

# 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

list of scans in raw files

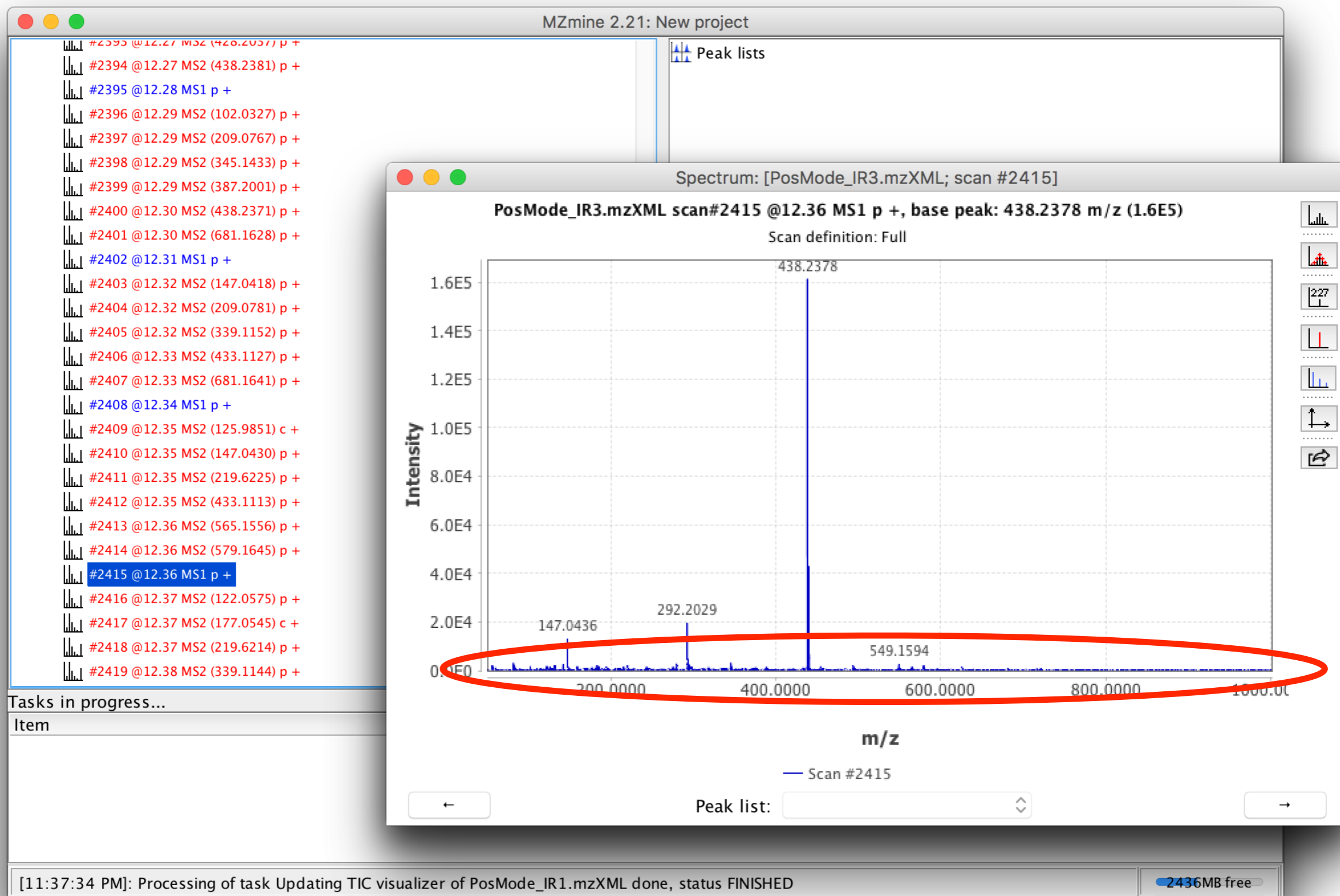
- MS scans in blue
- MS/MS scans in red
- # sequential number
- @ retention time
- MS level
- type of spectrum
  - p = profile
  - c = centroid
  - t = thresholded
- polarity of ionization
  - + = positive
  - - = negative
  - ? = unknown

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

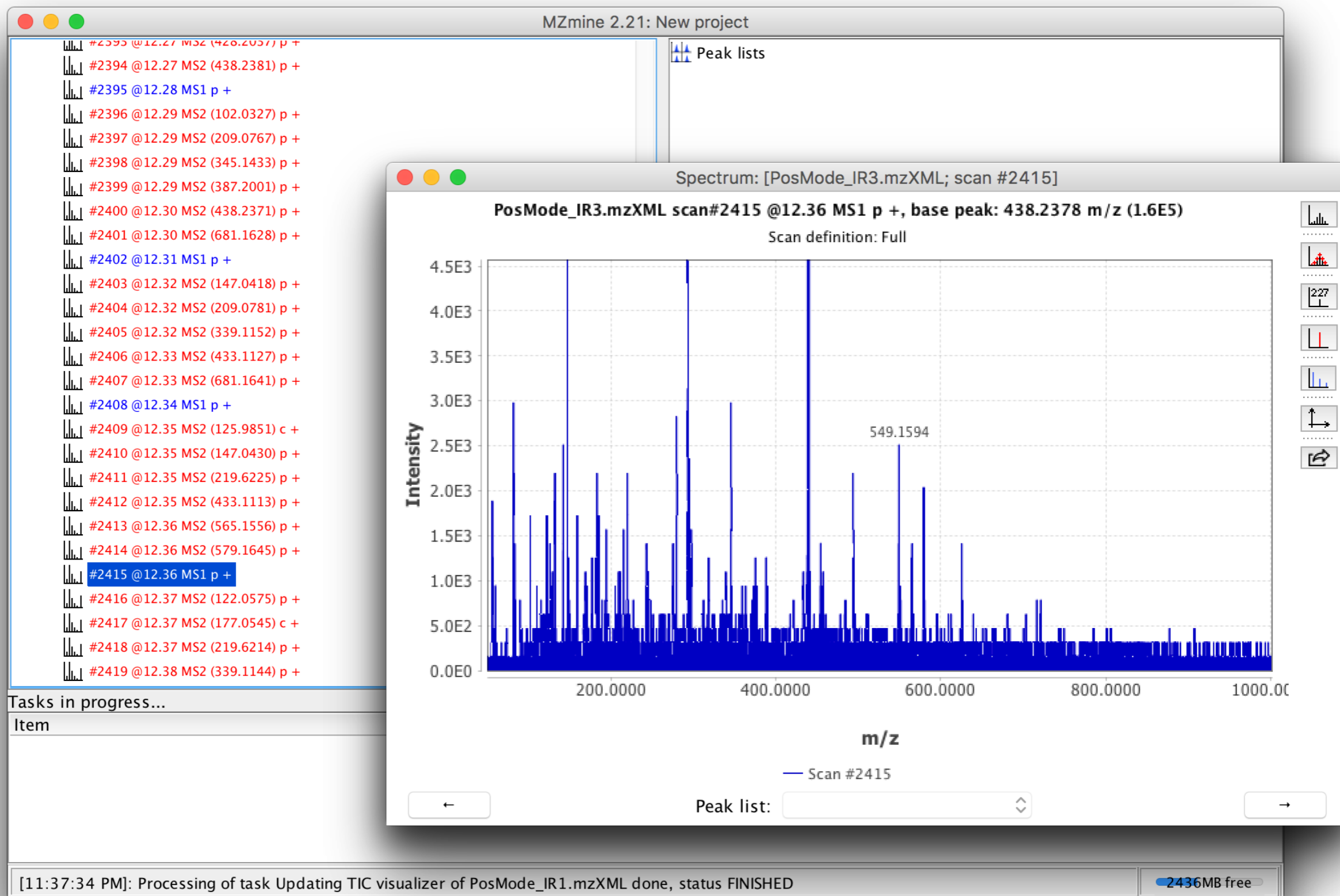
[11:37:34 PM]: Processing of task Updating TIC visualizer of PosMode\_IR1.mzXML done, status FINISHED

3248MB free

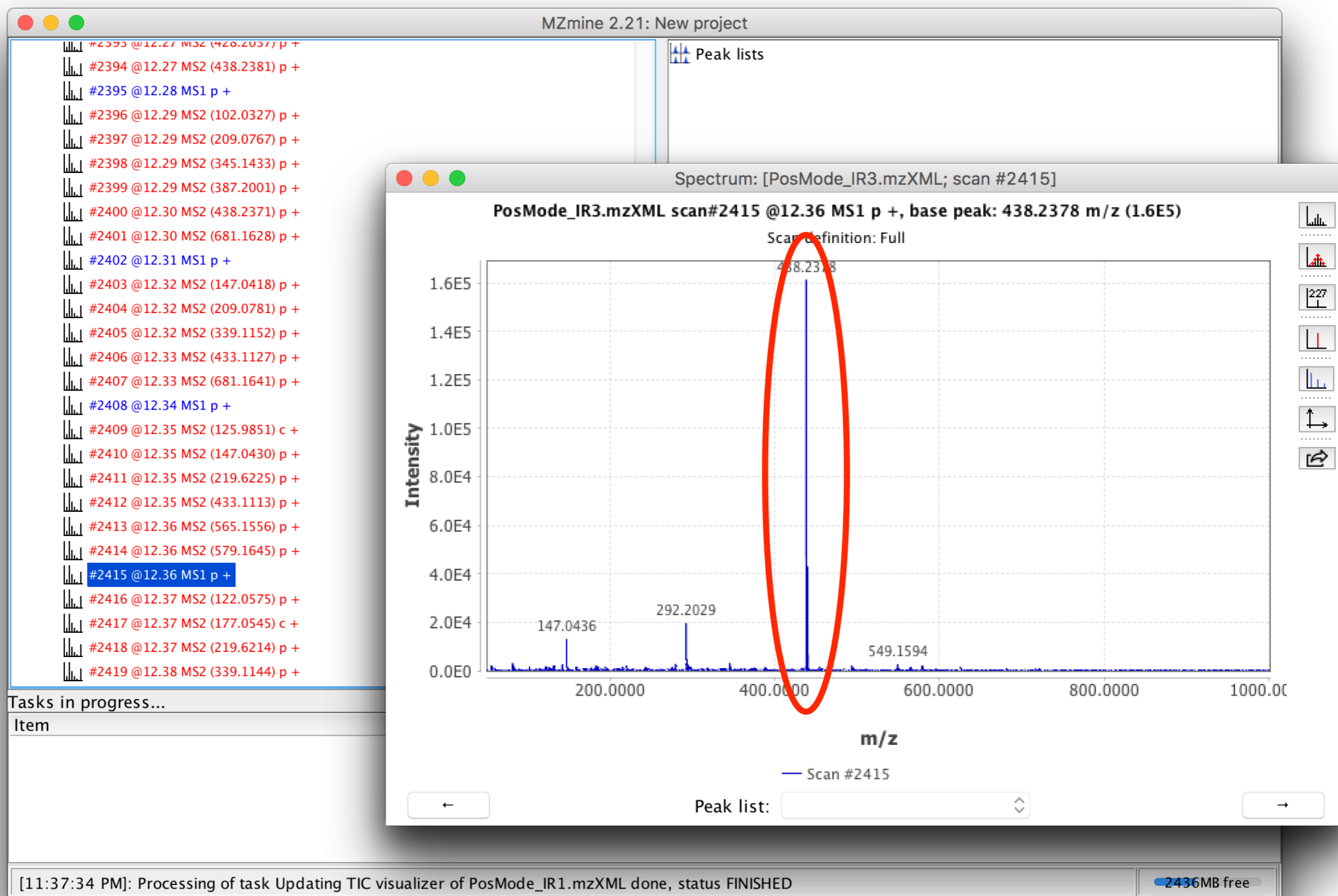
# One mass spectrum



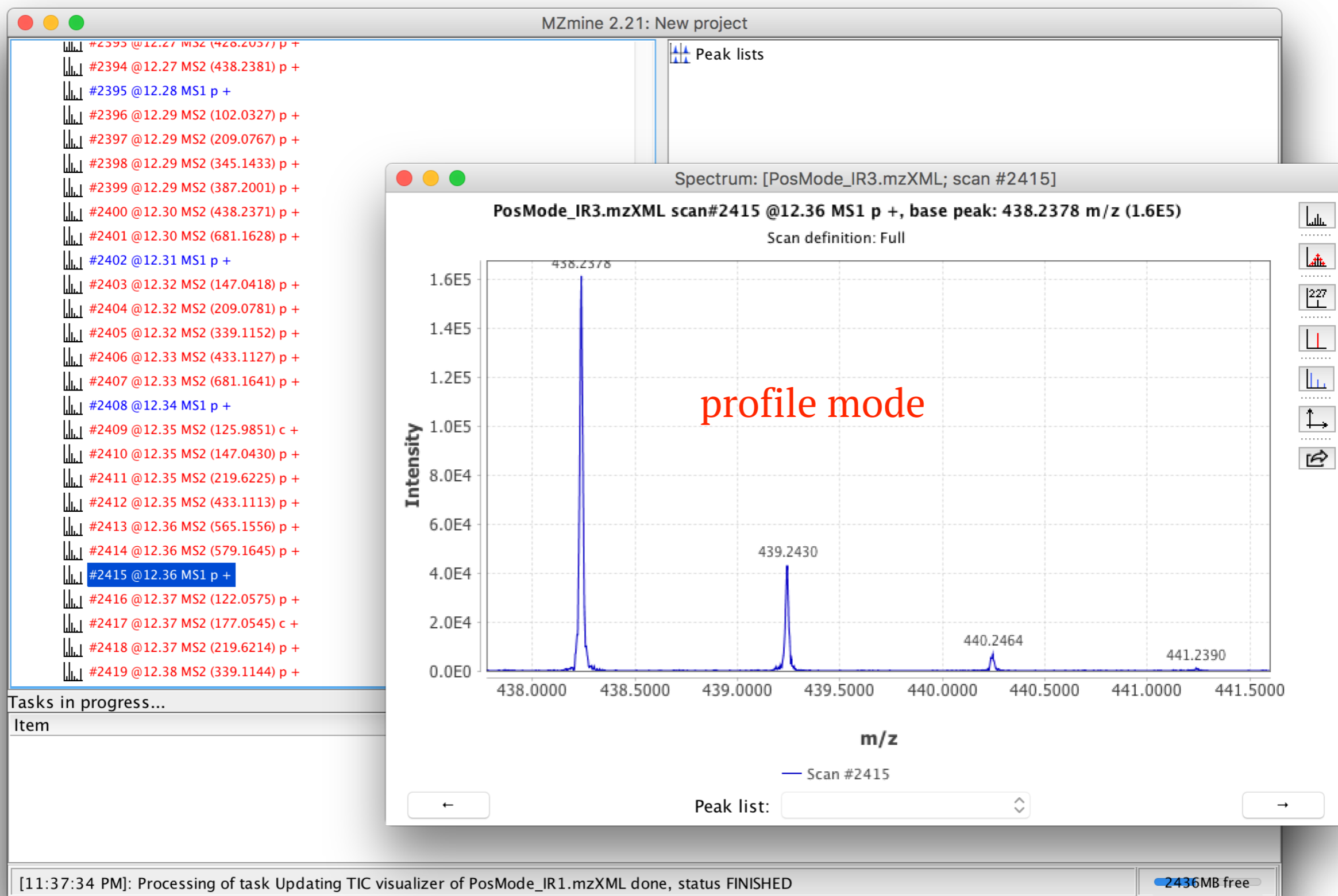
# One mass spectrum



# One mass spectrum



# Zoom in one mass spectrum





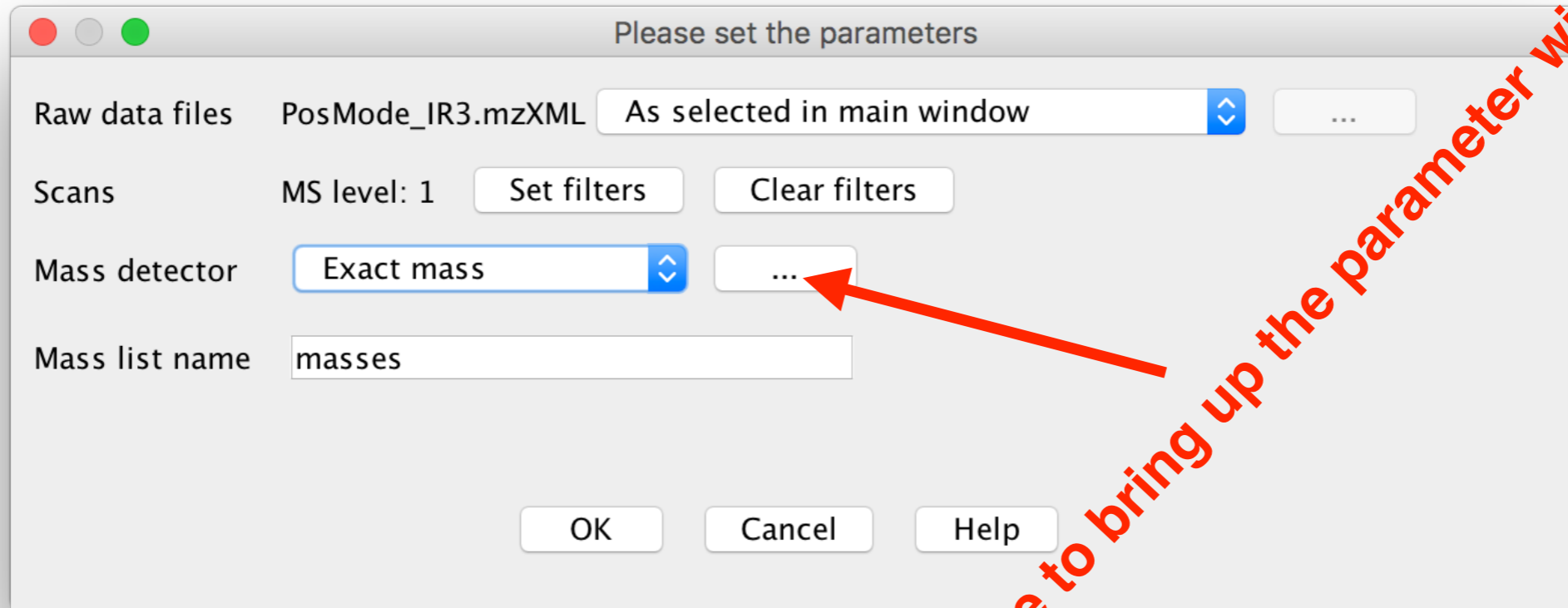
# Conversion to centroid mode

The screenshot shows the MZmine 2.21 software interface. The 'Raw data methods' menu is open, showing options for 'Raw data import', 'Order raw data files', 'Filtering', and 'Peak detection'. The 'Peak detection' sub-menu is also open, showing options for 'Mass detection', 'FTMS shoulder peaks filter', 'Chromatogram builder', 'GridMass - 2D peak detection', 'MS/MS peaklist builder', and 'Targeted peak detection'. The 'Mass detection' option is highlighted. The interface also shows a 'Peak lists' panel on the right and a 'Tasks in progress...' table at the bottom.

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

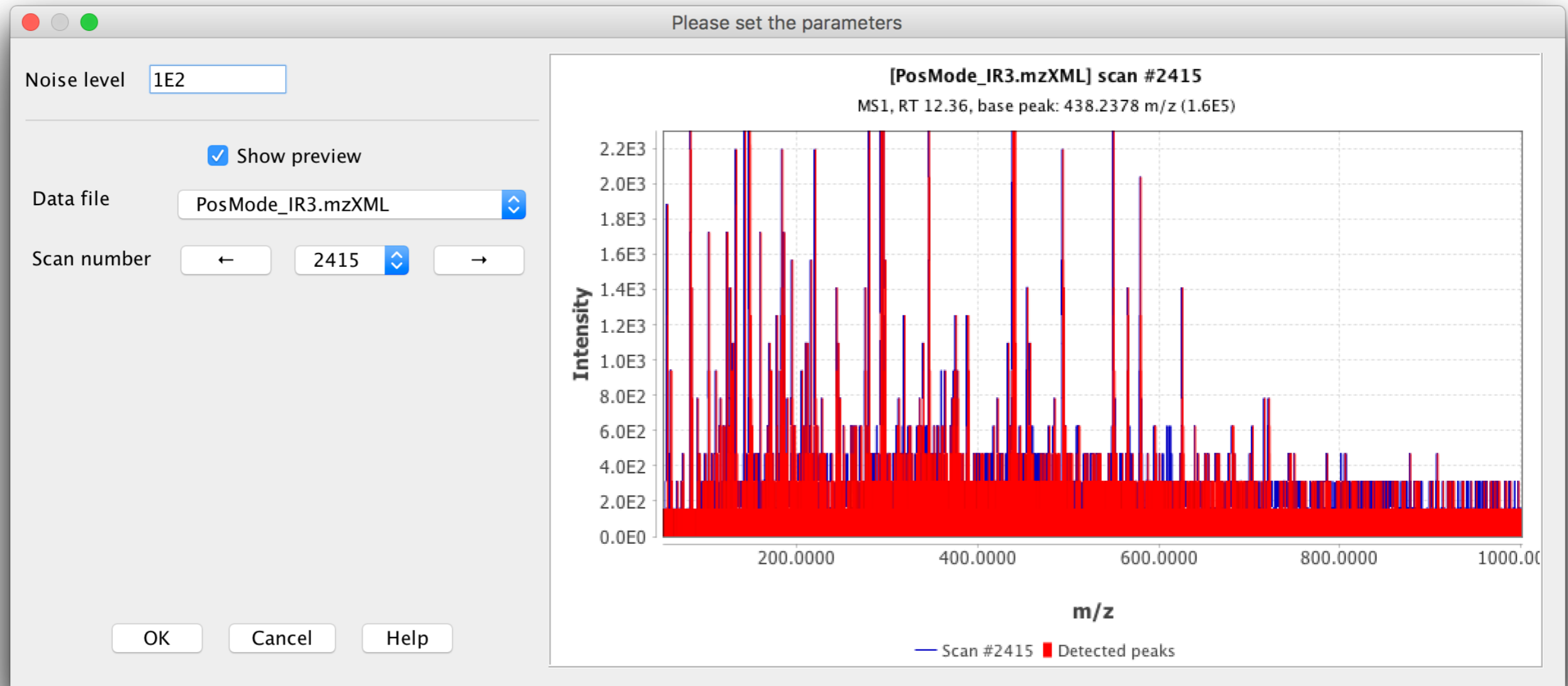
2601MB free

# Conversion to centroid mode



*click here to bring up the parameter window*

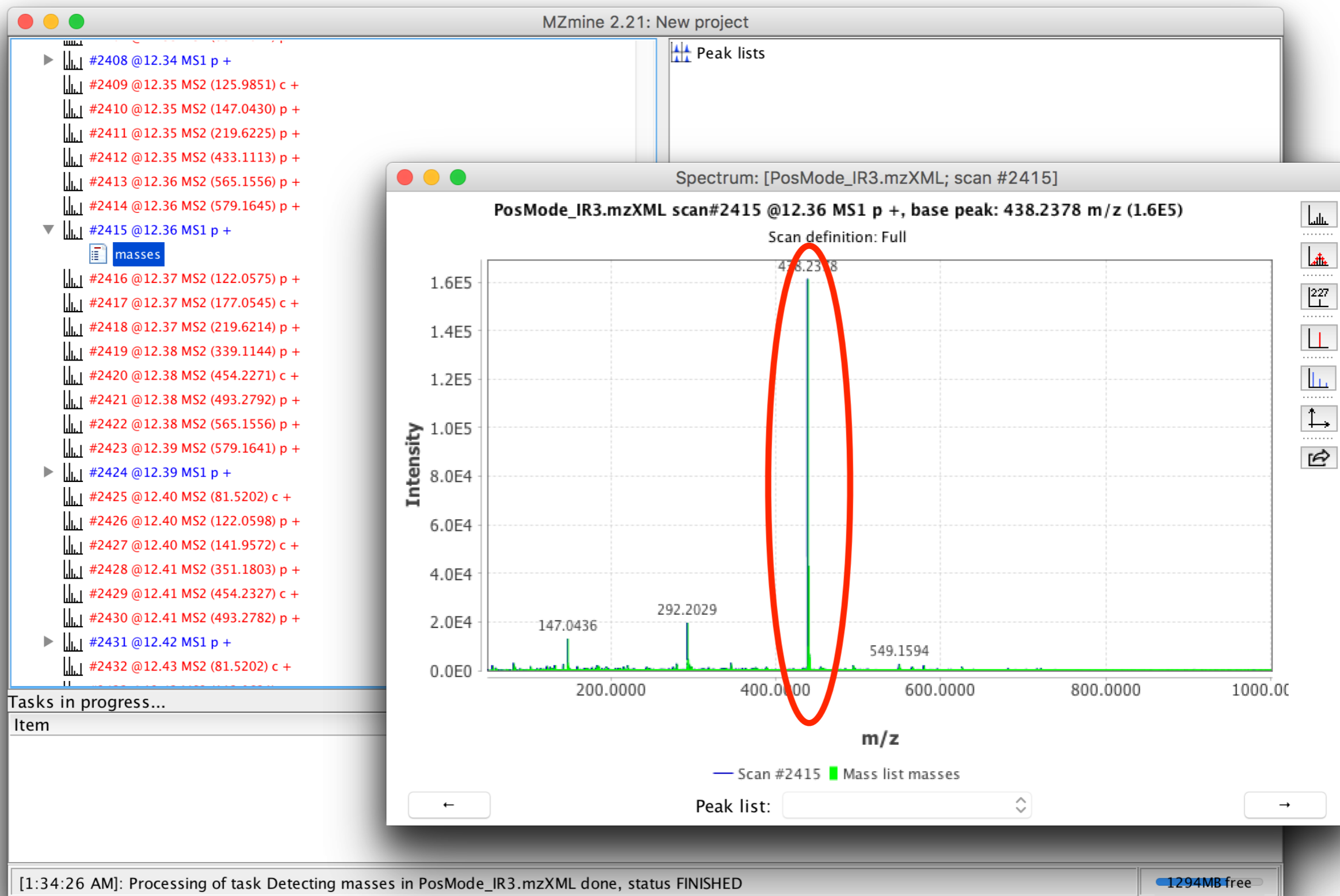
# Conversion to centroid mode



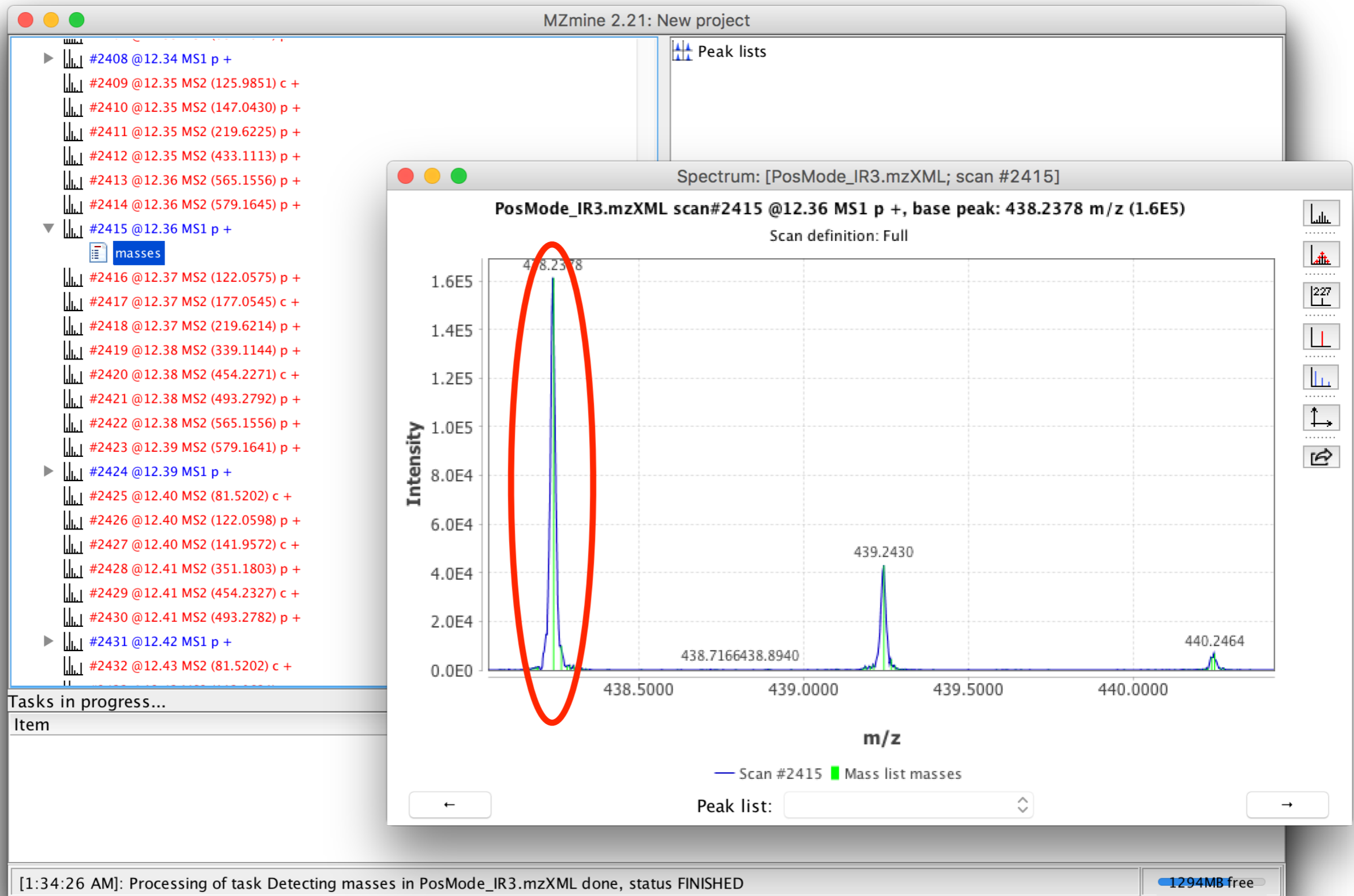
# Conversion to centroid mode

Mass detection in progress .....

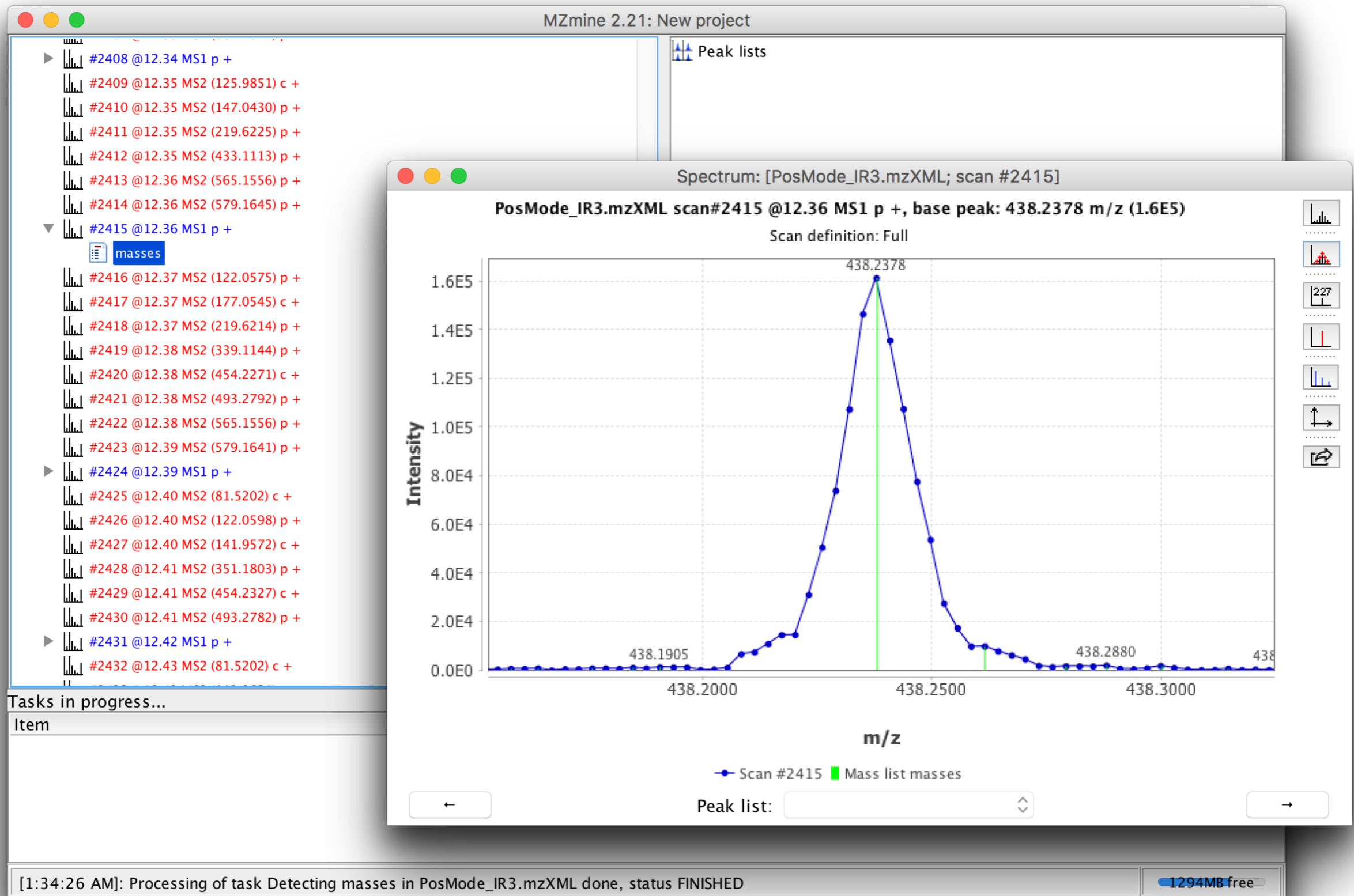
# Mass spectra in centroid mode



# Mass spectra in centroid mode



# Mass spectra in centroid mode

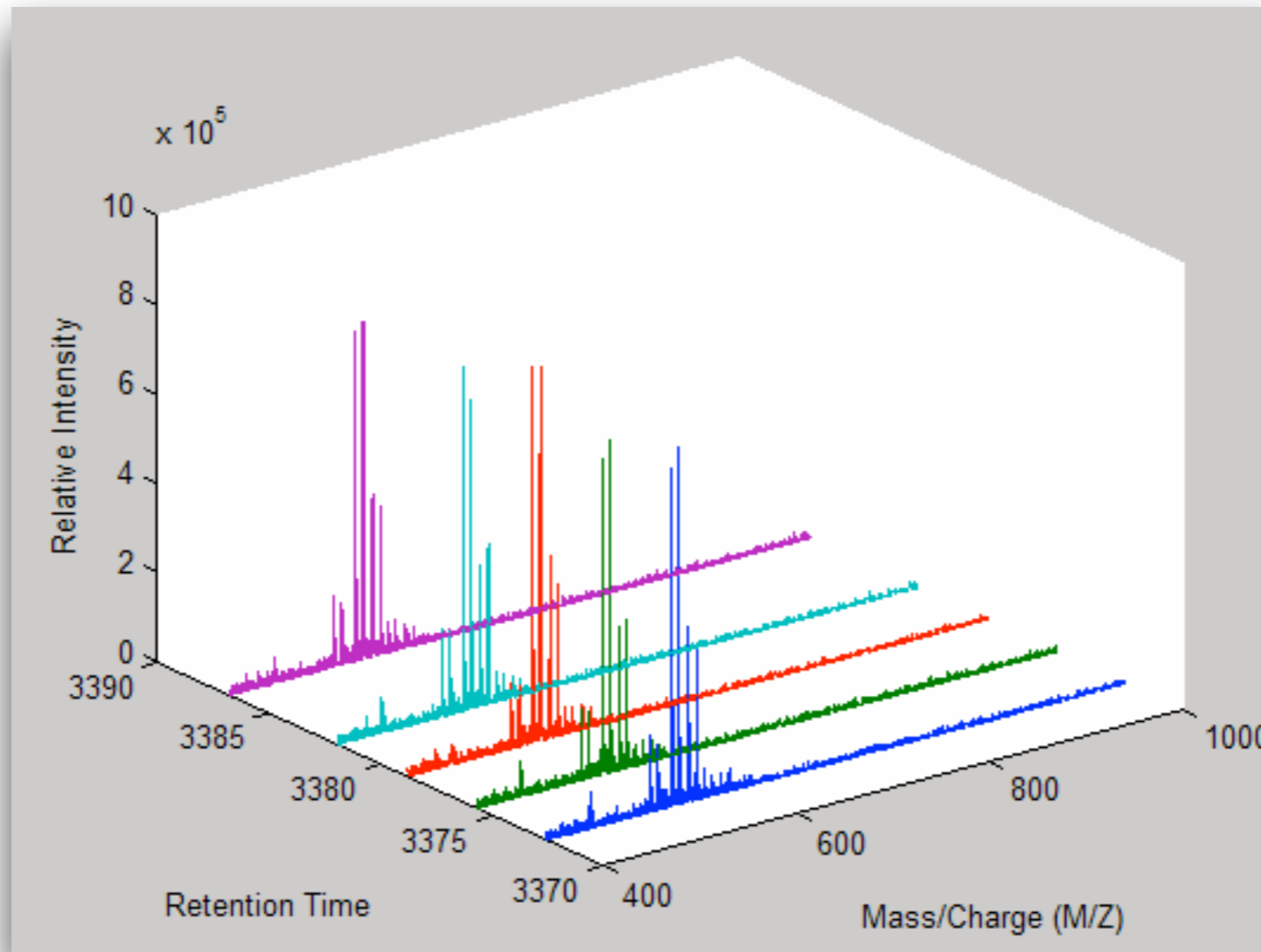


# Spectrum in centroid mode

- 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

The screenshot shows a software application window with a menu bar containing 'Project', 'Raw data methods', 'Peak list methods', 'Visualization', 'Windows', and 'Help'. The 'Visualization' menu is open, listing options: 'TIC/XIC visualizer', 'Spectra visualizer', '2D visualizer', '3D visualizer' (highlighted), 'MS/MS visualizer', 'Neutral loss visualizer', 'Scatter plot', 'Histogram plot', and 'Peak intensity plot'. The left pane shows a file explorer with 'Raw data files' containing several .mzXML files. The right pane is titled 'New project' and contains a 'Peak lists' section. At the bottom, a 'Tasks in progress...' table is visible, and a status bar shows a task completion message and a memory indicator.

Item	Priority	Status	% done
[1:34:26 AM]: Processing of task Detecting masses in PosMode_IR3.mzXML done, status FINISHED			

1956MB free

# Raw data in 3D

Please set the parameters

Raw data files      PosMode\_IR3.mzXML      As selected in main window      ...

Scans      MS level: 1      Set filters      Clear filters

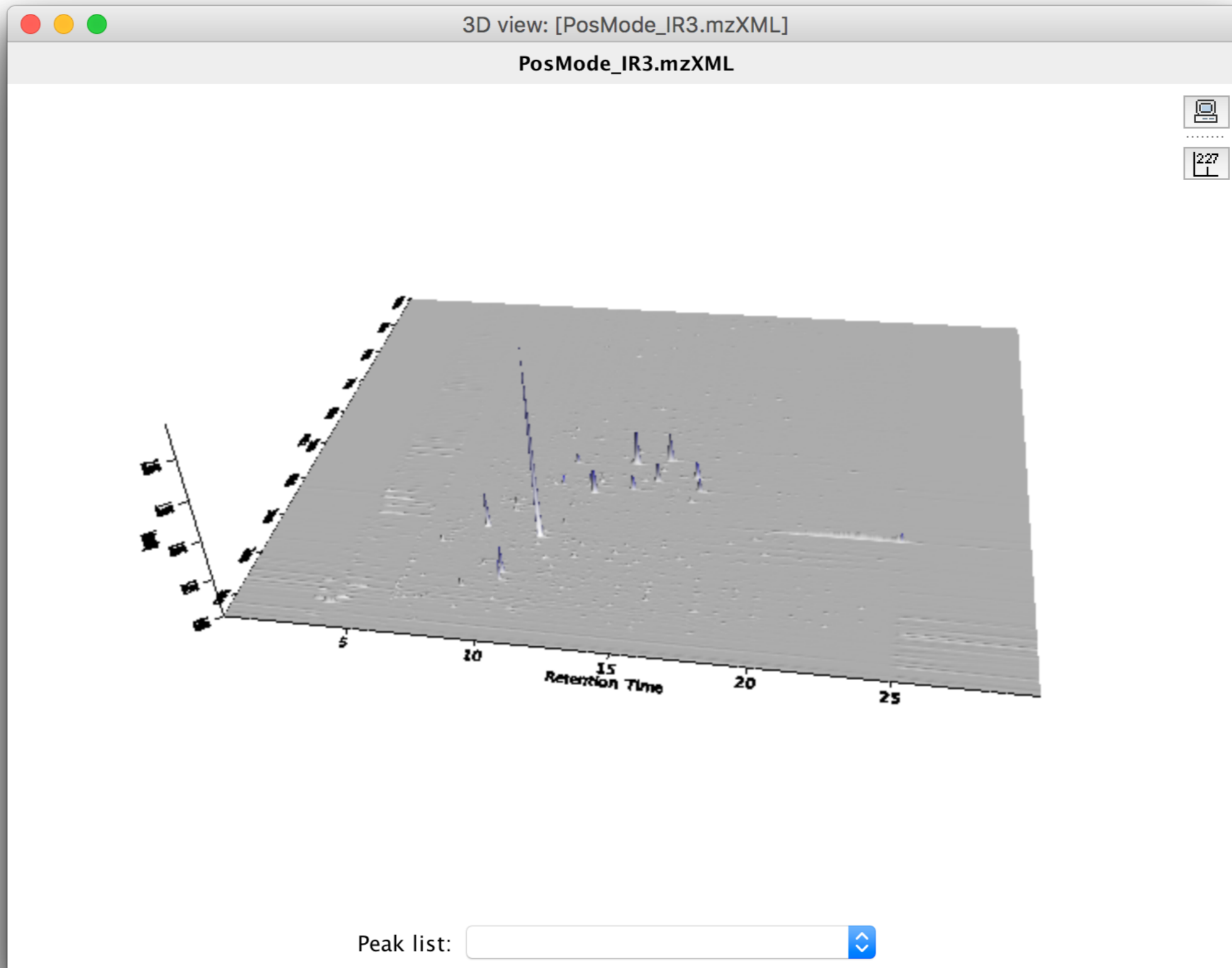
m/z      49.9916      -      1000.0115      Auto range      From mass      From formula

Retention time resolution      500

m/z resolution      500

OK      Cancel      Help

# 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

The screenshot shows a software interface with a menu bar at the top: Project, Raw data methods, Peak list methods, Visualization, Windows, and Help. The 'Visualization' menu is open, listing options: TIC/XIC visualizer, Spectra visualizer, 2D visualizer (highlighted), 3D visualizer, MS/MS visualizer, Neutral loss visualizer, Scatter plot, Histogram plot, and Peak intensity plot. On the left, a file list under 'Raw data files' includes PosMode\_IR3.mzXML (selected), PosMode\_IR2.mzXML, PosMode\_NR1.mzXML, PosMode\_IR1.mzXML, PosMode\_NR2.mzXML, and PosMode\_NR3.mzXML. On the right, a 'New project' window is open with a 'Peak lists' section. At the bottom, a 'Tasks in progress...' table is visible, and a status bar at the very bottom shows a task completion message and a memory indicator.

Item	Priority	Status	% done
Tasks in progress...			

[1:53:51 AM]: Processing of task Sampling 3D plot of PosMode\_IR3.mzXML done, status FINISHED

3144MB free

# Mass vs. retention time map

Please set the parameters

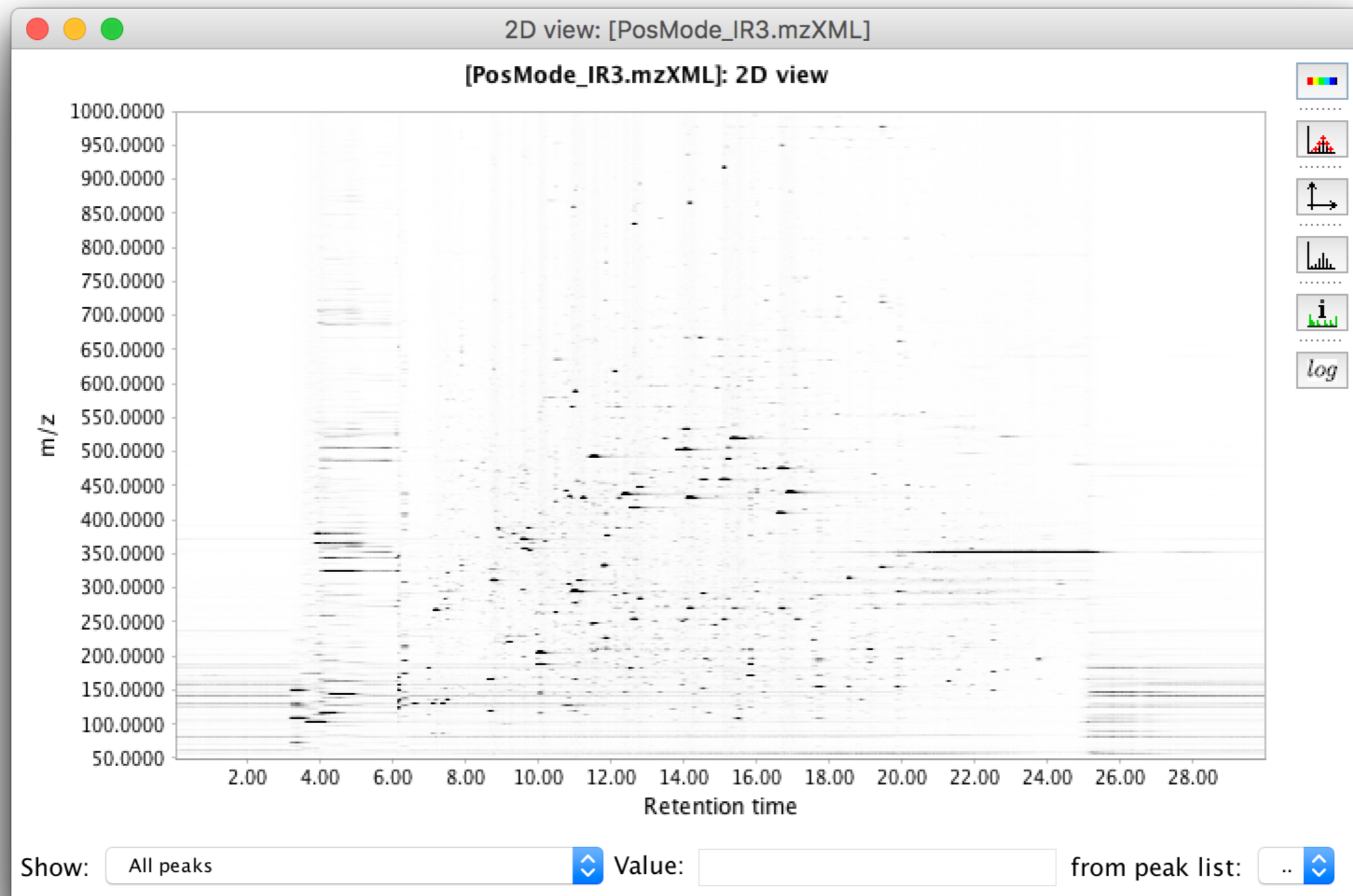
Raw data files PosMode\_IR3.mzXML As selected in main window ...

Scans MS level: 1 Set filters Clear filters

m/z 49.9916 - 1000.0115 Auto range From mass From formula

OK Cancel Help

# Mass vs. retention time map





# TIC

The screenshot shows a software application window with a menu bar and a main workspace. The menu bar includes 'Project', 'Raw data methods', 'Peak list methods', 'Visualization', 'Windows', and 'Help'. The 'Visualization' menu is open, showing options: 'TIC/XIC visualizer', 'Spectra visualizer', '2D visualizer', '3D visualizer', 'MS/MS visualizer', 'Neutral loss visualizer', 'Scatter plot', 'Histogram plot', and 'Peak intensity plot'. The 'TIC/XIC visualizer' option is highlighted. The main workspace is divided into two panes. The left pane, titled 'Raw data files', contains a list of files: 'PosMode\_IR3.mzXML', 'PosMode\_IR2.mzXML', 'PosMode\_NR1.mzXML', 'PosMode\_IR1.mzXML', 'PosMode\_NR2.mzXML', and 'PosMode\_NR3.mzXML'. The right pane, titled 'Peak lists', is currently empty. At the bottom of the window, there is a 'Tasks in progress...' section with a table. The table has columns for 'Item', 'Priority', 'Status', and '% done'. Below the table, a status bar shows a log message: '[2:12:26 AM]: Processing of task Updating 2D visualizer of PosMode\_IR3.mzXML done, status FINISHED' and a memory indicator: '8329MB free'.

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

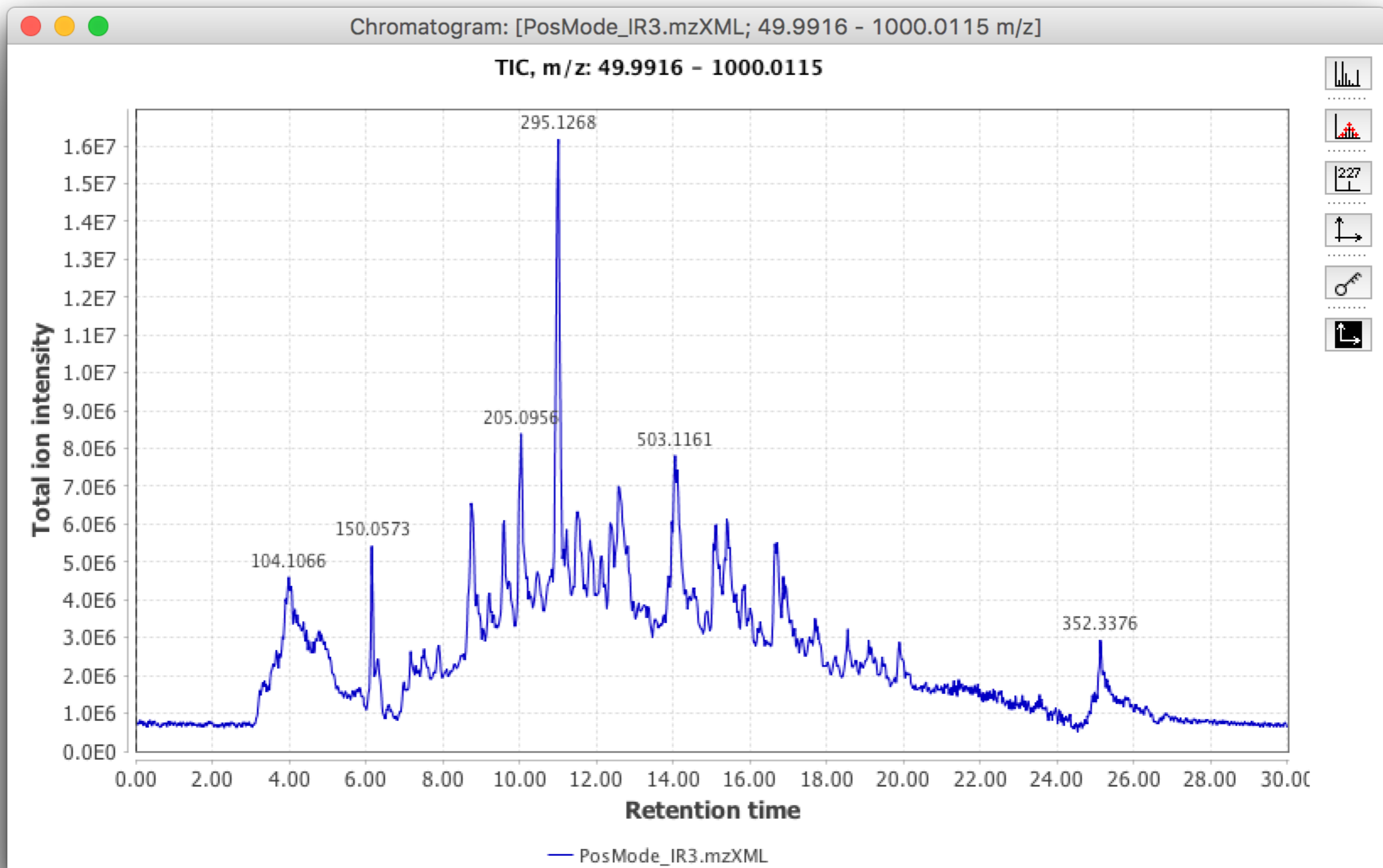
Peaks

All

Clear

OK Cancel Help

# TIC



# EIC

The screenshot displays the EIC software interface. At the top, there is a menu bar with 'Project', 'Raw data methods', 'Peak list methods', 'Visualization', 'Windows', and 'Help'. The 'Visualization' menu is open, showing options: 'TIC/XIC visualizer', 'Spectra visualizer', '2D visualizer', '3D visualizer', 'MS/MS visualizer', 'Neutral loss visualizer', 'Scatter plot', 'Histogram plot', and 'Peak intensity plot'. The 'TIC/XIC visualizer' option is highlighted. Below the menu bar, the main window is divided into two panes. The left pane, titled 'Raw data files', contains a list of files: 'PosMode\_IR3.mzXML', 'PosMode\_IR2.mzXML', 'PosMode\_NR1.mzXML', 'PosMode\_IR1.mzXML', 'PosMode\_NR2.mzXML', and 'PosMode\_NR3.mzXML'. The right pane, titled 'TIC/XIC visualizer.', contains a sub-pane 'Peak lists'. At the bottom of the window, there is a 'Tasks in progress...' section with a table:

Item	Priority	Status	% done

At the bottom of the window, a status bar shows the message: '[2:12:26 AM]: Processing of task Updating 2D visualizer of PosMode\_IR3.mzXML done, status FINISHED' and a memory indicator '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)

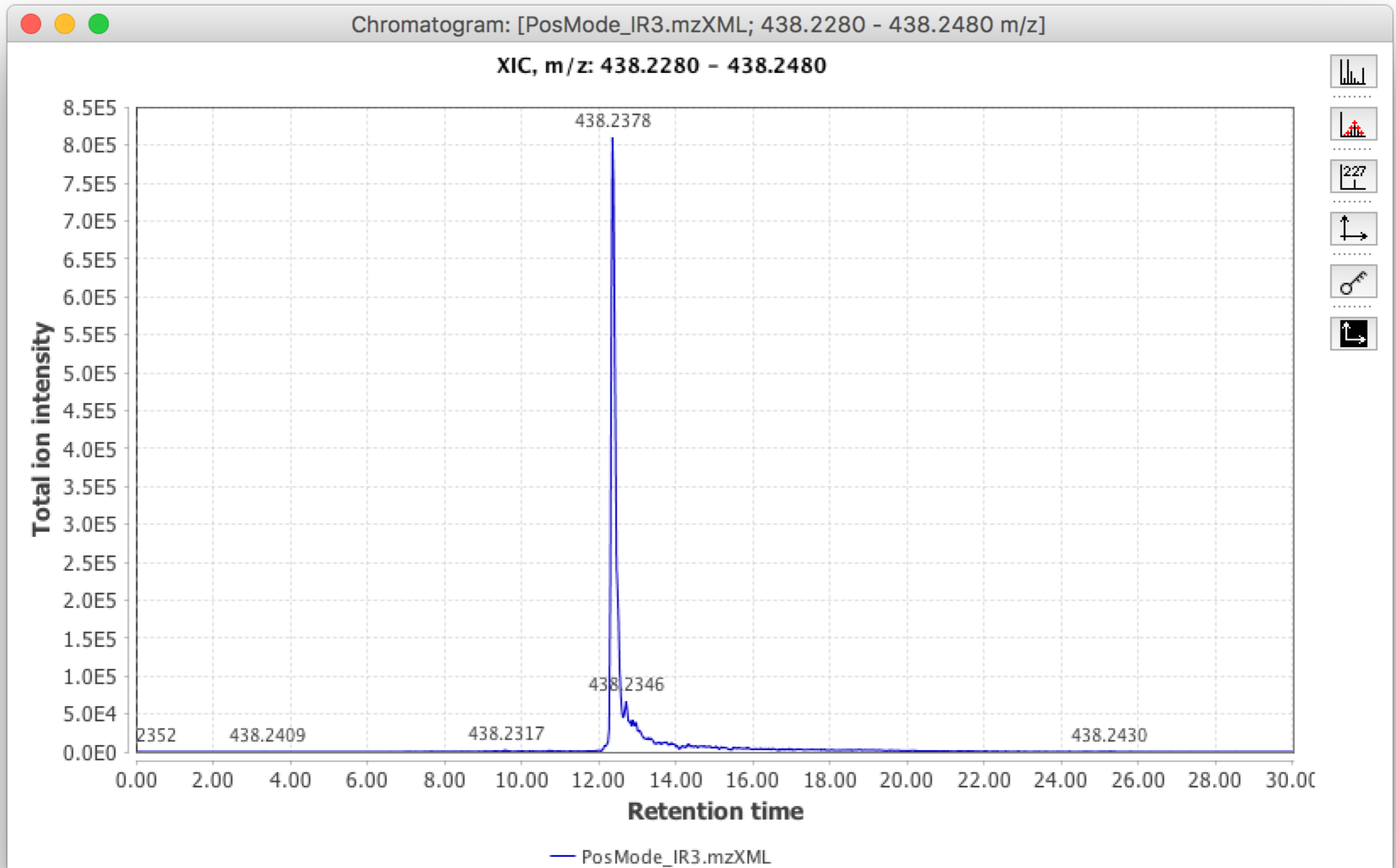
m/z 438.2280 - 438.2480 Auto range From mass From formula

Peaks

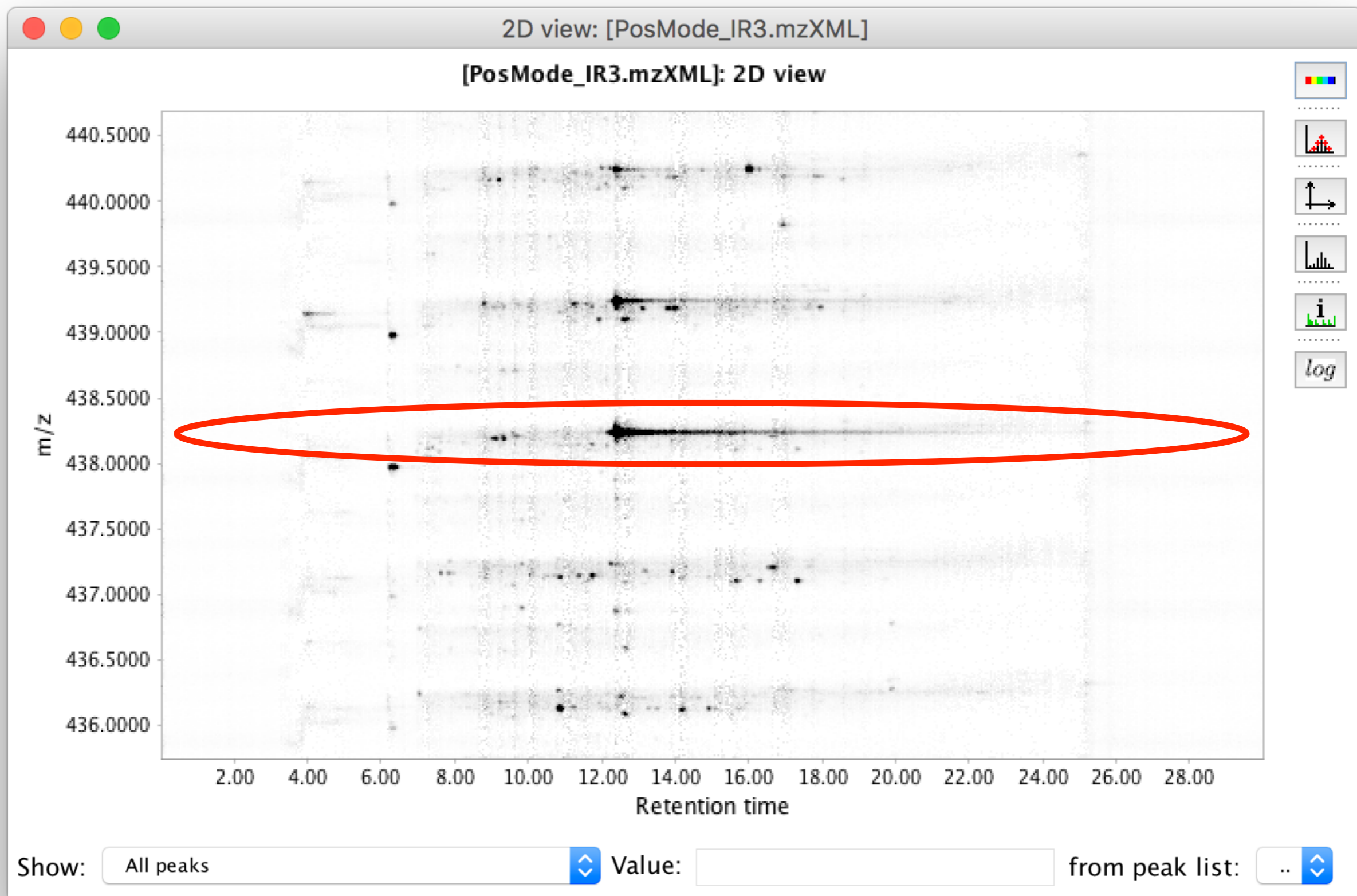
All Clear

OK Cancel Help

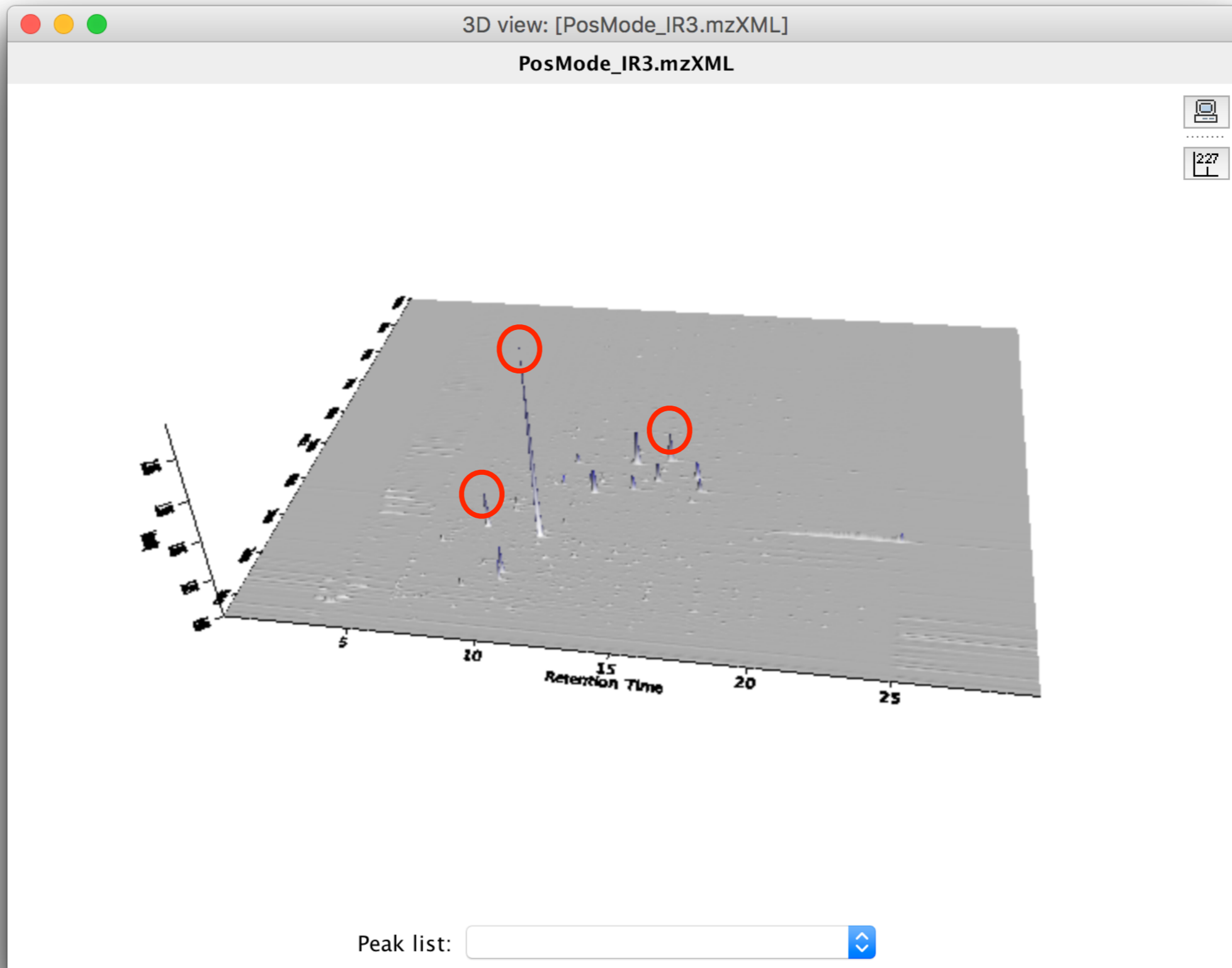
# EIC



# EIC



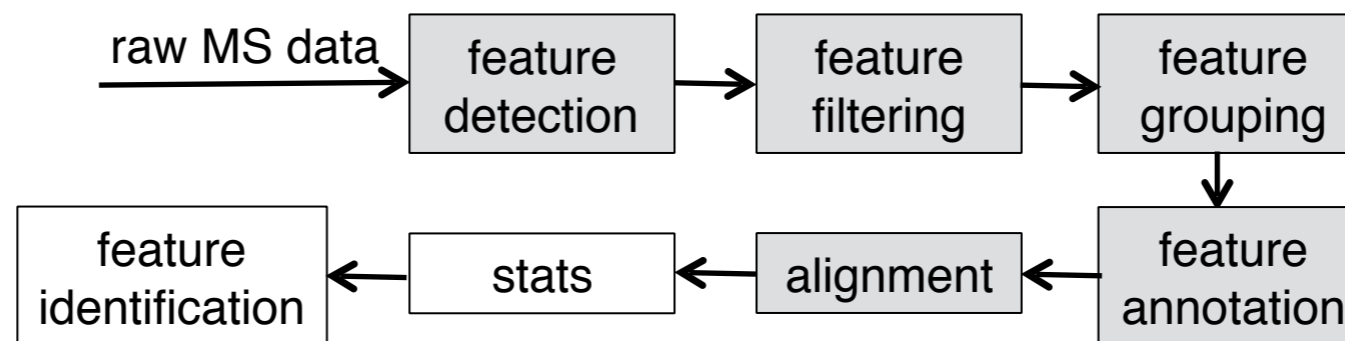
# Feature





# Feature

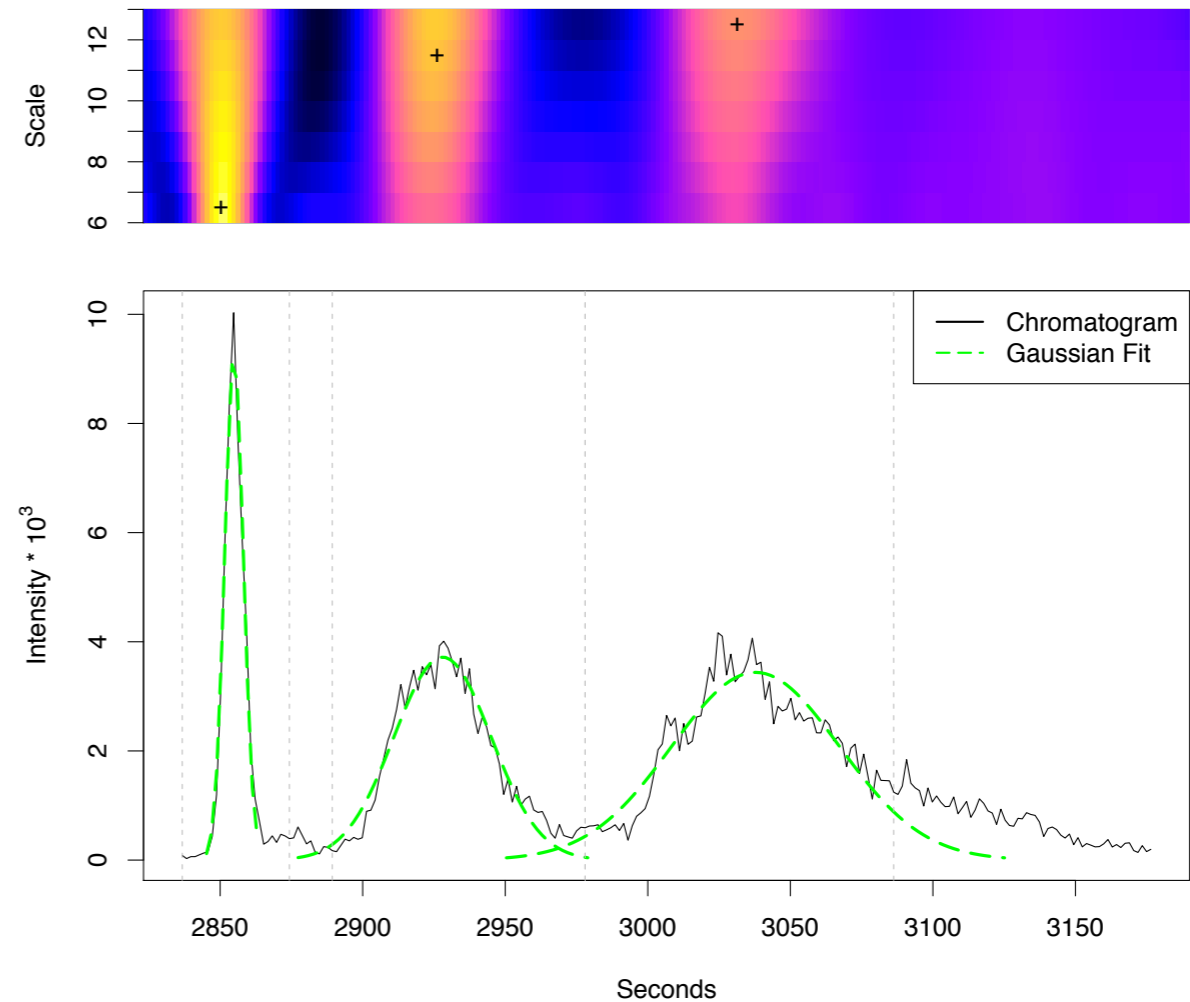
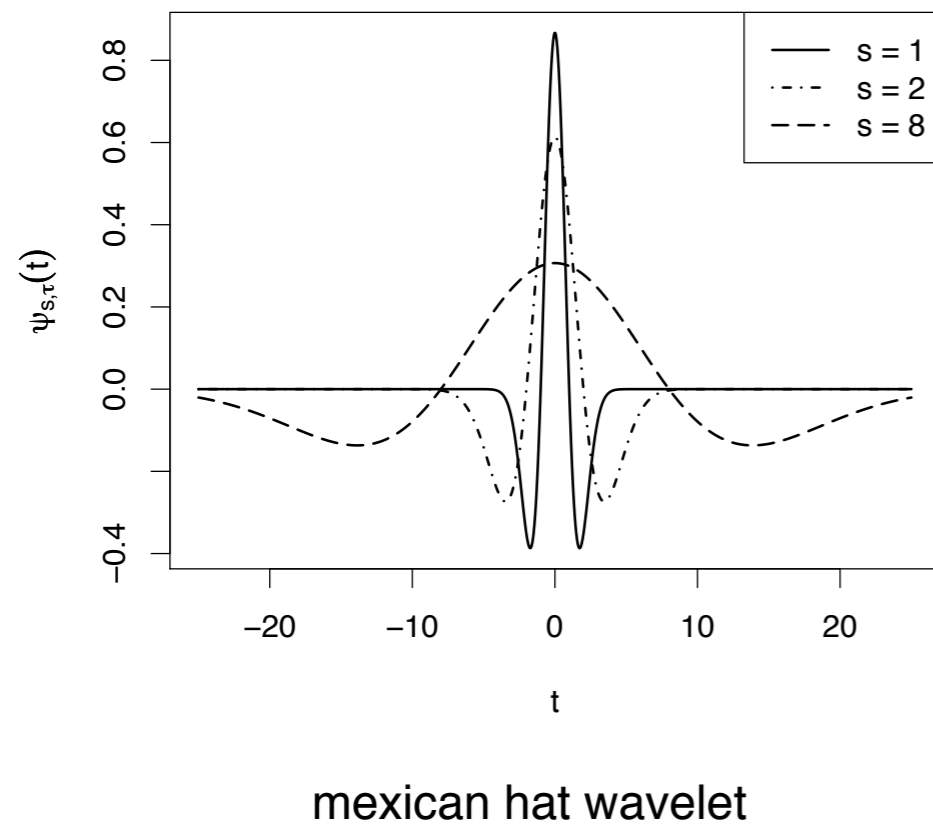
- **Feature:** A 3D signal induced by a single ion species (e.g.  $[M+H]^+$  or  $[M-H]^-$  of a compound)
- How to detect features?
  - by peak picking from EICs
- Data processing and analysis workflow



# Data Pre-processing

# Detection of chromatographic features

- Use wavelet transform



- Implemented in XCMS as the centWave method

# Detection of chromatographic features

Project Raw data methods Peak list methods Visualization Windows Help

Raw data files

- PosMode\_IR3.mzXML
- PosMode\_IR2.mzXML
- PosMode\_NR1.mzXML
- PosMode\_IR1.mzXML
- PosMode\_NR2.mzXML
- PosMode\_NR3.mzXML

Visualization menu:

- TIC/XIC visualizer
- Spectra visualizer
- 2D visualizer
- 3D visualizer
- MS/MS visualizer
- Neutral loss visualizer
- Scatter plot
- Histogram plot
- Peak intensity plot

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 155.0950 - 155.1150 Auto range From mass From formula

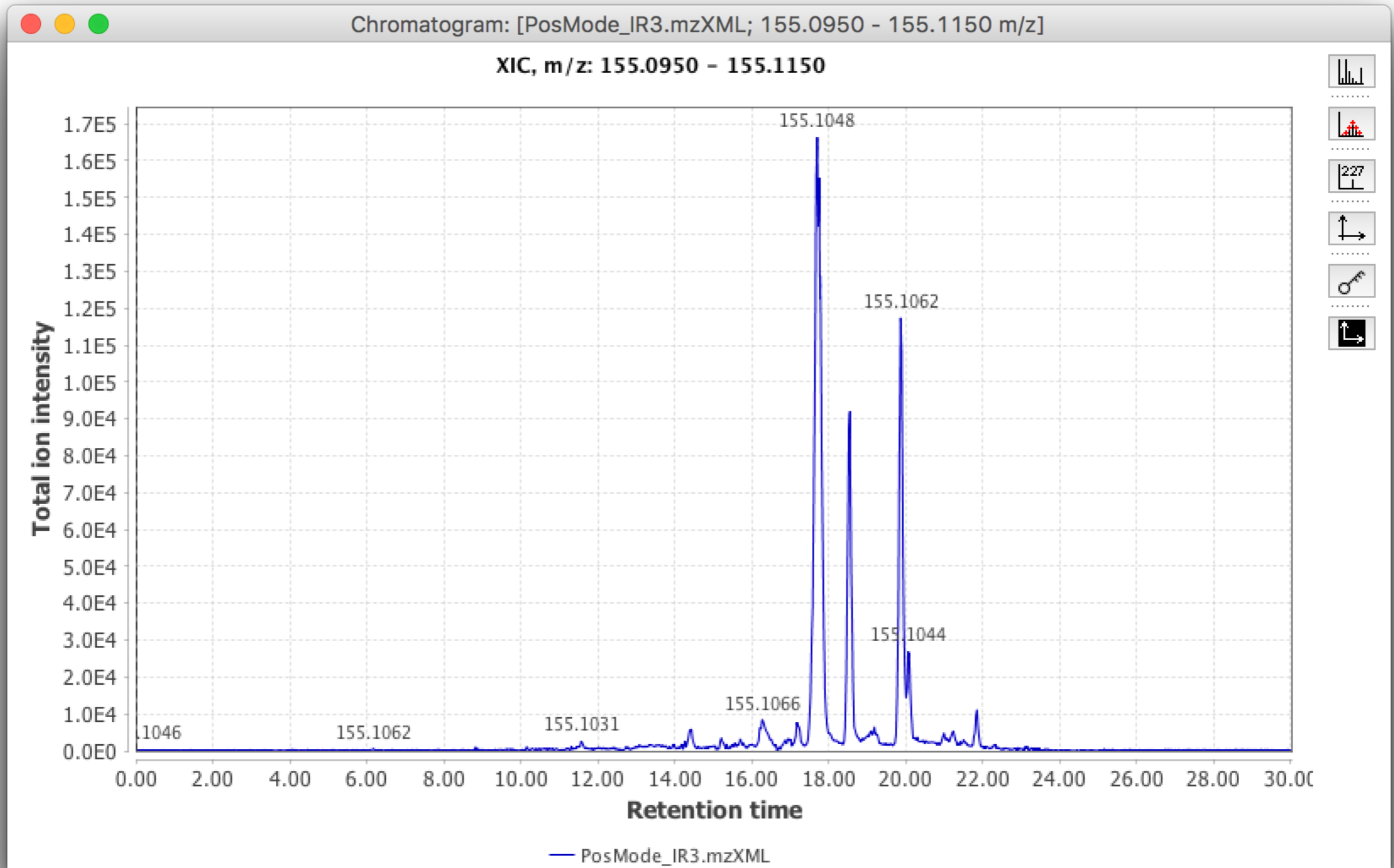
Peaks All Clear

OK Cancel Help

[2:12:26 AM]: Processing of task Updating 2D visualizer of PosMode\_IR3.mzXML done, status FINISHED

8329MB free

# Detection of chromatographic features



# Feature grouping and alignment

The screenshot displays a software application window with a menu bar (Project, Raw data methods, Peak list methods, Visualization, Windows, Help) and a system tray (99% battery, Mon Jul 18). The main window shows a list of raw data files under 'Raw data files':

- PosMode\_IR3.mzXML
- PosMode\_IR2.mzXML
- PosMode\_NR1.mzXML
- PosMode\_IR1.mzXML
- PosMode\_NR2.mzXML
- PosMode\_NR3.mzXML

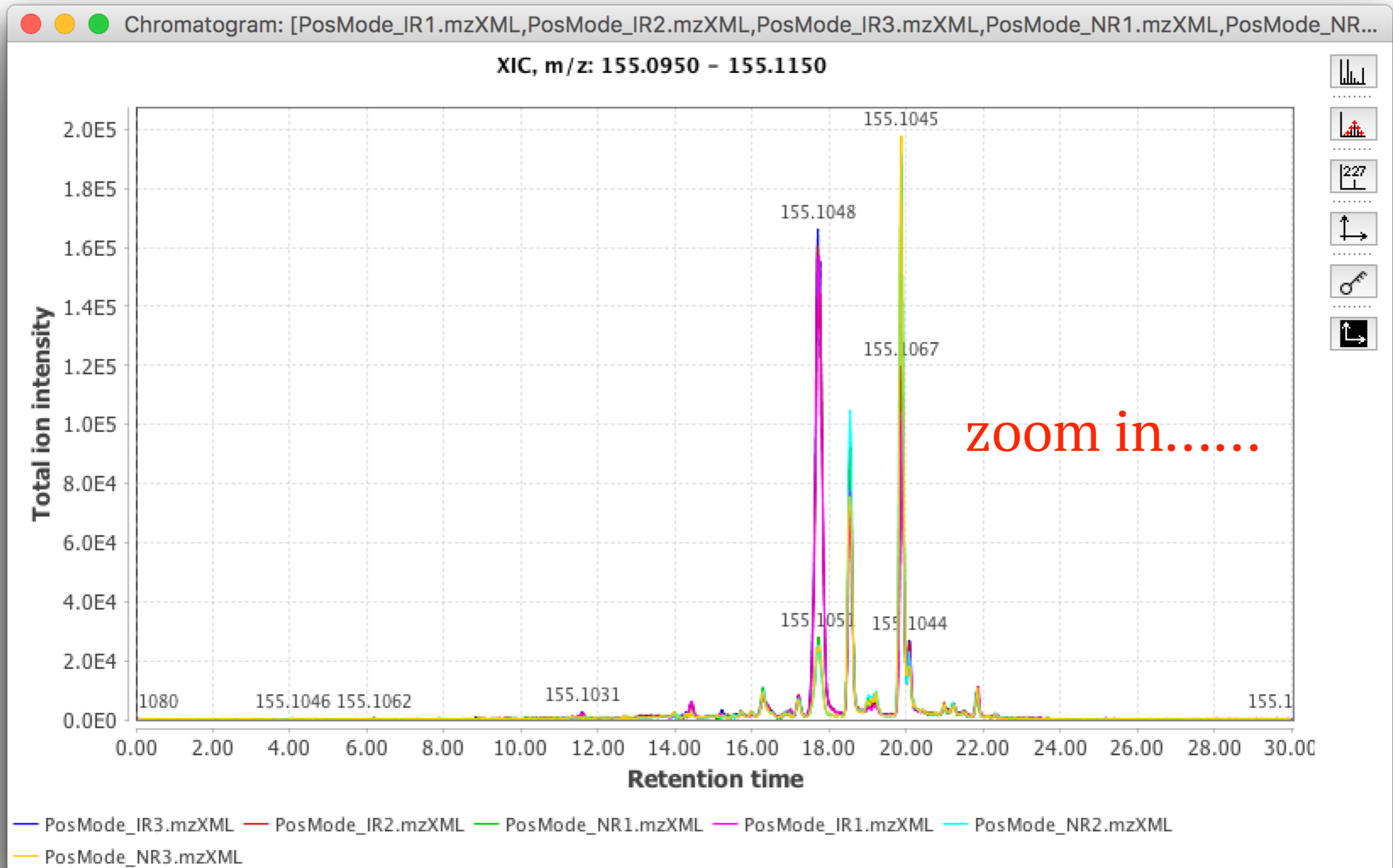
The 'Visualization' menu is open, listing options: TIC/XIC visualizer, Spectra visualizer, 2D visualizer, 3D visualizer, MS/MS visualizer, Neutral loss visualizer, Scatter plot, Histogram plot, and Peak intensity plot. A 'TIC/XIC visualizer.' window is also visible in the background.

In the foreground, a 'Please set the parameters' dialog box is open, showing the following settings:

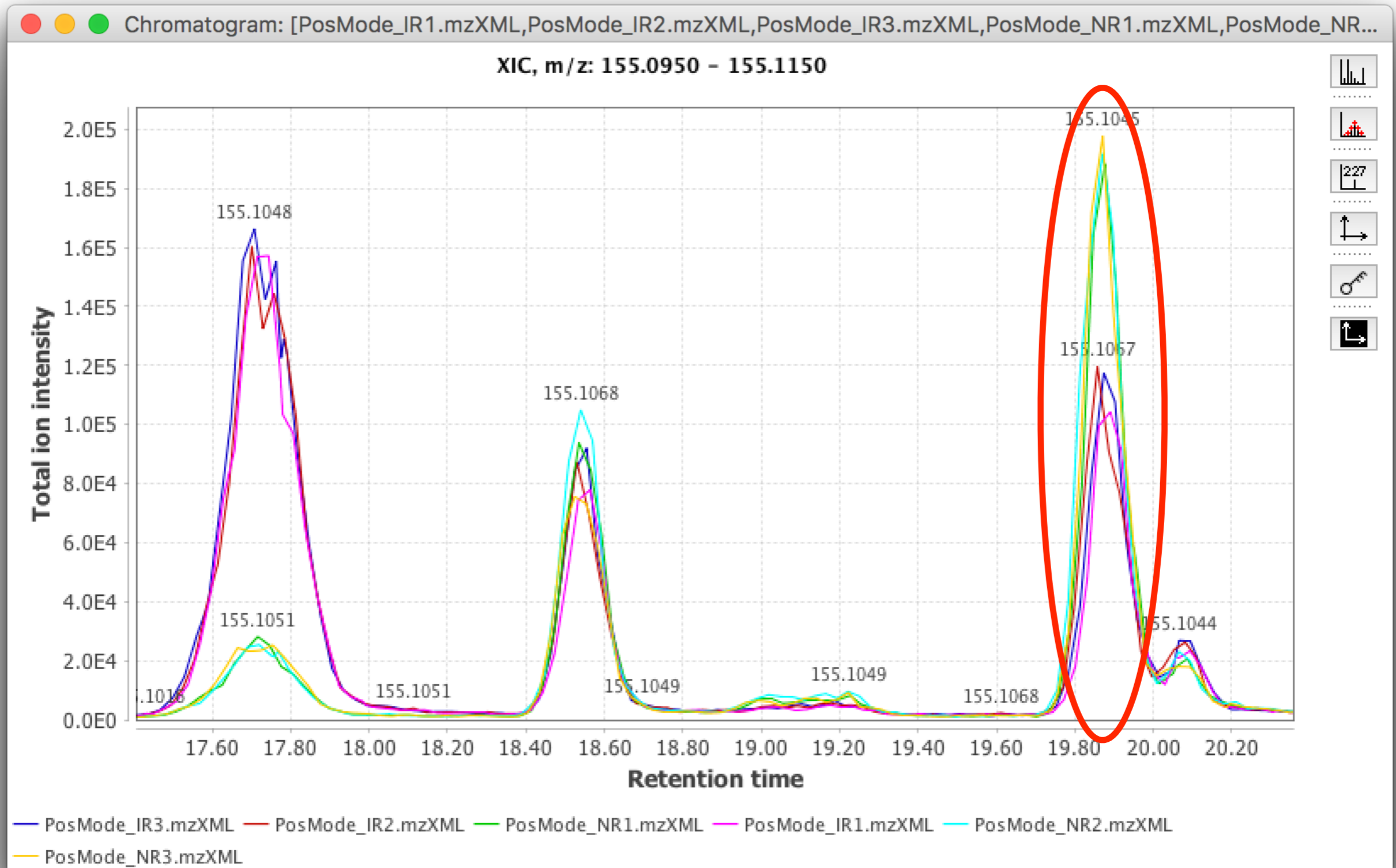
- Raw data files: PosMode\_IR3.mzXML (As selected in main window)
- Scans: MS level: 1 (Set filters, Clear filters)
- Plot type: Total ion current (TIC/XIC)
- m/z: 155.0950 - 155.1150 (Auto range, From mass, From formula)
- Peaks: (Empty list box) (All, Clear)

Buttons for OK, Cancel, and Help are at the bottom of the dialog. A status bar at the bottom shows: [2:29:26 AM]: Processing of task Updating 2D visualizer of PosMode\_IR3.mzXML done, status FINISHED. 157MB free.

# Feature grouping and alignment

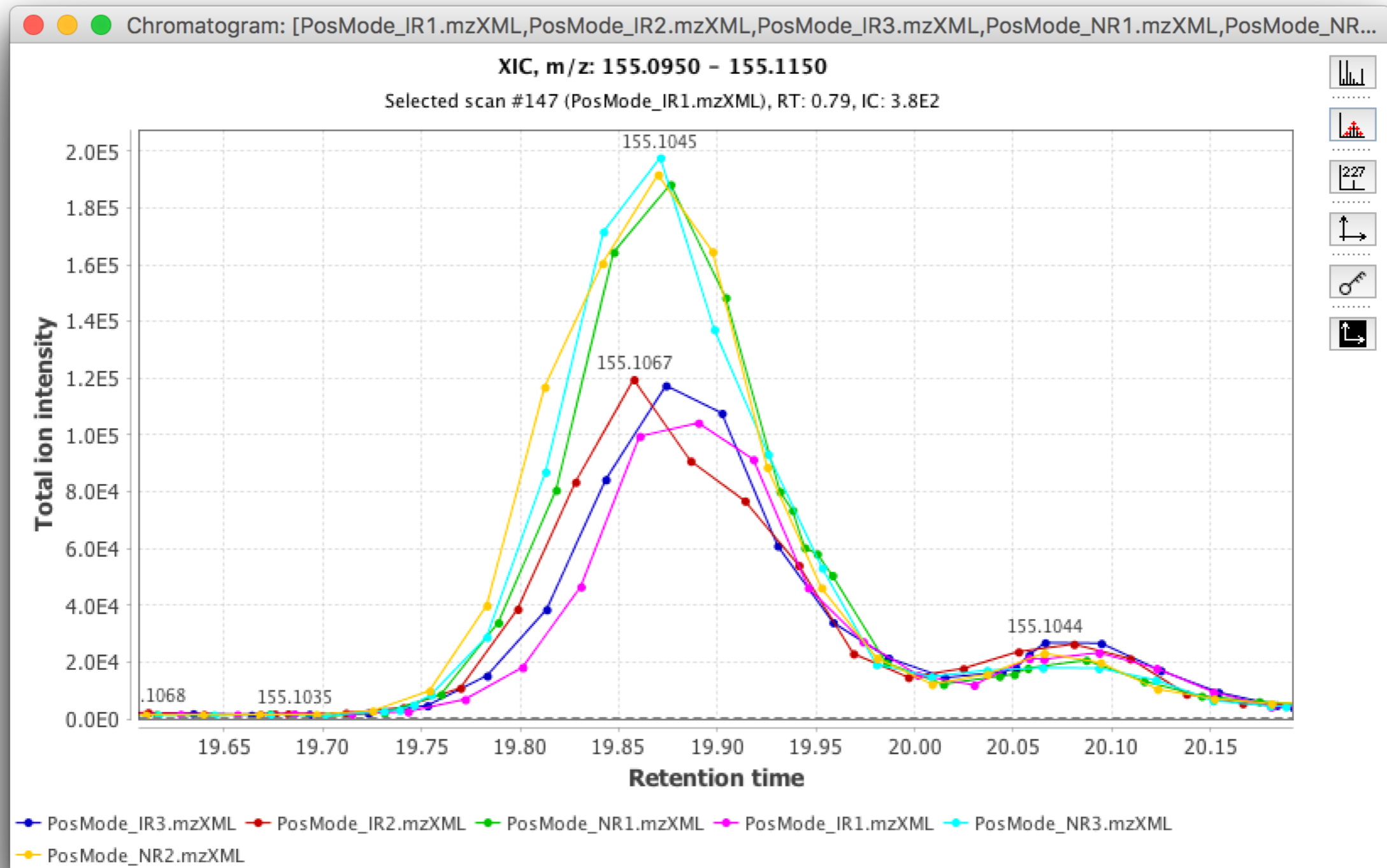


# Feature grouping and alignment





# Feature grouping and alignment

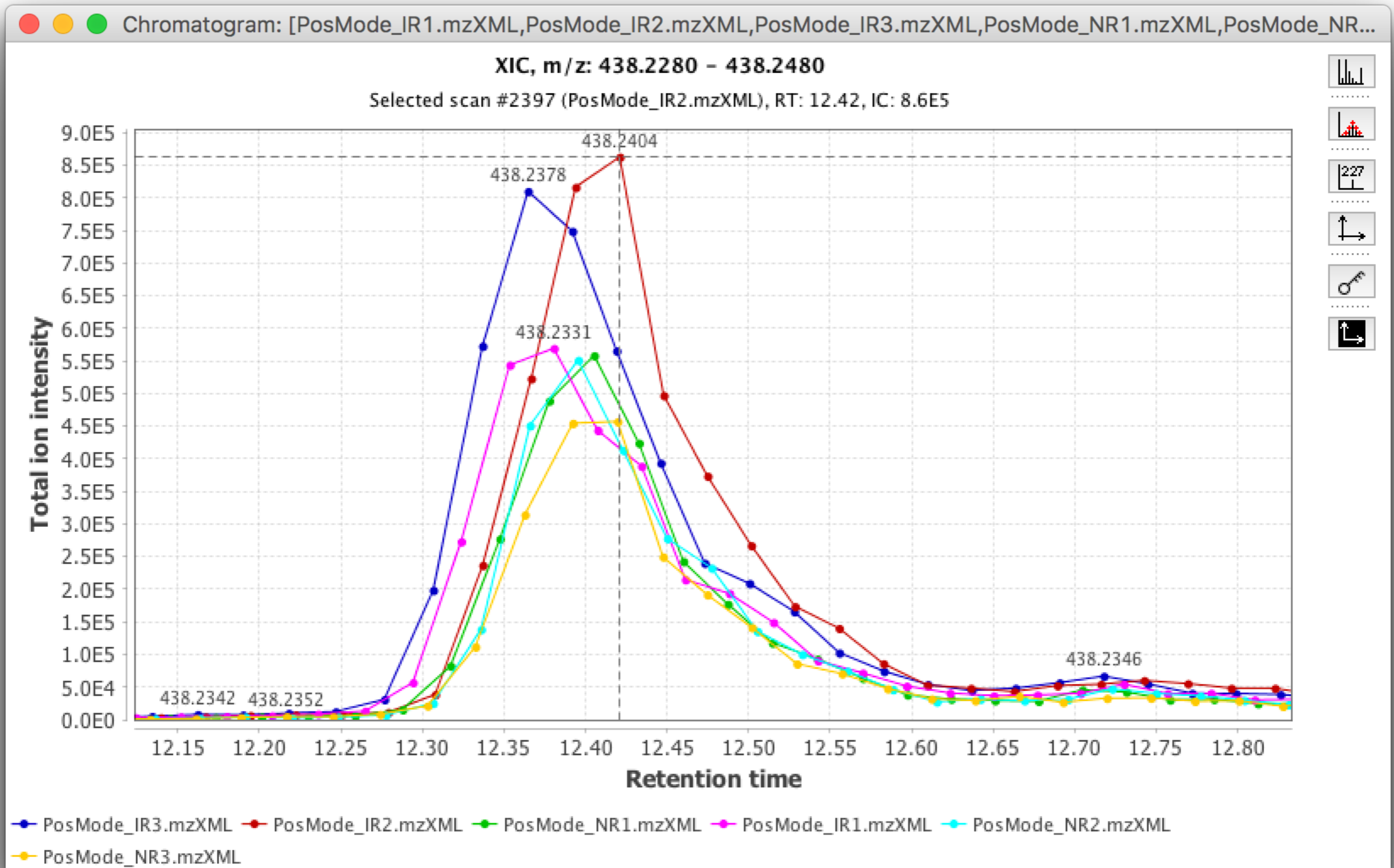


# Result of data pre-processing

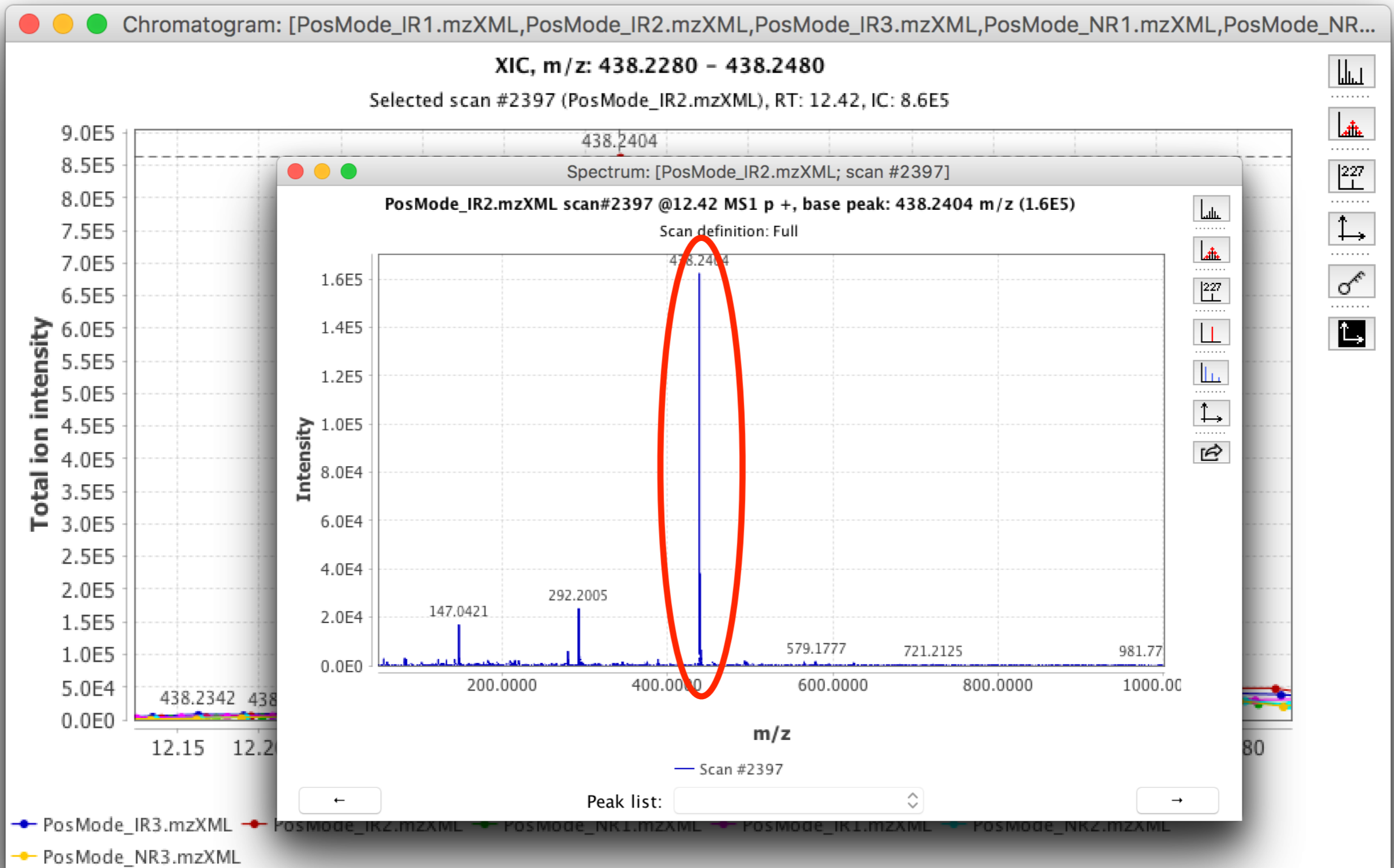
DB	Name	Mass	RT	platform	IN1	IN2	IN3	IN4	IN5	IN6
HMDB	1-Phenylethylamin	122.09745	24.97845	ES-	0.12862	0.1421305	0.1301326	0.1247924	0.1200045	0.1053275
HMDB	2-Ethylacrylic acid	101.06421	17.811575	ES-	0.0332025	0.0174262	0.0158166	0.0179326	0.0143742	0.0064953
HMDB	Canavanine	177.09653	10.338581	ES-	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



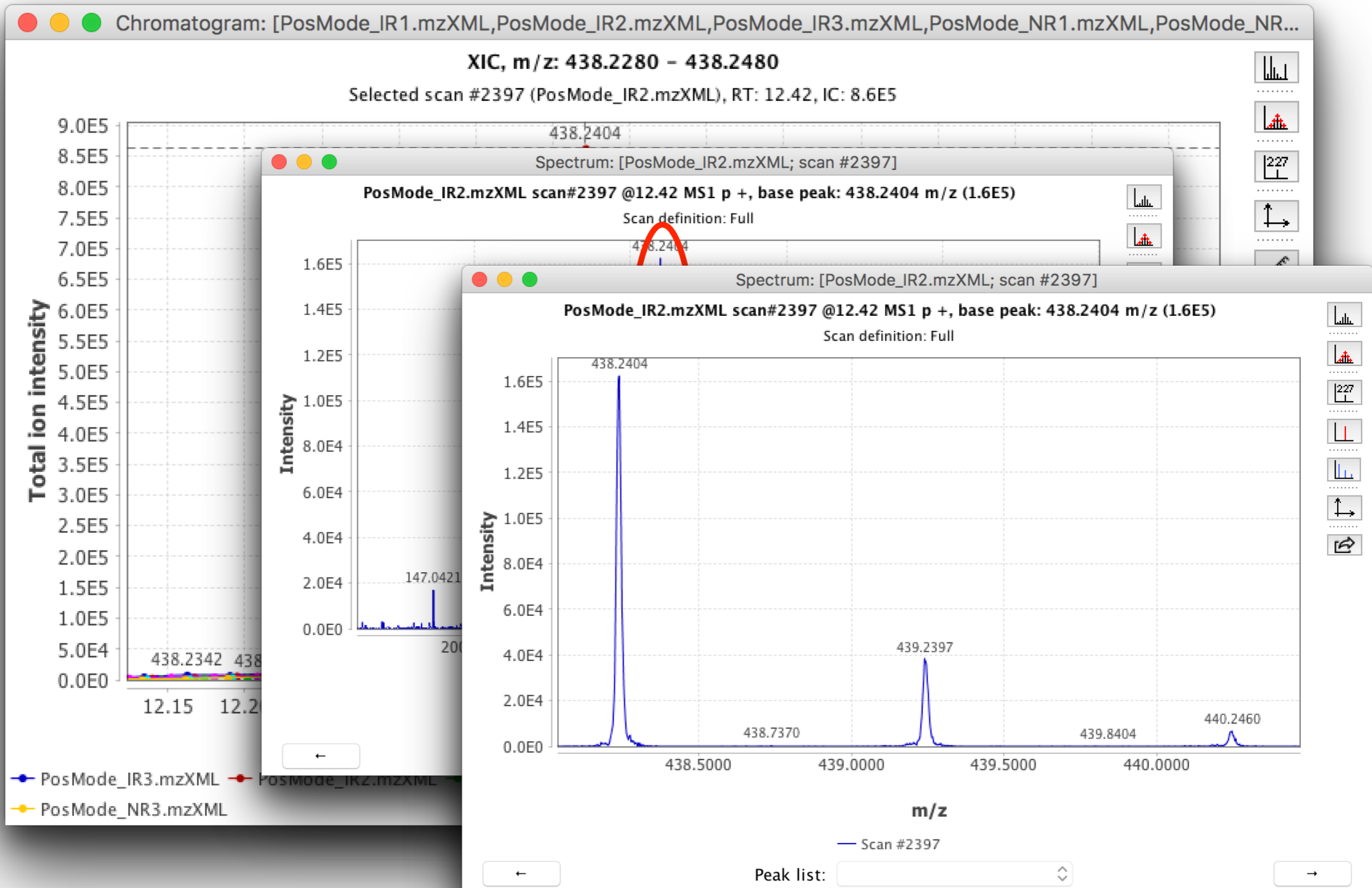
# Feature identification



# Feature identification



# Feature identification



# Feature identification

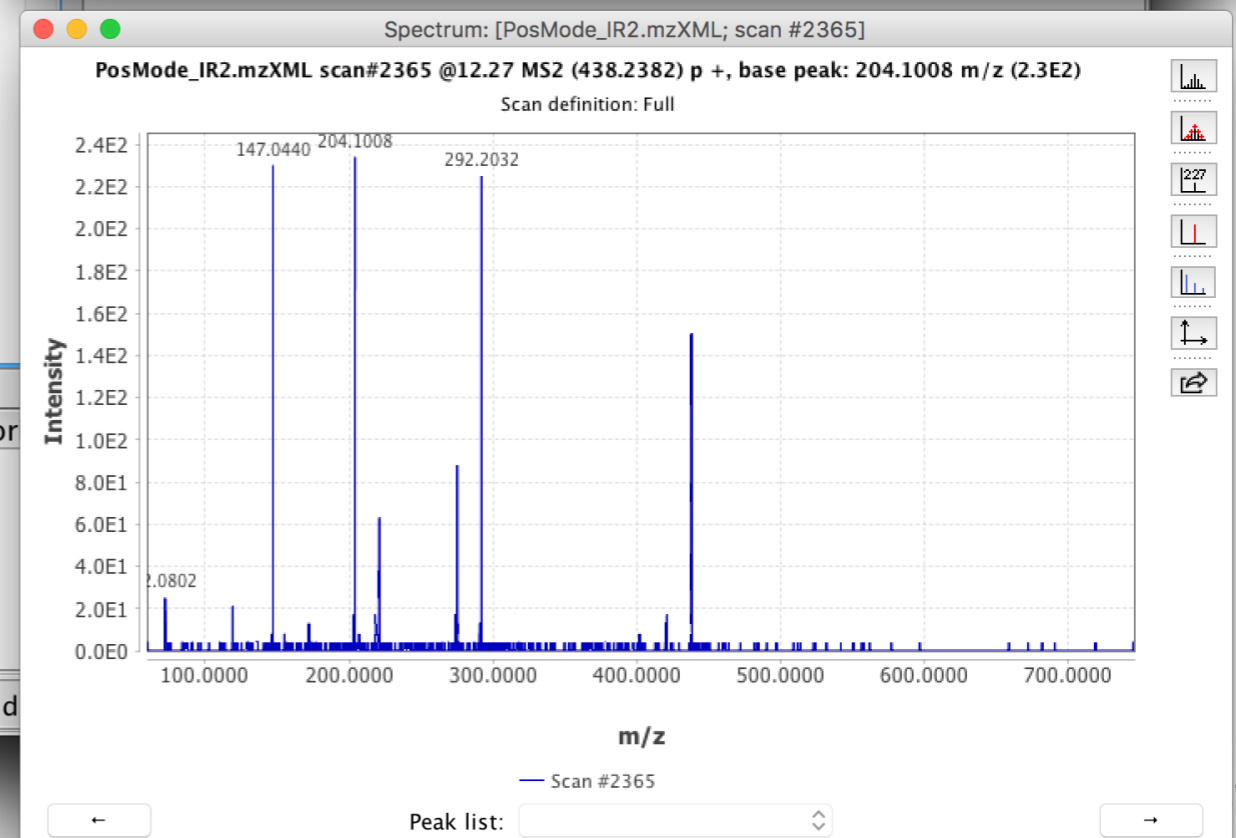
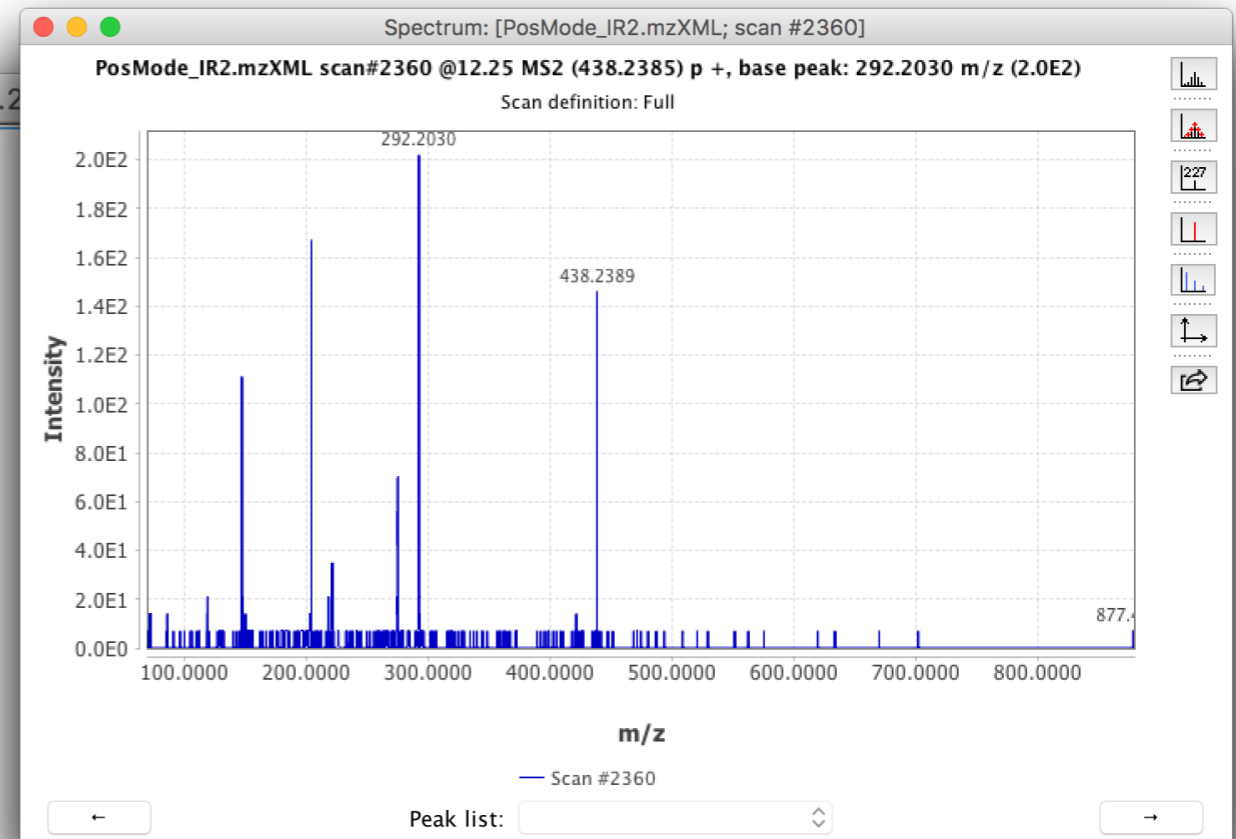
MZmine 2.2

- #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

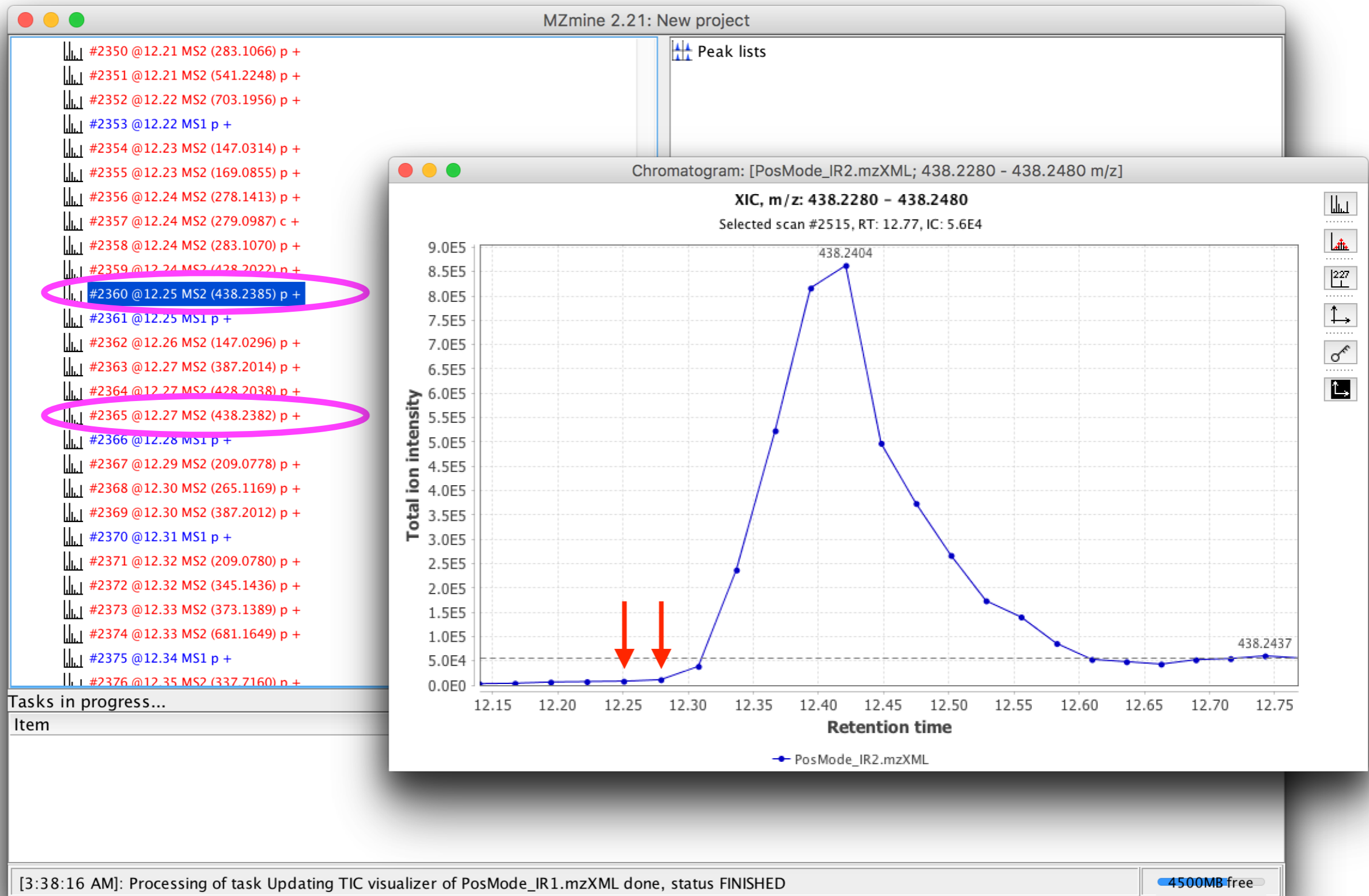


# Feature identification

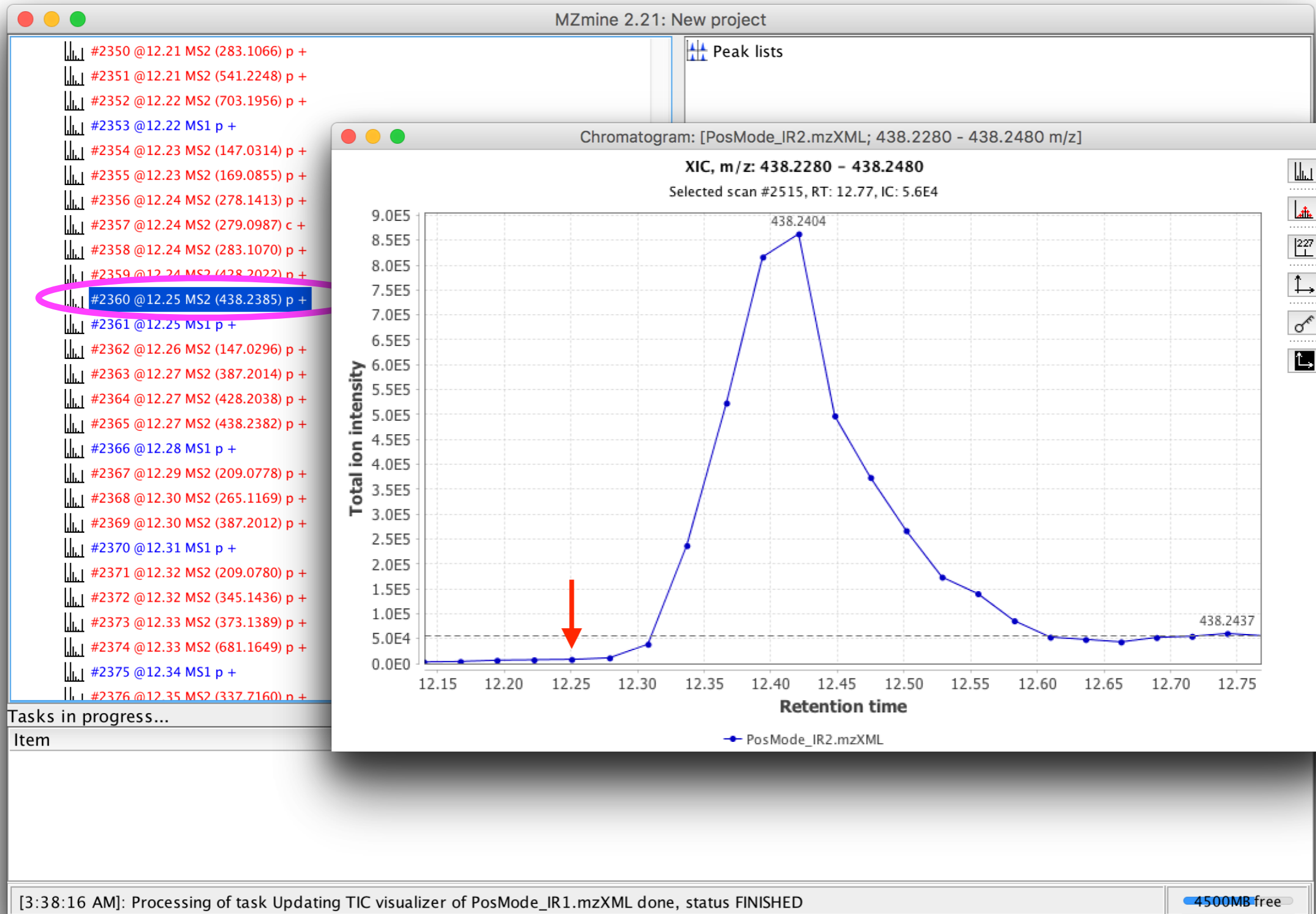
- Information we have for identification
  - M+H
  - Experimental isotopic identification
  - MS/MS



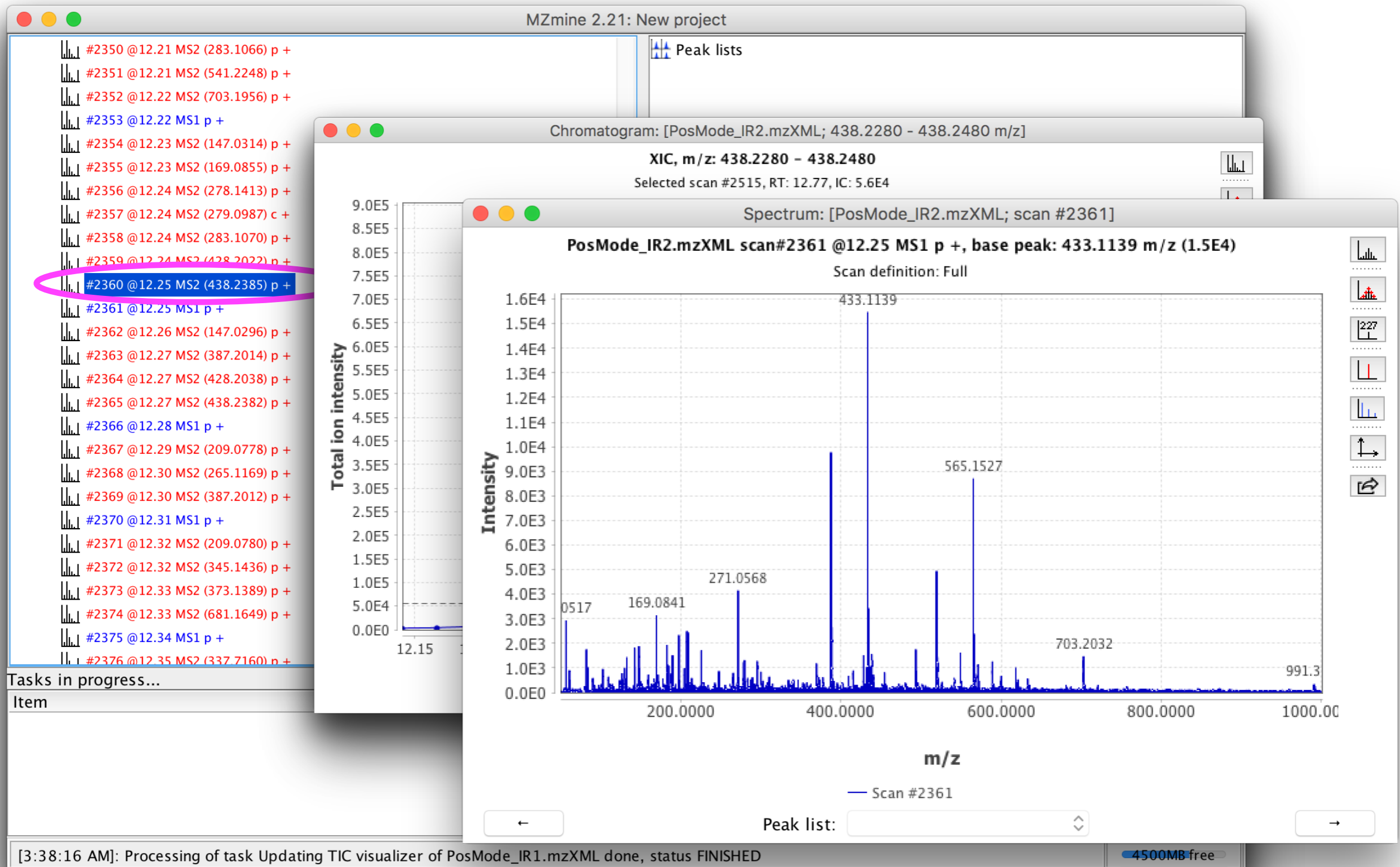
# Feature identification



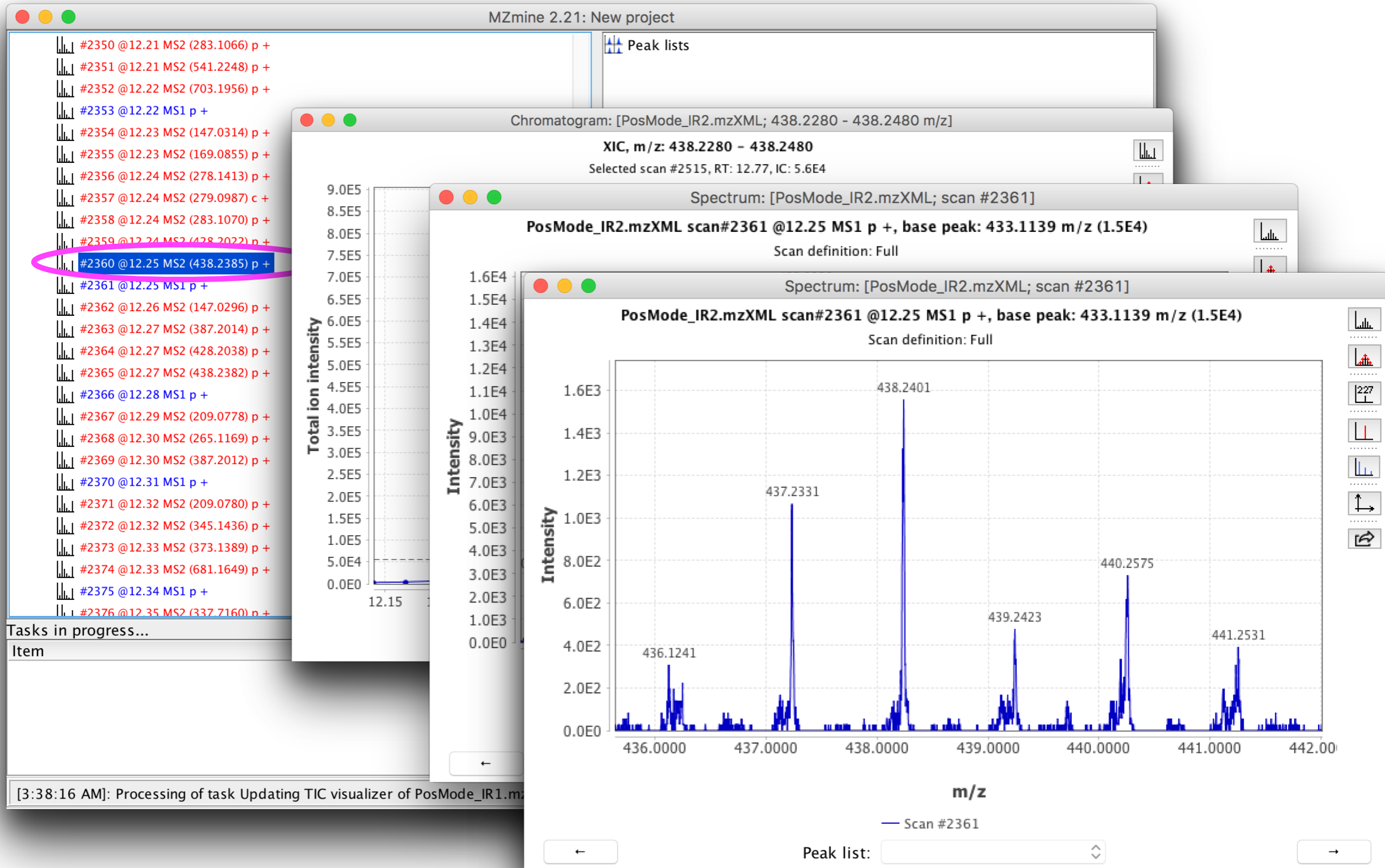
# Feature identification



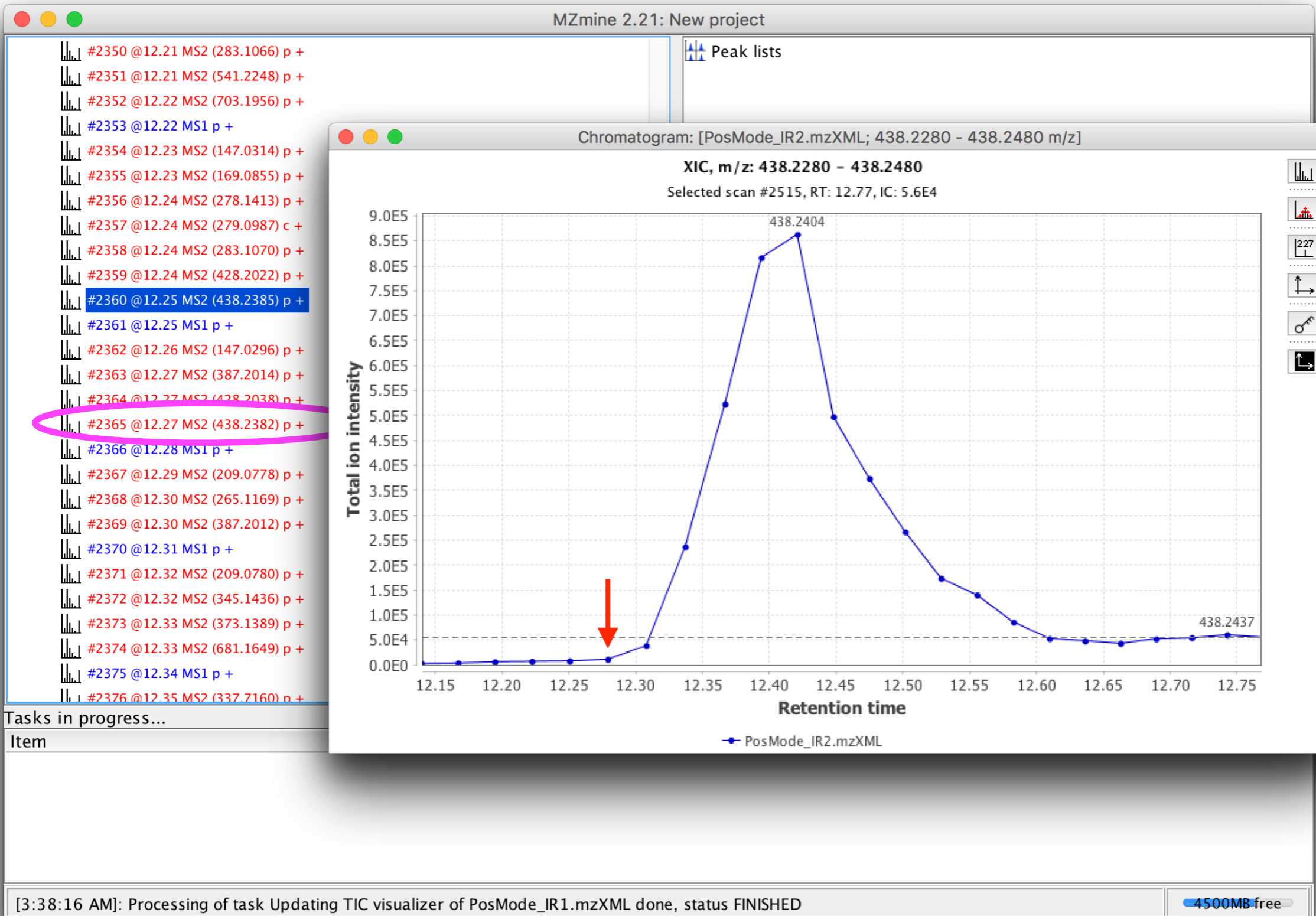
# Feature identification



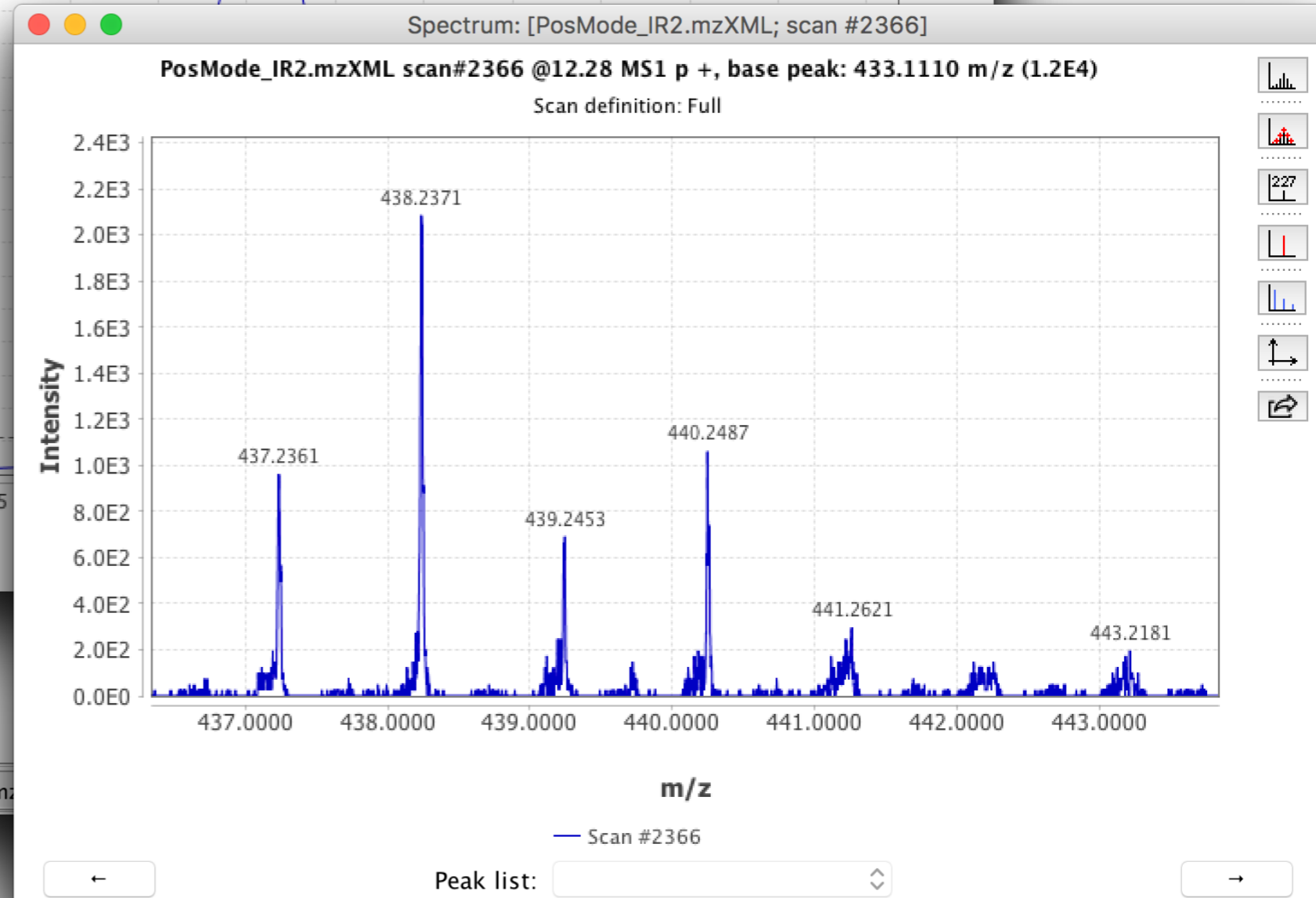
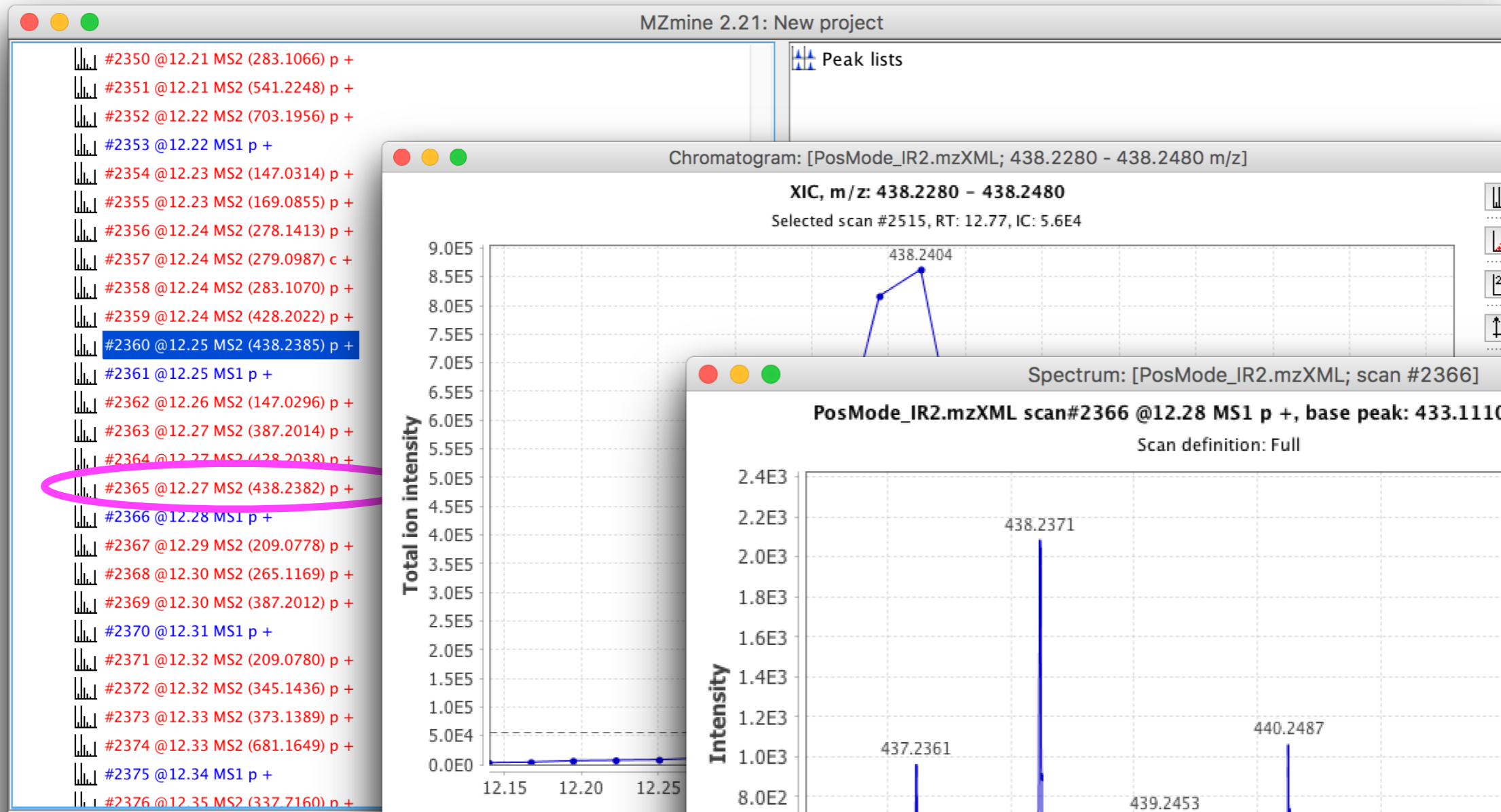
# Feature identification



# Feature identification



# Feature identification

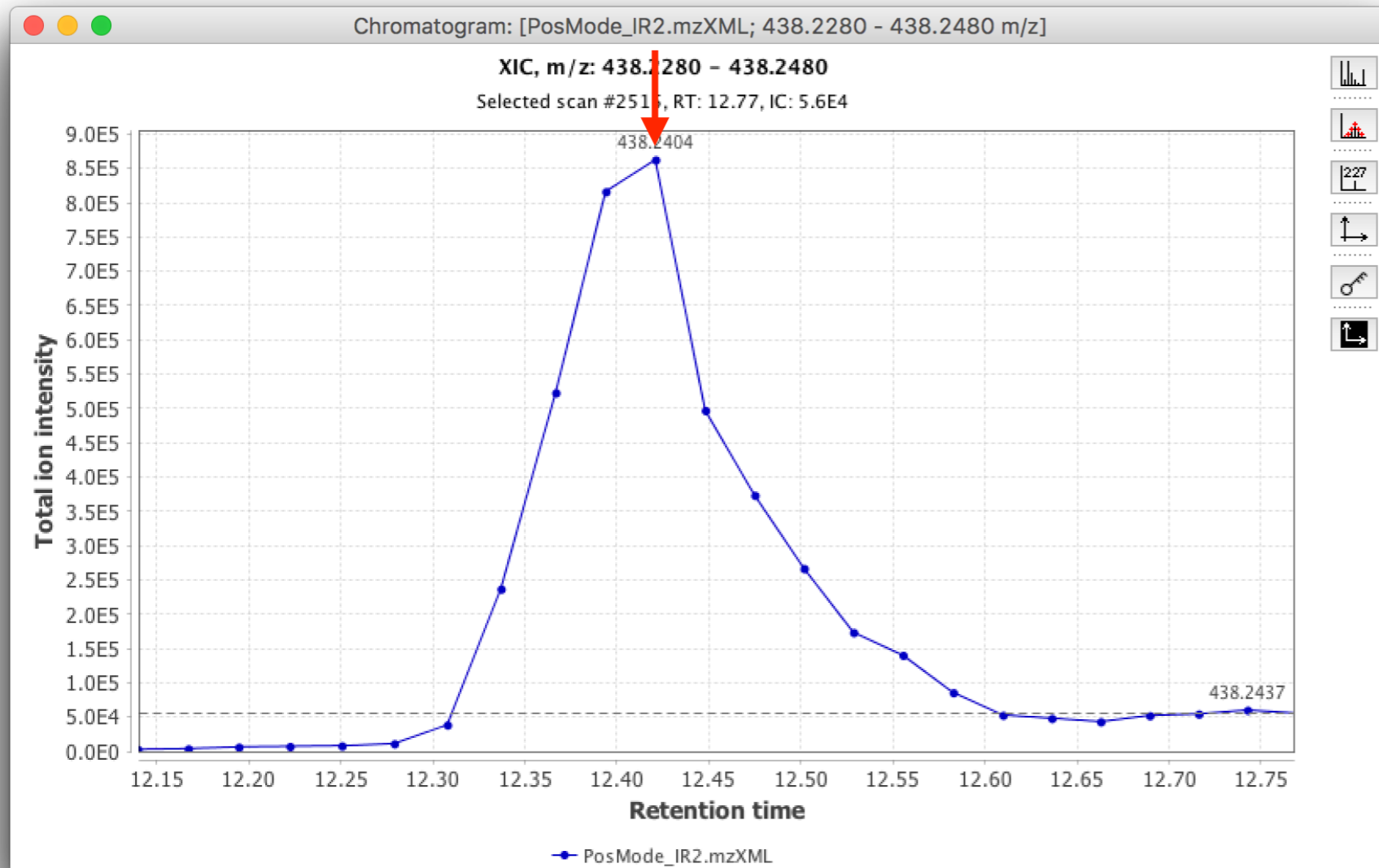


Tasks in progress...

Item

[3:38:16 AM]: Processing of task Updating TIC visualizer of PosMode\_IR1.mz

# Feature identification

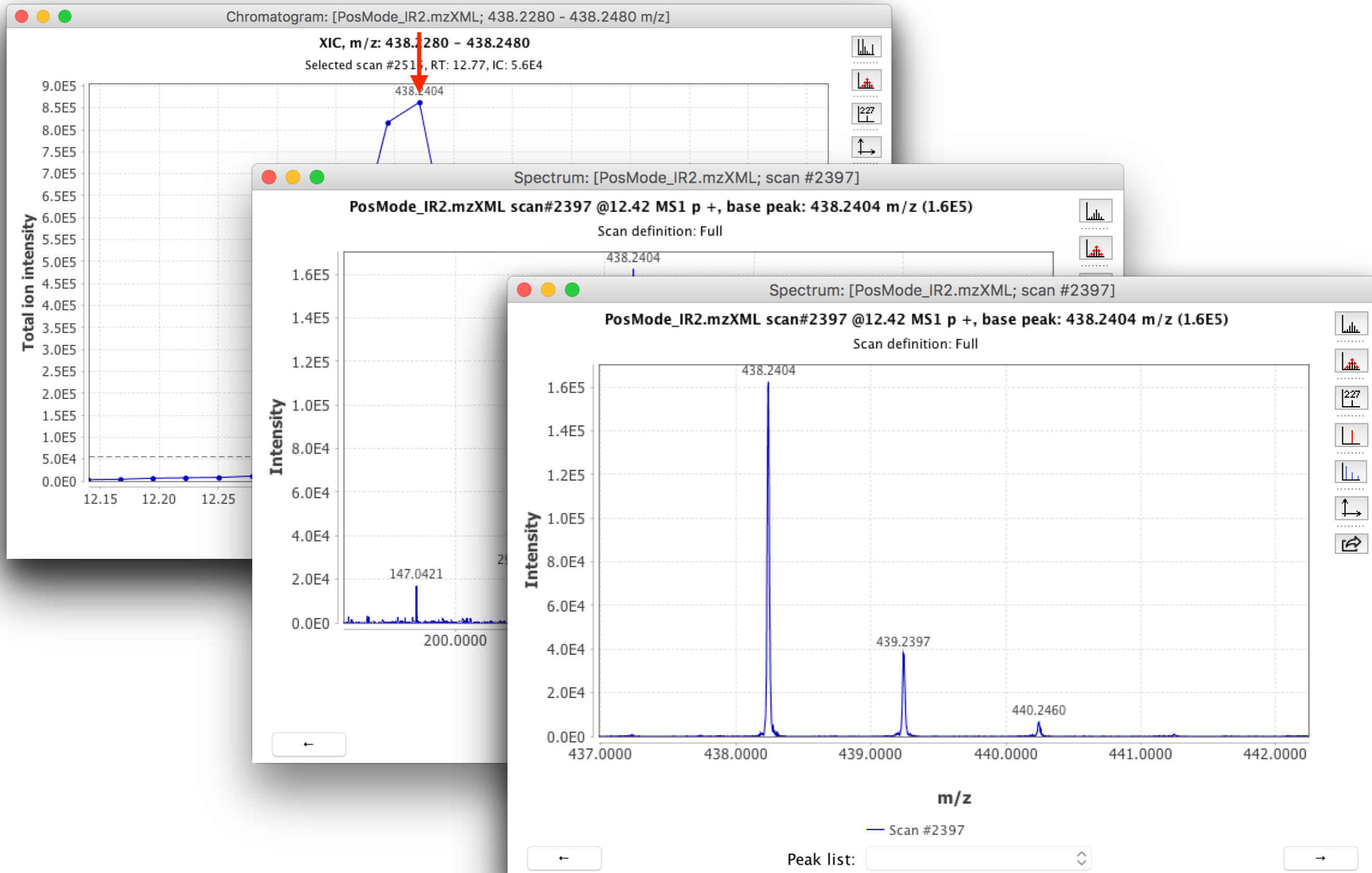


# Feature identification






# Feature identification



# Feature identification

← → ↻ [https://metlin.scripps.edu/metabo\\_search\\_alt2.php](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:

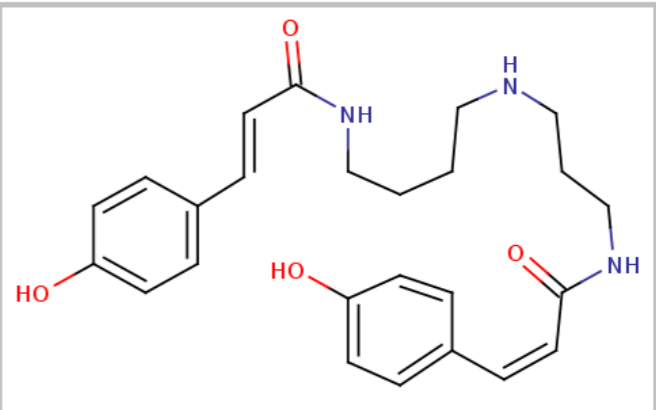
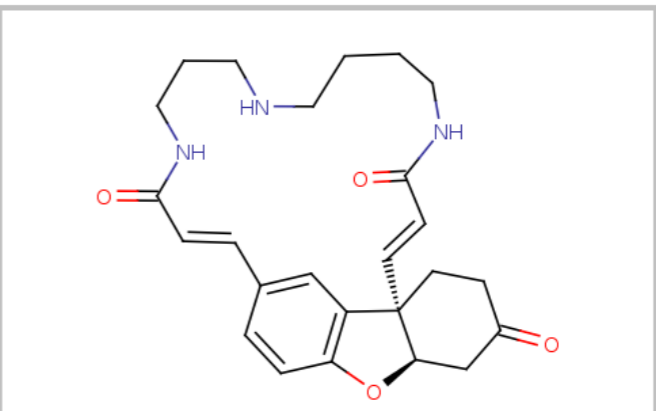
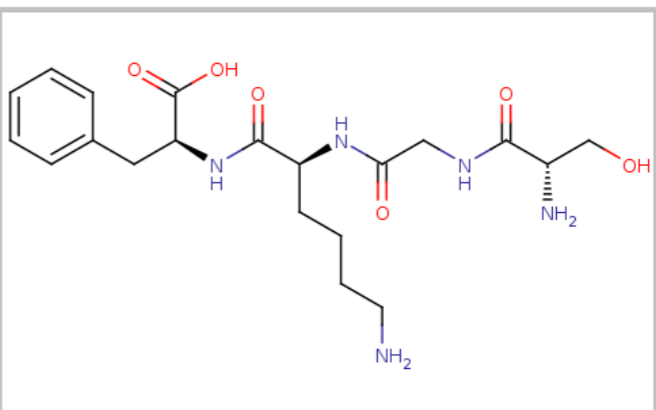
Neutral	M+H
Positive	M+NH4
Negative	M+Na
	M+H-2H2O
	M+H-H2O
	M+K
	M+ACN+H
	M+ACN+Na
	M+2Na-H
	M+2H
	M+3H
	M+H+Na
	M+2H+Na
	M+2Na
	M+2Na+H
	M+Li
	M+CH3OH+H

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

Remove peptides from search:

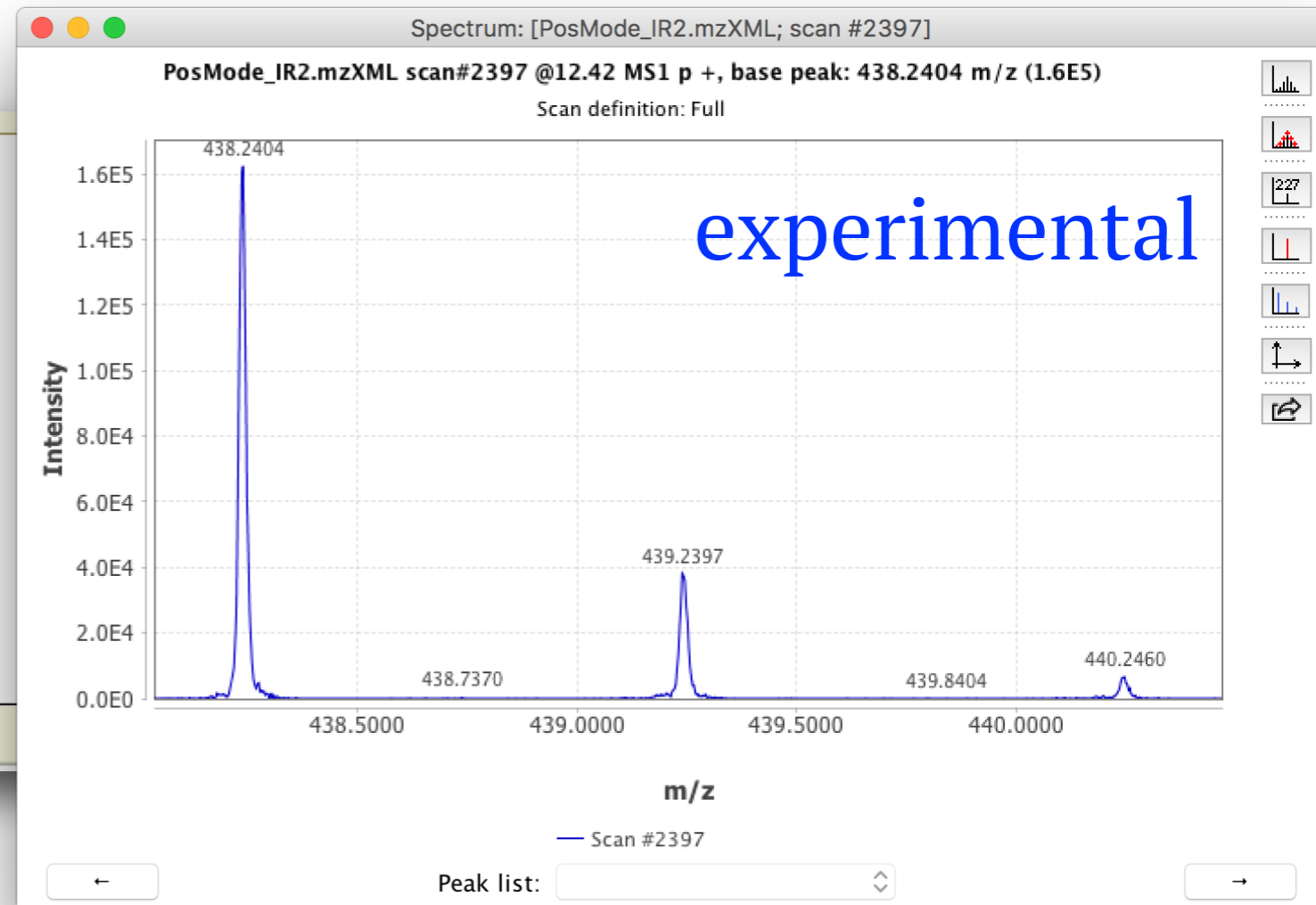
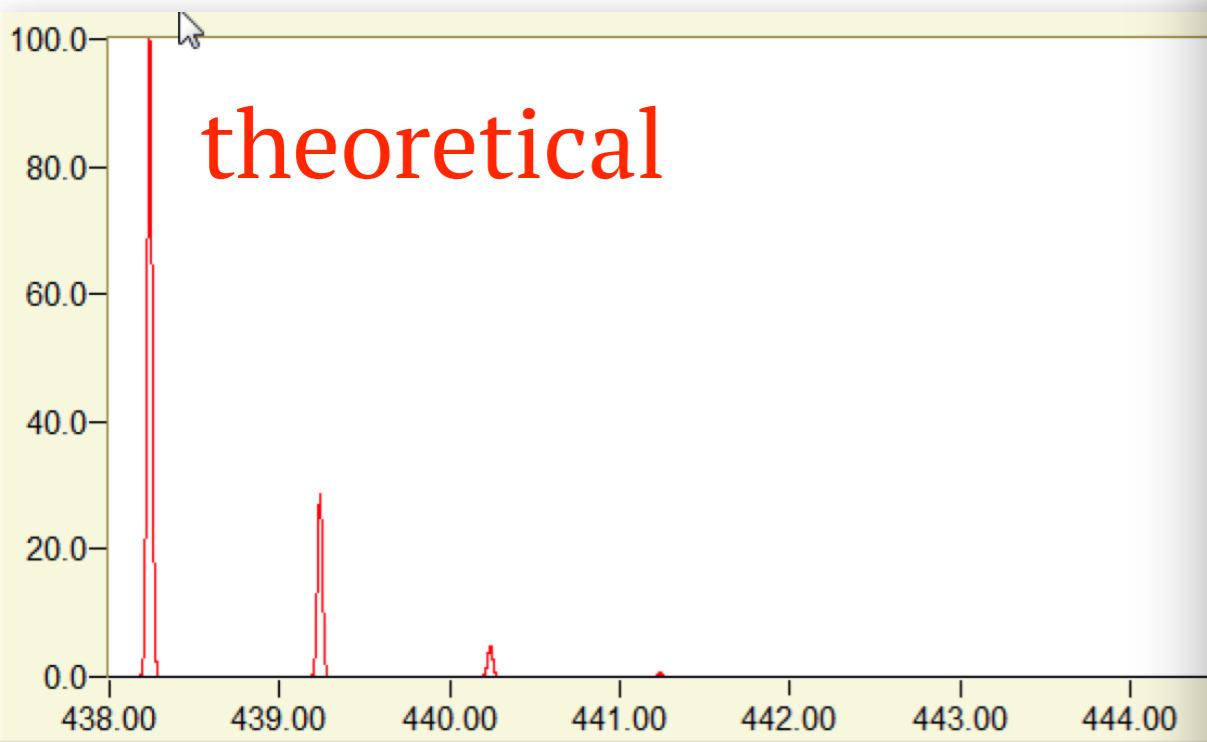
# Feature identification

Total: 87 Metabolites

METLIN ID	MASS	$\Delta$ ppm	NAME	MS/MS	STRUCTURE
89296 <input type="checkbox"/>	[M+H] <sup>+</sup> <u>m/z</u> 438.2387 M 437.2315	1	<b>N1,N10- Dicoumaroylspermidine</b> <i>Formula: C<sub>25</sub>H<sub>31</sub>N<sub>3</sub>O<sub>4</sub></i> <i>CAS: 65715-79-9</i>	NO	 The structure shows a central 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] <sup>+</sup> <u>m/z</u> 438.2387 M 437.2315	1	<b>Lunarine</b> <i>Formula: C<sub>25</sub>H<sub>31</sub>N<sub>3</sub>O<sub>4</sub></i> <i>CAS: 24185-51-1</i>	<a href="#">View</a>	 The structure shows a complex polycyclic molecule with a central coumarin core, a piperidine ring, and a long chain containing two secondary amine groups and a terminal amide group.
225643 <input type="checkbox"/>	[M+H] <sup>+</sup> <u>m/z</u> 438.2347 M 437.2274	7	<b>Ser Gly Lys Phe</b> <i>Formula: C<sub>20</sub>H<sub>31</sub>N<sub>5</sub>O<sub>6</sub></i> <i>CAS:</i>	NO	 The structure shows a linear peptide chain consisting of four amino acids: Phenylalanine, Serine, Lysine, and Glycine, with a hydroxyl group on the terminal glycine.
225567 <input type="checkbox"/>	[M+H] <sup>+</sup>	7	<b>Ser Gly Phe Lys</b>	NO	

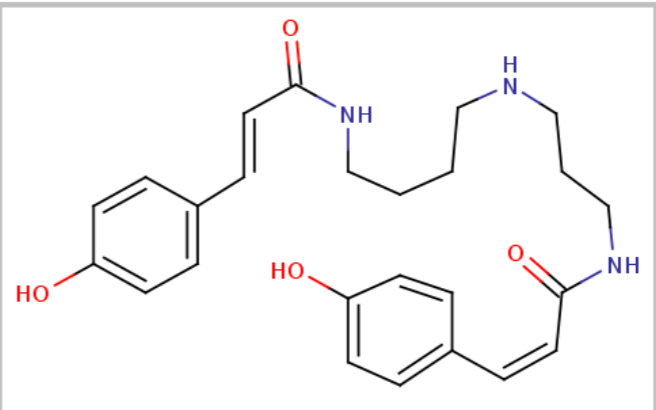
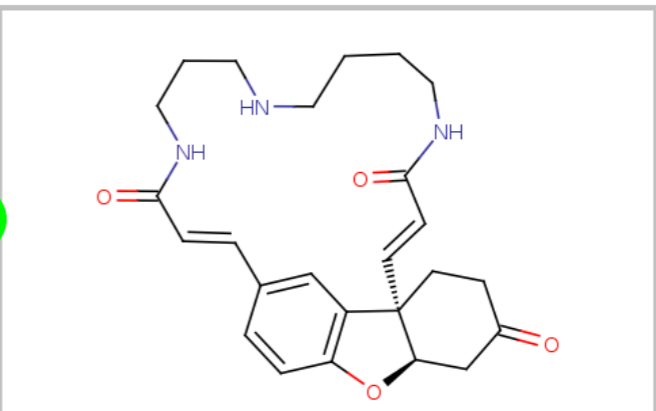
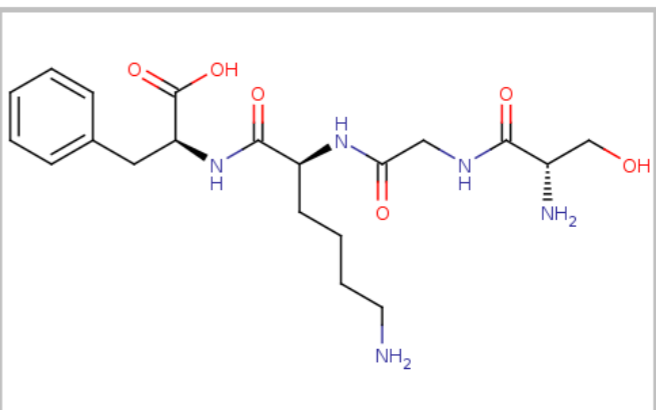
# Feature identification

- Compare isotopic distributions



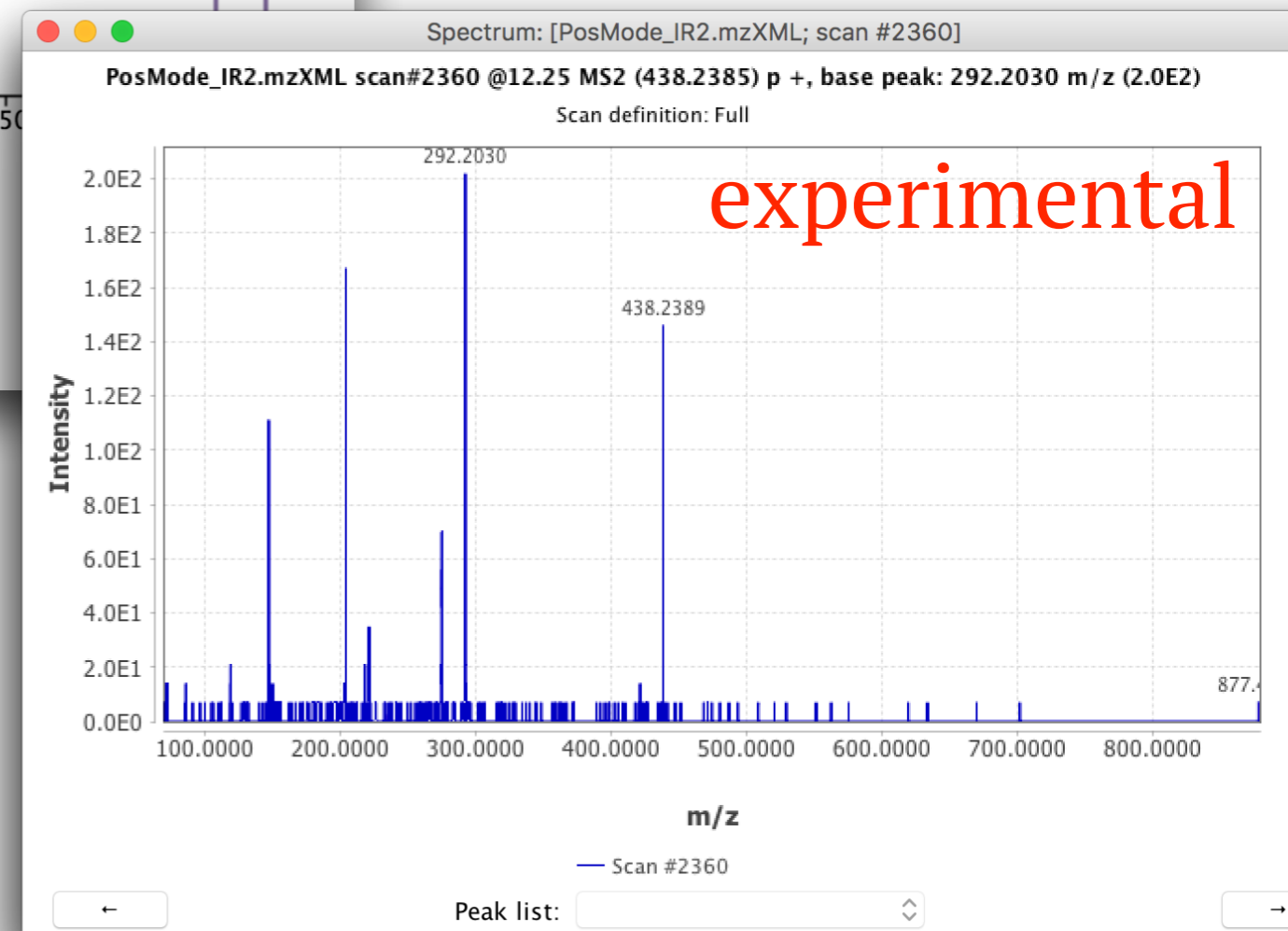
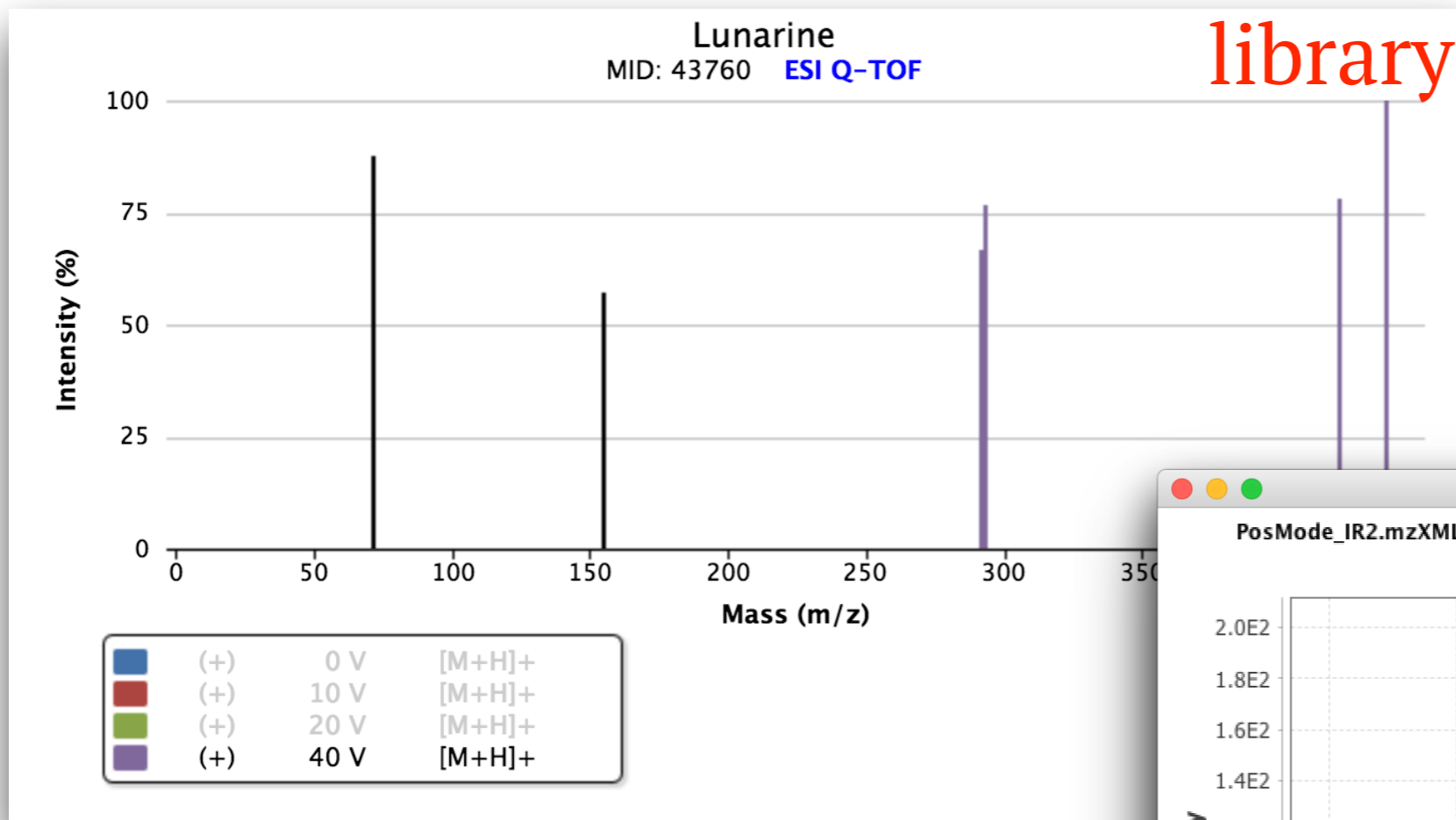
# Feature identification

Total: 87 Metabolites

METLIN ID	MASS	$\Delta$ ppm	NAME	MS/MS	STRUCTURE
89296 <input type="checkbox"/>	[M+H] <sup>+</sup> <u>m/z</u> 438.2387 M 437.2315	1	<b>N1,N10- Dicoumaroylspermidine</b> <i>Formula: C<sub>25</sub>H<sub>31</sub>N<sub>3</sub>O<sub>4</sub></i> <i>CAS: 65715-79-9</i>	NO	
43760 <input type="checkbox"/>	[M+H] <sup>+</sup> <u>m/z</u> 438.2387 M 437.2315	1	<b>Lunarine</b> <i>Formula: C<sub>25</sub>H<sub>31</sub>N<sub>3</sub>O<sub>4</sub></i> <i>CAS: 24185-51-1</i>	<a href="#">View</a>	
225643 <input type="checkbox"/>	[M+H] <sup>+</sup> <u>m/z</u> 438.2347 M 437.2274	7	<b>Ser Gly Lys Phe</b> <i>Formula: C<sub>20</sub>H<sub>31</sub>N<sub>5</sub>O<sub>6</sub></i> <i>CAS:</i>	NO	
225567 <input type="checkbox"/>	[M+H] <sup>+</sup>	7	<b>Ser Gly Phe Lys</b>	NO	

# Feature identification

- CompareMS/MS



**Thank you!**