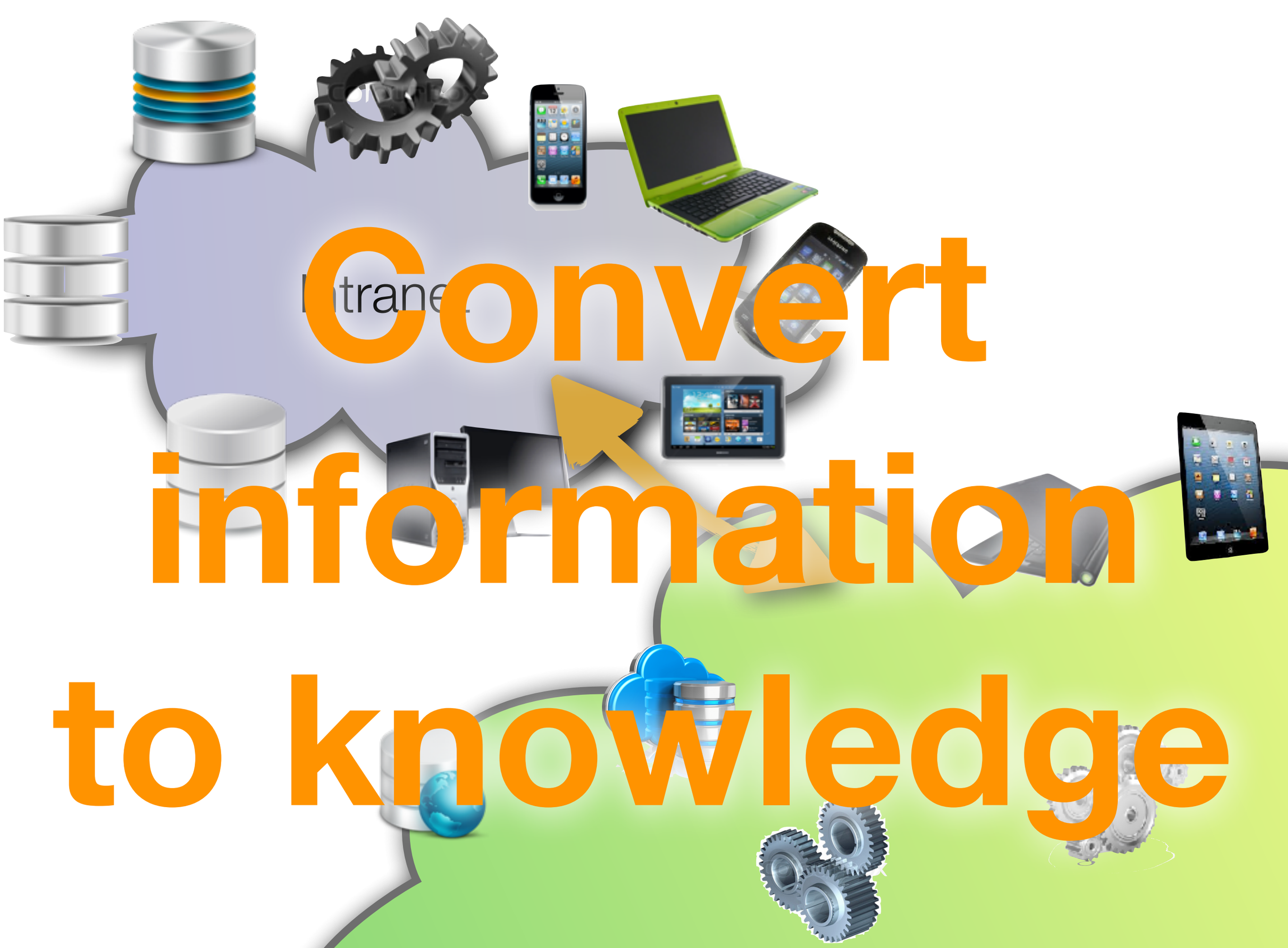
New challenges in the management of chemical information

Service of cheminformatics
luc.patiny@epfl.ch


Intranet
Convert
information
to knowledge

bp

LD_{50}

pKa

mp

IC_{50}

information

NMR

IR

logP

ms

Xray

QSAR

physical property

spectra

bond length

toxicity

fragmentation

knowledge

Data

Some available databases

ChemSpider

ChemExper

eMolecules

Zinc

http://depth-first.com/articles/2007/01/24/thirty-two-free-chemistry-databases/

Specialized databases

CSD (Cambridge structural database)

KNApSAcK

AIST: Spectral Database

NIST Chemistry web book

Downloadable databases

Crystallography Open Database

DrugBank

Pubchem

GDB-17

ChemBL

PDB

NMR shift DB

KEGG

Mass bank

Binding DB

Wikipedia

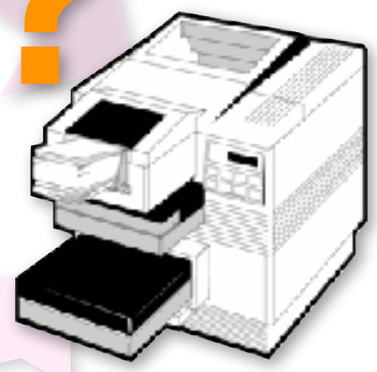
OChem : Online chemical database

in 5 years ?

in 10 years ?

FTP
USB

File Server



Print
PDF

in 20 years ?

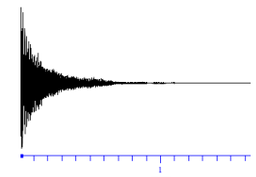
Install proprietary software
for each data format

Store “correctly” the chemical information

- Perennial and reprocessable !
- Only use **ALIVE** published **DESCRIBED** format
  - no proprietary format
    - no PDF, word, excel, …
    - chemical structures: molfile
    - spectra: JCAMP-DX

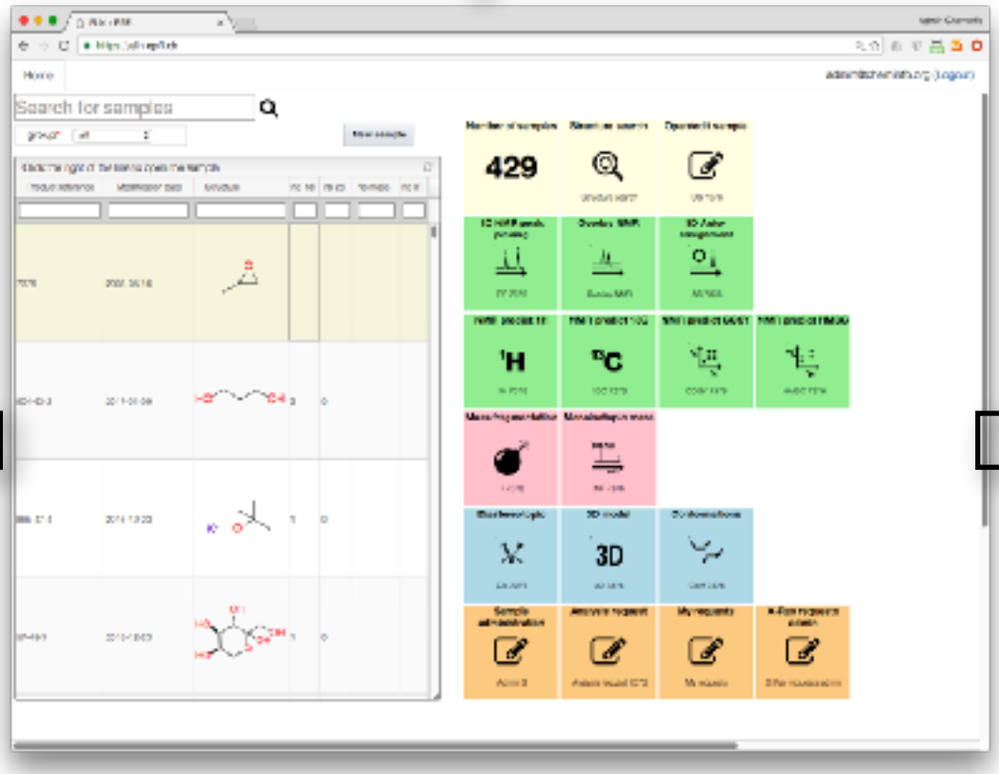
Store “correctly” the chemical information

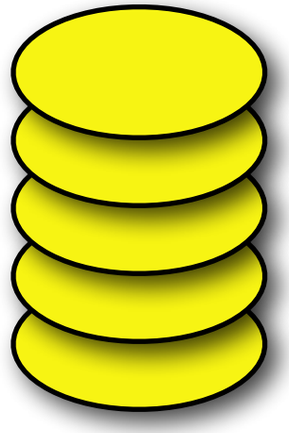
- bp : sign - low - high - pressure
  - 100-120 (20 torr) ; 10 (15 mmHg) ; >300 ; …
- NMR : 1H NMR (CDCl3): δ = 0.84 (3 H, t, J = 7.4 Hz, CH3), 0.94 (3 H, t, J = 7.4 Hz, CH3), 1.23 [6 H, q, J = 6.9 Hz, P(O)OCH2CH3], 1.51 (4 H, m, CH2), 2.20 (1 H, sextet, J = 6.6 Hz, CH), 3.80 (3 H, s, OCH3), 4.01 [4 H, m, P(O)OCH2CH3], 4.63 (1 H, d, JH,P = 17.1 Hz, NCHP), 6.88 (2 H, d, J = 8.6 Hz, CH).
- Traceability, flexibility : noSQL database



DB

Image analysis
Mass analysis
Similarity search
Predictions
…

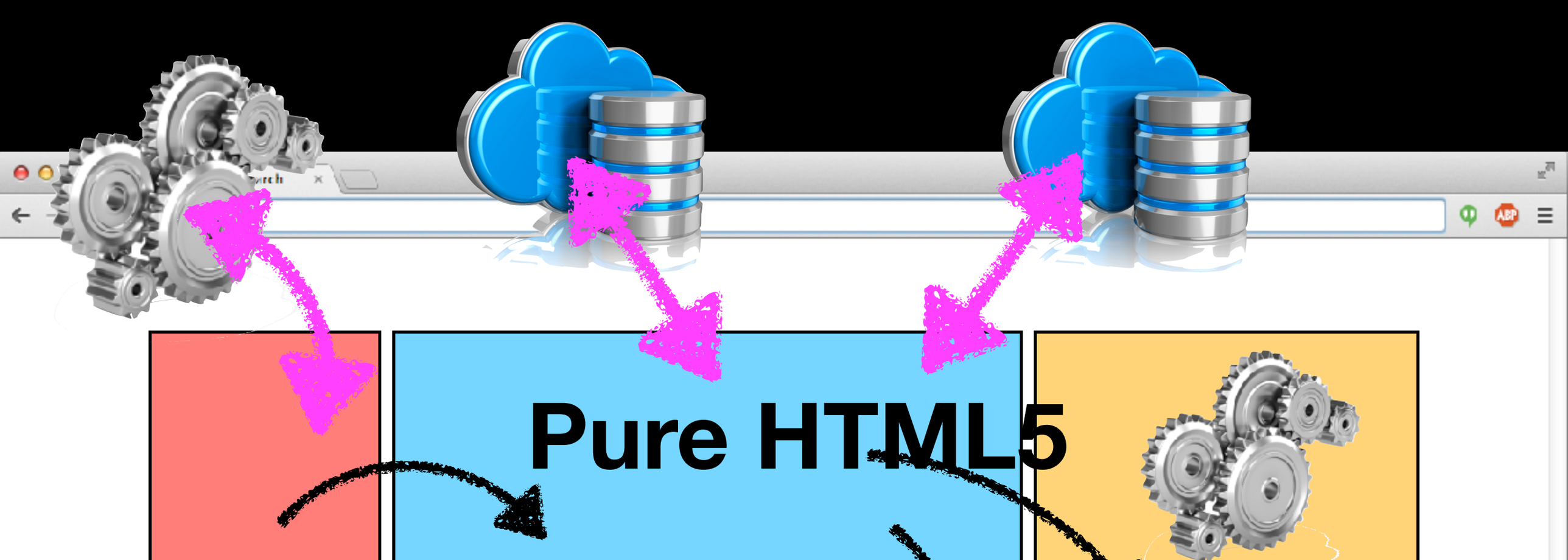




open data

Retrieving and processing data

The need for a flexible tool



Pure HTML5

- MacOS, windows, linux
- No installation
- open-source (GitHub)
- No licence fee

Knowledge

Research

Teaching

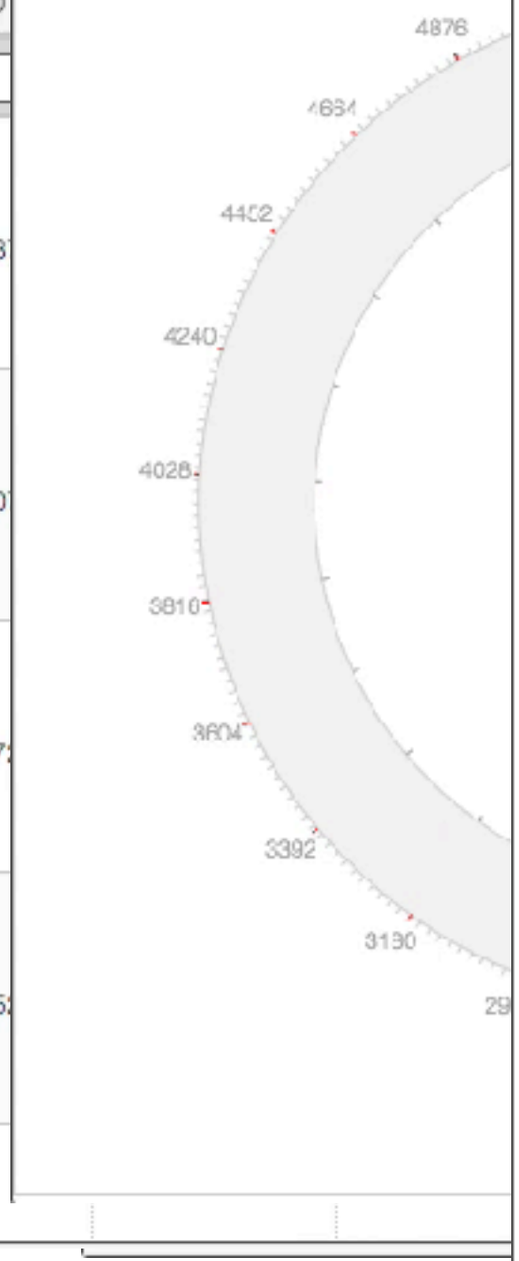
Unnamed module

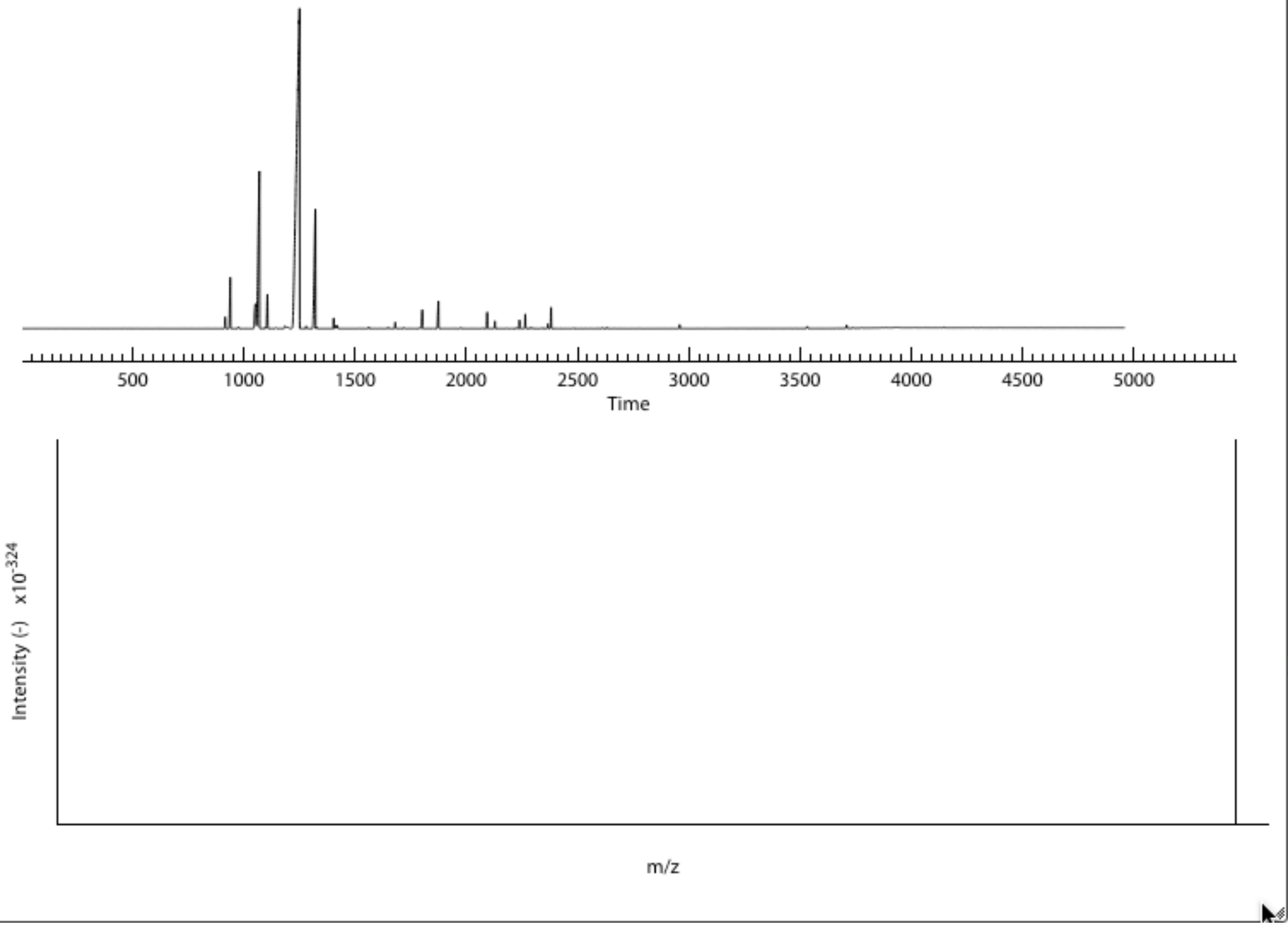
3D scatter plot

Protein sequence

Format: CODATA

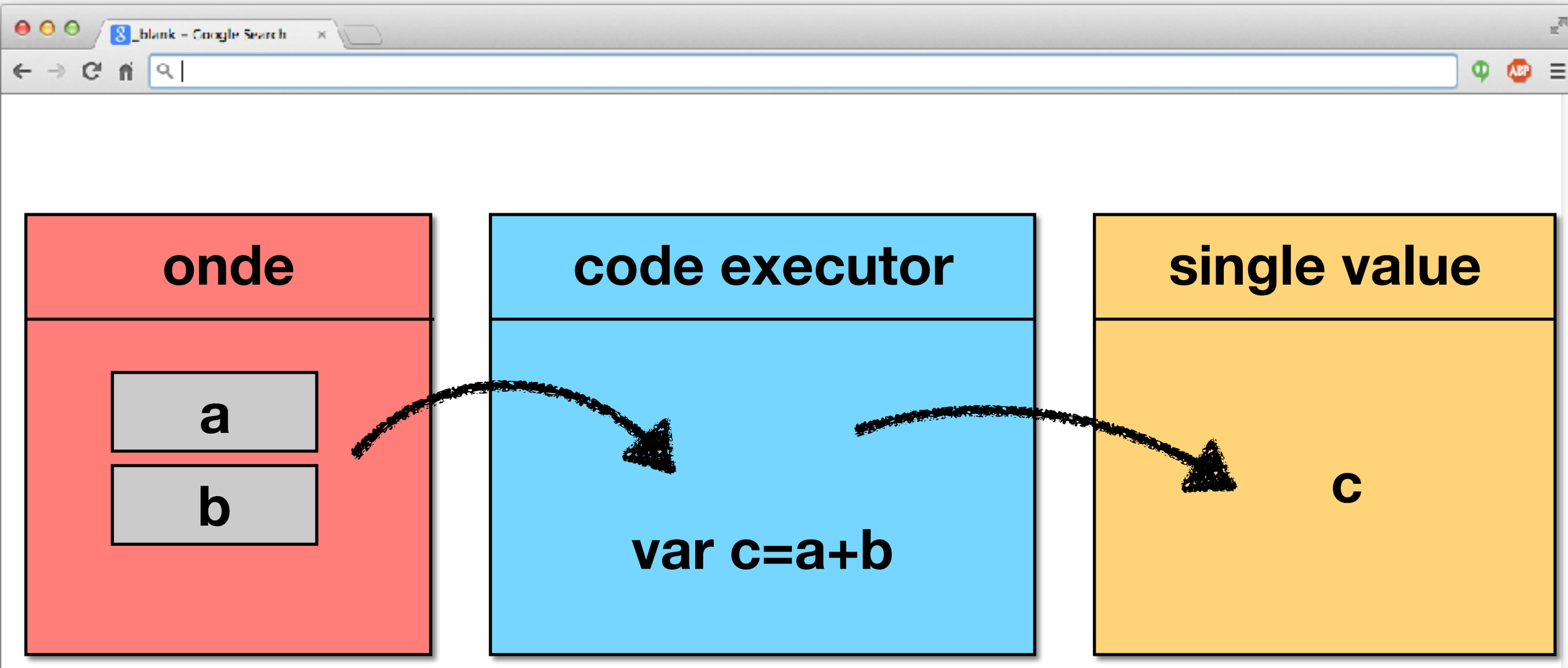
```
ENTRY            P918283
SEQUENCE
                      5         10        15        20        25        30
     1  M E T L C Q R L N V C Q D K I L T H Y E N D S T D L R D H I D Y W K H
                                E M G F K H I N H Q V V P T L A V S K N K A L
                                Y N S Q Y S N E K W T L Q D V S L E V Y L T A
```





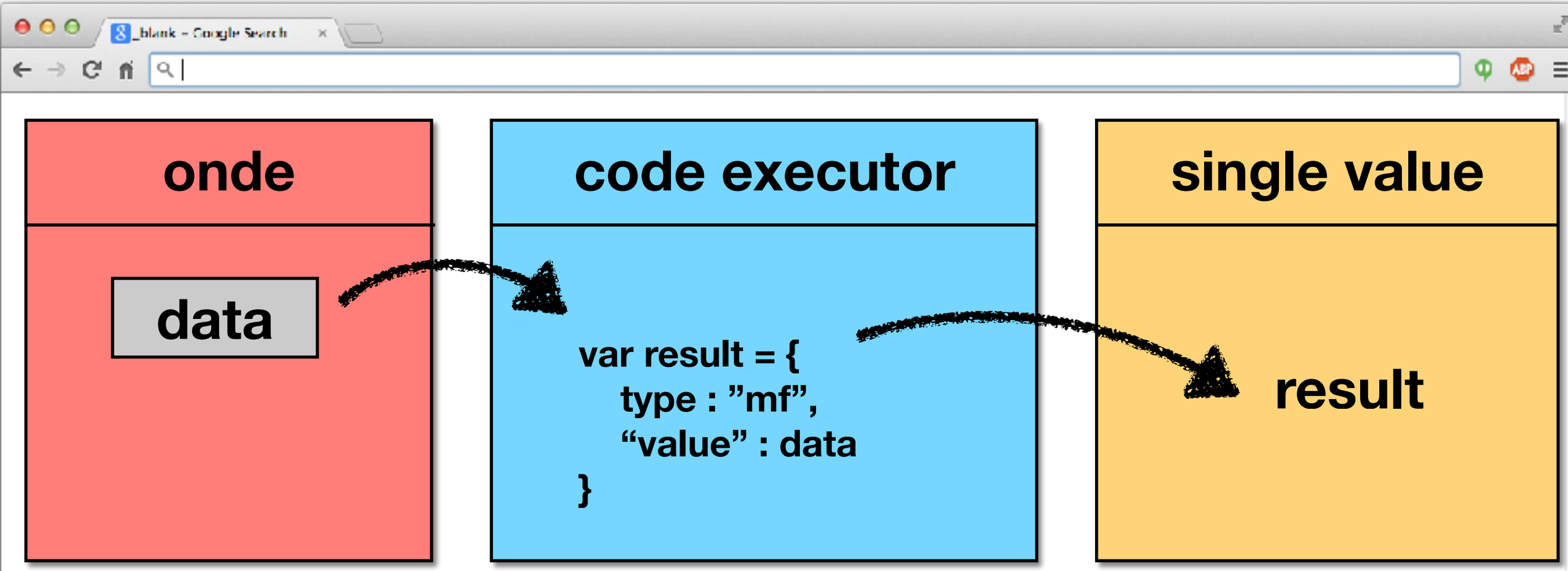
How does it works ?

Basic example : calculate a sum of 2 values



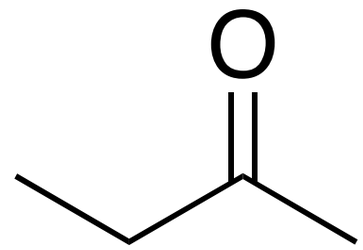
Demo

Chemical types - MF and SMILES



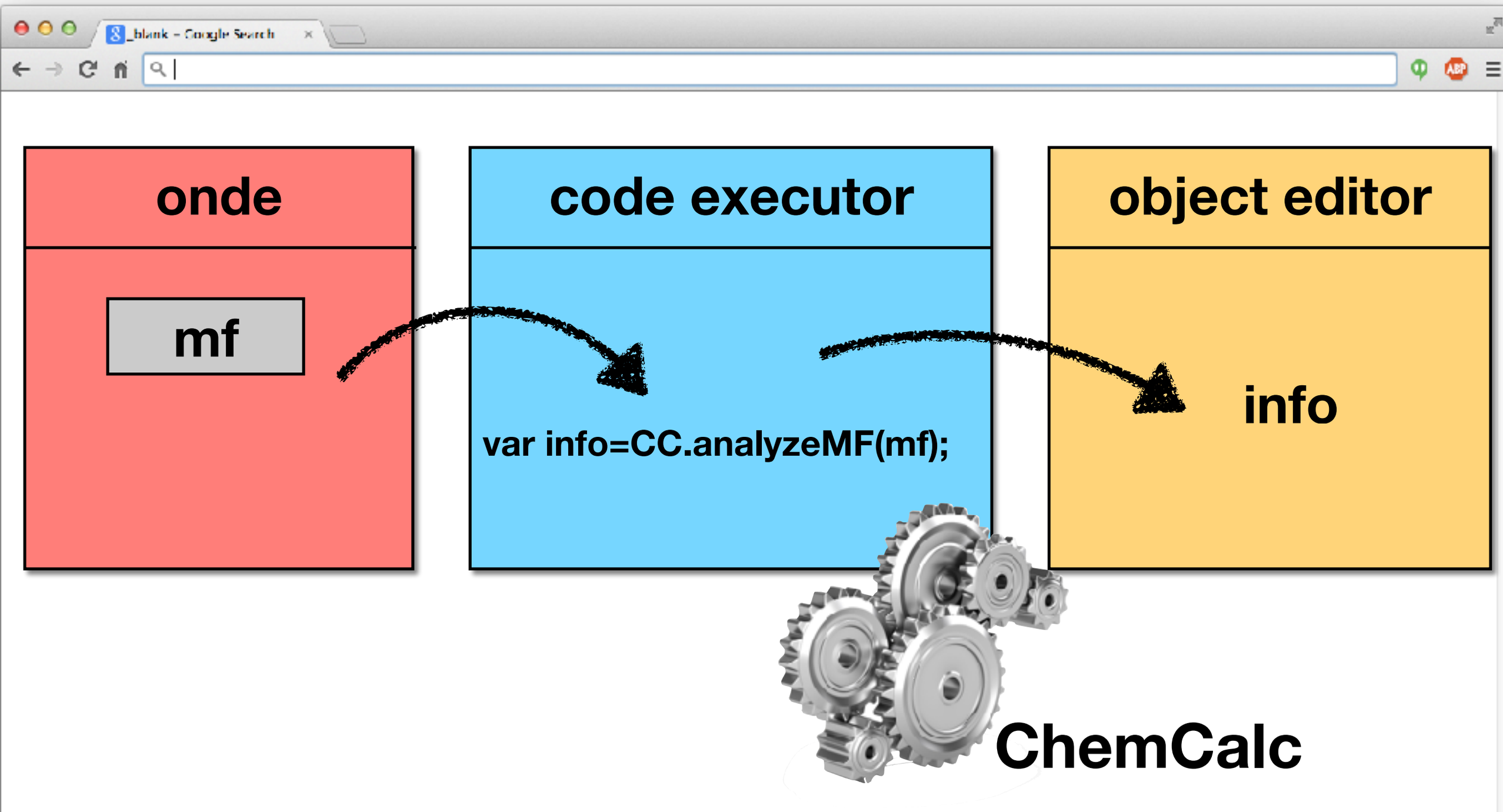
C6H15N → **$C_6H_{15}N$**

CCC(=O)C →



Demo

Using external libraries



Demo

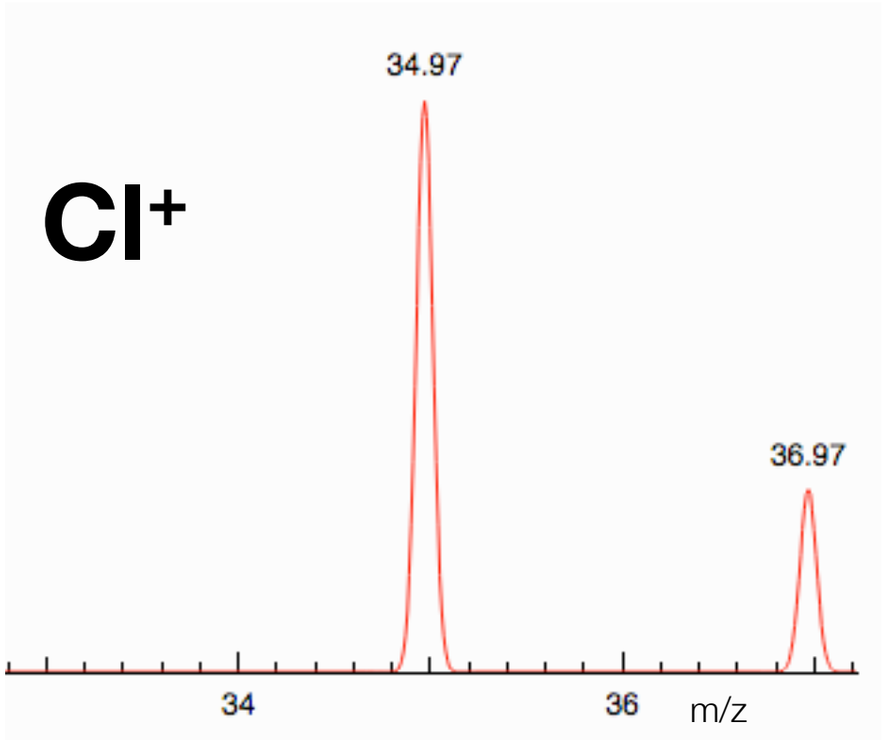
Molecular formula and mass

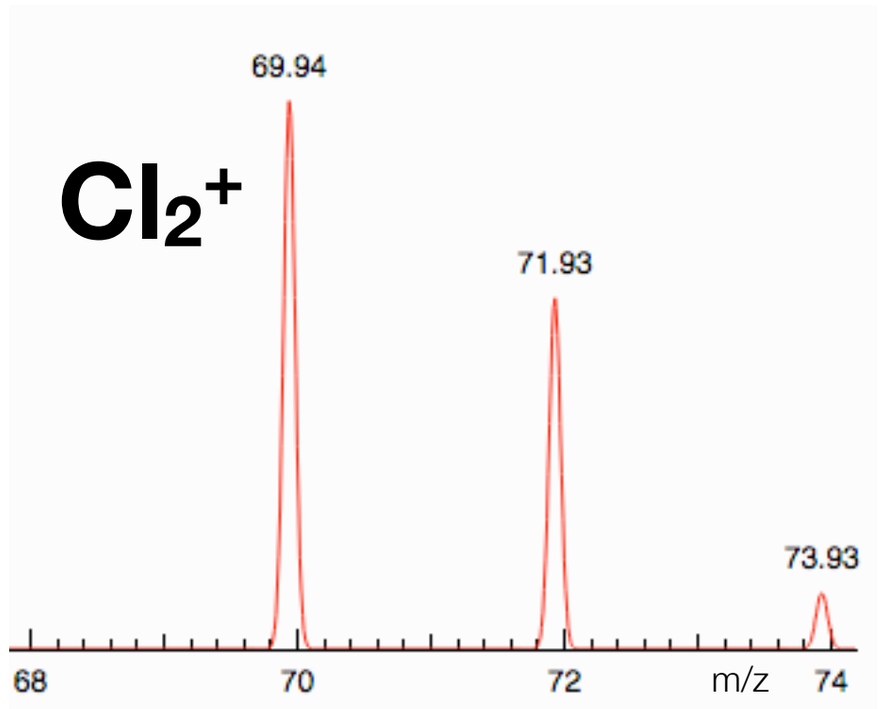
http://ms.cheminfo.org
http://www.chemcalc.org

Isotopic distribution

17
2
8
7
Cl
Chlorine
35.45

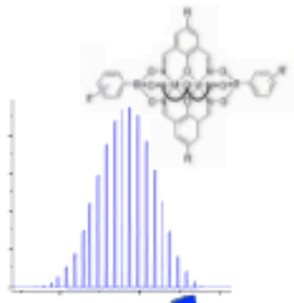
Isotope	mass (Da)	Abundance
^{35}Cl	34.97	75.8%
^{37}Cl	36.97	24.2%





ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Please cite US

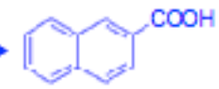


Calculations

Calculate your theoretical spectrum from a molecular formula

→ C_6H_6

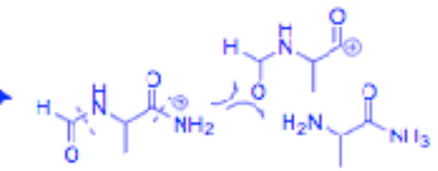
From a chemical structure

→ 

Find a MF from a Monoisotopic Mass

→ 78.047 → C_6H_6

MS/MS fragment calculation

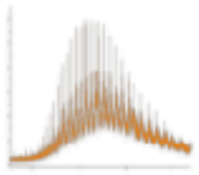
→ 

Peptides

Predict your CID/ETD spectra

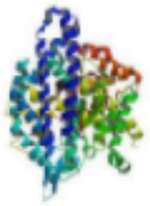
→ Ala-Gly-Ala-Trp

Match your Experimental data

→ 

Try our teaching tools

isic SSML



Match to list

Drag&Drop and match to contaminants

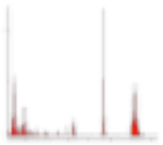
→ ***Easycont***

Similarity to spectrum

Extend your search using all elements

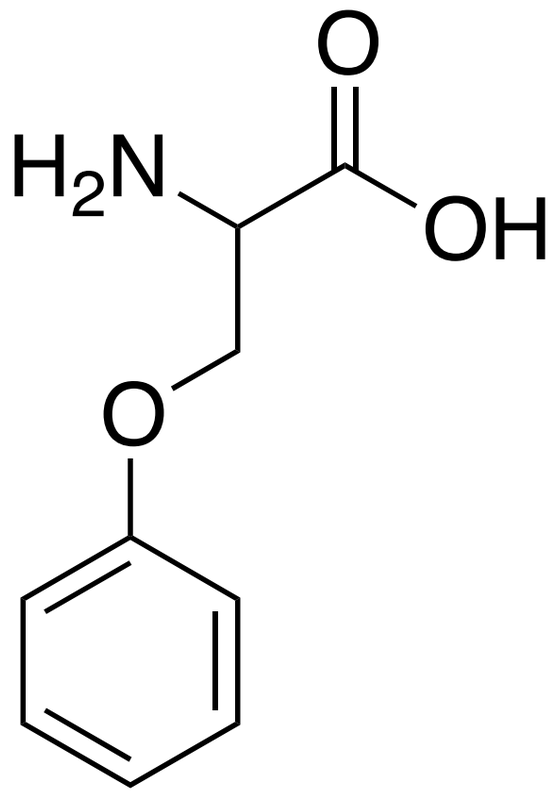
→ 

Drag&Drop your Experimental spectrum and Identify your molecular Formula

→ 

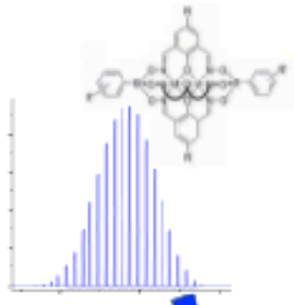
ChemCalc

- Use of groups (Ph, Ala, Gly, Fmoc)
  - HAlaGlyProOH
- Many components (separated by “.”)
  - C_{100} . C_{110} . C_{120}
- Specific isotope
  - $[^{13}C]_{100}$. $[^{13}C]_{50}[^{12}C]_{50}$
- Modification of the isotopic abundance
  - $C\{50,50\}_{10}C_{10}$
- Specification of the charge
  - $HAla_{10}OH^{+}$. $HAla_{10}OH^{++}$. $HAla_{10}OH\ (H^{+})2$
- Modification of a molecular formula
  - $HSer(H_{-1}Ph)OH$



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Please cite US

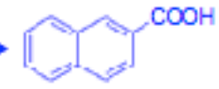


Calculations

Calculate your theoretical spectrum from a molecular formula

→ C_6H_6

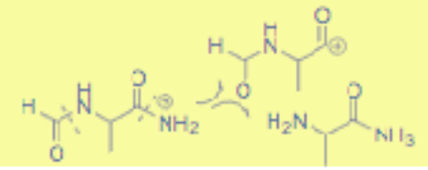
From a chemical structure

→ 

Find a MF from a Monoisotopic Mass

→ 78.047 → C_6H_6

MS/MS fragment calculation

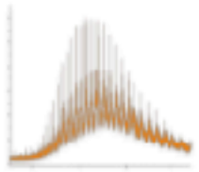
→ 

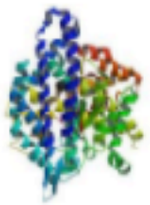
Peptides

Predict your CID/ETD spectra

→ Ala-Gly-Ala-Trp

Match your Experimental data

→ 



Try our teaching tools

Match to list

Drag&Drop and match to contaminants

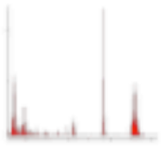
→ *Easycont*

Similarity to spectrum

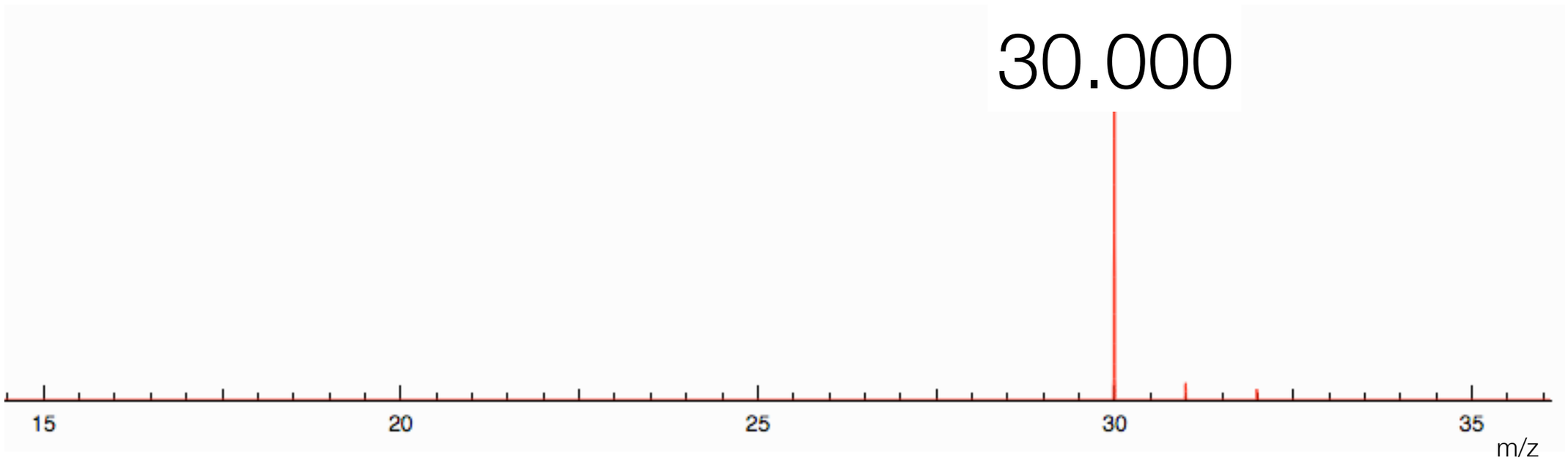
Extend your search using all elements

→ 

Drag&Drop your Experimental spectrum and Identify your molecular Formula

→ 

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

Statistics of pubchem

Pubchem SDF

Search by
EM and MF

Visualizer

FTP download
Calculate MF
MF stats
Exact mass

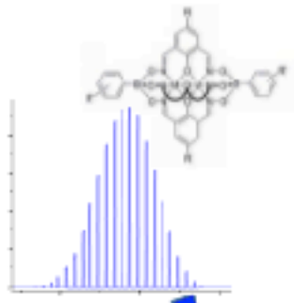
MongoDB

Webservice

https://github.com/cheminfo/node-pubchem

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Please cite US

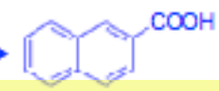


Calculations

Calculate your theoretical spectrum from a molecular formula

→ C_6H_6

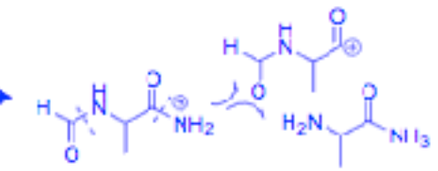
From a chemical structure

→ 

Find a MF from a Monoisotopic Mass

→ 78.047 → C_6H_6

MS/MS fragment calculation

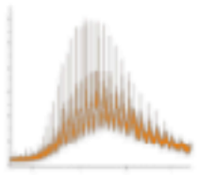
→ 

Peptides

Predict your CID/ETD spectra

→ Ala-Gly-Ala-Trp

Match your Experimental data

→ 

Try our teaching tools

Match to list

Drag&Drop and match to contaminants

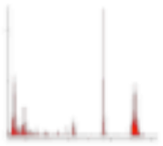
→ ***Easycont***

Similarity to spectrum

Extend your search using all elements

→ 

Drag&Drop your Experimental spectrum and Identify your molecular Formula

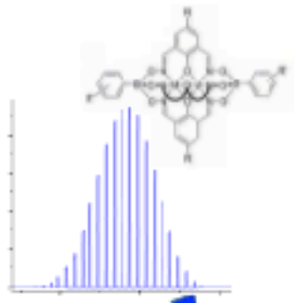
→ 

Contaminants / Custom database

- Two possible problems :
    - Mass spectra may contain contaminants
    - You want to look for a list of products
- A solution :
    - Edit a list of molecular formula in a spreadsheet
    - Match directly the list to your spectrum

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Please cite US

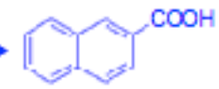


Calculations

Calculate your theoretical spectrum from a molecular formula

→ C_6H_6

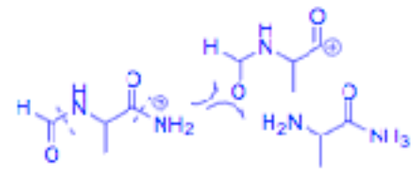
From a chemical structure

→ 

Find a MF from a Monoisotopic Mass

→ 78.047 → C_6H_6

MS/MS fragment calculation

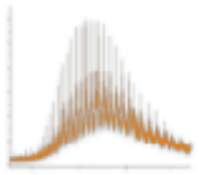
→ 

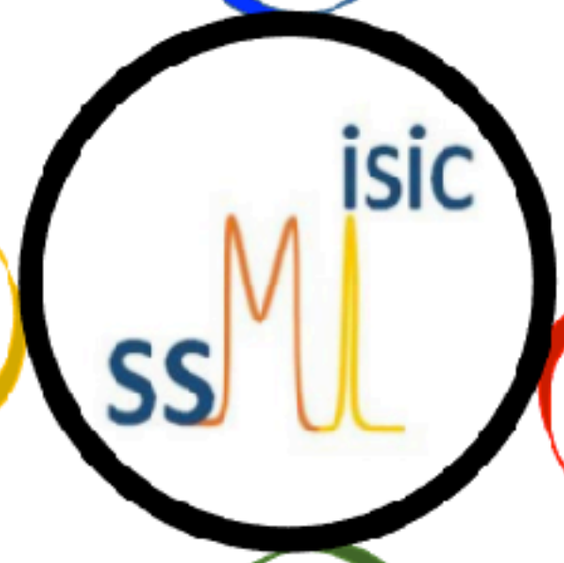
Peptides

Predict your CID/ETD spectra

→ Ala-Gly-Ala-Trp

Match your Experimental data

→ 



Try our teaching tools

Match to list

Drag&Drop and match to contaminants

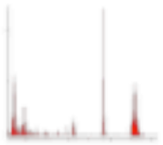
→ ***Easycont***

Similarity to spectrum

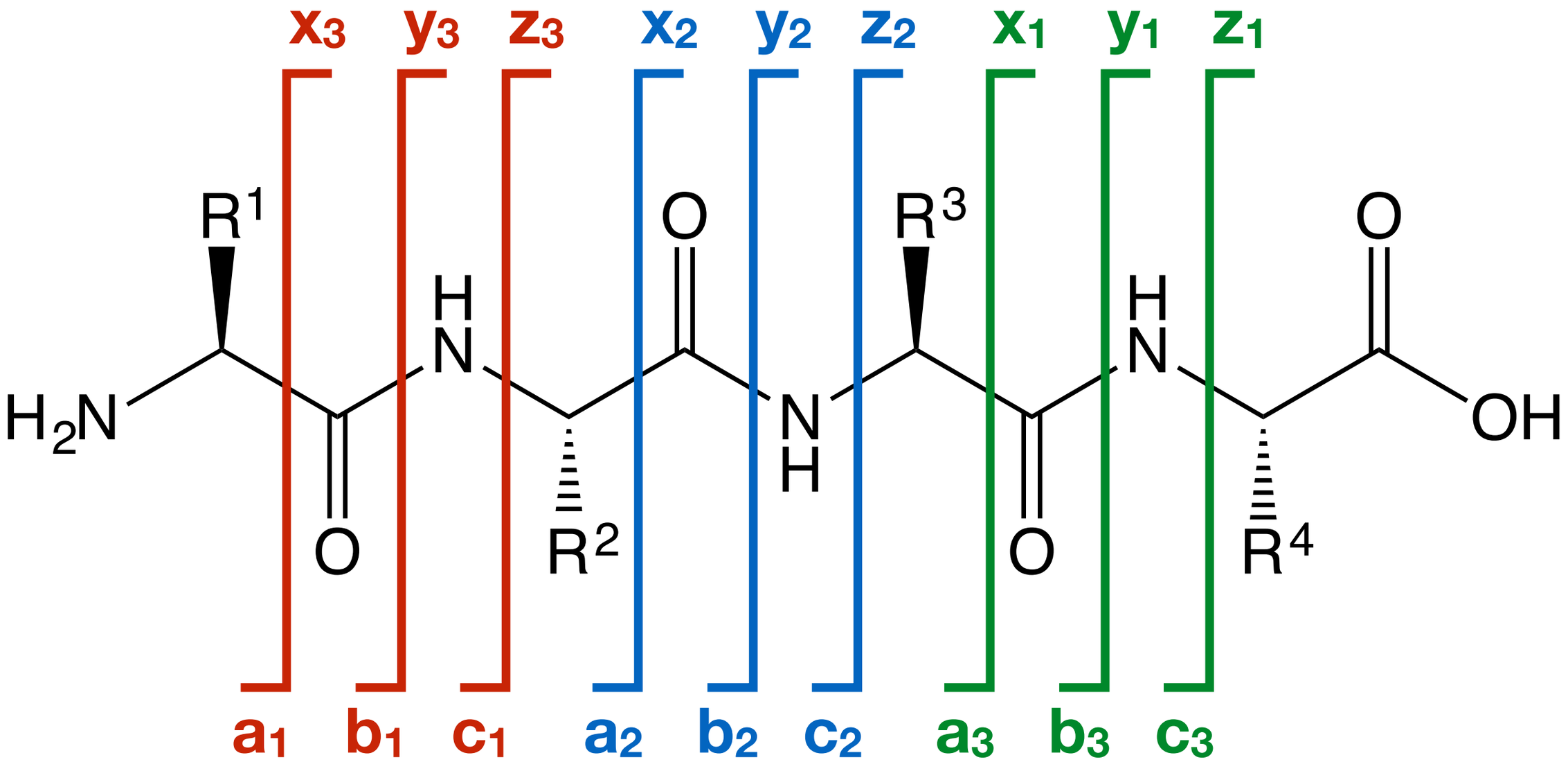
Extend your search using all elements

→ 

Drag&Drop your Experimental spectrum and Identify your molecular Formula

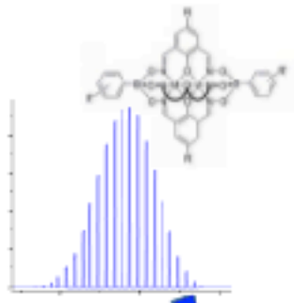
→ 

Peptide fragmentation



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Please cite US

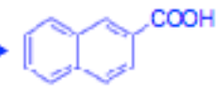


Calculations

Calculate your theoretical spectrum from a molecular formula

→ C_6H_6

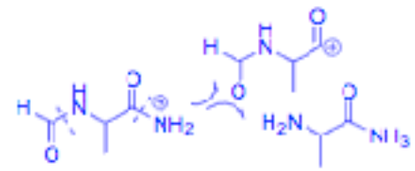
From a chemical structure

→ 

Find a MF from a Monoisotopic Mass

→ 78.047 → C_6H_6

MS/MS fragment calculation

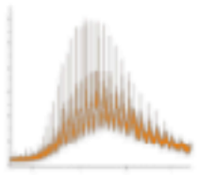
→ 

Peptides

Predict your CID/ETD spectra

→ Ala-Gly-Ala-Trp

Match your Experimental data

→ 

Try our teaching tools

Match to list

Drag&Drop and match to contaminants

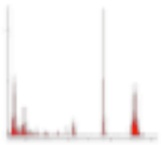
→ ***Easycont***

Similarity to spectrum

Extend your search using all elements

→ 

Drag&Drop your Experimental spectrum and Identify your molecular Formula

→ 

Cheminformatics

From 2D drawing to conformations

Draw

2D molfile

Display

conformers

Server

moloc
2D to 3D

Demo

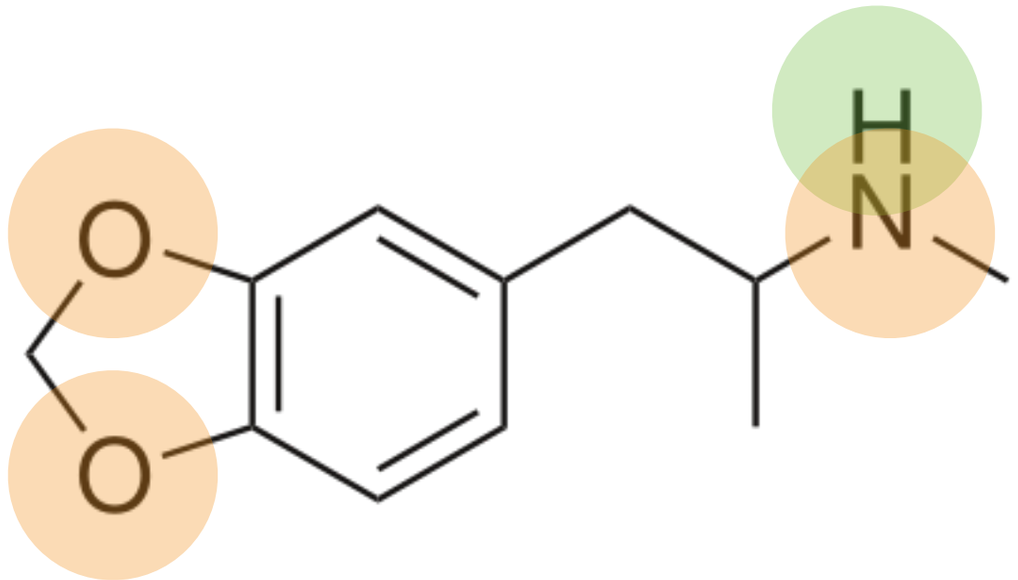
Lipinski rule of 5

✔ H bond donors ≤ 5

✔ H bond acceptors ≤ 10

✔ log P ≤ 5 **1.8**

✔ mw ≤ 500 **193**



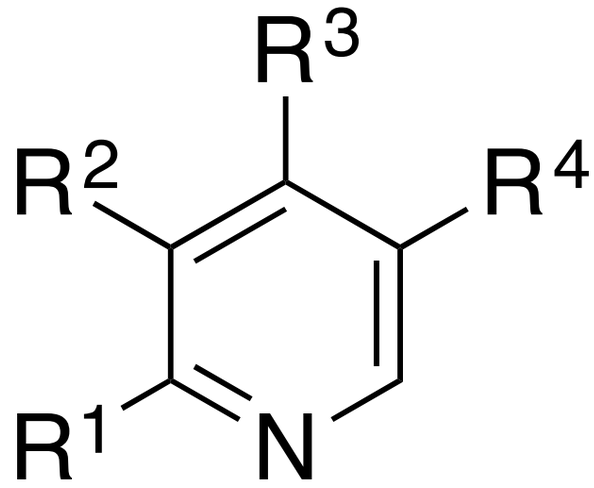
https://github.com/cheminfo/openchemlib-js

Demo 1

Demo 2

Virtual combinatorial library

- The problem
  - Search for products with specific properties :
    - logP
    - mw
    - H donor
    - H acceptor



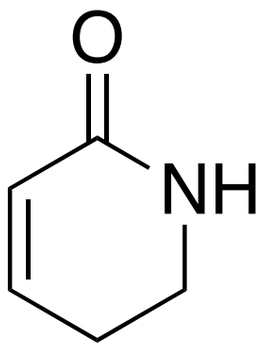
- A solution
  - Generate all possible molecules based on a template and fragments
  - Predict the properties
  - Filter using parallel coordinates

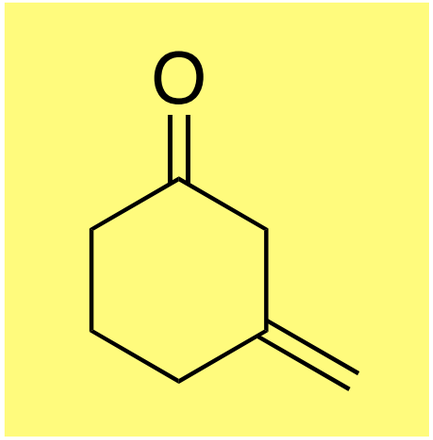
Vanderveen, J. R., Patiny, L., Chalifoux, C. B., Jessop, M. J., & Jessop, P. G. (2015). A virtual screening approach to identifying the greenest compound for a task: application to switchable-hydrophilicity solvents. *Green Chemistry*, *17*(12), 5182-5188.

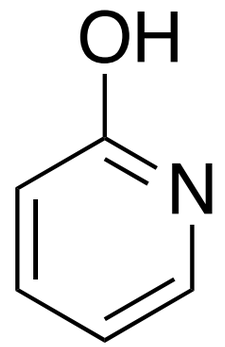
Demo

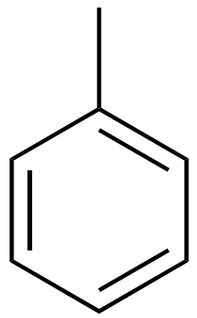
Substructure search - Wikipedia

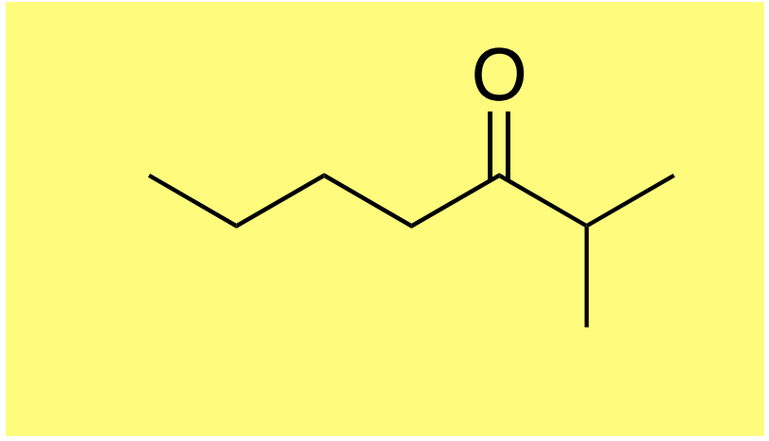
Substructure search ?

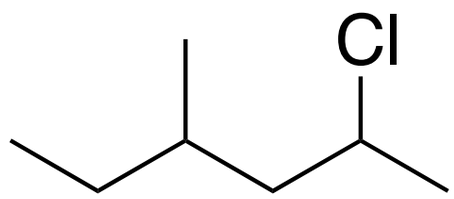


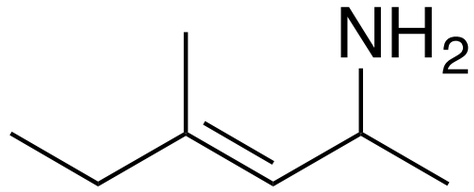


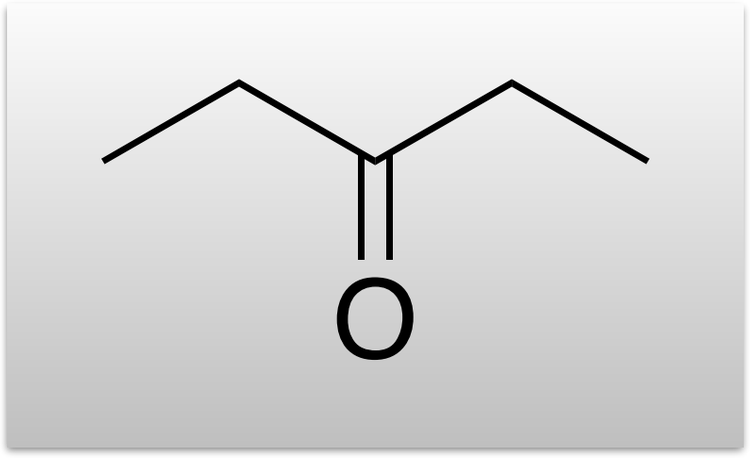












Ertl, P., Patiny, L., Sander, T., Rufener, C., & Zasso, M. (2015). Wikipedia Chemical Structure Explorer: substructure and similarity searching of molecules from Wikipedia. *Journal of cheminformatics*, *7*(1), 10.

Benzene - Wikipedia, the fr

https://en.wikipedia.org/wiki/Benzene

Luc Patiny

Create account Log in

WIKIPEDIA
The Free Encyclopedia

Article Ta

Read Edit View history

Search

Ber

From W

Th
No

Benze

C_6H_6
atom a
and hy

Benze
petrocl
([6]-an
abbrev
smell.
and cu
octane
of its m
carcinc

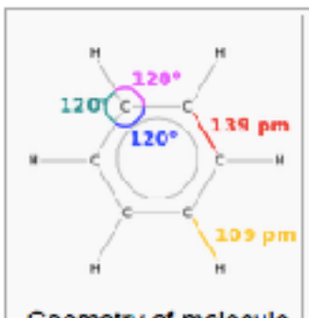
1 His
1
1
1.3 Early applications
2 Structure
3 Benzene derivatives
4 Production

s, see Benzene (disambiguation).

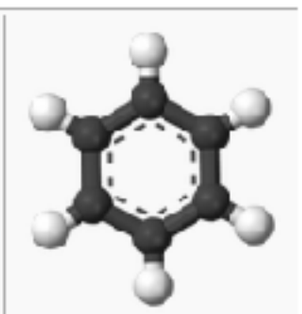
chemical formula
ring, with 1 hydrogen
ntain only carbon

most elementary
second [*n*]-annulene
. It is sometimes
le liquid with a sweet
h as ethylbenzene
because it has a high
rising a few percent
by benzene's

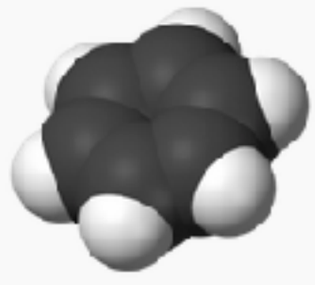
Benzene



Geometry of molecule



Ball and stick model of molecule



Benzene molecule

Names	
IUPAC name	Benzene
Systematic IUPAC name	Cyclohexa-1,3,5-triene
Other names	1,3,5-Cyclohexatriene, Benzol, Phene, Phenyl hydride
Identifiers	
CAS Registry Number	71-43-2 ✓
ChEBI	CHEBI:16716 ✓
ChEMBL	ChEMBL277500 ✓
ChemSpider	236 ✓
EC number	200-753-7
InChI	[show]
Jmol-3D images	Image
KEGG	C01407 ✓
PubChem	241
RTECS number	CY1400000
SMILES	[hide]
c1ccccc1	
INIL	J64922108F ✓
Properties	

c1ccccc1

Names	
IUPAC name	Benzene
Systematic IUPAC name	Cyclohexa-1,3,5-triene

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Every night …

Find

ChemBox
DrugBox

Extract

SMILES

Generate

JSON

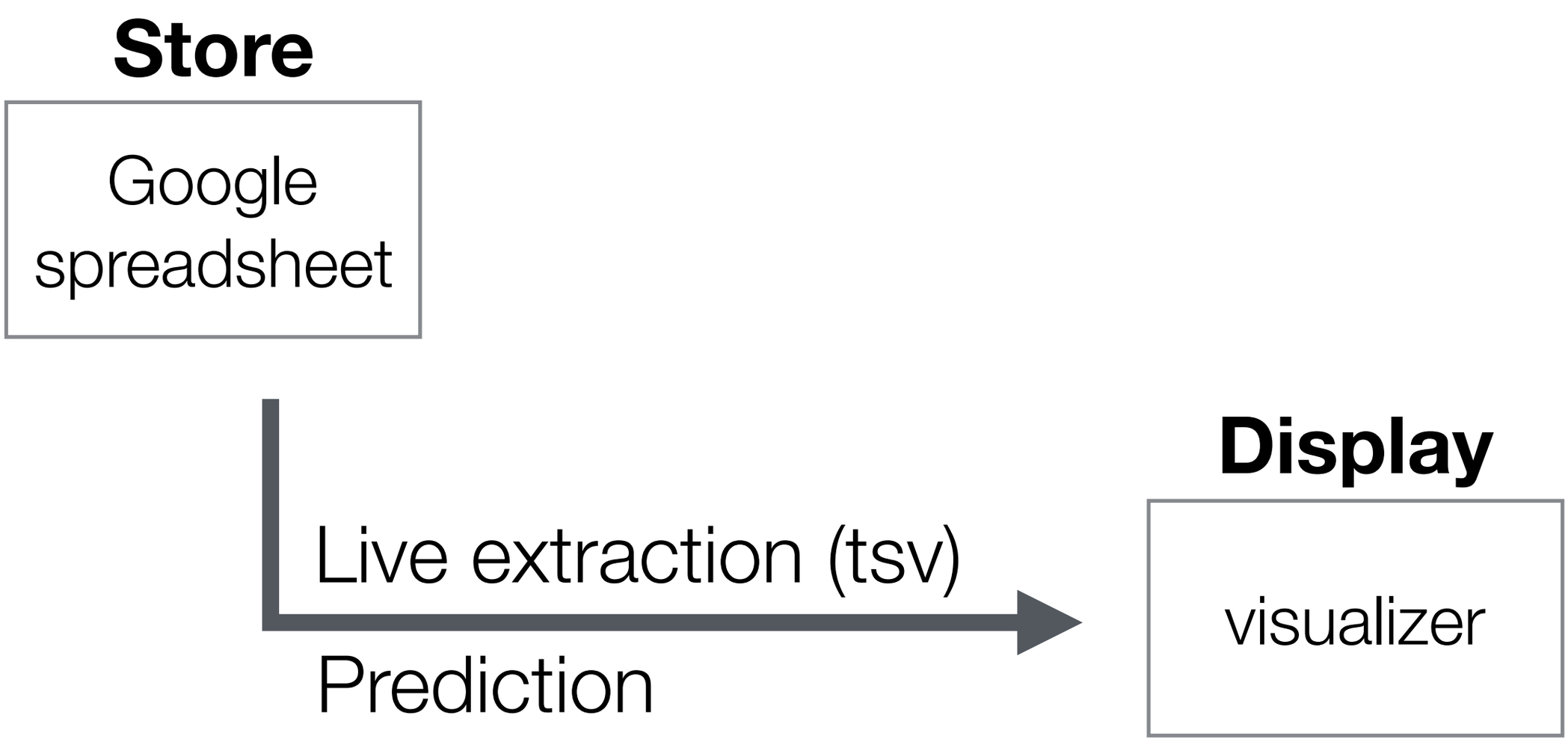
Publish

GitHub

https://github.com/cheminfo/wikipedia
http://www.cheminfo.org/wikipedia

Demo

Simple database of chemicals - open source malaria



http://malaria.cheminfo.org

"Open source drug discovery: highly potent antimalarial compounds derived from the tres cantos arylpyrroles." *ACS central science* 2, no. 10 (2016): 687-701.

Link to google docs

To the visualizer

Validation dataset for ab initio project

PubChem
internal database
92674432

Filter molecules
129808

UI to browse
and select

Convert to XYZ,
zip, download

moloc
add H
2D to 3D

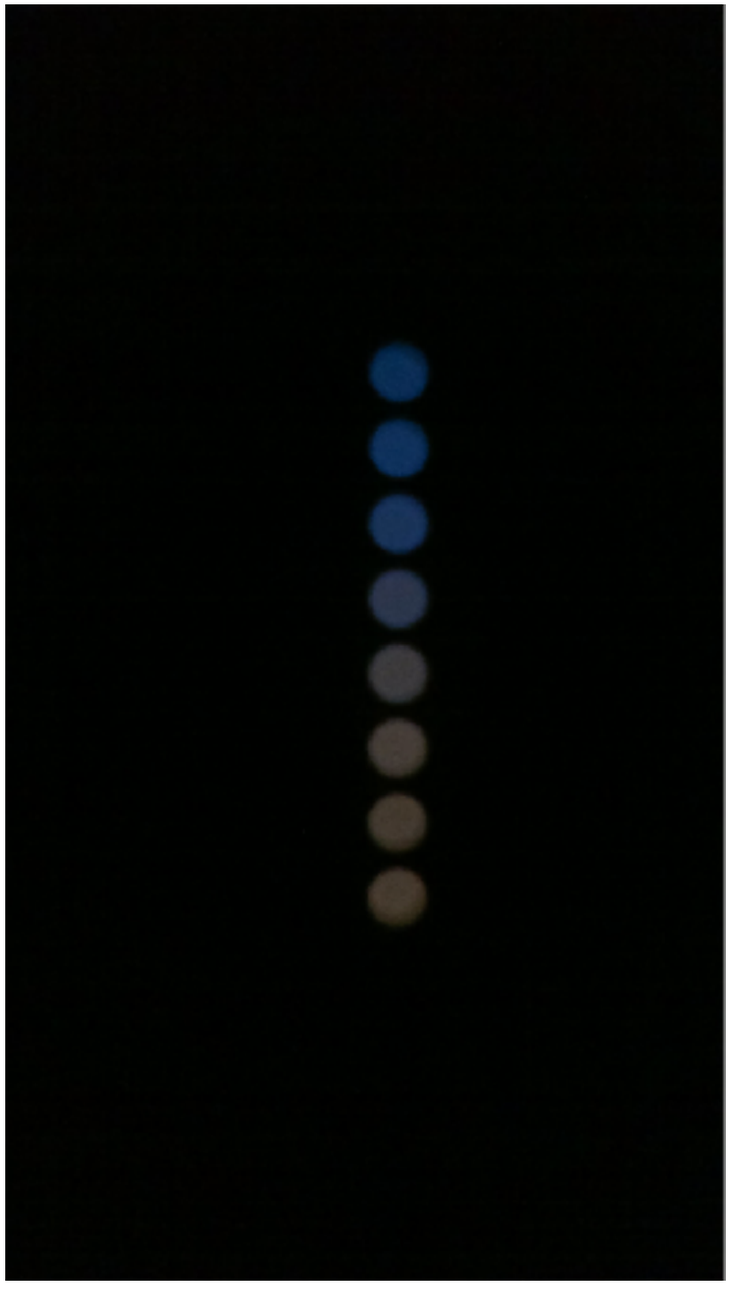
Demo

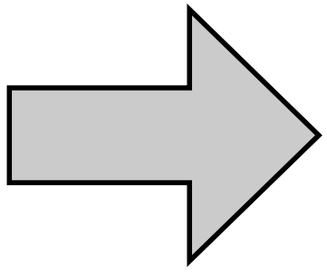
Image analysis

Image analysis : Bacteria in water

Demo

Image analysis: IC_{50}

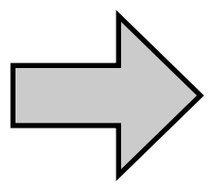


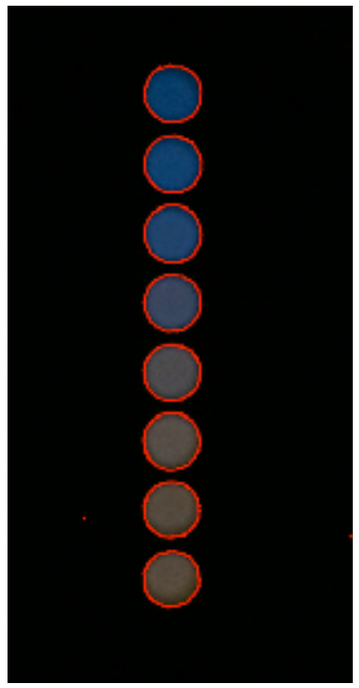


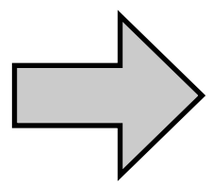
[ ]

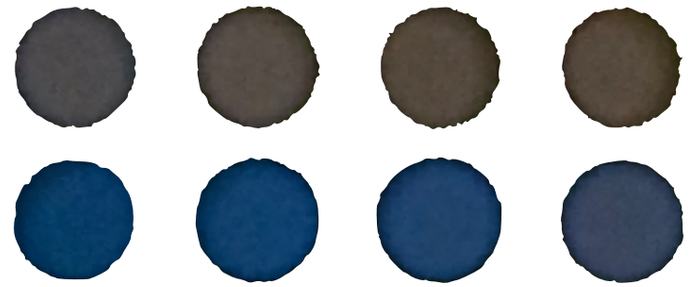
Bioluminescent sensor proteins for point-of-care therapeutic drug monitoring
Kai Johnsson *et al.*, Nature chemical biology 2014, 10, 598-603.

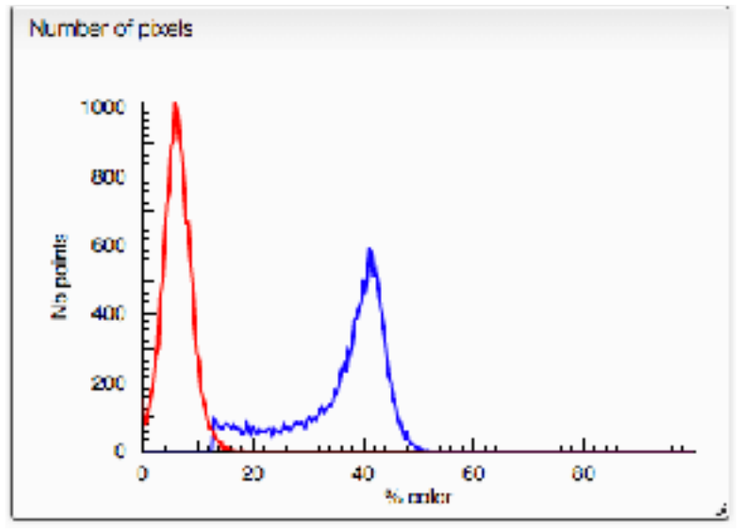
The problem : many steps

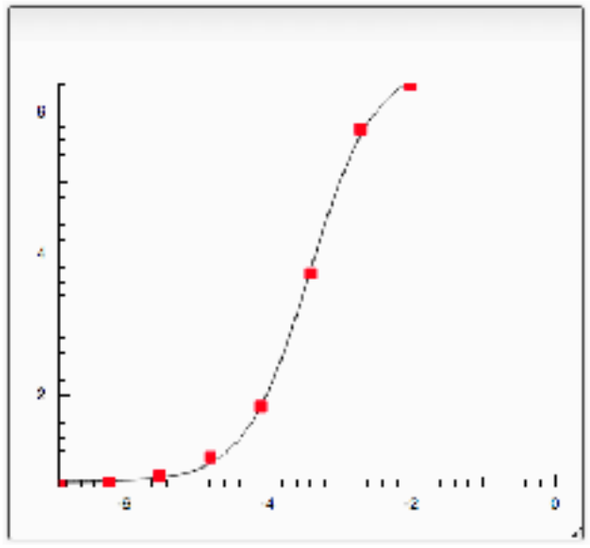


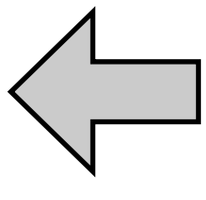












[ ]

Demo

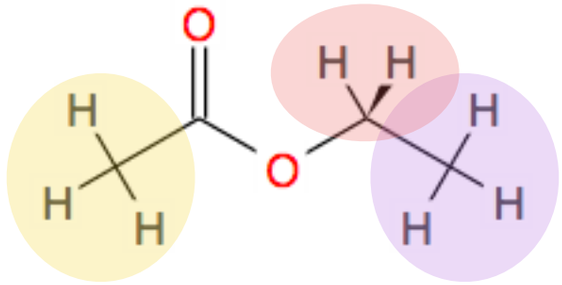
SEM / TEM image analysis

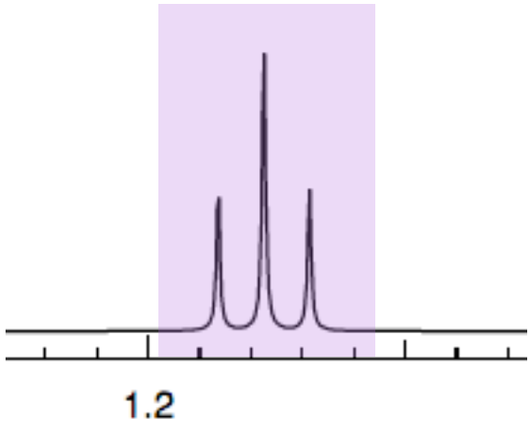
Demo

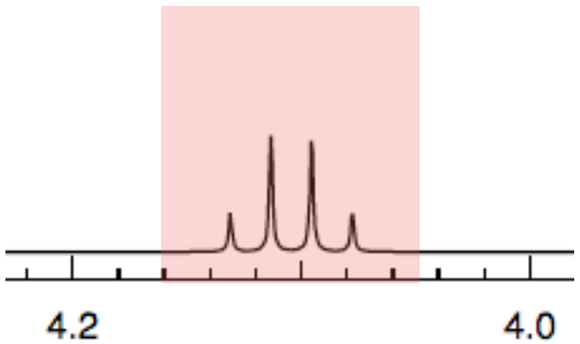
Nuclear magnetic resonance

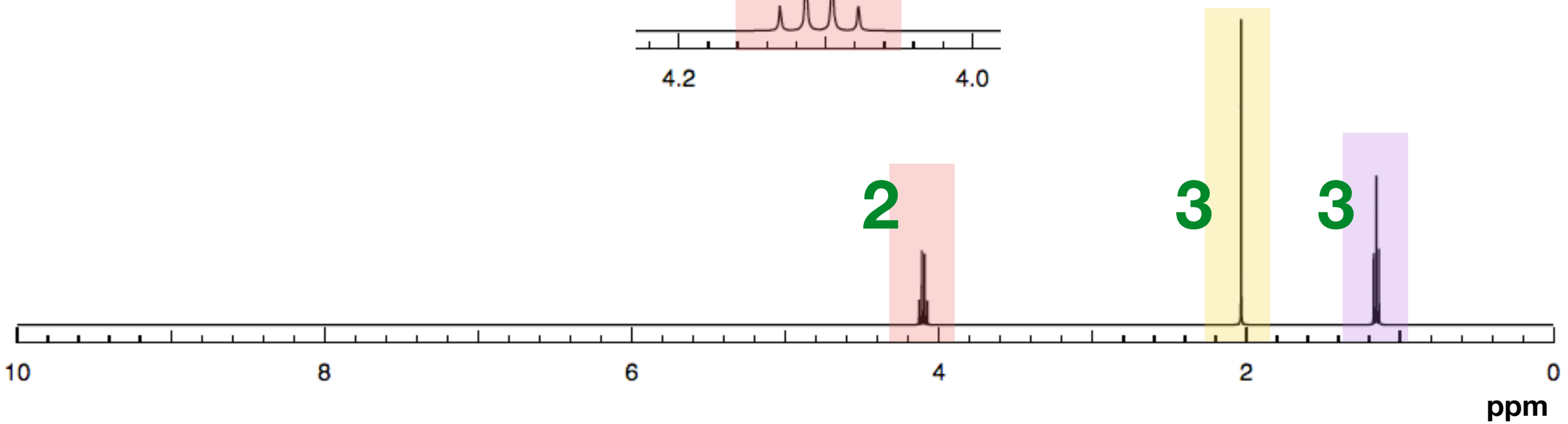
http://www.nmrdb.org

^{1}H NMR spectroscopy

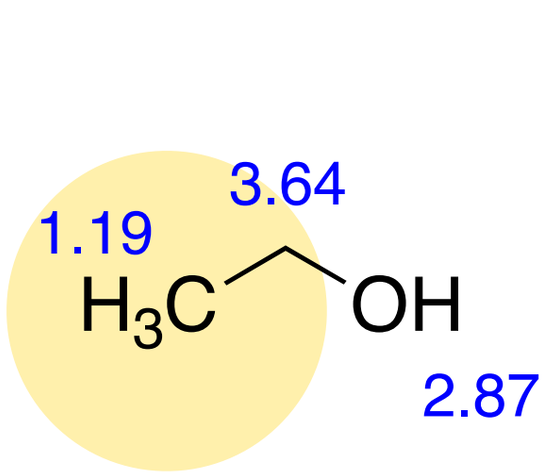


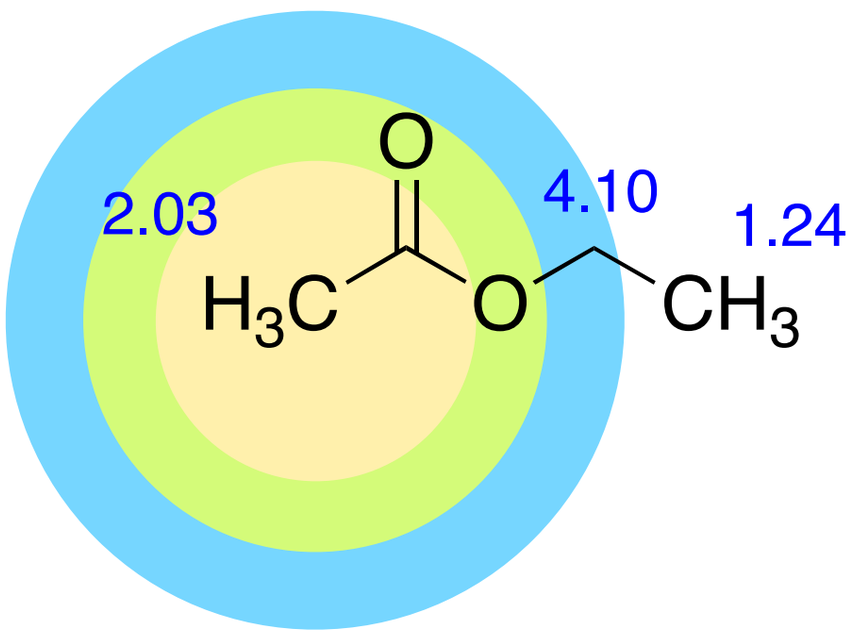


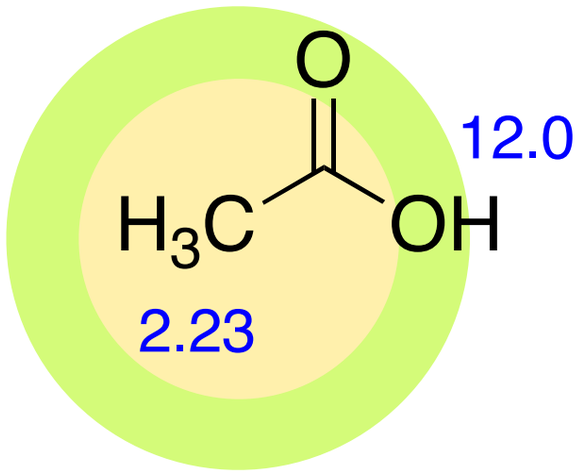


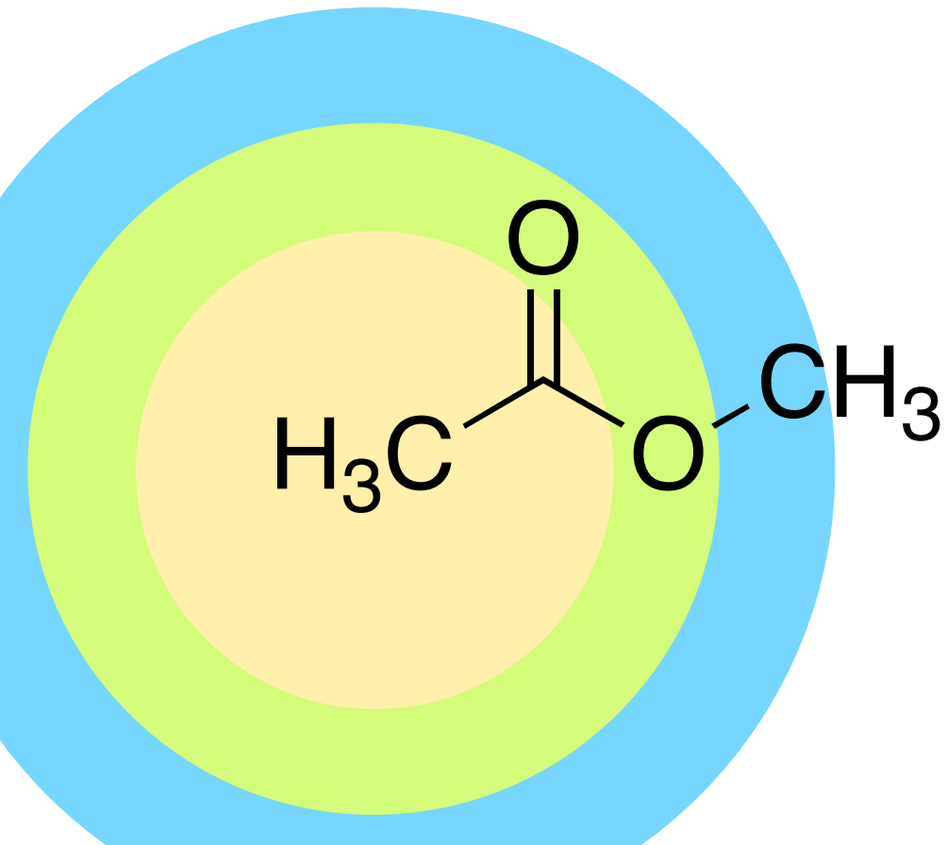


HOSE code









$$\delta = \frac{1.19+2.03+2.23}{3} = 1.81$$

$$\delta = \frac{2.03+2.23}{2} = 2.13$$

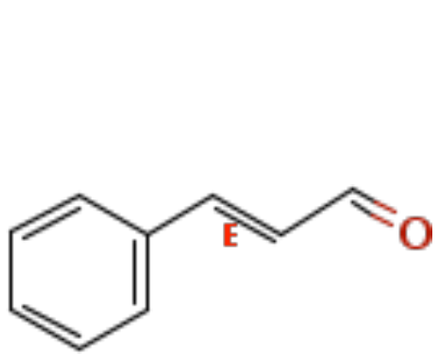
$$\delta = 2.03$$

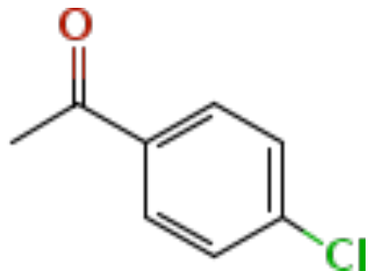
NMR self learning algorithm - Ask Ernö

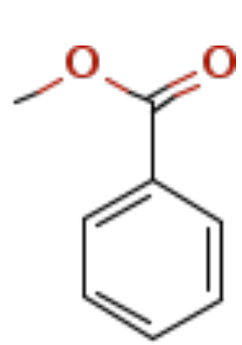
Creating a NMR chemical shifts predictor without using chemical shifts !?

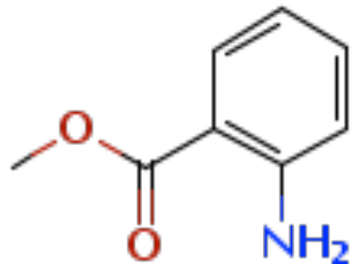
Castillo AM, Bernal A, Dieden R, Patiny L, Wist J. “Ask Ernö”: a self-learning tool for assignment and prediction of nuclear magnetic resonance spectra. Journal of cheminformatics. 2016 Dec 1;8(1):1-8.

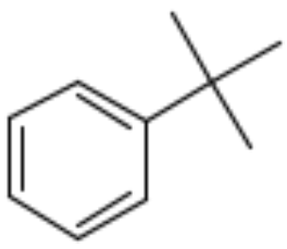
1. Creating a dataset: molfile / spectrum

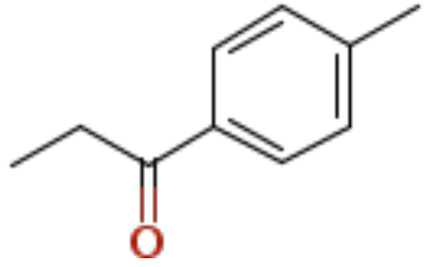


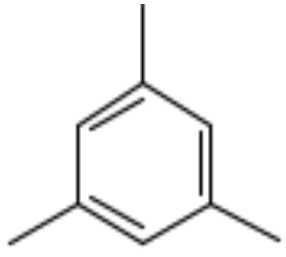


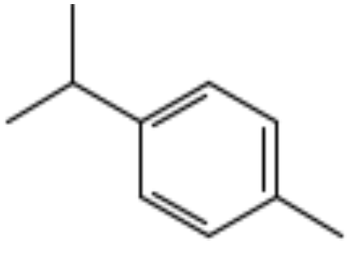


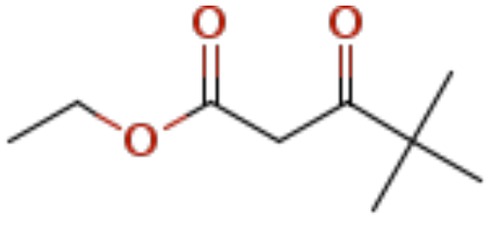






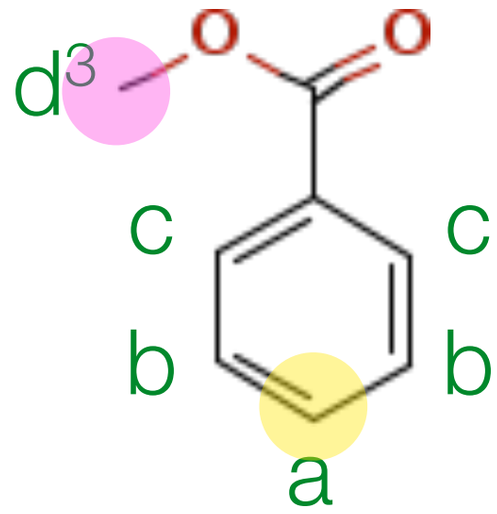




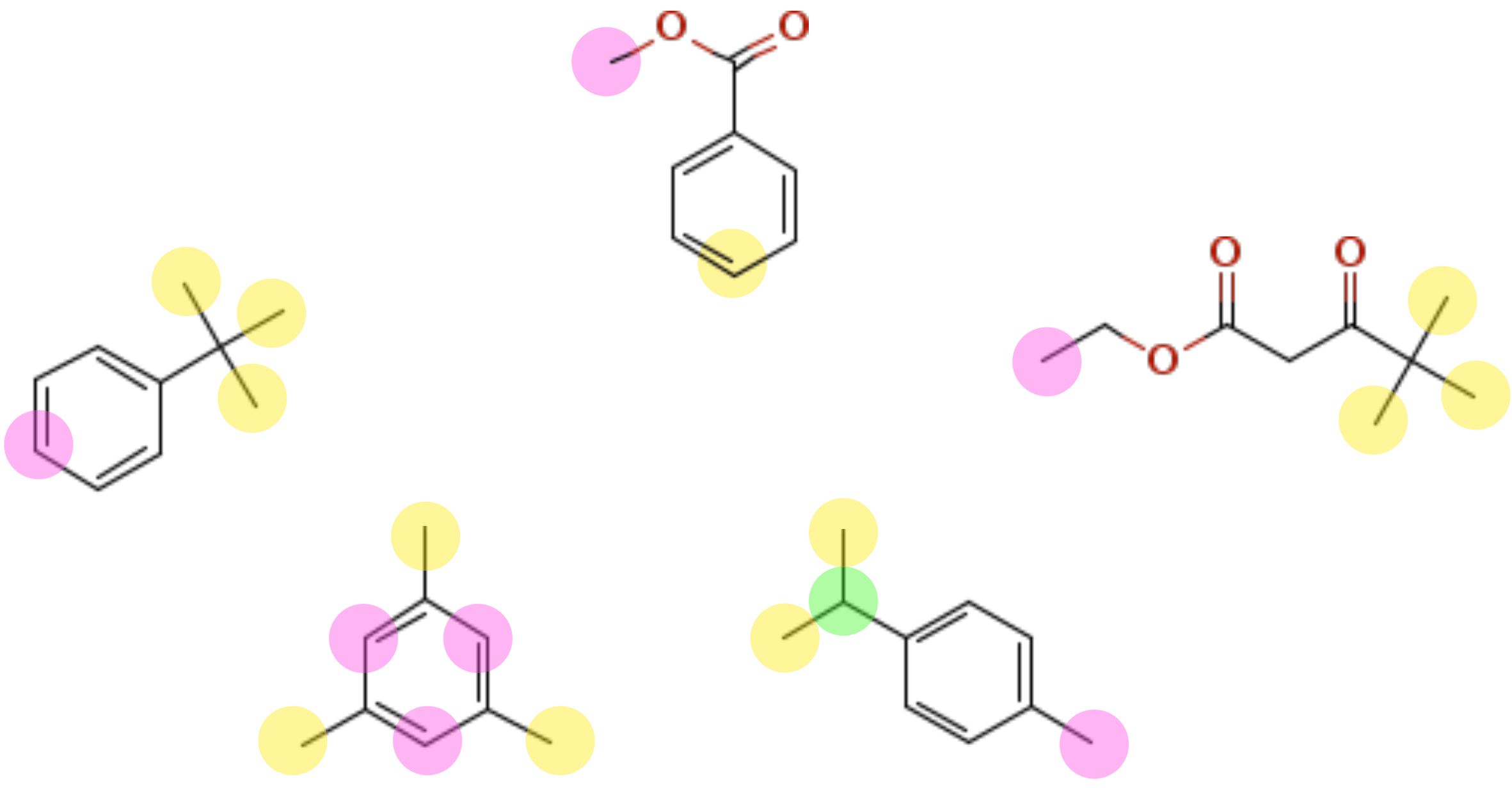


2341 molecules / NMR spectra

2. Number and kind of hydrogens



3. Unambiguous number of protons

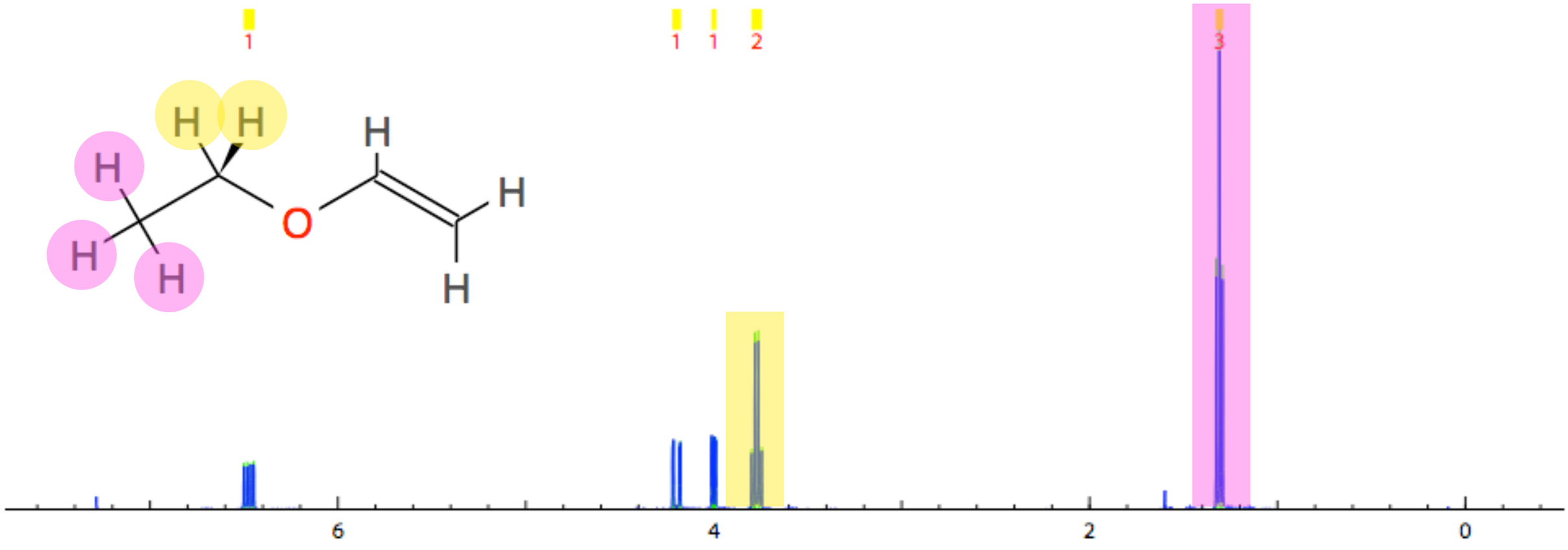


3. Peak picking and automatic integration

- Integration of the zones
- Removal of the NMR solvent / impurities

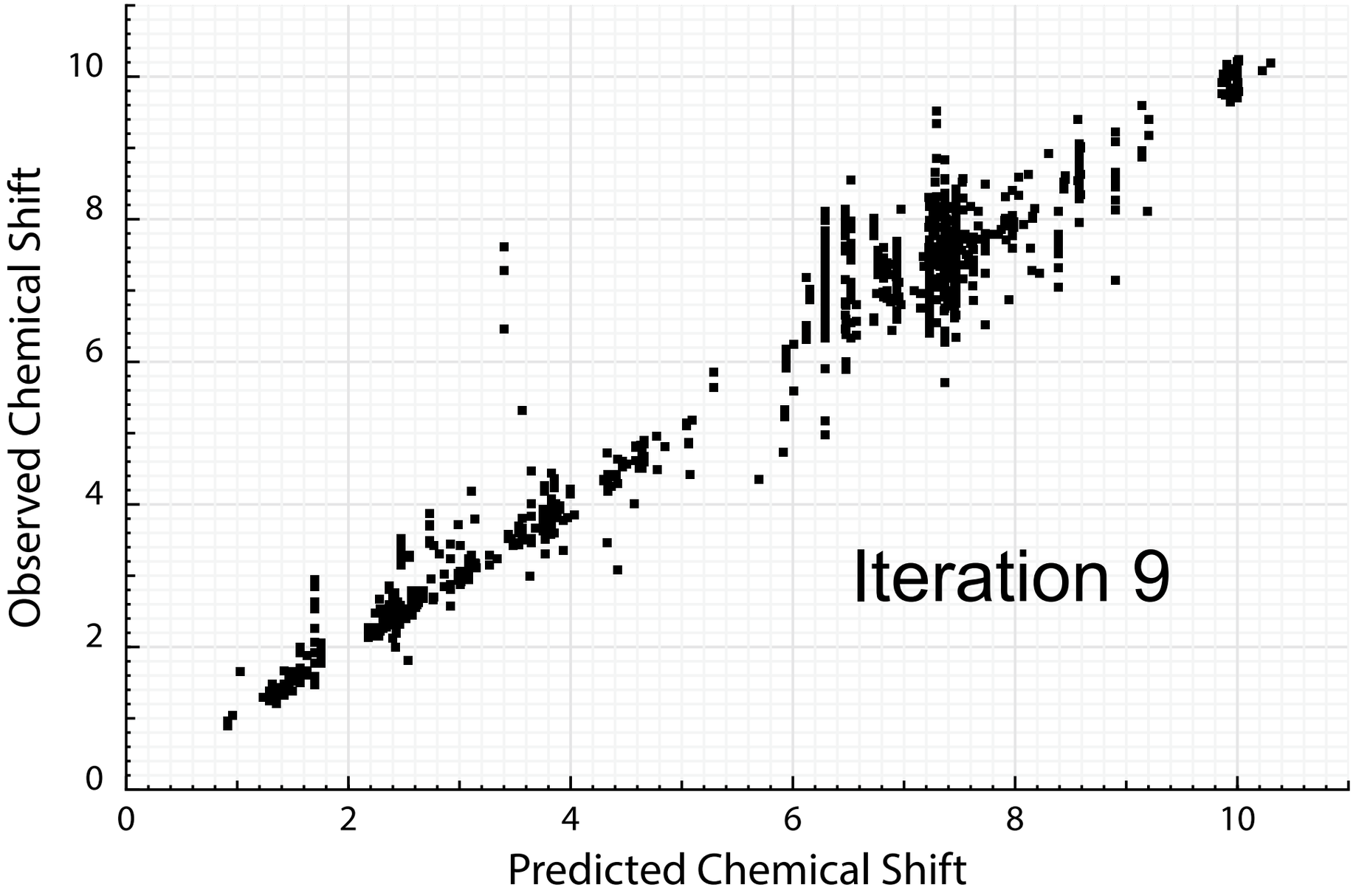
Demo

4. Assign non ambiguous protons



… and create corresponding HOSE codes

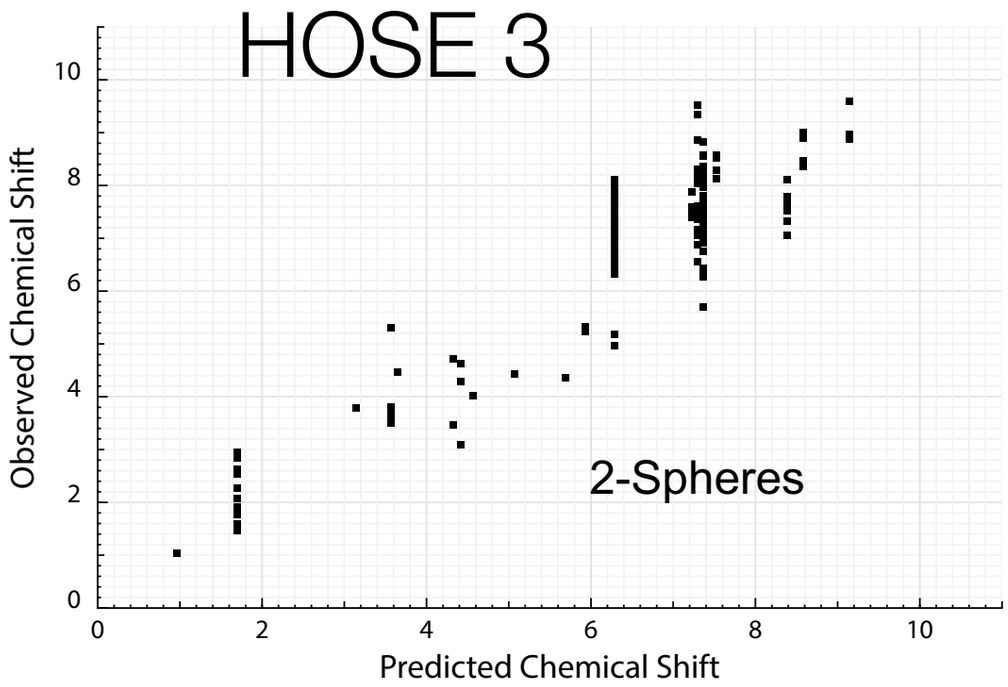
Analysis of the 2341 molecule / spectra set

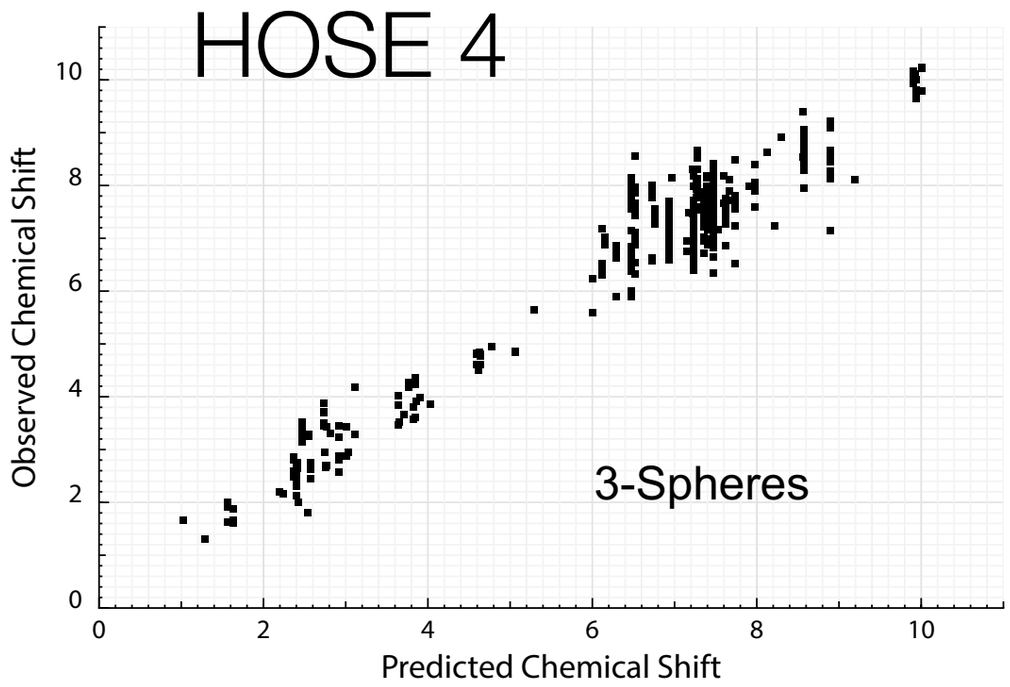


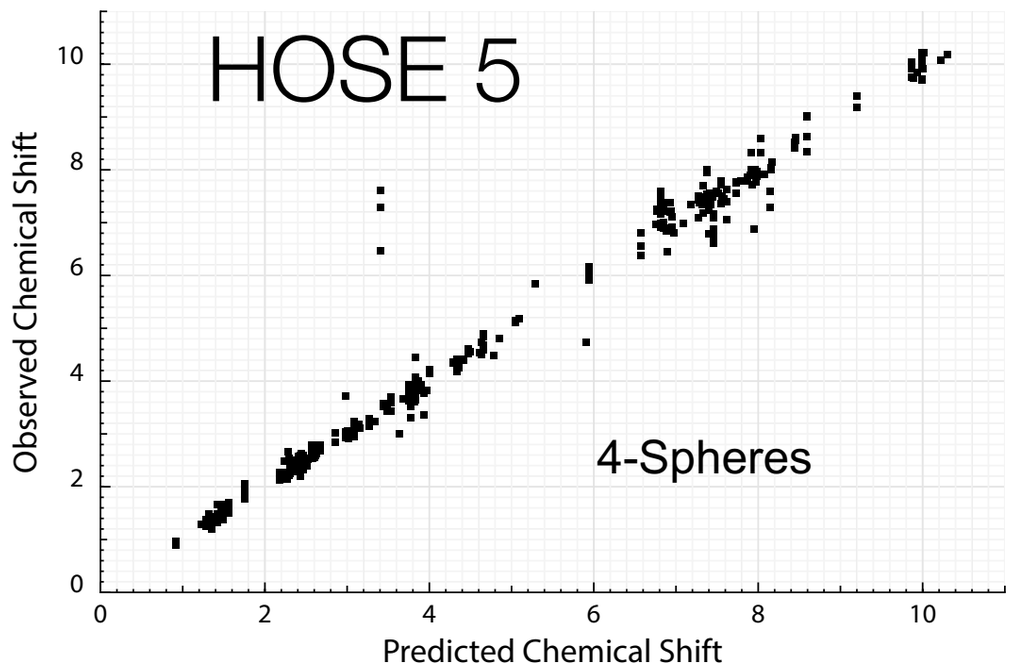
HOSE 3 : 63
HOSE 4 : 382
HOSE 5 : 916

Test set of 298 molecules

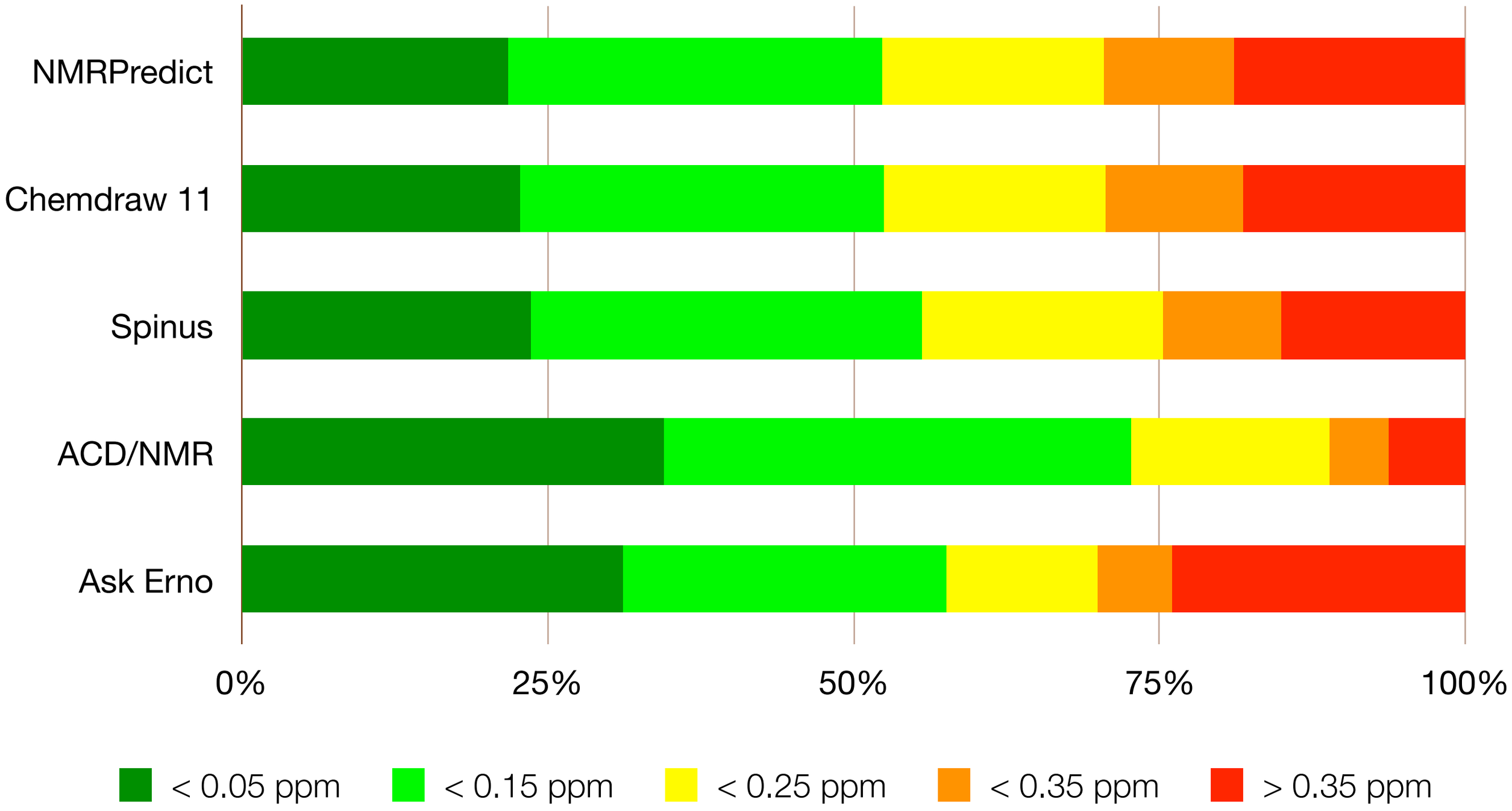
Quality of prediction based on HOSE code level





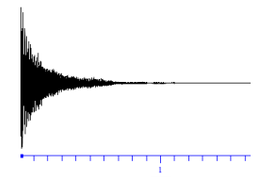


Prediction average errors



We need reliable structured data !

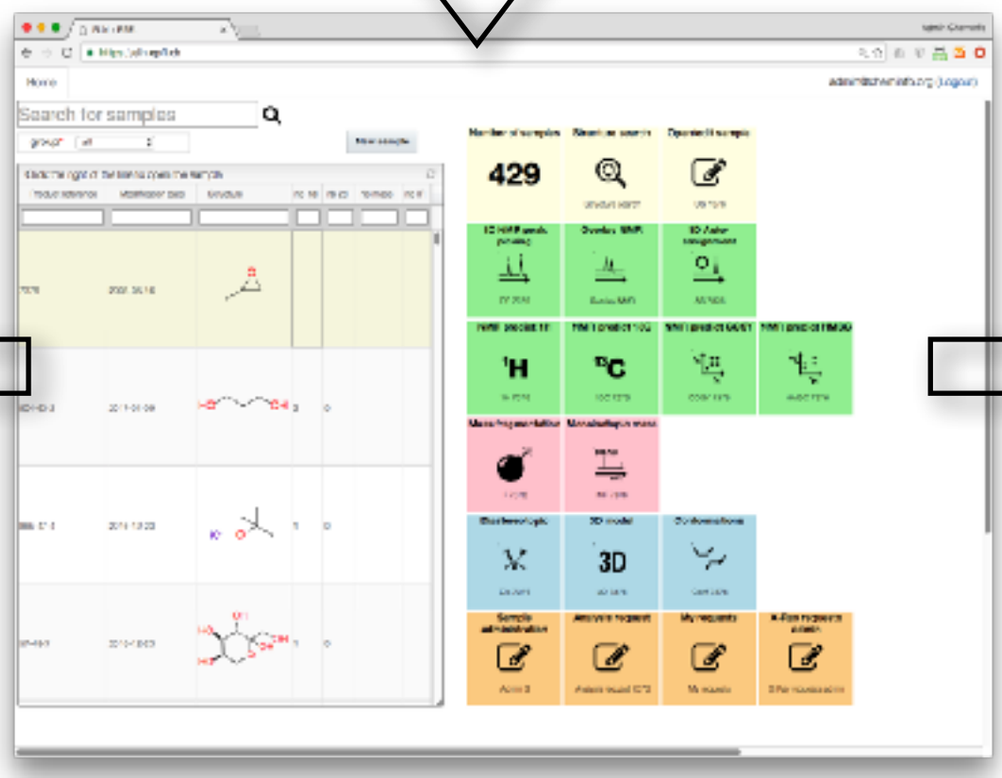
http://www.c6h6.org



DB

www.c6h6.org

Image analysis
Mass analysis
Similarity search
Predictions
…



open data

Demo

Conclusions

Advanced visualization platform

- Solves scientific problems in the browser
- Pure HTML5 / javascript
- Easily reusable javascript library
- Research, service and teaching
- Over 150 tools in 3 years

www.cheminfo.org

Client / Server approach

Databases

Processing engine

Tools

Server

Knowledge

Client

Thanks !

ChemCalc

- Alain Borel (Library EPFL)
- Laure Menin
- Michael Krompiec
- Marek Noga

NMRdb

- Julien Wist
- Andrès Castillo
- Andrès Bernal

script, visualizer, components

- Norman Pellet
- Michaël Zasso
- Daniel Kostro
- Jefferson Hernández
- Miguel Asencio
- Julien Wist
- Andrès Castillo