

Developing a Data Processing Pipeline for Extending a Comprehensive Tandem Mass Spectral Library

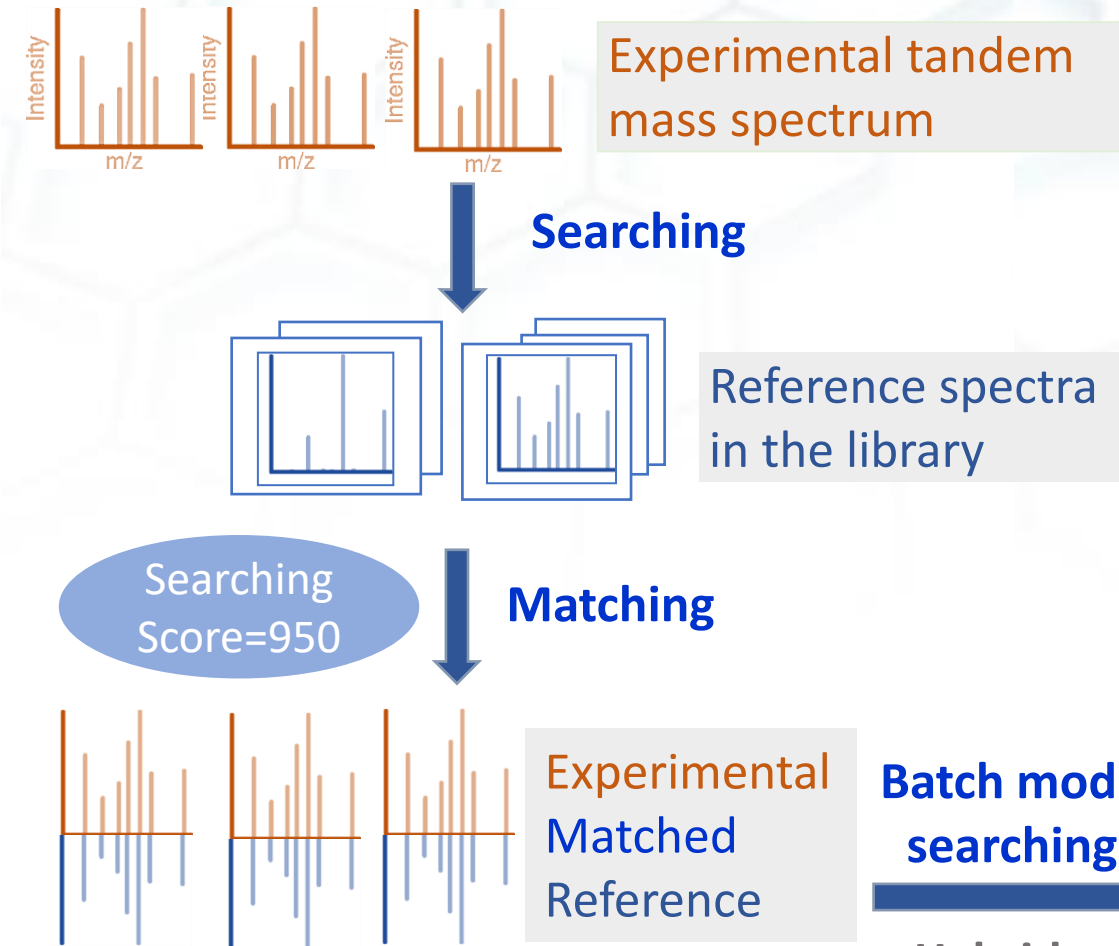


Xiaoyu Yang, Pedatsur Neta,

Yuxue Liang, Connie A. Remoroza, Yamil Simón-Manso, Kelly H. Telu,
Yuri Mirokhin, Dmitrii Tchekhovskoi, Alexey Mayorov, Tytus D. Mak,
Lewis Geer, Stephen E. Stein

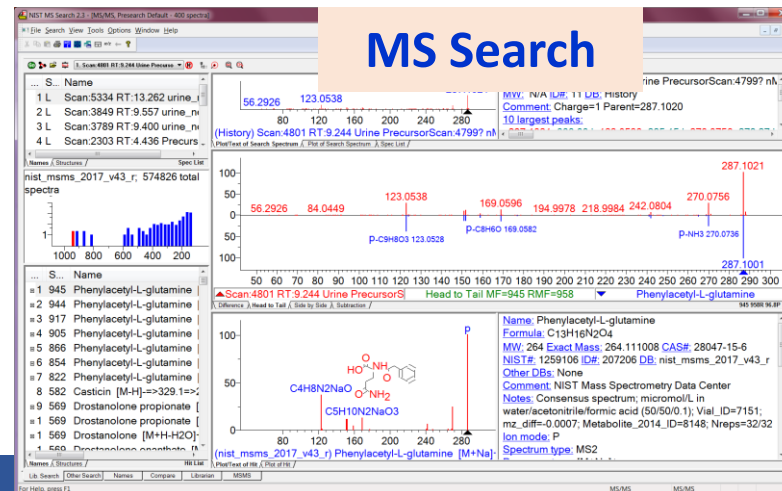
NIST Mass Spectrometry Data Center
Gaithersburg, MD

Tandem Mass Spectral Library



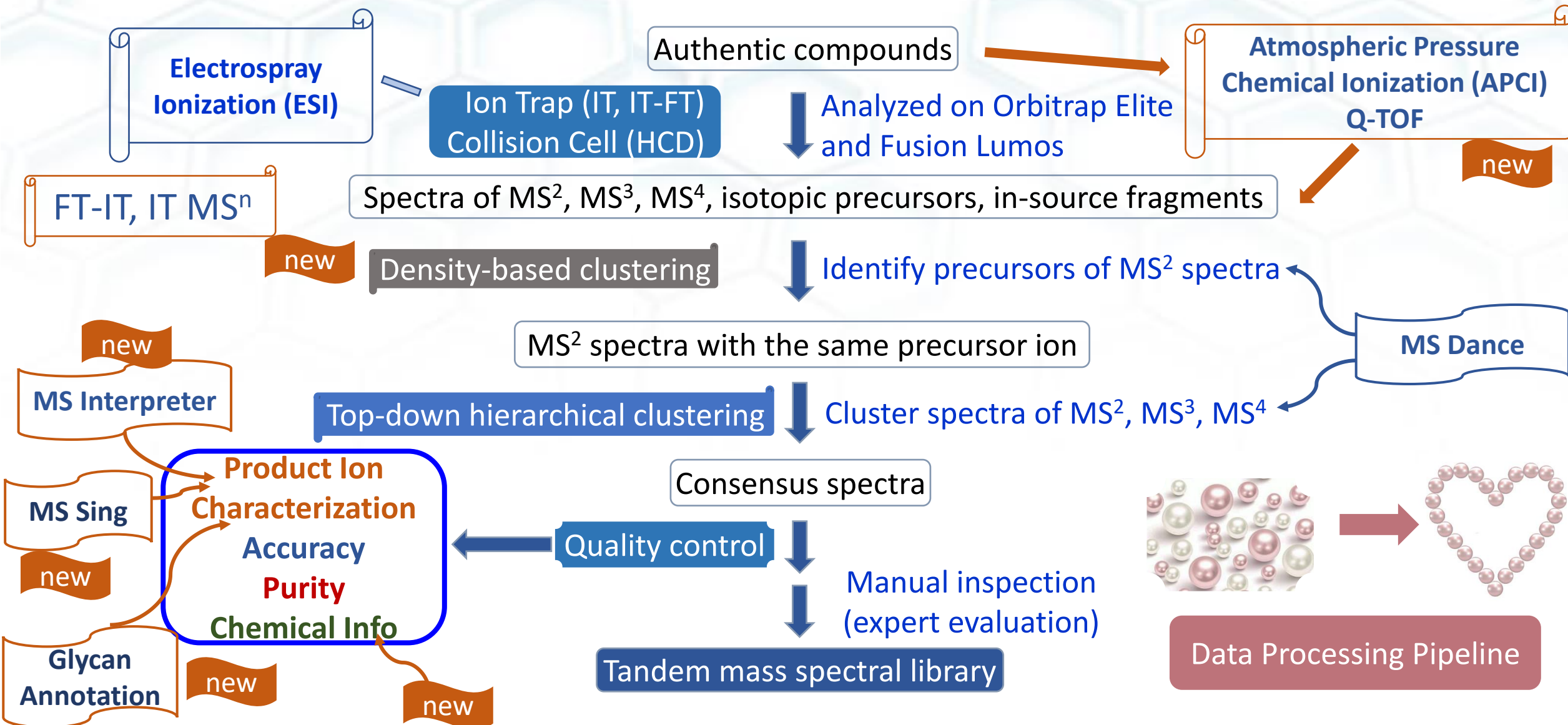
- ❖ Mass spectral library searching is a fast and reliable technique to identify compounds from LC/MS/MS data.
- ❖ Building a **comprehensive, high quality and reference** tandem mass spectral library is essential for the accurate compound identification.
- ❖ More compounds, more compound types;
- ❖ Different energies, precursor ions;
- ❖ Various mass spectrometers with different ion sources.

Batch mode searching
 Hybrid searching



I	J	K
Metabolite	Formula	Prec.Type
Diethyl succinate	C8H14O4	[M+H-C4H10O]+
L-.beta.-Homoserine	C4H9NO3	[M+H-H2O]+
7-Hydroxychromanone	C9H8O3	[M+H]+
3,4-Dihydroxy-L-phenyl	C9H11NO4	[M+H-NH3]+
5-Aminolevulinic acid	C5H9NO3	[M+H]+
Bestatin	C16H24N2O4	[M+H-C8H15O4N]+
Benzeneethanamine	C8H11N	[M+H]+
(+/-)-SKF 81297	C16H16ClNO2	[M+H-C8H7ClO2]+
25-Hydroxycholesterol	C27H46O2	[M+H-H4O2]+
25-Hydroxycholesterol	C27H46O2	[M+H-H4O2]+

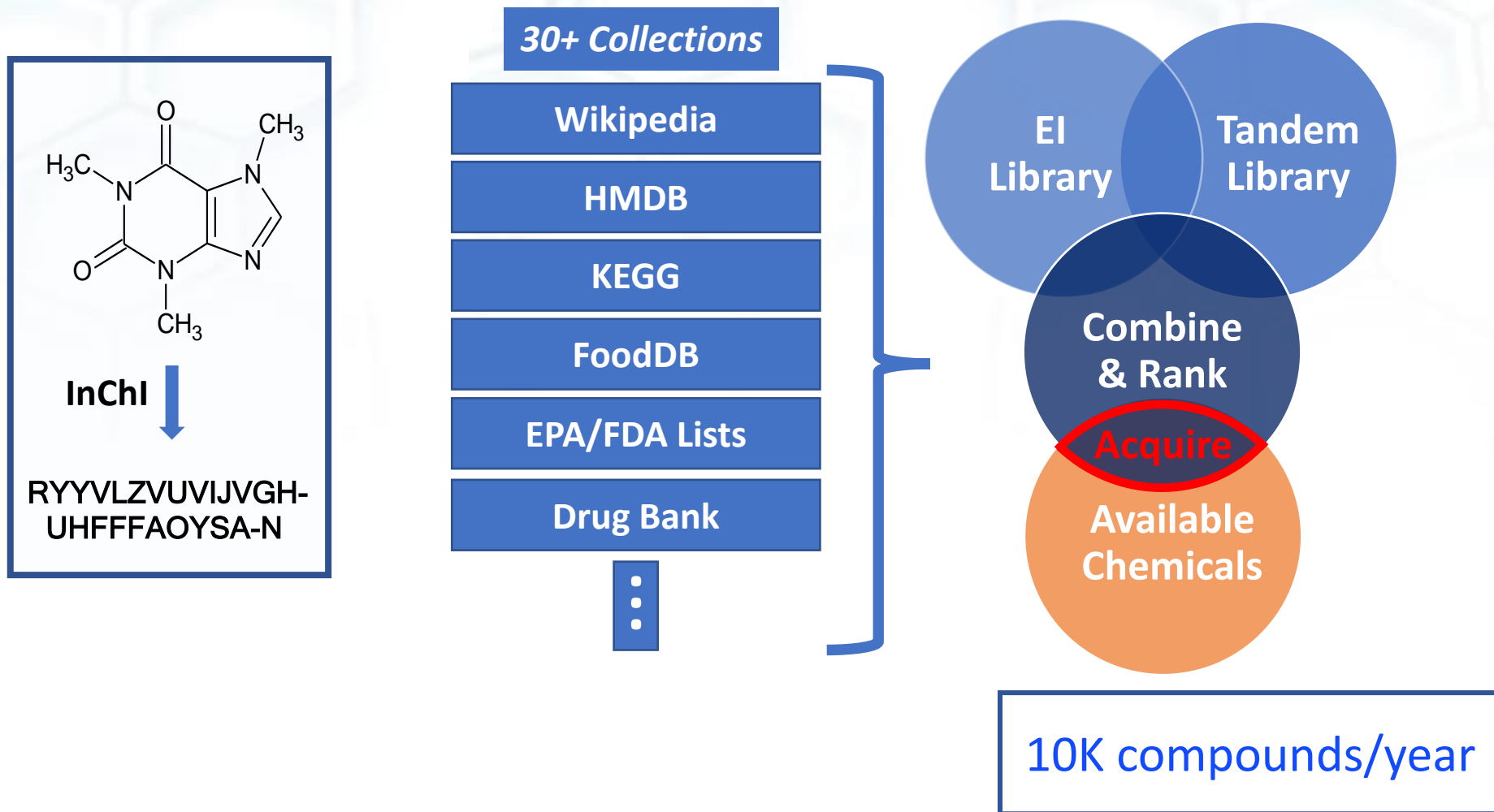
Procedure of Extending the NIST Tandem Mass Spectral Library



X. Yang, P. Neta, S. Stein. Extending A Tandem Mass Spectral Library to Include MS² Spectra of Fragment Ions Produced In-Source and MSⁿ Spectra. *J. Am. Soc. Mass Spectrom.* November 2017, 28: 2280–2287.

X. Yang, P. Neta, S. Stein. Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra. *Anal. Chem.*, 2014, 86: 6393-6400.

New Compound Selection Process



Big Data Analysis

50,000 spectra per compound
↓ for 30,000 compounds
1,500,000,000 spectra

Ion source: ESI, APCI

Instruments: HCD, IT, FT-IT; Q-TOF

Modes: Positive, Negative

Spectrum types: MS², MS³, MS⁴

Energies: 20 steps, 2%-320%

Precursors:

[M+H]⁺, [M+2H]²⁺, [M-H]⁻, [M-2H]²⁻;
[M+Na]⁺; Dimers, Trimers; Isotopic Precursors;
MSⁿ

In-source Fragments: [M+H-neutral]⁺; [M-H-neutral]⁻

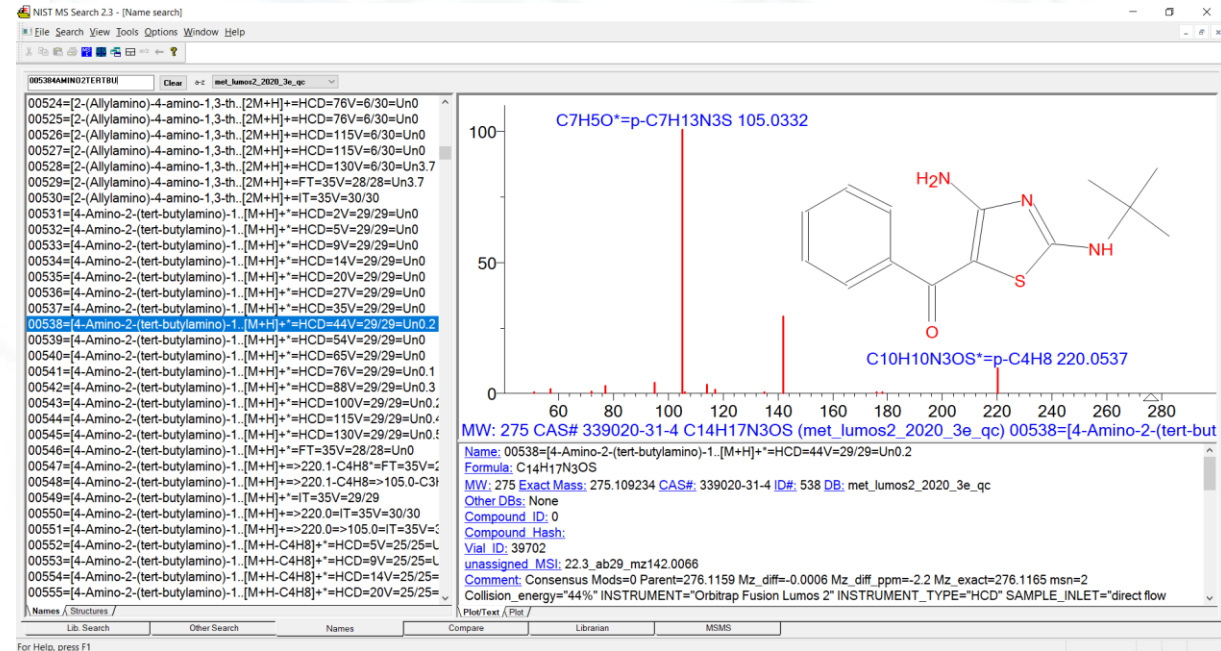
The Data Processing Pipeline

- Glycans
- Metabolites
- Peptides

- Fusion_Lumos
- Fusion_Lumos2
- Orbitrap_Elite
- QTOF

- merged_RT_mgf
- mzp
- mz_consensus
- QC
- raw

automation.cmd

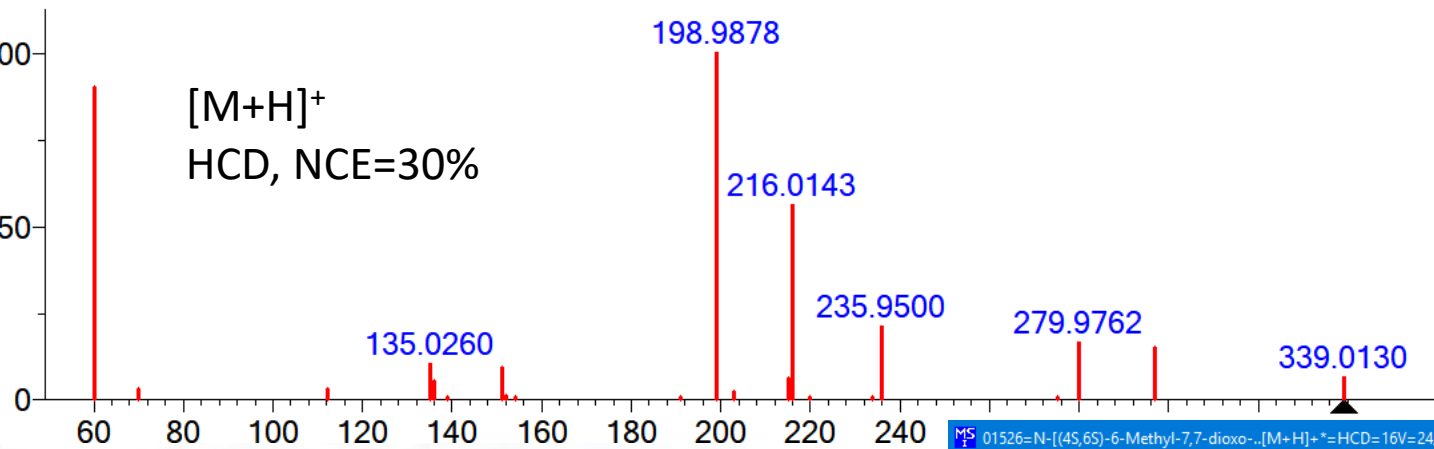


Annotating Product Ions

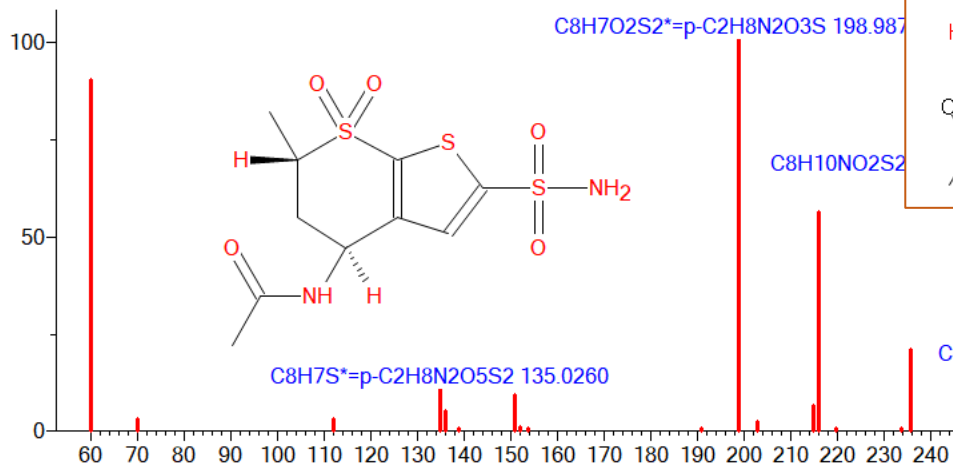
MS Interpreter 3.4

chemdata.nist.gov

[M+H]⁺
HCD, NCE=30%



N-[(4S,6S)-6-Methyl-7,7-dioxo-2-sulfamoyl-5,6-dihydrothiopyran-4-yl]acetamide

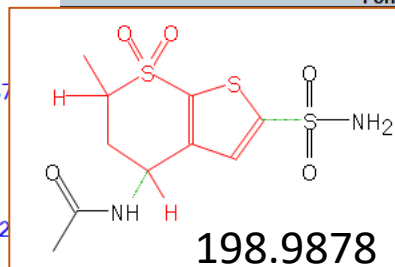


MS Sing

MS 01526=N-[(4S,6S)-6-Methyl-7,7-dioxo-2-sulfamoyl-5,6-dihydrothiopyran-4-yl]acetamide [M+H]⁺=HCD=16V=24/24=Un0 [M+H]⁺ - MS Interpreter

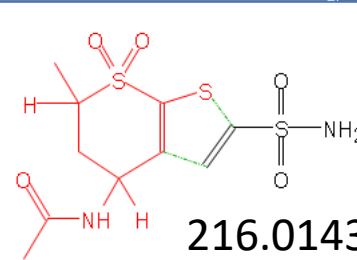
File Edit View Options Help

Formula Calculator

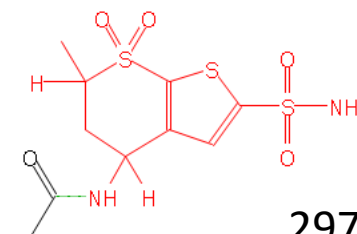


198.9878

-H₂O



216.0143



297.0032

Maximum Rate: 103(chain) @ 279.9766 m/z

exact m/z	formula	loss	type	H	rate	rel.rate	io
297.003197	C ₈ H ₁₃ N ₂ O ₄ S ₃	C ₂ H ₂ O	H-Add/Dissoc	+2	70	67	

Mass Spectrum for C₁₀H₁₄N₂O₅S₃(H); Mass = 339.0143; CAS = 147200-03-1; [M+H]⁺

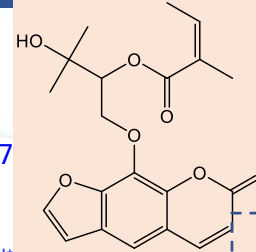
protonated 01526=N-[(4S,6S)-6-Methyl-7,7-dioxo-2-sulfamoyl-5,6-dihydrothiopyran-4-yl]acetamide [M+H]⁺=HCD=16V=24/24=Un0 [M+H]⁺



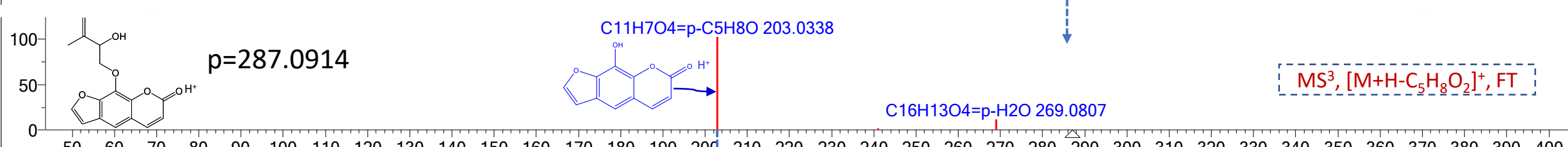
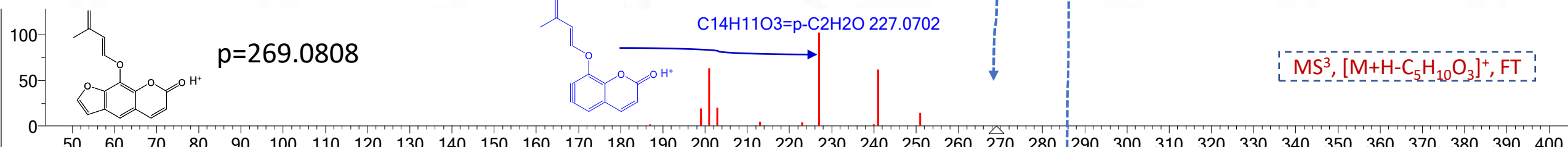
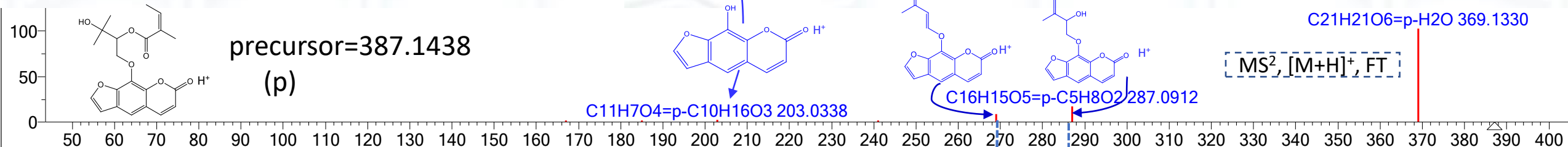
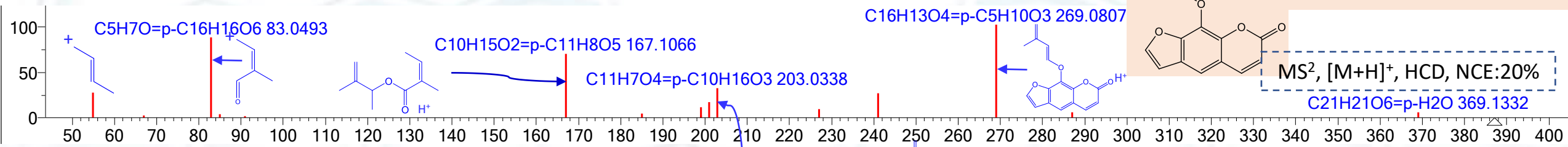
L-Click: Select R-Click: Menu I-DbClick: send or set parent. I-Drag: zoom

Product Ion Characterization for High Resolution MSⁿ Spectra

Tomazin
Formula: C₂₁H₂₂O₇



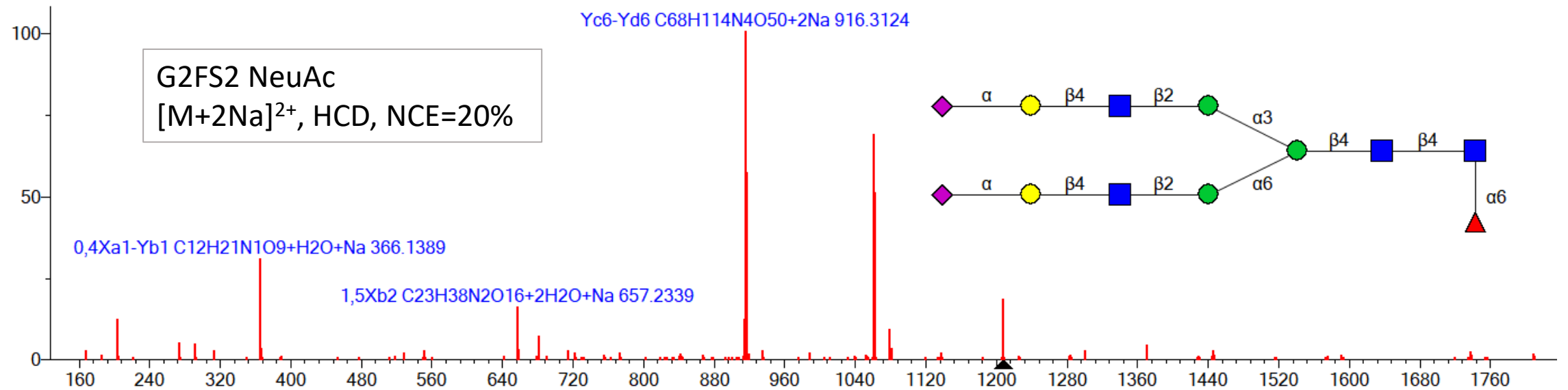
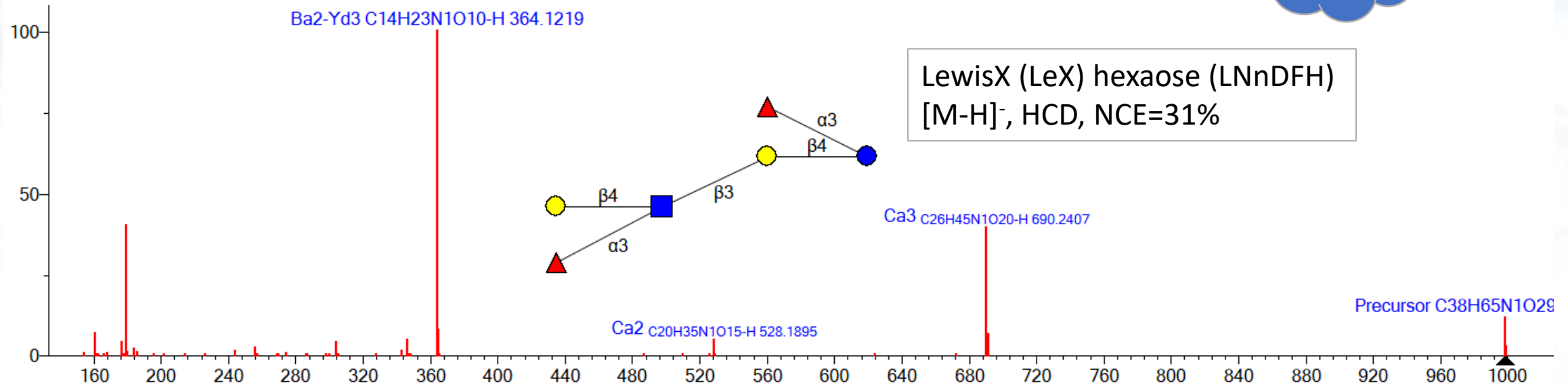
MS², [M+H]⁺, HCD, NCE:20%
C₂₁H₂₁O₆=p-H₂O 369.1332



NEW

MS⁴, [M+H-C₅H₈O₂-C₅H₈O]⁺, FT

Product Ion Annotation for Glycans

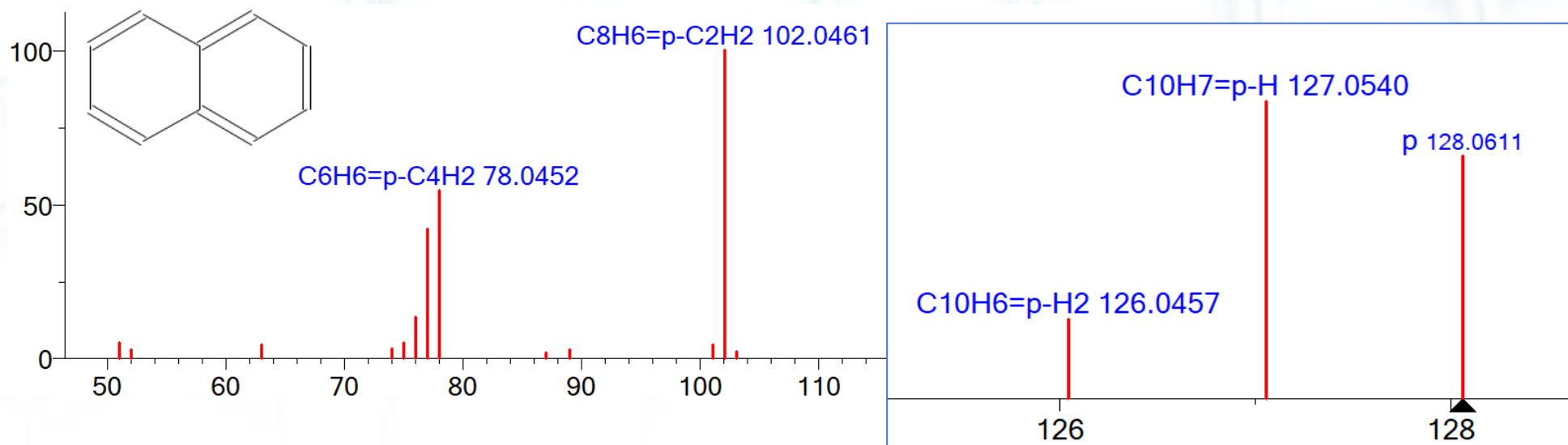


NEW

Q-TOF Spectra with APCI for Extractable and Leachable Compounds

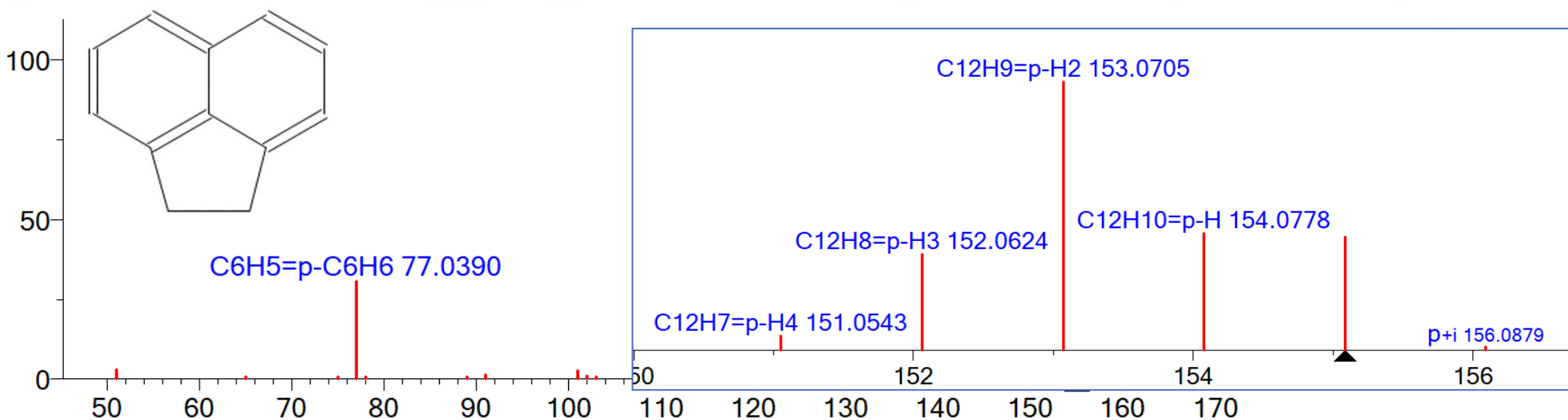
Naphthalene
[M]⁺.

Collision energy: 35V
Cone voltage: 175V

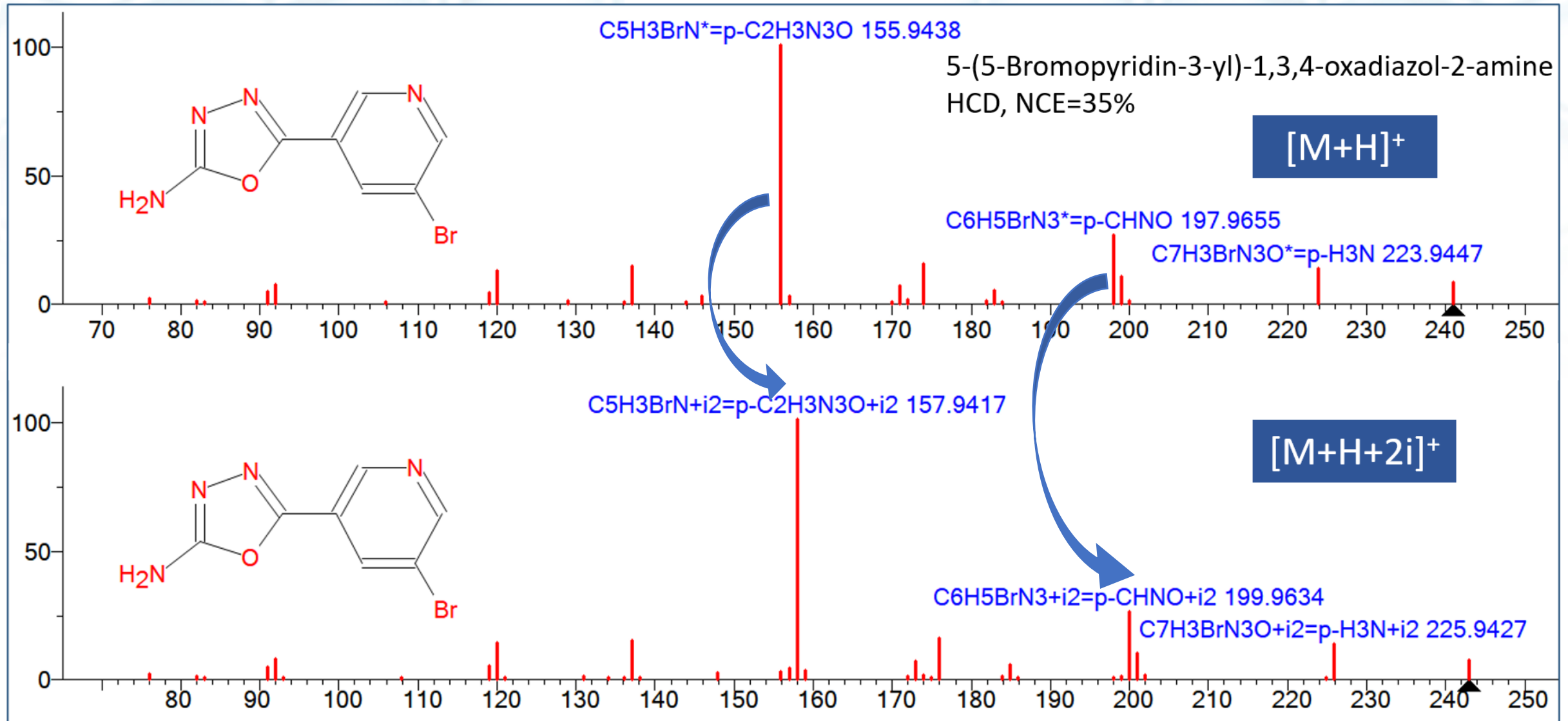


Acenaphthene
[M+H]⁺

Collision energy: 30V
Cone voltage: 175V

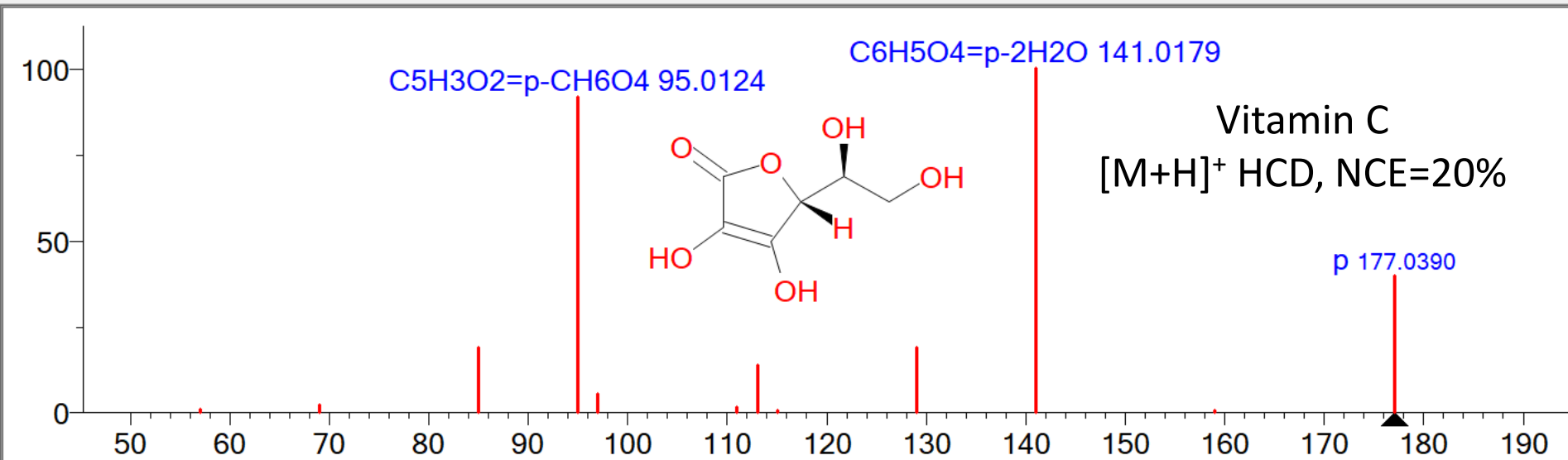


Using Pipeline for Quality Control of Spectra with Monoisotopic and Isotopic Precursors



Different Instruments, Precursors

- VITAMIN C
- Clear a-z hr_msms_nist2020_v42
- Vitamin C [M+H]⁺ HCD 2% P=177
 - Vitamin C [M+H]⁺ HCD 2% P=177
 - Vitamin C [M+H]⁺ HCD 10% P=177
 - Vitamin C [M+H]⁺ HCD 15% P=177
 - Vitamin C [M+H]⁺ HCD 20% P=177**
 - Vitamin C [M+H]⁺ HCD 25% P=177
 - Vitamin C [M+H]⁺ HCD 30% P=177
 - Vitamin C [M+H]⁺ HCD 35% P=177
 - Vitamin C [M+H]⁺ HCD 40% P=177
 - Vitamin C [M+H]⁺ HCD 40% P=177
 - Vitamin C [M+H]⁺ HCD 50% P=177
 - Vitamin C [M+H]⁺ HCD 60% P=177
 - Vitamin C [M+H]⁺ HCD 75% P=177
 - Vitamin C [M+H]⁺ HCD 90% P=177
 - Vitamin C [M+H]⁺ IT-FT 35% P=177
 - Vitamin C [M+H]⁺ QTOF 8V P=177
 - Vitamin C [M+H]⁺ QTOF 10V P=177
 - Vitamin C [M+H]⁺ QTOF 12V P=177
 - Vitamin C [M+H]⁺ QTOF 14V P=177
 - Vitamin C [M+H]⁺ QTOF 20V P=177
 - Vitamin C [M+H]⁺ QTOF 22V P=177
 - Vitamin C [M+H]⁺ QTOF 24V P=177
 - Vitamin C [M+H]⁺ QTOF 26V P=177
 - Vitamin C [M+H-2H₂O]⁺ HCD 2% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 5% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 5% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 15% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 20% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 25% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 30% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 30% P=141
 - Vitamin C [M+H-2H₂O]⁺ HCD 40% P=141



MW: 176 CAS# 50-81-7 C₆H₈O₆ (hr_msms_nist2020_v42) Vitamin C [M+H]⁺ HCD 20% 7 P=177

Name: Vitamin C
 Formula: C₆H₈O₆
 MW: 176 Exact Mass: 176.032088 CAS#: 50-81-7 NIST#: 1541592 ID#: 320379 DB: hr_msms_nist2020_v42
 Other DBs: None
 Compound ID: 0
 Compound Hash:
 Comment: NIST Mass Spectrometry Data Center
 Notes: Consensus spectrum; Nreps=27/27; Mz_diff=-1.7ppm; Vial_ID=363; Metabolite_2
 (50/50/0.1)
 Ion mode: P
 Instrument: Thermo Finnigan Elite Orbitrap
 Instrument type: HCD
 Ionization: ESI
 Collision energy: NCE=20% 7eV

HCD, IT-FT, IT, Q-TOF
 108 spectra, 13 MSⁿ
 8 precursor ions:
 [M+H]⁺, [M+H-2H₂O]⁺, [M+H-H₂O]⁺,
 [M-H]⁻, [2M-H]⁻, [M-H-C₂H₄O₂]⁻

NIST Tandem Mass Spectral Library 2020

30,999 compounds

1,320,389 spectra

185,608 precursor ions

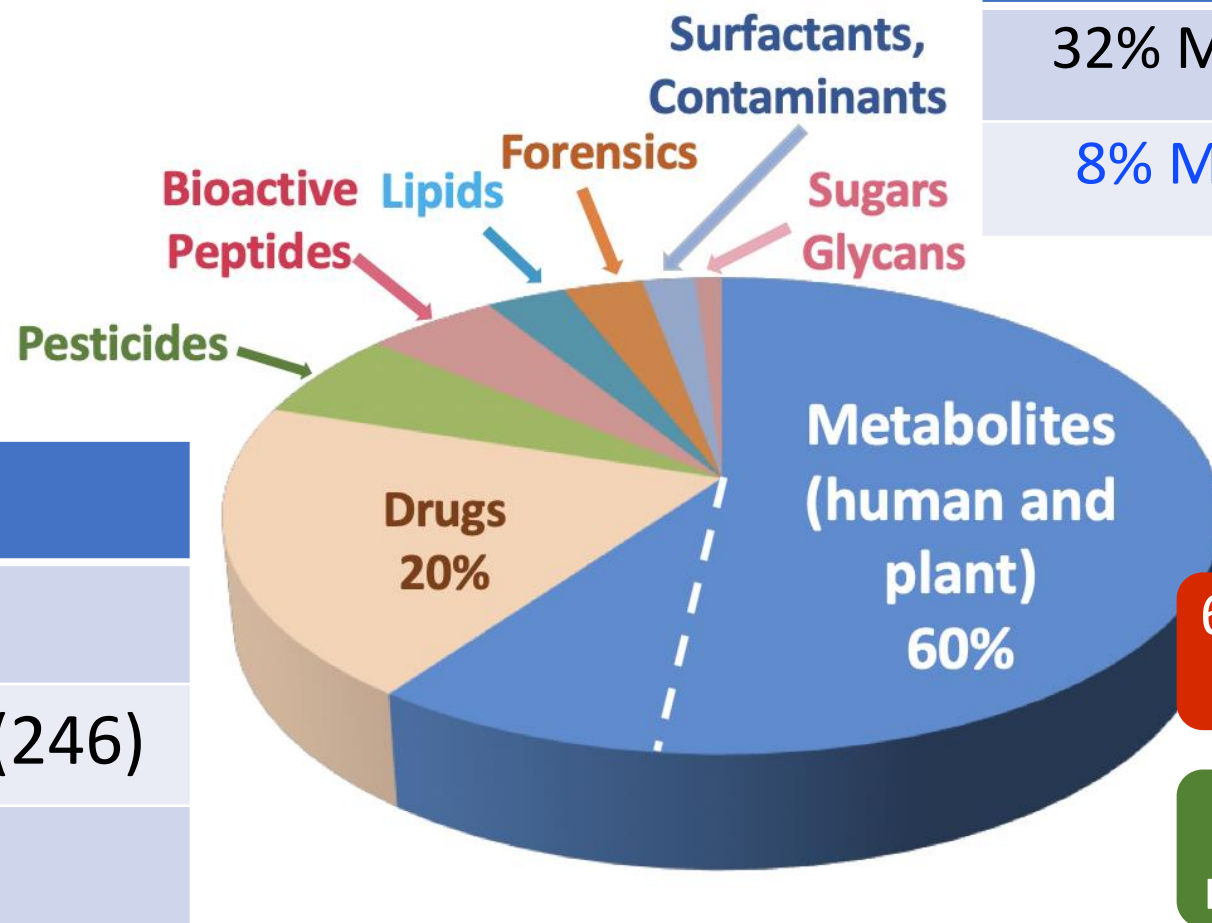
HRAM (27,840*)

Low Resolution (28,559)

APCI High Resolution Q-TOF (246)

Biological Peptides (1,904)

* numbers are the number of compounds



75%(+) 25%(-)

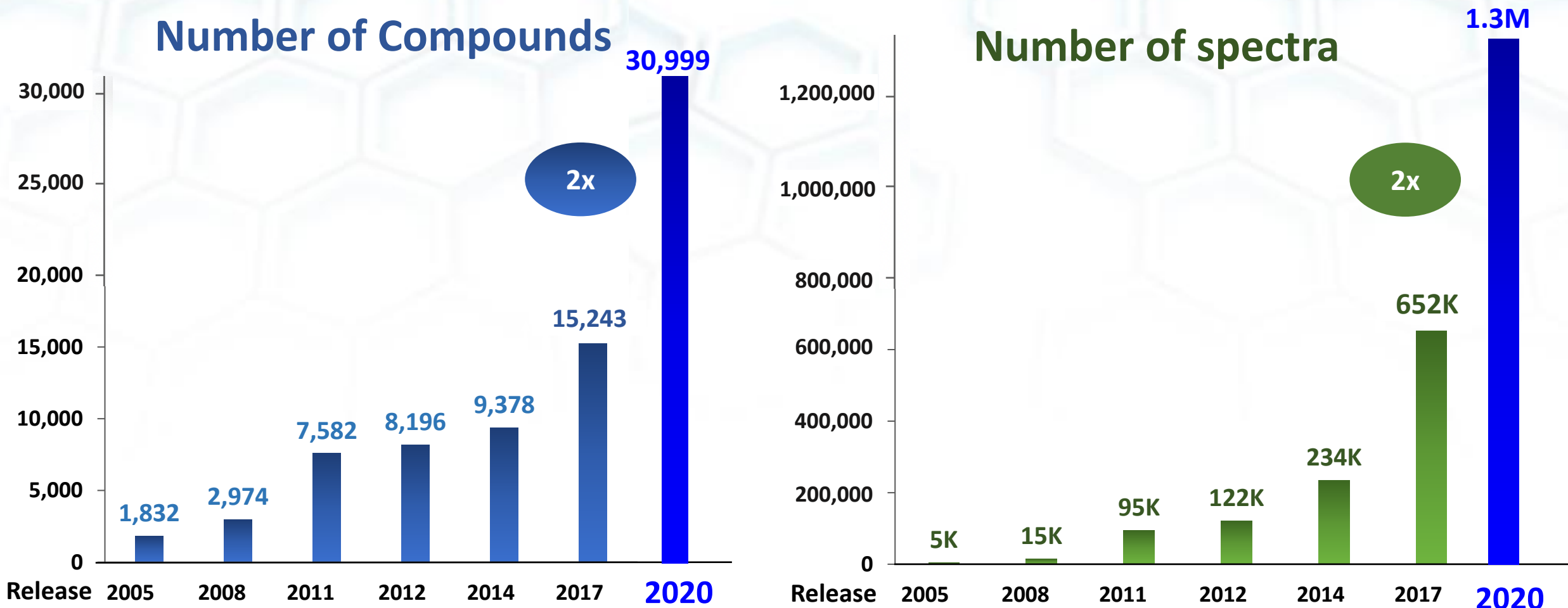
32% MS² in-source

8% MS³ and MS⁴

6,000 human metabolites

8,000 plant metabolites

NIST Tandem Mass Spectral Library 2020



Only in 2019: 7,841 new compounds, 256,532 new spectra

Library Application

- Metabolomics:
human plasma
human urine
human milk
E. Coli
CHO cell

- Proteomics
- Lipidomics
- Forensics
- Food
- Environmental studies



Identifying Human Metabolites by Searching the NIST Tandem MS Library



MS Search 2.4

NIST MS Search 2.4 - [MS/MS, Presearch Default - 400 spectra]

File Search View Tools Options Window Help

1. Scan:2006 RT:5.259 PrecursorScan:2004

#	S...	Name
1	A	Scan:2006 RT:5.259 Precursor
2	A	Scan:4801 RT:9.244 Precursor
3	A	Scan:2357 RT:4.535 Precursor
4	A	Scan:2349 RT:4.519 Precursor
5	A	Scan:2327 RT:4.480 Precursor
6	A	Scan:2329 RT:4.483 Precursor
7	A	Scan:2303 RT:4.436 Precursor
8	A	Scan:3789 RT:9.400 Precursor

Plot of Search Spectrum

Name: Scan:2006 RT:5.259 PrecursorScan:2004? nMSN:1547
MonoisPrMZfromRaw:0.0000 PrecursorCharge:1 PrecursorScanFTMS:1
FTResolution:30000 IBP:95614.56 ITot:381168.35 max2med:59.51
InjTime:100.00 HCD=25.00% IsolationMZ:290.1598 PrecursorAb:569302.56
MPY:1.00 ms1PrecursorTotAb:1176485853.64 ms1PrecursorInjTime:1.07
ms1PrecursorMZ:290.1601 ms1PrecursorMzAvg:290.1462

Plot of Search Spectrum

human plasma

library

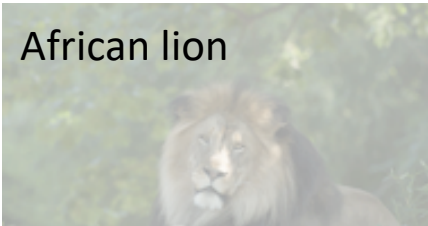
Scan:2006 RT:5.259 PrecursorScan:2004? Head to Tail MF=831 RMF=874 3-Methylglutarylcarnitine

Lib.	Score	DotProd	Rc
hr	831	874	95
hr	820	848	93
hr	819	876	95
hr	796	851	92
hr	789	858	93
hr	787	835	91
hr	777	802	88
hr	773	844	91
hr	770	798	90

Plot of Hit

Name: 3-Methylglutarylcarnitine
Formula: C₁₃H₂₃NO₆
MW: 289 Exact Mass: 289.152538 CAS#: 102673-95-0 NIST#: 1475080 ID#: 273102 DB: hr_msms_nist2020_v42
Other DBs: None
Compound ID: 0
Compound Hash:
Comment: NIST Mass Spectrometry Data Center
Notes: Consensus spectrum; Nreps=15/15; Mz_diff=1.4ppm; Vial_ID=12399; Metabolite_2016_11_29_ID=31280; micromol/L in water/acetonitrile/formic acid (50/50/1)

Applying Data Processing Pipeline in Building Glycan Library:



- Milk samples of human, bovine and other mammals
- UHPLC with HCD and IT-FT/ion trap with FTMS
- Identified glycans by searching the library with hybrid search
- Used the data processing pipeline to generate and annotate consensus spectra
- 2,605 positive and negative ion spectra of 219 oligosaccharides

- MS Interpreter 3.4
- NIST MS Search 2.4 (hybrid search included)
- NIST MS PepSearch (batch mode search of MS Search, hybrid search included)
- NIST Peptide Tandem Mass Spectral Libraries
(> 4M spectra)

Summary

- We developed a data processing pipeline to optimize data analysis for extending the NIST Tandem Mass Spectral Library 2020 with **31K compounds and 1.3M spectra** including **6K human metabolites, 8K plant metabolites, 2K drugs, 1K pesticides** etc.
- This library was also extended with spectra of
 - ❖ High resolution MSⁿ
 - ❖ High resolution Q-TOF mass spectra with APCI for extractable and leachable compounds
 - ❖ High resolution HCD, ion trap spectra of glycolipids
- The data processing pipeline can be used for building-your-own libraries.

Acknowledgements



Pedatsur Neta
Yuxue Liang
Connie A. Remoroza
Yamil Simón-Manso
Kelly H. Telu

Yuri Mirokhin
Dmitrii Tchekhovskoi
Oleg Toropov
Alexey Mayorov
Tytus D. Mak

Lewis Geer
Sanford Markey
William E. Wallace
Stephen E. Stein