ITMO UNIVERSITY

INFOCHEMISTRY SCIENTIFIC CENTER

Cheminformatics and synthetic biology: storing and searching chemical data

Prof. Sergey Shityakov

Infochemistry Scientific Center (ISC)

ITMO University

Saint-Petersburg, 2024

Chemical Data (DrugBank)

DRUGBANK Online | Explore | Data Library NEW | Academic Downloads | Interaction Checker

- **Identification**
  - Summary
  - Brand Names
  - Name
  - Accession Number
  - Background
  - Type
  - Groups
  - Structure
  - Weight
  - Chemical Formula
  - Synonyms
- **Pharmacology**
- **Interactions**
- **Products**
- **Categories**
- **Chemical Identifiers**
- **References**
- **Clinical Trials**
- **Pharmacoeconomics**
- **Properties**
- **Spectra**
- **Targets (10)**
- **Enzymes (9)**

Summary

Ibuprofen is an NSAID and non-selective COX inhibitor used to treat mild-moderate pain, fever, and inflammation.

Brand Names

Addaprin, Advil, Advil Cold and Sinus, Advil Congestion Relief, Advil PM, Advil Sinus Congestion and Pain, Alivio, Caldolor, Cedaprin, Children's Ibuprofen, Diphen, Duexis, Ibu, Ibutab, Junior Strengt READ MORE

Generic Name Ibuprofen

DrugBank Accession Number DB01050

Background

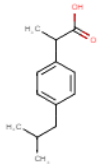
Ibuprofen is a non-steroidal anti-inflammatory drug (NSAID) derived from propionic acid and it is considered the first of the propionics.[7] The formula of ibuprofen is 2-(4-isobutylphenyl) propionic acid and its initial development was in 1960 while researching for a safer alternative for aspirin.[8] Ibuprofen was finally patented in 1961 and this drug was first launched against rheumatoid arthritis in the UK in 1969 and USA in 1974. It was the first available over-the-counter NSAID.[9]

On the available products, ibuprofen is administered as a racemic mixture. Once administered, the R-enantiomer undergoes extensive interconversion to the S-enantiomer *in vivo* by the activity of the alpha-methylacyl-CoA racemase. In particular, it is generally proposed that the S-enantiomer is capable of eliciting stronger pharmacological activity than the R-enantiomer.[24]

Type Small Molecule

Groups Approved

Structure



3D | Download | Similar Structures

Weight

Average: 206.2808
Monoisotopic: 206.13067982

Chemical Formula $C_{13}H_{18}O_2$

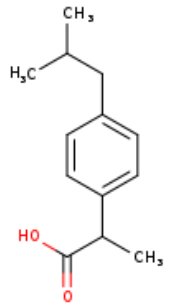
Synonyms
Show All Synonyms

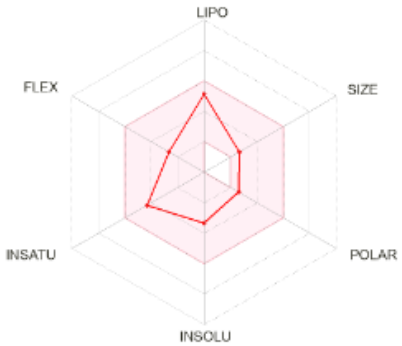
Ibuprofen | Ibuprofene | Ibuprofeno | Ibuprofenum | Ibuprophen

ITsMOre than a UNIVERSITY

Chemical Data (Swiss-ADME)

Ibuprofen





SMILES CC(Cc1ccc(cc1)C(C(=O)O)C)C

Physicochemical Properties	
Formula	C13H18O2
Molecular weight	206.28 g/mol
Num. heavy atoms	15
Num. arom. heavy atoms	6
Fraction Csp3	0.46
Num. rotatable bonds	4
Num. H-bond acceptors	2
Num. H-bond donors	1
Molar Refractivity	62.18
TPSA	37.30 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	2.17
Log $P_{o/w}$ (XLOGP3)	3.50
Log $P_{o/w}$ (WLOGP)	3.07
Log $P_{o/w}$ (MLOGP)	3.13
Log $P_{o/w}$ (SILICOS-IT)	3.15
Consensus Log $P_{o/w}$	3.00

Water Solubility	
Log *S* (ESOL)	-3.36
Solubility	9.09e-02 mg/ml ; 4.41e-04 mol/l
Class	Soluble
Log *S* (Ali)	-3.97
Solubility	2.23e-02 mg/ml ; 1.08e-04 mol/l
Class	Soluble
Log *S* (SILICOS-IT)	-3.44
Solubility	7.49e-02 mg/ml ; 3.63e-04 mol/l
Class	Soluble

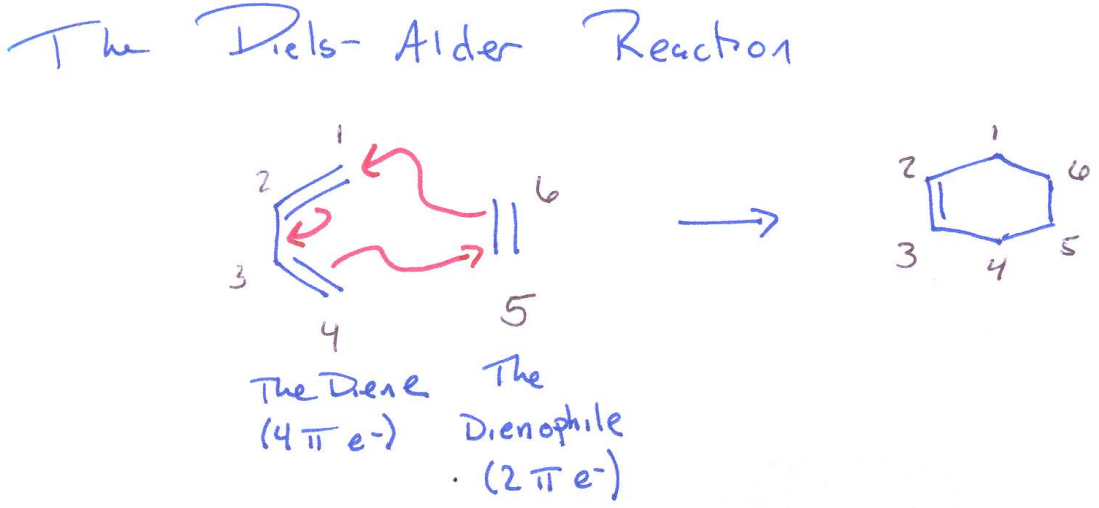
Pharmacokinetics	
GI absorption	High
BBB permeant	Yes
P-gp substrate	No
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	No
Log K_p (skin permeation)	-5.07 cm/s

Druglikeness	
Lipinski	Yes; 0 violation
Ghose	Yes
Veber	Yes
Egan	Yes
Muegge	Yes
Bioavailability Score	0.85

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 1 violation: MW<250
Synthetic accessibility	1.92

ITsMOre than a UNIVERSITY

Chemical Data

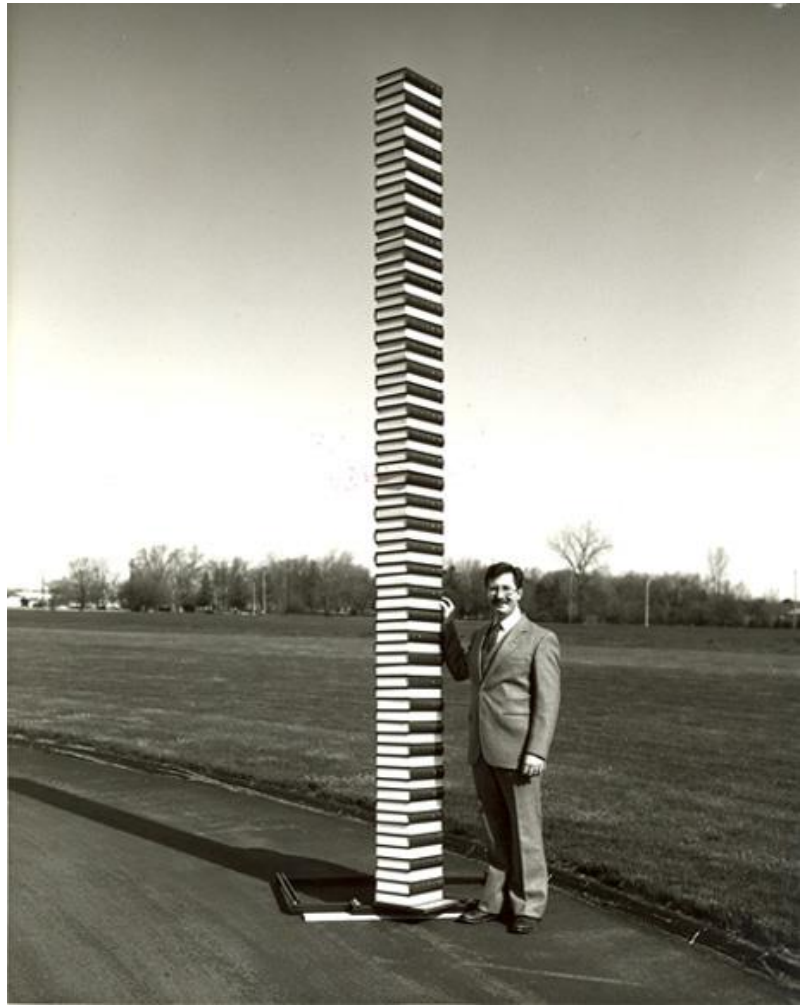


- Chemical data is special
- Chemical names are important (but inconvenient)
- Atoms connected by bonds can be thought of as a group of objects (atoms) that are connected together in a particular way (bonds)

Handling chemical data

What do we want to do with our chemical information?

- Display chemical compounds
  - 2D
  - 3D
- Search for
  - Structures
  - Substructures
  - Similar structures (2D or 3D)
  - Chemical reactions
- Name to structure conversion (and *vice versa)*



The tenth collective index of Chemical Abstracts consisted of 75 volumes and weighed 170 kg. It contained nearly 24 million entries.

Chemical file formats

- Every research institution has developed their own chemical structure format…
- Conversion between formats can be performed by programs like *babel* (http://openbabel.org)

```
> babel -L formats
abinit -- ABINIT Output Format [Read-only]
acr -- ACR format [Read-only]
adf -- ADF cartesian input format [Write-only]
adfout -- ADF output format [Read-only]
alc -- Alchemy format
arc -- Accelrys/MSI Biosym/Insight II CAR format [Read-only]
axsf -- XCrySDen Structure Format [Read-only]
bgf -- MSI BGF format
box -- Dock 3.5 Box format
bs -- Ball and Stick format
c3d1 -- Chem3D Cartesian 1 format
c3d2 -- Chem3D Cartesian 2 format
cac -- CAChe MolStruct format [Write-only]
caccrt -- Cacao Cartesian format
cache -- CAChe MolStruct format [Write-only]
cacint -- Cacao Internal format [Write-only]
can -- Canonical SMILES format
car -- Accelrys/MSI Biosym/Insight II CAR format [Read-only]
castep -- CASTEP format [Read-only]
ccc -- CCC format [Read-only]
cdx -- ChemDraw binary format [Read-only]
cdxml -- ChemDraw CDXML format
cht -- Chemtool format [Write-only]
cif -- Crystallographic Information File
ck -- ChemKin format
cml -- Chemical Markup Language
cmlr -- CML Reaction format
com -- Gaussian 98/03 Input [Write-only]
CONFIG -- DL-POLY CONFIG
CONTCAR -- VASP format [Read-only]
copy -- Copy raw text [Write-only]
crk2d -- Chemical Resource Kit diagram(2D)
crk3d -- Chemical Resource Kit 3D format
csr -- Accelrys/MSI Quanta CSR format [Write-only]
cssr -- CSD CSSR format [Write-only]
ct -- ChemDraw Connection Table format
```

Typical information in molecular structure files

Information about the whole molecule:

- molecule name
- journal article (for crystal structures)
- creator or author(s)

Information about each atom:

- atomic element (H, He, C, N, O, F, etc.)
- atom name (E.g. in an amino acid N, CA, CB, CO O, etc.)
- Cartesian coordinates (X, Y, Z) or Z-matrix atom number
- Atom charge (formal and/or partial)
- residue name (E.g. for a protein: Ala, Pro, etc.)
- temperature factor and occupancy for crystal structures

Bonding information:

- usually stored as a <u>connection table</u> which describes which atoms are bonded together.
- information about bond-orders (single, double, aromatic, etc.) is important, but it is not always stored in some file formats (E.g. pdb)

PDB file format

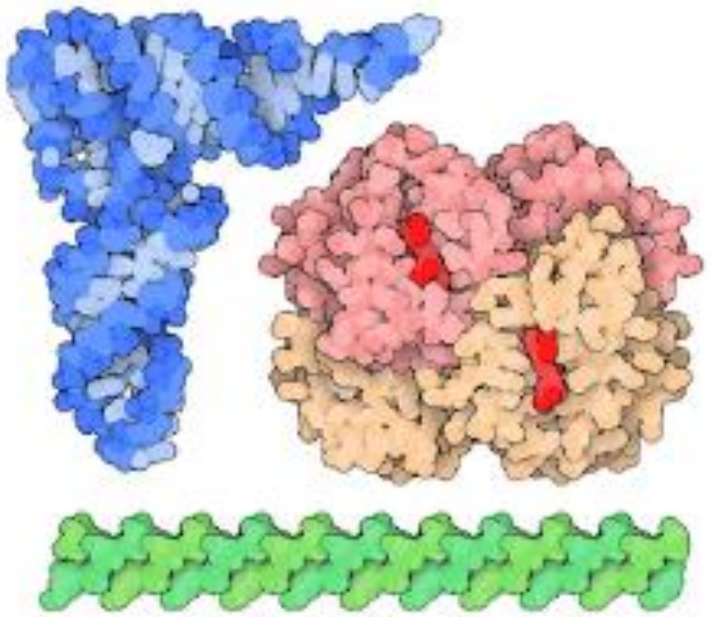
- The Protein Data Bank (PDB) file format was developed by the Brookhaven National Laboratory to store protein crystal structure information
- Used by many molecular modelling programs
- The PDB format has limitations:
  - Columns are of fixed size
  - Does not contain information about bond orders (these are recorded in a separate database)
- The Databank has developed new formats to replace the PDB format. e.g. the mmCIF format (Macromolecular Crystallographic Information File)

- Ref: http://www.rcsb.org/ pdb/info.html.

PDB

Atomic Coordinates: PDB Format

```
                                  Chain name
    Amino Acid                   /   Sequence Number
                \               /   /
     Element     \             /   /      -----Coordinates-----
                  \   \       /   /       X        Y        Z       (etc.)
ATOM       1  N    ASP L    1             4.060    7.307    5.186   ...
ATOM       2  CA   ASP L    1             4.042    7.776    6.553   ...
ATOM       3  C    ASP L    1             2.668    8.426    6.644   ...
ATOM       4  O    ASP L    1             1.987    8.438    5.606   ...
ATOM       5  CB   ASP L    1             5.090    8.827    6.797   ...
ATOM       6  CG   ASP L    1             6.338    8.761    5.929   ...
ATOM       7  OD1  ASP L    1             6.576    9.758    5.241   ...
ATOM       8  OD2  ASP L    1             7.065    7.759    5.948   ...
                \\
                 Element position within amino acid
```



Molecular Type	X-ray	NMR	EM	Multiple methods	Neutron	Other	Total
Protein (only)	146871	11954	7471	186	72	32	166586
Protein/Oligosaccharide	8676	31	1306	5	0	0	10018
Protein/NA	7750	277	2369	3	0	0	10399
Nucleic acid (only)	2445	1408	62	11	2	1	3929
Other	154	31	5	0	0	0	190
Oligosaccharide (only)	11	6	0	1	0	4	22
Total	165907	13707	11213	206	74	37	191144

ITsMOre than a UNIVERSITY

SD file format

- Developed by Molecular Design Limited (MDL).
- Can store 2D or 3D structures
- Can contain *query structures* which can contain variable atom and bond types. E.g an atom may be *either* nitrogen or carbon, or a bond could be *either* double or aromatic
- Can store additional information such as biological activity data associated with the molecule

```
-ISIS- 10229913002D

13 13 0 0 0 0 0 0 0 0999 V2000
 -0.0586   -1.1517    0.0000 C  0 0 1 0 0 0 0 0 0 0 0 0
 -1.7103   -0.5379    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  .
  .
  0.6069    1.4103    0.0000 C  0 0 2 0 0 0 0 0 0 0 0 0
  2.8138    1.3828    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  3.9207   -0.5379    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
 -3.9207    0.7414    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  1.2724    2.1586    0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
  3.9207    0.7414    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
1  2  1  0  0  0  0
1  3  1  0  0  0  0
1  4  1  1  0  0  0
   .
   .
1  5  1  6  0  0  0
6  10 1  0  0  0  0
10 13 1  0  0  0  0
12 16 1  0  0  0  0
15 18 2  0  0  0  0
M END
>  <Isis_internal_number> (2)
2

>  <chemical_name> (2)
Minaprine dihydrochloride

>  <smiles_code> (2)
c1(c2ccccc2)(cc(c(NCCN3CCOCC3)nn1)C).Cl.Cl

>  <Plate position> (2)
66

$$$$
```

Atoms

Bonds

Data

Record end

Chemical Line Notations

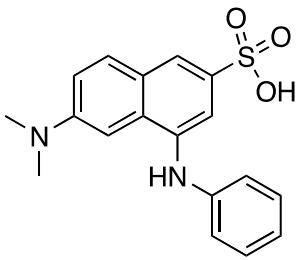
- Representations of molecules that fit on a single line. E.g. standard structural formulas. These work well for linear compounds, but less well for rings…

H_2O

Line notations are:

- Compact
- Generally human readable/understandable

There are many line notations



6-Dimethylamino-4-phenylamino-naphthalene-2-sulfonic acid

- Wiswesser line notation: 1L66J BMR& DSWQ IN1&1 – An early line notation *(1949)* that describes molecules as fragments. Used for databases but fell out of use because it is not very computer friendly
- Rosdal: 1=-5-=10=5,10-1,1-11N-12-=17=12,3-18S-19O,18=20O,18=21O,8-22N-23,22-24 – A linear representation of a connection table developed by Beilstein
- SMILES: CN(C)C1=CC=C2C (C(NC3=CC=CC=C3)=CC(S(=O)(O)=O)=C2)=C1 - Developed by Dave Weninger and Daylight Chemical Systems
- InChi – InChI=1S/C18H18N2O3S/c1-20(2)15-9-8-13-10-16(24(21,22)23)12-18(17(13)11-15)19-14-6-4-3-5-7-14/h3-12,19H,1-2H3,(H,21,22,23) - A compact chemical representation developed by IUPAC
- Sybyl Line Notation (SLN, Tripos)

SMILES

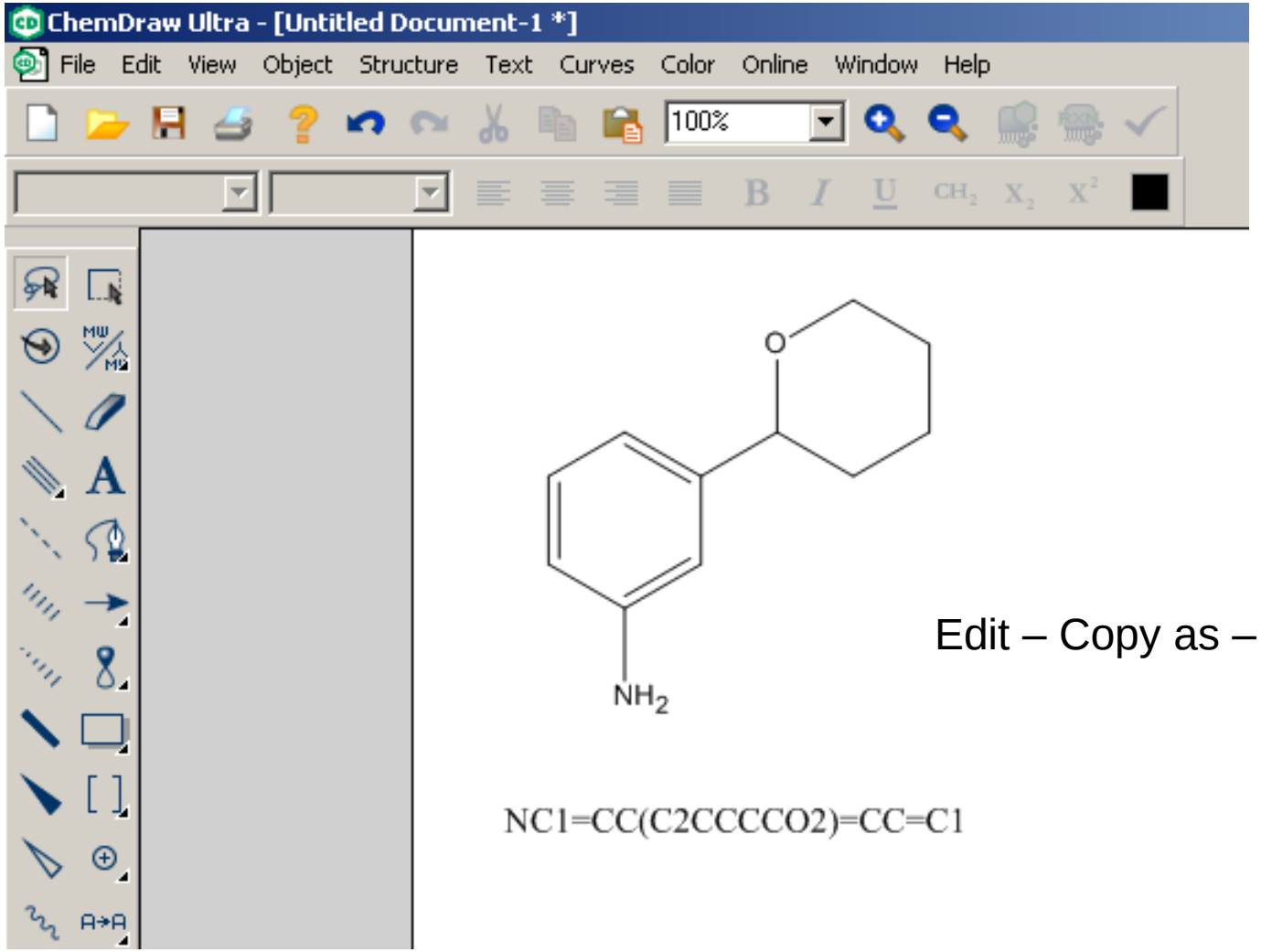
- SMILES is the most widely used and most useful chemical line notation
- Can be used as input/output in many programs

- Simple SMILES strings resemble standard chemical nomenclature. The atoms commonly found in organic molecules (B, C, N, O, P, S, F, Cl, Br, I) are represented by the atomic element symbol.
- Single bonds are implied between each atom.
- Hydrogen atoms are not usually shown but can be included in square brackets

Example smiles

Ethanol	CCO
Acetic acid	CC(=O)O
Cyclohexane	C1CCCCC1
Pyridine	c1cnccc1
Trans-2-butene	C/C=C/C
L-alanine	N[C@@H](C)C(=O)O
Sodium chloride	[Na+].[Cl-]

Generate SMILES using Chemdraw



Edit – Copy as – smiles

SMILES – simple examples

SMILES string	*Compound*
C	Methane
CC	Ethane
N	Ammonia
[NH3]	Ammonia
CCCCCCO	1-hexanol

SMILES - branches

- Branches are represented by enclosing the side-chain in parentheses ‘()’

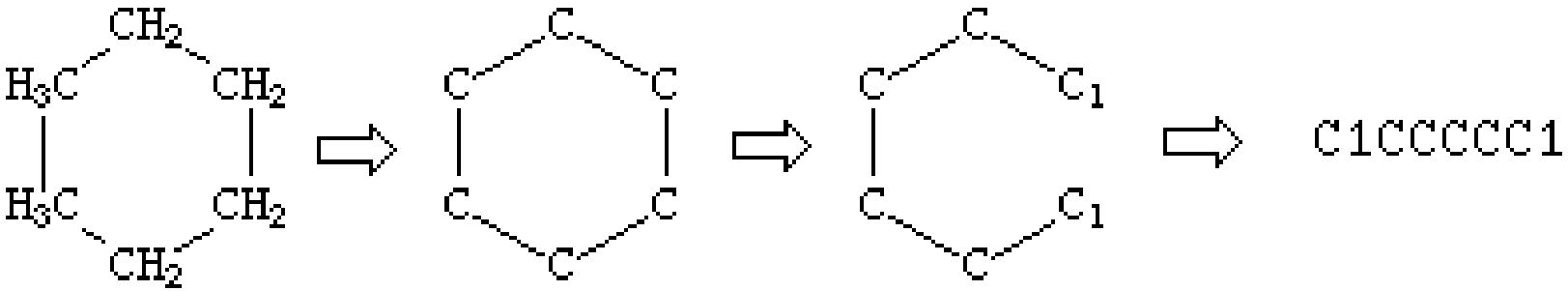
CC(=O)O	Acetic acid
OC(C)(C)C	*t*-butyl alcohol

SMILES - rings

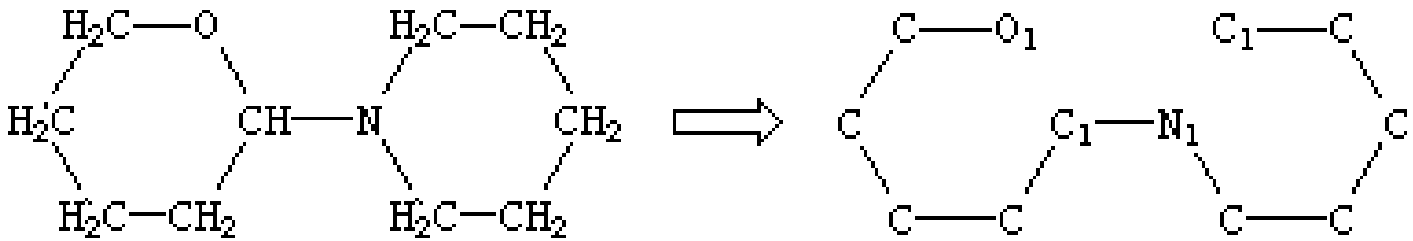
- Rings are specified by using numbers to create ‘ring closures’. The number follows after the atom.
- Lower case characters are used to specify aromatic rings

C1CCCCC1	cyclohexane
c1ccccc1	benzene
n1ccccc1	pyridine
c1cccc2c1cccc2	naphthalene

SMILES – rings

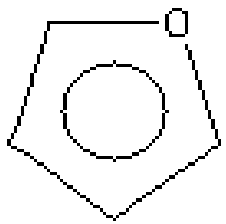


Cyclohexane – we need to break one bond in the ring

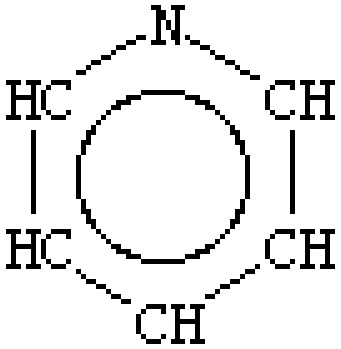


O1CCCCC1N1CCCCC1

SMILES – aromaticity



Furan
C1=COC=C1
c1cocc1



Pyridine
N1=CC=CC=C1
n1ccccc1

C1=CC=CC=C1 benzene but more usually **c1ccccc1**

SMILES - Additional notations

- SMILES contains additional features which can be used to describe chirality, double bond isomers (E, Z) and metal complexes.
- These are described in more detail at

http://www.daylight.com/learn/

http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html#RTFToCX1

SMILES - benefits

- Smiles is essentially a language with simple letters, bonds and rules
- They are extremely compact and use little storage space
- But I can write ethanol two ways
- CCO
- OCC
- The two ‘words’ are different. What can we do about this if we want to search databases?

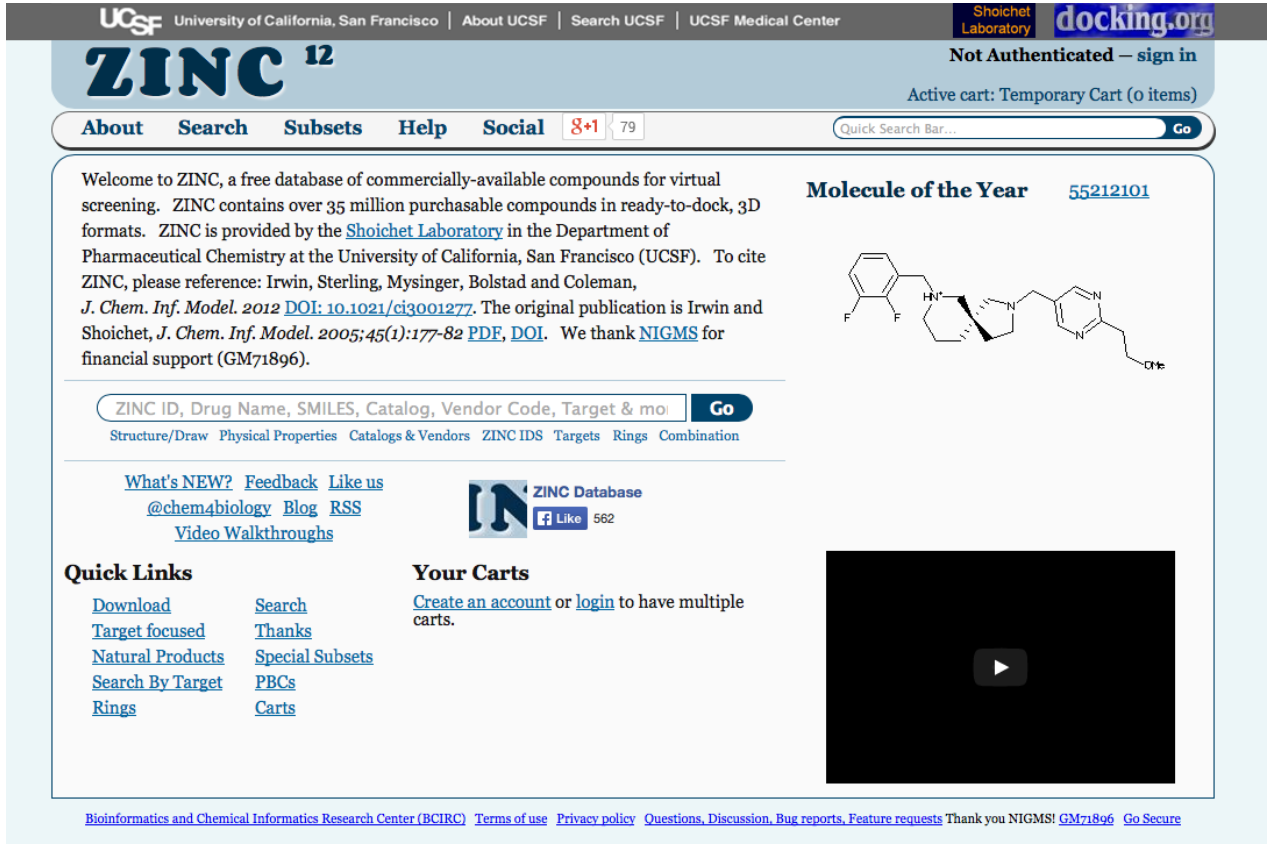
Chemical Databases

- Chemical databases are important in all stages of medicinal chemistry
- Databases may contain:
  - Chemical structure, reaction and synthetic data (e.g. Beilstein, Chemical Abstracts, the Merck Index)
  - Compound structure and synthesis information (e.g. An in-house compound registry)
  - Biological activity data such as in-house testing data or MDL Drug Data Report (MDDR)

CAS registers 89 million compounds and 39 million patent and journal articles

ZINC

The ZINC database (http://zinc.docking.org/) collects together commercially available compounds, converts them to 3D structures and creates a number of useful subsets for drug desing (druglike, leadlike, etc, etc).



Chemical structures are special

- The important distinction between chemical database software and other database programs used for holding text or images is that a chemical database must be able to *interpret chemical structures*

- In a chemical database it is desirable to be able to search for:
  - Individual exact compounds
  - Compounds containing a particular substructure
  - Compounds similar to a given structure

Thank you for your attention