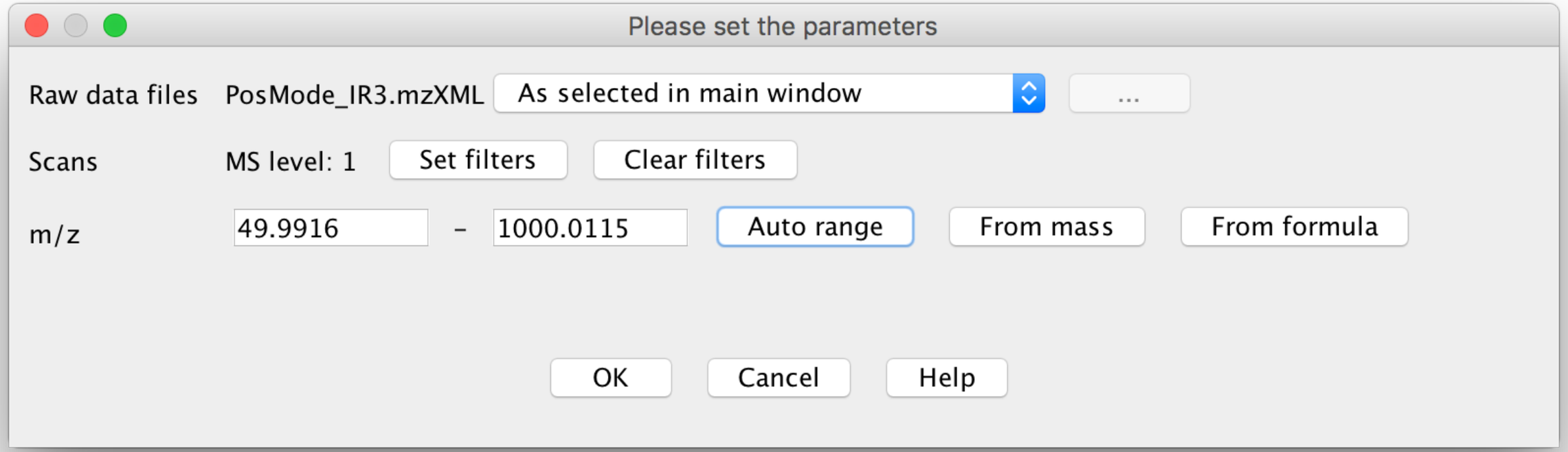
- Direct processing of the 3D data is NOT trivial

- 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)

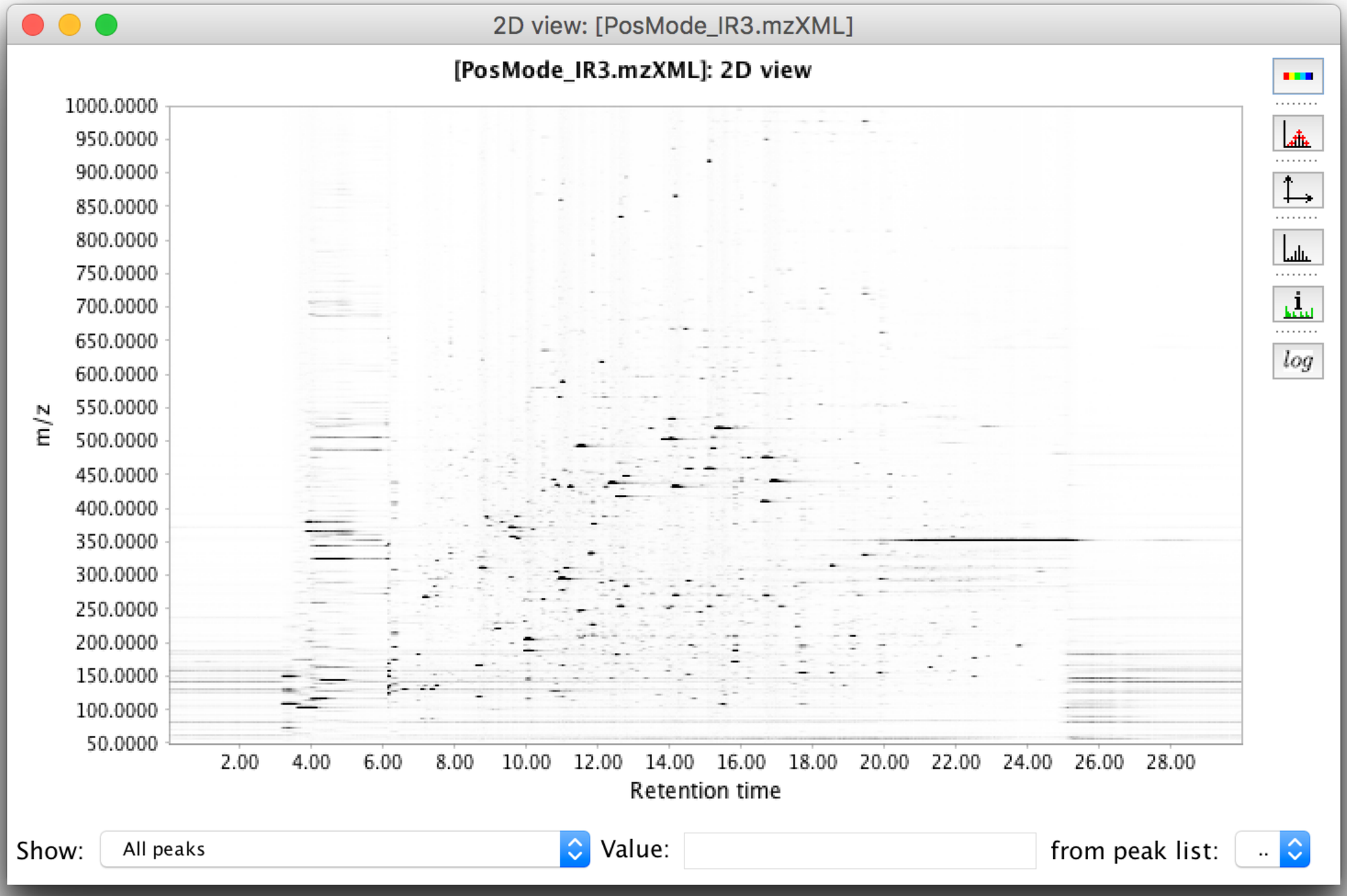
Mass vs. retention time map



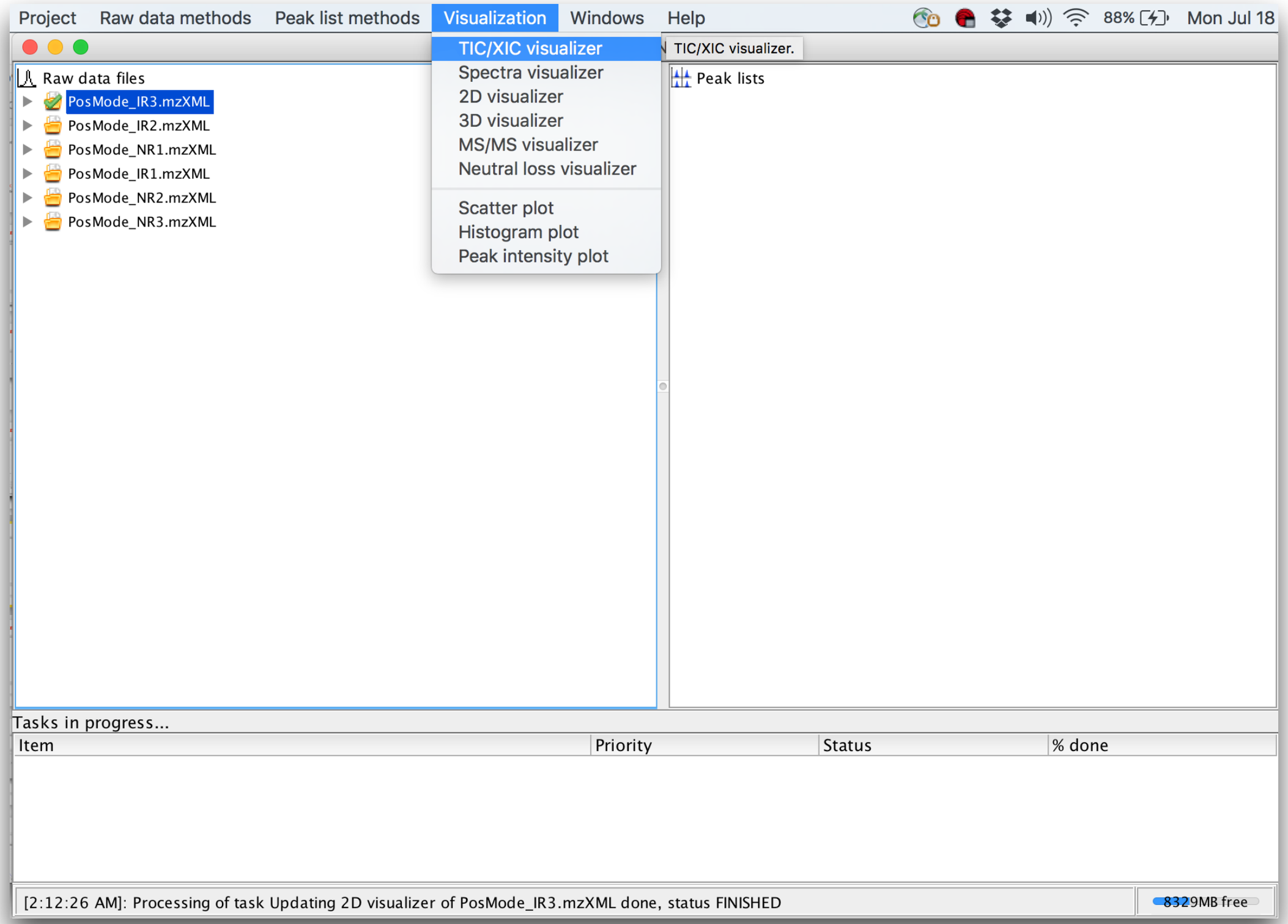
Mass vs. retention time map



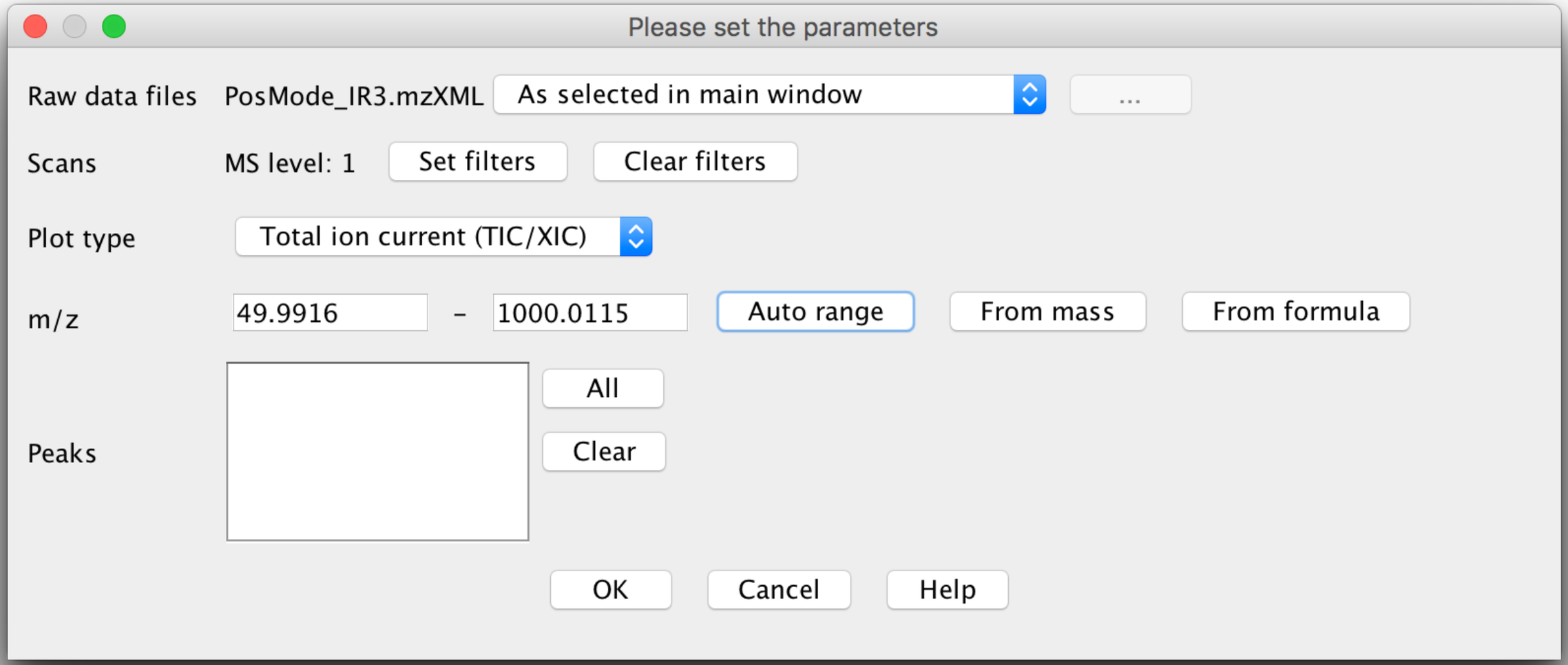
Mass vs. retention time map



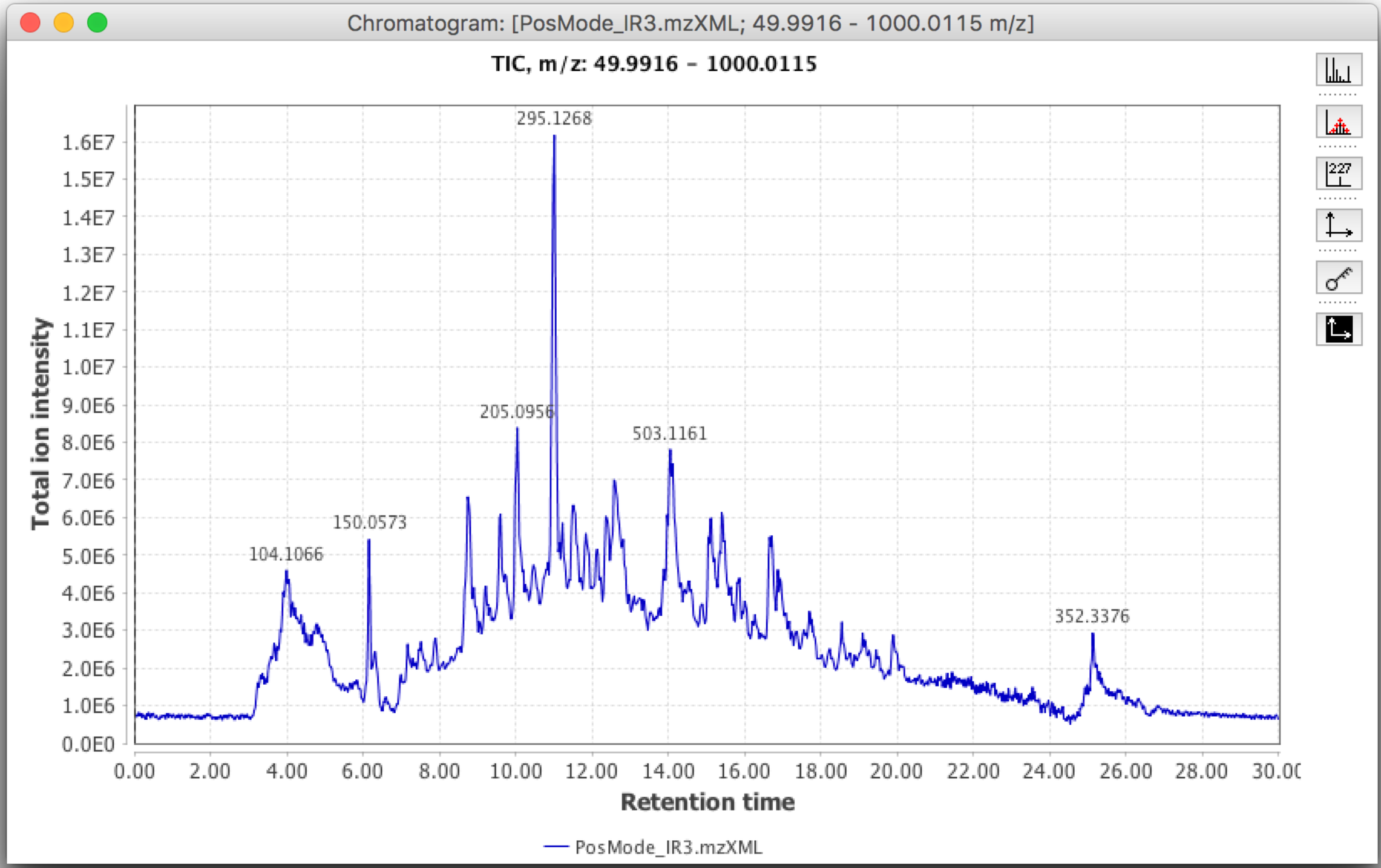
TIC


Project
Raw data methods
Peak list methods
Visualization
Windows
Help
88%
Mon Jul 18
TIC/XIC visualizer
Spectra visualizer
2D visualizer
3D visualizer
MS/MS visualizer
Neutral loss visualizer
Scatter plot
Histogram plot
Peak intensity plot
TIC/XIC visualizer.
Raw data files
PosMode_IR3.mzXML
PosMode_IR2.mzXML
PosMode_NR1.mzXML
PosMode_IR1.mzXML
PosMode_NR2.mzXML
PosMode_NR3.mzXML
Peak lists
Tasks in progress...
Item
Priority
Status
% done
[2:12:26 AM]: Processing of task Updating 2D visualizer of PosMode_IR3.mzXML done, status FINISHED
8329MB free

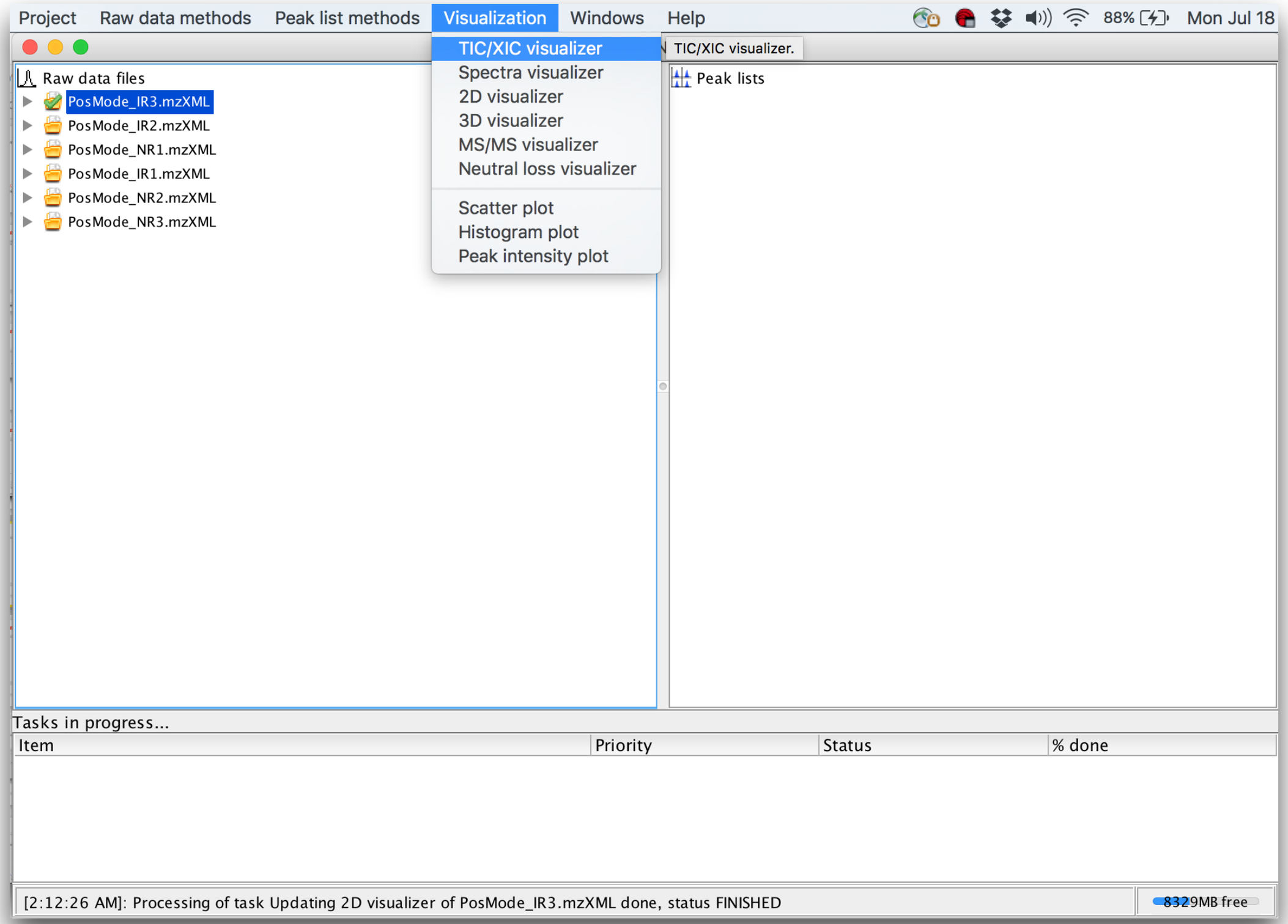
TIC


Please set the parameters
Raw data files
PosMode_IR3.mzXML
As selected in main window
...
Scans
MS level: 1
Set filters
Clear filters
Plot type
Total ion current (TIC/XIC)
m/z
49.9916
-
1000.0115
Auto range
From mass
From formula
Peaks
All
Clear
OK
Cancel
Help

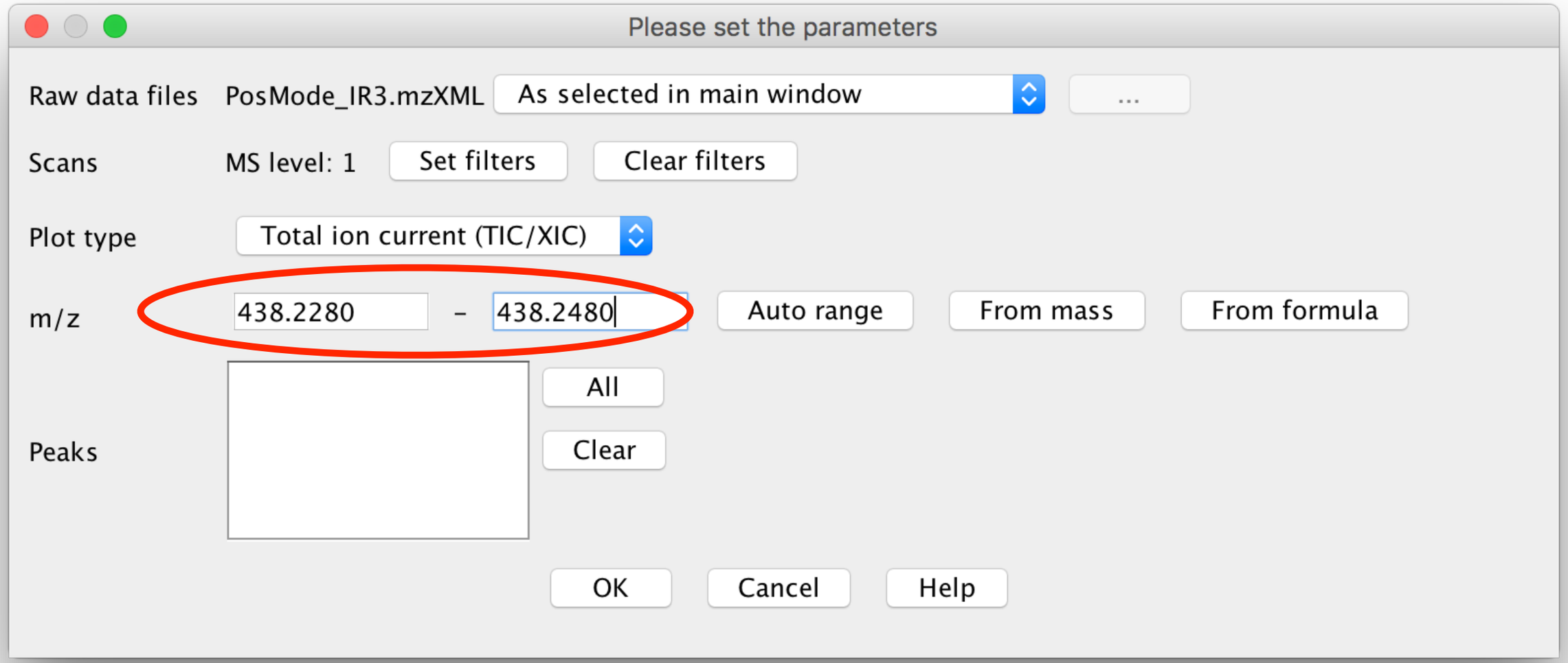
TIC


Chromatogram: [PosMode_IR3.mzXML; 49.9916 - 1000.0115 m/z]
TIC, m/z: 49.9916 - 1000.0115
295.1268
205.0956
503.1161
150.0573
104.1066
352.3376
Total ion intensity
1.6E7
1.5E7
1.4E7
1.3E7
1.2E7
1.1E7
1.0E7
9.0E6
8.0E6
7.0E6
6.0E6
5.0E6
4.0E6
3.0E6
2.0E6
1.0E6
0.0E0
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
22.00
24.00
26.00
28.00
Retention time
PosMode_IR3.mzXML

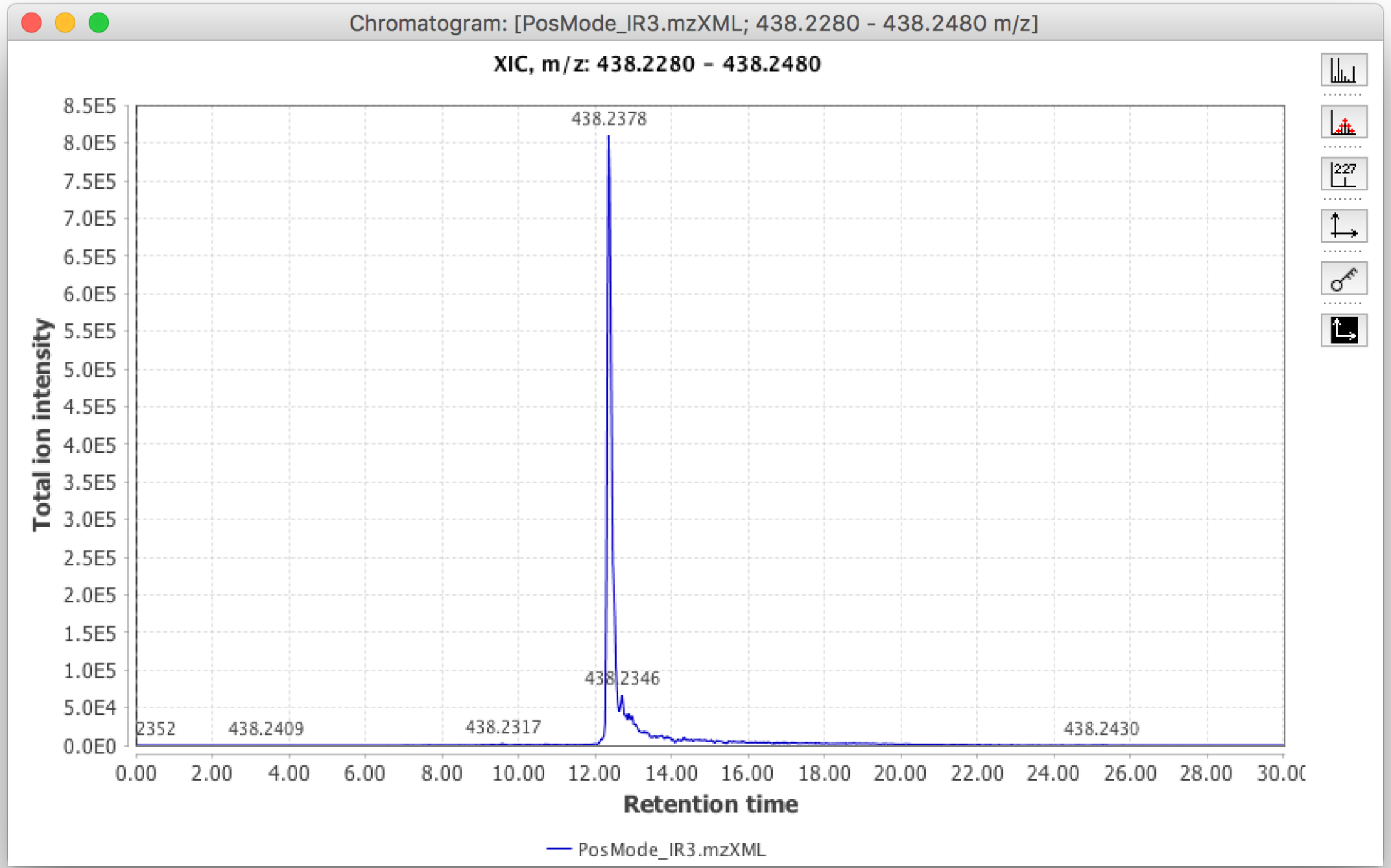
EIC


Project
Raw data methods
Peak list methods
Visualization
Windows
Help
88%
Mon Jul 18
TIC/XIC visualizer
Spectra visualizer
2D visualizer
3D visualizer
MS/MS visualizer
Neutral loss visualizer
Scatter plot
Histogram plot
Peak intensity plot
TIC/XIC visualizer.
Raw data files
PosMode_IR3.mzXML
PosMode_IR2.mzXML
PosMode_NR1.mzXML
PosMode_IR1.mzXML
PosMode_NR2.mzXML
PosMode_NR3.mzXML
Peak lists
Tasks in progress...
Item
Priority
Status
% done
[2:12:26 AM]: Processing of task Updating 2D visualizer of PosMode_IR3.mzXML done, status FINISHED
8329MB free

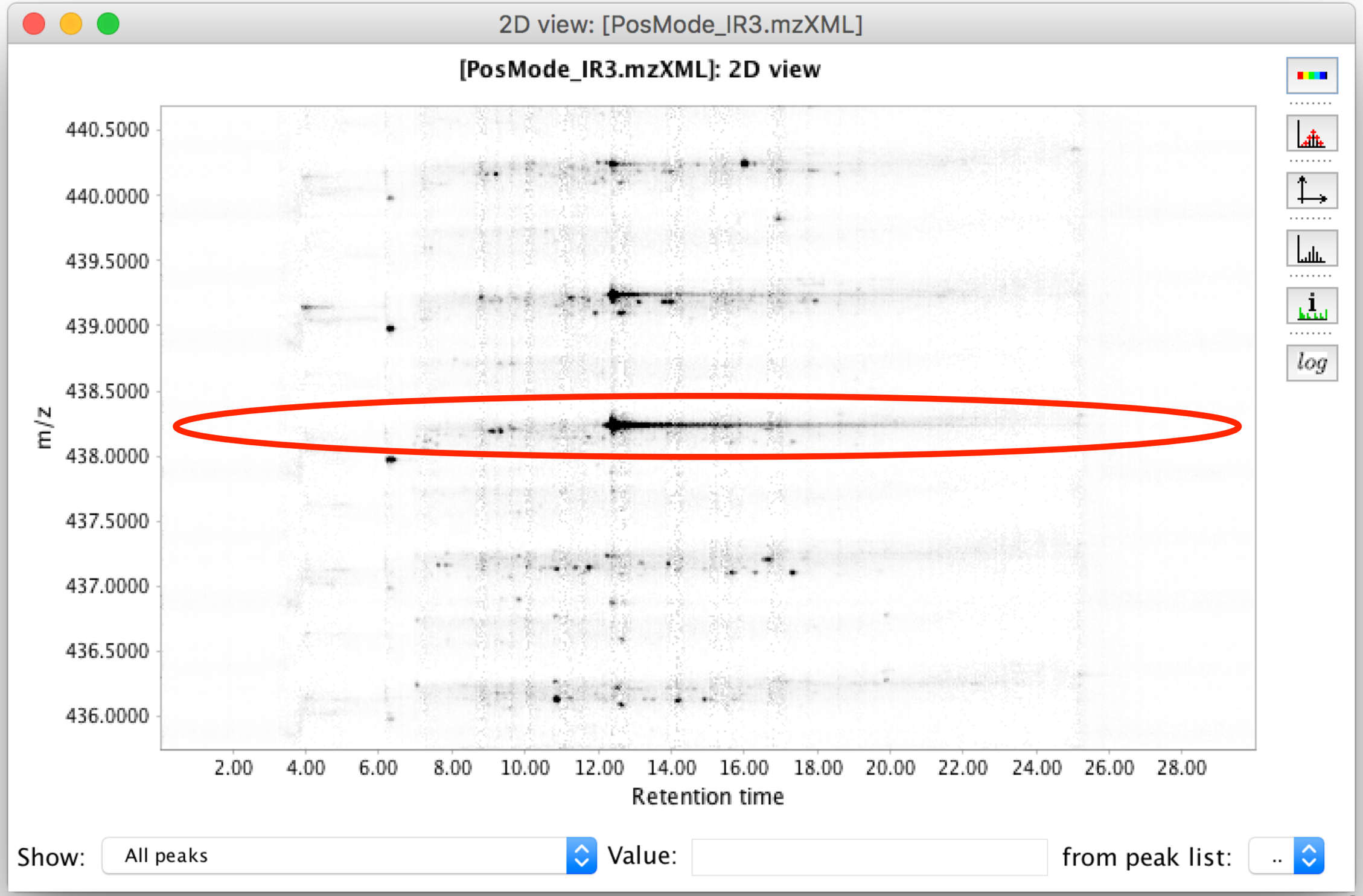
EIC


Please set the parameters
Raw data files
PosMode_IR3.mzXML
As selected in main window
...
Scans
MS level: 1
Set filters
Clear filters
Plot type
Total ion current (TIC/XIC)
m/z
438.2280
-
438.2480
Auto range
From mass
From formula
Peaks
All
Clear
OK
Cancel
Help

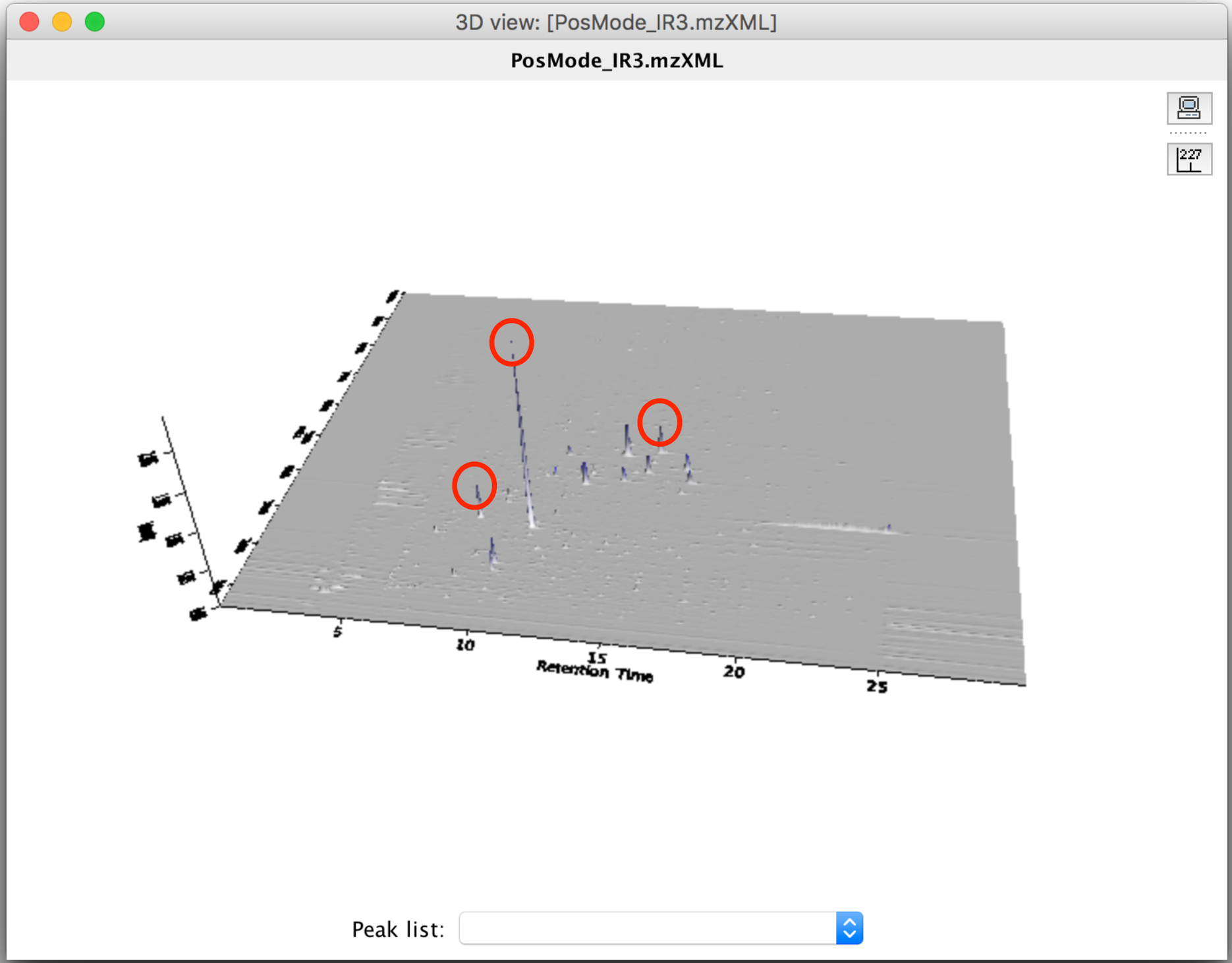
EIC


Chromatogram: [PosMode_IR3.mzXML; 438.2280 - 438.2480 m/z]
XIC, m/z: 438.2280 - 438.2480
Total ion intensity
8.5E5
8.0E5
7.5E5
7.0E5
6.5E5
6.0E5
5.5E5
5.0E5
4.5E5
4.0E5
3.5E5
3.0E5
2.5E5
2.0E5
1.5E5
1.0E5
5.0E4
0.0E0
438.2378
438.2346
2352
438.2409
438.2317
438.2430
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
22.00
24.00
26.00
28.00
Retention time
PosMode_IR3.mzXML

EIC


2D view: [PosMode_IR3.mzXML]
[PosMode_IR3.mzXML]: 2D view
440.5000
440.0000
439.5000
439.0000
438.5000
438.0000
437.5000
437.0000
436.5000
436.0000
m/z
2.00
4.00
6.00
8.00
10.00
12.00
14.00
16.00
18.00
20.00
22.00
24.00
26.00
28.00
Retention time
log
Show:
All peaks
Value:
from peak list:
..

Feature

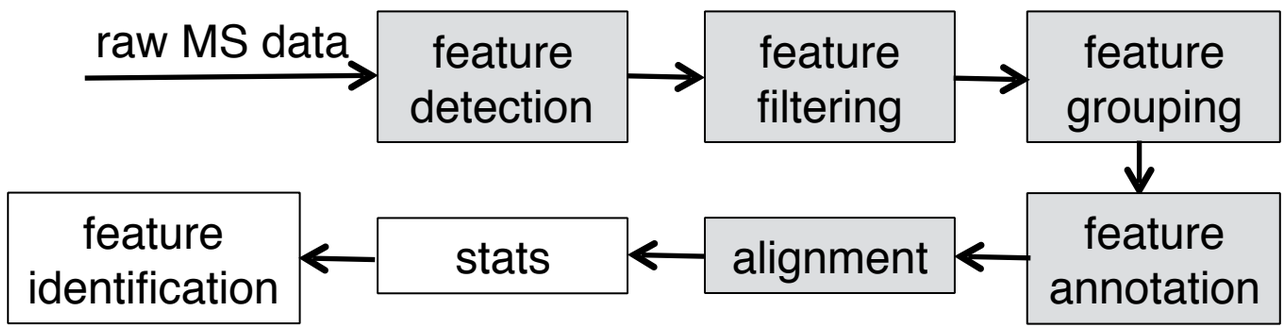

3D view: [PosMode_IR3.mzXML]
PosMode_IR3.mzXML
227
5
10
15
20
25
Retention Time
Peak list:

Feature

- Feature: A 3D signal induced by a single ion species (e.g. $[M+H]^+$ or $[M-H]^-$ of a compound)

- How to detect features?
  - by peak picking from EICs

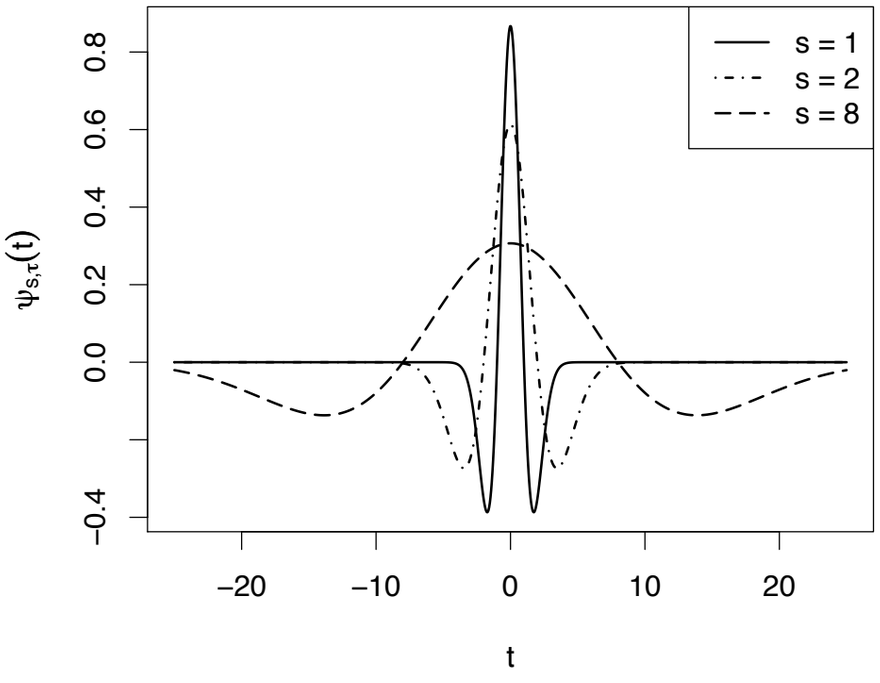
- Data processing and analysis workflow



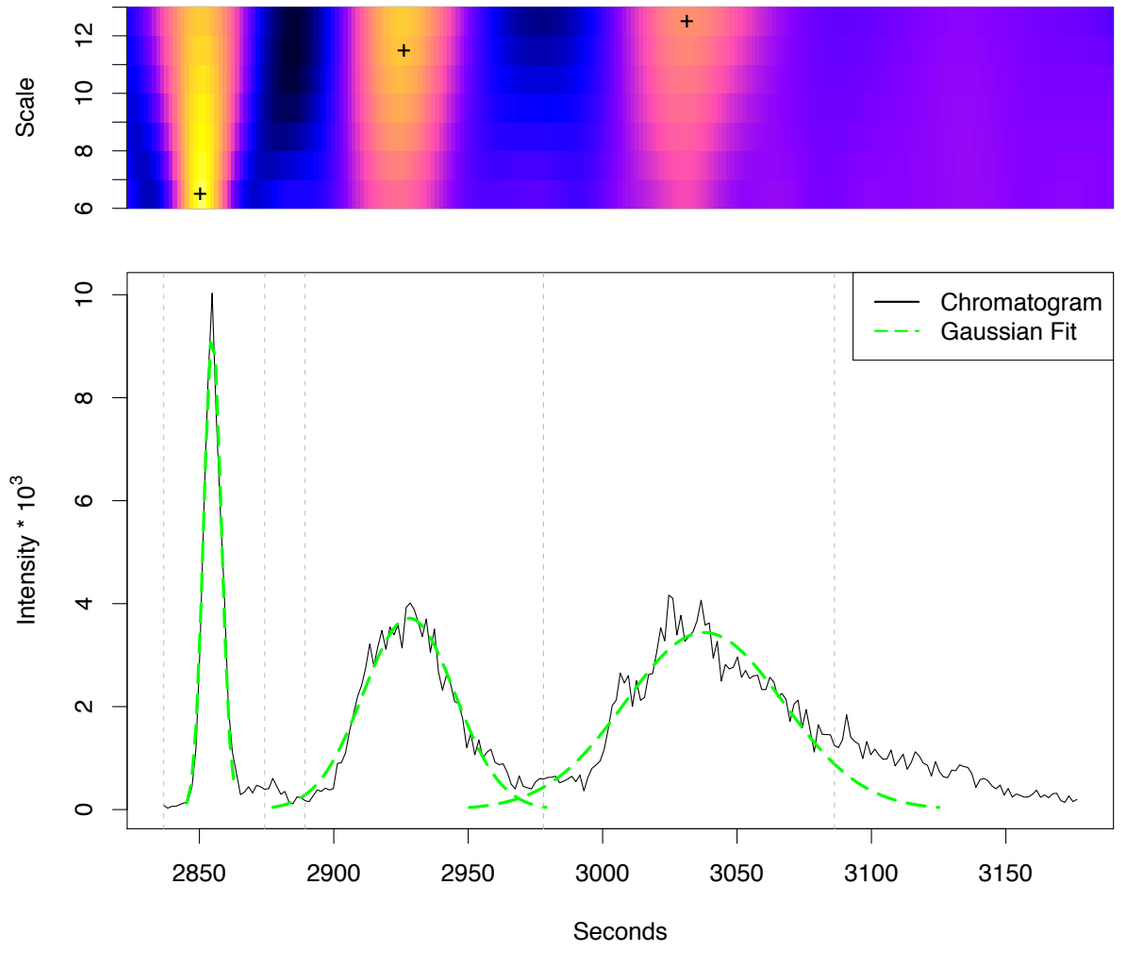
Data Pre-processing

Detection of chromatographic features

- Use wavelet transform



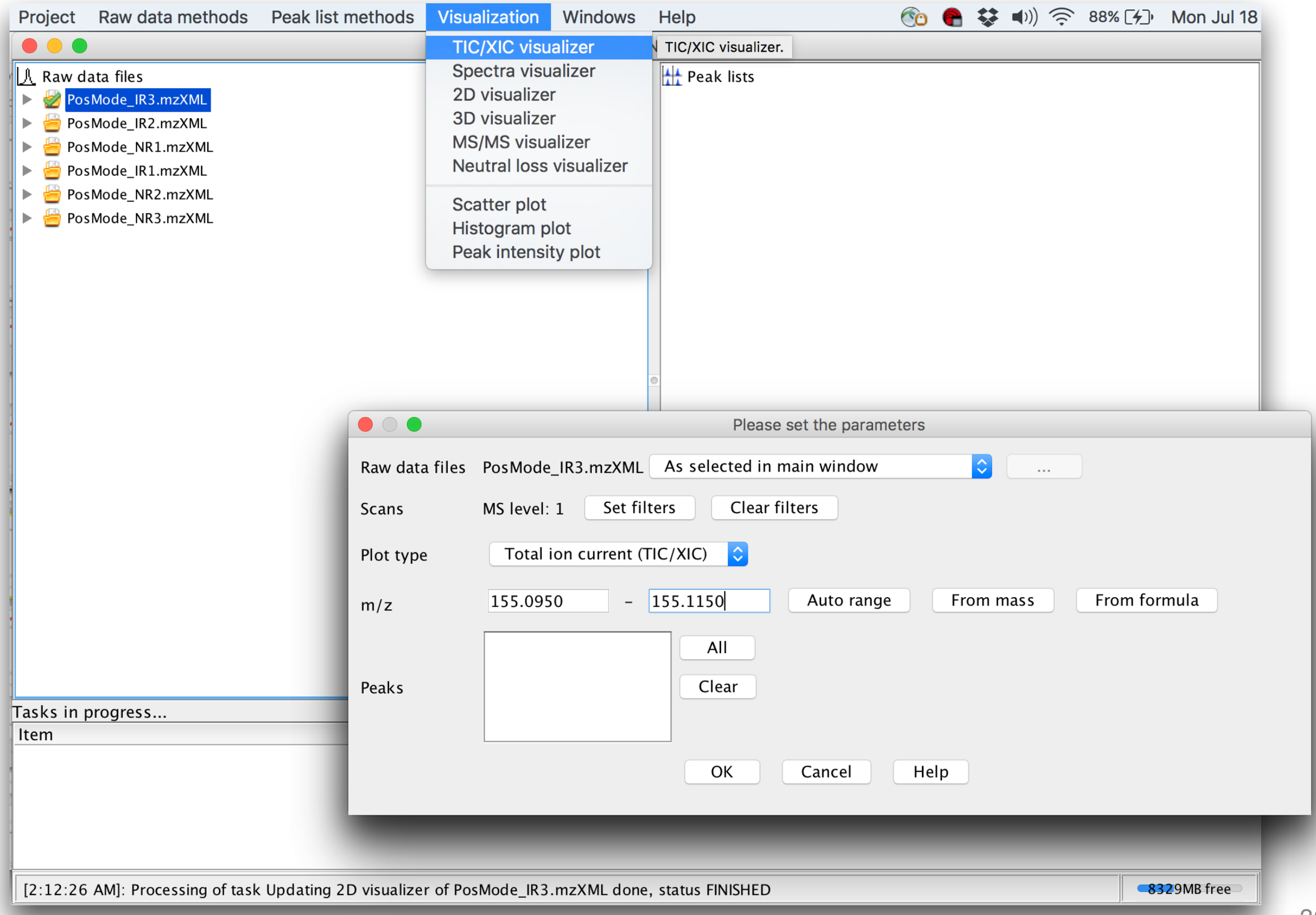
mexican hat wavelet



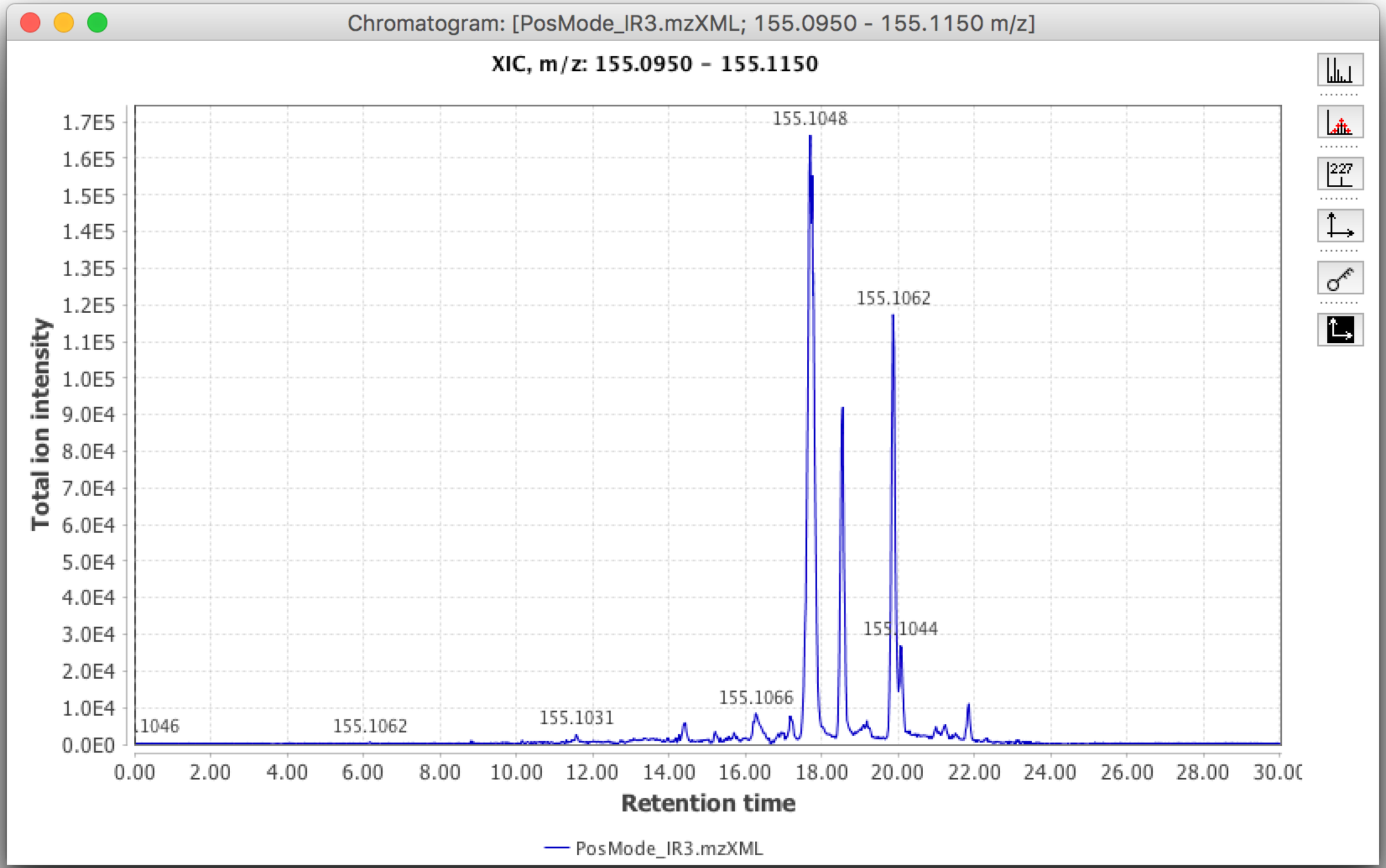
- Implemented in XCMS as the centWave method

Tautenhahn, R.; Bottcher, C.; Neumann, S., Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics* **2008,** *9*, 504.

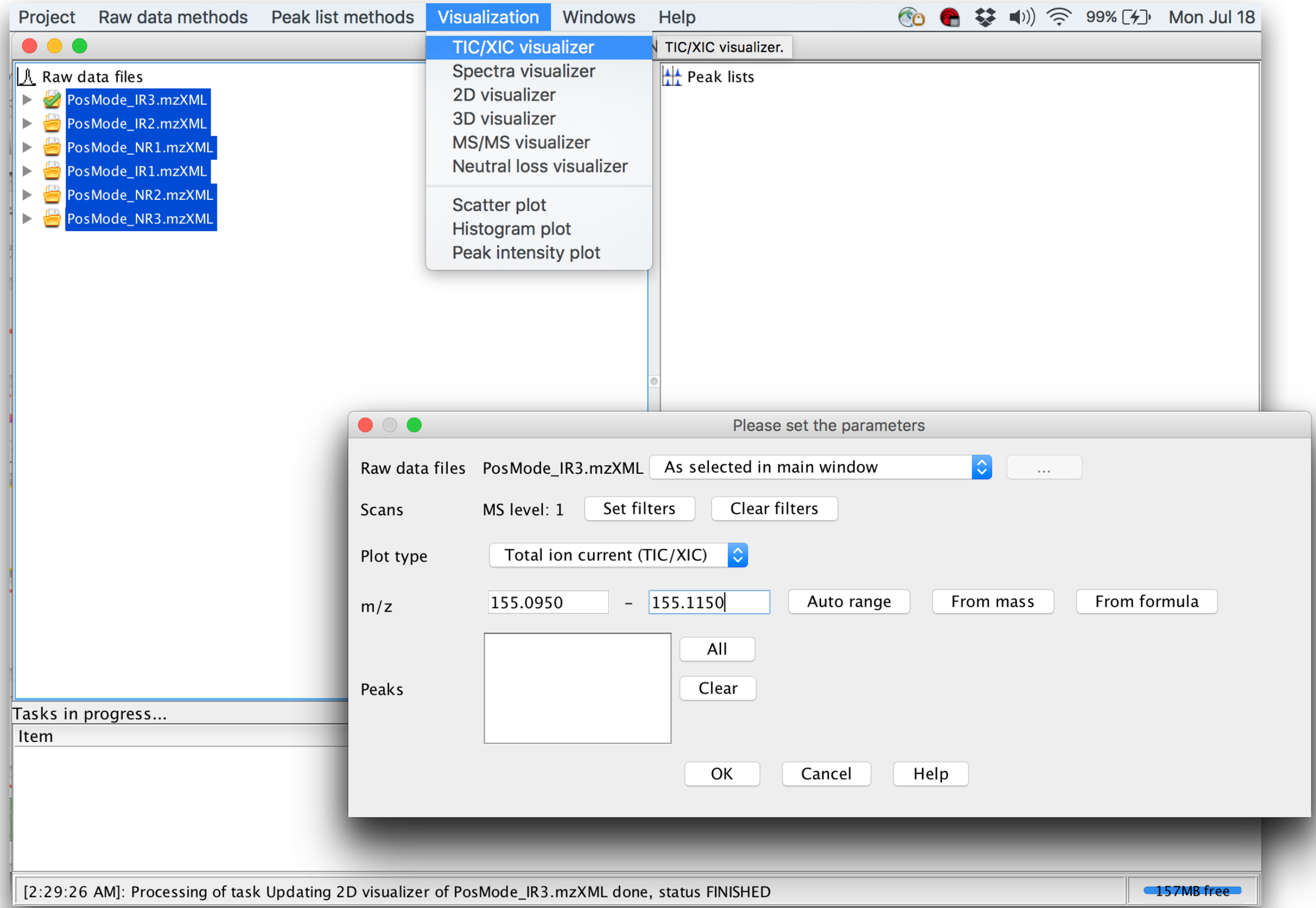
Detection of chromatographic features



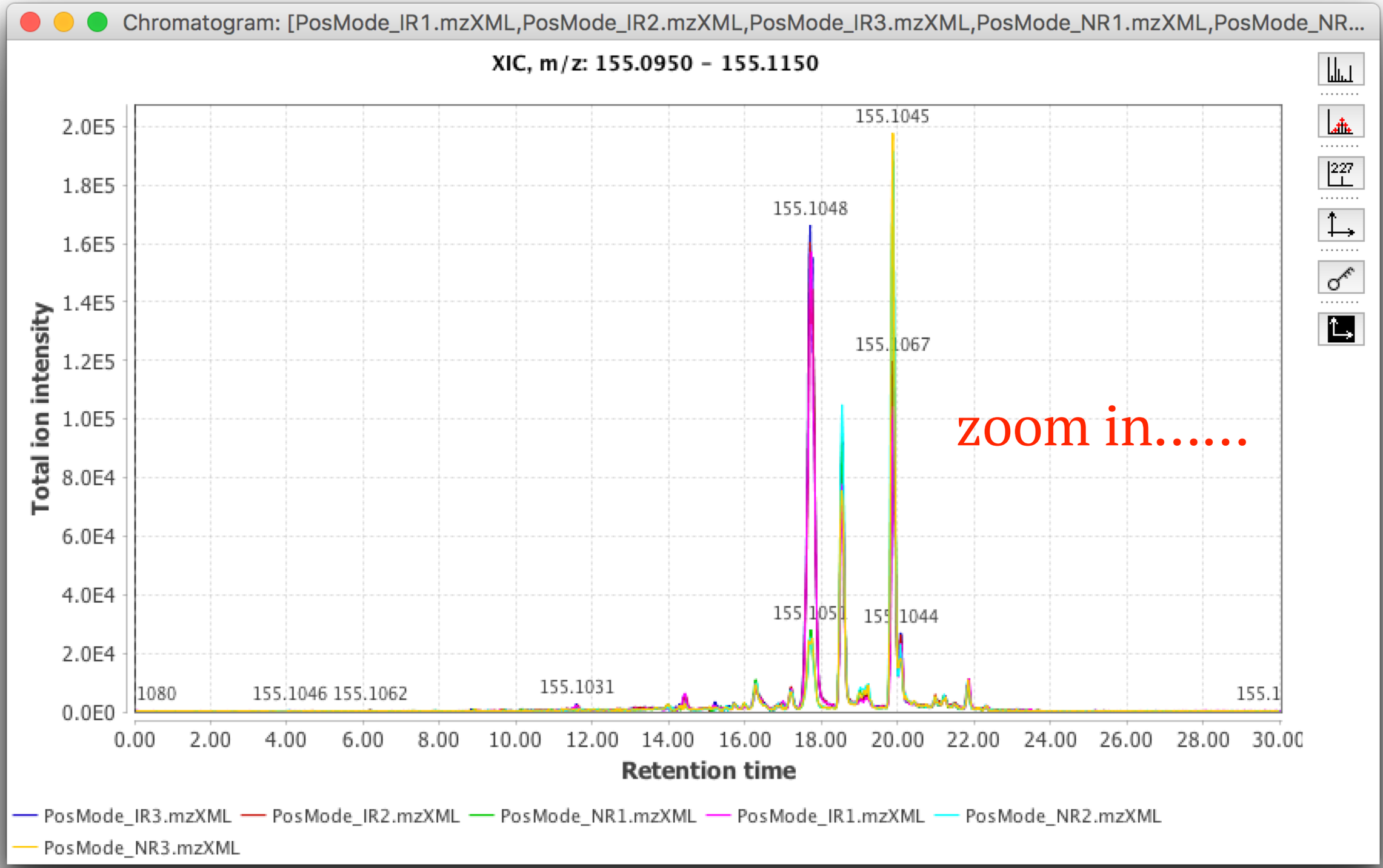
Detection of chromatographic features



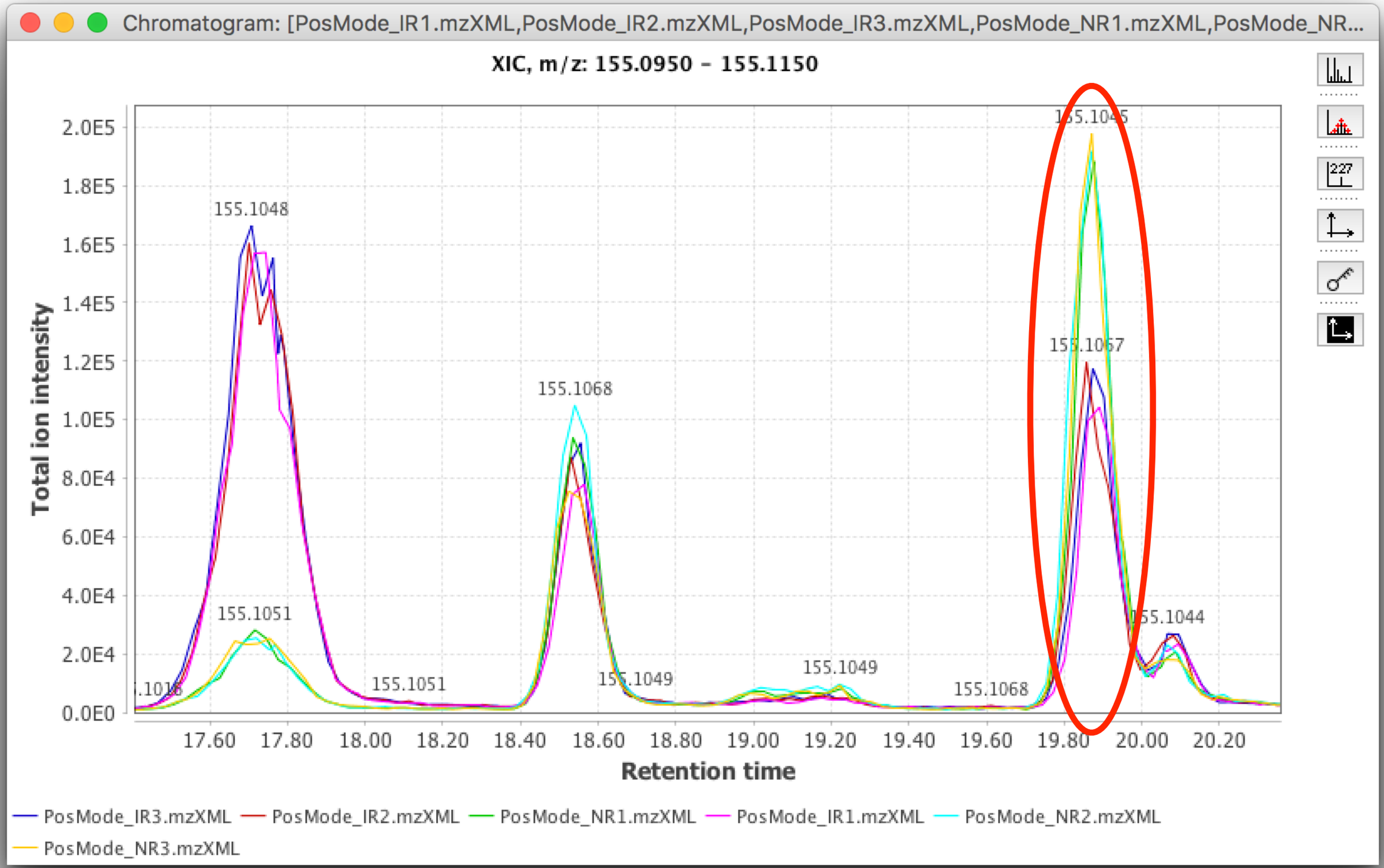
Feature grouping and alignment



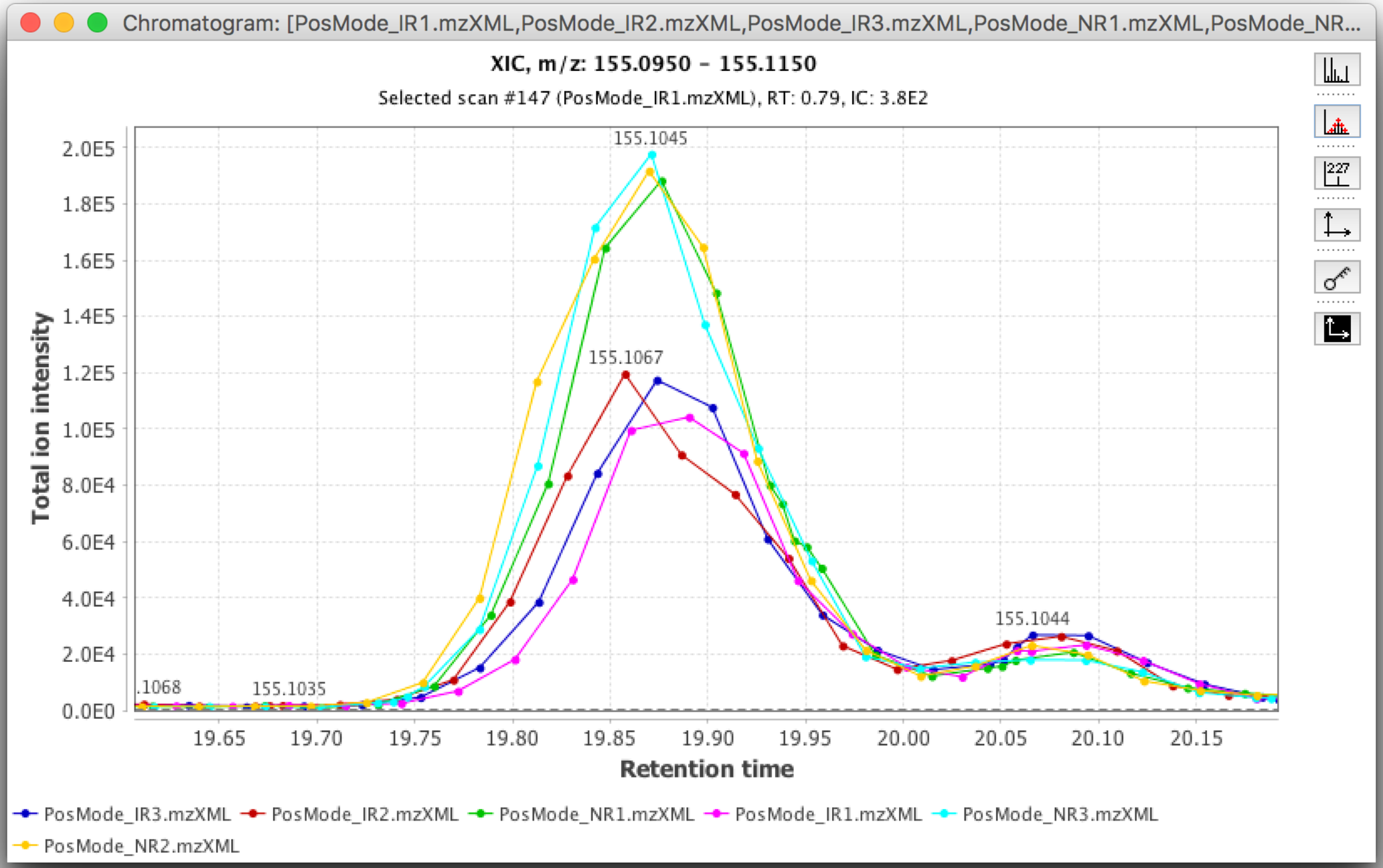
Feature grouping and alignment



Feature grouping and alignment



Feature grouping and alignment



Result of data pre-processing

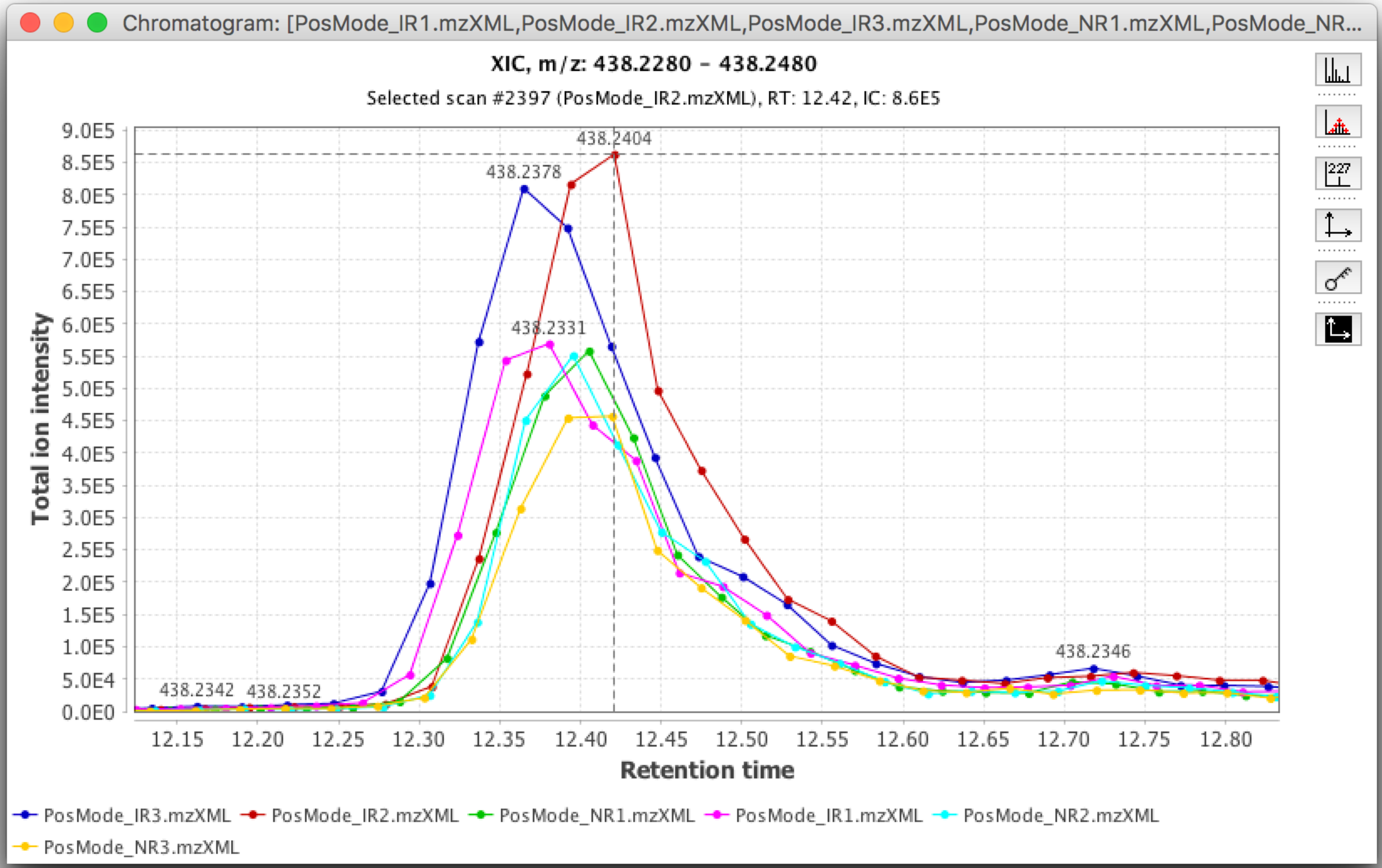
DB	Name	Mass	RT	platform	IN1	IN2	IN3	IN4	IN5	IN6
HMDB	1-Phenylethylamin	122.09745	24.97845	ES-	0.12862	0.1421305	0.1301326	0.1247924	0.1200045	0.1053275
HMDB	2-Ethylacrylic acid	101.06421	17.811575	ES-	0.0332025	0.0174262	0.0158166	0.0179326	0.0143742	0.0064953
HMDB	Canavanine	177.09653	10.338581	ES-	0.0141136	0.0134146	0.0182777	0.0193855	0.0245958	0.0011908
HMDB	Diketogulonic acid	193.03069	4.7050639	ES-	0.0209463	0.0203901	0.0165056	0.0189088	0.0137482	0.017231
HMDB	Iso-Valeraldehyde	87.080171	11.164359	ES-	0.6558109	0.2742277	0.2651933	0.3093793	0.2101024	0.0541026
in-house	3,4-Dehydro-Dprol	114.04431	3.5491023	ES-	0.2900544	0.287811	0.2290651	0.2754269	0.2314117	0.2061301
in-house	4-hydroxy-proline	132.05326	3.5958634	ES-	0.5584389	0.7353401	0.5273908	0.4412898	0.5074794	0.5423602
in-house	Malic acid	133.01996	3.9406386	ES-	0.0555016	0.0461576	0.0290383	0.0390783	0.0380952	0.0308288
in-house	2,3,4-Trihydroxybu	135.04472	3.5763487	ES+	0.0223984	0.0146371	0.0150894	0.0097238	0.0116862	0.0116129
in-house	2,3-Diaminopropic	105.07016	3.3202935	ES+	0.024859	0.0207034	0.0225235	0.0201288	0.0226763	0.0226569
in-house	4-Methy2-oxovaler	129.07306	16.624045	ES+	0.1341287	0.2458095	0.2138968	0.2383272	0.1646037	0.2156238
in-house	5-Aminopentanoic	116.0542	3.9125471	ES+	0.015214	0.0157145	0.0152048	0.0139855	0.0148445	0.0151512
in-house	Acetylcarnitine	204.12263	3.8790521	ES+	0.503742	0.4063954	0.3690539	0.3346704	0.1894332	0.267591
HMDB	11-beta-hydroxyan	483.25453	21.64161	ES+	0.0352862	0.0143528	0.0117155	0.0149876	0.0110671	0.003493
HMDB	13-Hydroperoxylin	313.23515	21.000715	ES+	0.012489	0.0124697	0.0117186	0.0120185	0.0129048	0.0116153
HMDB	17-Hydroxylinolen	295.22749	19.925457	ES+	0.0141132	0.0156397	0.0151444	0.0142477	0.0153367	0.015173
HMDB	2,4-Diaminobutyri	119.0844	3.8790898	ES+	0.0636478	0.0838566	0.0635174	0.067999	0.0942851	0.0625007
HMDB	2,6 dimethylheptar	302.23203	18.02586	ES+	0.0031349	0.0042189	0.0027814	0.0082044	0.002749	0.0032303
HMDB	2-Ethylhydracrylic	119.07199	15.226531	ES+	0.0236145	0.0239315	0.0242947	0.0237831	0.0239368	0.0242611
HMDB	2-Ketohexanoic ac	131.07027	3.7353582	ES+	0.0038071	0.0051703	0.0041894	0.0056894	0.0057567	0.0036369

Feature identification

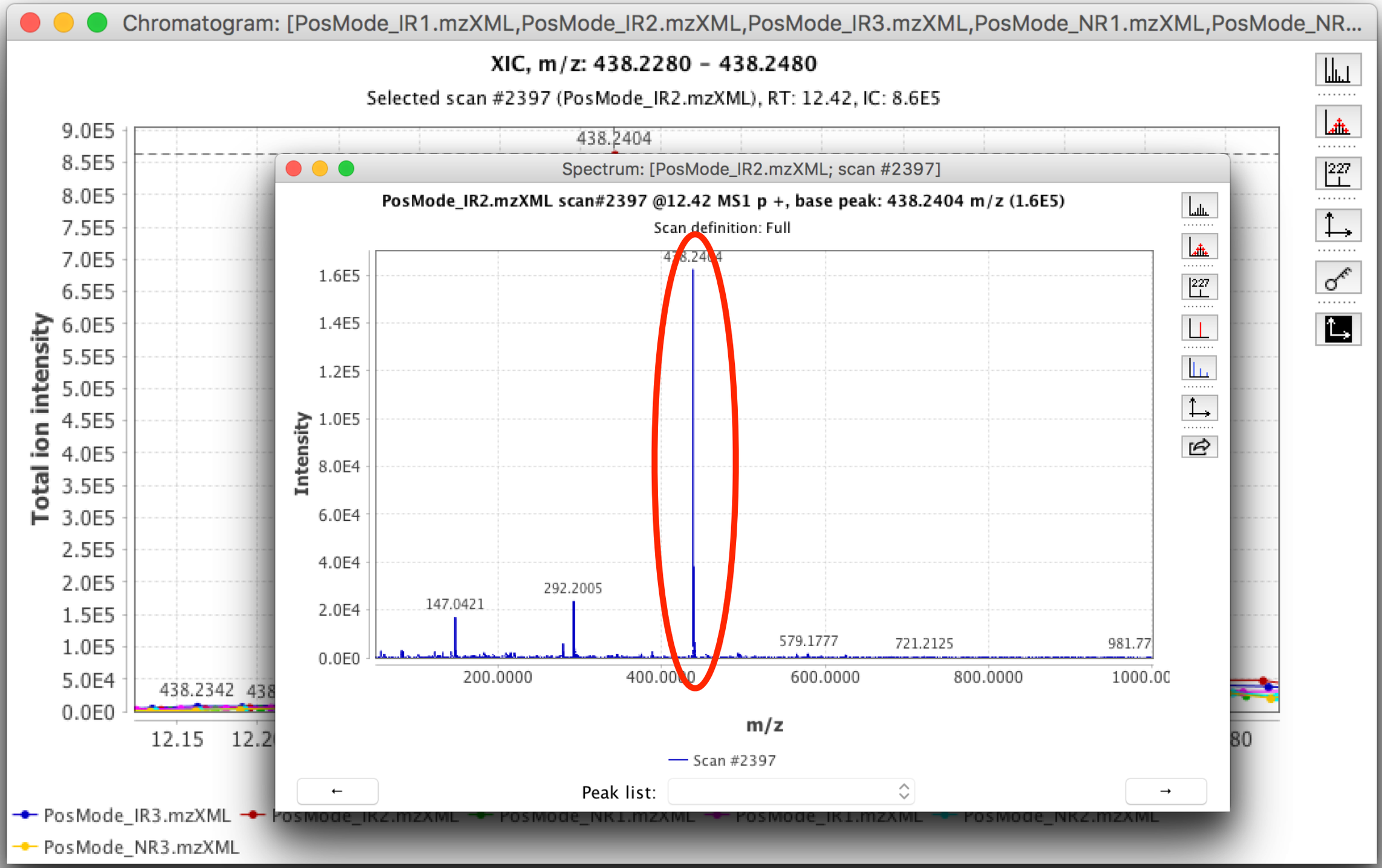
- Apply statistics and machine learning to detect discriminating peaks

- Identify discriminating peaks

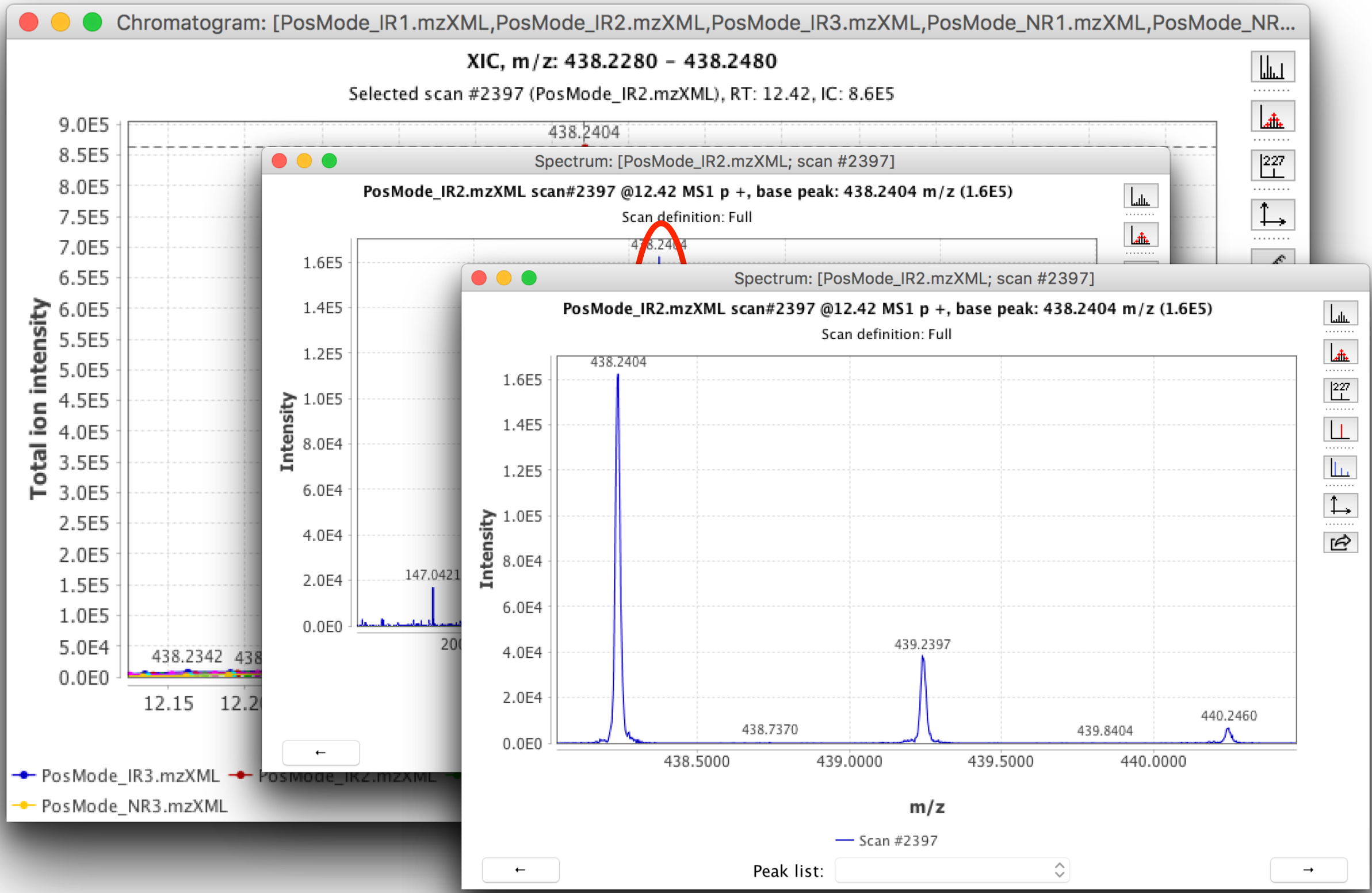
Feature identification



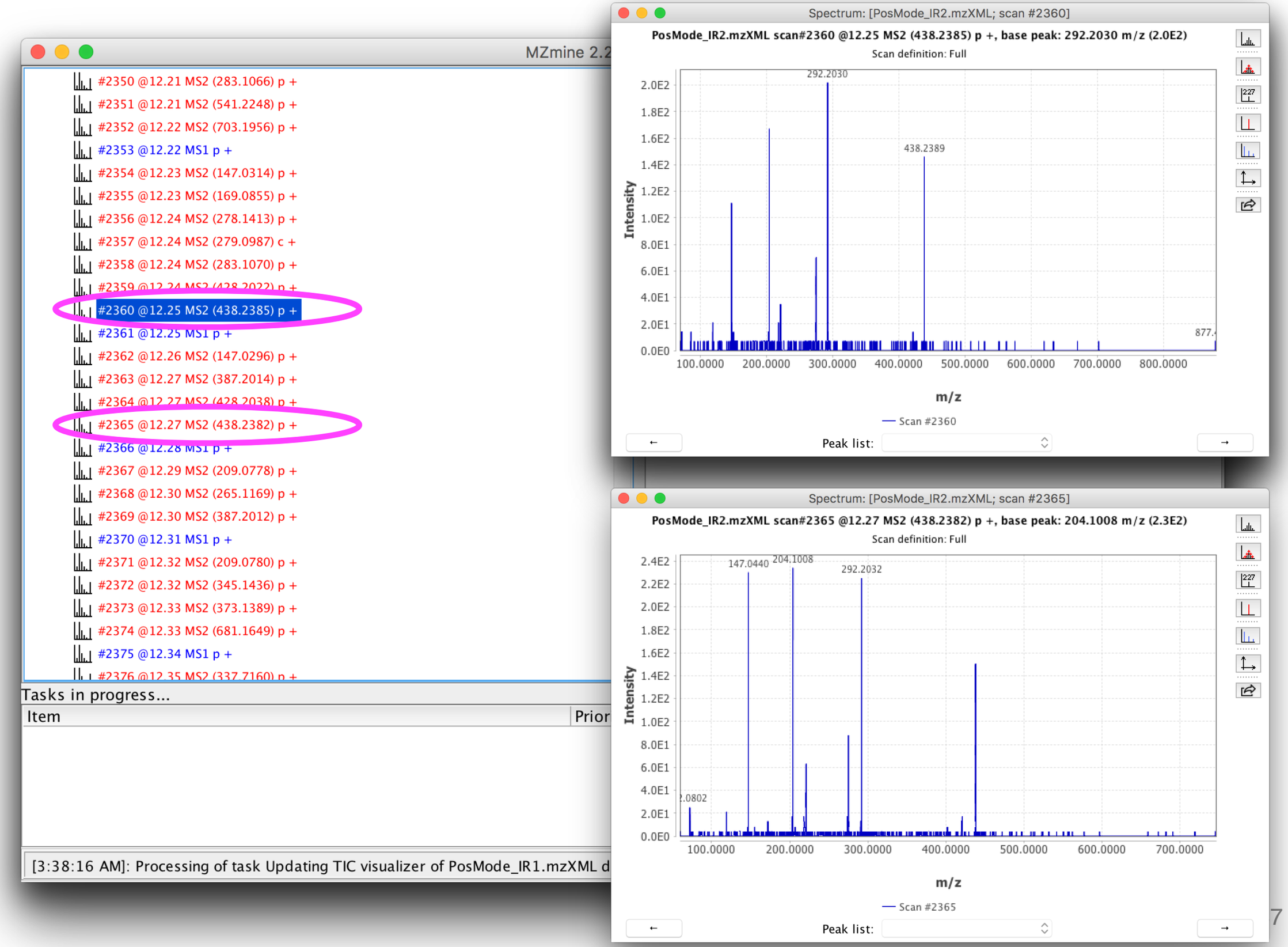
Feature identification



Feature identification



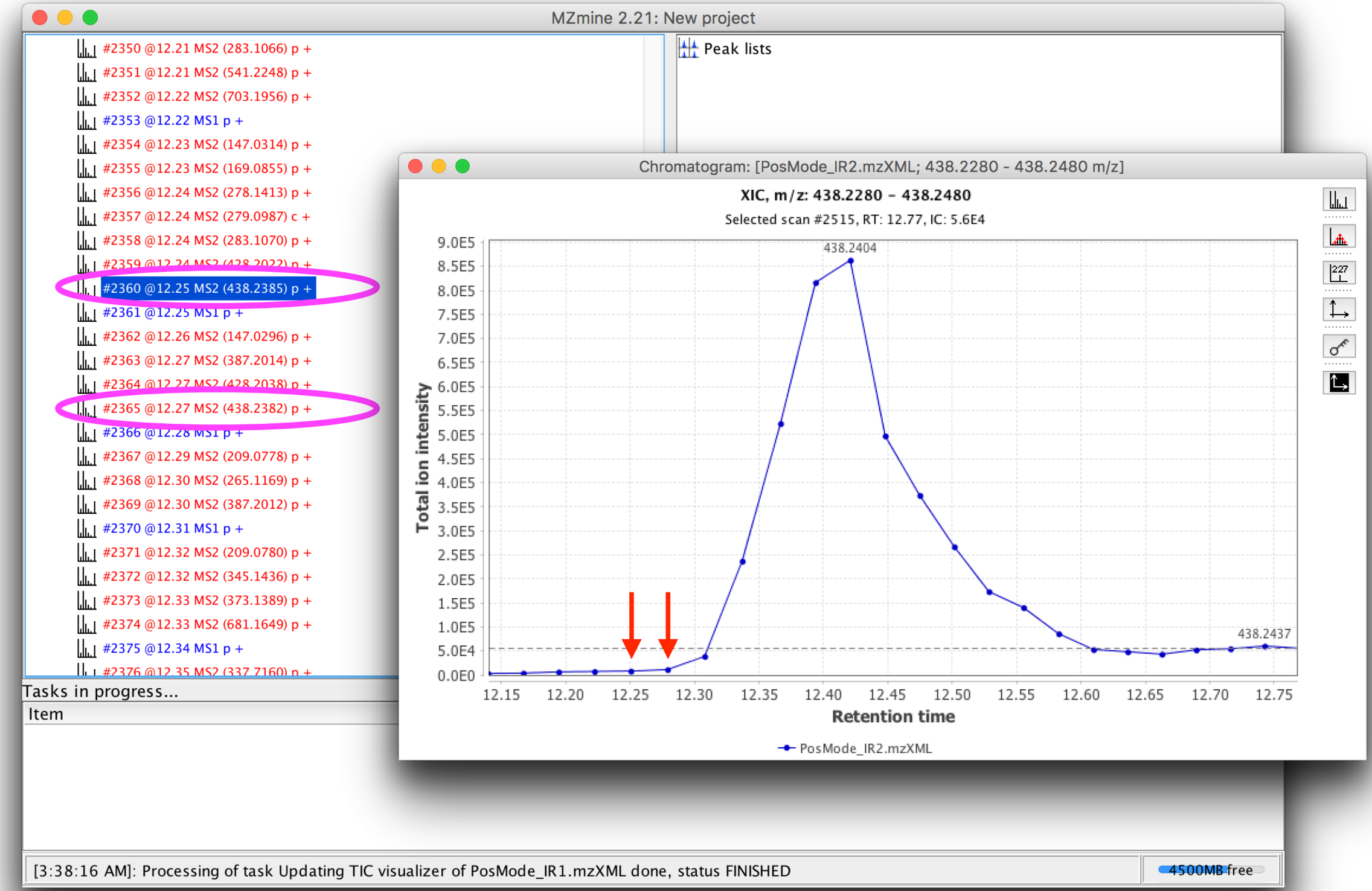
Feature identification



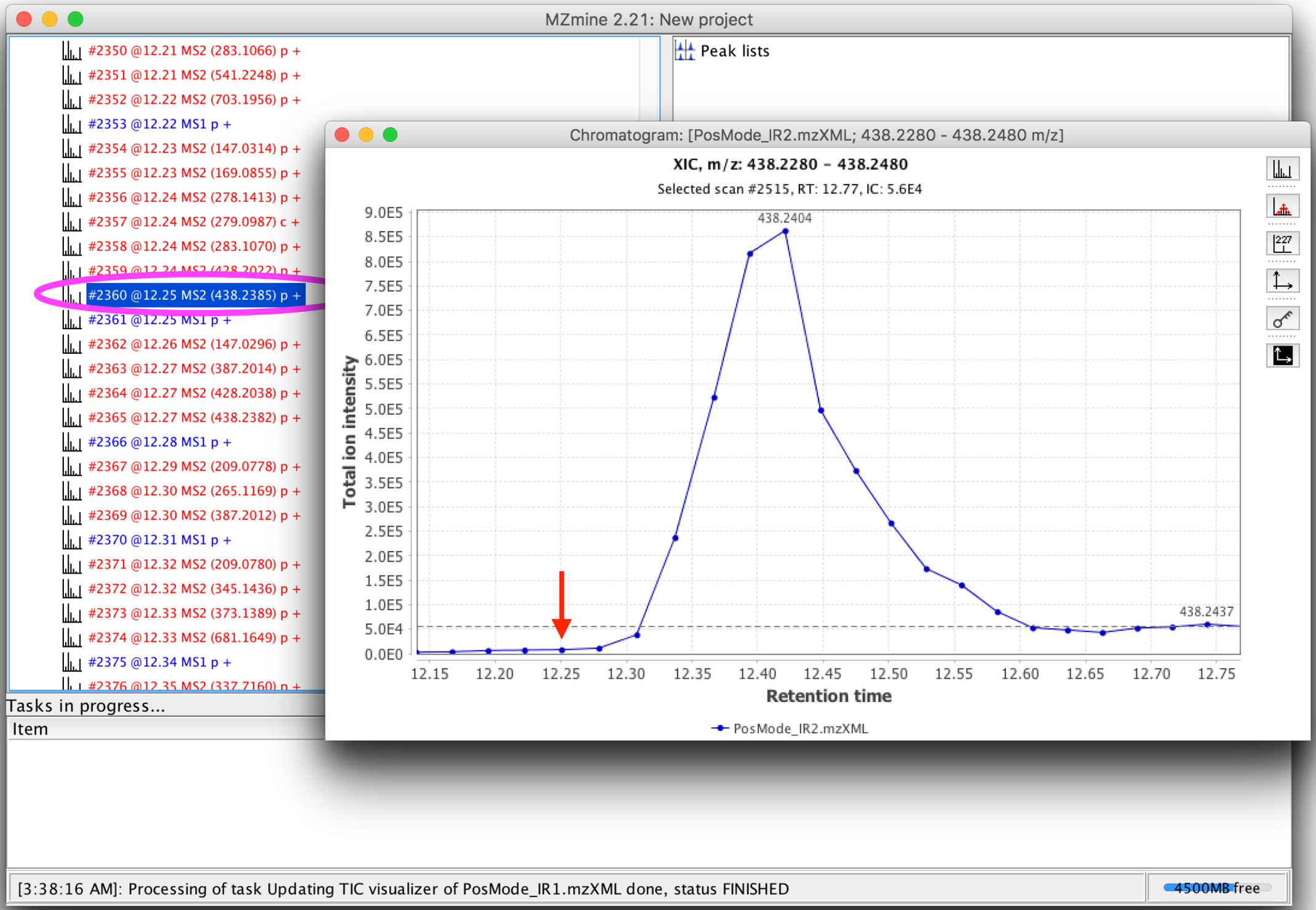
Feature identification

- Information we have for identification
  - M+H
  - Experimental isotopic identification
  - MS/MS

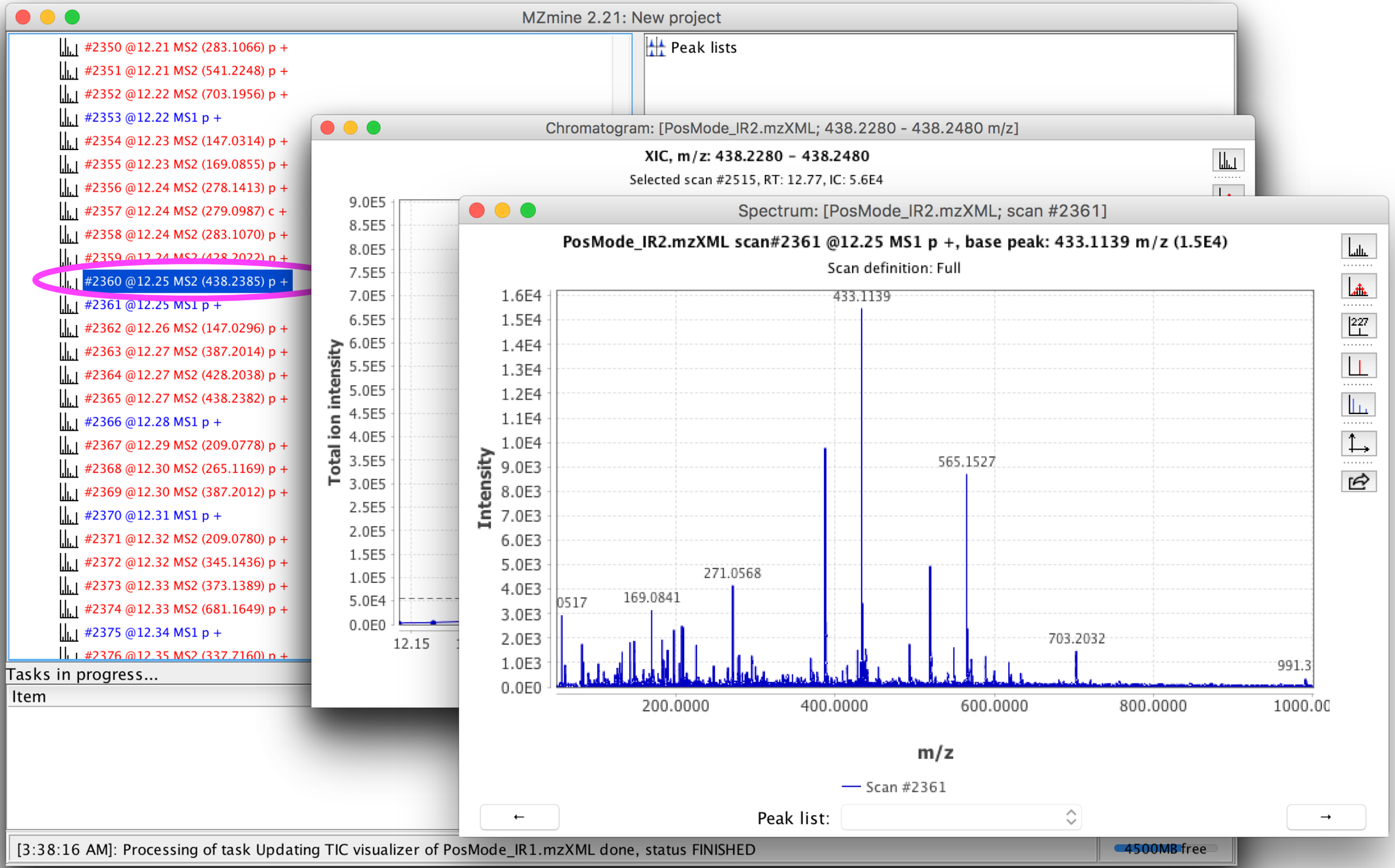
Feature identification



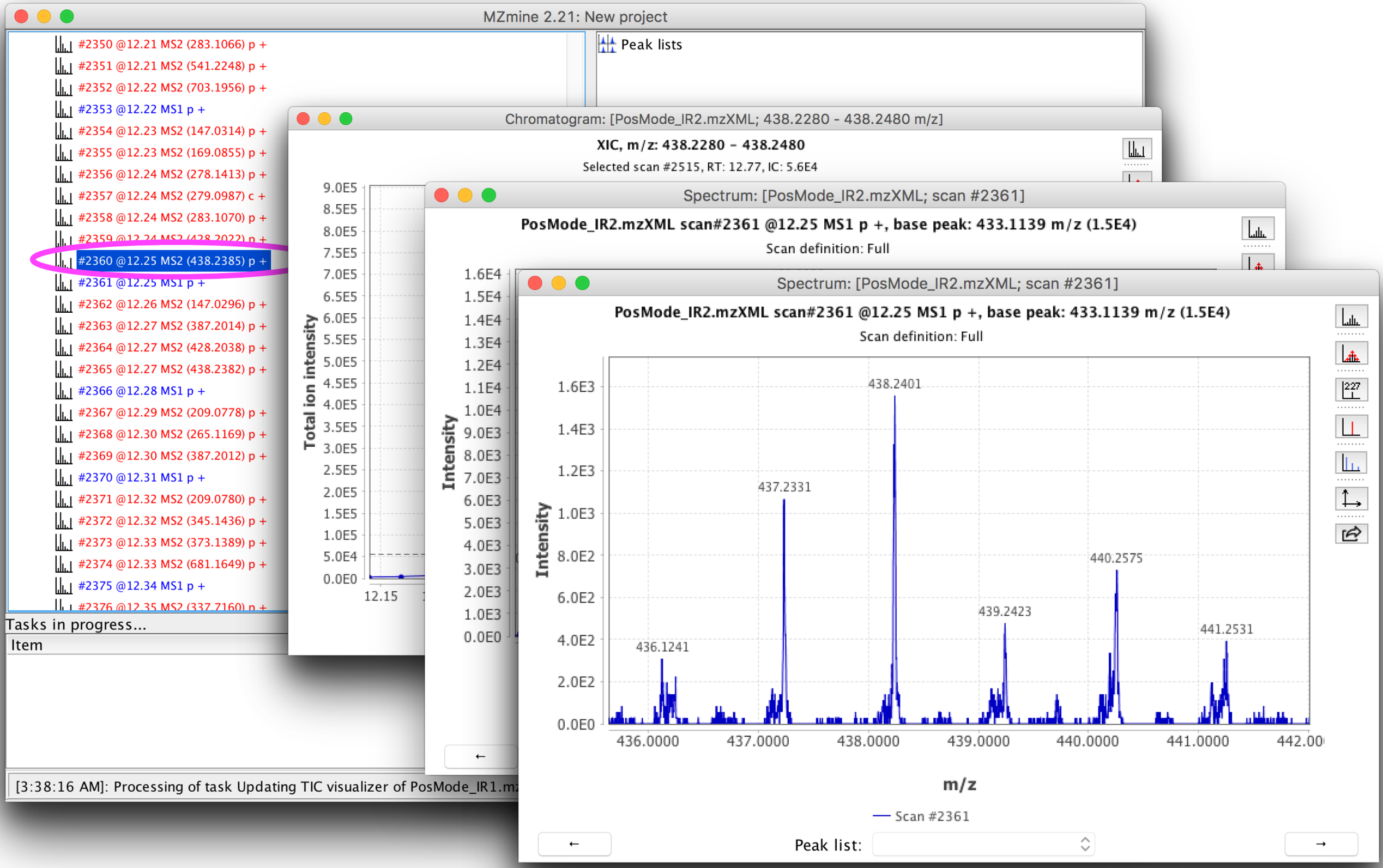
Feature identification



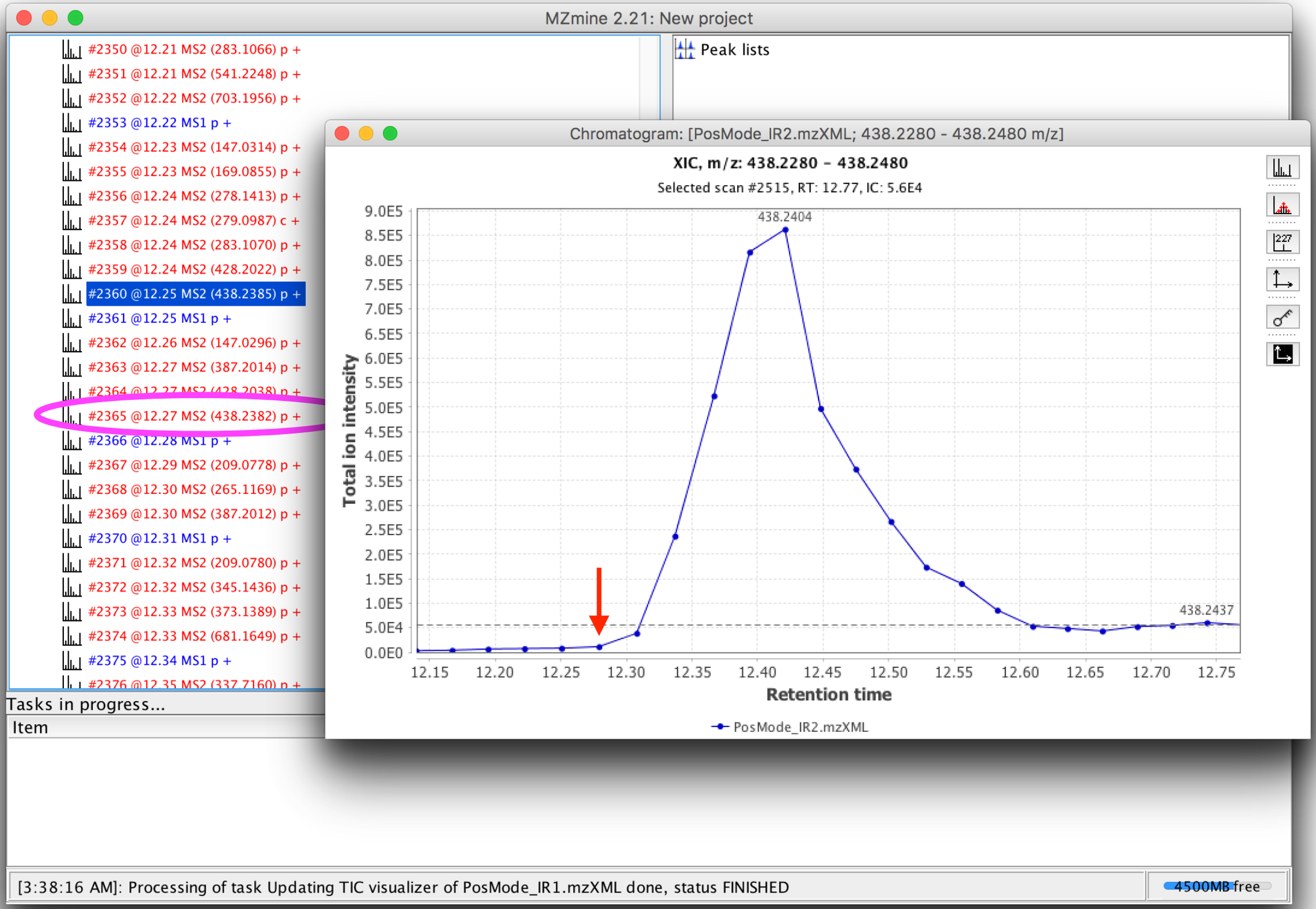
Feature identification



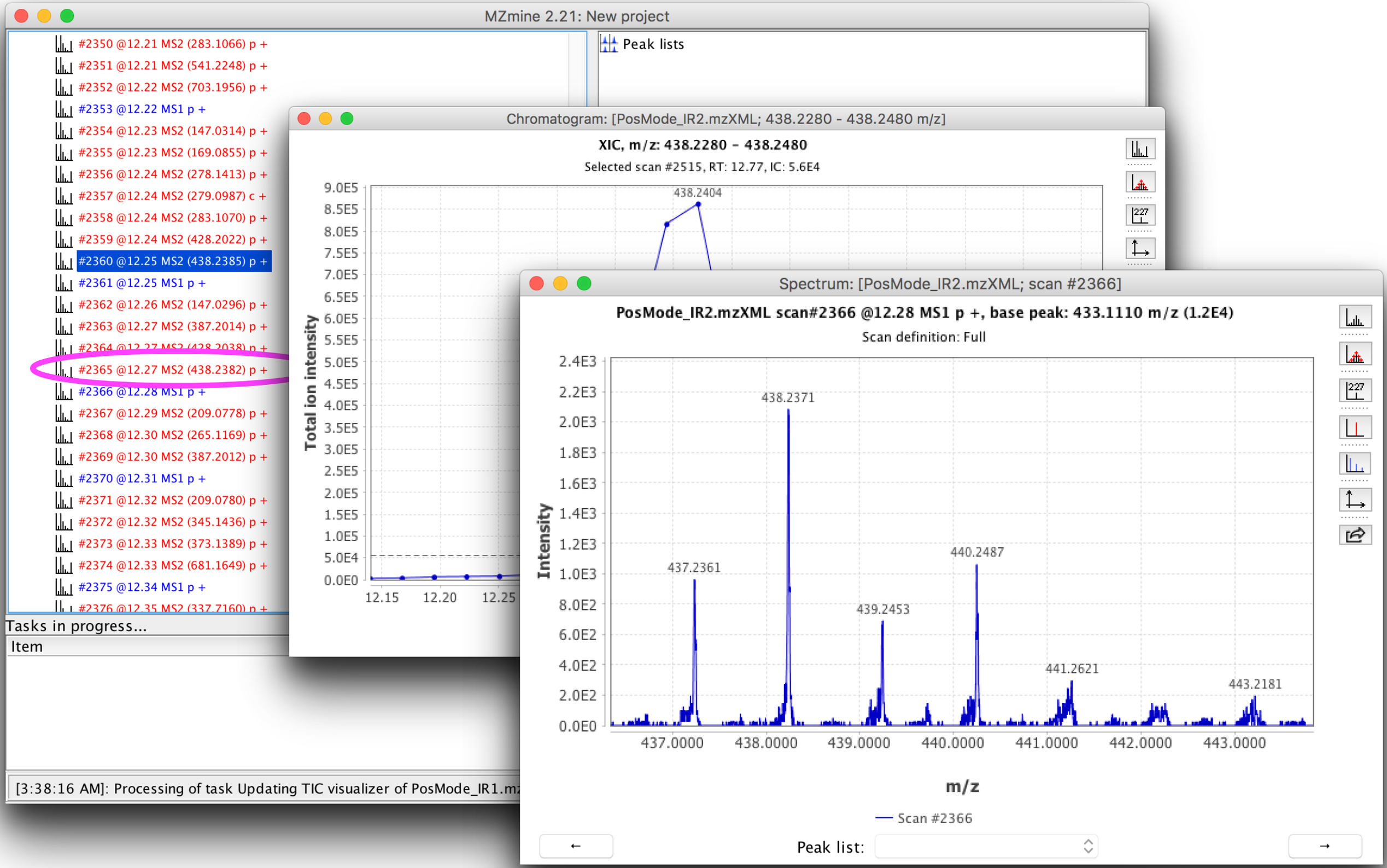
Feature identification



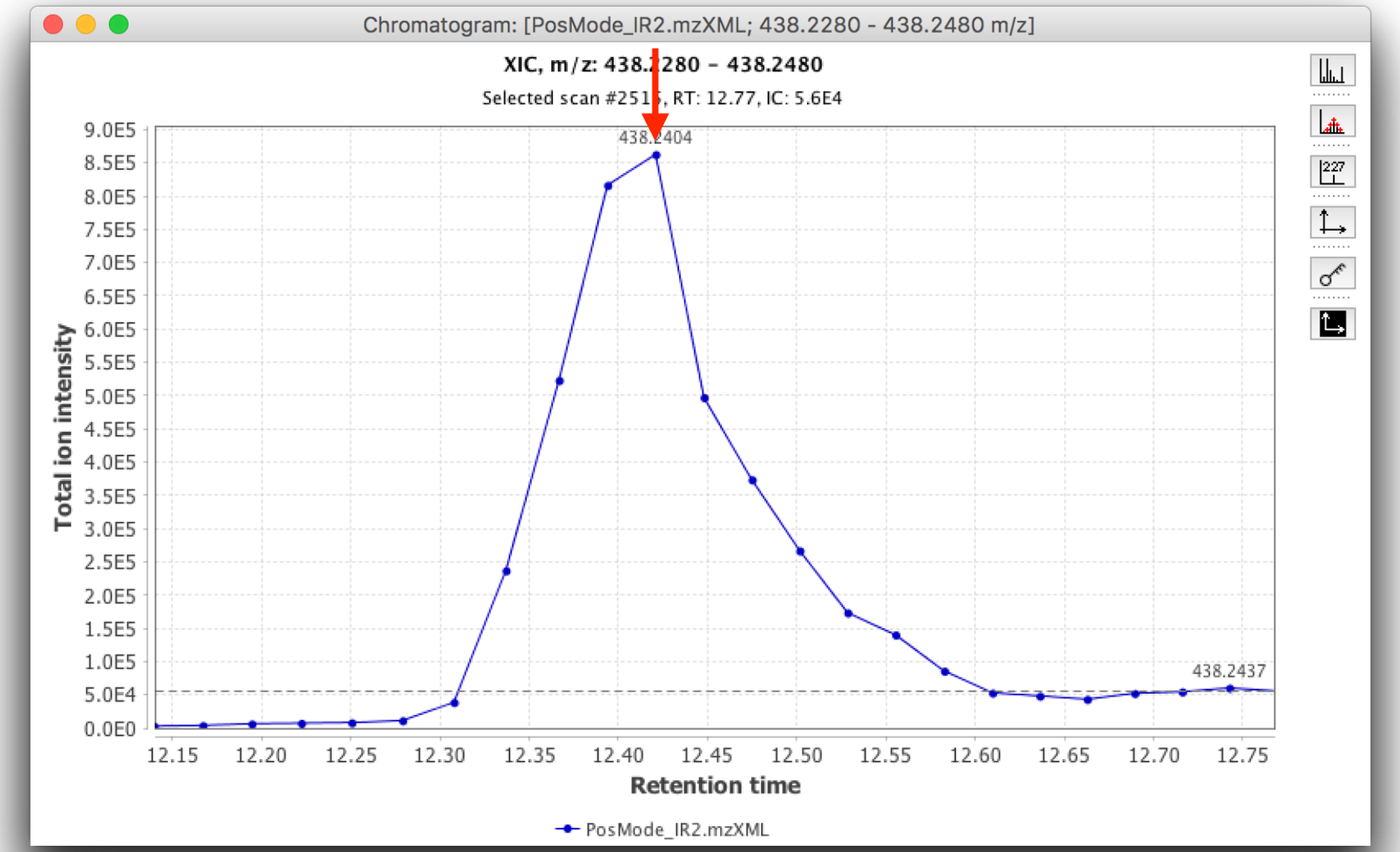
Feature identification



Feature identification



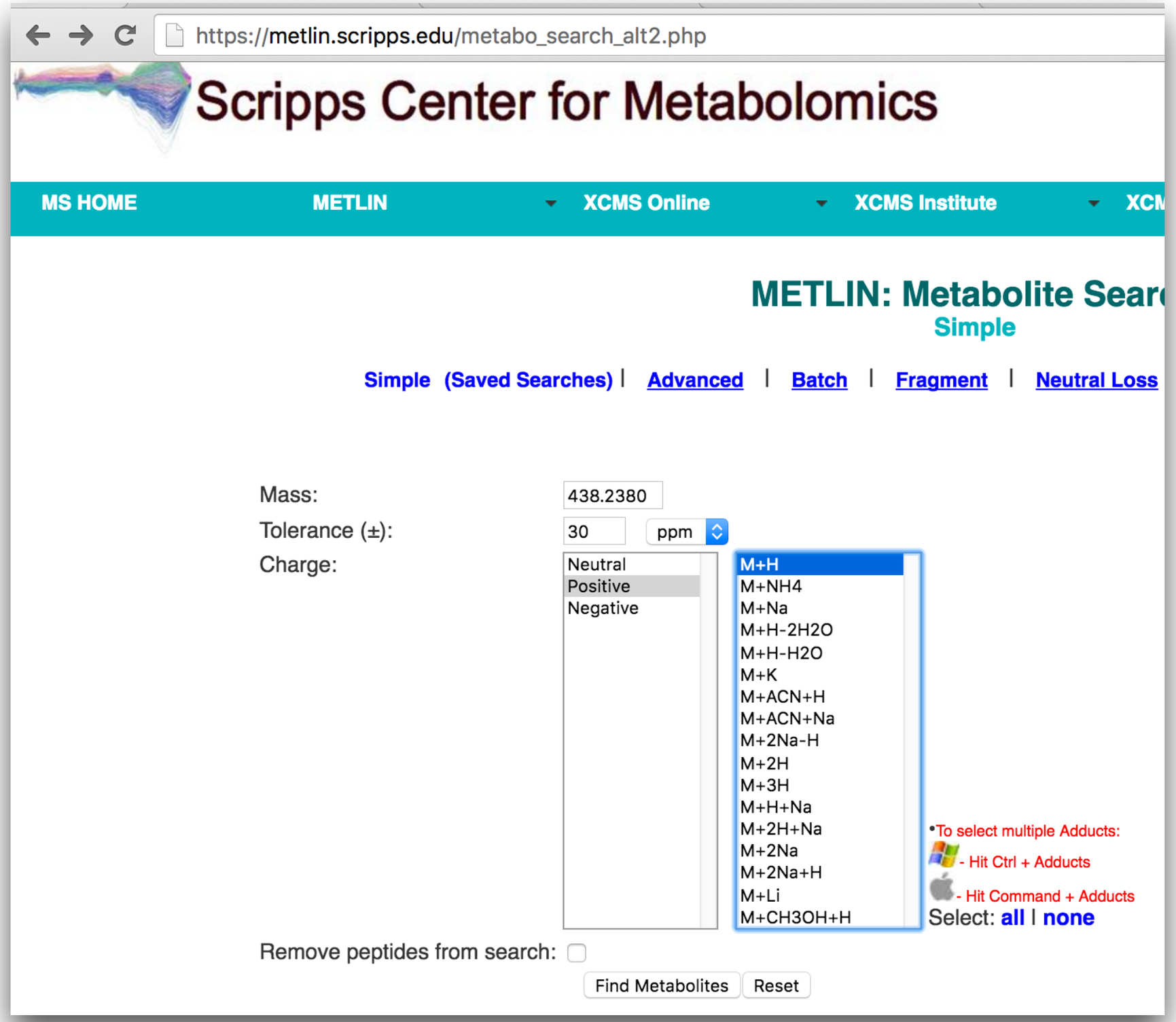
Feature identification



Feature identification

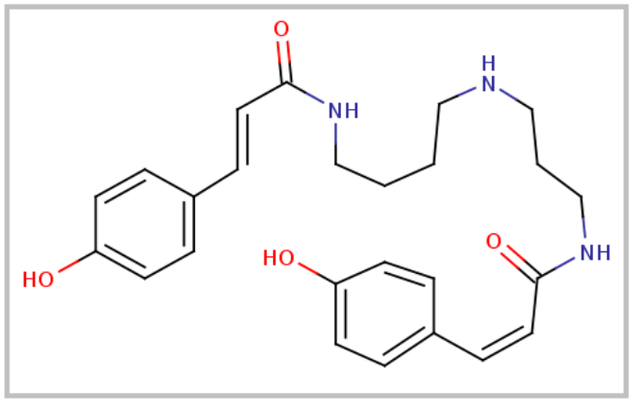
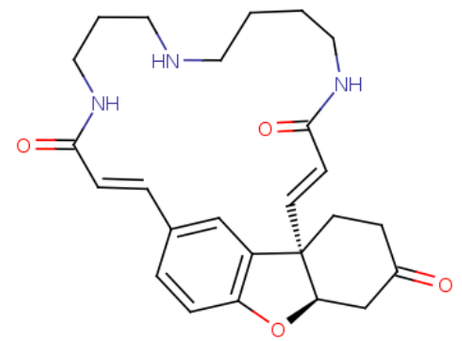
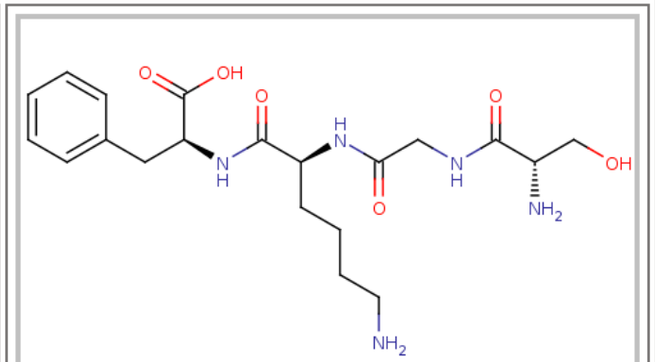
Feature identification

Feature identification



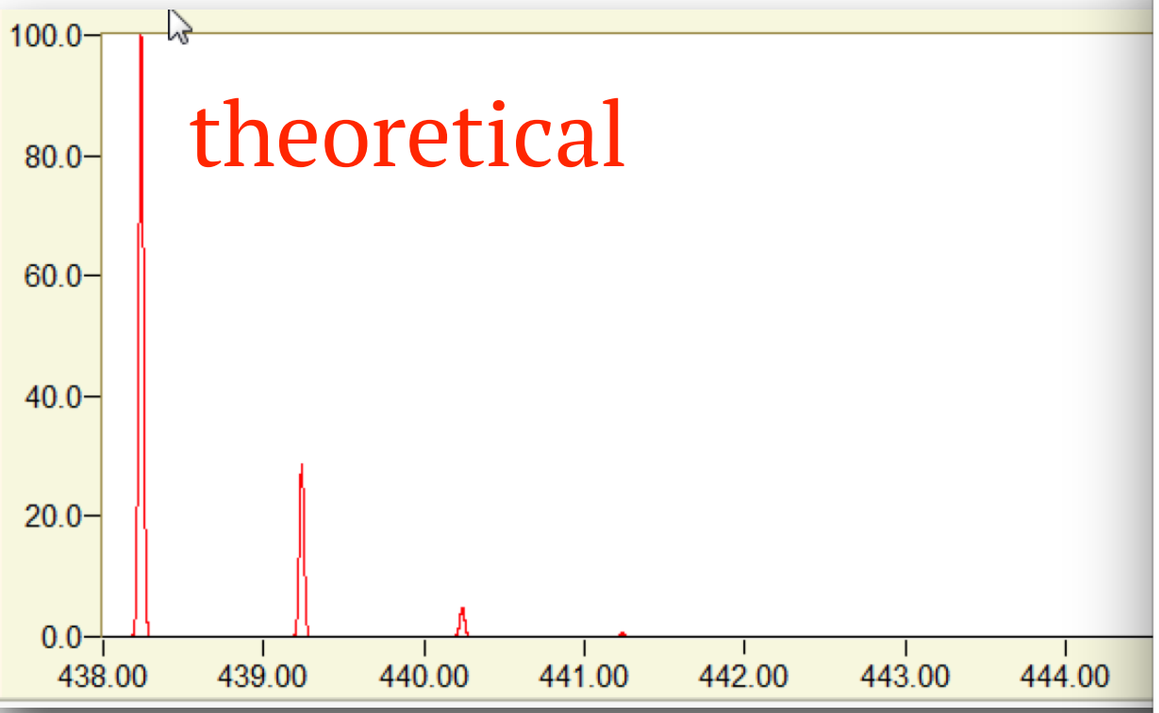
Feature identification

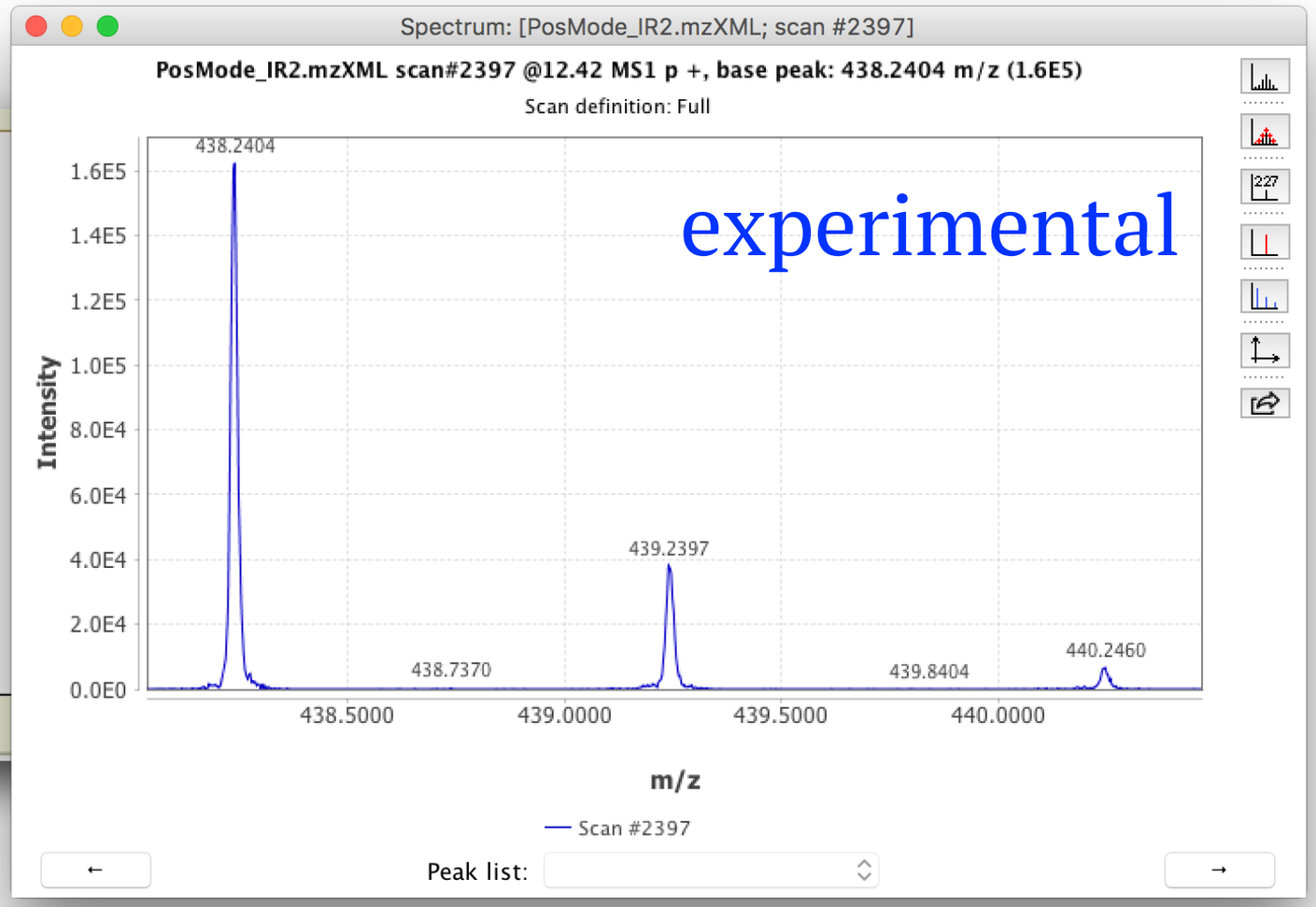
Total: 87 Metabolites

METLIN ID	MASS	Δppm	NAME	MS/MS	STRUCTURE
89296	$[M+H]^+$ **m/z** **438.2387** M 437.2315	1	**N1,N10-Dicoumaroylspermidine** *Formula:* C25H31N3O4 *CAS:* 65715-79-9	NO	
43760	$[M+H]^+$ **m/z** **438.2387** M 437.2315	1	**Lunarine** *Formula:* C25H31N3O4 *CAS:* 24185-51-1	View	
225643	$[M+H]^+$ **m/z** **438.2347** M 437.2274	7	**Ser Gly Lys Phe** *Formula:* C20H31N5O6 *CAS:*	NO	
225567	$[M+H]^+$	7	**Ser Gly Phe Lys**	NO	

Feature identification

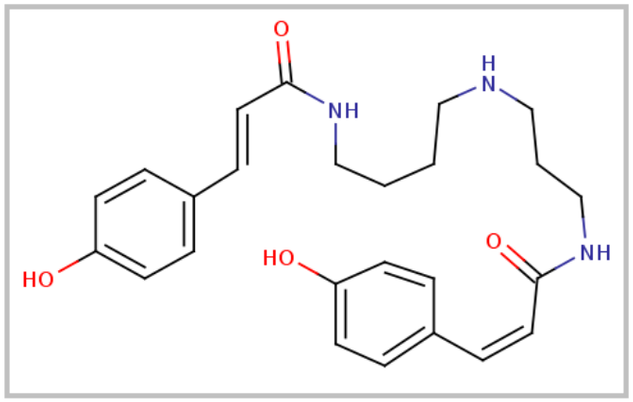
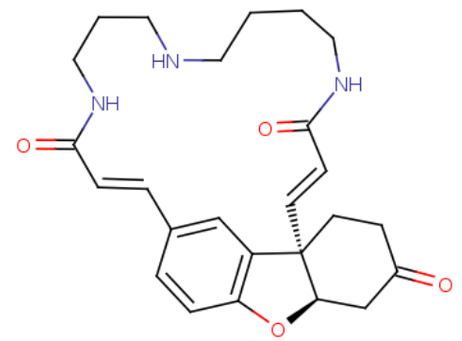
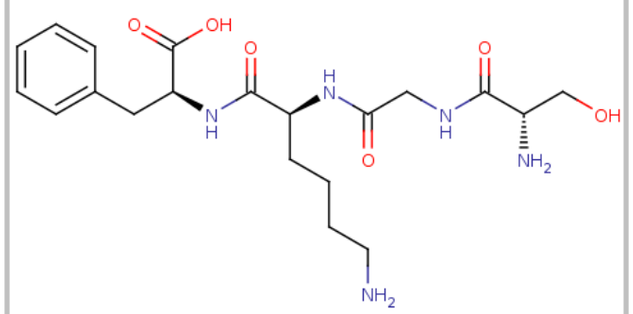
- Compare isotopic distributions





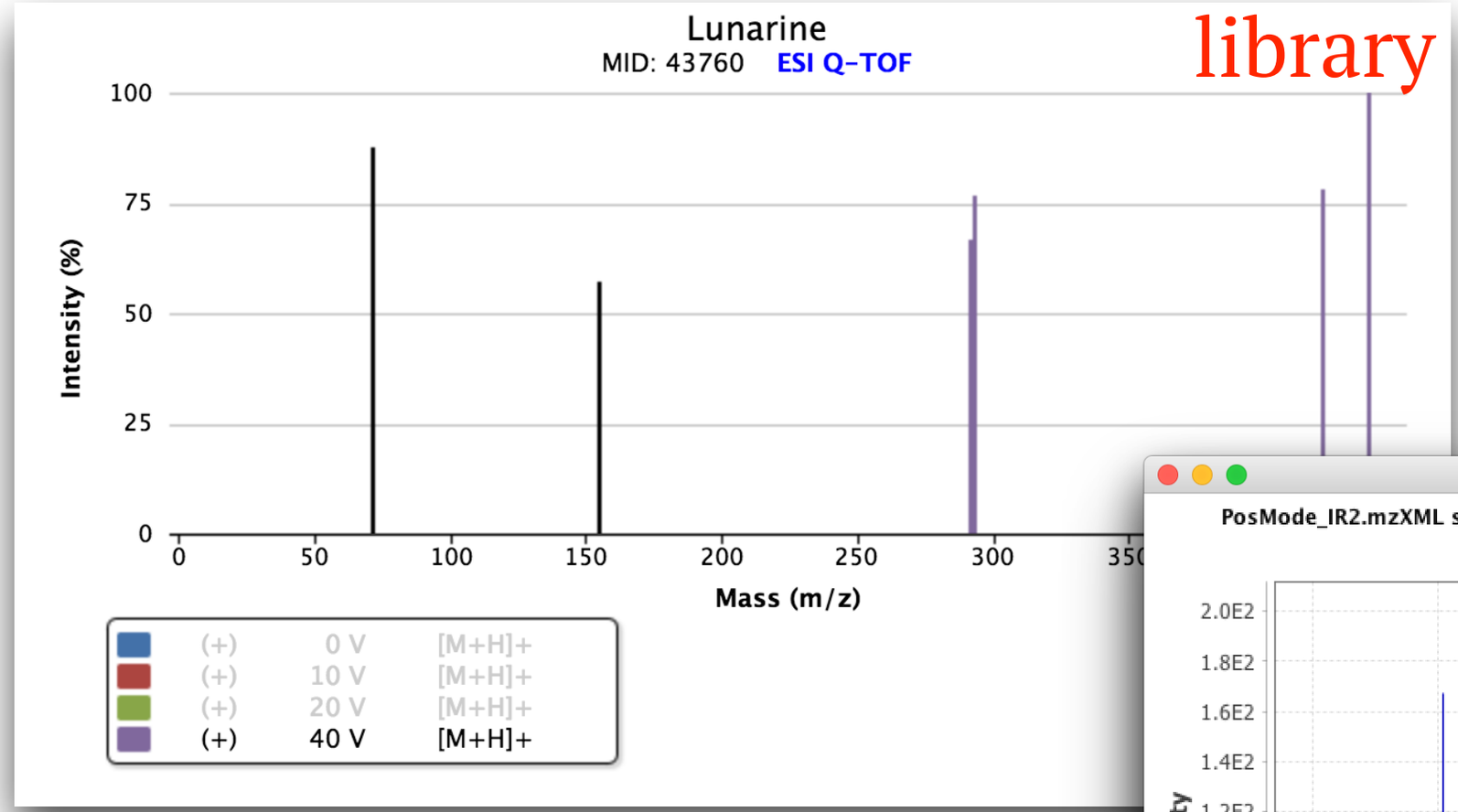
Feature identification

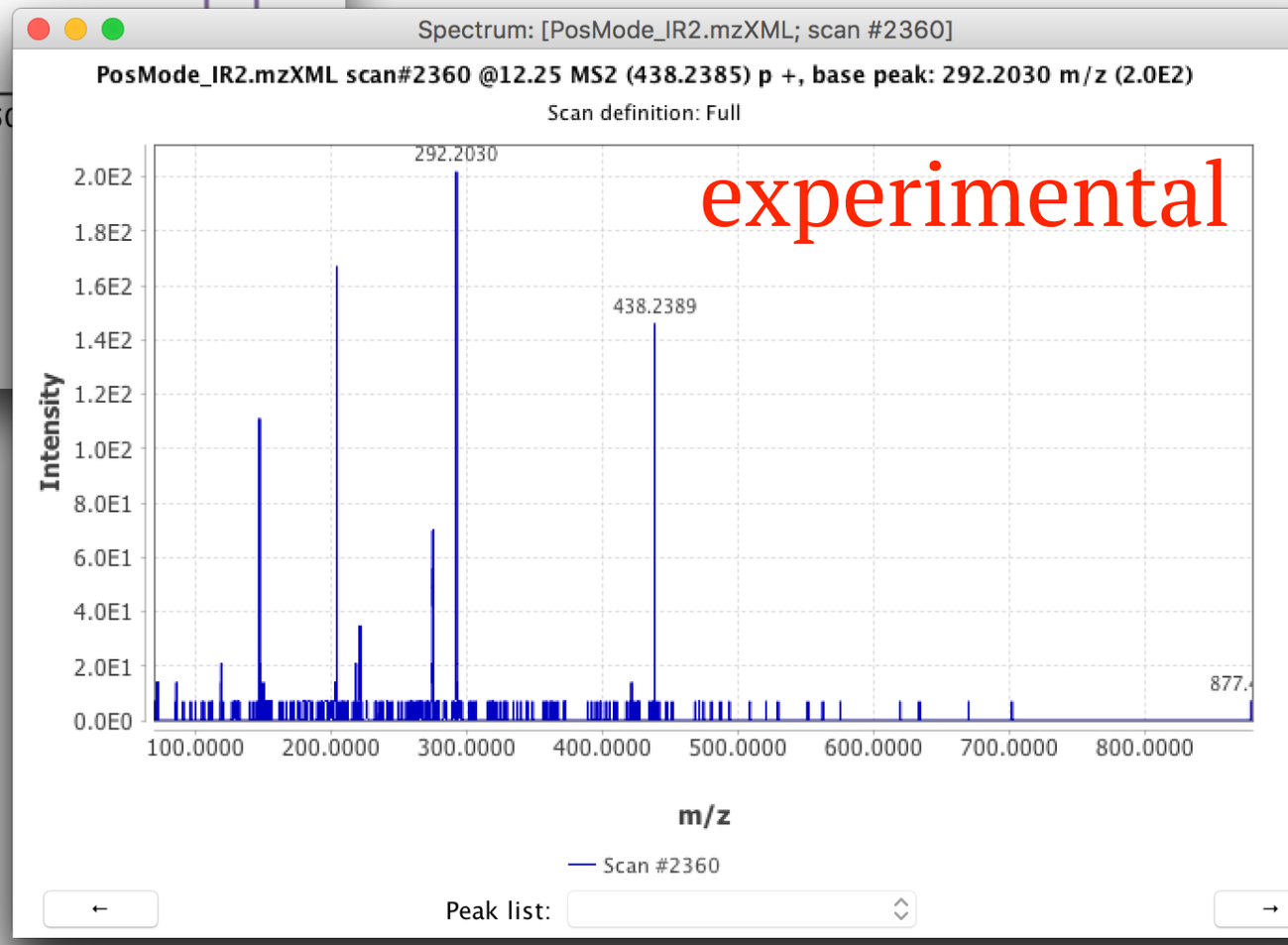
Total: 87 Metabolites

METLIN ID	MASS	Δppm	NAME	MS/MS	STRUCTURE
89296	$[M+H]^+$ **m/z** **438.2387** M 437.2315	1	**N1,N10-Dicoumaroylspermidine** *Formula:* C25H31N3O4 *CAS:* 65715-79-9	NO	
43760	$[M+H]^+$ **m/z** **438.2387** M 437.2315	1	**Lunarine** *Formula:* C25H31N3O4 *CAS:* 24185-51-1	View	
225643	$[M+H]^+$ **m/z** **438.2347** M 437.2274	7	**Ser Gly Lys Phe** *Formula:* C20H31N5O6 *CAS:*	NO	
225567	$[M+H]^+$	7	**Ser Gly Phe Lys**	NO	

Feature identification

- CompareMS/MS





Thank you!