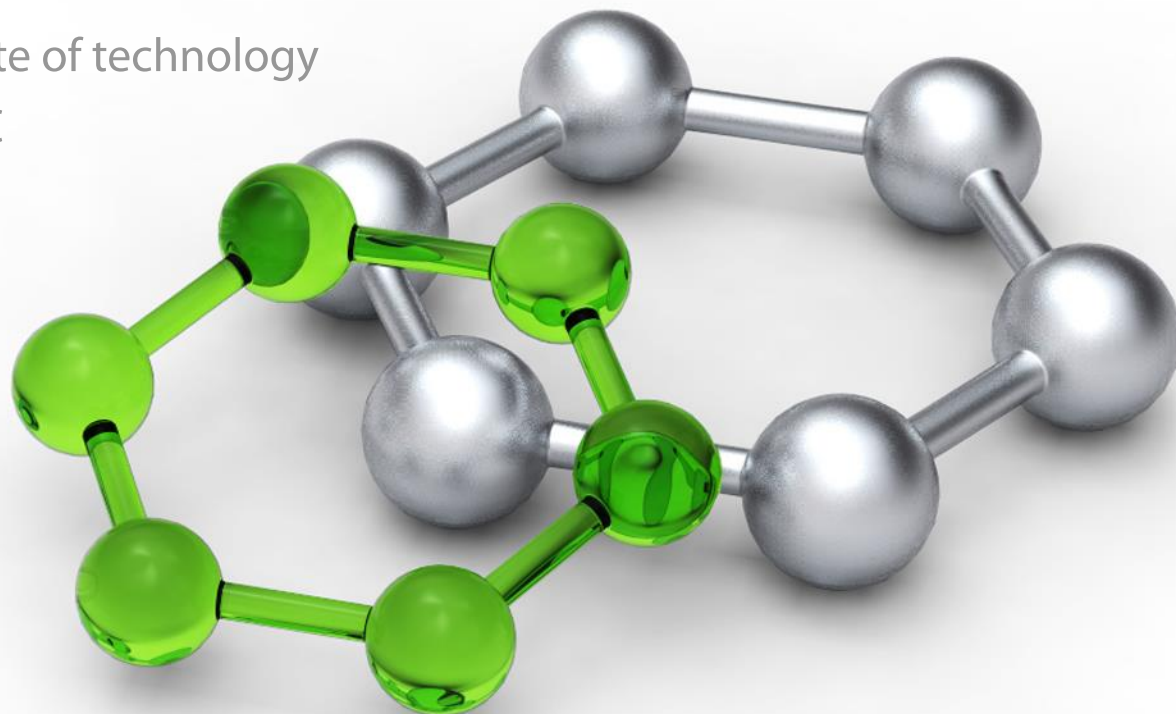




CEITEC

central european institute of technology
BRNO | CZECH REPUBLIC



Datové formáty pro zápis molekul

Model molekuly pro počítačové zpracování

Atomy:

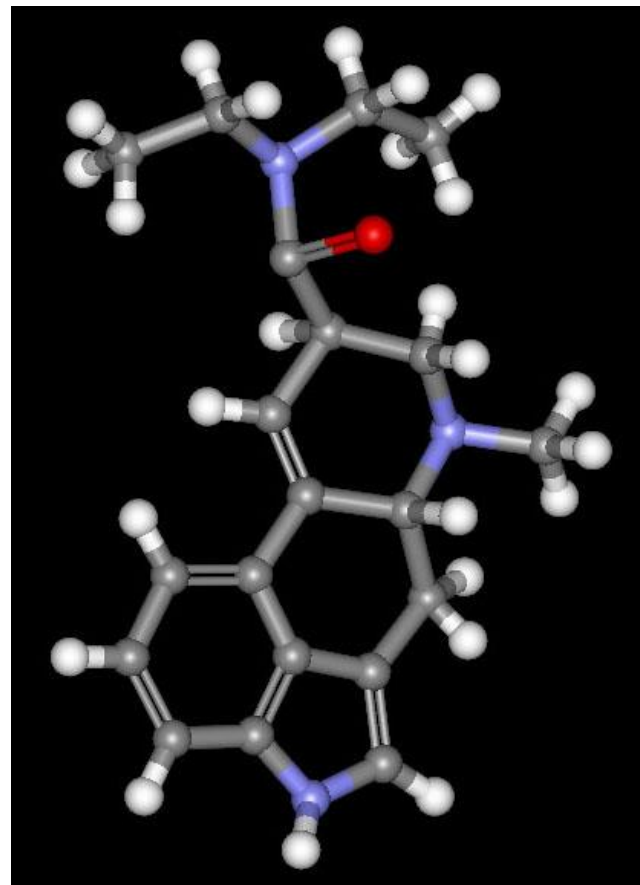
Body v prostoru

U každého uveden chemický symbol prvku

Vazby:

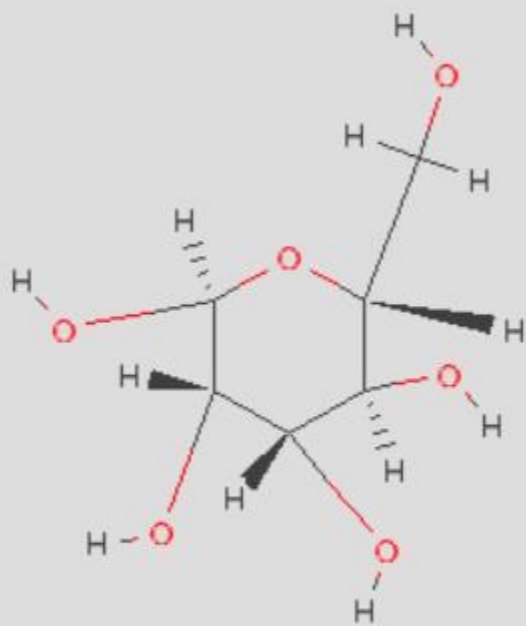
Dvojice atomů, které jsou vázány

Násobnost vazby

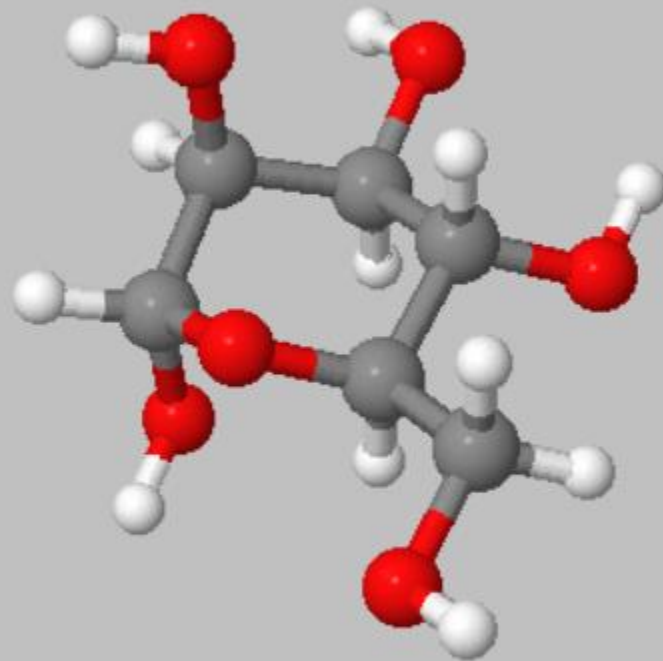


Model molekuly pro počítačové zpracování

2D struktura



3D struktura



Databáze

Malé molekuly:

- PubChem
- DrugBank
- LigandExpo

Proteiny a nukleové kyseliny:

- Protein Data Bank

Databáze

Příklady:

- PubChem: najdeme TNT
- DrugBank: najdeme ibalgin
- LigandExpo: najdeme chlorofyl
- Protein Data Bank: najdeme hemoglobin

Datové formáty

3D formáty:

- SDF/MOL formát
- PDB formát
- mmCIF formát

2D formáty:

- SMILES, SMIRKS, SSMARTS
- InChi, InChiKey
- CHUCKLES, CHORTLES, and CHARTS

Zápis molekuly v počítači - MOL a SDF soubor - organické molekuly

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

Cvičení

Najděte a stáhněte MOL(nebo SDF) soubor s 3D strukturou cyklohexanu.

Prohlédněte si tento soubor.

Cvičení

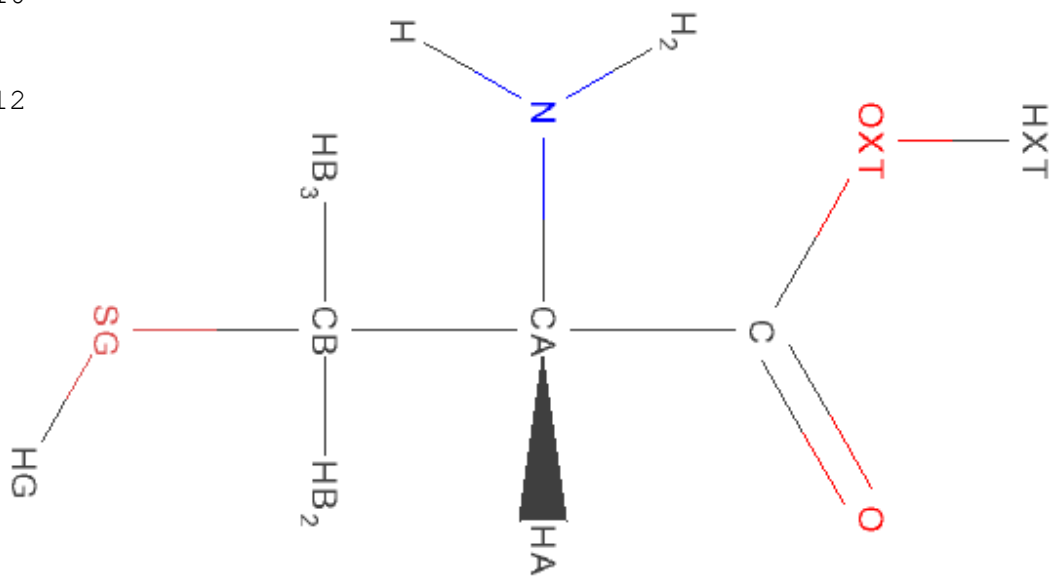
Najděte a stáhněte MOL soubor s 3D
strukturou aspirinu v Pubchem a LigandExpo.
Prohlédněte si soubory.

Zápis molekuly v počítači – PDB soubor – proteiny (čitelný, ale od roku 2016 nahrazen CIF/mmCIF formátem)

ATOM	1	N	CYS A	1	22.585	13.716	37.715	1.00	10.00	N
ATOM	2	CA	CYS A	1	22.372	13.468	39.168	1.00	10.00	C
ATOM	3	C	CYS A	1	21.806	14.686	39.893	1.00	10.00	C
ATOM	4	O	CYS A	1	22.614	15.553	40.277	1.00	10.00	O
ATOM	5	CB	CYS A	1	23.683	13.019	39.828	1.00	10.00	C
ATOM	6	SG	CYS A	1	25.202	13.440	38.921	1.00	10.00	S
ATOM	7	OXT	CYS A	1	20.565	14.747	40.076	1.00	10.00	O
ATOM	8	H	CYS A	1	22.963	12.902	37.230	1.00	10.00	H
ATOM	9	H2	CYS A	1	23.171	14.537	37.565	1.00	10.00	H
ATOM	10	HA	CYS A	1	21.614	12.654	39.253	1.00	10.00	H
ATOM	11	HB2	CYS A	1	23.739	13.412	40.869	1.00	10.00	H
ATOM	12	HB3	CYS A	1	23.651	11.923	40.031	1.00	10.00	H
ATOM	13	HG	CYS A	1	26.013	13.162	39.329	1.00	10.00	H
ATOM	14	HXT	CYS A	1	20.212	15.505	40.527	1.00	10.00	H

CONNECT	1	2	8	9		
CONNECT	2	1	3	5	10	
CONNECT	3	2	4	7		
CONNECT	4	3				
CONNECT	5	2	6	11	12	
CONNECT	6	5	13			
CONNECT	7	3	14			
CONNECT	8	1				
CONNECT	9	1				
CONNECT	10	2				
CONNECT	11	5				
CONNECT	12	5				
CONNECT	13	6				
CONNECT	14	7				

END



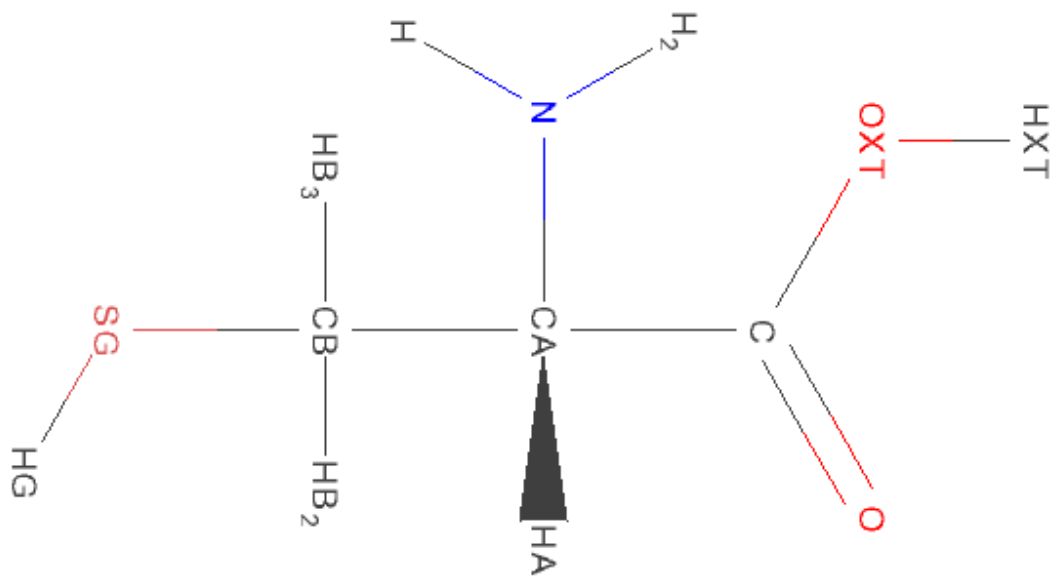
Zápis molekuly v počítači – fragment CIF souboru - proteiny

```

CYS N   N   N 0 1 N N N 22.585 13.716 37.715 1.585  0.483 -0.081 N   CYS 1
CYS CA  CA  C 0 1 N N R 22.372 13.468 39.168 0.141  0.450  0.186 CA  CYS 2
CYS C   C   C 0 1 N N N 21.806 14.686 39.893 -0.095 0.006  1.606 C   CYS 3
CYS O   O   O 0 1 N N N 22.614 15.553 40.277 0.685 -0.742  2.143 O   CYS 4
CYS CB  CB  C 0 1 N N N 23.683 13.019 39.828 -0.533 -0.530 -0.774 CB  CYS 5
CYS SG  SG  S 0 1 N N N 25.202 13.440 38.921 -0.247 0.004  -2.484 SG  CYS 6
CYS OXT OXT O 0 1 N Y N 20.565 14.747 40.076 -1.174 0.443  2.275 OXT CYS 7
CYS H   H   H 0 1 N N N 22.963 12.902 37.230 1.928 -0.454  0.063 H   CYS 8
CYS H2  HN2 H 0 1 N Y N 23.171 14.537 37.565 1.693  0.682 -1.065 H2  CYS 9
CYS HA  HA  H 0 1 N N N 21.614 12.654 39.253 -0.277 1.446  0.042 HA  CYS 10
CYS HB2 1HB H 0 1 N N N 23.739 13.412 40.869 -0.114 -1.526 -0.630 HB2 CYS 11
CYS HB3 2HB H 0 1 N N N 23.651 11.923 40.031 -1.604 -0.554 -0.575 HB3 CYS 12
CYS HG  HG  H 0 1 N N N 26.013 13.162 39.329 -0.904 -0.965 -3.145 HG  CYS 13
CYS HXT HXT H 0 1 N Y N 20.212 15.505 40.527 -1.326 0.158  3.186 HXT CYS 14

```

#



Cvičení

Najděte a stáhněte PDB a mmcif soubor cytochromu P450.

Prohlédněte si tyto soubory.

Zápis 2D struktury pomocí 3D formátů

Výhody:

?

Nevýhody:

?

Zápis 2D struktury pomocí 3D formátů

Výhody:

- Nejobecnější zápis struktury molekuly
- Snadno použitelné jako vstup pro algoritmy, pracující se strukturou

Nevýhody:

- Zabírá hodně místa
- Není vhodné pro některé speciální typy úkolů.

Formáty specifické pro zápis 2D struktury molekuly

- **SMILES**, SMIRKS, SSMARTS
- InChi, InChiKey
- CHUCKLES, CHORTLES, and CHARTS

atd

Chemical Description

Name	alpha-D-mannopyranose
Synonyms	alpha-D-mannose; D-mannose; mannose
Formula	C ₆ H ₁₂ O ₆
Formal charge	0
Molecular weight	180.156 g/mol
Component type	D-saccharide, alpha linking

Chemical features

Atom count	24
Chiral atom count	5
Chiral atoms	C1 C2 C3 C4 C5
Observed leaving atoms	O1 HO1 HO2 HO3 HO4 HO6
Bond count	24
Aromatic bond count	0

Chemical Identifiers

Systematic name (ACDLabs)	alpha-D-mannopyranose
Systematic name (OpenEye OEToolkits)	(2S,3S,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol

Chemical Descriptors

Stereo SMILES (CACTVS)	<chem>OC[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O</chem>
SMILES (CACTVS)	<chem>OC[CH]1O[CH](O)[CH](O)[CH](O)[CH]1O</chem>
Stereo SMILES (OpenEye)	<chem>C([C@@H]1[C@H]([C@@H]([C@@H]([C@H](O1)O)O)O)O)O</chem>
InChI descriptor	InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5+,6+/m1/s1
InChIKey descriptor	WQZGKKKJIJFFOK-PQMKYFCFSA-N

2.1.2 InChI



InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1

Computed by InChI 1.0.5 (PubChem release 2019.06.18)

► PubChem

2.1.3 InChI Key



BQJCRHHNABKAKU-KBQPJGBKSA-N

Computed by InChI 1.0.5 (PubChem release 2019.06.18)

► PubChem

2.1.4 Canonical SMILES



CN1CCC23C4C1CC5=C2C(=C(C=C5)O)OC3C(C=C4)O

Computed by OEChem 2.1.5 (PubChem release 2019.06.18)

► PubChem

2.1.5 Isomeric SMILES



CN1CC[C@]23[C@@H]4[C@H]1CC5=C2C(=C(C=C5)O)O[C@H]3[C@H](C=C4)O

Computed by OEChem 2.1.5 (PubChem release 2019.06.18)

Formáty pro zápis 2D struktury molekuly - SMILES

SMILES znamená následující:

**Simplified Molecular Input Line Entry
Specification**

= zakódování struktury molekuly do řetězce.

Dále uvedu stručný popis SMILES.

Podrobnější informace najdete např. zde:

<http://www.daylight.com/dayhtml/smiles/>

Formáty pro zápis struktury molekuly - SMILES – kódování atomů - syntaxe

Syntaxe v jazyce*, specifikujícím SMILES:

atom : '[' <mass> symbol <chiral> <hcount> <sign<charge>> ']' ;

Popis:

symbol	chemická značka atomu
	* = nspecifikovaný atom
<sign<charge>>	znaménko a náboj
<mass>	atomová hmotnost
<chiral>	chiralita (nebudeme používat)
<hcount>	počet vázaných vodíků

* analogie DTD.

Formáty pro zápis struktury molekuly

- SMILES – kódování atomů - příklady

Obrázek	SMILES string	Popis
S	[S]	Elementární síra
CH ₄	C	Methan (C vázaný s tolika H, aby měl plně obsazenou valneční vrstvu)
H ₂ S	S	Sirovodík (S vázaný s tolika H, aby měl plně obsazenou valneční vrstvu)
HO ⁻	[OH-], [OH-1]	Hydroxidový anion
²³⁵ U	[235U]	Izotop uranu s at. hmot. 235
+2	[+2]	Nespecif. atom s nábojem 2+

Formáty pro zápis struktury molekuly

- SMILES – kódování vazeb - syntaxe

Syntaxe v jazyce, specifikujícím SMILES:

bond : *<empty>* | '-' | '=' | '#' | ':' ;

Popis:

<i><empty></i>	libovolná vazba (při níž je valenční vrstva plně obsazena)
-	jednoduchá vazba
=	dvojná vazba
#	trojná vazba
:	aromatická vazba

Formáty pro zápis struktury molekuly

- SMILES – kódování vazeb - příklady

Obrázek	SMILES string	Popis
$\text{CH}_3\text{-CH}_3$	CC, C-C, [CH3]-[CH3]	Ethan
$ \begin{array}{c} \text{H} \backslash \\ \text{C} = \bar{\text{O}} \\ \text{H} / \quad \bar{\quad} \end{array} $	C=O, O=C	Formaldehyd
$\text{H-C}\equiv\text{N}$	C#N, N#C	Kyanovodík
$\text{CH}_2=\text{CH}_2$	C=C (lze i cc)	Ethen
$\text{CH}_2=\text{CH-CH}=\text{CH}_2$	C=C-C=C (lze i cccc)	1,3-butadien
?	ccc	Nelze odhadnout typ vazeb

Formáty pro zápis struktury molekuly - SMILES – kódování větvení - syntaxe

Syntaxe v jazyce, specifikujícím SMILES:

```
branch :      '(' <chain> ')'
            | '(' <chain> <branch> ')'
            | '(' <branch> <chain> ')'
            | '(' <chain> <branch> <chain> ')' ;
```

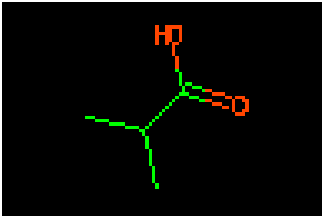
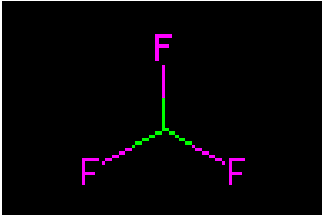
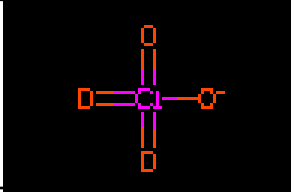
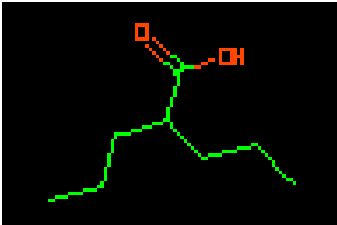
Popis:

<chain> řetězec

<branch> větev

Formáty pro zápis struktury molekuly

- SMILES – kódování větvení - příklady

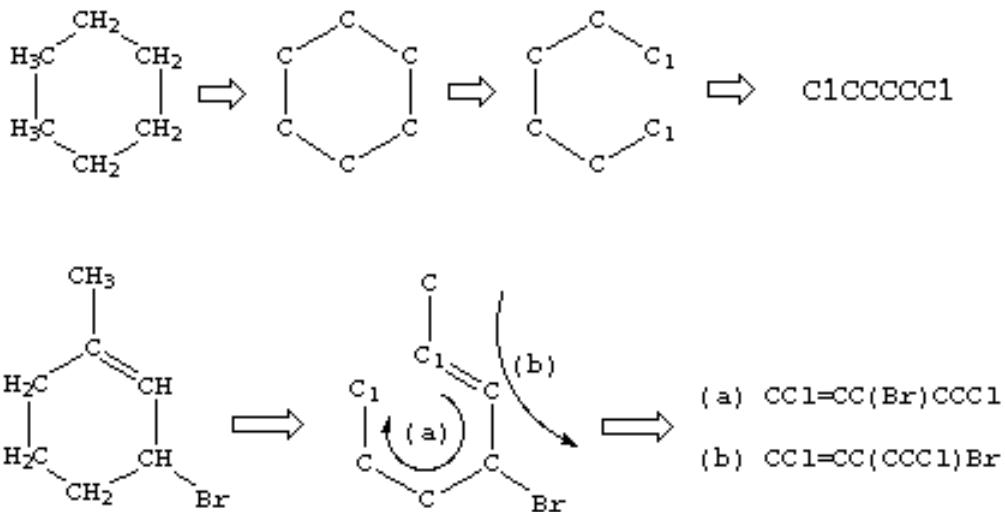
Obrázek	SMILES string	Popis
	<chem>CC(C)C(=O)O</chem>	Isobutanová kyselina
	<chem>FC(F)F,</chem> <chem>C(F)(F)F</chem>	Fluoroform
	<chem>O=Cl(=O)(=O)[O-],</chem> <chem>Cl(=O)(=O)(=O)[O-]</chem>	Perchlorátový anion
	<chem>CCCC(C(=O)O)CCC</chem>	4-heptanová kyselina

Formáty pro zápis struktury molekuly - SMILES – kódování cyklů

Zvolíme v cyklu libovolnou vazbu a její koncové atomy označíme číslem.

Cyklus v místě dané vazby přeručíme a zapíšeme ho jako lineární sekvenci atomů.

Příklady:



Formáty pro zápis struktury molekuly

- SMILES - zhodnocení

Výhody SMILES:

- Komprimace místa
- Možnost zápisu molekuly pomocí regulárního výrazu

Nevýhody SMILES:

- Nejednoznačnost (neexistuje „korektní“ pořadí atomů, 1 fakt lze zapsat více způsoby).
- Nutnost vytvoření úplného výpisu předtím, než lze na molekulu aplikovat nějaký algoritmus (izomorfismus, cykly atd.)

Formáty pro zápis struktury molekuly

- SMILES – zhodnocení II

Využití SMILES:

- Názvosloví a automatické generování názvů.
- Vyhledávání částí molekul pomocí regulárních výrazů.

Rozšíření SMILES:

Pokročilejší verzí SMILES stringů jsou SMARTS stringy. Jsou definovány stejně jako SMILES + obsahují navíc další pravidla. Podrobněji o SMARTS:

<http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>

Cvičení

Najděte SMILES zápis aspirinu.

V jakých databázích je dostupný?

Chemical Description

Name	alpha-D-mannopyranose
Synonyms	alpha-D-mannose; D-mannose; mannose
Formula	C6 H12 O6
Formal charge	0
Molecular weight	180.156 g/mol
Component type	D-saccharide, alpha linking

Chemical features

Atom count	24
Chiral atom count	5
Chiral atoms	C1 C2 C3 C4 C5
Observed leaving atoms	O1 HO1 HO2 HO3 HO4 HO6
Bond count	24
Aromatic bond count	0

Chemical Identifiers

Systematic name (ACDLabs)	alpha-D-mannopyranose
Systematic name (OpenEye OEToolkits)	(2S,3S,4S,5S,6R)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol

Chemical Descriptors

Stereo SMILES (CACTVS)	<chem>OC[C@H]1O[C@H](O)[C@@H](O)[C@@H](O)[C@@H]1O</chem>
SMILES (CACTVS)	<chem>OC[CH]1O[CH](O)[CH](O)[CH](O)[CH]1O</chem>
Stereo SMILES (OpenEye)	<chem>C([C@@H]1[C@H]([C@@H]([C@@H]([C@H](O1)O)O)O)O)O</chem>
InChI descriptor	InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5+,6+/m1/s1
InChIKey descriptor	WQZGKKKJIJFFOK-PQMKYFCFSA-N

InChI

- InChI = IUPAC International Chemical Identifier
- tento formát je unikátní pro všechny chemické substance
- byl vytvořen jako standard IUPAC v roce 2005
- je volně použitelný a šiřitelný pod licencí LGPL
- uchovává více informací než SMILES
- je stále čitelný pro člověka, který má dostatečnou praxi

InChI – jaké informace uchovává

- Identifikátor popisuje chemickou látku z několika pohledů:
 - atomy a vazby
 - tautomerie (může být vynechána, pokud není relevantní pro danou látku)
 - isometrie
 - stereometrie
 - elektrický náboj.

Proces překladau do InChI

- Algoritmus konverze struktury do InChI probíhá ve třech stupních:
 - 1) Normalizace - v tomto stupni se odstraní všechny redundantní informace
 - 2) kanonizace – v tomto kroku se každému atomu přiřadí jedinečné číslo
 - 3) Posledním stupněm je serializace, která generuje řetězec znaků.

Formát a vrstvy

- Každé InChI je uvozeno řetězcem „InChI=“
- Poté následuje číslo použité verze (v současné době „1“)
- Pak následuje písmeno S, splňuje-li toto InChI standard.
- Zbývající informace jsou rozděleny do šesti vrstev a podvrstev
- Každá tato vrstva obsahuje jiné specifické informace
- Oddělovačem vrstev je „/“ a začíná charakteristickým prefixem, s výjimkou vrstvy hlavní.

Hlavní vrstva

Musí být obsažena v každém InChI

- Sumární vzorec: nejprve jsou zapsány uhlíky, poté vodíky, následně ostatní atomy, které jsou v abecedním pořádku
- Vazby atomů (prefix: „c“): popisuje vazby mezi jednotlivými atomy v pořadí, v jakém byly očíslovány, tyto vazby jsou obsaženy pouze jednou
- Vazby atomů vodíku (prefix: „h“): popisuje, ke kterým atomům jsou navázány atomy vodíku

Vrstva nábojů

Protony (prefix: „p“): využívá se, pokud jsou v molekule kladné náboje

Elektrony (prefix: „q“): využívá se, pokud jsou v molekule záporné náboje.

Stereochemická vrstva

- dvojn  vazby a kumuleny (prefix: „b“)
- tetrahedrick  stereometrie atomů a allenů (prefix: „t“ „m“)
- jiný typ stereometrick  informace (prefix: „s“).

Izotopová vrstva

- (prefix: „i“, „h“)
- Dále využívá prefixů stereometrické vrstvy, pokud se jedná o izotopickou stereochemii.

Pevná vrstva H

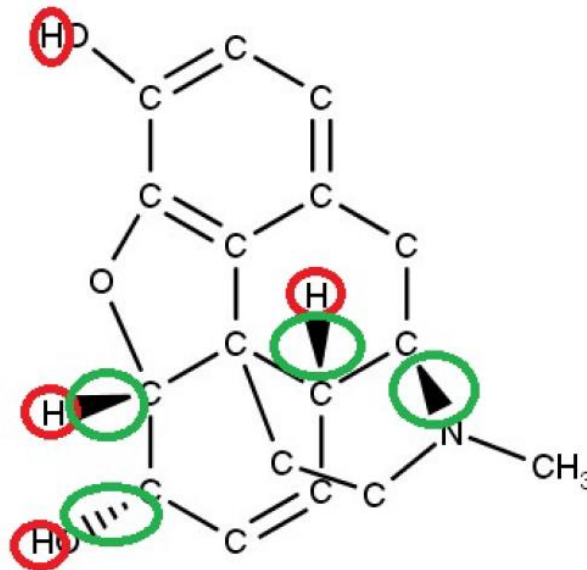
- (prefix: „f“)
- Tato vrstva se již ve standardním InChI nevyužívá, protože kumulovala informace z výše uvedených vrstev.

Znovu připojitelná vrstva

- (prefix: „r“)
- Tato vrstva se již ve standardním InChI nevyužívá, protože kumulovala informace z výše uvedených vrstev.

Příklad

- V InChI má molekula morfinu identifikátor:
InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14(15)17)8-11(10)18/h2-5,10-11,13,16,19-20H,6-8H2,1H3/t10-,11+,13-,16-,17-/m0/s1.
- **Červené písmo** se shoduje s vodíky na obrázku
Tato část je součástí hlavní vrstvy a musí být vždy přítomná.
- **Zelený text** jsou stereometrické informace - jsou označeny zelenými kolečky



InChIkey

- InChI je celkem dlouhý identifikátor nedeklarované délky
- to komplikuje jeho ukládání a další práci s ním
- proto byla vyvinuta alternativa, která z něj vychází
- kondenzovaný 27 znaků dlouhý InChIKey
- HASHInChI (algoritmus SHA-256)
- Pokud je vytvořen InChIKey z InChI, které je standardní, je standardní i InChIKey
- Díky použitému algoritmu je velmi malá pravděpodobnost duplicity mezi strukturami.

InChIkey

- InChIKey je rozdělen do několika částí:
AAAAAAAAAAAAAAAAA-BBBBBBBBFV-P
- „A“ označuje prvních 14 znaků a je vytvořeno zahashováním vazebných informací o molekule. Je zakončen pomlčkou.
- „B“ popisuje dalších osm znaků a je vytvořeno zahashováním zbytku InChI
- „F“ následuje znak identifikující druh InChIKey
- „V“ je identifikátor verze, v současné době se využívá „A“ pro první verzi, do budoucna se počítá s pokračováním abecedy pro verze další
- „P“ je protonový identifikátor

Příklad:

InChIKey molekuly morfinu:

BQJCRHHNABKAKU-KBQPJGBKSA-N

Cvičení

Najděte InChI zápis aspirinu.

V jakých databázích je dostupný?