MUNI | RECETOX

E2020 – Soft skills II – Information Literacy
3. Databases

01.03.2022

Mgr. Ludovic Mayer
ludovic.mayer@recetox.muni.cz

I want to start working/writing, but …

- What resources to use?
- Where can I find them?
- How can I find what interest me?
- What to do if I don’t find what I need?
- How to formulate a query understood by the database?

Content

- Lecture divided into 2 parts:
  - 1st part: Theoretical – to understand how is the information stored
    – how to access it
  - 2nd part: Practical – you are at the helm and you find the information!

VPN

- MUNI VPN
  - MUNI
  - Eduroam

https://it.muni.cz/en/services/wireless-wi-fi-connection

Wireless Wi-Fi Connection

A majority of Masaryk University's premises is covered by Wi-Fi network enabling internet connection from laptops, tablets, and mobile phones. MU uses *Eduroam*, a world-wide network, which enables internet connection at most of academic institutions around the world based on the principle of unified login.

#1 Eduroam

(Main Network)

UČO[1)]@muni.cz
secondary password[2)]

Eduroam is the main wireless network at MU enabling connection to internet. This network is being used by a variety of academic institutions around the world, so you can often connect to it automatically even on study or business trips.

#2 MUNI

(Auxiliary Network)

UČO[1)]
secondary password[2)]

Be careful, data transfer via this network is not secured! MUNI network functions as an auxiliary network. You should only use it if having trouble with Eduroam network.

Database

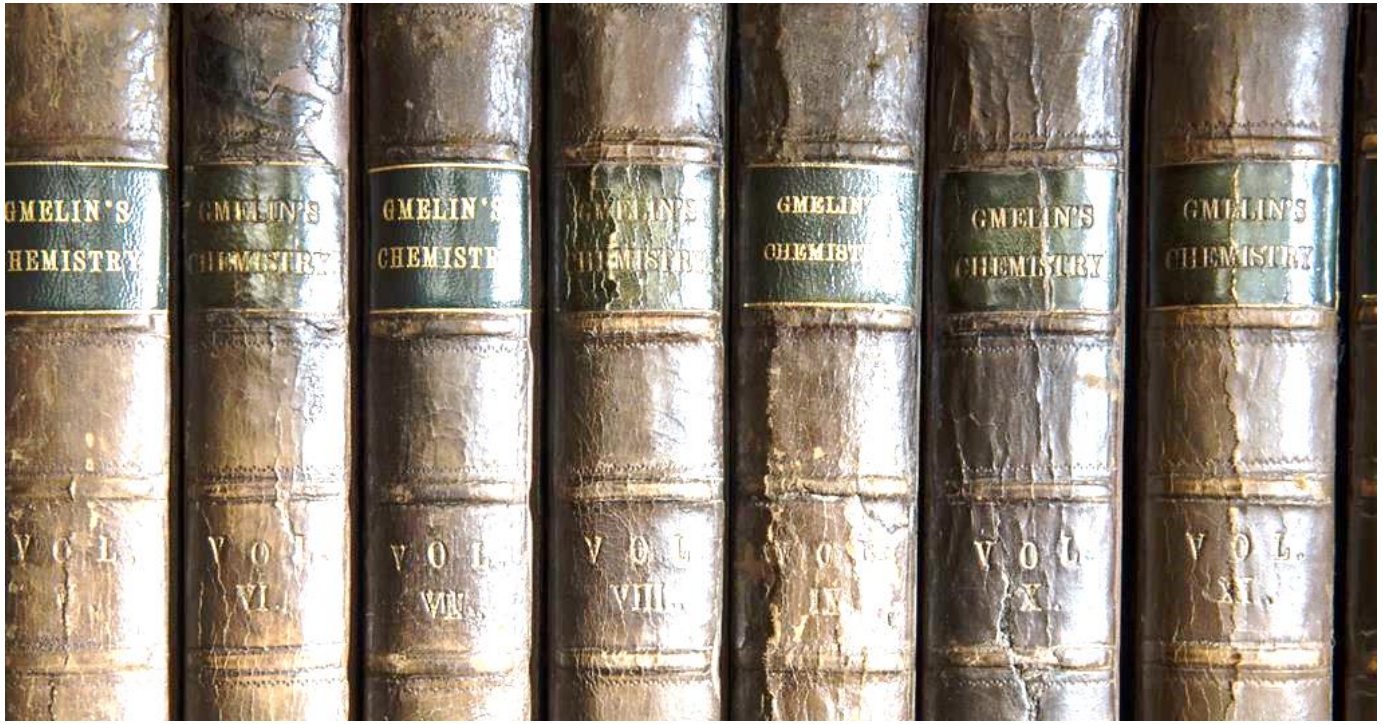
"a usually large collection of data organized especially for rapid search and retrieval (as by a computer)" *(Webster dictionary)*

Exist for multiples types of information

- Scientific Articles and Journals (WoS – Scopus *as seen in 2.Scientometry*)
- Chemical and other information
- Specific properties and parameters

Database: From “paper” to today

- First chemical database
  - First edition published in 1817, last edition: 8th edition 1990’s



Nowadays most of them are digitalized and the information is available online

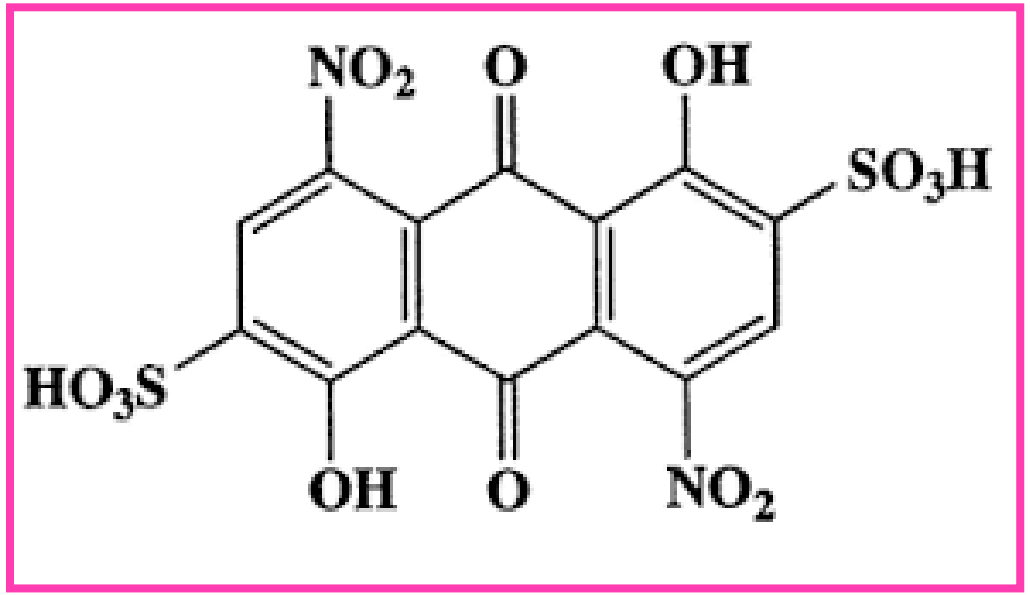
MUNI | RECETOX

Chemical databases

- Multiple existing databases based on different criterias
  - according to chemical structures
  - literature
  - crystallographic
  - spectroscopic: infrared, absorbance, nuclear magnetic resonance, …
  - reactions
  - thermodynamic
  - others, …

Chemical databases

- Often based on **chemical structures searches**
- Chemical structures are easy to read by humans (visual reading, …)



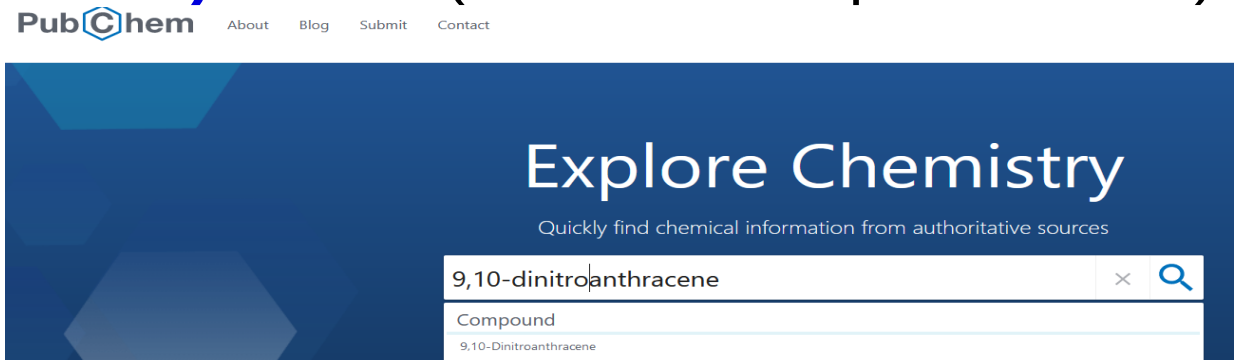
1,5-Dihydroxy-4,8-dinitro-9,10-dioxo-9,10-dihydroanthracene-2,6-disulfonic acid

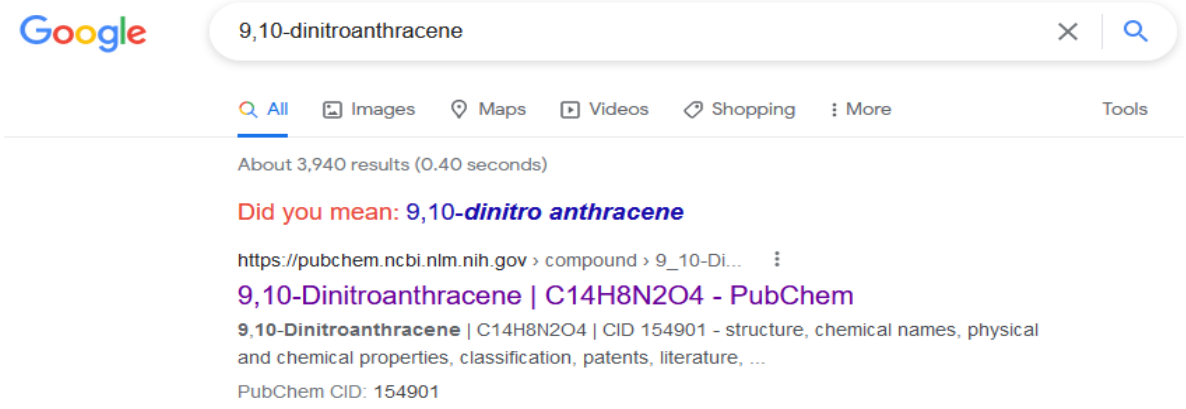
- How to explain it to computer / software / app ?

Chemical structures and their representation

– In order to look for a structure and for the computer to understand

a) **Name** (or similar unique identifier)





b) **Connectivity matrices** (e.g.MDL molfile, PDB, CML, … = Computer languages related to chemistry created specifically for computation

c) **Linear strings**: e.g. SMILES/SMARTS, SLN, WLN, InChi = easy computer language which computer can decipher the structure

MUNI | RECETOX

Chemical name (or similar unique identifier)

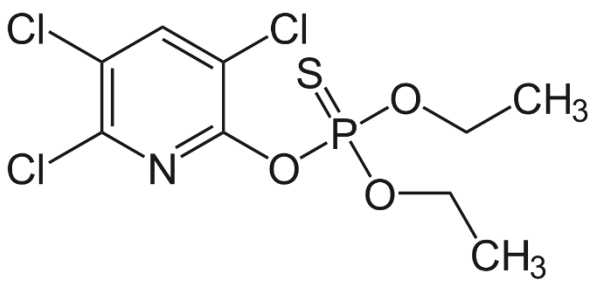
- Each compound can have multiple names
  - Common name (trivial, semi-trivial, systematic, business, …)
    - **Chlorpyrifos** (Insecticide, banned since 2020 in Europe)
  - Proper chemical name: (= IUPAC name, *International Union of Pure and Applied Chemistry*)
    - **0,0-diethyl 0-(3,5,6-trichloro-2-pyridinyl)-phosphorothioate**
  - Other names (sometimes names of commercial products they are featured in)
    - **Chlorpyrifos-ethyl, Brodan, Bolton insecticide, Cobalt, …**

Chemical name (or similar unique identifier)

- In order to remove possible errors and mistakes:
  - A unique numerical identifier was created =
  - **<u>CAS RN</u>** *(Chemical Abstracts Service Registry Number)*
  - Assigned to every chemical substance
  - From 1800's to today, registry account from more than 193 million compounds

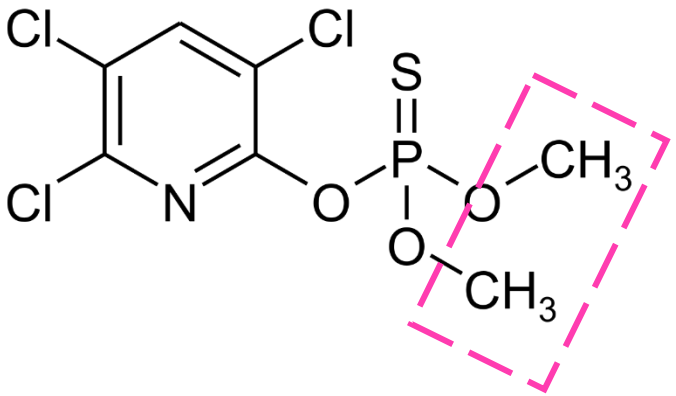
Chemical name (or similar unique identifier)

– Chlorpyrifos: CAS RN = 2921-88-2

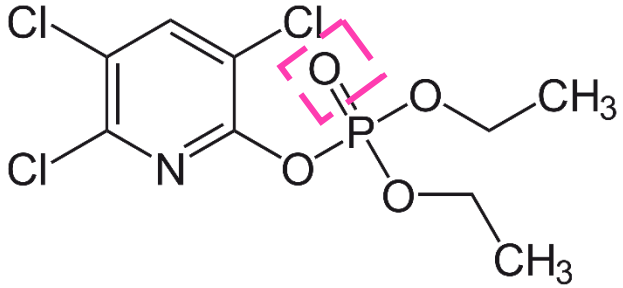


<u>Metabolites and derivatives:</u>

– Chlorpyrifos-methyl = 5598-13-0



– Chlorpyrifos-oxon: 5598-15-2



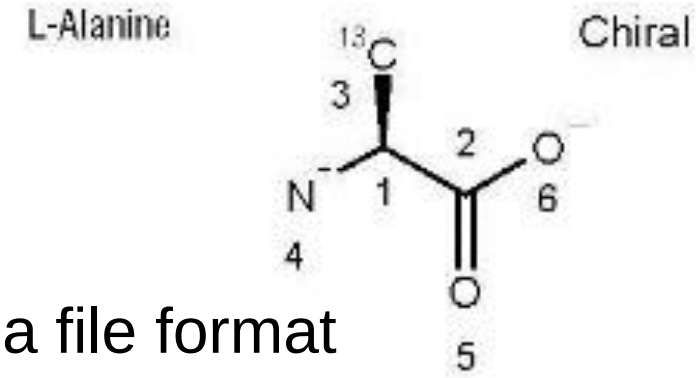
MUNI | RECETOX

Connectivity Matrix

- **Computer language**
- Chemical Table File (CT File)
  - Family of text based chemical file formats that describe molecules and chemical reactions
  - Numerous file format exist

- CT File is an open format
  - Lists each atom in a molecule, with the x-y-z coordinates of that atom, and the bonds amongst atoms
  - Just need to register on this website to access them: https://discover.3ds.com/ctfile-documentation-request-form

Connectivity Matrix



Molfile:

- An MDL Molfile is a file format
- Contains information about:
  - **atoms,**
  - **bonds, (=connectivity)**
  - **charges**
  - **coordinates** of a molecule

- Recognized by most cheminformatics software systems/applications

L-Alanine	**Title line** (can be blank but line must exist)	**Header Block (3 lines)**
ABCDEFGH09071717443D	**Program / file timestamp line** (Name of source program and a file timestamp)	
Exported	**Comment line** (can be blank but line must exist)	
6 5 0 0 1 0 3 V2000	**Counts line**	**Connection table**
-0.6622 0.5342 0.0000 C 0 0 2 0 0 0 0.6622 -0.3000 0.0000 C 0 0 0 0 0 0 -0.7207 2.0817 0.0000 C 1 0 0 0 0 0 -1.8622 -0.3695 0.0000 N 0 3 0 0 0 0 0.6220 -1.8037 0.0000 O 0 0 0 0 0 0 1.9464 0.4244 0.0000 O 0 5 0 0 0 0	**Atom block** (1 line for each atom): x, y, z (in angstroms), element, etc.	
1 2 1 0 0 0 1 3 1 1 0 0 1 4 1 0 0 0 2 5 2 0 0 0 2 6 1 0 0 0	**Bond block** (1 line for each bond): 1st atom, 2nd atom, type, etc.	
M CHG 2 4 1 6 -1 M ISO 1 3 13	**Properties block**	
M END	**END line** (NOTE: some programs don't like a blank line before M END)	**END**

MUNI | RECETOX

Connectivity Matrix

– Same exist for Proteins = **PDB format**

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

Connectivity Matrix without connectivity

X-Y-Z file

- No information about bonds (covalent, hydrogens, VdW, …) admits a greater flexibility

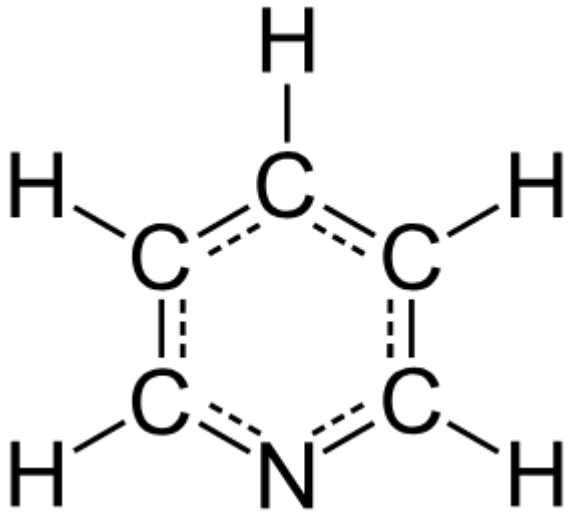
- Typical XYZ format specifies the molecule geometry
  - First line = number of atoms with Cartesian coordinates
  - Second line = a comment
  - Third and following line = atomic coordinates

```
<number of atoms>
comment line
atom_symbol11 x-coord11 y-coord11 z-coord11
atom_symbol12 x-coord12 y-coord12 z-coord12
...
atom_symbol1n x-coord1n y-coord1n z-coord1n
<number of atoms>
comment line
atom_symbol21 x-coord21 y-coord21 z-coord21
atom_symbol22 x-coord22 y-coord22 z-coord22
```

Connectivity Matrix without connectivity

Pyridine

– Formula: C_5H_5N




```
11

C       -0.180226841      0.360945118     -1.120304970
C       -0.180226841      1.559292118     -0.407860970
C       -0.180226841      1.503191118      0.986935030
N       -0.180226841      0.360945118      1.29018350
C       -0.180226841     -0.781300882      0.986935030
C       -0.180226841     -0.837401882     -0.407860970
H       -0.180226841      0.360945118     -2.206546970
H       -0.180226841      2.517950118     -0.917077970
H       -0.180226841      2.421289118      1.572099030
H       -0.180226841     -1.699398882      1.572099030
H       -0.180226841     -1.796059882     -0.917077970
```


MUNI | RECETOX

Linear string: SMILES

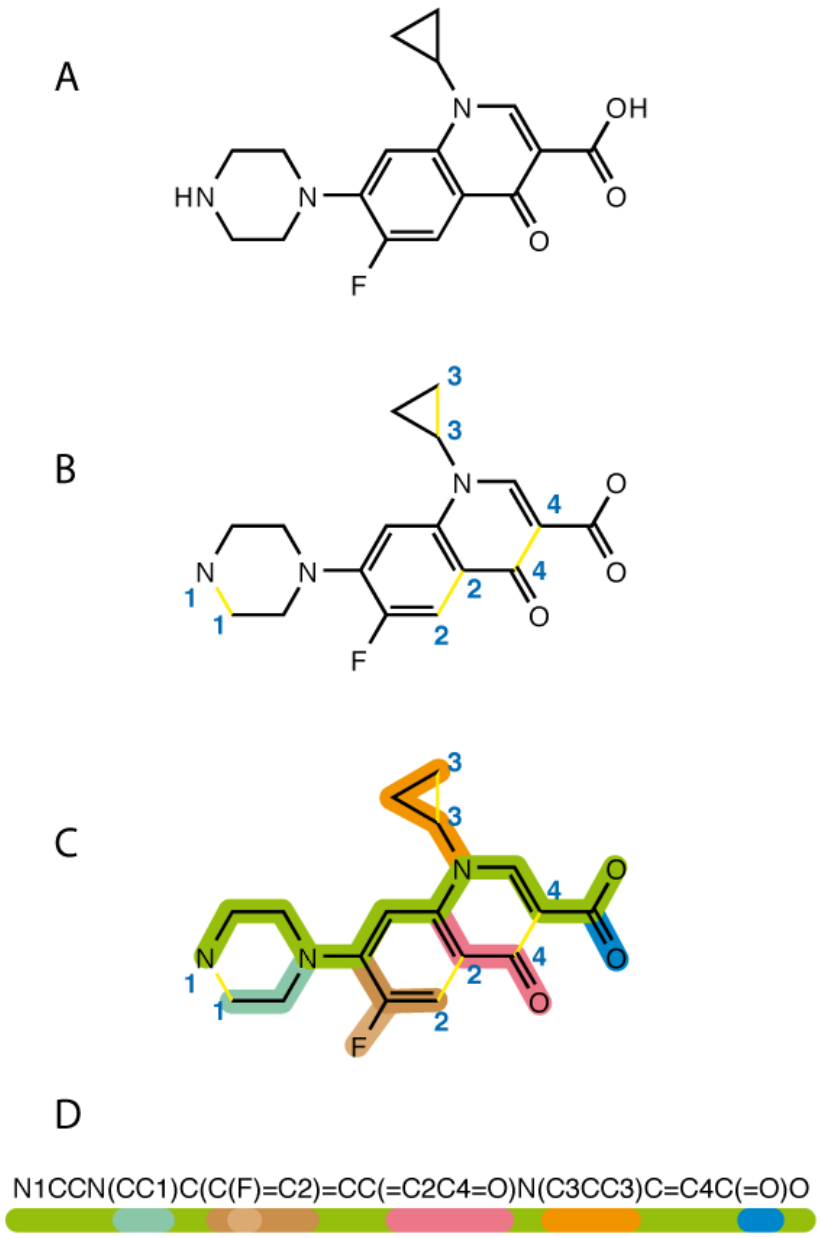
<u>Linear string</u>: represents structures as a linear string of characters

Simplified Molecular Input Line Entry Specification (SMILES)

- Chemical notation allowing user to represent a chemical structure
- Easily read, understood and used by computer
- Contains connectivity, but no longer 2D or 3D coordinates

Linear string: SMILES

- How does it work?
  - Every atoms are supported
  - Upper-case for aromatic atoms, lower-case for non-aromatic atoms
  - Bonds:
    - single bond
    - = double bond
    - # triple bond
    - * aromatic bond
    - . disconnected structures

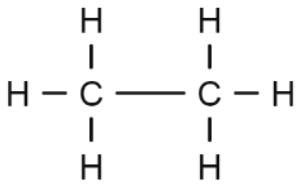
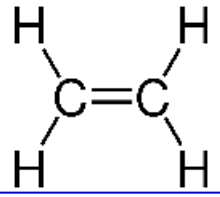
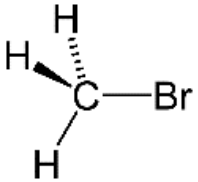


MUNI | RECETOX

Linear string: SMILES

Simple chain molecule

(Hydrogen suppressed = no need to put hydrogen in it, software understand that they are here)

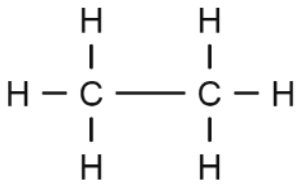
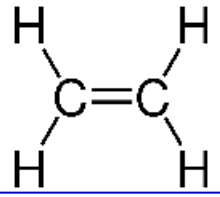
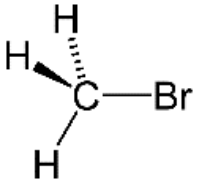
SMILES	Formula	Name	Structure
CC	CH_3CH_3	Ethane	
C=C	CH_2CH_2	Ethene	
	CH_3Br	Bromomethane	

MUNI | RECETOX

Linear string: SMILES

Simple chain molecule

(Hydrogen suppressed = no need to put hydrogen in it, software understand that they are here)

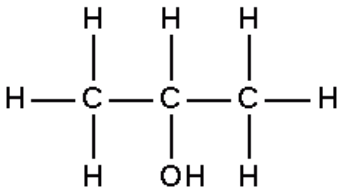
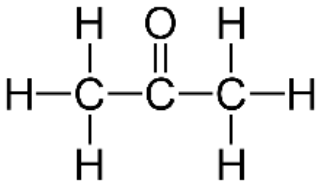
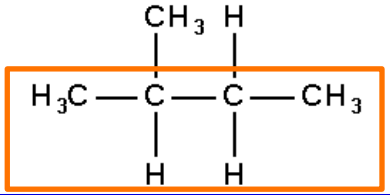
SMILES	Formula	Name	Structure
CC	CH_3CH_3	Ethane	
C=C	CH_2CH_2	Ethene	
CBr	CH_3Br	Bromomethane	

MUNI | RECETOX

Linear string: SMILES

<u>Branches</u>

(in parentheses = a branche placed right after the atom it is connected to)

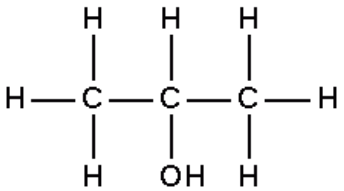
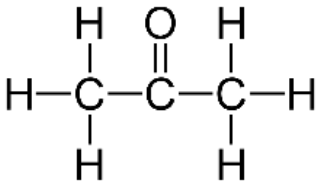
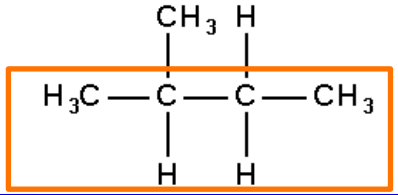
SMILES	Formula	Name	Structure
CC(O)C	$CH_3CHOHCH_3$	2-propanol	
CC(=O)C	CH_3COCH_3	2-propanone (acetone)	
	$CH_3CH_3CHCH_2CH_3$	2-methylbutane (Isopentane)	

MUNI | RECETOX

Linear string: SMILES

<u>Branches</u>

(in parentheses = a branche placed right after the atom it is connected to)

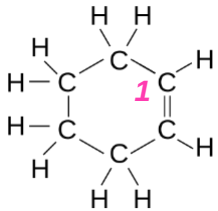
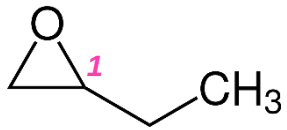
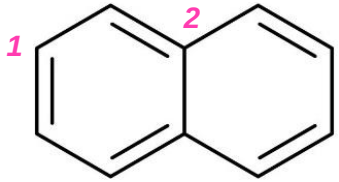
SMILES	Formula	Name	Structure
CC(O)C	$CH_3CHOHCH_3$	2-propanol	
CC(=O)C	CH_3COCH_3	2-propanone (acetone)	
CCC(C)C	$CH_3CH_3CHCH_2CH_3$	2-methylbutane (Isopentane=	

MUNI | RECETOX

Linear string: SMILES

<u>Rings:</u>

(Use number to identify opening and closing of ring atom)

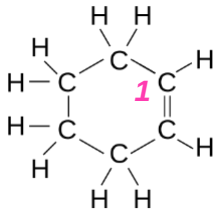
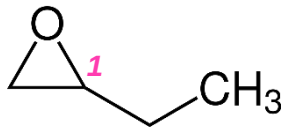
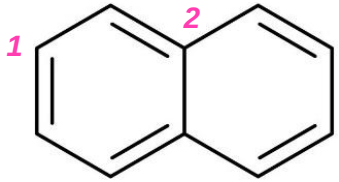
SMILES	Formula	Name	Structure
C=1CCCCC1 Also C*1*C*C*C*C*C1	$CHCHCH_2CH_2CH_2CH$	Cyclohexene	
	$CH_2(O)CHCH_2CH_3$	Ethyloxirane	
	CHCHCHCHCHCHCHCHCH	Naphtalene	

MUNI | RECETOX

Linear string: SMILES

<u>Rings:</u>

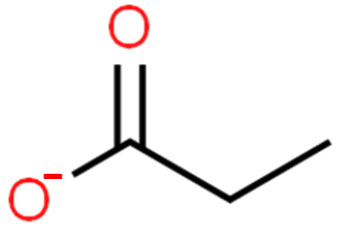
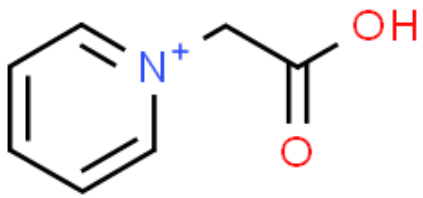
(Use number to identify opening and closing of ring atom)

SMILES	Formula	Name	Structure
C=1CCCCC1 Also C*1*C*C*C*C*C1	$CHCHCH_2CH_2CH_2CH$	Cyclohexene	
C1OC1CC	$CH_2(O)CHCH_2CH_3$	Ethyloxirane	
c1cc2ccccc2cc1	CHCHCHCHCHCHCHCHCH	Naphtalene	

Linear string: SMILES

<u>Charged atoms:</u>

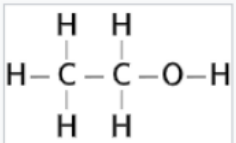
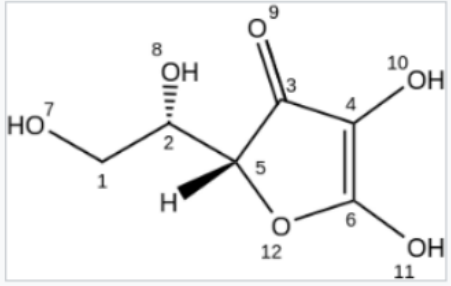
(Atoms followed by brackets which enclose the charge on the atom, maybe be explicitly stated ({-1}) or not ({-}))

SMILES	Name	Structure
CCC(=O)O{-1} Or CCC(=O)O{-}	Ionised form of propanoic acid	
c1ccccn{+1}1CC(=O)O	1-Carboxylmethyl pyridinium	

Linear string: InChI

InChI

- **In**ternational **Ch**emical **I**dentifier
- Introduced by IUPAC as a **standard** in 2006
- Contains different layers:
  - general formula,
  - hydrogens, charges,
  - stereochemistry,
  - isotopes,…

Structural formula	standard InChI
 ethanol	InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3
 L-ascorbic acid	InChI=1S/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

MUNI | RECETOX

Representation of structure - Conclusions

- Chemical names – connectivity matrices – linear strings
- Multiple choice exist, none of them wrong, some more popular than others

File Extension	MIME Type	Proper Name	Description
alc	chemical/x-alchemy	Alchemy Format	
csf	chemical/x-cache-csf	CAChe MolStruct CSF	
cbin, cascii, ctab	chemical/x-cactvs-binary	CACTVS format	
cdx	chemical/x-cdx	ChemDraw eXchange file	
cer	chemical/x-cerius	MSI Cerius II format	
c3d	chemical/x-chem3d	Chem3D Format	
chm	chemical/x-chemdraw	ChemDraw file	
cif	chemical/x-cif	Crystallographic Information File, Crystallographic Information Framework	Promulgated by the International Union of Crystallography
cmdf	chemical/x-cmdf	CrystalMaker Data format	
cml	chemical/x-cml	Chemical Markup Language	XML based Chemical Markup Language.
cpa	chemical/x-compass	Compass program of the Takahashi	
bsd	chemical/x-crossfire	Crossfire file	
csm, csml	chemical/x-csml	Chemical Style Markup Language	
ctx	chemical/x-ctx	Gasteiger group CTX file format	
cxf, cef	chemical/x-cxf	Chemical eXchange Format	
emb, embl	chemical/x-embl-dl-nucleotide	EMBL Nucleotide Format	
spc	chemical/x-galactic-spc	SPC format for spectral and chromatographic data	

Questions?

– Contact me anytime via email: ludovic.mayer@recetox.muni.cz