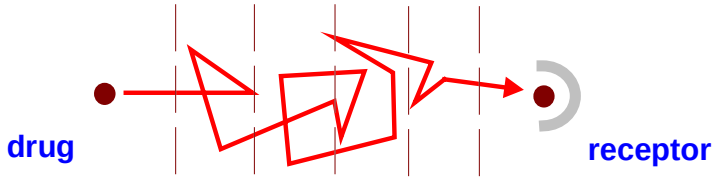
Nonlinear QSAR and 3D QSAR

Hugo Kubinyi

Germany

E-Mail kubinyi@t-online.de
HomePage www.kubinyi.de

Nonlinear Lipophilicity-Activity Relationships



Possible Reasons for Nonlinear Lipophilicity-Activity Relationships

- kinetic control of drug transport in biological systems
- equilibrium control of drug distribution
- steric hindrance, allosteric effects
- different pharmacokinetics and/or metabolism
- solubility, micelle formation
- end product inhibition of enzymes
- drug receptor occupation

Parabolic Model (C. Hansch, 1964)

$\log 1/C = a (\log P)^2 + b \log P + c$
$\log 1/C = a \pi^2 + b \pi + c$
$\log 1/C = a (\log P)^2 + b \log P + c \sigma + d$
$\log 1/C = a \pi^2 + b \pi + c \sigma + d E_s + c$, etc.

Probability Model (J. McFarland, 1970)

$\log 1/C = a \log P - 2a \log (P + 1) + c$

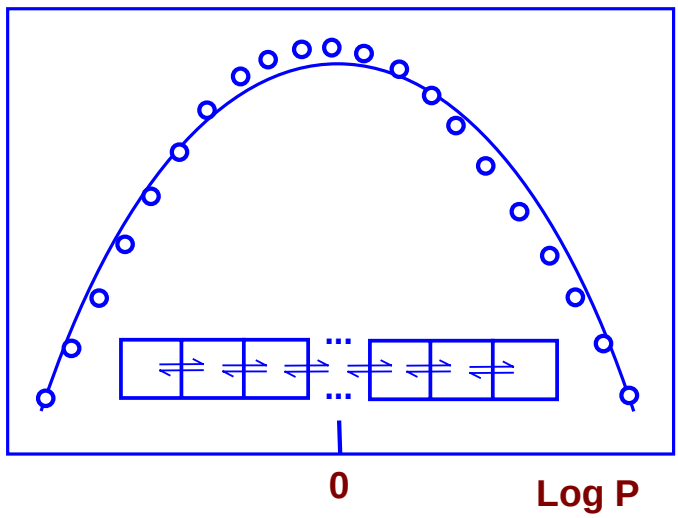
Equilibrium Model (R. Hyde, 1975)

$\log 1/C = a \log P - \log (aP + 1) + c$

Bilinear Model (H. Kubinyi, 1976)

$\log 1/C = a \log P - b \log (\beta P + 1) + c$

Hansch Multicompartment Model (1969)



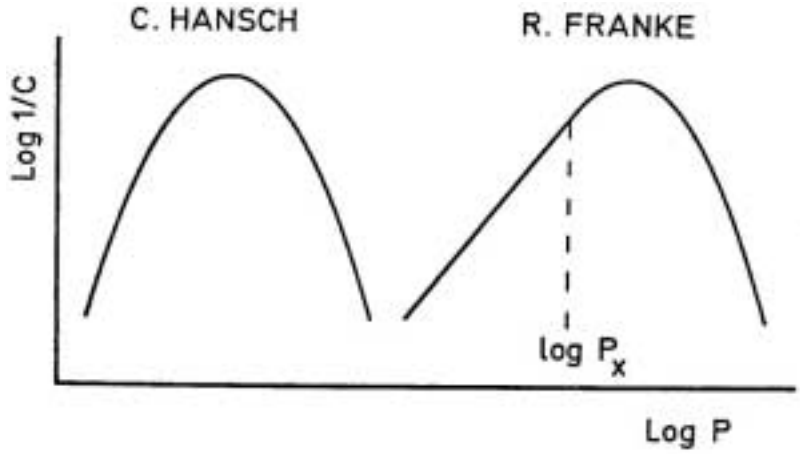
Definition

$P = k_1/k_2$
(correct)

Hypothesis

$k_1.k_2 = 1$
(wrong)

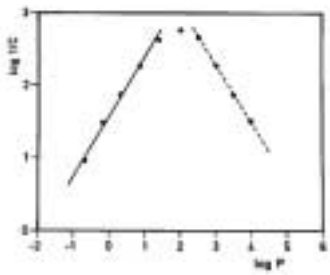
Hansch and Franke Models



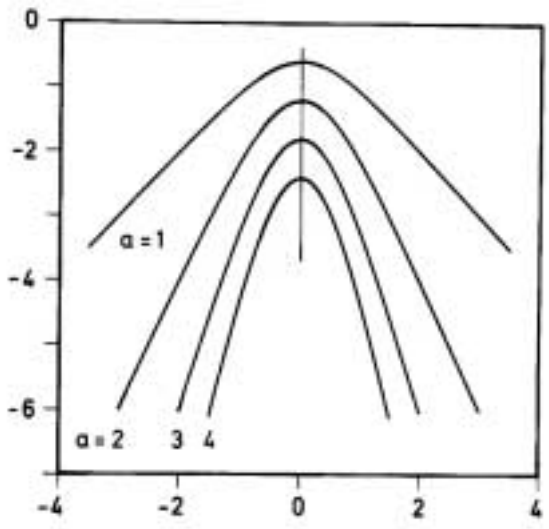
$\log 1/C = a \log P + c$ $(\log P < \log P_x)$

$\log 1/C = \alpha (\log P)^2 + ß \log P + \gamma$ $(\log P > \log P_x)$

McFarland Probability Model (1970)



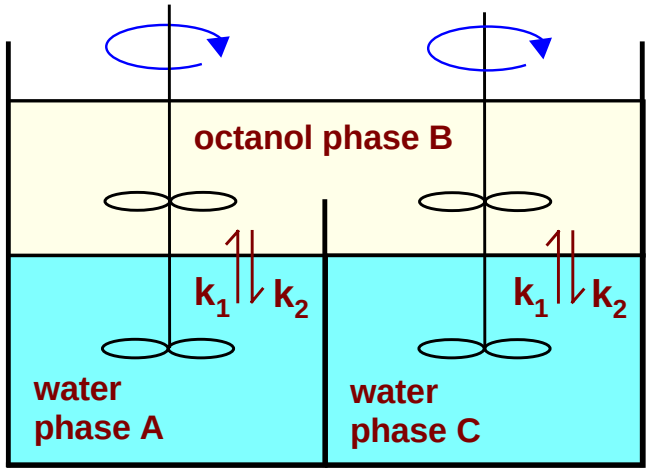
Neurotoxic activity of alcohols, rat, i.p. application



Substance Distribution in a Three-Compartment System

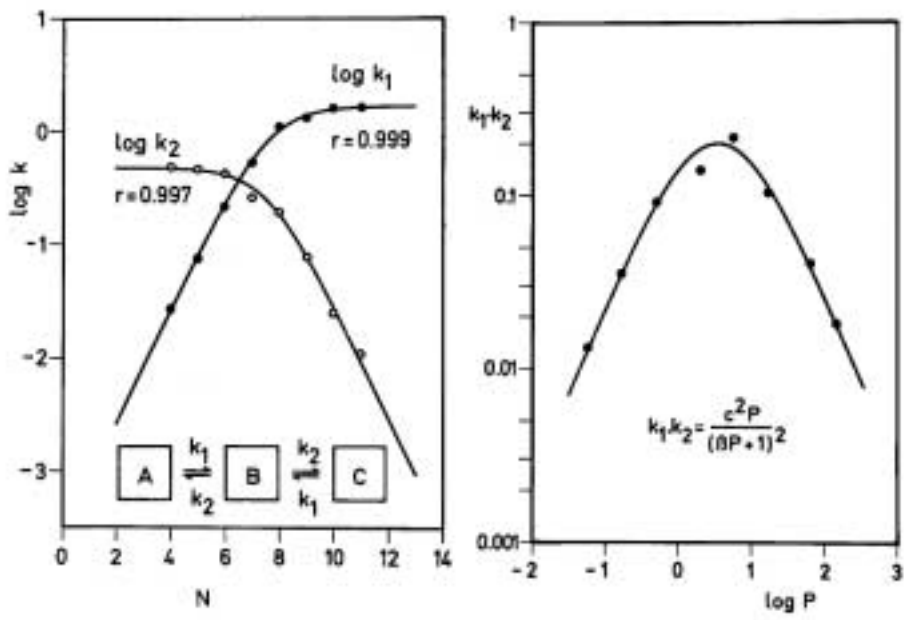
Rate constants of drug transport, k_1 and k_2

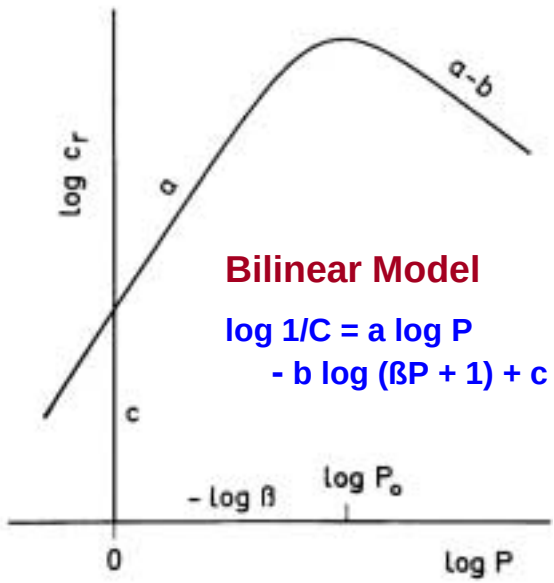
B. C. Lippold and G. F. Schneider, Arzneim.-Forsch. 25, 843 (1975)



Homologous Quaternary N-Alkylammoniumbromides, Water/n-Octanol/Water, + NaBr, experimental k_1 and k_2 values in h^{-1}.

Number of Carbon Atoms of Alkyl Group	k_1	k_2
4	0.027	0.490
5	0.076	0.470
6	0.217	0.425
7	0.534	0.264
8	1.112	0.196
9	1.340	0.078
10	1.620	0.025
11	1.650	0.011





Bilinear Model

log 1/C = a log P
- b log (ßP + 1) + c

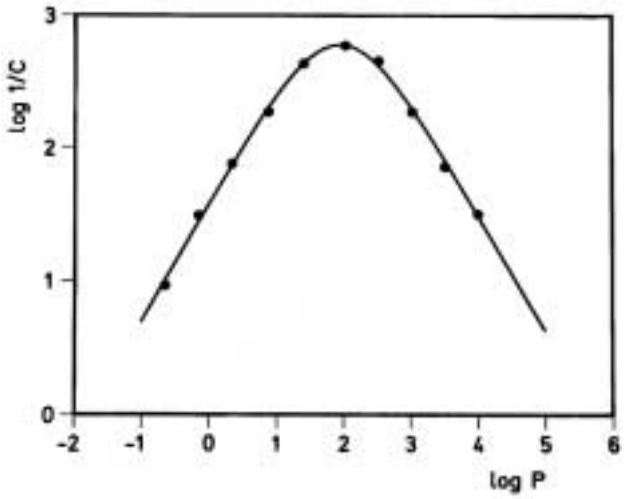
Advantages of the bilinear model

- better fit of the linear left and right sides
- better description of the lipophilicity optimum

Disadvantages of the bilinear model

- iterative estimation of the nonlinear parameter ß
- Loss of one degree of freedom (4 parameters, instead of 3)

Neurotoxic Activity of Primary n-Alcohols (rat, i.p. application)

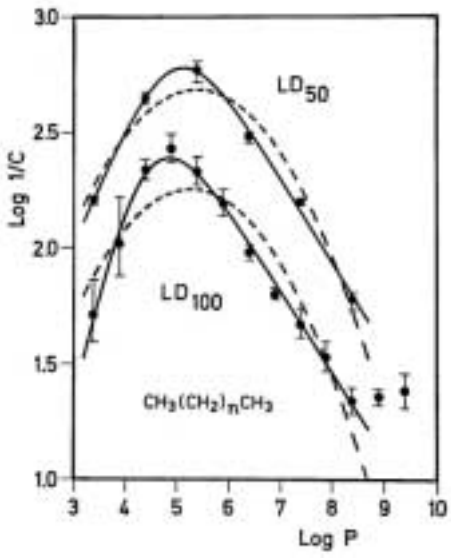


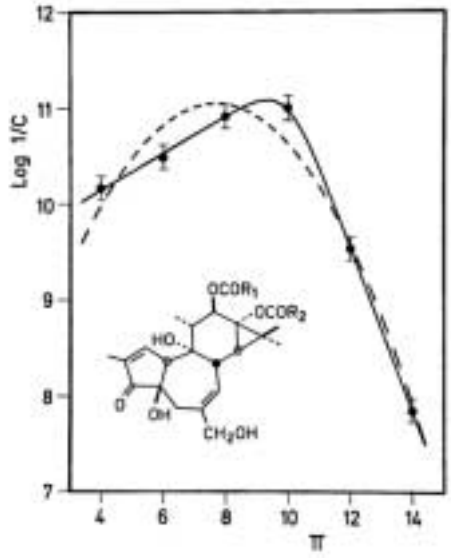
Log 1/C =
0.892 (±0.05) log P
- 1.766 (±0.10)
log (ßP + 1) - 1.586

log ß = - 1.933
log P optimum = 1.94

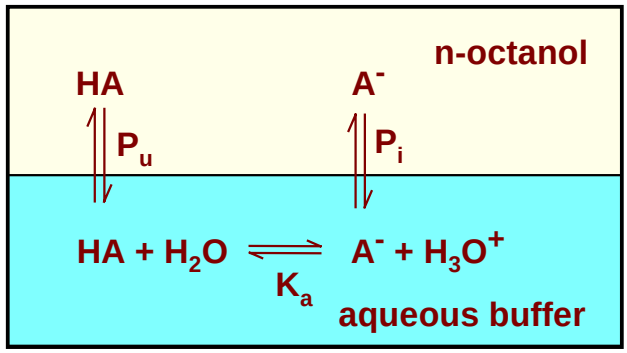
(n = 10; r = 0.996;
s = 0.041; F = 637.6)

Nonlinear Lipophilicity-Activity Relationships





Dissociation und Ionization



Henderson-Hasselbalch Equation $[H_3O^+] = K_a.(f_u/f_i)$

$pH = pK_a + \log (f_i/f_u) = pK_a + \log f_i - \log f_u$

f_u = fraction of neutral form (undissoziated acid or base)
$f_i = \alpha$ = fraction of charged form (= $1 - f_u$)

Hugo Kubinyi, www.kubinyi.de

Dissociation of Acids and Ionization of Bases

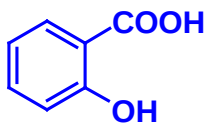
$$P_{app} \approx \frac{[AH]_{org}}{[AH]_{aq} + [A^-]_{aq}} = (1 - \alpha)\, P_u = \frac{P_u}{1 + 10^{pH - pK_a}}$$

Acids: $\log P_{app} = \log P_u - \log (1 + 10^{pH - pK_a})$

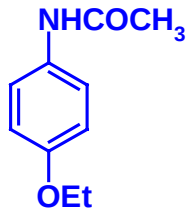
$$P_{app} \approx \frac{[B]_{org}}{[BH^+]_{aq} + [B]_{aq}} = (1 - \alpha)\, P_u = \frac{P_u}{1 + 10^{pK_a - pH}}$$

Bases: $\log P_{app} = \log P_u - \log (1 + 10^{pK_a - pH})$

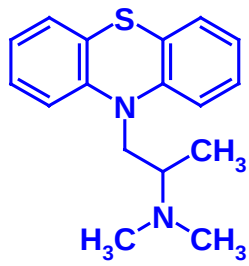
Distribution of Acids, Bases and Neutral Compounds



1, Salicylic Acid
pK_a = 3.0



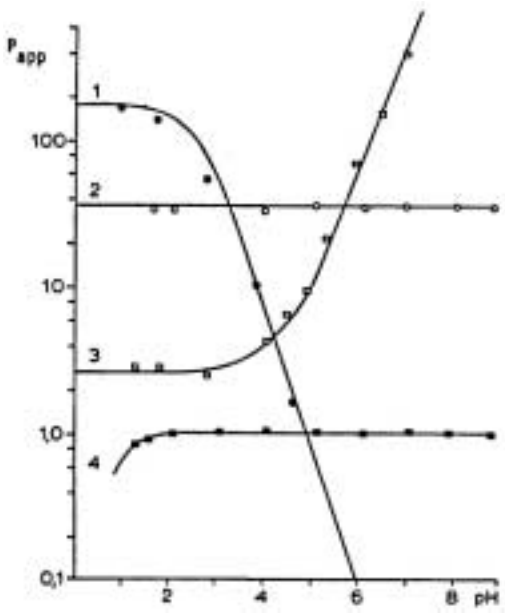
2, Phenacetin
neutral



3, Promethazine
pK_a = 9.1

4, Caffeine
pK_a = 0.6

Hugo Kubinyi, www.kubinyi.de

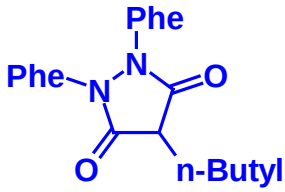


Approximation for bases

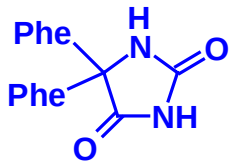
for pH > pK_a
log P_{app} = log P_u

for pH < pK_a
log P_{app} ≈
log P_u - pK_a + pH

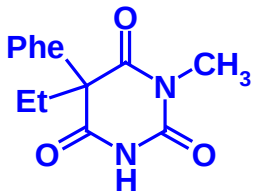
for pH << pK_a
log P_{app} = log P_i



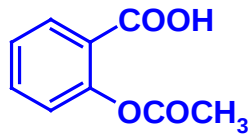
1, Phenylbutazone
pK_a = 4.5



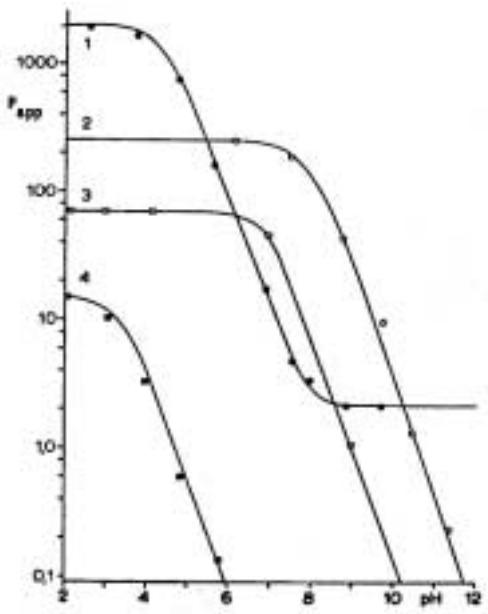
2, Diphenylhydantoin
pK_a = 8.3



3, N-Methyl-phenobarbital
pK_a = 7.4



4, Acetylsalicylic Acid
pK_a = 3.5



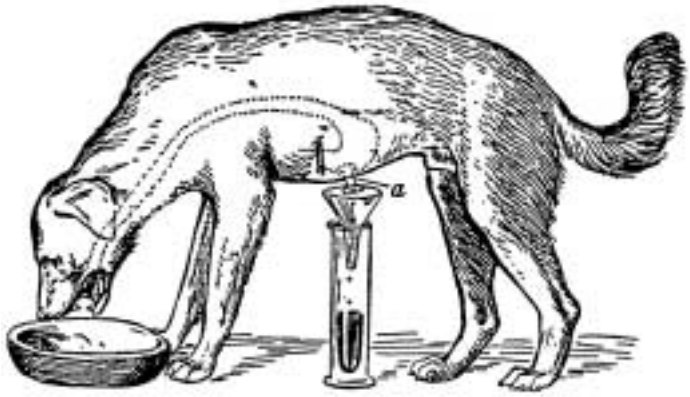
Approximation for acids

for pH < pK_a
log P_{app} = log P_u

for pH > pK_a
log P_{app} ≈
log P_u – pH + pK_a

for pH >> pK_a
log P_{app} = log P_i

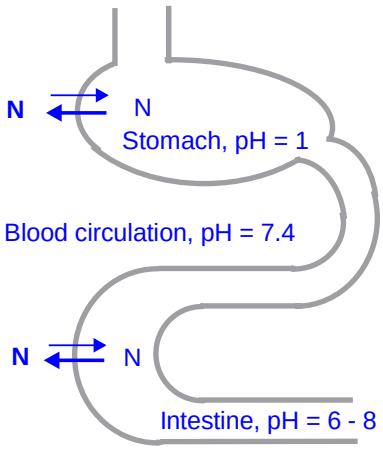
pH Dependence of the Absorption of Acids and Bases



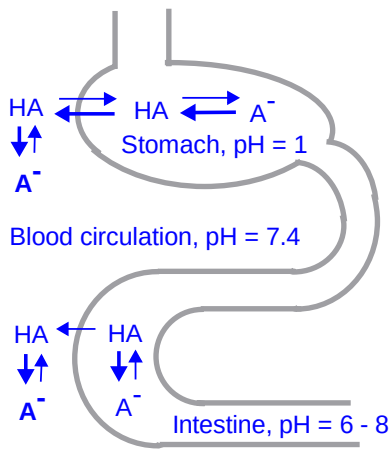
A strong base like strychnine, pK_a = 8.3, is not toxic for a dog with a pylorus ligation.

Gastrointestinal Absorption of Neutral Compounds and Acids

a) Neutral drug

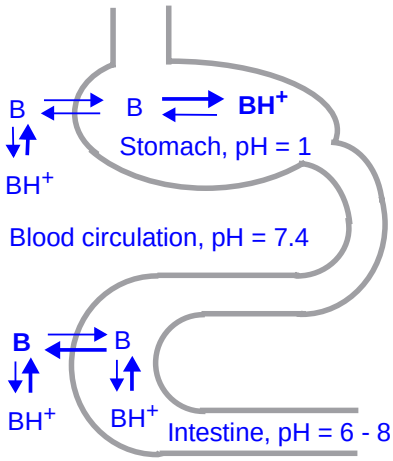


b) Acid, pK_a = 4

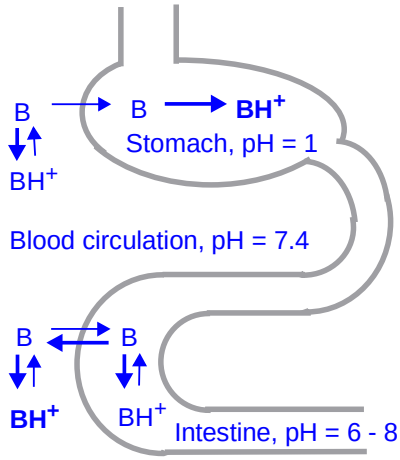


Gastrointestinal Absorption of Weak and Strong Bases

c) Weak base, pK_a = 5



d) Strong base, pK_a = 9



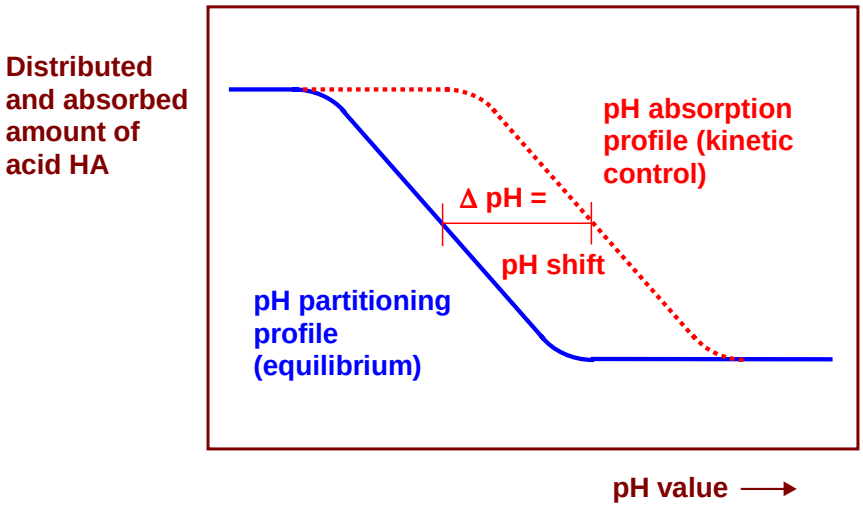
Buccal Absorption of Acids and Bases

Compound	pH	log P_{app}	log k_{abs}
Propranolol	5.08	-1.04	-2.19
	6.02	-0.10	-1.71
log P = 3.33	7.00	0.88	-1.22
pK_a = 9.45	7.93	1.81	-0.79
	8.94	2.70	-0.53
	9.93	3.21	-0.35
p-Hexylphenyl-	4.00	4.20	-0.46
acetic acid	5.00	3.63	-0.54
	6.00	2.72	-0.72
log P = 4.25	7.00	1.72	-1.07
pK_a = 4.36	8.00	0.72	-1.44
	9.00	-0.28	-1.78

$$\log k_{abs} = 0.45\ (\pm 0.05)\ \log P_{app} - 0.45\ (\pm 0.05)\ \log (0.0016 \cdot P_{app} + 1) - 1.69$$

$$(n = 12;\ r = 0.988;\ s = 0.102)$$

The “pH Shift” in the Absorption of Lipophilic Acids and Bases



3D-QSAR Approaches

CoMFA, Comparative Molecular Field Analysis

(Richard Cramer et al., 1988)

Select training and test sets of comparable diversity.

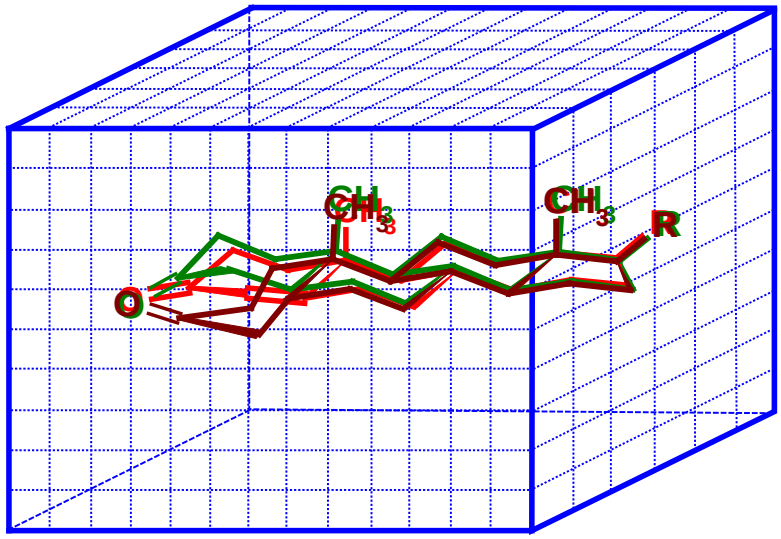
Generate 3D structures of all molecules of the data set.

Establish orientation rules for superposition of the molecules, using e.g. the "active analog approach".

Align the molecules according to their pharmacophore and surface properties, using e.g. the program SEAL.

Insert the molecules in a box and generate a grid that covers also a sufficiently large volume around the molecules.

Box and Grid



Calculate property fields for every molecule, in each grid point, by using probe atoms or groups (steric and electrostatic effects are most often separately treated; in addition, hydrophobic effects as well as hydrogen bond donor and acceptor properties may be considered).

Grid points with low variance may be neglected or variable / region selection can be performed.

Perform PLS analysis, using an optimum number of latent variables (vectors) to correlate biological activities.

Check the internal predictivity of the PLS model by a stepwise elimination of objects (crossvalidation). If necessary, repeat these steps.

Predict the biological activities of the test set molecules.

Electrostatic, Steric and "Similarity" Fields

Coulomb potential:

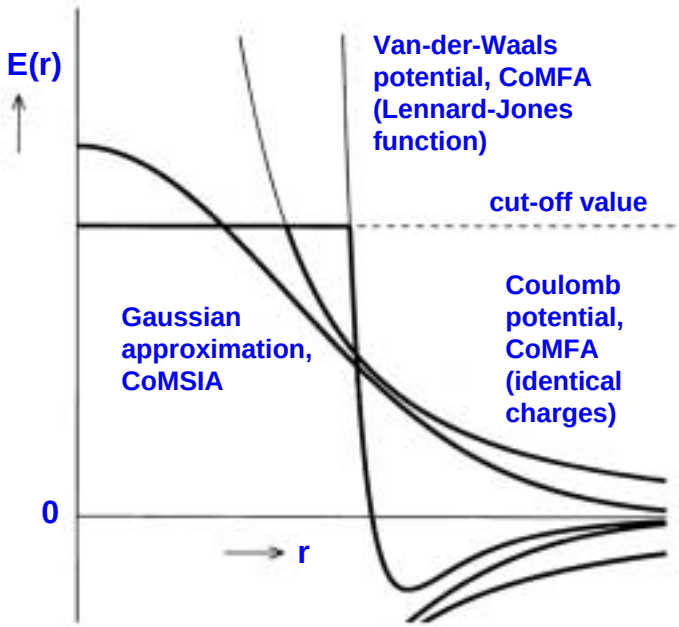
$$E_C = \sum_{i=1}^{n} \frac{q_i q_j}{D r_{ij}}$$

Lennard-Jones potential:

$$E_{vdW} = \sum_{i=1}^{n} \left(A_j \, r_{ij}^{-12} - C_j \, r_{ij}^{-6} \right)$$

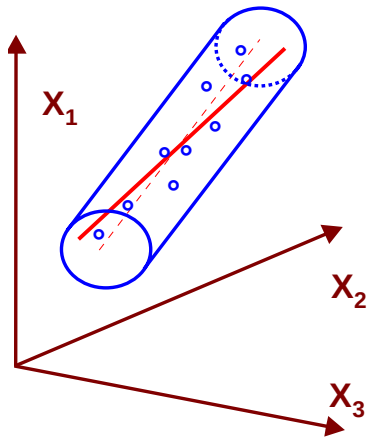
SEAL similarity coefficients:

$$A_F = -\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \, e^{-\alpha r_{ij}^2} \quad ; \quad w_{ij} = w_E q_i q_j + w_S v_i v_j + \ldots$$



CoMFA and CoMSIA Potentials

PLS (Partial Least Squares) Analysis



From $u = kt$

(k = constants $k_1, k_2 \ldots k_j$;
j = number of PLS vektors)

follows:

$$BA_i = a_1S_{i1} + a_2S_{i2} + a_3S_{i3} + \ldots + a_mS_{im} + b_1E_{i1} + b_2E_{i2} + b_3E_{i3} + \ldots + b_mE_{im}$$

Y vector

$$\begin{matrix} BA_1 \\ BA_2 \\ BA_3 \\ \vdots \\ \vdots \\ BA_n \end{matrix}$$

X matrix

$$\begin{matrix} S_{11} & S_{12} & S_{13} & S_{14} & \ldots & S_{1m} & E_{11} & E_{12} & \ldots & E_{1m} \\ S_{21} & S_{22} & S_{23} & S_{24} & \ldots & S_{2m} & E_{21} & E_{22} & \ldots & E_{2m} \\ S_{31} & S_{32} & S_{33} & S_{34} & \ldots & S_{3m} & E_{31} & E_{32} & \ldots & E_{3m} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ S_{n1} & S_{n2} & S_{n3} & S_{n4} & \ldots & S_{nm} & E_{n1} & E_{n2} & \ldots & E_{nm} \end{matrix}$$

PLS analysis

SAMPLS analysis

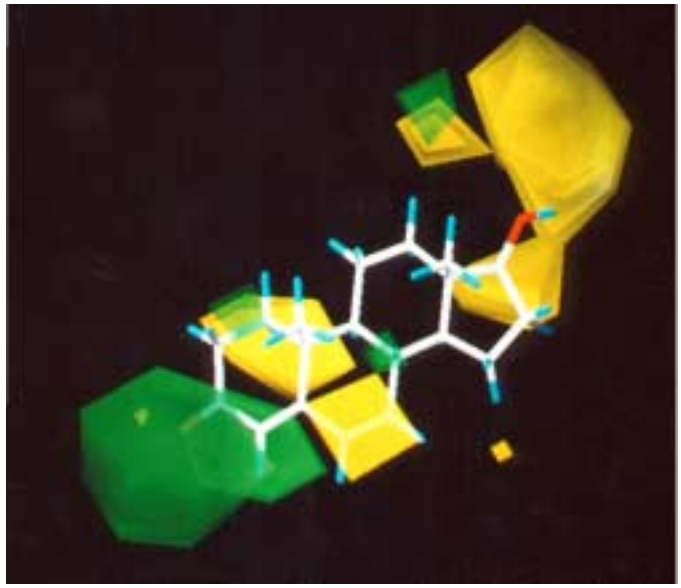
$$\begin{matrix} u_{11} \\ u_{12} \\ u_{13} \\ \vdots \\ \vdots \\ u_{1n} \end{matrix} \quad \begin{matrix} t_{11} \\ t_{12} \\ t_{13} \\ \vdots \\ \vdots \\ t_{1n} \end{matrix} \quad , \quad \begin{matrix} t_{21} \\ t_{22} \\ t_{23} \\ \vdots \\ \vdots \\ t_{2n} \end{matrix} \quad , \ldots$$

$$\begin{matrix} c_{11} & c_{12} & c_{13} & \ldots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \ldots & c_{2n} \\ c_{31} & c_{32} & c_{33} & \ldots & c_{3n} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \ldots & c_{nn} \end{matrix}$$

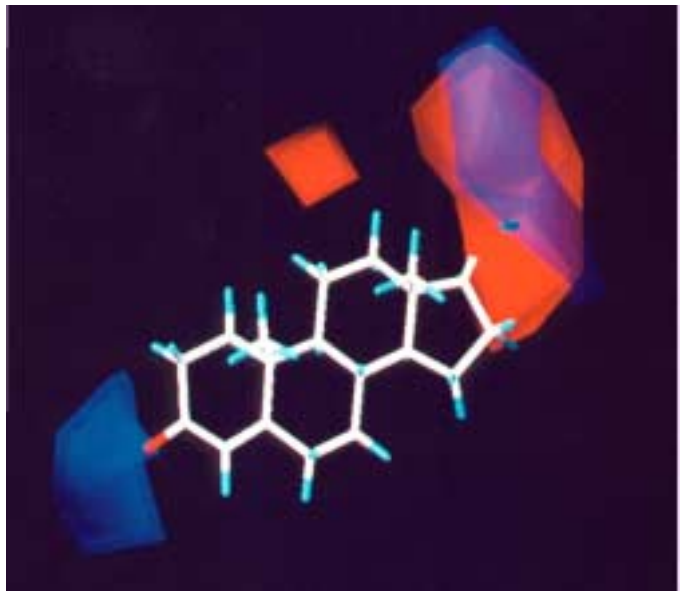
PLS vectors
(latent variables)

covariance matrix

Steric CoMFA Contour Map



Electro-static CoMFA Contour Map



Recommendations for CoMFA Analyses („good practice“)

Type of biological data (affinities, inhibition constants)
Variance and error range of biological data
Selection of start geometries (flexible molecules)
Method for calculation of charges should be cited
Pharmacophore for superposition of the molecules
Description of the alignment (atom-by-atom, field-based)
Scaling and weighting of fields
Number of PLS vectors („Occam‘s razor“)
Variable or region selection
Crossvalidation - LOO or groups
Crossvalidation only for internal prediction -
Prediction of a test set !

QSAR and 3D-QSAR, Scope and Limitations

a) Free Wilson Analysis

\+ easy to perform, most often unique solutions
\+ clear separation of substituent effects
\+ helps in the derivation of Hansch models
\+ combination with Hansch analysis to a mixed approach

\- needs at least two sites of chemical variation
\- many parameters, only few degrees of freedom
\- very narrow QSAR models, no "outside" predictions

b) Hansch Analysis

+ correlates activities with physicochemical properties
+ "outside" predictions are possible

- only applicable to congeneric series
- works best with simple variation of aromatic substituents
- considers only 2D structures
- no unique solutions
- risk of chance correlations (number of variables!)
- risk of failure in "too far outside" predictions

c) 3D QSAR (CoMFA and CoMSIA)

+ considers 3D structures of the ligands
+ applicable to more heterogeneous data sets
+ elektrostatic, steric, hydrophobic and hydrogen bond fields
+ 3D maps of favorable and unfavorable interactions

- uncertainties about the bioactive conformation
- uncertainties about different binding modes of ligands
- cut-off levels (can be avoided in CoMSIA)
- variable selection yields fragmented contour maps
- high risk (if not guarantee) of chance correlations
- only applicable to *in vitro* data