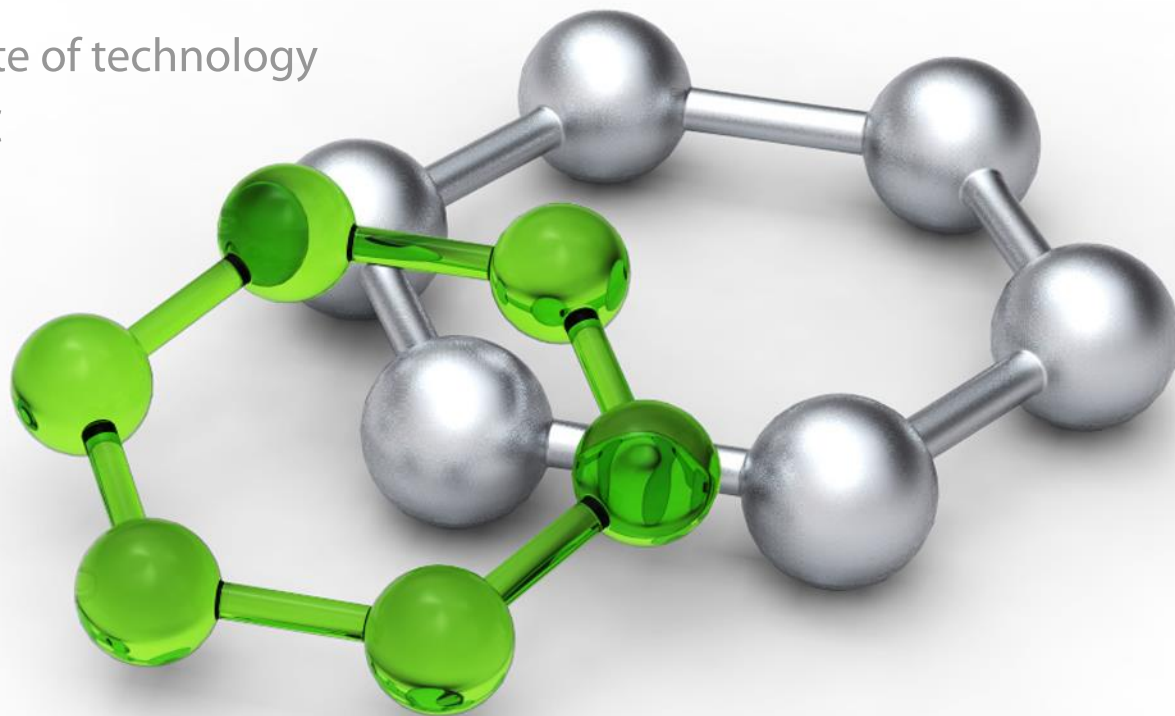




CEITEC

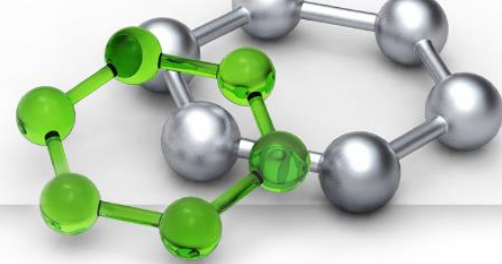
central european institute of technology  
BRNO | CZECH REPUBLIC



Chemoinformatika – úvod

Radka Svobodová

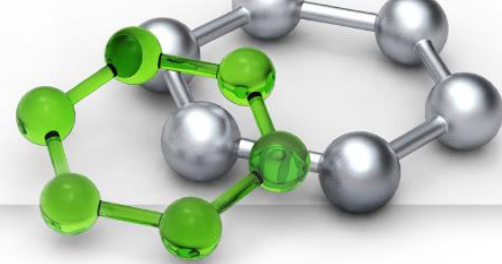
# Proč nahrazovat nebo doplňovat experiment výpočtem?



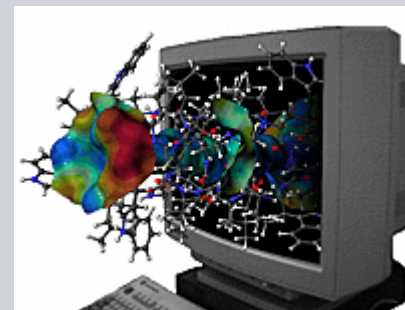
- Vyhneme se práci s toxickými, výbušnými a radioaktivními látkami
- Můžeme pracovat i s nestabilními látkami
- Ušetříme náklady za chemikálie a za realizaci experimentu
- Ušetříme čas experimentálním chemikům :-)
- ...



# Chemoinformatika

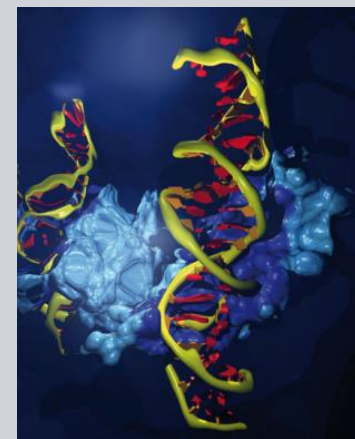


„Chemoinformatika využívá infromatických a algoritmických přístupů pro řešení chemických problémů. Převážně se zaměřuje na získání informací z databází malých nebo středně velkých molekul (léků, organických látek, ...).“

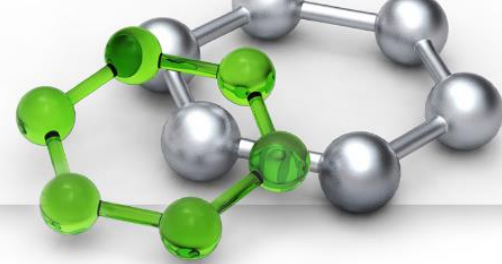


Vzniká v devadesátých létech dvacátého století.

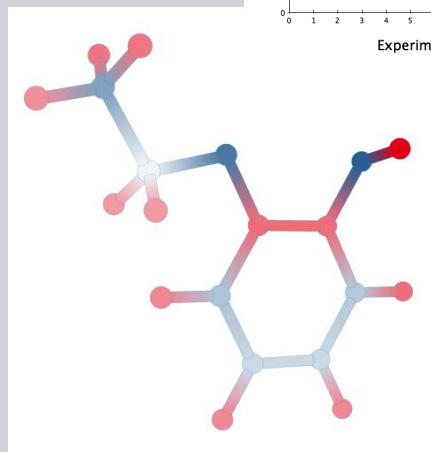
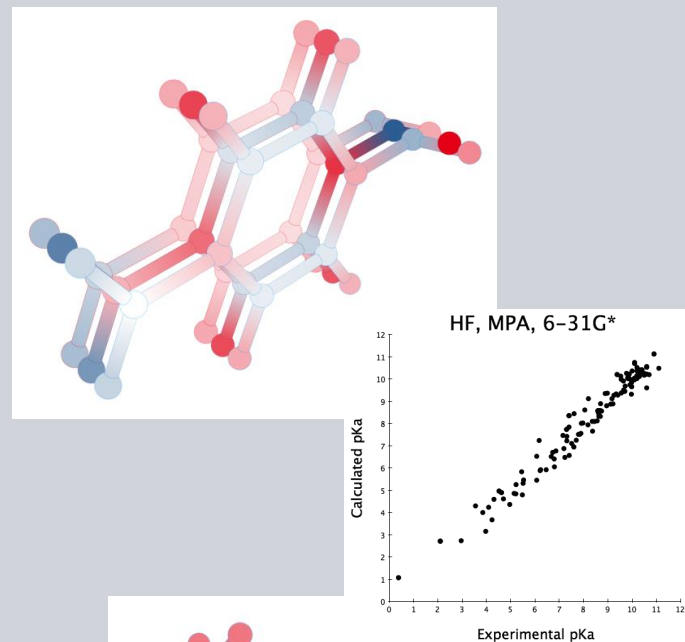
Rozvoj spojen s dostupností velkého množství dat o molekulách léků apod. a s potřebami farmaceutického průmyslu



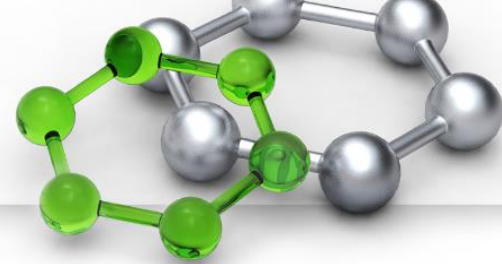
# Chemoinformatika – hlavní oblasti



- Podobnostní vyhledávání v databázích
- Výpočty a aplikace deskriptorů
- QSAR / QSPR
- Vytváření a aplikace virtuálních knihoven molekul
- Virtuální screening



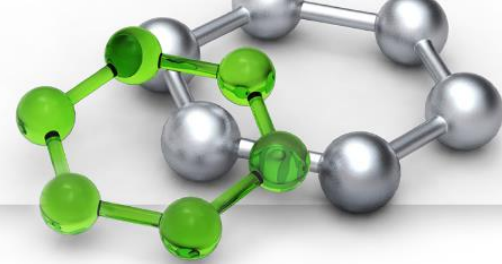
# Jak zapsat molekulu v počítači?



- Zjistit, které informace molekulu popisují
- Zapsat je do počítače



# Které informace popisují molekulu?



Počty atomů?

Málo

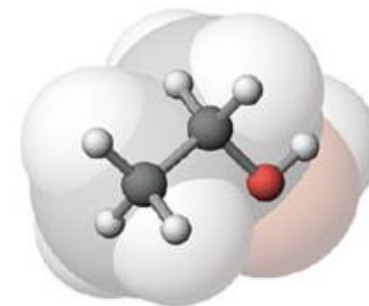
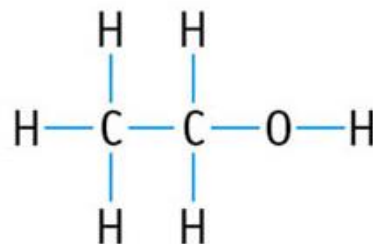
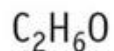
Počty atomů a umístění vazeb?

Lepší

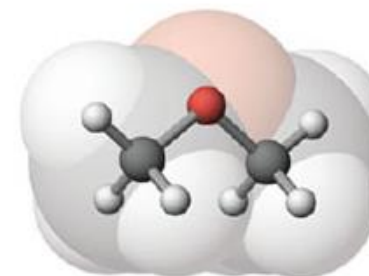
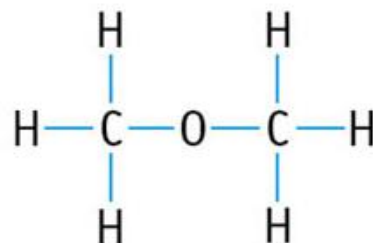
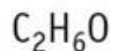
Počty atomů, umístění vazeb a poloha atomů v prostoru?

Ano

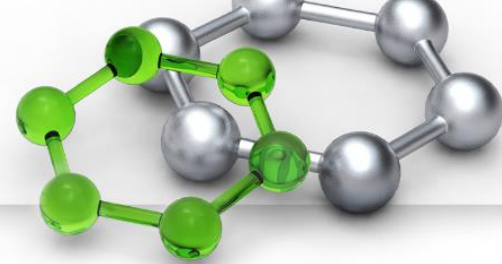
Ethanol



Dimethyl ether



# Model molekuly pro počítačové zpracování



## Atomy:

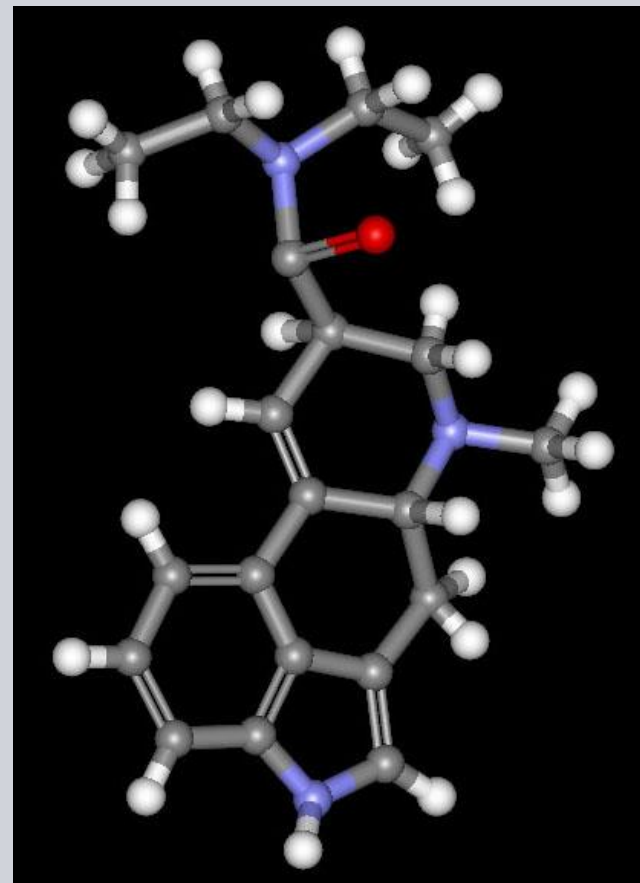
Body v prostoru

U každého uveden chemický symbol prvku

## Vazby:

Dvojice atomů, které jsou vázány

Násobnost vazby



# Zápis molekuly v počítači

Počet atomů

Počet vazeb

První atom je uhlík

```
-ISIS- 09270222202D
13 13 0 0 0 0 0 0 0 0999 V2000
-3.4639 -1.5375 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-3.4651 -2.3648 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.7503 -2.7777 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.0338 -2.3644 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.0367 -1.5338 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.7521 -1.1247 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.7545 -0.2997 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-2.0413 0.1149 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-3.4702 0.1107 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-1.3238 -1.1186 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.6125 -1.5292 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.6167 -2.3542 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.1000 -1.1125 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 2 2 0 0 0 0
6 7 1 0 0 0 0
3 4 2 0 0 0 0
7 8 1 0 0 0 0
7 9 2 0 0 0 0
4 5 1 0 0 0 0
5 10 1 0 0 0 0
2 3 1 0 0 0 0
10 11 1 0 0 0 0
5 6 2 0 0 0 0
11 12 2 0 0 0 0
6 1 1 0 0 0 0
11 13 1 0 0 0 0
M END
```

První tři čísla jsou x, y a z souřadnice atomů

První vazba je mezi atomy 1 a 2 a jde o dvojnou vazbu

CC(=O)Oc1ccc(cc1)C(=O)O



```

21 21 0 0 0 0 0 0 0 0 0 1 V2000
 18.7769 -15.2504 -0.1032 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 18.7571 -16.6359 -0.1252 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.5868 -14.5409 -0.1114 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.5465 -17.3106 -0.1545 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.3767 -15.2158 -0.1421 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.3559 -16.6013 -0.1633 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.6081 -13.0313 -0.0880 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 20.0592 -14.5322 -0.0715 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.5247 -18.7799 -0.1764 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 15.1150 -14.4620 -0.1527 N 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 20.0742 -13.3140 -0.0089 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 21.1073 -15.1564 -0.0523 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.4750 -19.3759 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 18.5697 -19.4030 -0.2650 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 14.0496 -15.0560 -0.1515 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 15.1330 -13.2425 -0.1568 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 19.7111 -17.2054 -0.1194 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 15.3860 -17.1427 -0.1873 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 17.6136 -12.6451 -1.1298 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 16.7057 -12.6567 0.4410 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 18.5209 -12.6823 0.4410 H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2 1 1 0 0 0 0
 3 1 2 0 0 0 0
 4 2 2 0 0 0 0
 5 3 1 0 0 0 0
 6 4 1 0 0 0 0
 6 5 2 0 0 0 0
 3 7 1 0 0 0 0
 1 8 1 0 0 0 0
 4 9 1 0 0 0 0
 5 10 1 0 0 0 0
 8 11 2 0 0 0 0
 8 12 2 0 0 0 0
 9 13 2 0 0 0 0
 9 14 2 0 0 0 0
10 15 2 0 0 0 0
10 16 2 0 0 0 0
17 2 1 0 0 0 0
18 6 1 0 0 0 0
19 7 1 0 0 0 0
20 7 1 0 0 0 0
21 7 1 0 0 0 0

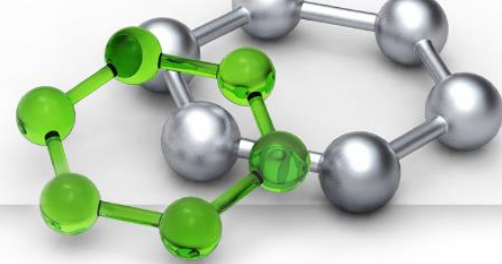
```

M END

Kvízová otázka:

**Nakresli tuto  
molekulu.  
Jak se daná  
molekula  
jmenuje?**

# Současné databáze molekul



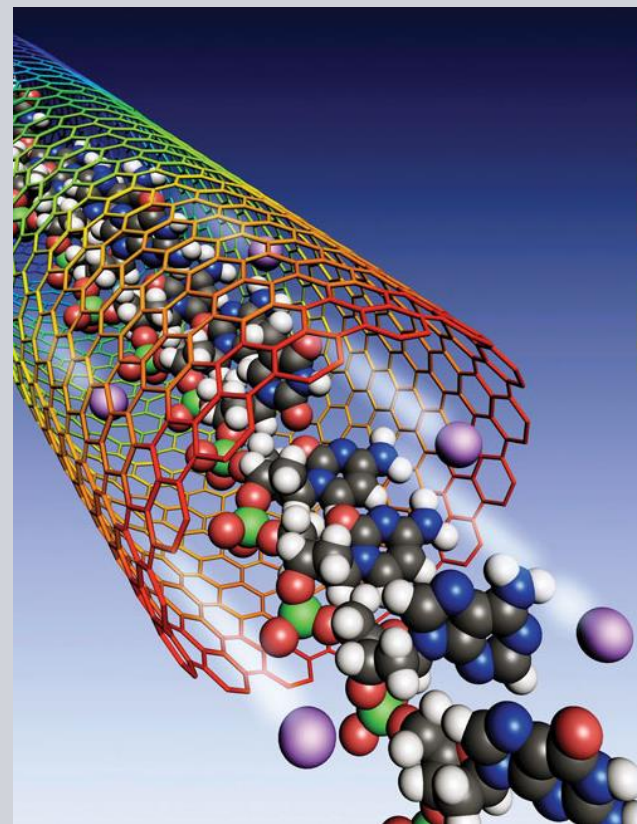
Prožíváme „**informační boom**“ v oblasti dat o molekulách

**Důvod:** Vysoce výkonné techniky strukturní analýzy

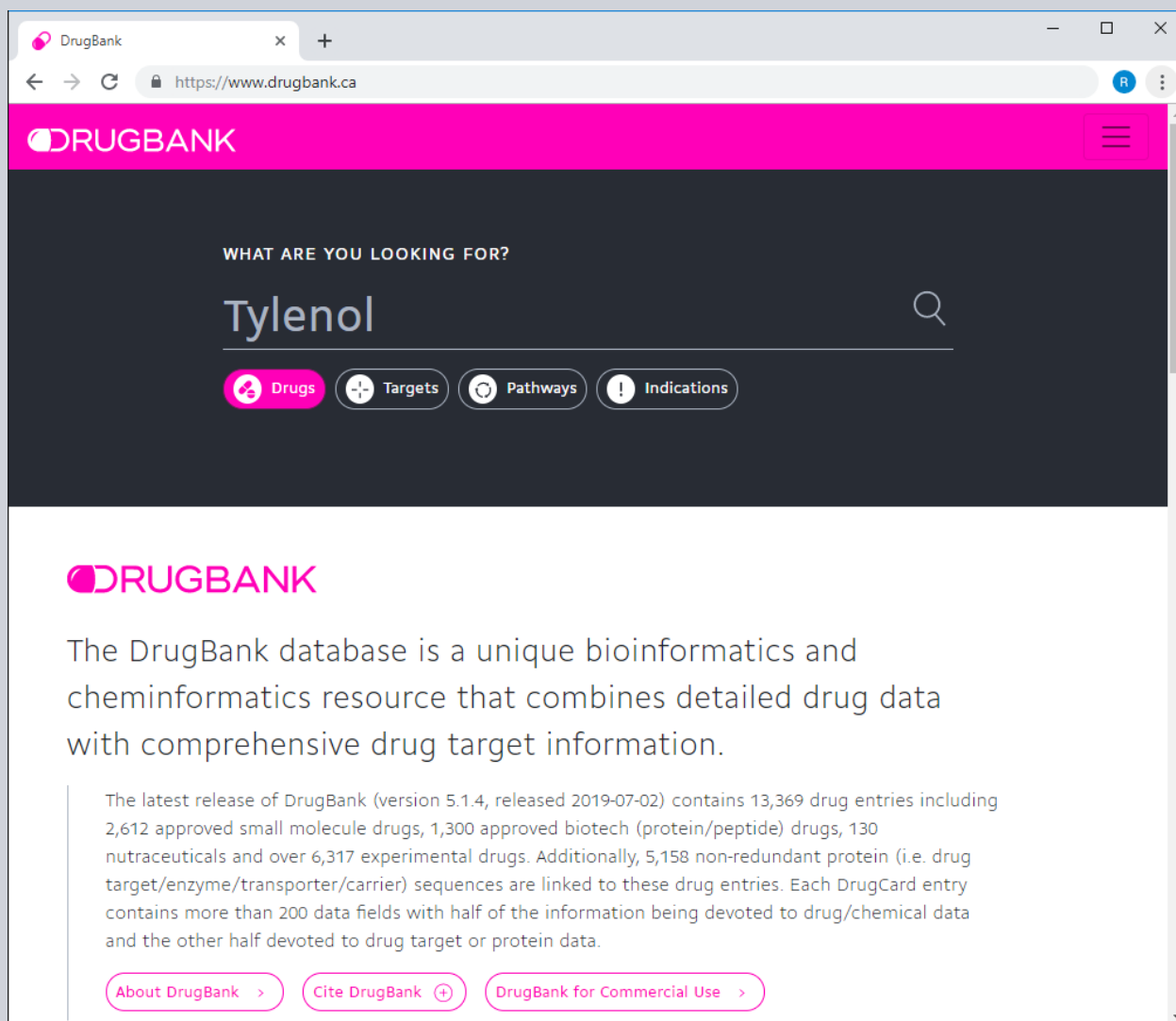
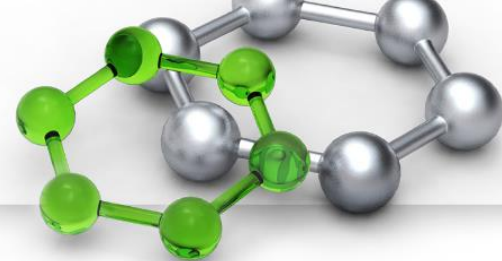
**Důsledky:**

- Máme k dispozici miliony struktur malých molekul (organické molekuly, léky, ...)
- Známe struktury více než 160 000 proteinů a více než 100 000 000 organických molekul
- Jsme schopni zjistit informaci o genomu jednoho člověka za pár dnů

**Většina těchto informací je veřejně přístupná :-)**



# DrugBank – ukázka databáze léků



The screenshot shows the DrugBank website interface. At the top, there is a pink navigation bar with the DrugBank logo and a menu icon. Below this is a dark grey search area with the text "WHAT ARE YOU LOOKING FOR?" and a search input field containing "Tylenol". To the right of the search input is a magnifying glass icon. Below the search input are four buttons: "Drugs" (highlighted in pink), "Targets", "Pathways", and "Indications".

**DRUGBANK**

WHAT ARE YOU LOOKING FOR?

Tylenol

Drugs Targets Pathways Indications

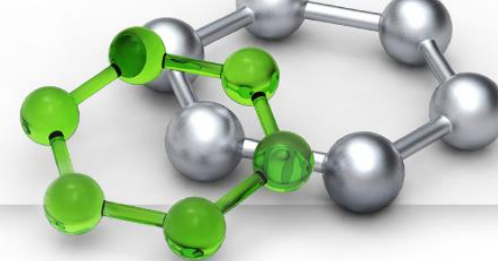
**DRUGBANK**

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

The latest release of DrugBank (version 5.1.4, released 2019-07-02) contains 13,369 drug entries including 2,612 approved small molecule drugs, 1,300 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6,317 experimental drugs. Additionally, 5,158 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.

[About DrugBank](#) [Cite DrugBank](#) [DrugBank for Commercial Use](#)

# DrugBank – ukázka databáze léků

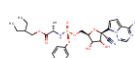


**DRUGBANK**

Drugs

Remdesivir Targets (2) Enzymes (3) Transporters (6)

**IDENTIFICATION**

<b>Name</b>	Remdesivir
<b>Accession Number</b>	DB14761
<b>Description</b>	Remdesivir, or GS-5734, is an adenosine triphosphate analog first described in the literature in 2016 as a potential treatment for Ebola. <sup>1</sup> In 2017, its activity against the coronavirus family of viruses was also demonstrated. <sup>2</sup> Remdesivir is also being researched as a potential treatment to SARS-CoV-2, the coronavirus responsible for COVID-19. <sup>4,8</sup>
	Remdesivir was granted an FDA Emergency Use Authorization on 1 May 2020. <sup>11</sup> This is not the same as an FDA approval. <sup>12</sup>
<b>Type</b>	Small Molecule
<b>Groups</b>	Investigational
<b>Structure</b>	

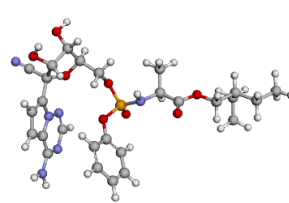
Download:

**DRUGBANK**

Drugs

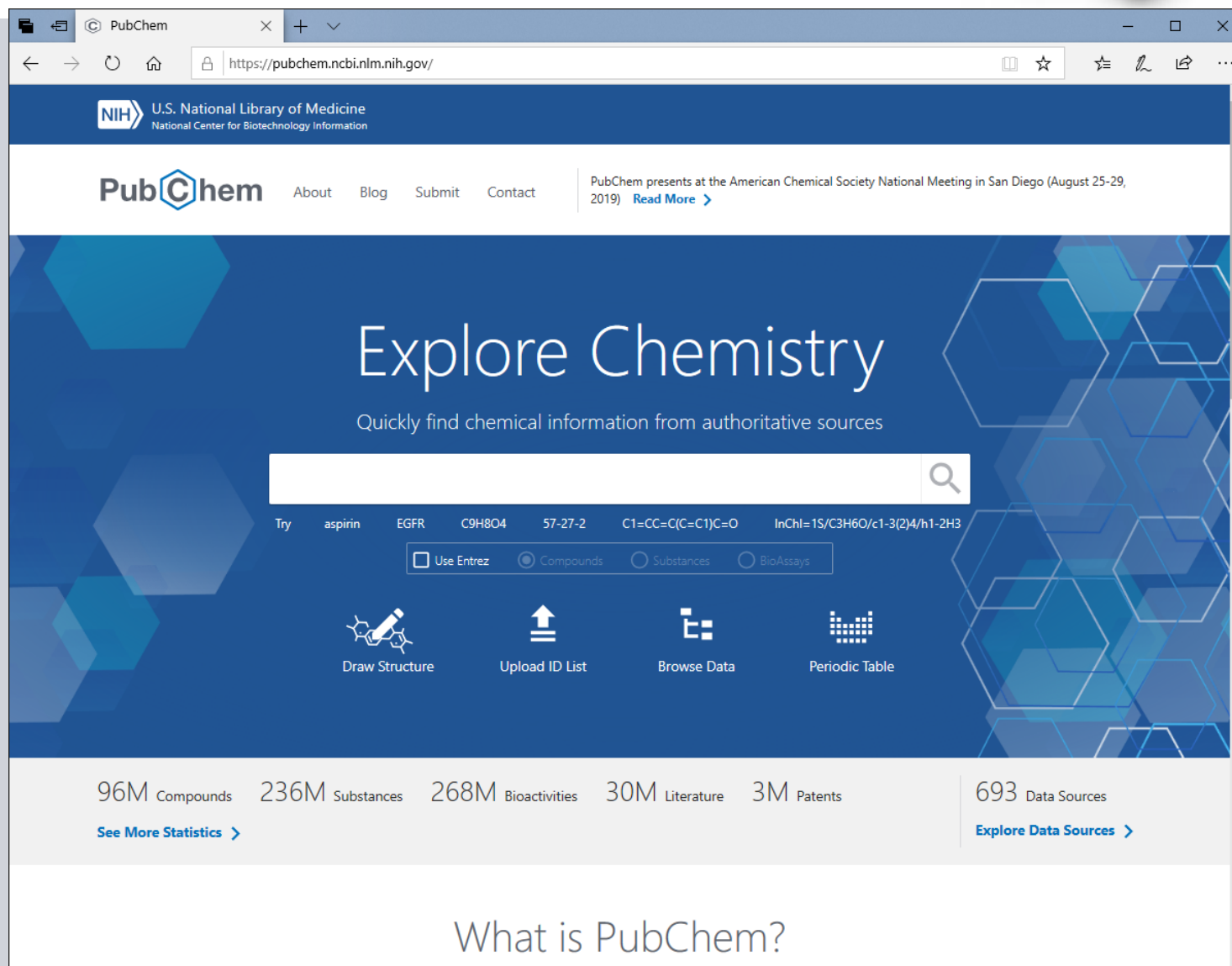
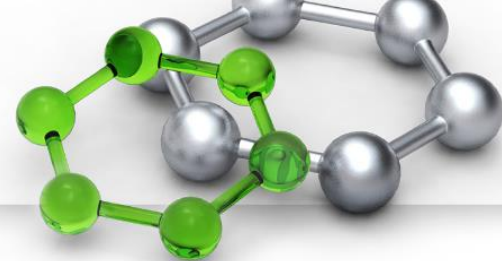
Are you a drug manufacturer or company looking to submit data directly to DrugBank?  
[Register Today](#)

3D structure for Remdesivir (DB14761)



Download:

# PubChem – ukázka databáze organických molekul



PubChem

U.S. National Library of Medicine  
National Center for Biotechnology Information

PubChem About Blog Submit Contact

PubChem presents at the American Chemical Society National Meeting in San Diego (August 25-29, 2019) [Read More >](#)

## Explore Chemistry

Quickly find chemical information from authoritative sources

Try aspirin EGFR C9H8O4 57-27-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3(2)4/h1-2H3

Use Entrez  Compounds  Substances  BioAssays

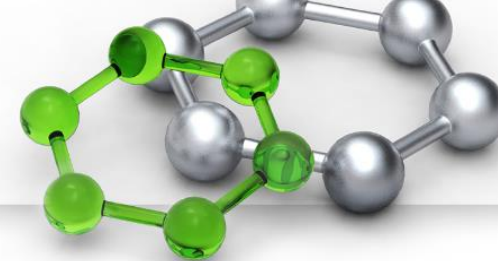
Draw Structure Upload ID List Browse Data Periodic Table

96M Compounds 236M Substances 268M Bioactivities 30M Literature 3M Patents 693 Data Sources

[See More Statistics >](#) [Explore Data Sources >](#)

What is PubChem?

# PubChem – ukázka databáze organických molekul



COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

**NIH** National Library of Medicine  
National Center for Biotechnology Information

**PubChem** About Blog Submit Contact Search PubChem

COMPOUND SUMMARY

# Morphine


PubChem CID: 5288826

Structure:

2D 3D Crystal

Find Similar Structures

Chemical Safety:

 Irritant

Share Tweet Email Cite Download

CONTENTS

- Title and Summary
- 1 Structures
- 2 Names and Identifiers
- 3 Chemical and Physical Properties
- 4 Spectral Information
- 5 Related Records
- 6 Chemical Vendors
- 7 Drug and Medication Information
- 8 Pharmacology and Biochemistry
- 9 Use and Manufacturing

# Ligand Expo – ukázka databáze ligandů



Browser tabs: Ligand Expo, Ligand Dep, SUC\_D3L1, SUC\_D3L1, New Tab

Address bar: ligand-expo.rcsb.org/ld-search.html

**RCSB PDB** PROTEIN DATA BANK  
RCSB PDB | Contact Us

**Ligand Expo**

Home Search Browse Download Ligand Expo Help

## Chemical Component Search Tools

Use the forms below to search for chemical components within the PDB Component Dictionary.

- Search for chemical components by 3-letter component identifier code, molecular name, molecular formula, SMILES description, or InChi/InChiKey chemical description.  
You can also check to see if a 3-letter code is being held by a deposition in progress.
- Chemical substructure searches can also be conducted by starting from a chemical drawing created within the MarvinSketch tool.  
Either start with a SMILES description or chemical data file (see drop-menu for acceptable formats), or draw a 2D chemical structure from scratch (**Launch** without input). It can also generate chemical component definitions from your 2D structure.
- Search for instances of a chemical component throughout the PDB. The **Display** option allows you to simply see a list of PDB codes, or to download these coordinates in PDB, MOL/SDF and mmCIF formats.
- You can also search for analogs to the standard amino acids, nucleotides, popular drugs, and common aromatic ring systems by using the *Browse* feature in the top menu bar.

**Your query results are also searchable!** Each hit from your initial query will contain links to continue searching by similar name, chemical formula, or structure (SMILES).

**MOLECULAR NAME, FORMULA, AND DESCRIPTOR SEARCH OPTIONS** ?

Search term:  Search type:

**SKETCH INPUT AND/OR STRUCTURE SEARCH OPTIONS** ?

File name:  No file selected. File format:

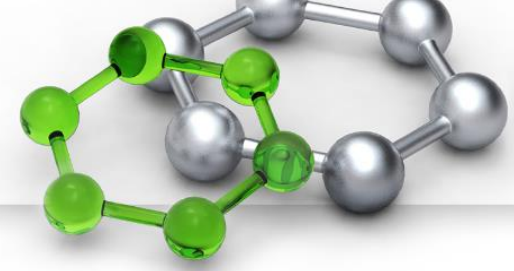
-- OR --

SMILES string:  Display size:

**SEARCH FOR INSTANCES OF CHEMICAL COMPONENTS BY 3-LETTER ID CODE** ?

Component ID code:  Display:

# Ligand Expo – ukázka databáze ligandů



ligand-expo.rcsb.org/reports/B/BCL/index.html

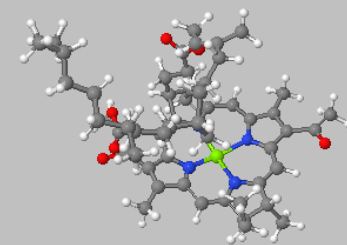
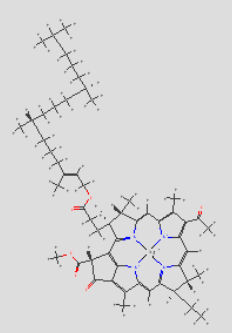
**RCSB PDB**  
PROTEIN DATA BANK  
RCSB PDB | Contact Us

**Ligand Expo**

Home Search Browse Download Ligand Expo Help

Chemical Details **Geometry** Atom Nomenclature Downloads Related Resources

### PDB Chemical Component BCL



Ideal Model JSmol

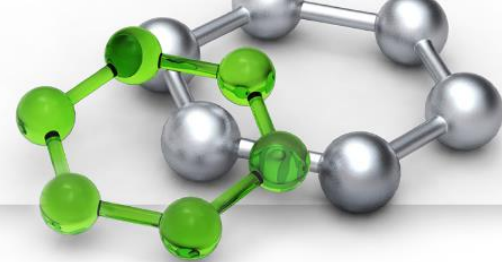
### Chemical Description

<b>Name</b>	BACTERIO <b>CHLOROPHYLL</b> A
<b>Formula</b>	C55 H74 Mg N4 O6
<b>Formal charge</b>	0
<b>Molecular weight</b>	911.504 g/mol
<b>Component type</b>	NON-POLYMER

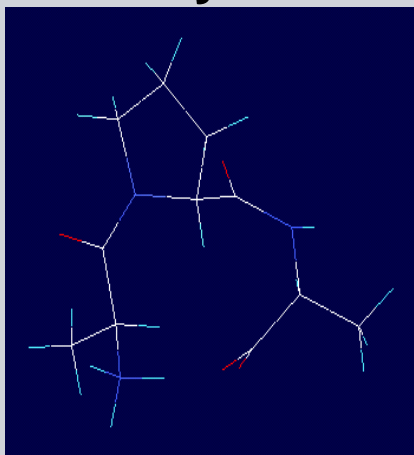
**Ambiguous Chemistry Warning** The chemical description of this component is not well described in this definition. Descriptors and chemical names should be used with caution.



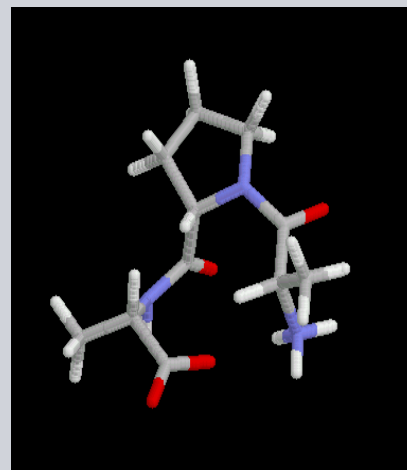
# Vizualizace malé molekuly v počítači



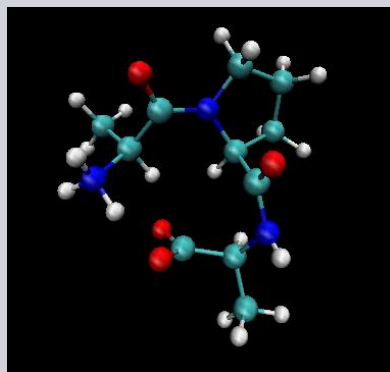
**Drátový model:**



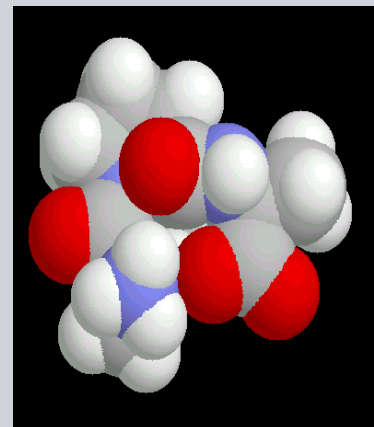
**Tyčinkový model:**



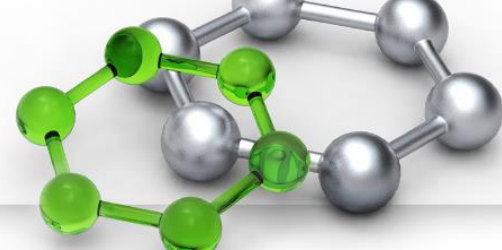
**Tyčinky a kuličky:**



**Kalotový model (CPK):**



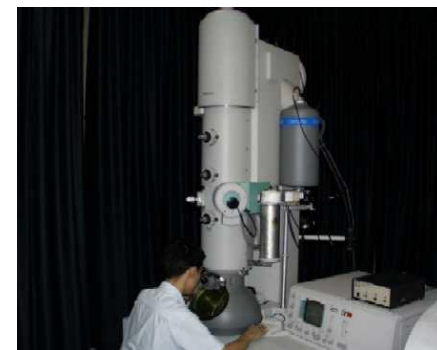
# Protein Data Bank – zdroje dat



89% Rentgenová  
krystalografie



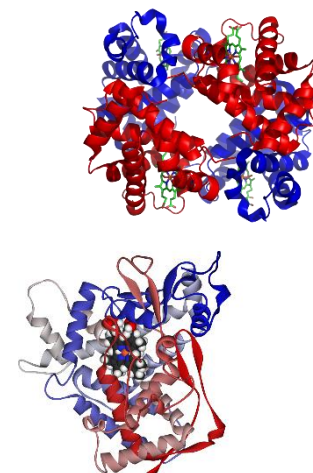
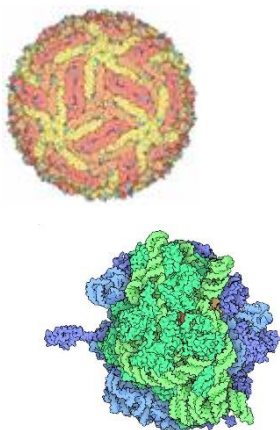
10% NMR  
spektroskopie



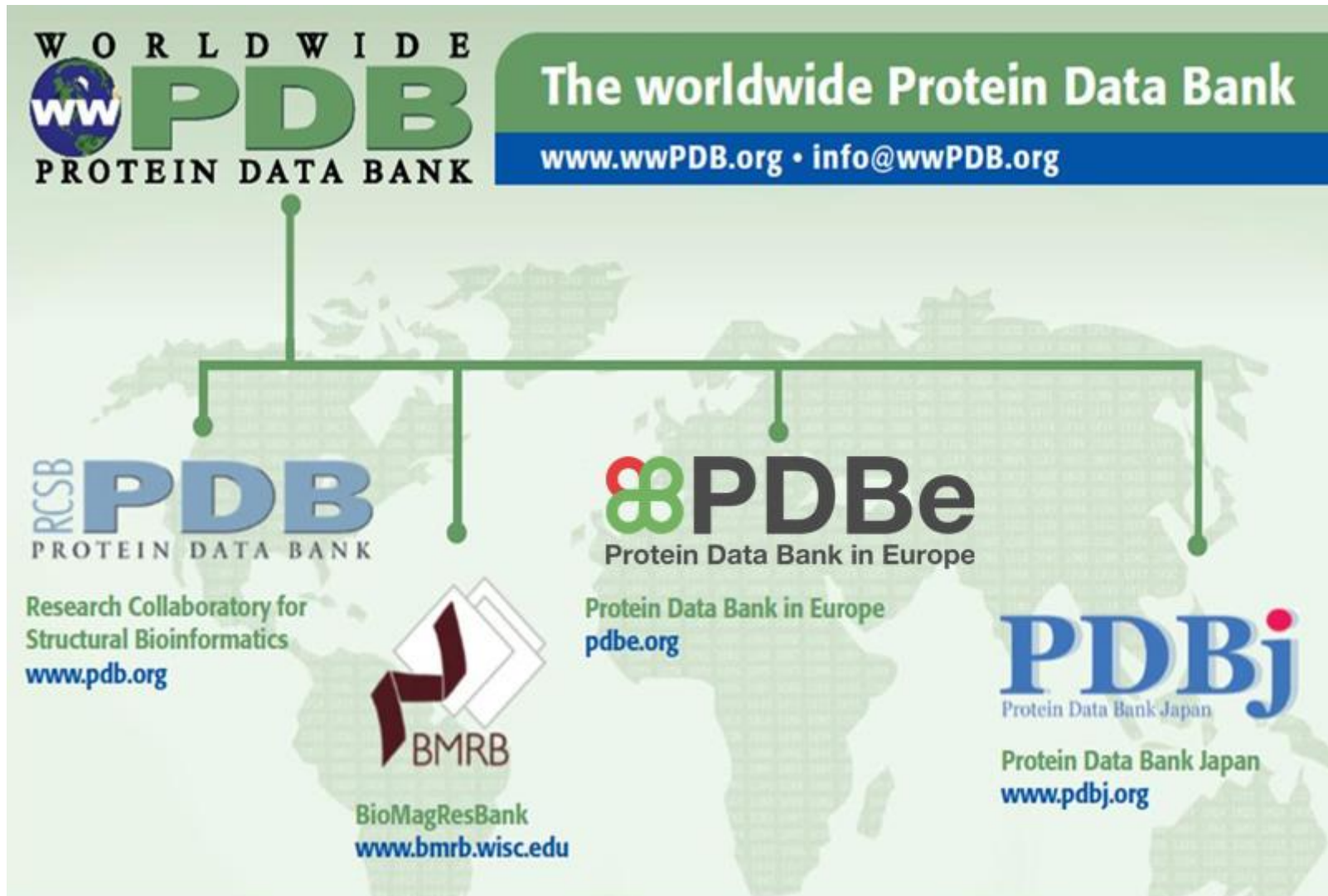
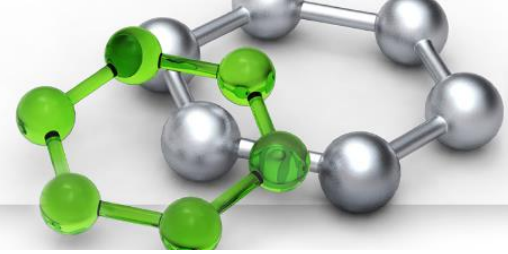
1% kryo-elektronová  
mikroskopie

3D struktura

...								
ATOM	46	C	GLY	A	70	51.536	23.360	40.507
ATOM	47	O	GLY	A	70	50.947	22.279	40.325
ATOM	48	N	ILE	A	71	50.965	24.532	40.270
ATOM	49	CA	ILE	A	71	49.595	24.644	39.786
...								

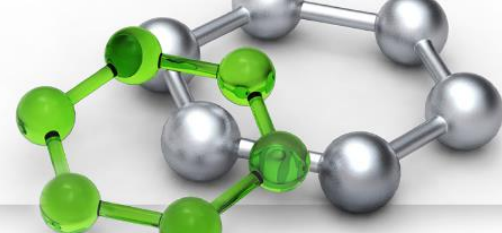


# Protein Data Bank



**> 160k biomacromolecular structures**

# Protein Data Bank – ukázka databáze proteinů



Screenshot of the Protein Data Bank (PDB) website interface. The browser address bar shows the URL: <https://www.ebi.ac.uk/pdbe/>. The page header includes the EMBL-EBI logo and the text "Protein Data Bank in Europe Bringing Structure to Biology". A search bar is visible with a search button and examples: hemoglobin, BRCA1\_HUMAN. The main navigation menu includes: PDBe home, Deposition, PDBe services, PDBe training, Documentation, About PDBe, Share, and Feedback. The page content is organized into several sections:

- Featured structure:** Pygmalion 5x2g, dated 1st September 2019. The text states: "The September image in our 2019 calendar is inspired by a molecular system that can edit DNA and the story of a statue coming to life." A "Read more..." link is provided.
- News:** "Links added to raw experimental data at PDBe" (2 August, 2019), "Improve your previously released PDB coordinates with OneDep" (1 August, 2019), "A celebration of the PDB Art project" (26 July, 2019), and "Mandatory mmCIF format for crystallographic depositions to the PDB" (1 July, 2019).
- Events:** "Art Exhibition: Molecules of Life" by Kendrew Foyer, EBI South Building Wellcome Genome Campus (10 Sep 2019 to 27 Sep 2019), "EBI Structural bioinformatics course" (16 Sep 2019 to 20 Sep 2019), and "EBI Exploring Biological Sequence course" (10 Oct 2019).
- Popular:** A list of links including PDBe-KB, EMsearch, PDBeFold, PDBePISA, PDBeChem, Sequence search, PDBe REST API, EM resources, NMR resources, EMPIAR, Coordinate Server, and PDB Component Library.
- Latest archive statistics:** As of 11 September 2019 the PDB contains 155830 entries (latest PDB entries, chemistry, biology) and EMDB contains 9016 entries (latest map releases, latest header releases, latest updates).
- Tweets by @PDBeurope:** A tweet from David Armstrong (@DaveASci) retweeted by Protein Data Bank, dated 13h, mentioning a training course in Medellin, Colombia.

At the bottom of the page, a cookie consent banner is visible: "This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our [Privacy Notice](#) and [Terms of Use](#). I agree, dismiss this banner".

# Protein Data Bank – ukázka databáze proteinů



Search the PDB archive < PD PDB 3hyu structure surr X + -

European Bioinformatics Institute [GB] https://www.ebi.ac.uk/pdbe/entry/pdb/3hyu

EMBL-EBI Protein Data Bank in Europe Bringing Structure to Biology

Services Research Training About us

Search Examples: hemoglobin, BRCA1\_HUMAN Advanced search

Feedback

## PDBe > 3hyu

Crystal structure of the altitude adapted hemoglobin of guinea pig.  
Source organism: *Cavia porcellus*

Primary publication:  
Structure of the altitude adapted hemoglobin of guinea pig in the R2-state.

Pairet B, Jaenicke E  
PLoS ONE 5 e12389 (2010)  
PMID: 20811494

X-ray diffraction  
1.67Å resolution  
Released: 23 Jun 2010  
Model geometry Fit model/data

Quick links

- 3hyu overview
- Citations
- Structure analysis
- Function and Biology
- Ligands and Environments
- Experiments and Validation
- View
- Downloads
- 3D Visualisation

PDB-REDO

The sliders below show the change in model quality between original PDB entry and the PDB-REDO entry

Model Geometry Fit model/data PDB-REDO

### Function and Biology

Biochemical function: heme binding

Biological process: oxygen transport

Cellular component: hemoglobin complex

Sequence domains:

- Haemoglobin, alpha-type
- Haemoglobin, beta-type
- Globin
- Globin/Protoglobin
- Globin-like superfamily

### Ligands and Environments

2 bound ligands:

2 x HEM 4 x PO4

No modified residues

### Experiments and Validation

Metric	Percentile Ranks	Value
Rfree		0.201
Clashscore		3
Ramachandran outliers		0
Sidechain outliers		0.4%
RSRZ outliers		3.1%

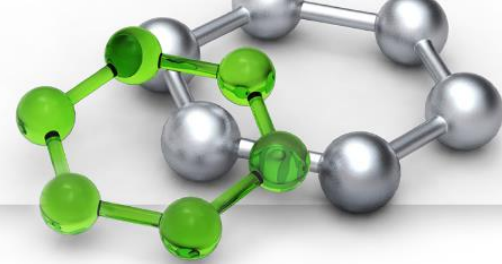
Structure analysis

Assembly composition: hetero tetramer (preferred)

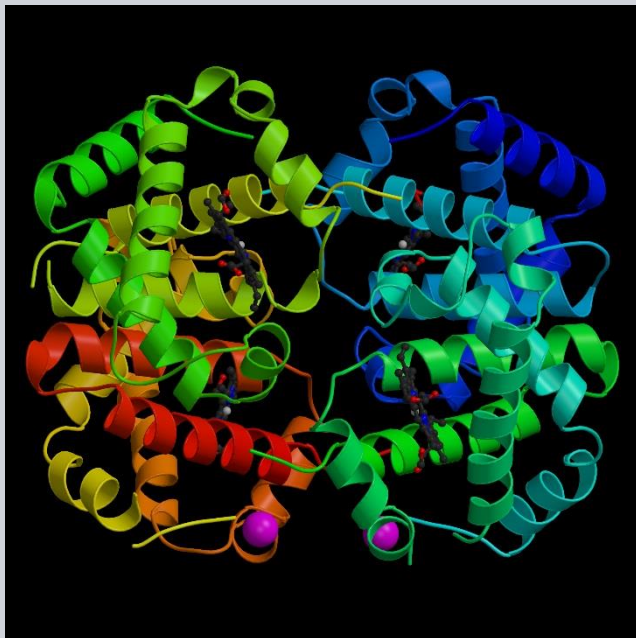
Entry contents: 2 distinct polypeptide molecules

Macromolecules (2 distinct): Hemoglobin subunit alpha

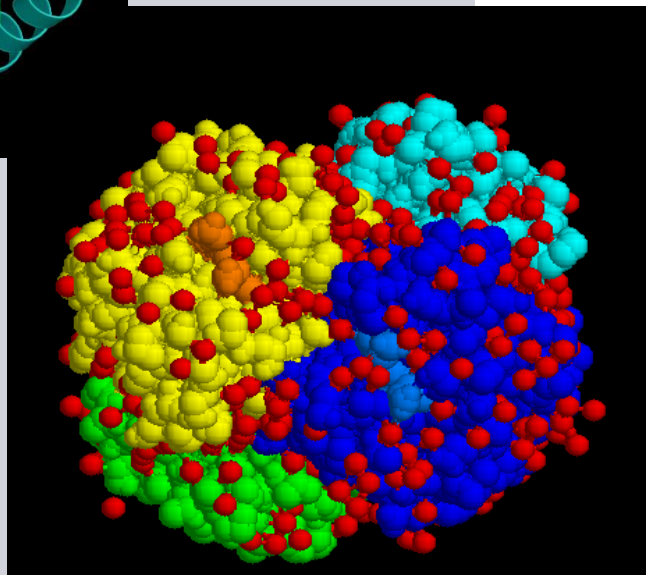
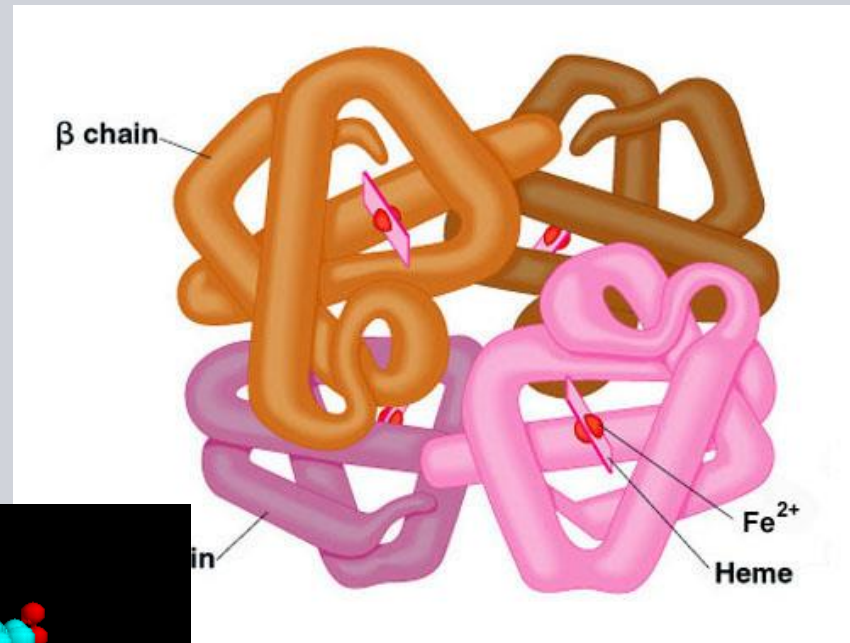
# Vizualizace biomolekuly v počítači



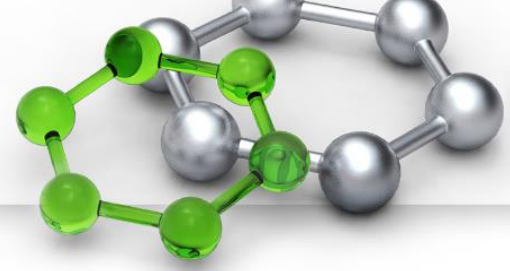
Cartoon model:



Schématický model:



# Proces vývoje léku



Uvedení nového léku na trh stojí v průměru 900 milionů dolarů a trvá více než 10 let.

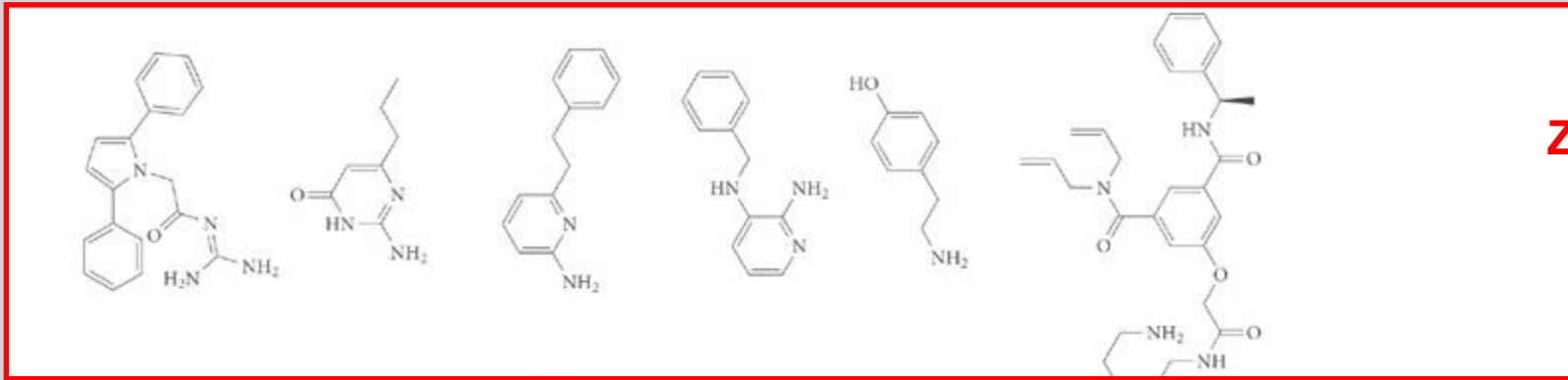
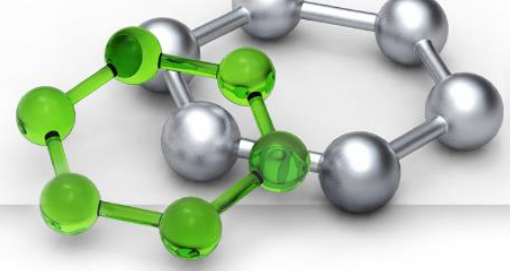
Farmaceutické společnosti často zkoumají a testují 10 000 – 30 000 rozličných látek předtím, než je jedna z nich úspěšně uvedena na trh.

Látky jsou nejdříve **navrženy v základním výzkumu.**

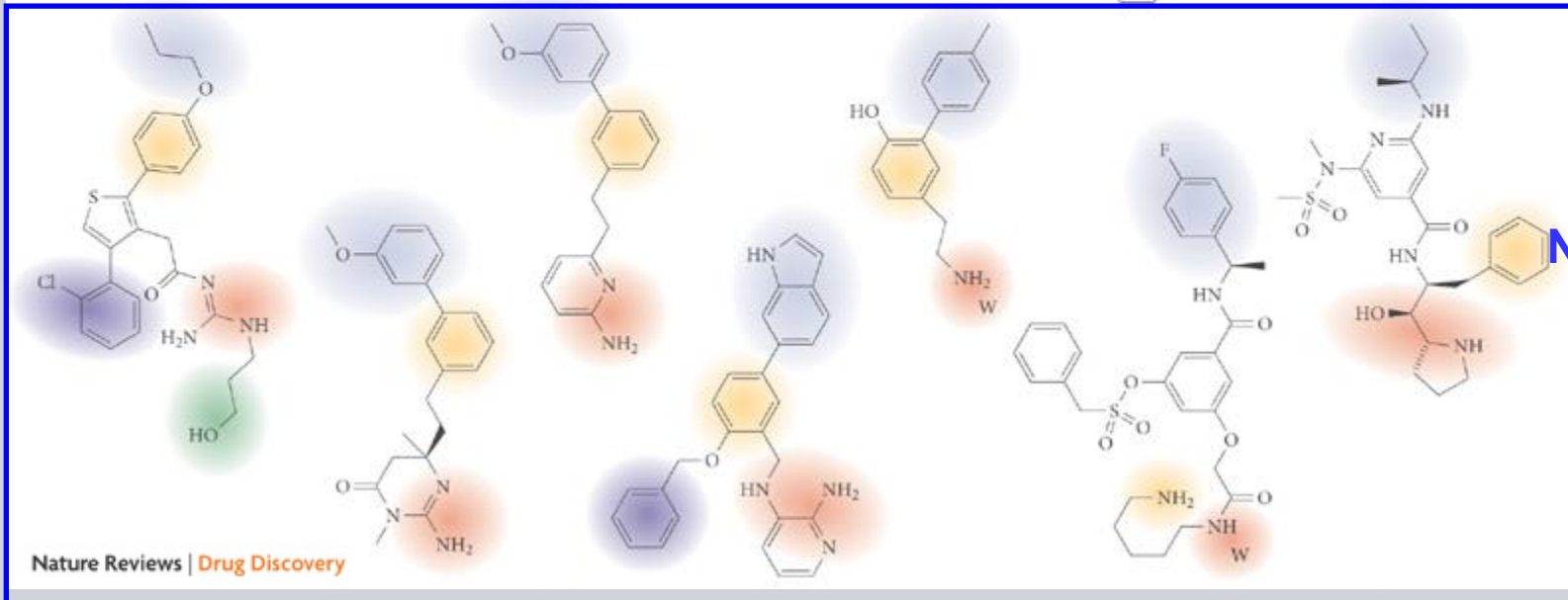
Poté musí projít předklinickými a klinickými zkouškami.

Většina nepostoupí dále, ale ty, které postoupí, mohou nabídnout šanci na kvalitnější život pacientů.

# Návrh léku (drug design)



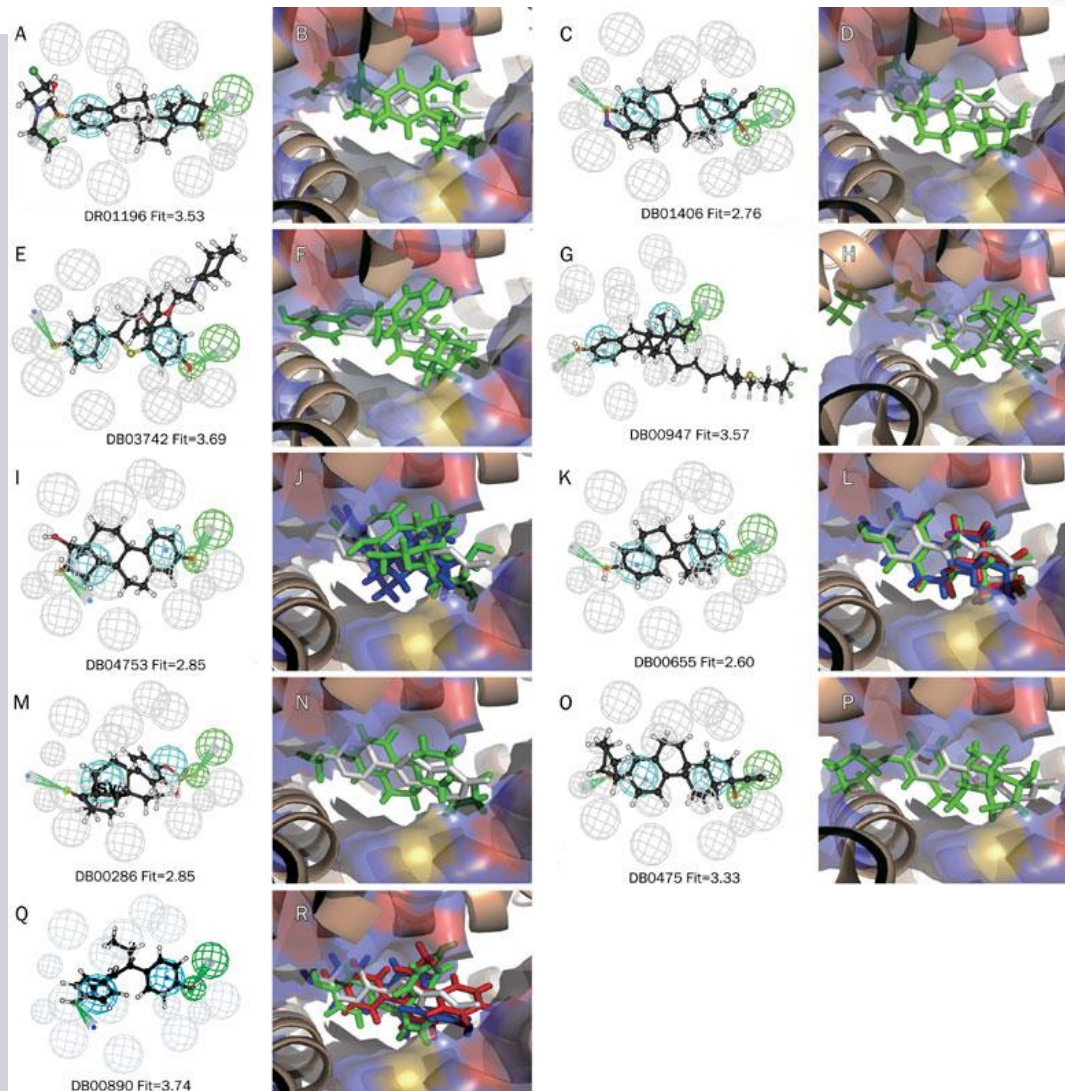
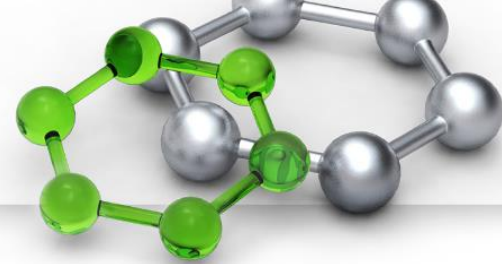
Známé  
léky



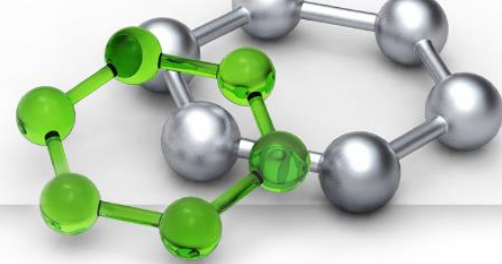
Nově  
navržené  
léky



# Návrh léku (drug design)



# Návrh léku (drug design)

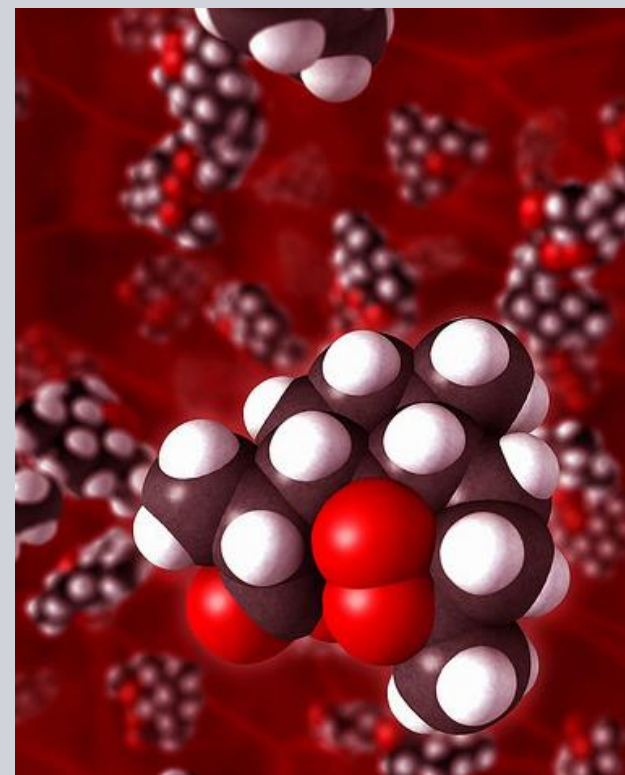


Na základě několika známých molekul léků můžeme vytvořit (ručně nebo automaticky) rozsáhlé sady molekul.

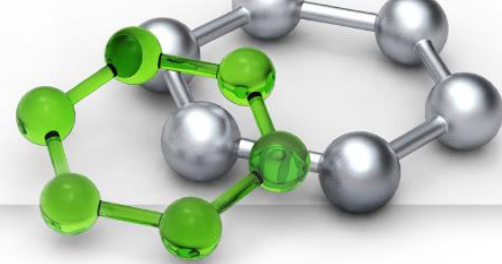
Tyto sady se nazývají virtuální knihovny a obsahují desítky, stovky i miliony molekul.

Několik z molekul ve virtuální knihovně může být velmi účinnými léky.

**Ale které to jsou ???**



# Jak zjistit, která z navržených molekul je lékem?



Navržené molekuly existují jen na papíře nebo v počítači a nebyly zatím syntetizovány.

Nemáme tedy naměřeny jejich fyzikální a chemické vlastnosti ani nevíme nic o jejich aktivitě.

Jak tedy určit, která z nich bude vhodným lékem?

## Máme dvě možnosti:

a) Molekuly syntetizovat a jejich vlastnosti i aktivitu změřit.

b) Vlastnosti i aktivitu molekul odhadnout (predikovat) na základě jejich struktury.

Chemoinformatika