# Quantitative Structure Activity Relationships QSAR and 3D-QSAR

## Introduction

- **Aims**
- **To relate the biological activity of a series of compounds to their physicochemical parameters in a quantitative fashion using a mathematical formula**
- **Requirements**
- **Quantitative measurements for biological and physicochemical properties**

- **Physicochemical Properties**

- **Hydrophobicity of the molecule**
- **Hydrophobicity of substituents**
- **Electronic properties of substituents**
- **Steric properties of substituents**

} **Most common properties studied**

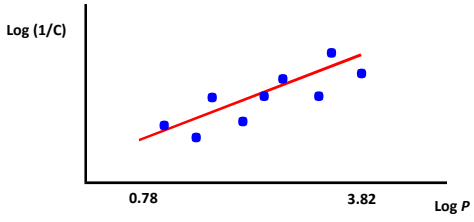
## Hydrophobicity of the Molecule

Partition Coefficient $P$ = $\frac{\text{[Drug in octanol]}}{\text{[Drug in water]}}$

High $P$ ⇒ High hydrophobicity

## Hydrophobicity of the Molecule

- Activity of drugs is often related to $P$
  e.g. binding of drugs to serum albumin
  (straight line - limited range of log $P$)

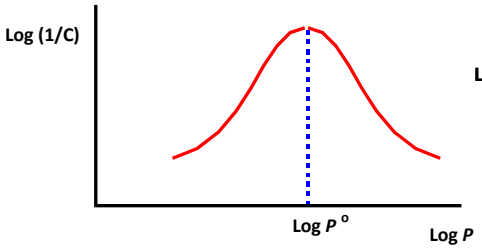



$\text{Log}\left(\frac{1}{C}\right) = 0.75 \log P + 2.30$

- Binding increases as log $P$ increases
- Binding is greater for hydrophobic drugs

# Hydrophobicity of the Molecule

**Example 2** General anaesthetic activity of ethers
(parabolic curve - larger range of log *P* values)




$$\text{Log}\left(\frac{1}{C}\right) = -0.22(\log P)^2 + 1.04 \log P + 2.16$$

Optimum value of log *P* for anaesthetic activity = log $P^o$

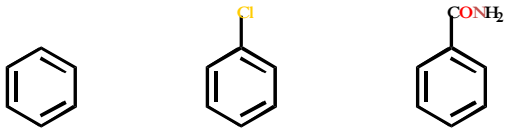
# Hydrophobicity of Substituents

- the substituent hydrophobicity constant ($\pi$)

## Notes:

- A measure of a substituent's hydrophobicity relative to hydrogen
- Tabulated values exist for aliphatic and aromatic substituents
- Measured experimentally by comparison of log *P* values with log P of parent structure

**Example:**




Benzene (Log *P* = 2.13)
Chlorobenzene (Log *P* = 2.84)
Benzamide (Log *P* = 0.64)

$\pi_{Cl} = 0.71$ $\pi_{CONH_2} = -1.49$

- Positive values imply substituents are more hydrophobic than H
- Negative values imply substituents are less hydrophobic than H

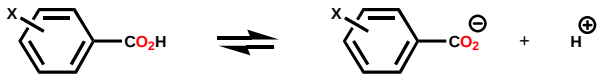
# Electronic Effects

## Hammett Substituent Constant (σ)

**Notes:**

- The constant (σ) is a measure of the e-withdrawing or e-donating influence of substituents
- It can be measured experimentally and tabulated

(e.g. σ for aromatic substituents is measured by comparing the dissociation constants of substituted benzoic acids with benzoic acid)




X=H $K_H$ = Dissociation constant = $\frac{[PhCO_2^-]}{[PhCO_2H]}$

# Steric Factors

## Taft's Steric Factor ($E_s$)

- Measured by comparing the rates of hydrolysis of substituted aliphatic esters against a standard ester under acidic conditions

$E_s = \log k_x - \log k_o$ $k_x$ represents the rate of hydrolysis of a substituted ester

$k_o$ represents the rate of hydrolysis of the parent ester

- Limited to substituents which interact sterically with the tetrahedral transition state for the reaction
- Cannot be used for substituents which interact with the transition state by resonance or hydrogen bonding
- May undervalue the steric effect of groups in an intermolecular process (i.e. a drug binding to a receptor)

## Steric Factors

**Molar Refractivity (*MR*)** - a measure of a substituent's volume

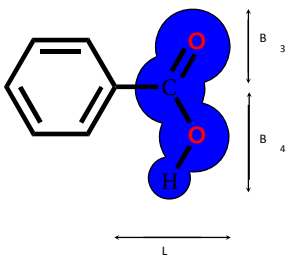
$$MR = \underbrace{\frac{(n^2 - 1)}{(n^2 - 2)}}_{\text{Correction factor for polarisation (n=index of refraction)}} \times \underbrace{\frac{\text{mol. wt.}}{\text{density}}}_{\text{Defines volume}}$$

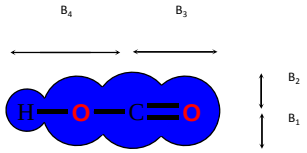
## Steric Factors

**Verloop Steric Parameter**

- calculated by software (STERIMOL)
- gives dimensions of a substituent
- can be used for any substituent

**Example - Carboxylic acid**



## The role of the molecular descriptors

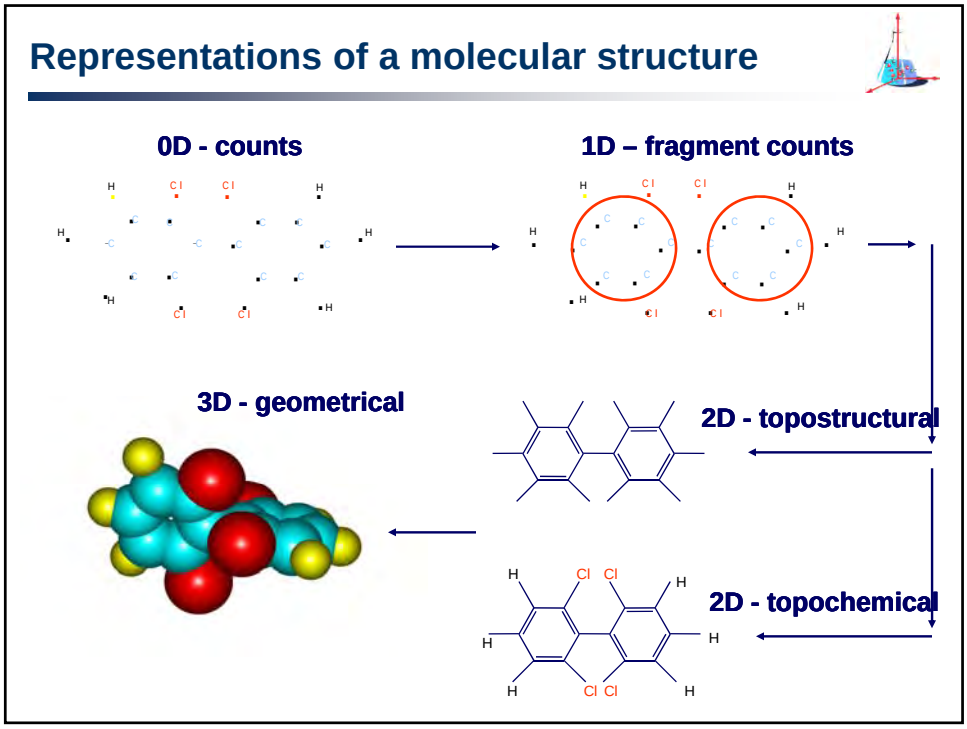
### Environmental properties

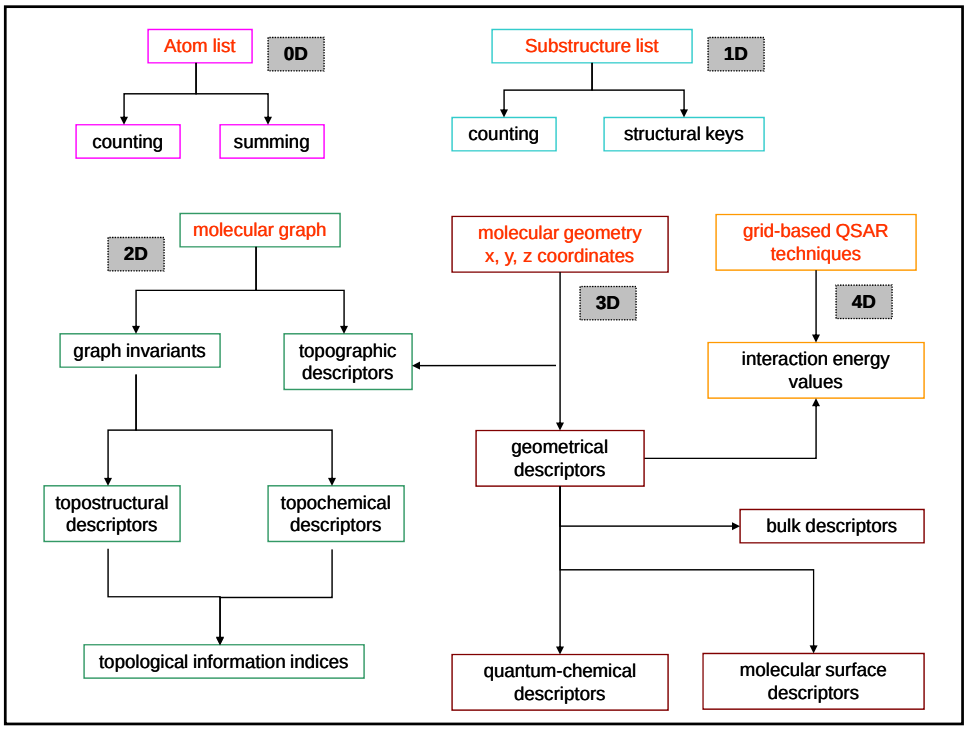
- biodegradation
- bioconcentration
- BOD
- COD
- half - life time
- mobility
- atmospheric persistance
- .....

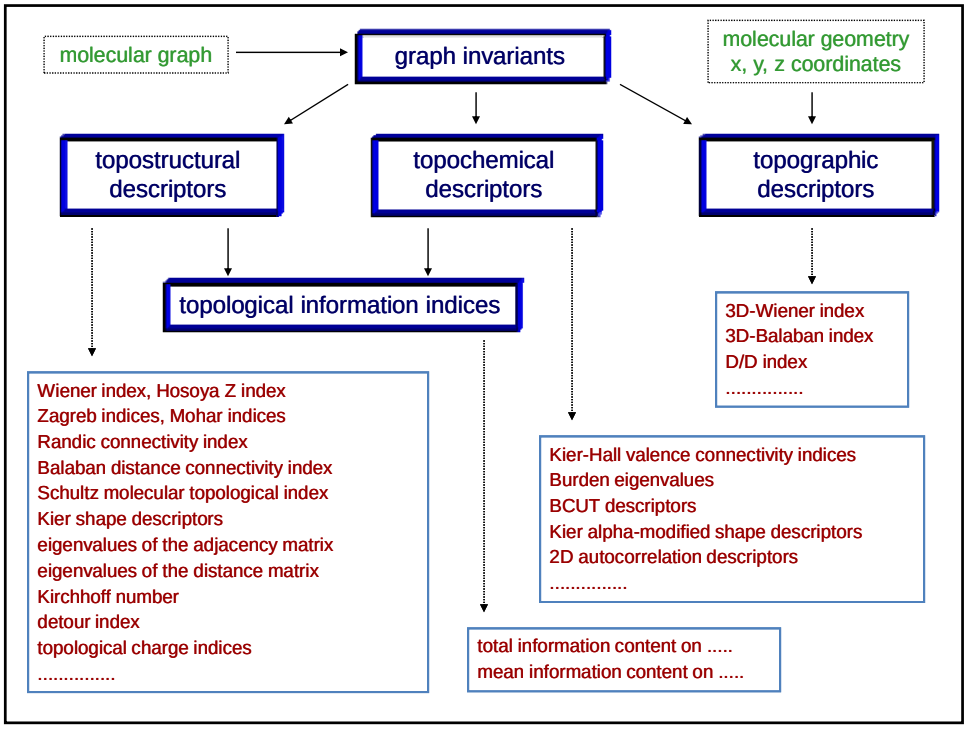
## The role of the molecular descriptors

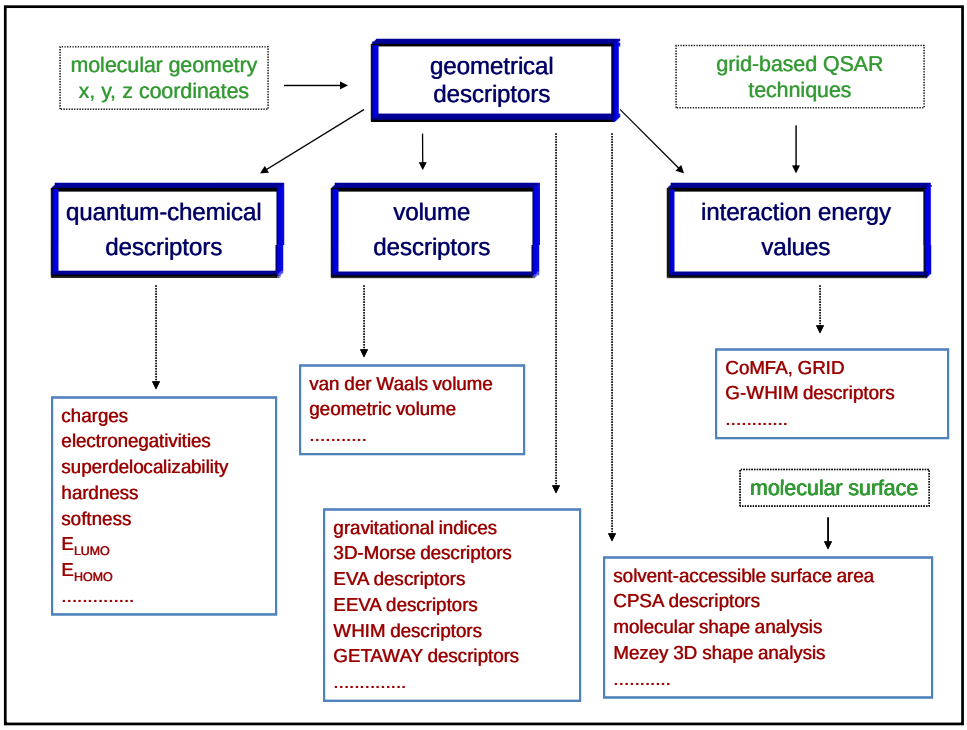
### .... and more

- conductivity
- retention time
- reological behaviours
- .....


Representations of a molecular structure
0D - counts
1D – fragment counts
3D - geometrical
2D - topostructural
2D - topochemical


Atom list
0D
counting
summing
Substructure list
1D
counting
structural keys
2D
molecular graph
graph invariants
topographic descriptors
topostructural descriptors
topochemical descriptors
topological information indices
molecular geometry x, y, z coordinates
3D
geometrical descriptors
bulk descriptors
quantum-chemical descriptors
molecular surface descriptors
grid-based QSAR techniques
4D
interaction energy values

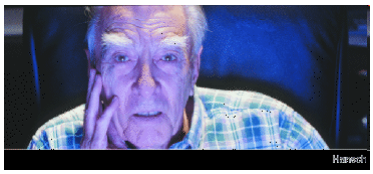

molecular graph
graph invariants
molecular geometry
x, y, z coordinates
topostructural descriptors
topochemical descriptors
topographic descriptors
topological information indices
3D-Wiener index
3D-Balaban index
D/D index
.....
Wiener index, Hosoya Z index
Zagreb indices, Mohar indices
Randic connectivity index
Balaban distance connectivity index
Schultz molecular topological index
Kier shape descriptors
eigenvalues of the adjacency matrix
eigenvalues of the distance matrix
Kirchhoff number
detour index
topological charge indices
.....
Kier-Hall valence connectivity indices
Burden eigenvalues
BCUT descriptors
Kier alpha-modified shape descriptors
2D autocorrelation descriptors
.....
total information content on .....
mean information content on .....


molecular geometry
x, y, z coordinates
geometrical descriptors
grid-based QSAR techniques
quantum-chemical descriptors
volume descriptors
interaction energy values
charges
electronegativities
superdelocalizability
hardness
softness
$E_{LUMO}$
$E_{HOMO}$
.....
van der Waals volume
geometric volume
.....
CoMFA, GRID
G-WHIM descriptors
.....
molecular surface
gravitational indices
3D-Morse descriptors
EVA descriptors
EEVA descriptors
WHIM descriptors
GETAWAY descriptors
.....
solvent-accessible surface area
CPSA descriptors
molecular shape analysis
Mezey 3D shape analysis
.....

## Some historical notes

Based on these descriptors, 90 years later, Corwin Hansch proposed the first QSAR approach.

**Lipophilic, electronic and steric descriptors for ortho-, meta-, and para-substituents**



1964

Corwin HANSCH

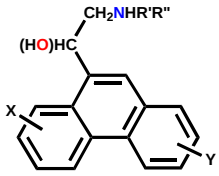
## Hansch Equation

- **A QSAR equation relating various physicochemical properties to the biological activity of a series of compounds**
- **Usually includes log *P*, electronic and steric factors**
- **Start with simple equations and elaborate as more structures are synthesised**
- **Typical equation for a wide range of log *P* is parabolic**

$$\text{Log}\left(1/C\right) = -k_1(\log P)^2 + k_2 \log P + k_3\sigma + k_4 E_s + k_5$$

## Hansch Equation

**Example:** Antimalarial activity of phenanthrene aminocarbinols




$$\text{Log}\left(\frac{1}{C}\right) = -0.015\,(\log P)^2 + 0.14\log P + 0.27\Sigma\pi_X + 0.40\Sigma\pi_Y + 0.65\,\Sigma\sigma_X + 0.88\Sigma\sigma_Y + 2.34$$

**Conclusions:**

- Activity increases slightly as log *P* (hydrophobicity) increases (note that the constant is only 0.14)
- Parabolic equation implies an optimum log $P^o$ value for activity
- Activity increases for hydrophobic substituents (esp. ring Y)
- Activity increases for e-withdrawing substituents (esp. ring Y)

## Hansch Equation

**Choosing suitable substituents**

Substituents must be chosen to satisfy the following criteria;

- A range of values for each physicochemical property studied
- Values must not be correlated for different properties (i.e. they must be orthogonal in value)
- At least 5 structures are required for each parameter studied

| Substituent | H | Me | Et | n-Pr | n-Bu |
|---|---|---|---|---|---|
| π | 0.00 | 0.56 | 1.02 | 1.50 | 2.13 |
| *MR* | 0.10 | 0.56 | 1.03 | 1.55 | 1.96 |

} Correlated values. Are any differences due to π or MR?

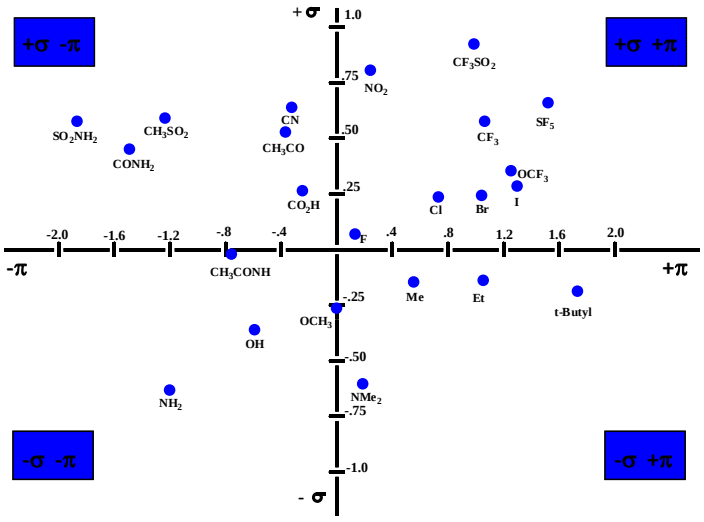
| Substituent | H | Me | OMe | $NHCONH_2$ | I | CN |
|---|---|---|---|---|---|---|
| π | 0.00 | 0.56 | -0.02 | -1.30 | 1.12 | -0.57 |
| *MR* | 0.10 | 0.56 | 0.79 | 1.37 | 1.39 | 0.63 |

} No correlation in values Valid for analysing effects of π and MR.

# Craig Plot

**Craig plot shows values for 2 different physicochemical properties for various substituents**

**Example:**



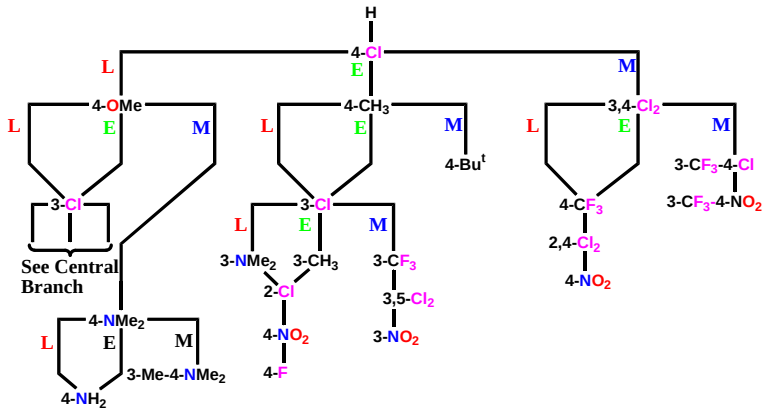

# Craig Plot

- **Allows an easy identification of suitable substituents for a QSAR analysis which includes both relevant properties**
- **Choose a substituent from each quadrant to ensure orthogonality**
- **Choose substituents with a range of values for each property**

# Topliss Scheme

**Used to decide which substituents to use if optimising compounds one by one (where synthesis is complex and slow)**

**Example: Aromatic substituents**




# Topliss Scheme

## Rationale

**Replace H with *para*-Cl (+π and +σ)**

Act. ↑ | Little change | Act. ↓

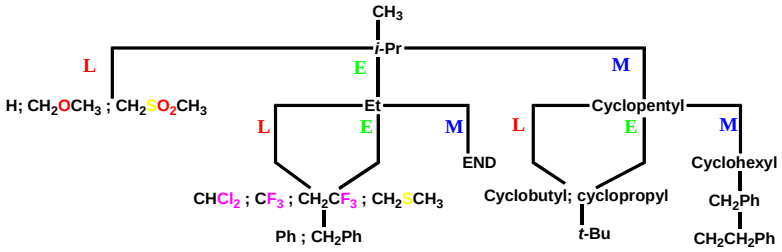
**+π and/or +σ advantageous** → **Add second Cl to increase π and σ further**

**Favourable π unfavourable σ** → **Replace with Me (+π and -σ)**

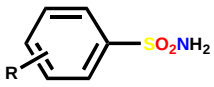
**+π and/or +σ disadvantageous** → **Replace with OMe (-π and -σ)**

**Further changes suggested based on arguments of π, σ and steric strain**

## Topliss Scheme
### Aliphatic substituents



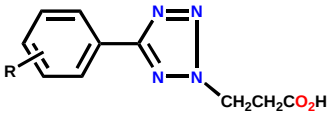

## Topliss Scheme
### Example




| Order of Synthesis | R | Biological Activity | High Potency |
|---|---|---|---|
| 1 | H | - | |
| 2 | 4-Cl | M | |
| 3 | 3,4-$Cl_2$ | L | |
| 4 | 4-Br | E | |
| 5 | 4-$NO_2$ | M | * |

M= More Activity
L= Less Activity
E = Equal Activity

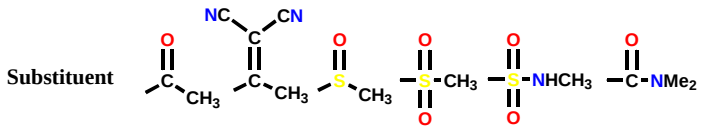
## Topliss Scheme
### Example




| Order of Synthesis | R | Biological Activity | High Potency |
|---|---|---|---|
| 1 | H | - | |
| 2 | 4-Cl | L | |
| 3 | 4-MeO | L | |
| 4 | 3-Cl | M | * |
| 5 | 3-$CF_3$ | L | |
| 6 | 3-Br | M | * |
| 7 | 3-I | L | |
| 8 | 3,5-$Cl_2$ | M | * |

M= More Activity
L= Less Activity
E = Equal Activity

## Bio-isosteres




| | | | | | | |
|---|---|---|---|---|---|---|
| π | -0.55 | 0.40 | -1.58 | -1.63 | -1.82 | -1.51 |
| $\sigma_p$ | 0.50 | 0.84 | 0.49 | 0.72 | 0.57 | 0.36 |
| $\sigma_m$ | 0.38 | 0.66 | 0.52 | 0.60 | 0.46 | 0.35 |
| *MR* | 11.2 | 21.5 | 13.7 | 13.5 | 16.9 | 19.2 |

- Choose substituents with similar physicochemical properties (e.g. CN, $NO_2$ and COMe could be bio-isosteres)
- Choose bio-isosteres based on most important physicochemical property

(e.g. COMe & SOMe are similar in $\sigma_p$; SOMe and $SO_2Me$ are similar in π)

## Free-Wilson Approach

**Method**

- The biological activity of the parent structure is measured and compared with the activity of analogues bearing different substituents
- An equation is derived relating biological activity to the presence or absence of particular substituents

$$\text{Activity} = k_1X_1 + k_2X_2 + \ldots k_nX_n + Z$$

- $X_n$ is an indicator variable which is given the value 0 or 1 depending on whether the substituent (n) is present or not
- The contribution of each substituent (n) to activity is determined by the value of $k_n$
- Z is a constant representing the overall activity of the structures studied

## Free-Wilson Approach

**Advantages**

- No need for physicochemical constants or tables
- Useful for structures with unusual substituents
- Useful for quantifying the biological effects of molecular features that cannot be quantified or tabulated by the Hansch method

**Disadvantages**

- A large number of analogues need to be synthesised to represent each different substituent and each different position of a substituent
- It is difficult to rationalise why specific substituents are good or bad for activity
- The effects of different substituents may not be additive

(e.g. intramolecular interactions)

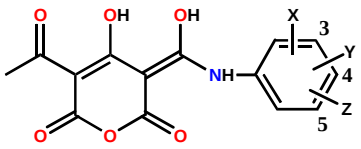
## Free-Wilson / Hansch Approach

**Advantages**

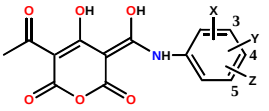
- **It is possible to use indicator variables as part of a Hansch equation - see following Case Study**

## Case Study

**QSAR analysis of pyranenamines (SK & F)**
**(Anti-allergy compounds)**



## Case Study




**Stage 1** 19 structures were synthesised to study π and σ

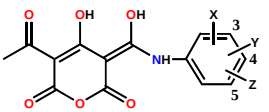
$$\text{Log}\left(1/C\right) = -0.14\Sigma\pi - 1.35(\Sigma\sigma)^2 - 0.72$$

$\Sigma\pi$ and $\Sigma\sigma$ = total values for π and σ for all substituents

**Conclusions:**

- Activity drops as π increases
- Hydrophobic substituents are bad for activity - unusual
- Any value of σ results in a drop in activity
- Substituents should not be e-donating or e-withdrawing (activity falls if σ is +ve or -ve)

## Case Study




**Stage 2** 61 structures were synthesised, concentrating on hydrophilic substituents to test the first equation

**Anomalies**

**a) 3-NHCOMe, 3-NHCOEt, 3-NHCOPr.**
**Activity should drop as alkyl group becomes bigger and more hydrophobic, but the activity is similar for all three substituents**

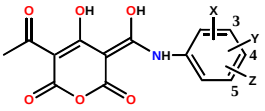
**b) OH, SH, $NH_2$ and NHCOR at position 5 : Activity is greater than expected**

**c) $NHSO_2R$ : Activity is worse than expected**

**d) 3,5-$(CF_3)_2$ and 3,5$(NHMe)_2$ : Activity is greater than expected**

**e) 4-Acyloxy : Activity is 5 x greater than expected**

## Case Study



Theories

**a) 3-NHCOMe, 3-NHCOEt, 3-NHCOPr.**
**Possible steric factor at work. Increasing the size of R may be good for activity and balances out the detrimental effect of increasing hydrophobicity**

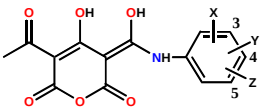
**b) OH, SH, $NH_2$, and NHCOR at position 5**
**Possibly involved in H-bonding**

**c) $NHSO_2R$**
**Exception to H-bonding theory - perhaps bad for steric or electronic reasons**

**d) 3,5-$(CF_3)_2$ and 3,5-$(NHMe)_2$**
**The only disubstituted structures where a substituent at position 5 was electron withdrawing**

**e) 4-Acyloxy**
**Presumably acts as a prodrug allowing easier crossing of cell membranes.**
**The group is hydrolysed once across the membrane.**

## Case Study



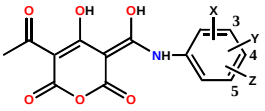
**Stage 3** **Alter the QSAR equation to take account of new results**

$$\text{Log}\left(\frac{1}{C}\right) = -0.30\Sigma\pi - 1.35(\Sigma\sigma)^2 + 2.0(F-5) + 0.39(345\text{-HBD}) - 0.63(NHSO_2) + 0.78(M-V) + 0.72(4\text{-OCO}) - 0.75$$

Conclusions

| | |
|---|---|
| (*F*-5) | Electron-withdrawing group at position 5 increases activity (based on only 2 compounds though) |
| (3,4,5-HBD) | HBD at positions 3, 4,or 5 is good for activity<br>Term = 1 if a HBD group is at any of these positions<br>Term = 2 if HBD groups are at two of these positions<br>Term = 0 if no HBD group is present at these positions<br>Each HBD group increases activity by 0.39 |
| ($NHSO_2$) | Equals 1 if $NHSO_2$ is present (bad for activity by -0.63).<br>Equals zero if group is absent. |
| (*M*-*V*) | Volume of any *meta* substituent. Large substituents at *meta* position increase activity |
| 4-O-CO | Equals 1 if acyloxy group is present (activity increases by 0.72).<br>Equals 0 if group absent |

## Case Study



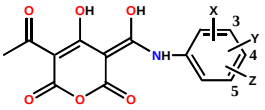

**Stage 3** Alter the QSAR equation to take account of new results

$$\text{Log}\left(\frac{1}{C}\right) = -0.30\Sigma\pi - 1.35(\Sigma\sigma)^2 + 2.0(F-5) + 0.39(345\text{-HBD}) - 0.63(NHSO_2) + 0.78(M-V) + 0.72(4\text{-OCO}) - 0.75$$

**Note**

The terms (3,4,5-HBD), ($NHSO_2$), and 4-O-CO are examples of indicator variables used in the free-Wilson approach and included in a Hansch equation

## Case Study




**Stage 4**

37 Structures were synthesised to test steric and *F*-5 parameters, as well as the effects of hydrophilic, H-bonding groups

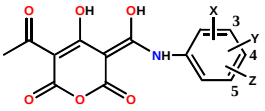
**Anomalies**

Two H-bonding groups are bad if they are *ortho* to each other

**Explanation**

Possibly groups at the *ortho* position bond with each other rather than with the receptor - an intramolecular interaction

## Case Study



**Stage 5** **Revise Equation**

$$\text{Log}\left(\frac{1}{C}\right) = -0.034(\Sigma\pi)^2 - 0.33\Sigma\pi + 4.3(F-5) + 1.3(R-5) - 1.7(\Sigma\sigma)^2 + 0.73(345-\text{HBD}) - 0.86(\text{HB-INTRA}) - 0.69(\text{NHSO}_2) + 0.72(4-\text{OCO}) - 0.59$$

**NOTES**

a) Increasing the hydrophilicity of substituents allows the identification of an optimum value for $\pi$ ($\Sigma\pi = -5$). The equation is now parabolic ($-0.034\ (\Sigma\pi)^2$)

b) The optimum value of $\Sigma\pi$ is very low and implies a hydrophilic binding site

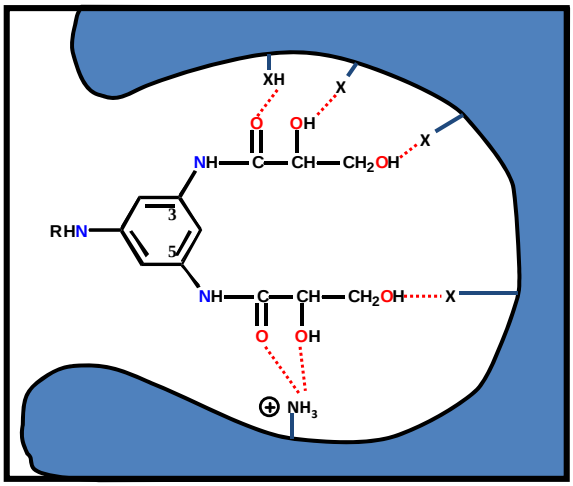
c) *R*-5 implies that resonance effects are important at position 5

d) HB-INTRA equals 1 for H-bonding groups *ortho* to each other (act. drops -086)
equals 0 if H-bonding groups are not *ortho* to each other

e) The steric parameter is no longer significant and is not present

## Case Study

**Stage 6** **Optimum Structure and binding theory**



## Case Study

**NOTES on the optimum structure**

- It has unusual $NHCOCH(OH)CH_2OH$ groups at positions 3 and 5
- It is 1000 times more active than the lead compound
- The substituents at positions 3 and 5
  - are highly polar,
  - are capable of hydrogen bonding,
  - are at the *meta* positions and are not *ortho* to each other
  - allow a favourable *F*-5 parameter for the substituent at position 5
- The structure has a negligible $(\Sigma\sigma)^2$ value

## 3D-QSAR

**Notes**

- Physical properties are measured for the molecule as a whole
- Properties are calculated using computer software
- No experimental constants or measurements are involved
- Properties are known as 'Fields'
- Steric field - defines the size and shape of the molecule
- Electrostatic field - defines electron rich/poor regions of molecule
- Hydrophobic properties are relatively unimportant

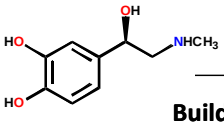
**Advantages over QSAR**

- No reliance on experimental values
- Can be applied to molecules with unusual substituents
- Not restricted to molecules of the same structural class
- Predictive capability

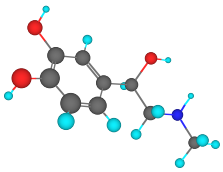
## 3D-QSAR

**Method**

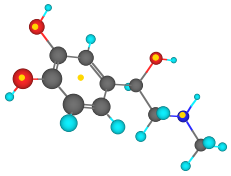
- Comparative molecular field analysis (CoMFA) - Tripos
- Build each molecule using modelling software
- Identify the active conformation for each molecule
- Identify the pharmacophore




Build 3D model



Active conformation

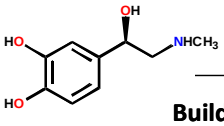


Define pharmacophore

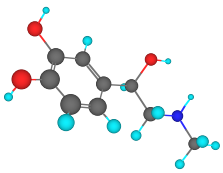
## 3D-QSAR

**Method**

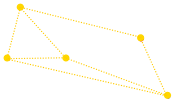
- Comparative molecular field analysis (CoMFA) - Tripos
- Build each molecule using modelling software
- Identify the active conformation for each molecule
- Identify the pharmacophore




Build 3D model



Active conformation

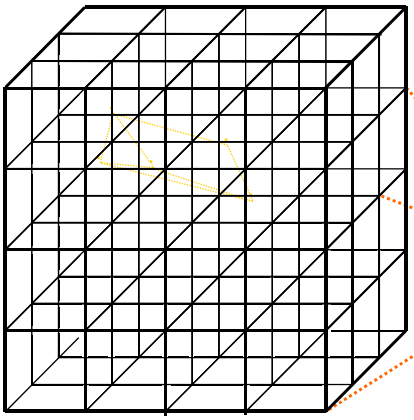


Define pharmacophore

## 3D-QSAR

**Method**

- Place the pharmacophore into a lattice of grid points

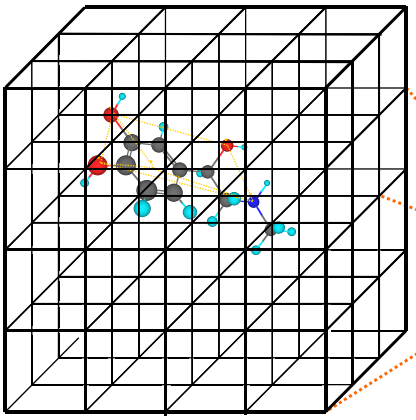



- Each grid point defines a point in space

## 3D-QSAR

**Method**

- Position molecule to match the pharmacophore

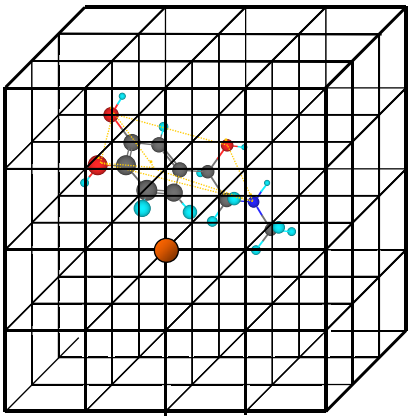



- Each grid point defines a point in space

## 3D-QSAR

**Method**

- A probe atom is placed at each grid point in turn



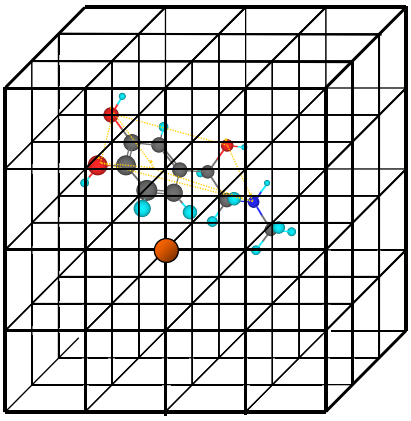
Probe atom

- Probe atom = a proton or $sp^3$ hybridised carbocation

## 3D-QSAR

**Method**

- A probe atom is placed at each grid point in turn



Probe atom

- Measure the steric or electrostatic interaction of the probe atom with the molecule at each grid point

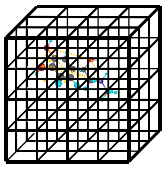
# 3D-QSAR

## Method

- The closer the probe atom to the molecule, the higher the steric energy
- Define the shape of the molecule by identifying grid points of equal steric energy (contour line)
- Favorable electrostatic interactions with the positively charged probe indicate molecular regions which are negative in nature
- Unfavorable electrostatic interactions with the positively charged probe indicate molecular regions which are positive in nature
- Define electrostatic fields by identifying grid points of equal energy (contour line)
- Repeat the procedure for each molecule in turn
- Compare the fields of each molecule with their biological activity
- Identify steric and electrostatic fields which are favorable or unfavorable for activity

# 3D-QSAR

## Method



Tabulate fields for each compound at each grid point

| Compound | Biological activity | Steric fields (S) at grid points (001-998) | | | | | Electrostatic fields (E) at grid points (001-098) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | S001 | S002 | S003 | S004 | S005 etc | E001 | E002 | E003 | E004 | E005 etc |
| 1 | 5.1 | | | | | | | | | | |
| 2 | 6.8 | | | | | | | | | | |
| 3 | 5.3 | | | | | | | | | | |
| 4 | 6.4 | | | | | | | | | | |
| 5 | 6.1 | | | | | | | | | | |

Partial least squares analysis (PLS)

QSAR equation  Activity = aS001 + bS002 +.....mS998 + nE001 +.....+yE998 + z

# 3D-QSAR

## Method

- Define fields using contour maps round a representative molecule