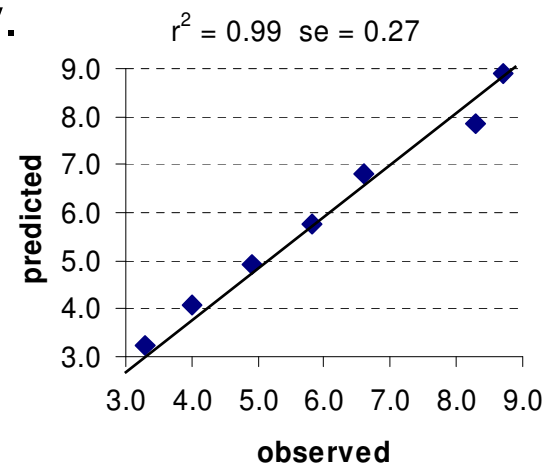


# QSAR, QSPR, statistics, correlation, similarity & descriptors

The tools of trade for the computer based *rational drug design*, particularly if there is no structural information about the *target* (protein) available.

QSAR equations form a quantitative connection between chemical structure and (biological) activity.

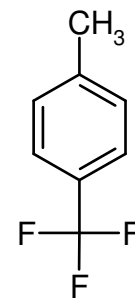
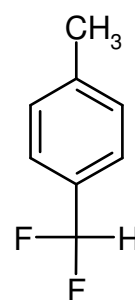
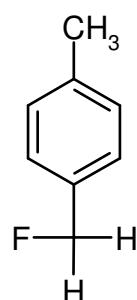
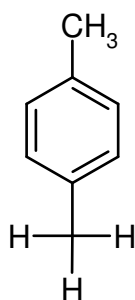
$$\log(1/C) = k_1 \cdot P_1 + k_2 \cdot P_2 + \dots + k_n \cdot P_n$$



The presence of experimentally measured data for a number of known compounds is required, e.g. from high throughput screening.

# Introduction to QSAR (I)

Suppose we have experimentally determined the binding constants for the following compounds



$K_i$  [ $10^{-9}$  mol  $l^{-1}$ ]

1550

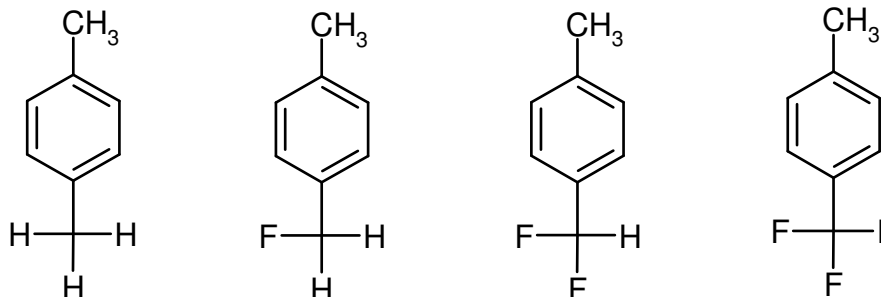
250

5.0

2.0

Which feature/property is responsible for binding ?

# Introduction to QSAR (II)



$K_i$  [ $10^{-9}$  mol l $^{-1}$ ]

1550

250

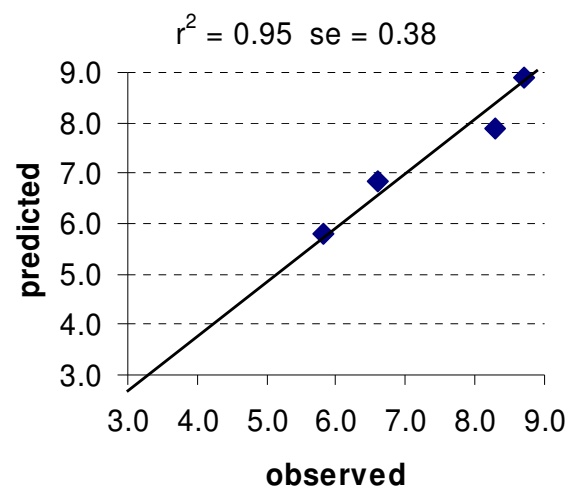
5.0

2.0

Using the number of fluorine atoms as descriptor we obtain following regression equation:

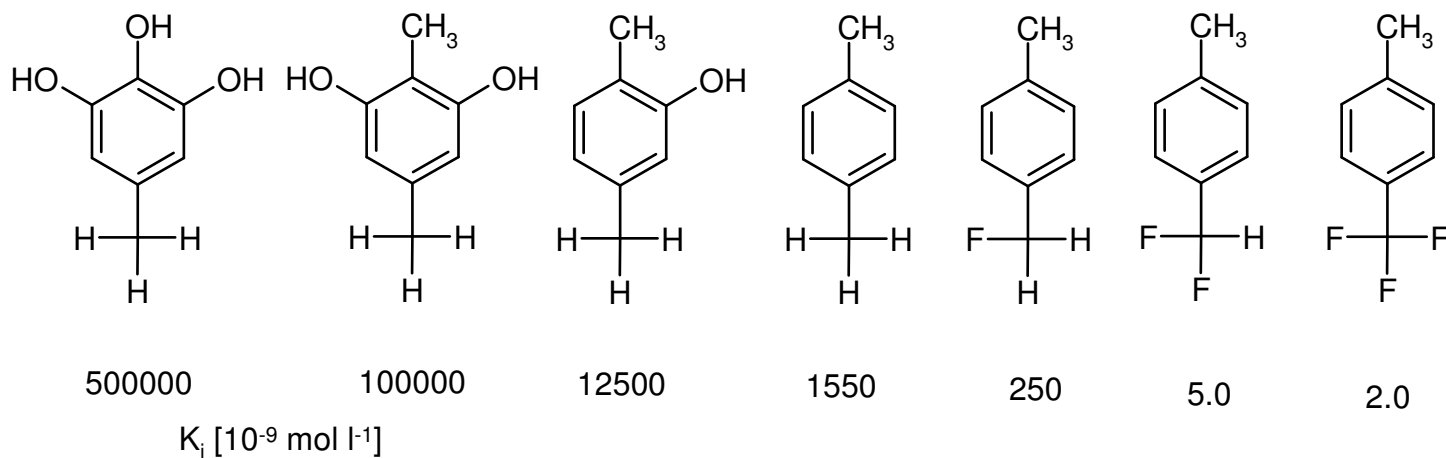
$$\log(1/K_i) = a \cdot n_{\text{fluorine}} + b$$

$$\log(1/K_i) = 1.037 \cdot n_{\text{fluorine}} + 5.797$$



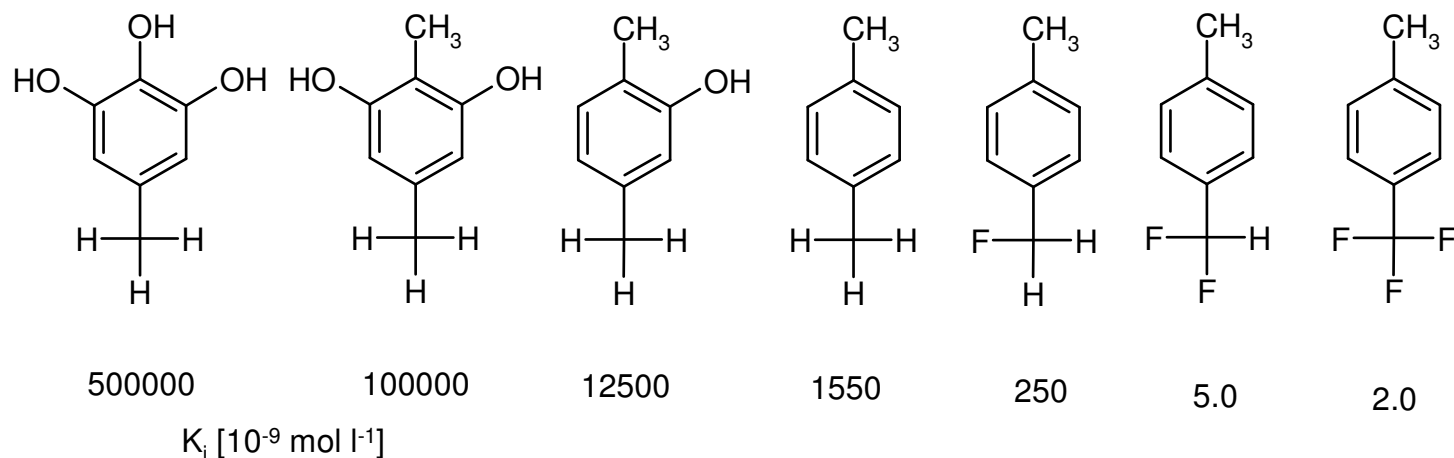
# Introduction to QSAR (III)

Now we add some other compounds



Which features/properties are now responsible for binding ?

# Introduction to QSAR (IV)

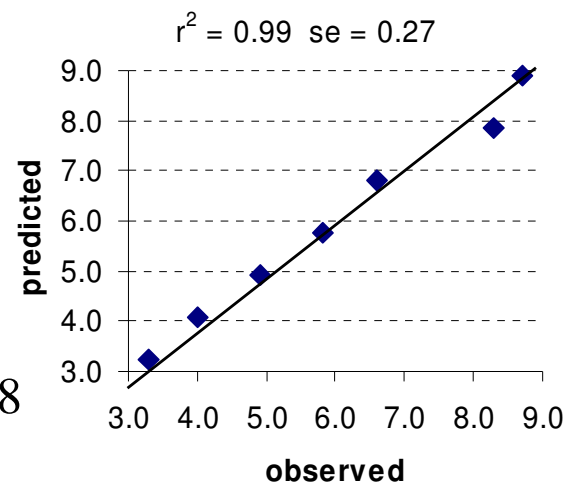


We assume that following descriptors play a major role:

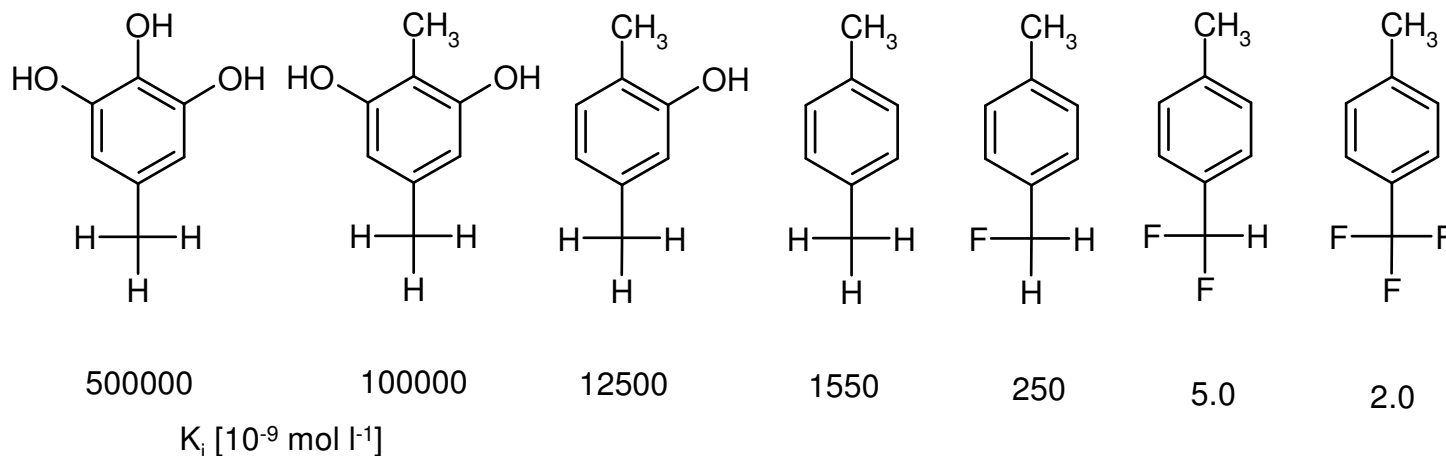
- number of fluorine atoms
- number of OH groups

$$\log(1/K_i) = a_1 \cdot n_{\text{fluorine}} + a_2 \cdot n_{\text{OH}} + b$$

$$\log(1/K_i) = 1.049 \cdot n_{\text{fluorine}} - 0.843 \cdot n_{\text{OH}} + 5.768$$



# Introduction to QSAR (V)



$$\log(1/K_i) = 1.049 \cdot n_{\text{fluorine}} - 0.843 \cdot n_{\text{OH}} + 5.768$$

$$r^2 = 0.99 \quad se = 0.27$$

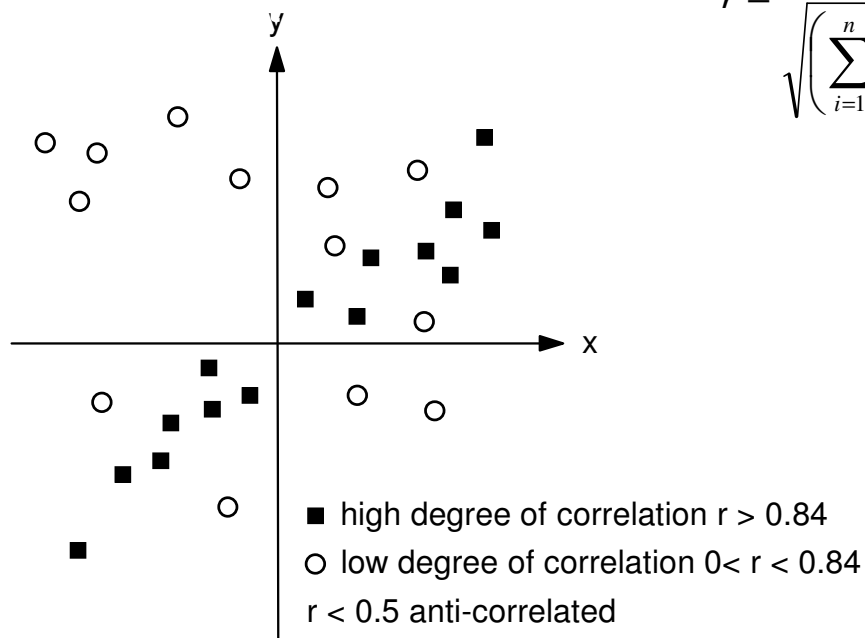
Is our prediction sound or just pure coincidence ?

→ We will need statistical proof (e.g. using a test set,  $\chi^2$ -test, p-values, cross-validation, boots trapping, ...)

# Correlation (I)

The most frequently used value is  
Pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)}} \in [-1...1]$$



→ A plot tells more than pure numbers !

# Definition of terms

QSAR: quantitative structure-activity relationship

QSPR: quantitative structure-property relationship

activity and property can be for example:

$\log(1/K_i)$  constant of binding

$\log(1/IC_{50})$  concentration that produces 50% effect

physical quantities, such as boiling point, solubility, ...

aim: prediction of molecular properties from their structure without the need to perform the experiment.

→ *in silico* instead of *in vitro* or *in vivo*

advantages: saves time and resources

# Development of QSAR methods over time (I)

1868 A.C.Brown, T.Fraser:  
Physiological activity is a function of the chemical constitution (composition)

but: An absolute direct relationship is not possible,  
only by using differences in activity.

Remember:

1865 Suggestion for the structure of benzene by  
A. Kekulé. The chemical structure of most organic  
compounds at that time was still unknown !

1893 H.H.Meyer, C.E.Overton  
The toxicity of organic compounds is related to their  
partition between aqueous and lipophilic biological  
phase.

# Development of QSAR method over time (II)

1868 E.Fischer

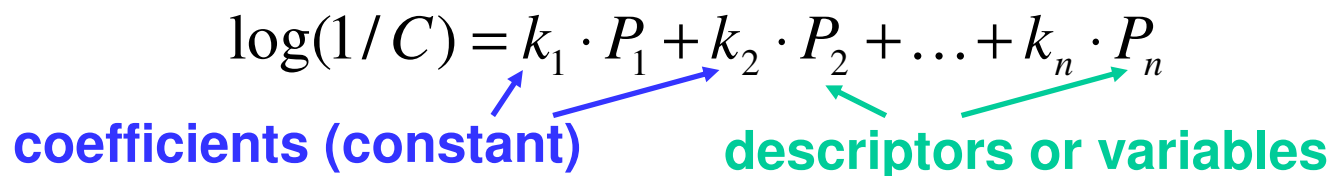
Key and lock principle for enzymes. Again no structural information about enzymes was available!

1930-40 Hammet equation: reactivity of compounds  
physical, organic, theoretic chemistry

**1964** C.Hansch, J.W.Wilson, S.M.Free, F.Fujita  
birth of modern QSAR-methods  
**Hansch analysis** and **Free-Wilson analysis**

$$\log(1/C) = k_1 \cdot P_1 + k_2 \cdot P_2 + \dots + k_n \cdot P_n$$

**coefficients (constant)**      **descriptors or variables**



linear free energy-related approach

# Descriptors

Approaches that form a mathematical relationship between numerical quantities (descriptors  $P_i$ ) and the physico-chemical properties of a compound (e.g. biological activity  $\log(1/C)$ ), are called QSAR or QSPR, respectively.

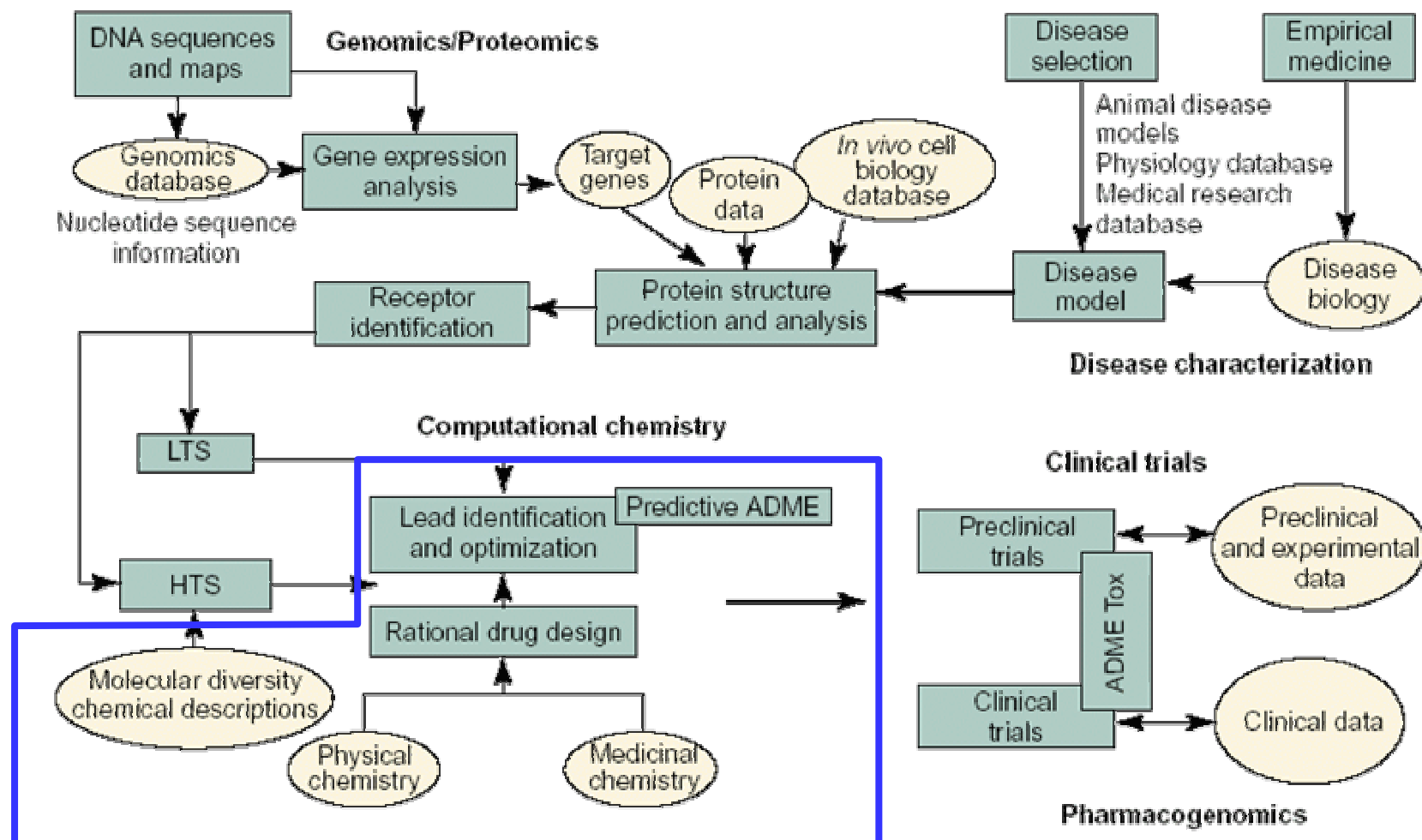
$$\log(1/C) = k_1 \cdot P_1 + k_2 \cdot P_2 + \dots + k_n \cdot P_n$$

Furthermore, descriptors are used to quantify molecules in the context of diversity analysis and in combinatorial libraries.

In principle any molecular or numerical property can be used as descriptors

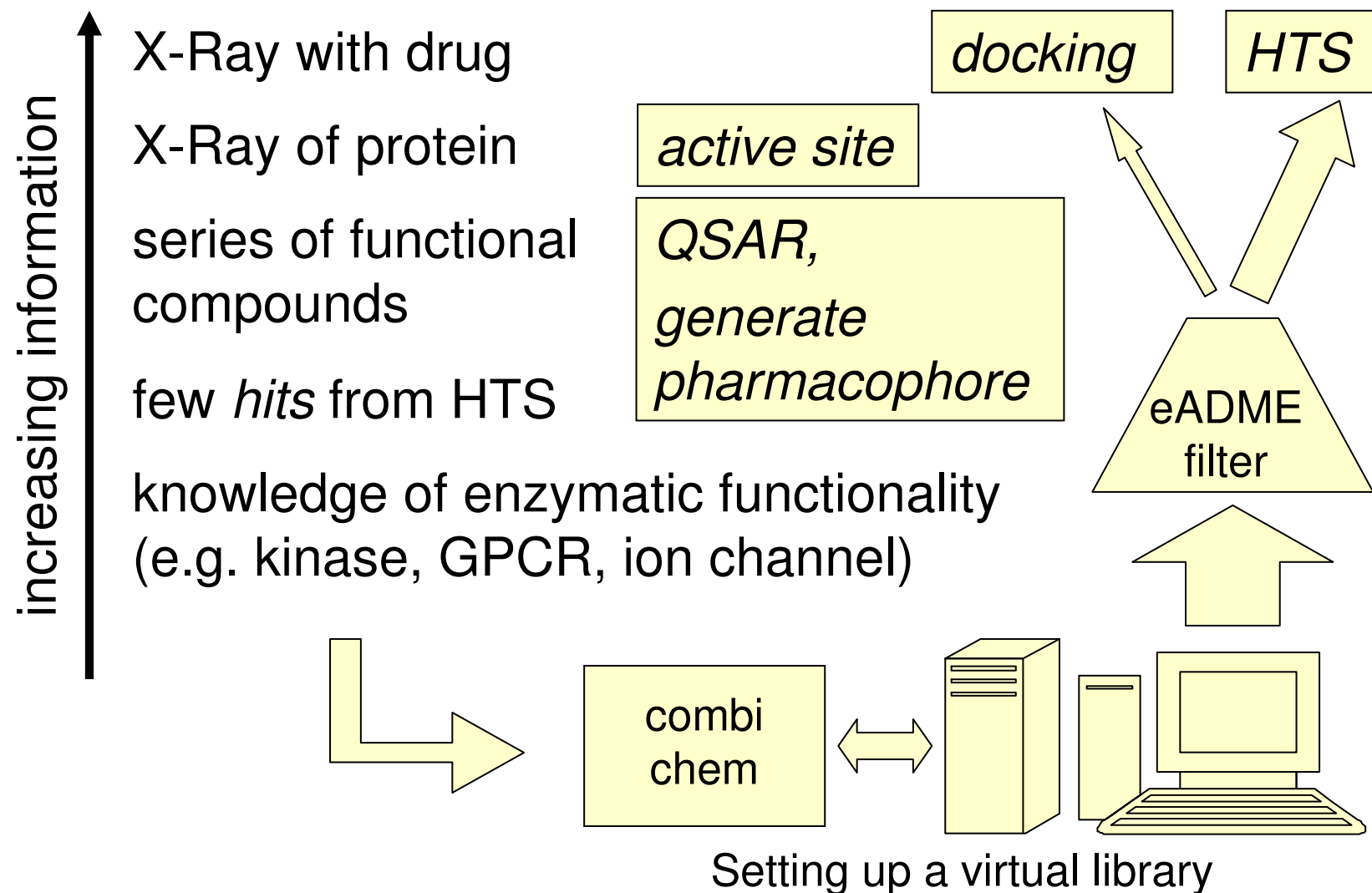
More about descriptors and their classification see  
<http://www.codessa-pro.com/descriptors/index.htm>

# Flow of information in a *drug discovery pipeline*



*Drug Discovery Today*

# Compound selection



# (Some) descriptors based on molecular properties used to predict ADME properties

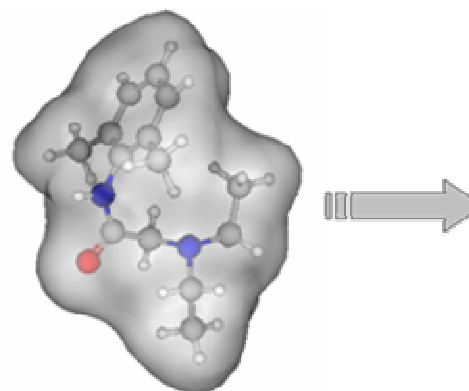
logP water/octanol partitioning coefficient

Lipinski's rule of five

topological indices

polar surface area

similarity / dissimilarity



Index	Descriptor	Value	Descriptor	Value	Descriptor	Value	Descriptor	Value	Descriptor	Value
1	MolWt	309.42	MolWt	309.42	MolWt	309.42	MolWt	309.42	MolWt	309.42
2	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
3	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
4	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
5	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
6	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
7	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
8	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
9	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
10	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
11	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
12	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
13	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
14	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
15	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
16	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
17	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
18	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
19	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
20	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
21	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
22	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
23	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
24	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00
25	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00	TPSA	40.00

QSAR quantitative structure activity relationship

QSPR quantitative structure property rel.

# „1D“ descriptors (I)

For some descriptors we need only the information that can be obtained from sum formula of the compound. Examples:

molecular weight, total charge, number of halogen atoms, ...

Further 1-dimensional descriptors are obtained by the summation of atomic contributions. Examples:

sum of the atomic polarizabilities

refractivity (*molar refractivity*,  $M_R$ )

$$M_R = (n^2 - 1) MW / (n^2 + 2) d$$

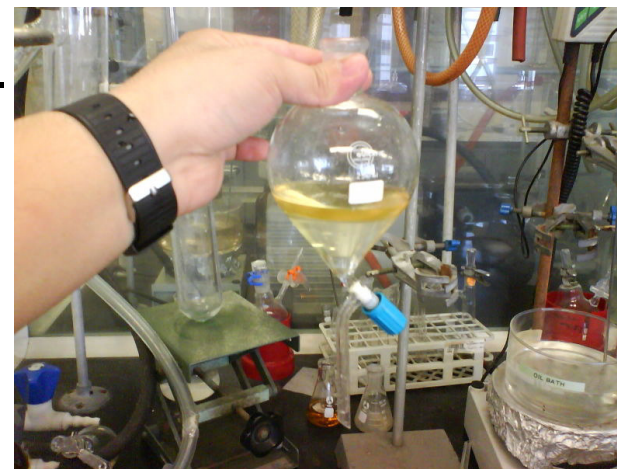
with refractive index  $n$ , density  $d$ , molecular weight  $MW$

Depends on the polarizability and moreover contains information about the molecular volume ( $MW / d$ )

# logP (I)

The *n*-octanol / water partition coefficient, respectively its logarithmic value is called logP.

Frequently used to estimate the membrane permeability and the bioavailability of compounds, since an orally administered drug must be enough **lipophilic** to cross the lipid bilayer of the membranes, and on the other hand, must be sufficiently water soluble to be transported in the blood and the lymph.



hydrophilic  $-4.0 < \log P < +8.0$  lipophilic

citric acid  $-1.72$

iodobenzene  $+3.25$

„typical“ drugs  $< 5.0$

# logP (II)

An increasing number of methods to predict logP have been developed:

Based on molecular fragments (atoms, groups, and larger fragments)

**ClogP** Leo, Hansch et al. *J.Med.Chem.* **18** (1975) 865.  
problem: non-parameterized fragments  
(up to 25% of all compounds in substance libraries)

based on atom types (similar to force field atom types)

**SlogP** S.A. Wildman & G.M.Crippen *J.Chem.Inf.Comput.Sci.*  
**39** (1999) 868.

**AlogP, MlogP, XlogP...**

Parameters for each method were obtained using a mathematical fitting procedure (linear regression, neural net,...)

Review: R.Mannhold & H.van de Waaterbeemd,  
*J.Comput.-Aided Mol.Des.* **15** (2001) 337-354.

## logP (III)

Recent logP prediction methods more and more apply whole molecule properties, such as

- molecular surface (polar/non-polar area, or their electrostatic properties = electrostatic potential)
- dipole moment and molecular polarizability
- ratio of volume / surface (globularity)

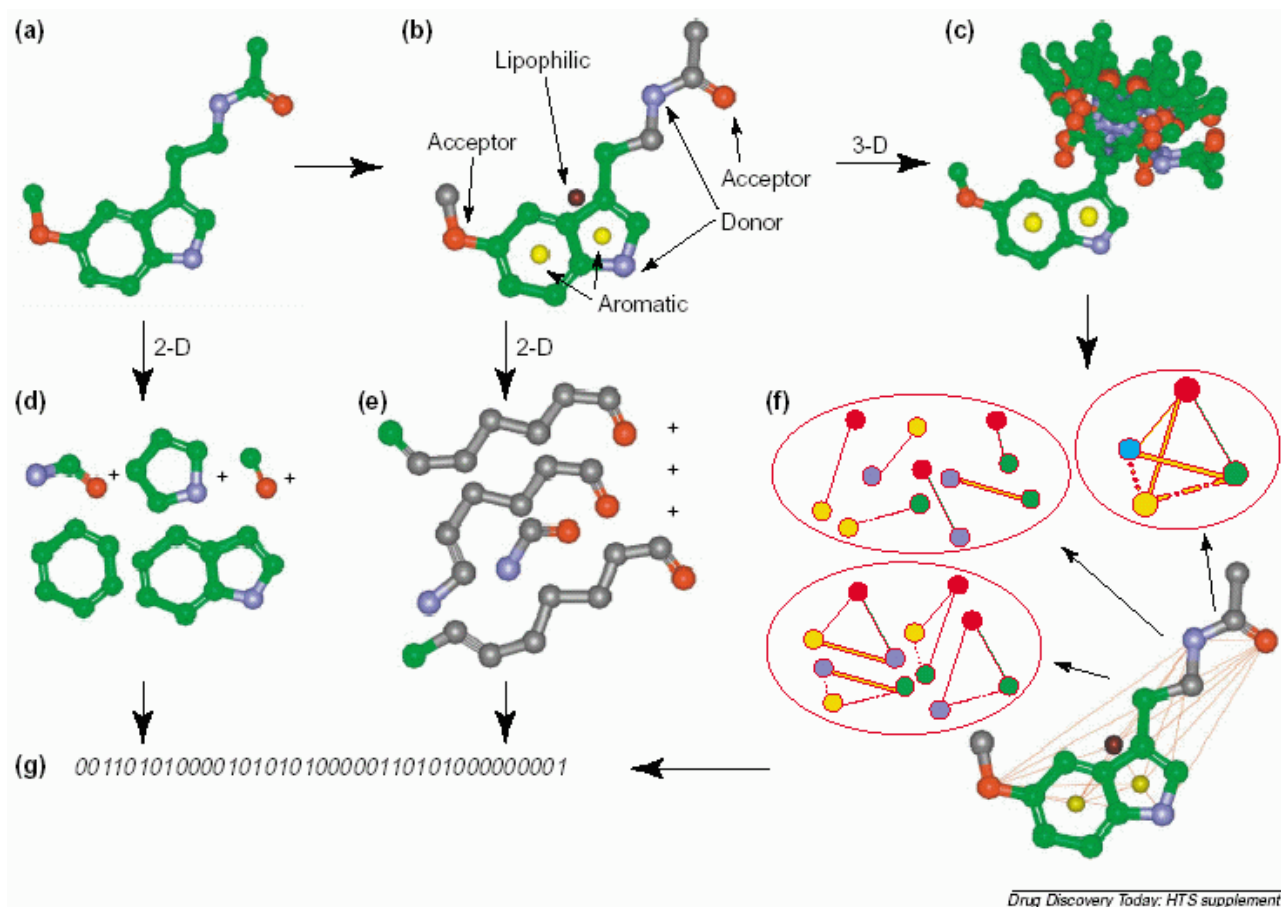
Example: Neural net trained with quantum chemical data  
logP T. Clark et al. *J.Mol.Model.* **3** (1997) 142.

## „1D“ descriptors (II)

Further atomic descriptors use information based on empirical atom types like in force fields. Examples:

- Number of halogen atoms
- Number of  $sp^3$  hybridized carbon atoms
- Number of H-bond acceptors (N, O, S)
- Number of H-bond donors (OH, NH, SH)
- Number of aromatic rings
- Number of COOH groups
- Number of ionizable groups ( $NH_2$ , COOH)
- ...
- Number of freely rotatable bonds

# Fingerprints



**Figure 2.** Schematic illustration of primary methods used in molecular fingerprint creation. (a) Create 2-D and 3-D model of molecule; (b) deconstruct the molecule into pharmacophoric elements; (c) generate conformational models; (d) deconstruct the molecule into topological/substructural elements; (e) determine distance between pharmacophoric groups using bond counts; (f) determine 2-, 3- or 4-center distance combinations of pharmacophoric groups for each conformer; and (g) determine the presence or absence of each descriptor element and combine to create a binary fingerprint.

binary *fingerprint* of a molecule

# Lipinski's Rule of 5

Combination of descriptors to estimate intestinal absorption.  
Insufficient uptake of compounds, if

Molecular weight > 500

slow diffusion

$\log P > 5.0$

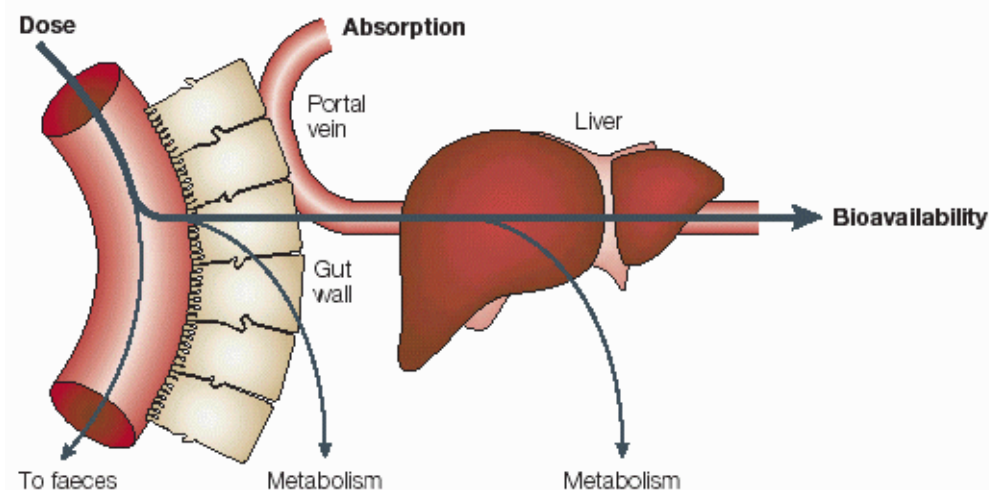
too lipophilic

> 5 H-bond donors (OH and NH)

too many H-bond with the head

>10 H-bond acceptors (N and O atoms)

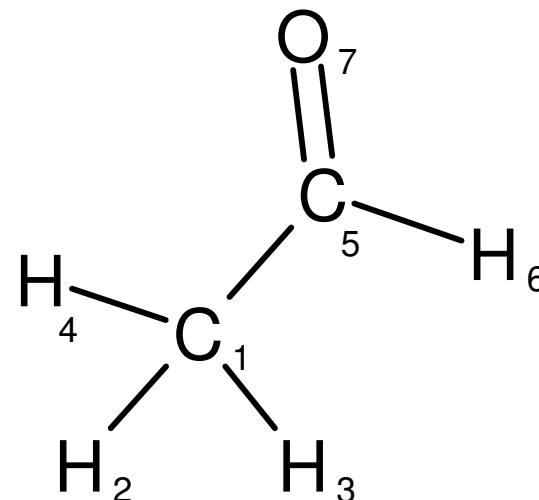
groups of the membrane



C.A. Lipinski et al. *Adv. Drug. Delivery Reviews* **23** (1997) 3.

# 2D descriptors (I)

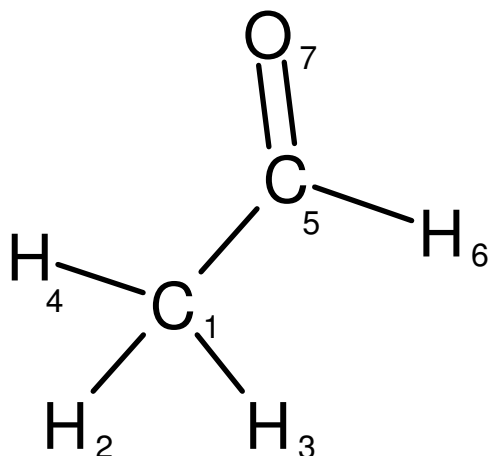
Descriptors derived from the configuration of the molecules (covalent bonding pattern) are denoted **2D descriptors**. Since no coordinates of atoms are used, they are in general **conformationally independent**, despite containing topological information about the molecule. C.f. representation by SMILES



	adjacency matrix M							distance matrix D						
C1	0	1	1	1	1	0	0	0	1	1	1	1	2	2
H2	1	0	0	0	0	0	0	1	0	2	2	2	3	3
H3	1	0	0	0	0	0	0	1	2	0	2	2	3	3
H4	1	0	0	0	0	0	0	1	2	2	0	2	3	3
C5	1	0	0	0	0	1	1	1	2	2	2	0	1	1
H6	0	0	0	0	1	0	0	2	3	3	3	1	0	2
O7	0	0	0	0	1	0	0	2	3	3	3	1	2	0

## 2D descriptors (II)

The essential topological properties of a molecule are the degree of branching and the molecular shape.



An  $sp^3$  hybridized carbon has got 4 valences, an  $sp^2$  carbon only 3.

Thus the ratio of the actual branching degree to the theoretically possible branching degree can be used as descriptor as it is related to the saturation.

## 2D descriptors (III)

Common definitions:

$Z_i$  ordinary number (H=1, C=6, N=7, LP=0)

$h_i$  number of H atoms bonded to atom  $i$

$d_i$  number of non-hydrogen atoms bonded to atom  $i$

Descriptors accounting for the degree of branching and the flexibility of a molecule:

### Kier & Hall Connectivity Indices

$p_i$  sum of s and p valence electrons of atom  $i$

$v_i = (p_i - h_i) / (Z_i - p_i - 1)$  for all non-hydrogen (heavy) atoms

# Kier and Hall Connectivity Indices

$Z_i$  ordinary number (H=1, C=6, LP=0)

$d_i$  number of heavy atoms bonded to atom  $i$

$p_i$  number of  $s$  and  $p$  valence electrons of atom  $i$

$v_i = (p_i - h_i) / (Z_i - p_i - 1)$  for all heavy atoms

Chi0    0th order     $\chi_0 = \sum_i \frac{1}{\sqrt{d_i}}$  for all heavy atom with  $d_i > 0$

Chi1    1st order     $\chi_1 = \sum_i \sum_{j>i} \frac{1}{\sqrt{d_i d_j}}$  for all heavy atoms if  
 $i$  is bonded to  $j$

Chi0v  
Valence index     $\chi_{0v} = \sum_i \frac{1}{\sqrt{v_i}}$  for all heavy atoms with  $v_i > 0$

# Kier and Hall Shape Indices (I)

$n$  number of heavy atoms (non-hydrogen atoms)

$m$  total number of bonds between all heavy atoms

$p_2$  number of paths of length 2

$p_3$  number of paths of length 3 from the distance matrix **D**

Kappa1  $\kappa_1 = \frac{n(n-1)^2}{m^2}$

Kappa2  $\kappa_2 = \frac{(n-1)(n-2)^2}{p_2^2}$

Kappa3  $\kappa_3 = \frac{(n-1)(n-3)^2}{p_3^2}$  for even  $n$   
 $\kappa_3 = \frac{(n-3)(n-2)^2}{p_3^2}$  for odd  $n$

# Kier and Hall Shape Indices (II)

Relating the atoms to  $sp^3$ -hybridized carbon atoms yields the Kappa alpha indices

$$\alpha = \sum_i^n \frac{r_i}{r_c - 1} \quad \begin{array}{l} r_i \text{ covalence radius of atom } i \\ r_c \text{ covalence radius of an } sp^3 \text{ carbon atom} \end{array}$$

$$\kappa_{\alpha 1} = \frac{s(s-1)^2}{(m+\alpha)^2} \text{ with } s = n + \alpha$$

element	hybridization	$\alpha$
C	$sp^3$	0
C	$sp^2$	-0.13
C	$sp$	-0.22
N	$sp^3$	-0.04
N	$sp^2$	-0.20
N	$sp$	-0.29
O	$sp^3$	-0.04
P	$sp^3$	+0.43
S	$sp^3$	+0.35
Cl		+0.29

# Balaban, Wiener, and Zagreb Indices

$n$  number of heavy atoms (non-hydrogen atoms)

$m$  total number of bonds between all heavy atoms

$d_i$  number of heavy atoms bonded to atom  $i$

$w_i = \sum_{i \neq j} D_{ij}$  Sum of the off-diagonal matrix elements of atom  $i$  in the distance matrix **D**

BalabanJ 
$$\frac{m}{m - n + 1} \sum \frac{1}{\sqrt{w_i w_j}}$$

WienerJ (pfad number)  $\frac{1}{2} \sum_i^n w_i$  Correlates with the boiling points of alkanes

Wiener polarity  $\frac{1}{2} \sum_i^n w_i$  if  $D_{ij} \geq 3$

Zagreb index  $\sum_i d_i^2$  for all heavy atoms  $i$

# What message do topological indices contain?

topological indices are associated with the

- degree of branching in the molecule
- size and spacial extension of the molecule
- structural flexibility

Usually it is not possible to correlate a chemical property with only one index directly

Although topological indices encode the same properties as fingerprints do, they are harder to interpret, but can be generated numerically more easily.

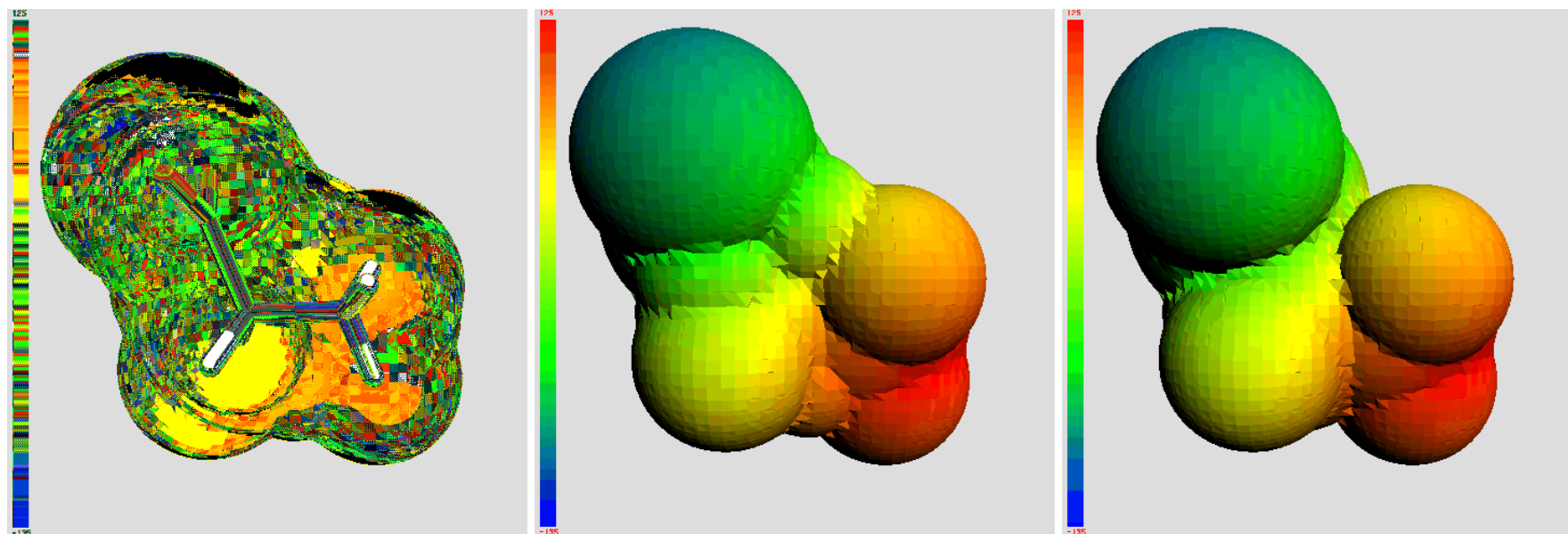
# 3D descriptors

Descriptors using the atomic coordinates (x,y,z) of a molecule are therefore called **3D descriptors**.

As a consequence they usually **depend on the conformation**.

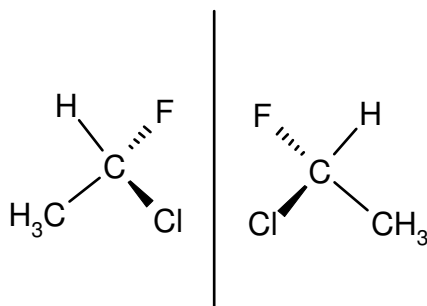
Examples:

van der Waals volume, molecular surface, polar surface, electrostatic potential (ESP), dipole moment



# Chirality Descriptors

Most biological interactions are stereospecific e.g. ligand binding



share identical 1D and 2D-descriptors

Ideas for including chirality:

- Using differences of the van der Waals volume or the electrostatic potential after superposition (rotation)
- Adding +1/-1 to chiral centers in the adjacency matrix while computing topological descriptors
- Modifying the sign of 1D-descriptors (electronegativity, size, polarizability,...) with respect to the enantiomer

Lit: G.M.Crippen *Curr.Comput.-Aided Drug Des.* **4** (2008) 259-264.

# Quantum mechanical descriptors (selection)

Atomic charges (*partial atomic charges*) No observables !

Mulliken population analysis

electrostatic potential (ESP) derived charges

dipole moment

polarizability

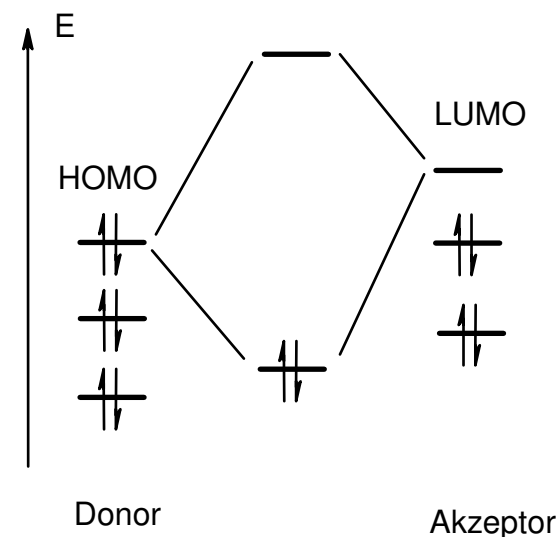
HOMO / LUMO

energies of the frontier orbitals  
given in eV

covalent hydrogen bond acidity/basicity

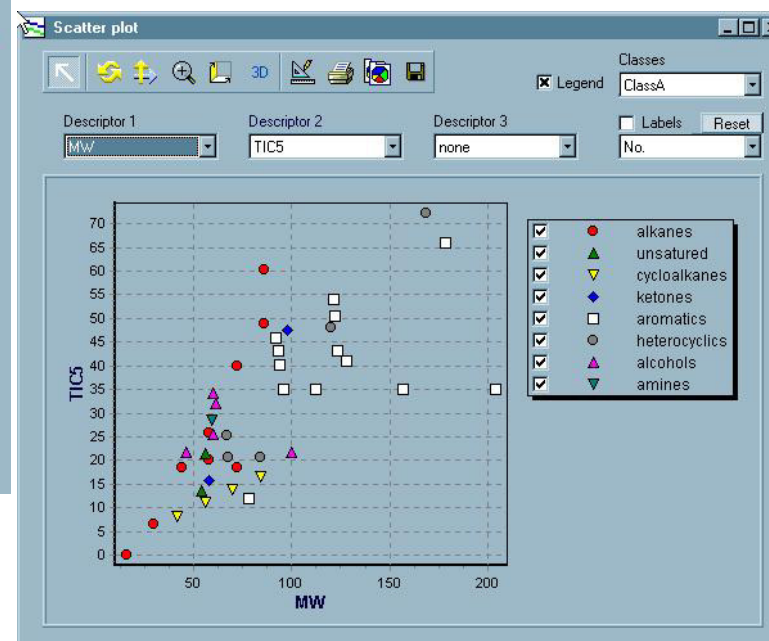
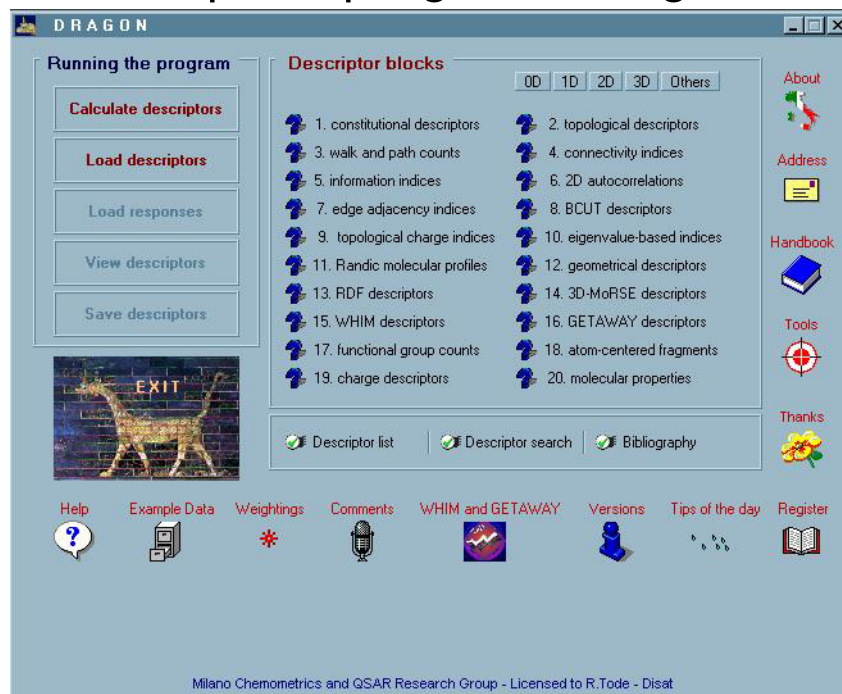
difference of the HOMO/LUMO energies compared  
to those of water

Lit: M. Karelson et al. *Chem.Rev.* **96** (1996) 1027



# (e)DRAGON

a computer program that generates >1400 descriptors

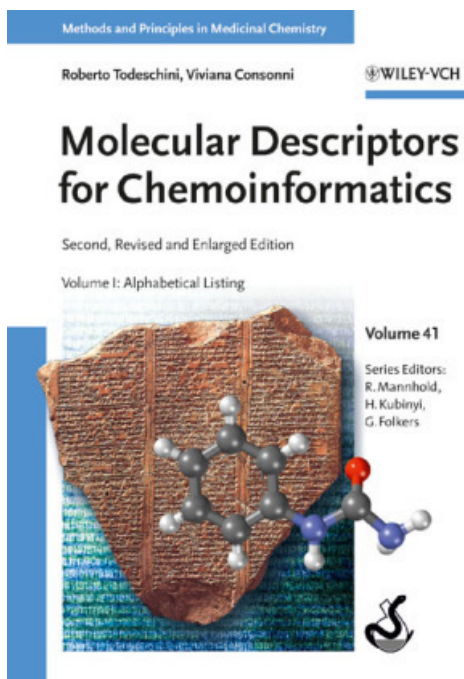


Roberto Todeschini

<http://www.vcclab.org/lab/edragon/>

Requires 3D-structure of molecules as input

# Further information about descriptors

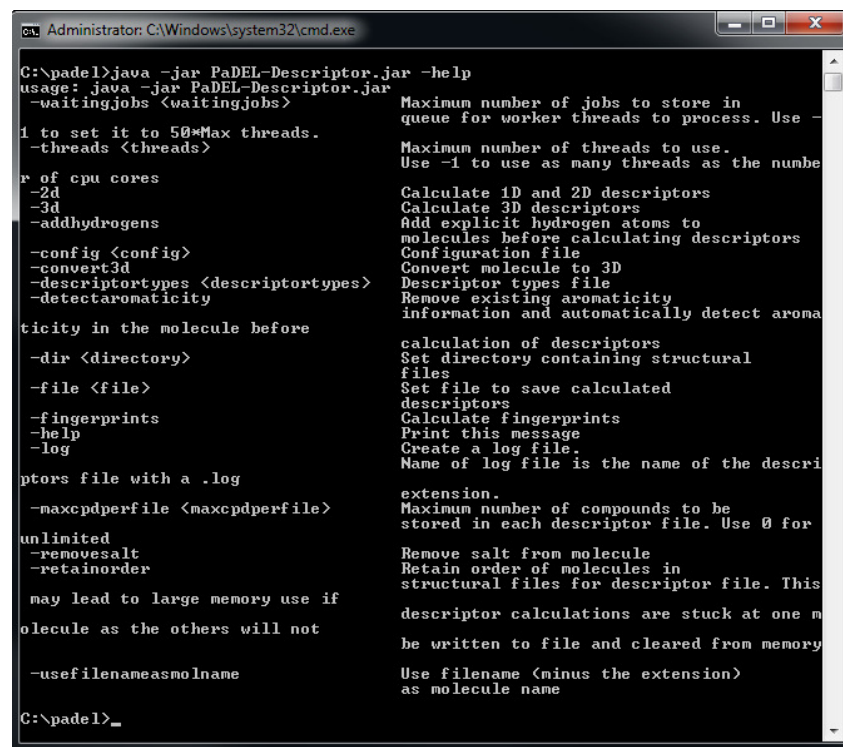
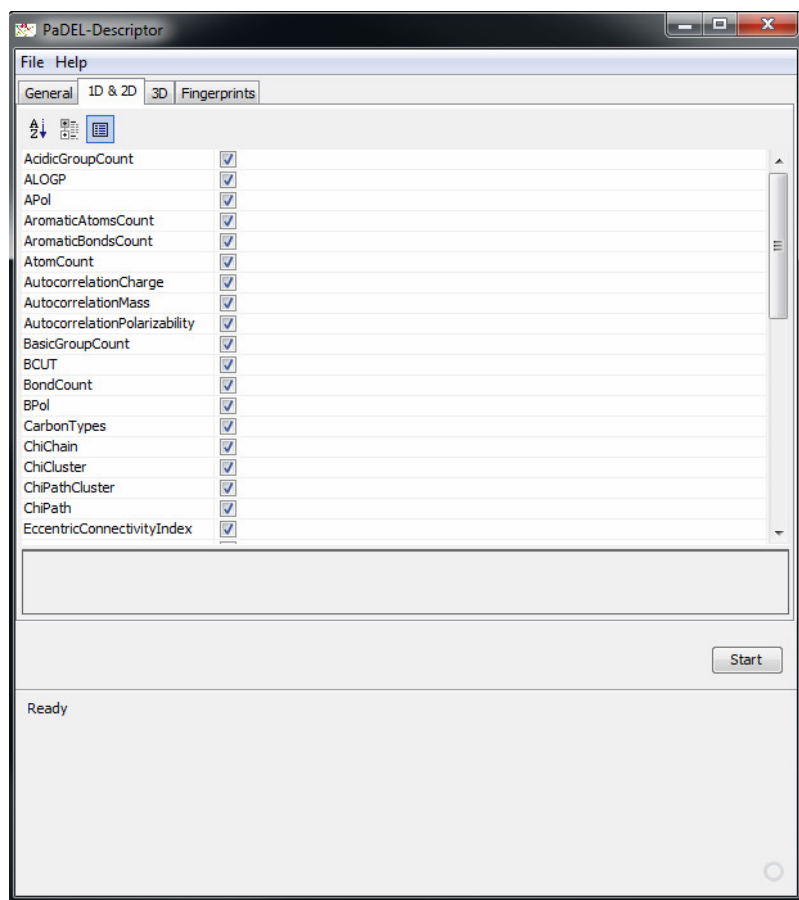


Roberto Todeschini, Viviana Consonni

*Handbook of Molecular Descriptors*,  
Wiley-VCH, 2nd ed. (2009)  
1257 pages

CODESSA Alan R. Katritzky, Mati Karelson et al.  
<http://www.codessa-pro.com>

# PaDEL-Descriptor



Open Source Software (JAVA)

Chun Wei Yap

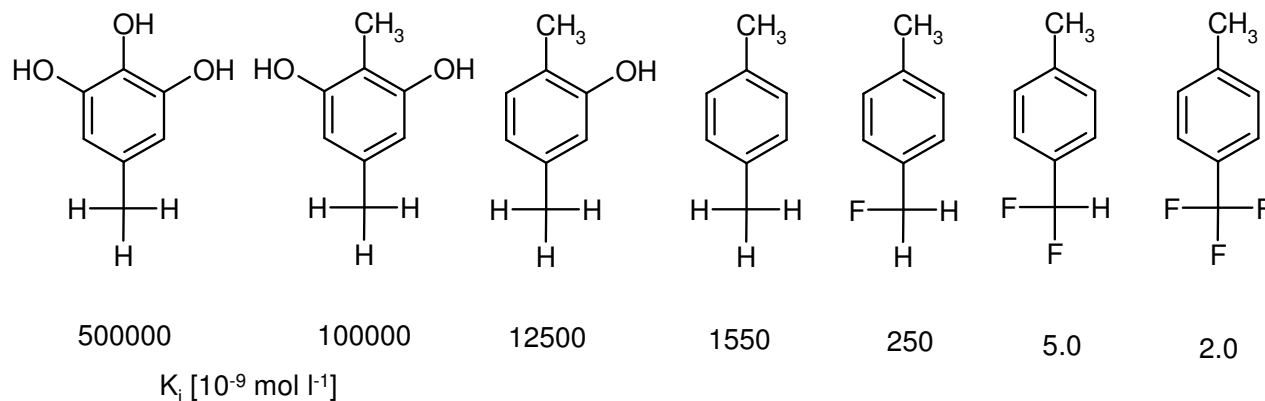
<http://www.yapcwsoft.com/dd/padeldescriptor/>

C.W. Yap *J.Comput.Chem.* **32** (2011) 1466-1474.

# Chosing the right compounds (I)

To derive meaningful QSAR predictions we need

- A sufficient number of compounds **statistically sound**
- Structurally diverse compounds **tradeoff between count and similarity**



How similar are compounds to each other ?

→ Clustering using distance criteria  
that are based on the descriptors

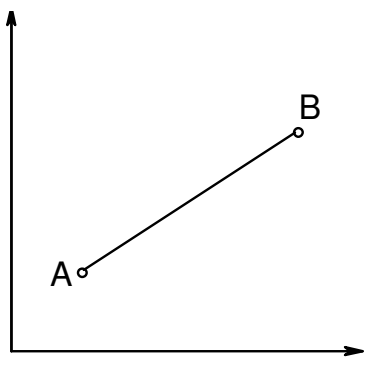
# Distance criteria and similarity indices (I)

$\chi_A$  fulfilled property of molecule  $A$

$|\chi_A \cap \chi_B|$  intersection of common properties of  $A$  and  $B$

$|\chi_A \cup \chi_B|$  unification of common properties of  $A$  and  $B$

## Euklidian distance



formula

$$D_{A,B} = \sqrt{\sum_{i=1}^N (x_{iA} - x_{iB})^2}$$

definition

$$D_{A,B} = \sqrt{|\chi_A \cup \chi_B| - |\chi_A \cap \chi_B|}$$

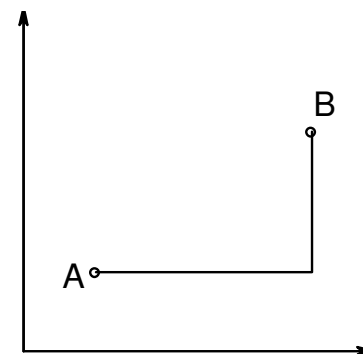
range

$\infty$  to 0

other names

—

## Manhattan distance



$$D_{A,B} = \sum_{i=1}^N |x_{iA} - x_{iB}|$$

$$D_{A,B} = |\chi_A \cup \chi_B| - |\chi_A \cap \chi_B|$$

$\infty$  to 0

City-Block, Hamming

# Distance criteria and similarity indices (II)

## Soergel distance

$$D_{A,B} = \sum_{i=1}^N |x_{iA} - x_{iB}| / \sum_{i=1}^N \max(x_{iA}, x_{iB})$$

$$D_{A,B} = |\chi_A \cup \chi_B| - |\chi_A \cap \chi_B| / |\chi_A \cup \chi_B|$$

1 to 0

—

## Tanimoto index

$$S_{A,B} = \left( \sum_{i=1}^N x_{iA} x_{iB} \right) / \left( \sum_{i=1}^N (x_{iA})^2 + \sum_{i=1}^N (x_{iB})^2 - \sum_{i=1}^N x_{iA} x_{iB} \right)$$

$$S_{A,B} = |\chi_A \cap \chi_B| / |\chi_A \cup \chi_B|$$

−0.333 to +1 (continuous values)

0 to +1 (binary on/off values)

Jaccard coefficient

For binary (dichotomous) values the Soergel distance is complementary to the Tanimoto index

# Distance criteria and similarity indices (II)

## Dice coefficient

$$S_{A,B} = \left( 2 \sum_{i=1}^N x_{iA} x_{iB} \right) / \left( \sum_{i=1}^N (x_{iA})^2 + \sum_{i=1}^N (x_{iB})^2 \right)$$

$$S_{A,B} = 2|\chi_A \cap \chi_B| / (|\chi_A| + |\chi_B|)$$

−1 to +1

0 to +1

Hodgkin index

Czekanowski coefficient

Sørensen coefficient

monotonic with the  
Tanimoto index

## Cosinus coefficient

$$S_{A,B} = \left( \sum_{i=1}^N x_{iA} x_{iB} \right) / \sqrt{\sum_{i=1}^N (x_{iA})^2 \cdot \sum_{i=1}^N (x_{iB})^2}$$

$$S_{A,B} = |\chi_A \cap \chi_B| / \sqrt{|\chi_A| |\chi_B|}$$

0 to +1 (continuous values)

0 to +1 (binary on/off values)

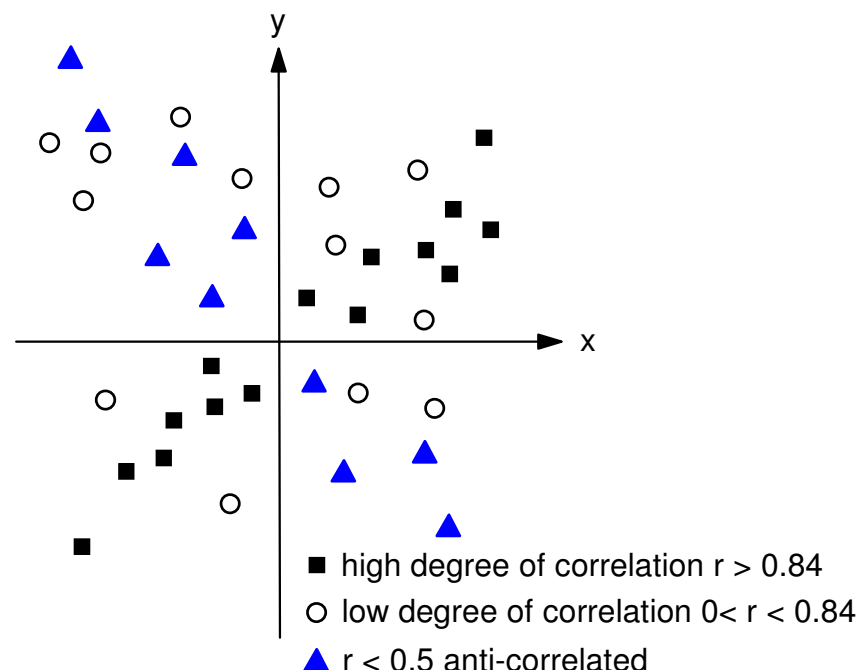
Carbo index

Ochiai coefficient

Highly correlated to the  
Tanimoto index

# Correlation between descriptors (I)

Descriptors can also be inter-correlated (colinear) to each other  
→ redundant information should be excluded



Usually we will have a wealth of descriptors (much more than the available molecules) to choose from. To obtain a reasonable combination in our QSAR equation, multivariate methods of statistics or other selection procedures must be applied.

## Correlation between descriptors (II)

How many descriptors can be used in a QSAR equation ?

Rule of thumb:

per descriptor used, at least 5 molecules (data points)  
should be present

otherwise the possibility of finding a coincidental  
correlation is too high.

(Ockham's razor: fit anything to anything)

Therefore:

Principle of parsimony

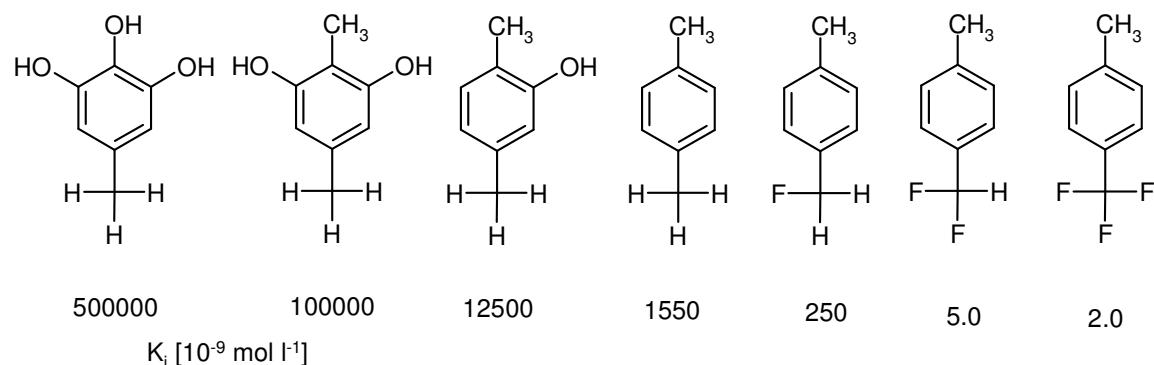
# Deriving QSAR equations (I)

After removing the inter-correlated descriptors, we have to determine the coefficients  $k_i$  for those descriptors that appear in the QSAR equation.

Such multiple linear regression analysis (*least square fit* of the according coefficients) is performed by statistics programs

There are several ways to proceed:

1. Using the descriptor that shows the best correlation to the predicted property first and adding stepwise descriptors that yield the best improvement (**forward regression**)



$$\log(1/K_i) = 1.049 \cdot n_{\text{fluorine}} - 0.843 \cdot n_{\text{OH}} + 5.768$$

# Deriving QSAR equations (II)

2. Using all available descriptors first, and removing stepwise those descriptors that worsen the correlation fewest  
(**backward regression/elimination**)

3. Determining the best combination of the available descriptors for given number of descriptors appearing in the QSAR equation (2,3,4,...) (**best combination regression**)

This is usually not possible due to the exponential runtime

Problem of forward and backward regression:

Risk of local minima

Problem: Which descriptors are relevant or significant?

Determination of such descriptors, see lecture 6