

# Welcome!

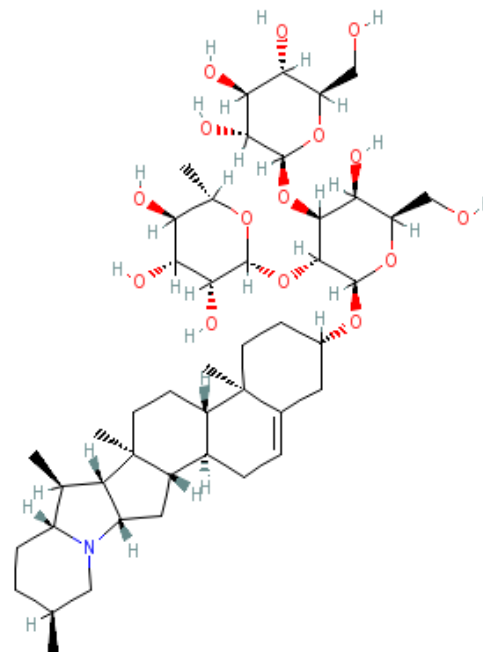
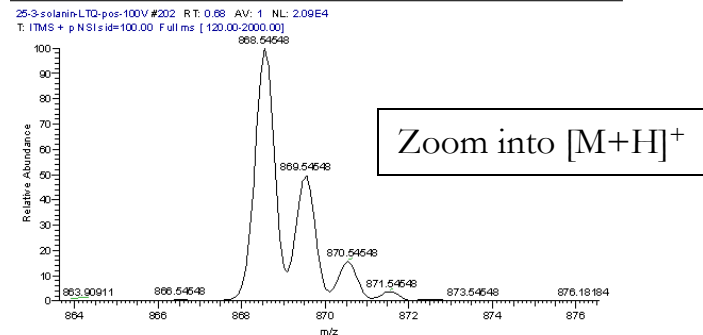
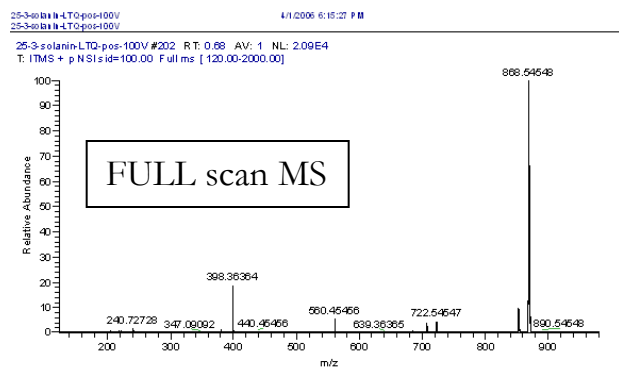
Mass spectrometry meets cheminformatics  
WCMC Metabolomics Course 2013  
Tobias Kind

Course 2: Mass spectral and molecular  
data handling

# Molecules and mass spectra

Dense relationship between molecular structure and mass spectra

- Important to handle molecular structures
- Important to handle mass spectra and chromatograms (GC-MS, LC-MS)



ESI (pos) mass spectrum  
with zoom into isotopic pattern

[Solanine](#)  
([InChIKey=ZGVSETXHNHBTRK-OTYSSXIJBP](#))

# How are mass spectra stored?

More than 50 vendor specific formats are known.  
For every MS, LC-MS, GC-MS a single file format.  
Mostly very complex data streams (formats).



Tower of Babel – Source: Brueghel/WIKI

For simple electron impact (EI) spectra  
 $m/z$  and intensity list sufficient

## Example MSP Files

Name: Cocaine  
Formula: C17H21NO4  
MW: 303  
CAS#: 50-36-2; EPA#: 113834  
DB#: 32675  
Num Peaks: 87  
14 8; 15 15; 27 18; 28 15; 29 15;  
30 11; 32 19; 39 32; 40 12; 41 68;  
42 234; 43 16; 44 41; 45 10; 50 30;  
51 121; 52 12; 53 41; 54 27; 55 78;  
56 36; 57 43; 58 12; 59 50; 65 29;  
66 15; 67 58; 68 63; 69 17; 70 30;  
71 9; 74 6; 75 8; 77 355; 78 39;  
79 40; 80 36; 81 125; 82 999; 83 367;  
84 36; 91 47; 92 11; 93 51; 94 366;  
95 50; 96 249; 97 111; 98 10; 100 11;  
105 296; 106 30; 107 18; 108 54; 109 12;  
110 18; 114 4; 118 9; 119 36; 120 22;  
121 10; 122 88; 123 15; 124 11; 135 6;  
138 7; 140 10; 150 27; 151 4; 152 38;  
153 7; 154 14; 155 23; 166 32; 179 4;  
180 19; 181 59; 182 716; 183 83; 184 8;  
198 95; 199 12; 272 69; 273 14; 303 172;  
304 37; 305 5;

Metadata like  
CAS, MW, Formula

$m/z$  - intensity pairs

For complex MS/MS data, accurate masses,  
ionization voltage and instrument method needed

## Example Thermo Finnigan RAW file:

```
data_dependent_02 #1 RT: 0.0082

Total Ion Current:      2268344.00
Scan Low Mass:         150.00
Scan High Mass:        1000.00
Scan Start Time (min): 1.01
Scan Number:           33
Base Peak Intensity:   100761.00
Base Peak Mass:        180.95
Scan Mode:              + c Full ms [150.00-1000.00]

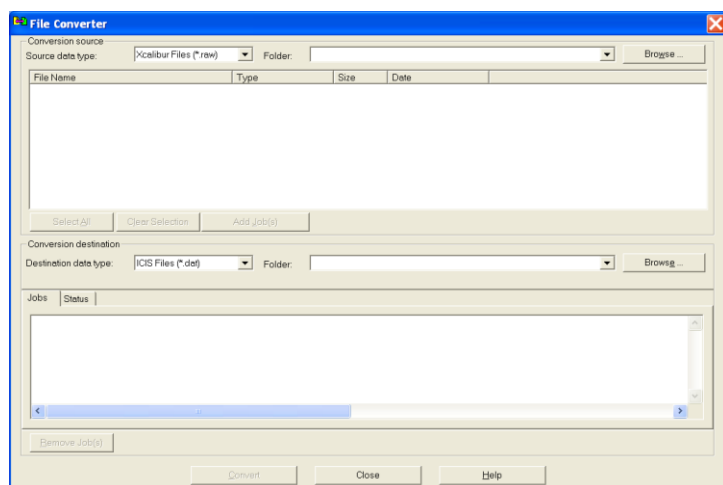
Instrument Data:
=====
Micro Scan Count:      3
Ion Injection Time (ms): 199.98
Scan Segment:          1
Scan Event:            1
Elapsed Scan Time (sec): 1.89
API Source CID Energy: 0.00
Resolution:            Low
Average Scan by Inst:  No
BackGd Subtracted by Inst: No
Charge State:          0
```

# Inter-conversions of mass spectra

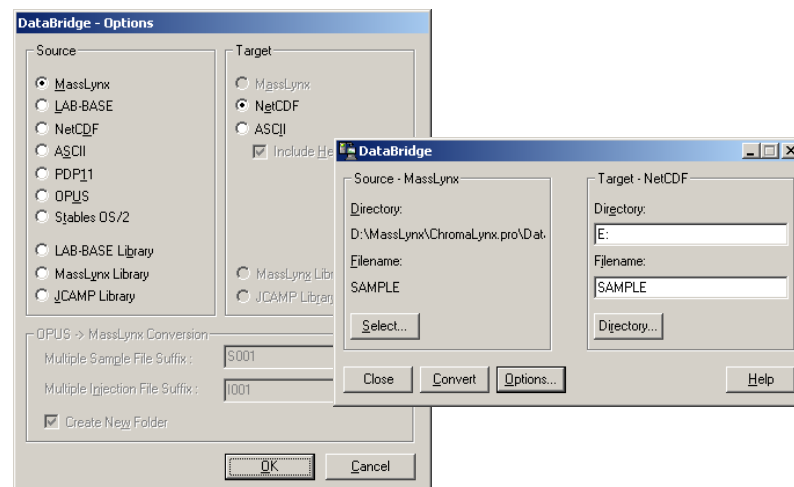
Issue: Its an extreme hassle, data may get lost, may require license

Solution: Open exchange formats (JCAMP, netCDF, mzXML)

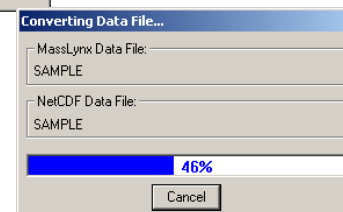
Problem: how to convert complex mass spectral MS experiments?



Thermo FileConvert



Waters DataBridge

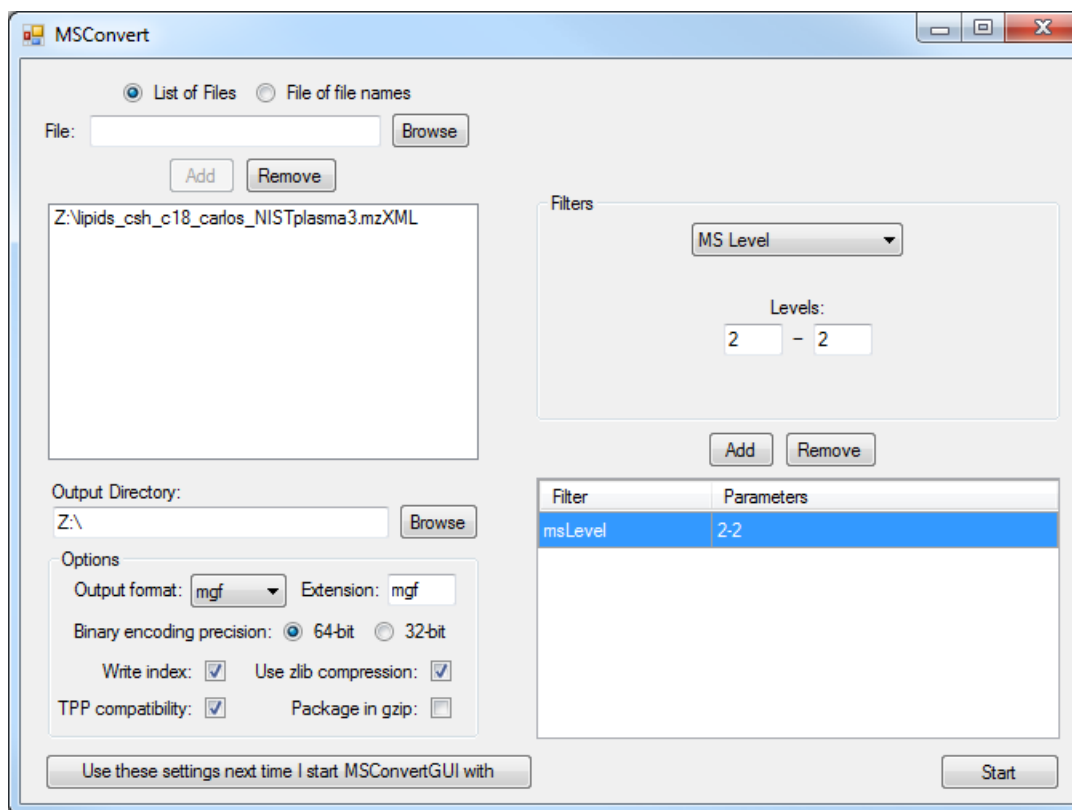


See helper applications [MassTransit](#)

See helper applications [ms-utils.org](#)

See helper applications [Lib2NIST](#)

# ProteoWizard for almost all vendor software



Input vendors supported: ABI, Agilent, Bruker, Thermo, Waters  
Output formats supported: mzML, mzXML, MGF (MS/MS), ASCII

# Mass Spectra – Importance of Metadata

Name: Roxithromycin

Formula: C<sub>41</sub>H<sub>76</sub>N<sub>2</sub>O<sub>15</sub>

MW: 836 CAS#: 80214-83-1 NIST#: 1005429 ID#: 2064 DB: nist\_msms

Other DBs: None

Comment: Draisci R. J CHROMATOGR A 926 (1) 97-104 2001

Instrument type QqQ/triple quadrupole

Spectrum type ms2

Compound type M

Precursor type [M+H]<sup>+</sup>

Precursor m/z 837.53

Collision energy 25 eV

Instrument PE Sciex API III Plus

Ionization ESI

Ion mode P

Collision gas Ar

Pressure gas target thickness 3.00x10<sup>+15</sup> atoms/cm<sup>2</sup>

5 largest peaks:

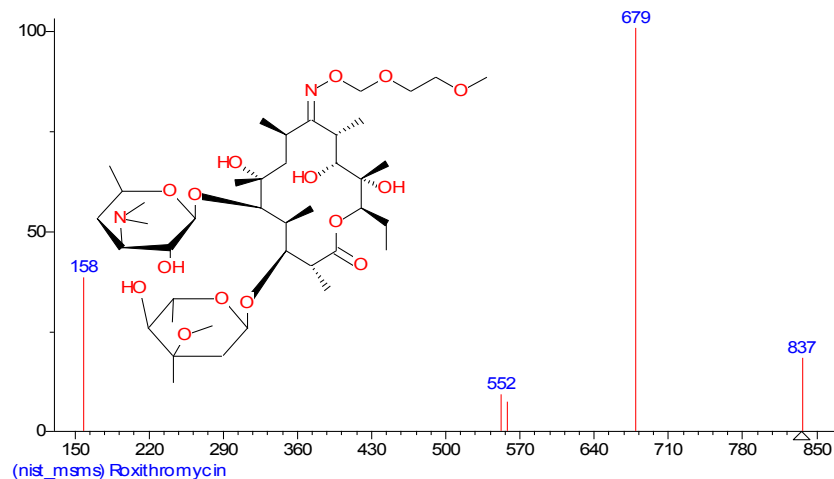
679	999		158	380		837	180		552
90		558		70					

5 m/z Values and Intensities:

158	380		552		90		558
70		679	999		837	180	

Synonyms:

no synonyms.



Different MS techniques deliver different mass spectra  
Information must be captured (best via XML)

# Open Exchange formats for mass spectra

Why? You're in a successful lab using multiple vendor mass spectrometers.

Why? You want to share and receive mass spectra from colleagues.

Why? Future grants will require depositing of mass spectra in repositories.

## Common exchange formats for GC-MS

- [JCAMP-DX](#) format for mass spectrometry
- [netCDF](#) format for hyphenated data (LC-MS, GC-MS)
- NIST MSP and MassBank record format (GC-MS)

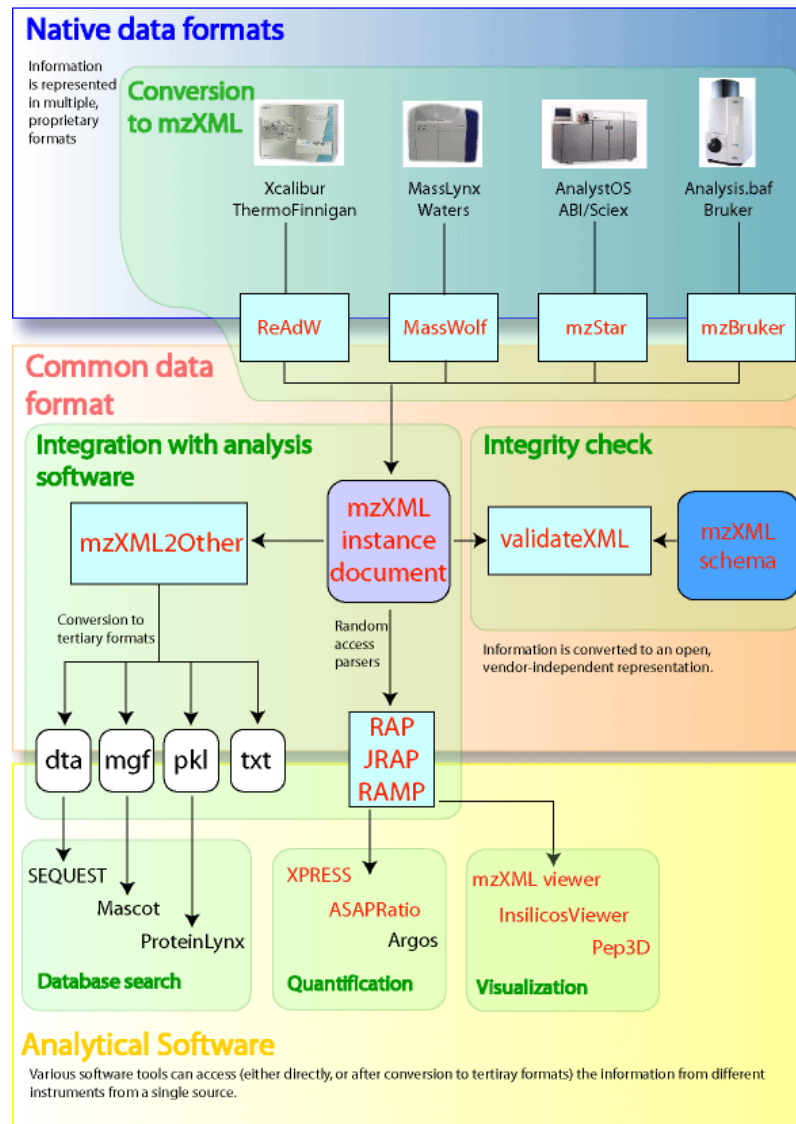
## Common exchange formats for LC-MS/MS

- mzML for LC-MS/MS
- mzXML for (LC-MS and MS/MS)
- [MassBank](#) record format – well defined

→ Ask vendors for multiple export options, proprietary formats are no good

→ Format converters are only temporary solutions

# mzXML format for LC-MS/MS data



Dta, mgf, pkl files hold MS/MS spectra for database search

Picture Source:  
Seattle Proteome Center (SPC)  
NHLBI Proteomics Center at the  
Institute for Systems Biology  
<http://www.proteomecenter.org>



# How does mzXML look like?

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<msRun
  xmlns="http://sashimi.sourceforge.net/schema/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://sashimi.sourceforge.net/schema/
http://sashimi.sourceforge.net/schema/MsXML.xsd"
  scanCount="4140"
  startTime="PT120.030000S"
  endTime="PT5880.790000S">
  <parentFile fileName="raft0020.mzXML"
    fileType="RAWData"
    fileSha1="da39a3ee5e6b4b0d3255bfef95601890afd80709"/>
  <instrument manufacturer="ThermoFinnigan"
    model="LCQ Classic"
    ionisation="ESI"
    msType="Ion Trap">
    <software type="acquisition"
      name="ICIS"
      version="8.4"/>
  </instrument>
  <dataProcessing>
    <software type="conversion"
      name="dat2xml"
      version="0.1"/>
  </dataProcessing>
  <scan num="1"
    msLevel="1"
    peaksCount="959"
    retentionTime="PT120.030000S"
      startMz="400.0000"
      endMz="1400.0000"
    lowMz="400.3742"
    highMz="1399.3711"
      basePeakMz="534.2230"
      basePeakIntensity="913904.0000"
      totIonCurrent="31883915.0000">
    <peaks
      precision="32">Q8gv5kaBhgBDyLU0RpCAAEPJNhBGPfgAQ8m6CEcGnQBDyhmYP4AAAEPK
p9RGM/QAQ8sQIEXgEABDy2RGRgC8AEPL67pGs04AQ8xrDKW/EABDzLrgRw8kAEPNDf5GA
cgAQ82t2kaDSgBDzjg8RWwyABErVXqRn/oAeSteQhHMewARK2RED+AAABErBf0R0AdAEStz
QhHBX4ARK3IZEca2QBErgrWRmooAESulAA/gAAARK5apEcuAABErnURijkAESuk+BGzO4A
RK7Bykc2RgBERuvgrRo+0AA==</peaks>
  </scan>
</msRun>
```

compressed data



## General Structure of XML data

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<msRun ..>
  <instrument>
    ...
  </instrument>
  <dataProcessing>
    ...
  </dataProcessing>
  <scan num="1">
    ...
  </scan>
  <scan num="2">
    ...
  </scan>
  <index name="scan">
    <offset id="1">849</offset>
    <offset id="2">11405</offset>
    <offset id="3">12072</offset>
    <offset id="4">20708</offset>
    ...
  </index>
</msRun>
```

# MGF – Mascot Generic Format for MS/MS

```
BEGIN IONS
TITLE=804.40 [Da] ;MGDG 18:0/18:0
Comments: PEPMASS=804.40; MGDG 18:0/18:0; [M+NH4]+);
PEPMASS=804.40
CHARGE=1+
RTINSECONDS=2.811
225.063 1.31
283.227 1957.16
284.258 81.77
298.387 14.79
299.111 65.01
300.196 16.16
310.894 9.66
311.290 3995.51
785.723 3687.35
786.441 3839.42
786.814 981.80
814.903 0.71
END IONS
...
```

← Required

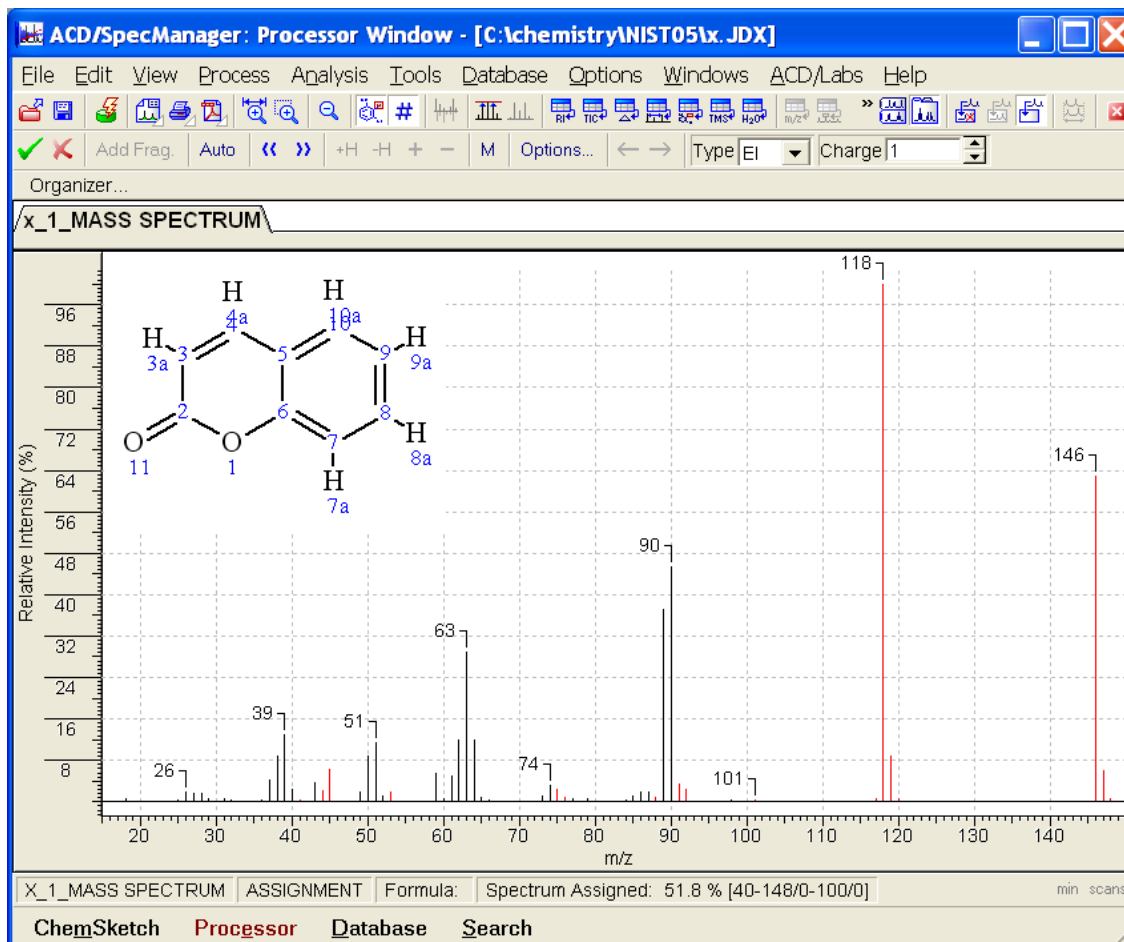
← Required

m/z and abundance pairs  
(not normalized)

← Required

Most common format for MS/MS search, can hold 10000s of spectra  
Files can be large (with noise) → slow search performance  
For MS/MS search export only 100 most abundant MS/MS peaks

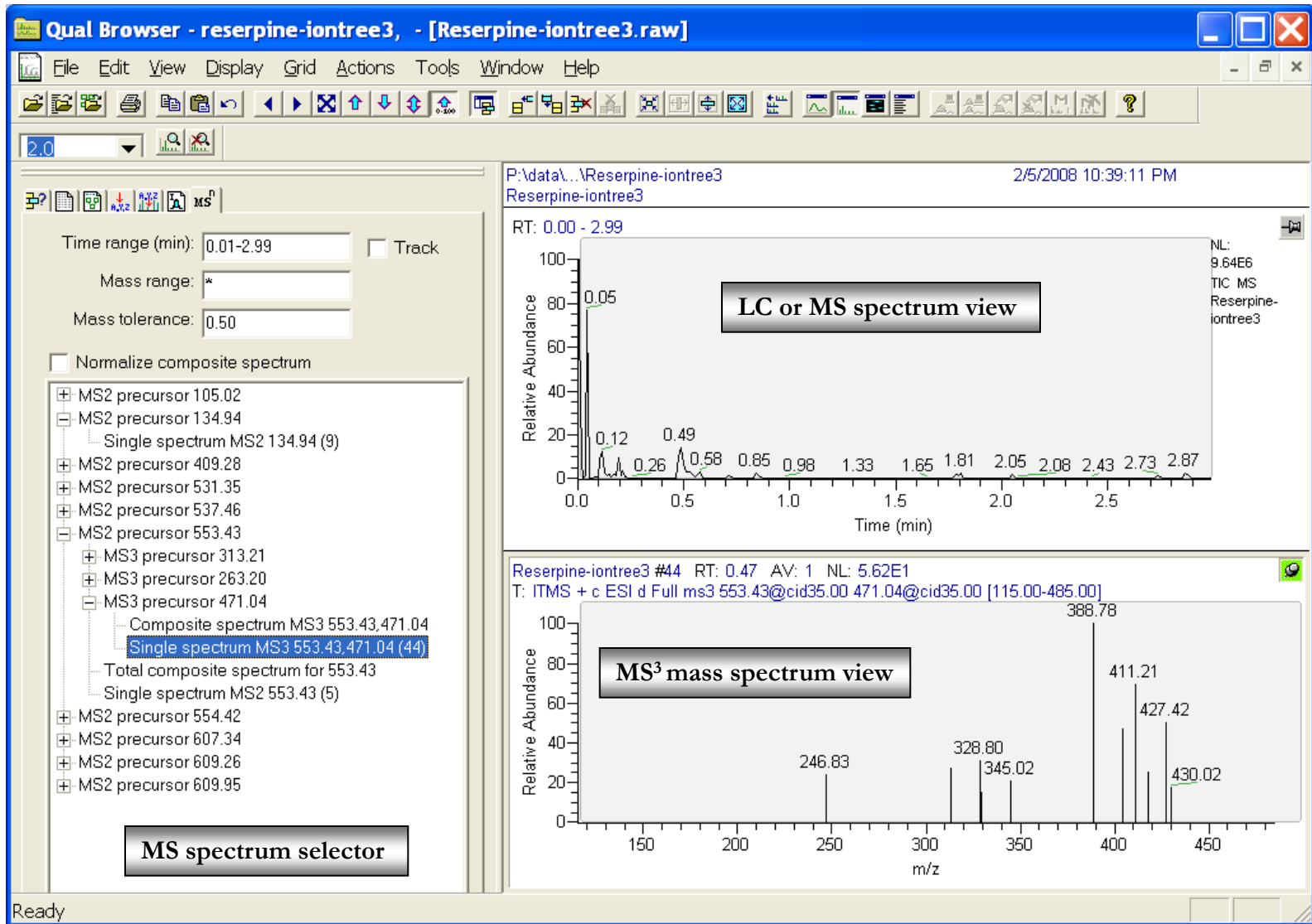
# Mass spectral data handling ACD/SpecManager



- Can handle multiple formats
- Can do spectral annotations
- Can store spectra in database

See also [HighChem MassFrontier](#)  
See also [NIST MS Search](#)

# MS data handling - Thermo XCalibur example



# BioClipse showing JCAMP file

The screenshot displays the BioClipse software interface with the following components:

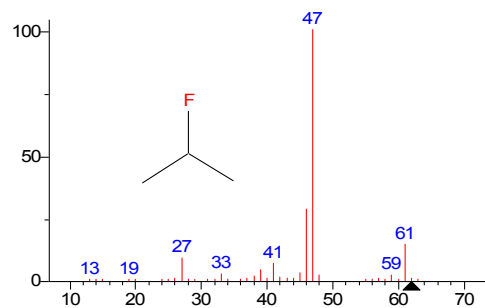
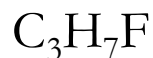
- BioResource Navigator:** A tree view on the left showing the file structure, including folders like 'chemical-structures', 'configuration', 'plugins', 'SampleData', 'jmolScripts', 'js', 'pdb', 'sequences', and 'spectra'. The 'spectra' folder contains files 'aug07.dx' and 'cpd01.jdx'.
- Peak Spectrum View:** A mass spectrum plot titled 'CPD01 entry 1'. The x-axis is labeled 'jcampdx:' and ranges from 25 to 75. The y-axis is labeled 'jcampdx:' and ranges from 0 to 12,500. Three peaks are labeled with their m/z values: 29.0, 44.0, and 73.0.
- Peak Table View:** A table at the bottom left showing peak data. The table has columns for 'xaxis', 'yaxis', and several empty columns. The data rows are as follows:

xaxis	yaxis				
27.00	1248.00				
28.00	2067.00				
29.00	5538.00				
30.00	351.00				
31.00	702.00				
32.00	39.00				
36.00	78.00				
38.00	39.00				
39.00	39.00				
40.00	39.00				
41.00	39.00				
- Spectrum Metadata View:** A tree view on the right showing the metadata for the spectrum. The keys and values are:
  - JCamp General Keys
    - Title: CPD01 entry 1
    - Owner: COPYRIGHT UNKNOW
    - Origin: UNKNOWN ORIGIN
  - JCamp Substance Keys
    - CAS Registry No: 0000
  - JCamp Instrument Keys
  - Non JCamp Keys
    - rti: 0.0000
    - substance\$mass\$mw: 0
    - jcamp\$datatype: MASS SPECTRUM
    - SPECTRUMID: sid\_ede1fc5e69a3a1
- Console / Properties:** A table at the bottom right showing the properties of the selected file (cpd01.jdx).

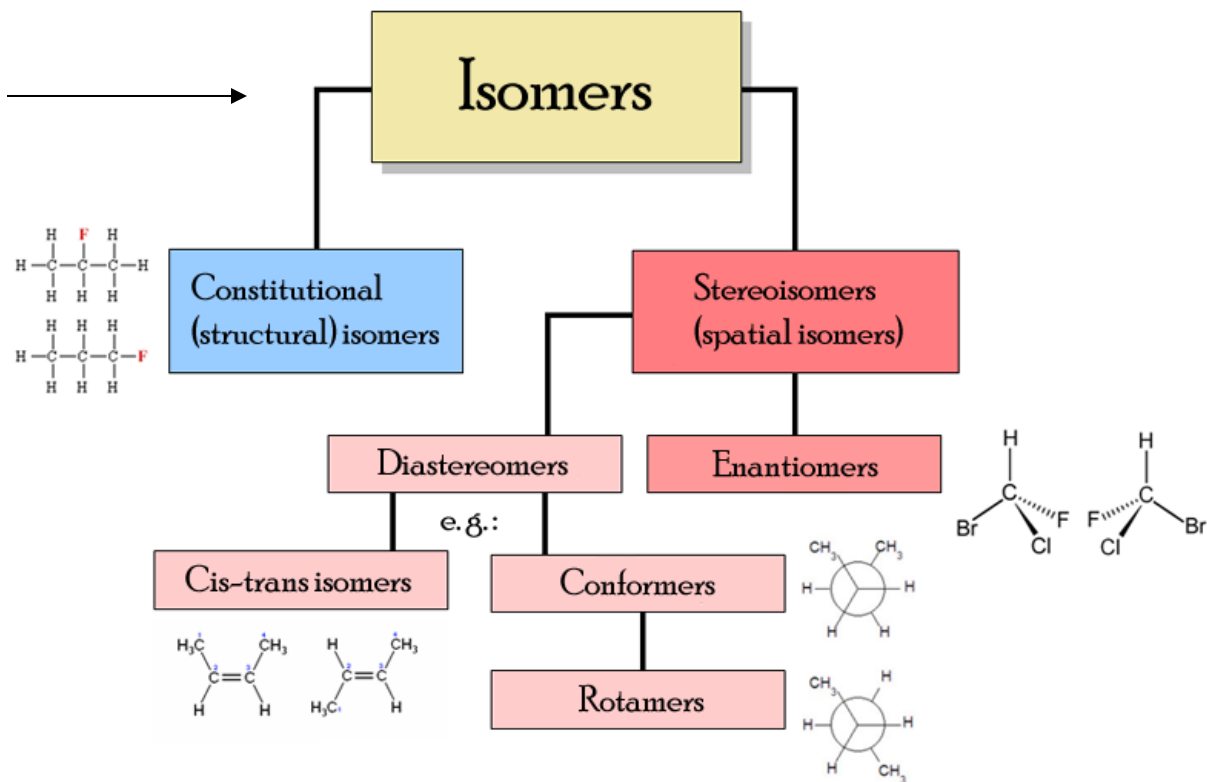
Property	Value
Format	JCAMP-DX Spectrum
Name	cpd01.jdx
Parent	spectra
Path	C:\chemistry\bioclipse\Sam...
Persisted by	net.bioclipse.model.resour...
Size	379 bytes
Type	Spectrum File

# Organic Chemistry Reminder

Molecular Formula

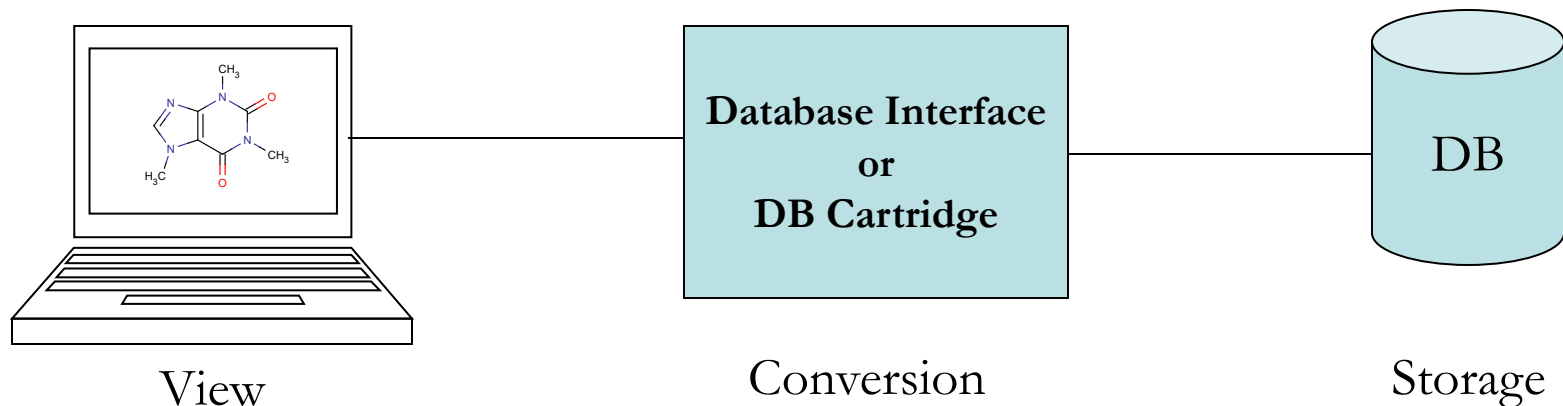


(mainlib) Propane, 2-fluoro-

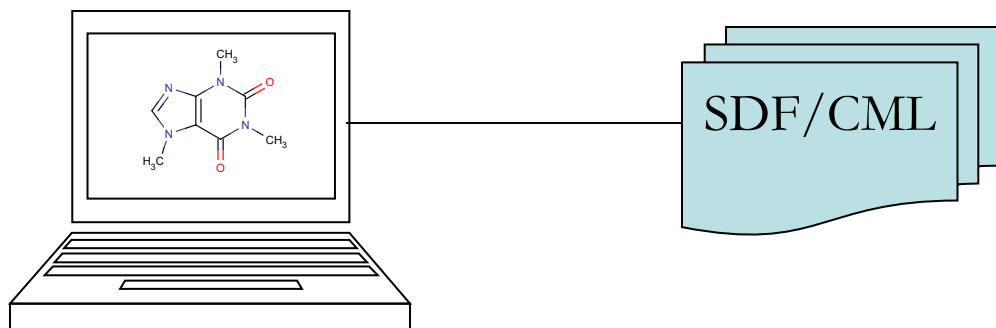


## Where are structures stored? (same for spectra)

A) In databases – for millions of structures



B) In structure files (text files) – for few structures



# How are structures stored?

...here cometh the (true) tower of Babel again  
...more than 100 different file formats in use



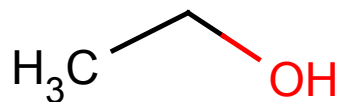
Tower of Babel – Source: Brueghel/WIKI

Structure formats can store 1D, 2D and 3D coordinate information and metadata

CCO

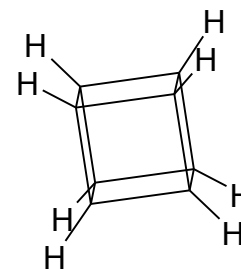
**1D**

InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3  
InChIKey=[LFQSCWELJHOTTHZ-UHFFFAOYAB](#)



**2D**

InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3  
InChIKey=[LFQSCWELJHOTTHZ-UHFFFAOYAB](#)

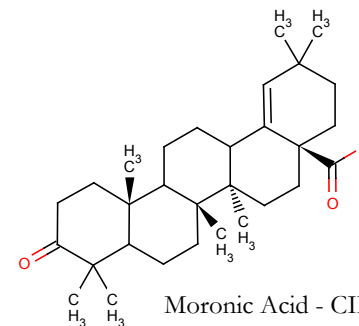


**3D**

InChI=1/C8H8/c1-2-5-3(1)7-4(1)6(2)8(5)7/h1-8H  
InChIKey=[TXWRERCHRDBNLG-UHFFFAOYAL](#)



# Chemical Structure Handling



Most common structure formats you need to know:

[SMILES](#)/SMARTS - **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification

[SDF](#)/MOL - **S**tructure **D**ata **F**ile

[InChI](#)/InChIkey - IUPAC **I**nternational **C**hemical **I**dentifier

[PDB](#) - **P**rotein **D**ata **B**ank

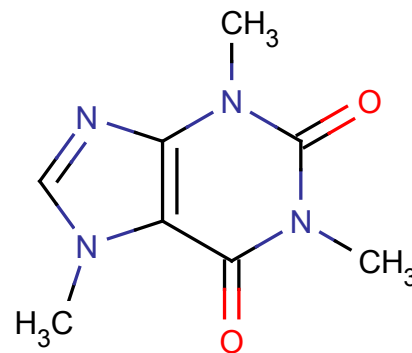
[CML](#) - **C**hemical **M**arkup **L**anguage

**Some problems:**

- Data format needs to be based on Open Standard (problem with SMILES, ok with CML)
- Stereo and aromatic bond information needs to be saved (ok with SDF)
- Format needs to be small in space for millions of compounds (ok with SMILES)
- SMILES notation needs to be unique (problem with SMILES)
- Structure representation should be portable and based on Open Standard (ok with CML)

# Chemical Structure Identifiers

Structure Identifiers are needed for uniquely identifying structures  
Important for searching chemical structures in text and databases



Structure Name – IUPAC name or common name

**1,3,7-trimethylpurine-2,6-dione**

CAS RN – Chemical Abstracts identifier

**58-08-2**

PubChem ID – PubChem Compound ID

**CID: 2519**

InChIKey – Short representation of InChI

**InChIKey=RYYVLZVUVIJVGH-UHFFFAOYAW**

InChI – IUPAC **I**nternational **C**hemical **I**dentifier

**InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3**

# SMILES structure format

Positive: Good for storing structures in single line

Fast text based search possible; human readable

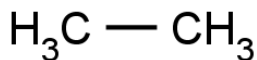
Negative: Many different SMILES codes exist

SMILES for same structure can be different (canonical or unique SMILES needed)

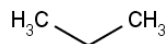
C



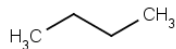
CC



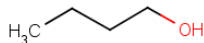
CCC



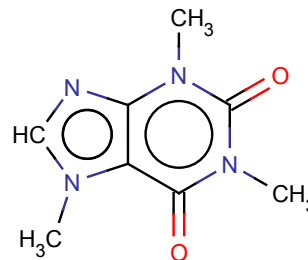
CCCC



CCCCO



CCCCN



[InChI=1/C8H10N4O2/c1-10-4-9-6-5\(10\)7\(13\)12\(3\)8\(14\)11\(6\)2/h4H,1-3H3](#)

**All those SMILES codes represent caffeine**

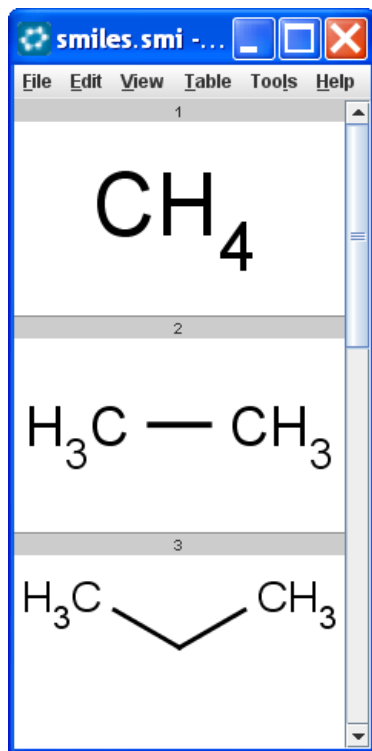
[c]1([n+](c1)[CH3])[c]([c]2([c]([n+]1[CH3])[n][cH][n+]2[CH3]))[O-])[O-]  
CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12  
Cn1cnc2n(c(=O)n(c(=O)c12  
Cn1cnc2c1c(=O)n(c(=O)n2C  
N1(C)C(=O)N(C)C2=C(C1=O)N(C)C=N2  
O=C1C2=C(N=CN2C)N(C(=O)N1C)C  
CN1C=NC2=C1C(=O)N(C)C(=O)N2C

Caffeine SMILES Source [InChI1 FAQ](#)

# SDF/MOL structure format

Positive: established standard format; good for storing structures safely  
can store 3D structure; can store metadata (boiling points, toxicity, mass spectra)

Negative: large file size, need compression



OpenBabel02240823422D

```
1 0 0 0 0 0 0 0 0999 V2000
  0.0000  0.0000  0.0000 C  0  0  0  0  0
M END
$$$$
```

OpenBabel02240823422D

```
2 1 0 0 0 0 0 0 0999 V2000
  0.0000  0.0000  0.0000 C  0  0  0  0  0
  0.0000  0.0000  0.0000 C  0  0  0  0  0
  1  2  1  0  0  0
M END
$$$$
```

OpenBabel02240823422D

```
3 2 0 0 0 0 0 0 0999 V2000
  0.0000  0.0000  0.0000 C  0  0  0  0  0
  0.0000  0.0000  0.0000 C  0  0  0  0  0
  0.0000  0.0000  0.0000 C  0  0  0  0  0
  1  2  1  0  0  0
  2  3  1  0  0  0
M END
$$$$
```

← Creator

← Coordinates for 3D

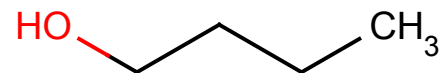
← Connection of atoms

# CML structure format

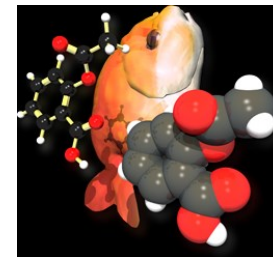
Positive: Open Standard format; good for storing structures safely  
machine readable

Negative: huge files; redundant information; needs compression

```
<?xml version="1.0" ?>
<molecule id="m1">
  <atomArray>
    <atom id="a1" elementType="C"
      x2="2.6673582436560714" y2="0.30800000000000006" />
    <atom id="a2" elementType="C"
      x2="1.3336791218280362" y2="-0.46199999999999997" />
    <atom id="a3" elementType="C"
      x2="4.440892098500626E-16" y2="0.30800000000000016" />
    <atom id="a4" elementType="C"
      x2="-1.3336791218280348" y2="-0.46200000000000002" />
    <atom id="a5" elementType="O"
      x2="-2.6673582436560705" y2="0.30799999999999997" />
  </atomArray>
  <bondArray>
    <bond atomRefs2="a1 a2" order="1" />
    <bond atomRefs2="a2 a3" order="1" />
    <bond atomRefs2="a3 a4" order="1" />
    <bond atomRefs2="a4 a5" order="1" />
  </bondArray>
</molecule>
```



# Tools for chemical structure conversion



Example: Free [OpenBabel](#) – can handle around 100 formats

OpenBabelGUI

File View Help

---- INPUT FORMAT ----

smiles -- SMILES format

Format Info

CONVERT

---- OUTPUT FORMAT ----

sdf -- MDL MOL format

Use this format for all input files (ignore file extensions)

C:\chemistry\mopac7\mopac2007\

Start import at molecule # specified

End import at molecule # specified

Input below (ignore input file)

CCO

Continue with next object after error, if possible

Attempt to translate keywords

Delete hydrogens (make implicit)

Add hydrogens (make explicit)

Add Hydrogens appropriate for pH model

Convert dative bonds e.g. [N+][O-]=O to N(=O)=O

Center Coordinates

Combine mols in first file with others having same name

Convert only molecules matching SMARTS:

Convert only molecules NOT matching SMARTS:

Output file

Output below only (no output file)

1 molecule converted

OpenBabel02240816342D

```
3 2 0 0 0 0 0 0 0999 V2000
0.0000 0.0000 0.0000 C 0 0 0 0 0
0.0000 0.0000 0.0000 C 0 0 0 0 0
0.0000 0.0000 0.0000 O 0 0 0 0 0
1 2 1 0 0 0
2 3 1 0 0 0
M END
$$$$
```

Project Cost

This calculator estimates how much it would cost to hire a team to write this project from scratch. [More >](#)

Include	Code Only
Codebase	210,820 LOC
Effort (est.)	55 Person Years
Avg. Salary	\$ 38000/year
<b>\$2,093,732</b>	

OpenBabel is community developed ( PC,LINUX,MAC)

See also [ChemAxon molconvert](#)

# Handling molecules on your PC – Instant-JChem

**Your Projects**

**Query - Grid view**

Cdid: =

Structure: Substructure

Double click to sketch structure

Chemical Terms Filter:

MolWeight: <

Formula: =

IUPAC name: =

Donors: =

Acceptors: =

Rot bonds: =

DB regid: =

DB name: =

XLogP: <

MOLI  
BioCyc  
NIST Chemistry WebBook  
KEGG

**Data Search**

Clear Show All Search

**Molecule and Metadata**

...	Cdid	Structure	MolWeight	Formula	IUPAC n...	Donors	Acceptors	Rot bonds	DB regid	DB name	XLogP
1	1,001		203.24	C9H17NO4	(2-acetoxy-3-carboxyloxypropyl)-trimethyl-ammonium	0	4	5	(-)o-acetylca mitine	BioCyc	3.41
2	1,002		156.14	C7H8O4	5,6-dihydroxycyclohexa-1,3-diene-1-carboxylic acid	3	4	1	1-2-CIS-DI-O H-BENZOAT E	BioCyc	-0.31
3	1,003		75.11	C3H9NO	1-aminopropan-2-ol	2	2	1	1-AMINO-PR OPAN-2-OL	BioCyc	-0.84
4	1,004		169.07	C3H8NO5P	(3-amino-2-oxo-propoxy) phosphonic acid	3	6	4	1-AMINO-PR OPAN-2-ONE -3-PHOSPHA TE	BioCyc	-0.84
5	1,005		202.55	C6H3ClN2O4	1-chloro-2,4-dinitrobenzene	0	4	0	1-CHLORO-2 4-DINITROBE NZENE	BioCyc	2.33
6	1,006		163.18	C7H9N5	9-ethyl-9H-purin-6-amine				1-ETHYLADENINE		0.09
7	1,007		146.14	C6H12O5	2,3-dihydroxy-3-methylpentanoic acid	2	4	2	1-KETO-2-M ETHYLVALE RATE	BioCyc	0.22

Pubchem demo: 1,000 out of 1,000 rows.

Best way to handle structures on your PC/MAC  
Up to one million molecules ok on slow PC

Download [Instant-JChem](#)

## The Last Page - What is important to remember



There are different exchange formats for mass spectral data

→ netCDF, JCAMP, mzXML

Metadata must be stored together with mass spectra

Mass spectra should be published in machine readable format (not on paper)

Open Data formats for mass spectral data (in XML) are important

There are different exchange formats for chemical structures

→ SMILES, SDF, MOL, PDB, InChIKey, PDB, CML

→ Databases IDs and InChIKeys should be submitted with each profiling report.