05−MathDAMP−TwoGroups

This notebook provides a template for the comparison two groups of replicate datasets with the *Math*DAMP package. First, the respective group's datasets are averaged and compared in a similar way as demonstrated for two datasets in the 03−MathDAMP−TwoDatasets notebook. Additionally, *t*−test is performed for the groups of corresponding signal intensities from all datasets to locate statistically significant differences.
Additional notebooks from the *Math*DAMP package (04−MathDAMP−Outliers.nb and 06−MathDAMP−MultipleGroups.nb) provide templates for identifying outliers in a group of datasets and for the comparison of multiple groups of replicate datasets. The notebook 02−MathDAMP−Elements.nb demonstrates the basic functionality of the *Math*DAMP package.

Step 1 : Loading the Data

First, the *Math*DAMP package has to be loaded. Please assign the path leading to *Math*DAMP files to the `MathDAMPPath` variable. Due to the size of the datasets and results the global variable `$HistoryLength` is set to `1` to save memory. 1GB of physical memory may be necessary to execute this notebook.

```
$HistoryLength = 1;
MathDAMPPath = "/home/baran/math/ms/MathDAMP.1.0.0/";
<< (MathDAMPPath <> "MathDAMP.m")
```

```
MathDAMP version 1.0.0 loaded (2006/04/26)
```

```
This program is distributed in the hope that it will be useful, but WITHOUT
  ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

Datasets acquired by capillary electrophoresis coupled to a time−of−flight mass spectrometer (CE−TOFMS) will be used for the demonstration in this notebook. The *.bdt datafiles were generated using a separate in house software from *.csv datafiles exported by the Analyst QS software (for Agilent TOFMS). The csv data were binned to a 0.02 m/z units resolution along the m/z axis, baselines were subtracted from the individual electropherograms (as by the `DAMPSubtractBaselines` function with default options), noise was removed (as with the `DAMPRemoveNoise` function with default options), the data were binned to 1 m/z units resolution and saved in a binary format as *.bdt datafiles.

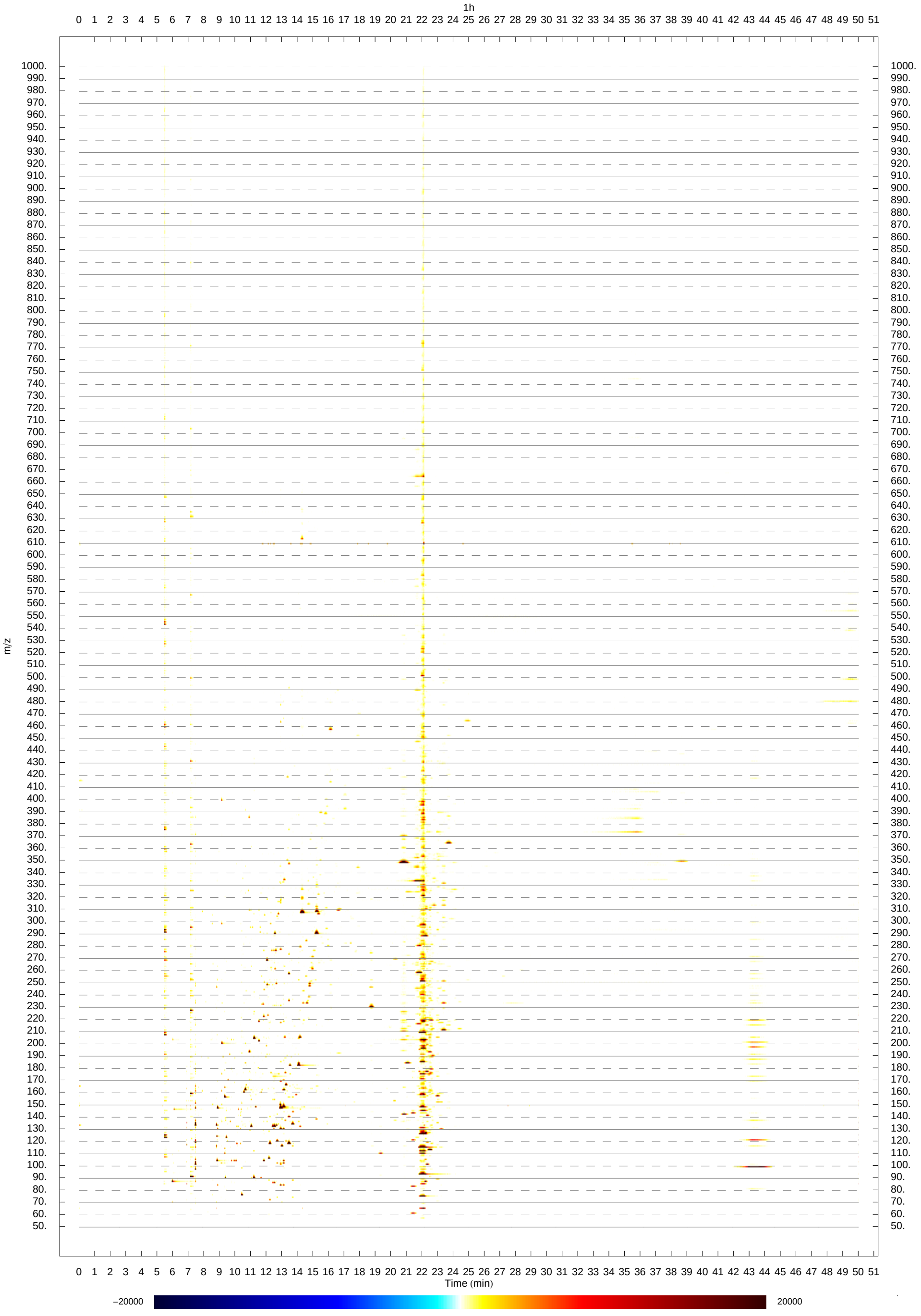
```
fnames = FileNames["/home2/baran/data/examples/" <> # <> "/*.bdt"] & /@ {"1h", "4h"};
{ctrl, smpl} = DAMPImportBDT[#, StringReplace[#, __ ~~ "/" ~~ str__ ~~ "/" ~~ __ → str]] & /@ # & /@ fnames;
NumberForm[MemoryInUse[], DigitBlock → 3]
```

```
301,988,872
```

■ Optional : Exploring the data, locating the peak of the internal standard in the reference dataset

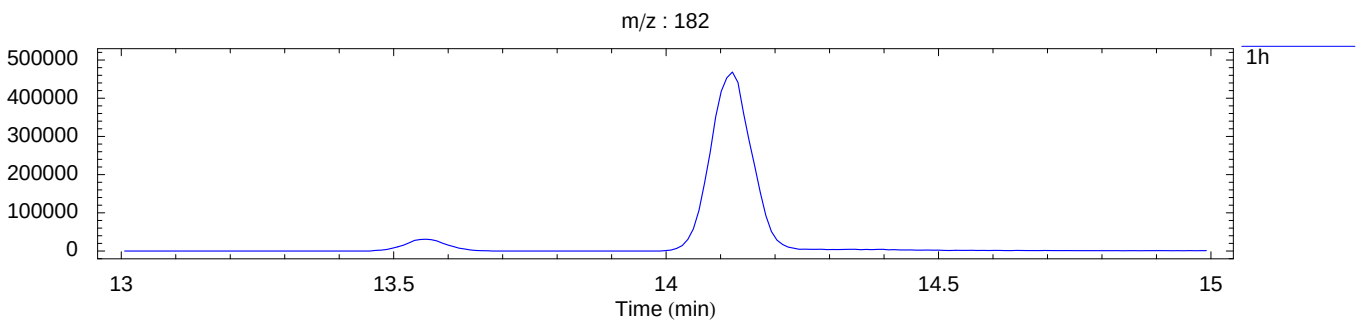
For a preliminary exploration, the loaded data may be visualized on density plots. The dimensions of the signal intensity matrix of the datasets are too large to grant at least one computer screen pixel per datapoint. Some signals may not be therefore visible on the density plot. For a high resolution view, please export the density plot to a postscript format or replot a larger version. The `Export` function may be used for saving the plot in a postscript format or the plot may be selected and the selection saved as EPS via the Edit menu entry. The postscript files may subsequently be converted to pdf format using tools as ghostscript for example.
Because of significant differences between the dimensions of datasets acquired by quadrupole and time−of−flight mass spectrometers, different default appearance of the density plots is desirable for CE−TOFMS data (sparser tickmarks and gridlines along the m/z axis, plot dimensions, etc.). The `DAMPCETOFMSDensityPlotOptions` variable groups the options which modify the appearance of the density plot for CE−TOFMS data.

```
DAMPDensityPlot[ctrl[[1]], MaxScale → 20000, Sequence @@ DAMPCETOFMSDensityPlotOptions];
```


1h
m/z
Time (min)
−20000
20000

If the location of the peak of the internal standard will be specified explicitly, it is necessary to locate it in the reference dataset (dataset to which the remaining datasets will be aligned and normalized).

```
DAMPPlotChromatogram[{ctrl[[1]]}, 182, PlotOptions → {PlotRange → {{13, 15}, All}}];
```



Step 2 : Performing the Differential Analysis

The function `DAMPTwoGoups` performs the comparison of two groups of replicate datasets. `DAMPNormalizeGroup` function is used internally to align and normalize the datasets along with the annotation tables. Please refer to the MathDAMP.nb notebook for more details about the implementation of functions `DAMPNormalizeGroup` and `DAMPTwoGroups`. Execute ?*FunctionName* to list a brief description of the respective function's available options.

```
?DAMPNormalizeGroup
```

```
DAMPNormalizeGroup[msdatas,options] aligns msdatas (a list of datasets) and normalizes them according to the areas of the peaks of the internal standard and
  external normalization coefficients (optional). The results are returned as a list of rules: {NormalizedDatasets->...,AlignedAnnotationTables->...}
Options:
Reference - position of the reference dataset within msdatas to which the remaining msdatas will be normalized (default: 1)
AlignmentTimeRange - peak picks from the reference dataset falling within
  this timerange only (specified as {starttime,endtime} in minutes) will be used for alignment (default: All)
RepresentativePeakOptions - options to be passed to the DAMPSelectRepresentativePeaks function to filter
  the initial peak picks (default: {PeaksPerChromatogram->5,PeaksPerInterval->8,IntervalSize->.5})
PeakPickingOptions - options to be passed to the DAMPPickPeaks function (default: {Threshold->5000})
PeakLayoutPlotOptions - options to be passed to the DAMPPlotPeakLayout function (default: {})
FitShiftFunctionOptions - options to be passed to the DAMPFitShiftFunction function (default: {})
AnnotationTables - a list of annotation tables to be aligned to the reference msdata (default: None)
OutputTimeRange - time range to which the resulting normalized datasets should be cropped (default: All)
ExternalNormalizationCoefficients - list of coefficients by which the signal intensities in msdatas
  will be multiplied. The number of coefficients in the list must equal the number of datasets in msdatas (default: None)
Resolution - resolution to which the datasets were binned along the m/z dimension. The annotation
  tables passed through the AnnotationTables option will be binned the same way (default: 1)
InternalStandard - the internal standard for signal intensity normalization may be specified in one of two ways: 1) short name (
  3rd column) from the first annotation table in the list passed via the AnnotationTables option. In this case the position of
  the internal standard will be extrapolated from the aligned annotation table and the vicinity blindly integrated 2) specifying
  the m/z and integration time range (as in the reference dataset) explicitly: {mz,{starttime,endtime}}. (default: None)
AutoISIntegrationVicinity - if the location of the internal standard is extrapolated from the aligned annotation table, this option
  determines the vicinity (in minutes) of the predicted retention/migration time to be blindly integrated (default: {-.25,.25})
SaveMemory - if set to true, signal intensities are rounded to integers in internal calculations and results (default: True)
```

```
?DAMPTwoGroups
```

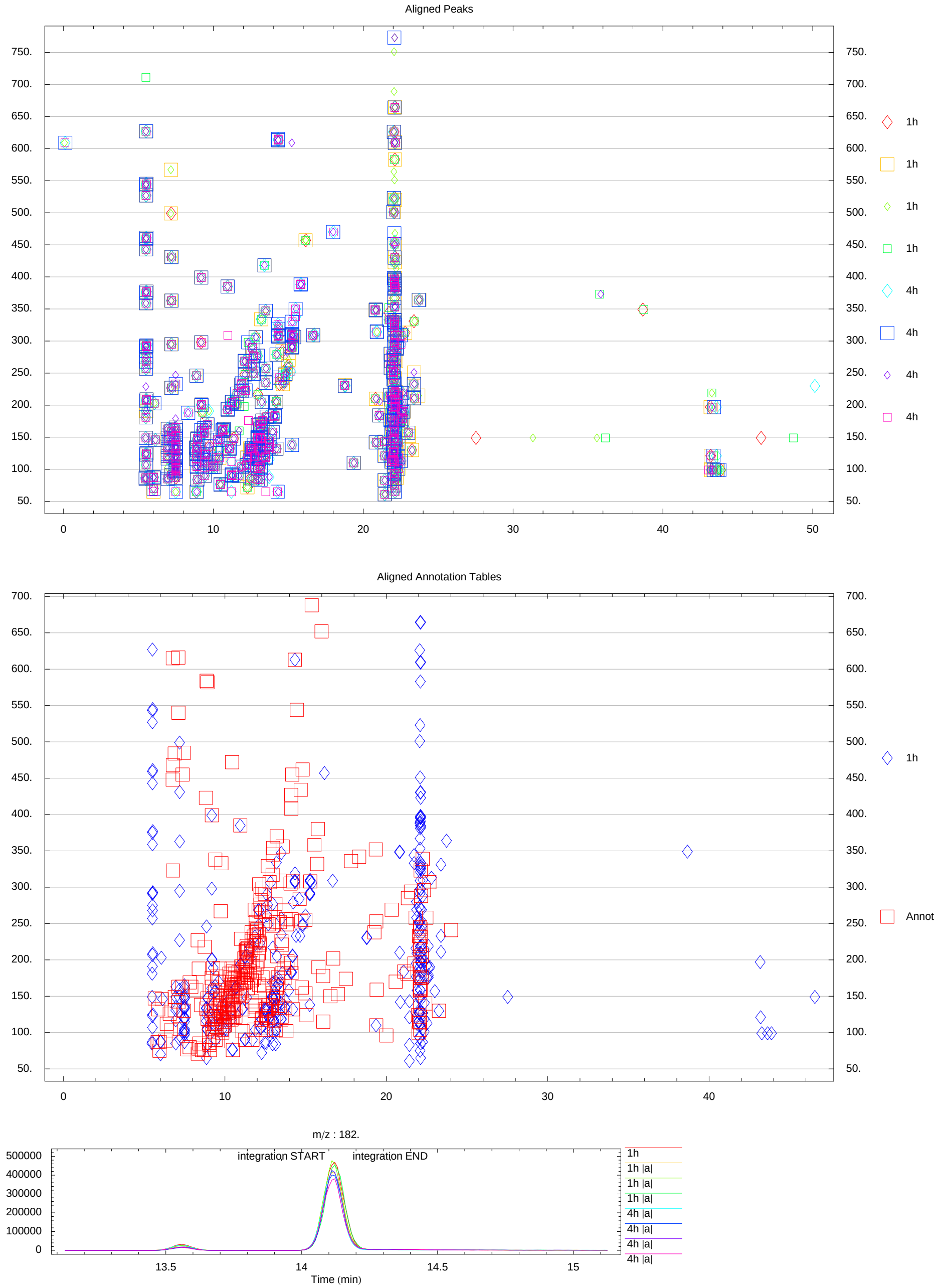
```
DAMPTwoGroups[msdatas1,msdatas2,options] generates datasets representing the absolute, relative, and absolute×relative differences between the averaged
  datasets of msdatas2 datasets and msdatas1 datasets. Additionally, a dataset representing t-scores between the groups of corresponding signal
  intensities from msdatas1 and msdatas2 datasets is generated. The results are returned as a list of rules: {NormalizedDatasets->...,AveragedGroup1->...,
  AveragedGroup2->...,Absolute->...,Relative->...,AbsRel->...,FilteredAbsRel->...,TScores->...,AlignedAnnotationTables->...,GroupCounts->...,GroupNames->...}
Options:
NormalizeGroupOptions - a list of options for the DAMPNormalizeGroup function which is used internally to normalize the datasets (default: {})
ThresholdForRelative - relative difference in the relative result will be set to 0 if neither of the
  two corresponding signal intensities (from the averaged datasets) is equal to or greater than this threshold (default: 0)
GroupNames - names to assign to groups (will be combined into the SampleName of the results). If set to
  Automatic, the SampleName of the first dataset from every group is used (default: Automatic)
AbsRelTrendFilter - determines the minimum number of individual corresponding signal intensities from every group which must follow the same trend as their
  averages to remain in the absolute×relative result. This is intended to filter out results originating from individual spikes or outliers (default: 2)
```

Most of the options for the `DAMPTwoGroups` and `DAMPNormalizeGroup` functions are specified explicitly in the command below to allow easy editing of the options. The annotation table for the cation mode CE−MS analysis is used. This table was assembled according to the CE−TOFMS analysis of a mixture of standard compounds. Methioninesulfone is used as the internal standard. Its short name (in the annotation table) 363 is passed to the `DAMPNormalizeGroup` function via the `InternalStandard` option. The location of the peak of the internal standard will be extrapolated from the aligned annotation table. Overlaid electropherograms of the vicinities of the expected peaks of the internal standard are plotted along with indicators of the beginning and the end of blindly integrated regions for visual confirmation. To specify the location of the peak explicitly, use the notation `{mz,{starttime,endtime}}` instead of the short name. In this case it would be `{182,{14,14.6}}` (according to the electropherogram at the end of the optional section).
The first dataset from the `ctrl` list will be used as the reference dataset (as specified by the options `Reference->1`).

The `DAMPTwoGroups` function returns the absolute, relative, and absolute×relative difference between the averages of the datasets of both groups. Additionally, a *t*−score map is generated by performing a *t*−test for the groups of corresponding signal intensities. In the case below, only peaks picked in the migration time range 8 − 23 min in the reference sample are used for the align−ment. This leads to a better alignment of the peaks in this range at the expense of the alignment of the stack of peaks with migration times around 25 min.

```
Clear[rslt];
rslt = DAMPTwoGroups[ctrl, smpl,
   NormalizeGroupOptions → {Reference → 1, AlignmentTimeRange → All, InternalStandard → 363, AutoISIntegrationVicinity → {-.2, .2},
     PeakPickingOptions → {Threshold → 5000}, RepresentativePeakOptions → {PeaksPerChromatogram → 5, PeaksPerInterval → 5, IntervalSize → .5},
     FitShiftFunctionOptions → {GapPenalty → {3, .5}}, PeakLayoutPlotOptions → DAMPCETOFMSPeakLayoutOptions,
     AnnotationTables → {DAMPLoadAnnotationTable[MathDAMPPath <> "/iab_cems_cation.csv"]}, OutputTimeRange → {4, 27},
     ExternalNormalizationCoefficients → None}, ThresholdForRelative → 0, GroupNames → Automatic, AbsRelTrendFilter → 2];
```

Picked Peaks
750.
700.
650.
600.
550.
500.
450.
400.
350.
300.
250.
200.
150.
100.
50.
0
10
20
30
40
1h
1h
1h
1h
4h
4h
4h
4h
Representative Peaks
600.
550.
500.
450.
400.
350.
300.
250.
200.
150.
100.
50.
0
10
20
30
40
1h
1h
1h
1h
4h
4h
4h
4h

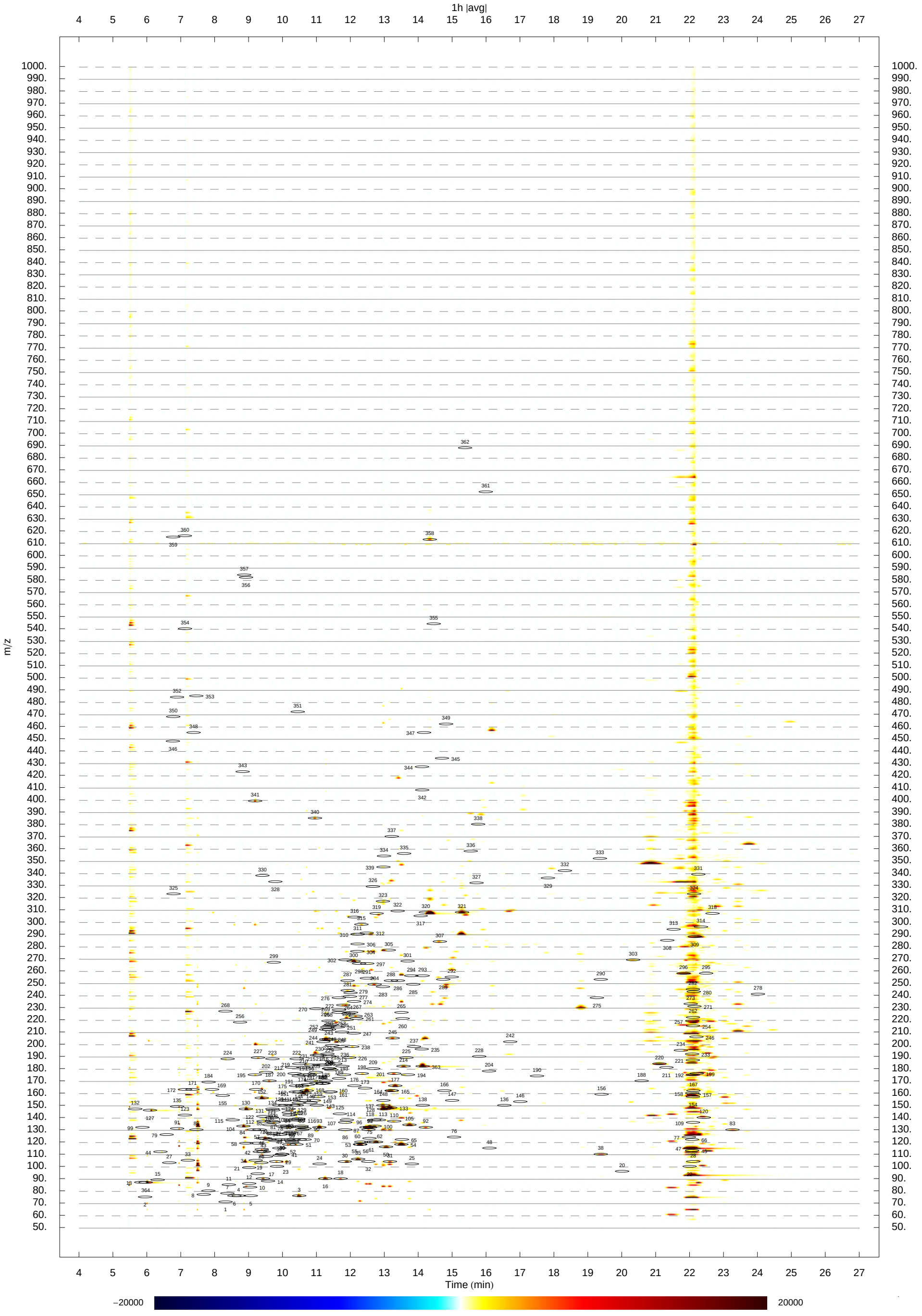


IS normalization coefficients : {1., 1.10901, 1.02771, 1.04268, 1.16882, 1.18705, 1.15713, 1.24596}

Step 3 : Exploring the Results, Listing the Candidates

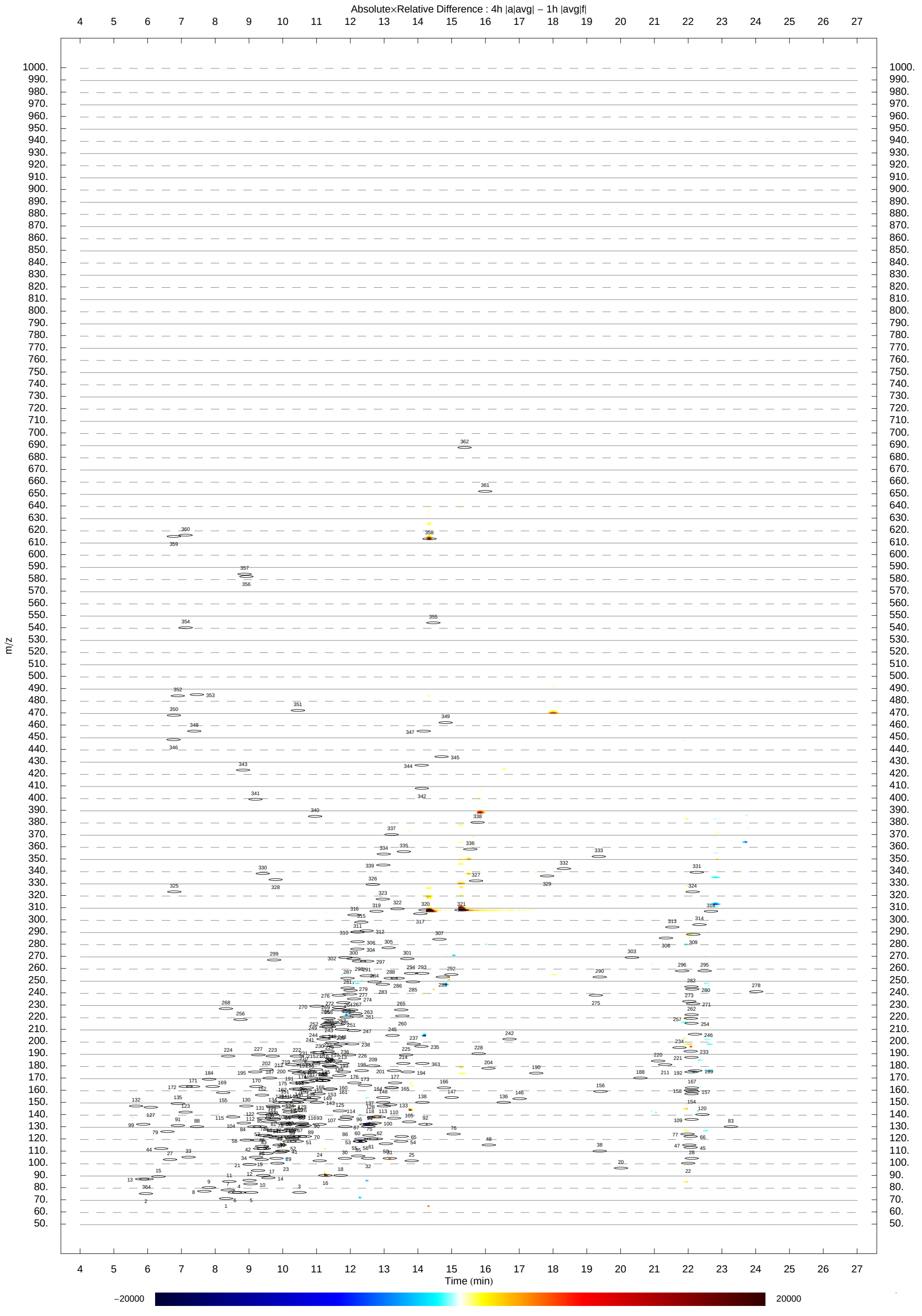
The normalized datasets may be explored on density plots (annotated). The averaged normalized datasets from the first group are shown on the plot below.

```
DAMPDensityPlot[AveragedGroup1 /. rslt, MaxScale → 20000,
  AnnotationTables → (AlignedAnnotationTables /. rslt), Sequence @@ DAMPCETOFMSDensityPlotOptions];
```


1h |avg|
m/z
Time (min)
−20000
20000

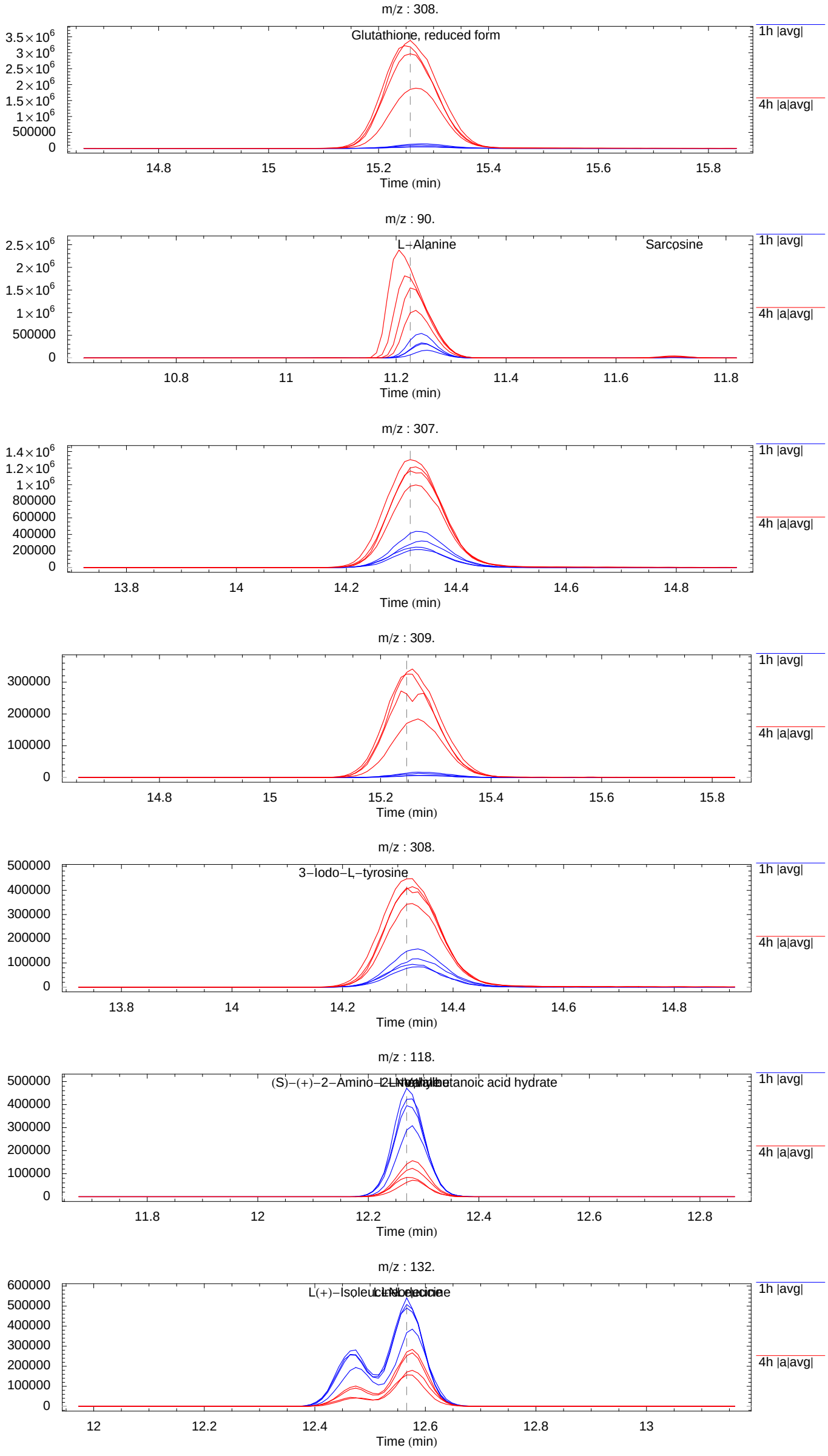
The *t*−score map or any of the result datasets calculated using the averages of the groups may be plotted to show differences between the groups. Additionally, the result datasets may be combined or used as filters against each other. For example, the absolute×relative difference (of the averages) result may be filtered to allow only those signals for which the *t*−score is above certain threshold.

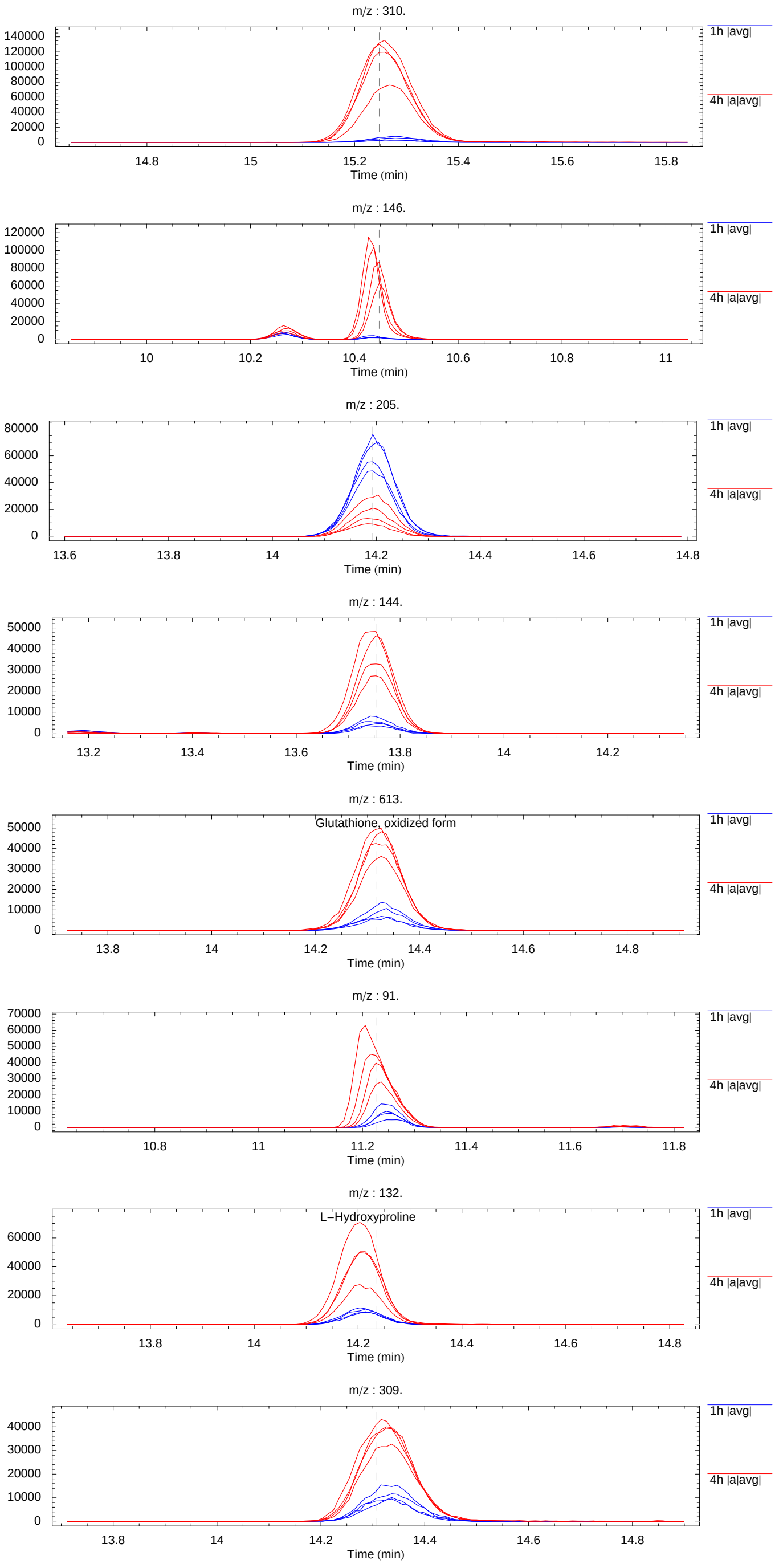
```
filtset = DAMPFilter[AbsRel /. rslt, DAMPSmooth[TScores /. rslt], 5];
DAMPDensityPlot[filtset, MaxScale → 20000, AnnotationTables → (AlignedAnnotationTables /. rslt), Sequence @@ DAMPCETOFMSDensityPlotOptions];
```


Absolute×Relative Difference : 4h |a|avg| − 1h |avg|f|
m/z
Time (min)
−20000
20000

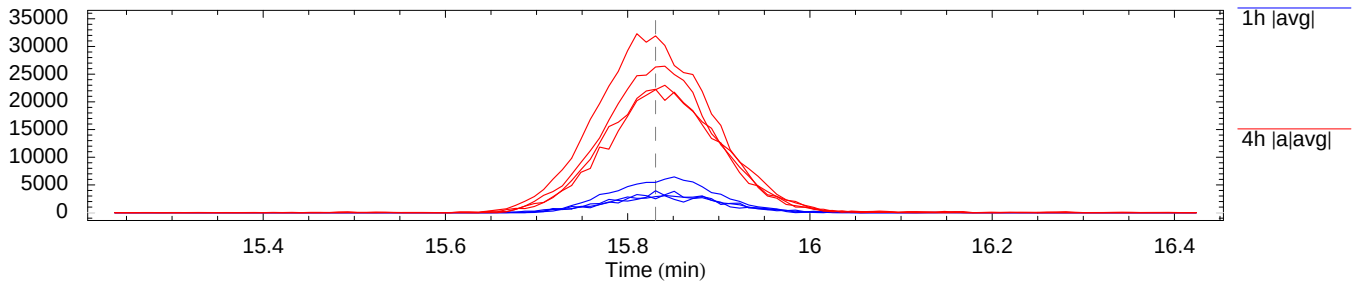
Overlaid electropherograms in the vicinities of the most significant differences (according to a particular result) may be plotted in descending order of significance for visual confirmation (and for the rejection of false positives). Below are the electropherograms of the top 50 candidates from the filtered absolute×relative difference result. The vertical dashed line indicates the position of the most significant difference according to the result dataset. (Further below are the top 50 candidates from the *t*−test result.)

```
plotcolors = Transpose[{DAMPGenColors[2], GroupCounts /. rslt}];
DAMPPlotCandidates[NormalizedDatasets /. rslt, filtset, PlotCount → 50,
  PlotChromatogramOptions → {PlotOptions → {PlotStyle → Join @@ (Table[#[[1]], {#[[2]]}] & /@ plotcolors)},
    AnnotationTable → (AlignedAnnotationTables /. rslt)[[1]], LegendData → Transpose[{plotcolors[[All, 1]], (GroupNames /. rslt)}]}];
```

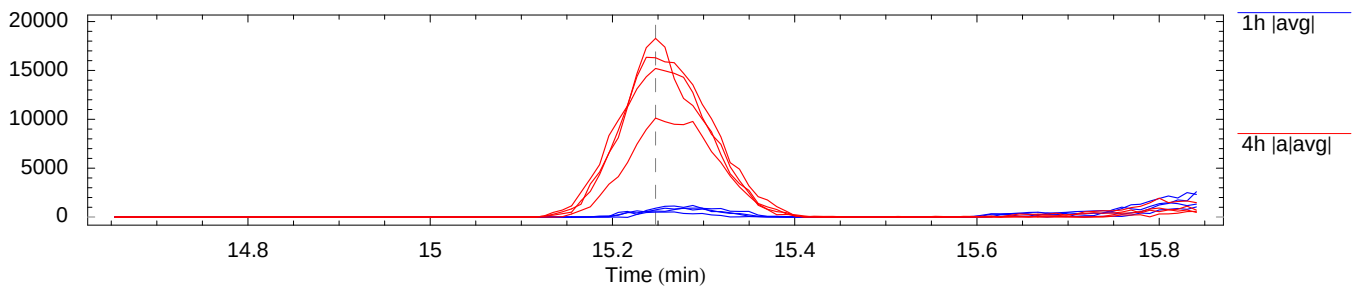



m/z : 310.
140000
120000
100000
80000
60000
40000
20000
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 146.
120000
100000
80000
60000
40000
20000
0
10
10.2
10.4
10.6
10.8
11
Time (min)
1h |avg|
4h |a|avg|
m/z : 205.
80000
60000
40000
20000
0
13.6
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 144.
50000
40000
30000
20000
10000
0
13.2
13.4
13.6
13.8
14
14.2
Time (min)
1h |avg|
4h |a|avg|
m/z : 613.
Glutathione, oxidized form
50000
40000
30000
20000
10000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 91.
70000
60000
50000
40000
30000
20000
10000
0
10.8
11
11.2
11.4
11.6
11.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 132.
L−Hydroxyproline
60000
40000
20000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 309.
40000
30000
20000
10000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|

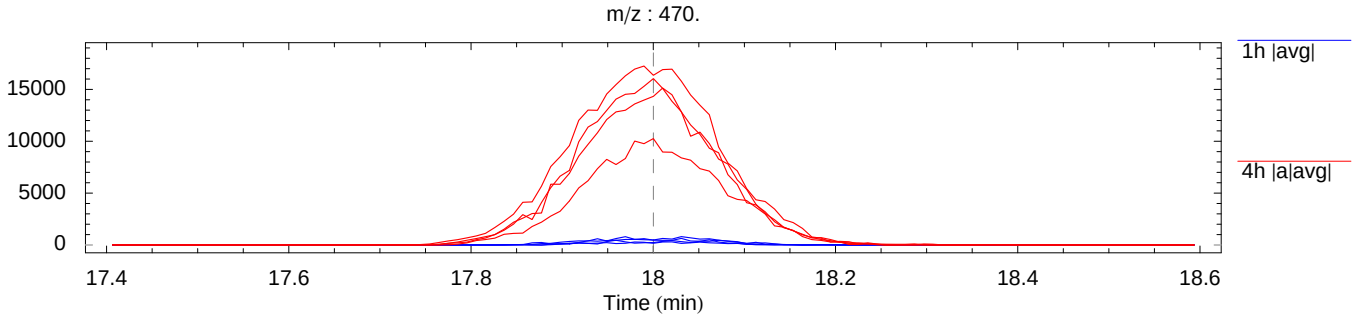
m/z : 388.



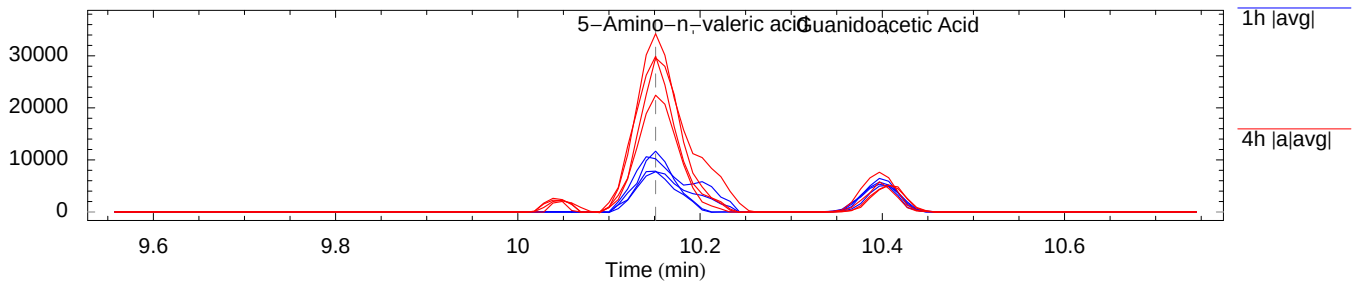
m/z : 311.



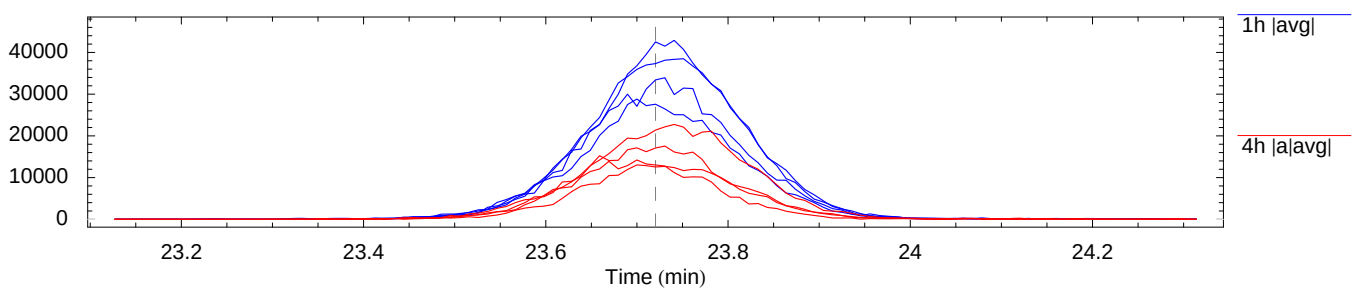
m/z : 470.



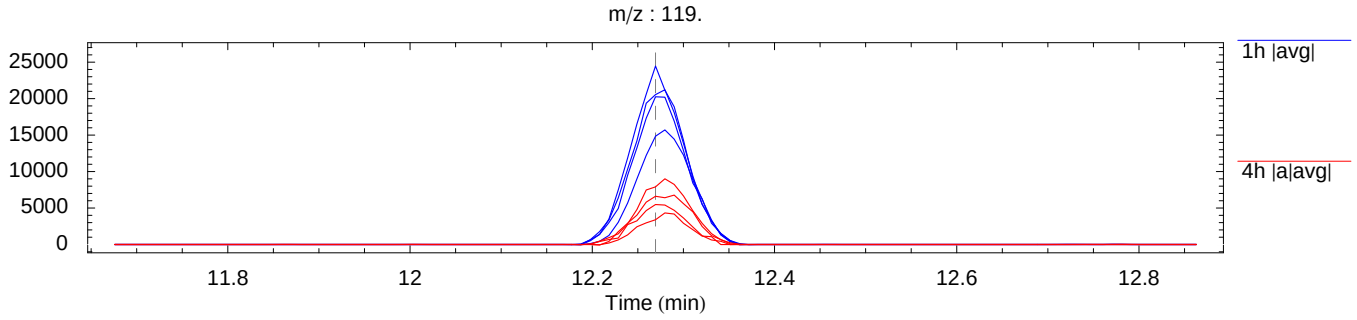
m/z : 118.



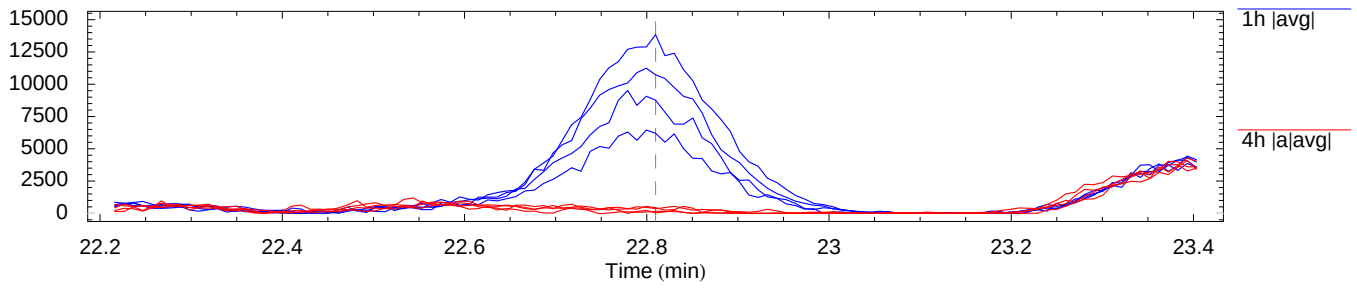
m/z : 364.



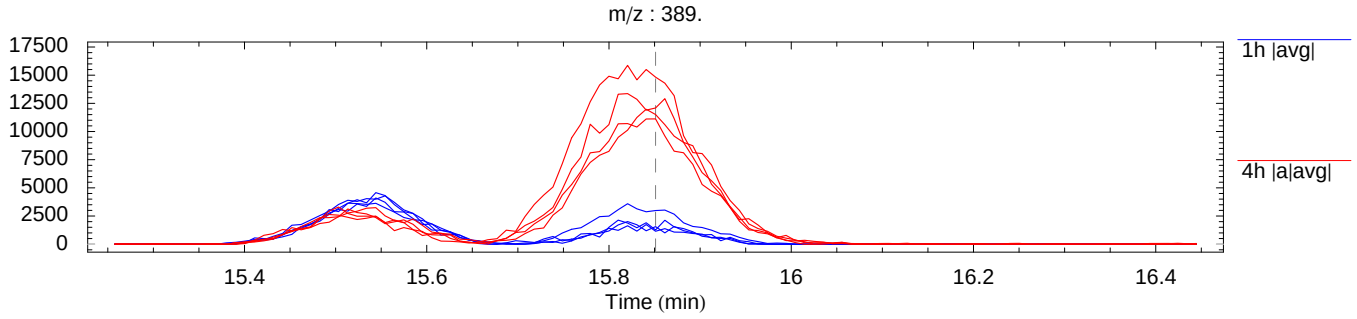
m/z : 119.



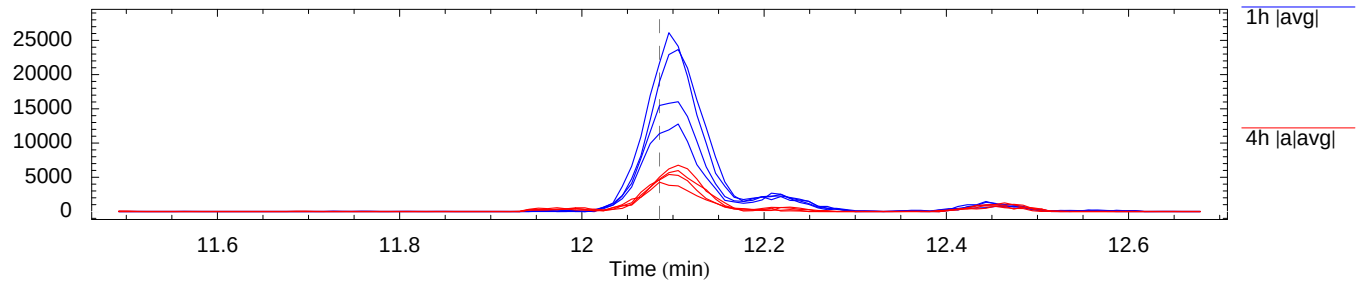
m/z : 313.



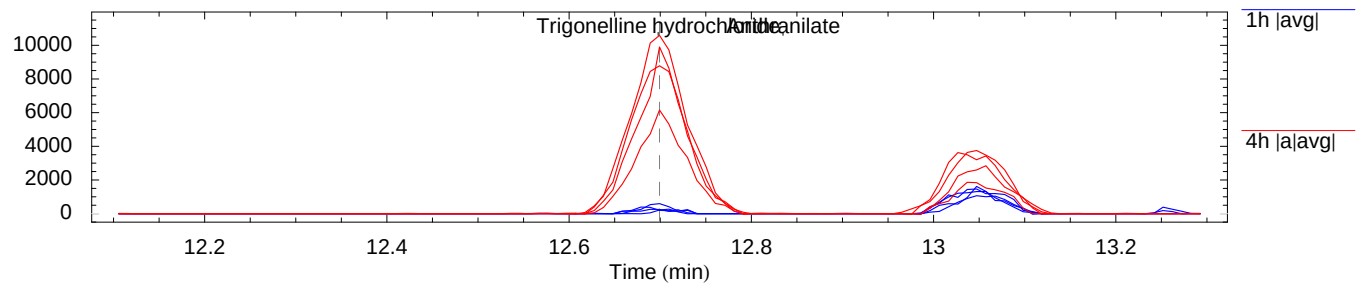
m/z : 389.



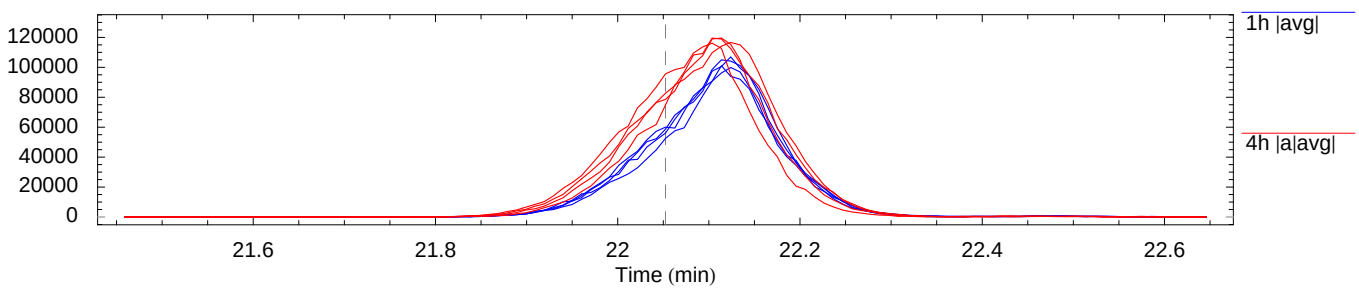
m/z : 248.



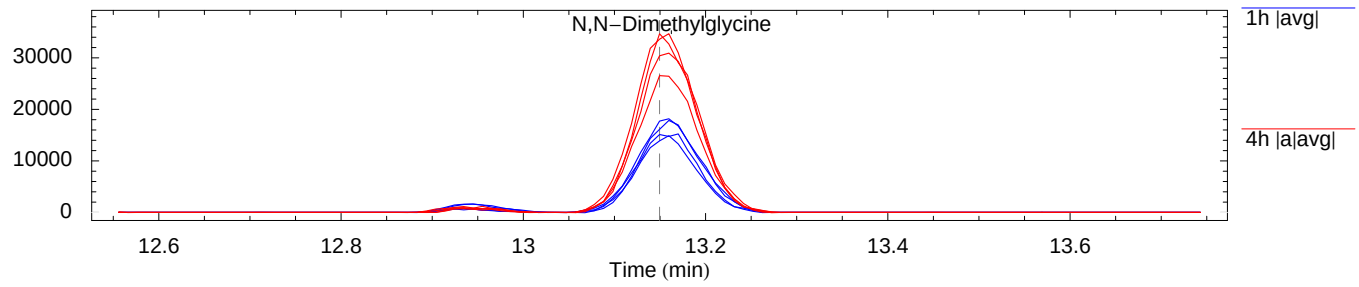
m/z : 138.



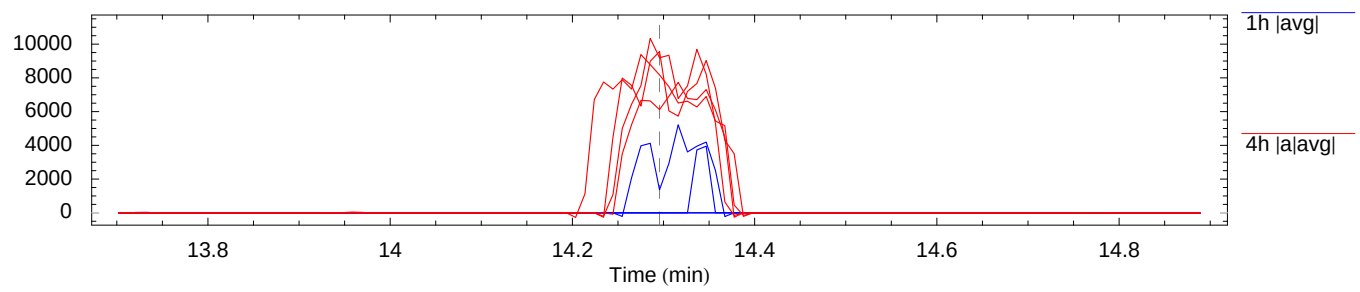
m/z : 196.



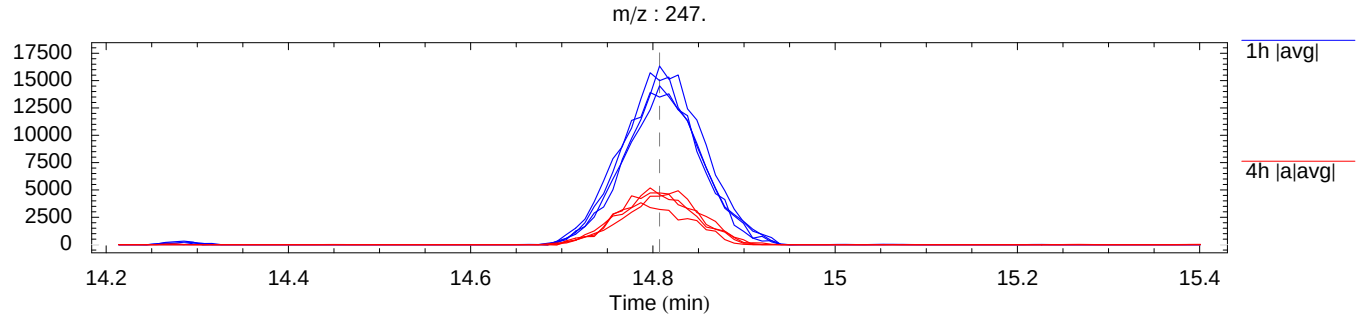
m/z : 104.



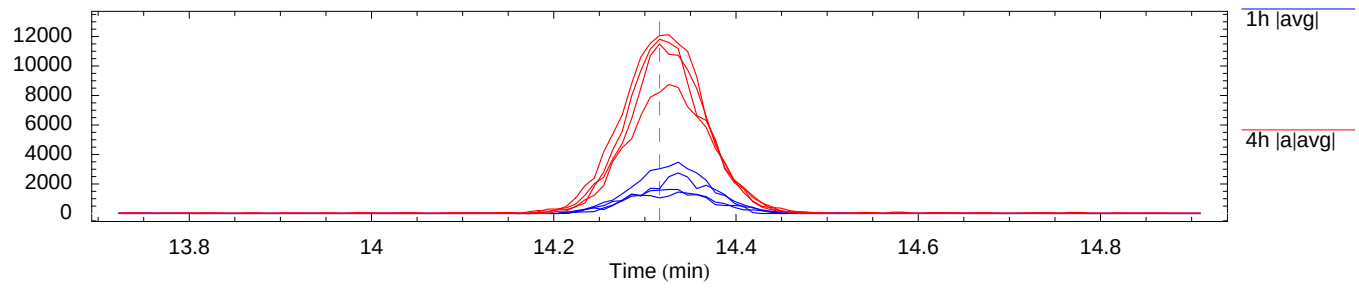
m/z : 65.



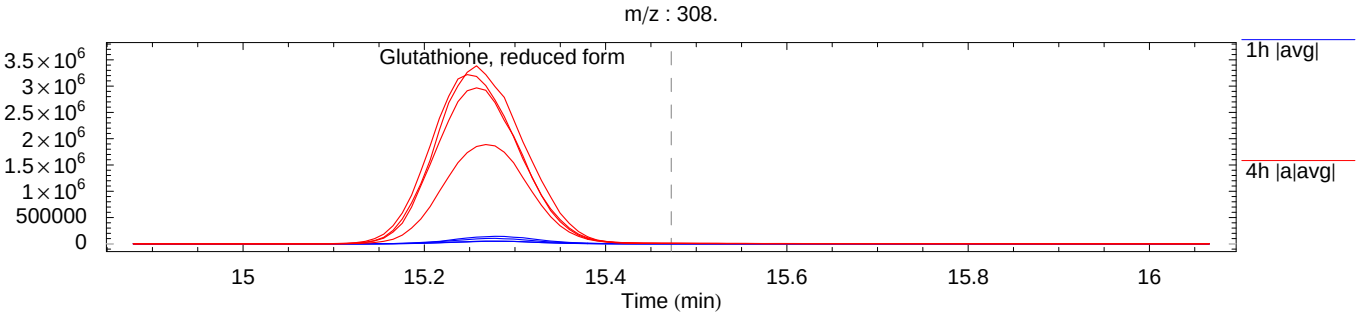
m/z : 247.



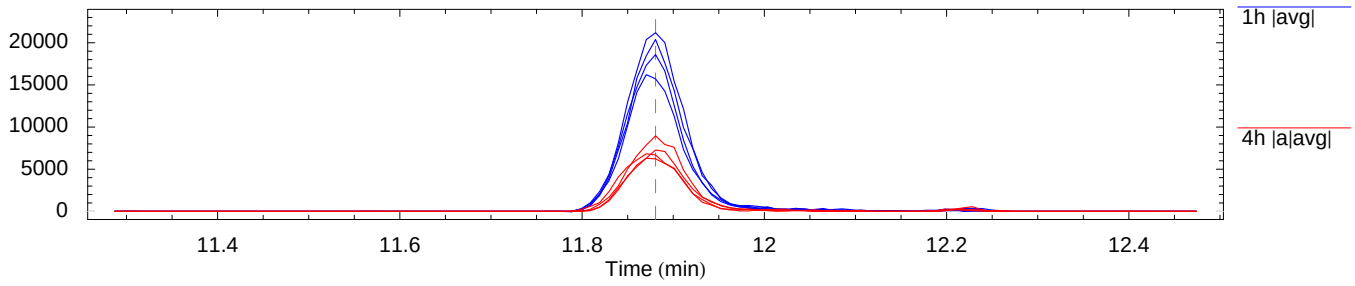
m/z : 614.



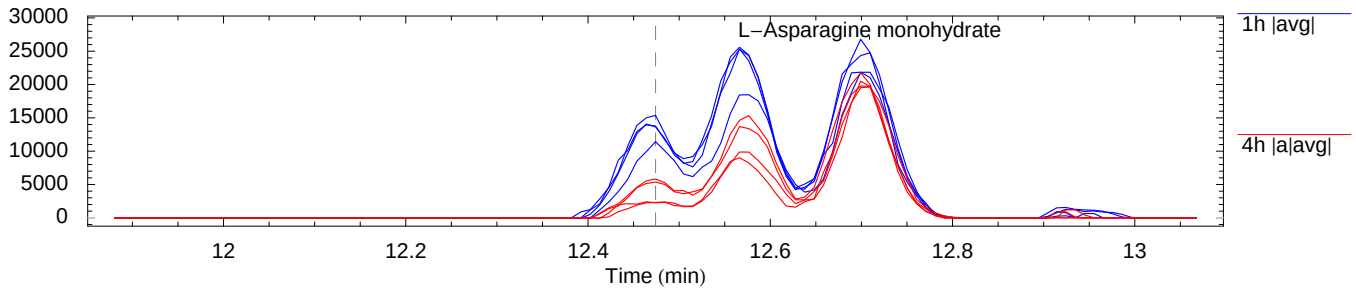
m/z : 308.



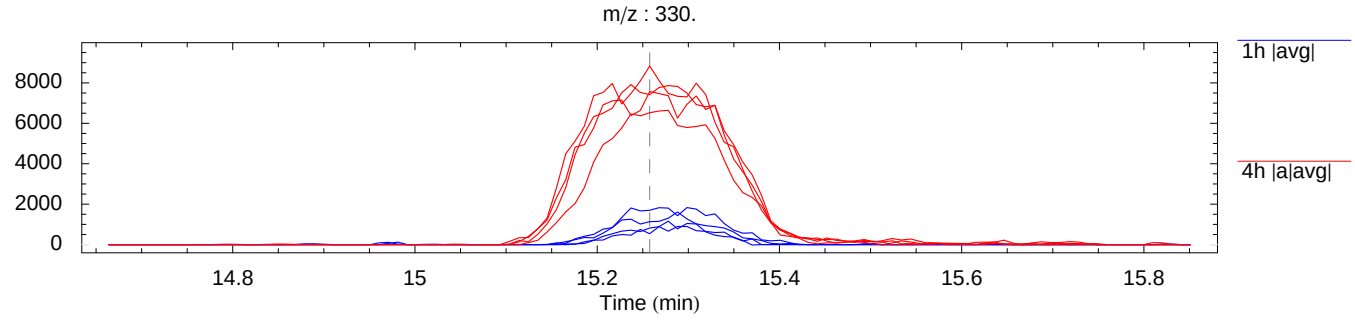
m/z : 222.



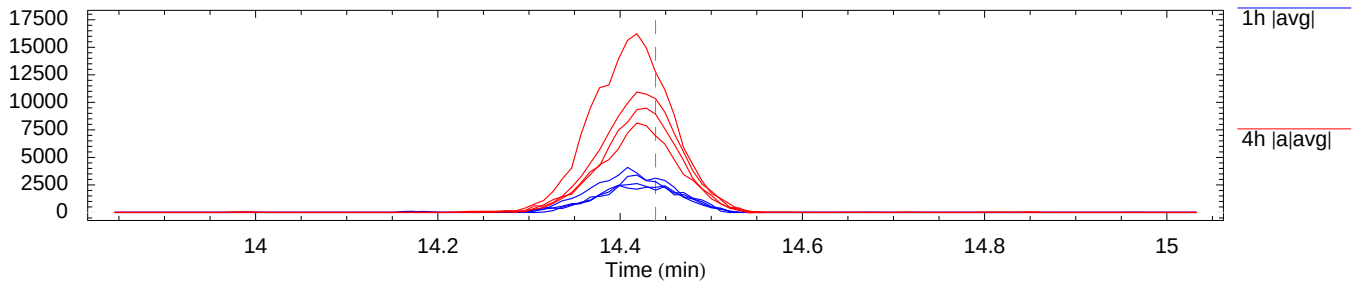
m/z : 133.



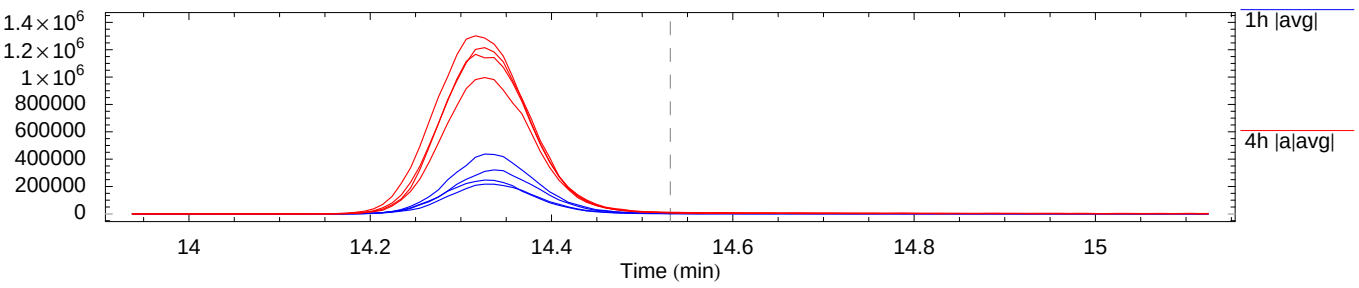
m/z : 330.



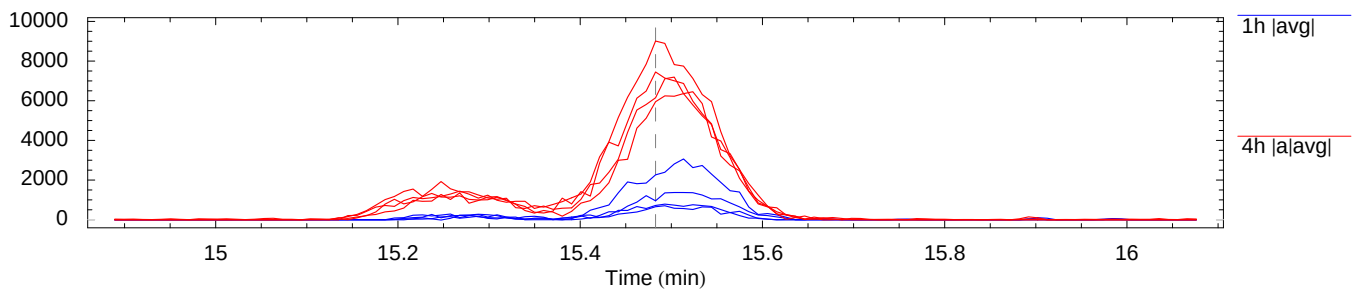
m/z : 243.



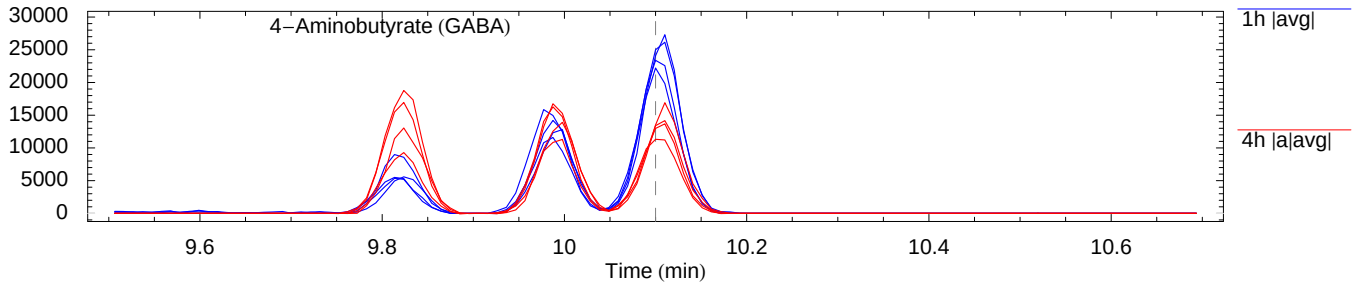
m/z : 307.



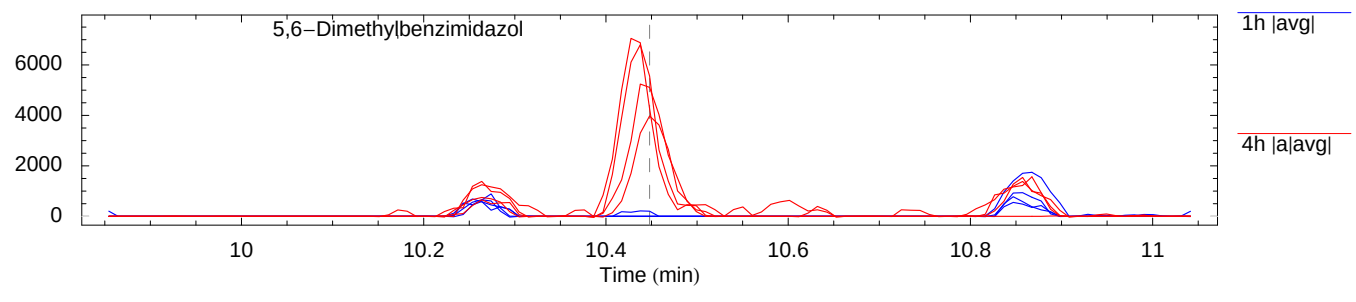
m/z : 350.

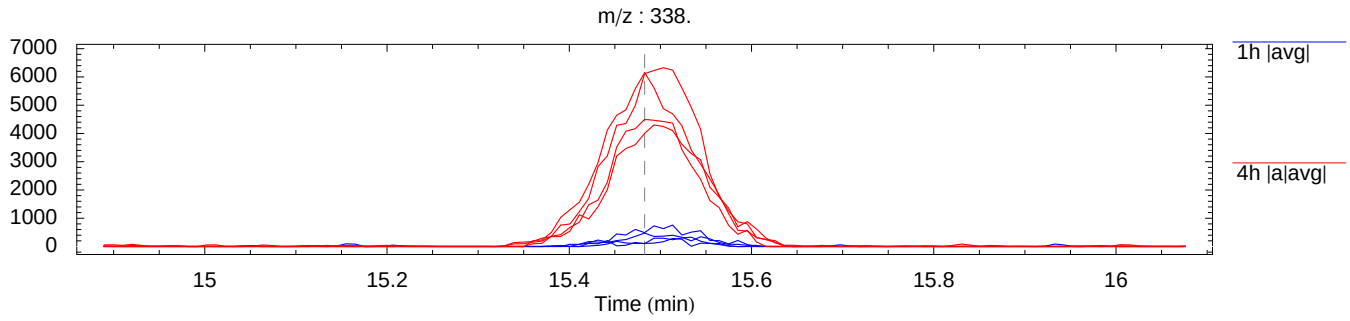


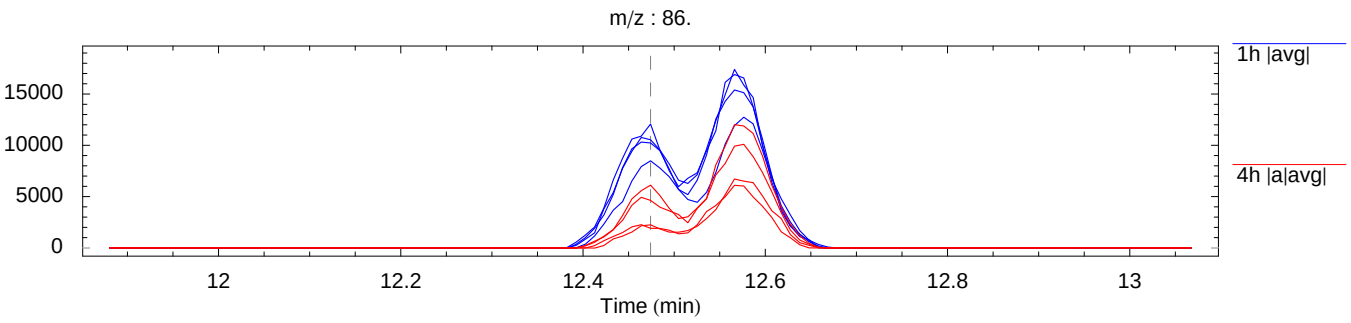
m/z : 104.

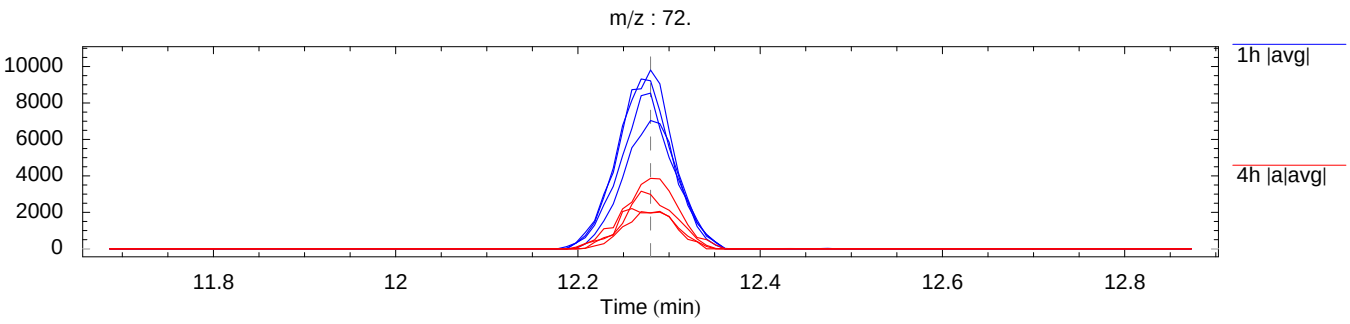


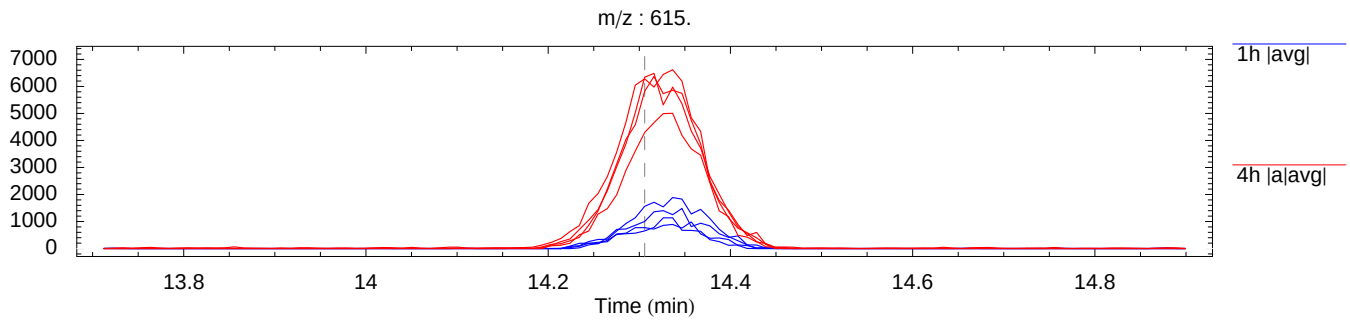
m/z : 147.

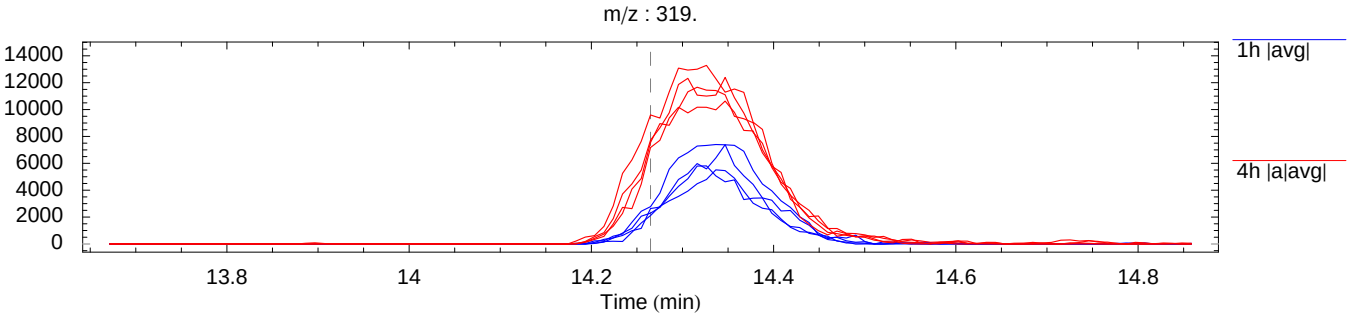


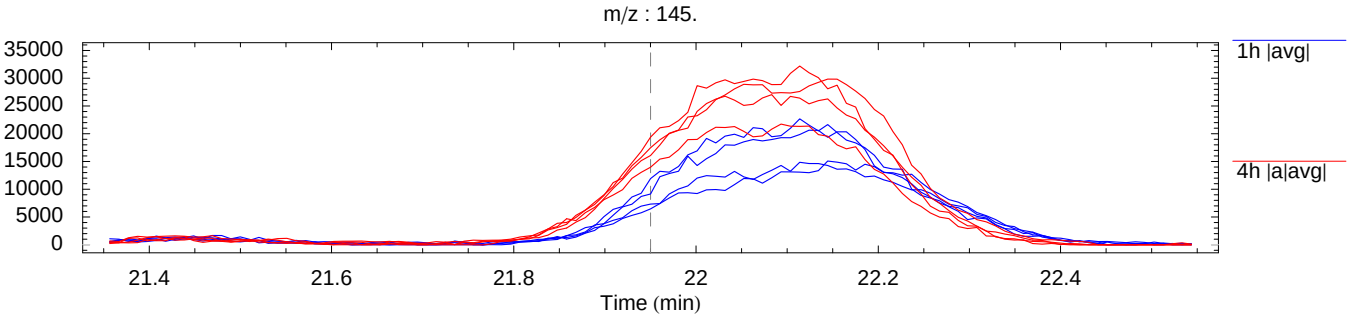

m/z : 338.
7000
6000
5000
4000
3000
2000
1000
0
15
15.2
15.4
15.6
15.8
16
Time (min)
1h |avg|
4h |a|avg|

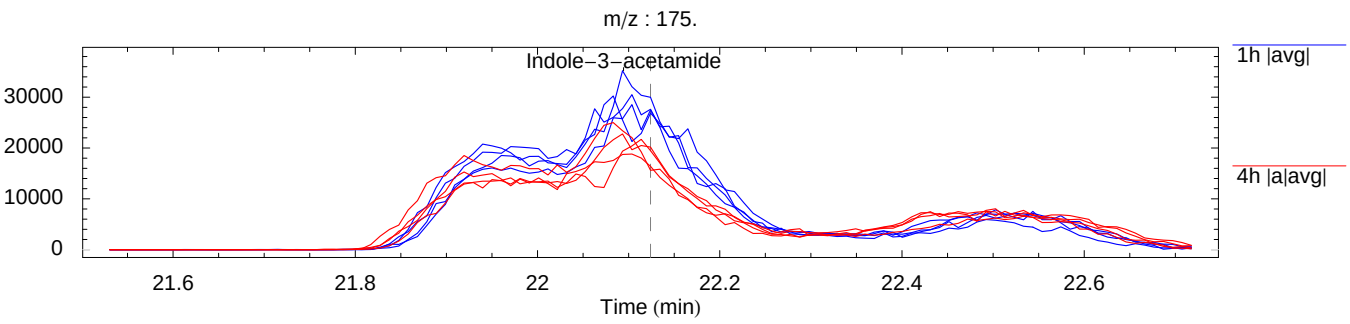

m/z : 86.
15000
10000
5000
0
12
12.2
12.4
12.6
12.8
13
Time (min)
1h |avg|
4h |a|avg|

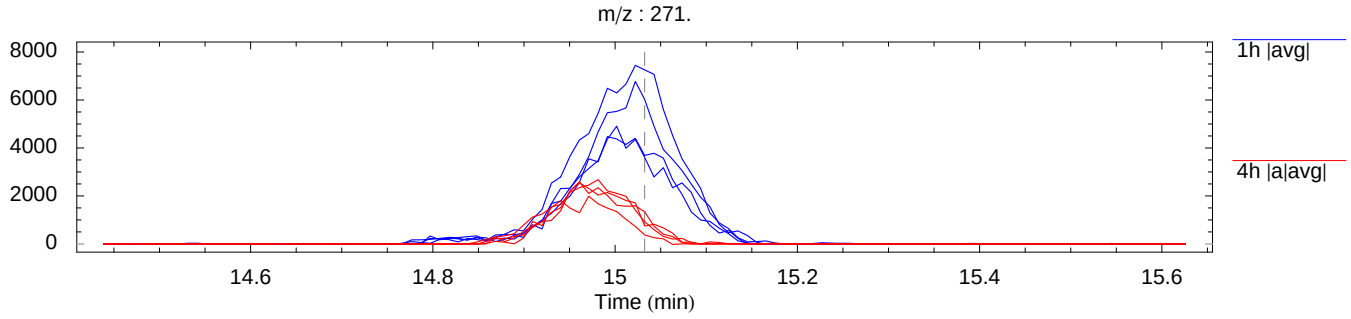

m/z : 72.
10000
8000
6000
4000
2000
0
11.8
12
12.2
12.4
12.6
12.8
Time (min)
1h |avg|
4h |a|avg|

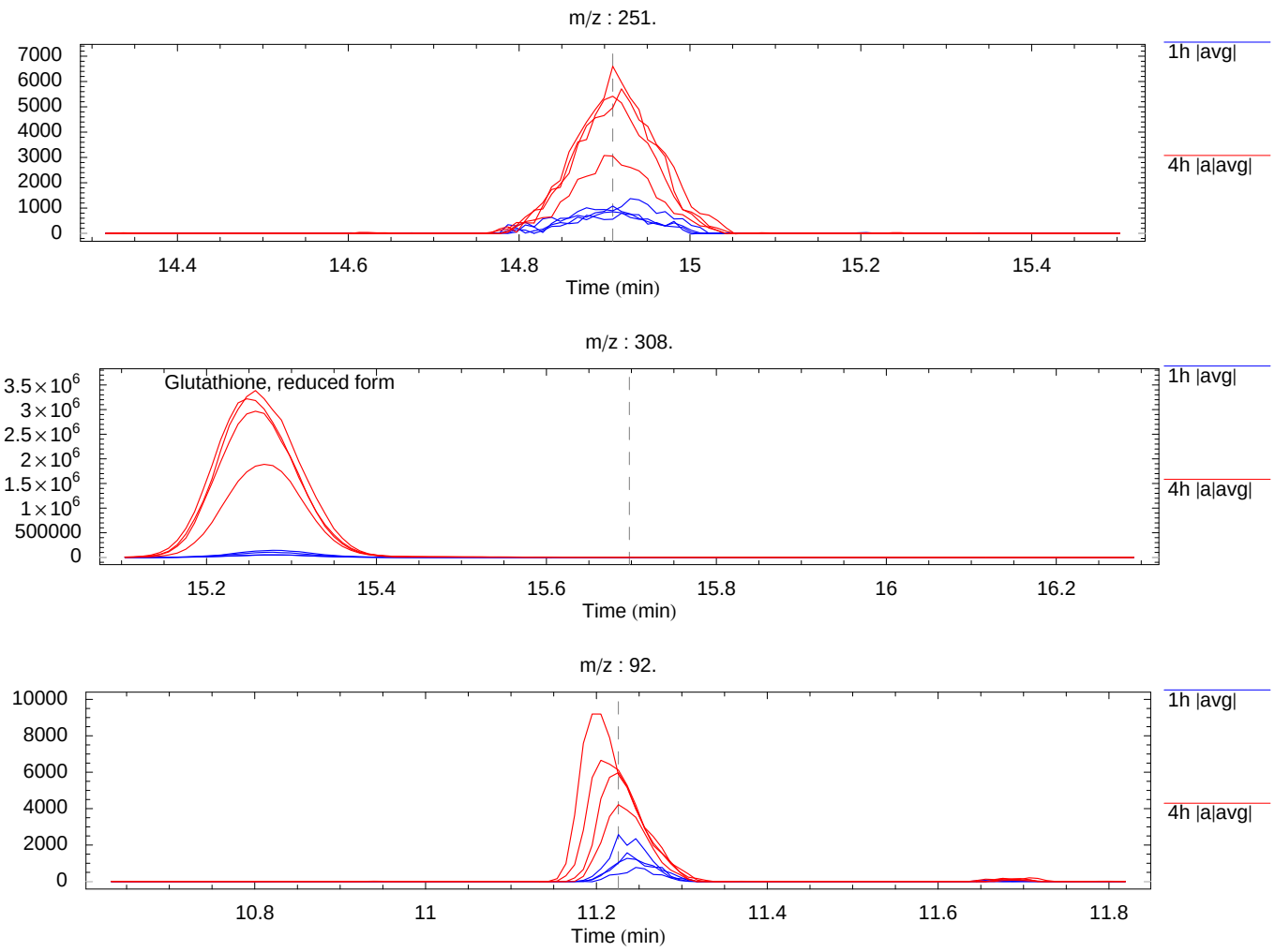

m/z : 615.
7000
6000
5000
4000
3000
2000
1000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|


m/z : 319.
14000
12000
10000
8000
6000
4000
2000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|


m/z : 145.
35000
30000
25000
20000
15000
10000
5000
0
21.4
21.6
21.8
22
22.2
22.4
Time (min)
1h |avg|
4h |a|avg|

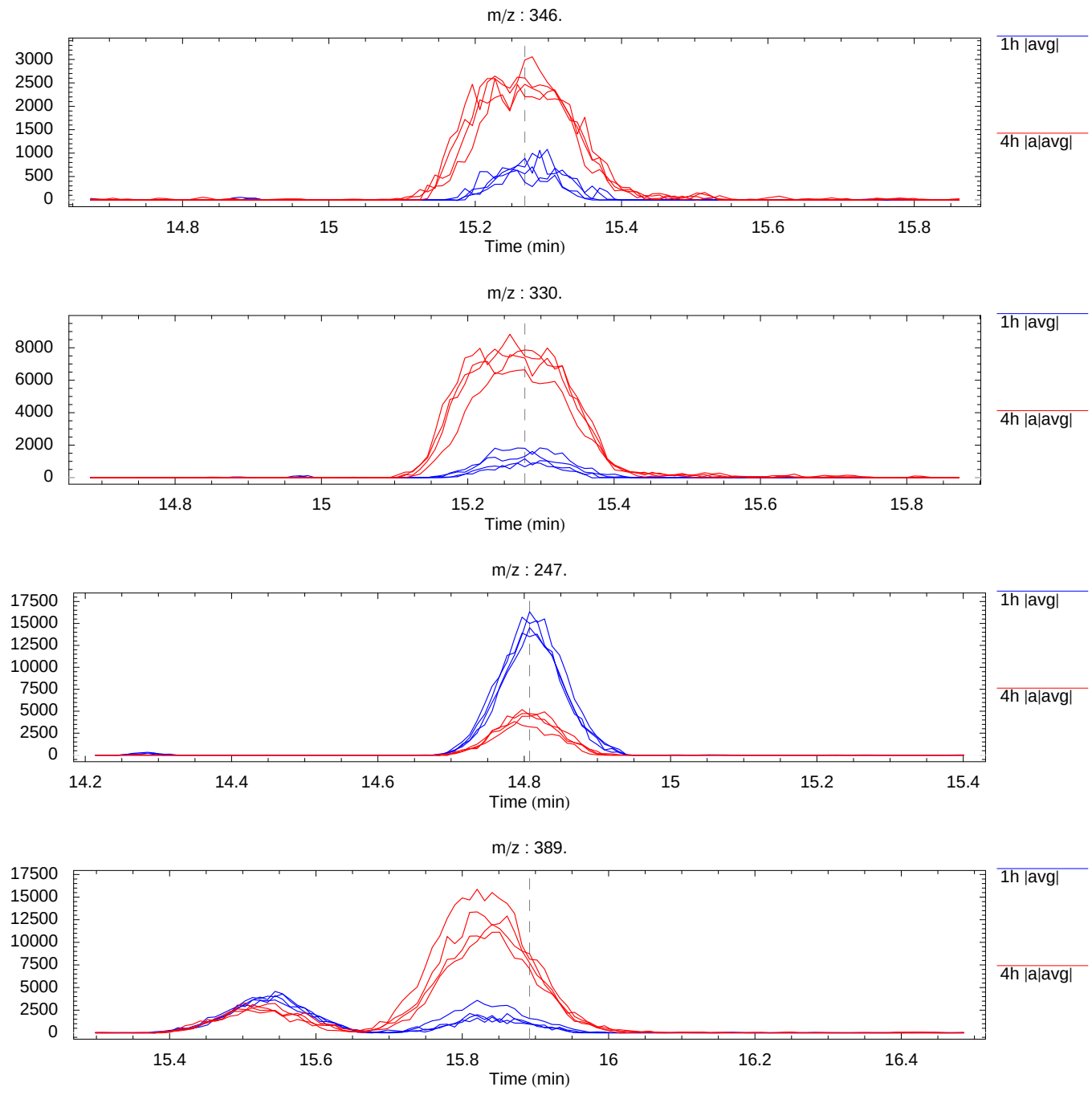

m/z : 175.
Indole−3−acetamide
30000
20000
10000
0
21.6
21.8
22
22.2
22.4
22.6
Time (min)
1h |avg|
4h |a|avg|

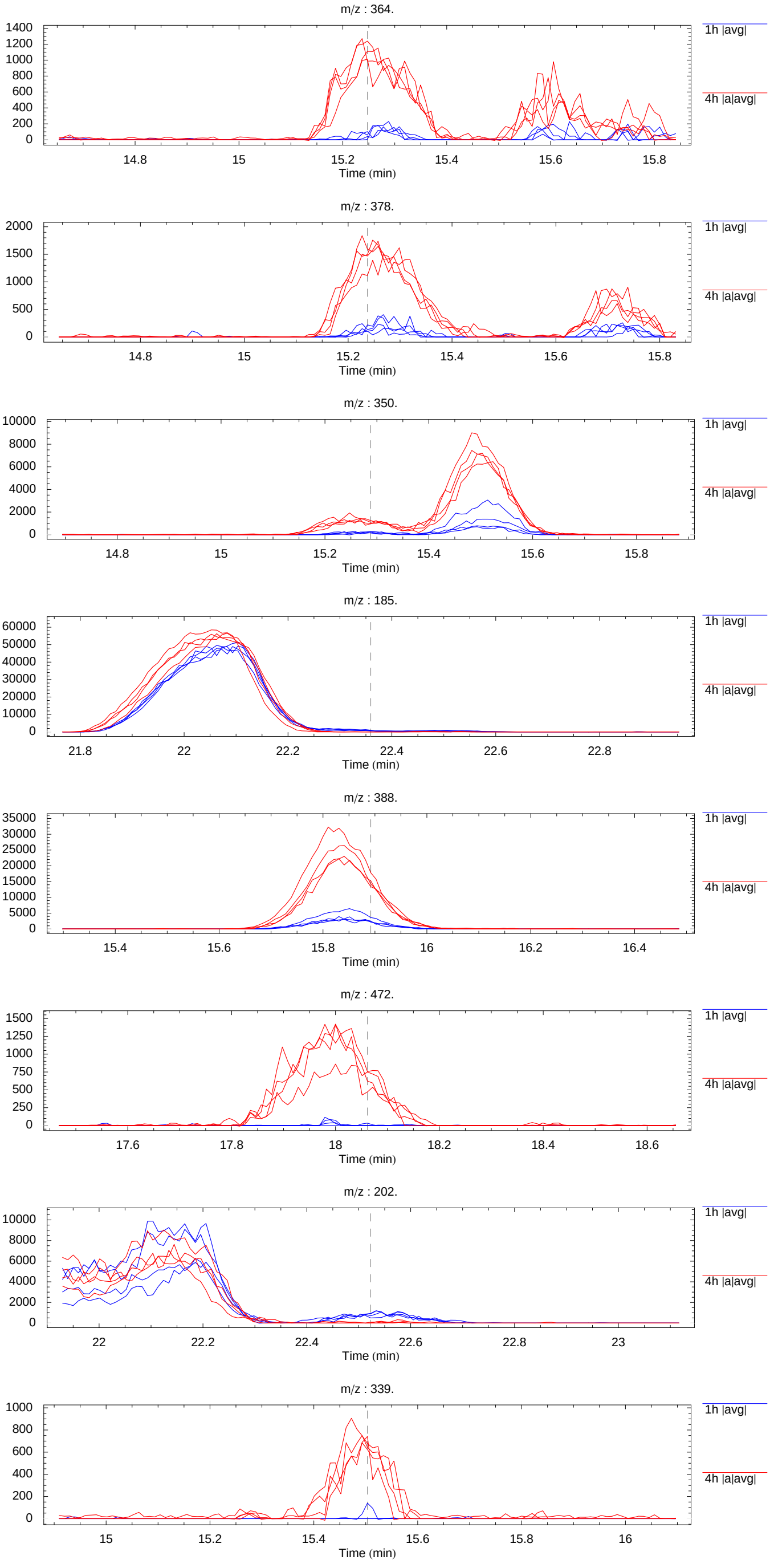

m/z : 271.
8000
6000
4000
2000
0
14.6
14.8
15
15.2
15.4
15.6
Time (min)
1h |avg|
4h |a|avg|



Different results datasets provide different ranking. For example, differences in small peaks usually do not score high in the absolute×relative difference result. On the other hand, presence of an outlying peak (in terms of signal intensities) in one dataset causes the drop of the difference in t−score ranking. Since different results have their strengths and weaknesses, it may prove beneficial to generate the lists of candidates based on multiple results. Below are electropherograms of candidate differences ranked according to the *t*−test result. Apart from the different ranking, smaller peaks achieve higher ranking in the list below when compared to ranking according to the absolute×relative result. However, the difference for L−Alanine (ranked second according to the absolute×relative result) is not among the first 50 candidates from the *t*−test result. Also note, that the vertical dashed line indicating the most significant difference often does not correspond to the peak top for the *t*−test result.

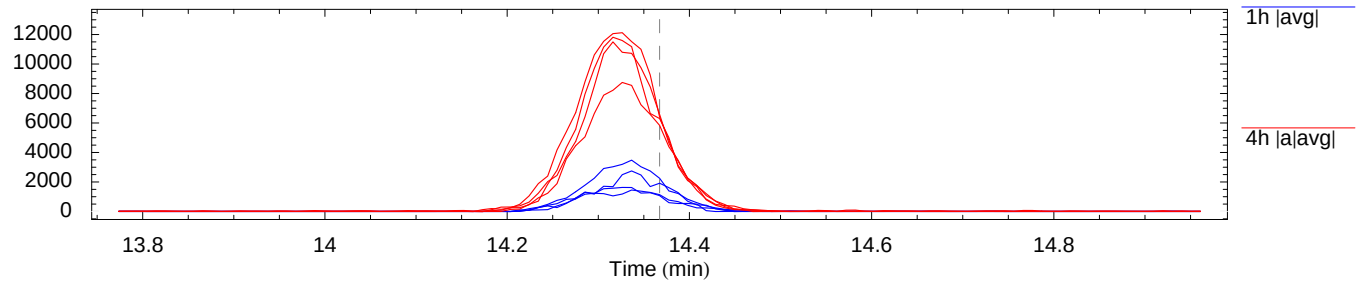
```
plotcolors = Transpose[{DAMPGenColors[2], GroupCounts /. rslt}];
DAMPPlotCandidates[NormalizedDatasets /. rslt, DAMPSmooth[TScores /. rslt], PlotCount → 50,
  PlotChromatogramOptions → {PlotOptions → {PlotStyle → Join @@ (Table[#[[1]], {#[[2]]}] & /@ plotcolors)},
    AnnotationTable → (AlignedAnnotationTables /. rslt)[[1]], LegendData → Transpose[{plotcolors[[All, 1]], (GroupNames /. rslt)}]}];
```



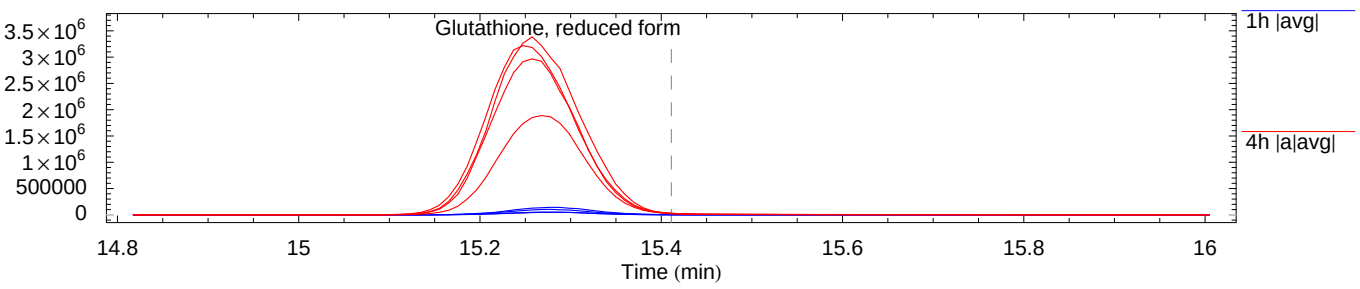

m/z : 364.
1400
1200
1000
800
600
400
200
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 378.
2000
1500
1000
500
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 350.
10000
8000
6000
4000
2000
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 185.
60000
50000
40000
30000
20000
10000
0
21.8
22
22.2
22.4
22.6
22.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 388.
35000
30000
25000
20000
15000
10000
5000
0
15.4
15.6
15.8
16
16.2
16.4
Time (min)
1h |avg|
4h |a|avg|
m/z : 472.
1500
1250
1000
750
500
250
0
17.6
17.8
18
18.2
18.4
18.6
Time (min)
1h |avg|
4h |a|avg|
m/z : 202.
10000
8000
6000
4000
2000
0
22
22.2
22.4
22.6
22.8
23
Time (min)
1h |avg|
4h |a|avg|
m/z : 339.
1000
800
600
400
200
0
15
15.2
15.4
15.6
15.8
16
Time (min)
1h |avg|
4h |a|avg|

m/z : 340.
1h |avg|
4h |a|avg|
Time (min)
m/z : 309.
1h |avg|
4h |a|avg|
Time (min)
m/z : 309.
1h |avg|
4h |a|avg|
Time (min)
m/z : 307.
1h |avg|
4h |a|avg|
Time (min)
m/z : 613.
Glutathione, oxidized form
1h |avg|
4h |a|avg|
Time (min)
m/z : 470.
1h |avg|
4h |a|avg|
Time (min)
m/z : 320.
1h |avg|
4h |a|avg|
Time (min)
m/z : 308.
3−Iodo−L−tyrosine
1h |avg|
4h |a|avg|
Time (min)

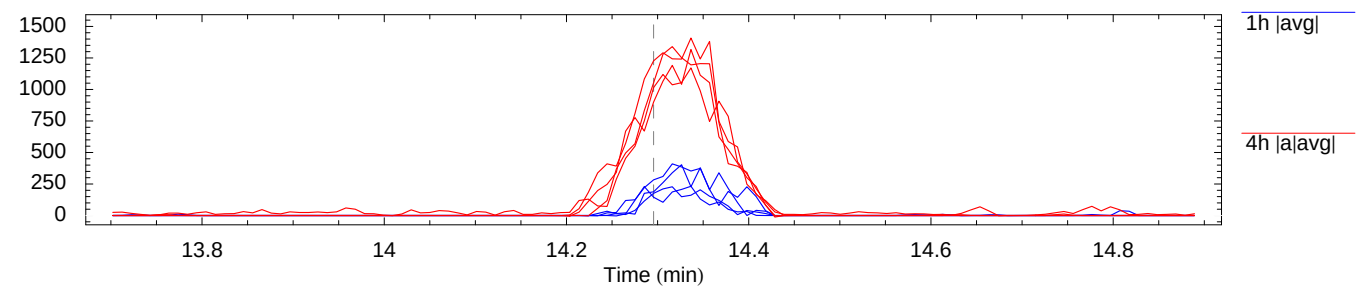
m/z : 614.



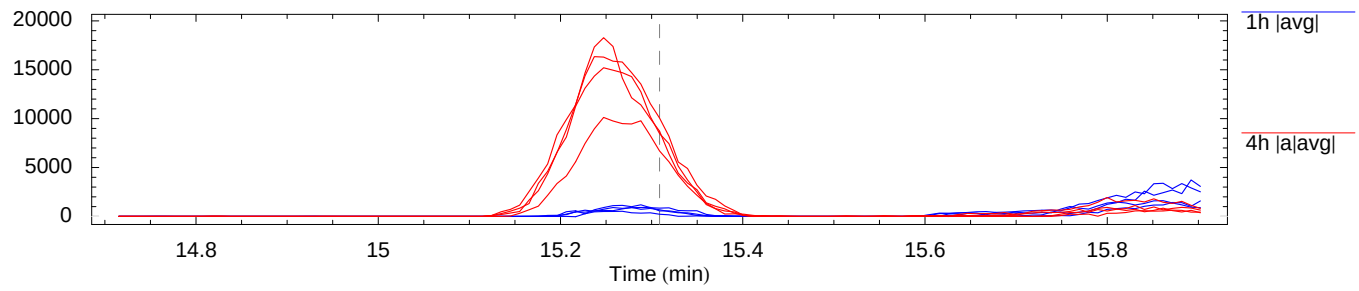
m/z : 308.



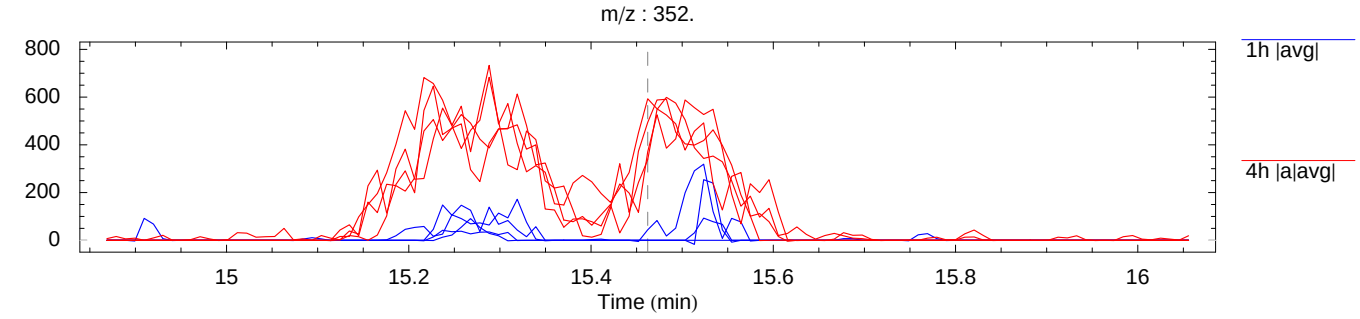
m/z : 616.



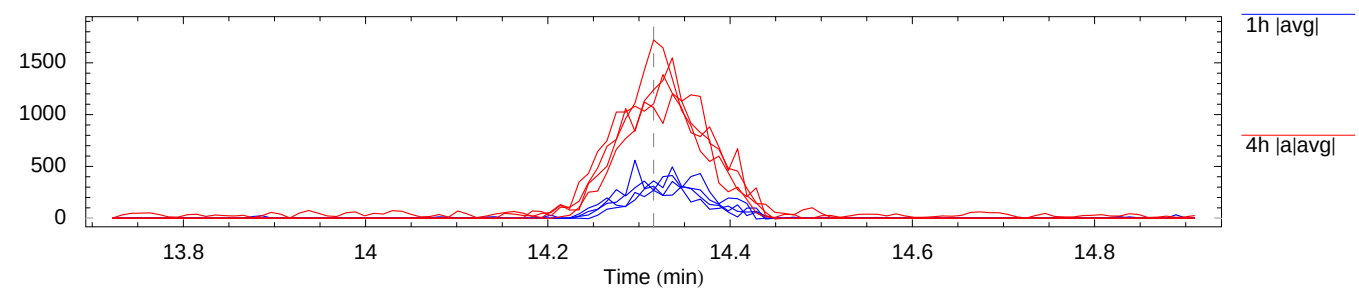
m/z : 311.



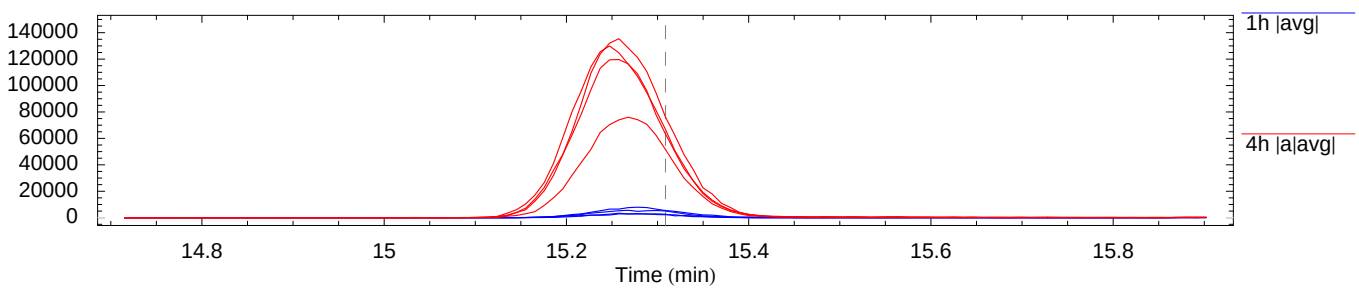
m/z : 352.



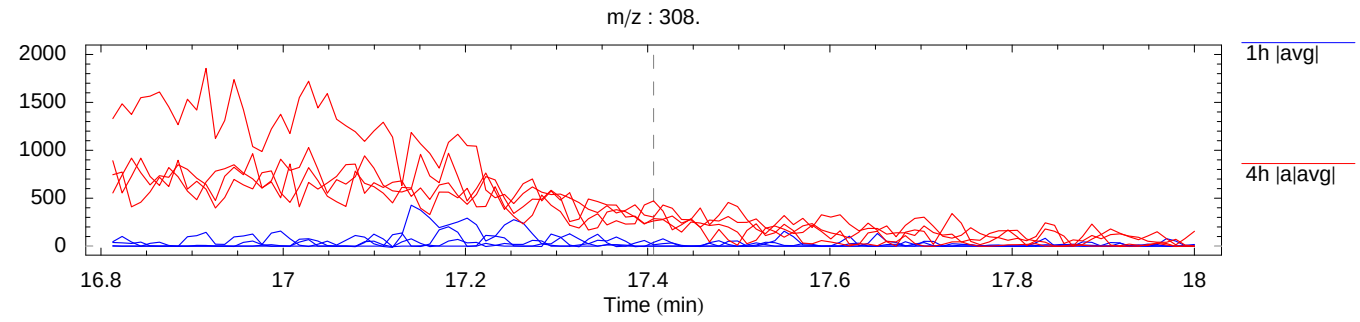
m/z : 484.

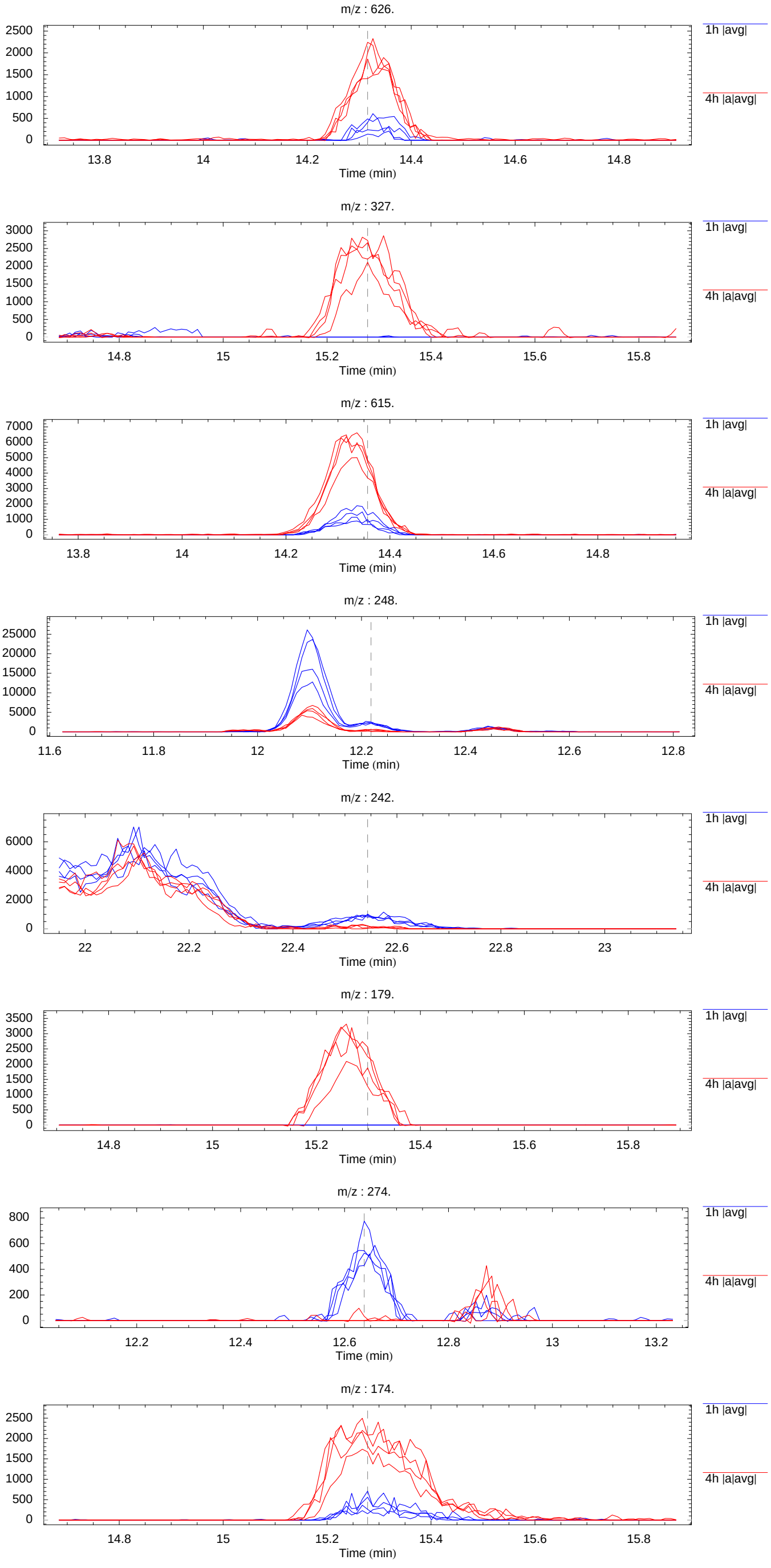


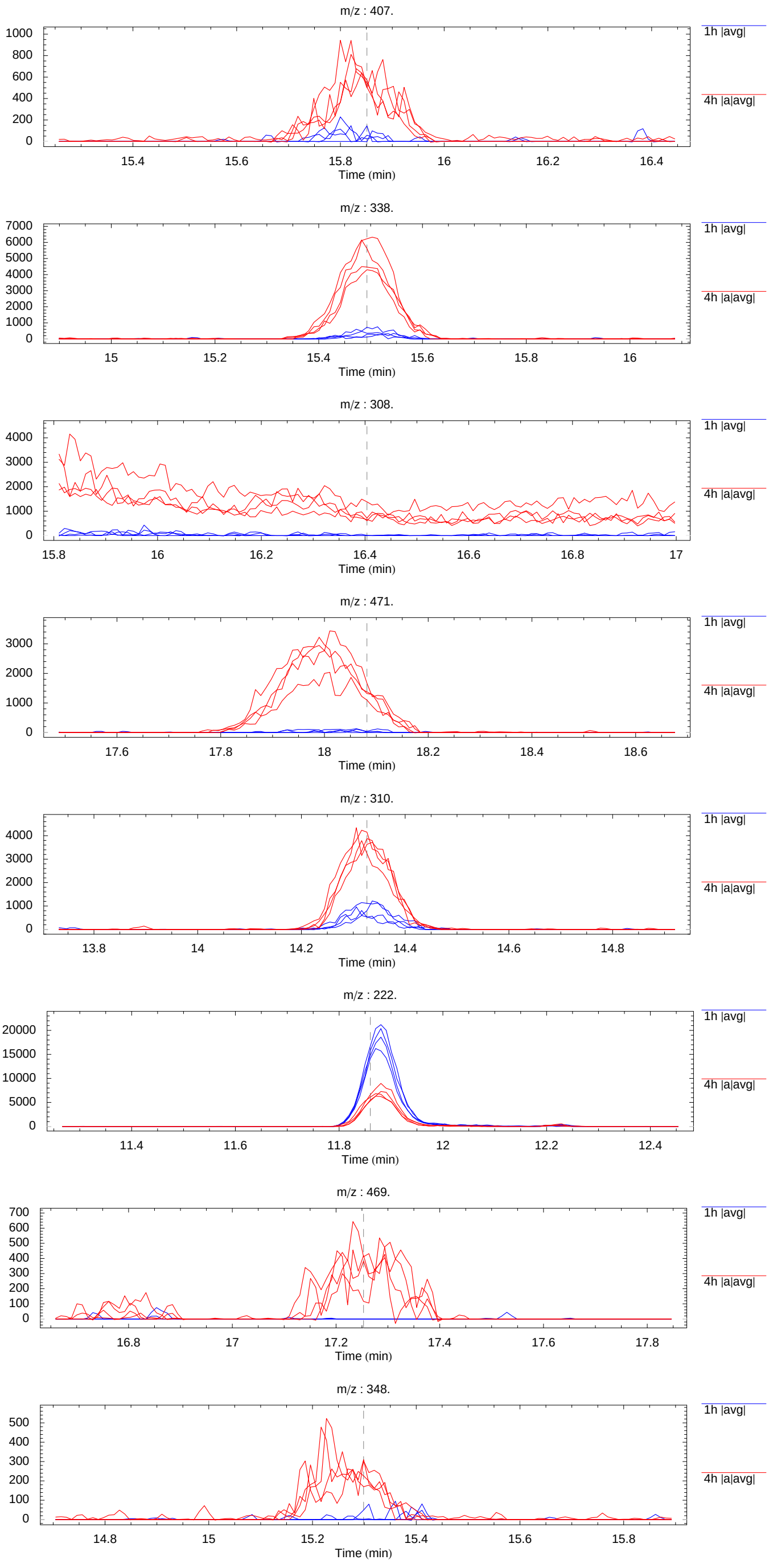
m/z : 310.

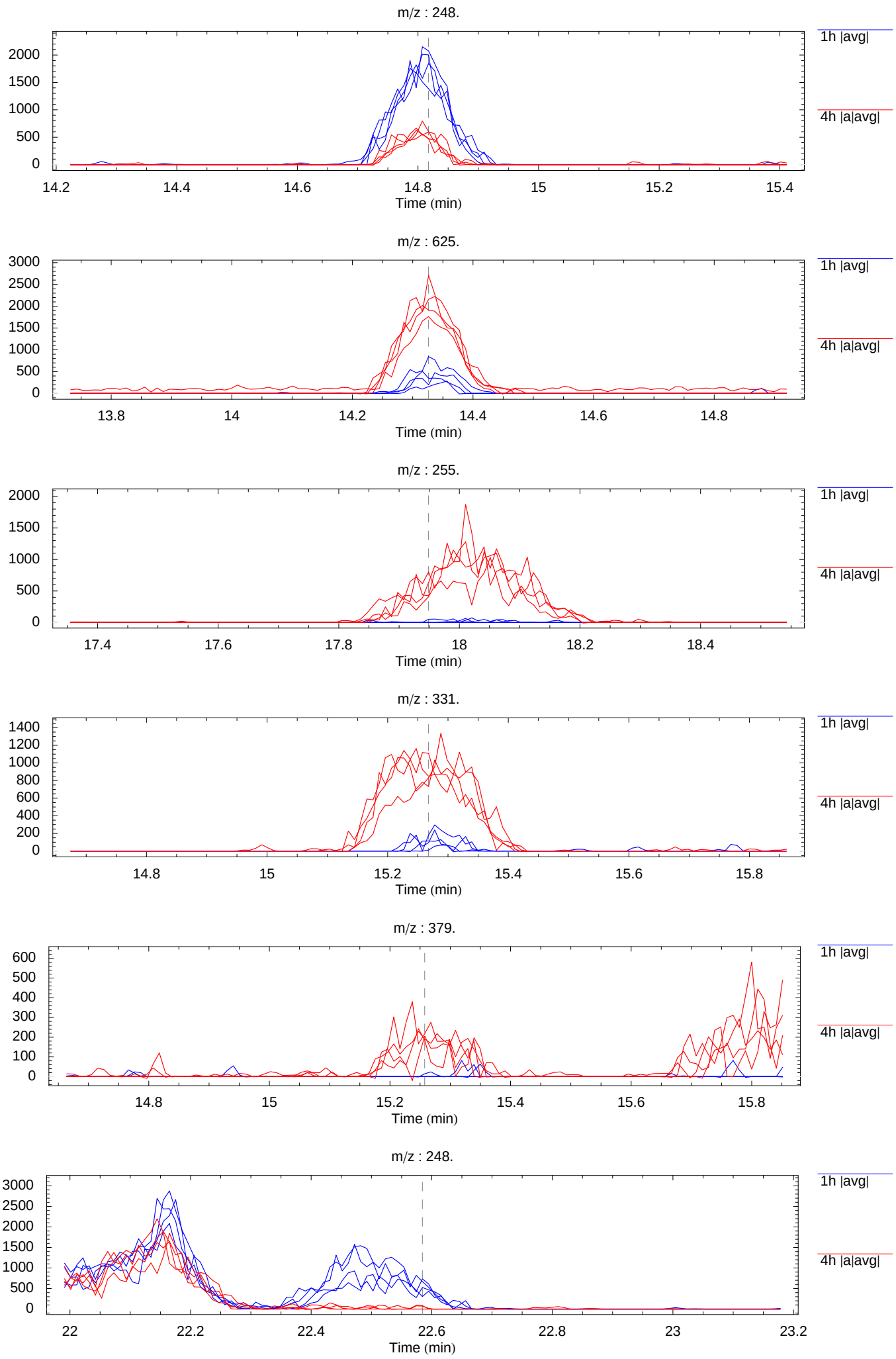


m/z : 308.




m/z : 626.
2500
2000
1500
1000
500
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 327.
3000
2500
2000
1500
1000
500
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 615.
7000
6000
5000
4000
3000
2000
1000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 248.
25000
20000
15000
10000
5000
0
11.6
11.8
12
12.2
12.4
12.6
12.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 242.
6000
4000
2000
0
22
22.2
22.4
22.6
22.8
23
Time (min)
1h |avg|
4h |a|avg|
m/z : 179.
3500
3000
2500
2000
1500
1000
500
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 274.
800
600
400
200
0
12.2
12.4
12.6
12.8
13
13.2
Time (min)
1h |avg|
4h |a|avg|
m/z : 174.
2500
2000
1500
1000
500
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|


m/z : 407.
1000
800
600
400
200
0
15.4
15.6
15.8
16
16.2
16.4
Time (min)
1h |avg|
4h |a|avg|
m/z : 338.
7000
6000
5000
4000
3000
2000
1000
0
15
15.2
15.4
15.6
15.8
16
Time (min)
1h |avg|
4h |a|avg|
m/z : 308.
4000
3000
2000
1000
0
15.8
16
16.2
16.4
16.6
16.8
17
Time (min)
1h |avg|
4h |a|avg|
m/z : 471.
3000
2000
1000
0
17.6
17.8
18
18.2
18.4
18.6
Time (min)
1h |avg|
4h |a|avg|
m/z : 310.
4000
3000
2000
1000
0
13.8
14
14.2
14.4
14.6
14.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 222.
20000
15000
10000
5000
0
11.4
11.6
11.8
12
12.2
12.4
Time (min)
1h |avg|
4h |a|avg|
m/z : 469.
700
600
500
400
300
200
100
0
16.8
17
17.2
17.4
17.6
17.8
Time (min)
1h |avg|
4h |a|avg|
m/z : 348.
500
400
300
200
100
0
14.8
15
15.2
15.4
15.6
15.8
Time (min)
1h |avg|
4h |a|avg|


m/z : 248.
1h |avg|
4h |a|avg|
Time (min)
m/z : 625.
1h |avg|
4h |a|avg|
Time (min)
m/z : 255.
1h |avg|
4h |a|avg|
Time (min)
m/z : 331.
1h |avg|
4h |a|avg|
Time (min)
m/z : 379.
1h |avg|
4h |a|avg|
Time (min)
m/z : 248.
1h |avg|
4h |a|avg|
Time (min)