Quantitative Structure Activity Relationship (QSAR) is involved in building mathematical models for correlating molecular structures with molecular properties. In this section we introduce the notion of molecular descriptors and present the QSAR model and its validation.

Author(s):   Hanoch Senderowitz (Predix Pharmaceutical), Claude Cohen (Synergix)

Prerequisites: None

Number of Pages: 191 (194 Screens)

Last updated: April 2004

🔊 Voice: available

The topic Introduction to QSAR contains the following 20 pages:
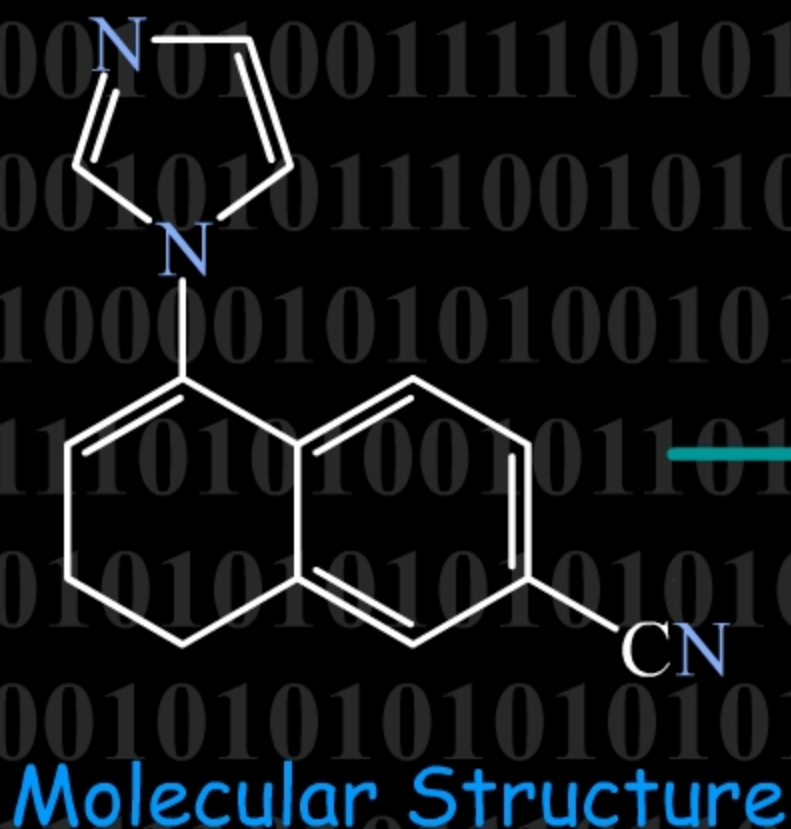
- Molecular Structure and Molecular Properties
    - Structure-Property Relationships: Example 1
    - Structure-Property Relationships: Example 2
    - Structure-Property Relationships: Example 3
- What is QSAR?
- What is QSPR?
- Focus on a Single Property at a Time
- Molecular Descriptors
- Examples of Molecular Descriptors
- The QSAR Equations
- Types of Molecular Descriptors
    - Molecular Descriptors: 1D
    - Molecular Descriptors: 2D

...

For the entire list, see the navigation panel.

F1.1.1 Molecular Structure and Molecular Properties

One of the most pervasive postulates in the life sciences is that all molecular properties are coded by and consequently result from molecular structure. Some examples of structure-property relationships are illustrated on the following pages.

Molecular Structure

- biological properties
- chemical properties
- physical properties
- electronical properties
- etc...

Paracetamol selectively inhibits the cyclooxygenase enzyme COX-3 found in the brain and spinal cord and consequently relieves pain and reduces fever.
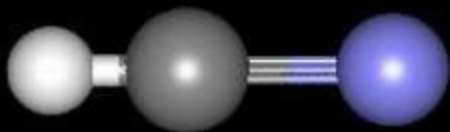
Structure

Property

Paracetamol

Relives Pain

F1.1.3 Structure-Property Relationships: Example 2

Cyanide exerts its toxicity by inhibiting cytochrome-c oxidase, the terminal enzyme of the respiratory chain, leading to insufficient utilization of oxygen and suffocation. Inhibition occurs through binding to the ferric ion of the cytochrome.

Saccharin (usually sold as sodium saccharin) binds to the sweet taste T1R3 receptor located in the plasma membrane of the sweet-taste sensory cells located in the taste buds. Binding of saccharin to T1R3 initiates a cascade of events in the taste-sensory cell that eventually releases a signaling molecule to an adjoining sensory neuron, causing the neuron to send impulses to the brain. In the brain, these signals cause the actual sensation of sweetness.

Structure
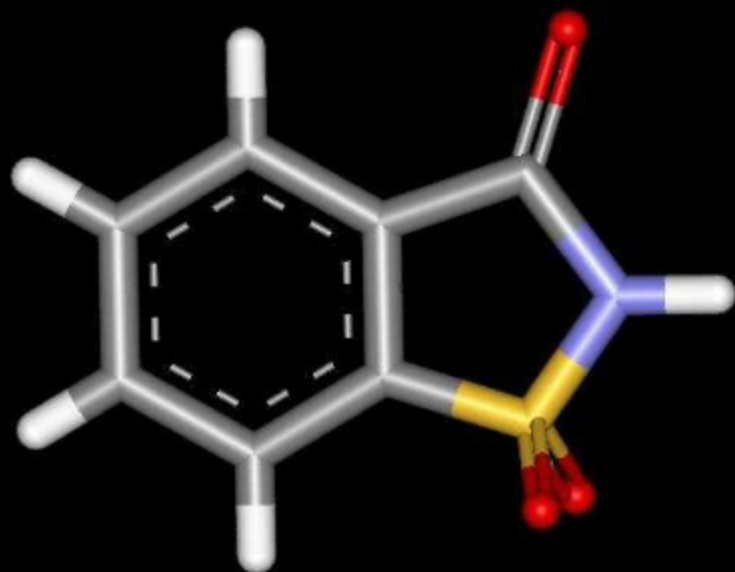
Saccharin

Property

Sweet Taste

F1.1.5 What is QSAR?

Molecules exert their biological effect by binding to their respective receptors, a phenomenon that in turn is governed by their molecular structures (and the molecular structure of the receptor). QSAR (Quantitative Structure Activity Relationship) attempts to formulate the relationship between structure and activity as a mathematical model.

$$\text{Biological effect} = f(\text{Molecular Structure})$$

Quantitative Structure Activity Relationships

F1.1.6 What is QSPR?

The biological effect is just one example of molecular properties. QSPR (Quantitative Structure Property Relationship) is an extension of QSAR and is designed to formulate the relationship between structure and any molecular property as a mathematical model. Other properties include for example: solubility, oral bioavailability, metabolic stability and cell permeability.
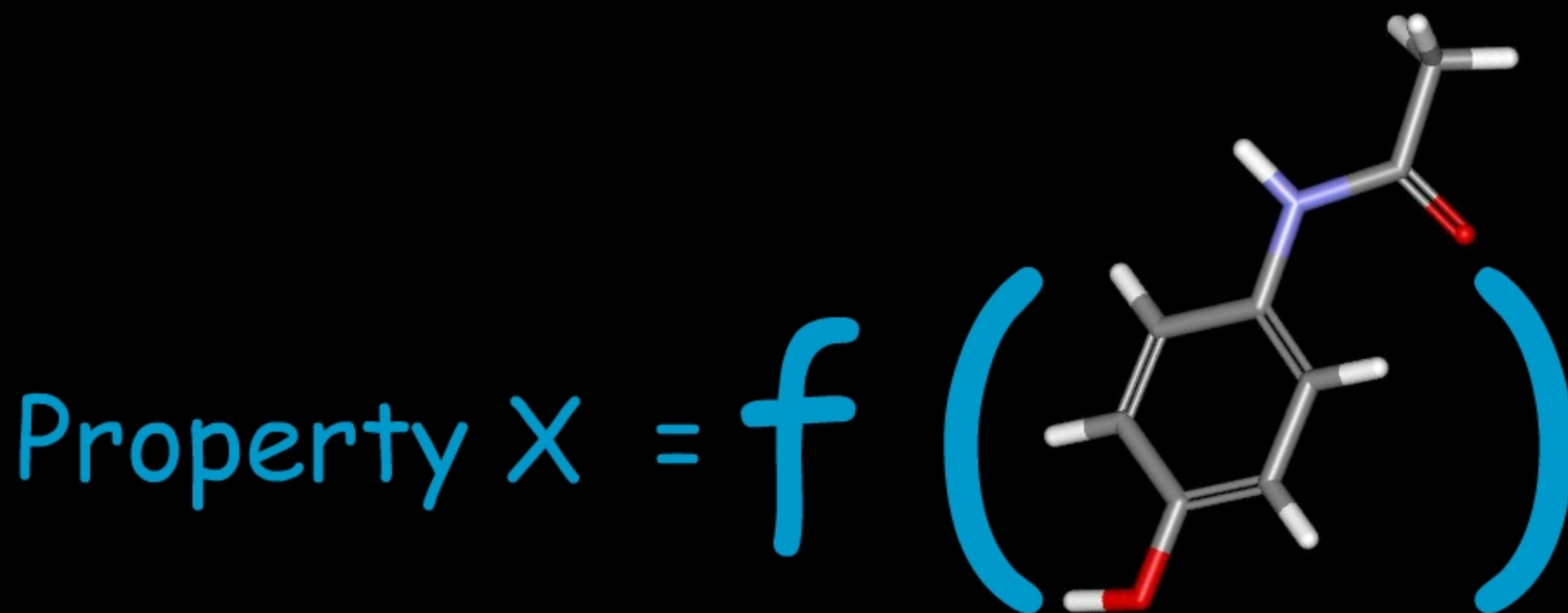
$$\text{Property X} = f\left(\quad \right)$$

F1.1.7 Focus on a Single Property at a Time

No single QSPR model can capture the direct connection between all the properties of a compound and its molecular structure; only a single property is handled at a time.
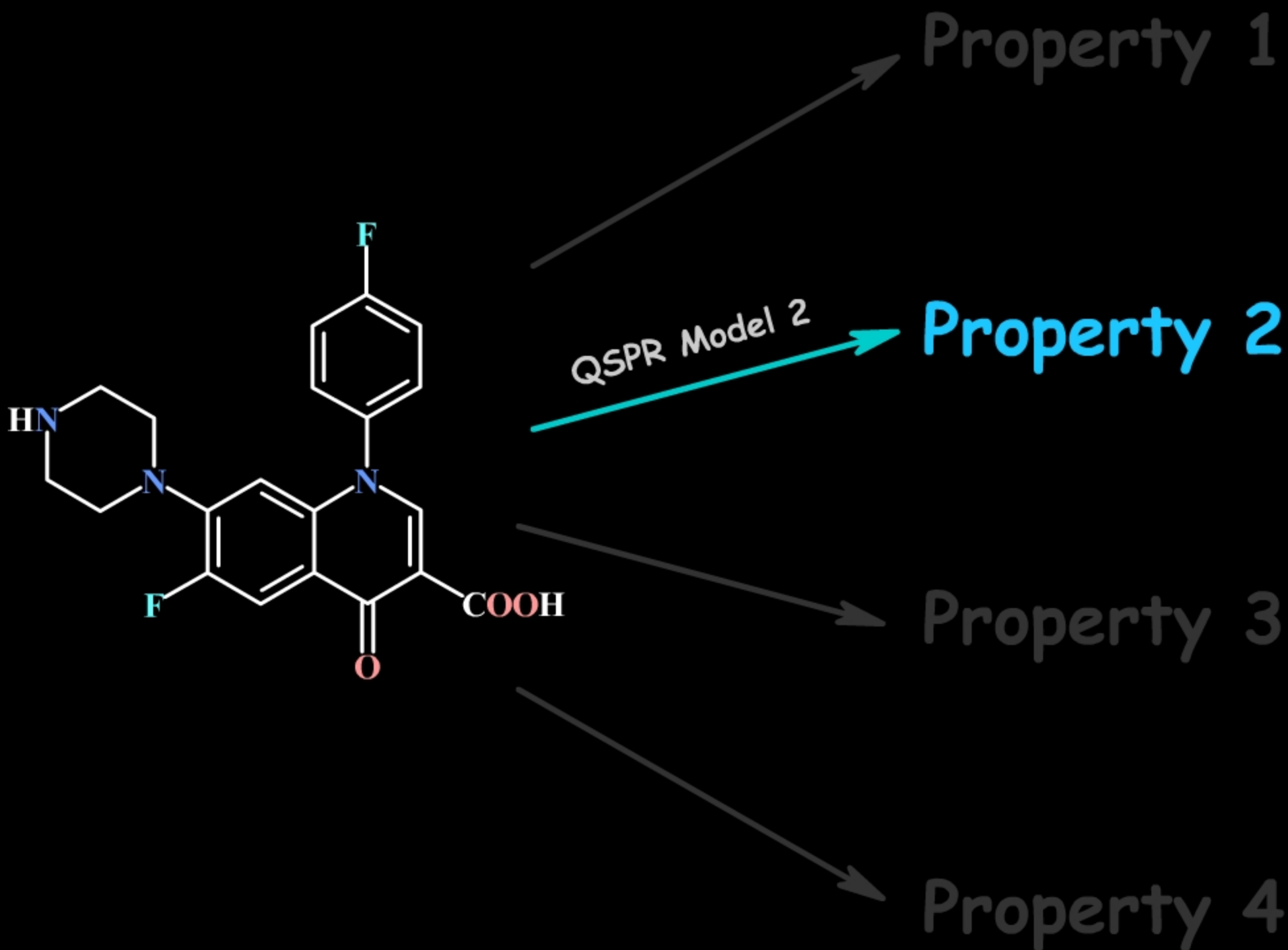
F1.1.8 Molecular Descriptors

Thus, the derivation of a direct relation with the molecular structure of one single property is extremely challenging. However, structural factors known as molecular descriptors that influence the molecular property can be identified. For this reason, the QSAR model correlates the property with molecular descriptors.
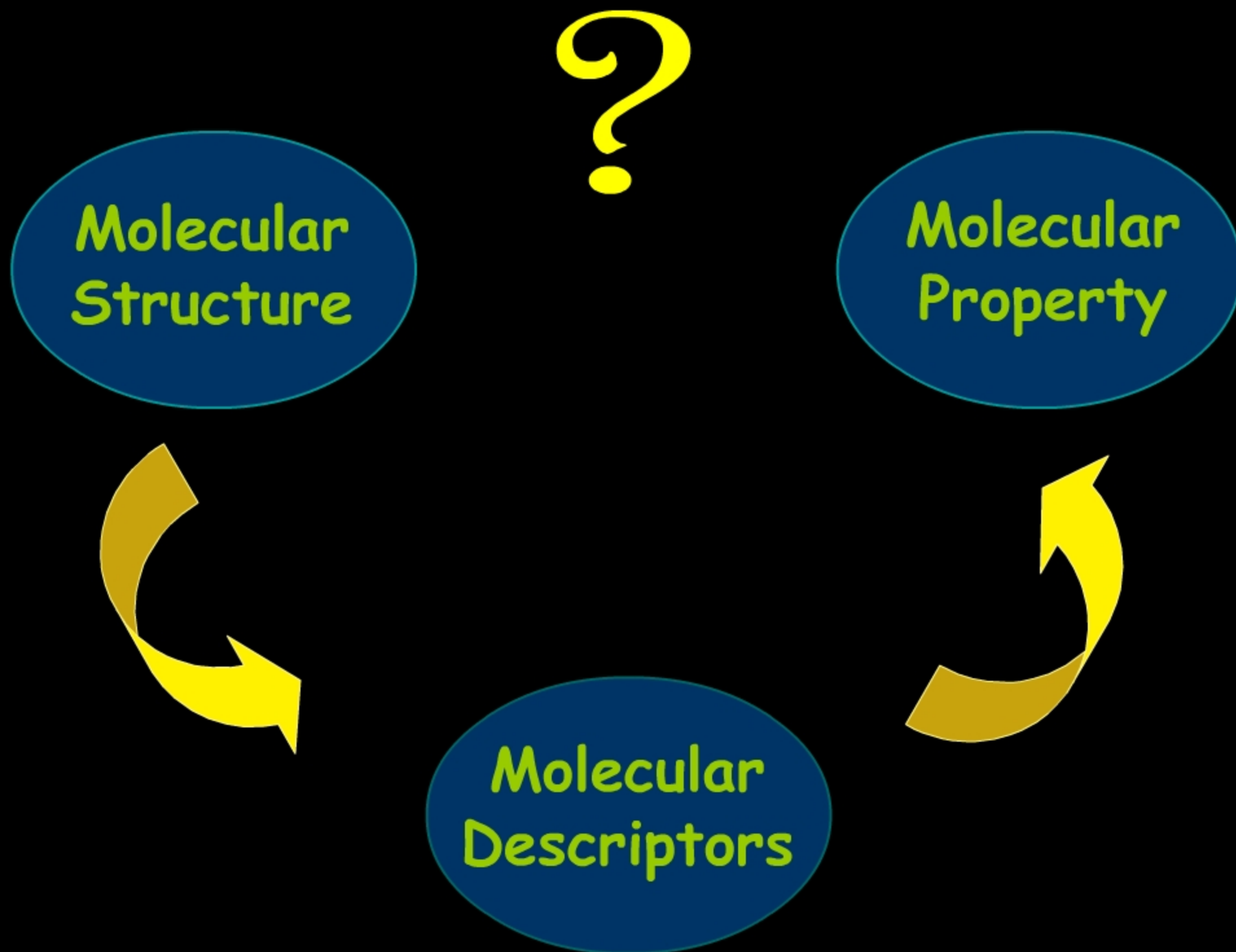
F1.1.9 Examples of Molecular Descriptors

Examples of molecular properties with their associated descriptors are listed in the following table. Later on in this chapter the nature and the meaning of some QSAR descriptors are presented.

Molecular Property	Descriptors
Lipophilicity	π, log P, R_M, f
Steric Properties	E_s, MR, MV, parachor
Electronic Properties	σ, R, F

F1.1.10 The QSAR Equations

All QSAR equations have a molecular property expressed as a function of specific descriptors. They differ in terms of the property they are attempting to correlate, the descriptors they use and the mathematical expression of the model.

$$\text{Oral bioavailability} = f_1 (\text{descriptors set1})$$

$$\text{Cell permeability} = f_2 (\text{descriptors set2})$$

$$\text{Toxicity} = f_3 (\text{descriptors set3})$$

$$\text{Metabolic stability} = f_4 (\text{descriptors set4})$$

$$\text{Receptor binding} = f_5 (\text{descriptors set5})$$

F1.1.11 Types of Molecular Descriptors

Molecular descriptors can be classified according to the dimensionality of the molecular structure from which they are derived. 1D descriptors are derived from the chemical formula, 2D descriptors are derived from a 2D (chemdraw-like) structure and 3D descriptors are derived from the 3-dimensional structure.

Saccharin

1D   2D   3D

$C_7H_5NO_3S$
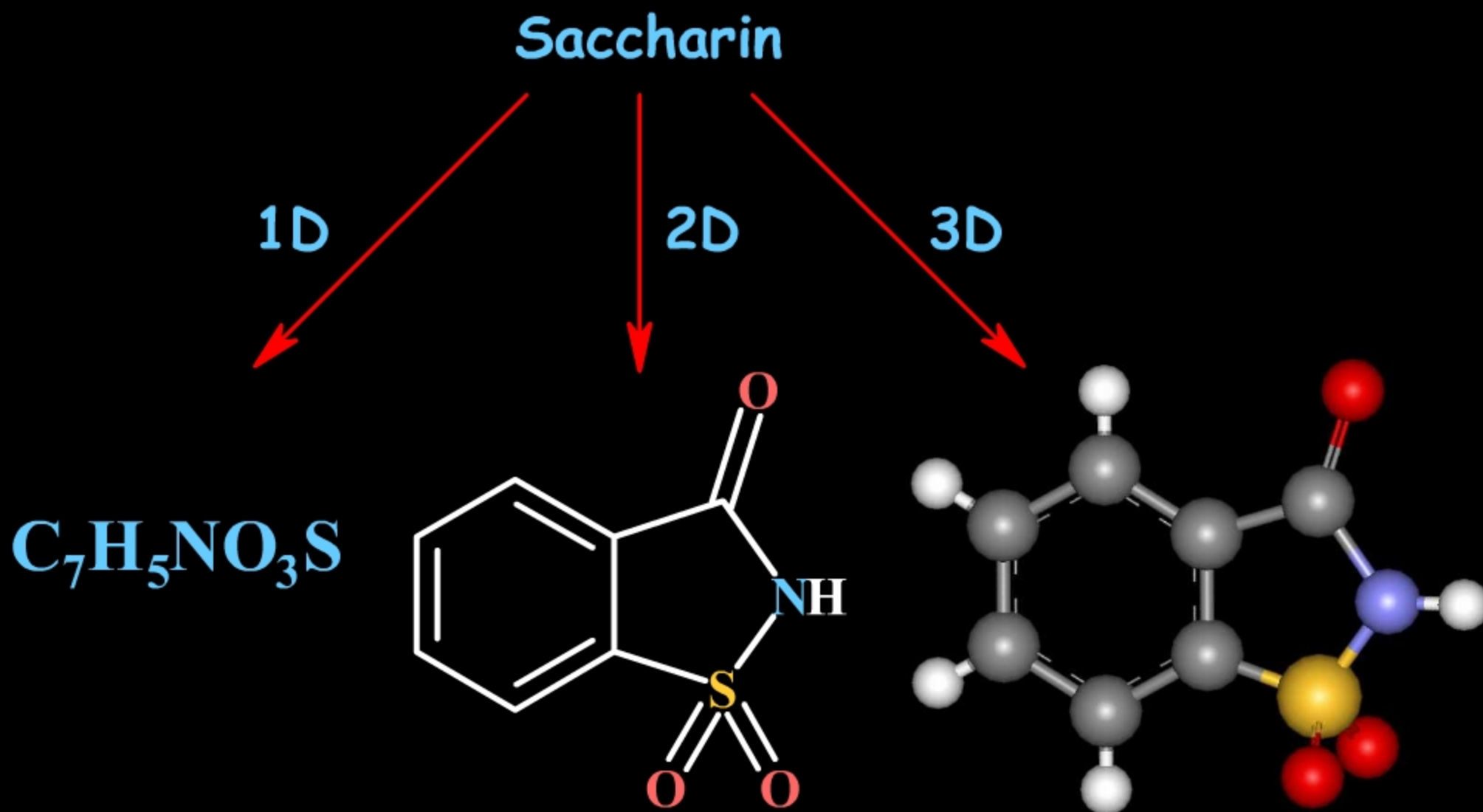
F1.1.12 Molecular Descriptors: 1D

The chemical formula constitutes a 1-Dimensional representation of the molecular structure from which 1D descriptors can be derived. Such descriptors are based exclusively on the type of atoms which make up the molecule.
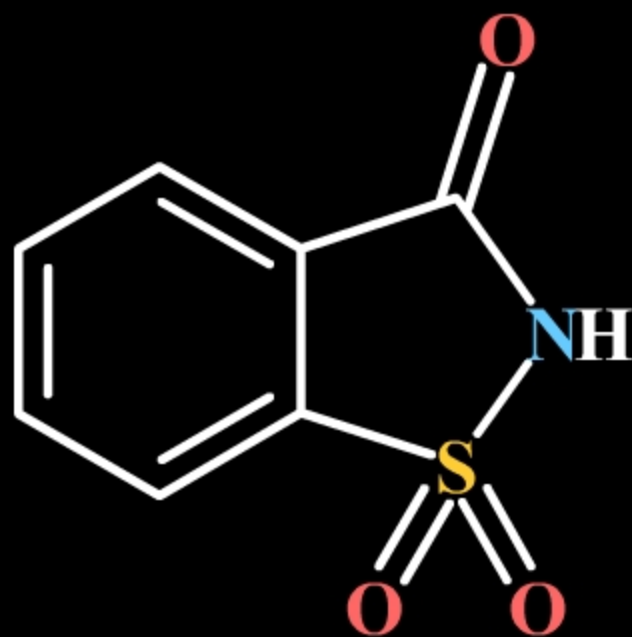
$$C_7H_5NO_3S$$

Nitrogen Atoms: 1

Oxygen Atoms: 3

Molecular Weight (gr/mol): 183.2

F1.1.13 Molecular Descriptors: 2D

A Chemdraw-like structure constitutes a 2-Dimensional representation of the molecular structure from which 2D descriptors can be calculated. In addition to types of atoms, 2D descriptors also incorporate the bonding pattern of the molecule.
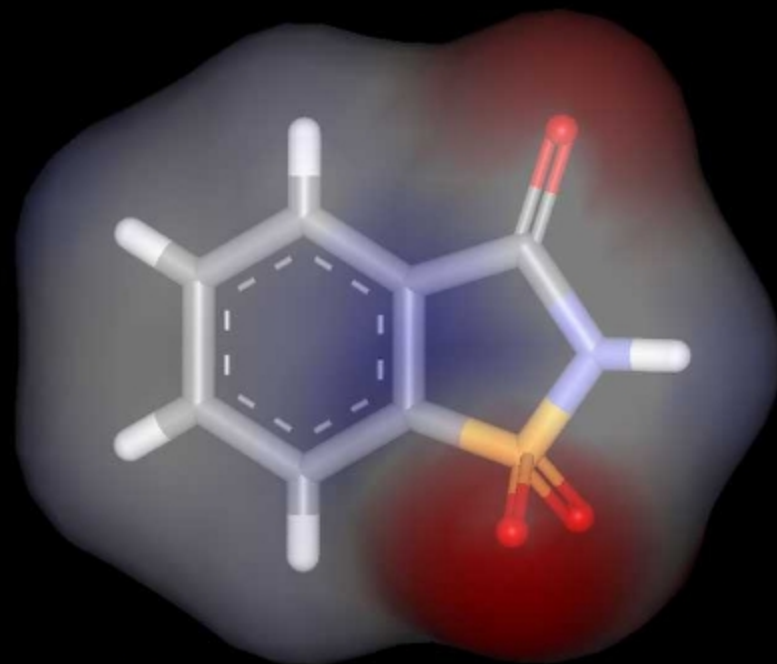

H-bond acceptors: 3

Rotatable bonds: 0

H-bond donors: 1

Number of rings: 2

F1.1.14 Molecular Descriptors: 3D

3D descriptors derived from a 3D molecular structure take the spatial arrangement of the atoms in the molecule into account.
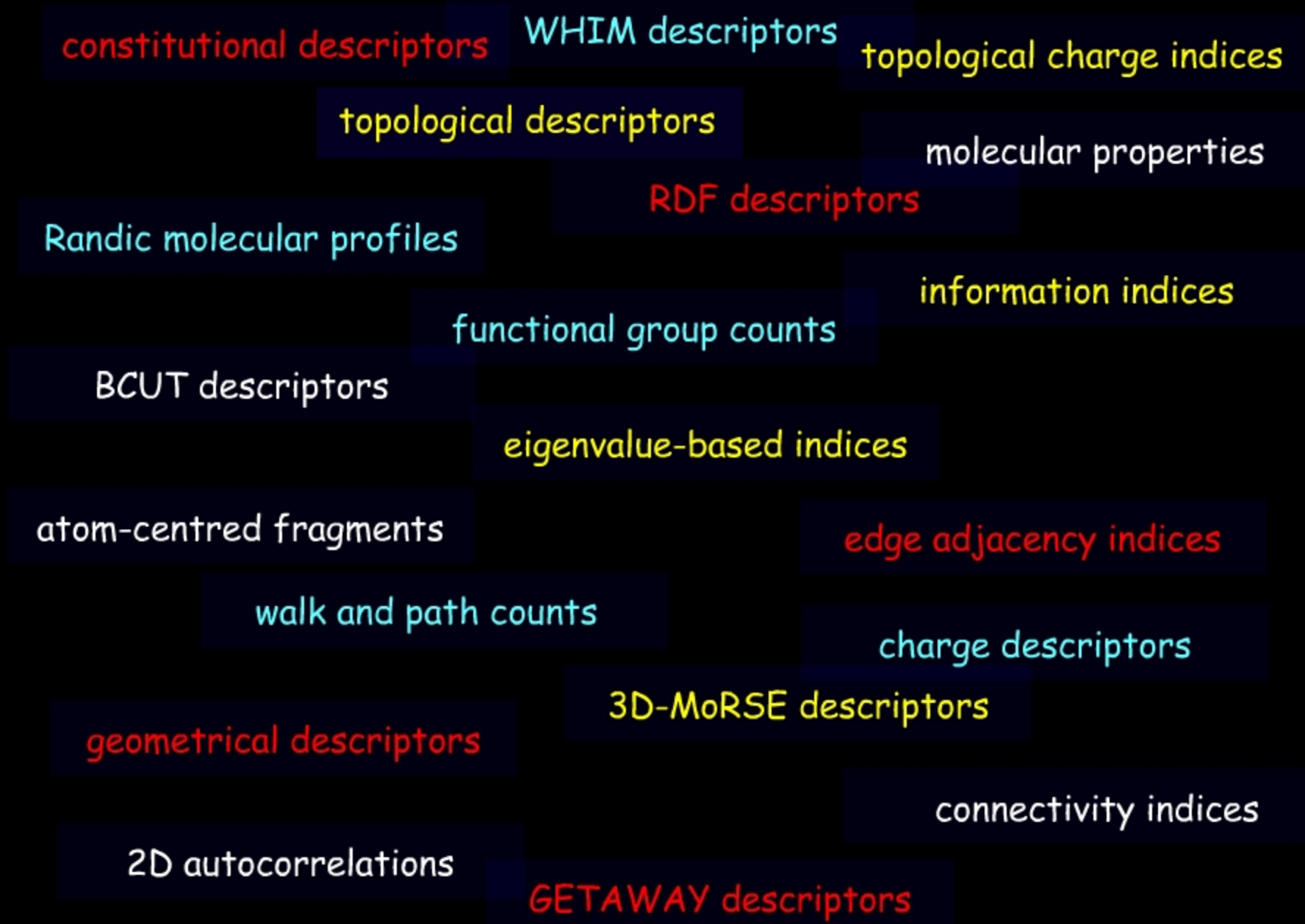

Molecular Volume: 117.6 cm^3

Dipole Moment: 2.2 Debyes

Surface Area: 168.6 Å^2

F1.1.15 A Multitude of Molecular Descriptors

The number of descriptors that can be derived from a molecular structure is virtually unlimited. Currently available software packages can calculate thousands of descriptors. For example the DRAGON program calculates 1612 descriptors distributed into 20 categories.

constitutional descriptors

WHIM descriptors

topological charge indices

topological descriptors

molecular properties

RDF descriptors

Randic molecular profiles

information indices

functional group counts

BCUT descriptors

eigenvalue-based indices

atom-centred fragments

edge adjacency indices

walk and path counts

charge descriptors

3D-MoRSE descriptors

geometrical descriptors

connectivity indices

2D autocorrelations

GETAWAY descriptors

F1.1.16 Biologically Relevant Descriptors

When constructing a QSAR model, the key is to use descriptors that are relevant to the specific property of interest. These "biologically relevant descriptors" help generate a model that differentiates between molecules that possess the property of interest and those that do not.
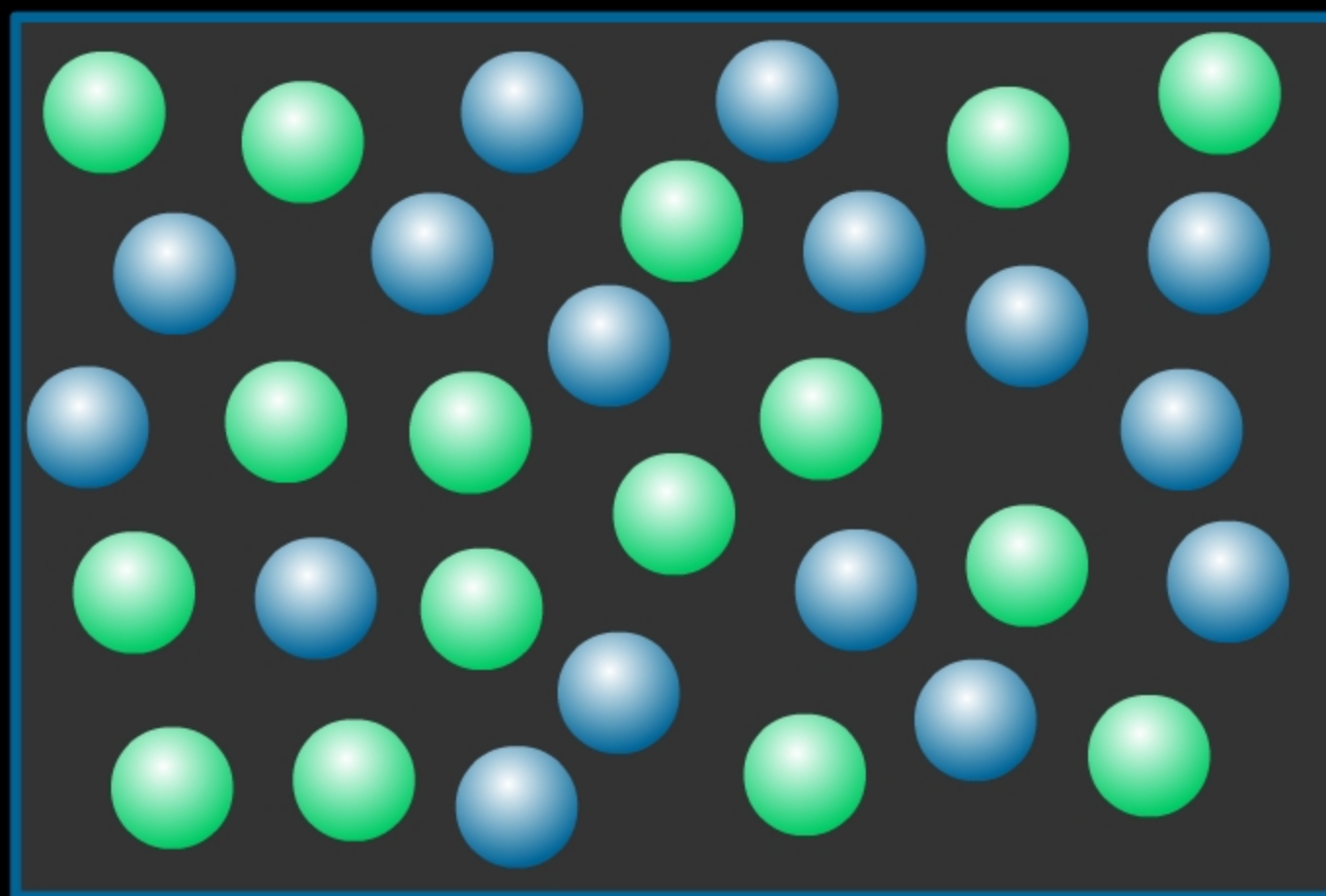
F1.1.17 Application of QSAR

QSAR models are built for three main reasons: to understand the relationship between structure and activity, design compounds with improved activity, and predict the activities of compounds prior to their synthesis. These reasons in fact adhere to the rational sequence of a QSAR analysis project where the first step is to understand the phenomenon, and then use this understanding to design new compounds.


Understanding   Design

A good model can reveal information about the receptor's binding site. For example a correlation with electronic descriptors may indicate that the biological activities could be due to the chemical reactivity of the compounds, or alternatively, a correlation with hydrophobic descriptors may reveal the existence of a hydrophobic pocket in the receptor.

$$\text{Biological effect} = f(\text{hydrophobic descriptors})$$

Hydrophobic Pocket
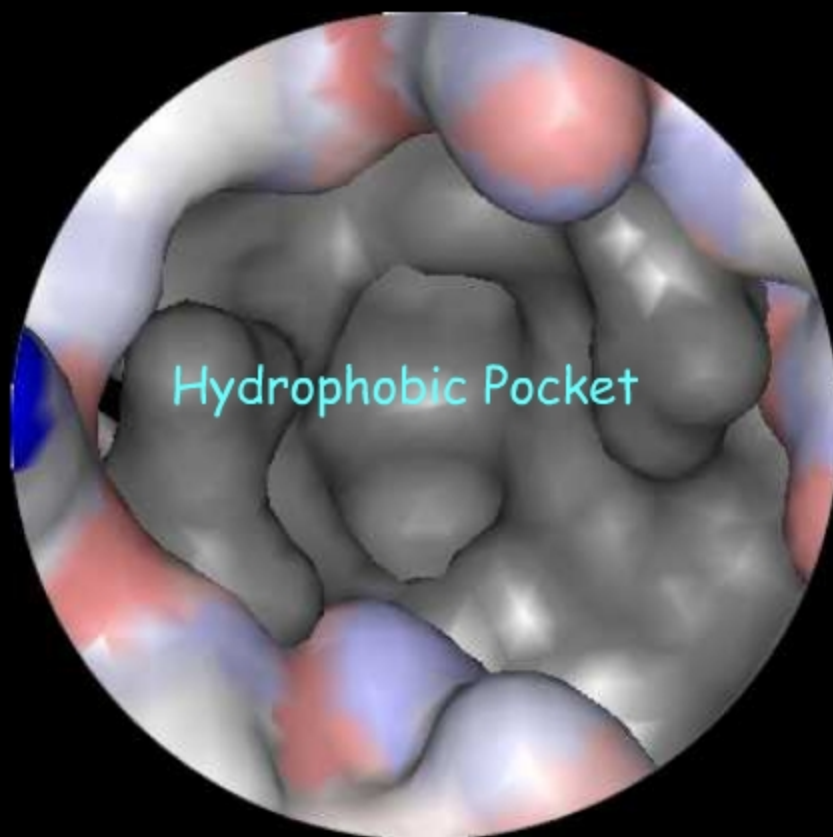
F1.1.19 Designing Compounds with Improved Activities

Once a QSAR model is obtained and reproduces the known data satisfactorily, it can be exploited to predict the biological activity of not yet synthesized analogs. This is of paramount importance in lead optimization and represents one of the most popular uses of the QSAR approach.
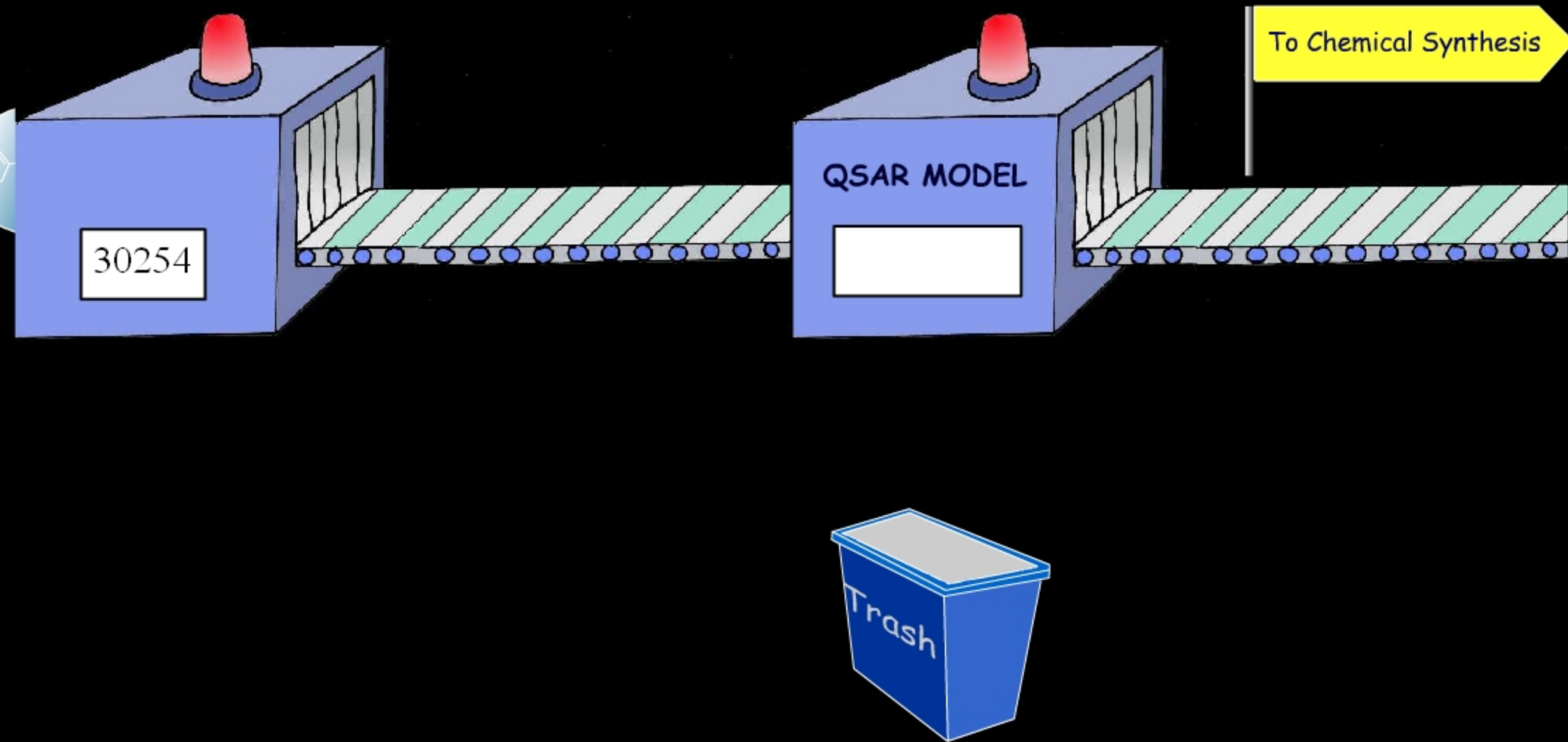
QSAR model

Compounds not
yet synthesized

Prediction of the
biological activities
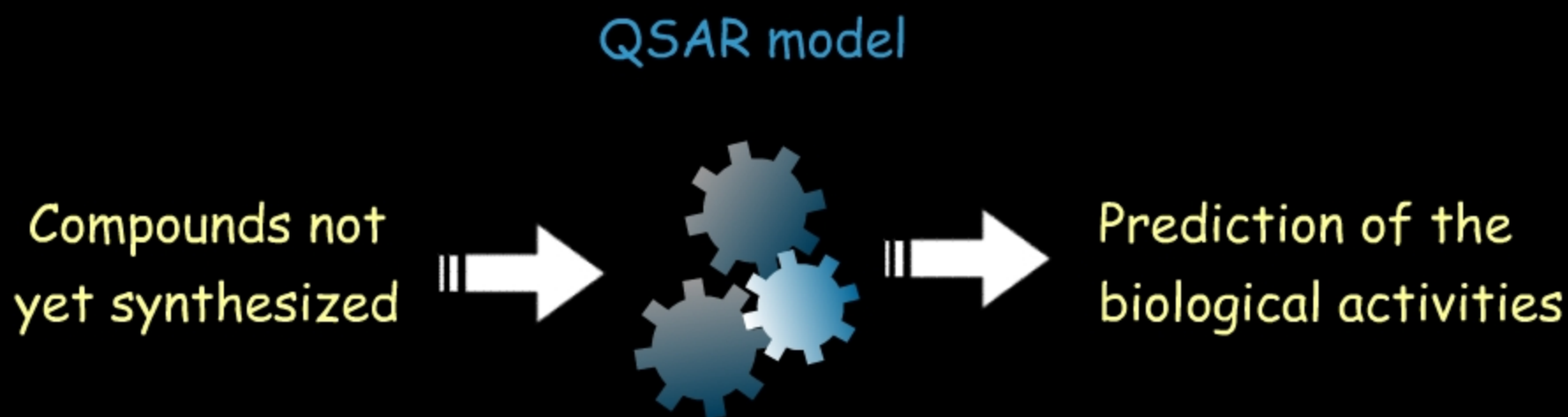
F1.1.20 Reducing a Virtual Library to a Practical Size

The recent explosion in combinatorial chemistry has added a new dimension to the QSAR approach by reducing a huge virtual library to a manageable size for combinatorial synthesis and high through-put screening.

The topic The Foundations of QSAR contains the following 28 pages:

- Birth of QSAR

- The Foundations of QSAR

- The Hammett Contribution

  - Dissociation Constants of Substituted Benzoic Acids

  - Dissociation of Substituted Phenylacetic Acids

  - Linear Free Energy Relationship

  - The Hammett Equation

  - The Meaning of ρ

  - The Meaning of σ

  - Examples of σ Constants

  - Predicting the pKa of Benzoic Acid Compounds

- Hansch Contribution

  - The Importance of Lipophilicity

…

For the entire list, see the navigation panel.

F1.2.1 Birth of QSAR

QSAR dates back to the 19th century with the work of Cros (1863) who first observed an inverse correlation between the toxicity of alcohols and their water solubility. Other important milestones include work by Crum-Brown and Frazer who related physiological action to chemical constitution (1868). A few years later Horst, Overton and Richet independently observed that the toxicity of organic compounds depended on their lipophilicity/solubility. This discovery was followed by research by Meyer and Overton, who proved that anesthetic potency correlated well with partition coefficients (1899).

- 1863 Cros   inverse correlation between toxicity and water solubility of alcohols

- 1868 Crum-Brown & Frazer   "physiological action" is a function of "chemical constitution"

- 1890's Horst & Overton   toxicity of organic compounds depend on their lipophilicity.

- 1893 Richet   "more they are soluble, less they are toxic"

- 1899 Meyer-Overton   partition coefficients correlate with anesthetic potency

F1.2.2 The Foundations of QSAR

During the first half of the 20th century, Louis Hammett laid the foundation for modern QSAR by correlating electronic properties of organic acids and bases with their equilibrium constants and reactivity. An important landmark in the development of QSAR took place in 1964 with the introduction of the Free-Wilson method and Hansch analysis. This section covers these three seminal contributions to QSAR in some detail.

- Louis Hammett

- Free-Wilson

- Corwin Hansch

F1.2.3 The Hammett Contribution

The dissociation of HA organic acids is a process by which a proton (H^+) is removed from the neutral compound, leaving behind a negatively charged species (A^-). The extent of the reaction is measured by the dissociation constant K. Louis Hammett observed that the dissociation constants of aromatic acids are influenced by the electronic properties of the substituents on the phenyl ring.

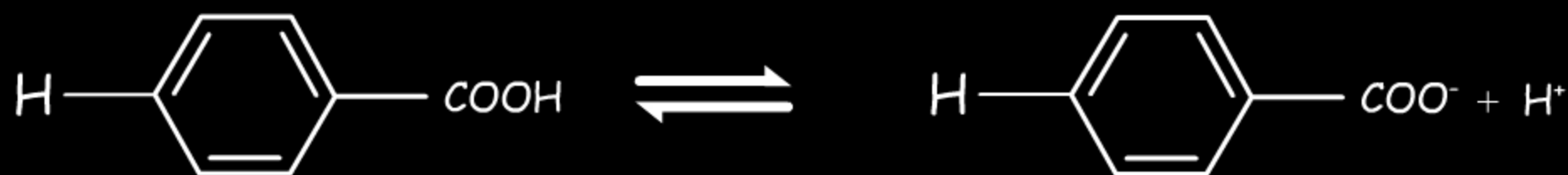$$HA \rightleftharpoons H^+ + A^-$$

$$K = \frac{[H^+][A^-]}{[HA]}$$

F1.2.4 Dissociation Constants of Substituted Benzoic Acids

The dissociation constants of substituted benzoic acids indicate that electron withdrawing groups increase dissociation while electron donating groups decrease it.

electron donating | electron withdrawing
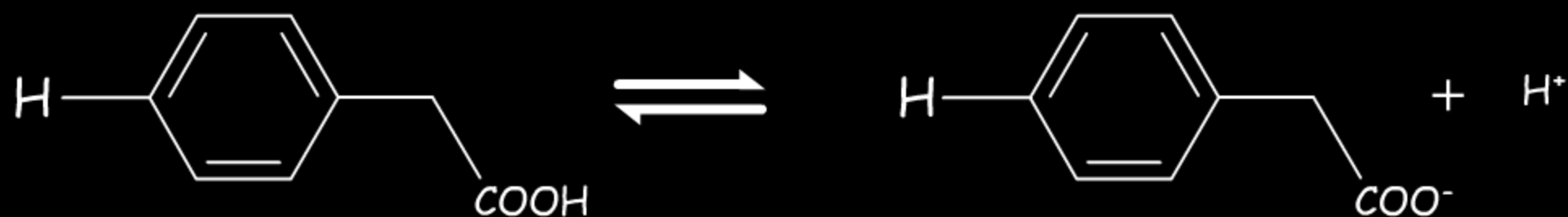
Dissociation Constant (10^{-5})

$K_0 = 6.2$

F1.2.5 Dissociation of Substituted Phenylacetic Acids

A similar effect exists for other equilibria such as substituted phenylacetic acids.

electron donating | electron withdrawing

Dissociation Constant (10^{-5})
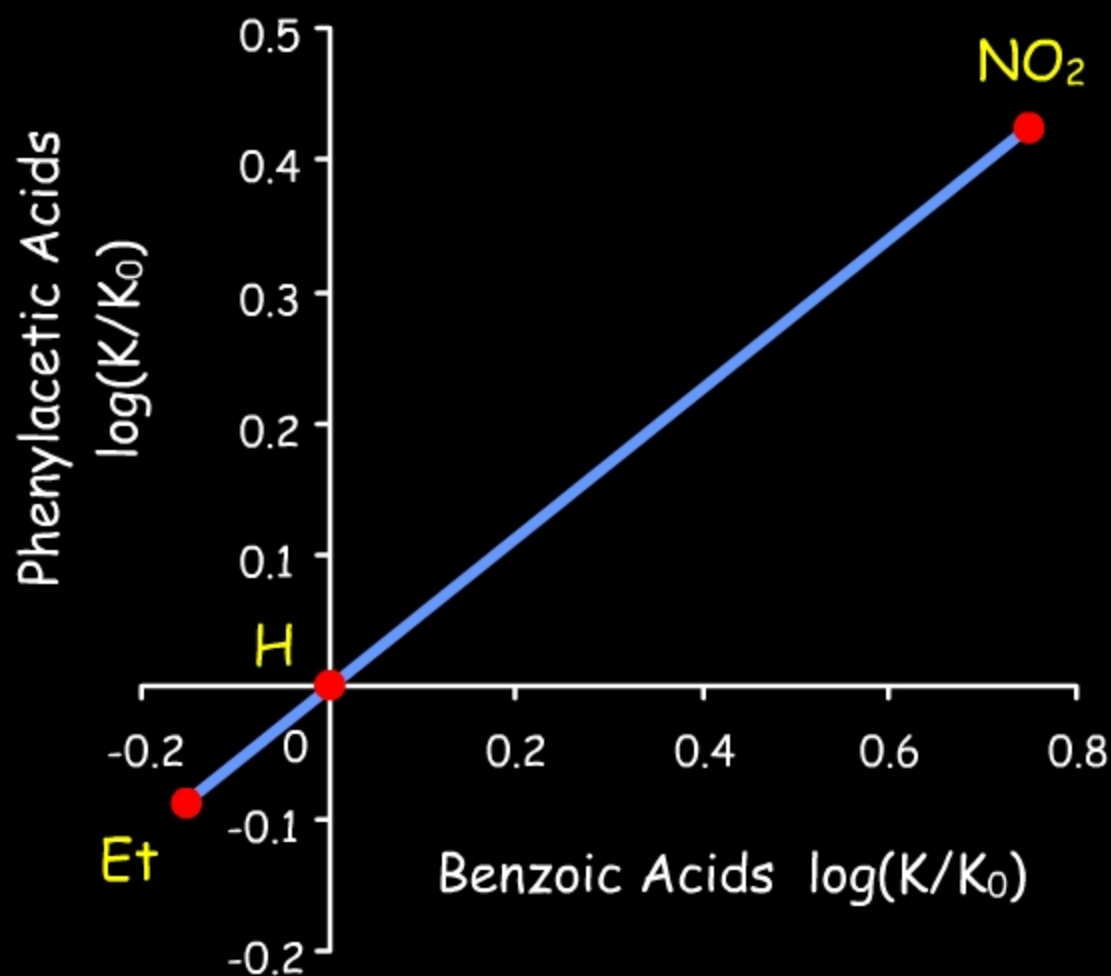
$K_0 = 5.2$

F1.2.6 Linear Free Energy Relationship

When plotting the quantity log(K/Ko) for benzoic acids on the X axis, where K and Ko refer to the unsubstituted and substituted compounds, respectively, and the corresponding values measured for the same set of substituents in phenylacetic acids on the Y axis, Hammett obtained a straight line. Because of the association between dissociation constants and free energies [$\Delta G = -RT \, Log(K)$] this phenomenon is known as the linear free energy relationship.

Benzoic Acid

R	K	log (K/K$_0$)
NO$_2$	37.05 x 10^{-5}	0.776
Et	4.4 x 10^{-5}	-0.15
H	6.2 x 10^{-5} (k$_0$)	0

Phenylacetic Acid

R	K	log (K/K$_0$)
NO$_2$	14.1 x 10^{-5}	0.43
Et	4.2 x 10^{-5}	-0.09
H	5.2 x 10^{-5} (k$_0$)	0

F1.2.7 The Hammett Equation

The straight line described on the previous page can be written as a linear equation, the Hammett equation. Note that ρ is related to a given scaffold (e.g. phenylacetic acids), whereas a σ is a descriptor of a substituent and describes its influence on the dissociation constant. It is positive for electron withdrawing substituents and negative for electron donating substituents.

$$y = \rho x$$

$$\log \frac{K}{K_o} = \rho \log \frac{K}{K_o}$$

$$= \rho \sigma$$

$$\log \frac{K}{K_o} = \sigma$$
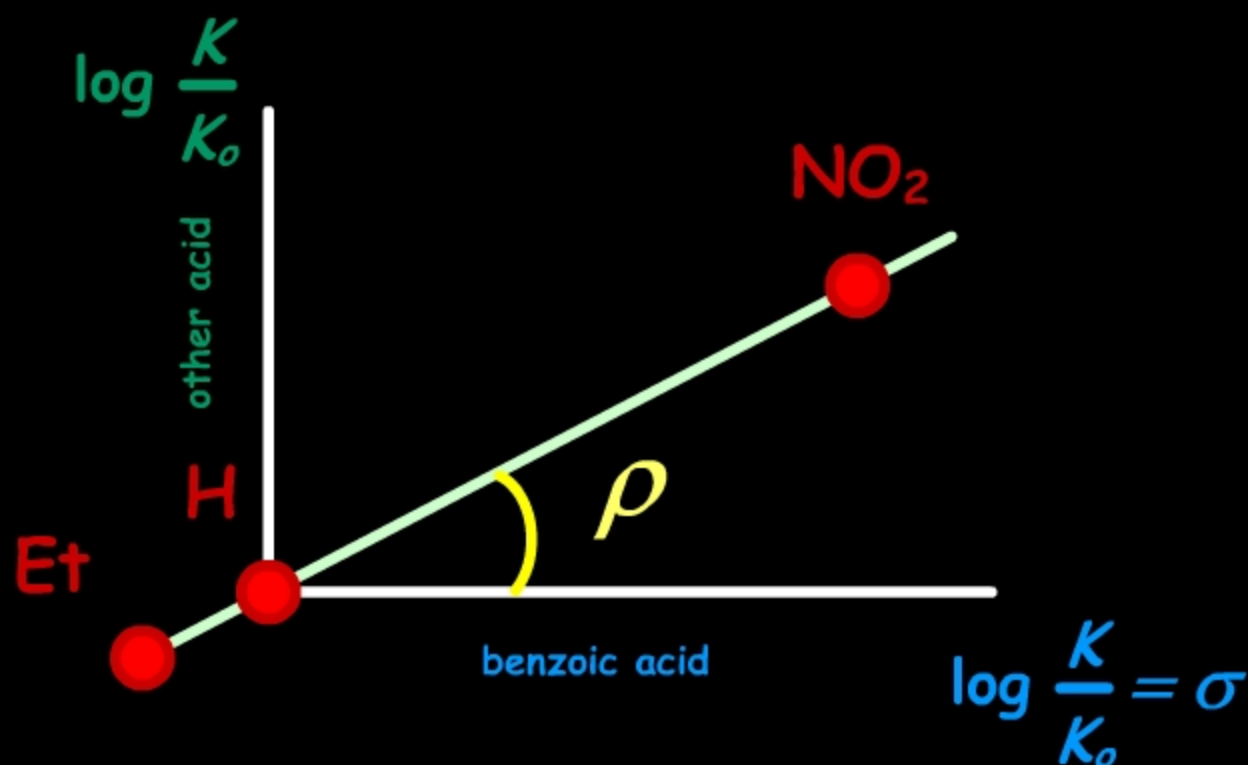
ρ pertains to a given equilibrium as compared to the benzoic acid equilibrium.

σ is a descriptor of a substituent

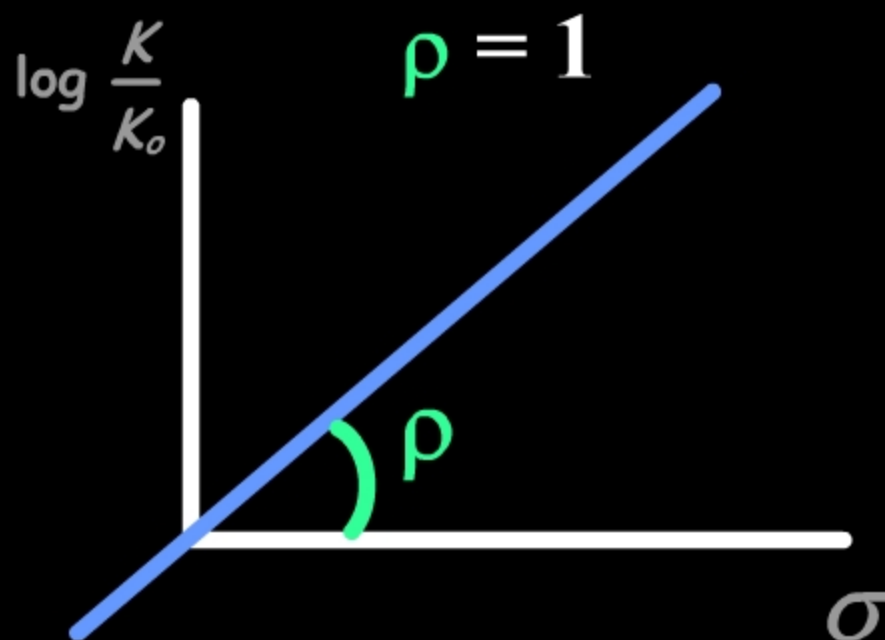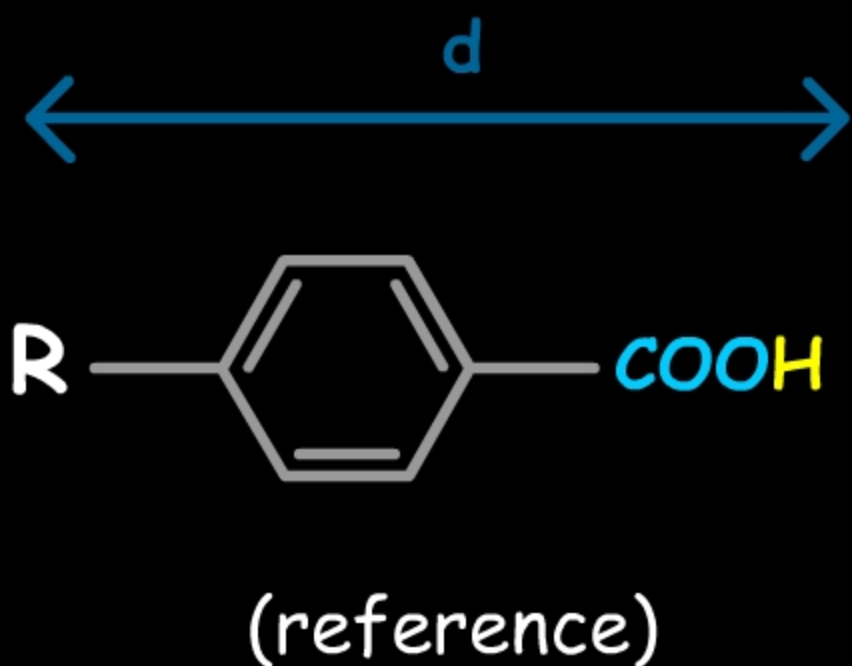# F1.2.8 The Meaning of ρ

ρ describes the magnitude of the effect a substituent can exert on the dissociation reaction of a given scaffold. As the distance between the substituent and the dissociated proton increases, its influence on the dissociation reaction decreases and so does the value of ρ.

- Benzoic Acid
- Phenylacetic Acid
- Phenylpropionic Acid

d

R —⟨benzene ring⟩— COOH

(reference)

$\log \dfrac{K}{K_o}$

$\rho = 1$

ρ

σ

F1.2.9 The Meaning of σ

σ describes the effect of substituents on the dissociation reaction. Substituents on the phenyl ring can increase or decrease the equilibrium constant by stabilizing or destabilizing the anionic form via the formation of a positive or negative partial charge at C1.
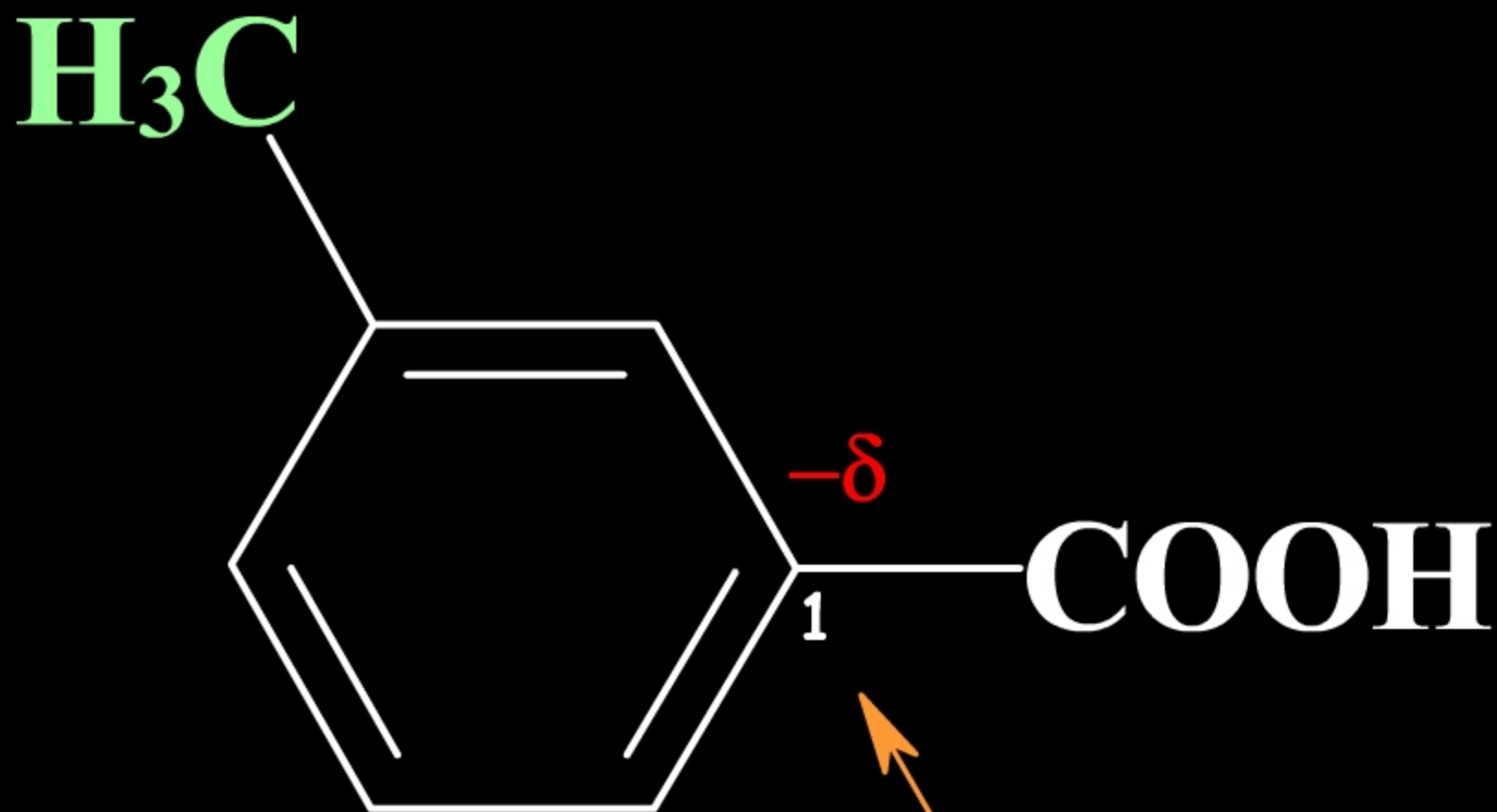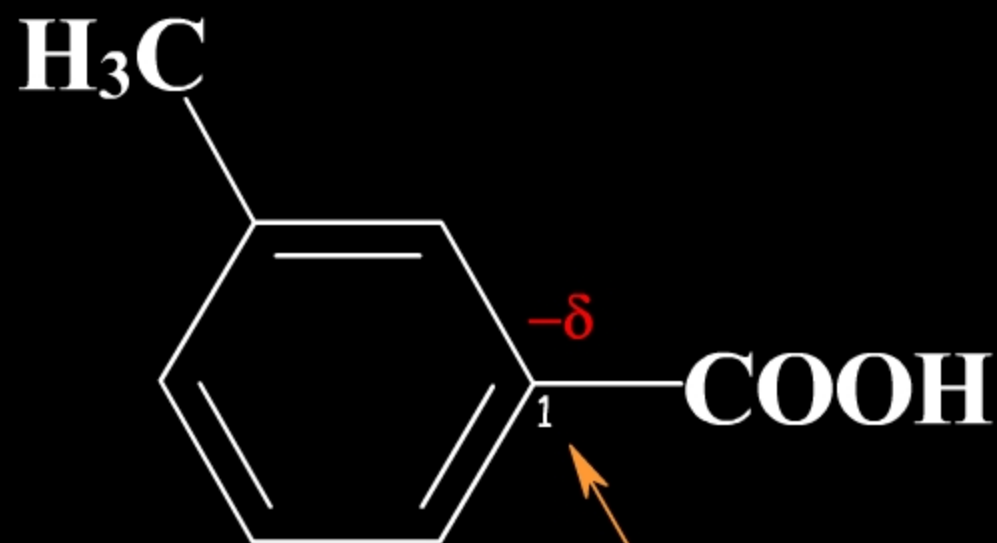

Destabilizes anionic form
Decreases dissociation

F1.2.10 Examples of σ Constants

Electron donating substituents have negative σ values, whereas positive σs correspond to electron withdrawing groups. Note that σ values differ depending on whether the substituent is meta or para (sigma values are clickable).
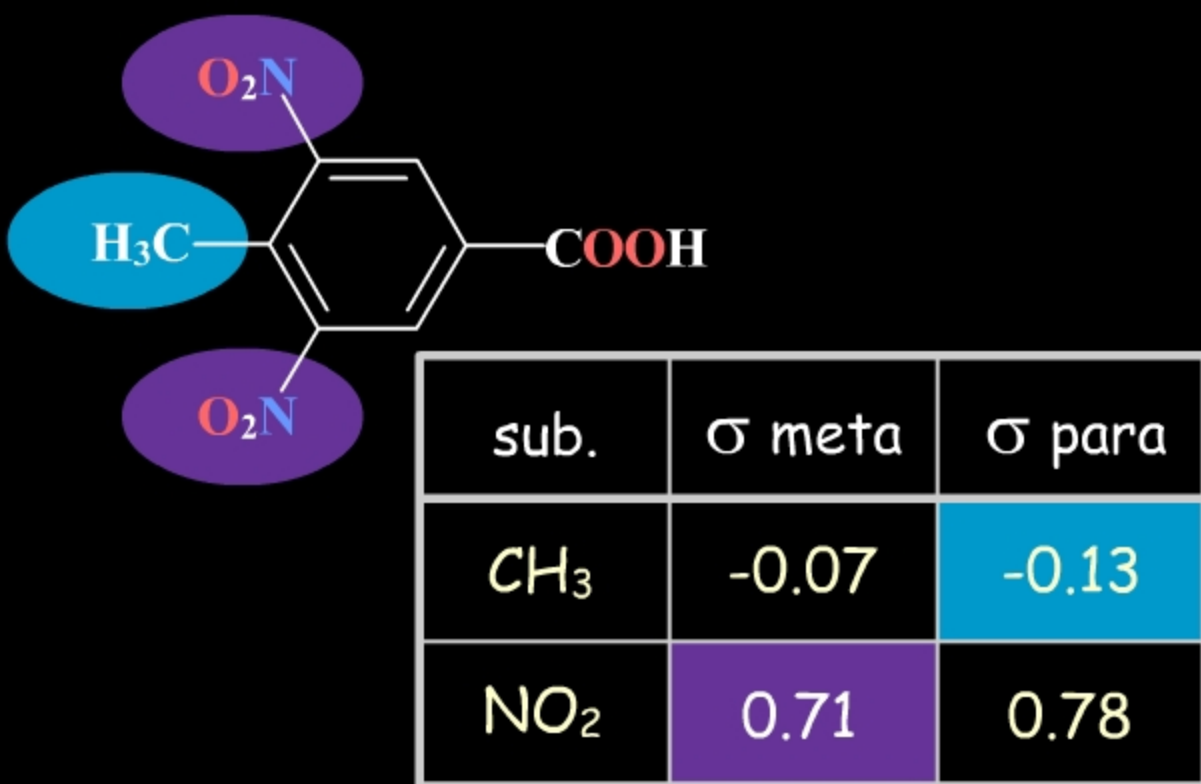
substituent	σ meta	σ para
CH_3	-0.07	-0.13
NH_2	-0.16	-0.66
CN	0.56	0.66
NO_2	0.71	0.78
OCH_3	0.12	-0.27

H_3C

$-\delta$

$\overset{1}{}$—COOH

Destabilizes anionic form
Decreases dissociation

F1.2.11 Predicting the pKa of Benzoic Acid Compounds

The Hammett equation is an example of a QSPR equation. It correlates a molecular property, the dissociation constant, with a set of molecular descriptors (σ and ρ). It can be used to predict the pKa of benzoic acid analogs. When a molecule has multiple substituents, the σ values are summed to yield the total value for the compound, as shown in the following example.

sub.	σ meta	σ para
CH$_3$	-0.07	-0.13
NO$_2$	0.71	0.78

$$\log \frac{K}{K_o} = \rho\sigma$$

$$\log K - \log K_o = \rho\sigma$$

$$-pK + pK_o = \rho\sigma$$

$$pK = pK_o - \rho\sigma$$

$$pK_{acid} = pK_{o\,(acid)} - \rho_{acid} \sum \sigma_{substituent}$$

Benzoic acid:

$$pK_0 = -\log(6.2 \times 10^{-5}) = 4.2 \qquad \rho = 1$$

$$pK_{acid} = 4.2 - 1.00\,(0.71 - 0.13 + 0.71) = 2.91$$

F1.2.12 Hansch Contribution

Corwin Hansch recognized the importance of lipophilicity for biological activity. Lipophilicity can be defined as the tendency of a compound to prefer an organic phase rather than an aqueous phase. Molecular lipophilicity is correlated with the presence of hydrophobic groups (e.g. aromatic rings) and with the absence of polar and ionizable groups.
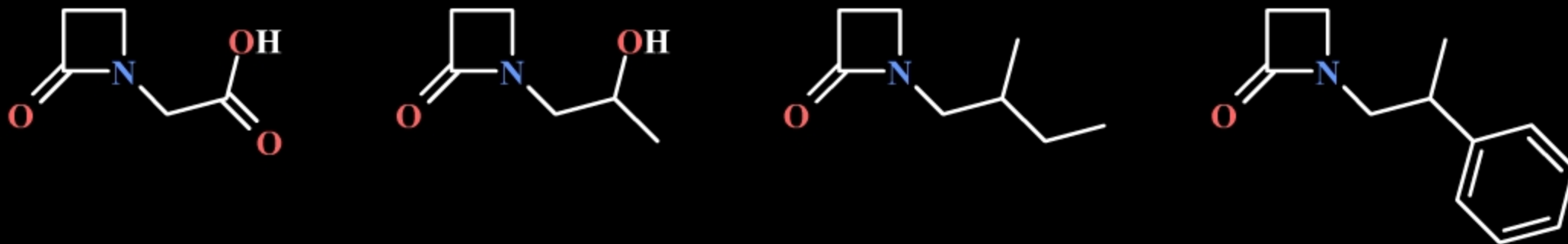
F1.2.13 The Importance of Lipophilicity

To exert its biological effect, a drug must reach its site of action. On its way, the compound encounters biological barriers. For example, the cell membrane is composed of a lipid bilayer, therefore lipophilic compounds are able to penetrate such barriers. The lipophilicity thus determines in part the amount of the compound that reaches the target site.

cell membrane
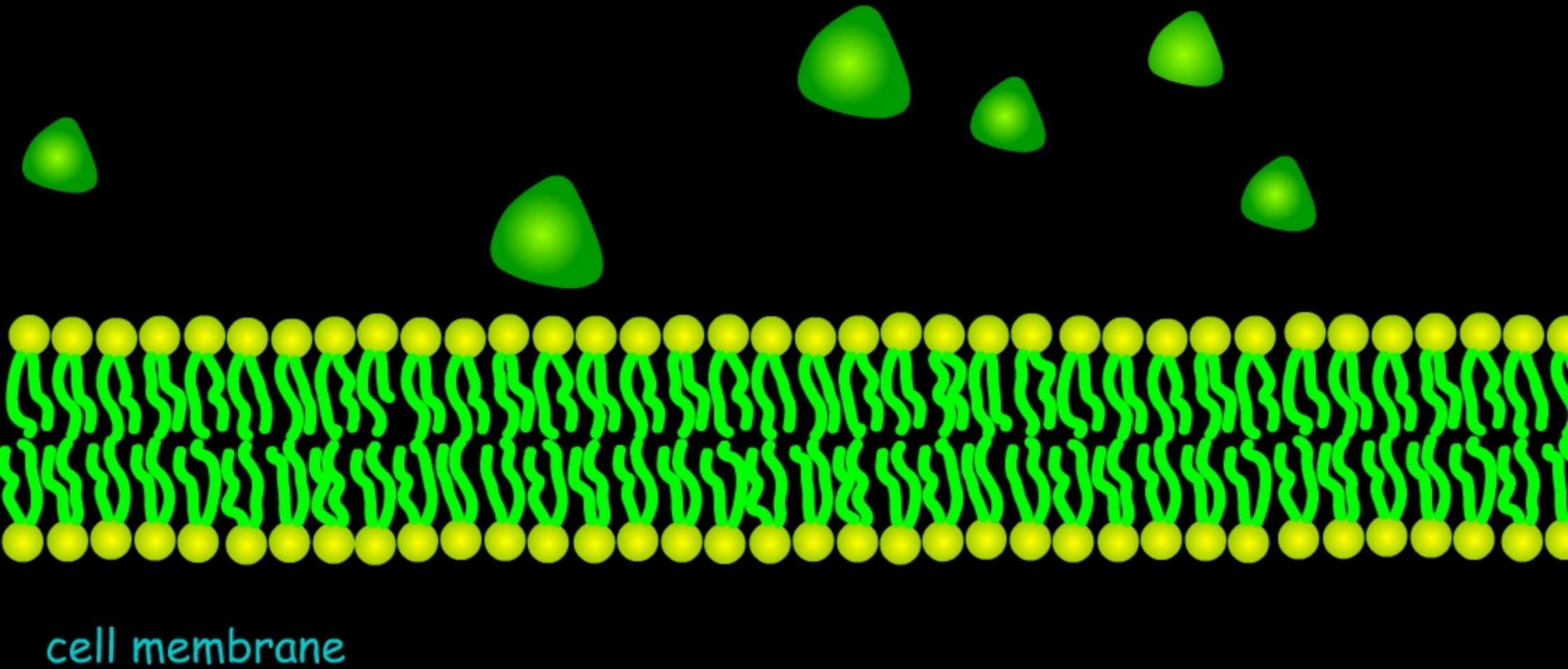
F1.2.14 LogP is a Measure of Compounds Lipophilicity

LogP describes the lipophilicity of a compound, and is the measure of the partitioning of this compound between a lipidic phase (usually 1-octanol) and an aqueous phase. It is represented by a partition coefficient P or by its log value. Compounds with P>1 (logP>0) are lipophilic whereas compounds with P<1 (logP<0) are hydrophilic.

Organic phase

Aqueous phase (buffer)
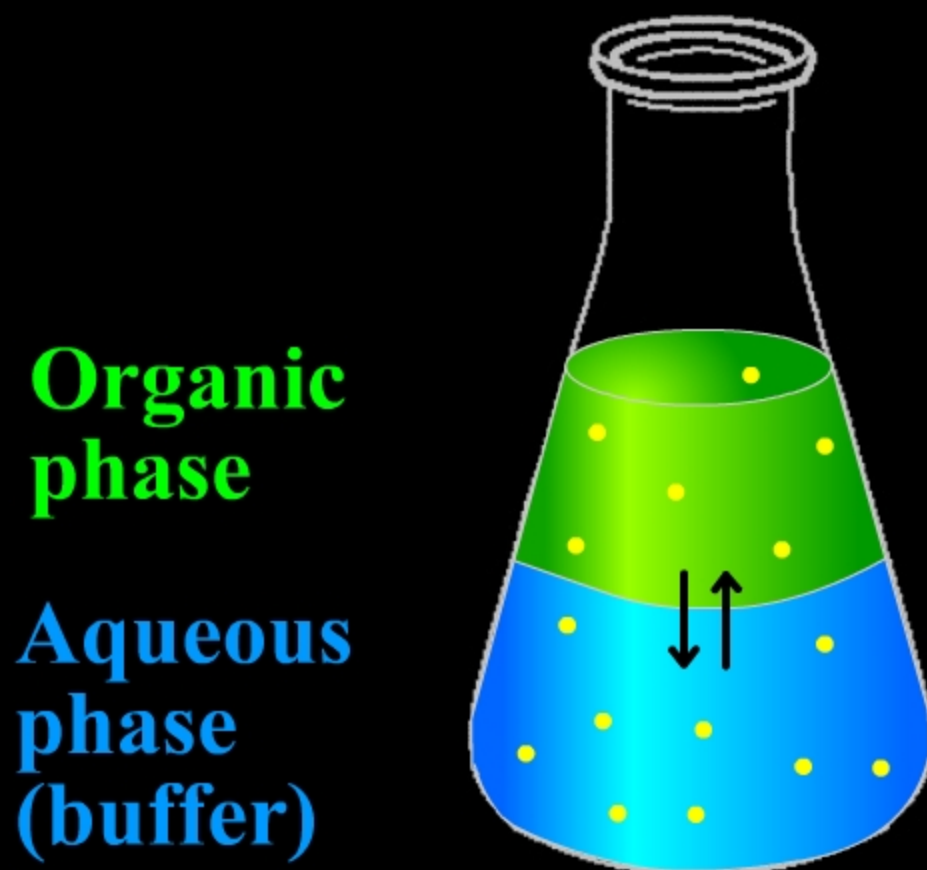
$$P = \frac{[drug]_{org}}{[drug]_{aq}}$$

P>1 (logP>0)   More lipophilic drug

P<1 (logP<0)   More hydrophilic drug

Hansch observed that in several compound series, the biological activities (inhibition of frog heart) correlated well with logP. The resulting QSAR expression turned out to be quite simple: a linear equation.

Hansch 1970

Inhibition of frog heart by miscellaneous chemicals

$$\text{Biological Activity} = f(\log P)$$

$$\text{Biological Activity} = 0.91 \log P + 0.14$$

$$n = 34 \; ; \; r^2 = 0.951$$

F1.2.16 Example of Correlation with LogP

A correlation with logP was found for the isonarcotic activities of diverse esters, alcohols, ketones, and ethers (structures, biological activities and logP are shown by clicking the button "training set"). The training set covers a large range of values for logP and the model generated is quite simple. In the equation $\log(1/C)$ represents the biological activity response.

correlation   training set

$$\log(1/C) = 0.731 \log P + 1.22$$
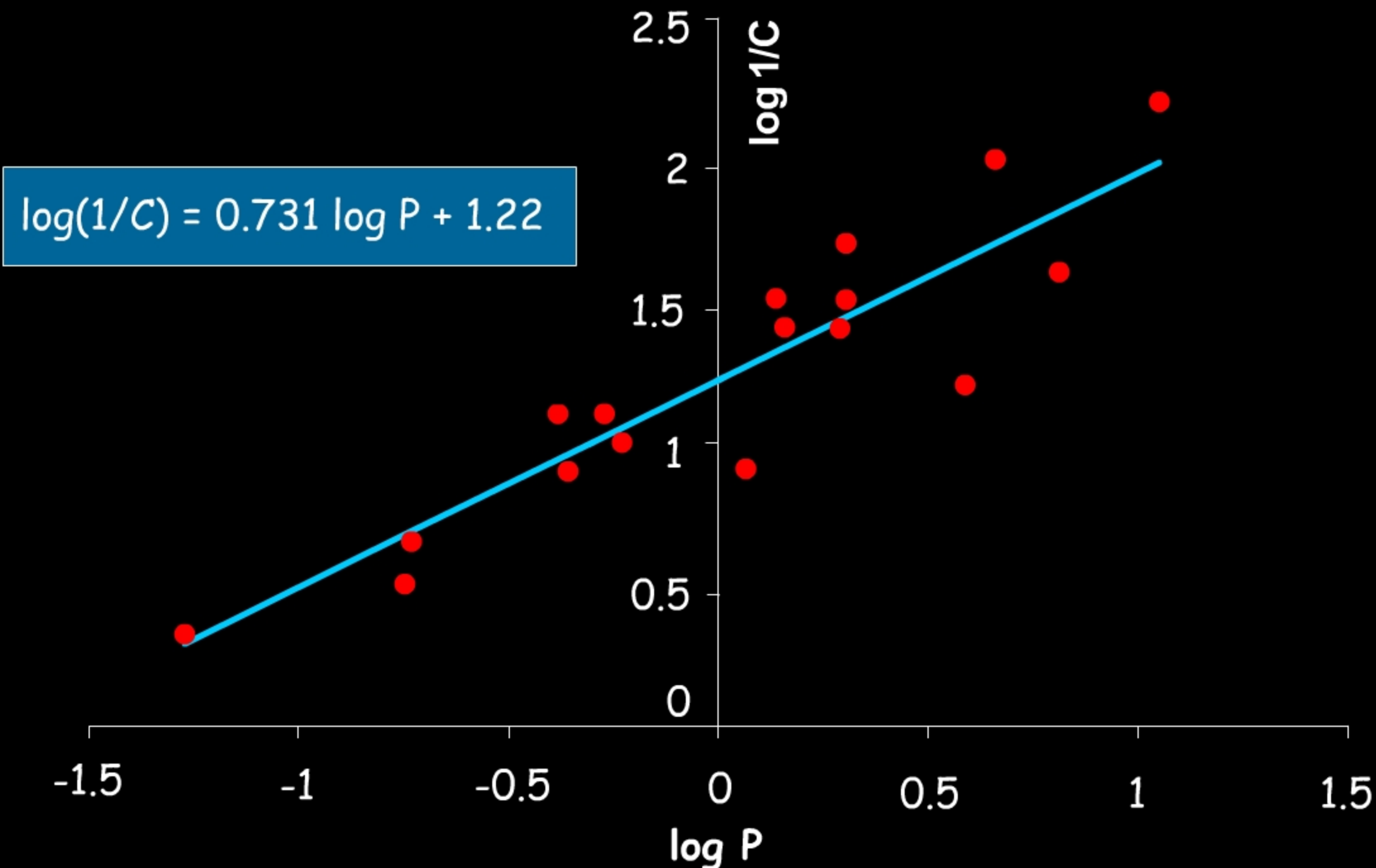
When attempting to ameliorate his initial idea (only one descriptor), Hansch further refined the QSAR equation and introduced other descriptors such as σ (electronic), Es (Taft), π (lipophilicity), MR (Molar refraction) that represent properties that may influence drug action. He also considered other mathematical equations likely to represent the behavior of the biological system. The analytical forms of some of the typical QSAR models he developed are indicated below, whereas the most common descriptors used in Hansch's analyses are presented in the following pages.

Typical Hansch QSAR Models

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

$$\log 1/C = a \pi^2 + b \pi + c$$

$$\log 1/C = a (\log P)^2 + b \log P + c \sigma + d$$

$$\log 1/C = a \pi^2 + b \pi + c \sigma + dE_s + c$$

F1.2.18 The π Descriptor

The π descriptor characterizes the lipophilicity of a substituent. It is defined as the difference between the logP of the substituted and unsubstituted compounds. Substituents that are more hydrophobic than H have π positive values, whereas substituents less hydrophobic than H have negative values. The π values for several common substituents are listed below, and an example is provided.

$$\pi = \log P - \log P_H$$

Example

log P=2.27   log P=2.13

$\pi_F = 2.27 - 2.13 = 0.15$
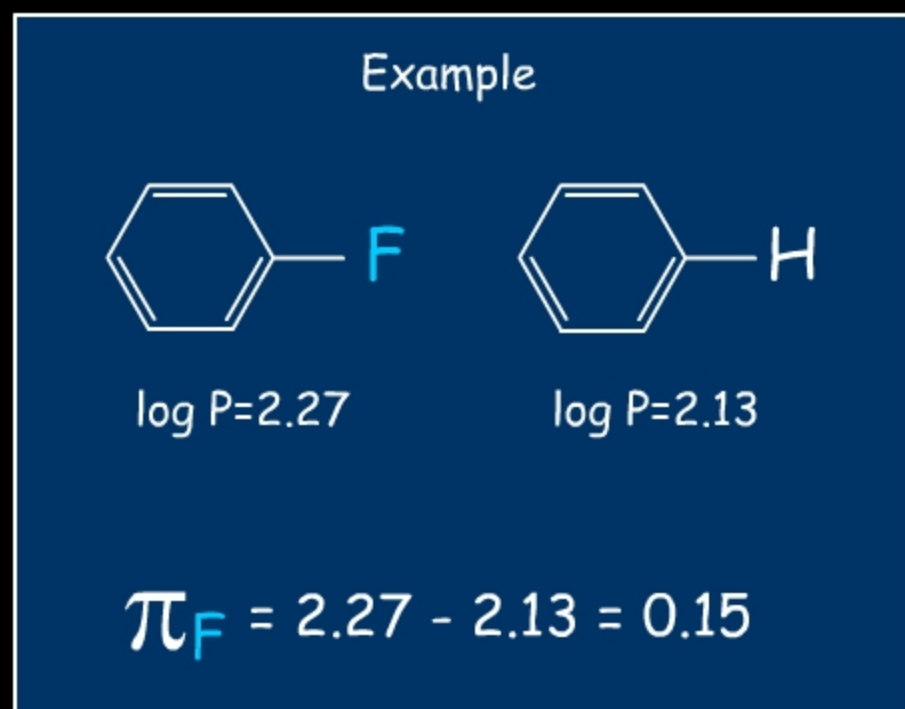
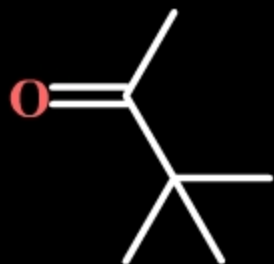| Substituent	π
H	0.00
F	0.15
Cl	0.70
Br	1.02
OH	-0.67
Me	0.52

F1.2.19 The MR Descriptor

Molar refractivity (MR) is a molecular descriptor that contains information on the compound's volume corrected by the refractive index (the ratio of the velocity of light in a vacuum to the velocity of light in the substance of interest). In the definition below, d is the density and n is the refractive index. Several MR values are listed below.
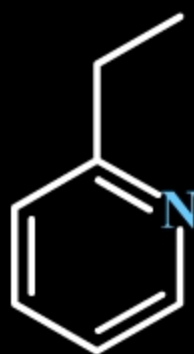
$$MR = \frac{(n^2-1)\ MW}{(n^2+1)\ d}$$
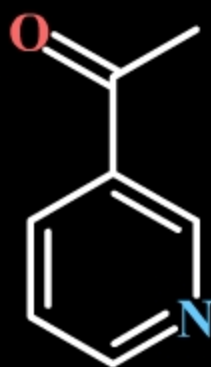
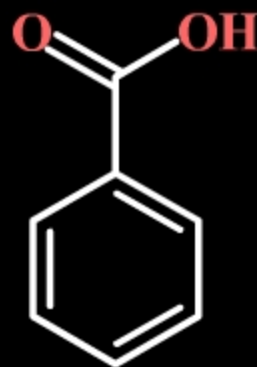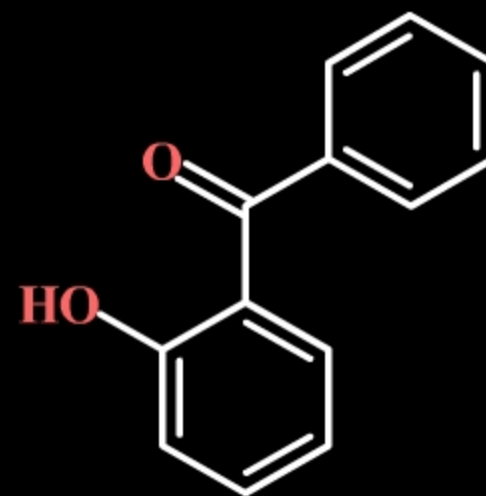| 34.37 | 29.84 | 33.89 | 33.18 | 57.92 |

F1.2.20 The Taft Descriptor (ES)

The Taft steric constant Es is an experimental value based on rate constants for a given model reaction. It is a measure of the steric effect exerted by a substituent on the equilibrium. In the definition below K_x and K_H are the rate constants for substituted and unsubstituted compounds. The bulkier the substituent, the more negative the Es. The table lists Es values for several common substituents.

$$E_s = \log K_X - \log K_H$$

X	Es
H	1.24
F	0.78
Cl	0.27
Br	0.08
I	-0.16
Me	0.00

F1.2.21 Meaning of Parabolic Dependence on LogP

Hansch's introduction of a parabolic term for logP was a breakthrough that helped make QSAR a more powerful tool. This additional term denotes the existence of an optimum in logP: molecules with higher values start to be trapped in hydrophobic membranes and cannot reach their site of action.

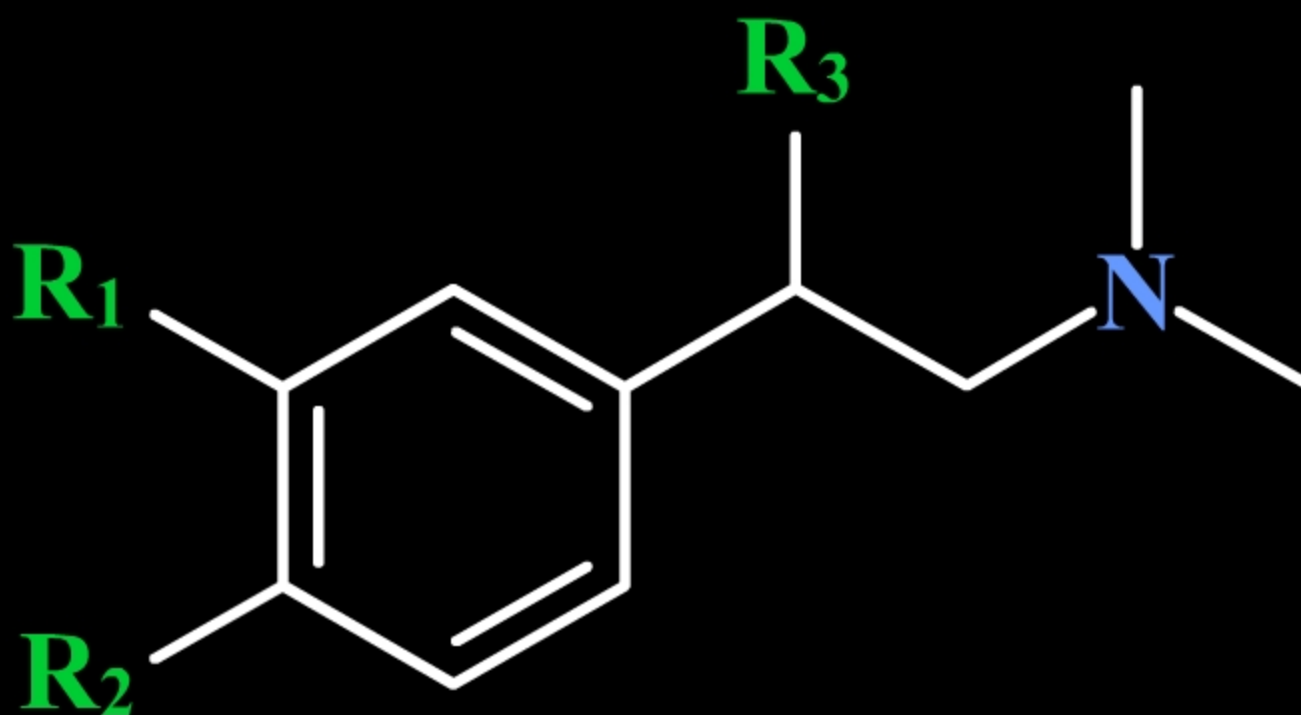◉ Graph   ◉ Equation
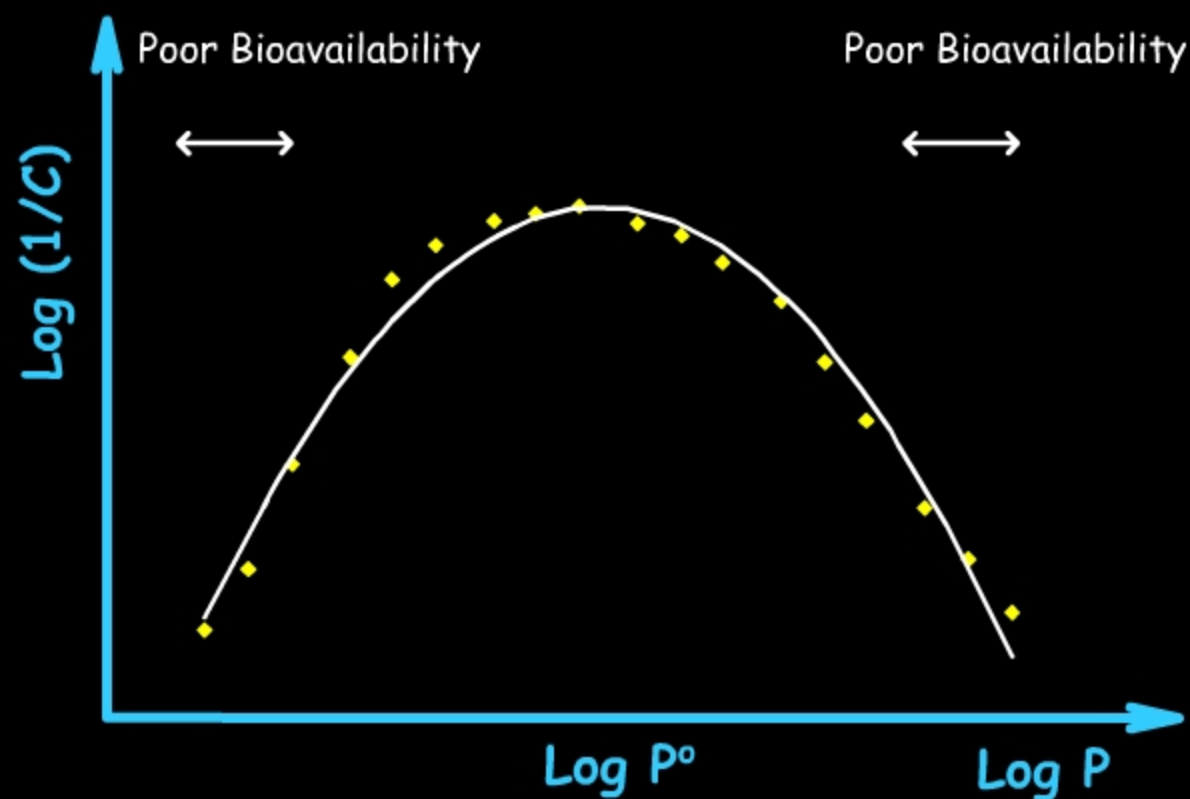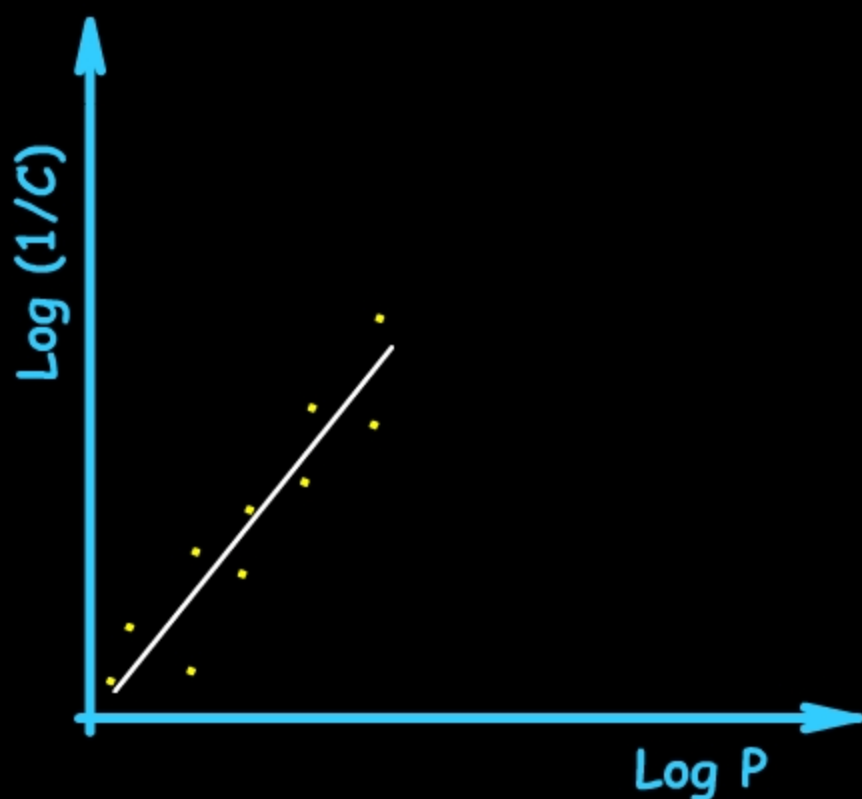
F1.2.22 The Free-Wilson Analysis

In 1964, Free and Wilson derived a mathematical model based on the hypothesis that overall biological activity is the sum of all elementary contributions of the substituents. This approach assumes that substituent effects are additive.

$$\log 1/C = R_1 \text{ Contribution} + R_2 \text{ Contribution} + R_3 \text{ Contribution}$$

F1.2.23 Indicator Variables and Substituent Weights

In the Free-Wilson model biological activity is represented by the following equation. Indicator Variables x_i (with values of 1 or 0) describe the presence or the absence of certain substituents at each scaffold position. The weight of each substituent is represented by the a_i coefficients. μ is the overall average activity.

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

R1	a_i
F	-0.301
Cl	0.207
Me	0.454

(F, Cl, Me)

$$F \qquad Cl \qquad Me$$
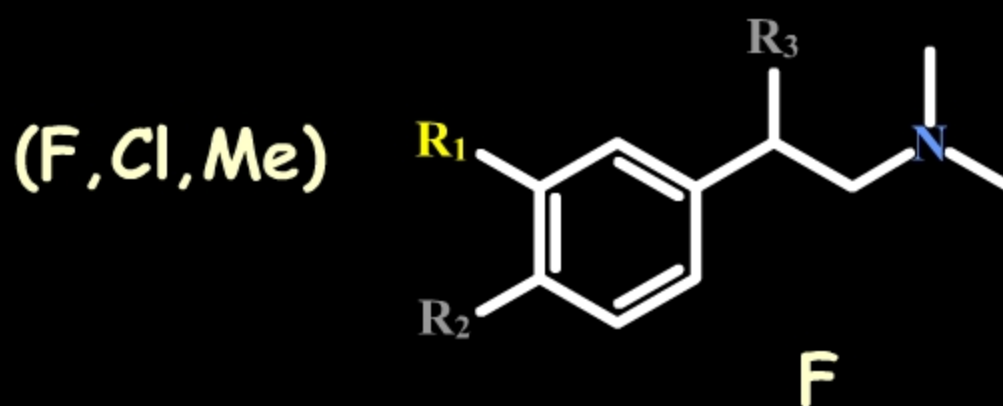
$$R_1 \text{ Contribution} = a_1 X_1 + a_2 X_2 + a_3 X_3$$

$$R_1 \ (F) = -0.301 + 0 + 0$$

$$R_1 \ (Cl) = 0 + 0.207 + 0$$

$$R_1 \ (Me) = 0 + 0 + 0.454$$

F1.2.24 Free-Wilson Structural Matrix

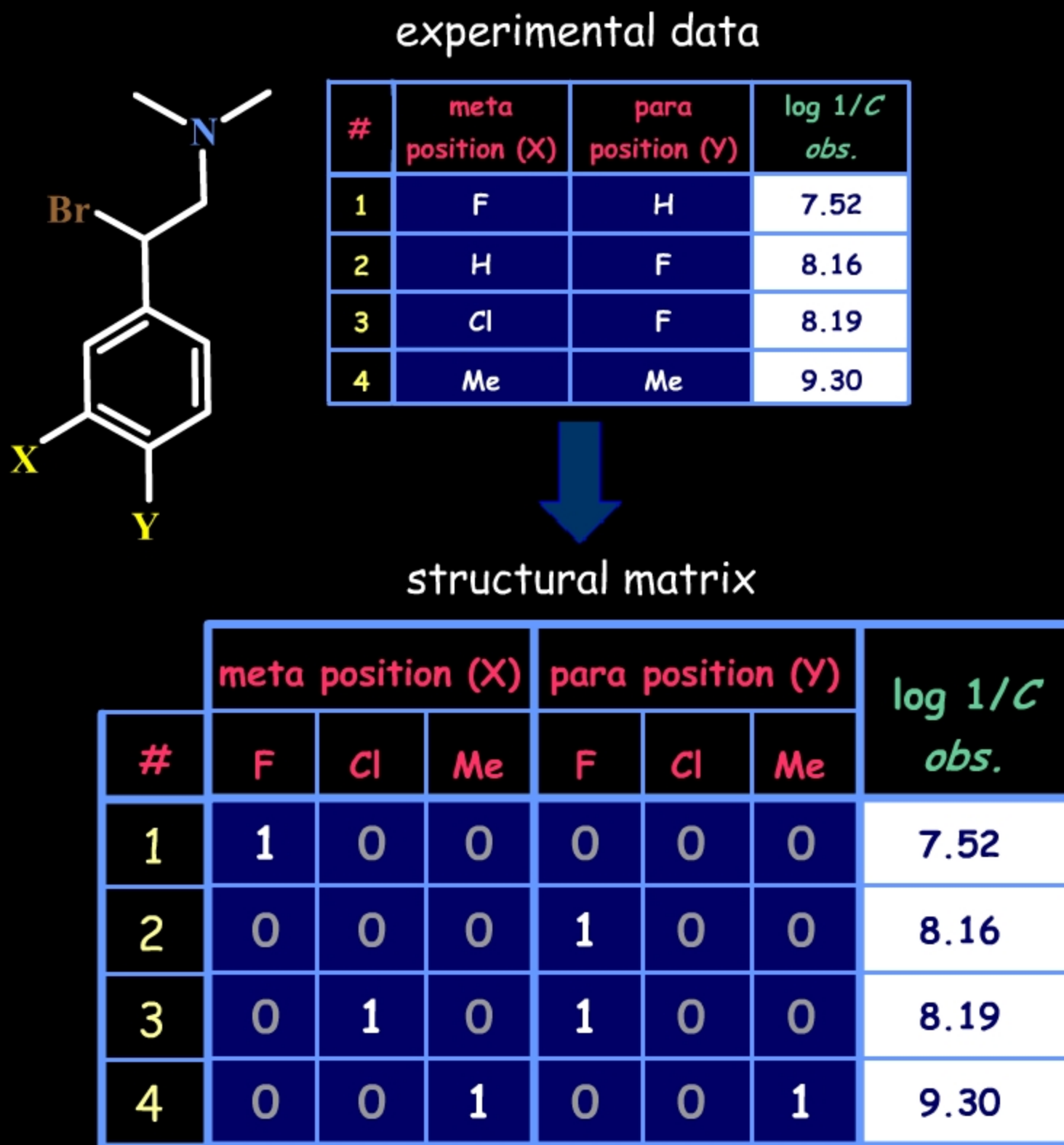The construction of the Free-Wilson QSAR model is based on a structural matrix composed of the x_i indicator variables.

experimental data

#	meta position (X)	para position (Y)	log $1/C$ obs.
1	F	H	7.52
2	H	F	8.16
3	Cl	F	8.19
4	Me	Me	9.30

structural matrix

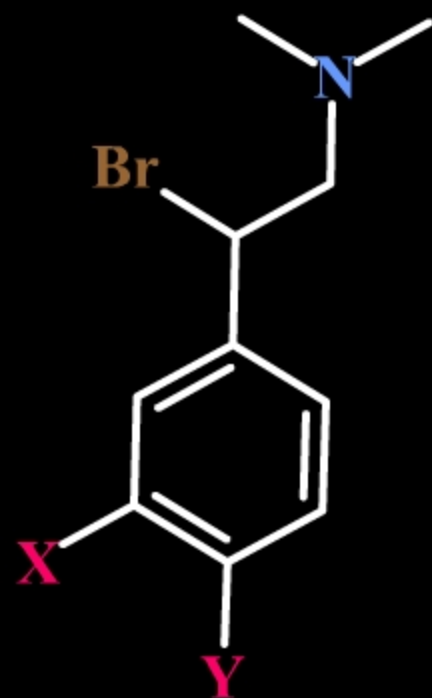| #	meta position (X)			para position (Y)			log $1/C$ obs.
	F	Cl	Me	F	Cl	Me	
1	1	0	0	0	0	0	7.52
2	0	0	0	1	0	0	8.16
3	0	1	0	1	0	0	8.19
4	0	0	1	0	0	1	9.30

F1.2.25 Example of Structural Matrix

An example of a matrix for a set of 22 disubstituted N,N-dimethyl-α-bromophenethylamines with antiadrenergic activities is given below.

○ molecule 1-11   ○ molecule 12-22

#	meta position (X)					para position (Y)					log 1/C obs.
	F	Cl	Br	I	Me	F	Cl	Br	I	Me	
1	0	0	0	0	0	0	0	0	0	0	7.46
2	0	0	0	0	0	1	0	0	0	0	8.16
3	0	0	0	0	0	0	1	0	0	0	8.68
4	0	0	0	0	0	0	0	1	0	0	8.89
5	0	0	0	0	0	0	0	0	1	0	9.25
6	0	0	0	0	0	0	0	0	0	1	9.30
7	1	0	0	0	0	0	0	0	0	0	7.52
8	0	1	0	0	0	0	0	0	0	0	8.16
9	0	0	1	0	0	0	0	0	0	0	8.30
10	0	0	0	1	0	0	0	0	0	0	8.40
11	0	0	0	0	1	0	0	0	0	0	8.46

F1.2.26 Example of Free-Wilson Equation

A mathematical procedure known as multiple linear regression is used to derive the QSAR equation from the structural matrix, ultimately leading to the values of a_i and μ for the 22 antiadrenergic agents.

matrix

$$
\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1
\end{array}
$$

multiple linear regression

$$\log 1/C = -0.301[m\text{-}F] + 0.207[m\text{-}Cl] + 0.434[m\text{-}Br] +$$
$$0.579[m\text{-}I] + 0.454[m\text{-}Me] + 0.340[p\text{-}F] +$$
$$0.768[p\text{-}Cl] + 1.020[p\text{-}Br] + 1.429[p\text{-}I] +$$
$$1.256[p\text{-}Me] + 7.821$$

$n = 22;\ r = 0.969$

F1.2.27 Predictability of the Model

The experimental and calculated values of the antiadrenergic molecules of the training set are indicated below and show that the Free-Wilson model reproduces the biological activities well. Moreover the equation can be used to predict the biological activities of new not yet synthesized analogs.

compound	log 1/C observed	log 1/C calculated	compound	log 1/C observed	log 1/C calculated
1	7.46	7.82	12	8.19	8.37
2	8.16	8.16	13	8.57	8.60
3	8.68	8.59	14	8.82	8.62
4	8.89	8.84	15	8.89	8.80
5	9.25	9.25	16	8.92	9.02
6	9.30	9.08	17	8.96	9.04
7	7.52	7.52	18	9.00	9.05
8	8.16	8.03	19	9.35	9.28
9	8.30	8.26	20	9.22	9.30
10	8.40	8.40	21	9.30	9.53
11	8.46	8.28	22	9.52	9.51

Compounds not yet synthesized ⇒ ⇒ Prediction of the Biological activity

F1.2.28 Understanding the Molecular Determinants

By examining the QSAR equation, we can draw several conclusions regarding the influence of the various substituents on activity. For example meta substituents contribute less than the para; the biological activity increases with electron donor ability etc...

$$\log 1/C = 7.821 - 0.301[m\text{-}F] + 0.207[m\text{-}Cl] + 0.434[m\text{-}Br] + 0.579[m\text{-}I] + 0.454[m\text{-}Me]$$
$$+ 0.340[p\text{-}F] + 0.768[p\text{-}Cl] + 1.020[p\text{-}Br] + 1.429[p\text{-}I] + 1.256[p\text{-}Me]$$

$$n = 22; \ r = 0.969$$

● **Biological activity increases with increasing lipophilicity**

● **Biological activity increases with electron donor ability**

● **Meta substituents contribute less than para substituents**

The topic Design of a QSAR Model contains the following 3 pages:

- Embarking on the Design of a QSAR Model
- The Four Steps
- An Iterative Process

F1.3.1 Embarking on the Design of a QSAR Model

The planning of a QSAR model must be carefully managed. In this section we will explore the methodology for designing a QSAR model in some detail, present the ideas and statistical concepts behind the QSAR model, the rules that need to be followed and the errors that should be avoided.

descriptors ?

parabolic ?

equations ?

linear ?

how many molecules ?

training set ?

correlation coefficient ?

What are the requirement ?

predictive ?

trend ?

normalize decriptors ?

molecule to synthesize ?
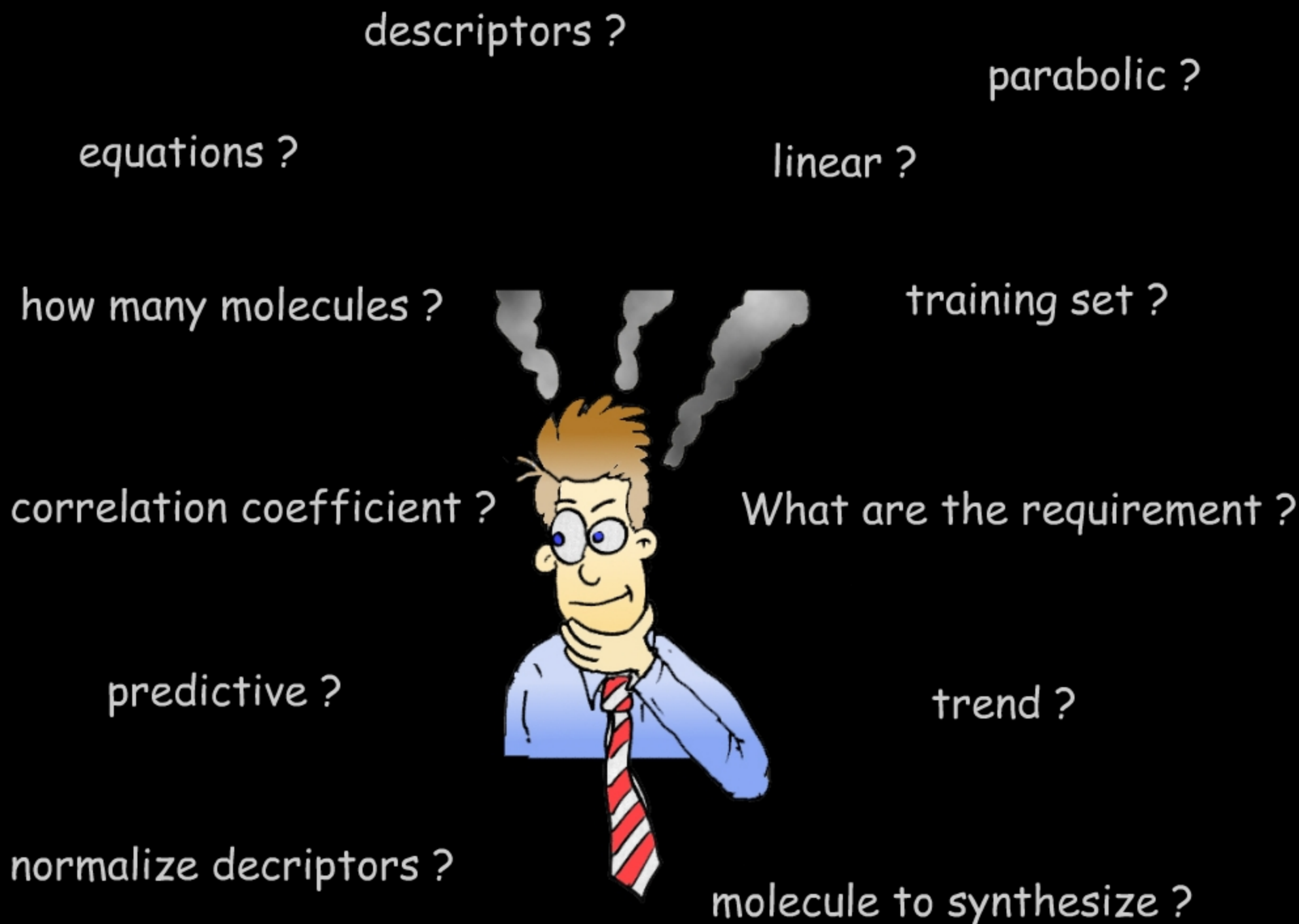
F1.3.2 The Four Steps

To construct a QSAR model the following steps should be followed: (1) assemble a sufficiently large and diverse set of compounds along with their biological activities; (2) select a set of descriptors which is likely to be related to the biological activity of interest; (3) formulate a mathematical equation that reflects the relationship between the biological activity and the chosen descriptors, and finally (4) validate the QSAR model.

- 1. Compounds Selection

- 2. Descriptors Selection

- 3. Building the QSAR model

- 4. Methods for Validating the model.

F1.3.3 An Iterative Process

Constructing a QSAR model is an iterative process. First, the QSAR equation is derived from an initial set of descriptors. Attempts are then made to improve this model by adding or removing descriptors and refining the mathematical equation, in an iterative fashion.

Compounds selection

Descriptors selection

The topic Compounds Selection: Step 1 contains the following 5 pages:

- Compounds Selection
- Predictions by Interpolation
- Example of Extrapolative Model
- Identification of Outliers
- Biological Activities in Terms of Log 1/C

F1.4.1 Compounds Selection

The selection of the compounds is the first step in building a QSAR model and consists of assembling a sufficiently large and diverse set of compounds with known biological activities. The molecules should be selected with great care in order to define a set of compounds that is homogenous and represents the system well.

Compounds selection

⬇

Descriptors selection

⬇

Building the QSAR model

⬇

Validating the model

F1.4.2 Predictions by Interpolation

The compounds selected for a QSAR analysis should cover a large range of values for those descriptors believed to be relevant to biological activity. This increases the probability that future compounds will have descriptors within this range and allow predictions to be interpolative rather than extrapolative. As a rule, interpolative predictions are more accurate than extrapolative predictions.

Poor compound selection

Better compound selection

Descriptor z

Descriptor n

Descriptor p

Descriptor s
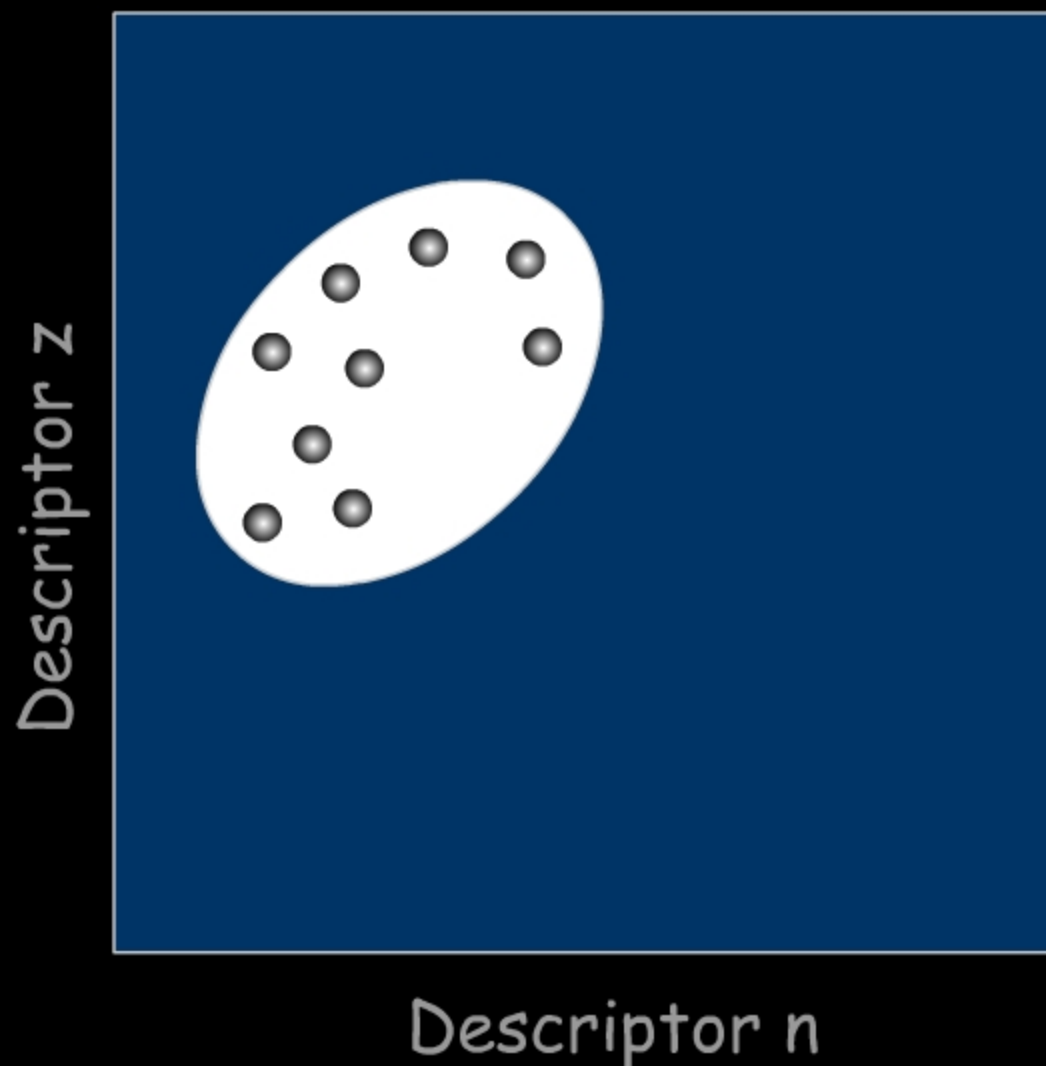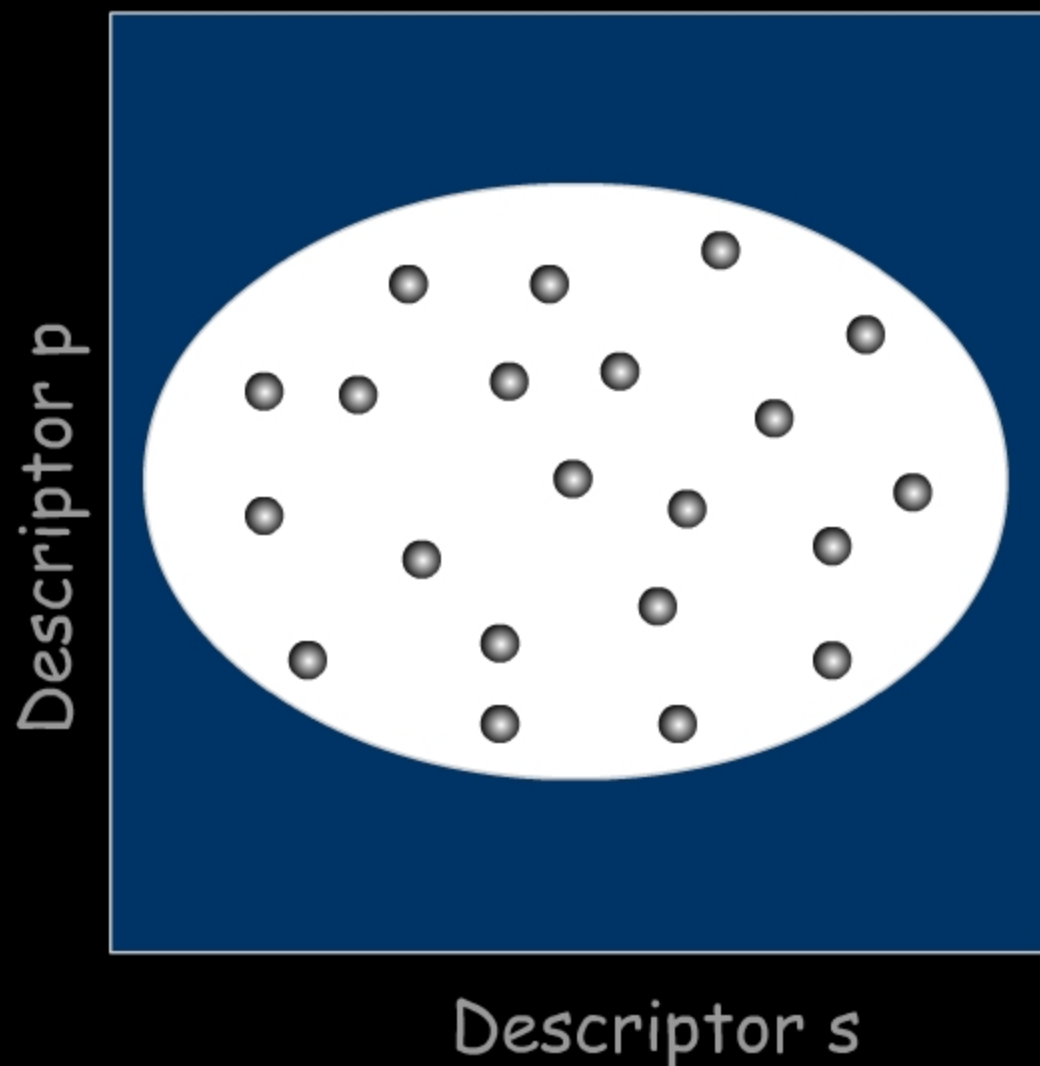
● Selected compound   ⬭ Interpolative zone   ◼ Extrapolative zone

F1.4.3 Example of Extrapolative Model

Extrapolating a model for values that are outside the range of the training set may lead to incorrect predictions. In the following example the experimental points lie in a straight line, however at higher values the model is more complex and no longer linear.

Start
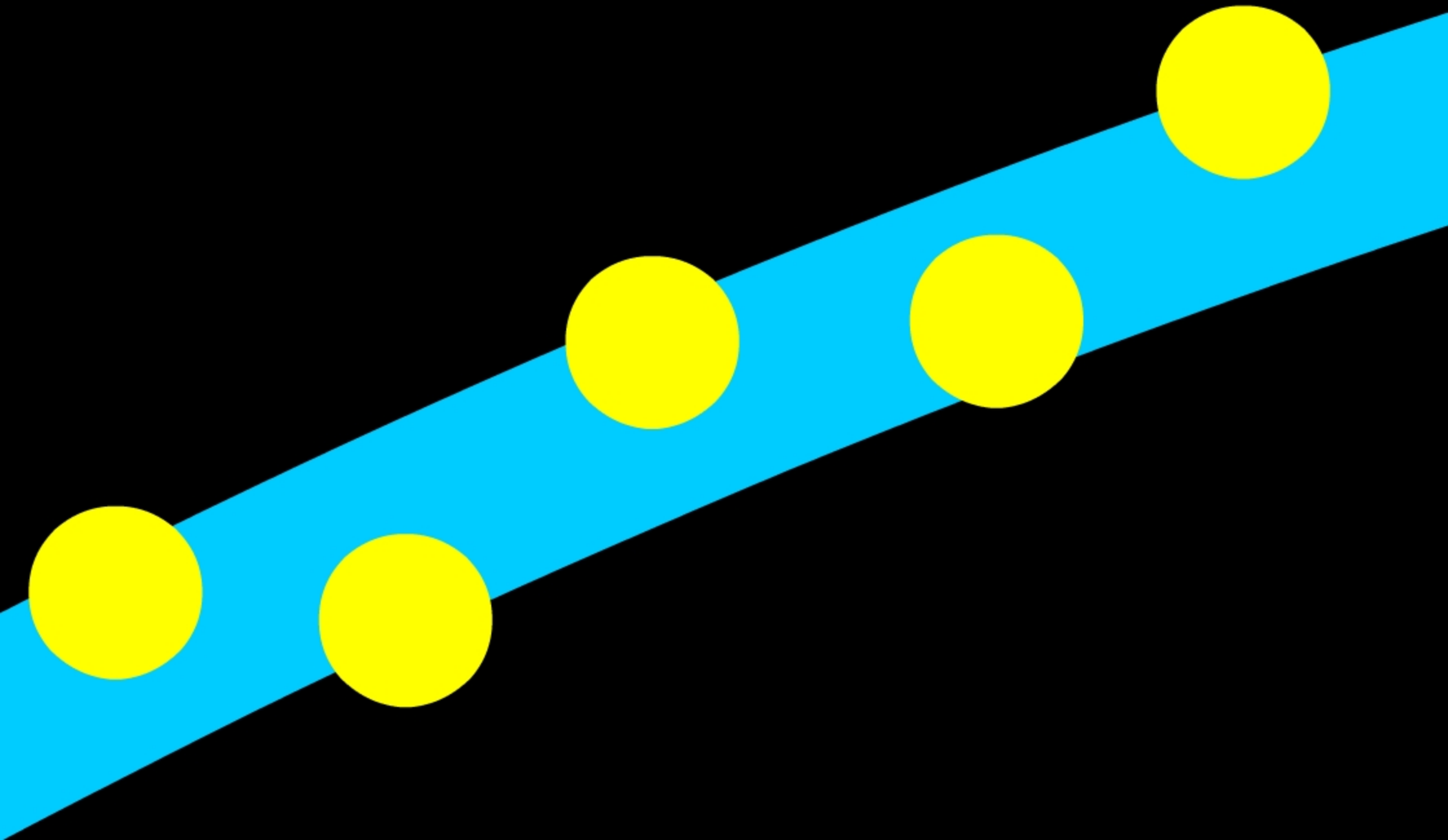
F1.4.4 Identification of Outliers

QSAR modeling is based on the assumption of homogeneity and an absence of influential outliers in the training set. An outlier can be a molecule acting according to a different mechanism of action, an improper biological activity as reported by another laboratory, or simply an incorrect value (experimental or typographic error). Repeat measurements of biological activities and using the greatest number of molecules helps reduce the distortions introduced by outliers.

○ with outlier   ○ without outlier
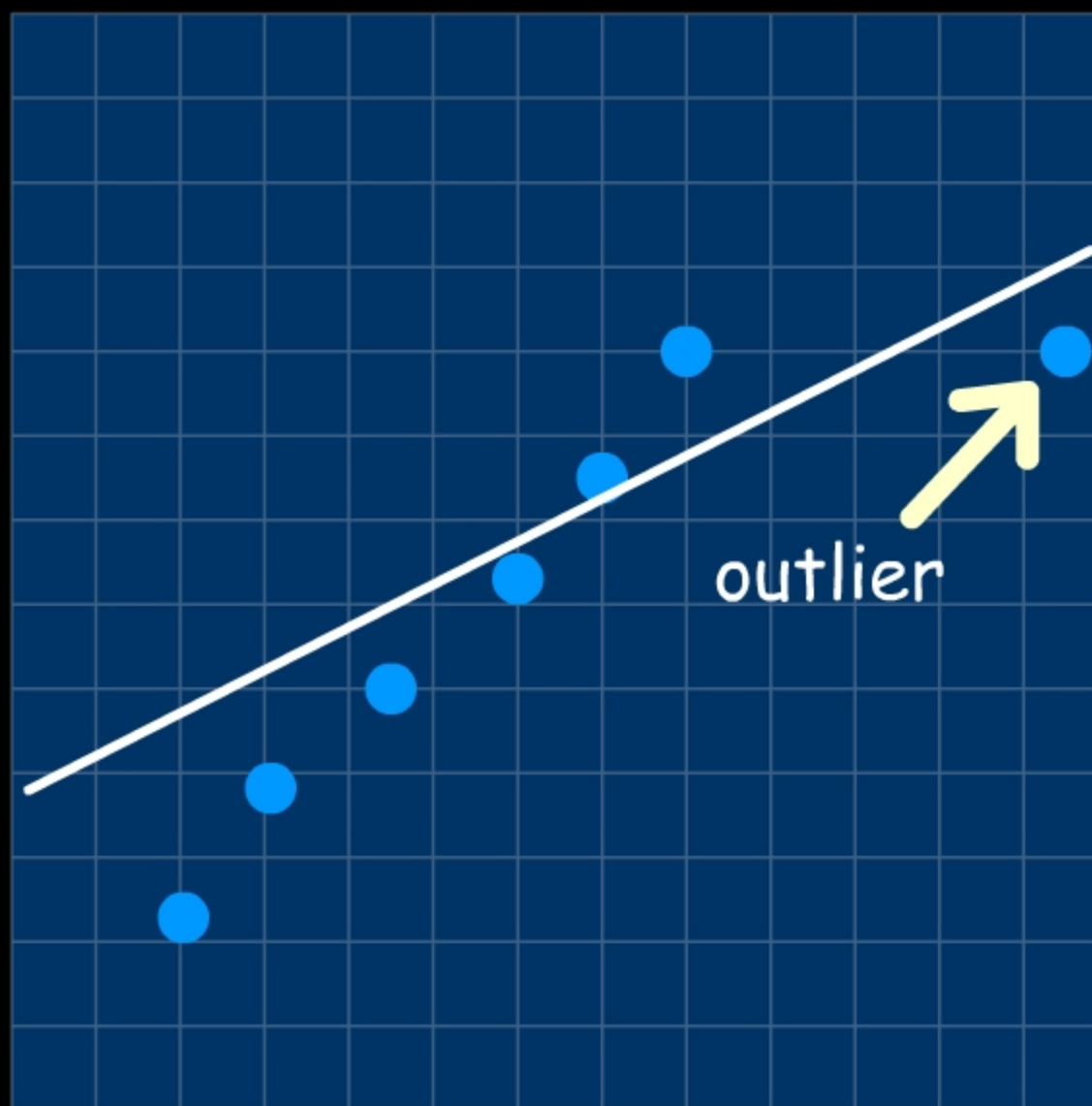
F1.4.5 Biological Activities in Terms of Log 1/C

To reflect the free energy variations that occur in the biological action, biological activities are expressed as the logarithm of the concentration of the compound (log[C] or log 1/[C]), where C is the concentration of the compound required to produce a given standard response. In the following set of equations E and S represent respectively, the enzyme and the substrate. Biological activities should be accurate and span at least 2-3 orders of magnitude.

$$E + S \rightleftharpoons ES$$

$$K = \frac{[ES]}{[E]\,[S]}$$

$$\Delta G = -2.3RT \log K \sim \log \frac{1}{[S]}$$

$$\Delta G \sim \log \frac{1}{[C]}$$

The topic Descriptors Selection: Step 2 contains the following 14 pages:

- Descriptors Selection
- Methods for Selecting Relevant Descriptors
    - Manual Selection of Descriptors
    - Automated Selection of Descriptors
    - Systematic Combination of Descriptors
    - Methods for Selecting a Subset of Descriptor
    - Forward Selection
    - Backward Elimination
    - Stepwise Regression
- Scaling Descriptors
- Correlation Between Descriptors
    - Example of Correlated Descriptors
    - Solution to the Problem of Correlated Descriptors
    ...

For the entire list, see the navigation panel.

F1.5.1 Descriptors Selection

As mentioned earlier in this chapter, the number of available descriptors for QSAR analyses is very large. A good model is based on a small number of well-chosen descriptors. When many descriptors are screened, a fortuitous correlation may occur. In the following pages important rules for the selection of relevant descriptors are presented.

Compounds selection

⬇

Descriptors selection

⬇

Building the QSAR model

⬇

Validating the model

F1.5.2 Methods for Selecting Relevant Descriptors

Relevant descriptors can be selected either manually or by using automated approaches. For each method, computer programs are available that help in the selection of relevant descriptors.
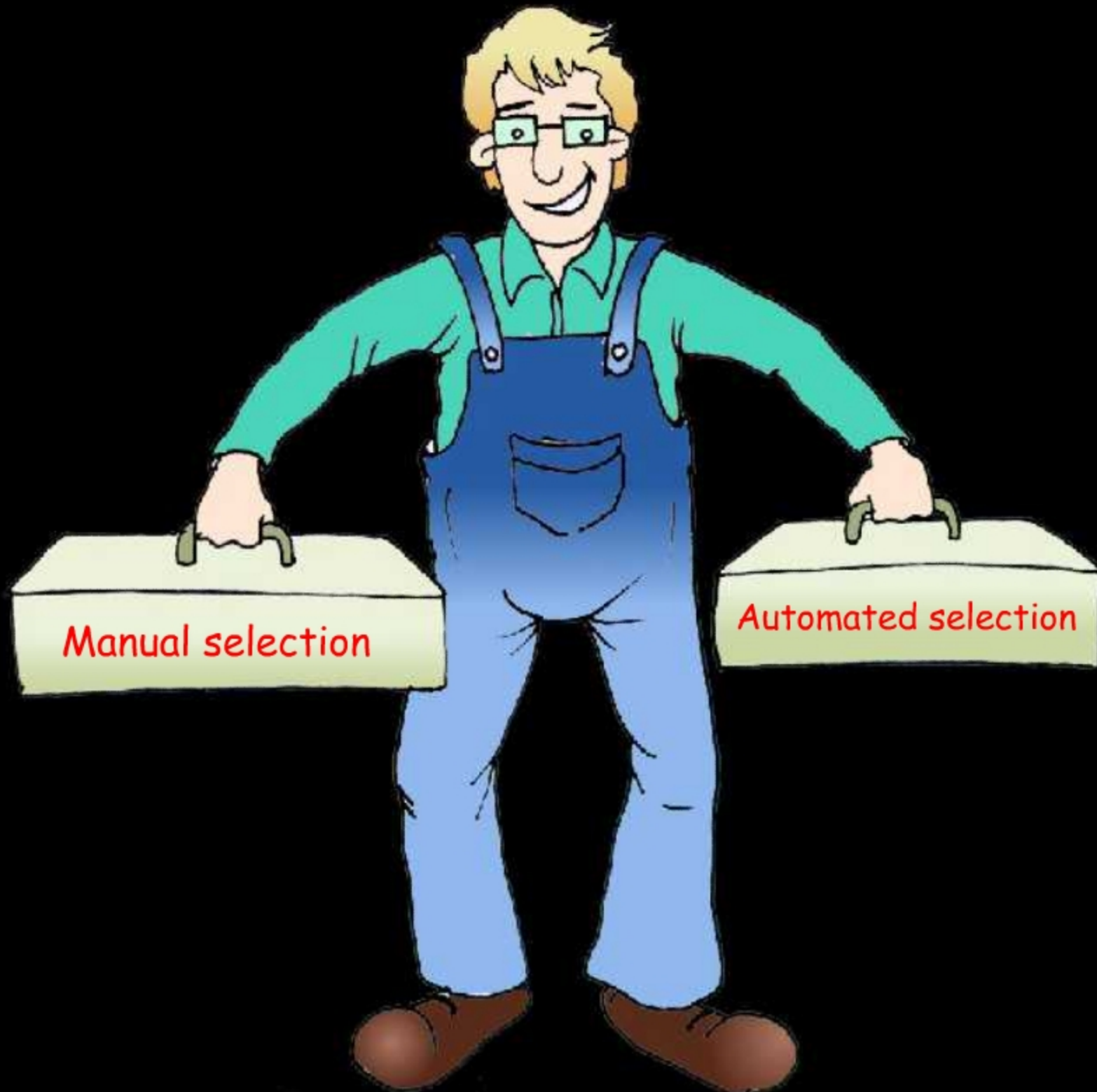
F1.5.3 Manual Selection of Descriptors

The manual method is based on a thorough understanding of the SAR and exploiting intuitions generated by the analyses. For example if preliminary analyses indicate that steric or hydrophobic substituents may increase activity, descriptors such as the molar refractivity (MR) and the hydrophobic substituent constant, π should be selected in the first place.

Interaction	Examples of corresponding descriptors
hydrophobic	π, log P
polar	σ
steric	E_s, MR

F1.5.4 Automated Selection of Descriptors

The second method looks at the selection of descriptors in an automated manner, using programs that score and rank them. Automated and manual methods can also be combined to select relevant descriptors and select those that are easy to interpret. Modern methods use genetic algorithms based on natural evolution principles (Darwin).
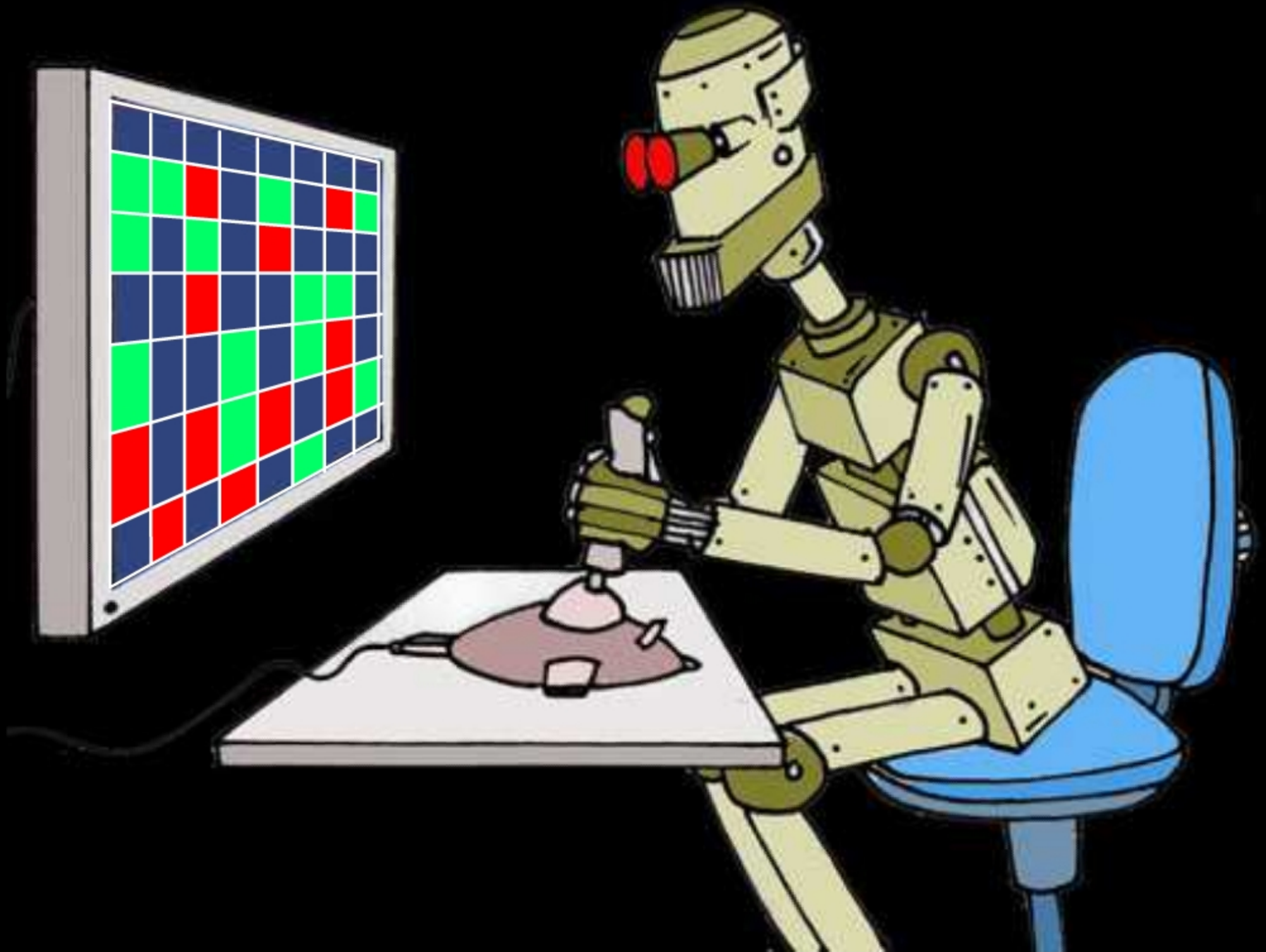
F1.5.5 Systematic Combination of Descriptors

In principle the identification of the best descriptors can be accomplished by a systematic evaluation of all their combinations. For each combination, a QSAR equation can be derived and then ranked. The highest-ranked equation will reveal the best subset of descriptors. However this systematic approach is not always feasible: for n descriptors (current software can process 2000), there are 2^n-1 different combinations (subsets). In the following pages we present automated methods that circumvent this difficulty.

○ Calculator   ○ Example

number of descriptors: [   ]   calculate

number of subsets: [   ]

F1.5.6 Methods for Selecting a Subset of Descriptor

"Forward regression", "backward elimination" and "stepwise regression" are methods for selecting a subset of descriptors from a large descriptor pool. The process starts with an initial subset of descriptors, then successive small alterations of this subset are made and assessed. If this modification improves the model, the change is accepted, otherwise it is rejected. The treatment is terminated when it is not possible to improve the model further.

F1.5.7 Forward Selection

The "forward selection" method starts with the single descriptor which best correlates with the dependent parameter. At each subsequent step the method adds the next most contributing descriptor. The process stops when the addition of a descriptor does not improve the model's performance as assessed by appropriate statistical indices.

Start

1 2 3

Descriptors

4 5 6

Descriptors

Score
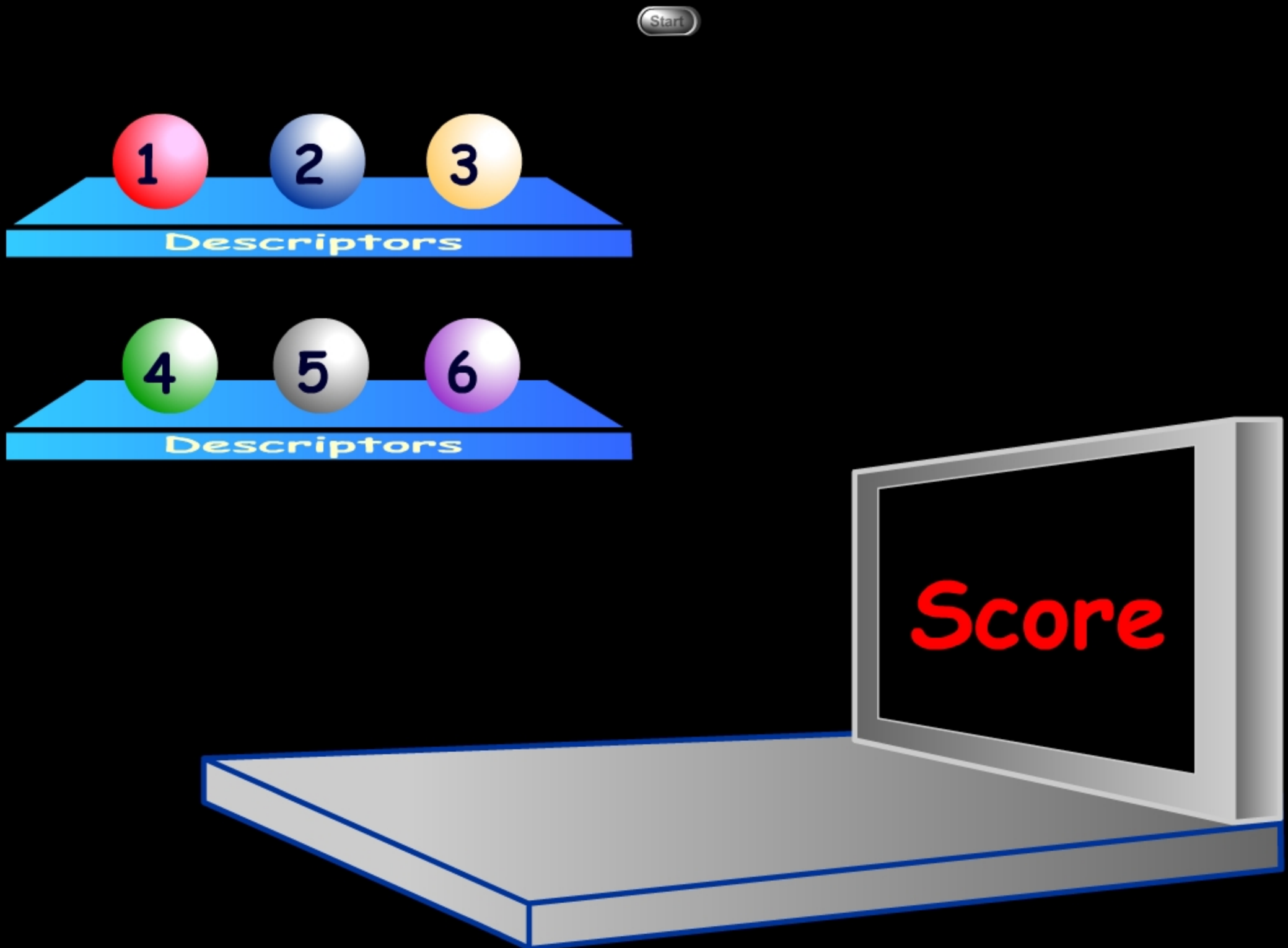
F1.5.8 Backward Elimination

The "backward elimination" method starts with a model that includes all the descriptors. At each step the method removes those descriptors that do not degrade the model's performance. The process is stopped when performance starts to decline as assessed by relevant statistical indices.

Start

Descriptors

Descriptors

Score

1 2 3 4 5 6

F1.5.9 Stepwise Regression

The "stepwise regression" method starts (like in forward selection) with the single descriptor that best correlates with the dependent parameter. At each subsequent step the method adds the next most contributing descriptor and can potentially remove non-contributing descriptors. The process is stopped when additional descriptors do not improve the model or when removing descriptors causes the model's performance to decline, as assessed by appropriate statistical indices.
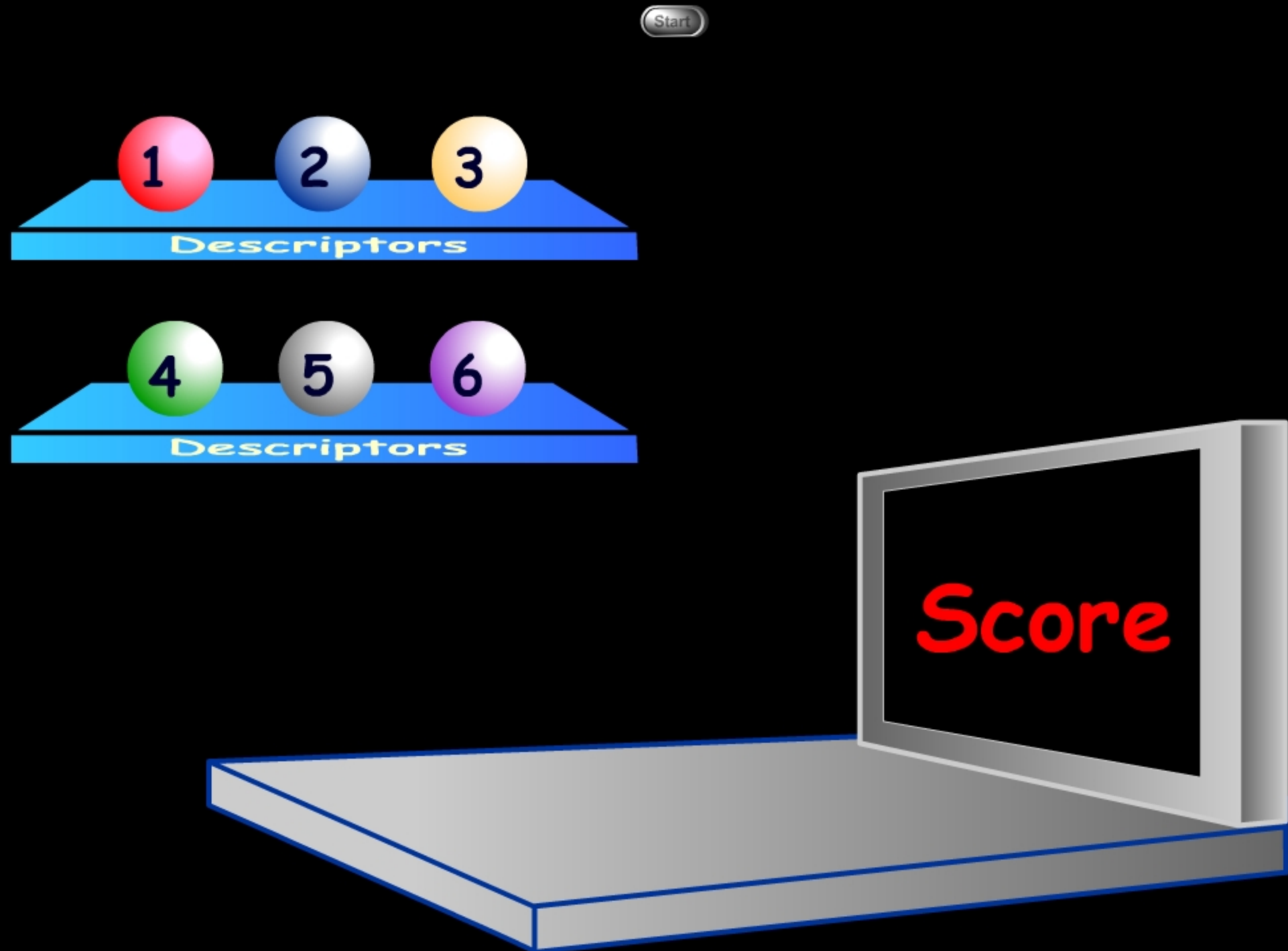
F1.5.10 Scaling Descriptors

Descriptors represent a broad range of physico-chemical properties. They need to be calibrated in order to provide a good balance of their respective influence when they are combined. Scaling treatment consists of a mathematical operation called "normalization" which sets boundaries for the variation of each descriptor.
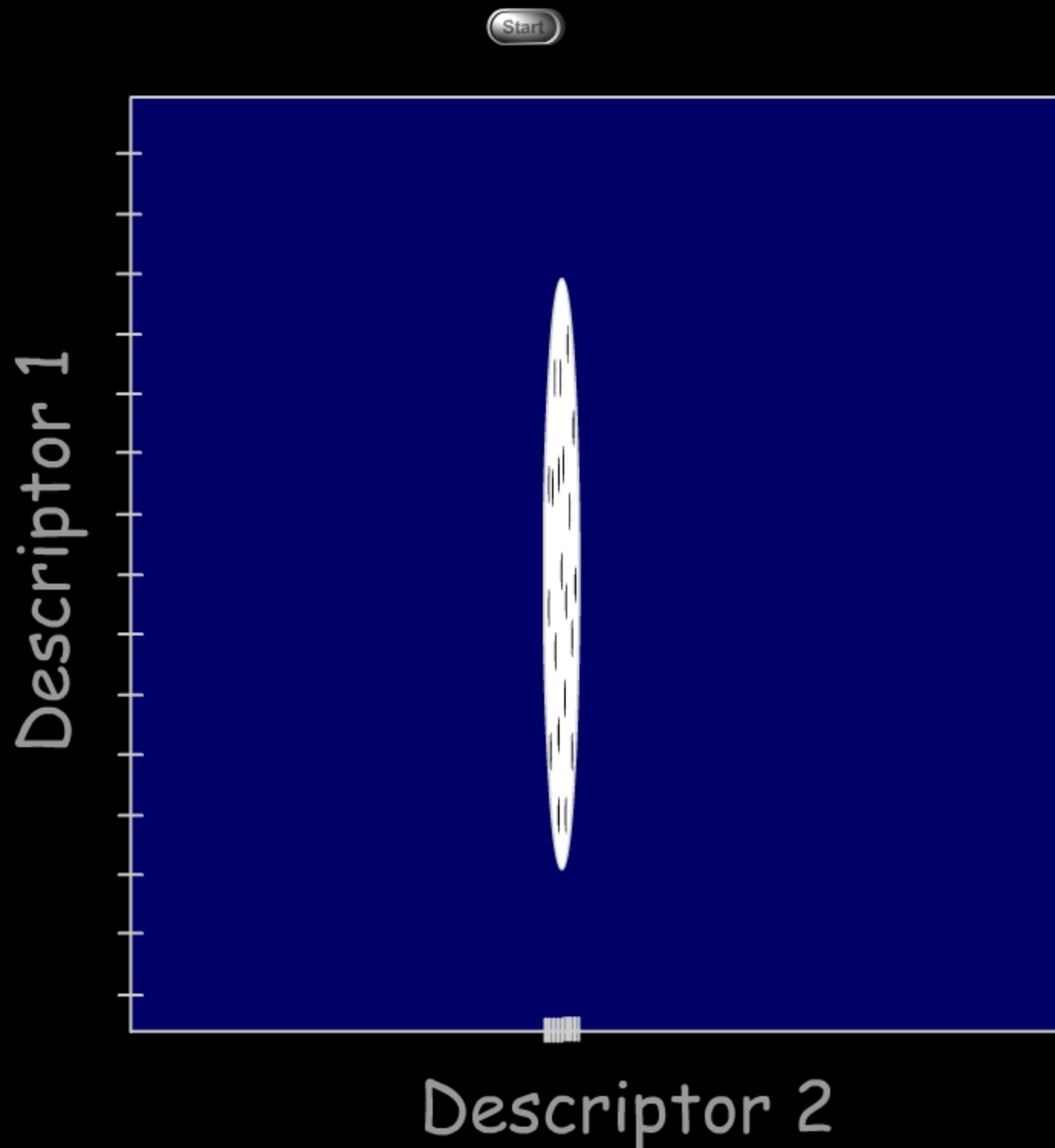
F1.5.11 Correlation Between Descriptors
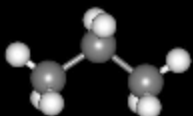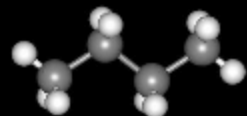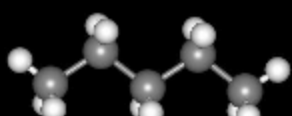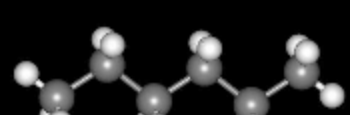
When two descriptors essentially convey the same information about a series of molecules they are said to be correlated. The use of correlated descriptors in the same equation must be avoided, because the information they characterize is over-represented when both are present. A "correlation matrix" provides useful information on the degree of correlation of different pairs of descriptors.

F1.5.12 Example of Correlated Descriptors

Consider for example the molecular weight and the number of carbon atoms as two descriptors characterizing a series of alkanes. These two descriptors are highly correlated, which can be shown graphically.
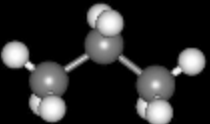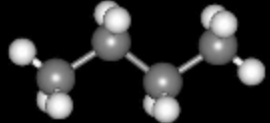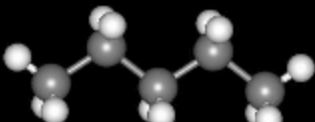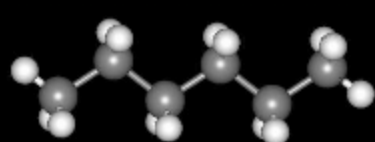
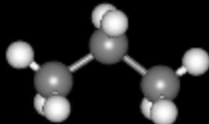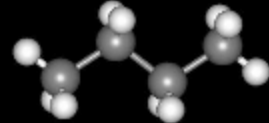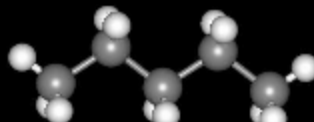| Structure	MW	# Carbons
	44.1	3
	58.1	4
	72.2	5
	86.2	6
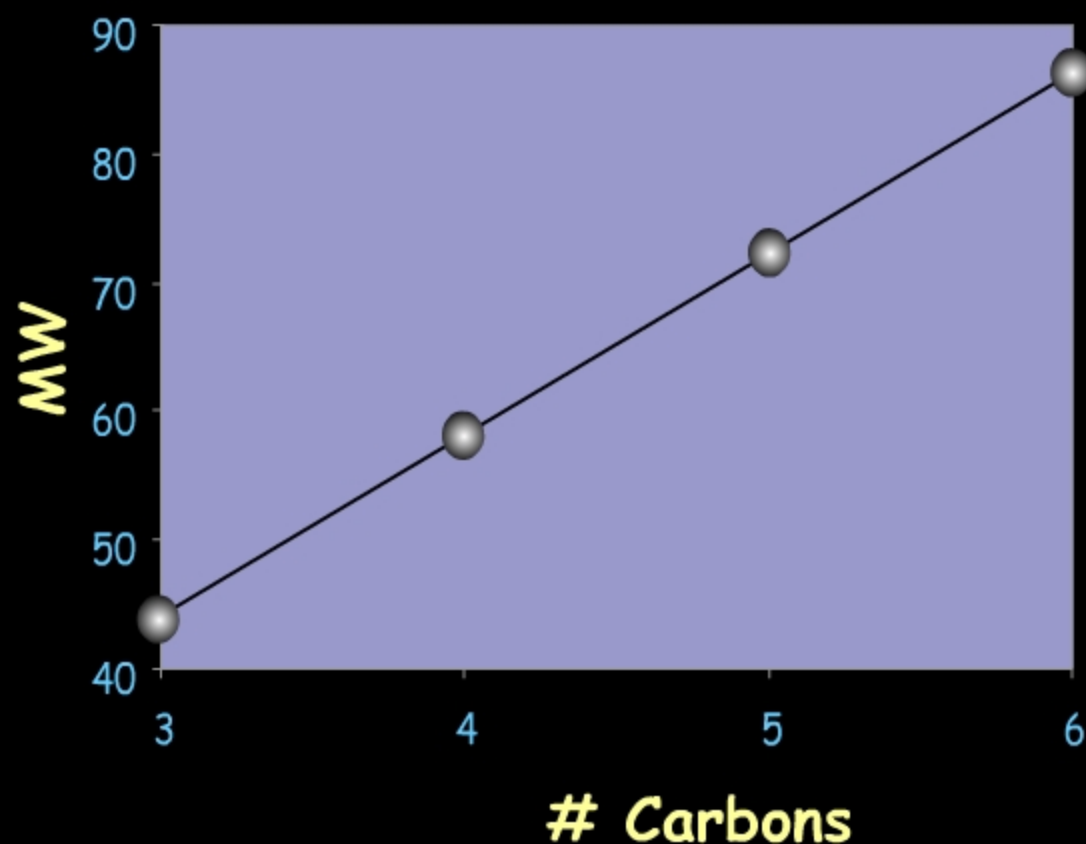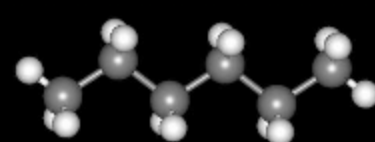
F1.5.13 Solution to the Problem of Correlated Descriptors

When two descriptors are highly correlated, the solution is to remove one of them. The descriptor that carries strong structural information is preferred and the less intuitive one is removed. An alternative solution consists of removing the descriptor that has the highest correlation with the other descriptors.

Structure	MW	# Carbons
	44.1	3
	58.1	4
	72.2	5
	86.2	6

Structure	MW
	44.1
	58.1
	72.2
	86.2

There is a general consensus that in a meaningful QSAR equation, the number of molecules in the training set should exceed the number of descriptors by a factor of 3 to 5.

n descriptors

molecules	activity	d_1	d_2	d_3	d_n
1	-0.23	-1.2	0.2	1.2	-0.2
2	0.50	3.5	0.5	3.5	54
3	2.10	5.1	2.1	21	53
4	-0.70			-7	23.3
5					9.3
6					7
7					6
8					-1.2
9	0.46			-3.2	30
.		.	.	.	.
p	0.12	4	0.1	12	-30

p molecules

$$p > 3n$$

The topic Deriving the Equation: Step 3 contains the following 24 pages:

For the entire list, see the navigation panel.

Step 3 consists of deriving the QSAR equation corresponding to the set of descriptors that were selected in the previous step.

Compounds selection

⬇

Descriptors selection

⬇

Building the QSAR model

⬇

Validating the model

F1.6.2 The Starting Point: The Study Table

The starting point for deriving a QSAR equation is the study table. It consists of a spreadsheet with molecules across the rows and molecular characteristics (biological activity, descriptors) down the columns. Typically, the first column indicates the molecular identification (e.g. compound number or name, 2D structure), the second column its activity, and subsequent columns the values of the corresponding descriptors.

Property of interest ↓   **Descriptors** ←——→

Compound	Activity	LogP	MR	MW	HOMO	Density
1	98	-4.03	87.10	332.2	-12.0	1.47
2	24	-3.68	76.53	324.4	-11.5	1.43
3	28	-4.34	91.23	290.3	-11.2	1.37
4	64	-5.19	100.2	310.1	-9.2	1.36
5	18	-5.59	91.32	291.5	-10.2	1.41
n	52	-4.83	72.12	340.3	-11.3	1.36

F1.6.3 Graphical Analysis of the Data

The study table should lead to graphical analyses. This step is of paramount importance and leaves room for "hunches" and preliminary interpretations. This is where the key questions are asked: is there an order? Are the points distributed according to known patterns? Can the recognized trends be translated into physico-chemical expressions? etc...
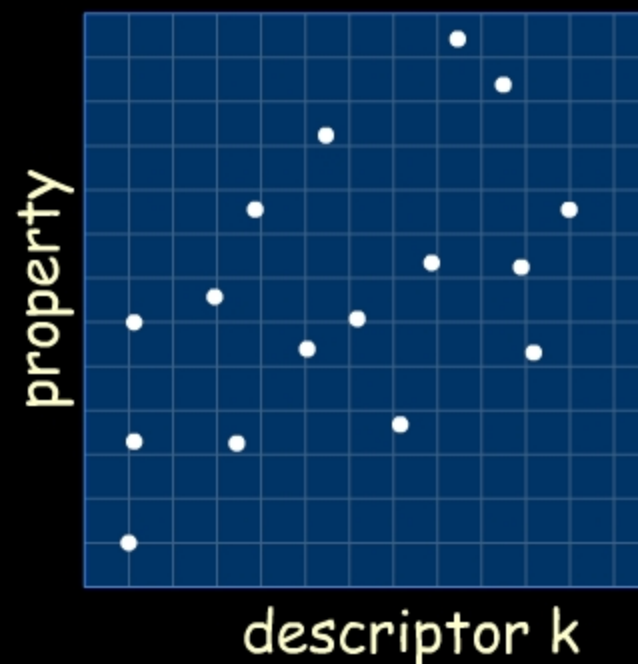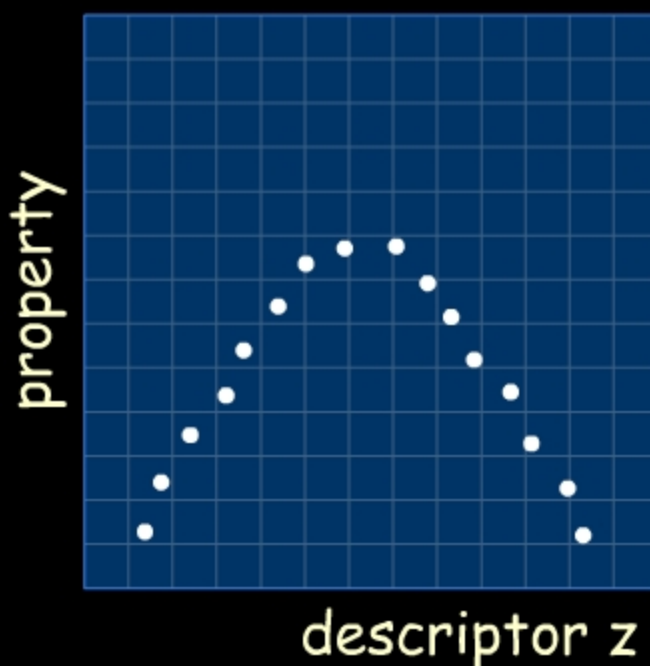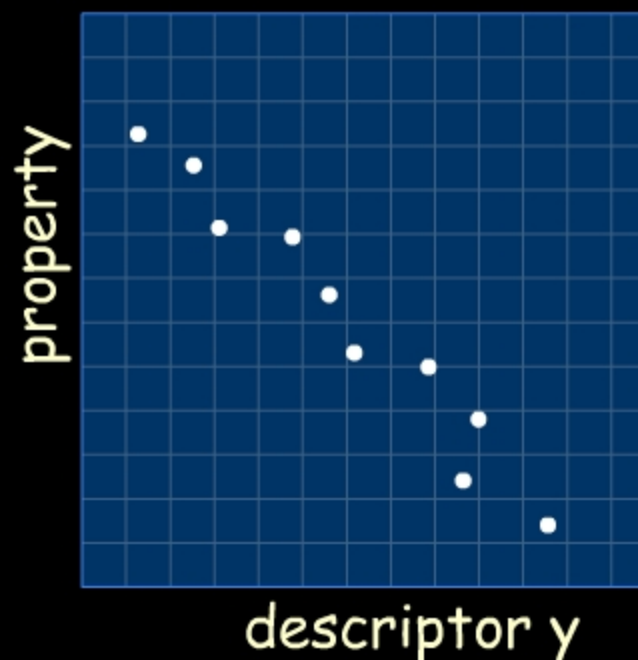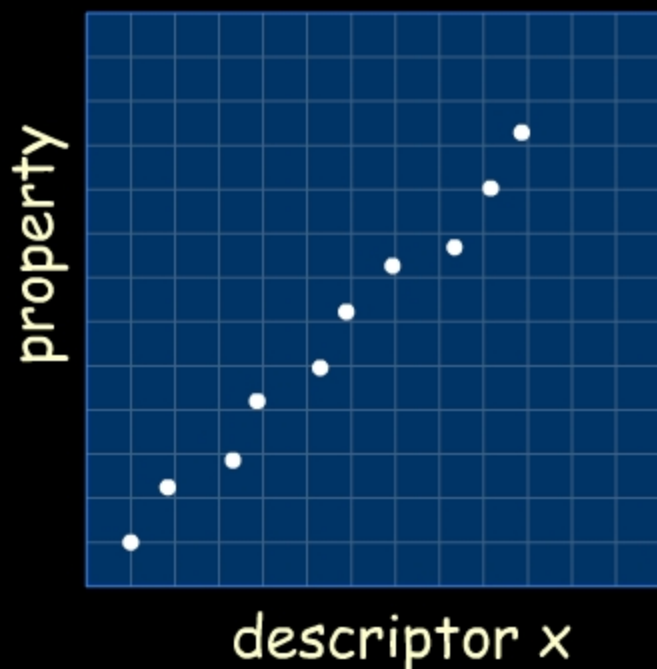
F1.6.4 Choice of the Mathematical Equation

After having identified trends in the system, the correlation process can begin. The initial analyses help guide the choice of the right mathematical equation. This equation should not be treated as a black-box; rather it should contain information that reflects the behavior and allows for interpretation of the system in a structural manner. Sound structural informational content in a QSAR equation is of utmost importance for formulating step 3.

F1.6.5 Complexity Levels and Data Overfitting

The next hurdle is the mathematical equation. At this stage the complexity of the model depends on both the form of the mathematical equation and the number of descriptors considered.

single linear regression

(Activity vs Descriptor — straight line)
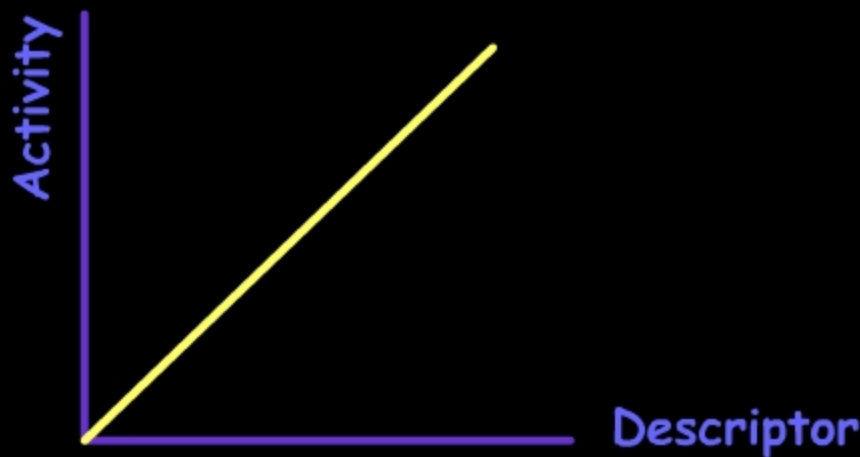
$$\text{Activity} = a\,(\text{descriptor}_1) + b$$

parabolic model

(Activity vs Descriptor — downward parabola)
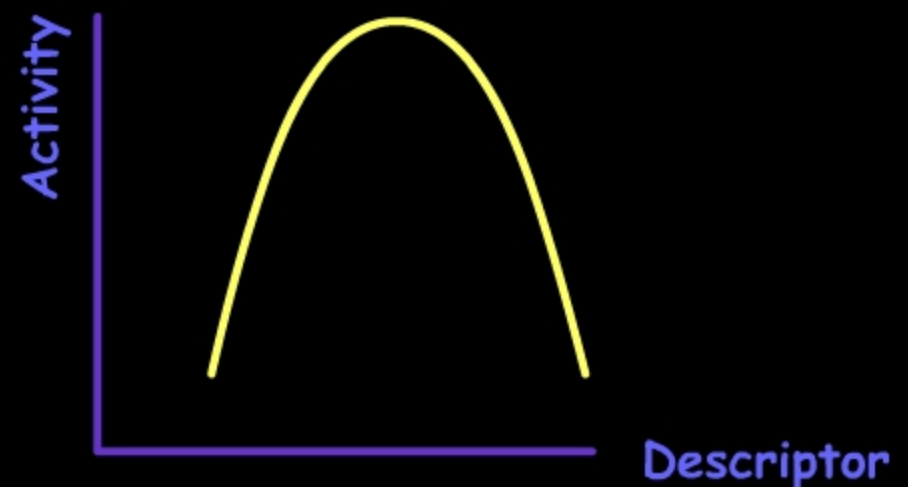
$$\text{Activity} = a\,(\text{descriptor}_1)^2 + b$$

multiple linear regression:

$$\text{Activity} = a(\text{descriptor}_1) + b(\text{descriptor}_2) + c(\text{descriptor}_3) + d \ldots$$

other models: parabolic, bilinear, probability, equilibrium etc...

F1.6.6 Mathematics are Very (too) Powerful

QSAR models can be skewed unintentionally by overly powerful mathematical choices. An equation that fits the data of a training set precisely can yield an equation that is perfect mathematically but meaningless for molecules other than those in the training set. For example if the training set consists of 20 molecules, it is always possible to select a set of 20 randomly chosen descriptors and solve the mathematical system for 20 equations and 20 unknowns. This error is known as data-overfitting.

Start

20 equations and 20 unknowns

$$\text{activity of } 1 = a_{1,1} d1 + a_{1,2} d2 + a_{1,3} d3 + \ldots + a_{1,20} d20$$

$$\text{activity of } 2 = a_{2,1} d1 + a_{2,2} d2 + a_{2,3} d3 + \ldots + a_{2,20} d20$$

$$\text{activity of } 3 = a_{3,1} d1 + a_{3,2} d2 + a_{3,3} d3 + \ldots + a_{3,20} d20$$

$$\text{activity of } 4 = a_{4,1} d1 + a_{4,2} d2 + a_{4,3} d3 + \ldots + a_{4,20} d20$$

$$\text{activity of } 20 = a_{20,1} d1 + a_{20,2} d2 + a_{20,3} d3 + \ldots + a_{20,20} d20$$

biological activities

unknowns

descriptors

F1.6.7 Illustration with an Example

To illustrate the data-overfitting problem, let's take a series of compounds for which the permeability through the blood brain barrier (BBB) has been found to be correlated with their logP and polar surface area. In the following graph we have plotted a hypothetical series of compounds in this space and color-coded them according to their BBB permeability. Compounds colored green are permeable whereas compounds colored red are not.
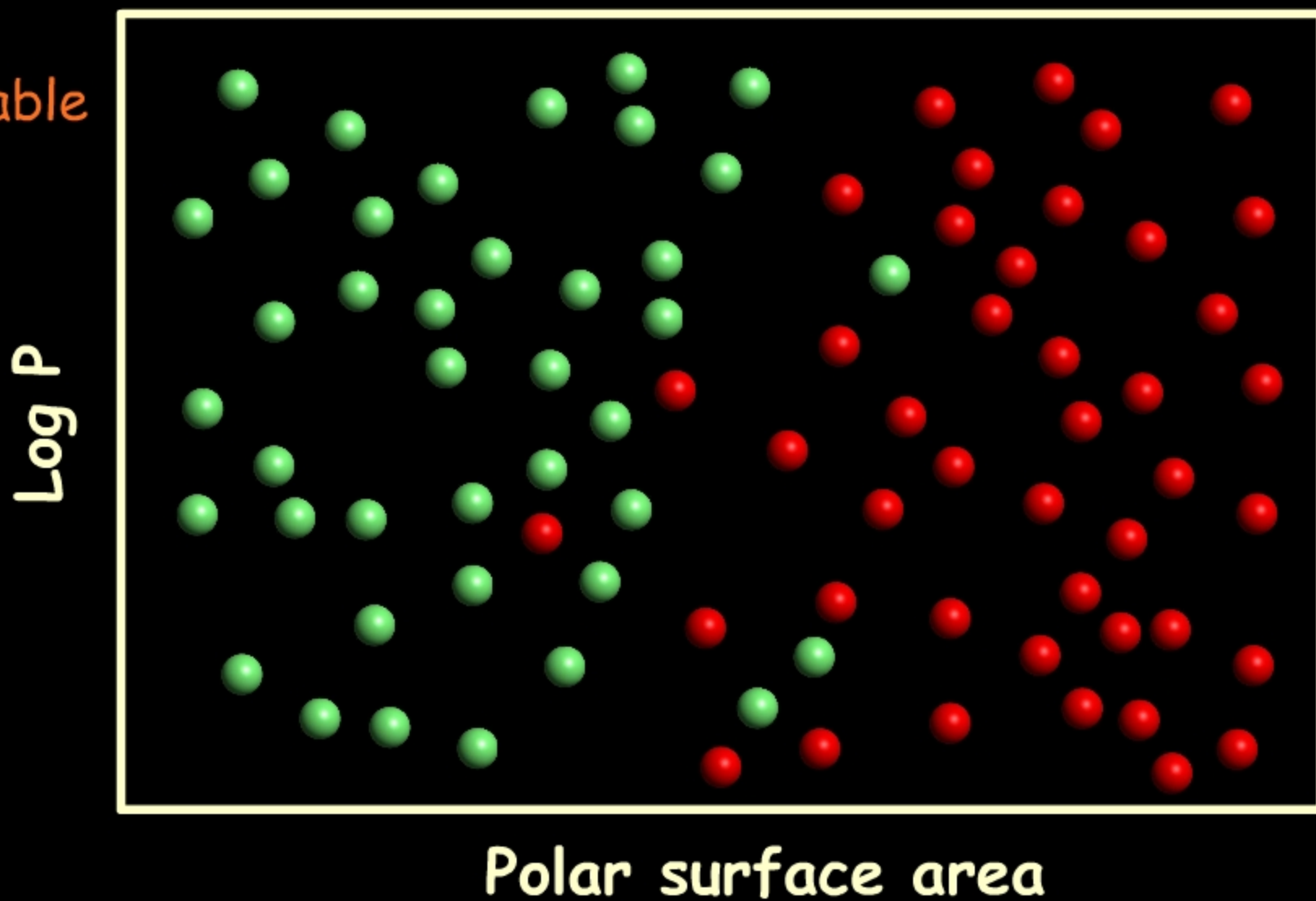
F1.6.8 A Simple Model

A linear model for differentiating between BBB permeable and BBB impermeable compounds can be formulated by drawing a straight line through the logP / Polar surface area space. Most of the compounds on the left side of the line are BBB permeable whereas most of the compounds on its right are BBB impermeable. As the model correctly classifies 45 out of the 50 compounds it has a success rate of 90%.

F1.6.9 A Complex Model

A model with an improved success rate can be generated by drawing a curved line across the logP / Polar surface area space. This model completely separates the BBB permeable compounds from the BBB impermeable compounds and thus has a success rate of 100%.
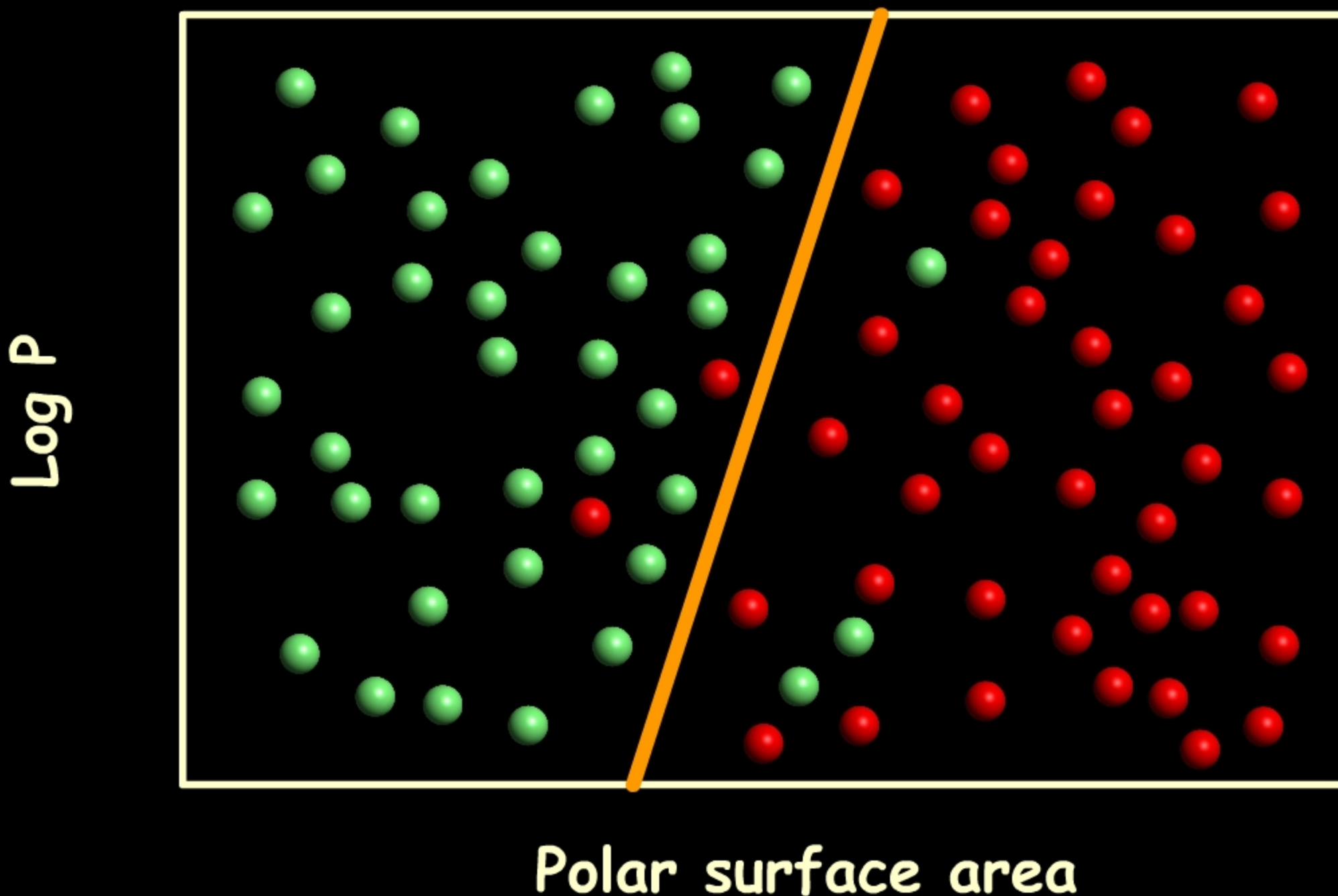
F1.6.10 Comparing the Two Models

Which of the two models better distinguishes BBB permeable from BBB impermeable compounds? Clearly the complex model has a higher success rate. However, by doing so it distorts its shape to correctly classify the outliers thereby completely reflecting the scatter of the training data - it is therefore an overfitted model. On the other hand, the simple model mislabels the outliers on the assumption that they are indeed outliers.
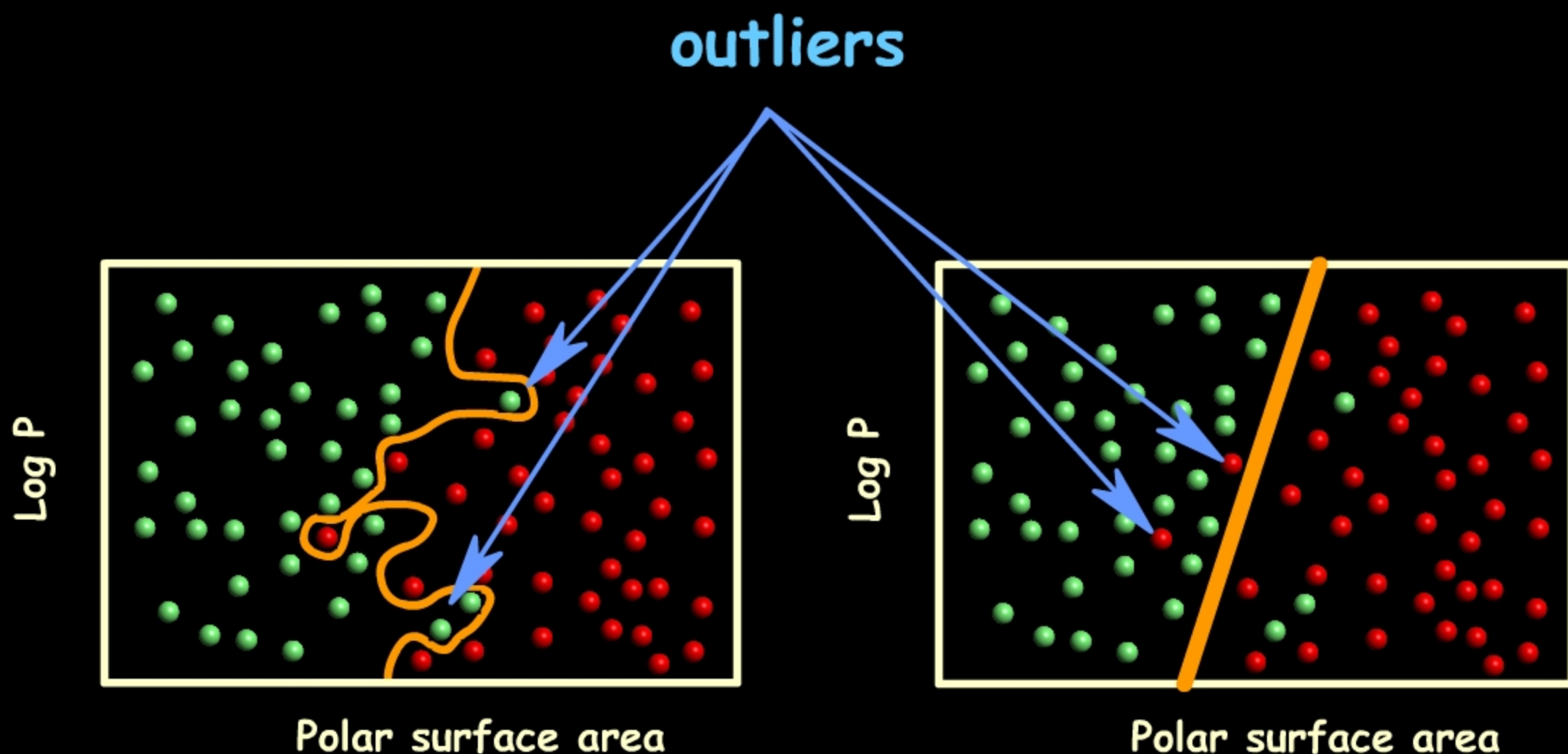
F1.6.11 Predictive Power of the Simple Model

The simple model predicts that all test compounds lying to the left of the line are BBB permeable and all those lying to the right of the line are BBB impermeable. Assuming that the test compounds are similar to the training compound, the prediction power of this model is expected to be high.

Predicted to be impermeable and is probably so

Log P

Predicted to be permeable and is probably so

Polar surface area
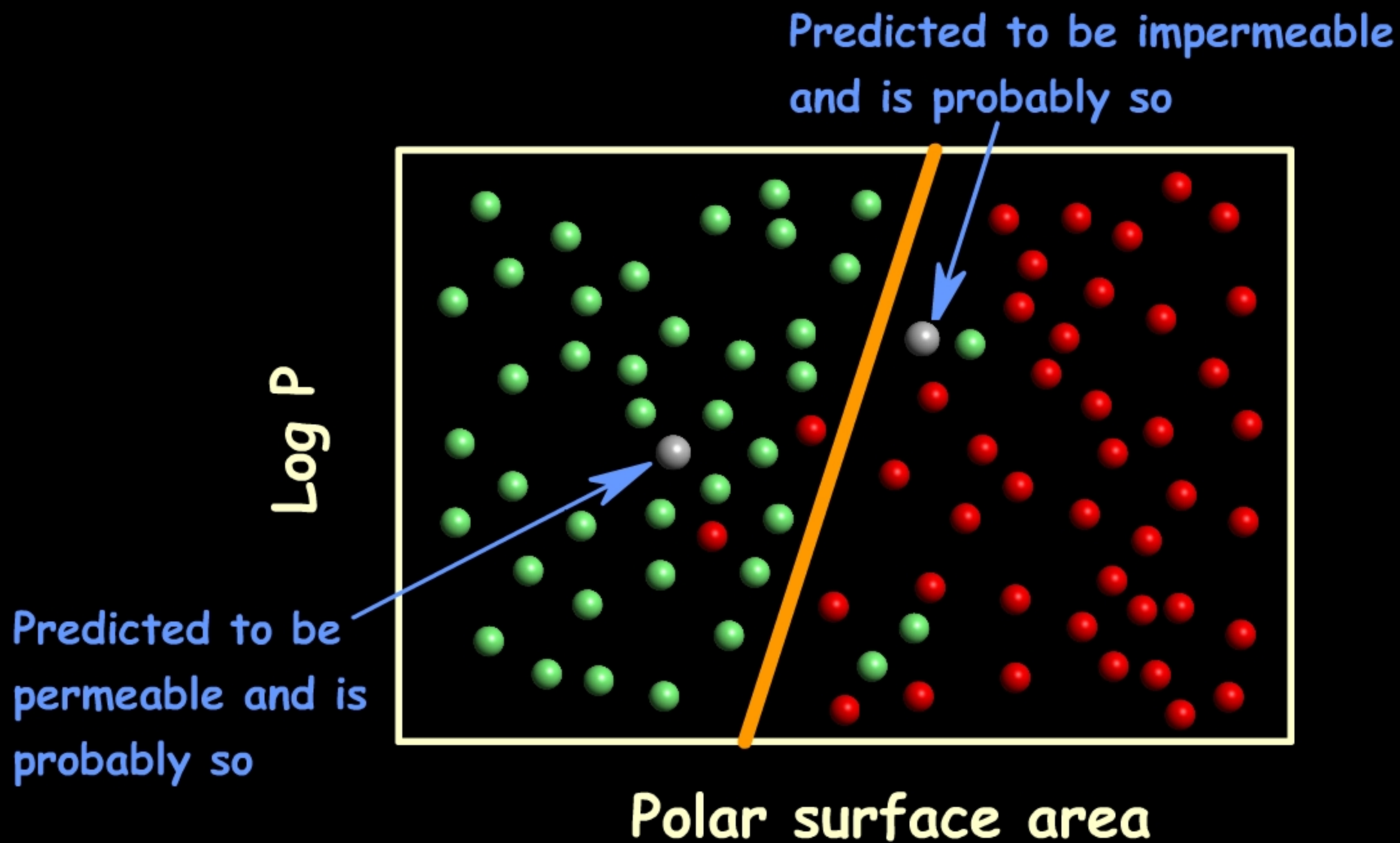
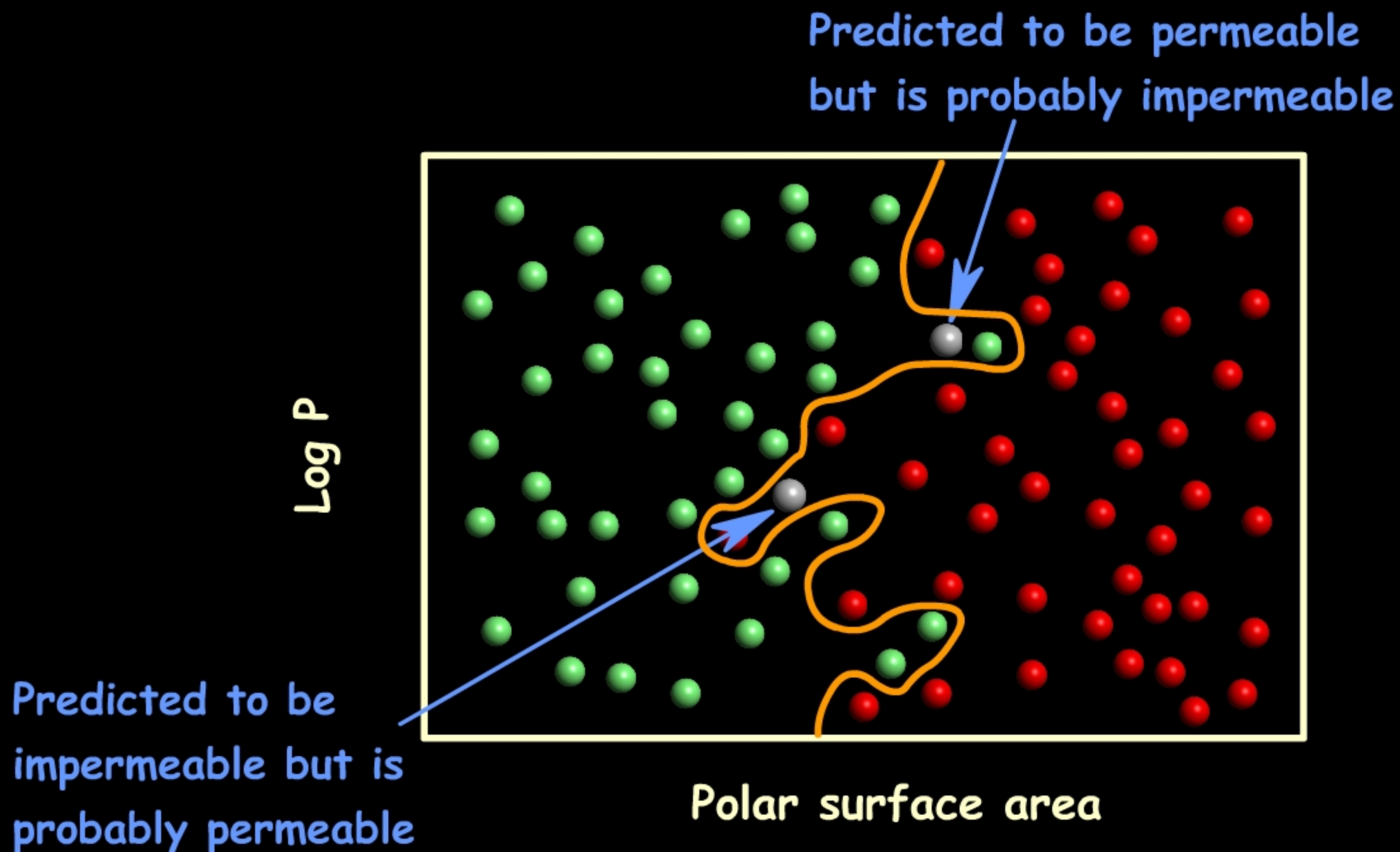## F1.6.12 Predictive Power of the Complex Model

The complex model also predicts that all test compounds lying to the left of the line are BBB permeable and all those lying to the right of the line are BBB impermeable. However, under the same assumption of similarity between test compound and training compound, many of its predictions are expected to be erroneous.

Predicted to be permeable but is probably impermeable

Predicted to be impermeable but is probably permeable

Log P

Polar surface area

F1.6.13 Complexity Dictated by Predictability of the Model

In the QSAR approach tailoring an equation to the peculiarities of a training set is not a problem. However, forcing the mathematics to fit too closely to the data may lead to meaningless models in terms of predictability (tools for assessing the predictability of a QSAR model will be presented in Step 4). The real issue is to stop the refinements early enough so that the predictive capabilities of the model are not lost.
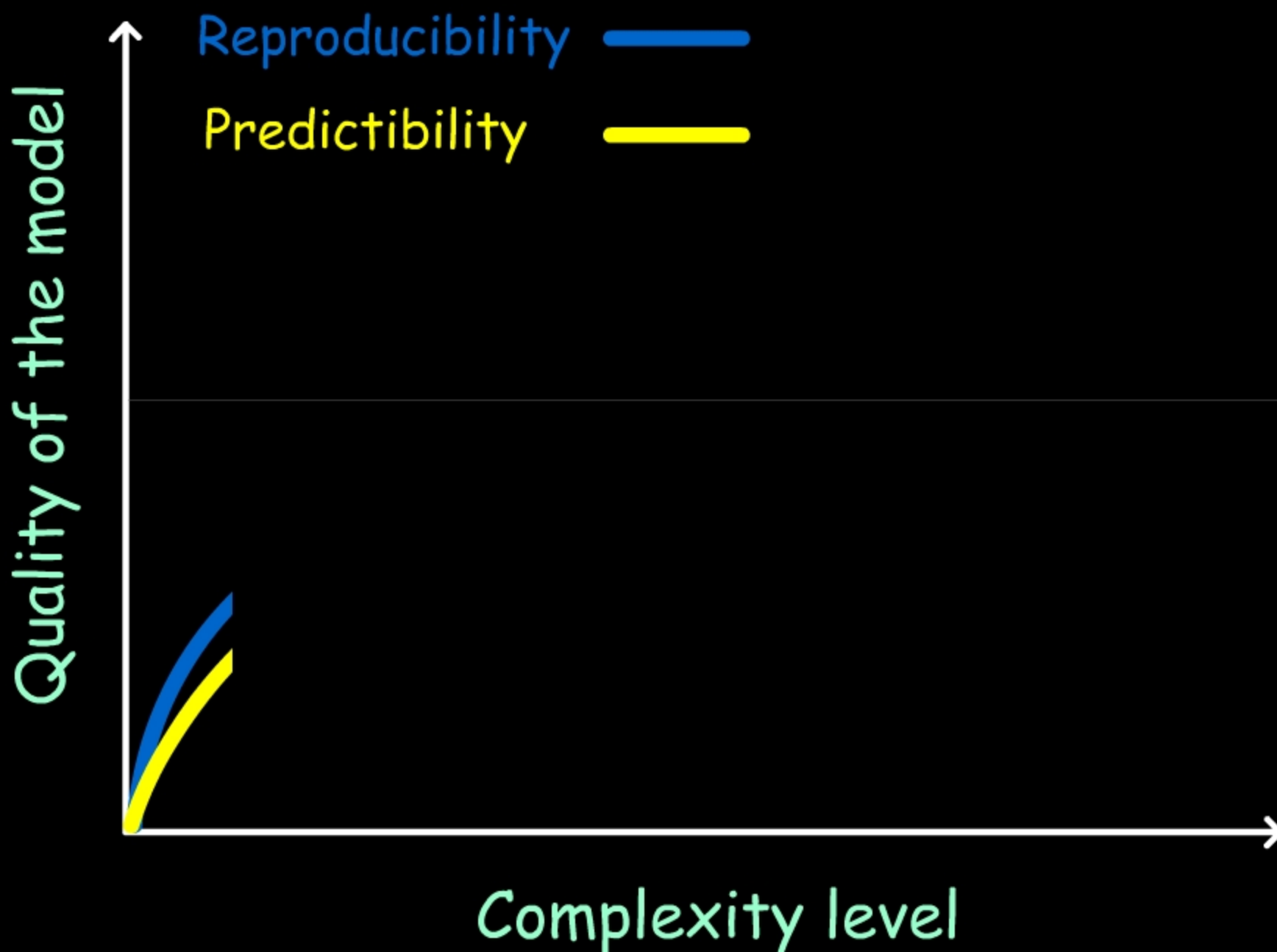
F1.6.14 Single Linear Equation: Mathematical Outline

The simplest form of a QSAR equation is a linear model with one descriptor. This simply yields the equation of a straight line of the form $y = b_0 + b_1 X$ where b_0 indicates the intercept of the line with the y axis and b_1 the slope of the line. b_0 and b_1 are calculated as described on the next page.
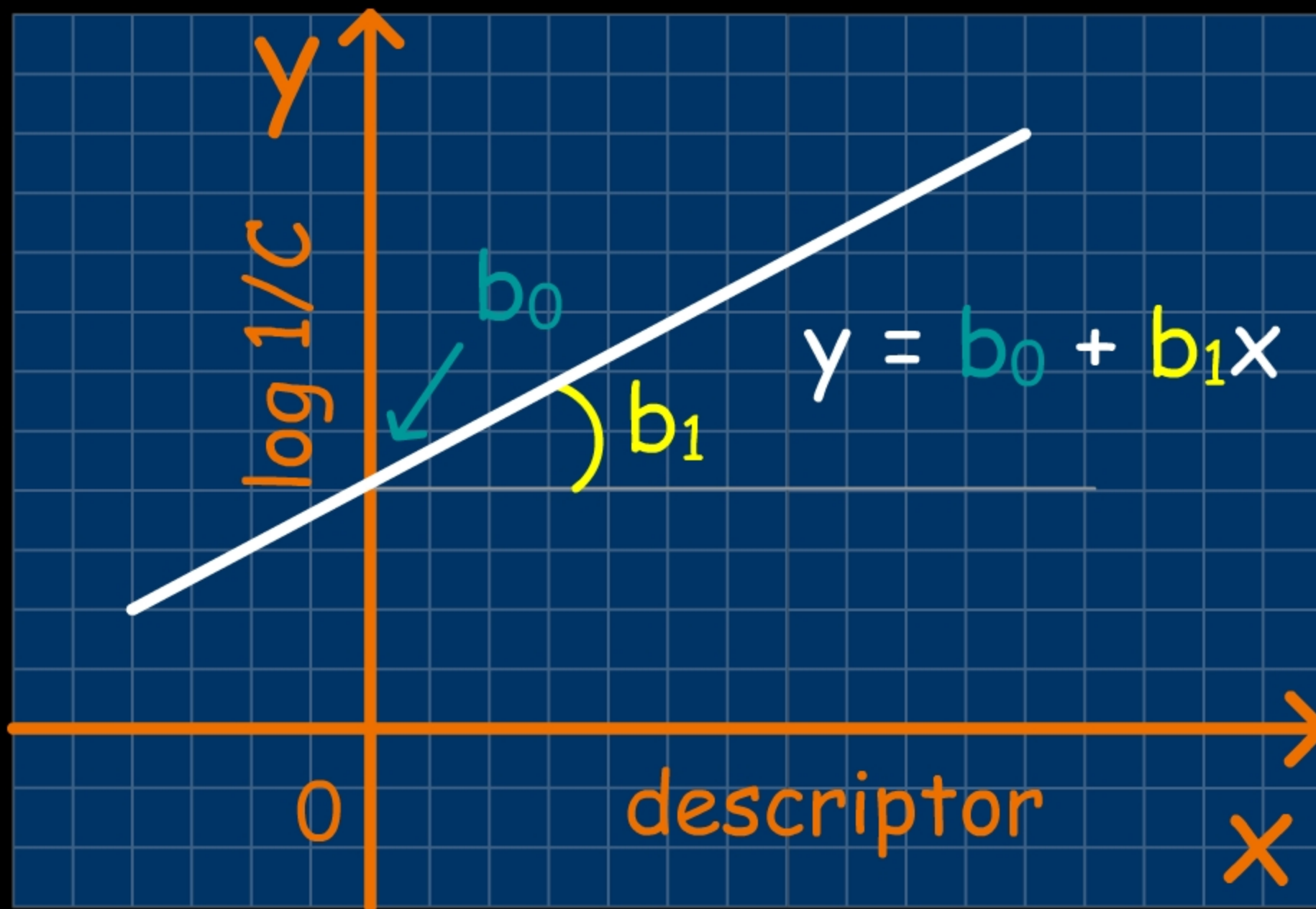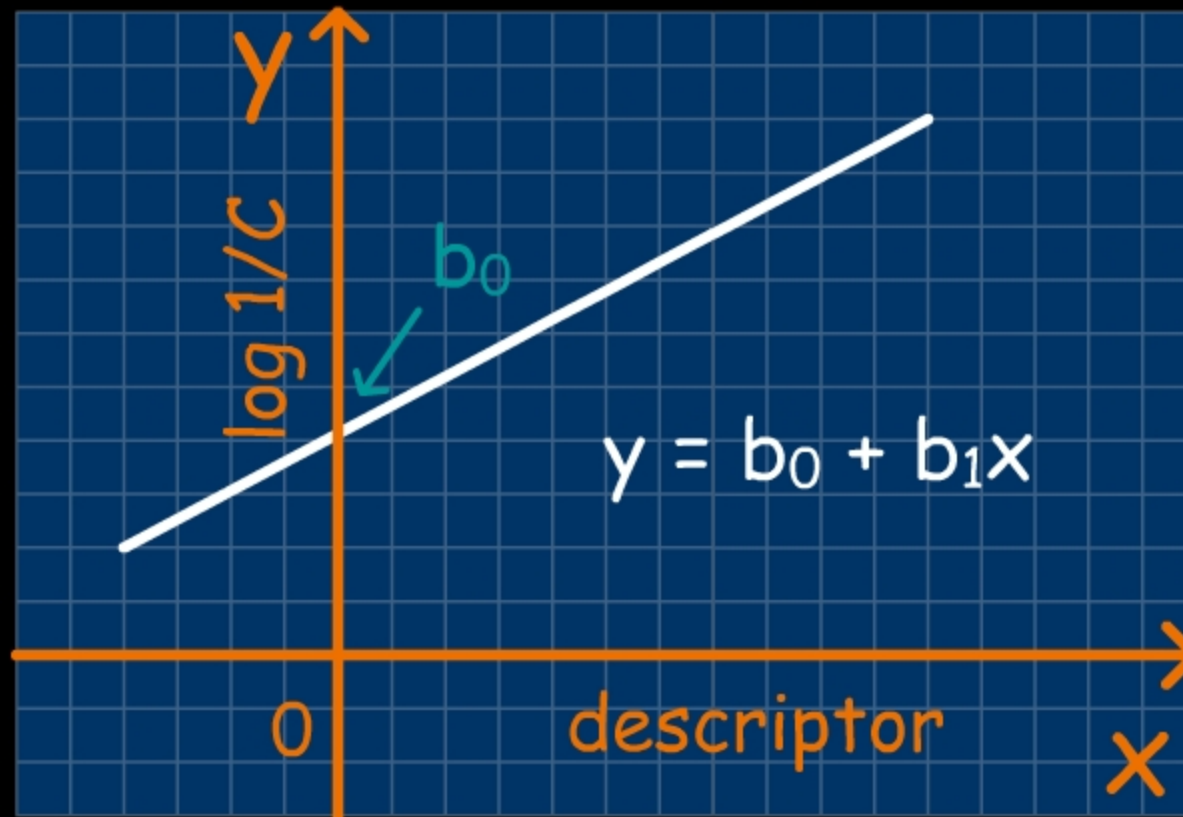
F1.6.15 Calculating b0 and b1

b_0 and b_1 are calculated using the two equations indicated below. The details of such calculations are presented for the Capsaicin example under the heading "Example of simple linear regression".

$$b_1 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \qquad\qquad b_0 = \bar{y} - b_1\bar{x}$$

F1.6.16 Multiple Linear Regression: Mathematical Outline

It is not always possible to correlate biological activities with a single descriptor (linear model with one descriptor). Given that biological action results from the combined influence of many factors, one can extend the QSAR model to multiple descriptors. Indeed, the observation that several parameters used simultaneously can lead to good models prompted the development of a method referred to as "multiple linear regression" (MLR). In this model linearity is maintained for each of the individual descriptors.

coefficients

$$\text{Activity} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n$$
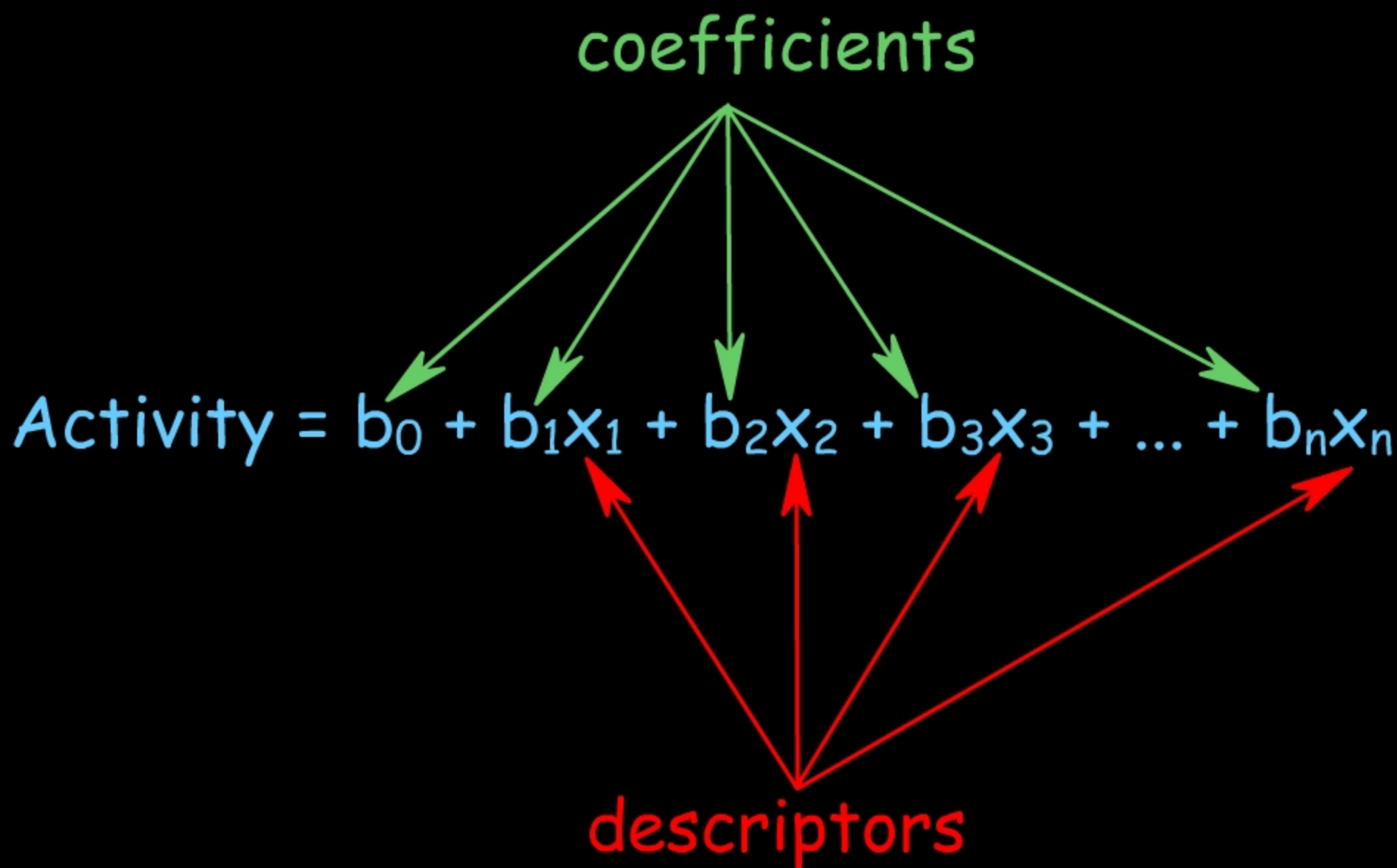
descriptors

F1.6.17 Example: MLR vs. Single Linear Models

The example of anticonvulsant compounds shown below demonstrates that each descriptor Es, σ and logP alone was not able to give a good correlation (r less than 0.40) with the biological activities. However, by using simultaneously logP and σ, a significant improvement was made (r=0.80). The addition of Es improves the model even more (r=0.95). This indicates that the biological properties result from the combined action of lipophilicity, steric and electronic effects.
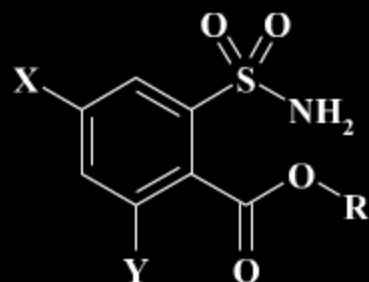
bad model ▮

good model ▮

model	r
log 1/C = 0.009 Es + 3.411	0.03
log 1/C = -0.626 σ + 3.314	0.27
log 1/C = -0.078 logP + 3.432	0.38
log 1/C = -0.210 logP - 2.214 σ + 3.154	0.80
log 1/C = 0.21 Es - 0.238 logP - 3.81 σ + 3.046	0.95

F1.6.18 The Mathematics of MLR: a Single Sample

In MLR we try to express activity as a linear combination of descriptors. We recognize the fact that in most cases, our fit to the experimental data will not be perfect and error is usually unavoidable. In the equations listed below, y (the activity) is a scalar; x_j is the value of the descriptor j and b_j its associated coefficient; e is the error. In the matrix notation, x^T is a row vector of the descriptors and b, a column vector of their associated coefficients.

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m + e$$

$$y = \sum_{j=1}^{m} b_jx_j + e$$

Matrix notation: $\quad y = x^Tb + e$

For the case of multiple compounds, the activity values are assembled into a vector y of length n, where n is the number of compounds. The descriptors are collected into an n by m matrix where n again is the number of compounds and m is the number of descriptors. The coefficients are collected into a vector of length m and the errors are collected into another vector of length n.

$$y = x\,b + e$$
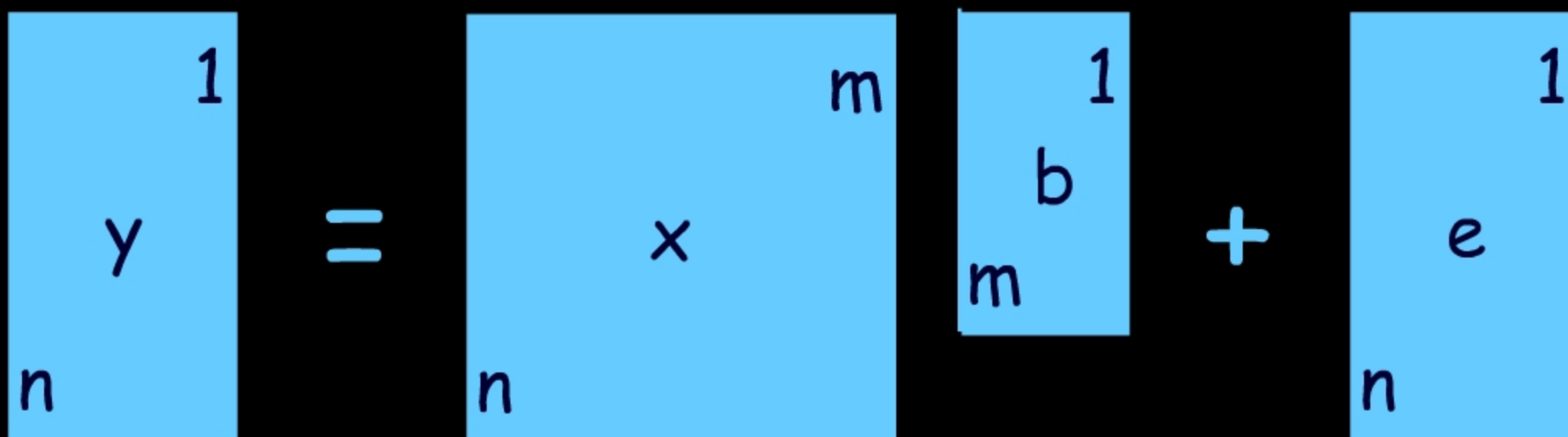
F1.6.20 The Solution of MLR

In the MLR formalism we search for the (unknown) set of coefficients b, which, when multiplied by the (known) descriptors, best approximates the (known) activity data (equation 1). A solution to this problem can be obtained through a matrix inversion procedure (equation 2).
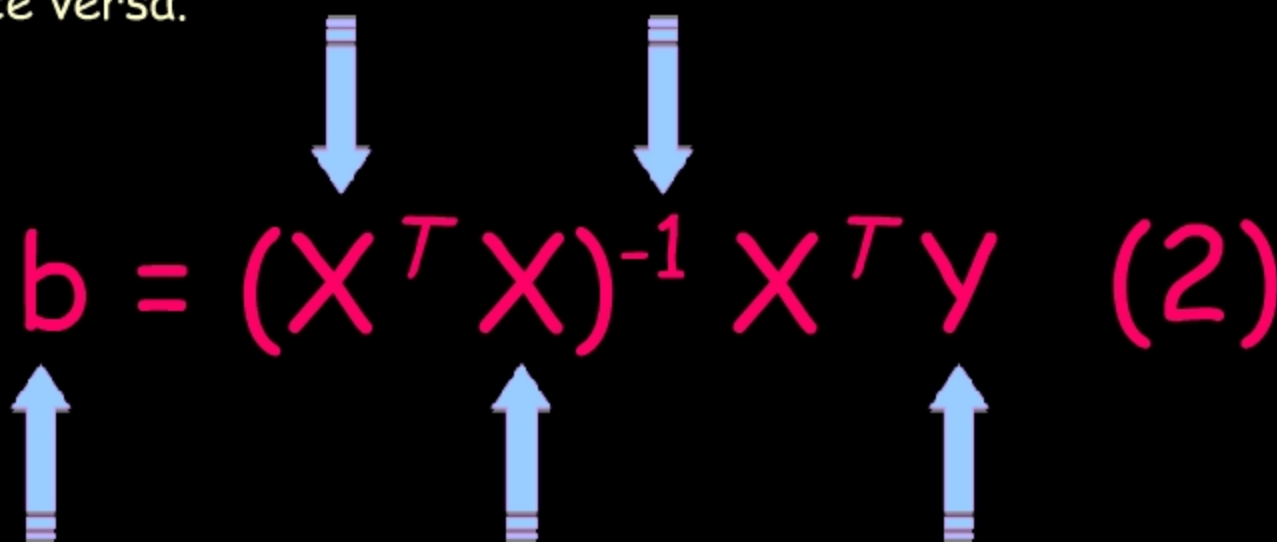
◉ coefficients   ◉ example

$$y = xb + e \qquad (1)$$

The transposed of the original descriptors matrix. A transposed matrix replaces columns with rows and vice versa.

The "-1" indicates matrix inversion

$$b = (X^T X)^{-1} X^T y \qquad (2)$$

The unknown vector of coefficients

The original descriptors matrix

The known vector of activities

F1.6.21 Analysis of the MLR Equation

One of the purposes of QSAR analyses is to understand the forces governing the activity of a particular class of compounds and to assist drug design. In the example shown below QSAR analyses reveal that the relative importance of the descriptors vary in the following order: logP > σ > Es > MR; therefore the biological activities are governed in the first place by hydrophobicity (logP) and polarity (σ) and to a lesser extent by steric effects (Es and MR).
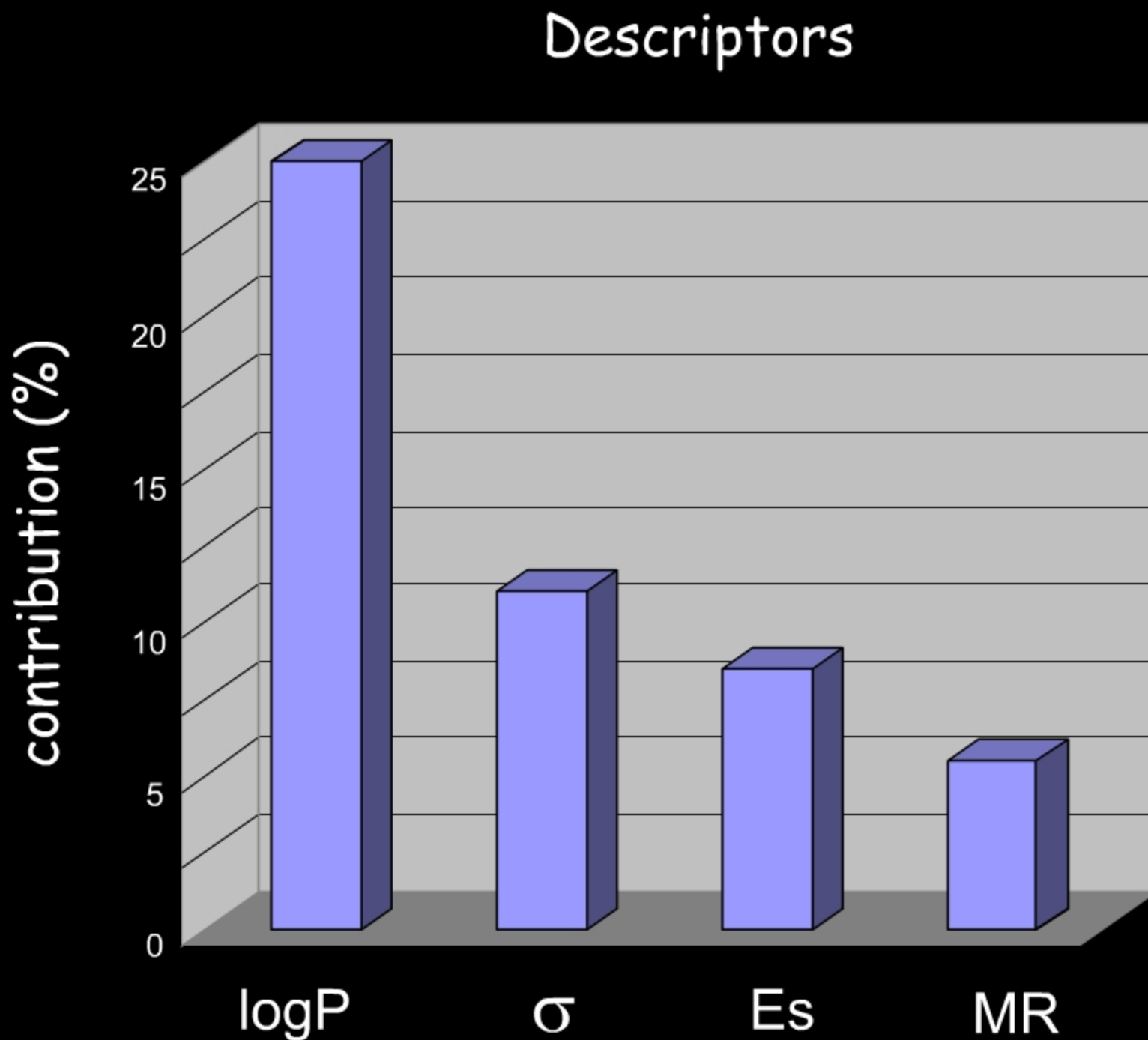
F1.6.22 Non-Linear Equations

A non-linear equation is an extension of a multiple linear regression. In some systems the linearity may not be sufficient to achieve a good correlation. Hansch was the first to introduce a parabolic term, and a complex biological process can be satisfactorily modeled by non-linear equations.
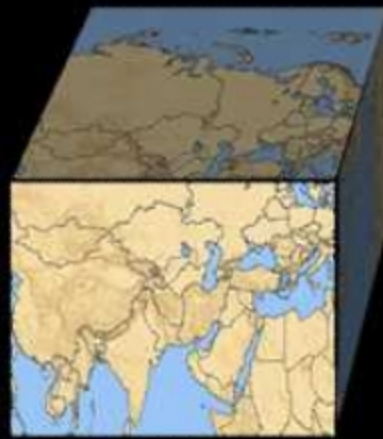

Planar   cubic   spherical   ellipsoidal

In the example below, the anticonvulsant activities of a set of molecules was initially found to be linearly correlated with logP. However, it is implausible to assume that the biological activity can increase indefinitely by increasing the lipophilicity of the molecules. It is known that highly lipophilic compounds cannot reach their site of action, because they are trapped in lipophilic environments. It is therefore more realistic to improve the initial model using a non-linear equation. The modified equation proved to be correct and revealed the existence of an optimum logP value, information that could not be derived from molecules with a small range of logP values.

linear model   non-linear model

$$\log (1/C) = 0.73 \, \log P + 2.5$$
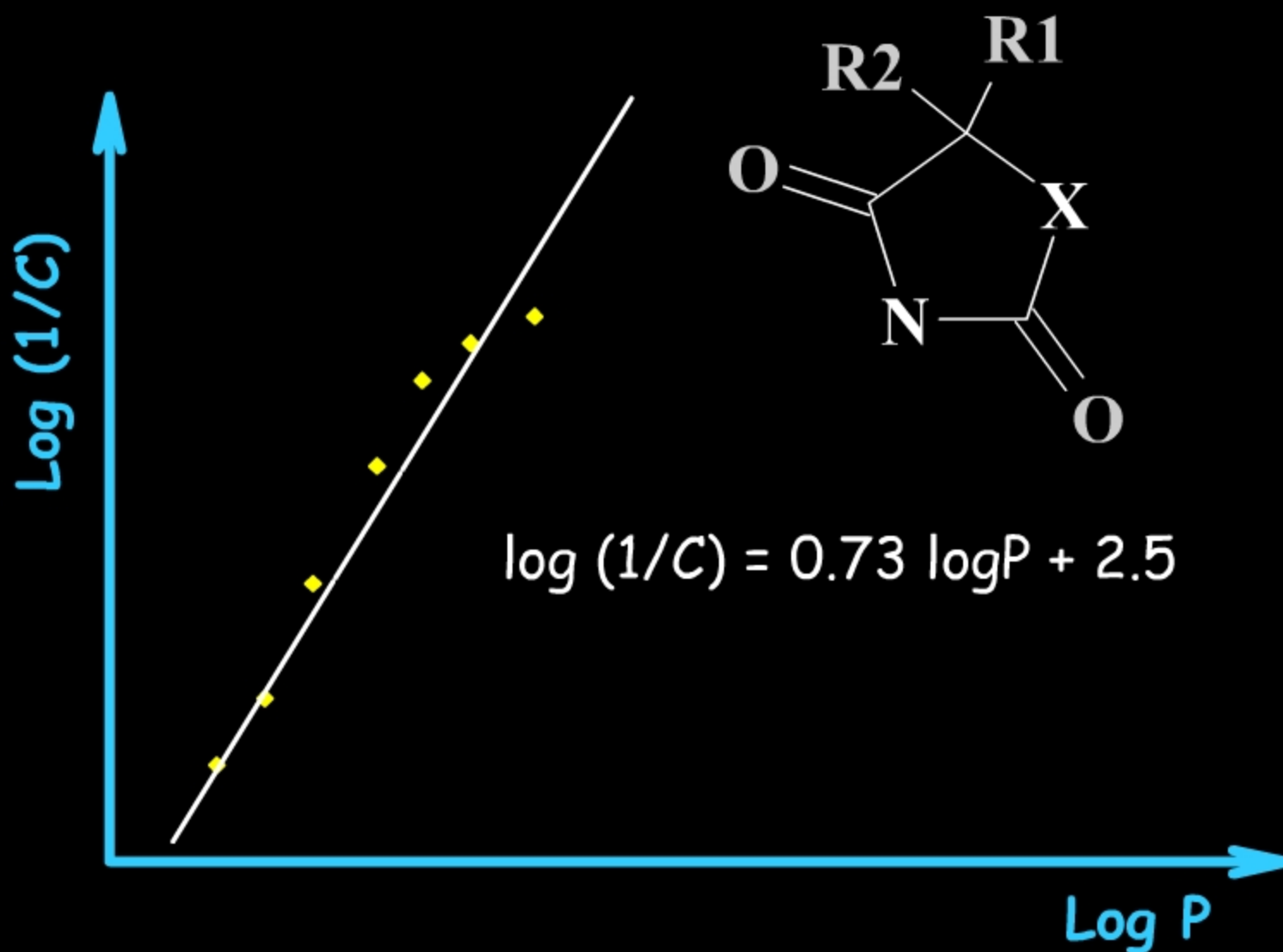
There are many reasons why the use of non-linear models is justified, including the kinetics of the drug transport, the equilibrium control of its distribution, allosteric effects, different pharmacokinetics, metabolism, solubility etc... The following are examples of non-linear models that have proved to be valid at least for special and complex biological systems.

Parabolic Model (Hansch)

$$\log 1/C = a\,(\log P)^2 + b\,\log P + c$$

Probability Model (McFarland)

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

Equilibrium Model (Hyde)

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

Bilinear Model (Kubinyi)

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

The topic Validating the Model: Step 4 contains the following 19 pages:

- Tools for Assessing the Quality of a Model
- Predictive and non-Predictive Models
- The Standard Deviation
- Correlation Index r^2
    - The Mathematics of r^2
    - TSS, the Total Variance
    - RSS, the Explained Variance
- t-test for Single Descriptors and Significance of r^2
    - Shape of t-distribution and Number of Molecules
    - Student's t-test Procedure
- F-test for Assessing the Significance of r^2
    - Performing the F-test
    - F-test Procedure

    ...

For the entire list, see the navigation panel.

F1.7.1 Tools for Assessing the Quality of a Model

Efficient tools are necessary for assessing the validity of a QSAR model. Numerical analyses or statistical methods provide a variety of indexes that serve to evaluate the quality of the model and its limitations. In the following pages we present some of these tools and explain how to use them.

Compounds selection

⬇

Descriptors selection

⬇

Building the QSAR model

⬇

Validating the model

$$\log 1/C = a (\log P)^2 + b \log P + c\, \sigma + d$$
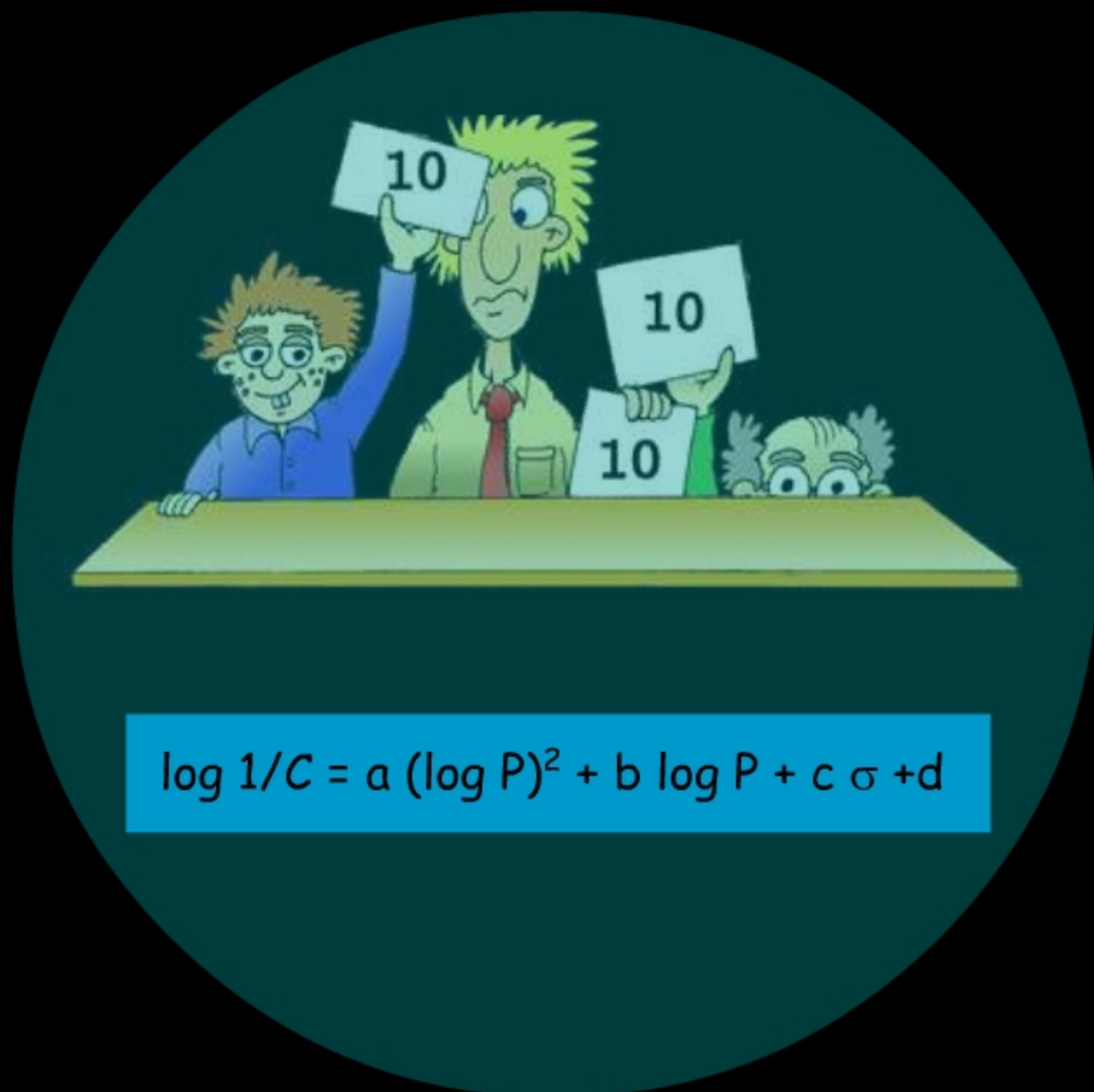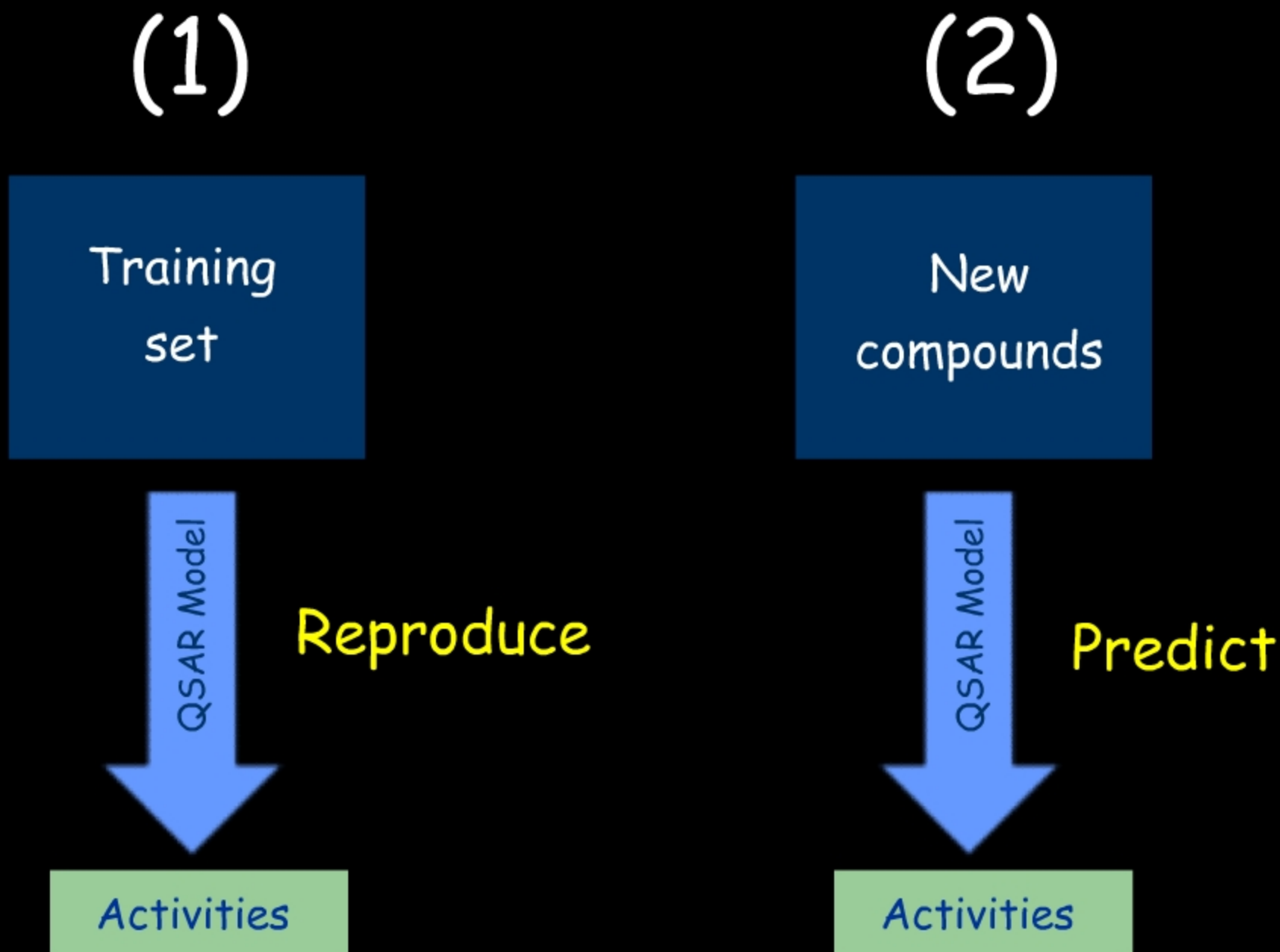
F1.7.2 Predictive and non-Predictive Models

Broadly speaking there are two groups of indices: (1) those that indicate how well the QSAR equation can "reproduce" the experimental data and (2) those that can tell how far the model can be extrapolated to new molecules.

F1.7.3 The Standard Deviation

The easiest way to "validate" a QSAR model is to calculate the standard error or standard deviation (SD or s), which is calculated as the average squared deviation of each number (the "residuals") from the mean. This index reflects how much the deviation between the data and the model is. The smaller the SD, the more the model is considered of good quality.

⦿ s calculation   ⦿ example

The Equation

F1.7.4 Correlation Index r²

The most frequently used index for evaluating the performance of a QSAR model is r² (squared correlation coefficient). r² measures the degree of correlation between the activity values calculated by the model and those measured experimentally. The value of r² can range between 0 (no correlation) to 1 (perfect correlation).

- r2=1
- r2=0.5
- r2=0

$$r^2 = 1$$

perfect correlation

Measured Activity (y-axis) vs Calculated Activity (x-axis)
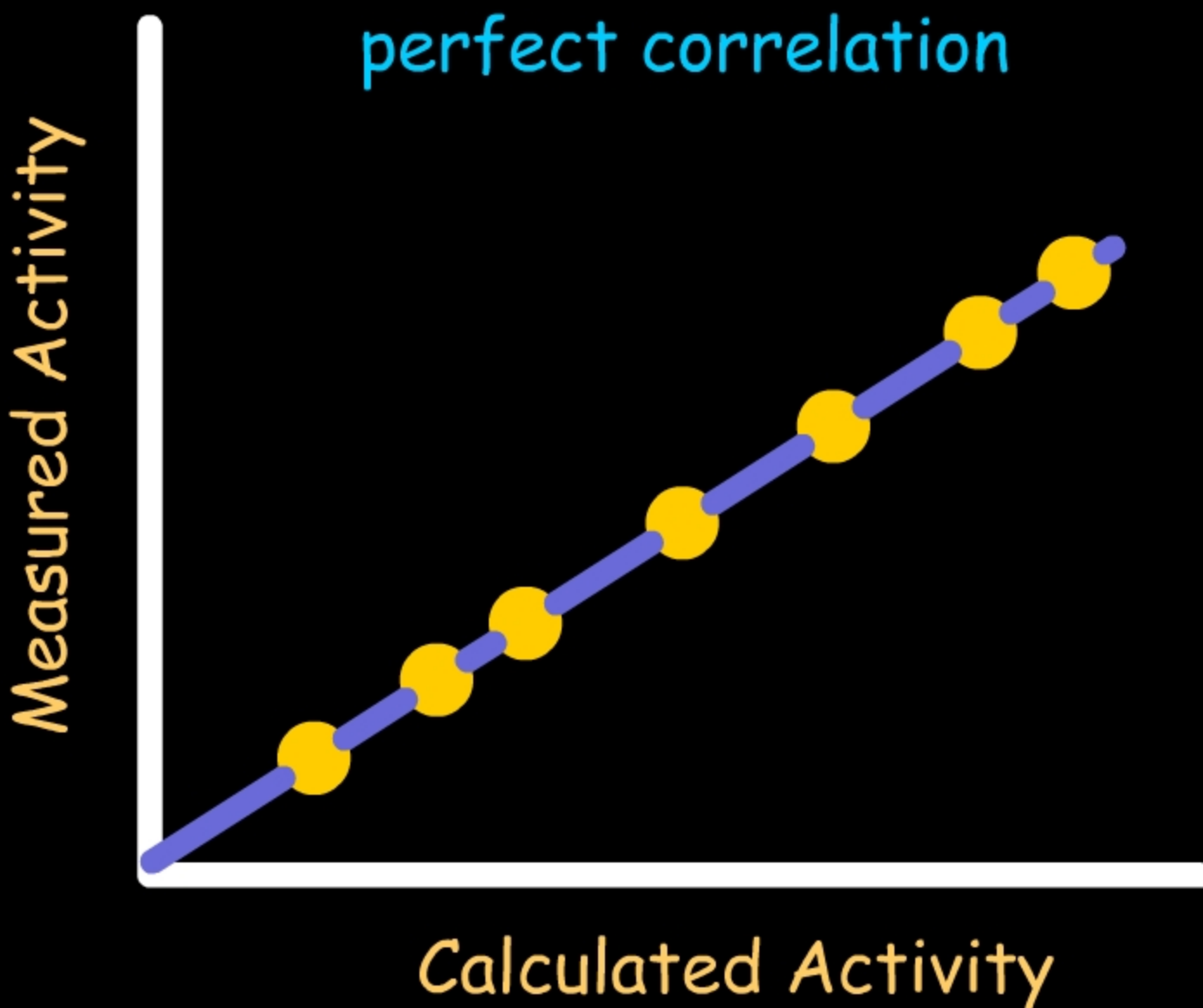
F1.7.5 The Mathematics of r²

Mathematically, r² is calculated by dividing the fraction of variance explained by the model (the "explained sum of squares", ESS) by the original variance (the "total sum of squares", TSS). ESS, the fraction of variance explained by the model is equal to the total variance (TSS) minus that portion of the variance which was not explained by the model (residual, RSS).

Original variance (Total Sum of Squares):

$$TSS = \sum_{i}^{N} (y_i - \bar{y})^2$$

TSS, the total variation in the dependent variable (y) is simply the spread of the data around the average.

Compound number
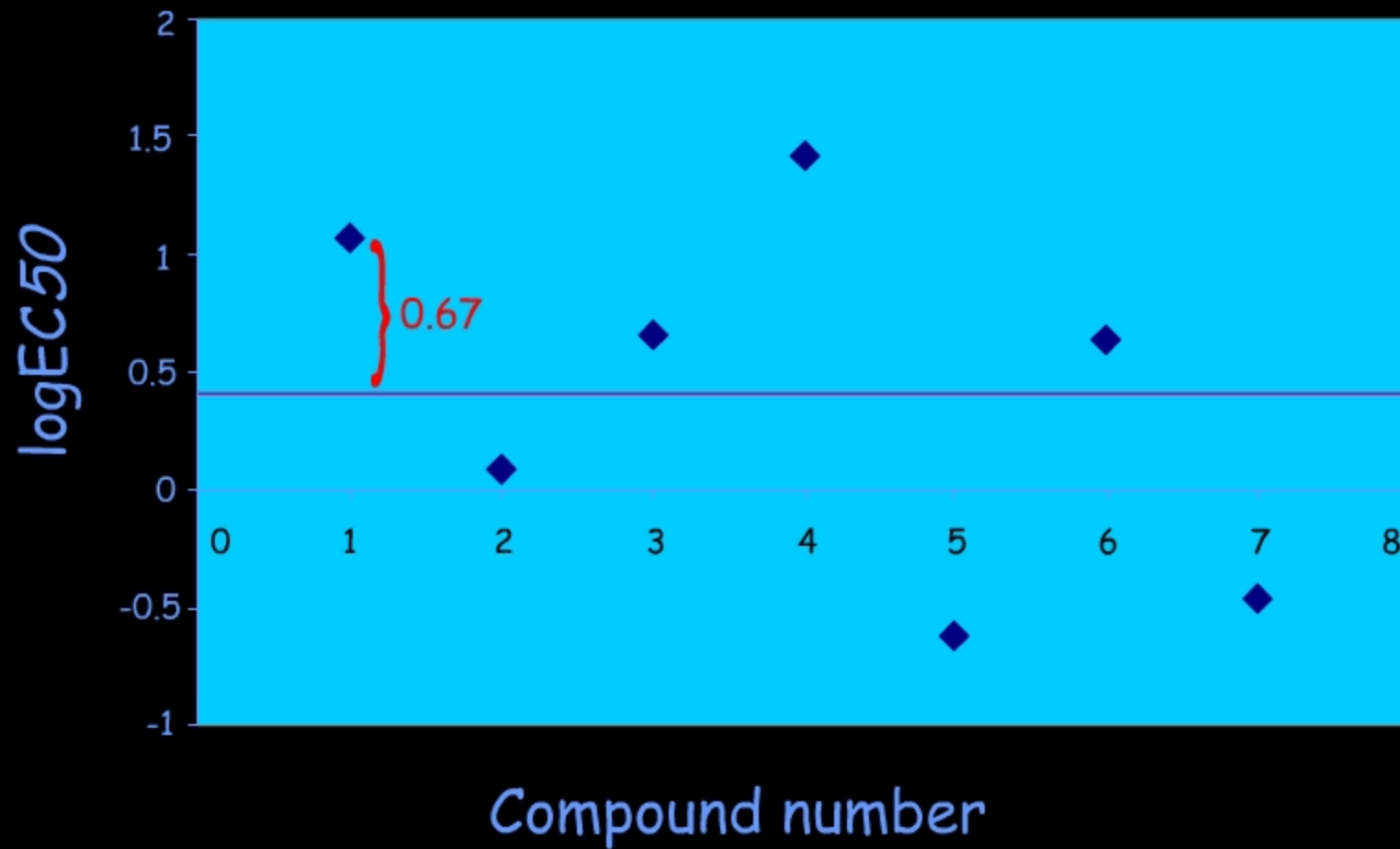
F1.7.7 RSS, the Explained Variance

In order to obtain RSS, the variance explained by the QSAR model, we start from the fact that the total variance is the sum of the explained and unexplained variances. Thus, the explained variance is the difference between the total variance and the unexplained variance. That portion of the variance which is left unexplained by the QSAR model (unexplained variance) can be obtained by finding the difference between the measured activity and the predicted activity (as given by the regression line).

$$\log EC_{50} = 0.794 - 0.817\pi$$
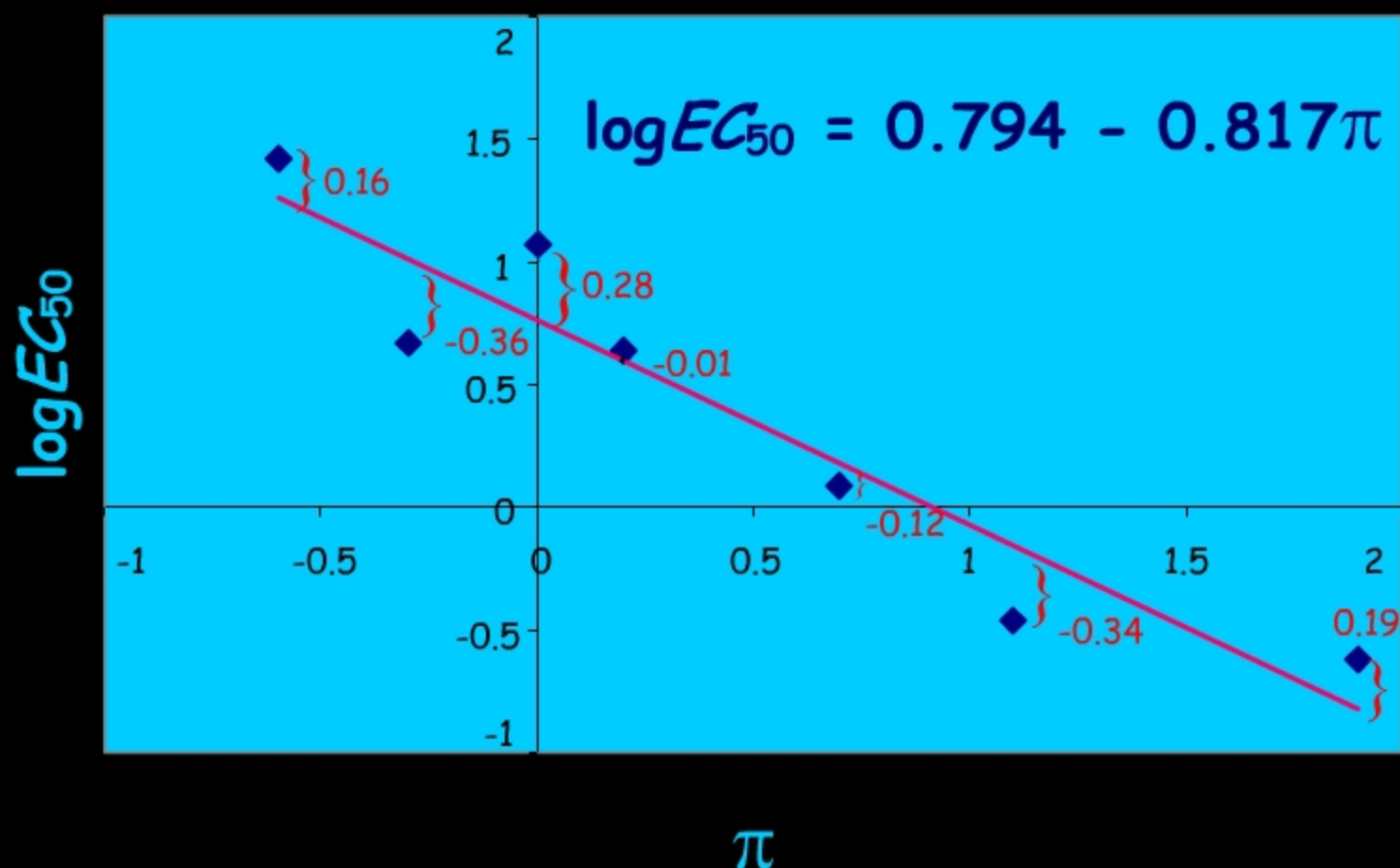
F1.7.8 t-test for Single Descriptors and Significance of r^2

r^2 alone is not sufficient to determine whether the relationship has occurred by chance; its significance can be calculated using the t-statistic for single descriptors as follows. We repeat the process of deriving of a QSAR equation and calculate the resulting r^2 values many times, each one using a different descriptor. If the number of molecules is large (> 30), the sampling distribution of the resulting r^2 values will have a normal (i.e., Gaussian) shape. If the number of molecules is small, it will have a shape known as a t-distribution.
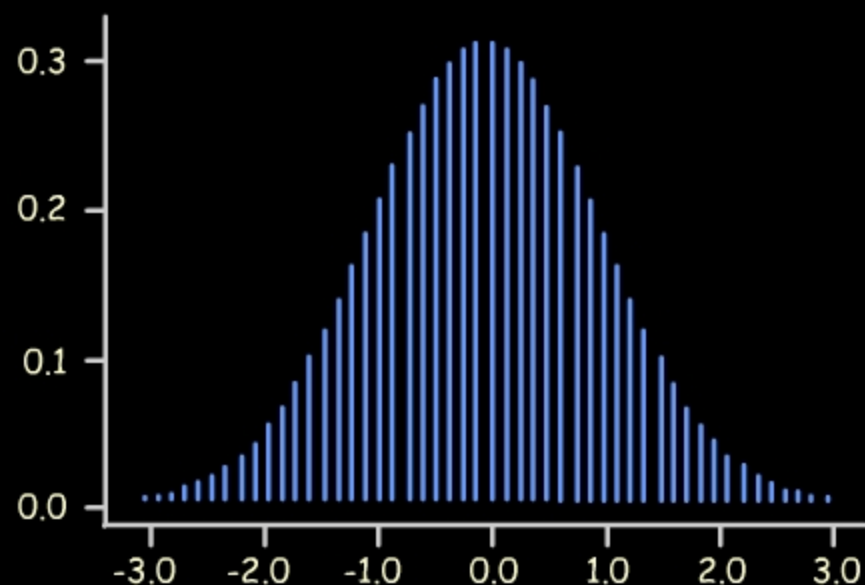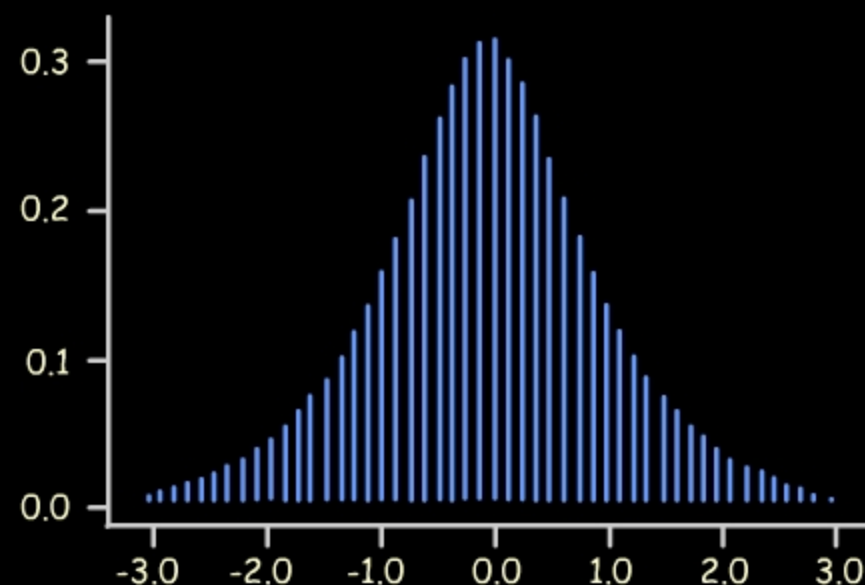
Values on the x-axis represent standard deviations from the mean located at X = 0.

F1.7.9 Shape of t-distribution and Number of Molecules

A value $r^2 = 1$ will always be obtained for a set of two molecules irrespective of the descriptor used for the QSAR analysis however, as the number of molecules increases, the probability of obtaining large r^2 values with irrelevant descriptors decreases. This probability corresponds to the area under the t-distribution curve (see below), away from the center (where $r^2 = 0$). The shape of the t-distribution therefore depends on the number of molecules used in the analysis.

t-distribution for 3 molecules
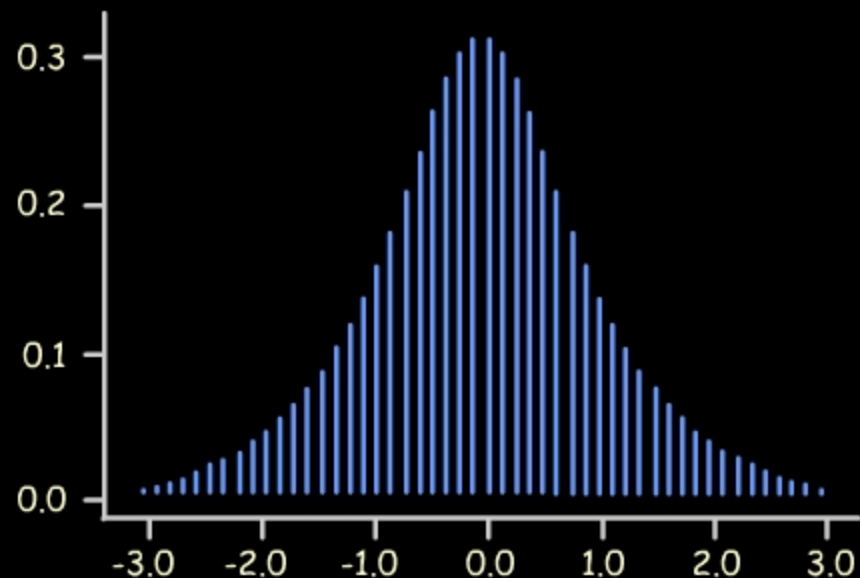
t-distribution for 30 molecules
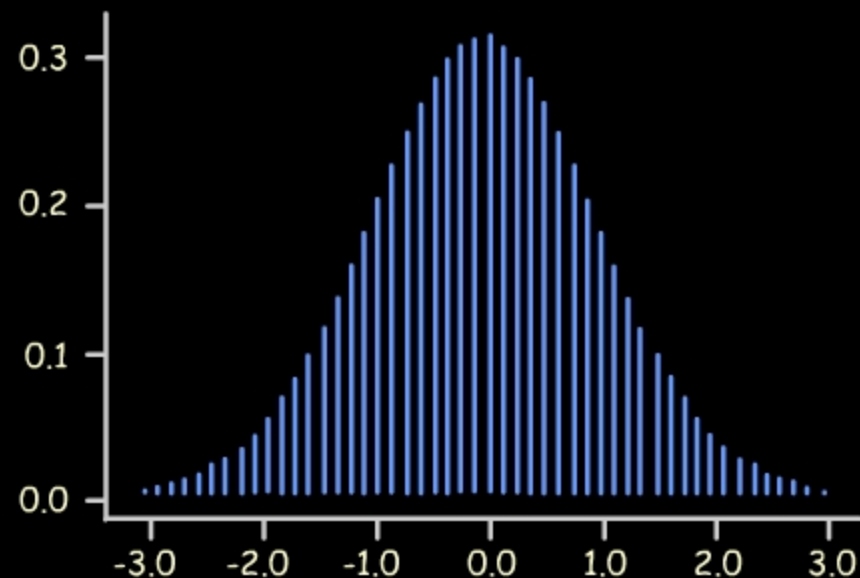
F1.7.10 Student's t-test Procedure

The Student t-test employs the t-distribution to test whether the correlation coefficient obtained from the QSAR analysis is significantly different from 0. The larger the t-value, the larger the probability that r^2 significantly differs from 0; that is, the larger the probability that the descriptor used for the analysis is relevant to the activity. Technically, the steps involved in the Student t-test are as follows.

1. Calculate t according to the above equation. $t = r\sqrt{\dfrac{N-2}{1-r^2}}$

2. Select a significance level (e.g., 0.05). (see step 2)

3. Look up the t value from a t-distribution derived for the correct number of data points (N) at the selected significance level.

4. If the calculated t-value is larger than the listed t-value, then the regression equation is significant at this significance level.

F1.7.11 F-test for Assessing the Significance of r²

The F-test is an extension of the t-test for the case of many descriptors. Like the t-test it tests (and hopefully rejects) the assumption that the model did not explain any of the original variance in the data set (i.e., ESS = 0). Like the t-test, the F-test uses an F-distribution which, similar to the t-distribution depends on the number of compounds and descriptors.

Molecules = 10
Descriptors = 4
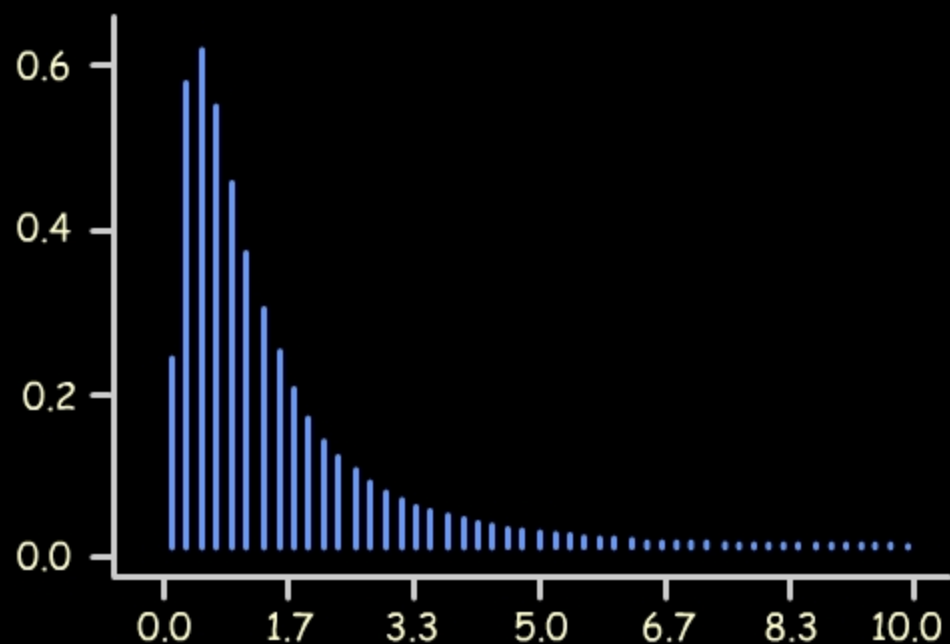
Molecules = 100
Descriptors = 10
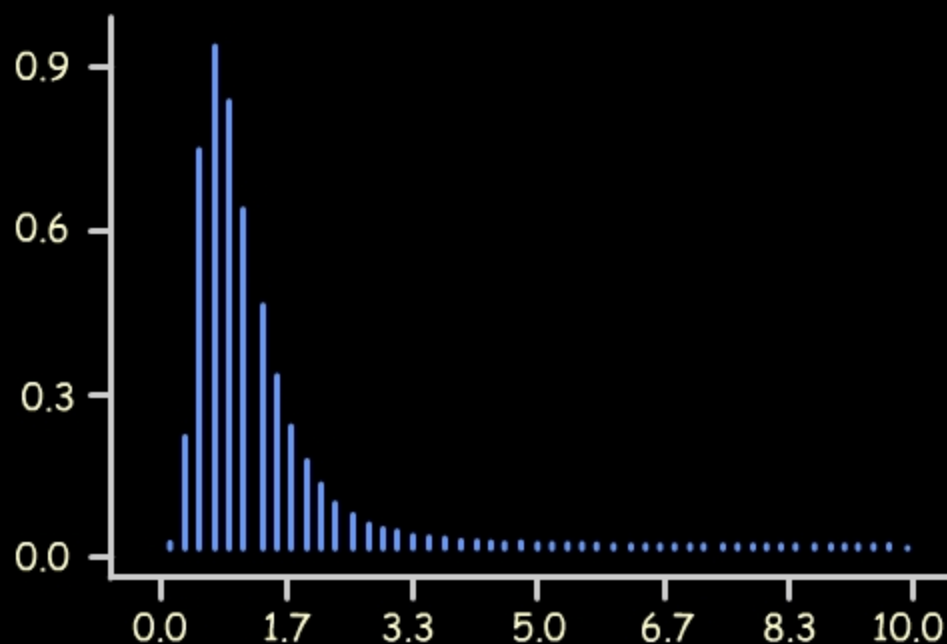
F1.7.12 Performing the F-test

The F-test employs the F-distribution to test whether the correlation coefficient obtained from the MLR analysis significantly differs from 0. The larger the F-value, the larger the probability that r^2 significantly differs from 0; i.e. the greater the probability that the descriptor used for the analysis is relevant to the activity. Technically, the steps involved in the F-test are as follows.

$$r^2 = \frac{ESS}{TSS} \quad \text{and} \quad r^2 = 1 - \frac{RSS}{TSS} \qquad F = \frac{ESS}{k} \; \frac{N-k-1}{RSS}$$

$$TSS = \frac{ESS}{r^2} \qquad \frac{1}{1-r^2} = \frac{TSS}{RSS}$$

$$\frac{ESS}{RSS} = \frac{r^2}{1-r^2}$$

Calculate F according to this equation: $$F = \frac{r^2(N-k-1)}{k(1-r^2)}$$

N - number of molecules, k - number of descriptors
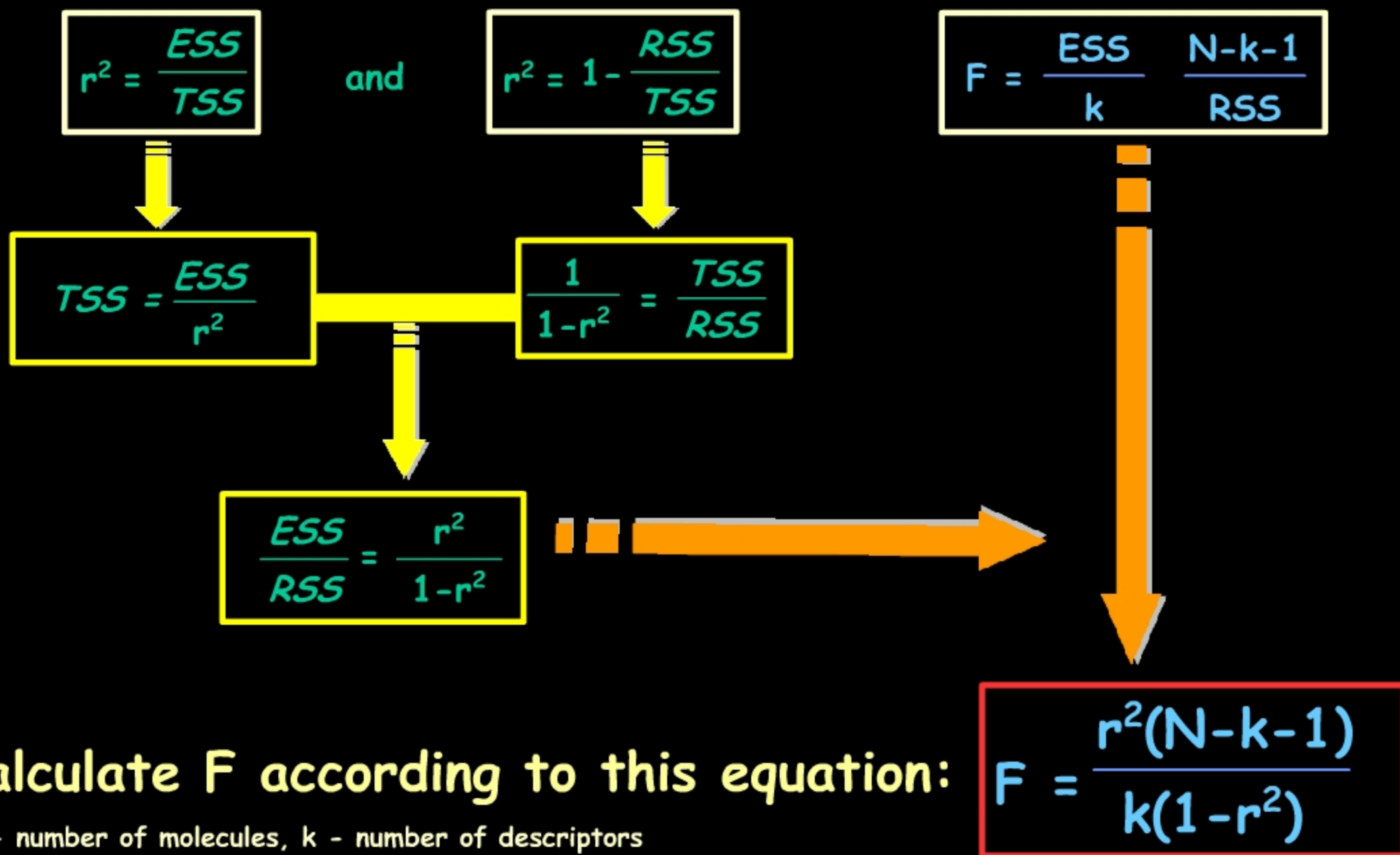
F1.7.13 F-test Procedure

The application of the steps involved in evaluating the significance of r^2 for the Capsaicin analogs using the F-test proceeds as follows:

● Calculate F: $F = \dfrac{r^2(N-k-1)}{k(1-r^2)}$; $r^2 = 0.92$; N=8; k=3

$$F = \frac{0.92(8-3-1)}{3(1-0.92)} = 15.33$$

● Select a significance level (p):   p = 0.01

● Look up the F value from an F-distribution with N=7, k = 1, p = 0.01: $F_{tab} = 7.59$

● The calculated F value (15.33) is larger than the tabulated F value (7.59). Thus, the correlation is significant at this level. The probability that the correlation is fortuitous is less than 1%.

F1.7.14 Assessing the Predictive Power of a Model

r^2, t and F are indices that can be generated to evaluate QSAR results. However, these parameters basically only tell us about the ability of the QSAR model to reproduce the data from which it was derived and not its aptitude to predict the activities of new compounds. Two methods are presented in the following pages to estimate the predictive power of a QSAR model.
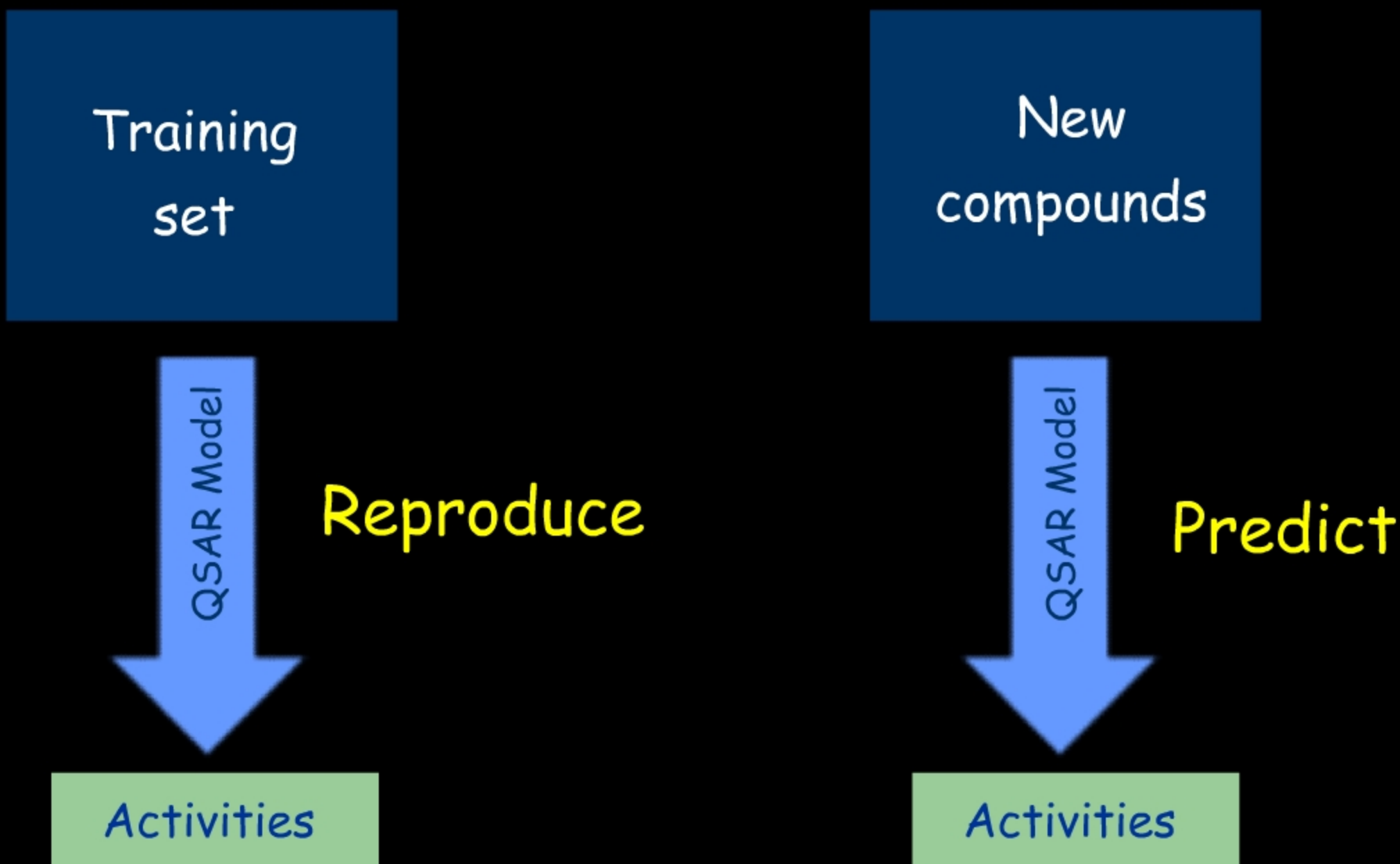
F1.7.15 The Test Set Method

The first method is known as the "test set method" and consists of partitioning the initial data into two sets, a preferred strategy when a large set of compounds is available. The initial data set is randomly divided into two parts; the first one is used to build a QSAR model and the second one to validate this model.
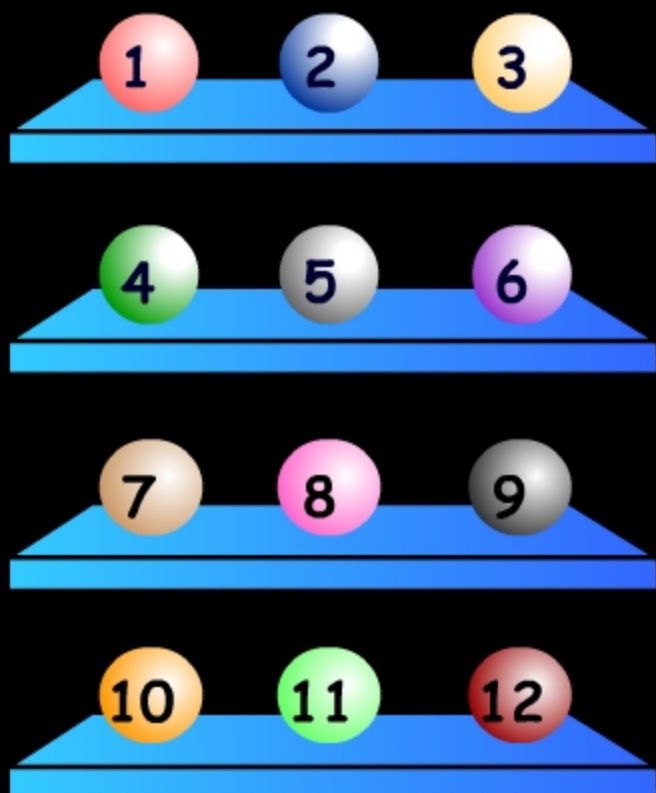
Training set   Test set

F1.7.16 The Cross Validation Method

The second method is known as "the cross validation method" - it is preferred when the size of the data set is too small. In this method the data are randomly divided into N e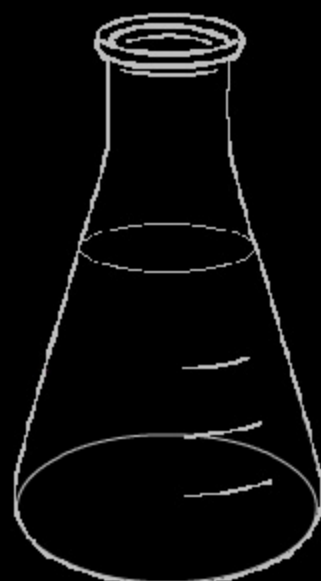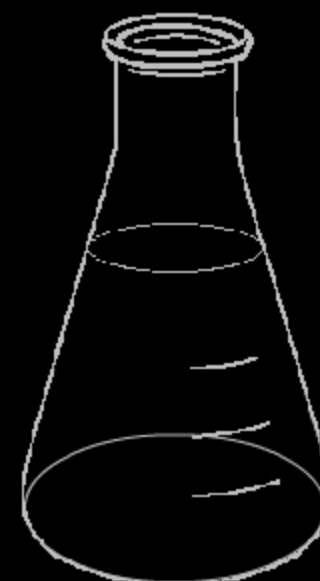qual parts; N-1 parts are used to build the model which is then used for the remaining N^{th} part to predict the activities of the corresponding molecules. The procedure is repeated until the activities of all compounds have been predicted independently.

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
compounds 1-10	Training set	Training set	Training set	Training set	Test set
compounds 11-20	Training set	Training set	Training set	Test set	Training set
compounds 21-30	Training set	Training set	Test set	Training set	Training set
compounds 31-40	Training set	Test set	Training set	Training set	Training set
compounds 41-50	Test set	Training set	Training set	Training set	Training set

F1.7.17 Limits of the Cross Validation Method

With the cross validation method, the QSAR model that is ultimately used to predict the activities of new compounds is derived from all N data points and is therefore different from the N partial QSAR models (i.e. those derived from the N-1 data points). Therefore cross validation does not provide us with the predictive power of a specific QSAR equation but rather with an estimate of our ability to make predictions for compounds similar to those used in our QSAR analysis.

Compounds used in original QSAR analysis

Prediction estimates made by cross validation apply

Prediction estimates made by cross validation do not apply

Descriptor 2

Descriptor 1
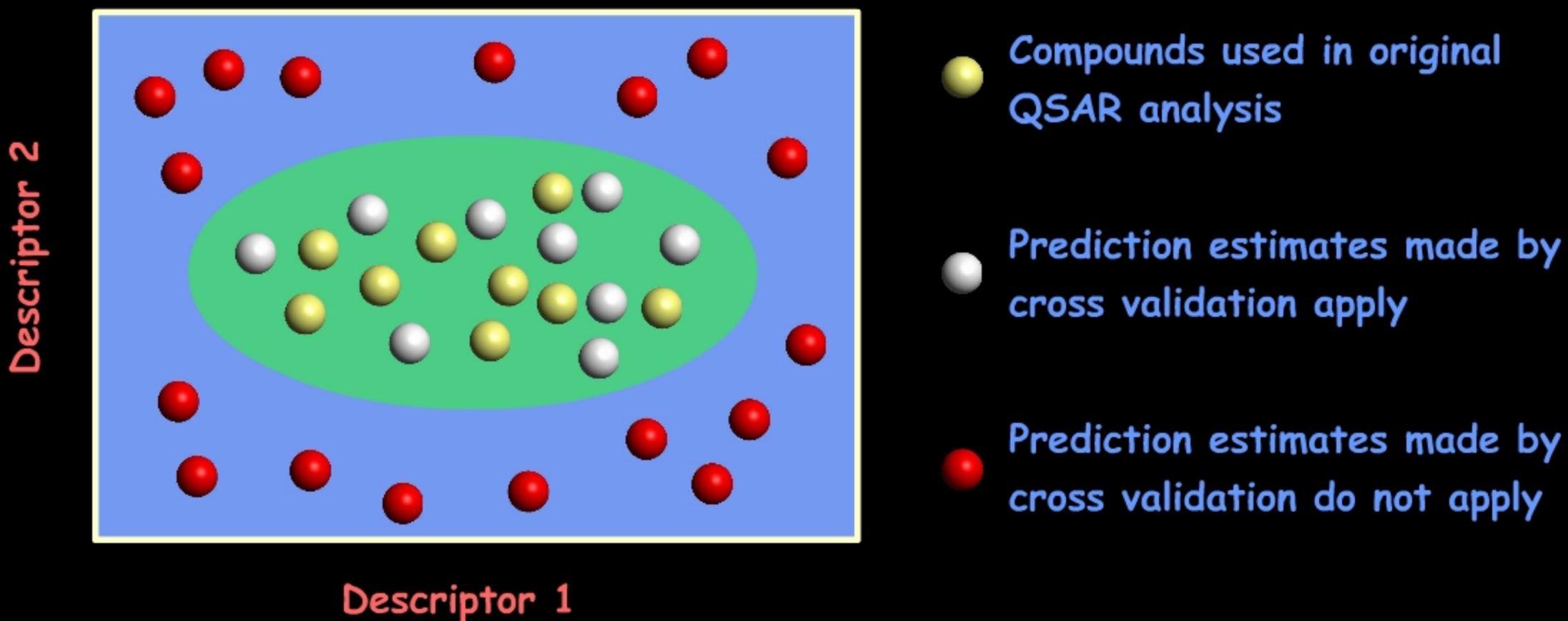
F1.7.18 The Predictive Index Q²

The predictive power of the model, termed Q², is computed by analogy with r², the difference being the use of the PRESS (predicted sum of squares) rather than the RSS (residual sum of squares) in the numerator. PRESS is calculated as the difference between the measured activity and the predicted activity for the test set compounds.

$$r^2 = 1 - \frac{RSS}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

$$RSS = \sum_{i=1}^{N} (y_{calc,i} - y_i)^2$$

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

$$PRESS = \sum_{i=1}^{N} (y_{pred,i} - y_i)^2$$

F1.7.19 Summary

When discussing mathematical tools available for assessing the quality of a QSAR model we saw that (1) the standard deviation is an isolated "absolute" index of local meaning; (2) with r^2 it is possible to compare different models, but this index is only mathematical - not statistical; (3) t and F have a statistical content that can be used for single and multiple linear regression respectively; however they only measure the ability of the QSAR model to reproduce the data from which it was constructed.

$$\log \frac{1}{C} = 1.14 \log P + 0.16$$

correlation coefficient for
assessing the quality of the model

F-value for assessing
the statistical significance

$$n = 25; \quad r^2 = 0.91; \quad s = 0.155; \quad F = 66.4; \quad Q^2 = 0.875$$

number of molecules

standard deviation

regression coefficient for
measuring the predictibility

The topic Example of Simple Linear Regression contains the following 11 pages:

- Example of Capsaicin Analogs
- Relevant Descriptors of Capsaicin Analogs
- The Capsaicin Study Table
- Graphical Analysis of Capsaicin Analogs
- Deriving a QSAR Linear Equation
- Experimental vs. Calculated Values
- Calculating r^2 for the Capsaicin analogs
- t-test for the Capsaicin Analogs
- F-test for a Series of the Capsaicin Analogs
- The QSAR Equation for the Capsaicin Analogs
- Predicting the Activities of Unknown Compounds

F1.8.1 Example of Capsaicin Analogs

Capsaicin analogs were studied for their analgesic properties and we will use this study to illustrate the derivation of a simple QSAR model. The biological activities (EC_{50}) were measured for some analogs as indicated below. The question is whether on the basis of these data, it is possible to develop a QSAR model and predict the biological activities of new compounds.
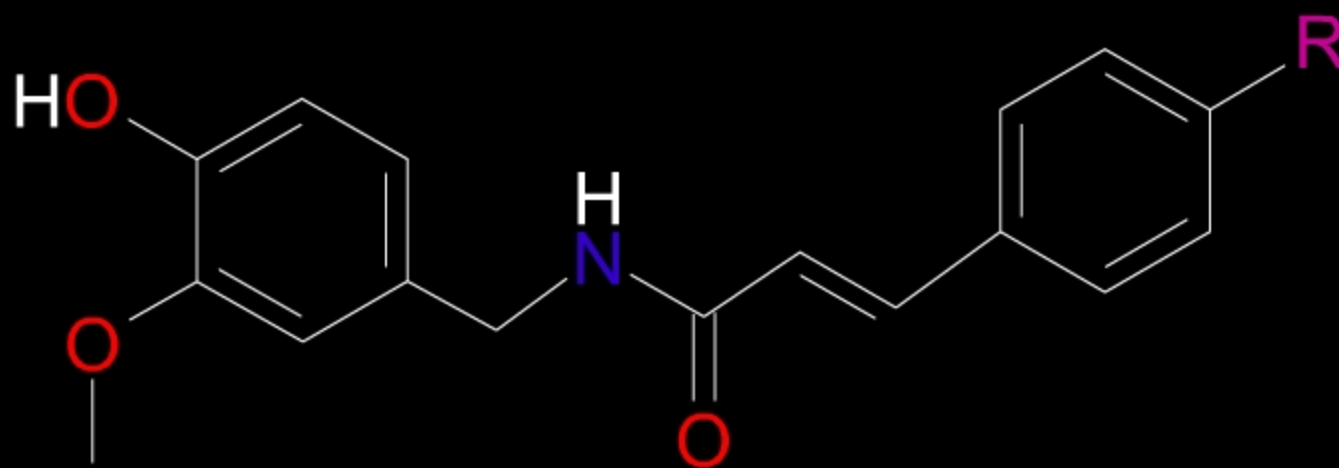
Compound	R	EC_{50} (mM)
1	H	11.80
2	Cl	1.24
3	NO_2	4.58
4	CN	26.50
5	C_6H_5	0.24
6	$N(CH_3)_2$	4.39
7	I	0.35

F1.8.2 Relevant Descriptors of Capsaicin Analogs

The selection of descriptors that correlate with the target biological activity is mandatory for the derivation of a meaningful QSAR model. For Capsaicin analogs, biological activity appears to be influenced by the lipophilicity of the substituent R. Following this assumption the descriptors deemed most suitable are the molar refractivity (MR) and the hydrophobic substituent constant π.

R
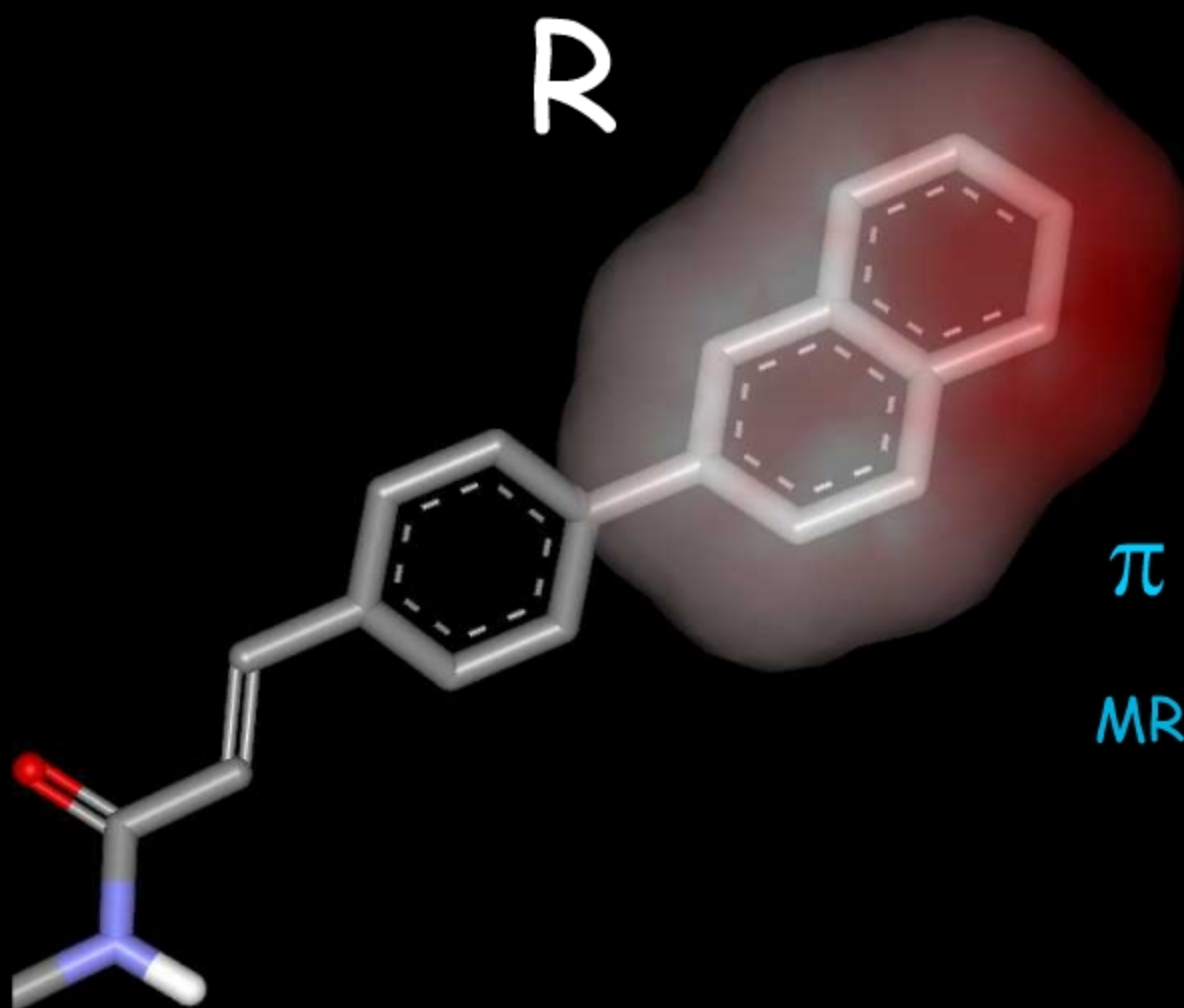
Lipophilicity Descriptors

π : encodes the lipophilic behavior

MR: contains information on the volume

F1.8.3 The Capsaicin Study Table

The following table summarizes the MR and π values which were calculated for the seven Capsaicin analogs. As discussed above, activities (EC_{50}) are expressed as their log values.

Compound	log EC_{50}	π	MR
1	1.07	0	1.03
2	0.09	0.71	6.03
3	0.66	-0.28	7.36
4	1.42	-0.57	6.33
5	-0.62	1.96	25.36
6	0.64	0.18	15.55
7	-0.46	1.12	13.94

F1.8.4 Graphical Analysis of Capsaicin Analogs

For Capsaicin analogs, if we plot the values from the study table for MR and π, respectively, there seems to be a weak correlation between the biological activity and the molar refractivity (MR). However, the hydrophobic substituent constant π shows a possible linear correlation.
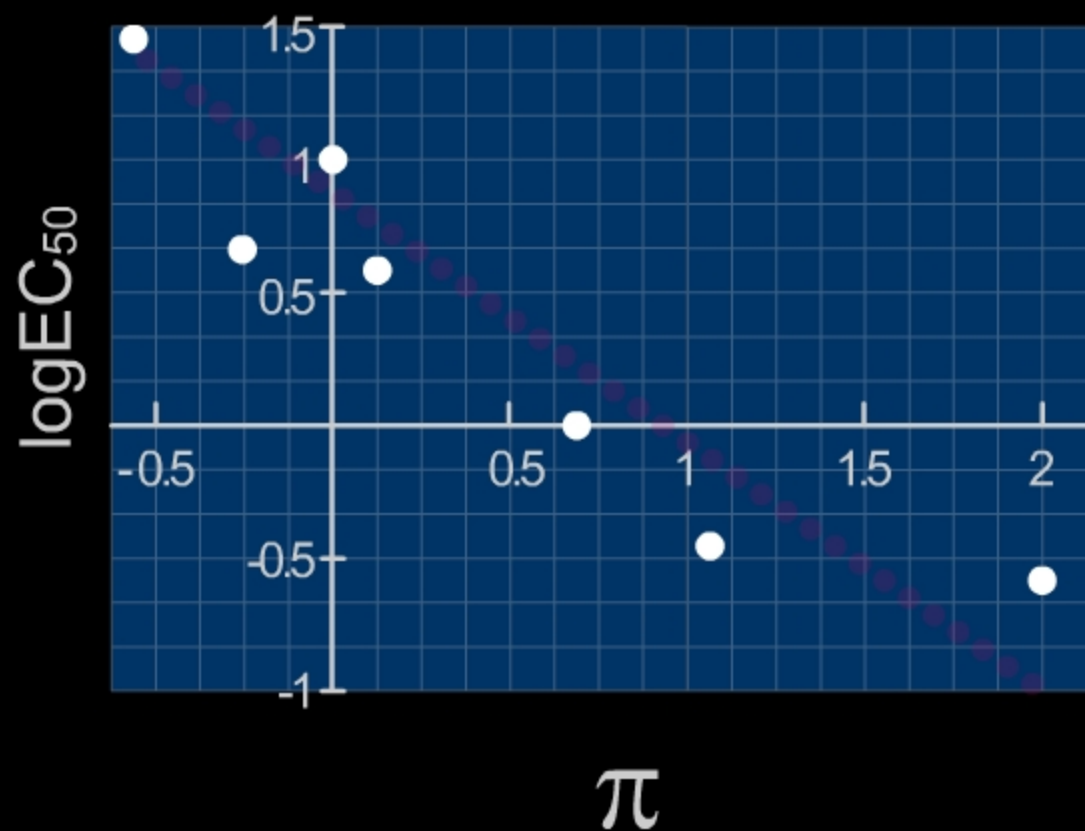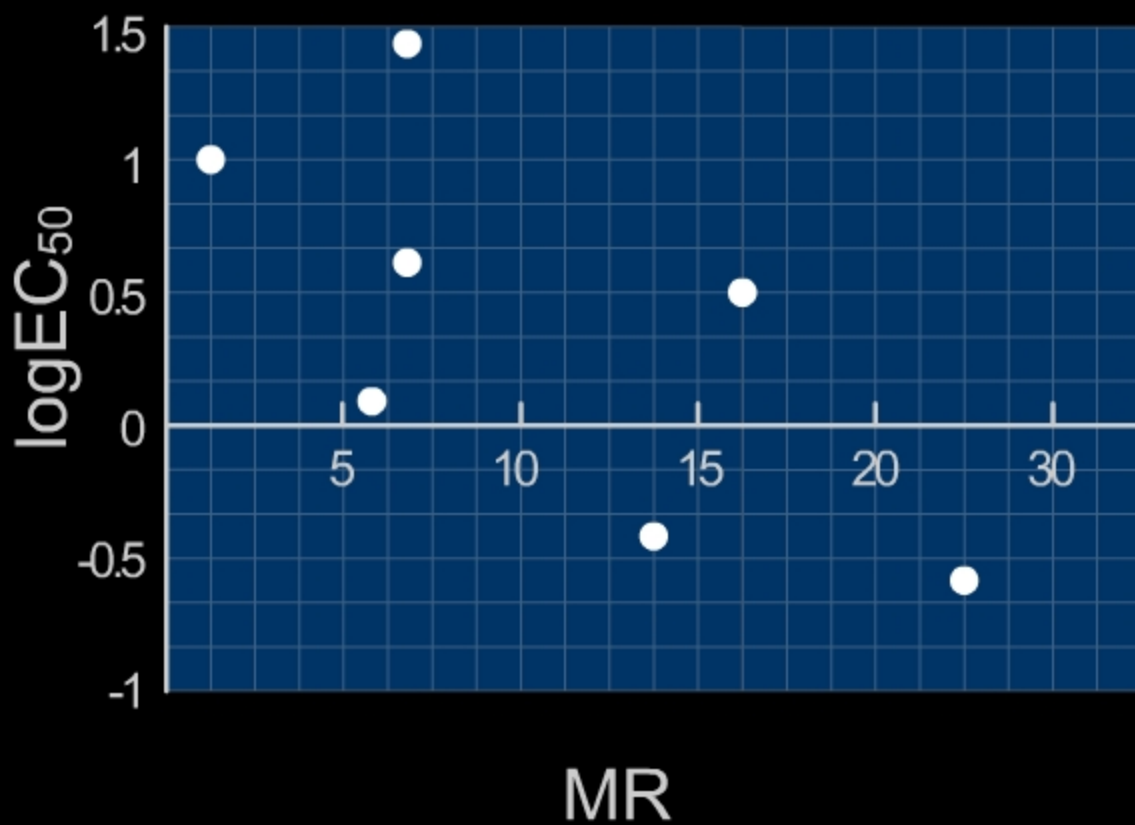
F1.8.5 Deriving a QSAR Linear Equation

The correlation between π and the biological activities is represented by the equation $y = b_0 + b_1X$, where b_0 is the intercept of the line with the y axis and b_1 the slope of the line. We show below how to calculate their numerical values.
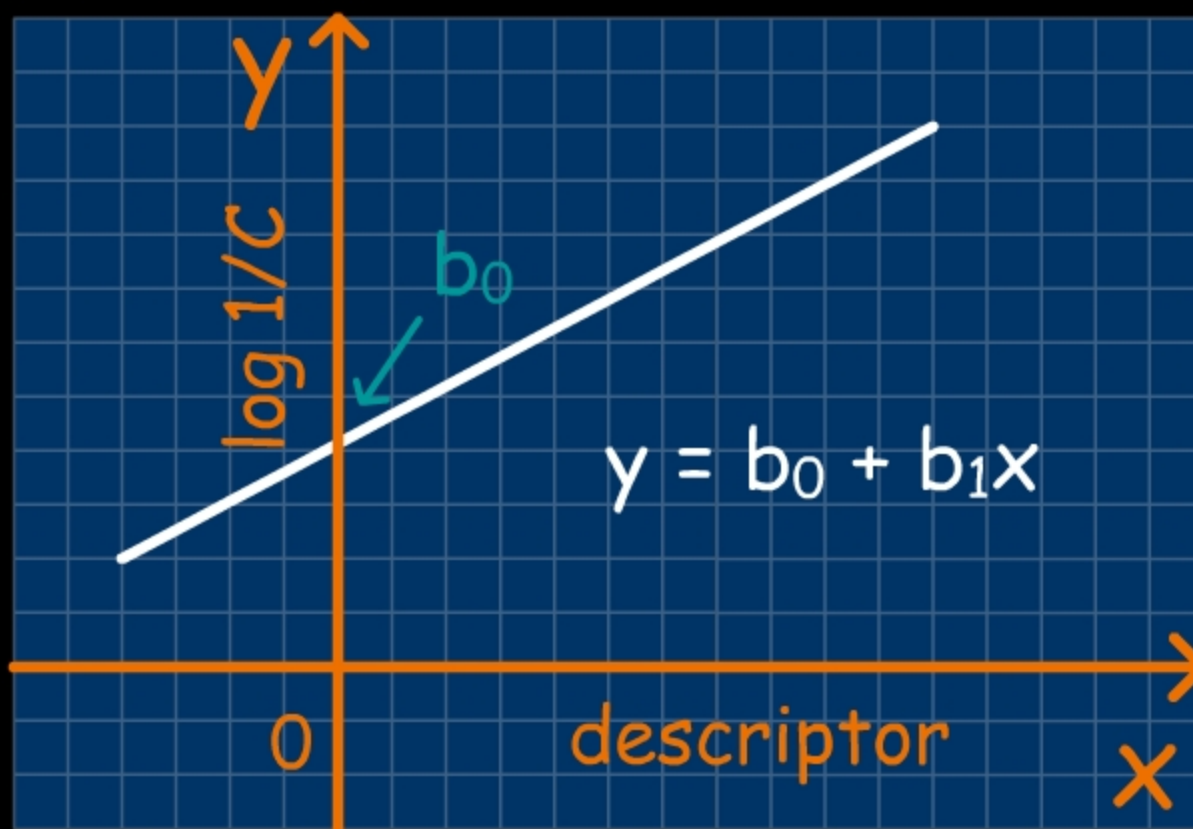
Reset   Continue

$$b_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad\qquad b_0 = \bar{y} - b_1\bar{x}$$

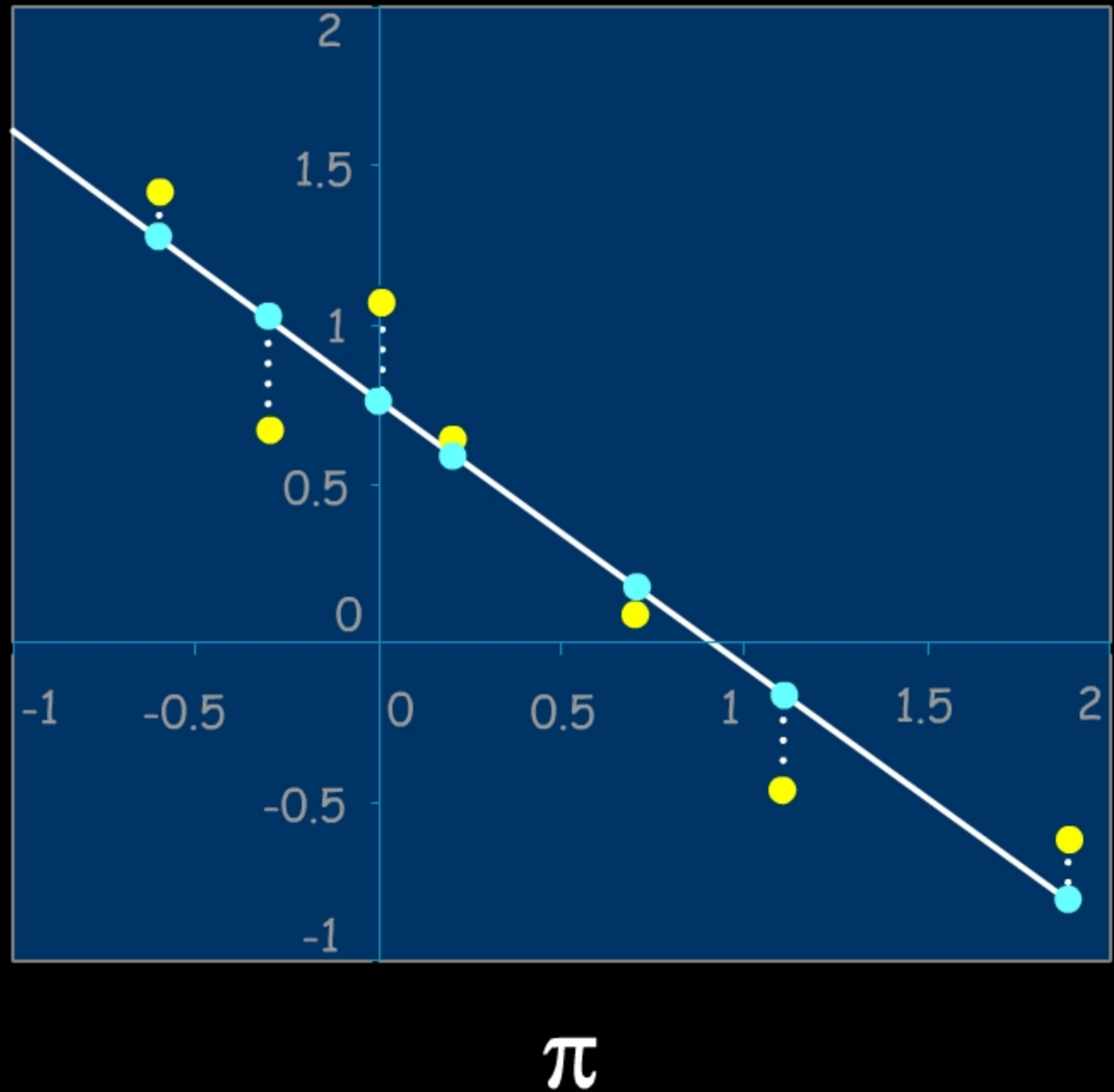| #	LogEC$_{50}$	π
1	1.07	0
2	0.09	0.71
3	0.66	-0.28
4	1.42	-0.57
5	-0.62	1.96
6	0.64	0.18
7	-0.46	1.12

F1.8.6 Experimental vs. Calculated Values

There is a difference between the experimental and the calculated values as shown below.

#	log EC$_{50}$ obs.	log EC$_{50}$ calc.
1	1.07	0.79
2	0.09	0.21
3	0.66	1.02
4	1.42	1.26
5	-0.62	-0.81
6	0.64	0.65
7	-0.46	-0.12

● Experimental

● Calculated

F1.8.7 Calculating r² for the Capsaicin analogs

For Capsaicin analogs, r^2 is calculated as follows.

#	log EC$_{50}$ obs.	log EC$_{50}$ calc.	Residual
1	1.07	0.79	0.28
2	0.09	0.21	-0.12
3	0.66	1.02	-0.36
4	1.42	1.26	0.16
5	-0.62	-0.81	0.19
6	0.64	0.65	-0.01
7	-0.46	-0.12	-0.34

$$TSS = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \quad RSS = \sum_{i=1}^{N} (Y_i - Y_{calc,\ i})^2$$

$$r^2 = \frac{TSS - RSS}{TSS}$$

$$\bar{Y} = \frac{1.07 + 0.09 + 0.66 + 1.42 + (-0.62) + 0.64 + (-0.46)}{7} = 0.4$$

$$TSS = (1.07-0.4)^2 + (0.09-0.4)^2 + (0.66-0.4)^2 + (1.42-0.4)^2 + (-0.62-0.4)^2$$
$$+ (0.64-0.4)^2 + (-0.46-0.4)^2 = 3.49$$

$$RSS = (0.28)^2 + (-0.12)^2 + (-0.36)^2 + (0.16)^2 + (0.19)^2 + (-0.01)^2 + (-0.34)^2 = 0.40$$

$$r^2 = \frac{3.49 - 0.40}{3.49} = \frac{3.09}{3.49} = 0.89$$

F1.8.8 t-test for the Capsaicin Analogs

The steps involved in evaluating the significance of r^2 are as follows:

○ t calculation   ○ t-table

●   **Calculate _t_:** $t = r\sqrt{\dfrac{N-2}{1-r^2}}$; $r^2 = 0.89$; $N = 7$

$$t = \sqrt{0.89}\,\sqrt{\dfrac{7-2}{1-0.89}} = \boxed{6.3604}$$

●   **Select a significance level (p).** $\boxed{p = 0.01}$

●   **Look up the t value from a t-distribution with _N_=7, p=0.01:** $\boxed{t = 2.998}$

●   **The calculated t value (6.3604) is larger than the tabulated _t_ value (2.998). Thus, the correlation is significant at this level. The probability that the correlation is fortuitous is less than 1%.**

F1.8.9 F-test for a Series of the Capsaicin Analogs

The steps involved for evaluating the significance of r^2 using the F-test proceed as indicated below. The F-test analyses finally indicate that a significant correlation is obtained and the probability of a chance correlation is less than 1%.

○ F calculation   ○ F-table

- Calculate F: $F = \dfrac{r^2(N-k-1)}{k(1-r^2)}$; $r2 = 0.89$; $N=7$; $k=1$

$$F = \frac{0.89(7-1-1)}{1(1-0.89)} = 40.45$$

- Select a significance level (p): $p = 0.01$

- Look up the F value from an F-distribution with $N=7$, $k = 1$, $p = 0.01$: $F = 12.25$

- The calculated F value (40.45) is larger than the tabulated F value (12.25). Thus, the correlation is significant at this level. The probability that the correlation is fortuitous is less than 1%.

QSAR studies reveal the importance of lipophilicity in the analgesic properties of a series of Capsaicin analogs as indicated by the good correlation found with the π descriptor. The correlation coefficient r^2 is 0.89 and analyses of the significance of the equation (t-test and F-test) show that there is less than a 5% chance that the relationship is due to chance. This validates the use of π as a descriptor for the structure-activity relationships.
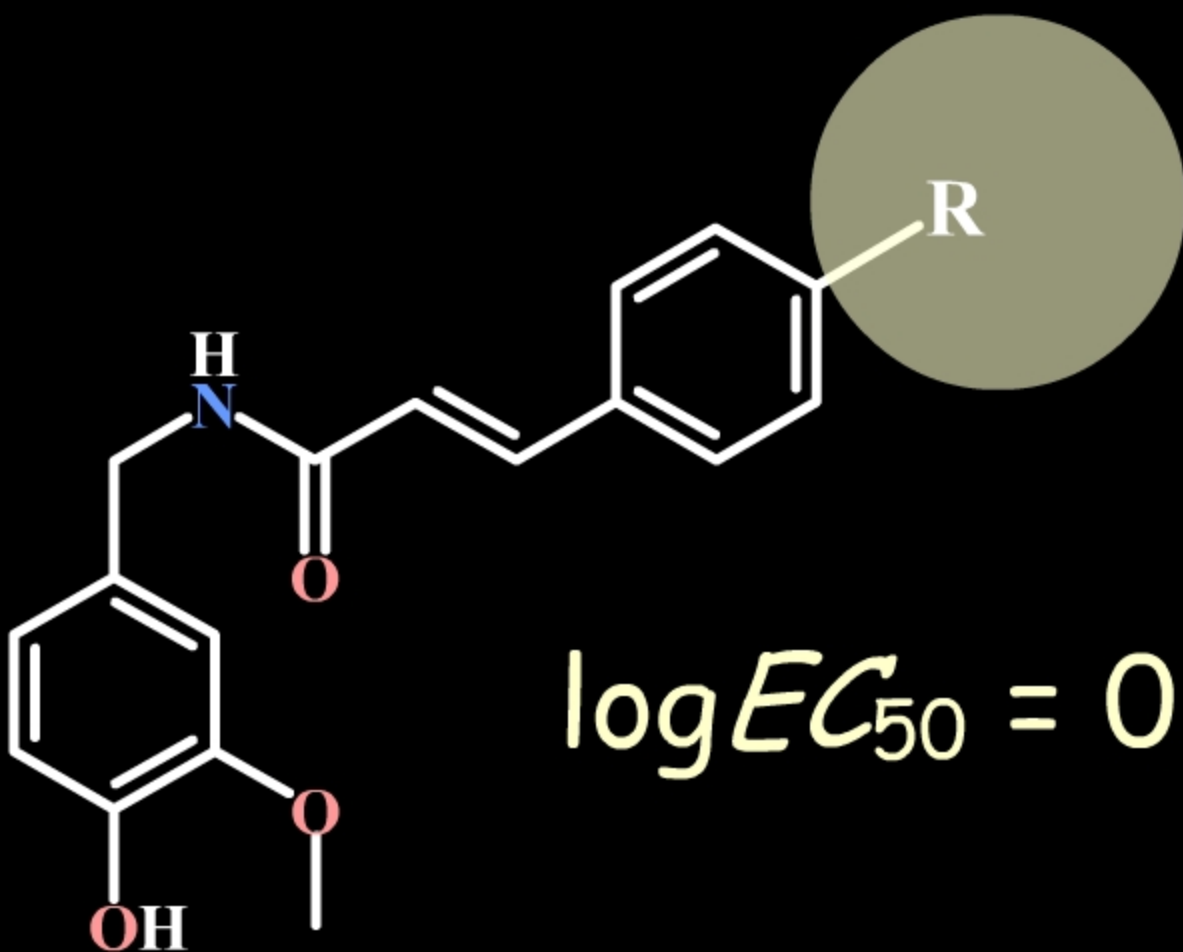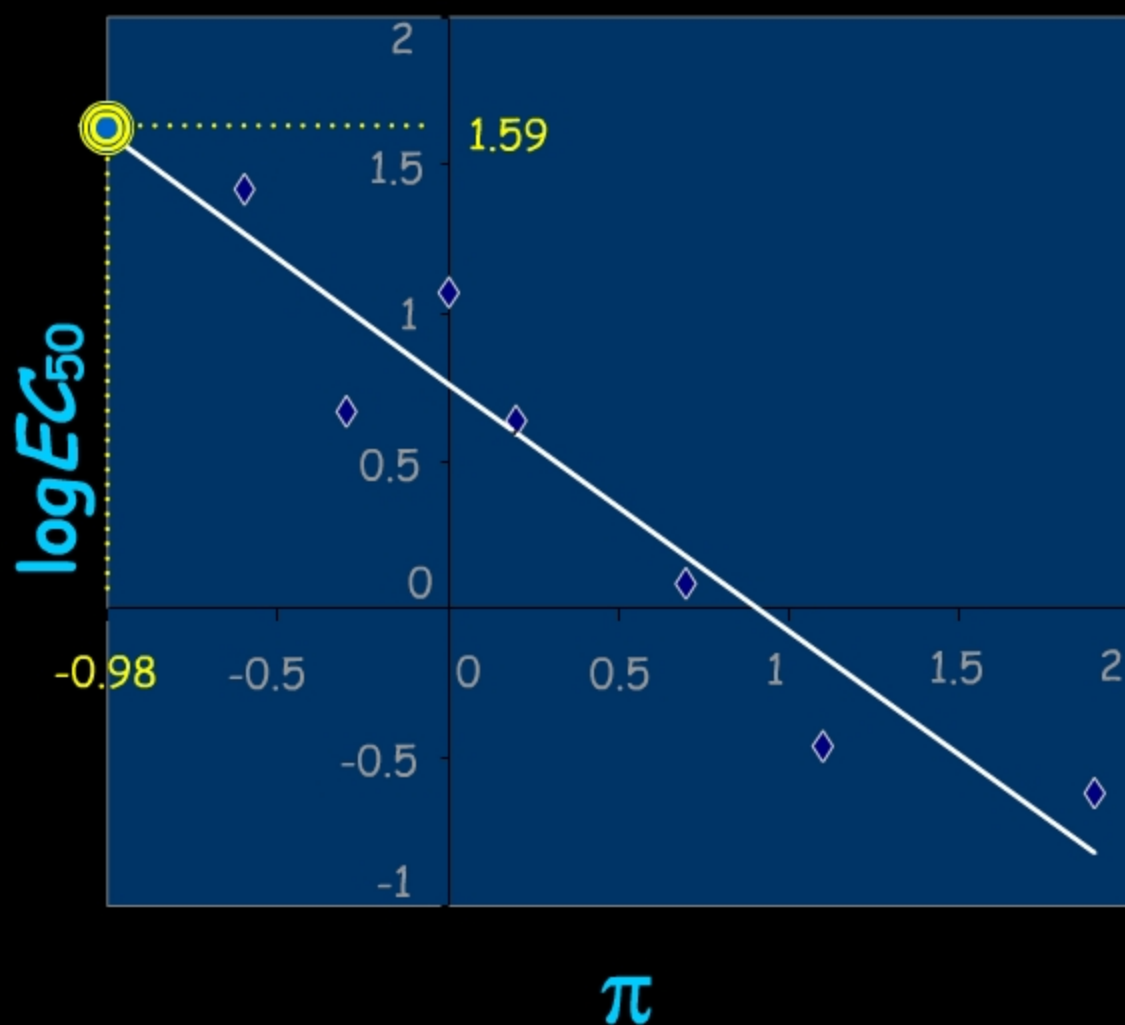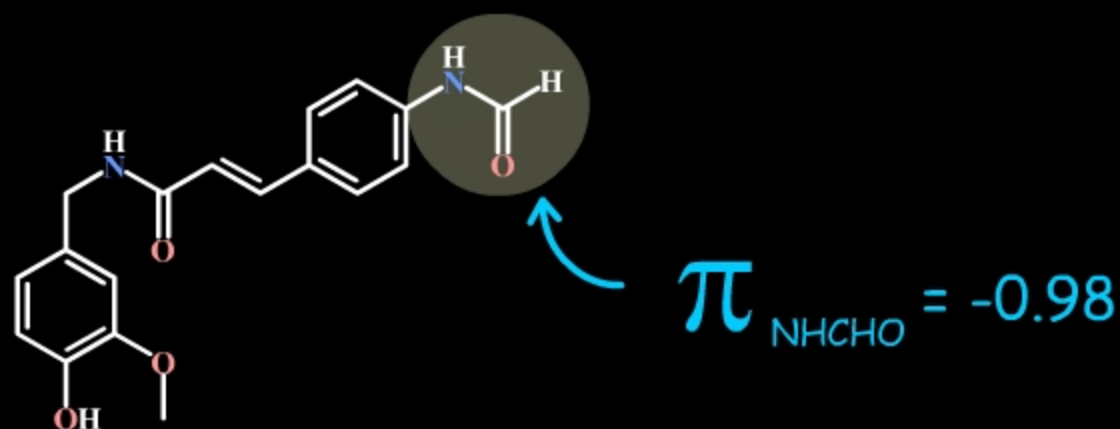
$$\log EC_{50} = 0.794 - 0.817 \times \pi$$

$$r^2 = 0.89; \quad s = 0.28; \quad t = 6.36; \quad F = 40.45$$

The derived QSAR model can be used to predict the biological activities of novel capsaicin analogs by introducing their corresponding π values in the QSAR equation. For example, the biological activity of the amide analog indicated below is predicted with an EC_{50} of 0.98 µM.

$$\pi_{NHCHO} = -0.98$$

$$\log EC_{50} = 0.794 - 0.817 \times \pi$$

$$\log EC_{50} = 0.794 - (0.817 \times -0.98)$$

$$\log EC_{50} = 1.59$$

$$predicted\ EC_{50} = 38.90\ \mu M$$