Chimica Farmaceutica

Application of quantitative structure-property relationships (QSPRs) to the modeling of different types of properties

SAPIENZA
UNIVERSITÀ DI ROMA

- **Main objective**: To learn how to establish and interpret QSPR models by multiple linear regression (MLR) approaches

- **Course contents**

  Introduction. Definition of QSPR/QSAR. Brief historical overview. QSPRs/QSARs by MLR. Molecular descriptors: type and selection. Data preparation for model development (homogeneity, representativity, normalization). The concept of training and test sets. Establishment of model equations. Model fitting. Outliers: detection, evaluation and elimination. Statistical criteria for model validation: internal, external and lateral validation. Quality assessment. Model robustness and model predictive ability. Advantages and disadvantages of MLR. Limitations of QSPR by MLR. Examples of application of MLR-QSPR: solvent effects on solvolysis reactions; prediction of soil sorption coefficients; human skin permeation. Course evaluation. Practical exercises on model building and evaluation.

Present situation in drug discovery

- ✓ **For every 1 drug** that reaches the market, ≈ **5.000 to 10.000** compounds are tested in preclinical trials, ≈ **250 drugs** are tested in preclinical animal trials and ≈ **5 drugs** in full scale human clinical trials. **Only 1 out of 5 drugs** entering clinical trials will gain approval by USA FDA

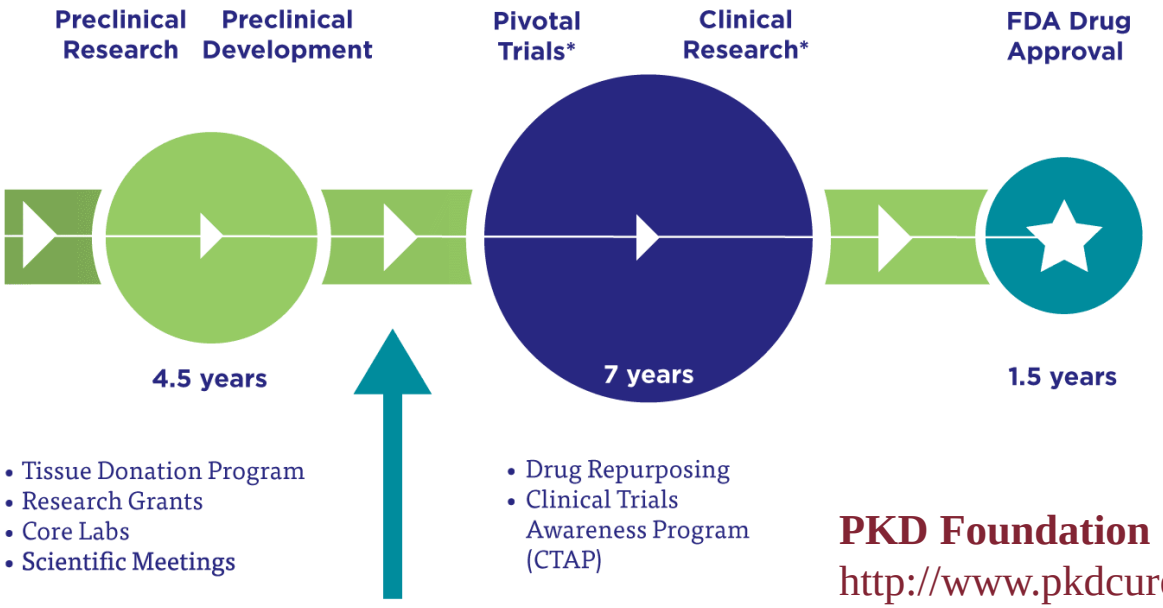
- ✓ Big pharmaceutical industries take **12 to 15 years** for the discovery of each new drug: ≈ 6 years in clinical trials + up to 6-9 more years for approval by FDA

- ✓ The costs involved from the discovery of a new drug to its introduction in the market are estimated to be about **0.8-1.6 billion $** ($10^9$) **!!!**

Glove, G, *AAPS Journal* **2007**, 9(3): E312-E316; Dickson, M., Gagnon, JP, *Nat. Rev. Drug Discovery*, **2004,** 3, 417-429; Dickson, M., *Discovery Medicine*, June **2009**, PhRMA & PKD Foundation, **2015**

Drug Discovery Timeline

Phases of Drug Development



Regulatory Review
Average Development Time: 17 years
Average Cost: $500 million - $1 billion
(After FDA drug approval, post-approval studies may occur.)

***Studies in humans**

PKD Foundation
http://www.pkdcure.org/research/drug-development/timeline (accessed on April 10, 2015)

PhRMA
http://www.phrma.org/media/multimedia/drug-discovery-timeline (video, 2:52) (accessed on April 10, 2015)

Present situation in drug discovery

- ✓ Each year throughout the world only **25** new drugs in average (2005-2013) enter the market, either NME or New BLA*: **41 in 2014**, 27 in 2013, 39 in 2012, 30 in 2011, 21 in 2010, 26 in 2009**,** 24 in 2008, 18 in 2007, 22 in 2006, 20 in 2005

(From Novel New Drugs, 2014 Summary, US FDA, Center for Drug Evaluation and Research (CDER), January 2015, www.fda.gov/drugs)

- ✓ This slight increase in the last few years confirms the R&D shift towards biologics, vaccines and monoclonal antibodies*

↓

Enormous pressure & fierce competition for new drugs

*Present situation in drug discovery

Key words used in the search in Web of Science	Period	Total # of publications /decade
"Lead compound or prototype"	1960-1969 2000-2007	178 25.584
"Design"	1960-1969 2000-2007	9.025 >100.000
"Molecular modeling"	1960-1969 2000-2007	1 3.965
"Synthesis"	1960-1969 2000-2007	32.685 >100.000
"Pharmacological essays"	1960-1969 2000-2007	0 1.205
"ADME" properties	1960-1969 2000-2007	0 504

In *Quim. Nova*, **2007**, 30, 6, 1456-1468

Strategies for the searching of new lead compounds and possibly new drugs

Lead Structure → a representative of a compound series with sufficient potential (as measured by potency, selectivity, pharmacokinetics, physicochemical properties, absence of toxicity and novelty) to progress to a full drug development program

Valler , M.J. and Green, D. *Drug Disc.Today*, **2000**, 5, 286

Strategies for the searching of new lead compounds and possibly new drugs

- Modification and improvement of already existing active molecules ("by chance" or "by structure-based or ligand-based design")

  "copiers"

- Systematic screening of sets of arbitrarily chosen compounds on selected biological assays

  industrious

- Retroactive exploitation of various pieces of biological information (from new discoveries in biology and medicine or just from fortuitous observations)

  intuitive

- Rational design based on the knowledge (or on a fair hypothesis) of the molecular cause of a certain biological response

  deductive

Strategies for the searching of new lead compounds and possibly new drugs

In drug development programs medicinal chemists search for:

- new pharmacophores*
- new chemical structures
- new drugs
- new mechanisms of action

Main goals:

✓ more active

✓ more selective

✓ less toxic molecules

✓ & as fast as possible

And is there a way of relating the structural characteristics of a compd and its biological activity?

Present situation in environmental sciences

A large number of substances have been manufactured and placed in the market for many years now, in Europe and all around the world, sometimes in very high amounts, ***but*** *no sufficient information has been given on hazards to human health and the environment*

Urgent need to introduce some regulations

REACH, European Community Regulation on Chemicals and their Safe Use (EC 1907/2006) was implemented from June 2007 on

Deals with the Registration, Evaluation, Authorization and Restriction of Chemical Substances

http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm

Present situation in environmental sciences

The aim of REACH is to identify, better and earlier, the intrinsic properties of chemical substances, and assess their hazards and risks, fate and effects

↓

Enhance the decision making processes regarding optimization, limitation or prevention of the disposal and/or of the recycling of solid wastes and synthetic chemicals, until they meet pre-set environmental criteria

↓

Improve protection of human health and the environment

http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm;
Handbook of Environmental Chemistry, **2001**, vol. 5E, 243-314

Present situation in environ mental sciences

- ✓ Many halogenated organic compounds such as polychlorinated biphenyls, polybrominated biphenyls, chlorinated aliphatic hydrocarbons, polychlorinated benzenes, polybrominated benzenes, polychlorinated anilines, polychlorinated nitro benzenes and phenols, and alkyl benzenes and phenols are found in the environment

- ✓ Most of them are persistent and show a tendency to accumulate, in biota*, soils and sediments and are also dispersed in the atmosphere.

J. Chem. Inf. Comput. Sci. **2004**, 44, 985-992

Present situation in environmental sciences

- ✓ The fate of these chemicals in the environment, *i.e.*, their partitioning mechanisms at aqueous-solid phase interfaces (water-soil, water-sediment, water-suspended solids, water-biosolids) and at solid or liquid-air interfaces, is controlled by their biological, chemical, and physical properties which are heavily dependent on their structures

So is there a way of relating the structural characteristics of a compd and its properties?

Definition of QSPR

QSPR ≡ Quantitative Structure-Property Relationship

In general terms, includes all statistical and mathematical methods by which various properties (physicochemical, biological, environmental, etc) are related with structural features

Definition

Particular case

QSAR* ≡ *Quantitative Structure-Activity Relationship

BR= *f* (various descriptors)

BR- biological response

QSPR ≡ Quantitative Structure-Property Relationship

Classical QSPR analyses Multiple Linear Regressions (Hansch Approach), Free-Wilson Analyses and Mixed Approaches

Non-Classical QSPR analyses Neural Networks, Decision Trees, Random Forests, Partial Least Squares, Linear Discriminant Analysis, Genetic Algorithms, etc.

3-D QSARs - *e.g.*, GRID/GOLPE and CoMFA

MLR-QSPR

One of the most common and powerful QSPR approaches is to use MLR to express a relationship between a **given property**, ***Y***, of a system and a set of independent **molecular** parameters or **descriptors** X_i, which encode chemical information and model different interaction mechanisms

$$Y = a_0 + \sum_i a_i X_i + \zeta$$

Y = log (1/*c*), log (1/*MIC*), log (1/IC_{50}), log (1/LD_{50}), log *k*, log *s*, or….

a_i - regression coefficients associated with each descriptor

ζ - residuals of the regression*

MLR-QSPR

And why is the dependent variable expressed in a logarithmic scale?

The distribution of a drug or a pollutant, for instance, corresponds to a **partitioning** between an aqueous and a non-aqueous phase

A partition is an equilibrium process and as such is related to ΔG:

$$\Delta G = -2.303\, RT \log K$$

For a given system's response to correspond to an energetic contribution so that it can be related to the interactions modelled by the descriptors, it should be expressed in a logarithmic scale…

log (1/*c*) are often used in QSAR so that smaller *c* values (higher 1/*c*) correspond to more active compounds

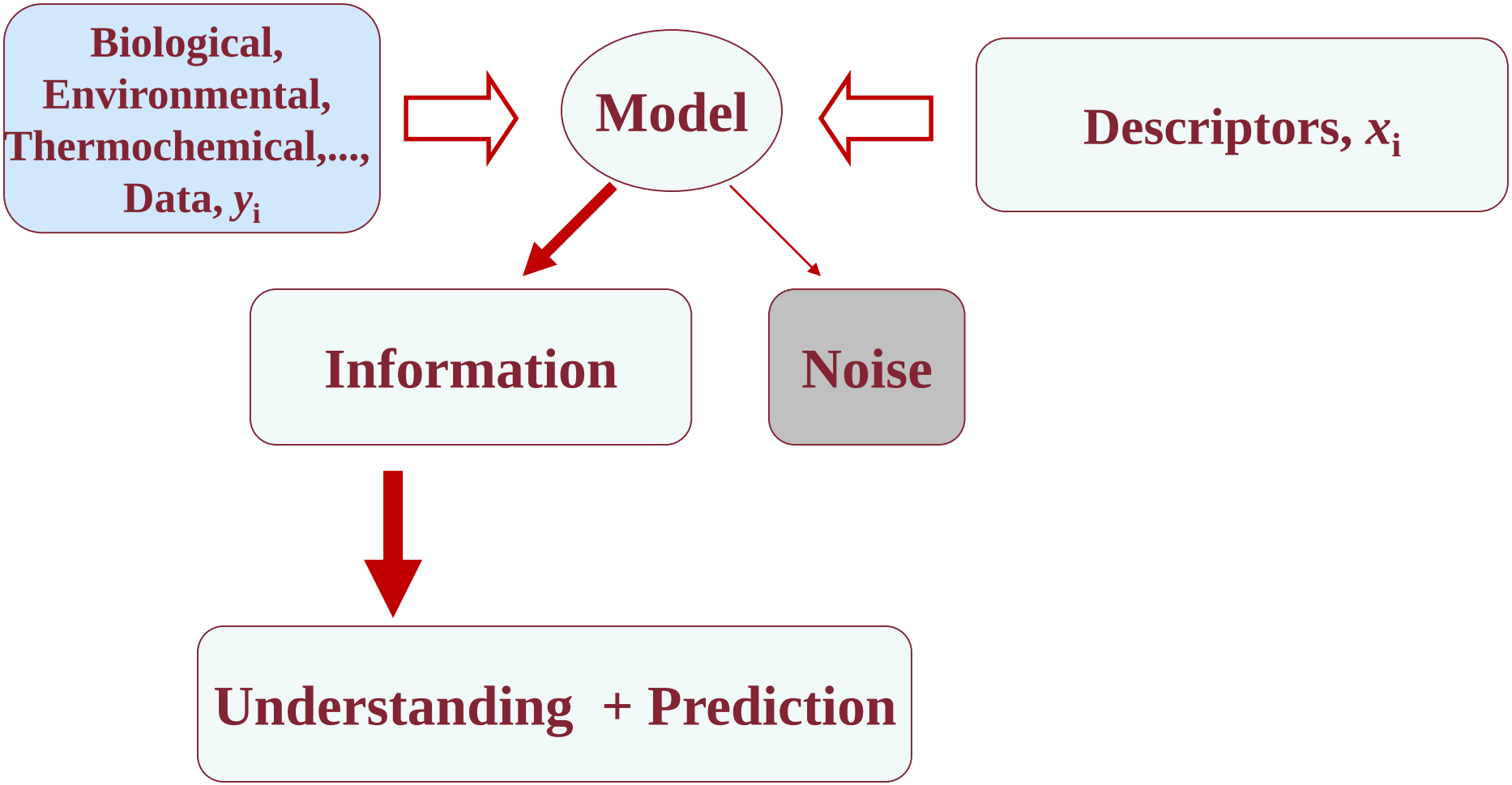
Brief historical overview of QSAR/QSPR

1868	**Crum-Brown & Fraser** ***First QSAR***	Physiological activity, Φ, as a function of chemical structure, C **$\Phi = f(C)$**
1893	Richet	Citotoxicity of some simple organic molecules, inversely related to solubility in H_2O
1901	Meyer & Overton	Narcotic action of some organic compounds, directly related to partition coefs. in oil/water
'40	**Albert, Bell & Roblin**	Importance of the ionization of weak bases and acids in bacteriostatic activity
1952	**Taft** $\log k\,(R\text{-}X) - \log k\,(H) = \rho\sigma$ (*) $\log k\,(R\text{-}X) - \log k\,(H) = \rho^*\sigma^* + \delta E_s$	Proposed an "extension" of the Hammett eq.* to aliphatic cpds to account for both polar and steric effects (E_s)
'60	**Zahradnik**	Applied the concept of the Hammett eq. to biological data: **$\log \tau_i - \log \tau_{Et} = \alpha\beta$**

Cont.-

Early ’60s	**Hansch & Fujita** ***First MLR-QSAR***	Combination of different physicochemical parameters in a linear additive manner $\log(1/c) = a \log P + b\,\sigma + c\,MR \ldots + cte$
1964	**Free-Wilson**	Correlation between the biological activity of a whole molecule and the presence of sub-structural fragments of known activity $BR = = \Sigma\, a_i\, x_i + \mu$
Late ’60s	**Hansch**	Later developments involved **1.** the formulation of parabolic models $\log(1/c) = a \log P + b (\log P)^2 + c\sigma + dE_s + \ldots + const.$
‘70	**Hansch & Free-Wilson models**	**2.** the use of a mixed approach (nice improvements for big data sets with large structural variations)
1977	**Kubinyi**	**3.** the formulation of non-linear, non- parabolic, models, such as the Kubinyi bilinear model $\log(1/c) = a \log P - b \log(\beta P + 1) + \ldots + const.$

Fundamental elements of a QSPR study



QSAR/QSPR Assumptions

- The compounds under study belong to a congeneric series
- All compounds within the series have the same mechanism of action
- The molecular structure is responsible for the observed activity/property
- The factors responsible for the observed biological /chemical response are represented by the descriptors used to encode the compounds' features

It is expected that a small change in chemical structure will be accompanied by a proportionally small change in biological activity (or any other property under study) and that the set of descriptors reveal these analogies*

QSPR ≡ Analogy Models

But in *QSAR* ➡ Similarity Paradox**

QSAR/QSPR Assumptions

- The compounds under study belong to a congeneric series
- All compounds within the series have the same mechanism of action
- The molecular structure is responsible for the observed activity/property
- The factors responsible for the observed biological /chemical response are represented by the descriptors used to encode the compounds' features
- It is expected that a small change in chemical structure will be accompanied by a proportionally small change in biological activity (or any other property under study) and that the set of descriptors reveal these analogies*
- The QSPR/QSAR established model only applies to cpds belonging to the same physico-chemical-biological space

(same applicability domain)

However, there are some differences between using QSAR/QSPR, for example, in pharmaceutical or in enviromental research

QSAR in drug design research	QSPR in environmental sciences
Objectives	
• Optimize biological activities of drugs	• Estimate rates of fate processes (sorption / desorption behaviors)*
• Understand the mechanisms of action	• Analyze processes and understand partitioning mechanisms
• Predict behaviors (activity) prior to any synthesis	• Predict behaviors, fate and effects (toxicity, genotoxicity, bioavailability)
• Find new and more active lead cpds	• Control hazards
Characteristics	
• Response in isolated systems • Effects are specific and well defined • Receptor known only in some cases	• Whole organism response • Net (global) effects • Receptor unknown in most cases
(Some) Techniques	
• Hansch approach • Multivariate data analysis • Computational molecular modeling	• Hansch approach • Multivariate data analysis • Molecular modeling <u>not applied</u>

Adapted from *Handbook of Environmental Chemistry*, **2001**, vol. 5E, 243-314

QSAR/QSPR Objectives

- **Identification** of the key factors that determine a certain biological/chemical response (**interpretative ability**)

- **Prediction** of the system's behaviour when convenient structural modifications are designed and introduced in a molecule, prior to any synthesis, microbiological assay or experimental measurement (**predictive ability**)

⇓

Scale economies of effort, time and expenditure

⇓

Very imp. to Pharmaceutical, Chemical, Food, Agrochemical, ...Industries

Selection of the "right" descriptors for a QSAR/QSPR

Descriptors

Ideally a **descriptor** models

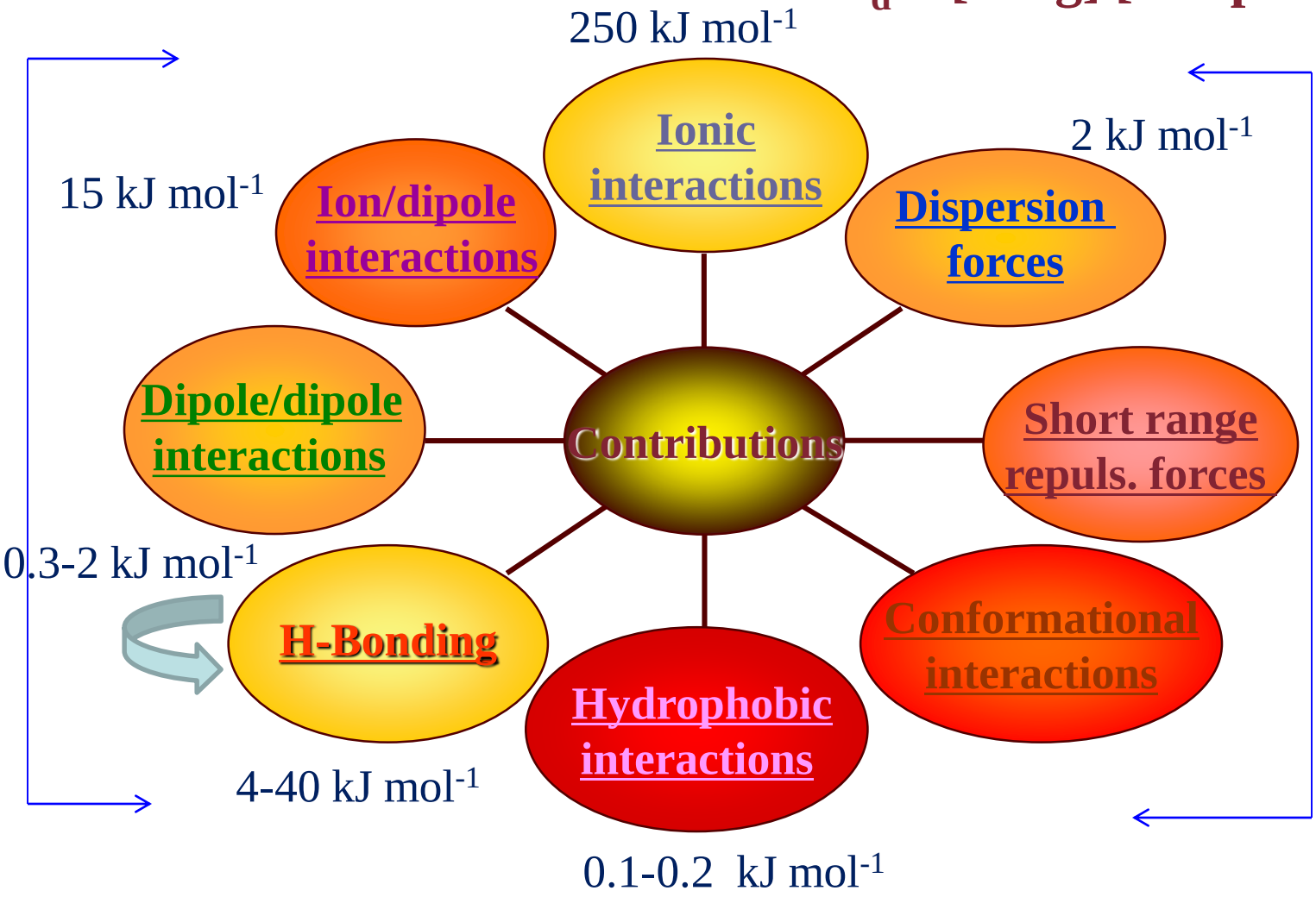
a certain type of **interaction** mechanism

In the case of a QSAR, the biological effect of a drug results from its **interaction** with a specific target (an enzyme, an ion channel, a nucleic acid, or any other biological macromolecule) and this interaction is determined by **intermolecular forces**

What type of intermolecular forces are these?

Intermolecular forces in d/r interactions

K_d = [drug] [receptor] / [complex]



K_d depends mainly on:

- Electrostatic d/r interactions
- Stereochemical d/r interactions

Descriptors

Molecular Descriptors

Structural or constitutional

MW, total # atoms, total # bonds…

Sub-structural or fragment

***L*, B_5, B_1, # nitro groups, # H-bond donors…**

Physicochemical

p*K*a, log *P*, solubility…

Electrostatic

μ, *c*, av. *I*, α….

Topological

Wiener index, 2D autocorrelation vector Connectivity indices….

Quantum Chemical

E_{tot}, E rep. e^-/e^-, E_{HOMO}, E_{LUMO}….

Geometrical

3D Wiener indices, 3D autocorrelation vectors, RDF, MV, MSA….

Thermodynamic

H_{vib}, H_{transl}, S_{vib}, S_{rot} ….

May be experimental or calculated, pure or composed

Descriptors

Some commercially available programmes for the calculation of molecular descriptors:

E-DRAGON: http://www.**vcclab**.org (free access; version 5.4 calculates more than 1600 molecular descriptors)

CD*K*Desc*UI*: http://www.rguha.net/code/java/cdkdesc.html

CODESSA Pro: http://www.codessa-pro.com

SPARC: http://sparc.chem.uga.edu/sparc/ (free access) (T_m, T_b, P_{vap}, S_w, $log\ K_{ow}$, K_{aw}, p*K*a)

MOLECULAR MODELING Pro Plus (MMPro): www.chemistry-software.com

Descriptors

- This parameterization of chemical structures or substructures is not only of great importance to QSPR studies but it has also much interest in definitions of Molecular Similarity and Diversity

- This type of information can be used in Molecular Modeling Studies or in Combinatorial Chemistry when one wants to generate a wide molecular diversity to improve chances of finding promising compounds

Descriptors

Most used descriptors in QSAR studies:

- Lipophilic parameters

  e.g., partition coefficients (log *P*), Hansch hydrophobic parameter (π), cromatographic parameters (*R*M, log *k'*)

Descriptors

Most used descriptors in QSAR studies:

- Lipophilic parameters
- Polarizability parameters
  e.g., molar refractivity (*MR*), excess molar refraction (R_2)

Descriptors

Most used descriptors in QSAR studies:

- Lipophilic parameters
- Polarizability parameters
- Electronic/electrostatic parameters
  e.g., Hammett σ ctes., dipole moments (μ), quantum chemical parameters, dipolarity/polarizability parameters (π_2^H)

Descriptors

Most used descriptors in QSAR studies:

- Lipophilic parameters
- Polarizability parameters
- Electronic/electrostatic parameters
- Steric parameters

  e.g., Taft E_s cte., Sterimol parameters (B_1, B_5, L)

Descriptors

Most used descriptors in QSAR studies:

- Lipophilic parameters
- Polarizability parameters
- Electronic/electrostatic parameters
- Steric parameters
- H-bonding effects parameters
  e.g., H-bond acidity and basicity parameters ($\Sigma\alpha_2^H$, $\Sigma\beta_2^H$)

Descriptors

Most used descriptors in QSAR studies:

- Lipophilic parameters
- Polarizability parameters
- Electronic/electrostatic parameters
- Steric parameters
- H-bonding effects parameters
- Indicator variables (*I*)
- And also *MW*, *MV*
- And in environmental processes also m.p., b.p. vapour pressure, p*K*a, water solubility, K_{aw}*

Descriptors

- There are more than 12.500 websites on QSARs

- The Pomona College Group Database (Leo and Hansch DB, now www.biobyte.com), has more that 17.000 QSARs from which about 8.500 are relative to biological systems and around 8.600 are related to Physical Organic Chemistry problems

C. Hansch, D. Hoekman, A. Leo, D. Weininger; C. D. Selassie *Chem. Rev.* **2002**, 102, 783-812

- There are various commercial programs with computational, statistical and graphic tools to perform QSAR /QSPR studies:

  Accelrys - http://www.accelrys.com/
  Codessa- http://www.semichem.com/
  HyperChem - http://www.hyper.com/
  ChewSW- http://www.chemsw.com/
  Tripos - http://www.tripos.com/
  Chemical Computing Group http://www.chemcomp.com/
  Unscrambler - http://www.camo.com/
  QSARINS - www.qsar.it (Paola Gramatica's group at University of Insubria)

  etc.

Descriptors

❖ **log *P*** (physicochemical parameter)

Lipophilicity is the property that has generated more interest in QSAR studies due to its direct relationship with solubility in aqueous phases, with permeation through membranes and with entropic contribution to drug-receptor binding

Def. It is defined as the partition of a cpd between a non aqueous and an aqueous phase

$$P' = c_{org} / c_{aq}$$

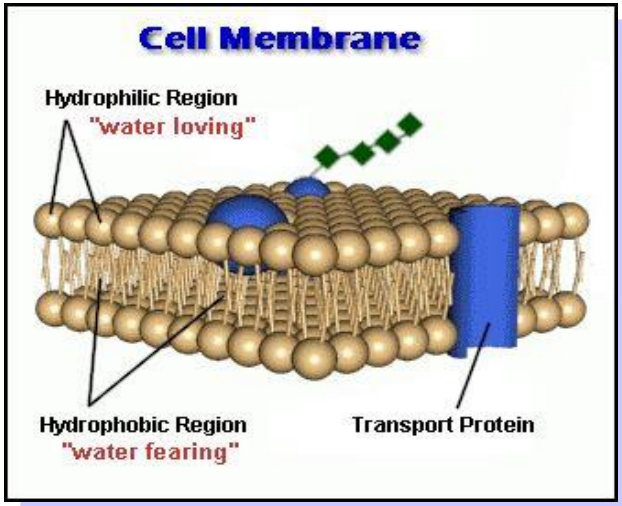
The most used organic phase in partition coefficients' studies is *n*-octanol

$P' \rightarrow P$ – *n*-octanol/water partition coefficient

Descriptors ❖ log *P*

Why are partition coefficients normally measured in *n*-octanol/water systems?

- *n*-octanol is a cheap solvent, relatively non-toxic and chemically non-reactive
- Is considered a "good" model of the lipidic constituents of biological membranes
  - ✓ Has a long alkyl chain (hydrophobic)
  - ✓ Has a polar hydroxylic group (hydrophilic)
  - ✓ OH group has an amphiphilic behaviour and is thus able to form H-bonds, like phospholipids and proteins in biological membranes



- ✓ octanol OH group is HBD & HBA, and is able to interact with polar groups of various solutes
- ✓ dissolves more organic cpds than alkanes, cycloalkanes or aromatic solvents
- ✓ it is transparent in UV, which facilitates quantitative measurements
- ✓ has $\downarrow p_{vapour}$,* which allows reproducible measurements, etc

Descriptors

❖ **log *P***

☻ Calculated partition coefficients also refer, in general, to *n*-octanol-water system

But the choice of *n*-octanol-water system as a mimetic system of biological membranes has been much debated since log *P* (oct/H_2O) is considered <u>not to</u> model adequately drug/receptor <u>specific</u> interactions in the lipidic bilayer due to being an isotropic (equal in all directions) lipophilicity parameter*

Solvents such as alkanes (inert solvents), particularly cyclohexane, chloride substituted hydrocarbons like chloroform (HBD) and propylene gycol dipelargonate - PGDP (HBA), have been proposed to model different membranes or parts of membranes and tissues

Differences in log *P* (Δlog *P*) measured in two solvent systems (*e.g.*, octanol-water and cyclohexane-water) have given information on the ability of a given cpd to form H-bonds

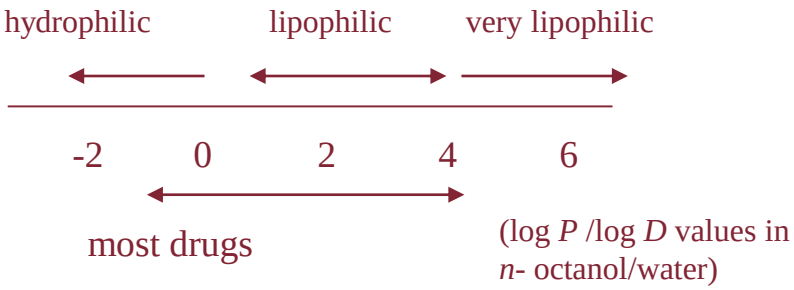
Descriptors

❖ log *P*

* However, from the 8.500 QSARs of biological systems of the Pomona College DB, **4614** have a term in log P (oct/H_2O)....

* This database has ≈ 30.000 log P (oct/H_2O) exp. values of which ~12.000 are considered reliable values and refer to the neutral species partition

Descriptors



❖ log *P*

Methods of log P determination

- Experimentally: classical ***shake-flask method*** (**S-F**) - not simple, expensive, & time-consuming* (-3 < log *P* < +3)**, but, if usable, is accurate and precise; **HPLC** is the preferred method in various labs, especially in industry, although there are many other methods to determine log *P*, for instance, Sirius® **potentiometric method** which allows measurements in extreme conditions (for very↓ and very↑ lipophilic cpds) **(Bosch, E., Martins, F, *et al. J. Chem. Eng. Data*, 2012, 57,330)**

- By calculation: one of the most advanced and reliable programs to estimate log *P* oct/H_2O is Clog *P**** (a group contribution method) (Leo, A. *Chem. Rev.*, **1993**, 93, 1281) which gives values that correlate well with exp. log *P*

$$\log P = 0.96\ (\pm 0.003)\ \mathrm{Clog}\ P + 0.08\ (\pm 0.008)$$

$N = 12107$; $r^2 = \mathbf{0.973}$; $sd_{fit} = 0.299$

Descriptors

❖ *MR* (physicochemical/stereochemical parameter)

The most used polarizability parameter has been the molar refractivity, *MR*

$$MR = MV\,[(n_D^2-1) / (n_D^2+2)]$$

MV – molar volume

n_D – refractive index

- ✓ Better than *MV* because n_D reflects polarizability ∴ *MR* reflects size and polarity/polarizability of a given group; however it says nothing about <u>shape</u>
- ✓ On the Pomona College DB there are **2553** QSARs based on the ***MR*** descriptor but only **422** involve ***MV***. Clog *P* also allows the estimate of *MR*
- ✓ Many QSAR studies on ligand-enzyme interactions have shown that substituents modelled by *MR* bind preferentially to polar areas whereas substituents modelled by π (Hansch lipophilicity parameter) bind to hydrophobic areas

Descriptors

❖ ***B_1, B_5, L*** (stereochemical parameters)*

Stereochemical effects are still difficult to describe due to the often lack of knowledge about tridimensional structures of drug binding sites

✓Some progress was attained with the definition of Verloop Sterimol parameters (1976). These are **calculated** parameters, in which *L* is a measure of the length of the substituent along the axis that connects it to the main molecule, and *B* parameters are orthogonal (also with *L*), being B_1 essentially a measure of the size (largely a steric effect) of the first atom in the substituent and B_5 is an attempt to define the effective volume of the whole substituent

✓To define space requisites of a given substituent, the program uses Van der Waals radii, distances and bond angle, and conformation estimates

✓**907** QSARs listed in the Pomona College DB use B_1, **728** use B_5 and **104** use L

Descriptors

❖ π_2^H, $\Sigma\alpha_2^H$, $\Sigma\beta_2^H$ (electrostatic/electronic)

The electronic properties of molecules can be described by an enormous variety of parameters (pK_a, μ, H-bond parameters, Hammett σ cte., etc.)

Lipophilicity parameters & *MR* → global properties of molecules

Electronic parameters → refer, <u>in general</u>, to a given atom or group

BUT π_2^H, $\Sigma\alpha_2^H$, $\Sigma\beta_2^H$ parameters **also** refer to total effects

✓ π_2^H - dipolarity/polarizability — from GLC measurements in polar stationary phases or from solvent/H_2O partition coefficients' measurements

✓ $\Sigma\alpha_2^H$ – H-bond total or effective acidity

✓ $\Sigma\beta_2^H$ – H-bond total or effective basicity

from solvent/H_2O partition coefficients' measurements

Descriptors

❖ Indicator variables

Used in MLR, to account for some characteristics that cannot be described by continuum variables and which lead to an unusual activity or to its absence within a set of cpds

- ✓ Correspond, normally, to a structural element, a substituent or other fragment, which produces (1) or not (0) a given effect
- ✓ Very useful in the first stages of a QSAR analysis and for large, heterogeneous and complex data sets: diferent sub-sets may start to be combined through them, until the true dependence between biological activity and physicochemical parameters is derived on the basis of a wider structural variation

*

Descriptors

But

Missing descriptors to describe, conveniently, important interactions in QSAR studies such as:

- The **partition** of drugs through membranes (octanol/water has proven not to be the ideal system to mimetize some biological membranes; work with liposomes, for instance, seems much more promising)
- The strength of **hydrogen bonding**
- The influence of **desolvation energies** in drug-receptor affinity

Descriptors

*The **selection** of **adequate** and **sufficient descriptors** to describe a given behavior is among one of the most **difficult tasks** for researchers*

Selection of Descriptors

What descriptors should be selected? *

- Those which are relevant to explain the variability in the response for the series of compounds being analyzed**
- Those that, being relevant, are not intercorrelated, so that there is no *redundancy* in the information described by the various descriptors***
- The smallest possible number to prevent *chance correlations* and thus facilitate the interpretation of the resulting models in physicochemical or mechanistic terms §
- Those which are interpretable and "reversible" § §

Selection of Descriptors

*

- ✓ By chance
- ✓ By intuition
- ✓ Exhaustive method (use all available descriptors in a sequential search (SS) …)
- ✓ Forward- or backward-stepping regression (very common in MLR)

And for <u>approaches other than MLR</u>, also by:

- ✓ Principal component analysis (*PCA*)
- ✓ Cluster analysis (*CA*)
- ✓ Genetic algorithms (*GA*)
- ✓ Neural networks (*NN*)
- ✓ Random Forests (RF)
- ✓ Kohonen self-organizing maps (SOM) ***etc.***

Preparation of data for QSPR development

One of the aspects that is sometimes overlooked in QSPR studies is the basic knowledge about the nature of the data to be analyzed

Are data

- Acurate?
- Precise?
- Complete?
- Consistent ?
- Representative?

So, First step in the development of any QSPR/QSAR model is the **Preparation of Data**

And how do we do this?

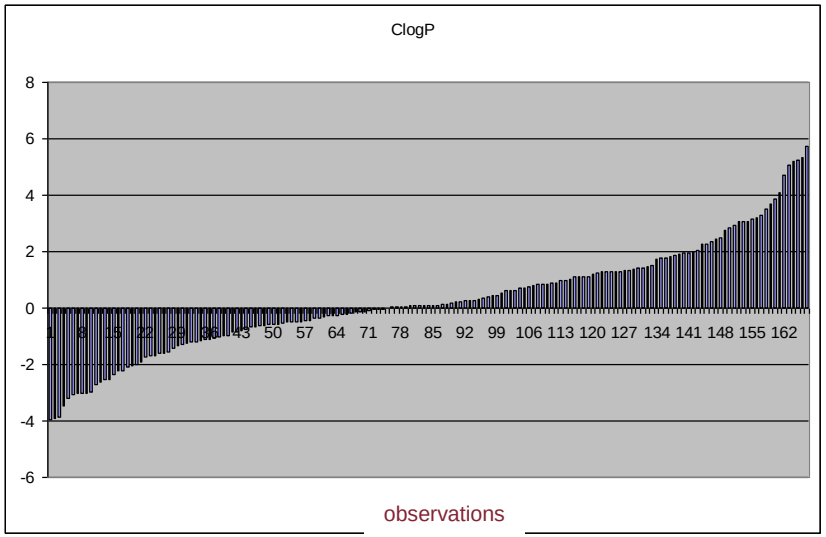
Data Preparation

1) Homogeneity

Data (Y_i e X_i) should present an **uniform, homogeneous distribution** (*i.e.*, without *clustering* or influential points)

2) Representativity

The tested compounds should be distributed throughout the **whole chemical/structural multidimensional space** – there should be observations representing values in the whole range of variation of the variables



Ex. values of *ClogP* (*Y*) of 167 compounds varying between - 4 e +6

Sometimes this implies a <u>mathematical transformation of the dependent variable</u>! *e.g.*,

$$Y' = \frac{1}{Y} \quad \text{or} \quad Y' = log\left(\frac{1}{Y}\right)$$

Data Preparation

3) Normalization (range scaling)

$$x_{ij}^{n} = \frac{x_{ij} - x_{j\,min}}{x_{j\,máx} - x_{j\,min}}$$

allows direct comparison of coefficients when they have different orders of magnitude* (normally, imposes a variation between 0 and 1)

x^{n}_{ij} is the new scaled value

Auto - scaling

$$x_{ij}^{n} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

with

$$\sigma_j = \sqrt{\left(\sum_{i=1}^{N} \frac{(x_{ij} - \bar{x}_j)^2}{N-1} \right)}$$

Data Preparation

4) Intercorrelation among descriptors

If $r^2(x_i, x_j) < 0.5$ then we can consider descriptors to be independent

(**obs.** *multicollinearity* should also be checked $\Rightarrow R^2$ of X_i against all other X_j must be < 0.8)

r^2	clog*P*	σ	B_5	L	I	μ
clog*P*		0.117	0.08	0.248	0.153	0.092
σ			0	0.03	0.008	0.008
B_5				0.002	0	0.058
L					0.641	0.002
I						0
μ						

L and *I* are intercorrelated

5) Detection of outliers

By

- graphical methods (*histograms*, *residuals plots*, *normal probability plots*)

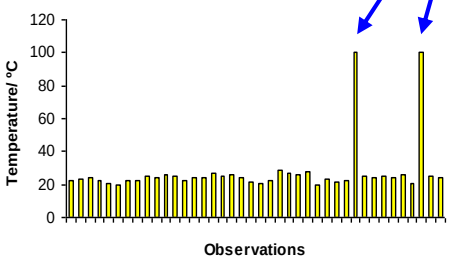
or

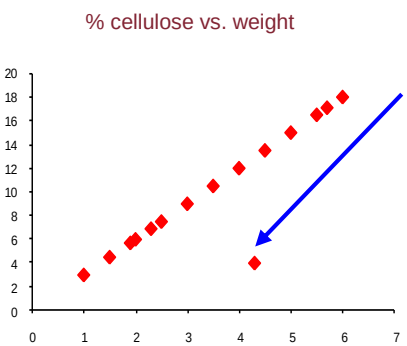
- analytical methods

Before fitting

Def. observation that stands out from the pattern of distribution of the data

outliers





Data Preparation

"Outliers constitute a serious problem in QSAR studies. Most often they are omitted from the data set without further comments, which is not a good practice. A lot of information might be derived from the careful inspection and consideration of the residuals of a multiple regression analysis and of the so-called outliers"

Kubinyi, Wolff, M.E., ed., Burger's Medicinal Chemistry and Drug Discovery, John Wiley & Sons: New York, 2003.

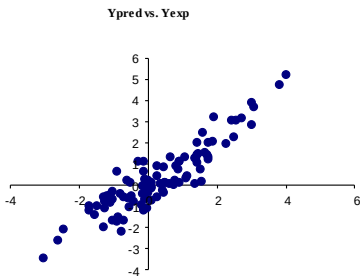
Data Preparation

5) Detection of outliers

After fitting

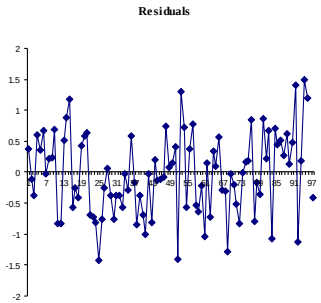
Graphical methods

- Y_{pred} *vs.* Y_{exp}



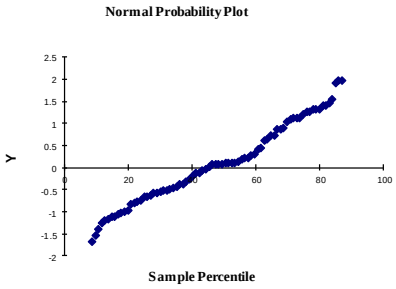
- linear trend
- $m \rightarrow 1$
- $a_0 \rightarrow 0$

- Residuals *vs.* time
- Residuals *vs.* Y
- Residuals *vs.* X_i



- random distribution without any pattern
- nil average
- constant variance

- Normal probability plot

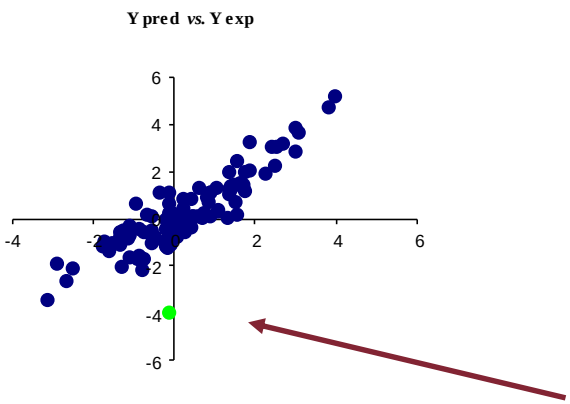


- nearly linear

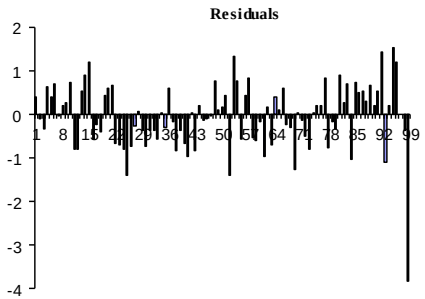
Data Preparation

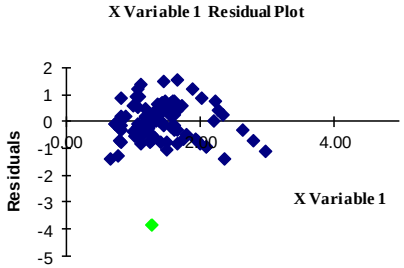
5) Detection of outliers

- Y_{pred} *vs.* Y_{exp}



- Residuals *vs.* time
- Residuals *vs.* Y
- Residuals *vs.* X_i





Data Preparation

5) Detection of outliers

After fitting

Analytical methods

Calculation: easy; fast; very common in the literature; BUT not very discriminatory

$|Y_{pred} - Y_{exp}| > 2\,s$

- Fit the data
- Calculate 2 s
- Calculate the differences $|Y_{pred} - Y_{exp}|$
- If dif. > 2 s $\Rightarrow$ remove observation
- Re-fit the data

Cook's Distance

$$D_i = \frac{\sum_i (\hat{Y} - \hat{Y}_i)^2}{p\,s^2}$$

$\hat{Y}, \hat{Y}_i$ - predicted observations for full data set & for data set without the ith observation

n – number of points; p - number of adjusted parameters; s^2 – model's variance without the ith observation

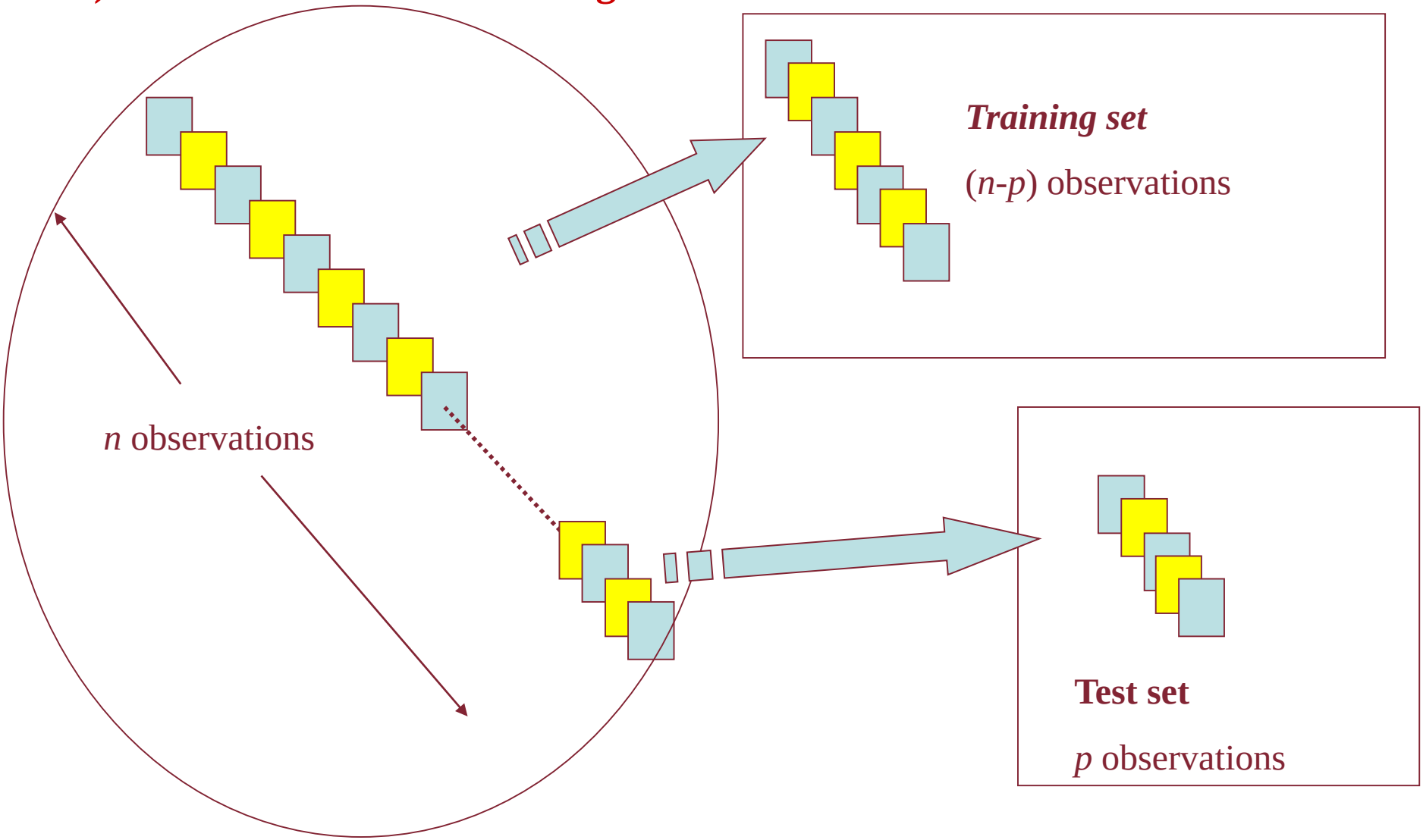
More sensitive, more discriminatory

Suppose you have more than 1 outlier
Consider ***n*** observations and ***k*** potential *outliers*:

- fit the data with all ***n*** observations
- fit **(*n*-*k*)** observations, each one without the potential outlier
- in each case, calculate a D value
- the suspicious observation is an outlier if $D_i > 4 / (n\text{-}p\text{-}1)$
- remove the outliers and re-fit the data
- check again if there is no outlier left

Data Preparation

6) Division of data into training and test sets



Data Preparation

6) Division of data into training and test sets

- ***Training set*** ⟶ used to <u>establish the model</u> equation

- ***Test set***, or prediction set ⟶ independent from the first one, used to <u>test the model</u>, allowing therefore an <u>external validation</u>

- An ideal division should result in a *test set* such that each of its observations should be nearby at least one *training set* observation (***similarity principle***)

Data Preparation

Common methods used to make the division between *training* and *test sets*

- ✓ Selection by chance
- ✓ Selection by data observation (H and R) - of the dependent variable and/or the independent variables
- ✓ Selection by dependent variable, Y_i, values (*e.g.*, activity values)
- ✓ Selection by *clustering* techniques
- ✓ Selection by D-optimal design algorithms
- ✓ Selection by *Self-Organizing Maps* (SOM)
- ✓ *etc.*

Data Preparation

In general, the choice of *training* and *test sets* should obey the following four criteria:

- *Training set* with adequate dimension (4-5 points per descriptor to avoid chance correlations)
- *Training set* should be as diverse as possible
- Representative points of *test* and *training sets* should be close but no *test set* points beyond *training set* points to avoid predictions outside the model's applicability domain
- Both sets should have an approximate dimension or at least a distribution 50-70% (*training set*) to 30-50% (*test set*)

Multiple Linear Regressions (MLR)

Evaluation of the model's significance and of the significance of the regression coefficients

Multiple Linear Regressions (MLR)

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$ **Model equation**

Some statistical definitions

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2$$

Sum of squares of residuals
(or variance in *y* <u>not</u> explained by the regression or unexplained sum of squares)

$$\hat{e}_i = y_i - \hat{y}_i$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Sum of squares of regression
(or variance in *y* explained by the regression or explained sum of squares)

$$\boxed{SST = SSR + SSE}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Sum of squares total (or total variance in *y*)

$\bar{y}$ – simple average of y_i values

Multiple linear regressions (MLR) – Analysis of Variance (ANOVA)

ANOVA	*df*	*SS*	*MS*
TOTAL	n-1	*SST*	
REGRESSION	p	*SSR*	*MSR* = *SSR* / p
RESIDUALS	n-p-1	*SSE*	*MSE* = *SSE* / (n-p-1)

df – degrees of freedom of SS term

p – number of variables

n – number of observations

R – correlation coefficient

R^2 – determination coefficient

s – standard deviation

1. $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

2. $$R^2_{adjusted} = 1 - \frac{(n-1)}{(n-p)}(1 - R^2)$$

R^2 – proportion of variance of data explained by regression model

$$\boxed{SST = SSR + SSE}$$

3. $$F = \frac{MSR}{MSE}$$

4. $$t_{Stat} = \frac{regression\ \ coefficien\ \ t}{standard\ \ deviation}$$

t- measure of the significance of individual terms in a regression eq.

F – measure of the overall significance of the regression model

MSE = variance of residuals = s^2

MSR = variance of regression

5. $$s = \sqrt{\frac{SSE}{(n\text{-}p\text{-}1)}}$$

s – absolute measure of the fit's quality

Multiple linear regressions (MLR)

F distribution table for a 95% confidence interval (CI): υ_1 is the number of variables, *p*, and υ_2 is given by *n*-*p*-1

	ν_1						
ν_2	1	2	3	4	5	10	∞
1	161.4	199.5	215.7	224.6	230.2	241.9	254.3
2	18.5	19.0	19.2	19.2	19.3	19.4	19.5
3	10.13	9.55	9.28	9.12	9.01	8.79	8.53
4	7.71	6.94	6.59	6.39	6.26	5.96	5.63
5	6.61	5.79	5.41	5.19	5.05	4.74	4.36
6	5.99	5.14	4.76	4.53	4.39	4.06	3.67
7	5.59	4.74	4.35	4.12	3.97	3.64	3.23
8	5.32	4.46	4.07	3.84	3.69	3.35	2.93
9	5.12	4.26	3.86	3.63	3.48	3.14	2.71
10	4.96	4.10	3.71	3.48	3.33	2.98	2.54
15	4.54	3.68	3.29	3.06	2.90	2.54	2.07
20	4.35	3.49	3.10	2.87	2.71	2.35	1.84
30	4.17	3.32	2.92	2.69	2.53	2.16	1.62
40	4.08	3.23	2.84	2.61	2.45	2.08	1.51
∞	3.84	3.00	2.60	2.37	2.21	1.83	1.00

Ex. $p = 3$ e $n = 12$

$F_{tab} = 4.07$

In Livingstone, D., *A practical guide to scientific data analysis*, Wiley & Sons Ltd, Chichester,, **2009**.

If, for a given CI, $F_{calc} > F_{tab}$ then *F* has statistical significance and the model eq. is significant at that particular CI

Multiple linear regressions (MLR)

A QSPR/QSAR model <u>can be accepted</u> if:

- The determination coefficient , R^2, is ≥ 0.60-0.70*
- The standard deviation is not much higher than the standard deviation of the biological data
- The *F* value is higher than that of a *F* distribution table for a given CI, for the same degrees of freedom
- All regression coefficients have statistical meaning at a significance level of 95%

A QSPR/QSAR model <u>should be rejected</u> if:

- The number of variables is very high (lost of physicochemical meaning)
- If standard deviation is lower than the error in the biological data ("over-prediction" by the model)

Multiple linear regressions (MLR)

To obtain a **robust QSPR/QSAR model equation, with statistical significance** and simultaneous good **interpretative** and **predictive abilities**, some conditions have to be met:

- Data should be reliable, homogeneous and representative*
- Variables should not be correlated, they should be orthogonal** ⇒ set a *intercorrelation matrix* between pairs of descriptors and determine r^2 (the determination coefficients between pairs of descriptors, r^2, should be < 0.5)***
- The model should contain a reduced number of variables in order to avoid chance correlations. The ratio between the # of cpds and the # of descriptors should be > 4-5 (that is, at least 4-5 points per each variable)
- The model should be consistent with the physicochemical or biochemical nature of the process under study

Multiple linear regressions (MLR)

A reliable and robust **QSPR/QSAR** model must be:

- Statistically significant and robust √
- **Validated** for a set of data not used to develop the model
- And its applicability domain should be well defined

Only QSPR/QSAR <u>validated</u> models may provide a sound mechanistic interpretation, which is particularly relevant for the development of new drugs/materials

Multiple linear regressions (MLR)

And how do we validate a QSPR/QSAR model?

Model fitting and quality assessment (internal and external validation)

Model Fitting

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

1. a_i are determined by minimization of sum of squares of residuals (*e.g.*, by the **least squares method**)

   Assumptions:

   - ✓ X_i without error
   - ✓ Y_i with $N(\mu; \sigma^2_{cte})$
   - ✓ $\varepsilon(Y_i)$ with $N(0; \sigma^2_{cte})$

2. Select best set of X_i, *e.g.*, by forward-stepwise method*

3. Best model
   - → **Internal validation** to ensure ***robustness***
   - → **External validation** to ensure ***predictive ability***

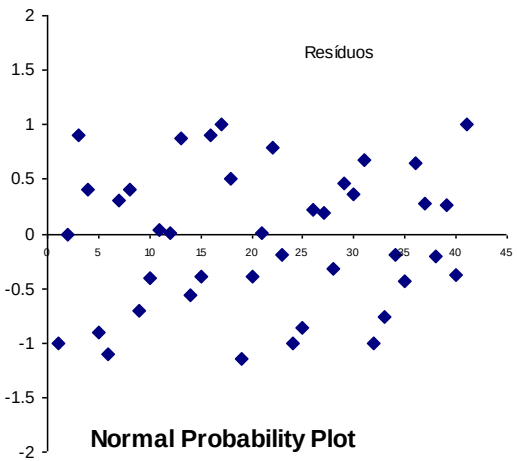
And how do we do this?

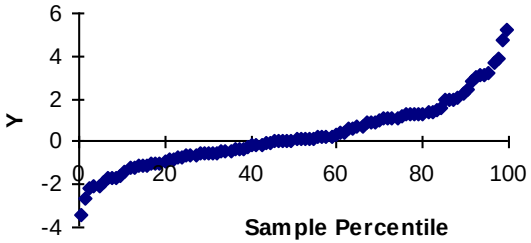
Internal Validation

Training set

1. Linearity evaluation (residuals analysis; normal probability plots)

- ✓ $m \longrightarrow 1$
- ✓ Random residuals distribution
- ✓ Residuals with nil average
- ✓ Residuals with constant variance
- ✓ Normal probability plot nearly linear





Internal Validation

Training set

2. Outliers' detection

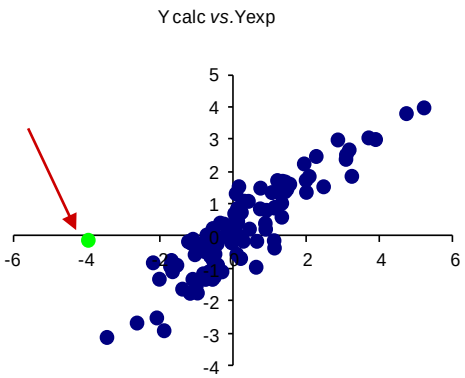
✓ Scattering plots Y_{calc} *vs.* Y_{exp}

✓ $| Y_{calc} - Y_{exp} | > 2\ SD$

✓ Cook's distance

$$D_i = \frac{\sum_i (\hat{Y} - \hat{Y}_i)^2}{p\ s^2}$$

$$D_i > 4/(n - p - 1)$$



Internal Validation

Training set

3. Statistical criteria

- ✓ $R^2 > 0.6$
- ✓ $SD \approx$ exp error of Y
- ✓ $SL > 95\ \%$ (Student's t distribution)
- ✓ $F \uparrow$ CV Q^2_{LMO} or $Q^2_{LOO} > 0.6$

$$Q^2 = 1 - \frac{\sum_{i=1}^{training} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{training} \left(Y_i - \overline{Y}_i \right)^2}$$

Internal Validation

Training set

Cross validation correlation coefficient, Q^2_{LMO}

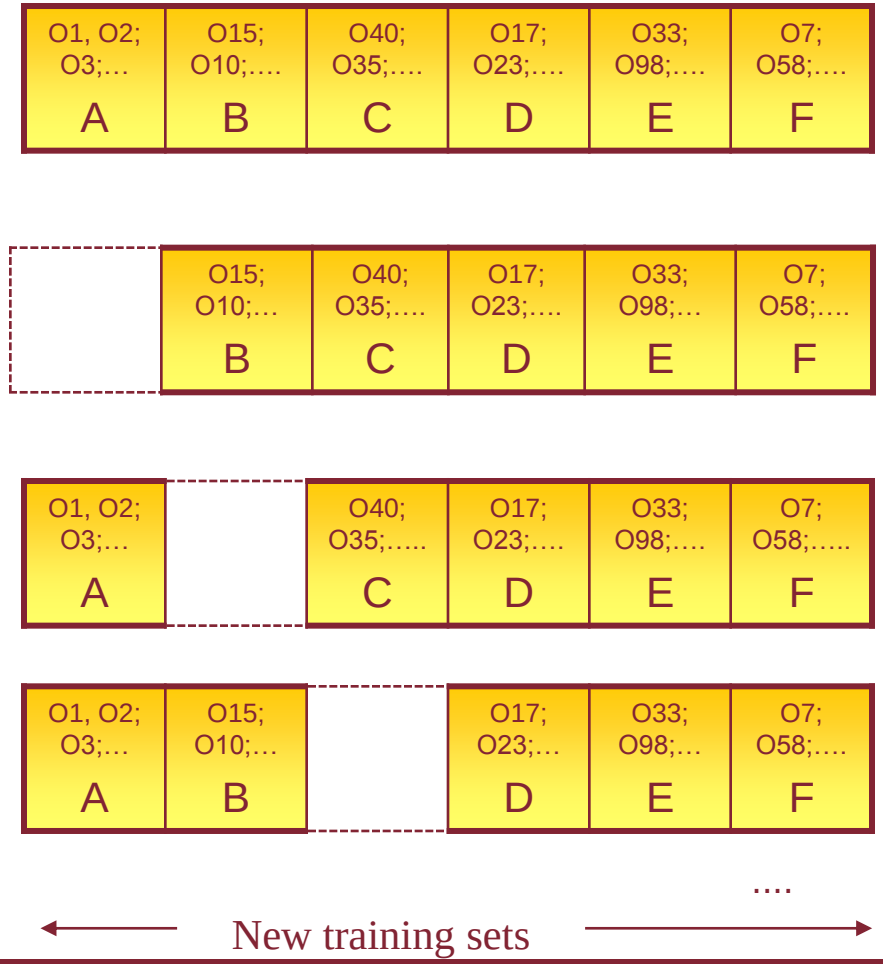
1. First we divide randomly the ***n*** observations in the training set in ***p*** sub-sets approximately with the same size

2. Then, we remove **one** of the ***p*** sub-sets and calculate the fitting with the remaining points

3. We proceed by calculating Q^2 (without A) ; Q^2 (without B); Q^2 (without C) …..

$$Q^2 = 1 - \frac{\sum_{i=1}^{training} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{training} \left(Y_i - \bar{Y}_i \right)^2}$$

4. At the end we determine Q^2_{LMO} (*i.e.*, the average of $Q^2_{w/o\ A}$; $Q^2_{w/o\ B}$; …..; $Q^2_{w/o\ F}$)



External Validation

Test set

1. Predict Y_i values for test set compounds, characterized by X_i, with training set coefficients, a_i

2. Plot Y_{calc} *vs.* Y_{exp} for test set (*scatter plot*)

3. Statistical criteria for test set

✓ $R^2 > 0.6$, $SD \downarrow$, $F \uparrow$, and $0.85 < m < 1.15$ (for test set regression)

✓ and R^2_0: $\frac{(R^2 - R_0^2)}{R^2} < 0,1$

✓ and AE, AAE ~ **0**

✓ $Q^2_{ext} > 0.5 / 0.7$

i- $$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{test} \left(Y_i - \bar{Y}_{training} \right)^2}$$

External Validation

Test set

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test} \left(Y_i - \hat{Y}_i \right)^2}{\sum_{i=1}^{test} \left(Y_i - \bar{Y}_{training} \right)^2}$$

ii-

✓. $r_m^2 = R^2 \left(1 - \sqrt{R^2 - R_0^2} \right) > 0.65$

✓. $\bar{r}^2_m$ (average between r^2_m for the y_{exp} *vs.* y_{calc} and the y_{calc} *vs.* y_{exp} regressions)

✓. $\Delta r^2_m \sim 0$

iii- $$CCC = \frac{2\sum_{i=1}^{n_{ext}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{extT}} (y_i - \bar{y})^2 + \sum_{i=1}^{n_{ext}} (\hat{y}_i - \bar{\hat{y}})^2 + n_{ext} (\bar{y} - \bar{\hat{y}})^2} > 0.85$$

iv- Careful analysis of *scatter plots* of Y_{calc} *vs.* Y_{exp}

Validation

Y- randomization *

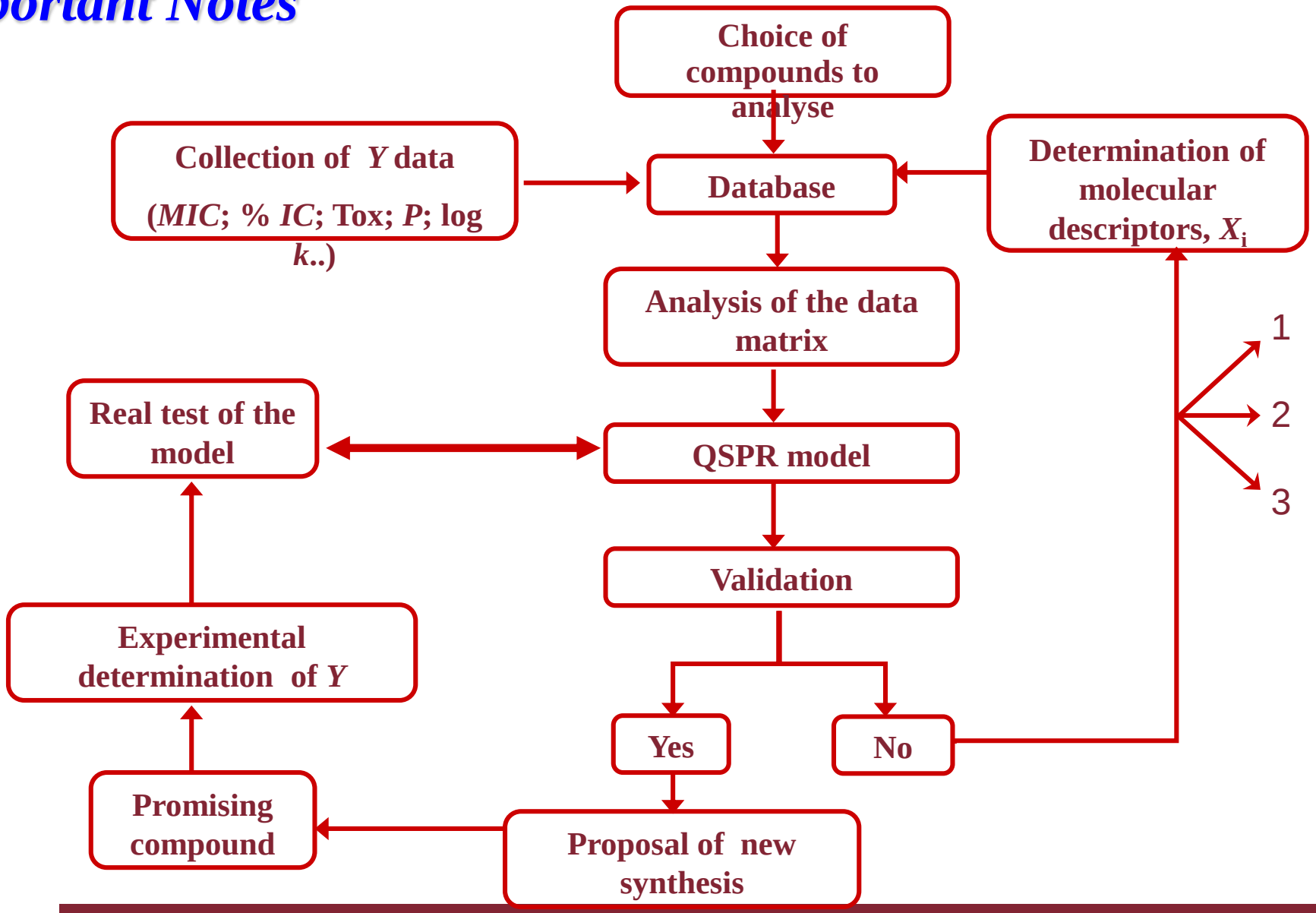
(to eliminate the possibility of "chance correlations" in "best" models)

If R^2 decreases significantly and RMSE increases significantly for the randomized models by comparison with the non-randomized model, that gives an indication of the robustness and reliability of the original non-randomized model.

- Todeschini's parameter

$$^{c}R_{p}^{2} = R\sqrt{(R^{2} - R_{r}^{2})} \quad > 0.5$$

Important Notes



Multiple linear regressions (MLR)

A reliable and robust **QSPR/QSAR** model must be:

- Statistically significant and robust √
- **Validated** for a set of data not used to develop the model √
- And its applicability domain should be well defined

Only QSPR/QSAR <u>validated</u> models may provide a sound mechanistic interpretation, which is particularly relevant for the development of new drugs

Important Notes

Robust and Predictive QSPRs/QSARs

⇓

- Reliable, homogeneous and representative data
- Relevant /meaningful descriptors
- Orthogonal descriptors (to avoid *redundancy*)
- Reduced number of descriptors (to avoid *chance correlations* & to facilitate interpretation)
- Detection and removal (and explanation) of outliers
- Suitable validation procedures (internal and external)

Important Notes

Robust and Predictive QSPRs/QSARs

⇓

- Well defined applicability domain*
- All equal, we should accept the simplest model (Ockham's rule)**
- Model consistency with physicochemical and/or biochemical nature of studied process
- Whenever possible, we should also make a **lateral validation** to understand the real structure-property relationship (*i.e.*, relate the new QSPR with other known, well-established and consistent QSPRs

Important Notes

Some MLR limiting conditions

- ✓ Compounds' number >> variables' number
- ✓ No intercorrelation between descriptors
- ✓ Variables without noise
- ✓ Continuous variables space
- ✓ Models only one *Y*

Some QSPR-MLR limiting conditions

- ✓ Non-observation of additivity and independence of descriptors
- ✓ Predictions (and interpretations) limited by the model's applicability domain
- ✓ Real structure-activity relations do not comply with the "simplicity" and linear nature of MLR models

Important Notes

"It is worth knowing if a QSPR model has the validated predictive power before it is applied to predict, let alone explain the SPR phenomenon of biological, pharmaceutical, environmental, or any other property of chemicals.(…) The philosophy of QSPR modelling is therefore: ***first validate and then explore*.**"

Tropsha, A., Gramatica, P., Gombar, V.K. *QSAR Comb.Sci.* **2003**, 22, 69-77

In fact only validated QSPR/QSAR models can provide a ***meaningful mechanistic interpretation***