QSAR = QUANTITATIVE STRUCTURE – ACTIVITY RELATIONSHIPS

We are searching for a relationship where a quantified biological activity is a function of structure or parameters which are connected with structure respectively

A= f (structure)

2 basic approaches of classical QSAR

•**regression analysis –** searches for a mathematical description of the function in most using linear or other regression

•**empirical methods –** search only for **extremes** (maxima or minima) of a given function without recognizing of its mathematical description thus the function remains a “black box”

Regression analysis

searches for an equation in form $A = a_0 + a_1x_1 + a_2x_2 + \ldots a_nx_n$,

where A is a quantified biological activity, x are parameters or descriptors derived from compounds´ structure, a , b are regression coefficients (a_0 …absolute term) acquired by calculation. In case of so called **Hansch method,** x are physico-chemical descriptors derived from the structure, in case of so called **Free-Wilson** approach x parameters express simple presence or non-presence of a particular substituent or structural fragment in the molecule.

Hansch method of regression analysis

$$A = a_0 + a_1x_1 + a_2x_2 + \ldots a_nx_n$$

A … a quantified **biological activity**, often in reciprocal or in logarithm in order to get linearity of equation

Examples:

- 1/MIC … reciprocal minimal inhibition concentration in antimicrobial compounds
- log ED_{50} ... logarithm of a dose which causes a desired effect in 50 % of testing subjects
- log LD_{50}... a parameter of acute toxicity; logarithm of a dose which causes death in 50 % of testing animals
- IC_{50} … a concentration of studied compound which lowers enzyme activity to its 50%
- log BB … express the ability of a compound cross the blood-brain barrier
- … etc.

a_1 ... a_n ... **regression coefficients** i.e. coefficients acquired by calculation using e.g. linear regression

“Classical” parameters x_1 ... x_n

- hydrophobic
- electronic
- steric

a) **Hydrophobic parameters –** in an equation often in square – they express ratio of solubility of a compound in lipids and in water; they often fundamentally impact compound activity particularly penetration through barrier systems of an organism e.g. **log P(octanol/water), log P(cyclohexane/water)** etc., parameter **R_m** from partition thin layer chromatography (TLC) on so called **reversed phase** (stacionary phase is lipophilic, mobile phase hydrophilic):

$$R_m = \log\left(\frac{1}{R_f} - 1\right),$$

further logarithm of **capacity factor log k´** from gas chromatography (GC) or reversed phase high-performance liquid chromatography (RP-HPLC)

$$\log k' = \log\left(\frac{(t_r - t_0)}{t_0}\right),$$

where t_r is retention time of a studied compound and and t_0 so called dead time of a column i.e. retention time of a compound which is not retained at the column (e.g. sodium nitrite is used in RP-HPLC on octadecylated silica gel) further (Hansch) **lipofilicity parameter π** – for series of compounds which contain various substituents on the same structural fragment (mostly often benzene ring)

$$\pi = \log \frac{P_X}{P_H} = \log P_X - \log P_H$$

where P_X is partition coefficient of the substituted compound and P_H partition coefficient of the unsubstituted one.

Calculated hydrophobic parameters

Except experimentally determined hydrophobic parameters are recently used estimations of such parameters acquired by means of calculations according to various algorithms. Among them, probably procedures for estimation of log P (octanol/water) by means of sum of log P increments belong to the simpliest ones, e.g. the formula of **Rekker** and **Nyss**

$$\log P = \sum^{i} a_i f_i,$$

where f_i called the fragment constant is log P of the particular fragment and a_i je is the count of occurrence of such fragment, or more precious estimation according to **Hansch** (and Leo) defined by formula

$$\log P = \sum^{i} a_i f_i + \sum^{j} b_j f_j ,$$

where f_i is the fragment constant, f_j is the correction factor which tries to respect the placement of a particular fragment in a moleule and its neighborhood and a_i and b_j are counts of occurrence of a given parameter. This calculated log P is frequently marked **Clog P**. However, much more complex procedures are recently used. They need computers and suitable software which in most enables also optimization of structure by means of molecular mechanics methods and calculations of some additional parameters for QSAR calculations (for PC e.q. Molgen, HyperChem). The conformity of calculated log P estimation with experimentally determined value is very different for various computing algorithms although a linear relationship between experimental and computed values suffices in many cases.

Mlog P

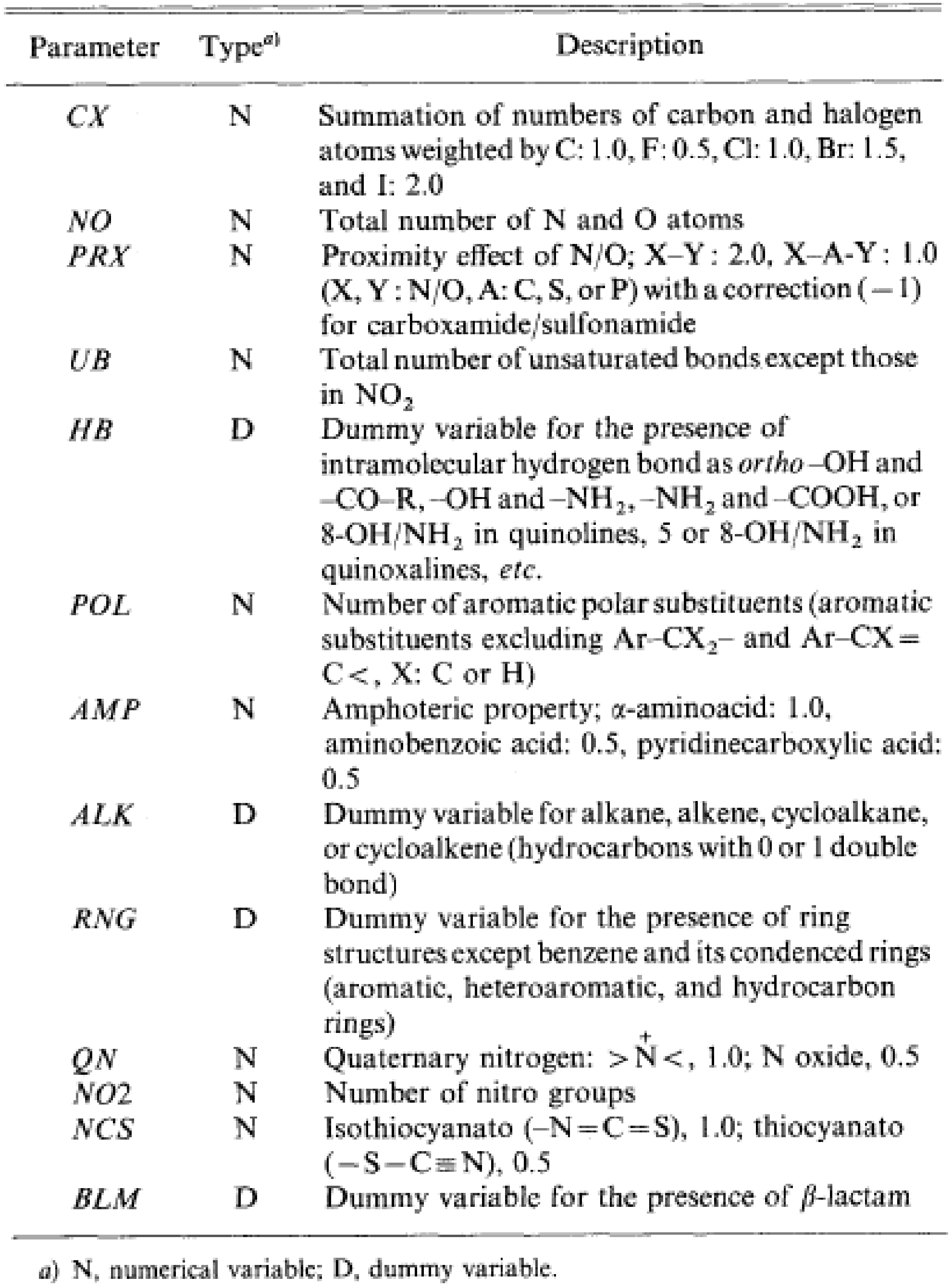
- derived by a statistic analysis of log P and structural data of 1230 compounds
- defined by formula

$$\begin{aligned}
\log P = {} & 1.244\,(CX)^{0.6} - 1.017(NO)^{0.9} + 0.406\,PRX \\
& \quad (t=60.5) \qquad\quad (t=58.5) \qquad\quad (t=33.8) \\
& -0.145(UB)^{0.8} + 0.511\,HB + 0.268\,POL - 2.215\,AMP \\
& \quad (t=9.5) \qquad\quad (t=5.9) \qquad (t=19.6) \qquad (t=19.5) \\
& +0.912\,ALK - 0.392\,RNG - 3.684\,QN + 0.474\,NO2 \\
& \quad (t=9.5) \qquad (t=13.1) \qquad (t=22.1) \qquad (t=10.8) \\
& +1.582\,NCS + 0.773\,BLM - 1.041 \\
& \quad (t=16.4) \qquad (t=5.0)
\end{aligned}$$

$$n=1230, \quad r=0.952, \quad s=0.411, \quad F_0(13,1216)=900.4$$

Mlog P (continued)

where



Parameter	Type[a]	Description
CX	N	Summation of numbers of carbon and halogen atoms weighted by C: 1.0, F: 0.5, Cl: 1.0, Br: 1.5, and I: 2.0
NO	N	Total number of N and O atoms
PRX	N	Proximity effect of N/O; X–Y : 2.0, X–A-Y : 1.0 (X, Y : N/O, A: C, S, or P) with a correction (−1) for carboxamide/sulfonamide
UB	N	Total number of unsaturated bonds except those in NO_2
HB	D	Dummy variable for the presence of intramolecular hydrogen bond as *ortho* –OH and –CO–R, –OH and $-NH_2$, $-NH_2$ and –COOH, or 8-OH/NH_2 in quinolines, 5 or 8-OH/NH_2 in quinoxalines, *etc.*
POL	N	Number of aromatic polar substituents (aromatic substituents excluding Ar–CX_2– and Ar–CX = C<, X: C or H)
AMP	N	Amphoteric property; α-aminoacid: 1.0, aminobenzoic acid: 0.5, pyridinecarboxylic acid: 0.5
ALK	D	Dummy variable for alkane, alkene, cycloalkane, or cycloalkene (hydrocarbons with 0 or 1 double bond)
RNG	D	Dummy variable for the presence of ring structures except benzene and its condenced rings (aromatic, heteroaromatic, and hydrocarbon rings)
QN	N	Quaternary nitrogen: $>\overset{+}{N}<$, 1.0; N oxide, 0.5
NO2	N	Number of nitro groups
NCS	N	Isothiocyanato (–N=C=S), 1.0; thiocyanato (−S−C≡N), 0.5
BLM	D	Dummy variable for the presence of β-lactam

[a] N, numerical variable; D, dummy variable.

An example of a QSAR relationship with a hydrophobic parameter only

Effect of some phenols as apoptose inductors in cancer cells

Hansch, C. et al.: Bioorg. Med. Chem. ***11****, 617 (2003)*

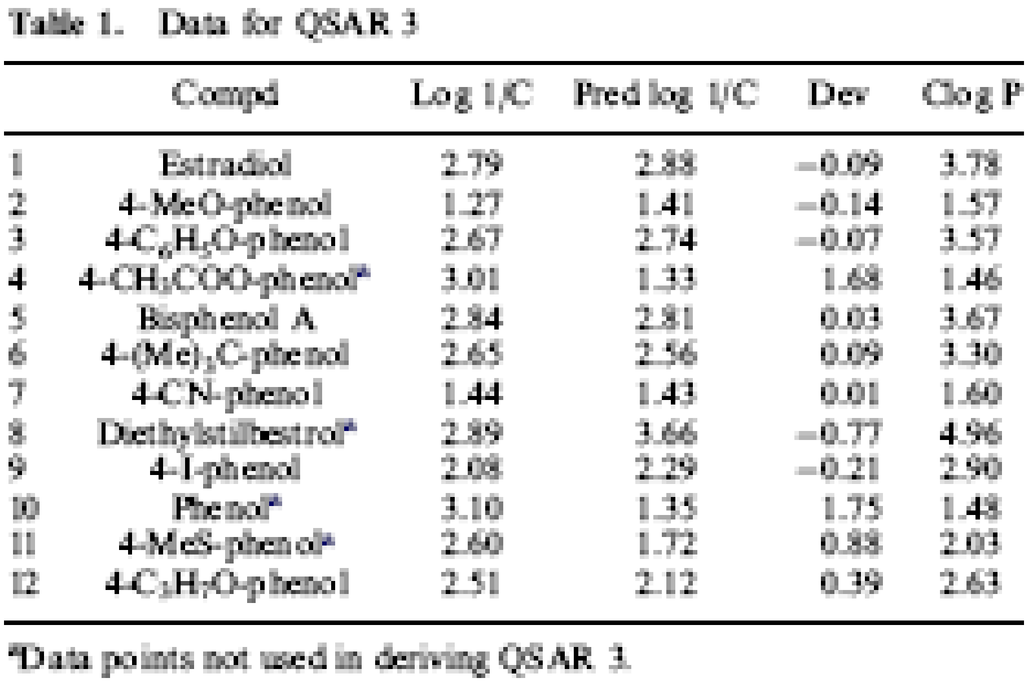


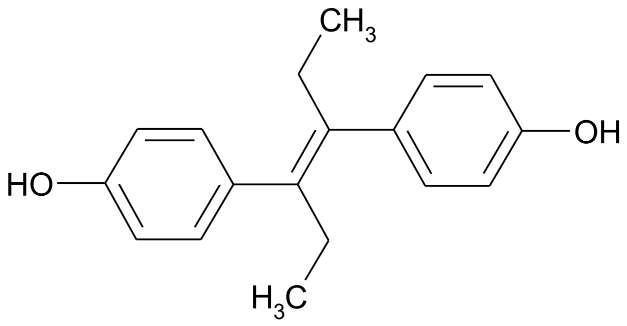
Table 1. Data for QSAR 3

	Compd	Log 1/C	Pred log 1/C	Dev	Clog P
1	Estradiol	2.79	2.88	−0.09	3.78
2	4-MeO-phenol	1.27	1.41	−0.14	1.57
3	4-C_6H_5O-phenol	2.67	2.74	−0.07	3.57
4	4-CH_3COO-phenol[a]	3.01	1.33	1.68	1.46
5	Bisphenol A	2.84	2.81	0.03	3.67
6	4-$(Me)_3C$-phenol	2.65	2.56	0.09	3.30
7	4-CN-phenol	1.44	1.43	0.01	1.60
8	Diethylstilbestrol[a]	2.89	3.66	−0.77	4.96
9	4-I-phenol	2.08	2.29	−0.21	2.90
10	Phenol[a]	3.10	1.35	1.75	1.48
11	4-MeS-phenol[a]	2.60	1.72	0.88	2.03
12	4-C_2H_5O-phenol	2.51	2.12	0.39	2.63

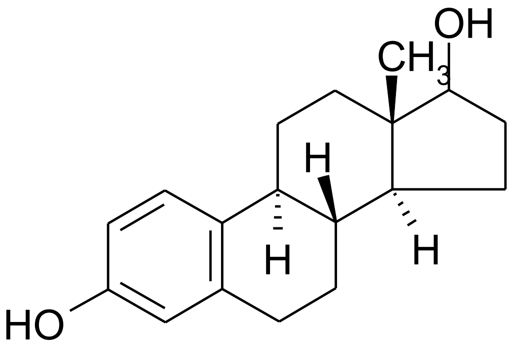
[a] Data points not used in deriving QSAR 3.

$$\log 1/C = 0{,}67(\pm 0{,}21)\mathrm{ClogP} + 0.37(\pm 0.63)$$

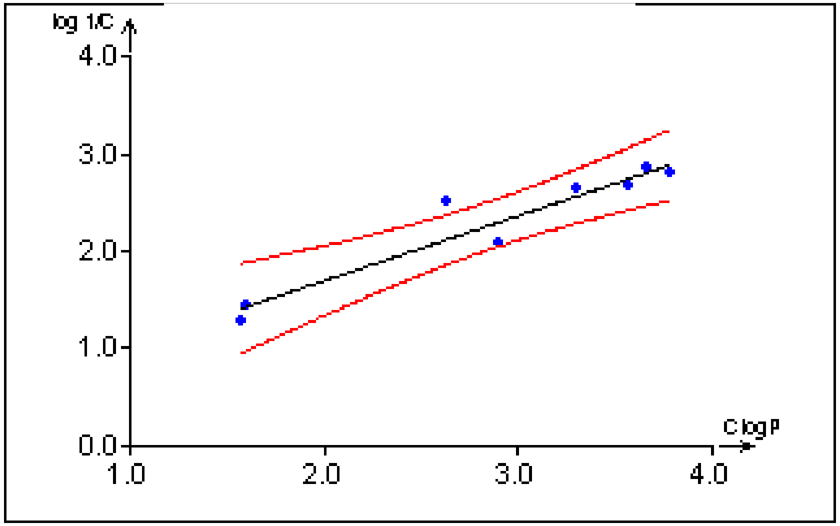
$$n = 8,\ r^2 = 0{,}910,\ s = 0{,}201,\ q^2 = 0{,}863$$



diethylstilbestrol (8)



estradiol (1)



b) Electronic parameters

-directly or indirectly linked with the electronic coat of a particular molecule

- **Hammet constants $\boldsymbol{\sigma}$** - for *m-* a *p*-substituted benzene; they express electron-donor (+M, +I) or elektron-accepting (-M, -I) properties of a substituent; derived constants: σ_m, σ_i, σ^*, similar Swain-Lupton constants $\mathscr{F}$, $\Re$
- parameters from **spectra** and other physical measurements – chemical shifts δ from NMR, wavelength of absorbtion maximum $\boldsymbol{\lambda}_{\mathbf{max}}$ from UV-VIS spectra, wavenumber $\boldsymbol{\nu}$ of a significant absorption band in IR spectra, half-wave potential $\mathbf{E}_{\mathbf{1/2}}$from polarography etc.; values must be significantly different for every member of a studied series
- calculated electronic parameters: polarity, polarizability, partial charge at a particular atom etc.

c) Steric parameters

- express “overall bulkiness” of a molecule or preferably of a particular substituent on a common skeleton

- **van der Waals radii $\mathbf{v}_{\mathbf{F}}$**
- **Taft steric constant $\mathbf{E}_{\mathbf{s}}$** derived by means of rate constants of alkanoic acids esters hydrolysis

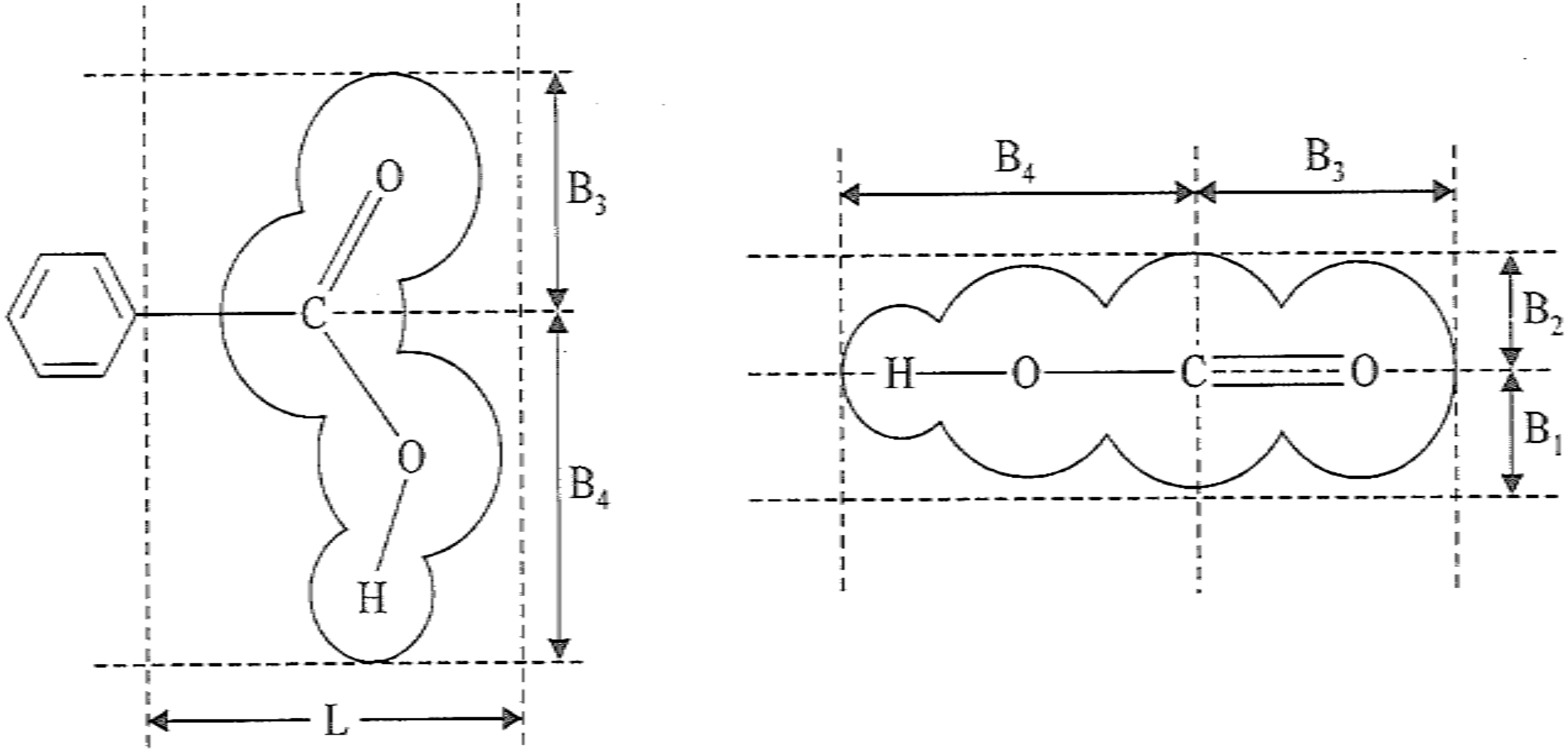
$$E_s = \log \frac{k_x}{k_h},$$

where k_x is the rate constant of hydrolysis of an ester of a particular alkanoic acid RCOOR´ and k_h the same constant for the corresponding acetic acid ester $CH_3COOR´$- a standard. E_s is not a purely steric parameter because it partially includes also electronic influnce (+I).

$E_s(CH_3) = 0$, for more bulky substituents $E_s < 0$, for less bulky ones $E_s > 0$

- Verloop steric parameters
    - for particular substituents
    - derived (measured) from computed molecule geometry (Sterimol)
    - L represents the length of the substituent and B_1 – B_4 designate the radii (i.e. longitudinal and horizontal)

(An example of carboxylic moiety in benzoic acid molecule)



Other parameters used in QSAR

- in most computed
- in most characterize the whole molecule
- often include 2 or 3 types of influence (hydrophobic+ electronic + steric)

“Classical”

•**parachor** $$P_r = \frac{M}{d} \gamma^{(1/4)}$$

where γ is surface tension, M molar weight and d density**.**

•**molar refraction** (= molecular refractivity) **MR** (also CMR); definition formula is known as Lorentz-Lorenz equation

$$MR = n^2 - \frac{1}{n^2} + \frac{2\mathrm{M}}{d} = \frac{(n^2 - 1)}{(n^2 + 2)} \frac{M}{d},$$

where n is refractivity index.

“Non-classical”

•**solvation energy –** if it is for water then **hydration energy ΔG_o^w**

•**molecule surface areas** of various kind: – polar van der Waals, non-polar, water accessible, dynamic polar (DPSA), topologic polar (TPSA) etc.

•**molecule volumes –** polar, water accessible etc.

Free- Wilson method of regression analysis

- searches for a relationship between a biological activity and presence or non-presence of some substituents or structural fragments in a molecule. Exactly it is **statistic separation of activity** into contributions of particular parts of a molecule i.e. aditivity of influence of substituents or other molecular fragments is assumed. Such a method leads to solution of equation systems of higher number of unknowns which are in simple cases to solve by means of matrix arithmetic otherwise by statistic software enabling multilinear regression (MLR).
- both Hansch and Free-Wilson methods could also be combined. A part of autonomous variables then express physico-chemical properties of compounds and other ones which are called **“indicator variables”** (symbol I) express presence or non-presence of particular molecular fragments. Usually there is only small count of indicator variables, often only one.

Empirical methods of QSAR

- preferred to use there where the mathematical description of the function A = f(structure) is not easily to find
- search only for extremes (maxims and/or minims) of given function; its mathematical description remains a “black box”
- while applied a synthetic chemist choices compounds to synthesize according to biological evaluation of previous ones

Optimization according one structural parameter: Fibonacci optimization

This method is based on the Fibonacci progression part of which is expressed in the Table. Compounds are ordered in accordance with the increasing value of a structural parameter which is assumed to influence the activity significantly. The number of compounds must conform to the number of points in some Fibonacci interval (see Table. If it is not so one of marginal compounds which are not probable to be the most active is excluded or on the contrary a fictive marginal compound is added. Compounds, which have numbers listed in the column C of the Table in a particular interval, are selected for synthesis. Their biological activities are determined and, in dependence of its results, the part of the given interval from one of marginal points to the less active compound is excluded. The resulted set of compounds is the next Fibonacci interval. Such selection is repeated until the most active compound is reached. This method enables to decrease the number of synthesized and tested compounds significantly e.g. instead of 589 compounds which were necessary to prepare and test to find the most active one, only 13 compounds are sufficient to synthesize and evaluate (see column C of the Table).

Table: Fibonacci optimization

Legend: A … number of compounds of a particular Fibonacci interval
B … order of compounds selected for synthesis and evaluation in a particular interval
C … total number of compounds needed for optimization

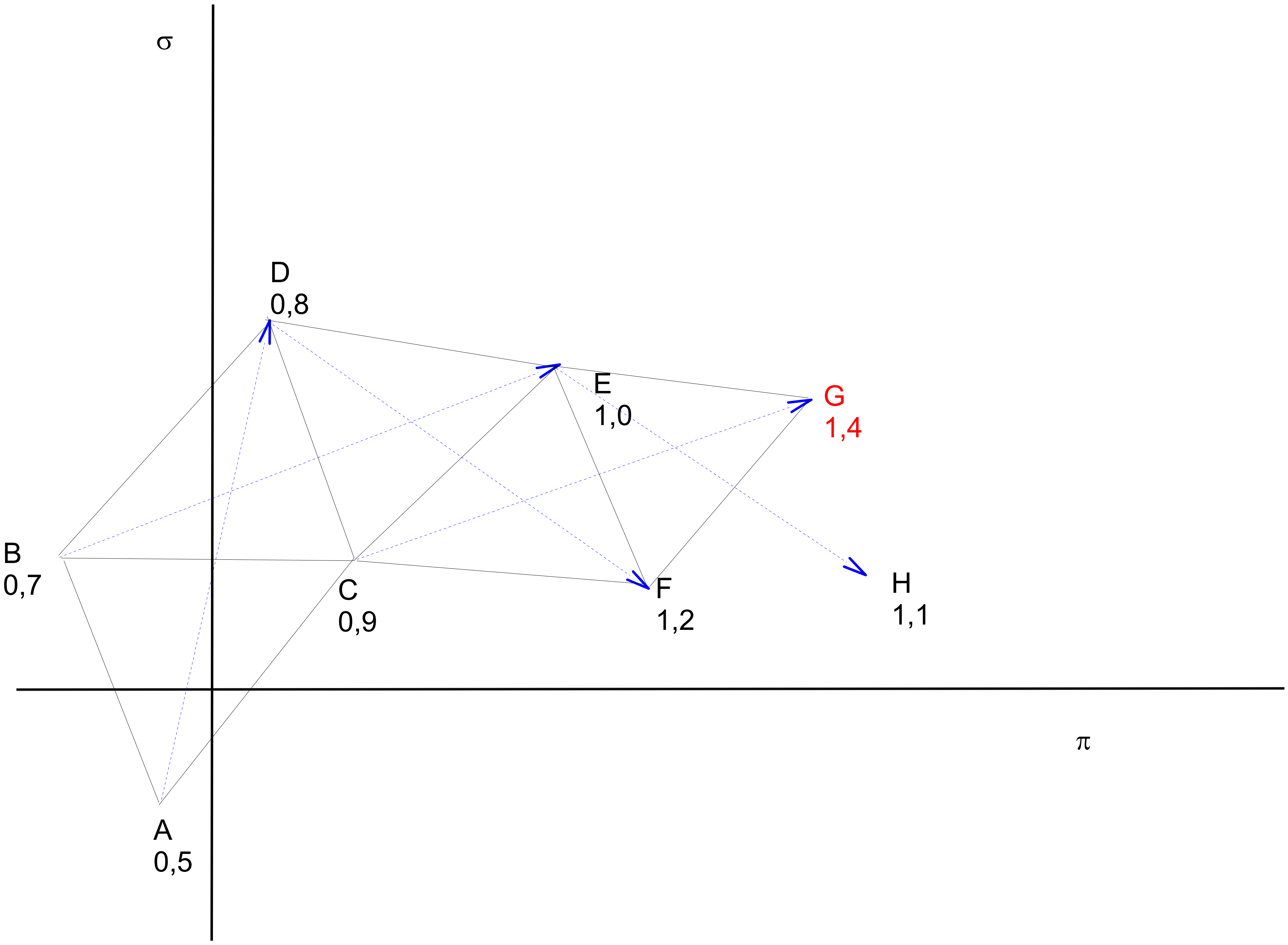
A	B	C	A	B	C	A	B	C
2	l and 2	2	20	8 and 13	6	143	55 and 89	10
4	2 and 3	3	33	13 and 21	7	222	89 and 144	11
7	3 and 5	4	54	21 and 34	8	366	144 and 233	12
12	5 and 8	5	88	34 and 55	9	589	233 and 377	13

Optimization according more structure parameters

Simplex method

Every compound can be characterized as a particular point in n-dimensional space in which the first coordinate is a **biological activity** and additional coordinates belong to **physical and physico-chemical properties** which are assumed to influence the activity. If we work in classical three-dimensional space i.e. if we optimize only two parameters we can perform such optimization also graphically on a chart paper. In fact we work in the projection into the plane of properties. Three compounds which are not far from themselves in the plane of properties are selected for (synthesis and) evaluation. Ideally their coordinates form an equilateral triangle. We compare activities of such three compounds. Now we draw a half-line from the point which belongs to the compound of the least activity through the center of join of two points of higher activities (alternatively through the point originated by division of this joint in reversal ratio of activities) and at the line we find the point which has the same distance from the joint to the point of the least activity but the opposite orientation. If no compound belong to thus found point we select for evaluation the nearest one. This point and two previous ones give us the next triangle which is put to the same optimization procedure. This procedure is repeated as long as the activity increases. Once the activity begins to decrease the compound with the highest reached activity can be recognized as the most active one.

Simplex method



Optimization according more structure parameters

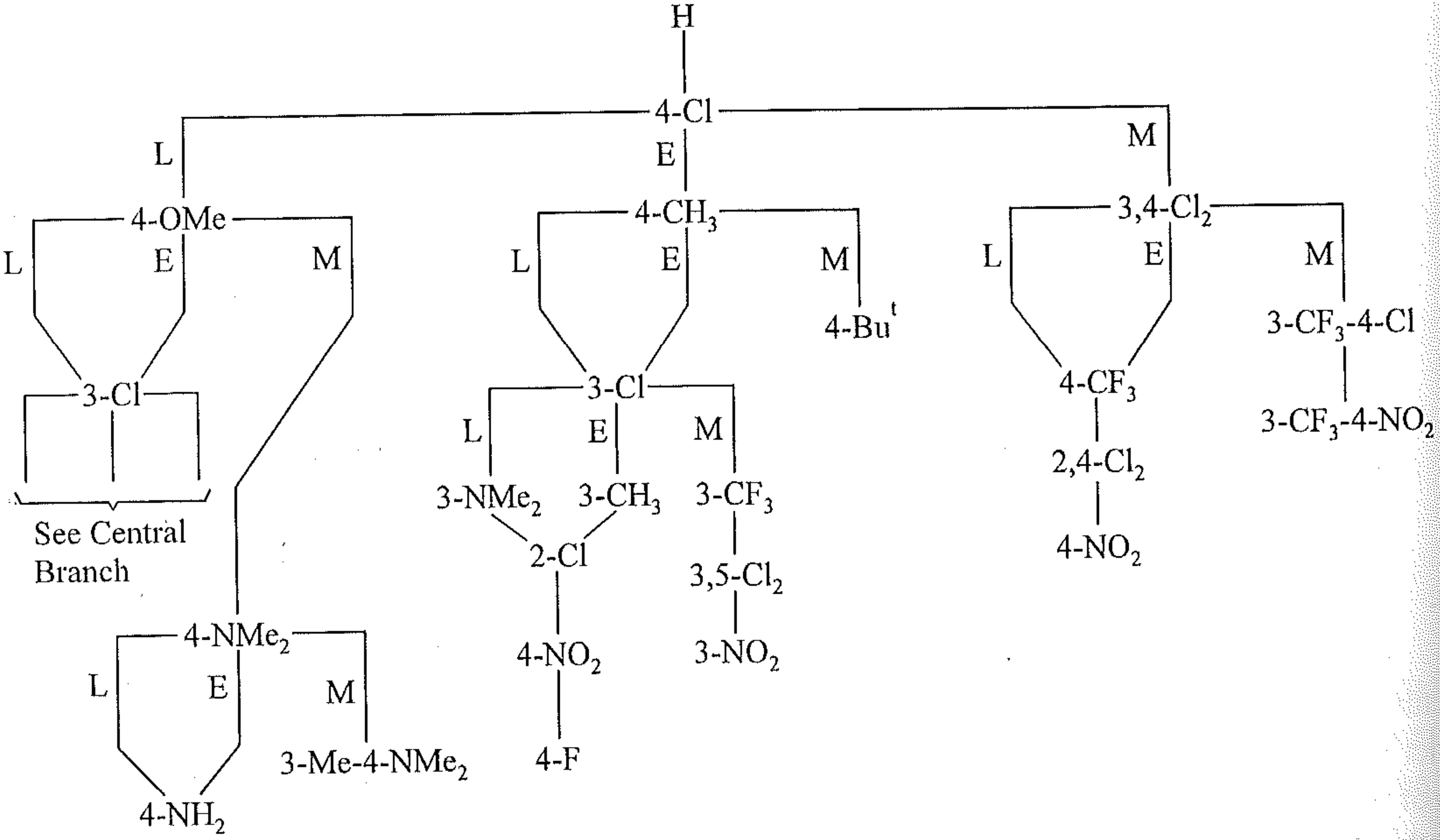
Optimization schemes

- sequences of rational intellectual processes of an medicinal chemist
- they have regard for hydrophobic, electron and steric parameters
- they are not universal: a novel one could be needed to formulate for a particular type of modifications of a particular structure

Scheme of modifications on phenyl (Topliss 1972):

- an active compound (= lead compound) having a moiety necessary for the activity (= a pharmacophore) bond on the unsubstituted benzene ring
- a pharmacophore cannot be modified unless the activity is lost but we can modify the benzene ring by any arbitrary substitution

Scheme of modifications on phenyl (Topliss 1972)



H

4-Cl

L E M

4-OMe

L E M

3-Cl

See Central Branch

4-NMe$_2$

L E M

4-NH$_2$

3-Me-4-NMe$_2$

4-CH$_3$

L E M

4-Bu^t

3-Cl

L E M

3-NMe$_2$

3-CH$_3$

2-Cl

4-NO$_2$

4-F

3-CF$_3$

3,5-Cl$_2$

3-NO$_2$

3,4-Cl$_2$

L E M

4-CF$_3$

2,4-Cl$_2$

4-NO$_2$

3-CF$_3$-4-Cl

3-CF$_3$-4-NO$_2$

Legend: E – equally active L – less active M – more active

Commentary to the scheme of modifications of substituents on phenyl

At first the **unsubstituted compound** and its **4-chloro derivative** are synthesized. Chlorine substitution lowers electron density in position one where the pharmacophore is bond and simultaneously increases lipophilicity (4-Cl: $\sigma = 0.23$; $\pi = 0.71$); if the 4-chloro derivative is more active both lipohilicity and electron-accepting properties can be further increased by further chlorine substitution. If the 4-chloro derivative is less active we can assume that electron density decrease influenced the activity negatively and **4-methoxy derivative** is prepared. It has almost the same lipophilicity as the unsubstituted compound but electron density in position one is higher (4-OCH_3: $\sigma = -0.27$, $\pi = 0.02$). If there is no significant difference between activities of unsubstituted compound and its 4-chloro derivative we can suppose that influences of electron density and lipophilicity act against each other and **4-methyl derivative** which has increased both is prepared (4-CH_3: $\sigma = -0.17$, $\pi = 0.56$). If activities of all compounds substituted in position 4 are lower than that of unsubstituted compound then there is evident that substitution in position 4 is sterically disadvantageous and compounds substituted in positions 2 and 3 can be prepared. Particular branches of this scheme can be continued until the compound with the optimal activity is reached.