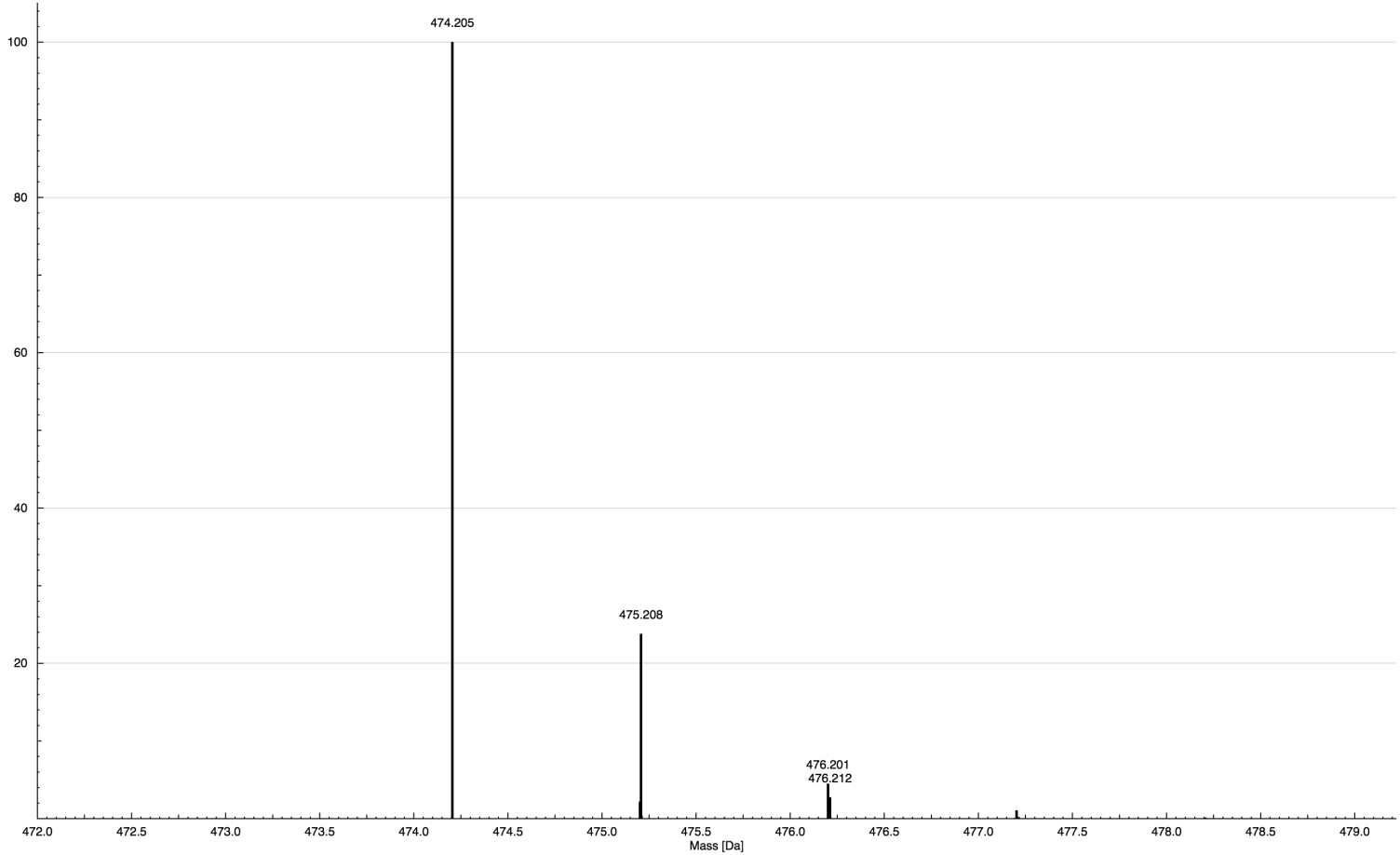
Exploring molecular formula from PubChem

Service of cheminformatics
michael.zasso@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

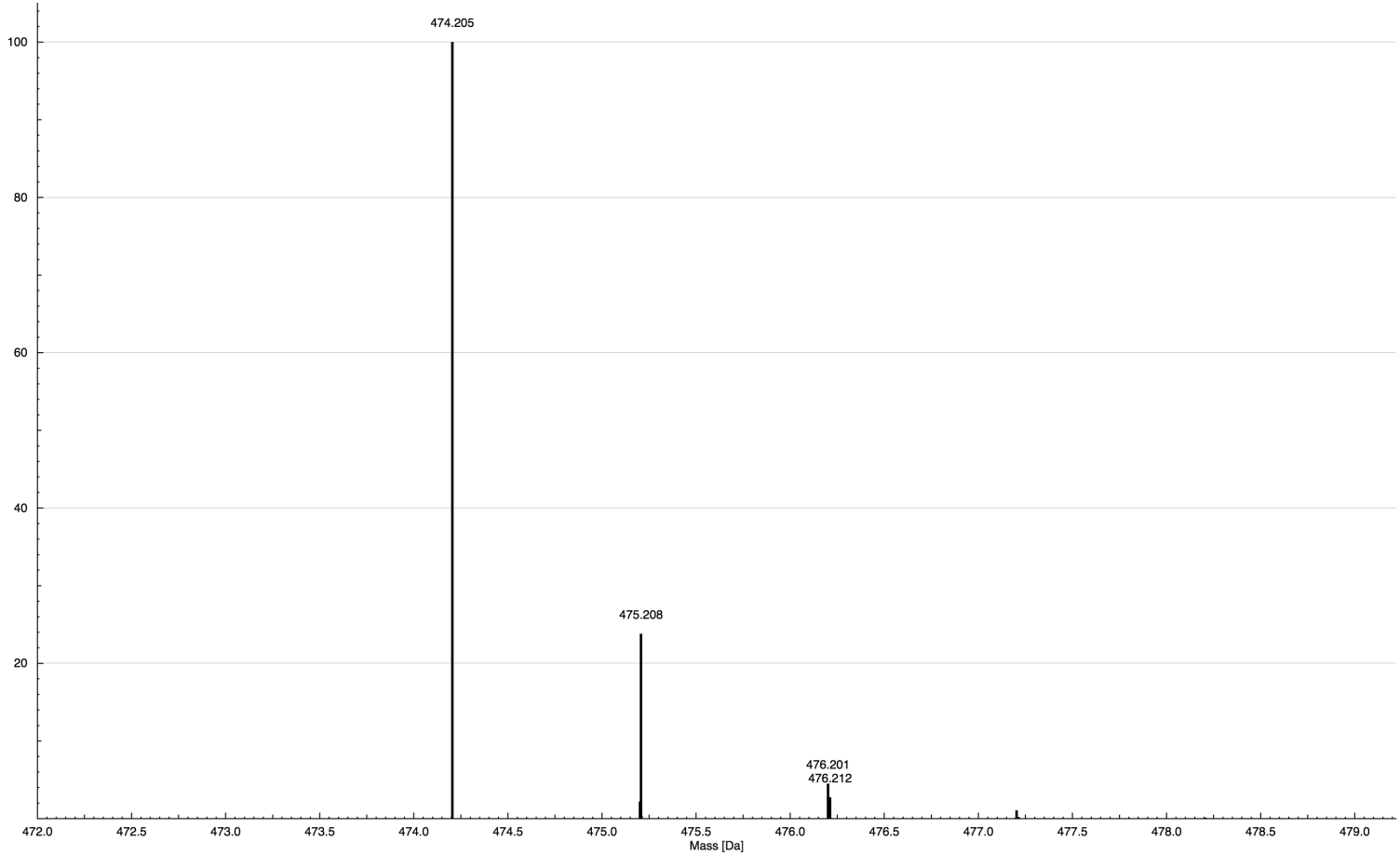
Molecular formula ?



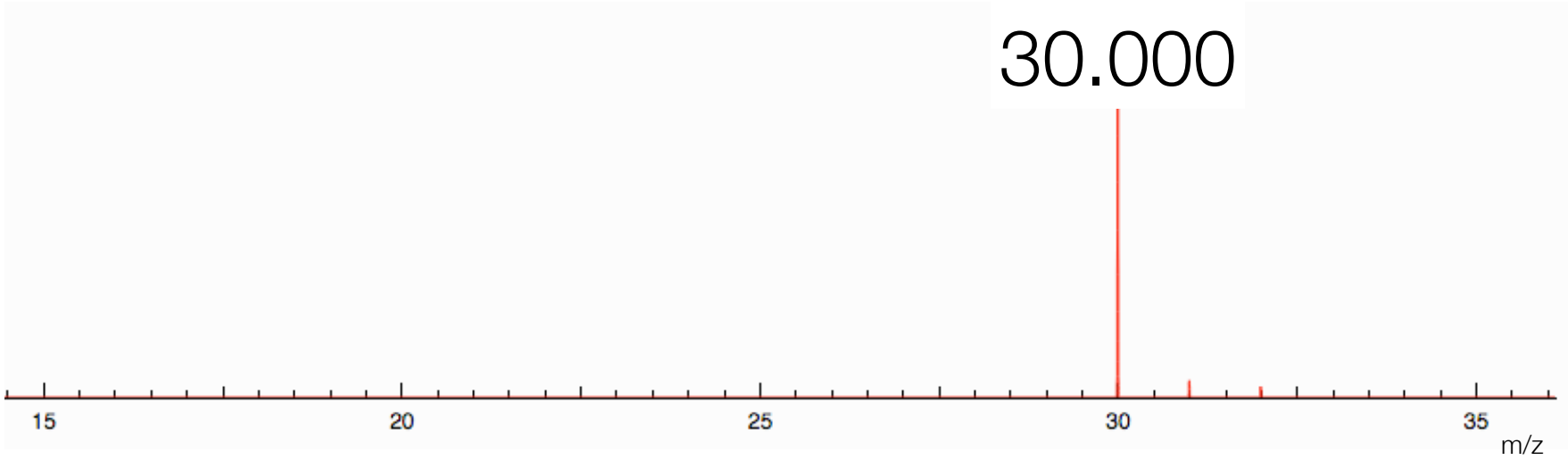
Two pieces of information

- Measured accurate mass: **474.20531 Da**

- Relative isotopic distribution: **100 / 27 / 9**



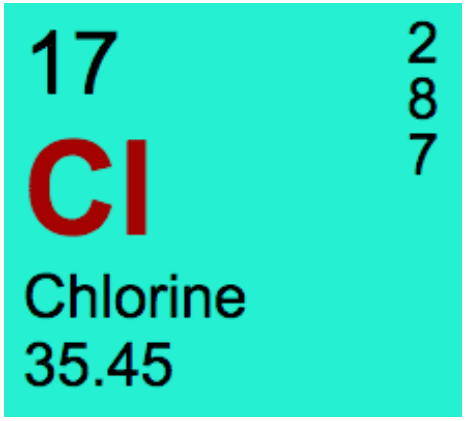
Monoisotopic mass - most abundant isotopes



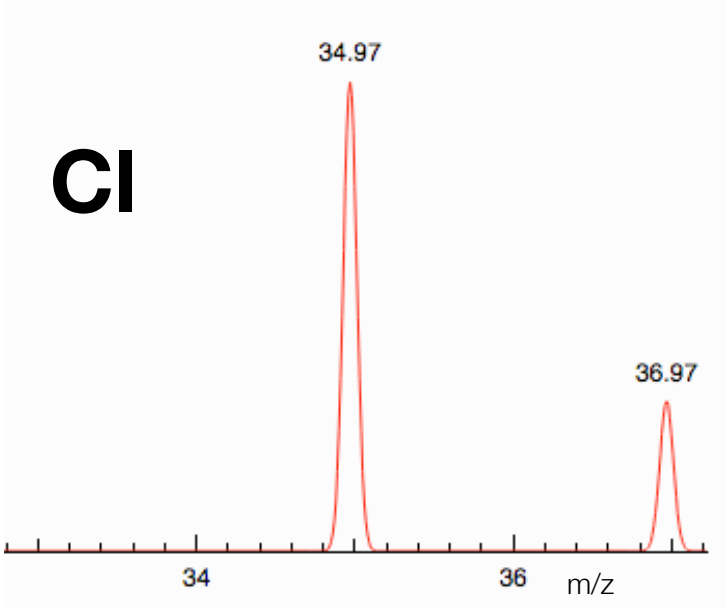
isotope	%	mass (Da)
^{1}H	99.99	1.0078
^{12}C	98.93	12.0000
^{14}N	99.64	14.0030
^{16}O	99.76	15.9949

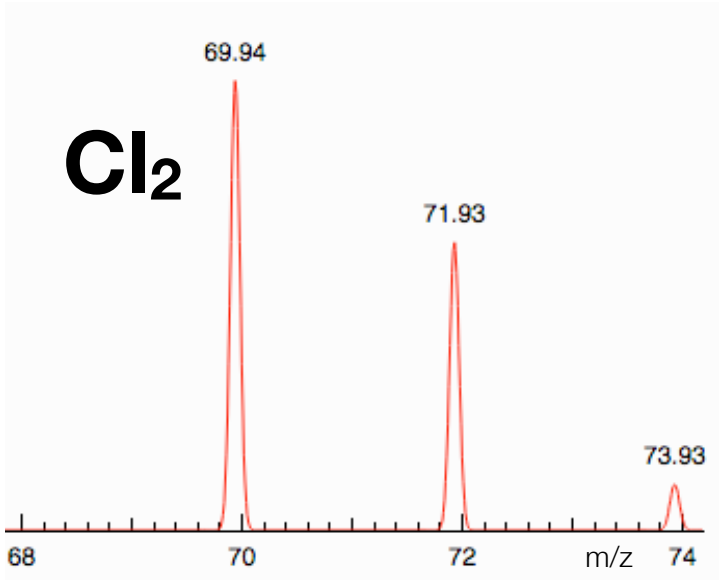
MF	mass (Da)	ppm
NO	29.9979	-2.011
CH_2O	30.0105	10.565
H_2N_2	30.0217	21.798
CH_4N	30.0343	34.374
C_2H_6	30.0469	46.95

Isotopic distribution



Isotope	mass (Da)	Abundance
^{35}Cl	34.97	75.8%
^{37}Cl	36.97	24.2%





Extract the most of our instrument

- Numbers given by the instrument are not absolute
    - Precision (ppm)
    - Error on relative abundance (isotopic distribution)

Let’s get back to our unknown product

- Measured accurate mass: **474.20531 Da**
- Precision of instrument: 5 ppm $\frac{474.20531 \times 5}{10^6} = 0.00237$
- Mass range: 474.20294 - 474.20768
- Elements range:

	C	H	N	O	P	S	F	Cl	Br
Min	1	0	0	0	0	0	0	0	0
Max	100	202	50	50	25	25	25	25	25

ChemCalc

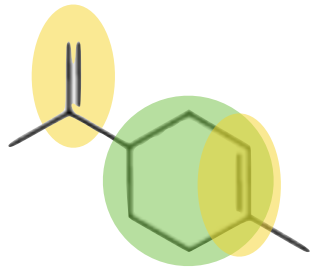
- From monoisotopic mass to list of possible molecular formulas
- Link
  - 10’798 candidates
  - Lots of stupid results

ChemCalc - filter results

- Degree of unsaturation to the rescue!

$$DoU = \frac{2 + 2C + N + P - X - H}{2}$$

- Link
  - Must be positive or 0
    - 619 candidates
  - Must be integer
    - No radical
    - 340 candidates



Limonene

$C_{10}H_{16}$

$$DoU = \frac{2 + 2 \cdot 10 - 16}{2} = \frac{6}{2} = 3$$

PUBCHEM

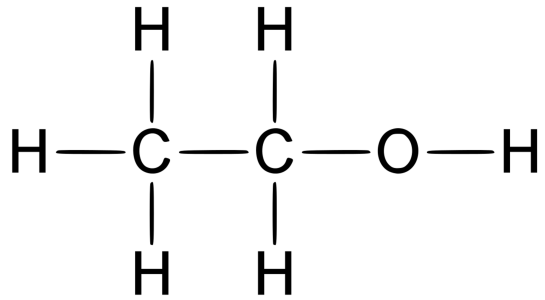
- Enormous free database: 91’679’191 structures
- MongoDB copy
  - Chemical structure
  - Molecular and exact mass
  - MF / atom information
- Weekly sync from FTP
- Hosted on GitHub: https://github.com/cheminfo/node-pubchem/

PUBCHEM

- A lot of mistakes / irrelevant data
    - U251 / Multiple fragments
- Keep unique molecular formulas:
    - One fragment
    - No charge
    - Only allowed elements: C, H, N, O, P, S, F, Cl, Br
    - Mass range: 100.5 - 1000.5 Da
    - 1’593’553 unique MFs

PUBCHEM - Statistics

- Mass range: 100.5 - 1000.5 Da
- Slot: 25 Da
- Element ratios
  - C/H, C/N, C/O, C/S, C/P, C/FClBr
  - O/P, O/S
  - CCNP/HFClBr (unsaturation)
- Log_2 scale: -8 to +8
- Results



Ratio			log_2
C/H	2/6	0.333	-1.58
C/O	2/1	2	1
CCNP/ HFClBr	4/6	0.666	-0.58

Sorting the MF based on PubChem statistics

- Simulation with 1000 random formulas from PubChem
- 450-500 Da
- Compute a score with the ratios

$$\sqrt[n]{\prod_{0}^{n} 0.8^{\frac{|ratio-mean|}{stdev}}}$$

Sorting the MF based on PubChem statistics

- Simulation with 1000 random formulas from PubChem
- 450-500 Da
- Compute a score with the ratios
- Take median of position of the correct MF in the list of candidates

ppm	median position	
	w/o ratio score	with ratio score
0.1	19	1
1	193	13
5	962	69
10	1915	138
50	9609	697
100	19457	1396

Apply to our unknown product

- Measured accurate mass: **474.20531 Da**

- Demo