

# Pokročilá chemoinformatika

Databáze, chemický prostor  
únor 2017

# Chemické databáze

# Chemické databáze

---

- Informace o molekulách, struktury molekul, vlastnosti, aktivity, ...
- **PubChem**
- **DrugBank**
- **ZINC**
- **ChEMBL**
- **ChemSpider**
- **PHYSPROP**  
<http://esc.syrres.com/fatepointer/search.asp>

# PubChem

The image shows a screenshot of the PubChem website in a browser window. The browser's address bar displays "pubchem.ncbi.nlm.nih.gov". The website's navigation menu includes "Databases", "Upload", "Services", "Help", "more", and "Today's Statistics". The main header features the "PubChem" logo. Below the logo are three tabs: "BioAssay", "Compound", and "Substance". A search bar is present with a "Go" button and links to "Limits" and "Advanced". A sidebar on the right contains various tool links: "BioAssay Tools", "Structure Search", "3D Conformer Tools", "Structure Clustering", "Classification", "Upload", "Download", and "PubChem FTP". A central message promotes the "PubChem Search Beta". A news banner at the bottom states: "New PubChem now uses the latest IUPAC recommendations for atomic mass and isotopic composition information. Read more...".

pubchem.ncbi.nlm.nih.gov

Databases > Upload Services > Help more > Today's Statistics >

PubChem

BioAssay ? Compound ? Substance ?

Go Limits Advanced

Try the [PubChem Search Beta](#)

**New** PubChem now uses the latest IUPAC recommendations for atomic mass and isotopic composition information. [Read more...](#)

more ... 

BioAssay Tools

Structure Search

3D Conformer Tools

Structure Clustering

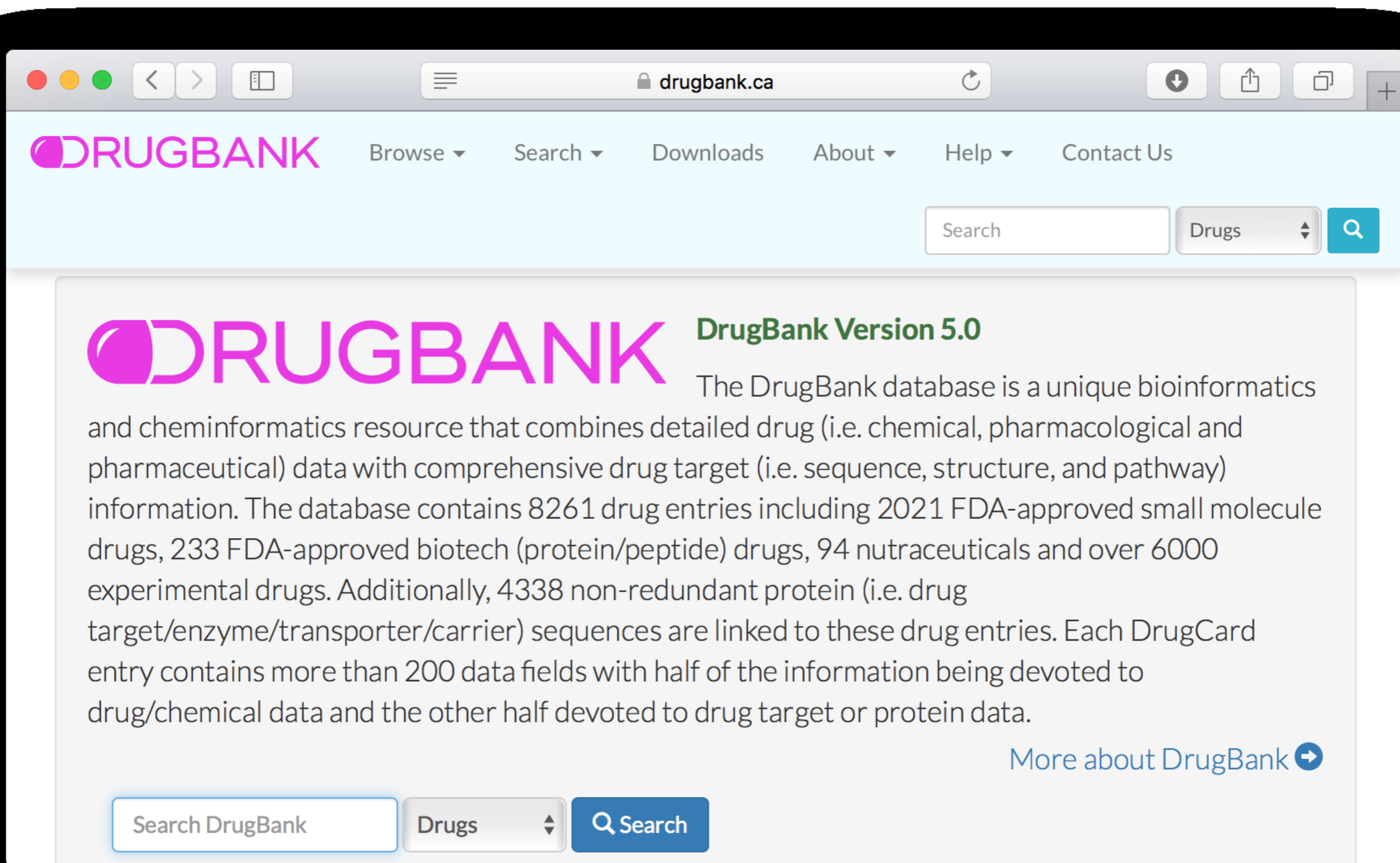
Classification

Upload

Download

PubChem FTP

# DrugBank



The image shows a screenshot of the DrugBank website homepage. The browser address bar displays "drugbank.ca". The navigation menu includes "Browse", "Search", "Downloads", "About", "Help", and "Contact Us". A search bar is present with a dropdown menu set to "Drugs" and a search icon. The main content area features the DrugBank logo and the heading "DrugBank Version 5.0". Below this, a paragraph describes the database as a unique bioinformatics and cheminformatics resource. At the bottom, there is a secondary search bar and a link to "More about DrugBank".

**DRUGBANK** **DrugBank Version 5.0**

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains 8261 drug entries including 2021 FDA-approved small molecule drugs, 233 FDA-approved biotech (protein/peptide) drugs, 94 nutraceuticals and over 6000 experimental drugs. Additionally, 4338 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

[More about DrugBank](#)

Search DrugBank Drugs Search

# ZINC

The screenshot shows the ZINC12 website interface. At the top, there is a browser window with the URL `zinc.docking.org`. The header includes the UCSF logo and navigation links: "University of California, San Francisco | About UCSF | Search UCSF | UCSF Medical Center". On the right, there is a "Shoichet Laboratory" logo and a "Not Authen" status. The main title "ZINC<sup>12</sup>" is prominently displayed. Below the title, there is a navigation menu with "About", "Search", "Subsets", "Help", and "Social" links, along with a "G+1 81" button and a "Quick Search Bar...". A message box on the left reads: "Please consider switching to [ZINC15](#), which is superior to ZINC12 in most ways. If you prefer ZINC12 after trying ZINC15, we would like to know why so that we can get you to make the switch. Read more (coming soon)". The main content area contains a welcome message: "Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the [Irwin](#) and [Shoichet](#) Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). To cite ZINC, please reference: Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model.* 2012 DOI: [10.1021/ci3001277](#). The original publication is Irwin and Shoichet, *J. Chem. Inf. Model.* 2005;45(1):177-82 [PDF](#), [DOI](#). We thank [NIGMS](#) for financial support (GM71896).". On the right, there is a section titled "Molecule of the Hour" with a chemical structure diagram of a complex molecule featuring a piperidine ring, a pyridine ring, and a central carbon atom bonded to a nitrogen atom and a double bond.

UCSF University of California, San Francisco | About UCSF | Search UCSF | UCSF Medical Center

Shoichet Laboratory

# ZINC<sup>12</sup>

Not Authen

Active cart: Temp

About Search Subsets Help Social G+1 81 Quick Search Bar...

Please consider switching to [ZINC15](#), which is superior to ZINC12 in most ways. If you prefer ZINC12 after trying ZINC15, we would like to know why so that we can get you to make the switch. Read more (coming soon)

Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the [Irwin](#) and [Shoichet](#) Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). To cite ZINC, please reference: Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model.* 2012 DOI: [10.1021/ci3001277](#). The original publication is Irwin and Shoichet, *J. Chem. Inf. Model.* 2005;45(1):177-82 [PDF](#), [DOI](#). We thank [NIGMS](#) for financial support (GM71896).

## Molecule of the Hour

ZINC ID, Drug Name, SMILES, Catalog, Vendor Code, Target & more Go

Structure/Draw Physical Properties Catalogs & Vendors ZINC IDS Targets Rings Combination

# ChEMBL

The screenshot displays the ChEMBL website interface within a browser window. The browser's address bar shows the URL `ebi.ac.uk`. The page header includes the EMBL-EBI logo and the ChEMBL logo. A breadcrumb trail indicates the current location: `EBI > Databases > Small Molecules > ChEMBL Database > Home`.

The main navigation area features a search bar labeled "Search ChEMBL..." and several filter tabs: "Compounds", "Targets", "Assays", and "Documents". Below these are more specific search options: "Ligand Search", "Target Search", "Browse Targets", "Browse Drugs", "Browse Drug Targets", and "Browse Indicators".

A sidebar on the left contains a list of links for various services and data sources, including "ChEMBL", "Downloads", "UniChem", "SureChEMBL", "Malaria Data", "ChEMBL-NTD", "ADME SARfari", "Web Services", "myChEMBL", "EBI RDF Platform", "FAQ", "Web status page", and "Funding".

The central content area is dominated by a chemical structure viewer. The viewer's toolbar includes icons for file operations (open, save, copy, paste), navigation (back, forward), zooming (in, out, reset), and other functions like 2D/3D toggles and a help icon. The main canvas shows a faint chemical structure, and a vertical legend on the right lists the elements H, C, N, O, S, and F.

# ChemSpider

The screenshot shows the ChemSpider website in a browser window. The address bar displays "chemspider.com". The page features a blue header with the ChemSpider logo and the tagline "Search and share chemistry". Navigation links for "Simple", "Structure", "Advanced", and "History" are visible. A search bar is prominently displayed with the placeholder text "Search" and a magnifying glass icon. Below the search bar, a list of search criteria is provided: "Systematic Name, Synonym, Trade Name, Registry Number, SMILES, InChI or CSID". The page also includes a "What is ChemSpider?" link and a "Search by chemical names" link. In the bottom right corner, there is a "Give Feedback" button with the RSC logo and the text "Generate Leads".

chemspider.com

ROYAL SOCIETY OF CHEMISTRY

# ChemSpider

Search and share chemistry

Simple Structure Advanced History

## Search ChemSpider

Matches any text strings used to describe a molecule.

Systematic Name, Synonym, Trade Name, Registry Number, SMILES, InChI or CSID ?

[What is ChemSpider?](#) [Search by chemical names](#)

Give Feedback  
Generate Leads

Open "www.chemspider.com" in a new tab

ChemSpider is a free chemical structure database. Systematic names



# PHYSPROP

The screenshot shows a web browser window with the URL `esc.syrres.com`. The page features the SRC logo with the tagline "Redefining possible" and a navigation menu with "Defense" and "Environment". The main heading is "FatePointers Search Module".

**Search**  
Search below by CAS RN#, substance name, SMILES, or chemical structure with ChemS3. Data sources may be omitted from the search with the buttons to the right.

Enter CAS RN#

or Substance Name (use \* as wildcard)

or SMILES Notation (use \* as wildcard)

**Sources**  
[\(Overview of available sources\)](#)

<input type="checkbox"/> Physprop	<input type="checkbox"/> Arizona Aquasol
<input type="checkbox"/> ClogP BioByte	<input type="checkbox"/> USDA PestProp
<input type="checkbox"/> HPVIS	<input type="checkbox"/> CHRIP (METI)
<input type="checkbox"/> OPP Inerts	<input type="checkbox"/> OPP Fate
<input type="checkbox"/> OPP RED	<input type="checkbox"/> U Minn BBD
<input type="checkbox"/> HSDB	<input type="checkbox"/> NIST
<input type="checkbox"/> HSDB Fate Data	<input type="checkbox"/> EFDB
<input type="checkbox"/> OECD SIDS	

**Substructure search using ChemS3**

Exact Search

# PDB databáze

EMBL-EBI

Services Research Training About us

Protein Data Bank in Europe

Examples: [hemoglobin](#), [BRCA1\\_HUMAN](#) [EMsearch](#)

PDBe home Deposition PDBe services PDBe training Documentation About PDBe Share Feedback

PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. [Read more about PDBe.](#)

## Featured structure

### Ebola GP fusion loop

1st January 2017

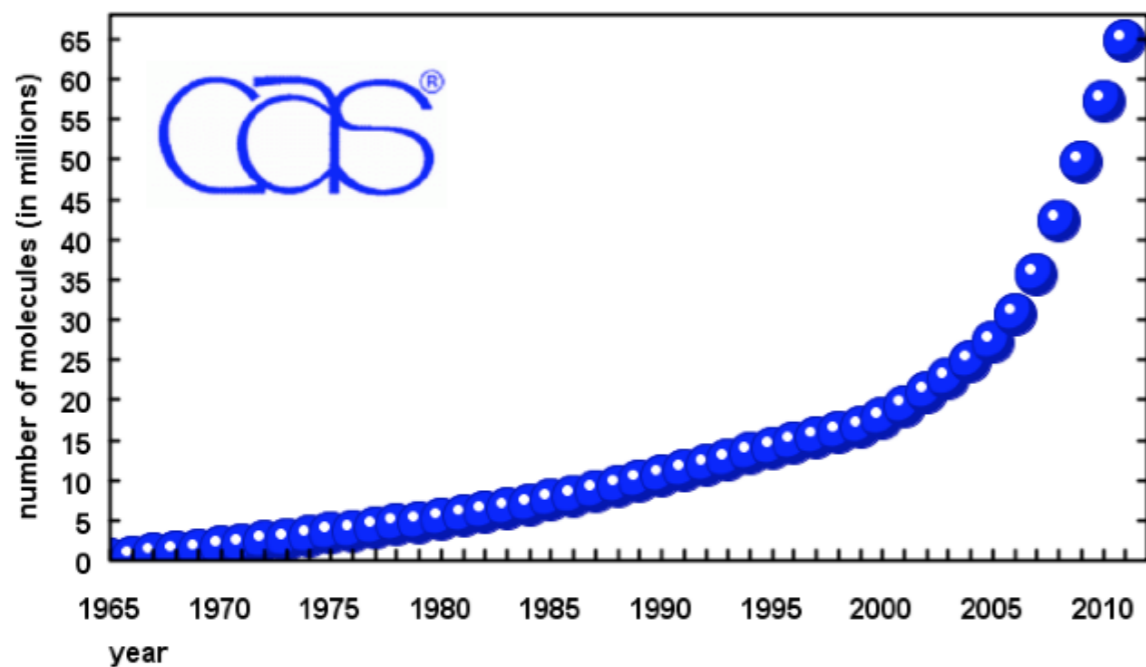


This featured structure explores the molecule from PDBe's 2017 calendar for January. The image shows a fragment of a protein from Ebola virus, based on PDB entry 2m5f, which is critical for the virus to infect its host.

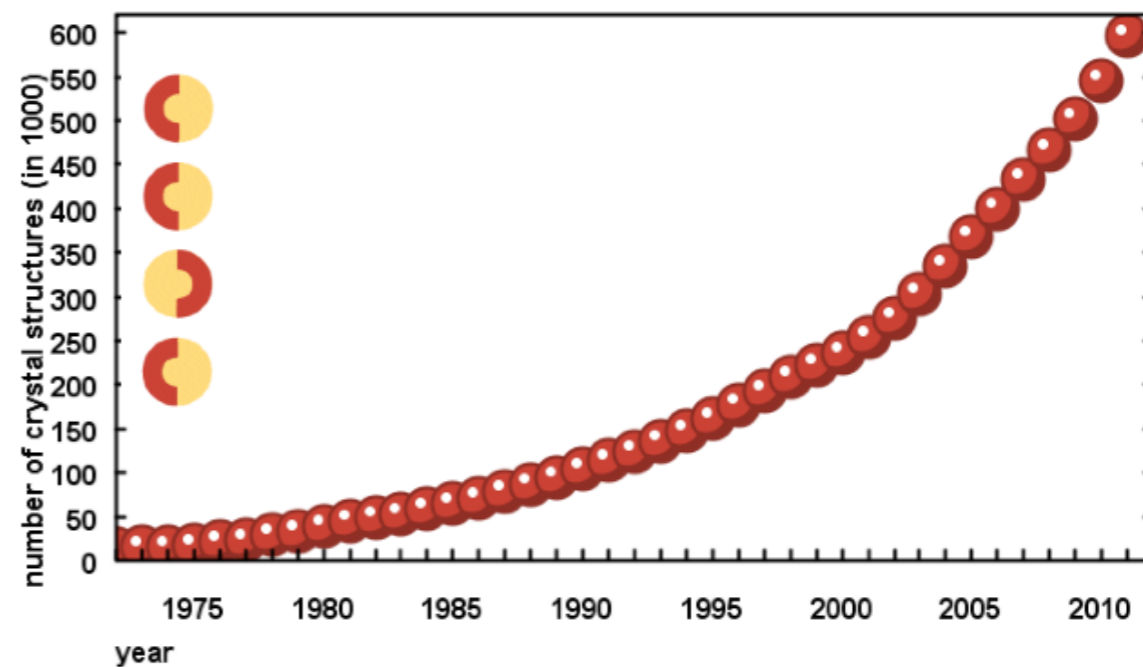
## Popular

- [EMsearch](#)
- [PDBeFold](#)
- [PDBePISA](#)
- [Sequence search](#)
- [PDBe REST API](#)
- [EM resources](#)
- [NMR resources](#)
- [EMPIAR](#)
- [Coordinate Server](#)
- [News](#)
- [Events](#)
- [Training](#)
- [Contact us](#)

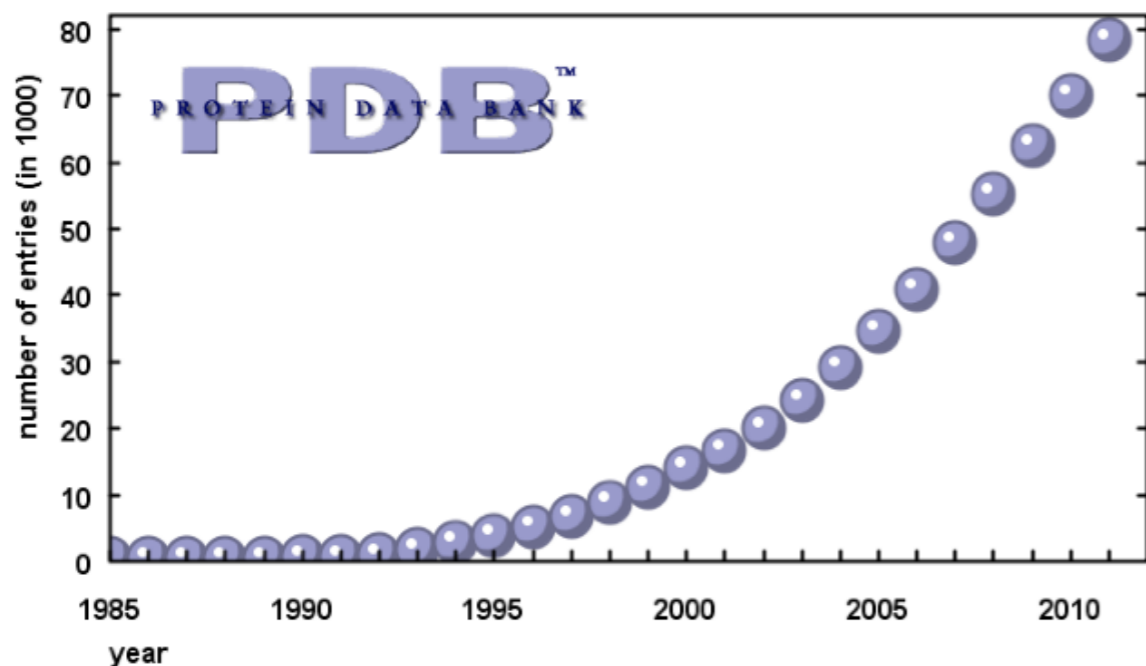
# Velikost základních chemických databází



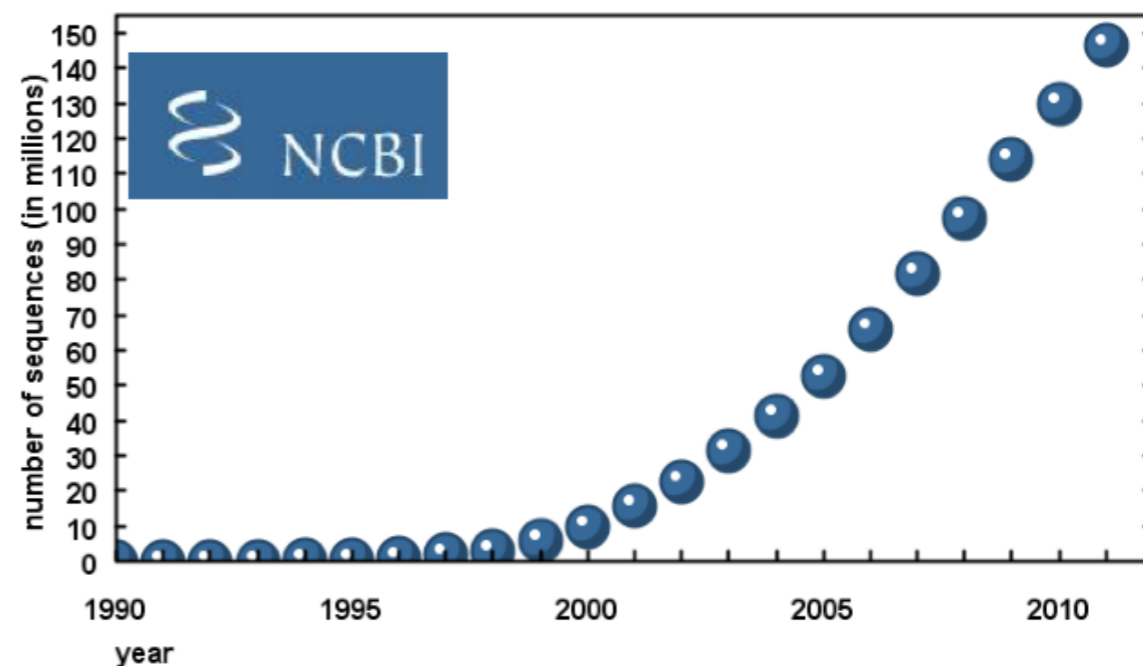
CAS – 65 million molecules



CCDC – 600'000 structures

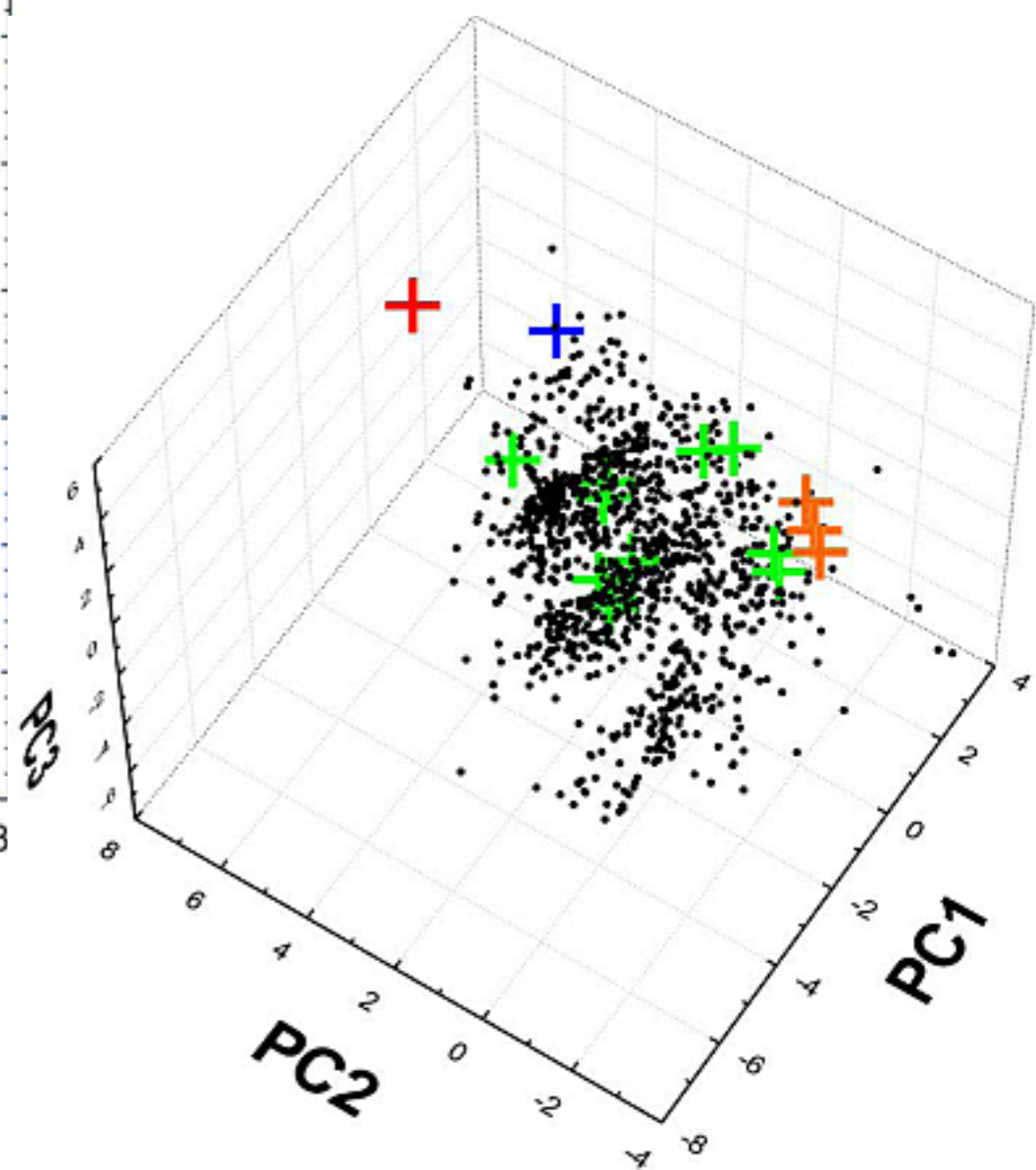
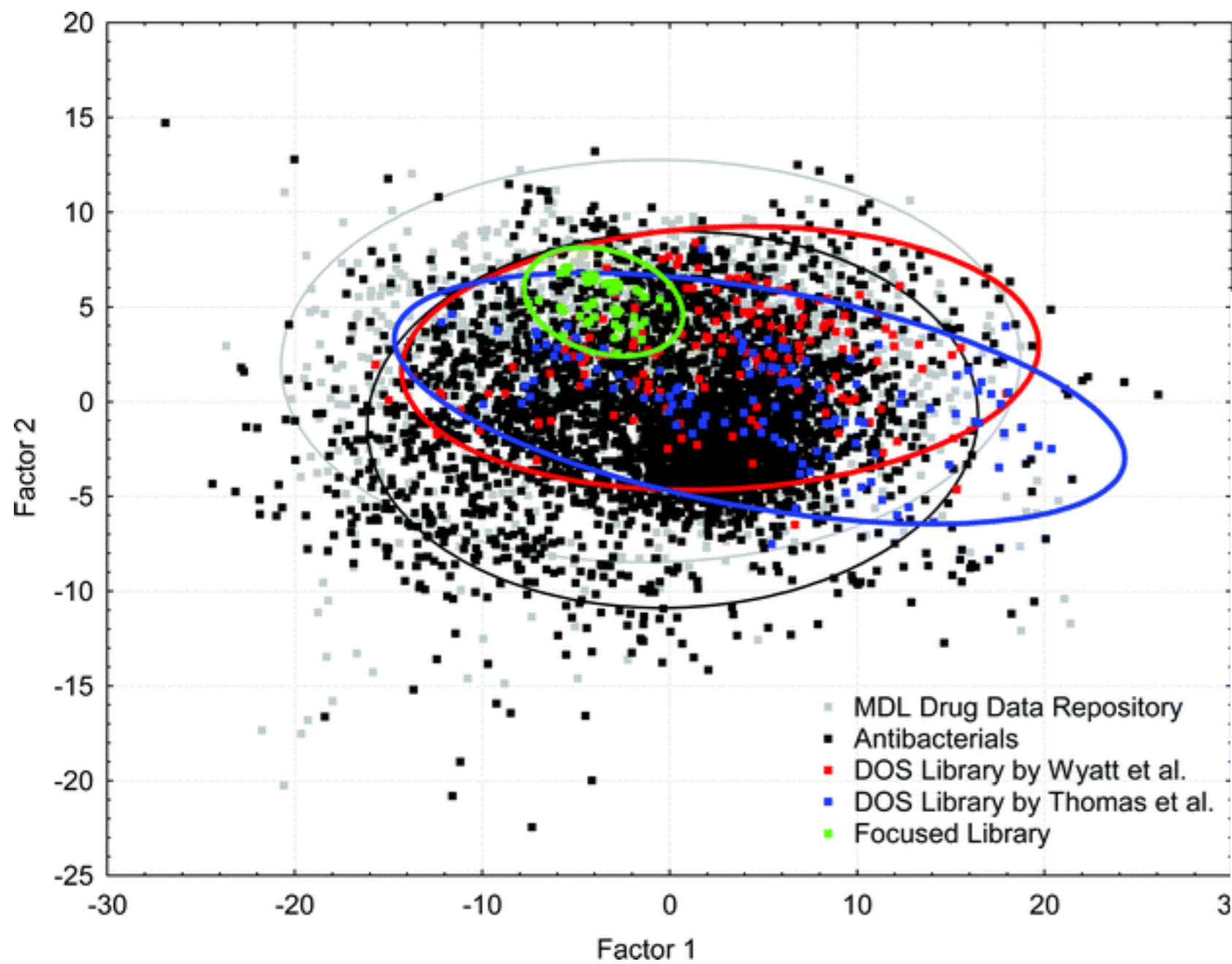


PDB – 78'000 proteins

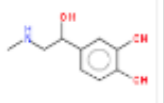
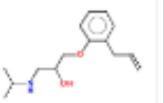
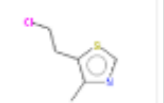
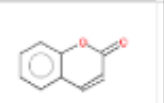
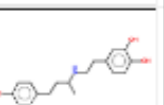


GenBank – 145 million sequences

# Chemický prostor (chemical space)



# Práce s chemickým prostorem

Molecule	logP	PSA	atoms	MTW	...
	-0.06	72.7	13	183.2	...
	2.58	41.5	18	249.3	...
	2.11	12.9	9	161.6	...
	2.01	20.2	11	146.1	...
	3.31	78.8	30	425.9	...

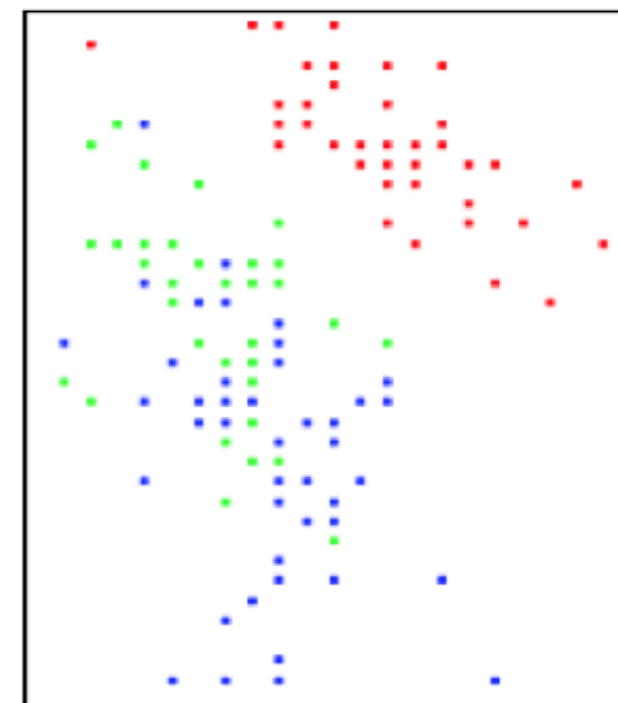


table with properties or fragments

dimensionality reduction

visualization

Formáty pro ukládání chemických,  
chemoinformatických a bioinformatických dat

Alchemy, Boogie, Cambridge CADPAC, Chem3D Cartesian 1, CSD CSSR, CSD GSTAT, Free Form Fractional, Gaussian Z-Matrix, Hyperchem HIN, Mac Molecule, Micro World, MM2 Ouput, MMADS, MOLIN, Mopac Internal, PC Model, Quanta, Spartan, Spartan Mol, Sybyl Mol2, Maccs 2d, UniChem XYZ, XED, AMBER PREP, Biosym , Cacao Cartesian, CHARMM, Chem3D Cartesian 2, CSD FDAT, Feature, GAMESS Output, Gaussian Output, MDL Isis, Macromodel, MM2 Input, MM3, MDL MOLfile, Mopac Cartesian, Mopac Output, PDB, ShelX, Spartan Semi-Empirical, Sybyl Mol, Conjure, Maccs 3d, XYZ

# Formáty pro ukládání struktury molek

---

- MOL (V2000, V3000), SDF  
<http://c4.cabrillo.edu/404/ctfile.pdf>
- MOL2
- PDB, mmCIF
- XYZ
- Smiles a InChI, InChIKey
- ASN.1 (textový a binární formát pro molekuly v PubChemu)



# Formáty pro ukládání informací

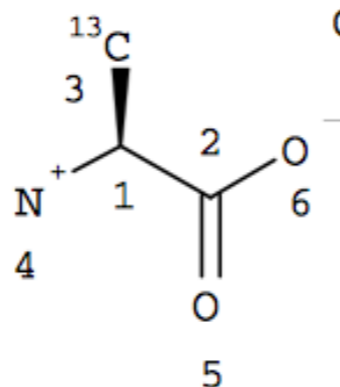
---

- SDF
- CSV
- XML

# MOL (V2000)

L-Alanine

Chiral



```

6  5  0  0  1  0
-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
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
M  CHG  2  4  1  6  -1
M  ISO  1  3  13
M  END

```

Blocks not used in this  
Ctab: Atom List, Stext

Counts Line

Atom Block

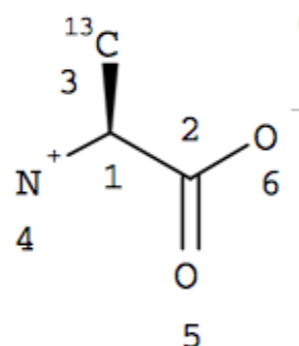
Bond Block

Properties Block

Connection  
Table (Ctab)

# MOL (V3000)

L-Alanine



L-Alanine

GSMACCS-II07189510252D 1 0.00366 0.00000 0

Figure 1, J. Chem. Inf. Comput. Sci., Vol 32, No. 3., 1992

0 0 0 0 0 999 V3000

M V30 BEGIN CTAB

M V30 COUNTS 6 5 0 0 1

M V30 BEGIN ATOM

M V30 1 C -0.6622 0.5342 0 0 CFG=2

M V30 2 C 0.6622 -0.3 0 0

M V30 3 C -0.7207 2.0817 0 0 MASS=13

M V30 4 N -1.8622 -0.3695 0 0 CHG=1

M V30 5 O 0.622 -1.8037 0 0

M V30 6 O 1.9464 0.4244 0 0 CHG=-1

M V30 END ATOM

M V30 BEGIN BOND

M V30 1 1 1 2

M V30 2 1 1 3 CFG=1

M V30 3 1 1 4

M V30 4 2 2 5

M V30 5 1 2 6

M V30 END BOND

M V30 END CTAB

M END

Header Block

← Comment Line

Counts Line

Atom Block

Bond Block

Connection  
Table (Ctab)

Blocks not used in this Ctab:  
Sgroup block, Rgroup block, 3D block

# Simplified molecular-input line-entry system SMILES

---

- vodíky (které mohou být snadno dopočítány, například v alkanech) se v notaci vynechávají a dopočítávají se
- atomy jdoucí za sebou jsou spojeny jednoduchou vazbou  
*příklady:* C (metan), CC (ethan), ..., CO (H<sub>3</sub>COH, methanol)
- dvojná vazba je znázorněna "=" a trojná "#"  
*příklady:* C=C (ethen), C=O (formaldehyd), C#C (ethyn), C#N (kyano)
- pomocí závorek "("") znázorňujeme větvení  
X(YW)Z... - na X je jednoduchou vazbou navázáno Y a Z, Y a W jsou spojeny jednoduchou vazbou, mezi Y nebo W a Z není žádná vazba  
*příklady:* CC(CC)CCC (2-ethylpentan), CC(Cl)C (2-chloropropan), CC(=O)C (aceton)
- pomocí čísel jsou označovány kruhy: C1 .... C1 (začátek a konec kruhu)  
*příklady:* C1CCCCC1 (cyklohexan), C1OC1 (oxiran)
- malými písmeny označujeme aromatické atomy  
*příklady:* c1ccccc1 (benzen), n1ccccc1 (pyridin)
- [NH] explicitně vyjádřený vodík, [O-] vyjádřený ion, [C@@H] vyznačená chiralita, ...
- tutoriál v angličtině:



# International Chemical Identifier InChI & InChIKey

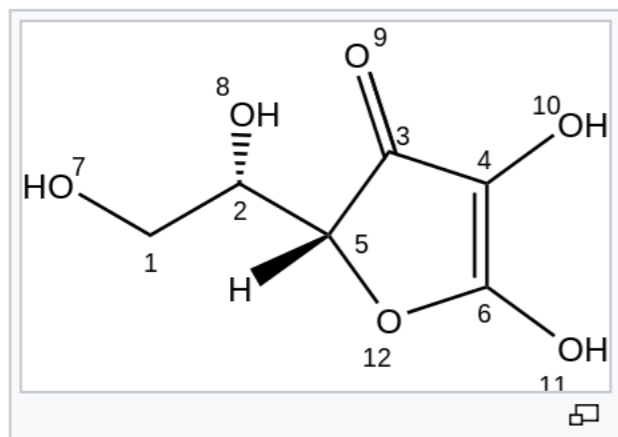
---

- Podobně jako SMILES se jedná o textový zápis molekuly, který se skládá z několika vrstev, které nemusí být vždy všechny zastoupeny

CH<sub>3</sub>CH<sub>2</sub>OH  
ethanol

InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3

InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3 (standard InChI)



L-ascorbic acid

InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

InChI=1S/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-8,10-11H,1H2/t2-,5+/m0/s1 (standard InChI)

- INChIkey je pak hash INChI  
z InChIKey nelze zpětně vytvořit InChI!

Nástroje pro práci se strukturami molekul

# OpenBabel

---

- Chemoinformatický nástroj pro práci s různými formáty molekul a dalšími pomocnými nástroji
- <https://openbabel.org>
- *Pro práci na wolfech použijte:*  
`module add openbabel`

# OpenBabel - konverze různých formátů

---

- Spouštíme v příkazové řádce pomocí **obabel** nebo **babel**
- Seznam podporovaných formátů  
**babel -L formats**
- Převod struktury mezi různými formáty (2.31)  
**obabel -ixxx molecule.xxx -oyyy -O molecule.yyy**  
kde **xxx** je vstupní a **yyy** je výstupní formát molekuly
- Převod struktury mezi různými formáty (<2.31)  
**obabel -ixxx molecule.xxx -oyyy molecule.yyy**



# OpenBabel - konverze různých formátů (windows)

The screenshot displays the OpenBabelGUI application window. The interface is divided into several sections:

- INPUT FORMAT:** A dropdown menu is set to "sdf -- MDL MOL format". Below it, the file path "C:\Users\virtualbox\Downloads\CID\_89594.sdf" is entered. A "CONVERT" button is prominently displayed in the center.
- OUTPUT FORMAT:** A dropdown menu is set to "smi -- SMILES format".
- Output file:** A section with a checked option "Output below only (no output file)".
- Conversion Options:** A central panel with various checkboxes for conversion settings, such as "Delete hydrogens (make implicit)", "Add hydrogens (make explicit)", and "Convert dative bonds".
- Input Data:** A text area on the left contains the SDF file content, including the molecule name "89594", the name "-OECHEM-09151311273D", and a list of atom coordinates and types.
- Output Data:** A text area on the right shows the resulting SMILES string: N1([C@@H](CCC1)Clcccnc1)C with the ID "89594".

# OpenBabel – příložením dvou struktur

---

- Spouštíme v příkazové řádce pomocí **obfit**
- program obfit potřebuje celkem 3 parametry, vzor, podle kterého bude přikládat (SMILES) a dvě struktury, první zafixuje a druhou se snaží hýbat
- obfit "N1([C@@H](CCC1)c1cccnc1)C" CID\_89594.sdf zinc\_1798.sdf

# Fingerprints a podobnost

## Podobnostní hledání

# Fingerprint

---

- Binární data informující o výskytu nějaké konkrétní skupiny

- 10010001010011110101001010001 ...

- **Příklad z openBabelu:**

```
>3rfm.pdb      256 bits set
```

```
0407002a 81807e18 60180100 47910f50 041c12c0 0200c110
0200a020 2000200c 86600b80 820f4be2 2c30800c 5007b800
1e01983e 01542801 853a00c0 001c000c 14801000 0e088001
00e02418 404e2301 e0000d40 383d8e78 238007c3 9770001c
00043801 c0a00200 68120600 10040100 0c004016 0046803b
e00c4200 23c12ea0
```

# Podobnost/vzdálenost

---

		molekula B		
		0	1	celkem
molekula A	0	$d$	$b$	$b + d$
	1	$a$	$c$	$a + c = A$
	celkem	$a + d$	$c + b = B$	$n$

$a$  je počet „1“, které má molekula A, ale které zároveň nemá molekula B

$b$  je počet „1“, které má molekula B, ale které zároveň nemá molekula A

$c$  je počet „1“, které má molekula A a které má zároveň i molekula B

$d$  je počet „0“, které má molekula A a které má zároveň i molekula B

$n$  je počet dvojnásobný počet fragmentů, platí  $n = a + b + c + d$

$A$  je celkový počet „1“ v molekule A

$B$  je celkový počet „1“ v molekule B

# Podobnostní koeficienty

**Table II** - Similarity coefficients studied.

Coefficients	Similarity expression	Source
1. Simple matching (SM)	$\frac{a+d}{a+b+c+d}$	Sokal and Michener, 1958
2. Rogers and Tanimoto (RT)	$\frac{a+d}{a+2b+2c+d}$	Rogers and Tanimoto, 1960
3. Anderberg (A)	$\frac{a}{a+2(b+c)}$	Anderberg, 1973
4. Russel and Rao (RR)	$\frac{a}{a+b+c+d}$	Russel and Rao, 1940
5. Jaccard (J)	$\frac{a}{a+b+c}$	Jaccard, 1901
6. Sorensen-Dice (SD)	$\frac{2a}{2a+b+c}$	Dice, 1945; Sorensen, 1948
7. Ochiai (O)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Ochiai, 1957
8. Ochiai II (OII)	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	Ochiai, 1957

# Podobnostní hledání v OpenBabelu

---

- `babel mysmiles.smi mymols.sdf -ofpt`

```
MOL_00000067 Tanimoto from first mol = 0.08888889
MOL_00000083 Tanimoto from first mol = 0.0869565
MOL_00000105 Tanimoto from first mol = 0.08888889
MOL_00000296 Tanimoto from first mol = 0.0714286
MOL_00000320 Tanimoto from first mol = 0.08888889
MOL_00000328 Tanimoto from first mol = 0.0851064
MOL_00000338 Tanimoto from first mol = 0.0869565
MOL_00000354 Tanimoto from first mol = 0.08888889
MOL_00000378 Tanimoto from first mol = 0.0816327
MOL_00000391 Tanimoto from first mol = 0.0816327
11 molecules converted
```

# Podobnostní hledání v PubChemu

The screenshot displays the PubChem website interface. At the top, the browser address bar shows `pubchem.ncbi.nlm.nih.gov`. The NCBI logo is on the left, and the PubChem logo with the tagline 'Structure Search' is in the center. A search bar contains the text 'PubChem Compound' and a dropdown arrow. To the right of the search bar is a 'Search' button and a 'Help' link. Below the search bar are links for 'Limits' and 'Advanced search'. A promotional banner reads 'Try the new PubChem Search.' Below this, there are several search method tabs: 'Search By:', 'Name/Text', 'Identity/Similarity' (which is selected), 'Substructure/Superstructure', 'Molecular Formula', '3D Conformer', and 'Saved Search'. Under the 'Identity/Similarity' tab, there are three sub-sections: 'Draw a Structure' (with a 'Launch' button and the text 'the PubChem editor to make a structure'), 'CID, SMILES, InChI', and 'Structure File'. At the bottom, there is a dropdown menu for 'Similar Compounds, score >= 80%' with a help icon.