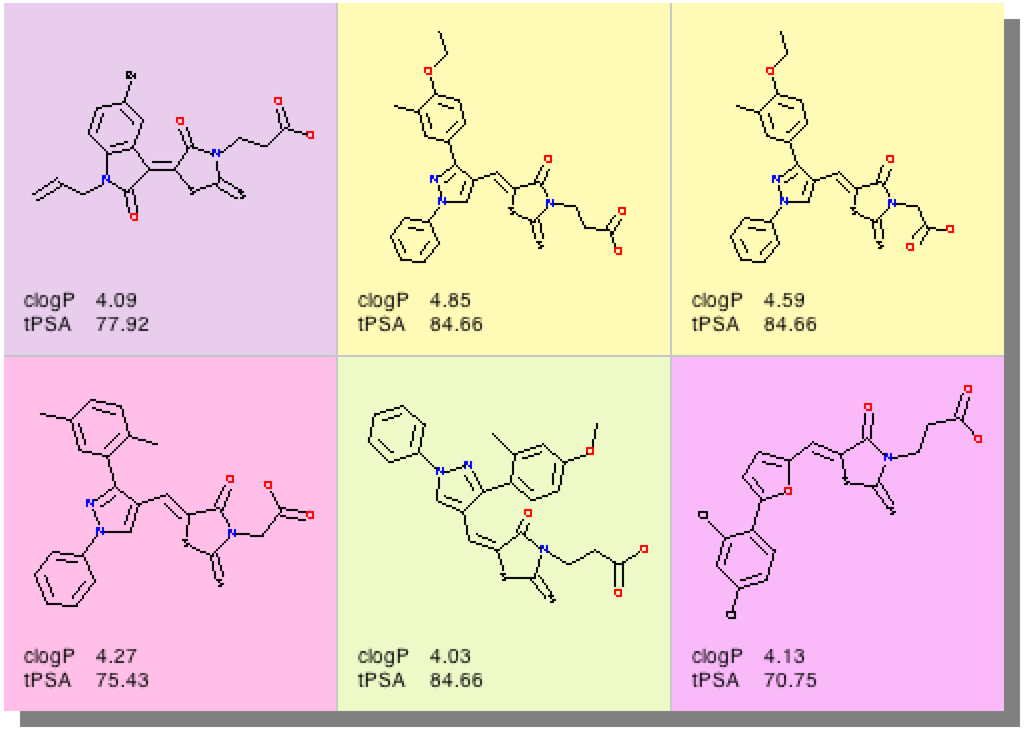
Cheminformatics

Brno. Stuctural Bioinoformatics



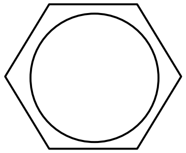
Cheminformatics

- What is cheminformatics?
- Reading chemical structures from .sdf and .mol2
- Chemical Tables (Sorting, Grid view, ..)
- Aromaticity
- Chirality
- Tautomer Enumeration
- Chemical Editor and View Modes
- Searching a Chemical Database (substructure & similarity)
- Formulating a Chemical Query
- Chemical Clustering and Trees
- Converting a Selection to 3D
- Property Prediction (LogP, LogS, DrugLikeness,…)

Cheminformatics: Introduction

- Explosive growth of the commercial chemicals.
- Millions of compounds available from 10-20 major vendors.
- Quickly changing: about 10-20% of a database may change every 3 months. New compounds emerge, old disappear
- Storage, Manipulations and Export of Chemical Libraries
- Predicted and Experimental Compound Properties
- Screening and SAR data
- Analog Design
- Virtual Compound Libraries
- Searching Chemical Libraries

Linear String Notation of Chemicals: Smiles and Smarts



c1ccccc1

Smiles - Chemical Structures:

- A good tutorial can be found at the Daylight site: http://www.daylight.com/dayhtml/smiles
- Common atoms are represented by element symbols: C,N,O,Cl, ..
- Rare elements, charges, isotopes, are shown like this [Au], [H+]
- Single bonds are not shown, double bonds are '=', tripple: '#'
- Branching is shown by parentheses (e.g. CC(=O)O)
- Ring closure is shown by matching digits (C1CCCC1)

Smarts - Chemical Patterns:

- [C,N,O] a list of possibilities, '*' - any atom, ~ any bond
- [C;R] in ring,

Smarts: Atoms and Bonds

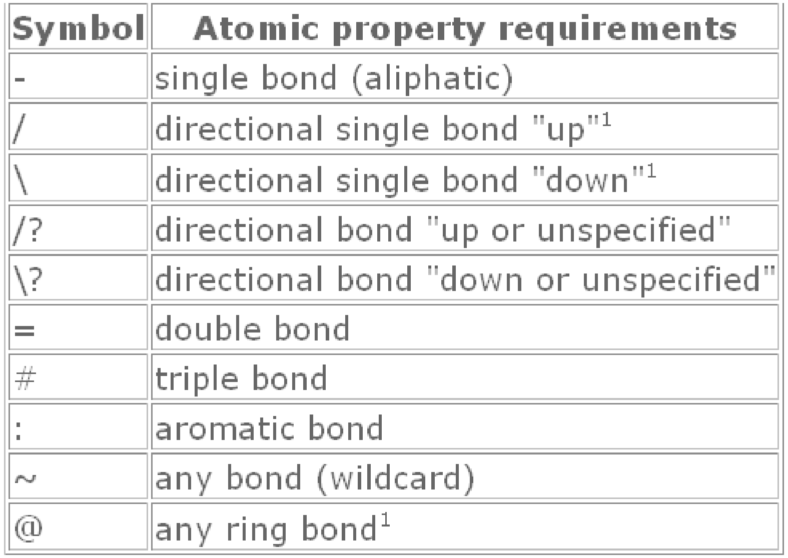
- http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

Atoms

* - any, a – aromatic, A – aliphatic, D – neighbors, Q – noncarbon,

H*n* – total H-count, R*n* -ring membership, **r***n* – ring size, **v***n* – valence, #*n* – atomic number, @ - chirality, @@ clockwise chirality

Bonds



Symbol	Atomic property requirements
-	single bond (aliphatic)
/	directional single bond "up"[1]
\	directional single bond "down"[1]
/?	directional bond "up or unspecified"
\?	directional bond "down or unspecified"
=	double bond
#	triple bond
:	aromatic bond
~	any bond (wildcard)
@	any ring bond[1]

Logical operations:

! – not

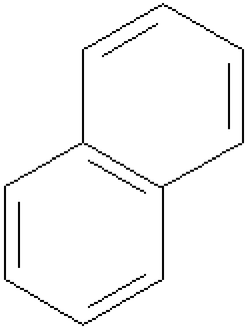
& - and

, (comma) – or

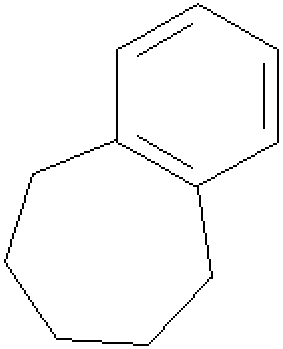
; - and (weak)

Chemical Searching, Smarts

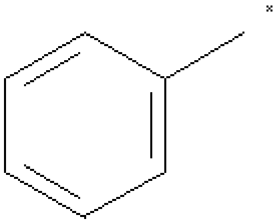
Hydrogens H0, H1, H2.., Valence: v5, Isotopes



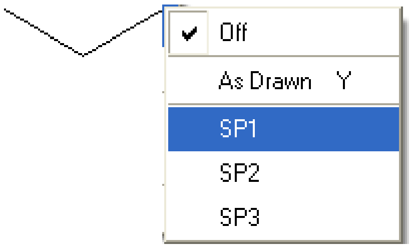
Ring Membership
R1, R2



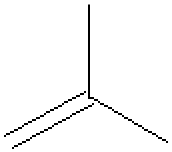
Ring Size
r5,r6



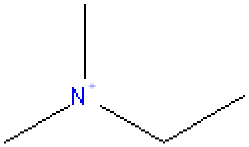
Attachment Point
[*] or [C*]



Hybridization
^3,^2,^1



Connectivity
D3,D2

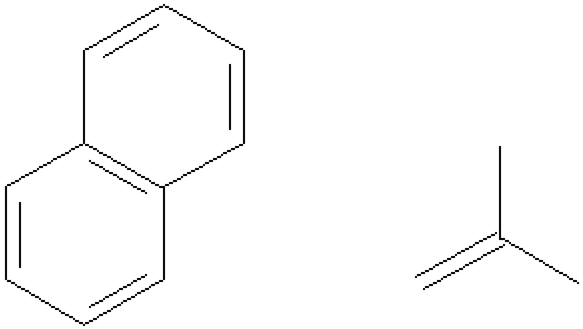


Charge
[Mg++] [N-]

Chemical Searching: several fragments

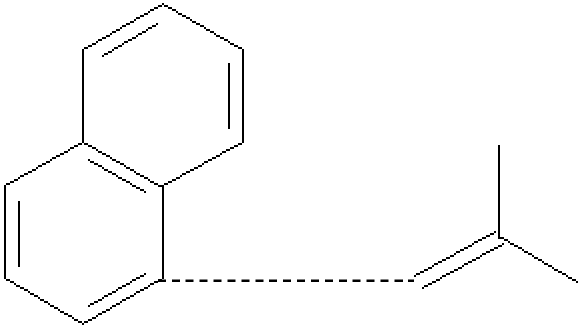
Just draw them side by side.

Smart: Smart1 dot Smart2, e.g. CCO.O



Chemical Searching: variable bond distance

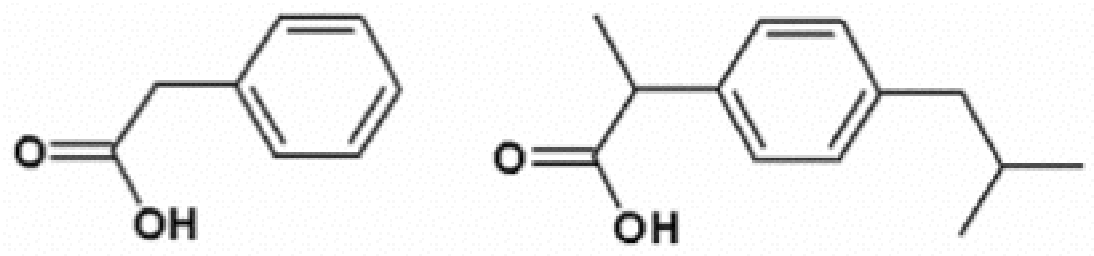
A special type of bond can be introduced that defines a range of interatomic distances in terms of the minimal number of chemical bonds



Chemical Searching: Properties

Chemical Searching: Similarity

- Exact Match
- Substructure Searching
- Pattern Searching
- Similarity Searching (Tanimoto of Fingerprints)

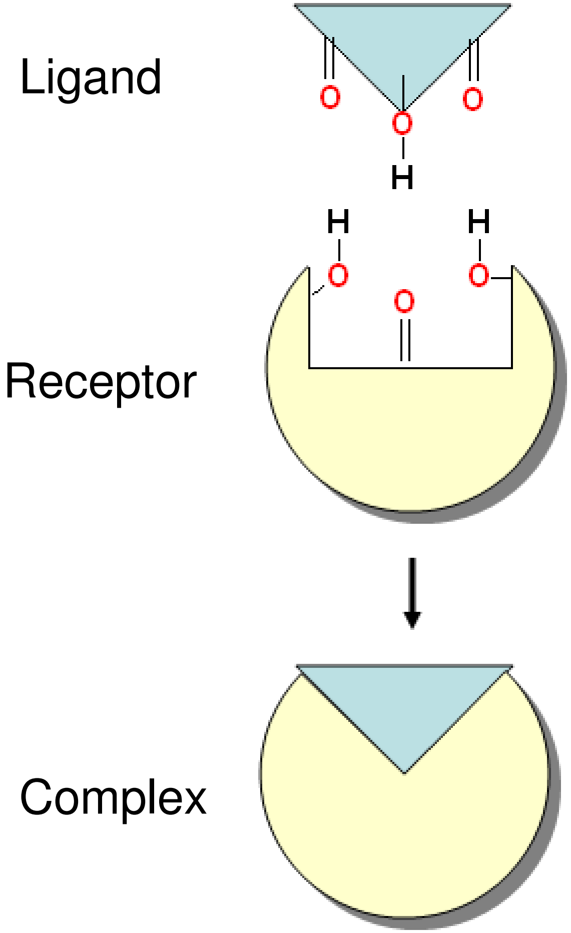


Chemical Similarity

- Divide both structures (A and B) into small fragment
- Merge fragment lists and form two “bit-strings”, e.g. 010001000111 and 101111011001
- Calculate a Tanimoto distance as nAB/nTotal
  nAB is the number of on-bits which are in common.
- Tanimoto distance is between 0.0 and 1.0

Pharmacophore Searching

A pharmacophore was first defined by Paul Ehrlich in 1909 as "a molecular framework” that carries the essential features responsible for a drug’s biological activity" (Ehrlich. Dtsch. Chem. Ges. 1909, 42: p.17).



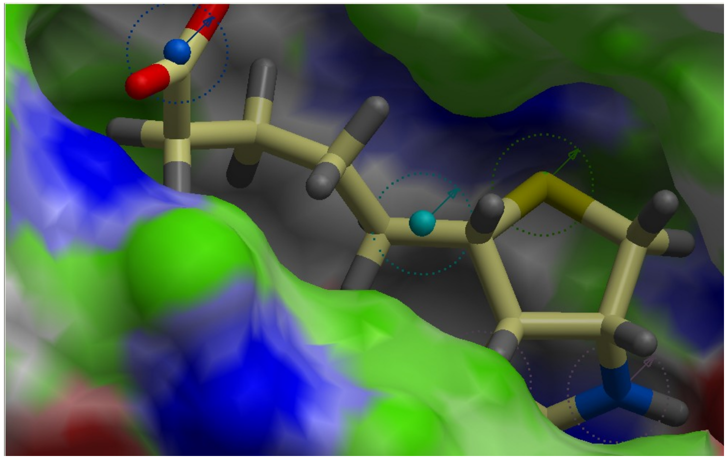
Hydrogen bond acceptors

Hydrogen bond donors

Charge

Hydrophobicity

Aromatic Ring Centers

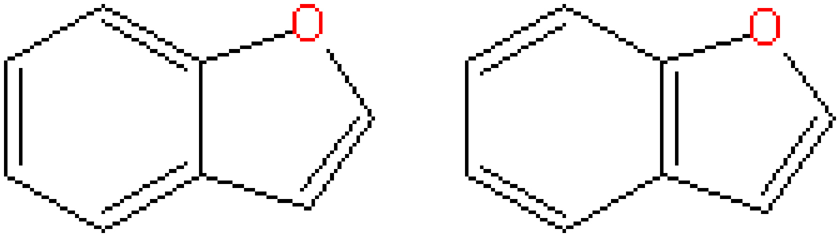


Converting to 3D

- Mol-files: 0D (x=y=z=0), 2D (z=0), 3D
- Necessary for conversion:
  - Correct bond orders
  - Formal charges
  - Stereo indicators (chirality, cis-trans for 0D)
  - ICM converts 2D pictures and optimizes them in MMFF94 force field.

Aromaticity

- Requires a *cyclic conjugated array of π orbitals in the same plane.*
- Huckel's Rules: The total number of electrons in the π system $4n+2$, where n=0,1,2
- Identical aromatic systems can be represented by different pattern of single and double bonds. ICM/Molcart matches aromaticity, not = and _.



Example: Two ways to draw bonds in an aromatic system

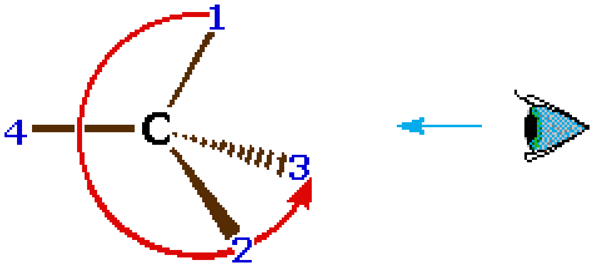
Chirality

- “I call any geometrical figure, or group of points, *chiral*, and say that it has chirality, if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself”. *Lord Kelvin, Baltimore Lectures, 1884*
- Chiral *centers*: Four ***different*** substituents.
- Four states: Unset; **R** (rectus); **S** (sinister) and RS (unknown, can be a racemic mixture).
- Chirality (R or S) can be shown with stereo-bonds.
- You may need to enumerate all ***enantiomers*** to find which one has biological activity
- There are more complex types of chirality (e.g. axial chirality)

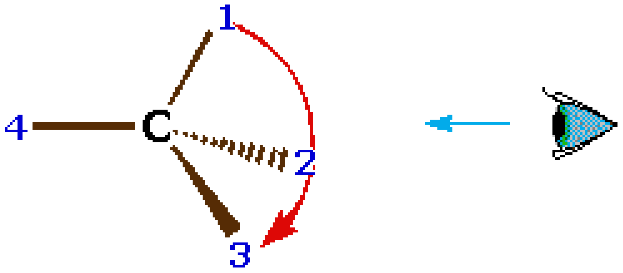
Stereoisomers

The Sequence Rule for Assignment of Configurations to Chiral Centers Assign sequence priorities to the four substituents by looking at the atoms attached directly to the chiral center.

The higher the atomic number of the immediate substituent atom, the higher the priority.



(*S*)

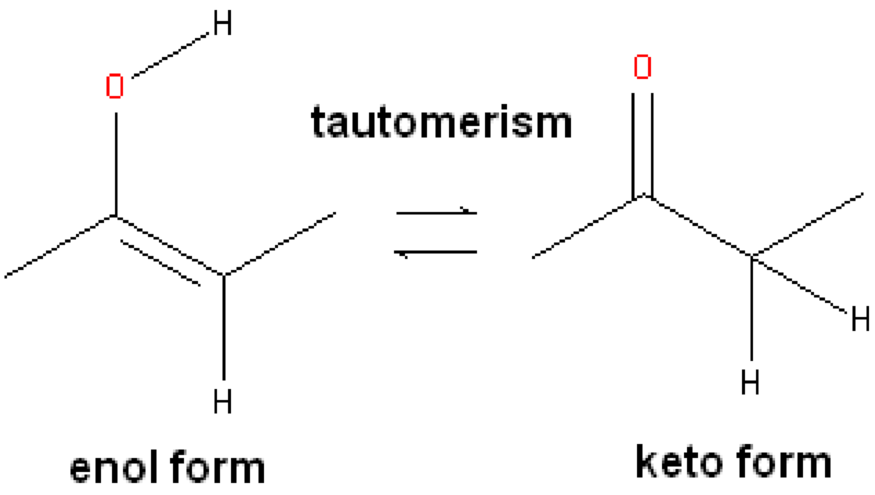


(*R*)

ICM will automatically determine the state and generate stereoisomers

Tautomers

Tautomers are formed by an interconvertible reaction called tautomerization whereby there is a formal migration of a hydrogen atom along with a switch of a single bond and an adjacent double bond.



During tautomerization a chemical equilibrium of the tautomers will be reached based on several factors, including, pH, temperature and solvent.

Vendor Compounds vs Natural

General Trends

Vendor chemistry

- Non-charges
- Non-chiral
- Simple
- Flexible

Natural compounds

- Charged
- Chiral
- Complex
- Rigid

Compound Properties, ADMET

- Potent compound in vitro fail because of:
- **A**bsorption (from gut, through membranes, to blood). FDp (Fraction of Dose in *p*ortal vein). Fh (*h*epatic vein)
  - Disposition: from plasma to cells
  - Elimination: mainly via liver and kidney
- **D**istribution (to tissues)
- **M**etabolism
- **E**xcretion
- **Tox**icity

Lipinski Rule of Five

Correlates with absorption and permeation

- <= 5 hydrogen-bond donors
- <= 5*2 HB-acceptors
- <= 500 molecular weight
- cLogP <= 5
- (extra: <= 5 torsions)

Properties (Contd)

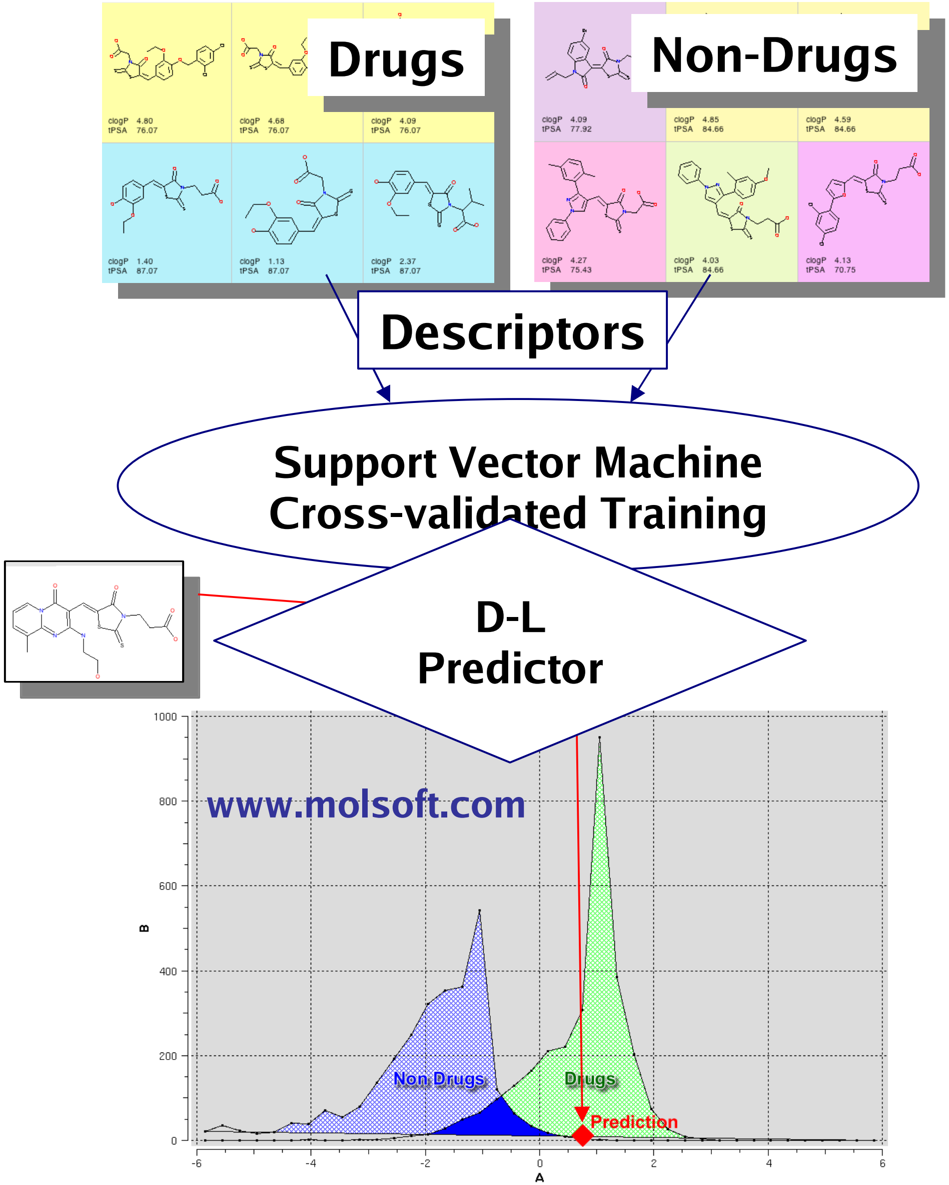
log S The aquous solubility of a compound significantly affects its absorption and distribution characteristics. Typically, a low solubility goes along with a bad absorption and therefore the general aim is to avoid poorly soluble compounds.

PSA Polar Surface Area

hERG potassium ion channels govern the repolarization phase of human ventricular action potentials. Many drugs or their metabolites cause hERG block, which can lead to cardiac arrhythmias and sudden death.

Cytochrome P450 (CYP) Isoenzymes
CYP isoenzymes are responsible for oxidative metabolism of many drugs, steroids and carcinogens. CYP isoenzymes are a group of heme-containing enzymes embedded primarily in the lipid bi-layer of the endoplastic reticulum of hepatocytes (liver cells).

Predicting Drug-Likeness



Drugs

Non-Drugs

Descriptors

Support Vector Machine
Cross-validated Training

D-L
Predictor

www.molsoft.com

Only 20 to 40% of the vendor database appear to be drug-like

Types of numerical problems

4) Two class (SVM)
5) Multi-class (SVM-multiclass)
6) Quantitative param., e.g. LogP (PLS, SVM-regression)

…

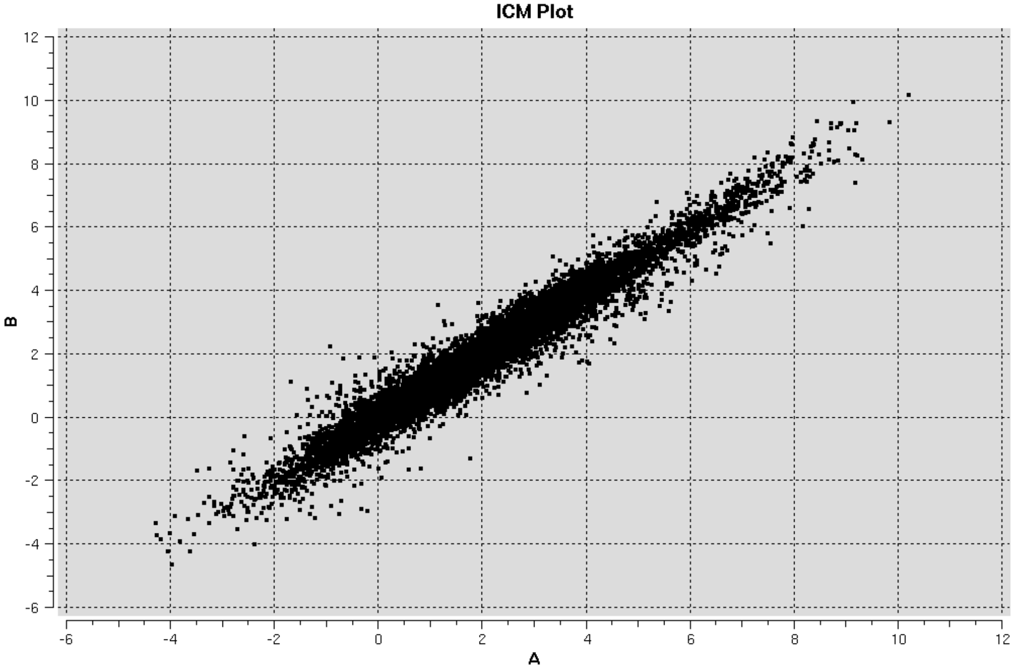
Concerns

- Insufficient data / Over-training
- Choice of descriptors
- Normalization of the descriptors
- Choice of the Non-Active training set

Drug-likeness prediction

- World Drug Index compounds were filtered and divided into 2 groups
- SVM trained on group 1
- Tested on group 2. 83% of the test group assigned correctly

logP -Predictions



We have trained Partial Least Squares (PLS) Regression model on the set of 13151 compounds with expiremental logP values.

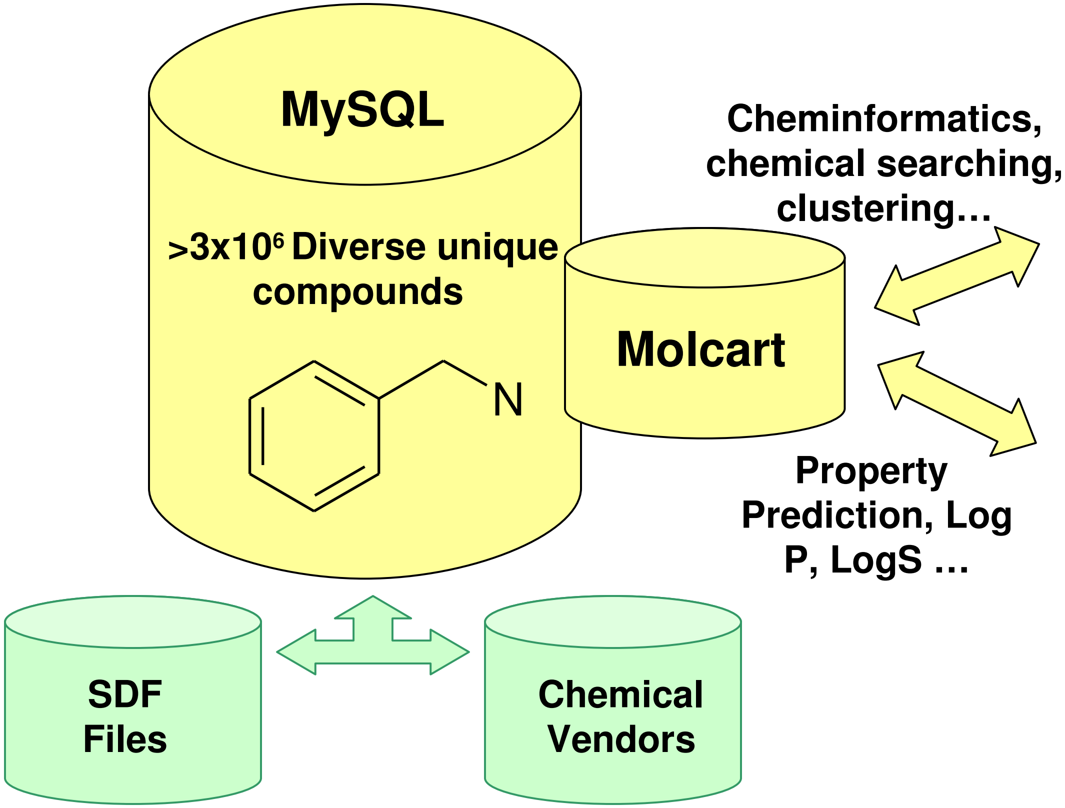
The correlation coefficient (r2) for fitting this training database is 0.98, and the standard deviation (rmsd) is 0.38.

The cross-validation test on randomly taken 50% of compounds as a train set and other 50% as a test set gives r2=0.94 and rmsd=0.61

In addition to good prediction quality this method is extremely fast and can be applied to large datasets.

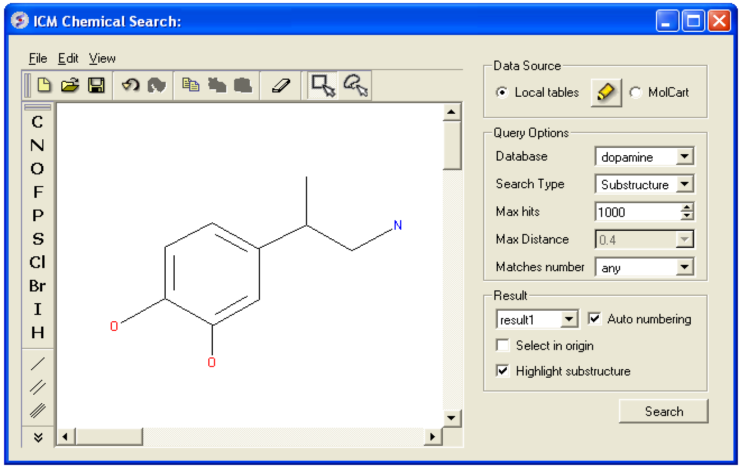
Molcart ®

Molcart ® is a state of the art enterprise wide chemical database management system.



Cheminformatics, chemical searching, clustering…

Property Prediction, Log P, LogS …



Data Source

● Local tables ○ MolCart

Query Options

Database: dopamine

Search Type: Substructure

Max hits: 1000

Max Distance: 0.4

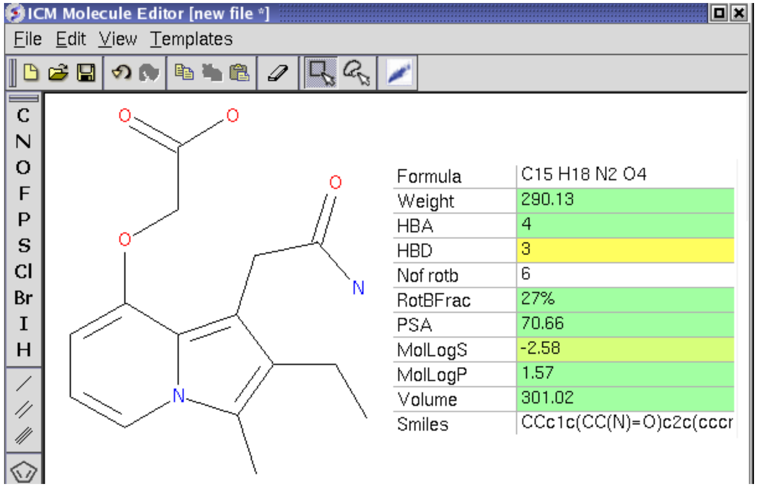
Matches number: any

Result

result1 ☑ Auto numbering

□ Select in origin

☑ Highlight substructure



Formula	C15 H18 N2 O4
Weight	290.13
HBA	4
HBD	3
Nof rotb	6
RotBFrac	27%
PSA	70.66
MolLogS	-2.58
MolLogP	1.57
Volume	301.02
Smiles	CCc1c(CC(N)=O)c2c(cccr

Compound databases of any size can be stored in Molcart and analyzed and searched using ICM cheminformatics and docking tools.