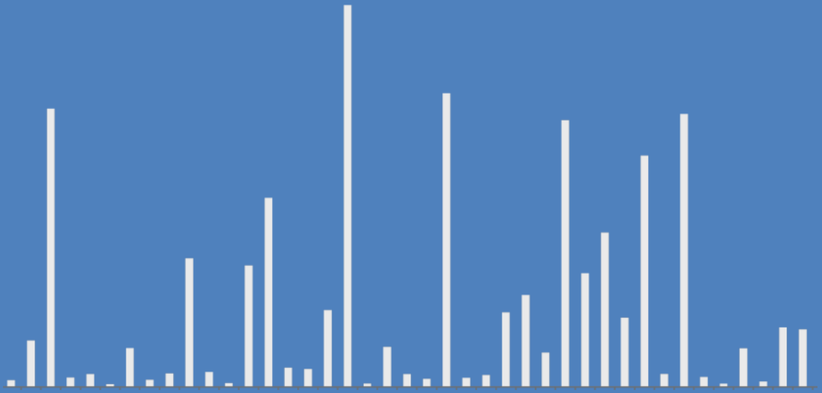
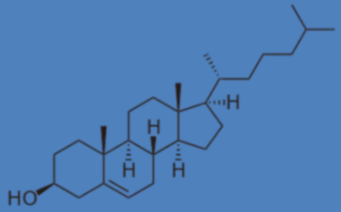
Building the NIST Tandem Mass Spectral Library 2014





Xiaoyu (Sara) Yang
Mass Spectrometry Data Center

Biomolecular Measurement Division Seminar
September 9, 2014

Outline

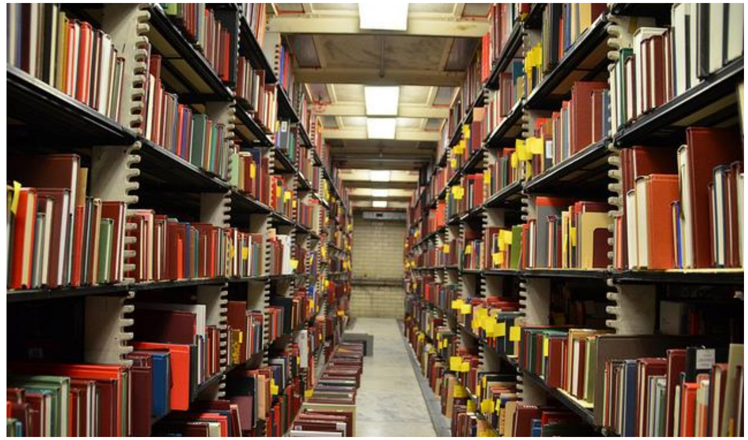
- **Methods of building the NIST tandem mass spectral library**

- **Quality Control**
  - **Peak annotation**
  - **Noise Removal**
  - **Chemical information consistency**

- **Major types of mass spectra**

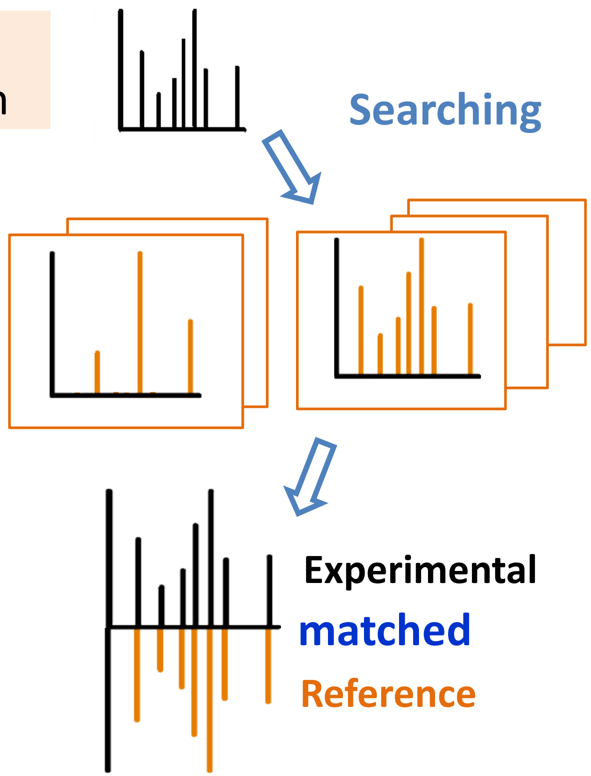
- **Major types of compounds**

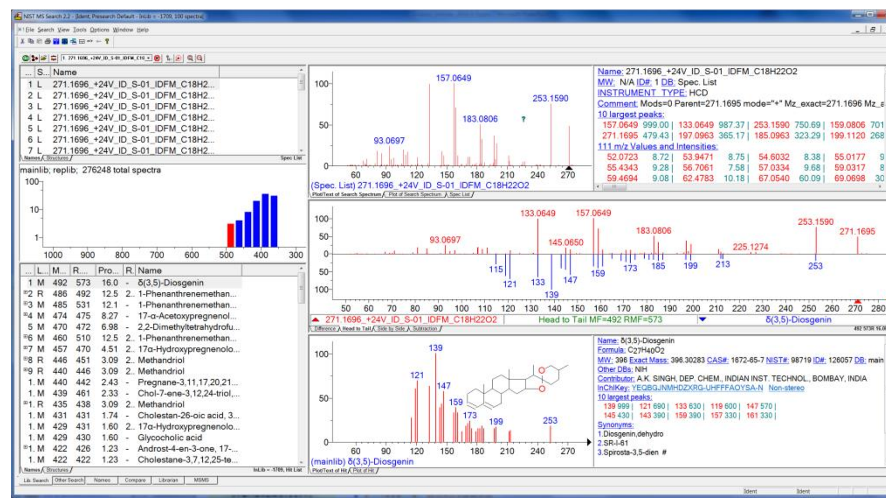
Mass Spectral Library Searching



Experimental mass spectrum

Reference Spectra in MS Library





MS Search

NIST Tandem Mass Spectral Library

Rationale and Objectives:

- Chemical identification through ***Electrospray ionization (ESI) tandem*** mass spectrometry (MS/MS) is becoming a routine technique in metabolomics, proteomics and other fields.
- The identification can be aided by matching the acquired tandem mass spectra against reference library spectra.
- We are developing a **comprehensive** library of **high quality** **reference** ESI tandem mass spectra for the identification of compounds through the ion fragmentations.

Goals:

- Develop a tandem mass spectral library of all biologically relevant metabolite ions.
- Provide the library in a form that is easily searchable using software tools.

Steps of Building the NIST Tandem Mass Spectral Library

Authentic samples

metabolites, drugs, peptides | lipids, pesticides, surfactants, glycans, sugars

↓

LC/MS/MS

Ion Trap (LTQ, IT/FTMS) | Collision Cell (HCD, QQQ, QTOF)

↓

First clustering

Cluster MS^2 spectra using precursor m/z

count-based clustering algorithm | cluster spectra of the similar precursor m/z values

↓

Second clustering

Create consensus spectra of MS^2, MS^3 and MS^4

adjusted dot product-based clustering algorithm | cluster spectra of the similar fragmentations

↓

Precursor type identification

Chemical structure, formula, name, synonyms, and CAS are consistent

- precursor purity
- mass accuracy

- peak annotation
- noise removal

Manual inspection

↓

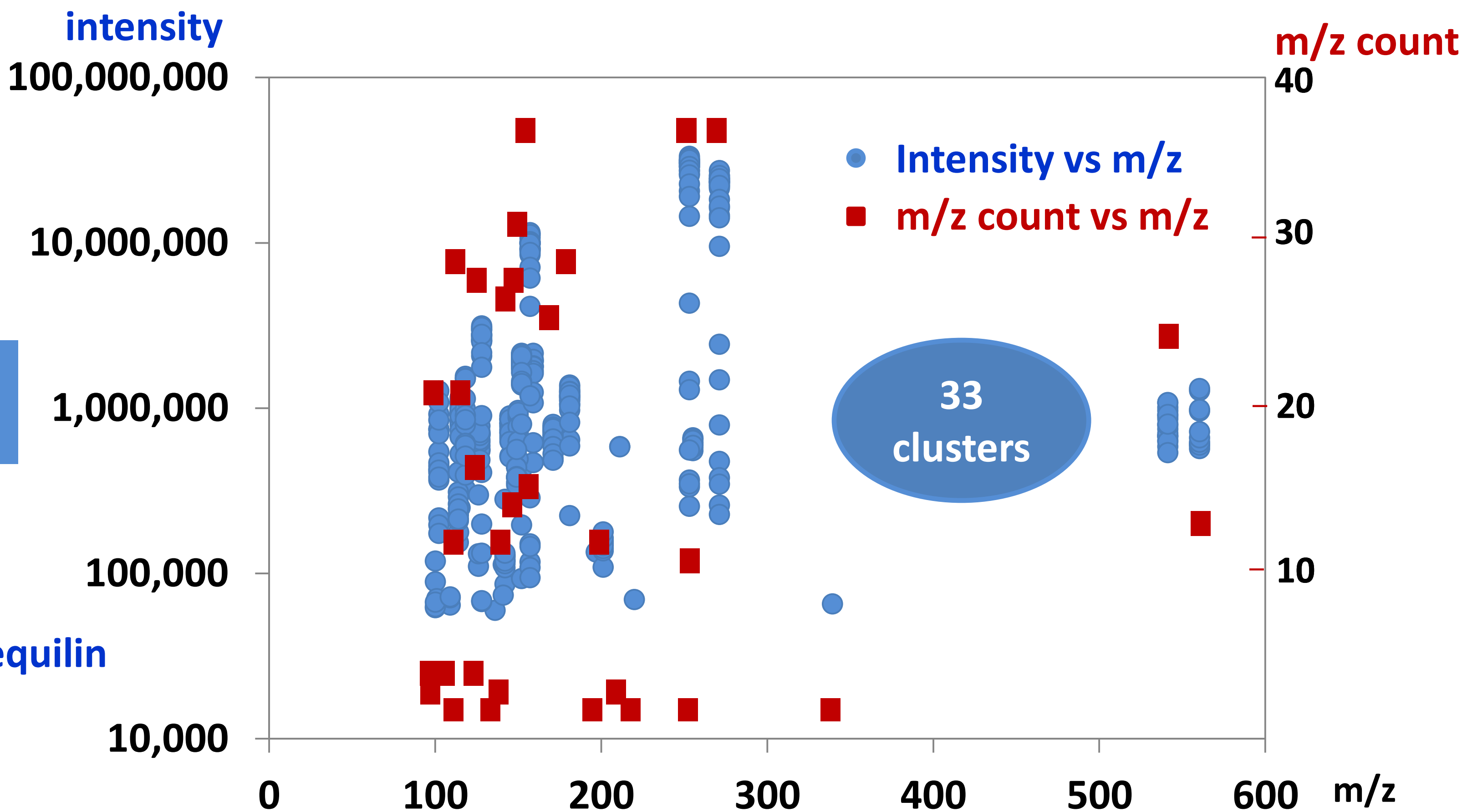
MSMS library with multiple precursor types

Count-based Algorithm for Clustering Precursors

Steps:

1. Count the number of precursors (m/z count) within 0.1 m/z;
2. Sort the precursors by the m/z count in descending order;
3. Group similar precursors into the same cluster by using the precursor with the highest m/z count as the cluster center;
4. Repeat step 3 until all the precursors are clustered.

437 spectra



β-Dihydroequilin
[M+H]+
HCD 10eV

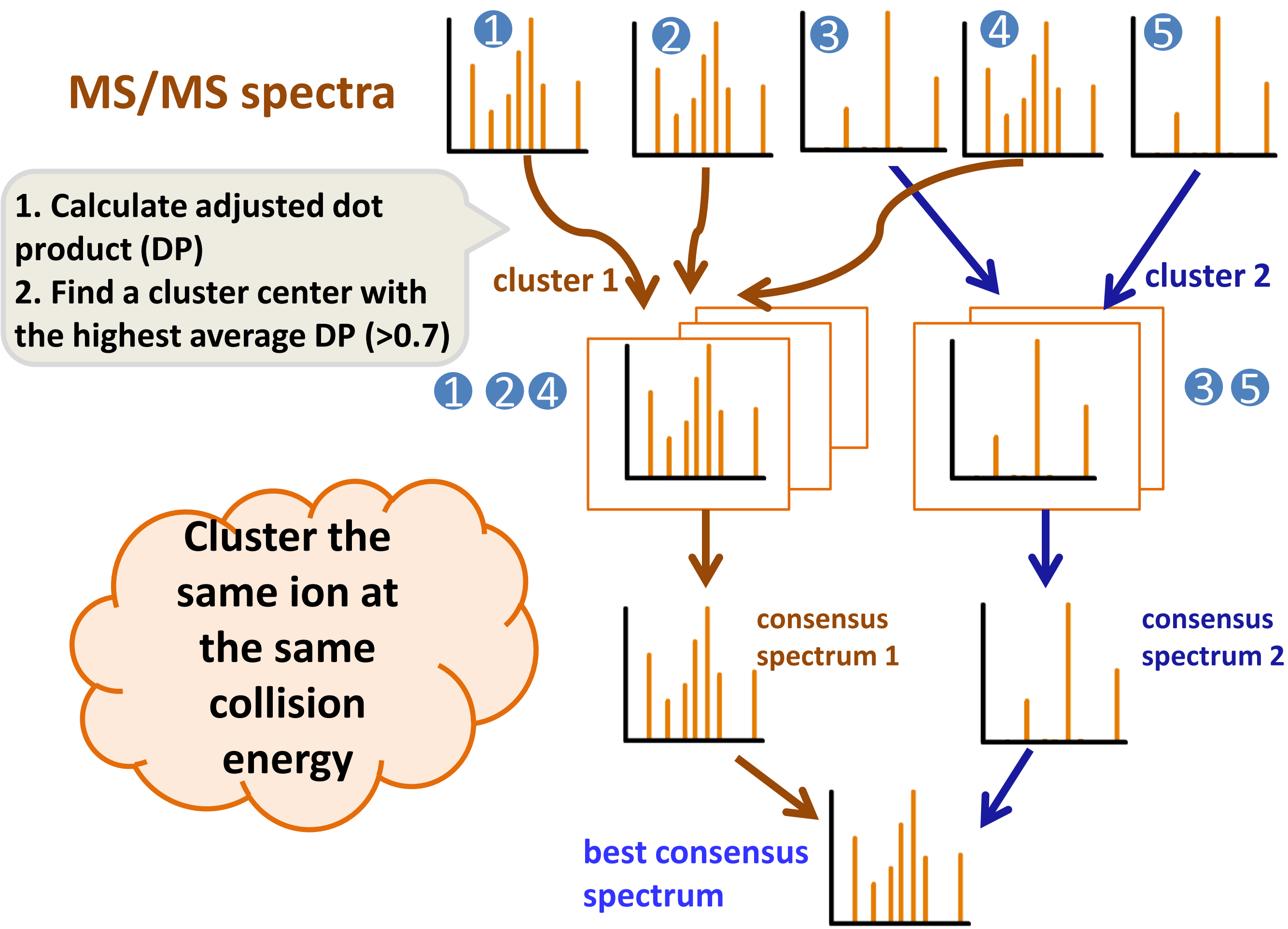
Clustering Algorithm for Generating Consensus Spectra

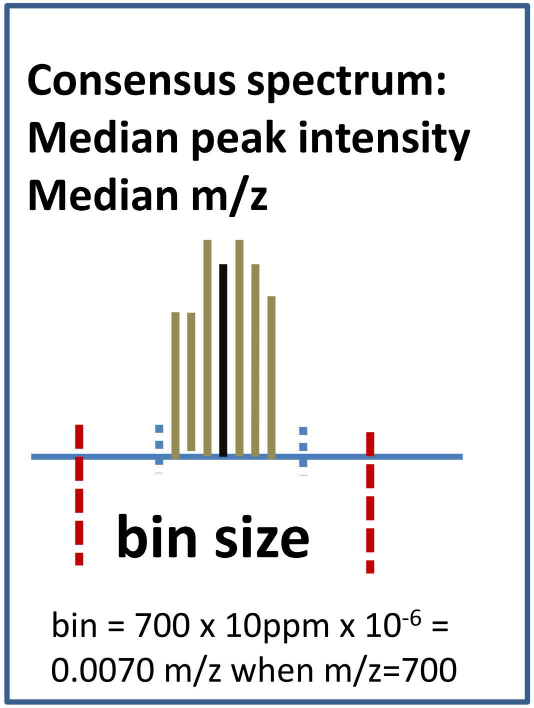
Steps:
1. Group similar spectra into the same cluster;
2. Generate one consensus spectrum from each cluster;
3. Pick the best consensus spectrum for the library.

Dot Product (DP) =

$$\sum \sqrt{I_1 * I_2} \;/\; \sqrt{\sum I_1 * \sum I_2}$$

I – peak intensity





Using Consensus Spectrum in the Library

- **Eliminated low quality spectra by spectral clustering.**
- **Improved the spectrum quality by using the median of the m/z and intensity values.**
- **Realistically represented the characteristic fragmentations.**

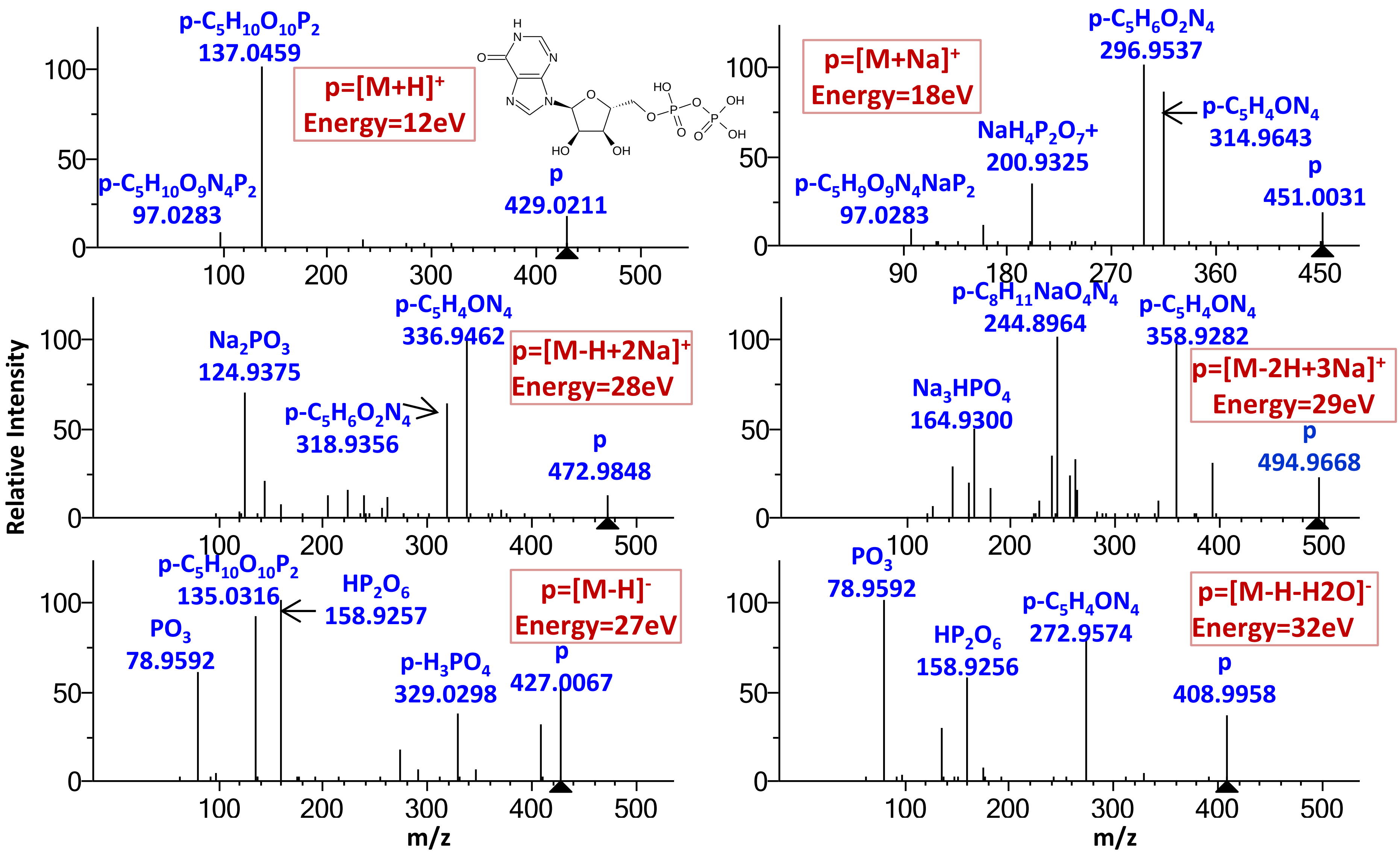
Same compound
Same precursor type
Same instrument
Same energy (cone, collision)
Same mode (+/-)
Same spectrum type (MS^2, MS^3, MS^4)

10-20 spectra / consensus spectrum
10-20 energy levels

What Precursor Types are in the NIST MSMS Library?

Compound Type	ESI Product	Ionic Species
Neutral Molecule	Positive Ions (129,662; **84%**)	$[M+H]^+$, $[M+2H]^{2+}$, $[2M+H]^+$, $[M+H-H2O]^+$, $[M+H-NH_3]^+$, $[M+H-OH]^+$, $[M+H+H_2O]^+$, $[M+NH_4]^+$, $[M+Na]^+$, $[M-H+2Na]^+$, $[M-2H+3Na]^+$, $[M+K]^+$, $[M-H+2K]^+$, $[M-2H+3K]^+$, $[M+Li]^+$, $[M-H+2Li]^+$, $[M-2H+3Li]^+$
	Negative Ions (23,638; **15%**)	$[M-H]^-$, $[M-2H]^{2-}$, $[2M-H]^-$, $[M-H-H_2O]^-$, $[M-H-NH_3]^-$, $[M-H+H_2O]^-$, $[M-H+NH_3]^-$
Organic Salt Cations	Positive Ions (1,751; **1%**)	$[Cat]^+$, $[Cat+H]^{2+}$, $[Cat-H_2O]^+$, $[Cat-NH_3]^+$, $[Cat+H_2O]^+$
	Negative Ions (40; **<0.1%**)	$[Cat-2H]^-$, $[Cat-2H-H_2O]^-$, $[Cat-2H-NH_3]^-$, $[Cat-2H+H_2O]^-$, $[Cat-2H+NH_3]^-$

Multiple Precursor Ions for More Flexible Identification



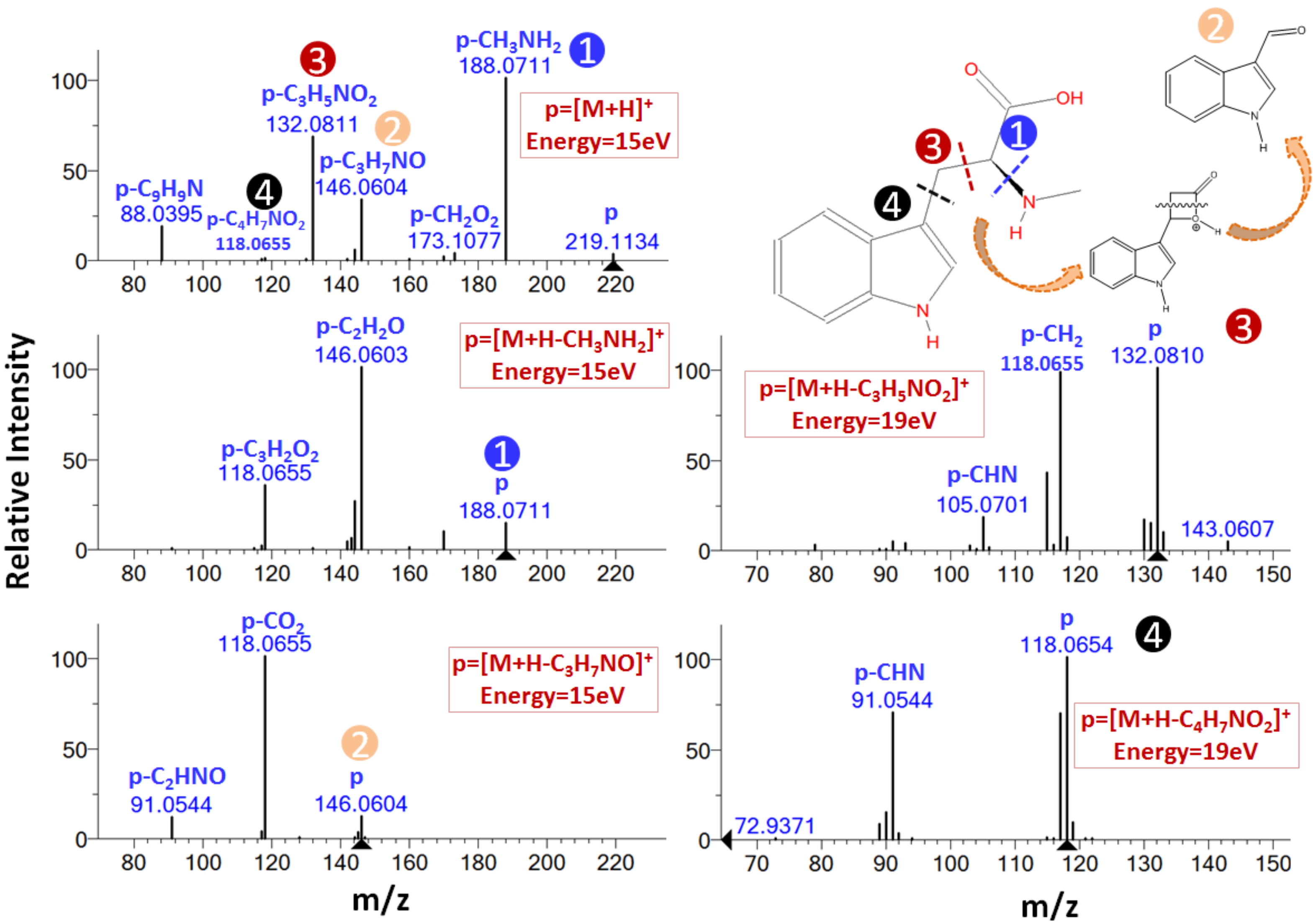
inosine 5'-diphosphate acquired on Orbitrap HCD

What Precursor Types are in the NIST MSMS Library?

Structure dependent losses:

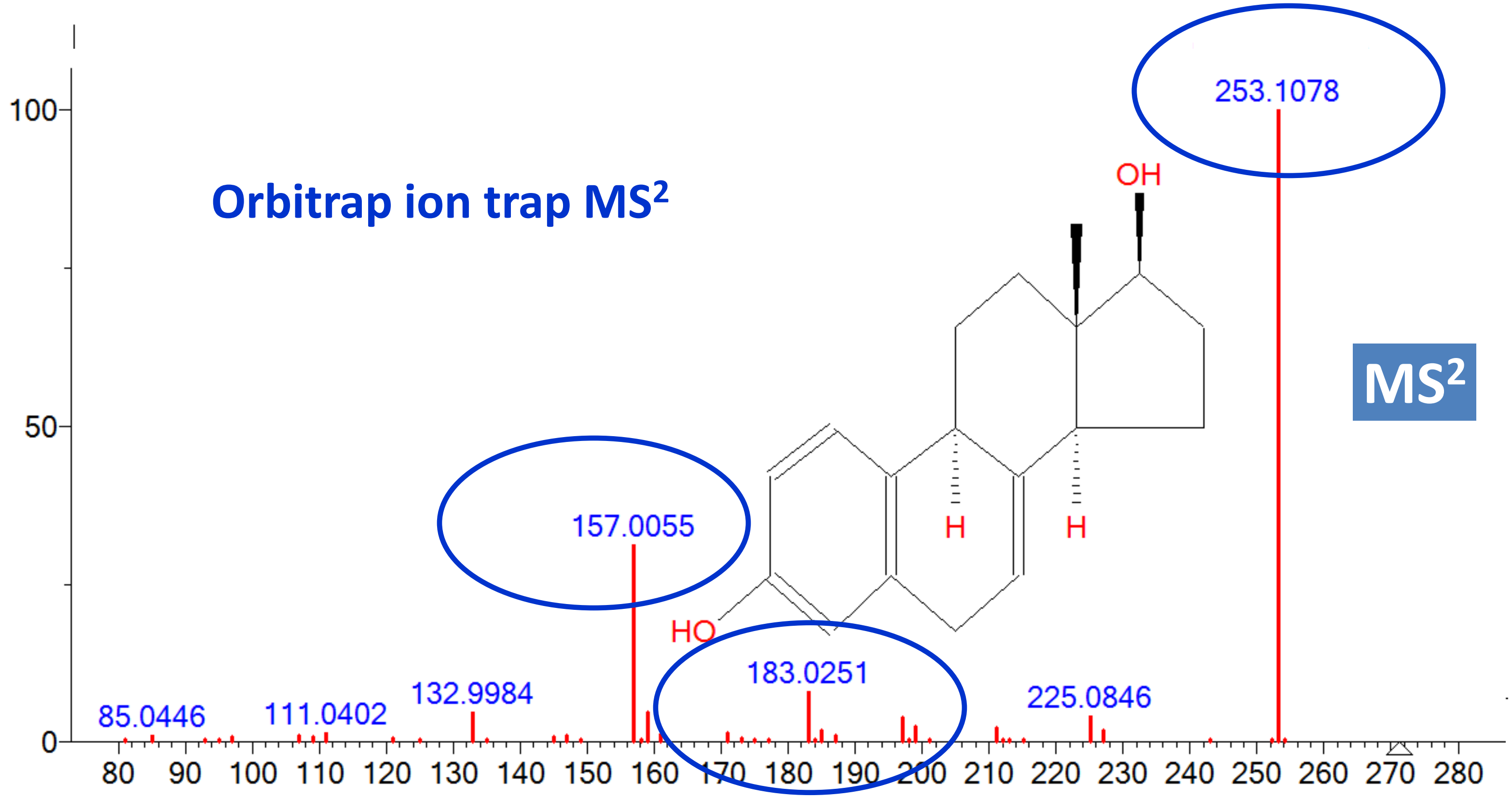
$2H_2O$, $3H_2O$, NH_3+H_2O,
H_2S, HCl, H_3PO_4, HCN, H_2,
CO, CO_2, HCOOH,
CH_4, CH_3, CH_3OH, CH_3SH,
C_2H_5OH,
…

In Source Fragmentation Confirms Metabolite Identification



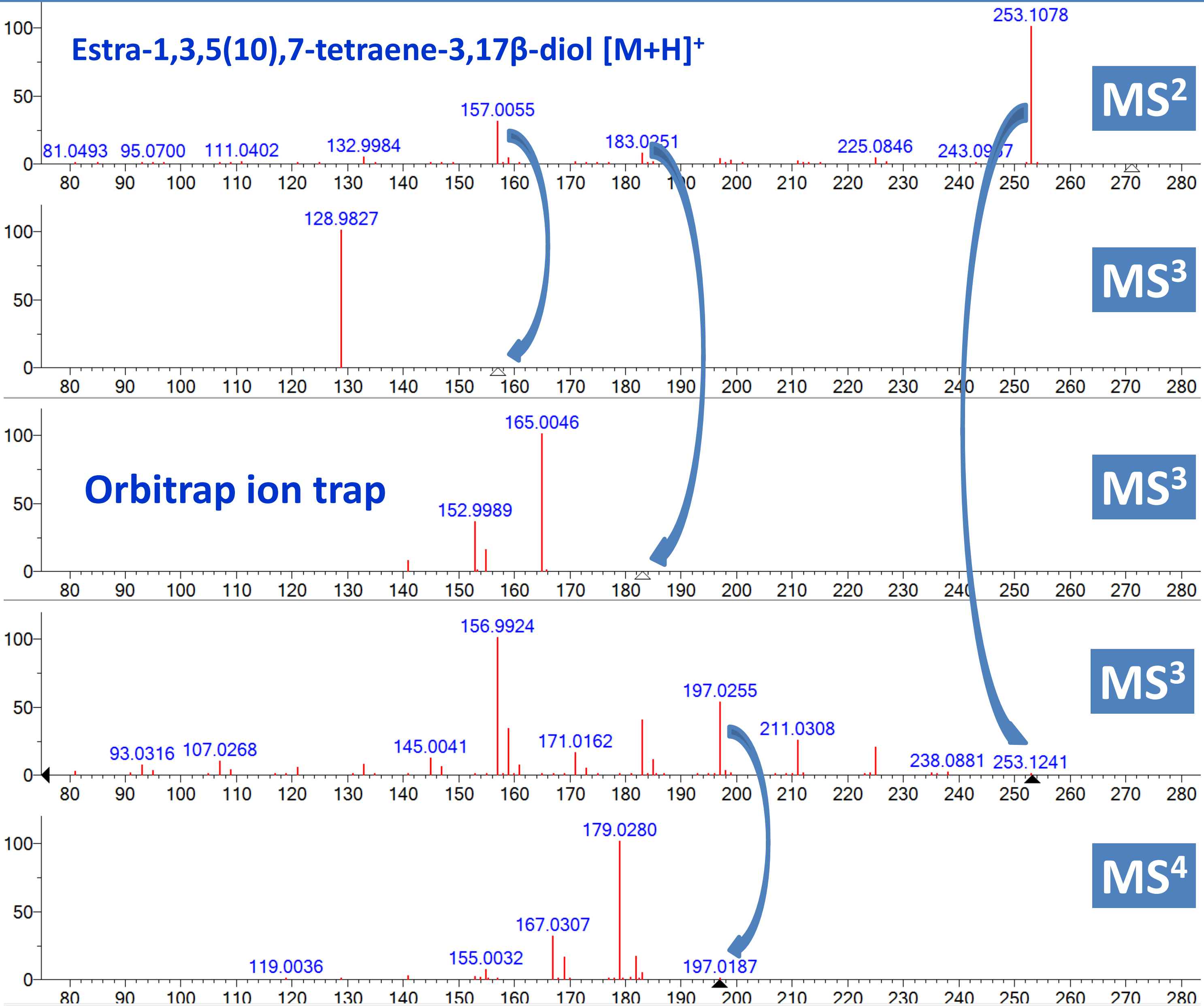
MS2 spectra acquired on Orbitrap HCD

MS^n



Estra-1,3,5(10),7-tetraene-3,17β-diol $[M+H]^+$

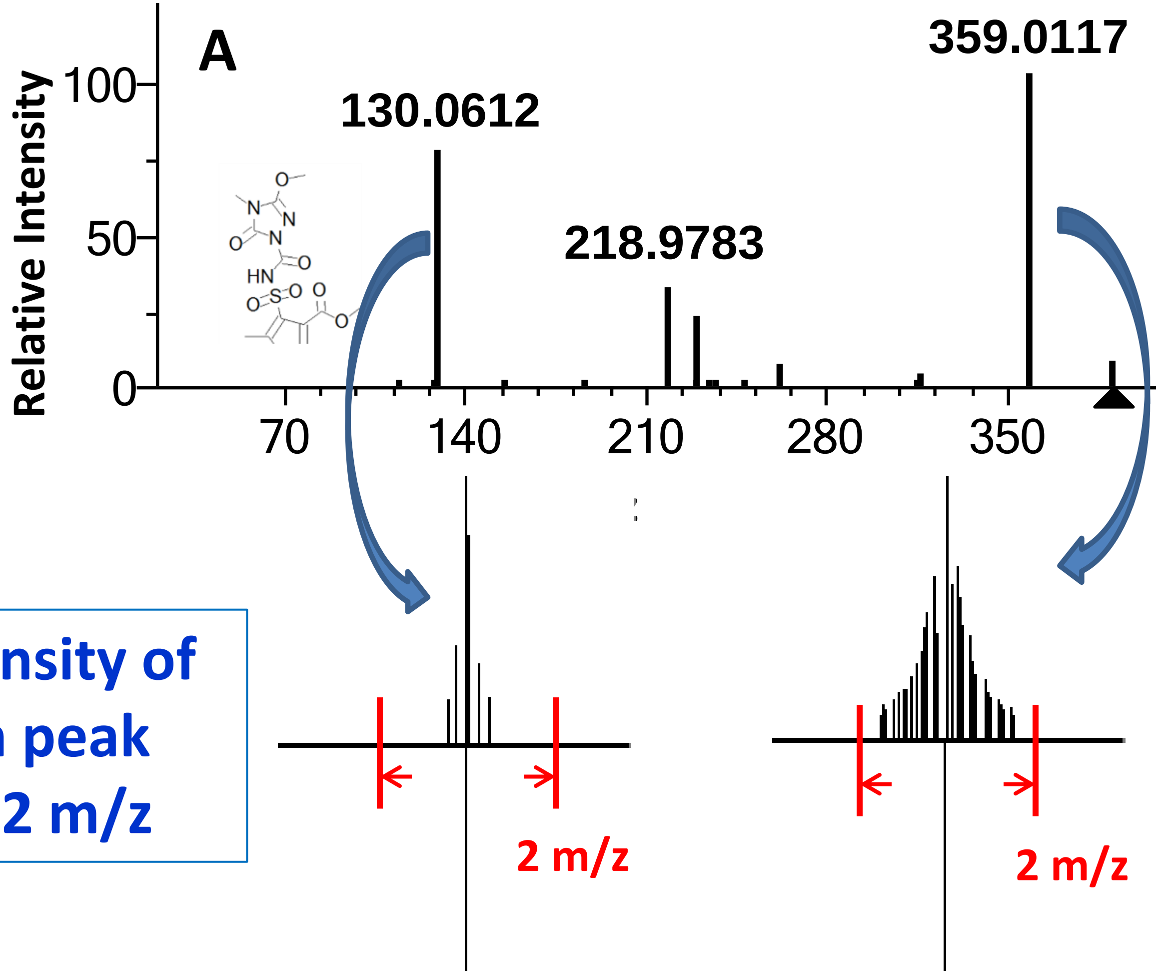
MS^n



Estra-1,3,5(10),7-tetraene-3,17β-diol [M+H]+
Orbitrap ion trap
MS2
MS3
MS3
MS3
MS4
253.1078
157.0055
81.0493
95.0700
111.0402
132.9984
225.0846
128.9827
165.0046
152.9989
156.9924
197.0255
211.0308
93.0316
107.0268
145.0041
171.0162
238.0881
253.1241
179.0280
167.0307
155.0032
119.0036
197.0187

Noise Removal

1. **Satellite peaks in Orbitrap HCD spectra: due to the Fourier transform ringing artifacts**



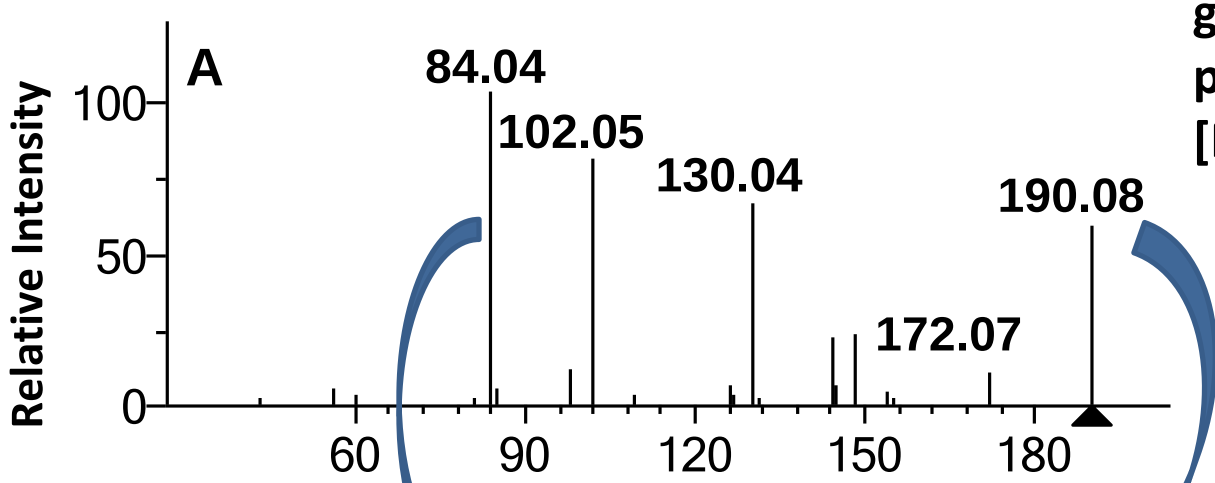
thiencarbazone -methyl $[M+H]^+$

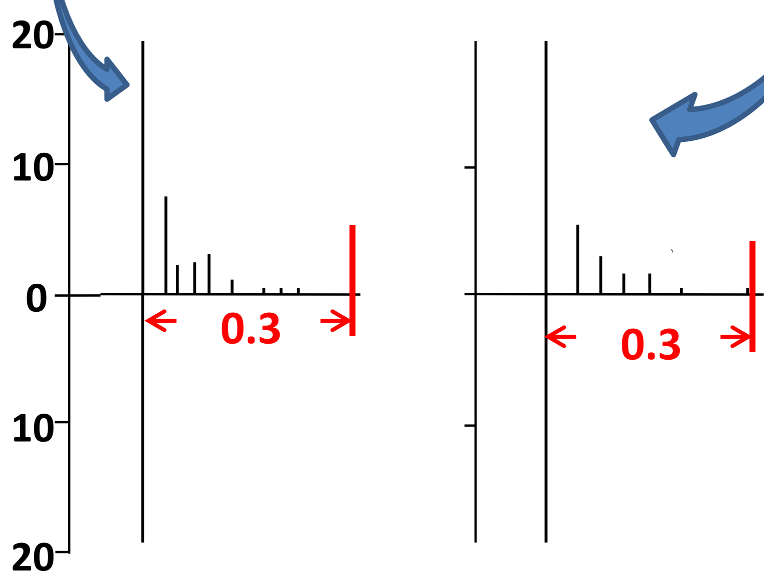
<1% intensity of the main peak within ± 2 m/z

Noise Removal

2. **Tailing peaks in QTOF spectra: due to the imperfect centroiding**

1-Eicosatrienoyl-sn-glycero-3-phosphoethanolamine $[M+H]^+$

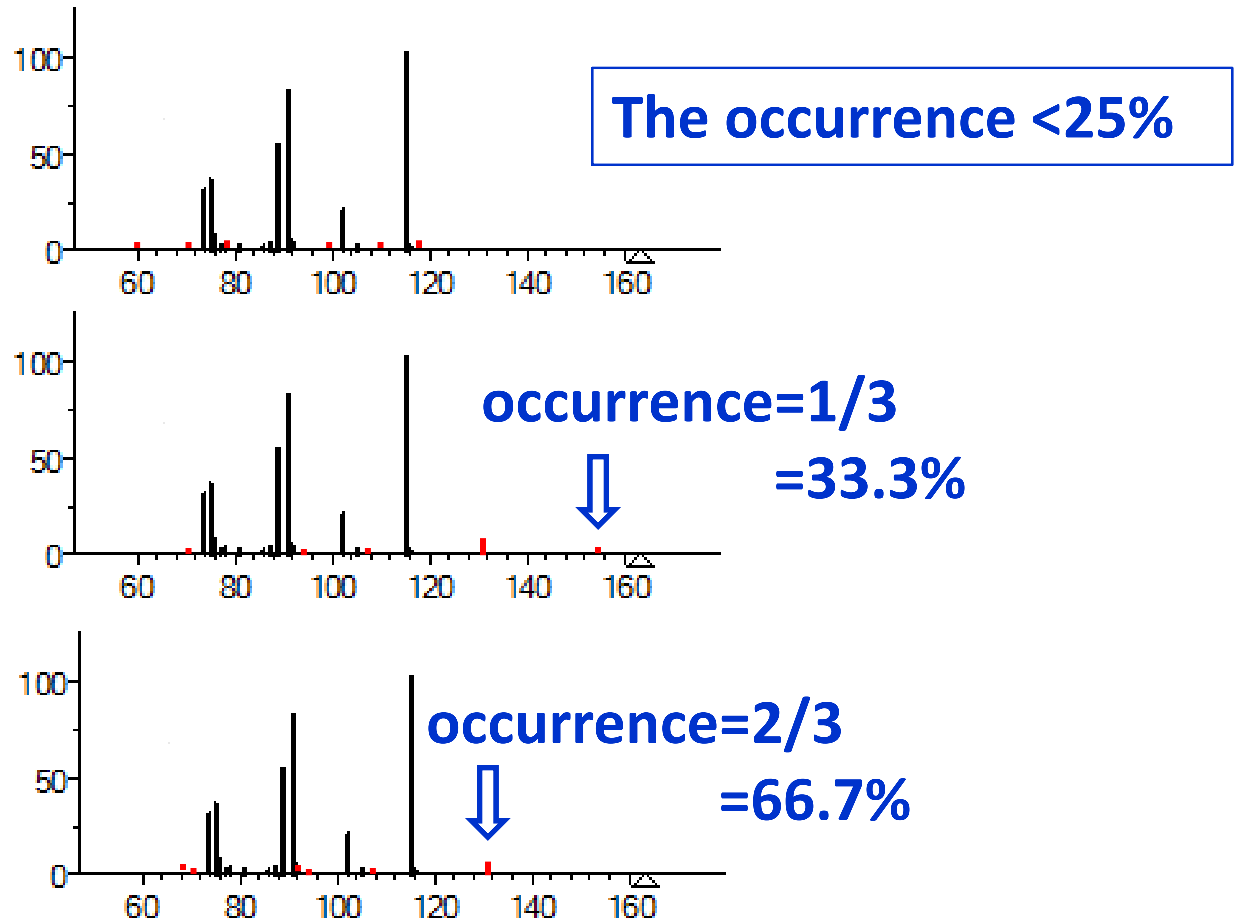




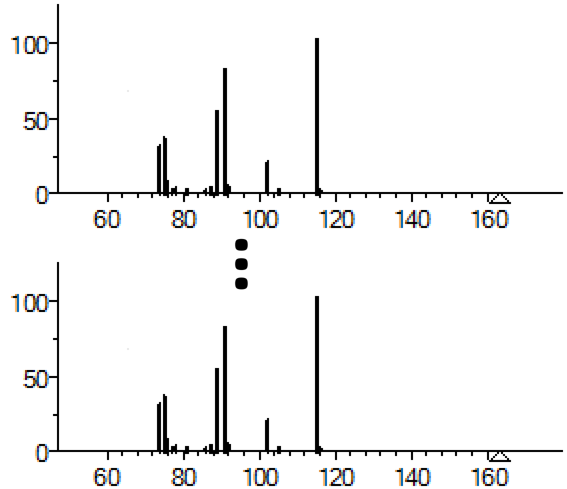
<20% intensity of the main peak after 0.3 m/z

Noise Removal

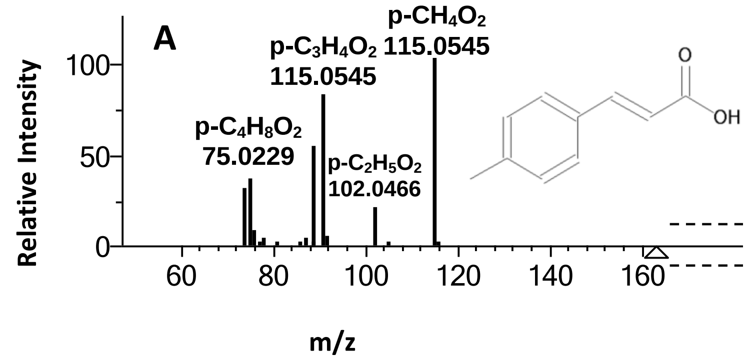
3. Random noise peaks: in all mass spectra due to unstable instrument, impurity…

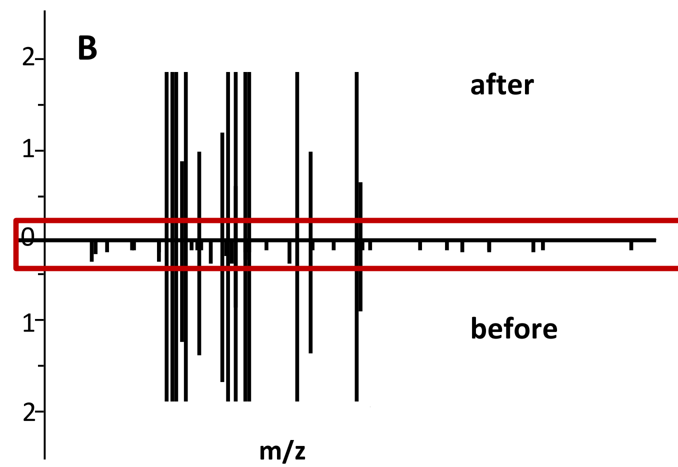


Noise Removal – an Example of Voting Algorithm



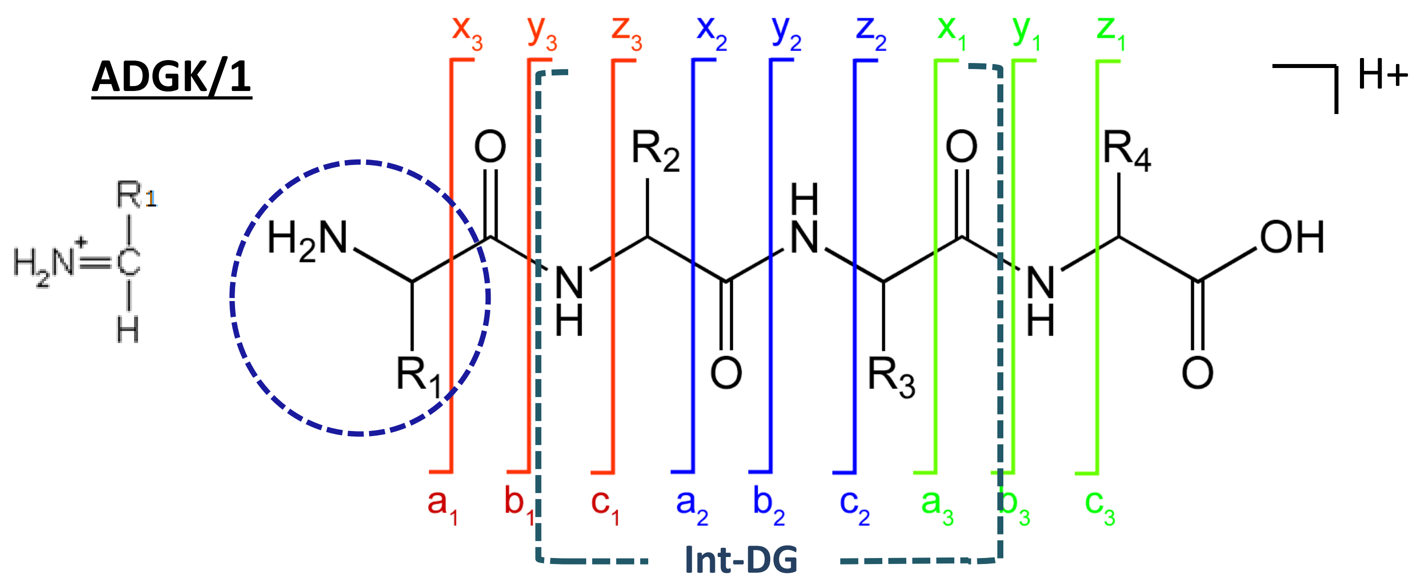
m/z	intensity	occurrence	m/z	intensity	occurrence
57.6951	3.10	2/12	91.0543	803.20	12/12
58.2664	2.30	2/12	91.1262	2.10	1/12
60.8356	2.10	1/12	91.6671	2.10	1/12
66.5419	2.00	1/12	91.9893	43.16	12/12
66.8604	2.00	1/12	95.6621	2.00	1/12
72.1006	2.10	1/12	100.4113	3.40	2/12
72.1606	3.10	2/12	102.0466	203.30	12/12
74.0151	298.60	12/12	105.0454	12.89	6/12
75.0229	355.04	12/12	105.4044	2.00	1/12
76.0307	77.12	12/12	110.2882	2.00	1/12
77.0386	11.69	5/12	115.0430	4.20	2/12
78.0464	34.27	10/12	115.0545	999.00	12/12
79.2907	2.00	1/12	115.9932	2.00	1/12
80.5368	2.00	1/12	116.0622	8.49	4/12
81.0335	13.09	6/12	116.2749	2.00	1/12
81.3809	2.00	1/12	118.1334	2.00	1/12
83.4835	3.30	2/12	129.0451	2.00	1/12
86.0151	15.98	7/12	134.7912	2.00	1/12





4-methylcinnamic acid $[M+H]^+$ HCD

Peak Annotation – Peptide (for low and high resolution MS/MS spectra)



- **p, y, b, a, internal fragments, neutral losses(-H_2O, -NH_3, -CO)**
- **y+10(CO-H_2O), a2-45($CONH_3$)**
- **52 immonium ions (e.g. IHA: $C_6H_7N_3O$ + H - CO --> $C_5H_8N_3$)**

Peak Annotation - Small Molecule (for high resolution MS/MS spectra)

- **Peaks were annotated with the most probable chemical formula consistent with the precursor formula**

formula valence=∑Count*(valence-2) + 2 + charge

⇧

count of each element

formula valence<=# of H; H:C ratio>=0.125

$$\text{Accuracy (ppm)} = \frac{|(\text{Observed m/z} - \text{Theoretical m/z})|}{\text{Observed m/z}} \times 10^6$$

10 ppm

$$\text{Unassigned (\%)} = \frac{\text{Sum of intensities of unassigned peaks}}{\text{Sum of intensities of all peaks}} \times 100$$

10 %

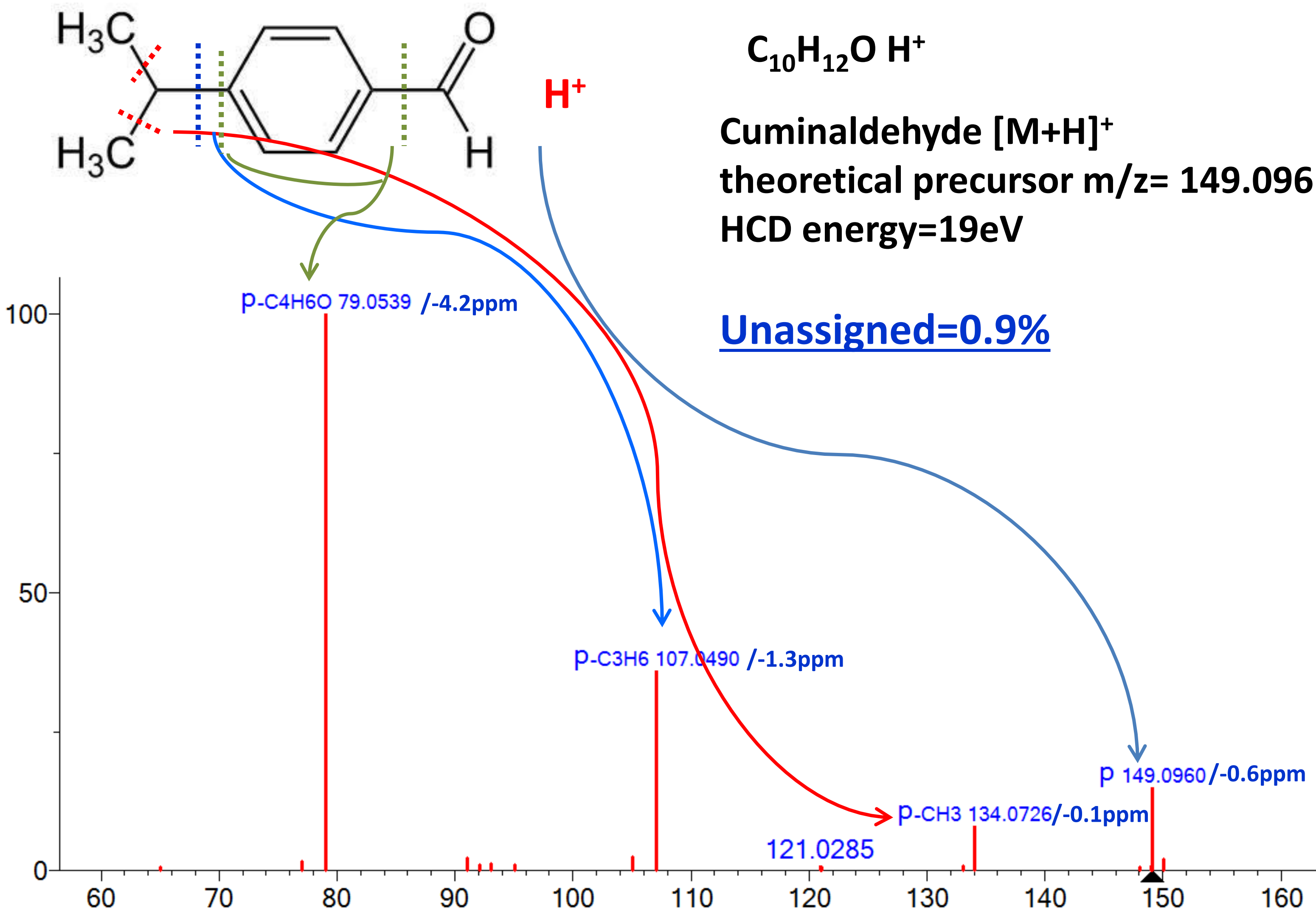
T. Kind and O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* 2007, 8:105 doi:10.1186/1471-2105-8-105

Peak Annotation - Small Molecule (for high resolution MS/MS spectra)

$C_{10}H_{12}O\ H^+$

Cuminaldehyde [M+H]$^+$
theoretical precursor m/z= 149.0961
HCD energy=19eV

Unassigned=0.9%



Anal. Chem. 2014, 86: 6393−6400

analytical chemistry

Article

pubs.acs.org/ac

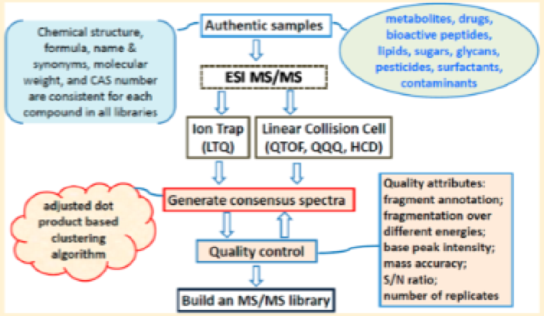
Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra

Xiaoyu Yang,* Pedatsur Neta, and Stephen E. Stein

Mass Spectrometry Data Center, National Institute of Standards and Technology, Mail Stop 8362, Gaithersburg, Maryland 20899, United States

Supporting Information

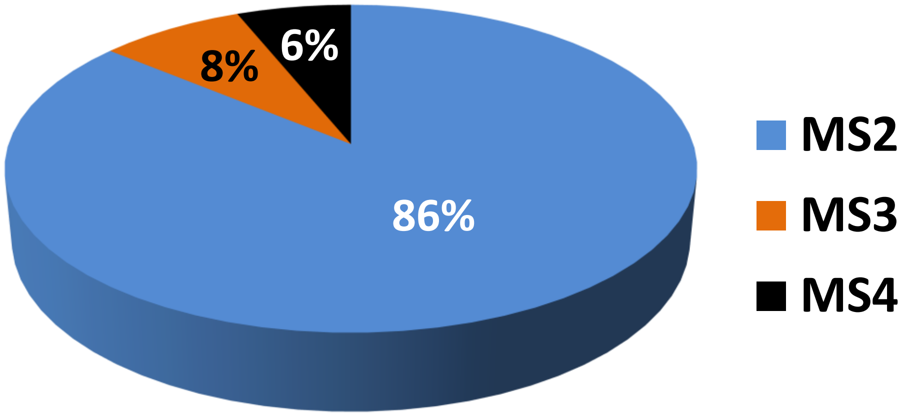
ABSTRACT: Electrospray ionization (ESI) tandem mass spectrometry coupled with liquid chromatography is a routine technique for identifying and quantifying compounds in complex mixtures. The identification step can be aided by matching acquired tandem mass spectra (MS^2) against reference library spectra as is routine for electron ionization (EI) spectra from gas chromatography/mass spectrometry (GC/MS). However, unlike the latter spectra, ESI MS^2 spectra are likely to originate from various precursor ions for a given target molecule and may be acquired at varying energies and resolutions and have characteristic noise signatures, requiring processing methods very different from EI to obtain complete and high quality reference spectra for individual analytes. This paper presents procedures developed for creating a tandem mass spectral library that addresses these factors. Library building begins by acquiring MS^2 spectra for all major MS^1 peaks in an infusion run, followed by assigning MS^2 spectra to clusters and creating a consensus spectrum for each. Intensity-based constraints for cluster membership were developed, as well as peak testing to recognize and eliminate suspect peaks and reduce noise. Consensus spectra were then examined by a human evaluator using a number of criteria, including a fraction of annotated peaks and consistency of spectra for a given ion at different energies. These methods have been developed and used to build a library from >9000 compounds, yielding 230,000 spectra.



Mass spectral reference libraries of electron ionization (EI) spectra are used extensively and routinely to identify compounds separated by gas chromatography.[1] For example, the current NIST/EPA/NIH Mass Spectral EI Library contains spectra for over 200,000 compounds and is a common, tightly integrated component in many gas chromatography/mass spectrometry (GC/MS) data systems.[2,3] Such use of reference libraries for the identification of electrospray ionization (ESI) tandem mass spectra (MS^2) has, however, been far more limited. While certain MS^2 reference libraries are available for specific applications, such as METLIN[4] for metabolomics, they are often limited to specific platforms or not integrated with an instrument data system.[4−10] Also unlike EI libraries,[2]

NIST has undertaken the production of a comprehensive ESI MS^2 library for a wide range of molecules,[11,12] intended for use on a variety of platforms and in a range of applications. Different methods are required for development of an ESI MS^2 library in comparison to those used for the odd-electron, positive ion, unit mass resolution MS^1 EI library. ESI MS^2 spectra are the result of even-electron transfer of ionic charge to neutral molecules in solutions at atmospheric pressure, frequently resulting in simple spectra with sparse fragmentation. The presence of multiple precursor ions and charge states for a single analyte is a necessary consequence of the ESI experiment in which protons or other cations or anions impart charge and

What Types of Mass Spectra are in the NIST MSMS Library?



<u>Instruments:</u>

Micromass Quattro Micro: Triple Quadrupole

Thermo Finnigan LTQ: IT/ion trap

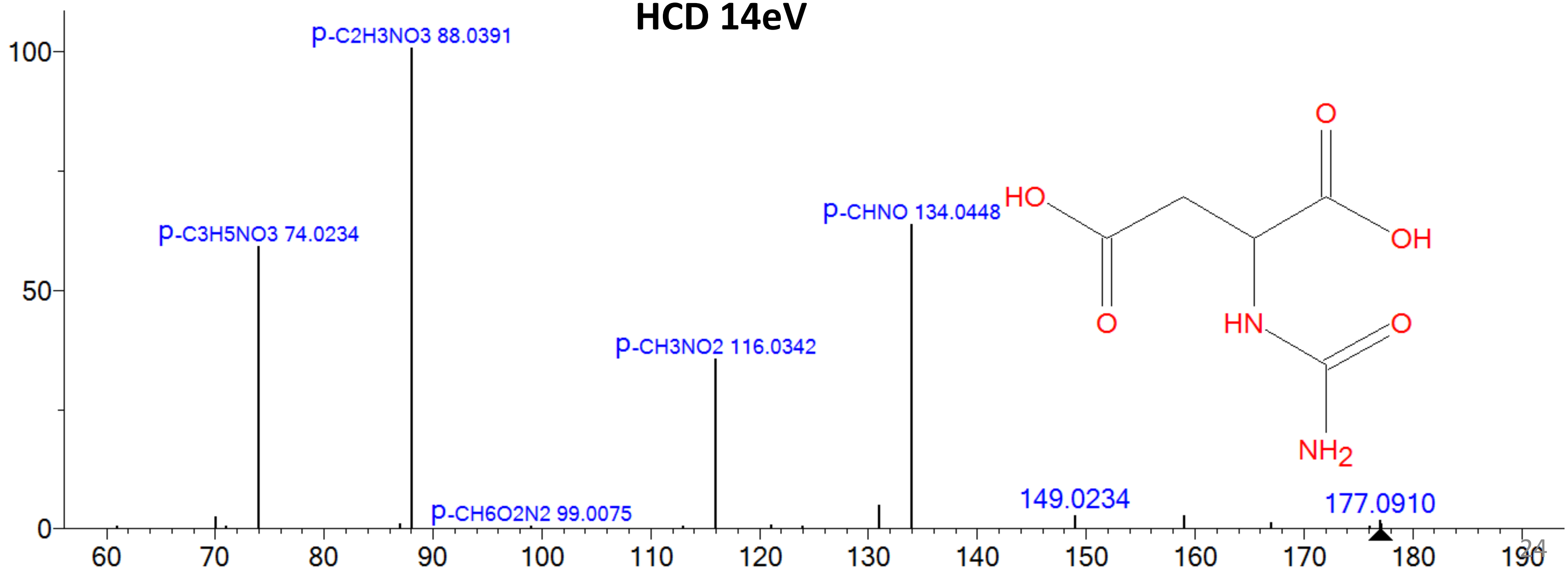
Agilent QTOF 6530: Q-TOF

Thermo Finnigan Elite Orbitrap: HCD
IT-FT/ion trap with FTMS
IT/ion trap

What Types of Compounds are in the NIST MSMS Library?

Metabolites ~50%

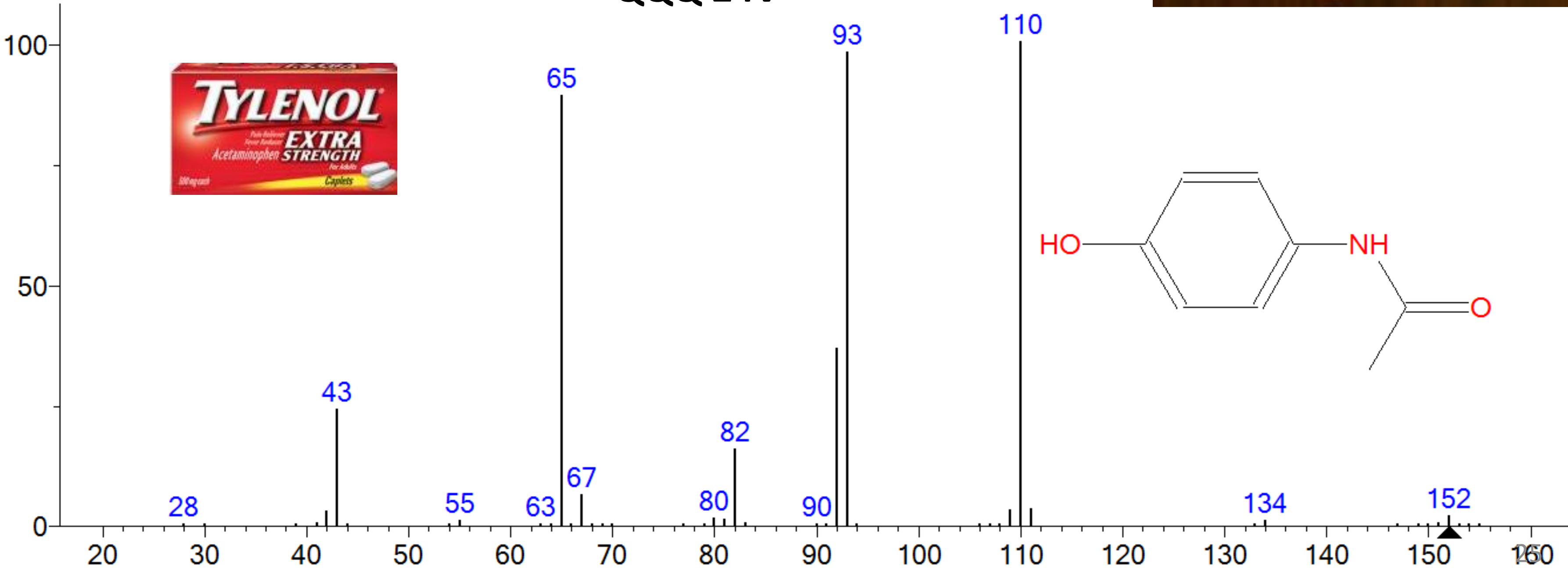
Ureidosuccinic acid
[M+H]+
HCD 14eV



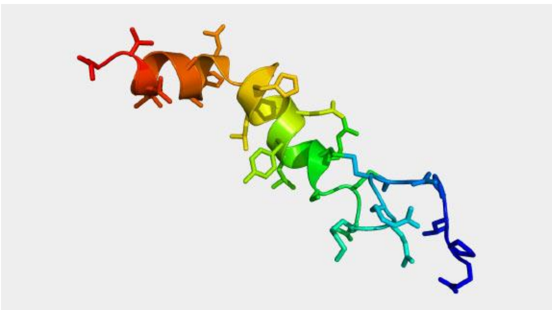
What Types of Compounds are in the NIST MSMS Library?

Drugs ~20%

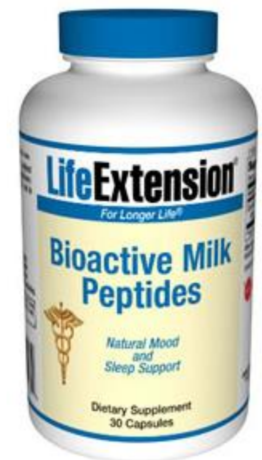
Acetaminophen
[M+H]+
QQQ 24V

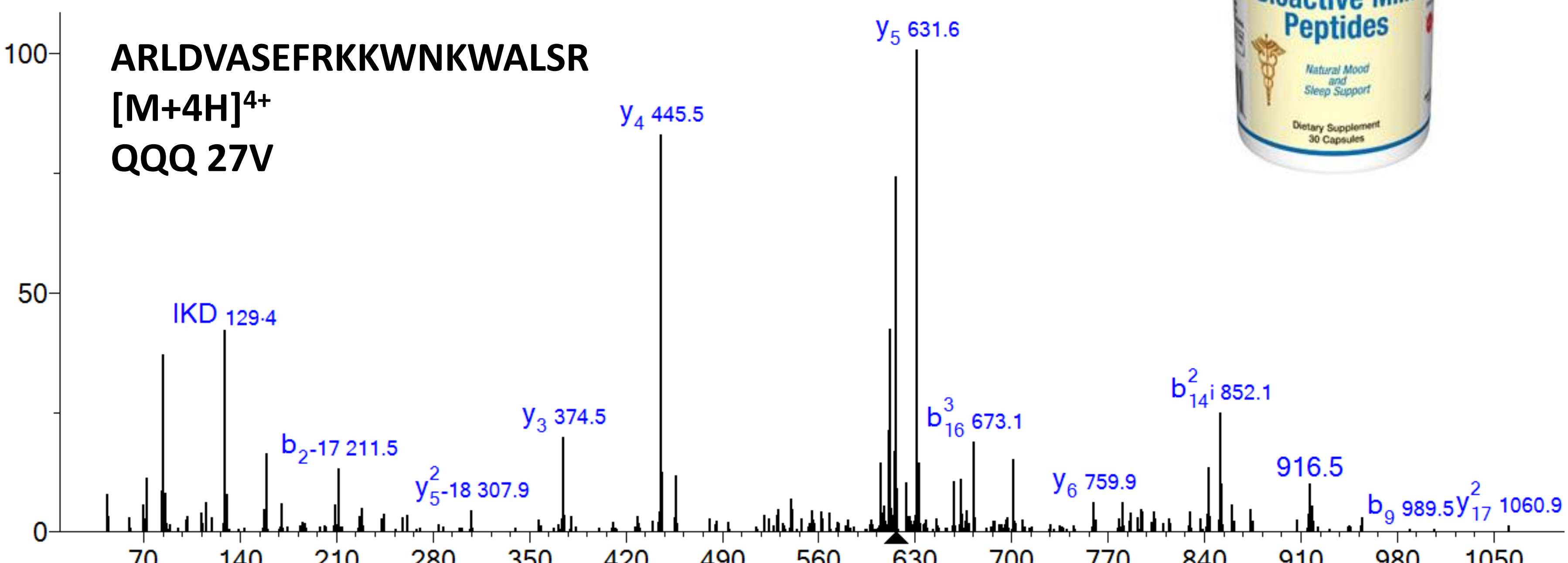


What Types of Compounds are in the NIST MSMS Library?



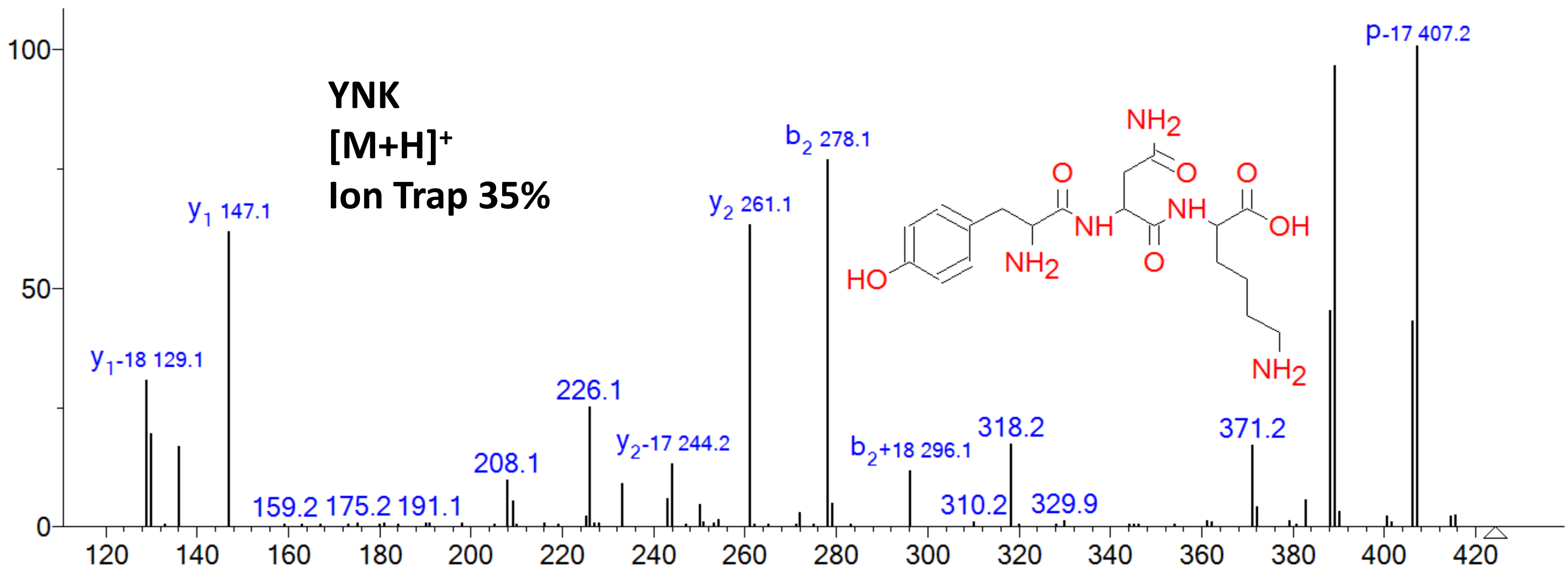
Bioactive peptides
~10%





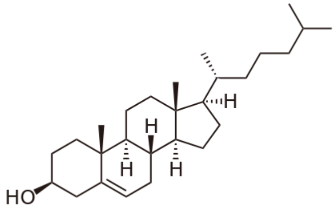
What Types of Compounds are in the NIST MSMS Library?

All amino acids (20)
All dipeptides (400)
All tryptic tripeptides (800)

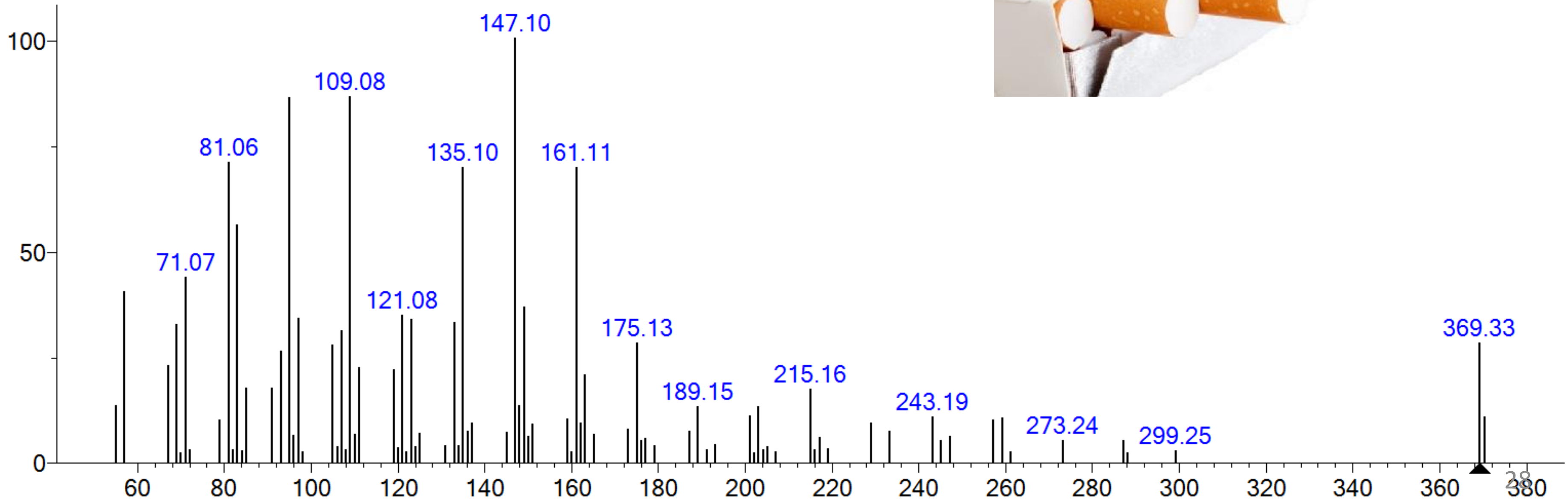


What Types of Compounds are in the NIST MSMS Library?

Lipids ~ 5%

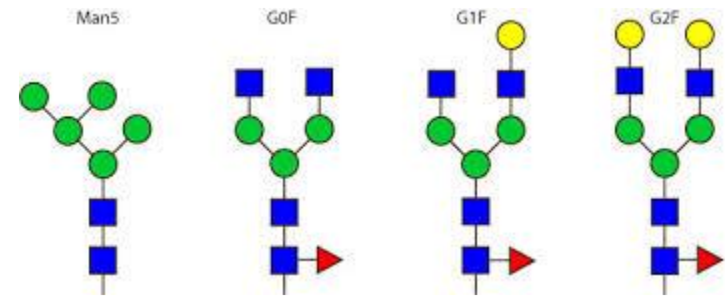


Cholesterol
[M+H-H2O]+
QTOF 20V

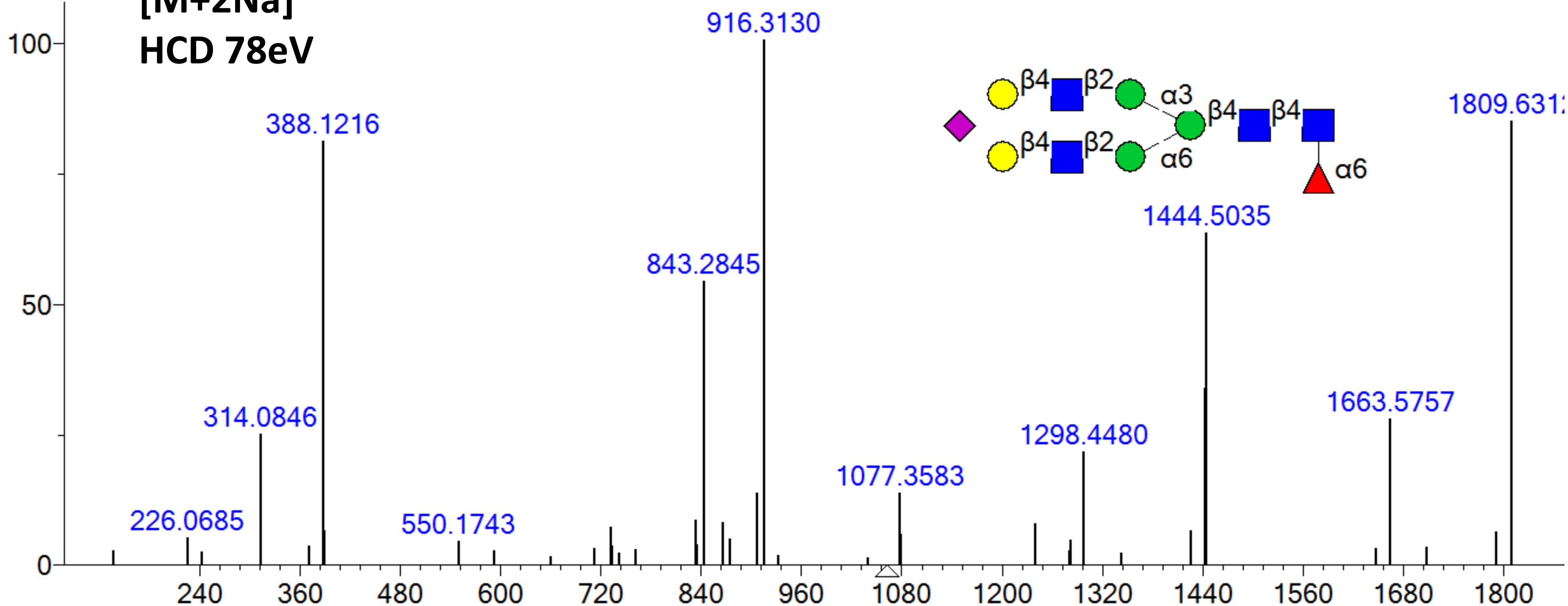


What Types of Compounds are in the NIST MSMS Library?

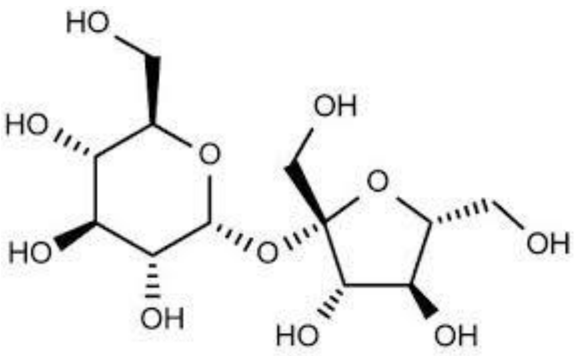
Glycans ~ 2%



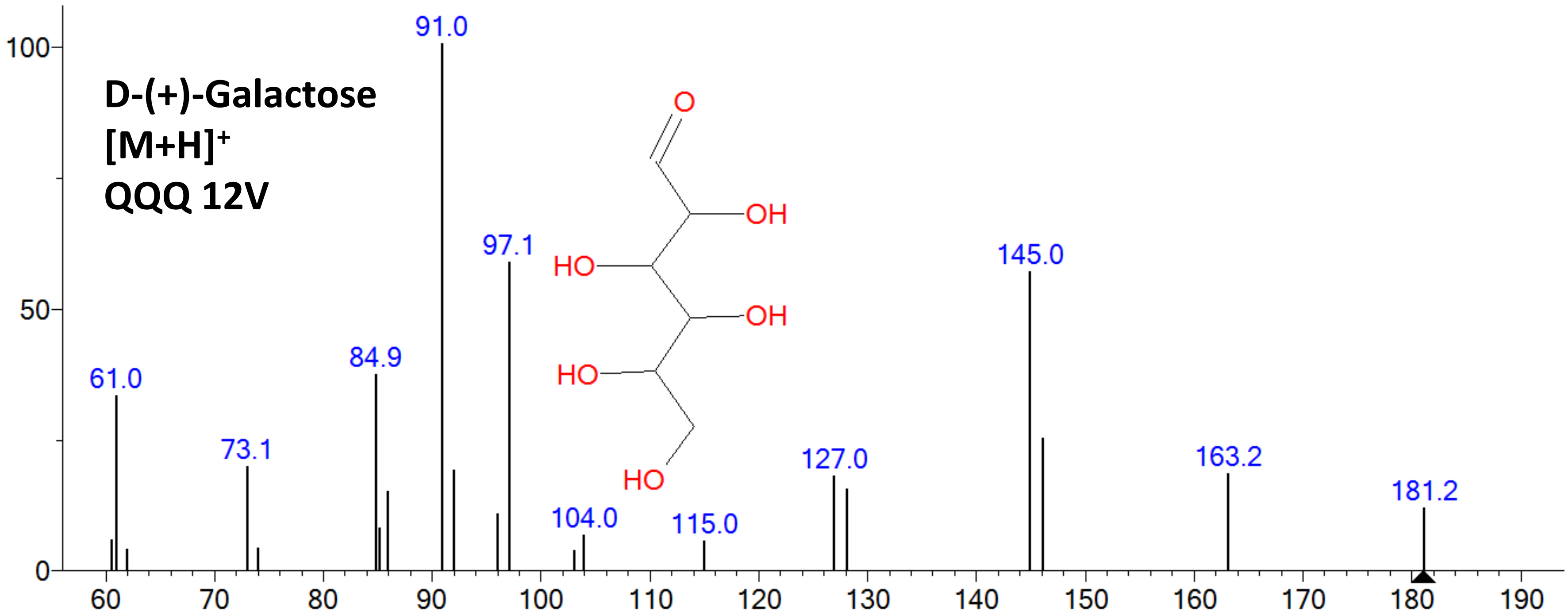
Glycan A1F-MIX
$[M+2Na]^{2+}$
HCD 78eV



What Types of Compounds are in the NIST MSMS Library?

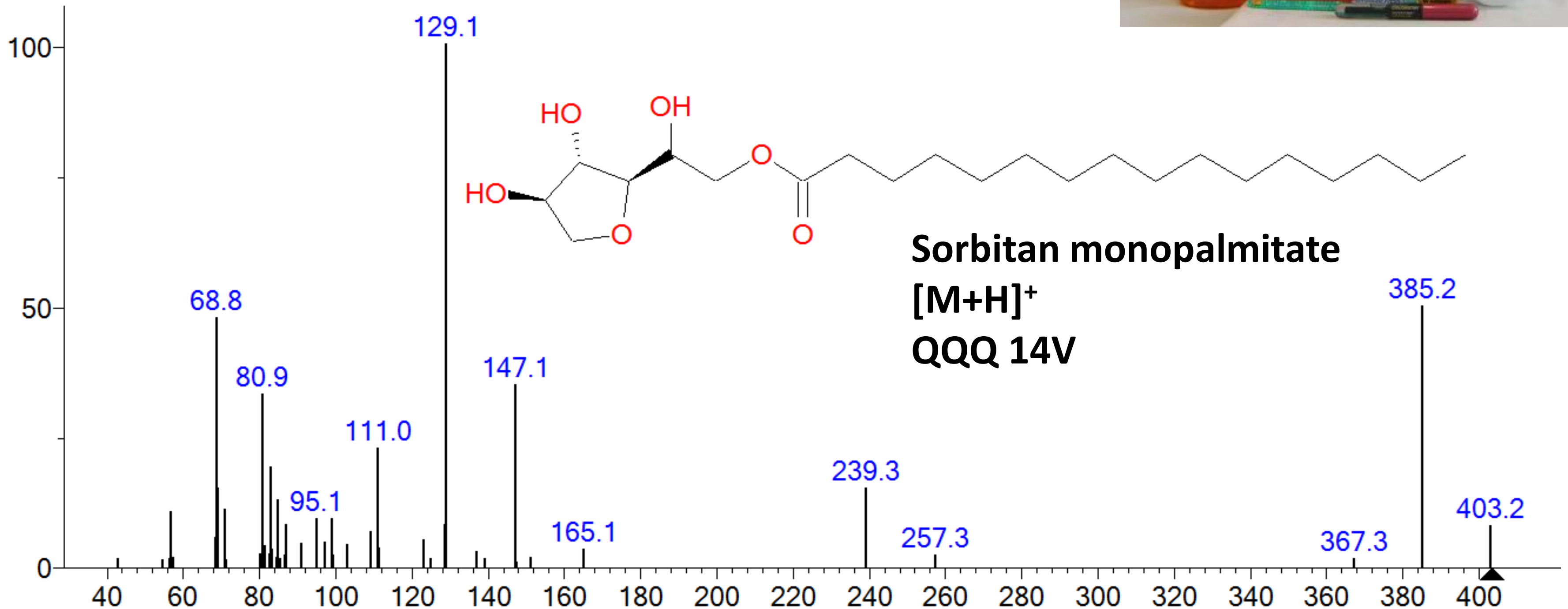


Sugars ~ 2%



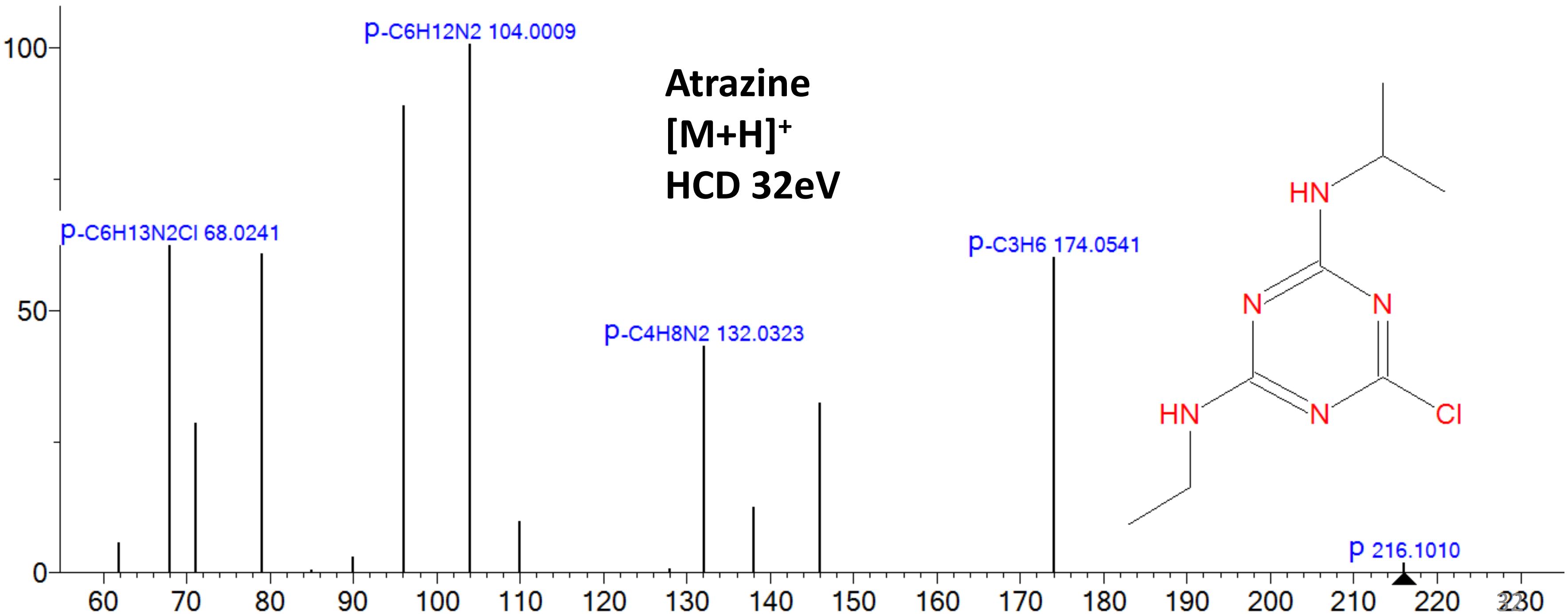
What Types of Compounds are in the NIST MSMS Library?

Surfactants and Contaminants

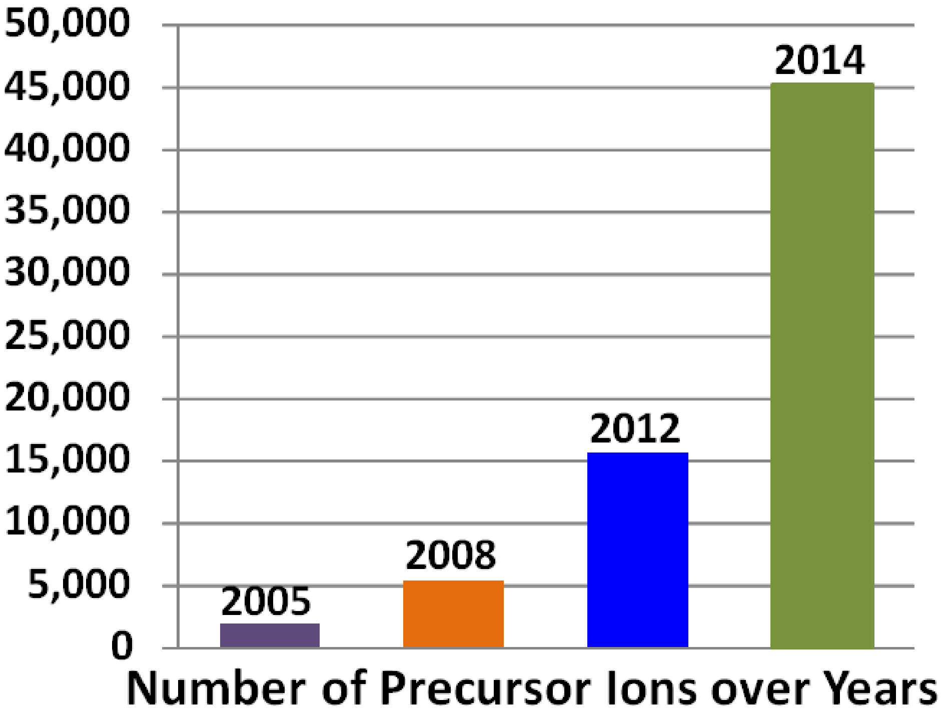


What Types of Compounds are in the NIST MSMS Library?

Pesticides



NIST Tandem Mass Spectral Library 2014



9,345 Compounds
45,298 Precursor Ions
234,284 Spectra
~90% Positive Ion Spectra
~10% Negative Ion Spectra

Instrument Type	Precursor Ions
Ion Trap	>40,000
Collision Cell (QTOF, QQQ, HCD)	>14,000

Conclusions

- Two level clustering algorithms (counter-based and distance-based) were developed and provide a robust means of generating consensus spectra of multiple precursor types for the NIST Tandem Mass Spectral Library.

- Quality control programs such as peak annotation and noise removal methods were developed and are used in building the reference quality NIST Tandem Mass Spectral Library.

- *NIST Tandem Mass Spectral Library 2014* can be applied in chemical identification in Metabolomics, Proteomics and other fields.

Plans for the future:

- Improve peak annotation program by using compound's structure and physical chemical properties.

- Develop more quality control methods to eliminate low quality spectra for the library.

Acknowledgements