

## Session 9:

# Protein-ligand interaction

Vicent Moliner & Iñaki Tuñón



### Quantitative structure-activity relationships (**QSAR**)

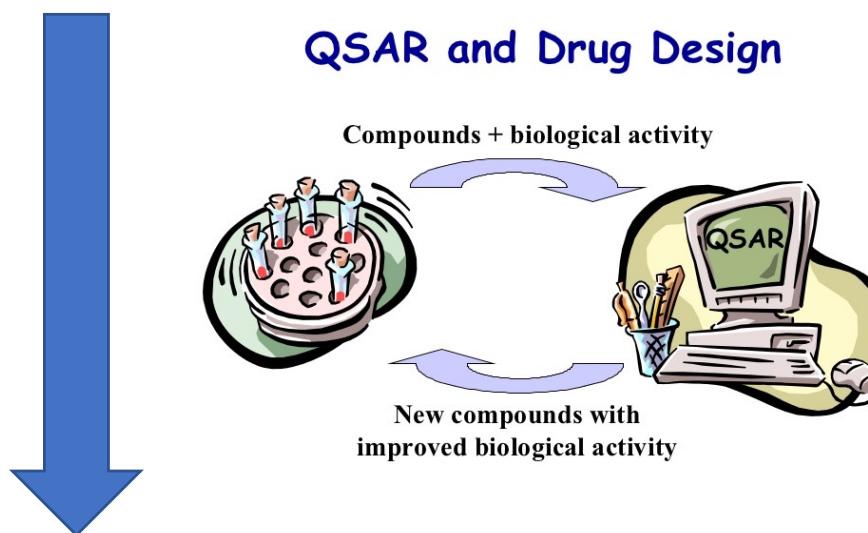
- 9.1 Introduction to QSAR
- 9.2 Chemical Information and Descriptors.
- 9.3 QSAR equations
- 9.4 Validation of a QSAR model
- 9.5 3D QSAR
- 9.6 QSAR Applications
- References

### Docking

- 9.7 Introduction protein-ligand interactions
- 9.8. Algorithms
- 9.9 Scoring Functions
- 9.10 Docking software packages.
- 9.11 Docking Applications
- References

### QSAR (Quantitative Structure Activity Relationship)

**Target:** to determine a mathematical relationship between the **biological activity** or another property (eg  $IC_{50}$ ,  $pK_a$ , ...) and one or more descriptive parameters (**descriptors**) related to the **structure** of the molecule (eg some physicochemical properties or the presence or absence of certain structural features).



to predict the activity of  
proposed new compounds



### Requirements of a QSAR model

Experimental data measured in an appropriate **group of compounds** and with sufficient precision to distinguish between them.

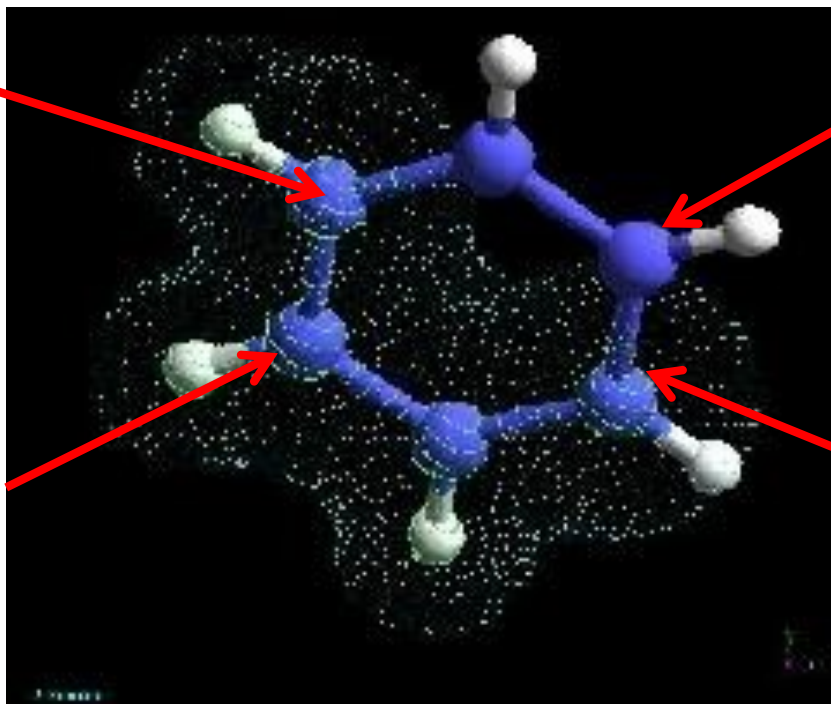
A group of parameters (**descriptors**) that can be easily obtained and that are expected to be related to biological activity.

A method to detect the **relationship** between the **parameters** and the experimental **activity**.

A method to **validate** the QSAR results.



### Why QSAR?



- 4 different positions in the benzene
- To try 10 different substituents



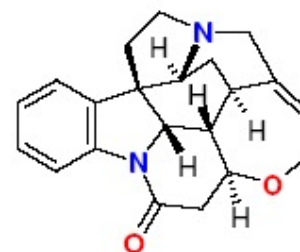
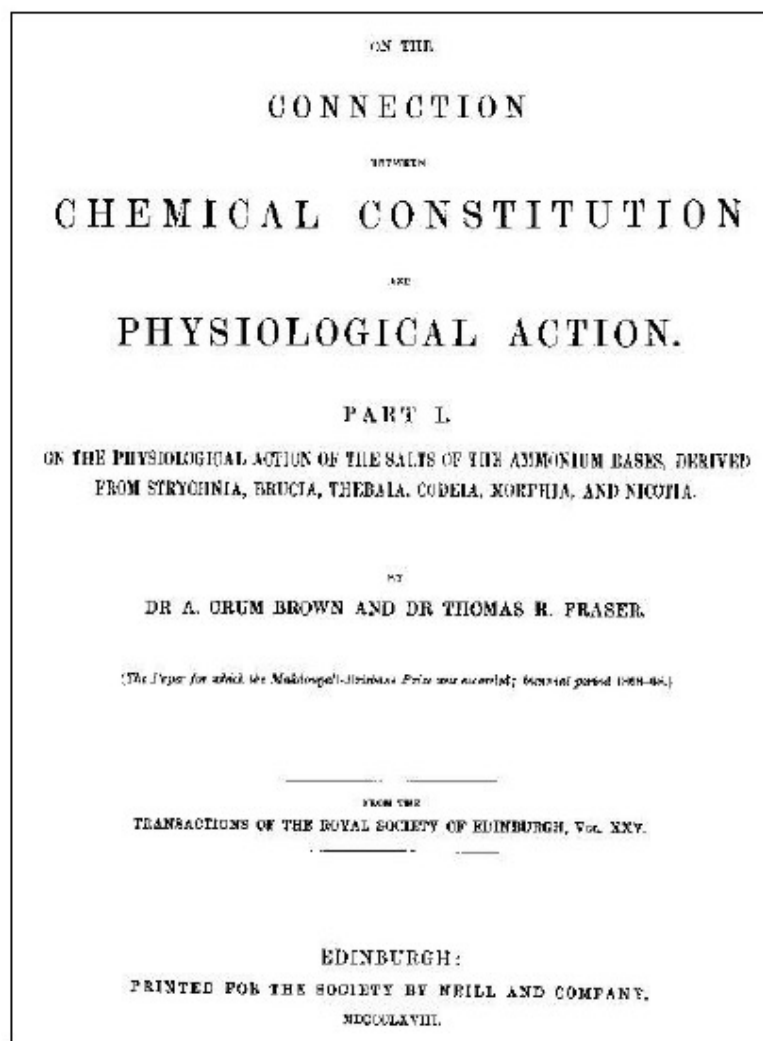
- Number of compounds to synthesize:

**10,000**

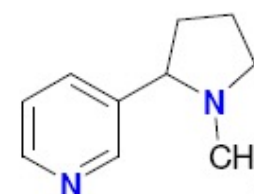
**QSAR:** to synthesize a small number of compounds and, analyzing the data obtained from their physicochemical properties, predict which of those 10,000 compounds may have a certain biological activity.

➤ is reduced to 3 substituents

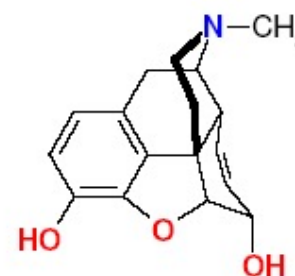
## 1<sup>st</sup> QSAR study:



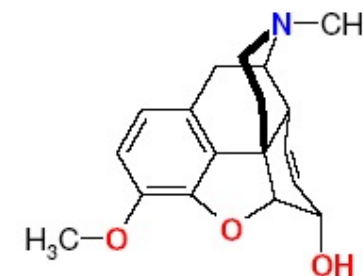
Estrichina



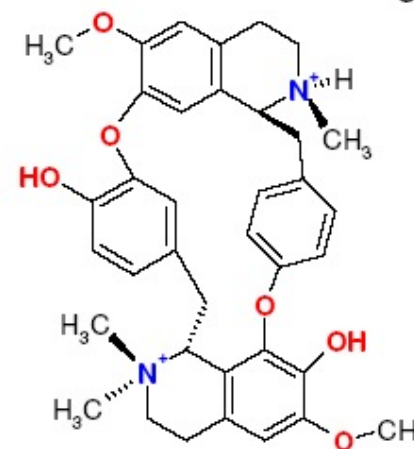
Nicotina



Morfina



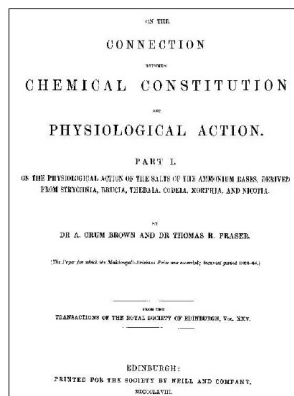
Codeina



d-tubocurarina

Brown A.C. and Fraser. T.R. *J. Anat. Physiol.* 1868, 2, 224-242

# 9.1 Introduction to QSAR



In 1868 Brown and Fraser postulated that there should be a correlation between 'physiological activities' and "chemical structures".

$$\Phi = f(C)$$

a measure of biological activity  
("physiological action")

a measure of chemical structure  
("chemical constitution").

**Problem:** to obtain  $f()$ , due to the difficulty of relating the changes in  $\Phi$  with the changes in  $C$  in a unique way.

➡ In 1868 Richardson shows that the toxicities of ethers and alcohols are inversely proportional to their solubility in water.

J. Richardson, *Medical Times and Gazette*, 1868, **2**, 703

- ➡ In 1893 Richet demonstrated the existence of a relationship between the narcotic effect of alcohols and their molecular weight. Furthermore, he correlated toxicity with aqueous solubility  
*C. Richet, C.R. Seances Soc. Biol., 1893, 9, 775*
  
- ➡ In 1899 Meyer and Overton found a linear relationship between the narcotic potency of organic compounds and their structures, specifically with the water / hydrocarbon partition coefficient.  
*H. Meyer, Arch. Experim. Pathol. und Pharmacol., 1899, 42, 109.*  
*E. Overton, Z Physik. Chem., 1897, 22, 189.*

## 9.1 Introduction to QSAR

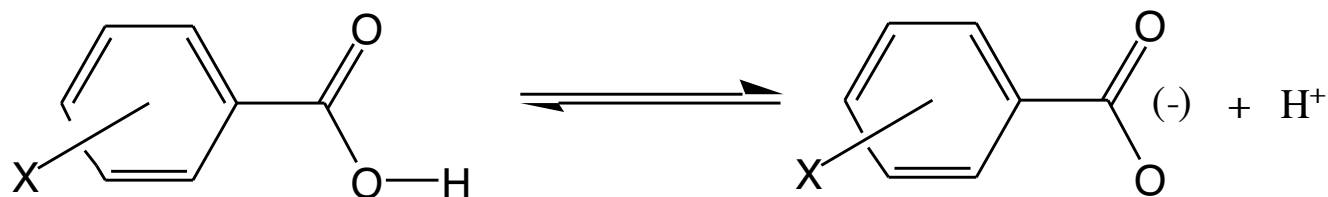


- ➔ In 1940 Hammett defined a relationship between rate (k) and equilibrium (K) constants for the dissociation of benzoic acids

$$\rho\sigma = \log (k_x/k_H) = \log (K_x/K_H)$$

describes the sensitivity of the reaction to substituents.

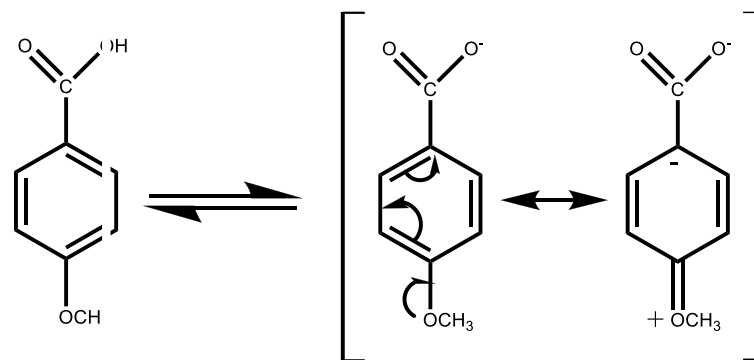
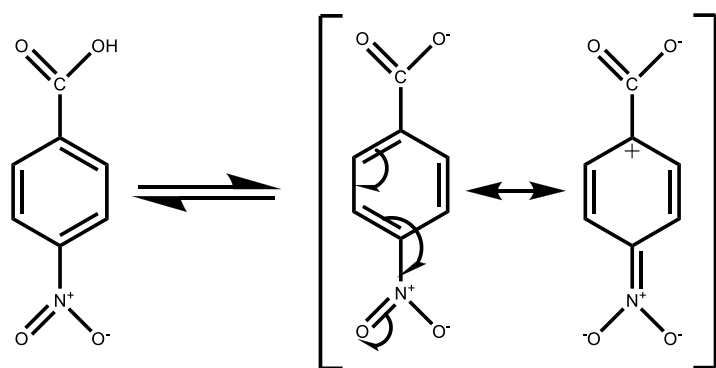
Hammett's constant: describes the electronic properties of the aromatic system substituents (X)



$$\rho\sigma = \log (k_x/k_H) = \log (K_x/K_H)$$

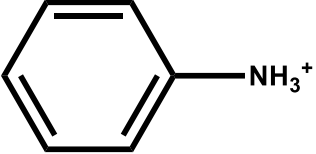
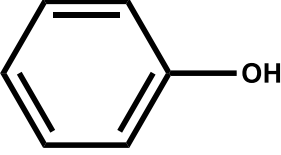
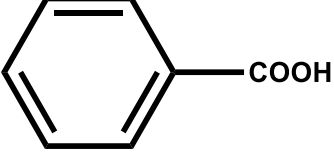
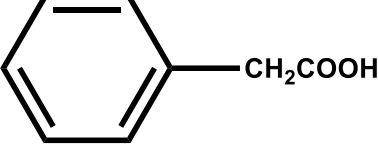
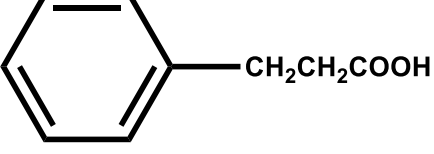
### The meaning of $\sigma$

Sustituyent	$\sigma_{\text{meta}}$	effect	$\sigma_{\text{para}}$	effect
-H	0.0	No	0.0	No
-NO <sub>2</sub>	0.71	Inductive	0.78	Inductive, Resonance
-Cl	0.37	Inductive	0.23	Inductive
-OCH <sub>3</sub>	0.12	Inductive	-0.27	Inductive, Resonance
-CH <sub>3</sub>			-0.13	Inductive



$$\rho\sigma = \log (k_x/k_H) = \log (K_x/K_H)$$

### The meaning of $\rho$

compound	$\rho$	distance H-cicle (in bonds)
	<b>2.90</b>	<b>2</b>
	<b>2.23</b>	<b>2</b>
	<b>1.00</b>	<b>3</b>
	<b>0.49</b>	<b>4</b>
	<b>0.21</b>	<b>5</b>

The further the dissociated H is, the less effect the substituents have

➡ In 1964 Free and Wilson postulated an additive model:

$$\log 1/C = \sum a_i + \mu$$

the contribution of  
substituent i

The activity of the parent  
compound

**C** = the concentration of drug required to achieve a certain level of biological activity. Most active drugs need low concentrations.



- ➔ In 1962 Hansch et al. published a model that correlates biological activity with lipophilicity:

$$\log 1/C = a \log P + b$$

**C** = the concentration of drug required to achieve a certain level of biological activity. Most active drugs need low concentrations.

Partition coefficient n-octanol / water =  $P_x = \frac{[\text{drug X}]_{\text{n-octanol}}}{[\text{drug X}]_{\text{water}}}$



$$\log 1/C = a \log P + b$$

**Log P** it is a measure of the hydrophobicity of the drug, as a measure of the ability to pass through the cell membrane.

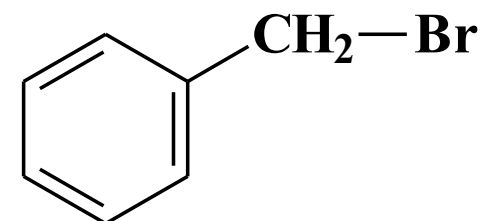
it reflects the relative solubility of the drug in octanol (as representative of the lipid bilayer of the cell membrane) and water (blood fluid and cell interior).

It can be determined experimentally.

It can be calculated by an additive approximation:

1 aromatic ring	0.780
7 H of carbons	1.589
1 C-Br bond	-0.120
1 alkyl carbon	0.195

Sum = 2.924 = calc. log P

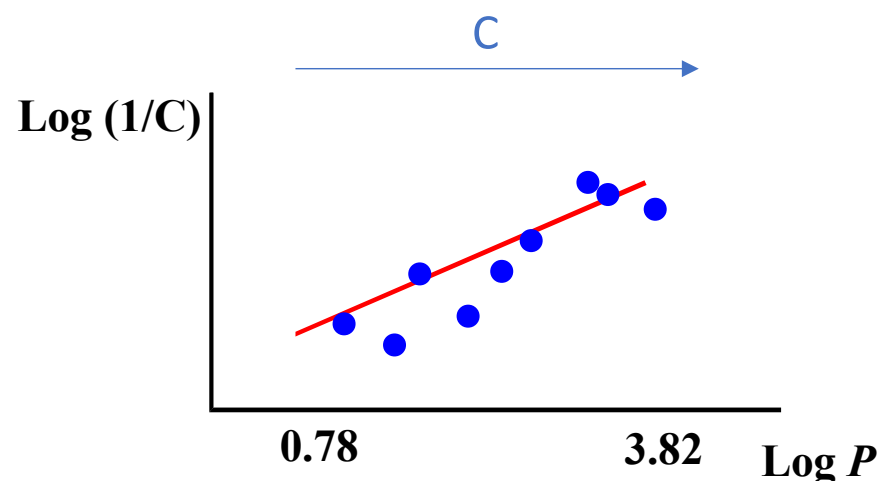


log P > 0 : lipid phase

log P < 0 : aqueous phase

$$\log 1/C = a \log P + b$$

Example: binding of drug to albumin has a linear behavior in a limited range of log P



$$\log 1/C = 0.75 \log P + 2.30$$

The binding increases when logP increases

Higher activity for hydrophobic drugs (lower C required to achieve the desired effect)

$\pi$  : Hydrophobicity parameter

$$\pi_x = \log \frac{P_x}{P_H}$$

C. Hansch, P.P. Maloney, T. Fujita and R.M. Muir, *Nature*, 1962, **194**, 178

$P_H$ : partition coefficient of the drug in a certain solvent (n-octanol)  
with respect to water

$P_x$ : the same but with the drug derivative (X)

$\pi > 0$  lipophilic character with respect to hydrogen

$\pi < 0$  hydrophilicity relative to hydrogen

The 1-octanol / water system is used as a reference. A membrane according to its specific characteristics can be modeled by other solvents:

inert  $\rightarrow$  alkane

amphiprotic  $\rightarrow$  1-octanol

proton donor  $\rightarrow$  chloroform

proton acceptor  $\rightarrow$  propylene glycol dipelargonate



$\pi$  : Hydrophobicity parameter

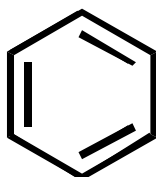
$$\pi_x = \log \frac{P_x}{P_H}$$

C. Hansch, P.P. Maloney, T. Fujita and R.M. Muir, *Nature*, 1962, **194**, 178

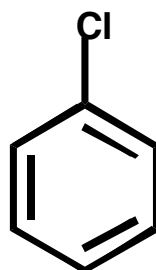
$P_H$ : partition coefficient of the drug in a certain solvent (n-octanol)  
with respect to water

$P_x$ : the same but with the drug derivative (X)

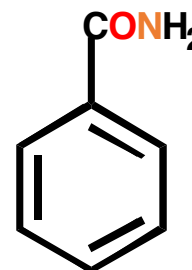
There are tabulated values of different substituents:



**Benzene**  
(Log  $P$  = 2.13)



**Chlorobenzene**  
(Log  $P$  = 2.84)

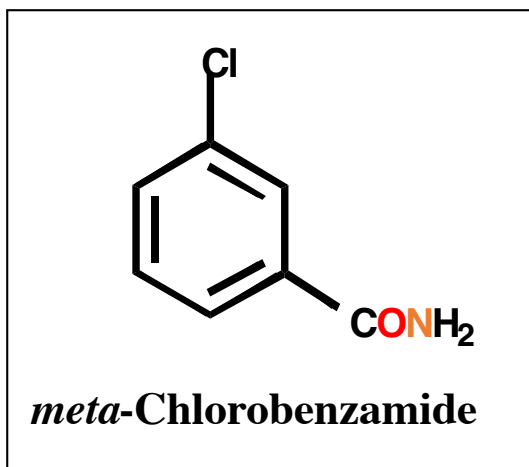


**Benzamide**  
(Log  $P$  = 0.64)

$\pi$  : Hydrophobicity parameter

$$\pi_x = \log \frac{P_x}{P_H}$$

It is possible to calculate log P using tabulated  $\pi$  values, which avoids the work of synthesis and experimental measurement

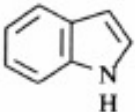

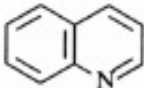

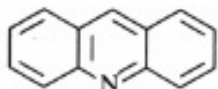
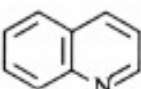
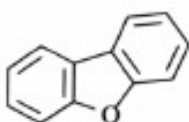
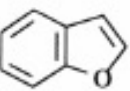
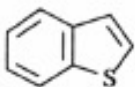

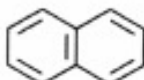


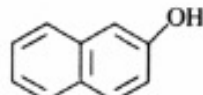
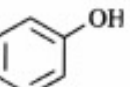


$$\begin{aligned}\text{Log } P_{(\text{theory})} &= \log P_{(\text{benzene})} + \pi_{\text{Cl}} + \pi_{\text{CONH}_2} \\ &= 2.13 + 0.71 - 1.49 \\ &= 1.35\end{aligned}$$

$$\text{Log } P_{(\text{observed})} = 1.51$$

## 9.1 Introduction to QSAR



Difference of Log P				$\pi$	
log P		— log P		=	2.14 — 0.75 = 1.39
log P		— log P		=	2.03 — 0.65 = 1.38
log P		— log P		=	3.40 — 2.03 = 1.37
log P		— log P		=	4.12 — 2.67 = 1.45
log P		— log P		=	3.12 — 1.81 = 1.31
log P		— log P		=	3.45 — 2.13 = 1.32
$2/3 \log P$				=	$2/3 (2.13) = 1.42$
log P		— log P		=	2.84 — 1.46 = <u>1.38</u>
ave. $1.38 \pm 0.046$					

The constant  $\pi$  is additive and depends on the molecular surroundings

The alkyl groups and the conjugated systems are less influencing

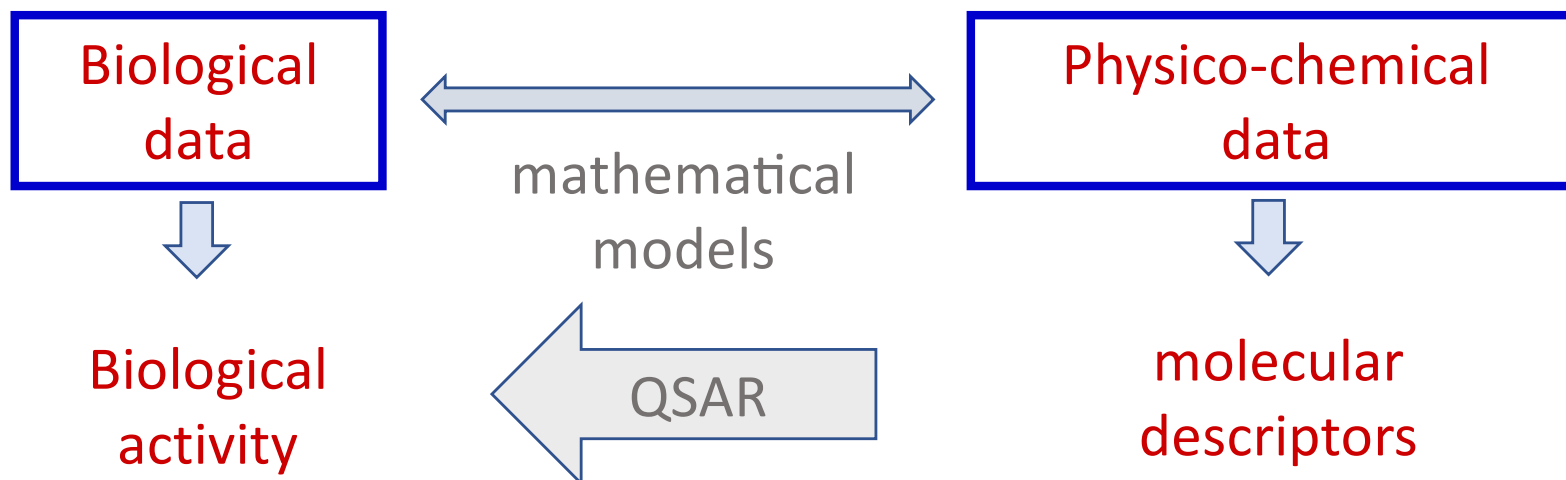
In general, the effects that increase hydrophobicity can be **correlated** with those factors that prevent the formation of acceptor hydrogen bonds with water.

- Electron attractor groups increase  $\pi$  when there are bridges of hydrogen involved by inductive effect
- Resonance effects in aromatic systems decrease the ability to form hydrogen bonds by delocalization of the pairs of free electrons increasing  $\pi$
- Steric effects are variable. If a group interferes with access to pairs of free electrons that can form hydrogen bonds with the water,  $\pi$  increases, but the grouping of hydrophobic groups will have the opposite effect

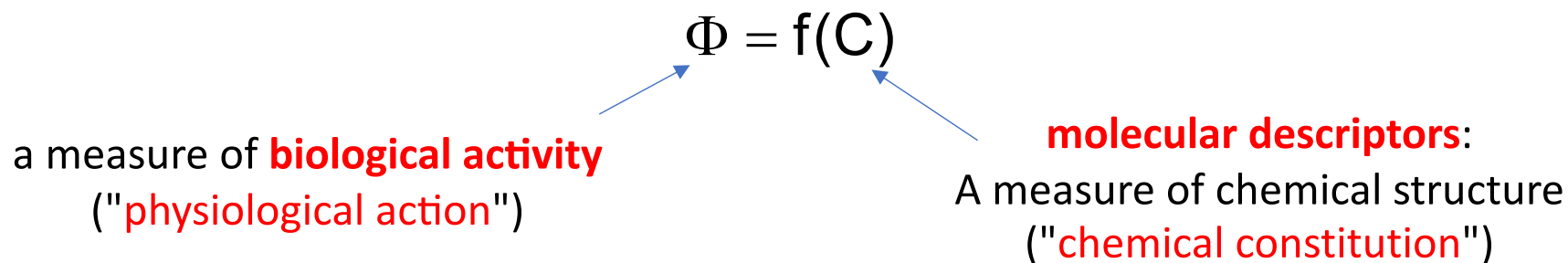
### **The essence of QSAR:**

$$\text{biological activity} = f(\text{physicochemical properties})$$





A **QSAR equation contains**, on the one hand, the **activities** to be determined and, on the other hand, the **molecular descriptors** that are related.



the **most common biological activities** are those that have to do with pharmacological activity:

- affinities to the enzyme's active site
- kinetic constants
- inhibition constants ( $K_i$ )
- $IC_{50}$  values
- pharmacokinetic parameters
- *in vivo* and *in vitro* biological activity values

Mean effective dose ( $ED_{50}$ )

Median lethal dose ( $LD_{50}$ )

Minimum Inhibition Concentration (MIC)

Mean effective concentration ( $EC_{50}$ )

Enzyme kinetics ( $k_{cat}$  /  $k_{uncat}$ )

Affinity to the enzyme's active site ( $K_M$ )



### **Required features of molecular descriptors:**

It must be correlated with the structural characteristics for a specific objective and show negligible correlation with other descriptors

It should be applicable to a wide class of compounds

It is better that it can be calculated quickly and does not depend on experimental properties

It should generate dissimilar values for structurally different molecules, even if the structural differences are small (minimal degeneracy)

It is always important that the descriptor has some form of physical interpretability to encode the query characteristics of the molecules studied.

Ability to assign descriptor values to the structure for visualization purposes



### DESCRIPTORS COMMONLY USED IN QSAR STUDIES

- General descriptors:
  - Melting point
  - Boiling point
  - Vapor pressure
  - Dissociation constants
  - Activation energy
  - Heat of reaction
  - Kinetic constants
  - Reduction potential
- Hydrophobic parameters:
  - Partition coefficient n-octanol / water (P)
  - Chromatographic coefficient (R<sub>m</sub>)
  - Solubility

### DESCRIPTORS COMMONLY USED IN QSAR STUDIES

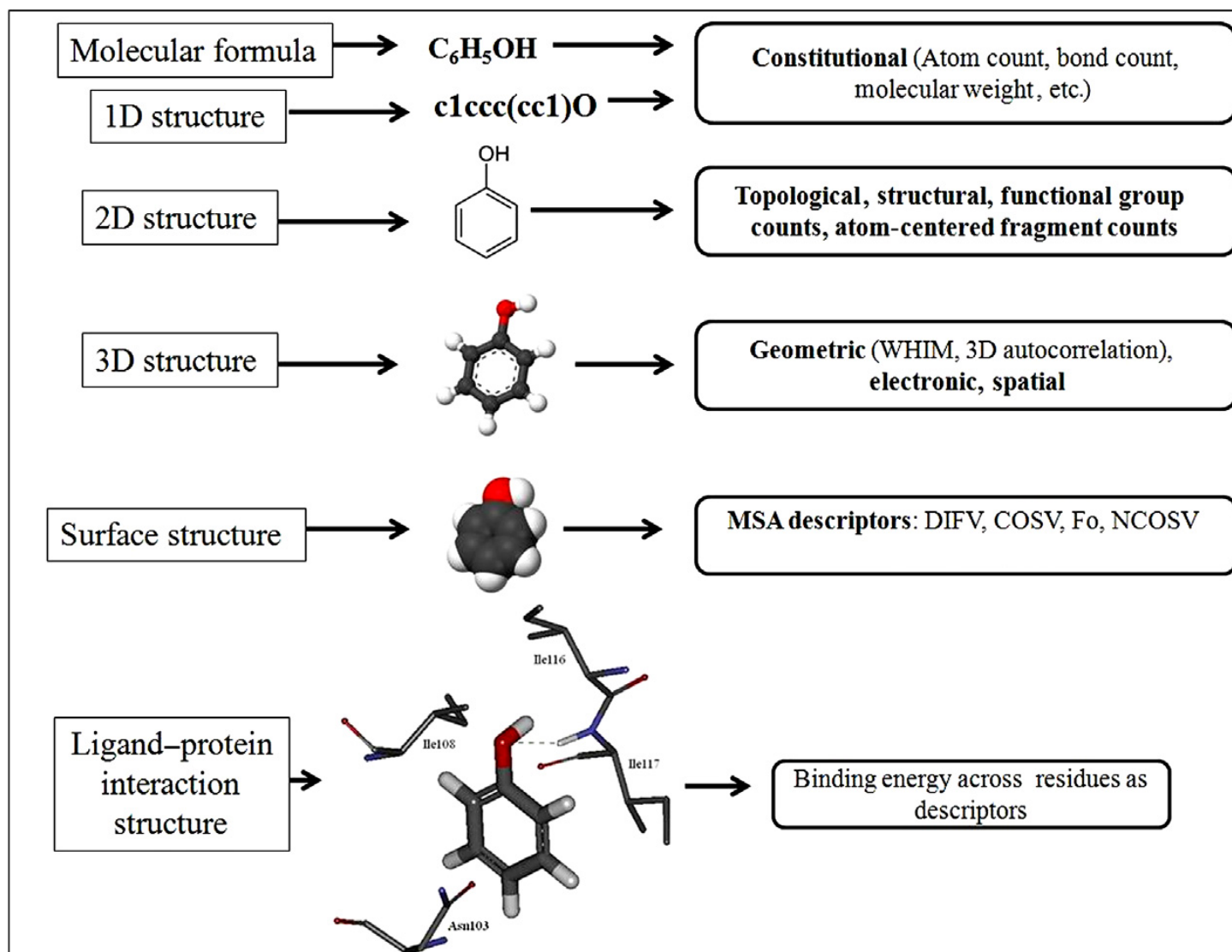
- electronic descriptors:
  - Hammett's constant ( $\sigma$ )
  - Taft constant for polar substituents ( $\sigma^*$ )
  - Field constant (R) and resonance constant (F) :  
$$\sigma = a F + b R$$
  - Ionization potential
  - Dielectric constant
  - Dipole moment
  - Hydrogen bridge bonding.
- Quantum chemical descriptors:
  - Partial charges
  - Electronic densities
  - Bond reactivity  $\pi$
  - Electronic polarizability
  - HOMO and LUMO energy

### DESCRIPTORS COMMONLY USED IN QSAR STUDIES

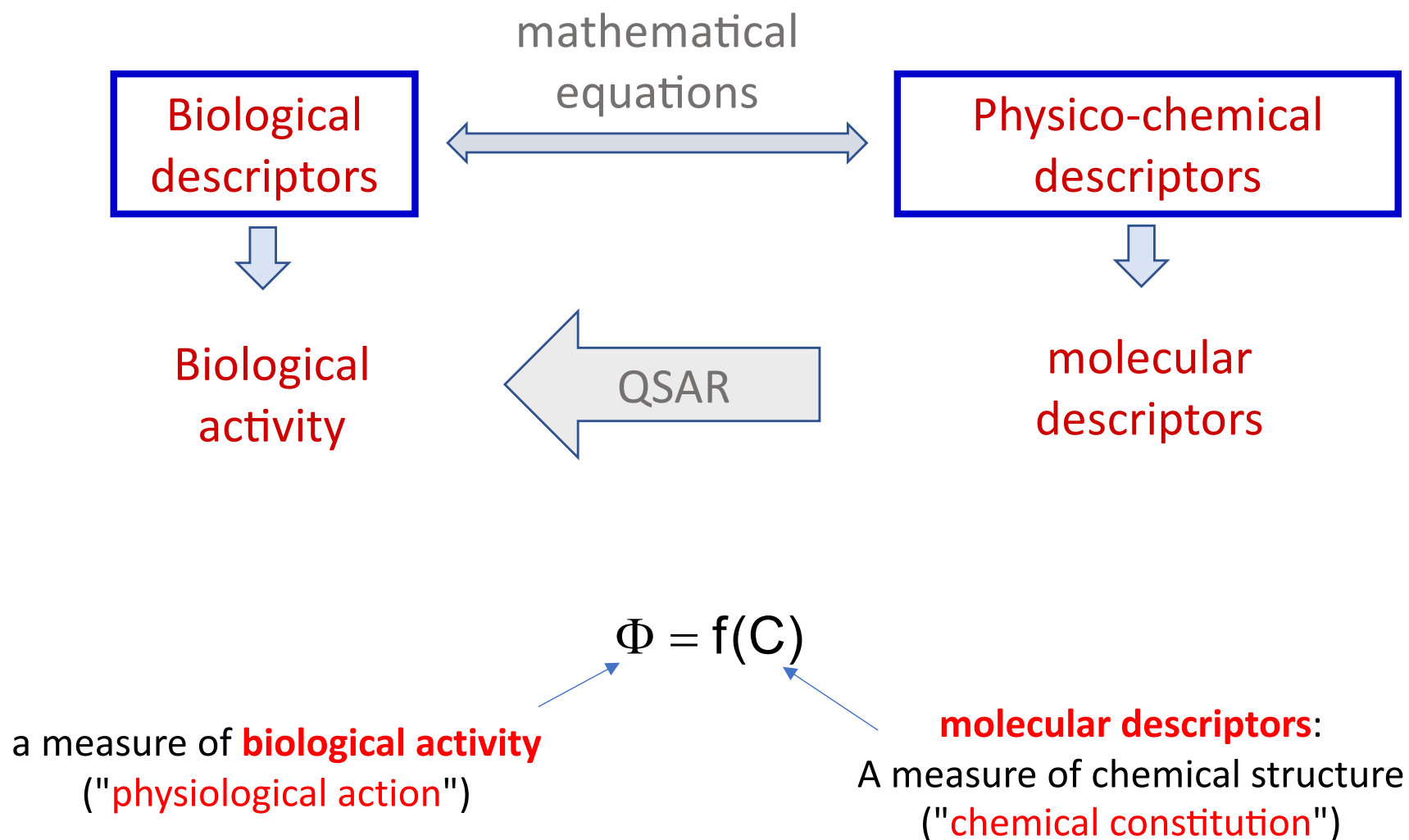
- steric descriptors:
  - Molar refractivity (MR).
  - Taft constant for steric substituents ( $E_s$ )
  - Molecular volume
  - Molecular shape
  - Molecular surface area
  - Van der Waals radii
- structural descriptors:
  - Fragments of atoms and bonds
  - Substructures
  - Substructure environment
  - Molecular weight
  - Number of atoms in a given structural element
  - Number of rings (in polycyclic compounds)
  - Molecular connectivity indices

### DESCRIPTORS COMMONLY USED IN QSAR STUDIES

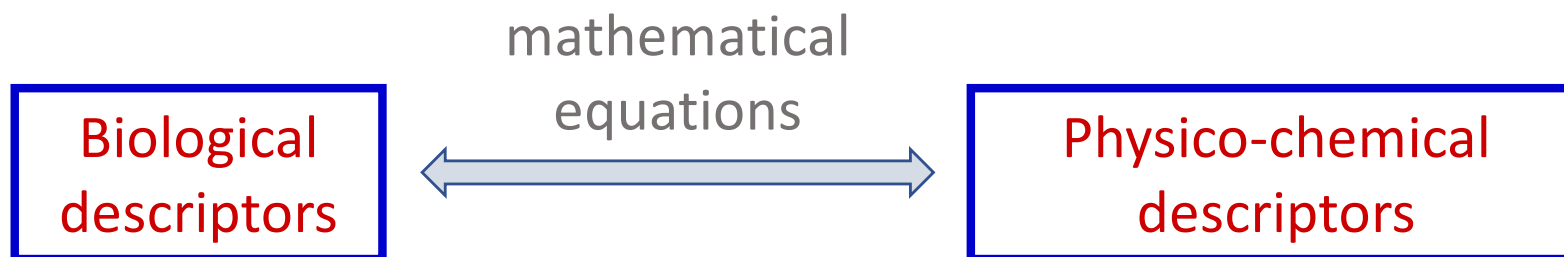
... from the different forms of molecular structure.



## 9.3 QSAR equations







arbitrary mathematical relationships, which can be represented by the generic equation:

$$P = f(D_1, D_2, D_3, \dots, D_n)$$

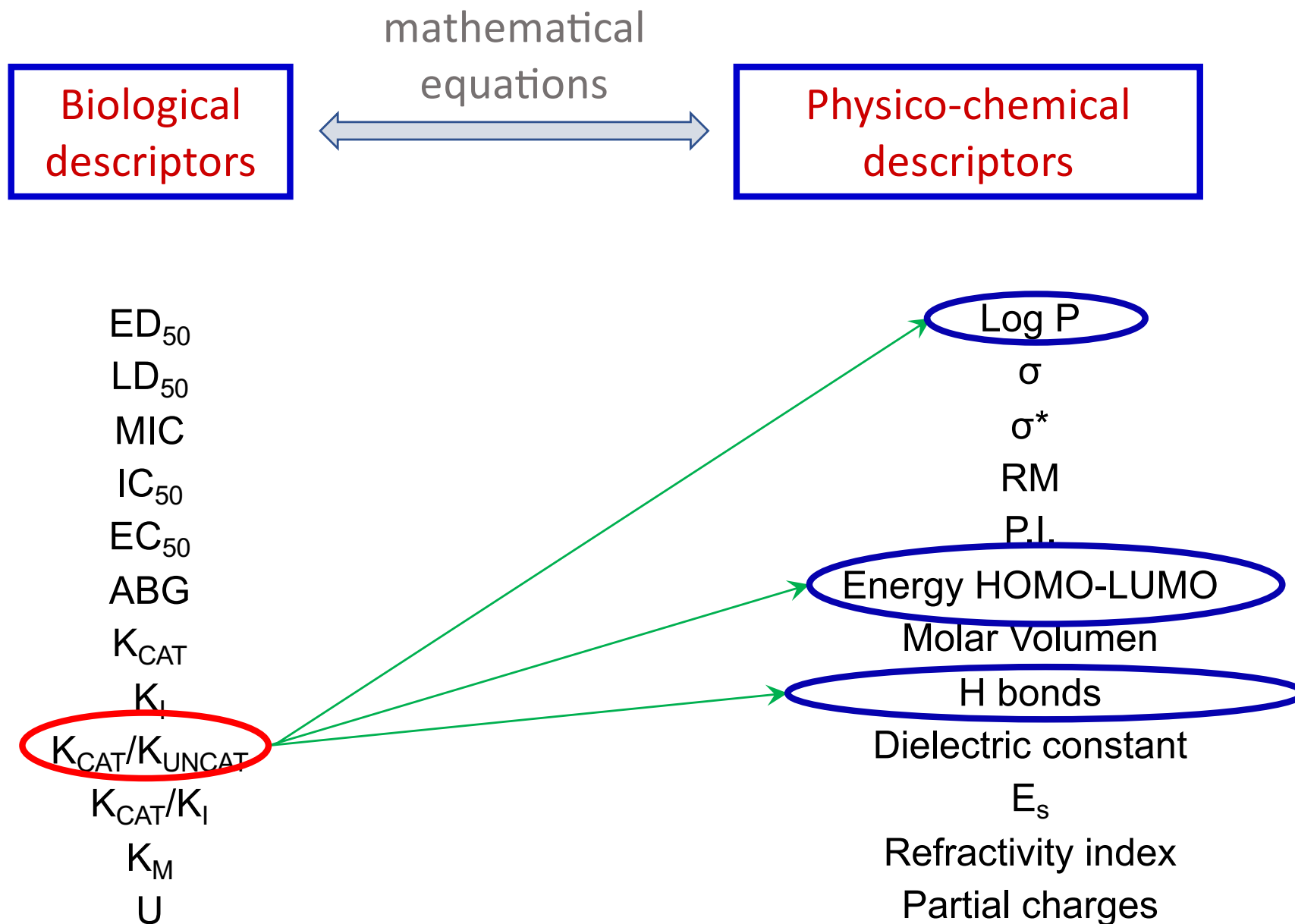
P: property or activity that you want to approximate

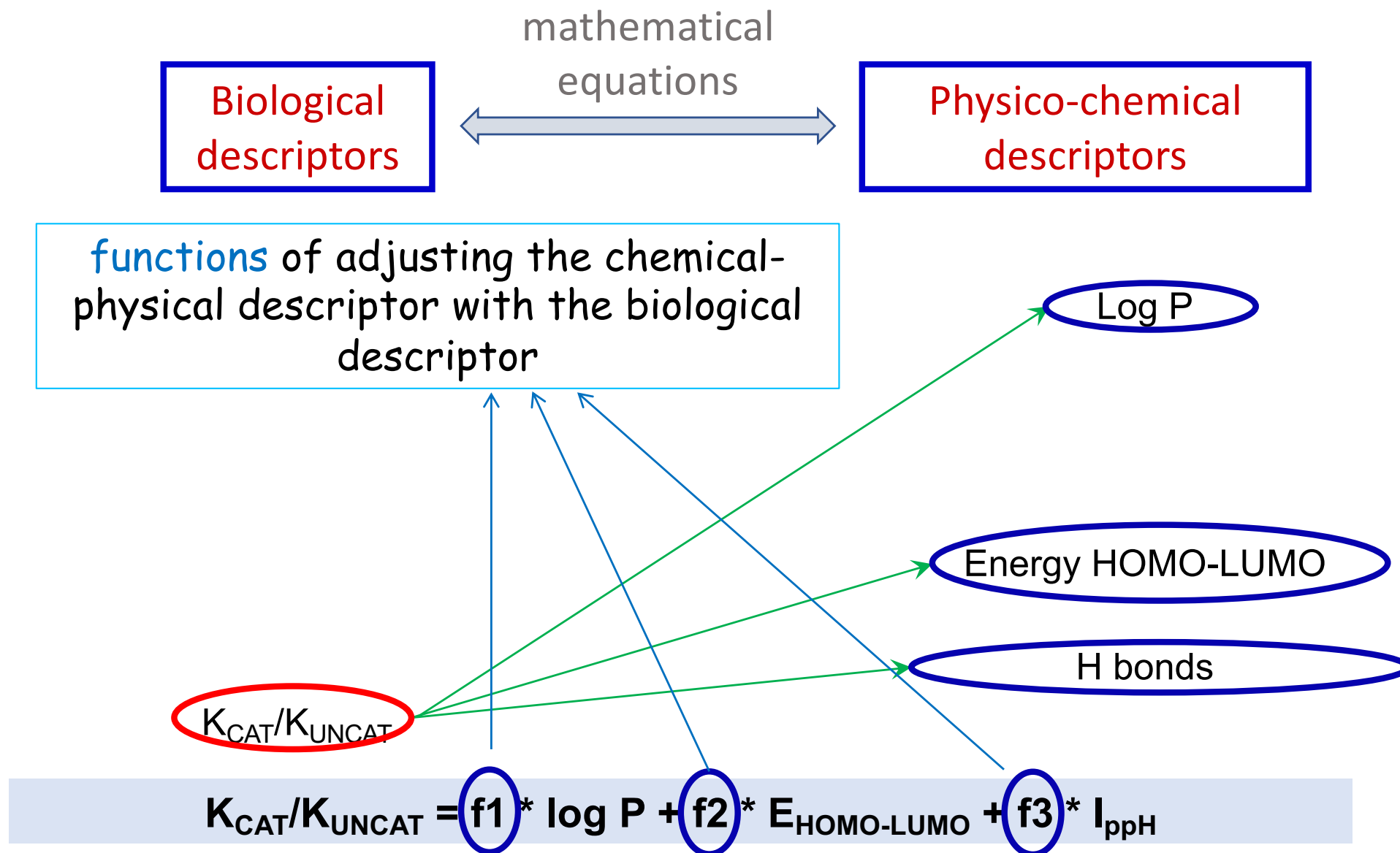
f( ): arbitrarily chosen mathematical function

$D_1, D_2, D_3, \dots$ : molecular, topological, electronic descriptors, prop. physical-chemical, etc., whose main requirement is that they be independent of each other

**main premise:** to build a model for the physical-chemical properties or biological activities are related to its structure

## 9.3 QSAR equations





## 9.3 QSAR equations



The use of a high number of descriptors raises problems such as:

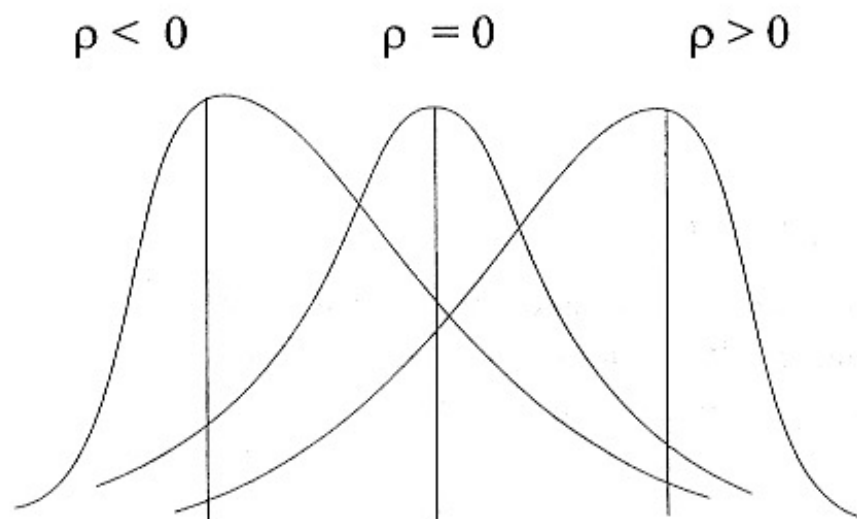
there are many possible combinations → long computing time

The selected set of compounds must give a good dispersion of values in the parameters: properties and descriptors (asymmetry and excess), which must not be correlated (correlation)

Asymmetry of the distribution

$$\rho = \frac{\sum_{j=1}^{N_c} n_j (O_j - \bar{O})^3}{n S^3}$$

$$S = \sqrt{\frac{\sum_{j=1}^{N_c} n_j (O_j - \bar{O})^2}{n}}$$



## 9.3 QSAR equations



The use of a high number of descriptors raises problems such as:

there are many possible combinations → long computing time

The selected set of compounds must give a good dispersion of values in the parameters: properties and descriptors (asymmetry and excess), which must not be correlated (correlation)

Pearson's correlation coefficients,  $r$ , helps us to know if they are correlated or not

$$r = \frac{\sum_{i=1}^n [x_i - \bar{x}][y_i - \bar{y}]}{\sqrt{\sum_{i=1}^n [x_i - \bar{x}]^2 \sum_{i=1}^n [y_i - \bar{y}]^2}}$$

$R = 0$  not correlated  
 $R = \pm 1$  perfectly correlated



Requirements of a QSAR model:

- Experimental data measured in an appropriate group and with sufficient precision to distinguish between them
- Descriptors that are readily available and expected to be related to biological activity
- A method to detect the relationship between parameters and experimental activity
- A method to validate QSAR results



Steps to generate a QSAR model:

1st. Approach of the objectives.

2nd. Determination of the biological activity of the series of compounds to be studied.

3rd. Describe the parameters to be correlated with biological activity.

4th. Establishment of the activity structure correlation and statistical analysis.

5th. Interpretation of the established relationship and prediction.

What molecules should we test?:

An (ordered) list of molecules !

Potential Candidates:

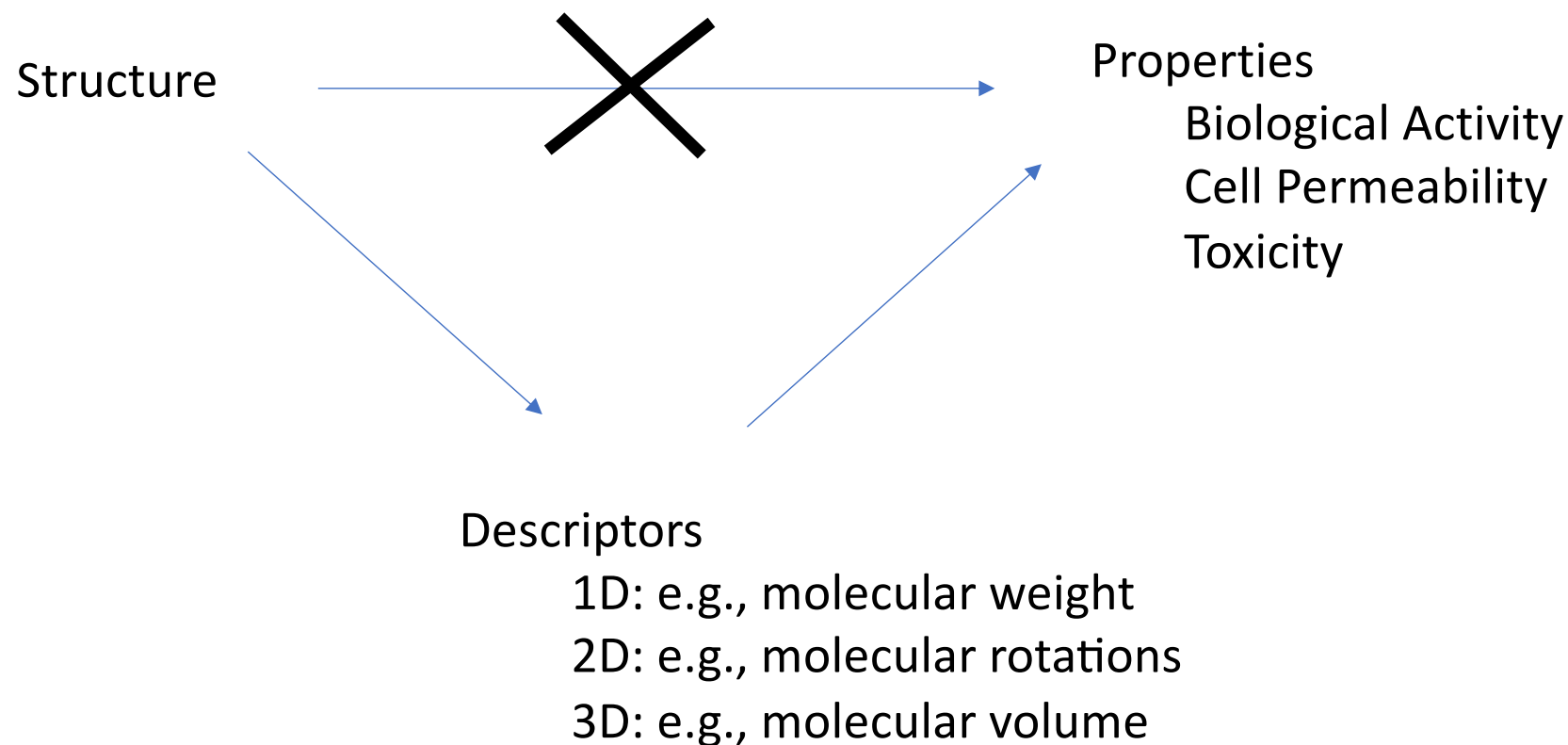
Corporate Database  
External Database  
Synthesis



Information:

Biological Activity  
Molecular Properties





## 9.3 QSAR equations



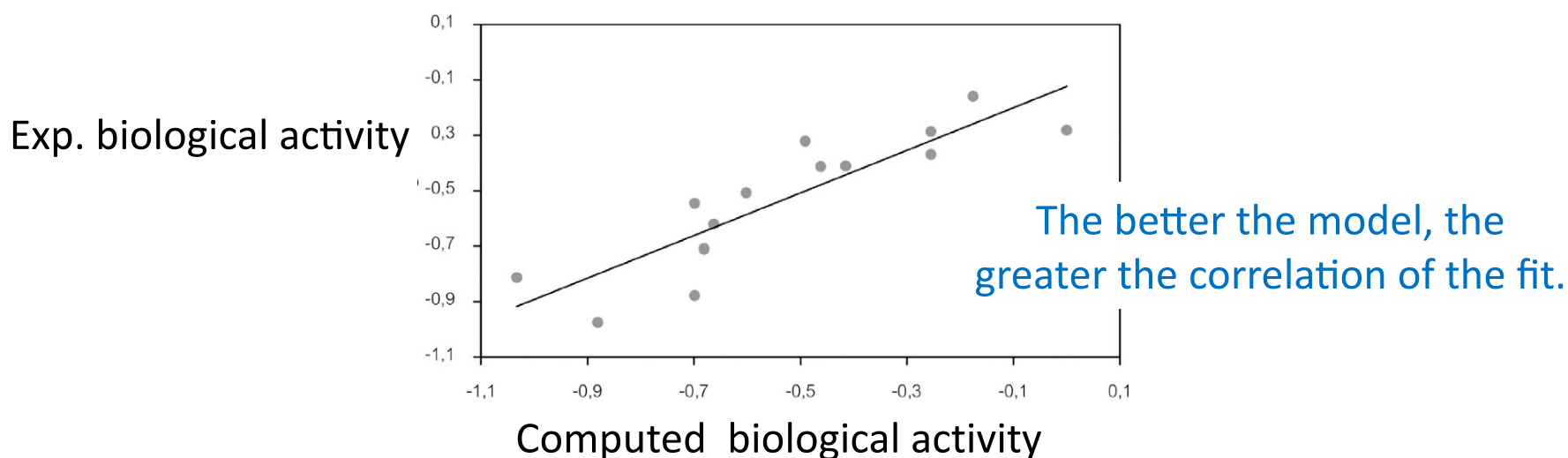
### Supervised procedures:

- Multiple Linear Regression (MLR)
- Discriminant analysis
- Partial Least Squares (PLS)
- Regression and classification trees
- Neural networks

### Not-supervised procedures:

- Principal Component Analysis (PCA)
- Cluster analysis
- Mapping no line

Once the expression of the QSAR model has been obtained, an adjustment of the experimental activities can be made with respect to those calculated by the model.



The graphical representation can help to discriminate those points that make the model worse

### Multiple Linear Regression (MLR)

One of the most used methods, given its simplicity, reproducibility and ease of interpreting the results obtained.

Handles several independent variables

There is a linear relationship between Y and several descriptors ( $x_i$ )

$$y = a_0 + a_1 X_1 + \dots + a_k X_k$$

$y$  = response or dependent variable

$x_i$  = descriptors (characteristics or independent variables) present in the model

$a_i$  = coefficients (where  $a_0$  is the constant term of the model)

Errors are minimized by least squares

Example:  $y = IC_{50}$  and  $x_i$  could be: molecular weight (MW)  $\log P$ ,  $K_{cat} / K_{uncat}$

$$IC_{50} = b_0 + b_1 * PM + b_2 * \log P + b_3 * K_{CAT}/K_{UNCAT}$$

### Multiple Linear Regression (MLR)

The least squares treatment ensures that the value of the coefficients minimizes the sum of the distances between the observed values and the ones calculated from the equation. It gives rise to a system of linear equations:

$$y_{c,j} = a_0 + \sum_{i=1}^k a_i x_{i,j} \quad d^2 = \sum_{j=1}^N (y_{c,j} - y_j)^2 \quad \frac{\partial d^2}{\partial a_i} = 0 \quad \begin{cases} [a_0] : a_0 N + a_1 \sum_{j=1}^N x_{1,j} + \dots + a_k \sum_{j=1}^N x_{k,j} = \sum_{j=1}^N y_j \\ [a_{\mu \neq 0}] : a_0 \sum_{j=1}^N x_{\mu,j} + a_1 \sum_{j=1}^N x_{\mu,j} x_{1,j} + \dots + a_k \sum_{j=1}^N x_{\mu,j} x_{k,j} = \sum_{j=1}^N x_{\mu,j} y_j \end{cases}$$

$$\begin{pmatrix} N & \sum_{j=1}^N x_{1,j} & \dots & \sum_{j=1}^N x_{k,j} \\ \sum_{j=1}^N x_{1,j} & \sum_{j=1}^N x_{1,j}^2 & \dots & \sum_{j=1}^N x_{1,j} x_{k,j} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^N x_{k,j} & \sum_{j=1}^N x_{1,j} x_{k,j} & \dots & \sum_{j=1}^N x_{k,j}^2 \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^N y_j \\ \sum_{j=1}^N x_{1,j} y_j \\ \vdots \\ \sum_{j=1}^N x_{k,j} y_j \end{pmatrix}$$

$$V_{k+1, k+1} \cdot W_{k+1, 1} = U_{k+1, 1} \quad W = V^{-1} \cdot U = (V^T \cdot V)^{-1} \cdot V^T \cdot U$$

### Multiple Linear Regression (MLR)

Several tools to analyze and validate the results:

Determination coefficient ( $R^2$ )

$$R^2 = 1 - \frac{\sum_{j=1}^N (y_j - y_{c,j})^2}{\sum_{j=1}^N (y_j - \langle y \rangle)^2}$$

$R^2 = 1$  ideal system

$R^2 \neq 1$  the fit moves away from a line

Correlation coefficient (R)

$$R = \sqrt{R^2}$$

Pearson's linear correlation coefficient (r)

allows establishing the relationship between two data sets from their normalized covariance

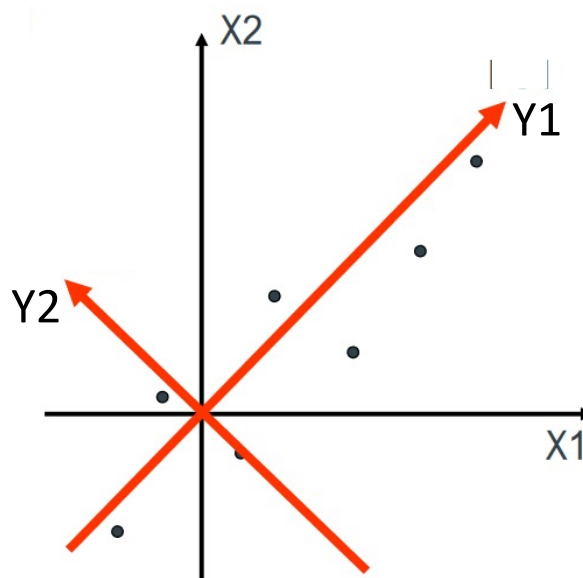
$$r = \frac{\sum_{j=1}^N (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_{j=1}^N (x_i - \langle x \rangle) \sum_{j=1}^N (y_i - \langle y \rangle)}$$

$$\begin{cases} r = +1 & \text{proportional correlation} \\ r = 0 & \text{no correlation} \\ r = -1 & \text{inverse correlation} \end{cases}$$

### Principal Component Analysis (PCA)

It is a technique used to **reduce** the dimensionality of a data set.

The PCA constructs a linear transformation that chooses a new coordinate system for the original data set, in which the largest variance of the data set is captured on the first axis, the second largest variance is the second axis, and so on

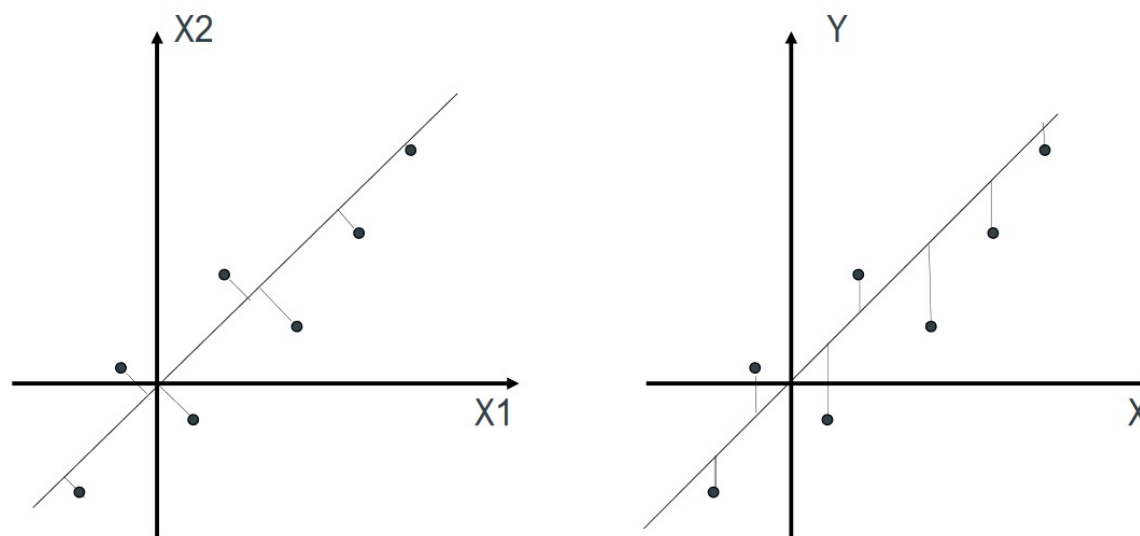


The Y1 and Y2 axes must be orthogonal

### Principal Component Analysis (PCA)

PCA minimizes the squares of orthogonal errors to the principal components

Linear regression minimizes the squares of the errors in the Y direction



The Y1 and Y2 axes must be orthogonal

### Principal Component Analysis (PCA)

The first step corresponds to standardizing the variables (descriptors) by subtracting the mean of each one and/or dividing the result by its standard deviation:

$$\{X_{i,j=1,N}\}_{i=1,k} \quad \{\langle X_i \rangle\}_{i=1,k} \quad \{S_{X_i}\}_{i=1,k} \quad \rightarrow \quad \{\chi_{i,j=1,N}\}_{i=1,k} \quad \chi_{i,j} = \frac{X_{i,j} - \langle X_i \rangle}{S_{X_i}/1}$$

From these new data we calculate the covariance matrix, which will be a square matrix of dimension k:

$$\{\chi_{i,j=1,N}\}_{i=1,k} \quad \{\langle \chi_i \rangle\}_{i=1,k} \quad \rightarrow \quad C = \begin{pmatrix} c_{1,1} & \dots & c_{1,k} \\ \vdots & \ddots & \vdots \\ c_{k,1} & \dots & c_{k,k} \end{pmatrix} \quad c_{\mu,\theta} = \frac{1}{N-1} \sum_{j=1}^N (\chi_{\mu,j} - \langle \chi_{\mu} \rangle)(\chi_{\theta,j} - \langle \chi_{\theta} \rangle)$$

$$C = cov_{k,k} = a_{k,N}^T \cdot a_{N,k} \cdot \frac{1}{N-1} \quad a_{N,k} = x_{N,k} - \mathbf{1}_{N,N} \cdot x_{N,k} \cdot \frac{1}{N-1} \quad \mathbf{1}_{i=1,N; j=1,N} = 1$$

The matrix C is then diagonalized obtaining a set of eigenvalues (D) and eigenvectors (P):

$$C_{k,k} = P_{k,k} \cdot D_{k,k} \cdot P_{k,k}^T$$

From the value of the eigenvalues, it is possible to decide whether to ignore that variable that has a small associated eigenvalue and thus reduce the number of variables



### Principal Component Analysis (PCA)

Alternatively, it is possible to perform a **transformation** on the sample data using the eigenvectors of those highest eigenvalues, thus reducing the dimensionality of the problem:  $a < k$

To do this, the **Q matrix** is constructed from the eigenvectors with the highest value and projecting the space of normalized variables:

$$T_{N,a} = \chi_{N,k} \cdot Q_{k,a}$$

The **new MLR model** is generated on this reduced set of variables ( $T_N, a$ )

To predict the activity of a new compound, we would also need the "k" descriptors that we would normalize with respect to the mean of each one and the new variables (T) would be generated, projecting with the eigenvectors (Q). The new "a" variables thus obtained would be introduced into the QSAR model.

Finally, the residual matrix (E) that captures the variation of the data can be estimated when reducing the dimension:

$$\chi'_{N,k} = T_{N,a} \cdot Q_{a,k}^T \quad E_{N,k} = \chi_{N,k} - \chi'_{N,k} \quad \chi_{N,k} = T_{N,a} \cdot Q_{a,k}^T + E_{N,k}$$

## 9.3 QSAR equations

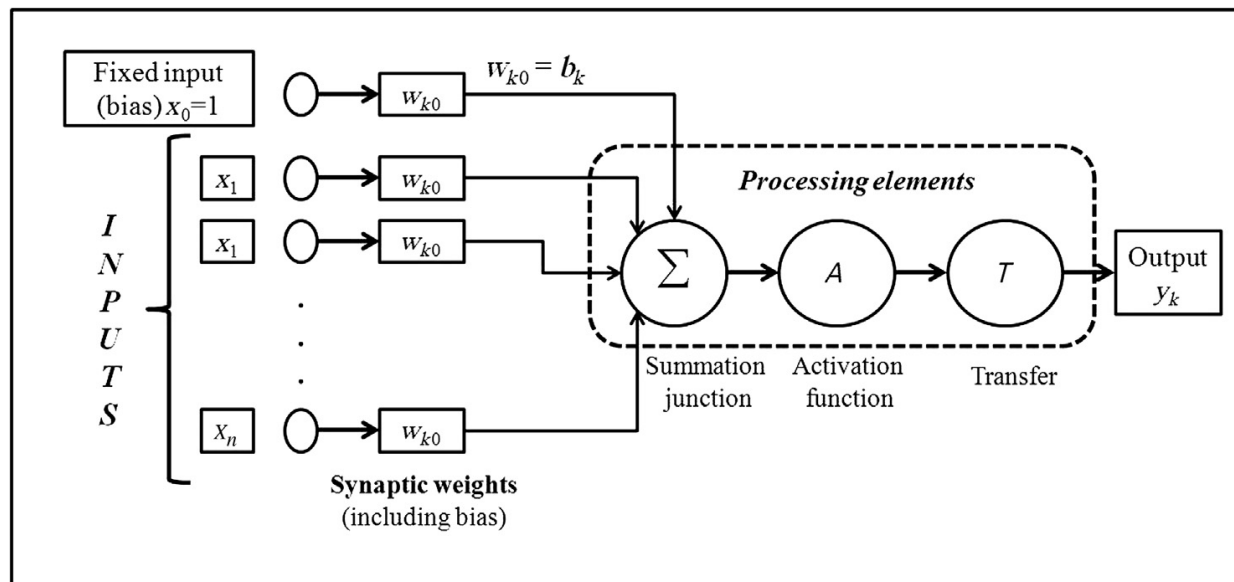


### Neural networks

An artificial neuron (**AN**) is a computational model inspired from the natural neurons. Each neuron receives a series of inputs through interconnections and emits an output.

Artificial neural networks (**ANNs**) connect artificial neurons arranged in layers in order to process information

This output is given by three functions: **summation**, **activation** and **transfer**.



schematic  
representation of a  
computer neuron

## 9.3 QSAR equations



In a computer neuron, four operations are performed:

- 1st. = the input and output function, which evaluates input signals from the neurons of the previous layer, determining the strength of each input, and passes the output signal to the neurons of the next layer.

- 2nd = the summation function, which calculates a total for the combined input signals according to the equation

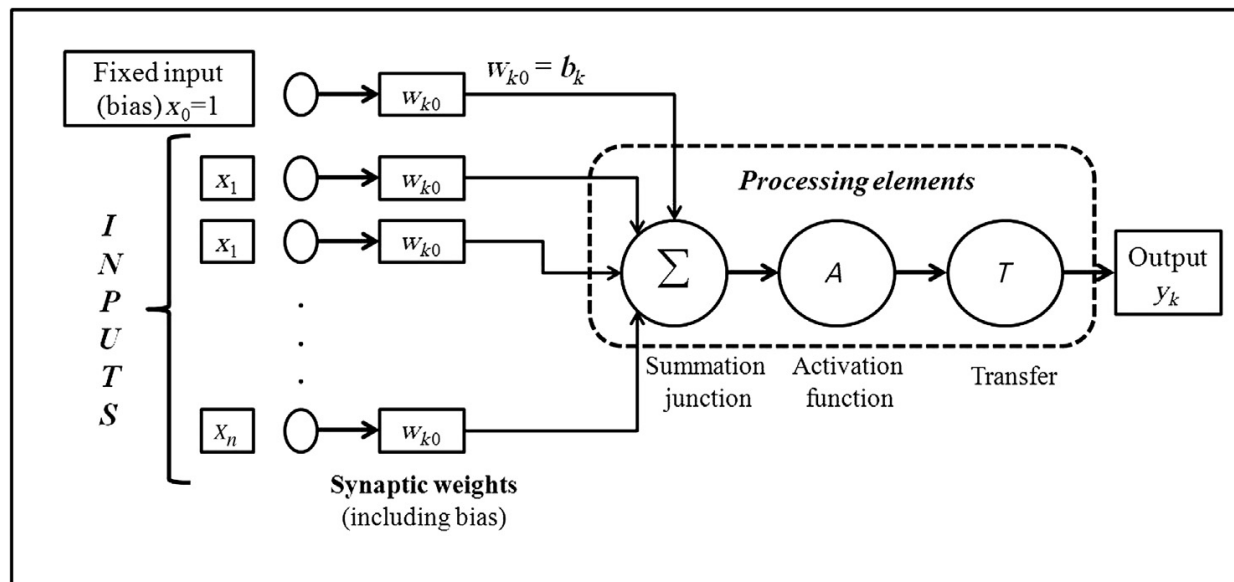
$$i_j = \sum w_{ji} o_i$$

$i_j$  = the net input in node  $j$  (layer  $\lambda$ ),

$o_i$  = the output of node  $i$  in the previous layer ( $\lambda - 1$ )

$w_{ji}$  = the weight associated with the nodes  $i$  and  $j$ .

- 3rd = the activation function, which allows the outputs to vary with respect to time.
- 4th = the transfer function, which maps the summed input to an output value.

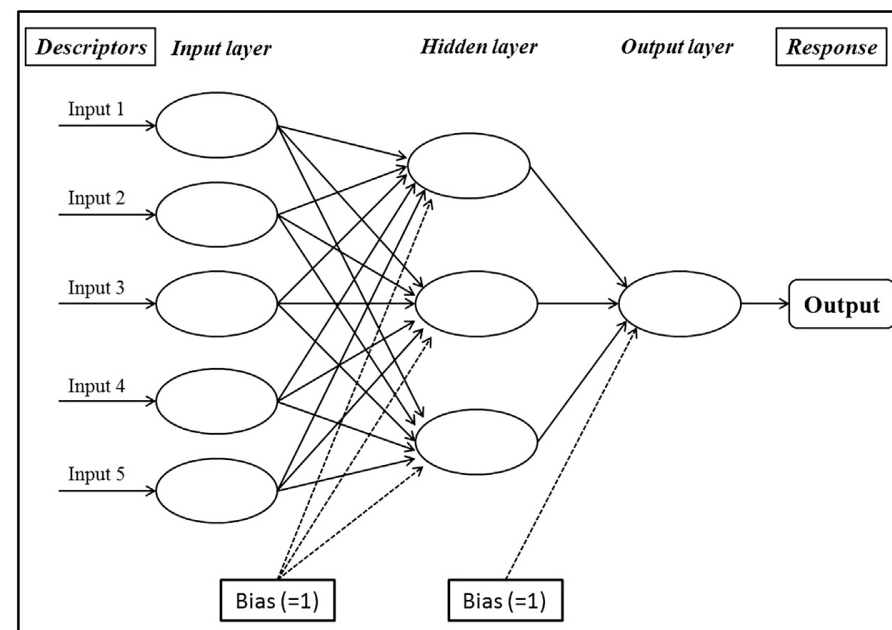


schematic  
representation of a  
computer neuron

### Neural networks

**Architecture of a three-layer ANN.** It is made up from **an input**, **an output**, and a **hidden layer**.

Each input layer node corresponds to a single independent variable, with the exception of the bias node. Similarly, each output layer node corresponds to a different dependent variable. Each node from the input layer is connected to a node from the hidden layer and every node from the hidden layer is connected to a node in the output layer.



There is usually some weight associated with every connection.

The method to control neural networks is by setting and adjusting weights between nodes. Initial weights are usually set at some random numbers and then they are adjusted during neural-network training. The neural-network fitting may suffer from overfitting, thus requiring strict tests of validation before they can be applied for prediction purposes.



The best validation process, although often unfeasible, is to divide the group of compounds in two: 1) one to derive the model  
2) test.

The quality of a QSAR model is evaluated with four parameters:

- number of compounds considered
- standard deviation of the regression ( $s$ )
- correlation coefficient ( $r$ )
- Fisher's test ( $F$ )

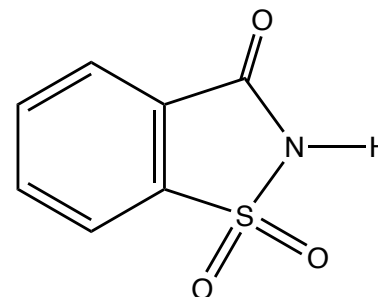
Another possibility is cross validation:

- The  $N$  compounds are divided into  $M$  groups
- A model is established using  $M-1$  groups and the properties of the compounds of the excluded group are predicted
- The process is repeated excluding each time one of the groups
- The standard deviation of the predictions is calculated

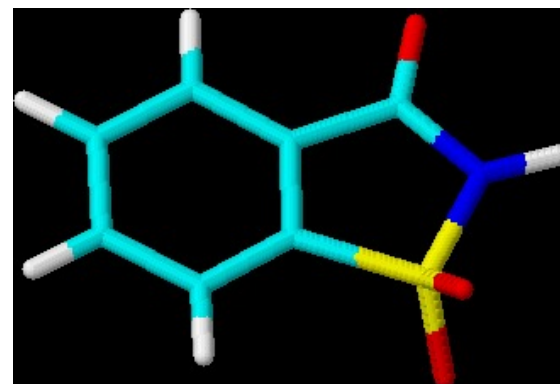
One 1D structure



One 2D structure

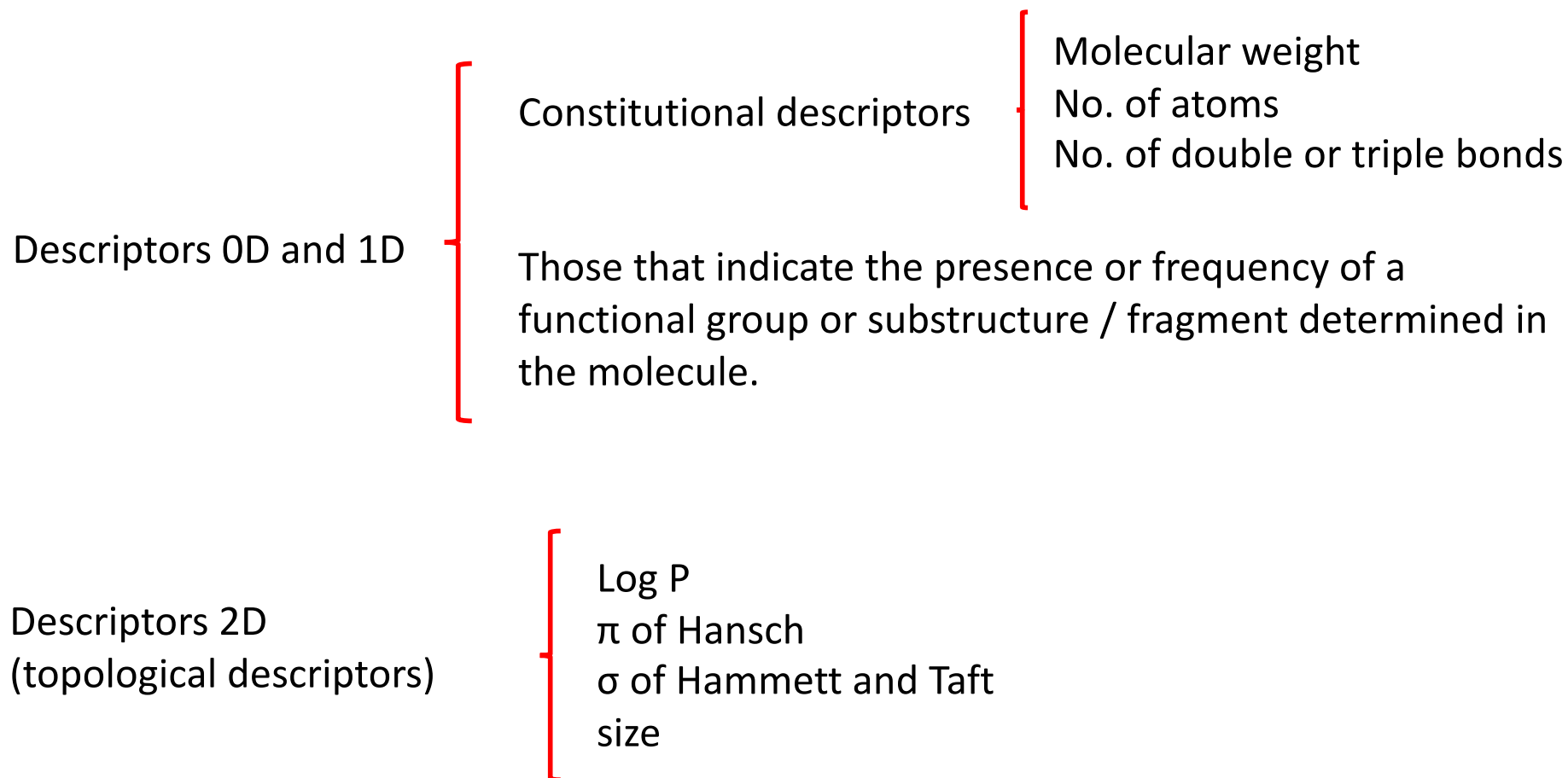


Many 3D structures



3D-QSAR is computationally more comprehensive and complex than 2D

1D-2D descriptors do not depend on conformation or orientation



### Advantages of 2D-QSAR

- Independence of value from conformation and orientation
- Optimized coordinates are not required or establishing orientation criteria for the calculation of descriptors
- Low computational cost
- No information is required on the structure of the receiver
- Very powerful for studies of a large number of structures ( $10^5$  to  $10^6$ )

### Disadvantages of 2D-QSAR

- Only molecules of similar structure can be studied
- Numeric descriptors in tables are not always valid (it is difficult to separate some properties from others)
- There are rare substituents not included in the tables
- It is necessary to synthesize a large number of molecules
- The equations do not directly suggest which new compound to synthesize

**In general:** - useful information but difficult to interpret  
- effective in identifying new lead compounds





### 3D-QSAR

- It considers the three-dimensional properties of the molecule to predict its biological response
- It involves a broad set of techniques and methods that correlate descriptors derived from the spatial representation of molecular structures.
- The most important aspects that can affect are: shape, size and electronic properties

### Common stages

- 1<sup>st</sup> : To determine the optimal (bioactive) conformation of the compound or ligand, either by experimental data (X-rays or NMR) or theoretically
- 2<sup>nd</sup> : To align the conformers of the 3D dataset
- 3<sup>rd</sup> : Calculate the descriptors
- 4<sup>th</sup> : Correlate the calculated descriptors with the experimental biological response of the studied compounds.

### **3D-QSAR:** general features

- The binding of the ligand with the receptor is considered directly related to the biological activity
- Molecular properties can be encoded by a specified number of descriptors - Compounds with similar structures have similar properties
- The structural properties that give rise to a biological response are determined by intermolecular forces (non-bonding), such as Lennard-Jones or electrostatics.
- The biological response depends only on the ligand and not on its metabolite (product)
- The lowest energy conformation is bioactive
- Receiver geometry is considered rigid
- The loss of translational and rotational degrees of freedom that occurs when binding occurs (related to entropy) is similar for all compounds
- The binding site of all the ligands is the same.

The 3D-QSAR methods can be classified based on a variety of criteria:

Basis of classification	Type	Examples of techniques
Based on employed chemometric techniques	Linear	CoMFA, CoMSIA, AFMoC, GERM, CoMMA, SoMFA
	Nonlinear	Compass
Based on the alignment criterion	Alignment-dependent	CoMFA, CoMSIA, MSA, RSA, GERM, AFMoC, HIFA, VFA, MQSM
	Alignment-independent	Compass, CoMMA, HQSAR, WHIM, GRIND, VolSurf, CoSA
Based on intermolecular modeling or the information employed to develop QSAR	Ligand-based	CoMFA, CoMSIA, MSA, RSA, Compass, GERM, CoMMA, SoMFA
	Receptor-based	AFMoC, HIFA

© Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press. 2015



**Comparative molecular field analysis (CoMFA)** is a molecular field-based, alignment-dependent, ligand-based method developed by Cramer et al.<sup>1</sup>

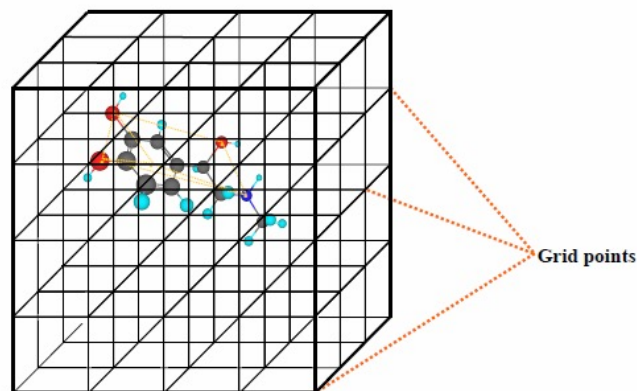
The objective of a CoMFA model is to derive a correlation between the **biological activity** of a series of molecules and their **3-D structures**. This correlation is derived from the superimposition of the molecules in their active conformation.

The **CoMFA** method performs the following steps:

- 1<sup>st</sup> To obtain the structures of the different molecules (ligands)
- 2<sup>nd</sup> To generate the bioactive conformation of each molecule by minimizing its energy, and the partial charges for each atom are obtained (QM, MM)
- 3<sup>rd</sup> To align all molecules, using manual or automated methods, in ways that presumably interact with the receptor
- 4<sup>th</sup> To locate in the center of a three-dimensional network with a vertex spacing (resolution) of 2 Å

1. Cramer III RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). i. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110(18):5959-67.

### Comparative molecular field analysis (CoMFA)



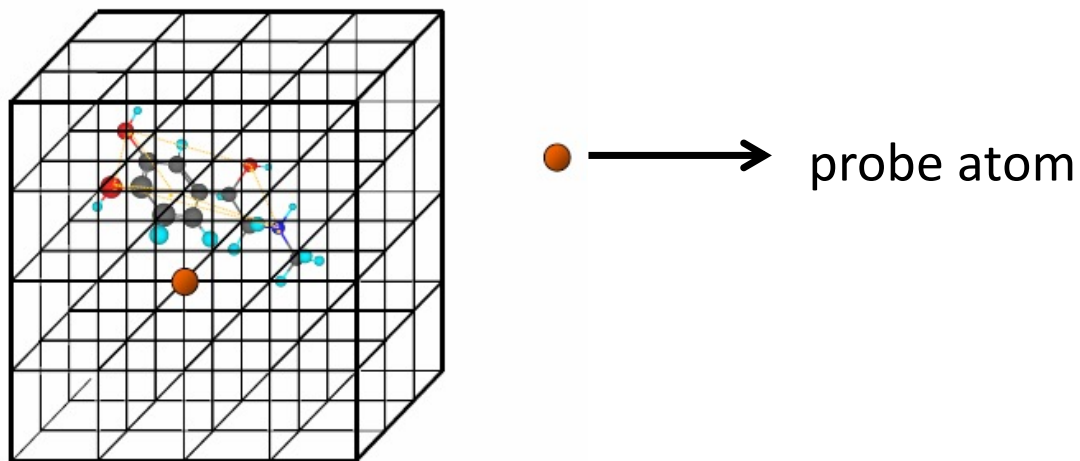
- The electrostatic and Lennard-Jones (van der Waals) energy is calculated on the different points of the network

$$V_{LJ} = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right] = \epsilon \left[ \left( \frac{r_o}{r} \right)^{12} - 2 \left( \frac{r_o}{r} \right)^6 \right] \quad r_o = 2^{\frac{1}{6}} \sigma \quad V = \frac{q_i q_j}{4\pi\epsilon r}$$

- For the calculation of the interaction energies, grids with an edge size of approx. of 4 Å greater than the molecular size, since otherwise a very high number of points would be generated.

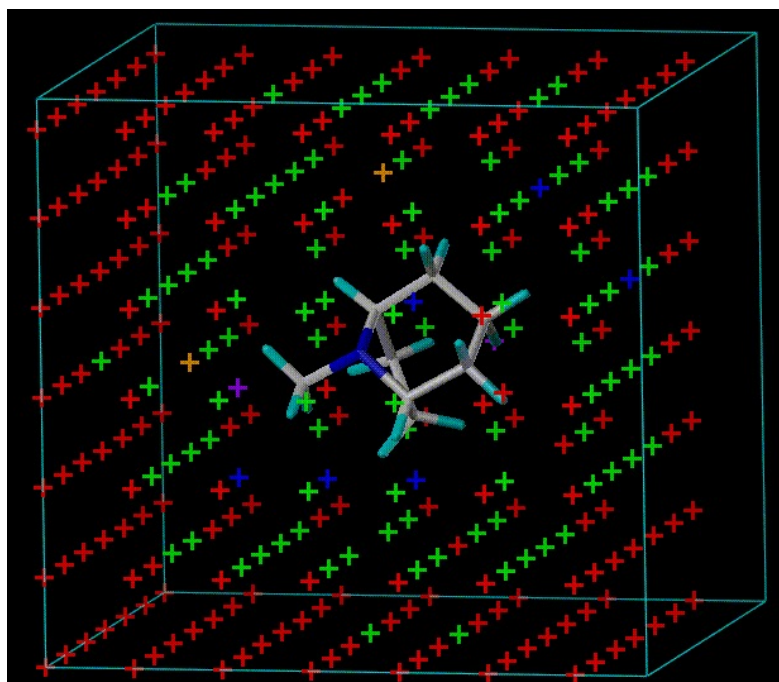
### Comparative molecular field analysis (CoMFA)

- The interaction energy is evaluated for each atom of the molecule, with respect to reference fragments (probe atom, “probes”) that are located in the different points of the grid
- In the case of CoMFA, a positive charge is used for the electrostatic (it simulates an  $H^+$  cation) and a carbon with  $sp^3$  hybridization for the steric ones.
- Cutoffs of approx.  $30 \text{ kcal mol}^{-1}$



### Comparative molecular field analysis (CoMFA)

- The different interaction energies obtained are used as a set of **descriptors** that correlate with the **activity** of the molecules using the **PLS technique**.



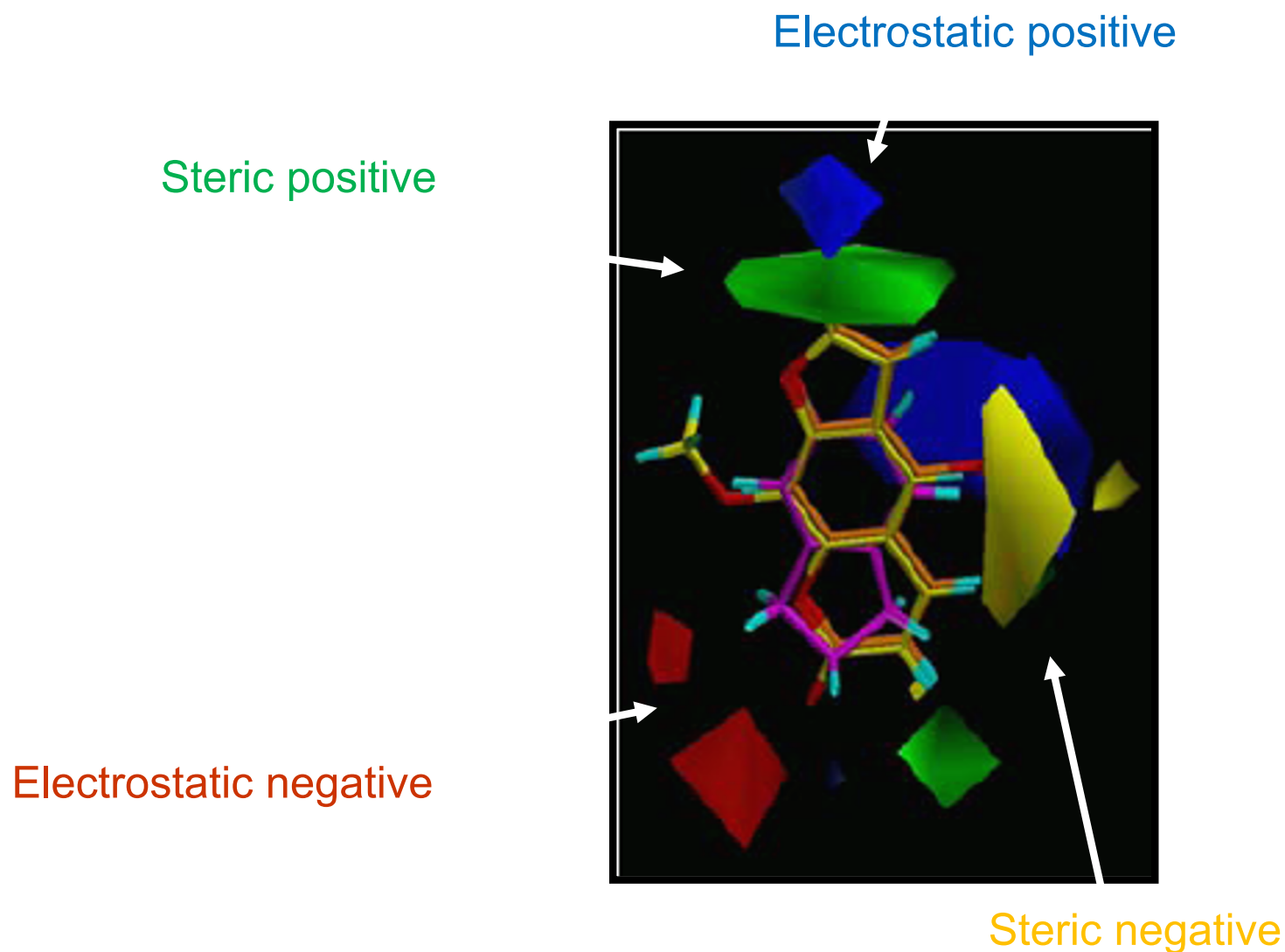
$$\text{Activity} = a\text{Steric} + b\text{Electric}$$

$$E_c = \sum_{i=1}^{N_{atoms}} \left[ \frac{Q_i Q_j}{D_{ij} R_{ij}} \right]$$

$$E_{vdw} = \sum_{i=1}^{N_{atoms}} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} \right]$$

### Comparative molecular field analysis (CoMFA)

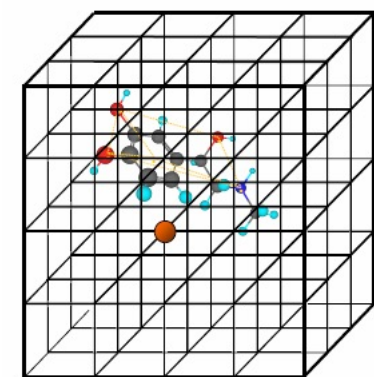
- The probe makes it possible to identify regions with attractive and repulsive steric and electrostatic interactions.





### Comparative molecular field analysis (CoMFA)

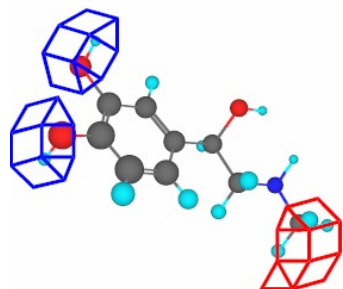
- The results can be expressed as correlation equations of linear combination of descriptors
- A visual interpretation is usually made when representing the results obtained by PLS



Producto	Actividad biológica	Campos estéricos (S) en los puntos grid (001-998)					Campos electrostáticos (E) en los puntos grid (001-998)				
		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										

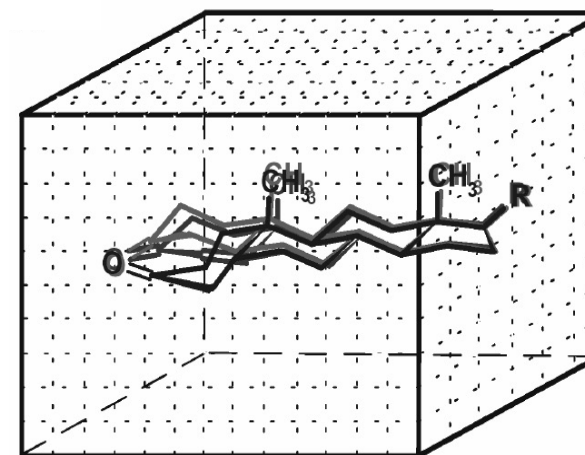
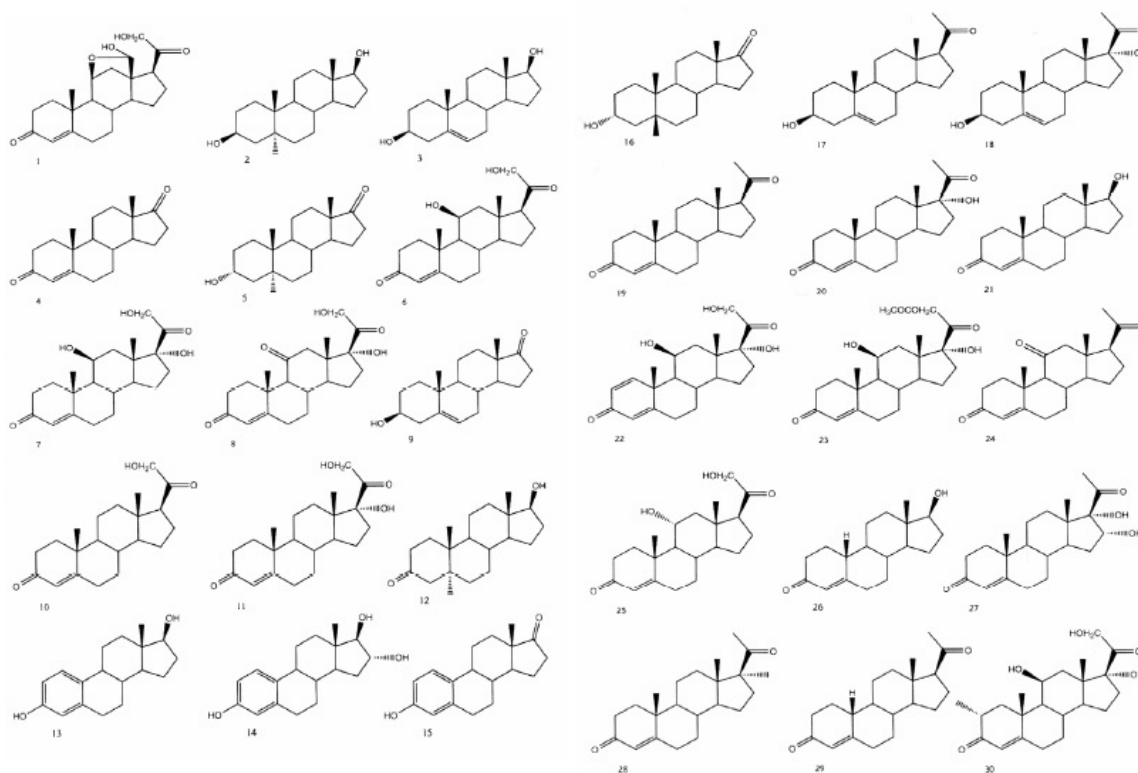


$$\text{Actividad} = aS001 + bS002 + \dots + mS998 + nE001 + \dots + yE998 + z$$



### Comparative molecular field analysis (CoMFA)

Example: In a study with 30 steroids with affinity for corticosteroid-affinity globulin (CBG) 21 were used as calibration set and a grid of separated points 1.5 Angstroms



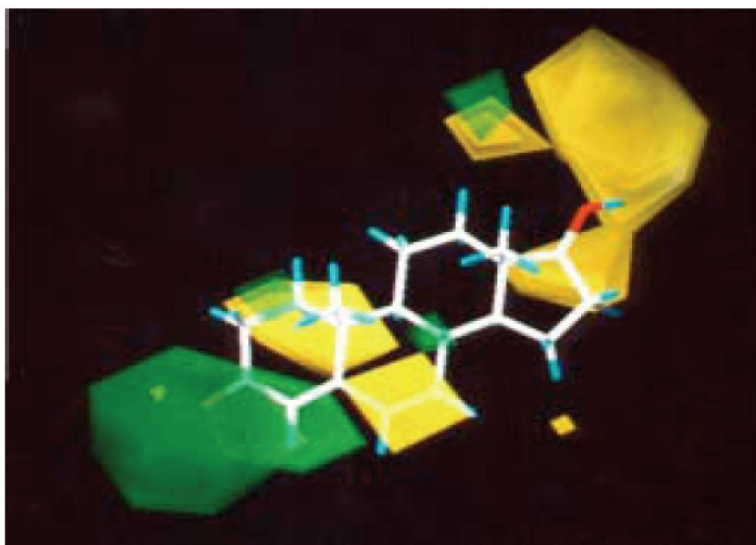
### Comparative molecular field analysis (CoMFA)

Once the fields were calculated, each steroid was defined by 4704 variables

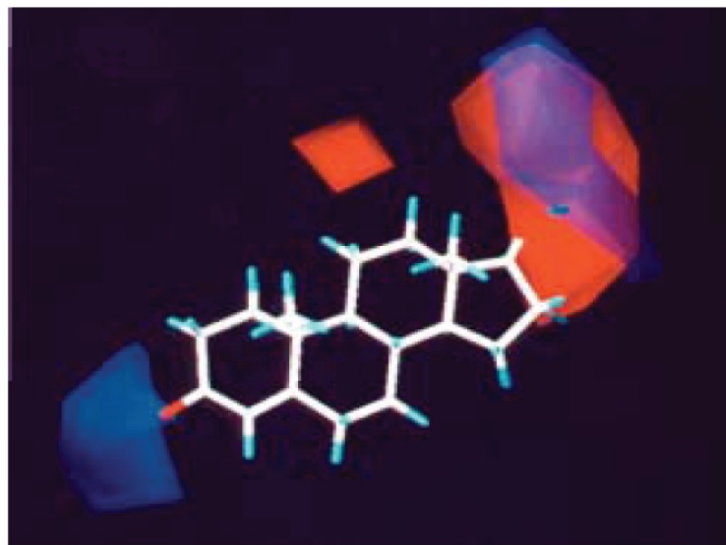
The negative electrostatic contour area indicates the area where the presence of groups with negative electron density (3-keto) is favorable.

Steroids with a more positive substituent such as OH have lower affinity

CoMFA steric contour map



CoMFA electrostatic contour map



The 3D-QSAR methods can be classified based on a variety of criteria:

Basis of classification	Type	Examples of techniques
Based on employed chemometric techniques	Linear	CoMFA, CoMSIA, AFMoC, GERM, CoMMA, SoMFA
	Nonlinear	Compass
Based on the alignment criterion	Alignment-dependent	CoMFA, CoMSIA, MSA, RSA, GERM, AFMoC, HIFA, VFA, MQSM
	Alignment-independent	Compass, CoMMA, HQSAR, WHIM, GRIND, VolSurf, CoSA
Based on intermolecular modeling or the information employed to develop QSAR	Ligand-based	CoMFA, CoMSIA, MSA, RSA, Compass, GERM, CoMMA, SoMFA
	Receptor-based	AFMoC, HIFA


© Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic Press. 2015

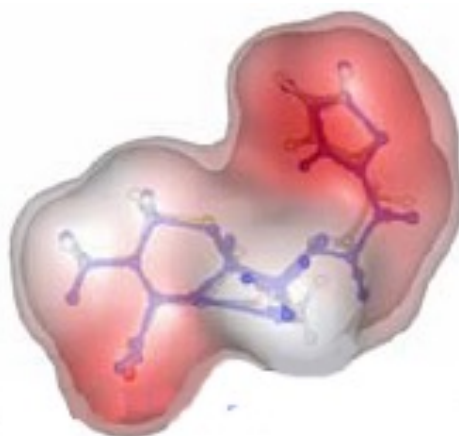
### Molecular Shape Analysis (MSA)

The **MSA** is an iterative process in which steps are repeated until the **molecular shape similarities** and other descriptors are checked and adjusted in order to generate a QSAR equation with optimal statistical significance.

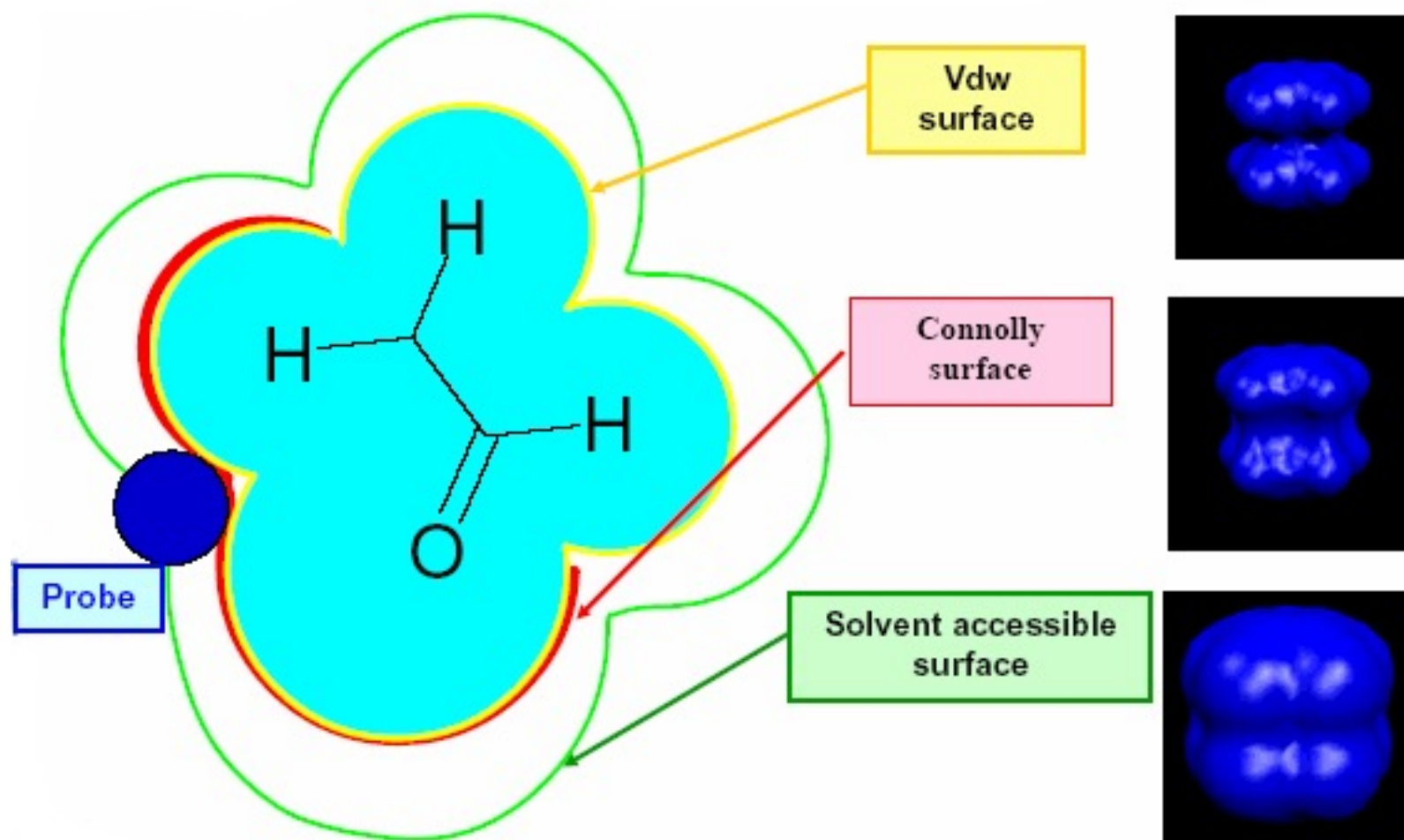
The goal of MSA is to generate a QSAR equation that incorporates **spatial molecular similarity data**.

As a global measure of the similarity of the molecular shape, the **superimposed steric volume** between a pair of molecules has been used.

Activity  Ligand shape accommodate to binding site



### Molecular Shape Analysis (MSA)



### **Molecular Shape Analysis (MSA)**

An MSA analysis consists of the following steps:

- 1<sup>st</sup> Conformational Analysis: Biologically Relevant. Generating conformers and energy minimization of each structure
- 2<sup>nd</sup> Hypothesize the active conformation: Minimum of the most active
- 3<sup>rd</sup> Select a candidate shape reference compound (normally the active one)
- 4<sup>th</sup> Perform Paired Molecular Overlays
- 5<sup>th</sup> Measure the similarity of molecular shapes through the calculation of the MSA descriptors (DIFFV, Fo, NCOSV, Shape RMS, COSV,...)
- 6<sup>th</sup> Determine other molecular characteristics (spatial, electronic, thermodynamic descriptors) to be added as independent variables along the MSA descriptors
- 7<sup>th</sup> Build a QSAR model through GFA, G/PLS methods
- 8<sup>th</sup> Use the optimized QSAR for ligand design



### Molecular Shape Analysis (MSA)

#### Limitations of 3D-QSAR-3D

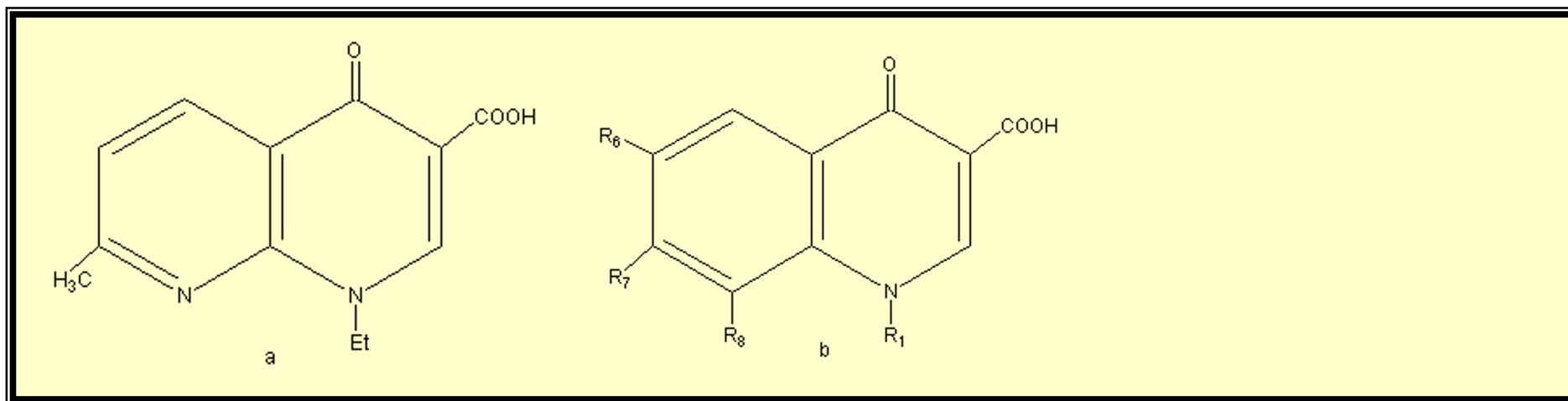
- Models can predict in the 3D space covered by the structures of calibration: if a position has only methyl and ethyl groups, the effect of higher alkyls can not be predicted

#### Advantages of 3D-QSAR

- The same protocol can be applied to different problems
- The method can handle groups of structurally distinct compounds and predict new compounds with characteristics that did not exist in the set of calibration
- Contour maps are directly related to important interactions drug-receptor
- If the structure of the receptor is known and the structures are aligned by placing them in the active site, the resulting contour maps will match the amino acids of the receptor that are important to the binding and will provide information on which are the main interactions



### Drug development



*Variation in position 1*

$\pi$  : hydrophobicity parameter

$$R_6 = F \quad R_7 = \text{piperaciline} \quad R_8 = H$$

$$\log (1/\text{MIC}) = -0.49 (L_1)^2 + 4.10 L_1 - 2.00$$

$$(n = 8; s = 0.13; r^2 = 0.91; F = 25.78)$$

$$L_{1(\text{opt})} = 4.2 \text{ \AA}.$$

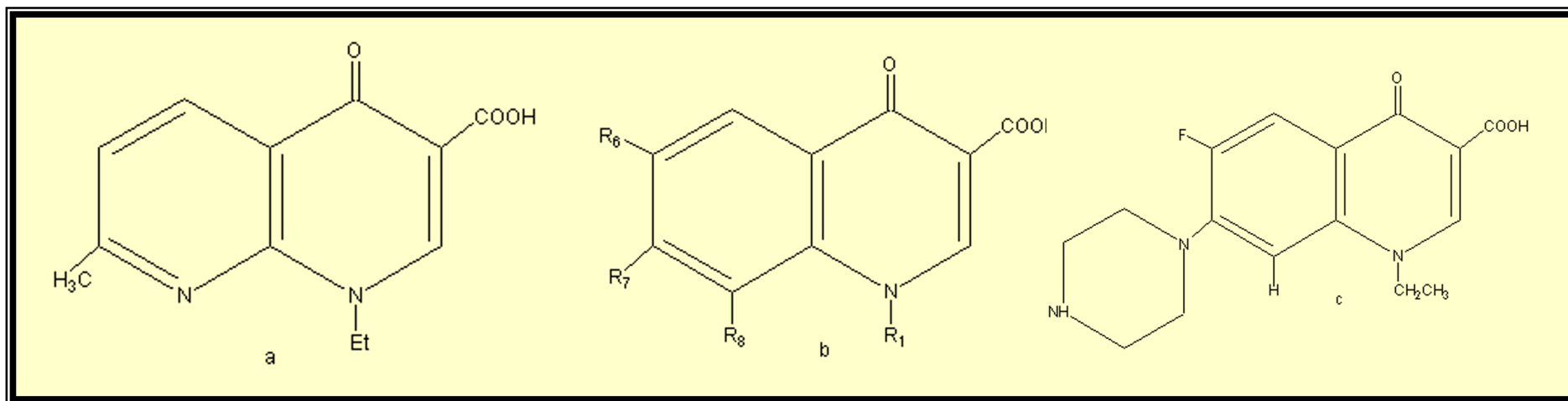
*Variation in position 6*

$$R_1 = \text{Et} \quad R_7 = R_8 = H, \quad R_6 = H, F, Cl, Br, I, NO_2, Me \text{ y } OMe$$

$$\log (1/\text{MIC}) = -3.32 (E_s(6))^2 - 4.37 E_s(6) + 3.92$$

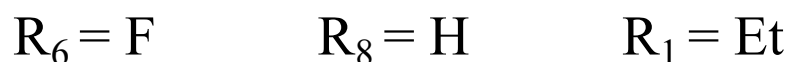
$$(n = 8; s = 0.11; r^2 = 0.98; F = 112.29) \quad E_s(6)_{\text{opt}} = -0.66, \text{ between } F \text{ and } Cl$$

### Drug development



*Variation in position 7*

$\pi$  : hydrophobicity parameter

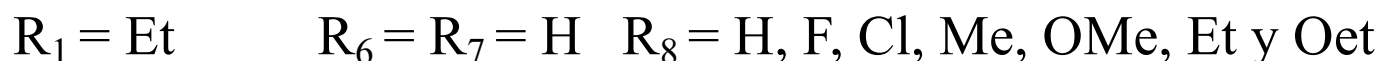


$$\log (1/\text{MIC}) = -0.24 (\pi_7)^2 - 0.68 \pi_7 - 0.71 I(7N\text{-CO}) + 5.99$$

( $n = 22$ ;  $s = 0.24$ ;  $r^2 = 0.89$ ;  $F = 47.97$ )

$$\pi_{7 \text{ opt}} = -1.38 \text{ \AA}.$$

*Variation in position 8*



$$\log (1/\text{MIC}) = -1.00 (B_4(8))^2 + 3.73 B_4(8) + 1.3$$

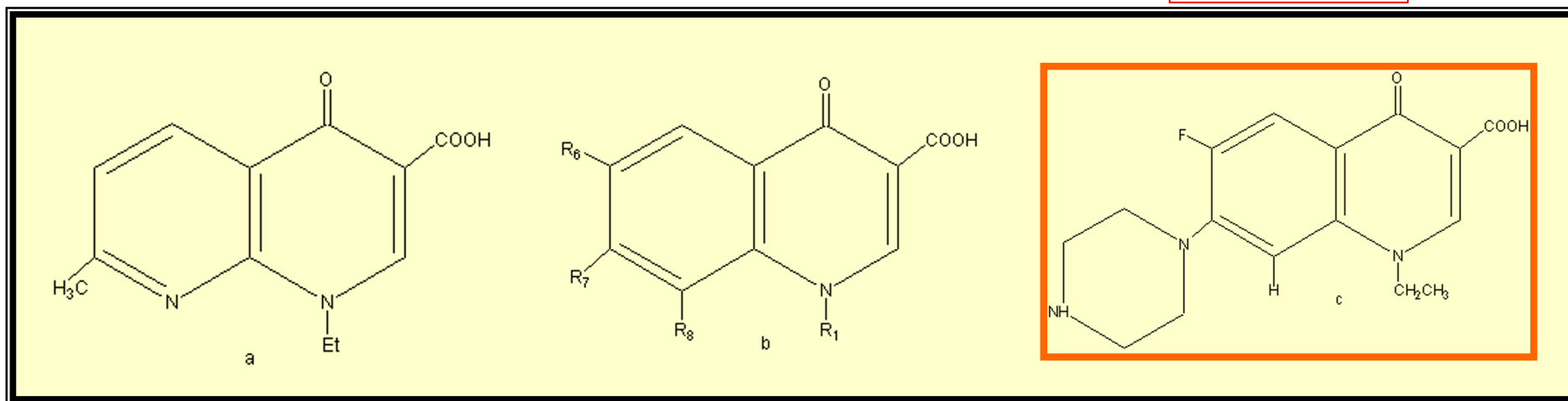
( $n = 7$ ;  $s = 0.22$ ;  $r^2 = 0.96$ ;  $F = 44.05$ )       $B_4(8)_{\text{opt}} = 1.83 \text{ \AA}$ , between Cl and Me

## 9.6 Applications of QSAR



### Drug development

#### Norfloxacin

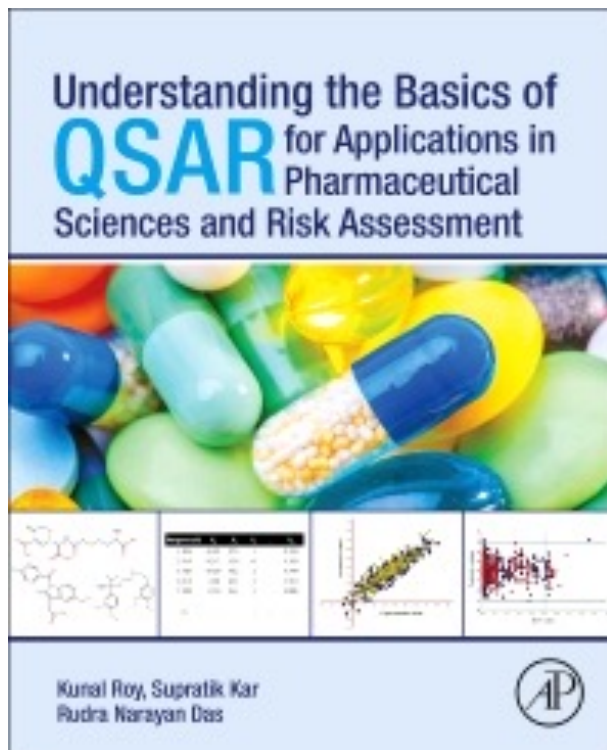


Nalidixic acid: Weak activity against urinary infections

$$\begin{aligned} \log (1/\text{MIC}) = & -0.36 (L_1)^2 + 3.04 L_1 - 2.50 (E_{s(6)})^2 \\ & + 0.99 I(7) - 0.73 I(7N\text{-CO}) - 1.03 B_4(8)^2 \\ & + 3.72 B_4(8) - 0.21 [\sum \pi(6,7,8)]^2 \\ & - 0.49 \sum \pi(6,7,8) - 0.68 \sum F(6,7,8) - 4.57 \\ (n = 71; s = 0.27; r^2 = 0.93; F = 64.07) \end{aligned}$$



an antibiotic, belonging to the class of fluoroquinolone antibiotics, used to treat urinary tract infections, gynaecological infections, inflammation of the prostate gland, gonorrhoeal and bladder infection.



ISBN: 978-0-12-801505-6

© 2015 Elsevier Inc. All rights reserved

Chem Soc Rev

REVIEW ARTICLE



[View Article Online](#)  
[View Journal](#) | [View Issue](#)



Cite this: *Chem. Soc. Rev.*, 2020, 49, 3525

## QSAR without borders

Eugene N. Muratov, <sup>ab</sup> Jürgen Bajorath, <sup>c</sup> Robert P. Sheridan, <sup>d</sup>  
Igor V. Tetko, <sup>e</sup> Dmitry Filimonov, <sup>f</sup> Vladimir Poroikov, <sup>f</sup> Tudor I. Oprea, <sup>ghi</sup>  
Igor I. Baskin, <sup>jk</sup> Alexandre Varnek, <sup>j</sup> Adrian Roitberg, <sup>l</sup> Olexandr Isayev, <sup>a</sup>  
Stefano Curtalolo, <sup>m</sup> Denis Fourches, <sup>n</sup> Yoram Cohen, <sup>o</sup> Alan Aspuru-Guzik, <sup>p</sup>  
David A. Winkler, <sup>qrst</sup> Dimitris Agrafiotis, <sup>u</sup> Artem Cherkasov <sup>\*v</sup> and  
Alexander Tropsha <sup>\*a</sup>

### Quantitative structure-activity relationships (**QSAR**)

9.1 Introduction to QSAR

9.2 Chemical Information and Descriptors.

9.3 QSAR equations

9.4 Validation of a QSAR model

9.5 3D QSAR

9.6 QSAR Applications

References

### Docking

9.7 Introduction to docking

9.8. Algorithms

9.9 Scoring Functions

9.10 Steps of docking

9.11 Docking software packages.

9.12 Docking Applications

References



Molecular docking represents an important technology for structure-based drug design.

Molecular docking is the study of how two or more molecular structures (e.g., drug and enzyme or protein) fit together.<sup>1</sup>

In a simple definition, docking is a computational technique aimed at the prediction of the most favorable ligand–target spatial configuration and an estimate of the corresponding complex free energy, although accurate scoring methods remain still elusive.

1. Kirkpatrick P. Virtual screening: gliding to success. Nat Rev Drug Disc 2004;3:299.

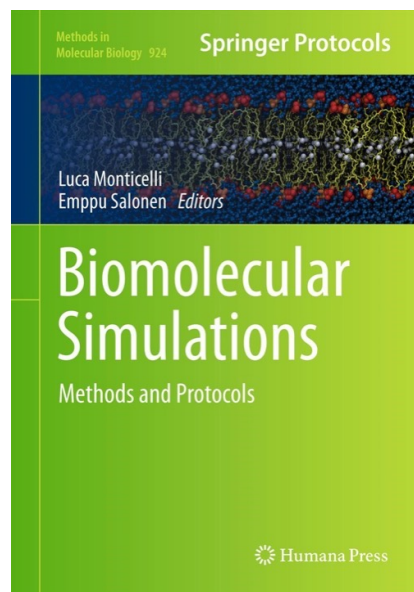
## 9.7 Introduction to protein-ligand interactions



Structure-based drug design docking studies have contributed to the discovery of a number of very important drugs:

Generic name	Manufacturer	Inhibit/Target
AG85, ag337, ag331	Agouron	Thymidylate synthase
Aliskiren	Novartis	Renin inhibitors
Amprenavir	GlaxoSmithKline	HIV protease
Boceprevir	Schering–Plough	Protease inhibitor used for treating hepatitis caused by hepatitis C virus (HCV)
Captopril	Bristol Myers–Squibb	Reversible inhibitor of angiotensin-converting enzyme (ACE)
Dorzolamid	Merck Sharp and Dohme	Carbonic anhydrase (hypercapnic ventilatory failure)
ER $\alpha$ and ER $\beta$	Information not available	Estradiol (E2) analogs
Indinavir	Merck	HIV protease
Inverase	Hoffman La Roche	HIV protease
LY-517717	Lilly/Protherics	Inhibitors of factor Xa serine protease
Nelfinavir	Hoffman La Roche	HIV protease
Nolatrexed dihydrochloride	Agouron	Thymidylate synthase (TS)
Norvir	Abbot	HIV protease
NVP-AUY922	Novartis	Heat shock protein 90 (HSP90)
Raltitrexed	AstraZeneca	Thymidylate
Raltegravir	Merck	HIV integrase
Rupintrivir	Agouron	Irreversible inhibitors of human rhinovirus (HRV) 3C protease
Saquinavir	Hoffman La Roche	HIV protease
TMI-005	—	Dual inhibitor of tumor necrosis factor- $\alpha$ (TNF $\alpha$ ) converting enzyme (TACE) and matrix metalloproteinases (MMPs)
Zanamivir	Gilead Sciences	Neuraminidase inhibitor





### Chapter 13. Molecular Docking Methodologies.

A. Bortolato, M. Fanton, J. S. Mason, S. Moro

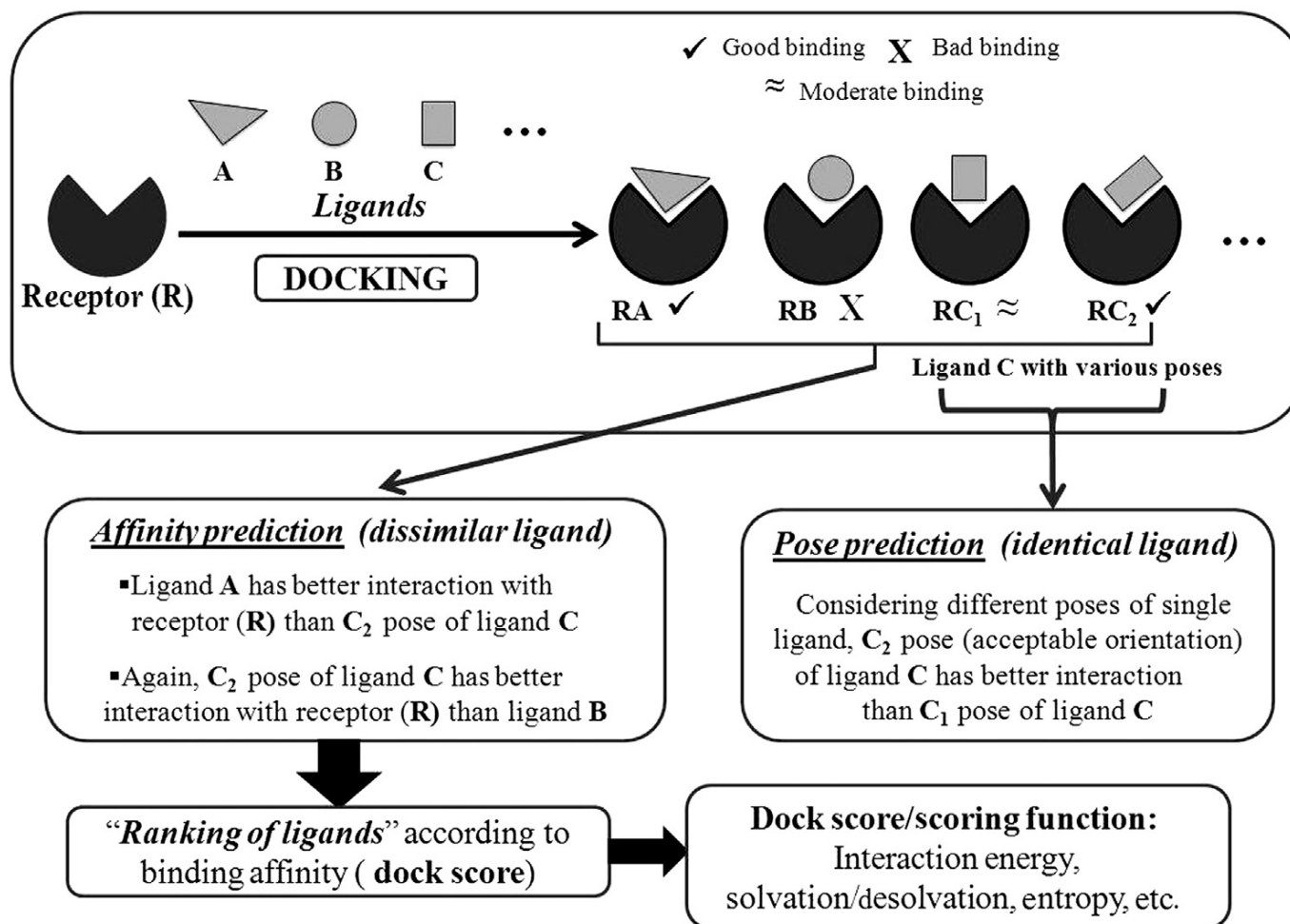
*Target-based methods provide important tools to help the identification of “bioactive regions” of chemical space relevant in drug discovery projects. The challenge is that these are tiny spots not evenly distributed in an almost infinite universe limited only by the chemist’s imagination.<sup>2,3</sup>*

2. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432:855–861

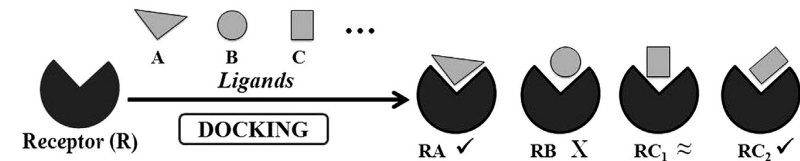
3. Muller G (2003) Medicinal chemistry of target family-directed masterkeys. Drug Discov Today 8:681–691



### Definition of fundamental terms of docking



### Definition of fundamental terms of docking



**Receptor:** a protein molecule or a polymeric structure in or on a cell that distinctively recognizes and binds a molecule (ligand) acting as a molecular messenger. When such ligands bind to a receptor, they cause some kind of cellular response.

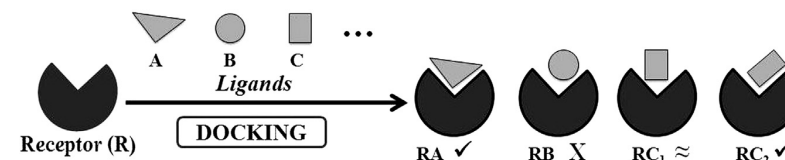
**Ligand:** the complementary partner molecule that binds to the receptor. Ligands are most often small drug molecules, neurotransmitters, hormones, lymphokines, lectins, and antigens, but they could also be another biopolymer or macromolecule (in the case of protein-protein docking).

**Docking:** a molecular modeling technique designed to find the proper fit between a ligand and its binding site (receptor).

**Dock pose:** A ligand molecule can bind with a receptor in a multiple positions, conformations, and orientations. Each such docking mode is called a dock pose.

**Binding mode:** orientation of the ligand relative to the receptor, as well as the conformation of the ligand and receptor when they are bound to each other.

### Definition of fundamental terms of docking

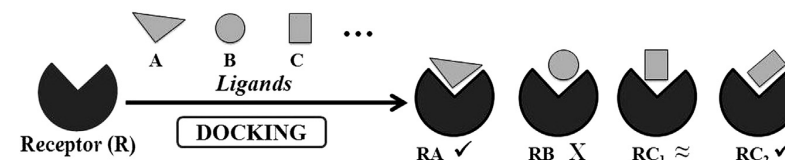


**Ranking:** the process of classifying which ligands are most likely to interact favorably to a particular receptor based on the predicted free energy of binding. After completion of docking, all ligands are consequently ranked by their respective dock scores (i.e., their predicted affinities). This rank-ordered list is then employed for further synthesis and biological investigation only for those compounds that are predicted to be most active.

**Pose prediction:** searching for the accurate binding mode of a ligand, which is typically carried out by performing a number of trials and keeping those poses that are energetically best. It involves finding the correct orientation and the correct conformation of the docked ligand due to their flexible nature.

**Dock score:** The process of evaluating a particular pose by counting the number of favorable intermolecular interactions such as hydrogen bonds and hydrophobic contacts. In order to recognize the energetically most favorable pose, each pose is evaluated based on its compatibility to the target in terms of shape and properties such as electrostatics and generate corresponding dock score.

### Definition of fundamental terms of docking



**Scoring or affinity prediction:** the energetically best pose or number of best poses found for each ligand, and comparing the affinity scores for different ligands give their relative rank ordering.

**Scoring functions** are generally divided into two main groups. One main group comprises knowledge-based scoring functions that are derived using statistics for the observed interatomic contact frequencies, distances, or both in a large database of crystal structures of protein-ligand complexes. The other group contains scoring schemes based on physical interaction terms. These so-called energy component methods are based on the assumption that the change in free energy upon binding of a ligand to its target can be decomposed into a sum of individual contributions:

$$\Delta G_{\text{bind}} = \Delta G_{\text{int}} + \Delta G_{\text{solv}} + \Delta G_{\text{conf}} + \Delta G_{\text{motion}}$$

$\Delta G_{\text{int}}$  = specific ligand-receptor interactions,

$\Delta G_{\text{solv}}$  = interactions of ligand and receptor with solvent,

$\Delta G_{\text{conf}}$  = conformational changes in the ligand and the receptor,

$\Delta G_{\text{motion}}$  = motions in the protein and the ligand during the complex formation.

### Essential requirements of docking

1. Receptor crystal structures → PDB

Identification of the bound-ligand in the cocrystal structure and the knowledge about its interaction with the corresponding protein's amino acid residues are very important before starting a docking study.

→ Receptor *homology modeling* and *threading* techniques

2. A set of ligands of interest → PDB (mol2 format);  
QM calculations;  
Data Banks (PubChem, ...);



**docking can be categorized into three main classes:**

1) protein-protein docking

2) protein-nucleic acid docking

3) protein-ligand docking → **structure-based drug design**



One of the main problems is the tremendous **complexity of the systems**: there are hundreds of thousands of degrees of freedom that have to be analyzed. Furthermore, the forces acting in the binding process are not exactly known yet, or are very difficult to calculate. We need the use of **algorithms for exploring conformations and orientations of a ligand with respect to receptor (protein)**. Ideally they should explore all degrees of freedom of the protein-ligand complex to ensure that all binding modes are included. In addition, two of the essential characteristics are **efficiency** and **speed**.

Algorithms to search for conformations can be categorized in three groups:



- a) **rigid-body docking**, where both the receptor and ligand are treated as rigid;
- b) **flexible ligand docking**, where the receptor is held rigid, but the ligand is treated as flexible;
- c) **flexible docking**, where both receptor and ligand flexibility is considered.

the most  
commonly  
used, but....

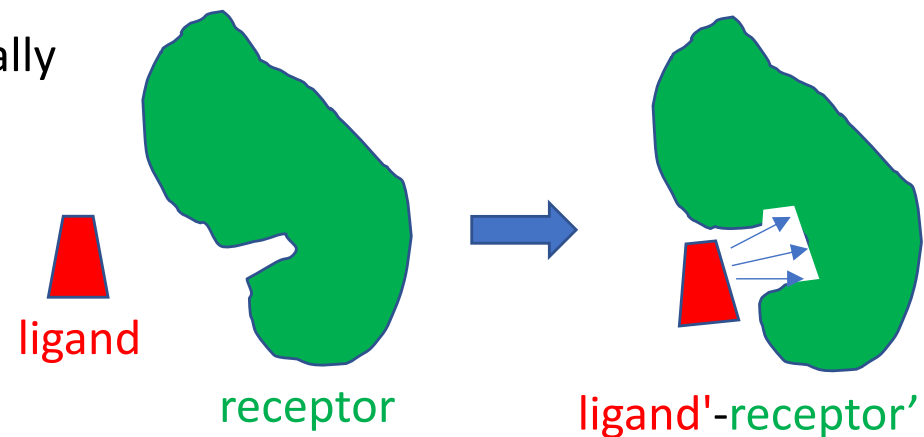
Ligand binding usually induces **protein conformational changes** or **induced fit** upon ligand binding in order to maximize energetically favorable interactions with the ligand

↓  
models in between b) and c)



Different degrees of receptor (protein) flexibility:

- (i) soft docking, → allows flexibility of the receptor and ligand by using a relaxed representation of the molecular surface
- (ii) side-chain flexibility,
- (iii) molecular relaxation,
- (iv) protein ensemble docking

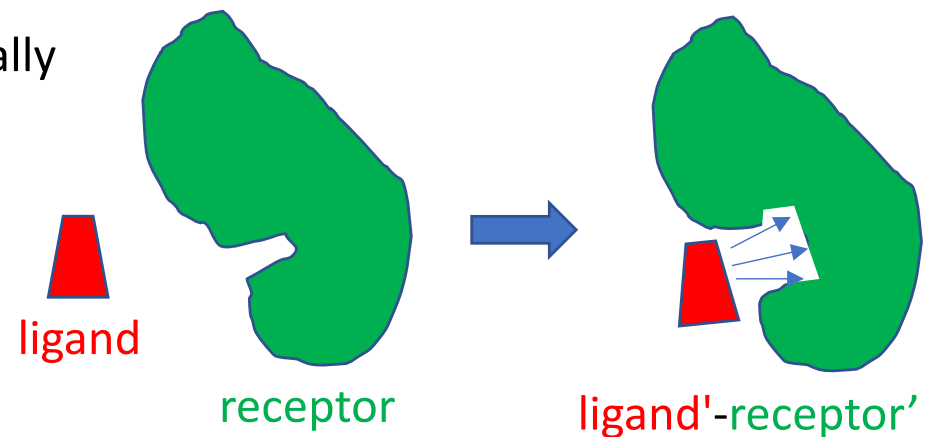





Ligand binding usually induces **protein conformational changes** or **induced fit** upon ligand binding in order to maximize energetically favorable interactions with the ligand



models in between b) and c)



Different degrees of receptor (protein) flexibility:

- (i) soft docking,
- (ii) side-chain flexibility,  allows active site side-chain flexibility
- (iii) molecular relaxation,
- (iv) protein ensemble docking

Ligand binding usually induces **protein conformational changes** or **induced fit** upon ligand binding in order to maximize energetically favorable interactions with the ligand

↓  
models in between b) and c)

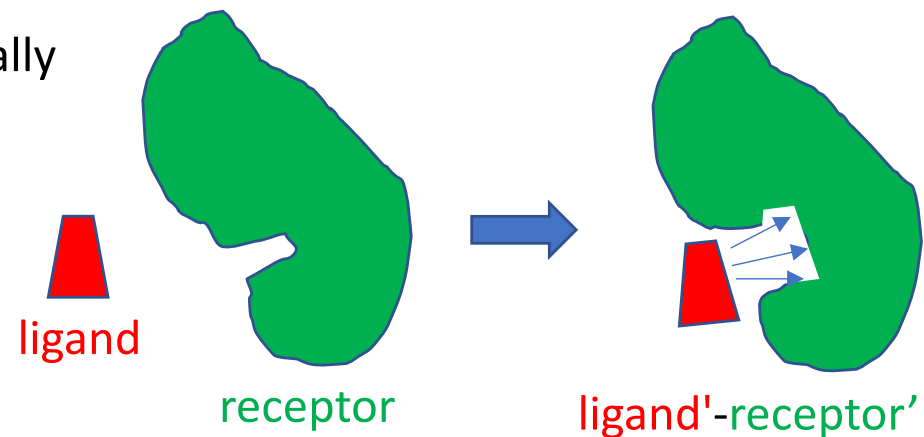


Different degrees of receptor (protein) flexibility:

- (i) soft docking,
- (ii) side-chain flexibility,
- (iii) molecular relaxation,
- (iv) protein ensemble docking



protein flexibility by first using rigid-body docking to place the ligand into the binding site and then relaxing the protein backbone and side-chain atoms nearby by MC, MD,...



Ligand binding usually induces **protein conformational changes** or **induced fit** upon ligand binding in order to maximize energetically favorable interactions with the ligand



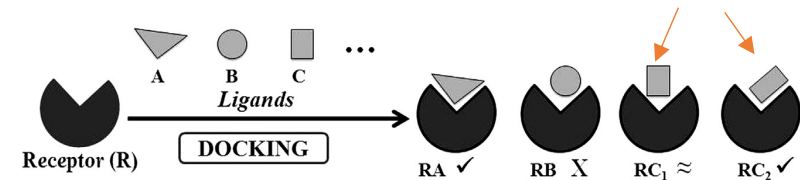
Different degrees of receptor (protein) flexibility:

- (i) soft docking,
- (ii) side-chain flexibility,
- (iii) molecular relaxation,
- (iv) protein ensemble docking



ensemble of protein structures to represent different possible conformational changes using MD, MC simulations, or experimentally from NMR or X-ray crystal structures

### ligand sampling and flexibility;



Given a protein target, the sampling algorithm generates possible ligand orientations or conformations (**poses**) around the selected binding site of the protein.

#### b.1) **shape matching**,

The ligand is placed using the criterion that the molecular surface of the placed ligand must harmonize the molecular surface of the binding site on the protein.

The conformation of the ligand is normally fixed during shape matching

#### b.2) **systematic search**,

flexible ligand docking, which create all the probable ligand binding conformations by exploring all degrees of freedom of the ligand.

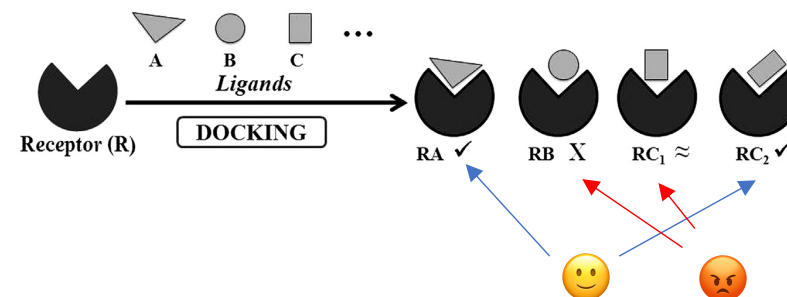
#### b.3) **stochastic algorithms**.

ligand-binding orientations and conformations are sampled by making random changes to the ligand at each step in the conformational space and the translational and rotational space of the ligand, respectively, by MC, MD, ...

## 9.9 Scoring functions



it is the fundamental element behind determining the accuracy of a protein-ligand docking algorithm



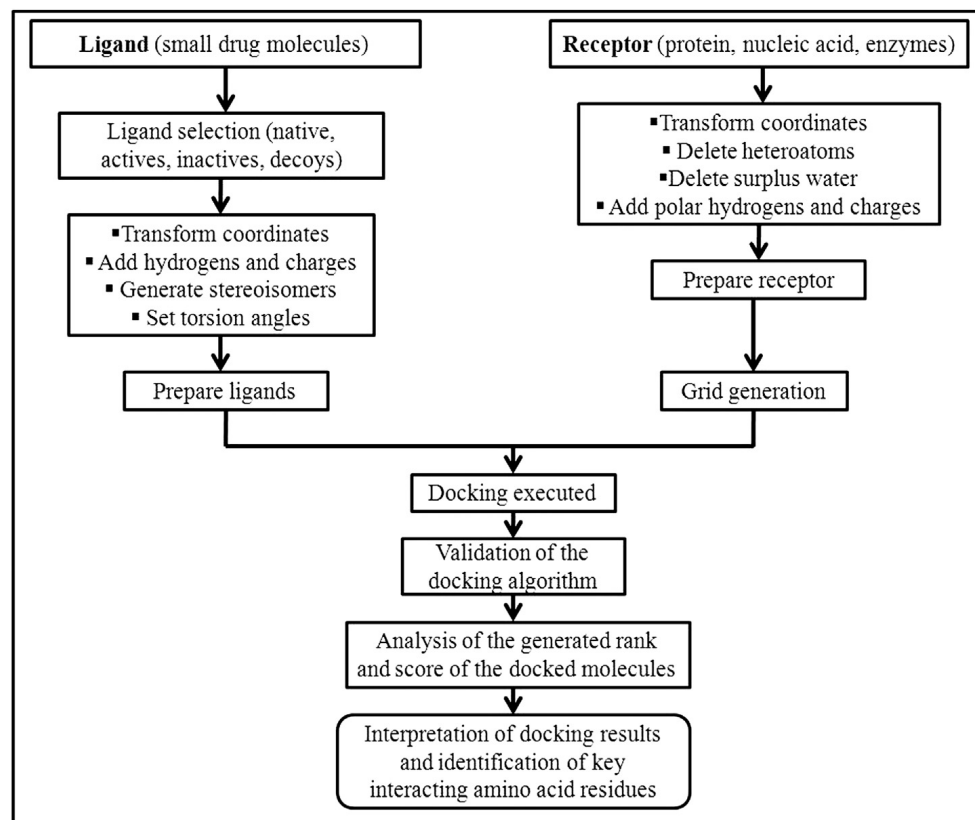
They represent the attempt to catalog and evaluate the different conformations of the ligand predicted by the search algorithm, in such a way that it allows us to discriminate possible bindings from others that are not.

The choice of the scoring function is fundamental in the binding method, since a rigorous description of the process is usually very computationally expensive and simplifications must be chosen ...



- 1) **FF scoring functions** : partitioning of the ligand-binding energy into individual interaction terms such as VDW energies, electrostatic energies, and bond stretching/bending/torsional energies, employing a set of derived FF parameters
- 2) **Empirical scoring functions** : The binding energy score is calculated by adding up a set of weighted empirical energy terms (such as VDW energy, electrostatic energy, hydrogen-bonding energy, desolvation term, entropy term, and hydrophobicity term)
- 3) **Knowledge-based scoring functions** : result from the structural information in experimentally determined protein-ligand complexes. The theory beneath the knowledge-based scoring functions is the PMF
- 4) **Consensus scoring** : it advances the probability of finding an accurate solution by amalgamating the scoring information from multiple scoring functions in anticipation of eliminating the inaccuracies of the individual scoring functions.
- 5) **clustering-based scoring method** : which includes the entropic effects by dividing generated ligand-binding modes into different clusters, calculated by the configurational space covered by the ligand poses or the number of ligand poses in the cluster

## 9.10 Steps of docking



1st. **Ligand preparation**: in this step all the duplicate structures should be removed, and options for ionization change, tautomer, isomer generation, and 3D generator must be set in the working software platform for the respective ligands.

2nd. **Protein preparation**: pKa, add hydrogens, contraions, waters...

3rd **Ligand-protein docking**: a grid box is generally generated at the centroid of the ligand bound to the active site of the receptor, but not always (**blind docking**). A set number of ligand poses should be saved for each conformation of the ligand and ranked according to their dock score function, and then their interaction with the receptor can be analyzed.

## 9.11 Docking software packages



Software	Algorithm and remarks	Software	Algorithm and remarks
AutoDock	AutoDock is a suite of automated docking tools capable of predicting how small molecules, such as substrates or drug candidates, bind to a receptor of a known 3D structure. The Lamarckian GA is used as the algorithm. Website: <a href="http://autodock.scripps.edu/">http://autodock.scripps.edu/</a>	FRED	The shape matching (Gaussian functions) algorithm is employed in the Fast Rigid Exhaustive Docking (FRED) software.
Discovery Studio	The conformational search of the ligand poses is performed by the MC trial method. Preprocessing of ligands is performed using the ligand fit program with selecting one of the energy grid out of three energy grids (PLP1, Dreiding, and CFF) available in Discovery Studio. The docking poses saved for each conformation of the compound are ranked according to their dock scores based on LigScore1, LigScore2, PLP1, PLP2, Jain, and PMF function. Website: <a href="http://accelrys.com/">http://accelrys.com/</a>	FTDock	Fourier Transform Docking (FTDock) is a free program that performs rigid-body docking on two biomolecules in order to predict their correct binding geometry.
DOCK	DOCK is a program that can examine possible binding orientations of protein–protein and protein–DNA complexes. It can be used to search databases of molecular structures for compounds that act as enzyme inhibitors or bind to target receptors. The shape matching (sphere images) algorithm is employed here. Website: <a href="http://www.cmpchem.ucsf.edu/kuntz/dock.html">http://www.cmpchem.ucsf.edu/kuntz/dock.html</a>	Glide	Glide is a fast and accurate docking program that addresses a number of problems, ranging from fast database screening to highly accurate docking. The descriptor matching/MC is the principal algorithm of Glide. The hierarchical filters in Glide ensure a fast and efficient reduction of large data sets to the few drug candidates that bind best with the target. Website: <a href="http://www.schrodinger.com/Glide/">http://www.schrodinger.com/Glide/</a>
DOT	Daughter Of Turnip (DOT) is a program for docking macromolecules to other molecules of any size. It can predict binding modes of small molecule–protein complexes. The intermolecular energies for all configurations generated by this search are calculated as the sum of electrostatic and VDW energies. Website: <a href="http://www.sdsc.edu/CCMS/DOT/">http://www.sdsc.edu/CCMS/DOT/</a>	GOLD	GOLD is a GA-based method for ligand protein docking. GOLD accounts for receptor flexibility through side-chain flexibility and, most important, ensemble docking. Website: <a href="http://www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx">http://www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx</a>
FADE and PADRE	Fast Atomic Density Evaluator (FADE) and Pairwise Atomic Density Reverse Engineering (PADRE) programs are designed to aid in the molecular modeling of proteins. In particular, the programs can rapidly elucidate features of interest such as crevices, grooves, and protrusions. The topographical information produced by FADE and PADRE can help researchers easily pinpoint the most prominent features of a protein, regions that are likely to participate in interactions with other molecules. In addition, it provides shape descriptors to aid in analyzing single molecules.	GRAMM	Global Range Molecular Matching (GRAMM) is a free program for protein docking. To predict the structure of a complex, it requires only the atomic coordinates of the two molecules (no information of the binding sites is needed). The molecular pairs may be two proteins, a protein and a smaller compound, two transmembrane helices, etc. The program performs an exhaustive 6D search through the relative translations and rotations of the molecules. Website: <a href="http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/">http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/</a>
FlexiDock	FlexiDock is a commercial software performs flexible docking of ligands into receptor binding sites. Website: <a href="http://www.tripos.com/software/fdock.html">http://www.tripos.com/software/fdock.html</a>	Hammerhead	Hammerhead is suitable for screening large databases of flexible molecules by binding to a protein of known structure. The approach is completely automated, from the elucidation of protein binding sites, through the docking of molecules, to the final selection of compounds.
FlexX	Incremental construction algorithm is employed in FlexX. The FlexX predicts the geometry of the protein–ligand complex and estimates the binding affinity. The two main applications of FlexX are complex prediction and VS. Complex prediction is used, when one have a protein and a small molecule binding to it but no structure of the protein–ligand complex is available. Website: <a href="http://www.biosolveit.de/flexx/">http://www.biosolveit.de/flexx/</a>	HINT	HINT is a software package that utilizes experimental solvent partitioning data as a basis for an empirical molecular interaction model. The program calculates empirical atom-based hydrophobic parameters that, in a sense, encode all significant intermolecular and intramolecular noncovalent interactions implicated in drug binding or protein folding.
		Liaison	Liaison is a commercial program for fast estimation of free energy of binding between a receptor and a ligand. The free energy of binding can be approximated by an equation in which only the free and bound states of the ligand are calculated. The method combines high-level molecular mechanics calculations with experimental data to build a scoring function for the evaluation of ligand–receptor binding free energies.
		LigandFit	The shape matching (moments of inertia) algorithm is employed.



## 9.11 Docking software packages



Software	Algorithm and remarks
MOE	MOE is a fast and accurate docking program. The dock poses were ranked according to the GBVI/WSA binding free-energy calculation and minimized using MMFF94x within a rigid receptor.
Molegro Virtual Docker	Molegro Virtual Docker is an integrated platform for predicting protein–ligand interactions. Molegro Virtual Docker handles all aspects of the docking process, from preparation of the molecules to determination of the potential binding sites of the target protein, and prediction of the binding modes of the ligands.
QSite	QSite is a mixed-mode QM/MM program for highly accurate energy calculations of protein–ligand interactions in the active site. The program is specifically designed for proteins and allows a number of different QM/MM boundaries for residues in the active site. QSite uses the power and speed of Jaguar to perform the quantum mechanical part of the calculations and OPLS-AA to perform the molecular mechanical part of the calculations.
Situs	Situs is a program package for the docking of protein crystal structures to single-molecule, low-resolution maps from electron microscopy or small-angle X-ray scattering.
SLIDE SuperStar	Descriptor matching algorithm is employed in SLIDE. SuperStar is a program for generating maps of interaction sites in proteins using experimental information about intermolecular interactions. The generated interaction maps are therefore fully knowledge-based. SuperStar retrieves its data from IsoStar, CCDC interaction database. IsoStar contains information about nonbonded interactions from both the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB).



The docking technology is successfully applied at multiple stages of the drug design and discovery process for three main purposes:

- 1) predicting the binding mode of a known active ligand,
- 2) identifying new ligands using VS,
- 3) predicting the binding affinities of allied compounds from a known active series.



The major specific applications of docking:

- the determination of the lowest free-energy structures for the receptor-ligand complex;
- calculation of the differential binding of a ligand to two different macromolecular receptors
- study of the geometry of a particular ligand-receptor complex.
- searching of a database and ranking of hits for lead generation and optimization for future drug candidate.
- to propose the modification of lead molecules to optimize potency or other properties.
- library design and data bank generation.
- screening for the side effects that can be caused by interactions with proteins,
- to check the specificity of a potential drug against homologous
- predicting protein-protein interactions.
- to create knowledge of the molecular association, which aids in understanding a variety of pathways taking place in the living system.
- **to reveal/design possible potential pharmacological targets.**

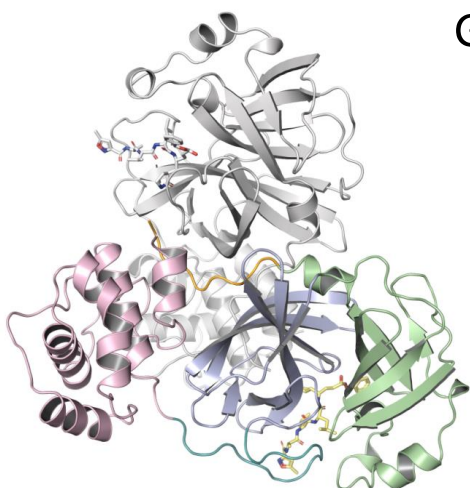
Journal of Chemical Information and Modeling 2021, in press

### **Benchmarking the ability of common docking programs to correctly reproduce and score binding modes in SARS-CoV-2 protease M<sup>pro</sup>**

Shani Zev, Keren Raz, Renana Schwartz, Reem Tarabeh, Prashant Kumar Gupta, Dan T. Major

testing the ability of several leading docking programs...

Glide, DOCK, AutoDock Vina, FRED and EnzyDock



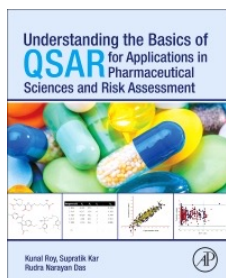
To correctly identify and score the binding mode of SARS CoV-2 M<sup>pro</sup> ligands in 193 crystal structures

Journal of Chemical Information and Modeling 2021, in press

### **Benchmarking the ability of common docking programs to correctly reproduce and score binding modes in SARS-CoV-2 protease M<sup>pro</sup>**

None of the codes are able to correctly identify the crystal structure in more than 36% of the cases for non-covalently bound ligands (Glide top performer), whereas for covalently bound ligands the top score was 45% (EnzyDock). These results suggest that one should perform in silico campaigns of M<sup>pro</sup> with care, and that more comprehensive strategies including ligand free energy perturbation might be necessary in conjunction with virtual screening and docking.

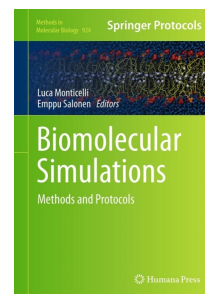
## Books:



Chapter 10. "Other related techniques"

ISBN: 978-0-12-801505-6

© 2015 Elsevier Inc. All rights reserved



Chapter 13. "Molecular Docking Methodologies"  
A. Bortolato, M. Fanton, J. S. Mason, S. Moro

ISBN 978-1-62703-016-8

© Springer Science+Business Media New York 2013

## Reviews:

Douglas B. Kitchen, Hélène Decornez, John R. Furr and Jürgen Bajorath "Docking and Scoring in Virtual Screening for drug Discovery: Methods and Applications" *Nature Reviews* 3, 935-949, 2004

a) R.A. Friesner et al. "Glide.." *J. Med. Chem.* 2004, 47, 1739-1749; b) R.A. Friesner et al. "Glide.." *J. Med. Chem.* 2004, 47, 1750-1759

Paul C. D. Hawkins, Gregory L. Warren, A. Geoffrey Skillman, Anthony Nicholls "How to do an evaluation: pitfalls and traps" *J Comput Aided Mol Des* (2008) 22:179–190

Cameron Mura, Charles E. McAnany "An Introduction to Biomolecular Simulations and Docking" *Molecular Simulation* (2014; special issue on simulations in molecular biology)

## Applications:

Donatella Callegari , ... Alessio Lodola "Comparative Analysis of Virtual Screening Approaches in the Search for Novel EphA2 Receptor Antagonists" *Molecules* 2015, 20, 17132-17151

Moustafa Gabr, Katarzyna Świderek "Discovery of a Histidine-Based Scaffold as an Inhibitor of Gut Microbial Choline Trimethylamine-Lyase" *ChemMedChem* 2020, 15, 2273 – 2279

Shani Zev, Keren Raz, Renana Schwartz, Reem Tarabeh, Prashant Kumar Gupta, Dan T. Major " Benchmarking the ability of common docking programs to correctly reproduce and score binding modes in SARS-CoV-2 protease Mpro" *Journal of Chemical Information and Modeling* 2021, in press