Molecular Descriptors

HOW TO REPRESENT CHEMICAL STRUCTURES IN NUMBERS?

Content

- Molecular Descriptors
- Types of Descriptors
- Physicochemical Descriptors
  - Hydrophobic
  - Electronic
  - Shape

Molecular Descriptors

Molecular Property	Corresponding Interaction	Parameters
Lipophilicity	hydrophobic interactions	log P, π, f, R_M, χ
Polarizability	van-der-Waals interactions	MR, parachor, MV
Electron density	ionic bonds, dipol-dipol interactions, hydrogen bonds, charge transfer interactions	σ, *R*, *F*, κ, quantum chemical indices
Topology	steric hindrance geometric fit	E_S, r_V, L, B, distances, volumes

Type of Molecular Descriptors

- **0D-descriptors** (i.e. constitutional descriptors, count descriptors)
- **1D-descriptors** (i.e. list of structural fragments, fingerprints)
- **2D-descriptors** (i.e. graph invariants)
- **3D-descriptors** (i.e. quantum-chemical descriptors, size, steric, surface and volume)
- **4D-descriptors** (i. e. GRID or CoMFA methods, Volsurf)

What should a descriptor be like?

- Should have structural interpretation
- Should have good correlation with at least one property
- Should preferably discriminate among isomers
- Should be possible to apply to local structure
- Should possible to generalize to "higher" descriptors
- Should be simple

What should a descriptor be like?

- Should not be based on experimental properties
- Should not be trivially related to other descriptors
- Should be possible to construct efficiently
- Should use familiar structural concepts
- Should change gradually with gradual change in structures
- Should have the correct size dependence, if related to the molecule size

Laying the foundation

- A.F.A. Cros (University of Strasbourg; 1863)
  - Increased toxicity of alcohols with decrease in water solubility
- H. H. Meyer (University of Marburg; 1890's) and Charles Ernest Overton (University of Zurich; 1890's) [working independently]
  - Toxicity of organic compounds depended on their lipophilicity
- Crum-Brown and Fraser
  - the physiological action of a substance was a function of its chemical composition and constitution
- Richet
  - inverse relationship between the cytotoxicity of a diverse set of simple organic molecules with water solubility.

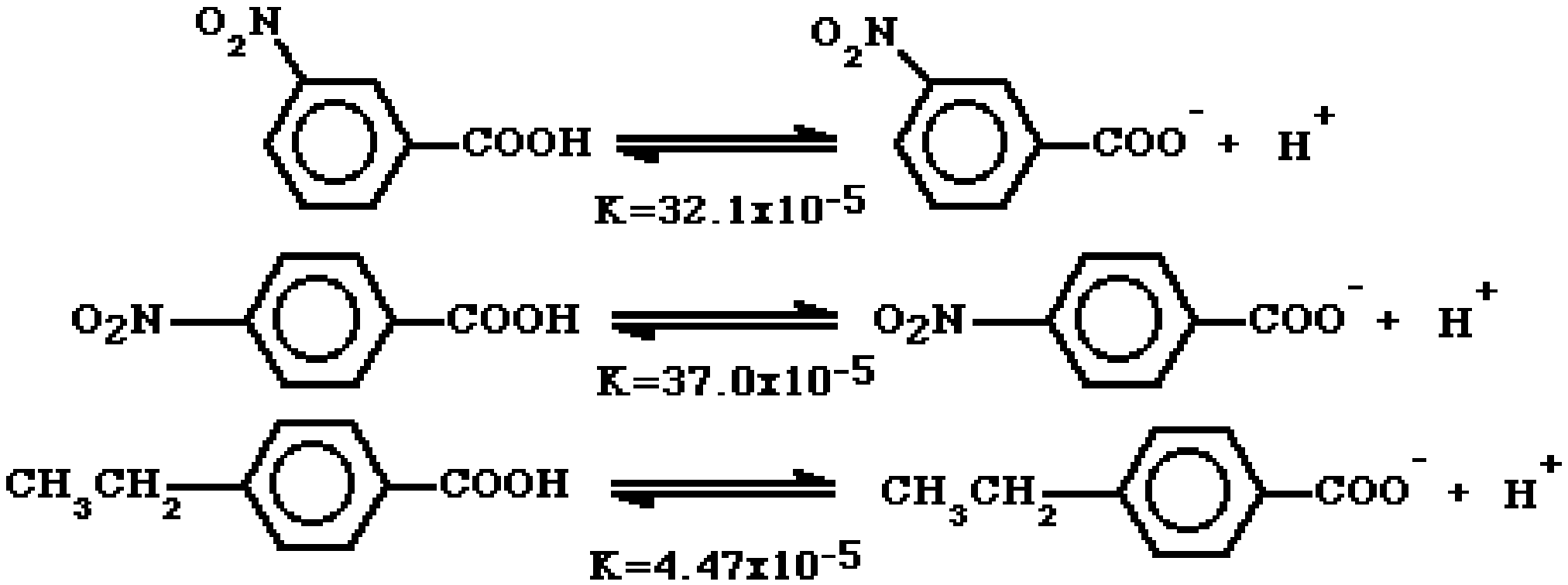
Laying the foundation

- Hammett
  - "sigma-rho" culture; to understand the effect of substituents on organic reactions
- Taft
  - devised a way to separate polar, steric, and resonance effects and introduced the first steric parameter, Es
- Hansch and Fujita
  - The contributions of Hammett and Taft together laid the mechanistic basis for the development of the QSAR paradigm.

Hammett Constant

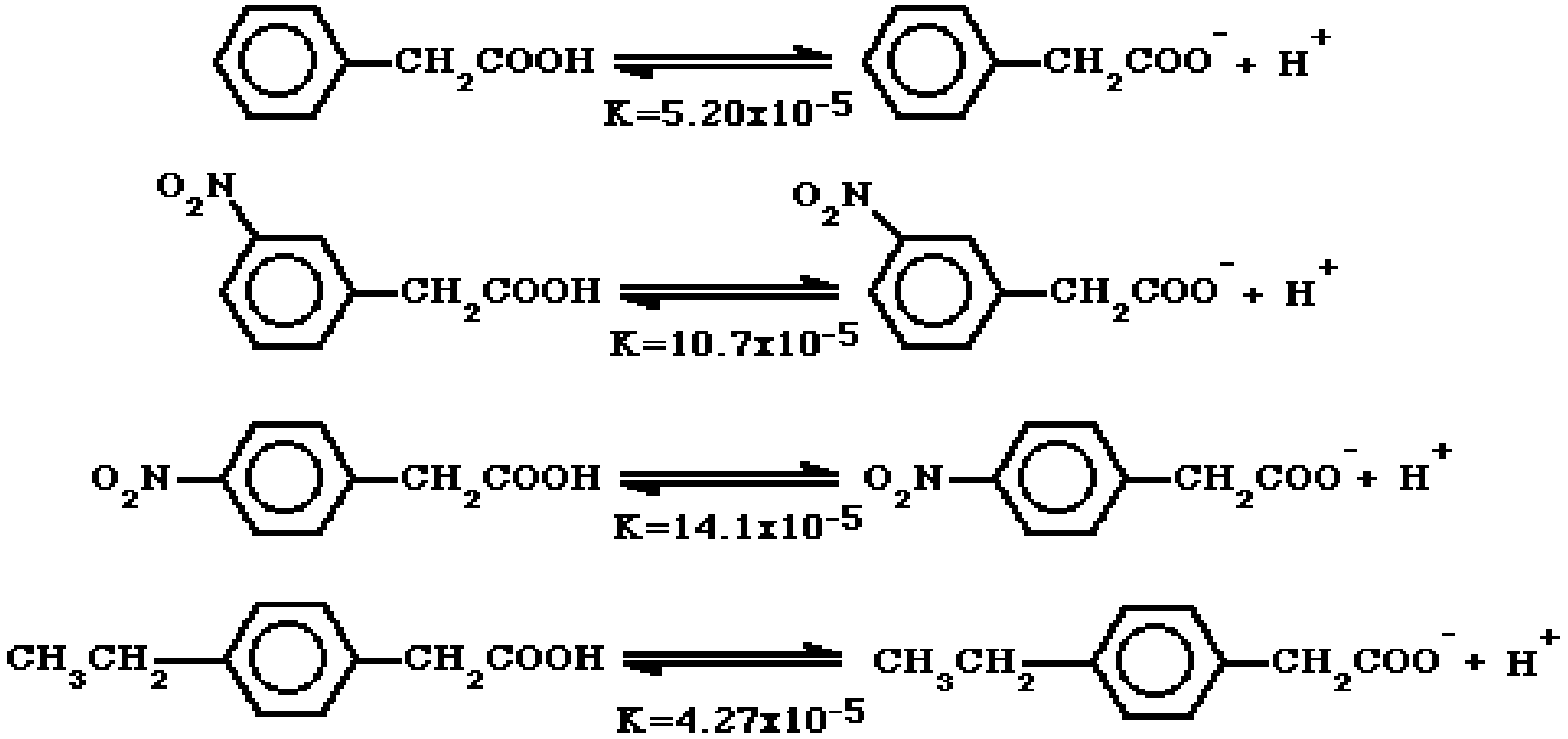
- Linear Free Energy Relationships
  - **Louis Hammett (1894-1987),** correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity
  - Measures the electron withdrawing or electron donating effects in comparison to benzoic acid & how affected its ionization)
- *Consider the dissociation of benzoic acid:*

Hammett Constant



Hammett Constant

Hammett observed similar substituent effects on the organic acids and bases dissociation like phenyl acetic acid.

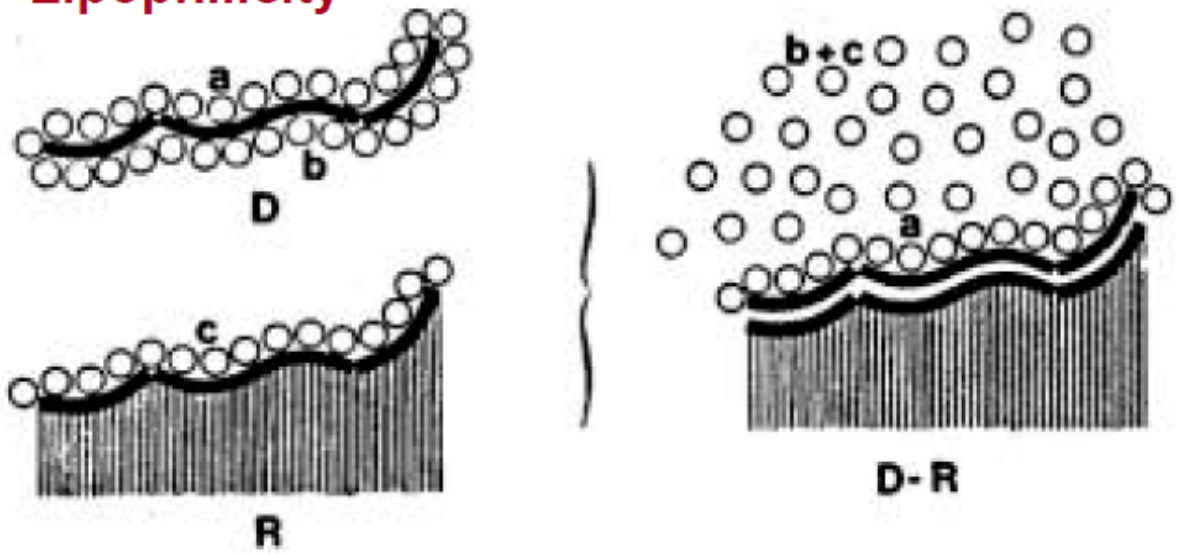


Lipophilicity Effects

- Just as the Hammett equation relates the electronic effects of substituents to reactions rates, Hansch believed that a linear free-energy relationship should exist for lipophilicity and biological activity
- Hansch suggested that the drug in the aqueous phase surrounding the cell made a random walk through the cell membrane (next slide), which is lipophilic, to interact with a particular site in the cell, the rate of which is dependent on the structure of the drug.

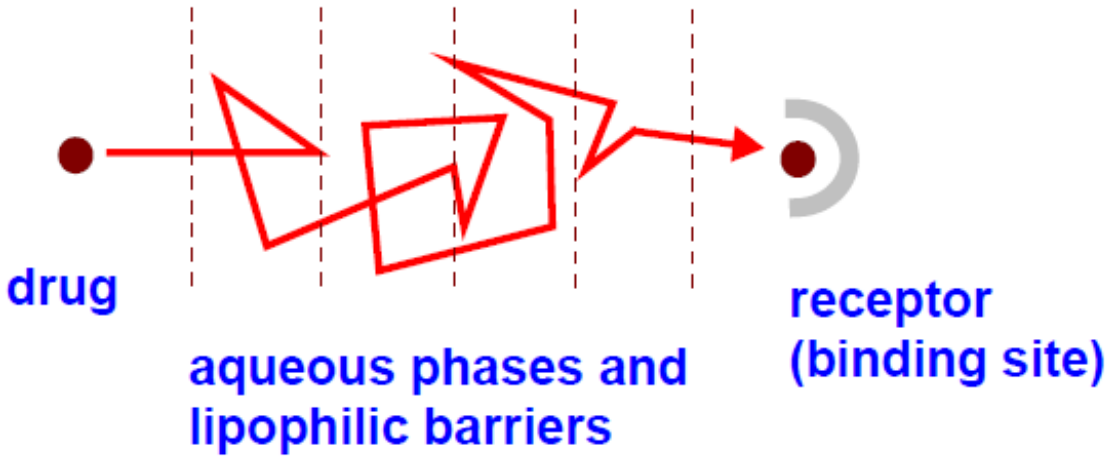
Lipophilicity and the random walk

Lipophilicity



Hydrophobic interaction between a drug and a binding site at a receptor

The "random walk" process



Calculation of Lipophilicity

As a measure of lipophilicity, Hansch proposed the partition coefficient, *P* as

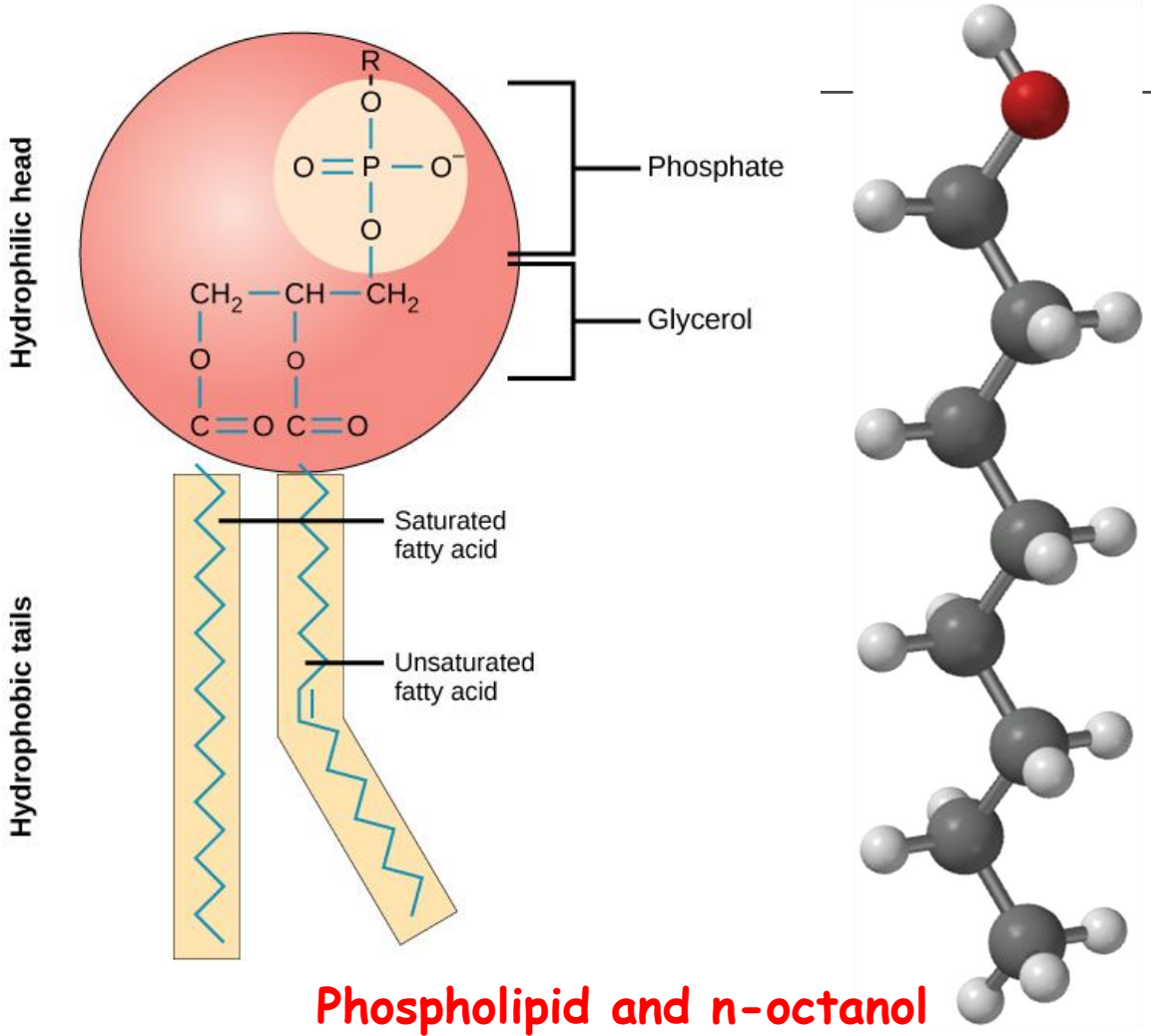
$$P = [\text{compound}]_{oct}/[\text{compound}]_{aq}\ (1 - \alpha)$$

where α is the degree of dissociation of the compound in water calculated from the ionization constant

P < 1, means that the compound is more soluble in water,

if a compound is more soluble in octanol, then *P* >1

Why n-Octanol is used?



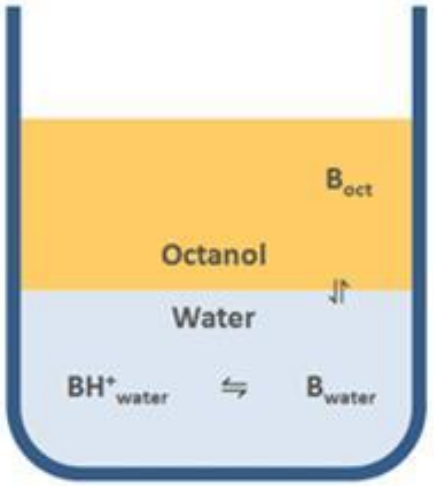
Phospholipid and n-octanol

N-Octonal/water as a standard system

- Membrane analogous structure
- Hydrogen bond donor and acceptor
- Practically insoluble in water
- No desolvation on transfer into organic phase
- Very low vapor pressure
- Transparent in the UV region
- Enormously large database of Log P values

LogP vs LogD

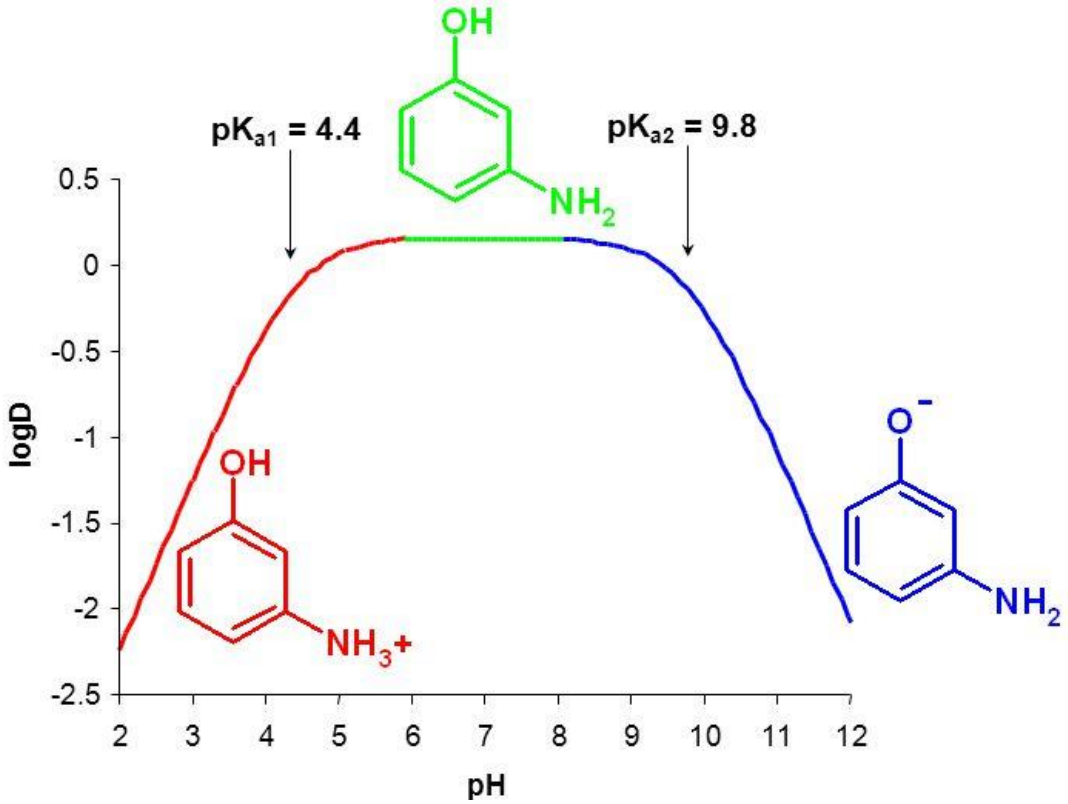
$$pK_a = \log \frac{[BH^+_{water}]}{[H^+][B_{water}]}$$



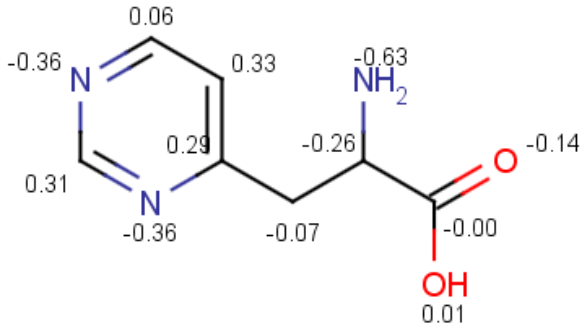
$$\log P = \log \frac{[B_{oct}]}{[B_{water}]}$$

$$\text{P = Partition Coefficient} = \frac{\text{Concentration of neutral species dissolved in partition solvent}}{\text{Concentration of neutral species dissolved in water}}$$

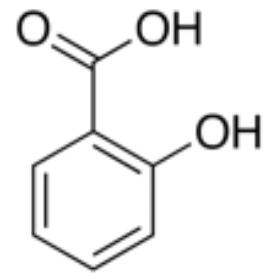
$$\text{D = Distribution Coefficient} = \frac{\text{Concentration of all species dissolved in partition solvent}}{\text{Concentration of all species dissolved in water}}$$



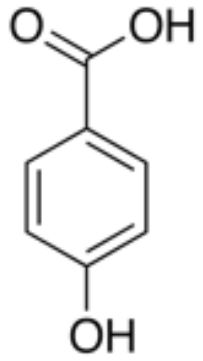
Calculation of Log P



- Atom-based –
  - Atom contribution
  - Eg: Alog P, Xlog P or Mlog P
- Fragment-based –
  - Group Contribution
- Knowledge-based-
  - By data mining approach- SVMs, Decision Trees or Neural Network

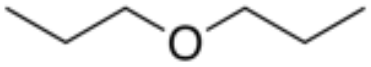


cLogP: 2.19

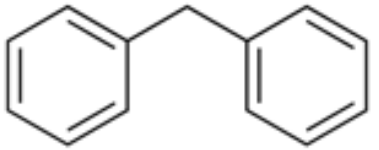


cLogP: 1.56

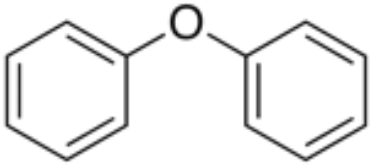
cLogP: 4.397



cLogP: 1.928



cLogP: 4.209



cLogP: 4.24

Calculation Log P

- Atom-based

$$\log P = \sum_i a_i f_i + \sum_j b_j f_j$$

SMILES: BrCc1ccccc1
ATOM #: 1.23.45678.
ISOC-ID: ..Za.aaaaa.
FRAG-ID: 1.....
H-COUNT: ..2..11111.
RING 1: ...a.aaaaa.

benzylbromide 2.3.1.

Class	Type	Log(P) Contribution Description	Comment	Value
Fragment	# 1	Bromide [Z]	Measured	0.480
Carbon		1 aliphatic isolating carbon	-	0.195
Carbon		6 aromatic isolating carbons	-	0.780
ExFragment	Hydrog	7 hydrogens on isolating carbons	-	1.589
ExFragment	Bonds	1 chain and 0 alicyclic (net)	-	-0.120
				2.924

logPstar = 2.92

Calculation/prediction of Log P/log D

Computational Methods:

- From Log S (solubility)-

$$\log P = \log S_o - \log S_w$$

$$= \log\left(\frac{S_o}{S_w}\right)$$

- Log D from Log P and pK_a

$$\log D \cong \log P$$ >>> For unionised molecules

$$\log D \cong \log P + \log (f^0)$$ >>> at a give pH. f^0 is mole fraction of the un-ionized form

$$\log D_{acids} \cong log\ P + pK_a - pH$$ >>> When pH – pK_a > 1 for acids

$$\log D_{acids} \cong log\ P + pK_a + pH$$ >>> When pK_a – pH > 1 for bases

Substituent constants π

- Just as substituent constants were derived by Hammett for the electronic effects of atoms and groups (**σ constants**),
- **Hansch** derived substituent constants for the contribution of individual atoms and groups to the partition coefficient.

Substituent Constant π

The lipophilicity substituent constant π for group X is given by

$$\pi_X = \log P_x - \log P_H = \log P_x/P_H$$

P_x is the partition coefficient of the compound with substituent X and P_H is for the parent compound (X = H)

This means $\pi_H = 0$

π is both

- additive (multiple substituents exert an influence equal to the sum of the individual substituents)
- constitutive (effect of a substituent may differ depending on the molecule to which it is attached)

Substituent constants π

J. B. Houston et al., J. Pharmacol. Exp. Ther. **189**, 244 (1974)

R-OCONH$_2$	P	log P	Δ log P = π_{CH_2}
Methyl	0.22	-0.66	
Ethyl	0.70	-0.15	} 0.51
Propyl	2.3	0.36	} 0.51
Butyl	7.1	0.85	} 0.49
Pentyl	22.5	1.35	} 0.50
Hexyl	70.8	1.85	} 0.50
Heptyl	230	2.36	} 0.51
Octyl	700	2.85	} 0.49
sec-Butyl	4.5	0.65	-0.20 *)
tert-Butyl	3.0	0.48	-0.37 *)

*) relative to *n*-butyl carbamate

π for some substituents

Substituent	CH_3	t-Bu	OH	OCH_3	CF_3	Cl	Br	F
π (aliphatic substituents)	0.50	1.68	-1.16	0.47	1.07	0.39	0.60	-0.17
π (aromatic substituents)	0.52	1.68	-0.67	-0.02	1.16	0.71	0.86	0.14

- Note that these sets of π values are not true constants and they are accurate only for the structures from which they have been derived.
- They are good approximations! for other structures, but remember that some adjustments have to be made to get accurate values

Steric Effects: Taft's equation

Since interaction of a drug with its receptor brings together two molecules, steric effects come into play.

Taft derived the steric parameter E_s as

$$E_s = \log k_{XCOOMe} - \log k_{CH_3COOMe} = \log k_x/k_o$$

The reference reaction is the acid catalyzed hydrolysis of α-substituted acetates (XCH_2COOMe).

This parameter is normally standardized for the methyl group ($X = CH_3$)

k_x represents the rate of hydrolysis of an aliphatic ester bearing the substituent X and k_o represents the rate of hydrolysis of the reference ester.

Taft's steric constants for some groups

Substituent	H	F	Me	Et	*n*-Pr	*n*-Bu	*i*-Pr	*i*-Bu	cyclopentyl
E_s	1.24	0.78	**0.0**	-0.07	-0.36	-0.39	-0.47	-0.93	-0.51

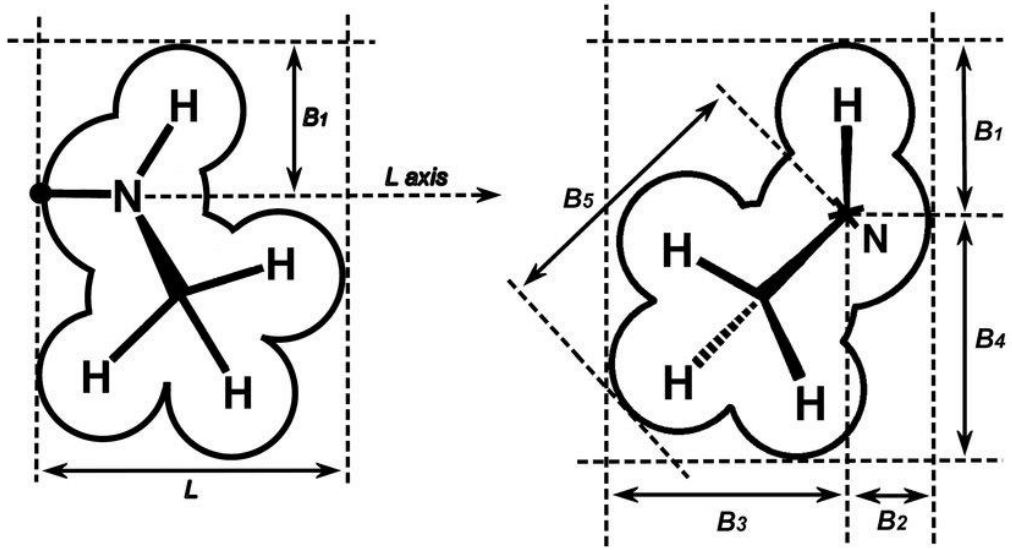
Note that the reference ester is X = Me, therefore E_s is 0.

Substituents such as H and F which are smaller than the Me group, result in a faster rate of hydrolysis $k_x > k_o$, making E_s positive

Substituents larger than Me reduce the rate of hydrolysis $k_x < k_o$ making E_s negative

STERIMOL Parameters

- A multiparametric method for characterizing the steric features of the substituents in more complex biological systems. (By Verloop)
- STERIMOL parameters are a set of five descriptors (L, B1, B2, B3, and B4)
  - L is the length of the substituent along the axis of the bond between the first atom of the substituent and the parent molecule.
  - width parameters B1-B4 are all orthogonal to L and form angles of 90 to each other



List of substituent constant parameters for common aromatic substituents

Substituent	π	MR	σ_m	σ_p	*L*	*B1*	*B2*	*B3*	*B4*	*B5*
H	0.00	0.103	0.00	0.00	2.06	1.00	1.00	1.00	1.00	1.00
CH_3	0.56	0.565	−0.07	−0.17	3.00	1.52	2.04	1.90	1.90	2.04
CH_2CH_3	1.02	1.030	−0.07	−0.15	4.11	1.52	2.97	1.90	1.90	3.17
CH_2OH	−1.03	0.719	0.00	0.00	3.97	1.52	2.70	1.90	1.90	2.70
CH_2CN	−0.57	1.011	0.16	0.01	3.99	1.52	4.12	1.90	1.90	4.12
CH_2Cl	0.17	1.049	0.11	0.12	3.89	1.52	3.46	1.90	1.90	3.46
CH_2Br	0.79	1.339	0.12	0.14	4.09	1.52	3.75	1.95	1.95	3.75
CH_2I	1.50	1.886	0.10	0.11	4.36	1.52	4.15	2.15	2.15	4.15
$CH_2C_6H_5$	2.01	3.001	−0.08	−0.09	3.63	1.52	6.02	3.11	3.11	6.02
$CH(CH_3)_2$	1.53	1.496	−0.07	−0.15	4.11	2.04	2.76	3.16	3.16	3.17
n-C_3H_7	1.55	1.496	−0.07	−0.13	5.05	1.52	3.49	1.90	1.90	3.49
n-C_4H_9	2.13	1.969	−0.08	−0.16	6.17	1.52	4.42	1.90	1.90	4.54
C_5H_{11}	2.67	2.426	−0.08	−0.16	7.11	1.52	4.94	1.90	1.90	4.94
C_6H_5	1.96	2.536	0.06	−0.01	6.28	1.70	1.70	3.11	3.11	3.11
$COCH_3$	−0.55	1.118	0.38	0.50	4.06	1.90	1.90	2.36	2.93	3.13
$CONH_2$	−1.49	0.981	0.28	0.36	4.06	1.60	1.60	2.42	3.07	3.07
COC_6H_5	1.05	3.033	0.34	0.43	4.57	2.36	5.98	3.11	3.11	5.98
OH	−0.67	0.285	0.12	−0.37	2.74	1.35	1.93	1.35	1.35	1.93