IOCB AS CR

Advanced *in silico* Drug Design

KFC/ADD

Chemical Libraries

Karel Berka, Ph.D.
Jindřich Fanfrlík, Ph.D.
Martin Lepšík, Ph.D.

Biotrend

UP Olomouc, 1.-3.2. 2016

elixir CZECH REPUBLIC

Motto

Q: What make compound a good drug?

A: What? Give it to the patient - if he survives and gets better, it was a good drug.

M. Paloncýová, ústní sdělení

Chemical Universe

- **C, H, O, N,**
- **P, S, F, Cl, Br, I**
- **MW < 500**

Chemical Universe

10^6 human proteins, RNA,…

targets

10^{40} - 10^{120} compounds

drugs

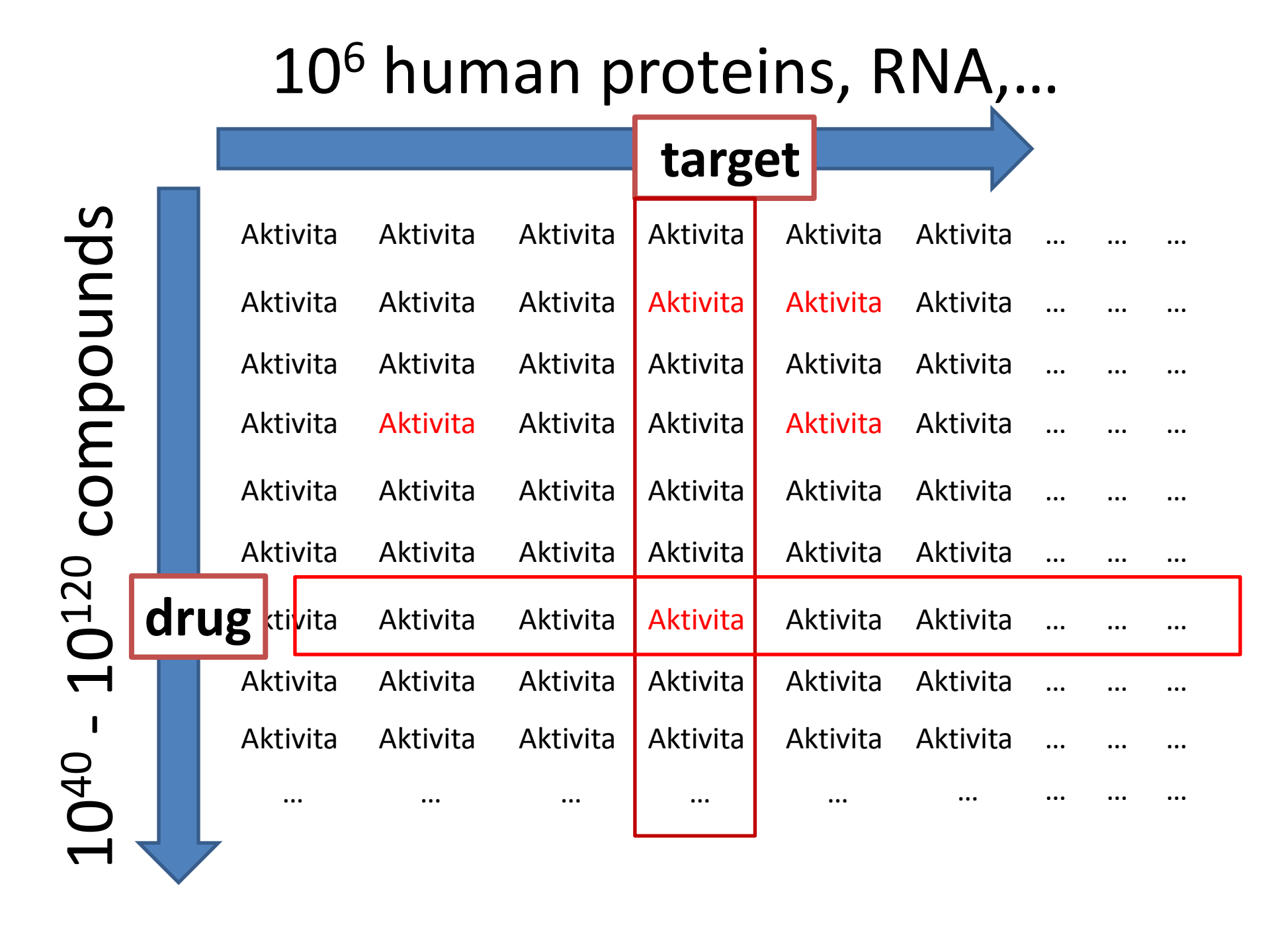
10^6 human proteins, RNA,...

targets

10^{40} - 10^{120} compounds

Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	Aktivita	...	...	...
...	...	...	...	...	...	...	...	...

drugs


10^6 human proteins, RNA,…
target
10^{40} - 10^{120} compounds
drug
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
ktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
Aktivita Aktivita Aktivita Aktivita Aktivita Aktivita … … …
… … … … … … … … …

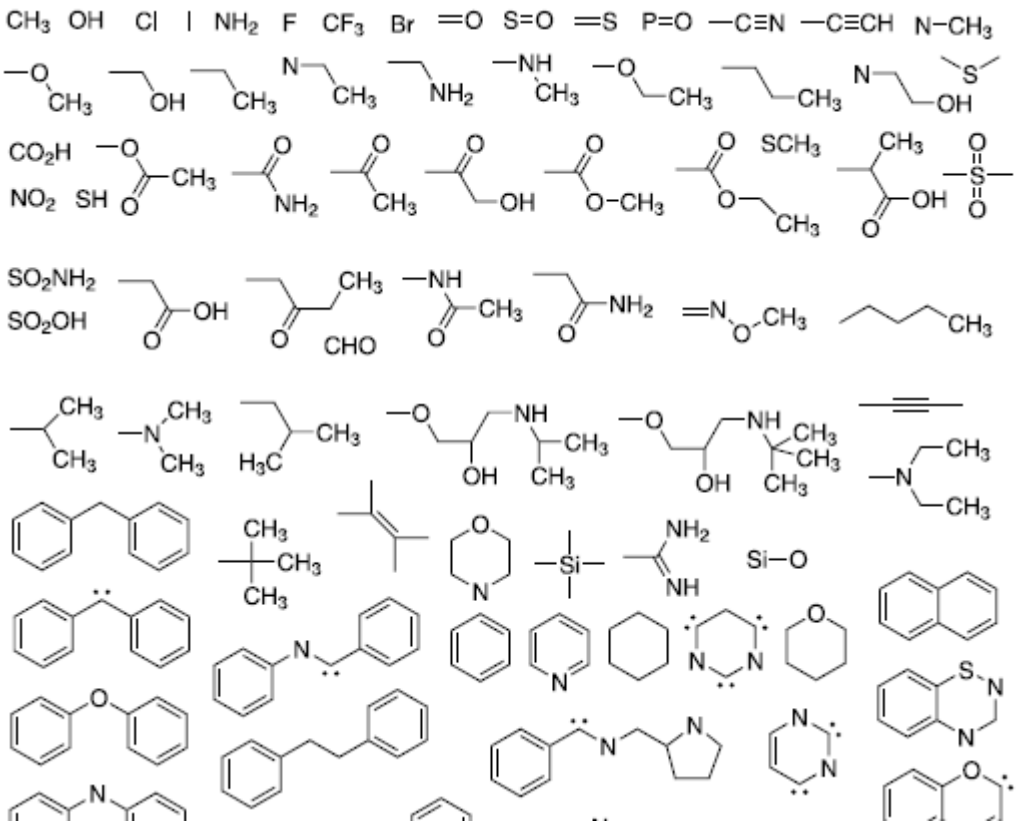
10^6 human proteins, RNA,…

10^{40} - 10^{120} látek

… … … … … … … … …
… … …
… … …
… … …
… … …
… … …
… … …
… … …
… … …
… … …

Chemical Libraries

- Large diversity
  - Lead search
- Specific
  - For combinatorial chemistry
- Typical motives:

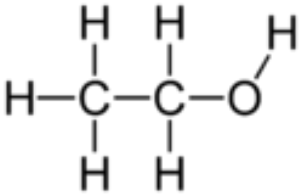


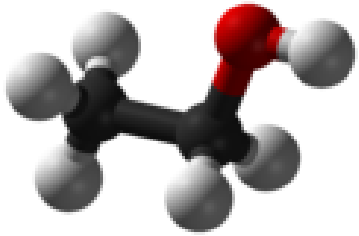
David C. Young - Computational Drug Design: A guide for computational and medicinal chemists. Wiley-Blackwell, New York, 2009, ISBN 978-0470126851

How to Store Chemical Compound

- Types:
- 1D, 2D, 3D

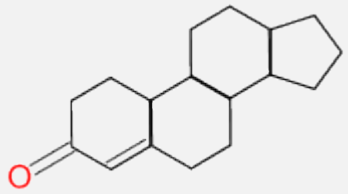
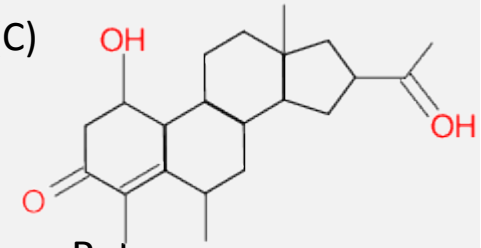
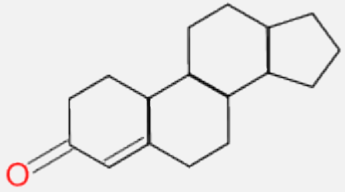
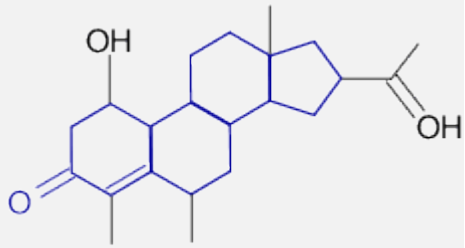
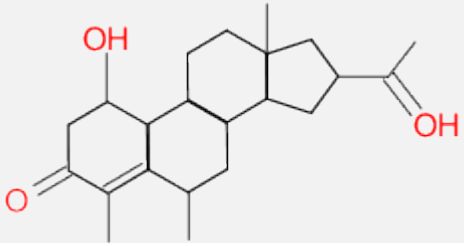
Ethanol: CCO





- What for:
- Similarity search
  - 2D and 3D similarity, motifs
- Predictions
  - pKa, logP/logD, charge distribution, logS, …

Chemical Information System

Operation	Classical information system		Chemical information system	
Storage	Name = ‘KSICHT’	Store text, numbers, pictures,…		Store chemical structures and information about them
Search	Search $Name	Returns: ‘KSICHT’	Search CC(=O)C4CC3C2CC(C)C1=C(C)C(=O)CC(O)C1C2CCC3(C)C4	Returns: 
Advanced searches	Find queries containing ‘chemist’	Returns: ‘chemist’ ‘taky chemist’ ~~‘bum’~~	Searc molecules containing: 	Returns: 
Questions	How to became chemist?	Returns: ‘Solve KSICHT’	Calculate logP(o/w) of: 	logP(o/w) = 2.62

1D Structure Representation

Stores molecule in **string** format

- **CAS number**
  - registered molecules only
- **SMILES**
  - simple format
- **InChI**
  - IUPAC format - more comprehensive

SMILES

Simplified Molecular-Input Line-Entry System

- Chemical graph

Atoms: organics with implicit H (B, C, N, O, P, S, X),

anorganics or isotopes - [Au], [2H],
charge – Ti+4 or Ti++++

aromates – small print (cccccc – benzene)

Bonds: simple – without sign (CC – ethane, O – water)

double (O=O), triple (N#N – nitrogen),
four ([Ga-]$[As+]),

ring breaking – numbers (%10)

(C1CCCCC1 – cyklohexane, c1cccc2c1cccc2 - naphtalene)

Branching: brackets (C(Cl)(Cl)Cl – chloroform)

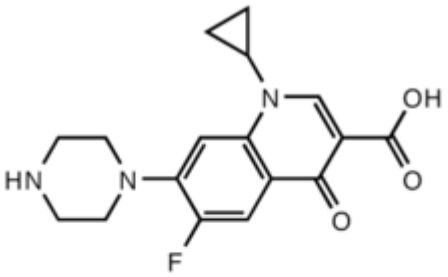
Stereochemistry: "/" a "\"

(F/C=C/F – trans-difluoroethen)

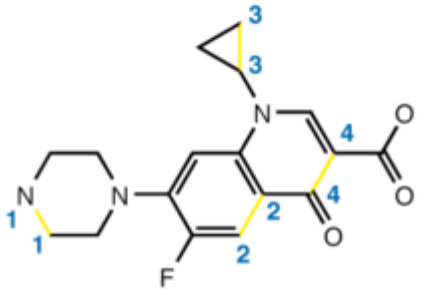
chiral atoms @ (c.clockwise) or @@ (clockwise)

(N[C@@H](C)C(=O)O – L-alanin)

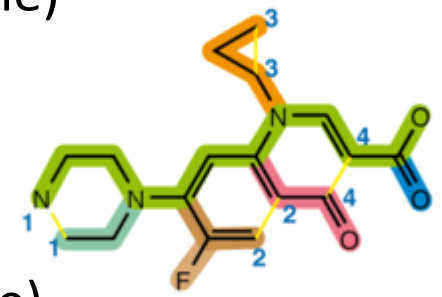
A



B



C



D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

SMARTS

- **SM**iles **AR**bitrary **T**arget **S**pecification
  - Selection of atomic regular expressions
- **Atoms** – symbol or atomic number [C],[#6],[C,c]
  - aromates small print [c],
  - Regular expression: * (any),A (aliphatics), a (aromate)
- **Bonds** – '-' (simple), '=' (double), '#' (triple), ':' (aromatic), '~' (any bond)
- **Connectivity** – X (different) a D (same) deskriptors - [CX4] carbon with 4 other atoms, [CD4] – quarter carbon
- **Cyclicity -** R descriptor - [CR] (aliphatic carbon atom in ring)
- **Logical operators** – (and= ; &) (or= ,) (not= !)

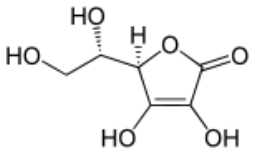
[N;H3;+][C;X4] (primary amine)

InChI

- IUPAC **In**ternational **Ch**emical **I**dentifier

CH_3CH_2OH
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)

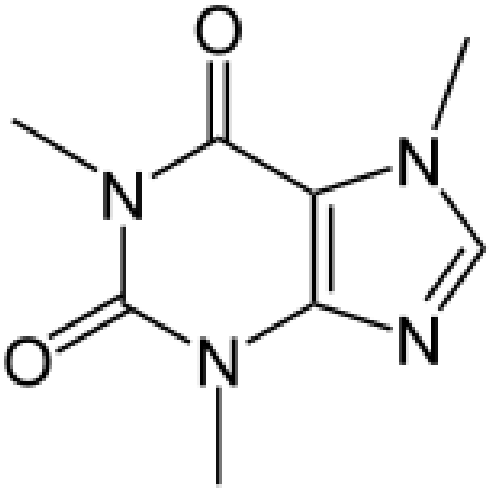
InChI=1(S standard)/chemical formula/c(atom connections)/h(hydrogens)/p(protons)/q(charges)/b(double bonds)/t ev. m(tetrahedral) /s(stereochemistry)/i(isotopes)/f/r

- **InChIKey** – for quicker searches
  - 27 long strings – 14 characters long hash of connectivity ~ 9 characters hash of other properties

Ethanol: LFQSCWFLJHTTHZ-UHFFFAOYSA-N
Ascorbic acid: CIWBSHSKHKDKBQ-JLAZNSOCSA-N

2D Structure Representation

- CHM – ChemDraw
- CDX – ChemDraw exchange file
- SDF, MOL(2)!



- Topology and connectivity

MOL/SDF

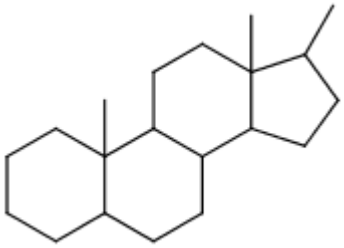
Row	Section	Description
1-3	header	
1		Name of molecule („benzene“)
2		Additional information
3		Comment
4-17	Connection table	
4		Sum of lines: 6 atoms, 6 bonds, ..., V2000 standard
5-10		atoms (1 row per atom): x, y, z, element, etc.
11-16		bonds (1 row per bond): 1. atom, 2. atom, type, etc.
17		properties
18	$$$$	End of molecule SDF can hold whole database

- MDL molfile, structure-data file

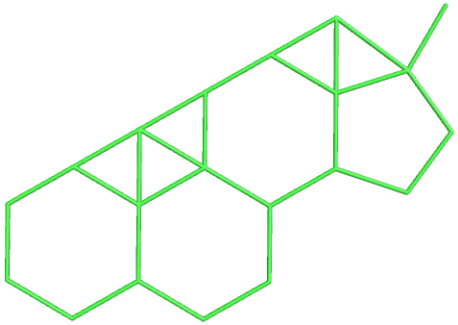
```
benzene
ACD/Labs0812062058

6 6 0 0 0 0 0 0 0 0 1 V2000
1.9050 -0.7932 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.9050 -2.1232 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7531 -0.1282 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7531 -2.7882 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-0.3987 -0.7932 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-0.3987 -2.1232 0.0000 C 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
M END
> <Molecular Weight>
499.61

$$$$
```

Ligand Structure Generation from 2D Stereochemistry

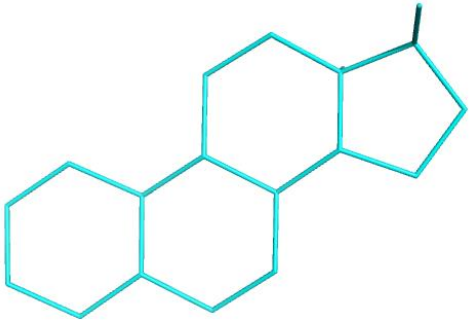
Experimentalist drawing

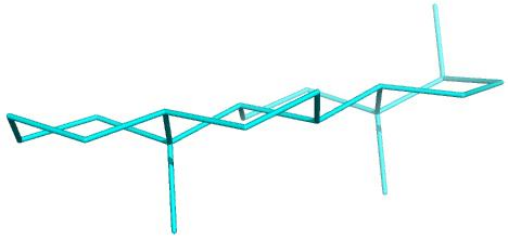


SDF

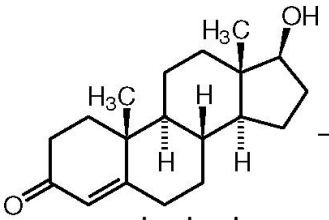


3D optimization

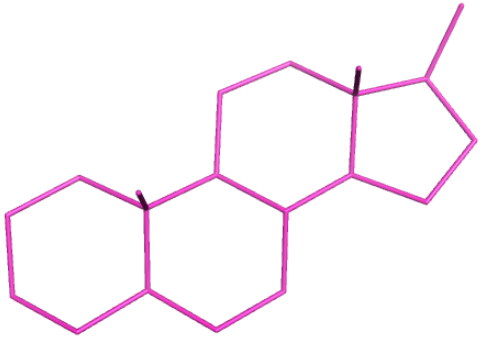


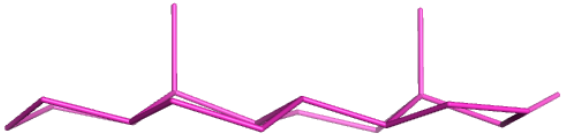


right stereoisomer



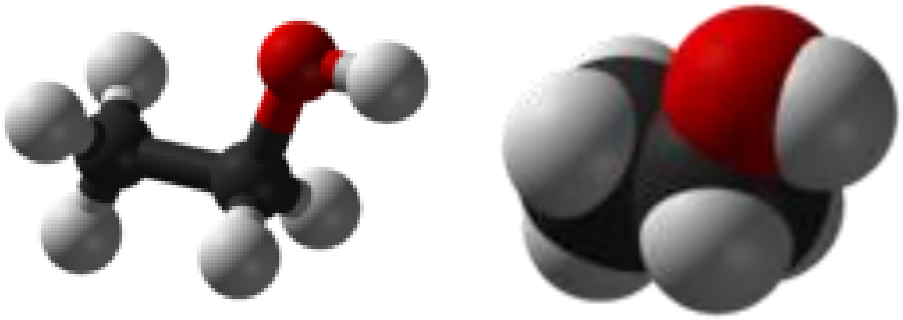
testosterone





3D Structure Representation

- MOL, SDF
- XYZ
- PDB



XYZ

- Free format
- Easy storage

row	Section	Description
1-2	Header	
1		Number of atoms
2		Comment
3-X	Block of atoms	(1 row per atom):
5-10		element, x, y, z
		More structures stored as multiple entries

```
5
methane molecule (in [[Ångström]]s)
C 0.000000 0.000000 0.000000
H 0.000000 0.000000 1.089000
H 1.026719 0.000000 -0.363000
H -0.513360 -0.889165 -0.363000
H -0.513360 0.889165 -0.363000
```

PDB - Protein DataBank file

```
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
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
...
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
...
```

Databases

- Small molecules
  - PubChem
  - ChEMBL
  - ChEBI
  - ChemSpider
  - ZINC ...
- Targets
  - PDB
  - BindingDB
  - Uniprot …

How to Search in Chemical Databases

- Text-wise (PDB - description)
- ID wise (PDBID, CAS, PubChemID, …)
- By Substructure (when looking for compounds with specific functional groups)
- By Superstructure (when looking for fragments)
- By Similarity (2D or 3D)
- By Properties (MW, number of atoms, LogP,…)