INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0719

A REVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELS

LE BAILLY DE TILLEGHEM, C. and B. GOVAERTS

This file can be downloaded from
http://www.stat.ucl.ac.be/ISpub

A review of Quantitative Structure-Activity Relationship (QSAR) models.

C. Le Bailly de Tilleghem[0] B. Govaerts[0]

August 2007

1 History of QSAR models

Quantitative Structure-Activity Relationships (QSARs) are mathematical models approximating the often complex link between chemical properties and biological activities of compounds. They are used in many different fields such as agrochemistry, pharmaceutical chemistry or toxicology. This paper reviews the existing literature about QSAR modelling, more particularly concerning the statistical tools used to evaluate their quality. Those are quite basic statistical tools, not necessarily appropiate and not necessarily well applied in practice but, the aim of this paper is to present the state of the art about statistics in QSAR.

The development of a QSAR is based on the assumption expressed more than a century ago by Crum-Brown and Fraser (1868), that the physiological action of a substance is a function of its chemical composition. This leads to the idea that similar structures have similar biological properties and a small change in chemical structure is accompanied by a proportionally small shift in biological activity.

During the following hundred years, researchers tried to formalise some of those relationships. Richardson (1868) showed that the toxicities of ethers and alcohols were inversely related to their water solubility. Richet (1893) demonstrated a relationship between the narcotic[1] effect of alcohols and their molecular weight and Overton (1897) and Meyer (1899) independently showed that the narcotic action of many compounds was dependent on their oil/water partition coefficients[2]. Then, great strides were being made in the delineation of substituent effects on organic reactions, mainly by the work of Hammett (1937) and Taft (1952). Those contributions constitute together the mechanistic basis for the development of the QSAR paradigm in the 60's by Corwin Hansch, the first chemist that really succeeds in quantifying relationships between compound physicochemical properties and biological activities. His series of papers about the relationships of plant growth regulators and their dependency on Hammett constants[3] and hydrophobicity laid the foundation of QSARs (Hansch *et al.*, 1962, 1963, 1964; Fujita *et al.*, 1964).

Since then, with the increasing knowledge in chemistry and the exploding power of computers, researchers are developing and using more and more QSAR models. QSARs development has now became an important branch of chemometrics, the science of the application of mathematical or statistical methods to chemical data.

[0] Institute of Statistics from the Université catholique de Louvain – 20, voie du roman pays, 1348 Louvain-la-Neuve, Belgium – lebailly@stat.ucl.ac.be – Supported by the IAP research network grant no P6/03 of the Belgian Government (Belgian Science Policy).

[1] **Narcotic** refers to a variety of substances that induced sleep.

[2] The **oil/water partition coefficient** is a measure of the hydrophobicity (repelled by water) and hydrophilicity (able to bond with water) of a substance.

[3] **Hammett constants** are measures of electron-withdrawing or electron-donating effects.

There are two common objectives of QSARs (Eriksson *et al.*, 2003). A first goal can be to understand the link between structure and activity and extract clues of which chemical properties are likely determinants for the biological activities of a compound. To achieve this objective, the models should be transparent and as simple as possible. A second aim can be to allow prediction of biological activity for untested and sometimes yet unavailable compounds. In this case, a QSAR model should have a good predictive power, whatever its complexity.

The modelling technique to apply in QSAR development depends on the kind of available data. A distinction is done by Livingstone (1995) between a SAR (Structure-Activity Relationship) and a QSAR (Quantitative Structure-Activity Relationship) model:

- A SAR is an association between a chemical substructure and a biological activity, a kind of structural alert. As an example, the presence of a carboxylic group or an amino group in a molecule is known to impart skin corrosion potential.
- A QSAR is a mathematical relationship between a quantitative measure of chemical structure, or a quantitative measure of a physicochemical property, and a biological activity. Two types of QSARs can be distinguished: classification models for which the response variable is on a categorical scale, and regression models, for which the response is continuous.

All statistical methods to model the link between a quantitative or a qualitative response and a set of quantitative and/or qualitative explanatory variables can then be applied to QSAR data. The most popular techniques for quantitative endpoints are simple or multiple linear regression, principal components regression, partial least squares regression, as well as regression trees and neural networks. If the activity property is qualitative, discriminent analysis, decision trees or distance-based similarity analysis are often applied to model its relationship with the physicochemical explanatory variables (Jaworska *et al.*, 2003).

The types of statistical models used in QSAR as well as the techniques to assess their qualities depend on the nature of the response. In this paper, only quantitative activity endpoints are considered. The different kinds of activity responses and chemical explanatory variables are detailed in Section 2 and 3 respectively. The different modelling techniques as well as the selection of an explanatory variables subset are reviewed in Section 4.

The main area of QSAR application is the pharmaceutical industry (Tong *et al.*, 2004). QSAR is now an inexorably embedded tool in drug discovery, from lead discovery to lead optimisation (the background of the thesis) (Hopfinger and Tokarski, 1997; Kubinyi *et al.*, 1998). Indeed, a successful new drug requires many properties, such as a low toxicity, and an appropriate absorption, distribution, metabolism and excretion (*ADME*) profile. The chemical entity should also be easily and, if possible, cheaply synthesised. If a drug candidate fails one of those drug-like properties, it can never come to the market. As, once developed, QSAR models are easily and rapidely applied to predict the activity endpoint for any new compound (not yet synthesised) if its chemical structure is known, there is increasing use of QSARs early in the drug discovery process as a tool to virtually screen large chemical databases. On the basis of the QSAR predictions, chemists can eliminate from further development those chemicals lacking drug-like properties and construct a priority list with the most promising compounds for further testing.

Authorities, industries and other institutions also use QSAR models for assessing the risks of chemicals released to the environment and allowing their regulation. For example, in 1996, the Environmental Protection Agency[4] used QSAR modelling for developing and implementing a screening and testing program for chemicals that may disrupt the endocrine system. More than 87 000 chemicals were initially selected for evaluation. It was of course impossible to synthesise and test all those compounds. QSAR models were used

[4] The **Environmental Protection Agency** is an agency charged by the United States federal government with protecting human health and with safeguarding the natural environment.

to virtually screen this huge database and rank the compounds for priority testing (Tong *et al.*, 2003).

The great advantage of QSAR models is that predicting activity instead of measuring it allows to save time and money in chemical management and to speed up managerial decision. In addition, QSAR models can be used to reduce, refine, or replace the use of animals for an experimental purpose. To avoid animal testing is an important request of international associations (Cronin *et al.*, 2003a,b) and QSAR models as well as *in vitro*[5] tests constitute alternative methods.

As with any estimated statistical model, predictions provided by QSAR models are always uncertain. The current challenge is no longer in fitting a model that is statistically able to predict the activity within the chemical training set, but in developing a model with the capability to accurately predict the activity of untested compounds. The expected quality of a QSAR model depends on the application domain and the achieved goals. But there are growing international concerns to standardise criteria to measure the quality of a QSAR model with scientific basis.

The European Economic Community (1996) presented, in a technical guidance document, a general framework in which QSARs can be used within the risk assessment process. In this document, it is pointed out that QSAR models are only approximating methods and that it is important to perform further analysis to validate the QSAR models. In order to ensure transparency of any QSAR model, the European Commission suggested to provide minimal information as presented in Table 1.

A number of principles for assessing the validity of QSARs were proposed at an international workshop on the "Regulatory Acceptance of QSARs for Human Health and Environmental endpoints", organised by the International Council of Chemical Associations (ICCA)[6] and the European Chemical Industry Council (CEFIC)[7], held in Setubal, Portugal, on 4-6 March, 2002 (Worth, 2002). But the workshop did not produce any guidance on how to interpret and apply these principles. .

The Organisation for Economic Co-operation and Development[8] (OECD) started a program in 2003 in order to enhance the use of QSARs in the regulatory assessment of chemicals (report from the expert group on (Q)SAR, OECD, 2004). The expert group worked on the basis of the Setubal principles and added a check list to provide guidance on the interpretation of those principles and to encourage consistency in their application to individual QSARs.

According to the OECD principles for (Q)SAR validation, "a (Q)SAR should be associated with the following information:

(1.) a defined endpoint,
(2.) an unambiguous algorithm,
(3.) a defined domain of applicability,
(4.) appropriate measures of goodness-of-fit, robustness and predictivity and
(5.) a mechanistic interpretation, if possible."

The intent of Principle 1 (defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different

[5] An **in vitro** (Latin: "within glass") test is an experimental technique where the experiment is performed in a test tube, or generally outside a living organism or cell.

[6] The **International Council of Chemical Associations** is the world-wide voice of the chemical industry, representing chemical manufacturers and producers all over the world. ICCA has a central role in the exchange of information within the international industry, and in the development of position statements on matters of policy

[7] The **European Chemical Industry Council** is aimed to maintain and develop a prosperous chemical industry in Europe. CEFIC promotes the best possible economic, social and environmental conditions to bring benefits to society with a commitment to the continuous improvement of all its activities including the safety, health and environmental performance

[8] The **Organisation for Economic Co-operation and Development** is an intergovernemental organisation in which representatives of 30 industrialised countries in North America, Europe and the Pacific, as well as the European Commission.

General information	Reference: Process modelled: Domain of model:
Y-variable (dependent variable)	Species Type: Other information: Test method Experimental procedure: End-point modelled Type: Reliability: Data source: Units:
X-variable (independent variable)	Descriptors ♮ of initial descriptors: List of initial descriptors: descriptors: ♮ of final descriptors: List of final descriptors: descriptors: Data source: Other remarks:
Model	Samples: ♮ initial compounds: ♮ final compounds: Presentation of data: Design of training set: Outliers: Technique: Model Statistics r-squared: q-squared (x-val): External validation: ratio ♮ compounds / ♮ descriptors(initial): ratio ♮ compounds / ♮ descriptors(final): Validation: Range of validity:
Accuracy	
Remarks	

Table 1: Minimal information for an approach based on QSAR models. European Economic Community (1996)

experimental conditions. It is therefore important to identify the experimental system that is being modelled by the (Q)SAR.

The intent of Principle 2 (unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. The issue of reproducibility of the predictions is covered by this principle. The different kinds of QSAR models as well as the choice of entering descriptors and other complexity parameters are reviewed in Section 4.

The need to define an applicability domain (Principle 3) expresses the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. The applicability domain is the set of molecules for which the QSAR model is valide. If a new molecule is outside the applicability domain, the prediction provided by the QSAR model may be unreliable as this molecule is not similar to the ones used to train the model. The different definitions of QSAR applicability domain are reviewed in Section 8.

The wording of the Principle 4 (appropriate measures of goodness-of-fit, robustness and predictivity) is intended to point out the distinction between the internal performance of a model (goodness-of-fit and robustness) and the predictive power of a model (external validation). Such statistical tools are reviewed in Sections 6 and 7.

It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR (Principle 5), or that there can even exist multiple mechanistic interpretations of a given model. The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intent of Principle 5 is to ensure that some consideration is given to the possibility of a mechanistic association between the physicochemical properties used in a model and the endpoint being predicted, and to ensure that this association is documented.

The OECD provides also a check list to help the application of those principles. This check list is presented in Table 2. When developing a QSAR model, it is recommended to fill in this list and join it to the document describing the model to ensure transparency as done by Gramatica in the report from the expert group on (Q)SAR, OECD (2004).

This international effort to transparently validate QSAR models is a consequence of the increasing interest in developing and using those modelling techniques. As the Figure 1 shows, the number of QSAR models publications has exponentially increased in the past twenty years.

QSARs have been subject to a number of excellent reviews: Cronin (2000); Dearden *et al.* (1997); Hulzebos *et al.* (1999); Walker *et al.* (2002). For a review of QSAR modelling in virtual screening and data mining, see Oprea *et al.* (2005). For applications of QSARs in modelling mutagenicity and carcinogenicity, see Benigni (2003) and in modelling toxicity and fate, see Cronin and Livingstone (2004).

2 The responses modelled in QSAR

In most QSARs definitions, the kind of modelled responses are said to be any biological activity of the compounds. In this paper, some frequent biological endpoints are reviewed.

QSARs have been developed for a large number of toxic endpoints in both pharmaceutical and environment fields. There is a great interest in modelling the deleterious effect of a substance on organisms, individual organs, cells or plants. One of these most important acute toxic endpoints is the $LC50$, *i.e.* the concentration of a compound that causes 50% lethality of the animal in a test batch. Several countries have

CONSIDERATIONS	**YES/NO**
1) Defined endpoint • Does the model have a clearly defined scientific purpose? (i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?) • Does the model have the potential to address (or partially address) a clearly defined regulatory need? (i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)? • Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period) • Are the units of measurement of the endpoint given?	
2) Defined algorithm • In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents? • In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used3?	
3) Domain of applicability • In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals? • In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment? • In the case of a (Q)SAR, are the descriptor and response variables with inclusion and/or exclusion rules that define the variable associated ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
4) Internal performance • Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables? • If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):a) is there an adequate description of the data processing?b) are the raw data provided? • Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)? • Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?(e.g. r^2 values and standard error of the estimate in the case of regression models) • Is the QSAR associated with any statistics based on cross-validation or resampling? • If yes, is the number or samples used indicated?	
5) Mechanistic basis • In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region) • In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)? • Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	

Table 2: Check list in applying the OECD principles. report from the expert group on (Q)SAR, OECD (2004)

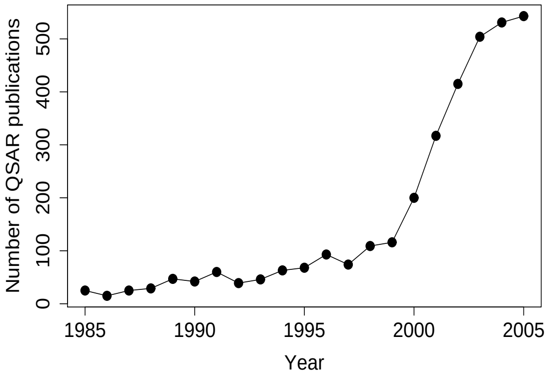

Figure 1: Number of references per year of publications which were retrieved using QSAR keyword with Pubmed (see http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed).

taken steps to ban the oral evaluation of $LC50$ on animals as it is judged cruel and sometimes unnecessary. As the OECD abolished the requirement of this oral test (Test guideline 401, 2003), alternative methods such as QSARs are required. For more details about toxic endpoints and measurements, see Cronin (2002) or Schultz *et al.* (2003)

To become a drug candidate, a compound should verify multiple other properties than being non-toxic. To be effective, a compound should optimise the crucial *ADME* properties. *ADME* stands for *Absorption*, *Distribution*, *Metabolisation*, and *Excretion*. Before a compound can become biologically active, it has to be taken into the bloodstream (*absorption*), to spread in the body, sometimes breaking barriers such as the blood-brain barrier, to finally arrive in the target organs or cells (*distribution*). Once the compound has worked its effect, it has to be broken down by biochemical reaction in the body (*metabolisation*) and the metabolised compound has then to be eliminated (*excretion*) in order to be not accumulated in the tissues.

In the past, a lot of drug investigations were failing in the last stages of the drug discovery and development process because of poor ADME properties. QSARs offer the opportunity to explore those properties even in the beginning of the process. That's why QSARs are increasingling used in pharmaceutical industries. For more details about QSARs predictions of ADME properties, see Selick *et al.* (2002).

The modelled endpoints can be of different nature: continuous, categorical ordinal or categorical nominal. The toxicity of a compound, for instance, can be continuously measured by its concentration that causes 50% lethality ($LC50$) as explained upper. The toxicity can also be an ordinal property like a compound is non-toxic, acceptably slightly toxic or too much toxic to become a new medicine. Finally, a response can be nominal as the fact that a compound can either break the blood-brain barrier or not. According to the nature of the response, the statistical tools applied to model it as well as the QSAR validation process are different. This paper only concentrates on continuous responses.

Whatever the type of response is, it should be clearly defined as recommended by the first OECD principle. The conditions in which the endpoint is measured should be detailed as well as the protocol(s) applied. This is an important requirement to ensure appropriate QSAR predictions by other users than the model builders.

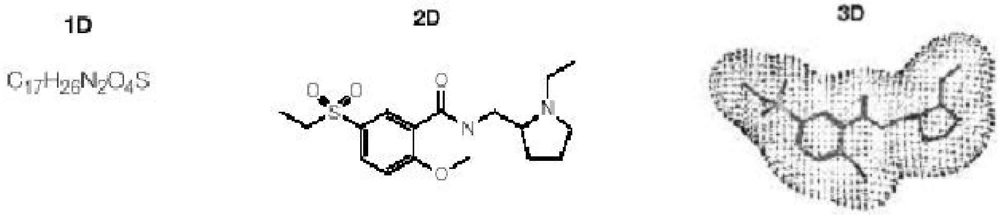

Figure 2: Number of dimensions taken into account in descriptors computation.

3 The explanatory variables entered in QSAR models

There are two approaches in modelling the activity-structure relationships: fragment-based QSAR and descriptor-based QSAR. The first approach consists in dividing each molecule of the database in all possible fragments (atoms or group of atoms) and modelling the activity of molecules only with the fragments counts. The second approach makes use of the whole structure of the compounds by modelling the activity as a function of chemical descriptors. Those descriptors could be simply the number of such atom, as well as properties characterising the links between atoms or the three-dimensional shape of the molecule. This paper focuses on chemical descriptors and the choice of descriptors that enter QSAR models.

Six years ago, Todeschini and Consonni (2000) have already listed more than 3000 molecular descriptors in their Handbook of molecular descriptors. Nowadays, around 8000 chemical properties can be used to describe a molecule. Descriptors are often classified according to the molecule dimensions taken into account as represented in Figure 2:

- 1D: one-dimensional linear representation of the molecule
  This class of descriptors contains simple indexes that can be deduced from the molecular formula, for instance, the number of such atoms or groups, the molecular weight[9] or the molar refractivity[10].

- 2D: two-dimensional planar representation of the molecule
  Two-dimensional descriptors regroup mainly topological and connectivity indexes. One of the most well-known 2D-descriptor is the octanol-water partition coefficient, K_{OW}, that is a measure of hydrophobicity[11] and hydrophilicity[12] of a substance.

- 3D: three-dimensional spatial representation of the molecule
  Three-dimensional descriptors summarise the geometry, the surface and the volume, as well as other electrostatic properties of the molecule such as the HOMO[13] or the LUMO[14]. The difference of the energies of the HOMO and LUMO can sometimes serve as a measure of the excitability of the molecule.

For more examples of descriptors, see (Livingstone, 2000; Todeschini and Consonni, 2000)

Some descriptors have to be measured in laboratories by performing experiences on the molecule, such as the acidity constant K_a. They are often referred to as *empirical descriptors*. These descriptors can be heavy to use in practice, especially if the goal of QSAR modelling is to speed up the drug discovery process. Hopefully, more and more descriptors can be computed exactly by applying new softwares such as ADAPT (Jurs, 2002; Stuper and Jurs, 1976), OASIS (Mekenyan and Bonchev, 1986), CODESSA (Katritzky *et al.*,

[9] The **molecular weight** can be calculated as the sum of the atomic weights of all the atoms constituing the molecule.

[10] **Molar refractivity** is a measure of the volume occupied by an atom or group

[11] **Hydrophobicity** is the property of a molecule that is repelled by water.

[12] **Hydrophilicity** is the property of a molecule that can bond with water.

[13] **HOMO** stands for Highest Occupied Molecular Orbital. It is the highest-energy orbital of a molecule with one or two electrons.

[14] **LUMO** stands for Lowest Unoccupied Molecular Orbital. It is the lowest-energy orbital with no electrons.

1994), and DRAGON (Consonni *et al.*, 2005) .

Those software can theoretically compute a set of descriptors for any molecule coded using a universal text representation called SMILES. The name *SMILES* stands for "Simplified Molecular Input Line Entry Specification". This universal text coding of a molecule was first introduced by Weininger (1988). It allows to specify the structure of chemical substances in a ASCII format that can be understood by most chemical softwares that compute descriptors. On Figure 3 are shown the SMILES coding for the acetic acid and the benzene molecules.

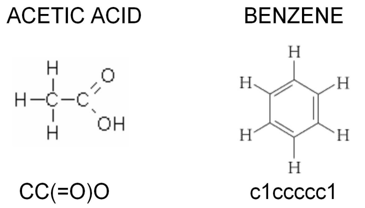

Figure 3: Two examples of SMILES coding: acetic acid and benzene molecules.

The nature of the descriptors may be either continuous (*e.g.* molecular weight), discrete (*e.g.* number of such atom), ordinal or nominal (*e.g.* solvent type). Continuous descriptors can be also transformed by centering or scaling, to ensure the same weight for each descriptor in the QSAR models. To avoid asymetric explanatory variables, it is not rare to work with the logarithm of descriptors. Other transformations are possible such as any power of the descriptors.

The number of available molecular descriptors is great. Allowing transformations and cross-products for interactions make the number of possible models exploding. And it is often frequent that the number of explanatory variables exceeds the number of observations, leading the failure of multiple linear regression models. With any kind of modelling techniques, using too many explanatory variables may cause overfitting and results in poor predictive power for new compounds. Not all available explanatory variables can enter the model and a choice between them has to be performed.

In practice, for some defined endpoint, important descriptors that influence this endpoint are known. For instance, it is well-known that the octanol-water partition coefficient, K_{OW}, influences greatly the $LC50$, *i.e.* the concentration of a substance that causes 50% lethality (report from the expert group on (Q)SAR, OECD, 2004). For those situations, one can hope to obtain good QSAR only with one or several of those few known descriptors. But, most of the time, the structure-activity relationship is much more complex, ambiguous and it is often necessary to use a greater number of chemical descriptors to be able to explain it. It is then necessary to apply descriptors selection methods to the huge set of available descriptors. But trying all the subsets of explanatory variables is impossible.

A first rapid screening can be made by discarding constant values and paired correlated variables that could lead to collinearity and unstability problems with multiple linear regression models. Then, there exists a number of various techniques to select a descriptors subset that will enter the regression models, such as stepwise regression, backward elimination, forward selection, simulated annealing, evolutionary and genetic algorithm. Subset variables selection constitutes the subject of a great part of QSAR papers (Kubinyi, 1994, 1996; Waller and Bradley, 1999; Whitley *et al.*, 2000). A recent comparison has given a demonstration of the advantages and success of genetic algorithm (Xu and Zhang, 2001).

Whatever is the subset variable selection method, the second OECD principle states that the descriptors choice should be transparent, saying which variables were first considered and finally which variables are kept in the model and why. Eriksson *et al.* (2003) noted that the choice of descriptors used is far more important than the specific modelling method employed. Nevertheless other methods than multiple linear regression are used in QSAR modelling, less sensitive to the correlation between the descriptors. The main used modelling methods are reviewed in the next section.

4 Different classes of QSAR models

The most common modelling techniques in QSAR to link the descriptors to the activity response are based on regression analysis. Among those techniques, Multiple Linear Regression (abbreviated *MLR*) is the most classical one. But, as explained upper, MLR is not adapted to the existing correlations between descriptors. Other alternatives exist. Multivariate projection methods to a subspace of orthogonal latent variables have became more and more popular in QSAR. Partial Least Squares Regression (*PLSR*) and Principal Components Regression (*PCR*) are two such projection techniques increasingly used. Another alternative, close to MLR, is the Ridge Regression (*RR*) who imposes a penalty to the size of the coefficient in the linear regression model. MLR, PLSR, PCR and RR are reviewed in details below. Other modelling techniques such as Regression Trees (*RT*), k-Nearest Neighbours (*kNN*) and Neural Networks (*NN*) are also briefly introduced.

The following notations are used :

- Y is the (random) activity response;
- $\mathbf{x}$ is a K-vector of (deterministic) descriptor variables (or any transformation of them, including power, cross-product, dummy variables encoding a categorical descriptor, constant term...);
- $\mathbf{x^c}$ and $\mathbf{z}$ are respectively the mean-centered and the standardised $\mathbf{x}$;
- $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ is the N-vector of observed responses;
- $\mathbf{y^c}$ is the mean-centered $\mathbf{y}$;
- $\mathbf{x_i}$ is the i^{th} observed K-vector of transformed descriptor variables ($i = 1, 2, \ldots, N$);
- $\mathbf{x_i^c}$ and $\mathbf{z_i}$ are respectively the mean-centered and the standardised $\mathbf{x_i}$ ($i = 1, 2, \ldots, N$);
- $\mathbf{X}$ is a $(N \times K)$ matrix where each row is an observed vector of transformed descriptor variables;
- $\mathbf{X^c}$ and $\mathbf{Z}$ are respectively the mean-centered and standardised matrix $\mathbf{X}$.

4.1 Multiple Linear Regression (MLR)

MLR assumes that the link between the activity response and the standardised descriptors is given by the equation

$$Y = \mathbf{z}'\boldsymbol{\beta} + \epsilon \tag{1}$$

where $\boldsymbol{\beta}$ is a K-vector of coefficients and ϵ is the random error term with zero mean, constant variance σ_ϵ^2 and independent from one observation to another.

This model is traditionally fitted to the data by least squares method, estimating the regression coefficients $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})$. If the inverse of the matrix $\mathbf{Z}'\mathbf{Z}$ exists, the solution of this minimisation problem is unique and depends explicitly on both observed responses and observed descriptors: $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. Using this least squares estimate, the predicted responses $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ are the orthogonal projections of the observed responses $\mathbf{y}$ on the space spanned by the columns of standardised transformed descriptors $\mathbf{Z}$.

MLR is very popular in QSAR modelling because of its simplicity and its well-known theoretical background (Cronin and Schultz, 2003; Schultz and Cronin, 2003). Through the estimated equation

$$\hat{Y} = \mathbf{z}'\hat{\boldsymbol{\beta}}\,, \tag{2}$$

the effect of each descriptor used on the response is understood as well as its amplitude. The hypothesis testing on the parameters allows to check if those effects are significant. The confidence interval and the prediction interval for the expected response and for the individual response respectively can be used to quantify the uncertainty of their estimations.

Concurrently to those advantages, MLR has its own limitations. A main drawback of MLR in the field of QSAR modelling is that it can not be used when the number of explanatory variables exceeds the number of observations. In addition, MLR assumes the predictor variables to be not linearly dependent, *i.e.* the rank of the matrix $\mathbf{Z}$ is K. If the data exhibit collinearity among the variables $\mathbf{z}$, the estimated regression coefficients get unstable. As a consequence, MLR can be used but a carefull selection within the set of available descriptors has to be performed as proposed in the previous section.

Examples of the application of MLR coupled with a variables selection by a genetic algorithm can be found in the report from the expert group on (Q)SAR, OECD (2004).

4.2 Principal Components Regression (PCR)

Principal Components Regression is an alternative to MLR when the explanatory variables are correlated (Draper and Smith, 1981; Myers, 1986). A Principal Components Analysis (*PCA*) is first applied to construct orthogonal latent variables, called the Principal Components (*PCs*), that can then enter a MLR model without any more collinearity problem.

The PCs are computed as linear combinations of the mean-centered descriptors, $\mathbf{t}_K = \mathbf{W}_K\mathbf{x^c}$, where the weight matrix $\mathbf{W}$ of dimension $(K \times K)$ is composed of the K eigenvectors of the covariance matrix of the observed mean-centered descriptors $\mathbf{X^c}$. One may use all the PCs (ordinary least squares results) as they are orthogonal or one may select only a few ones. A selection can be done by discarding the less important PCs and keeping only the $M < K$ first PCs such that the retained explained variance is large enough. The more the explanatory variables are correlated, the less PCs are necessary to recover most of the $\mathbf{X^c}$ information.

The M uncorrelated selected PCs, $\mathbf{t}_M$, are used to model the response:

$$Y = \mathbf{t_M}'\boldsymbol{\beta} + \epsilon = \mathbf{x^c}'\mathbf{W_M}'\boldsymbol{\beta} + \epsilon \tag{3}$$

where $\boldsymbol{\beta}$ is a M-vector of parameters that can be estimated using least squares. The corresponding predicted responses

$$\hat{\mathbf{y}} = \mathbf{T_M}(\mathbf{T_M}'\mathbf{T_M})^{-1}\mathbf{T_M}'\mathbf{y} = \mathbf{X^c}\mathbf{W_M}'(\mathbf{W_M}\mathbf{X^c}'\mathbf{X^c}\mathbf{W_M}')^{-1}\mathbf{W_M}\mathbf{X^c}'\mathbf{y} \tag{4}$$

are the orthogonal projections of the observed responses $\mathbf{y}$ on the space spanned by the M PCs. The effect of the original descriptors on the activity is not as direct as with MLR models. One has to inspect the scores and the loadings to interpret the PCs as a function of the original descriptors and analyse the most impacting descriptors.

While fitting a MLR model to explain the response by the M first PCs or all the PCs, one can then discard the PCs that have non-significant effect on the activity response or if there eliminations do not decrease the predictive power. Indeed, the PCs are only summarising the observed explanatory variables ignoring their potential link with the response. There is no guarantee that the M first PCs are the M best ones in predicting the response. Partial Least Squares Regression attempts to take also the response into account in the construction of the new regressors.

4.3 Partial Least Squares Regression (PLSR)

Partial Least Squares were first developed in 1966 in the field of econometrics by a Swedish statistician, Herman Wold (Wold, H., 1966). PLS Regression was finalised in 1982 by Herman Wold and his son, Svante Wold (Wold S. *et al.*, 1982). Svante Wold introduced this modelling technique in chemometrics (Wold and Dunn, 1983). Till that date, PLSR is increasingly used in QSAR modelling as it manages a great number of explanatory variables (even possibly greater than the number of available data) and the existing correlations between descriptors (Eriksson *et al.*, 2001). PLSR also does not assume that the explanatory variables are exact and 100% relevant for modelling the response. For all those reasons, PLSR is now far as popular as MLR.

The principle of PLSR is very similar to PCR as the link between the activity response and the descriptors is modeled through newly constructed latent variables. The specificity of PLSR is that those latent variables are recursively chosen to perform a simultaneous decomposition of the observed mean-centered descriptors $\mathbf{X^c}$ and the mean-centered observed responses $\mathbf{y^c}$ with the constraint that they explain as much as possible of the covariance between $\mathbf{X^c}$ and $\mathbf{y^c}$. More precisely, M latent variables are defined as linear combinations of the original descriptors, $\mathbf{t_M} = \mathbf{V_M}\mathbf{x^c}$, to model simultaneously $\mathbf{X^c}$ and $\mathbf{y^c}$:

$$\begin{cases} \mathbf{X^c}{=}\mathbf{T_M}\mathbf{P_M} + \mathbf{E} \\ \mathbf{y^c}{=}\mathbf{T_M}\mathbf{q_M} + \mathbf{f}. \end{cases} \tag{5}$$

$\mathbf{T_M}$ is the $(N \times M)$ matrix of scores and $\mathbf{P_M}$ and $\mathbf{q_M}$ are respectively the $(M \times K)$ matrix and M-vector of loadings.

An adequate recursive algorithm (Wold S. *et al.*, 1982; Martens, 1985; de Jong, 1993) is used to define the scores $\mathbf{T_M} = \mathbf{X^c}\mathbf{V_M}' = \mathbf{X^c}(\mathbf{W_M}'\mathbf{P_M})^{-1}\mathbf{W_M}'$ based on the singular value decomposition of $\mathbf{X^c}'\mathbf{y^c}$. As thoses scores are linear combinations of the original descriptors, the model for the response in (5) can be rewritten similarly to a MLR model:

$$\mathbf{y^c} = \mathbf{X^c}\mathbf{b_M} + \mathbf{f} \tag{6}$$

with the M-vector $\mathbf{b_M} = \mathbf{V_M}'\mathbf{q_M} = (\mathbf{W_M}'\mathbf{P_M})^{-1}\mathbf{W_M}'\mathbf{q_M}$.

Using the definition of the loadings $\mathbf{q_M} = (\mathbf{T_M}'\mathbf{T_M})^{-1}\mathbf{T_M}'\mathbf{y^c}$ provided by the PLS algorithm, the N responses are predicted by

$$\hat{\mathbf{y^c}} = \mathbf{T_M}(\mathbf{T_M}'\mathbf{T_M})^{-1}\mathbf{T_M}'\mathbf{y^c} = \mathbf{X^c}\mathbf{V_M}'(\mathbf{V_M}\mathbf{X^c}'\mathbf{X^c}\mathbf{V_M}')^{-1}\mathbf{V_M}\mathbf{X^c}'\mathbf{y^c}, \tag{7}$$

which is similar to MLR or PCR predictions. Those predictions can also be interpreted as the orthogonal projections of the observed responses $\mathbf{y^c}$ on the space spanned by the columns of the observed latent variables $\mathbf{T_M} = \mathbf{X^c}\mathbf{V_M}'$.

The problem of variable selection remains as in MLR modelling but PLSR solves the collinearity problem of the descriptors. A lot of different methods exist to choose the adequate number latent variables M. A quite often used technique in the field of QSAR modelling is the one of Baroni *et al.* (1993) called GOLPE (Generating optimal linear PLS estimations). It mixes a preliminary descriptors selection by means of D-optimal design in the loadings space and an iterative evaluation of the effects of the individual variables on the model predictivity.

The effect of original descriptors on the activity response is not as direct as with MLR models. But the scores and loadings can be used to interpret the latent variables as a function of the original descriptors and analyse the most impacting descriptors. For examples of PLSR applications to QSAR modelling, see for instance the work of Eriksson *et al.* (1997) or Eriksson *et al.* (2003)

4.4 Ridge Regression (RR)

Instead of selecting a variables subset and dropping descriptors from a MLR model, one can use a modification of the classical least squares MLR method, called Ridge Regression (*RR*). The assumed model of RR is exactly the same as the MLR model, $Y = \mathbf{z}'\boldsymbol{\beta} + \epsilon$. The difference stands in the least squares criteria that is penalised by a multiple of the sum of squared regression coefficients. For a fixed constant λ, the K least squares RR coefficients estimates are given by

$$\hat{\boldsymbol{\beta}}^{RR} = argmin_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta} \right\} = (\mathbf{Z}'\mathbf{Z} + \lambda I)^{-1}\mathbf{Z}'\mathbf{y}\,. \tag{8}$$

The activity response of a new molecule with transformed descriptors $\mathbf{x}$ is predicted by

$$\hat{y} = \mathbf{z}'\hat{\boldsymbol{\beta}}^{RR} = \mathbf{z}'(\mathbf{Z}'\mathbf{Z} + \lambda I)^{-1}\mathbf{Z}'\mathbf{y}\,. \tag{9}$$

This RR sum of squares is equivalent to the minimisation of the ordinary sum of squares under the constraint that the sum of the squared coefficients does not exceed a given size. This constraint protects the gradiants of the response surface in the direction of the smallest PCs in correlated $\mathbf{Z}$-space against potentially high variance. These directions are exactly the causes of trouble in MLR. That is how the additionnal penalty term has the effect to shrink the vector of coefficients and limits the two cited risks.

It is well-known that RR produces biased estimates of $\boldsymbol{\beta}$. RR modifies $\hat{\boldsymbol{\beta}}^{MLR}$ by introducing a bias, which decreases the variance of the estimates by more than the squared bias, so that the RR estimates of $\boldsymbol{\beta}$ have lower mean squared error (MSE) than $\hat{\boldsymbol{\beta}}^{MLR}$. More precisely, the ridge existence theorem (Vinod, 1981) shows that, for $0 < \lambda < \frac{2\sigma^2}{\boldsymbol{\beta}^T\boldsymbol{\beta}}$, $MSE(\hat{\boldsymbol{\beta}}^{RR}) < MSE(\hat{\boldsymbol{\beta}}^{MLR})$ where σ^2 is the residual variance. That's why RR performs pretty well in practice for prediction. Hastie *et al.* (2001) summarise: "For minimising prediction error, RR is generally preferable to MLR with variables subset selection, PCR and PLS. However the improvement over the latter two methods is only slight.".

For a comparison of RR, PLS and PCR in QSAR applications, see Hawkins *et al.* (2001).

4.5 Regression Tree (RT)

The principle of Regression Trees is really different from the previous regression methods (MLR, PLSR, PCR or RR) as the nature of the relationship between the activity response and the descriptors is not prespecified. RTs are nonparametric models.

The most simple technique for fitting a RT to QSAR data consists in recursively partitioning the data into successively smaller groups (called *nodes*) with binary splits based on a single descriptor (example: $x_j \leq c$ and $x_j > c$ for the j^{th} transformed descriptor and a constant c). At each step, splits for all of the descriptors are examined by an exhaustive search procedure and the best split is chosen. The idea of Breiman *et al.* (1984) is to define the best split at stage M as the one that minimises the total residual sum of squares over all the current M nodes, $\sum_{m=1}^{M}\sum_{i=1}^{N_m}(y_i - \hat{y}_i)^2$.

In the m^{th} node, the predicted activity $\hat{y}_i$ of the i^{th} molecule can be simply defined as the average value of the N_m observed activities of the compounds in that node. Indeed, this is the constant value that minimises the sum of squares. The activity response of a new molecule can be predicted by

$$\hat{y} = \sum_{m=1}^{M} \bar{y}_m I_m(\mathbf{x}) \tag{10}$$

where $\mathbf{x}$ is the K-vector of transformed descriptors, $I_m(\mathbf{x})$ is the indicator of the node (1 if the observation $\mathbf{x}$ belongs to the node m; 0 otherwise) and $\bar{y}_m$ is the average observed value at node m. One can also imagine to build local predictive model in each node of the tree, such as a MLR model, to obtain at the end a smoother regression surface.

The tree can recursively be splitted till all terminal nodes (called *leaves*) contain only one molecule. Techniques such as cross-validation can then be used to prune the overfitted tree to an optimal size (see Therneau and Atkinson (1997)).

The output is a tree diagram with the branches determined by the splitting rules and a series of terminal nodes that contain the mean response or a local model. An example of a regression tree with average response

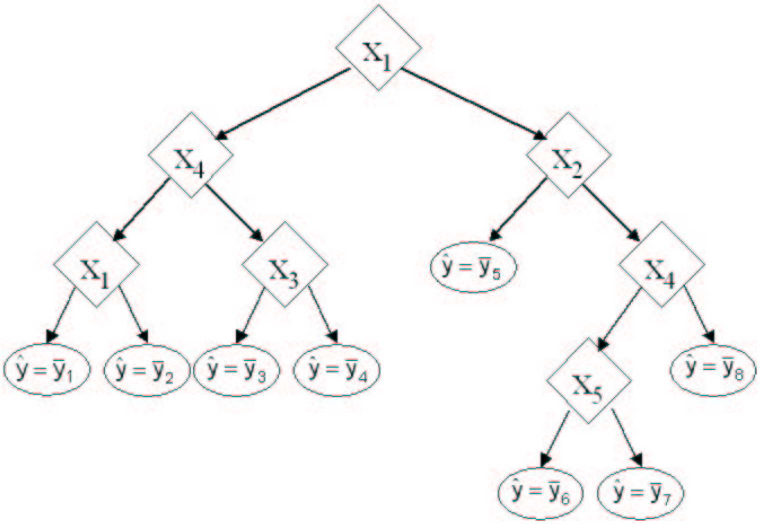

Figure 4: Schematic example of regression tree using the 5 explanatory variables x_1, x_2, x_3, x_4 and x_5 to recursively split the data set in 8 final leaves where the response is predicted by the mean response of observations belonging to that leave.

in the leaves is shown in Figure 4.

RT is a very flexible modelling technique. Indeed, there are various methods for growing the tree, allowing not only binary splitting, various criteria to optimise while splitting and also various tools for pruning (Hastie *et al.*, 2001). For all those variation of the presented basic RT, there is no *a priori* assumption about the kind of activity-structure relationship. RT is appreciated in QSAR applications as it can handle large data sets, it allows interactions and nonlinear relation between descriptors and response, and also because it produces sequences of prediction rules that are readily interpretable. Both continuous and categorical descriptors can be handeled when splitting.

For applications of RT to QSAR modelling, see Dzeroski (2001), Izrailev and Agrafiotis (2001) or Blockeel *et al.* (2004).

4.6 k-Nearest Neighbours (kNN)

The k-nearest neighbours method consists in predicting the activity of a molecule as the average (or weighted average) of the observed activity values of the k nearest molecules. kNN is greatly appreciated by QSAR practitionners as it is really intuitive. Indeed, its principle reflects the ancestral idea that similar compounds reveal similar activity property. Like RT, kNN does not make any *a priori* assumption about the nature of the activity-structure relationship. The descriptors are not directly used to model the activity but to define the neighbourhood. More precisely, for any molecule with observed descriptors $\mathbf{x}$, the predicted value is

$$\hat{y} = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i \,, \tag{11}$$

i.e. the average response of the k closest observations $\mathbf{x}_i$ in the neighbourhood $N_k(\mathbf{x})$ of $\mathbf{x}$.

Even if the kNN principle seems simple, its application in practice requires some crucial choices about the following parameters (Hastie *et al.*, 2001):

- Choice of k, the number of points belonging to the neighbourhood.
  The traditional rule of thumb suggests to adapt k to the number of observations: $k = \sqrt{N}$. One can

also use cross-validation to choose k to obtain the best bias-variance tradeoff.

- Choice of the distance between two observations.
  Some possible choices are the Euclidean distance, the Mahalanobis distance or the L_1-norm for quantitative descriptors and an indicator (Tutz, 1990) for qualitative descriptors (0 if the two observations share the same category, 1 otherwise). Those distances are precisely defined in Section 8.

- Choice of the weighting method.
  Computing the average of the observed activity in the neighbourhood provides the same weight to the k observations, the closest as the furthest observations of the neighbourhood. One can also weight the response values by a function of their distances or by using kernels.

Due to the curse of dimensionality, the kNN method performs very poorly in a high-dimensional descriptors space like it is often the case in QSAR modelling. Before applying the kNN method, it is necessary to perform a subset variables selection to eliminate irrelevant descriptors. Zheng and Tropsha (2000) proposed a variables selection based on simulated annealing to apply kNN to QSAR data. For another application of kNN to QSAR modelling, see Hoffman *et al.* (1999) or Baurin *et al.* (2002).

4.7 Neural Networks (NN)

The most common (Artificial) Neural Network model equation is of the form

$$f(\mathbf{x}) = G\Big(\sum_j (w_j g_j(\mathbf{x}))\Big) \tag{12}$$

where $\mathbf{x}$ is the K-vector of transformed descriptors. Like the neurons in the brain, the descriptors $\mathbf{x}$ (continuous or binary), the functions g_j and the function G are multiple layers that are interconnected to produce the prediction of the activity response by $f(\mathbf{x})$. When the number and the shape of functions g_j (often logistic functions) and the function G (often $tanh$) have been fixed, the weights w_i are estimated to minimise the squared prediction error. The gradient-based backpropagation method can achieve this task (Werbos, 1974) .

NN models are nonlinear models that can be represented as a network structure as depicted in Figure 5. There exist different architectures of NN according to the number of hidden layers, the number of nodes for each layer and the connections existing between all the different layers. The two main forms of NN are the feedforward NN and NN using radial basis functions.

Neural Networks are described with more details in Bishop (1995). For examples of QSAR Neural Network applications, see (Andrea and Kalayeh, 1991), (Devillers, 1996), (Duprat *et al.*, 1998), (Kovesdi *et al.*, 1999), (Manallack and Livingstone, 1999) or (Huuskonen *et al.*, 2000)

4.8 Choice of complexity parameters

All the presented models depend on some complexity or tuning parameters: the number of explanatory variables (descriptors) in MLR, the number of latent variables kept in PLSR or the number of PCs in PCR, the penalty parameter λ in RR, the level of pruning in RT, the number k of neighbours in kNN or the number of layers or nodes in NN. The idea is to try different values for those parameters and select the value that yields to the best model performance.

If the criteria measuring the model performance is only based on the fit to the training set (such as the squared error loss summarising the difference between the observed responses and the fitted values), then there is a clear risk of overfitting. Jaworska *et al.* (2005a,b) suggest to use other analytical criteria (Mallow Cp statistic, Akaike Information Criterion, Bayesian or Schwartz Information Criterion) or sample re-use techniques (Y-scrambling, cross-validation or bootstrap).

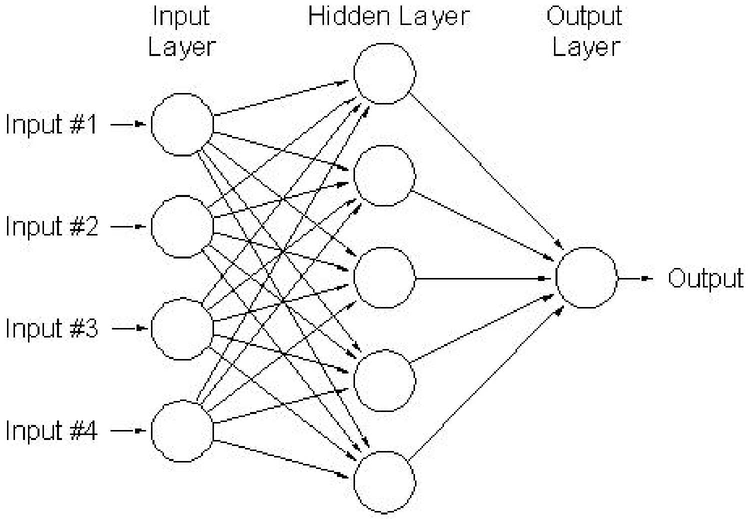

Figure 5: Schematic example of feedforward Neural Network (from the website *http* : *//smig.usgs.gov/SMIG/features*$_0$902*/tualatin*$_a$*nn.fig*3*.gif*).

With the recent computer advances, it's no more expensive to fit the data with a selected modelling technique using different complexity parameters. It is even also recommended to try different modelling technique and to choose the most performing one (Eriksson *et al.*, 2003).

5 Data collection and model fit

The biological data to which modelers have access are very limited in terms of chemical and biological space. Chemists often have to collect their own data for QSAR modelling. There is a lack of international dash to gather biological database, with adequate explanation of the way they have been obtained. This is one of the goal of the OECD principles.

The selection of the training set is crucial whatever is the field of application. The intuitive idea is to collect data similar to the data for which predictions will be made with the fitted QSAR model. The objective is to select molecules that well span the chemical domain of interest.

In theory, design of experiments should be used to generate the training set to ensure maximum variance in the descriptors space and produce a representative training set. But the powerful tools of experimental design can not directly be applied to construct a training set. Indeed, when a molecule is even not yet synthesised, the corresponding descriptors can be computed thanks to its SMILES coding, but the opposite is not possible as no chemist is able to construct a molecule with such given values of descriptors.

In many QSAR papers, nothing is said about the choice of the molecules in the training set. Most of the time the only remark is that the QSAR is fitted on molecules having the same mechanism of action (for instance, all esters, all antioxidant, all disinfectant,...). This ensures the adequacy of a unique QSAR to model the activity response of those molecules because a "universal" QSAR able to model chemicals with significantly different structures and modes of action may not seem plausible. A set of such predefined molecules is often called a *library* in chemical jargon.

A chemical library may be composed either of compounds having the same mechanism of action or compounds obtained by systematic reactions of a small number of starting compounds with a larger number of reagents (*combinatorial library*). The chemist may choose as a starting data set such a library. Often, as the compounds are not yet synthesised, the collection of compounds is referred to as a *virtual library.*

In practice, if this virtual library contains too much compounds to synthesise, the statistician may help the chemist to select a sub-library of reasonable size with diverse molecules. The statistical techniques to select a subset of compounds from a greater starting set is called multivariate design (Wold S. *et al.*, 1986) or statistical molecular design (Eriksson *et al.*, 1996; Linusson, 2000). Eriksson *et al.* (2000) proposed to apply fractional factorial designs and De Aguiar *et al.* (1995) proposed to apply D-optimal designs to obtain a representative data set from a virtual library. Those techniques of experimental design (Box *et al.*, 1978) are based on the descriptors computed for the compounds in the virtual library or on summaries of those descriptors such as the latent variables obtained with PLS or the principal components obtained with PCA.

In the first OECD principle, it is recommended to use the same assay protocol to obtain high quality data and to explain this protocol with the QSAR equation. Some preliminary analysis can be performed on the collected data to ensure their quality before applying the modelling techniques. One can first check if the standard deviation of experimental error is constant over the domain of interest. This is possible if replicates have been tested but this is not always the case.

When the quality of the collected data is insured, a QSAR model can be fitted. The data used to fit the QSAR model is called the *training set.* One can use the PLS or PCA techniques to project the training set in a space of smaller dimensions to analyse their distribution, looking for outliers or clusters. Compounds are often gathered in clusters and it has to be decided if a QSAR model for each cluster may be better than a global model on all data. The necessary homogeneity of the data and the absence of outlier can also be graphically analysed for any kind of models by plotting the observed responses (y_i, $i = 1, \cdots, N$) against the predicted ones ($\hat{y}_i$, $i = 1, \cdots, N$) or the residuals ($e_i = y_i - \hat{y}_i$, $i = 1, \cdots, N$) against the data numbers (i).

Different other tools can be used to detect outliers in the training set according to the QSAR model class. The standard instrument for MLR models is the *leverage* (Atkinson, 1985) that measures the distance from one molecule to the whole training set in the descriptors space. It is traditionally computed for the i^{th} compound of the training set as the i^{th} diagonal element of the Hat matrix

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' . \tag{13}$$

An observation with a leverage higher than $2K/N$ for great samples (Belsley *et al.*, 1980) or $3K/N$ for small samples (Vellman and Welsch, 1981) may be considered potentially as an influential point and the effect of removing such an observation from the training set has to be analysed. The leverage can be computed for any kind of model but is more meaningful for MLR and RR.

The *standardised residual in prediction* is another tool to detect outliers with MLR models. It is defined for the i^{th} observation in the training set as

$$e_i^{SRP} = \frac{e_i^{CV}}{\hat{\sigma}\sqrt{(1 - h_{ii})}} \tag{14}$$

where e_i^{CV} is the difference between the observed i^{th} response and its prediction using the model fitted on the rest of the data, $\hat{\sigma}$ is an estimate of the standard deviation of the residuals and h_{ii} is the leverage of the i^{th} observation. Molecules with a standardised residual in prediction greater than 2 or 3 are considered as outliers.

Leverages and standardised residuals in prediction can be simulateously represented on a graph called *William plot* to visualize outliers and define the applicability domain of the MLR model, as detailed in the sub-section 8.5.

For other classes of models, other distances (Euclidean, Mahalanobis, L_1, $DModX$, Hotteling's T^2) can be defined to measure the distance of a molecule to the training set and detect outliers. They are reviewed in section 8 on the applicability domain.

When outliers and influential points have been removed from the training set, further analyses have to be performed to assess the validity of the model. There is three main points in the validation of a QSAR model: the analysis of the model performance on the training set and on an external data set (OECD Principle 4), and the definition of its applicability domain (OECD principle 3), *i.e.* the group of compounds for which the QSAR model predictions are valid. A review of the QSAR literature about those three validation steps constitute the three next sections.

6 Internal model performance

There is an increasingly international interest to validate QSAR models (Cronin and Livingstone, 2004; report from the expert group on (Q)SAR, OECD, 2004). OECD principle 4 intends to make the distinction between the internal performance of a QSAR model characterised by goodness of fit and robustness measurements and its external performance characterised by its predictive power for new chemicals. The aim of this section is to review internal performance criteria for the practical application of OECD principle 4.

6.1 Validation of underlying assumptions

The assumptions underlying the QSAR model must be validated. In regression analysis, the residuals are often assumed to be independently normally distributed with zero mean and constant variance. To validate the normality assumption, an histogram or a normal probability plot of the observed residuals e_i's can be made. A graph of the observed residuals (e_i) against the predicted values ($\hat{y}_i$) validates the other undelying assumptions of the regression model (MLR, RR, PLSR and PCR).

6.2 Goodness of fit criteria

An intuitive idea is that a good model should provide predicted values similar to the observed ones. This can be visualised for any kind of models by plotting the observed responses (y_i, $i = 1, \cdots, N$) against the predicted ones ($\hat{y}_i$, $i = 1, \cdots, N$). The most well-known index for measuring the alignment of those points $(y_i, \hat{y}_i)$ is the coefficient of determination

$$R^2_Y = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \,. \tag{15}$$

This coefficient, bounded by 0 and 1, represents the percentage of the variation of the observed activity endpoint explained by the model. The higher it is, the better the fit is. This coefficient was introduced in the field of MLR but can be computed for any other kind of model with the same interpretation.

The main drawback of R^2_Y is that entering new explanatory variables in the model always increases its value. According to the parsimony principle, not too many variables should be entered in the model, especially if the number of observations is small. With MLR models, the R^2_Y is adjusted to take the number of parameters, K, into account as well as the number of observations, N:

$$R^2_{Y_{ADJ}} = R^2_Y - \frac{K-1}{N-K}(1 - R^2_Y) \,. \tag{16}$$

$R^2_{Y_{ADJ}}$ is upper-bounded by 1 but can decrease if the contribution of an additional variable is less than the impact on the degrees of freedom. The higher it is, the better the model is.

Other indexes are used with MLR models in the QSAR literature to compare the observed and the predicted data such as the standard error of estimate, s, or the Fisher statistics, F. The standard error of estimate, s, is the classical estimator for the residual standard deviation in MLR:

$$s = \sqrt{\frac{1}{N-K}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}\,. \tag{17}$$

If the model fits well the data, s should be small. If replicates have been performed, the experimental variance can be estimated and the estimated residual variance s^2 should be of the same order otherwise, this indicates that the model suffers from a lack of fit. The Fisher statistics, F is used in MLR analysis to test the global utility of the model by comparing the part of the response variation that is explained by the model to the part that left unexplained:

$$F = \frac{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2 - \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{K-1}}{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-K}} = \frac{\frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{K-1}}{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N-K}} = \frac{\frac{R_Y^2}{K-1}}{\frac{1-R_Y^2}{N-K}}\,. \tag{18}$$

The higher is F, the better is the fit of the model. The F statistics can be compared to the 95% quantile of the Fisher distribution, $F^{0.95}_{K-1,N-K}$, and if $F > F^{0.95}_{K-1,N-K}$, the MLR model is significantly useful.

With other QSAR models classes, such criteria can be generalised by an adequate measure of model complexity (K for MLR). With PLSR, the complexity can be measured by the number of latent variables and with RT, by the number of nodes.

To summarise, the higher are the indexes R_Y^2, $R^2_{Y_{ADJ}}$ and F, or the smaller is the residual standard deviation s, the better is the fit of the data by the model. But the model with the best fit is not surely the best model for prediction. If the model is changed to improve too much the fitting, the model will explain also the noise contained in the observed activity responses y_i's. A such overfitted model, even if it seems to be very good, may be useless to predict the activity response for new molecules not included in the training set. As the objective of the QSAR models developer is to apply them for the activity prediction of new compounds, even not yet synthesised, the developer should find a model with a maximised predictive power and not only with good fitting criteria. The predictive power of a model can be measured either on the training set (next subsection) or on an external set (next section).

6.3 Robustness and internal predictivity

The most important method in QSAR modelling to quantify the predictive power on the basis only of the training set is the *cross-validation* technique. The principle of cross-validation is to simulate predictions for new molecules not used in the fitting of the model. The training set is divided in distinct subgroups. A series of models are fitted on reduced datasets constructed by omitting each subgroup in its turn from the training set and each fitted model is applied on the left subgroup for predicting the Y response. In this way, a prediction is associated to every molecule of the original training set using a model trained on other data.

Let's denote $\hat{y}_i^{CV}$ the prediction of the activity response of the i^{th} molecule obtained by cross-validation. The classical index to summarise the internal predictive power is the cross-validated R_Y^2 defined as

$$Q_Y^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i^{CV})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}\,. \tag{19}$$

This is a cross-validation estimate of the percentage of the variance of the activity response explained by the model. The higher Q_Y^2 is, the more predictive the model is. The value of Q_Y^2 must be compared to the value of R_Y^2 (see equation (15)). If they are not of the same order and the difference $R_Y^2 - Q_Y^2$ is larger than 0.2 or 0.3, this indicates that the model overfits the training data or that there are some outliers (Eriksson *et al.*, 2003).

There is two main practices in defining the subgroups. The *leave-one-out* method considers each compound as a subgroup. Each compound is omitted in its turn. The model is fitted on the $N - 1$ other observations of the training set and the activity response of the remaining compound is predicted using the obtained model. To simulate predictions for new molecules not used in the fitting of the model, the leave-one-out principle is the most intuitive idea as the most information as possible is taken into account in the training of the model by removing only one observation. But leave-one-out is too optimistic and tends to overestimate the predictive power for independent new compounds.

The alternative is to delete more than one compound at each turn by defining subgroups with up to 50% of the data set. This method is called *leave-many-out.* The reasonable number of compounds to be included in each subgroup depends on the sample size. Let's denote G the number of groups, generally selected between 2 and 10. $\frac{N}{G}$ is the size of groups. Every group in its turn is left out of the training set and the model is fitted on the $N - \frac{N}{G}$ other compounds. This fitted model is then applied to the remaining group to predict the responses of its $\frac{N}{G}$ compounds.

Another technique to quantify the predictive power is the *bootstrap.* N compounds are drawn at random with replacement from the original dataset. Some of the original molecules might appear more than once in the resample while other molecules might not be included. The model is fitted on this resample and applied on the remaining compounds to predict the endpoint. Let's denote $\hat{y}_i^{(b)}$ the prediction of the activity response of the i^{th} molecule obtained with this bootstrap model. An estimate of the percentage of the variance of the activity response explained by the model is

$$Q_Y^{2\,(b)} = 1 - \frac{\sum_i (y_i - \hat{y}_i^{(b)})^2}{\sum_i (y_i - \bar{y}^{(b)})^2} \,. \tag{20}$$

This resampling is repeated a great number of time. Then the predictive power can be summarised by averaging the $Q_Y^{2\,(b)}$'s of each bootstrap model.

A final tool to insure that the model has not been obtained by chance correlation is the response permutation testing or Y-scrambling (Eriksson *et al.*, 2003). The N-vector of observed responses is permuted at random and the model is fitted on these permuted responses keeping intact the matrix of observed descriptors $\mathbf{X}$. The two indexes R_Y^2 and the cross-validated Q_Y^2 are computed. This permutation procedure is repeated a great number of times providing reference distributions of both R_Y^2 and Q_Y^2. If the R_Y^2 and Q_Y^2 of the initial model are higher than the values of the simulated distributions, this constitutes a stong indication that the initial model correlation was not obtained by chance.

The Q_Y^2, either computed by cross-validation (19) or by bootstrapping (20), is the most used index to quantify the internal predictive power of a QSAR model. Using leave-one-out cross-validation, Eriksson *et al.* (2001) suggested that $Q_Y^2 > 0.5$ is good and $Q_Y^2 > 0.9$ is excellent. But Golbraikh and Tropsha (2002) recommended to beware of Q_Y^2, especially if it is computed by leave-one-out cross-validation. Indeed, the only thing that is sure is that a small value of Q_Y^2 is the sign of a poor predictive power. But a high value of Q_Y^2 does not imply automatically a high predictive power for new compounds not used in the training set. The right way to estimate the predictive power is to test the model on a sufficiently large external data set.

7 External performance

7.1 Global external performance

According to Doweyko (2004), only about half of the QSAR models published in the last decade made reasonable predictions about test compounds not used to create the model. Tropsha *et al.* (2003) pointed out that the quality of QSAR models is often typically measured on the training set alone, but this approach does not necessarily generate good predictive QSAR models. The most demanding way to quantify the predictive power of a QSAR model should be based on an external data set, by making predictions for an independent set of data, not used in the model calibration. This provides a more rigorous evaluation of the model predictive capability for untested chemicals than cross-validation or bootstrap on the training set.

An external data set may be found in the QSAR literature but this is quite rare. Indeed, such data should concern the same endpoint, measured with the same protocol and the whole data set with all the computed descriptors should be known. This kind of data is often not publically available. If it is, there is always a risk of interlaboratory and assay variability among the different data sources.

If no similar data can be found in the QSAR literature, an external data set can be obtained by splitting the original data into a training set and a test set. The training set is the set of data which is used to construct the QSAR model and the test set is the set of data which is used to validate the QSAR model. The best splitting of the original data set can be found by an experimental design procedure (Marengo and Todeschini, 1992; Eriksson *et al.*, 2000).

Whatever the origin of the external set is, it has first to be checked that this external set is composed of compounds similar to the training data (*cfr* Section 8 on applicability domain) and contains no outlier. Then the estimated model can be applied to predict the activity response of each of the N' molecules in the external set. Those predictions ($\hat{y}'_i$, $i = 1, \cdots, N'$) are finally compared to the observed responses (y'_i, $i = 1, \cdots, N'$) graphically and using

$$R_Y^{2\,(ext)} = 1 - \frac{\sum_{i=1}^{N'} (y'_i - \hat{y}'_i)^2}{\sum_{i=1}^{N'} (y'_i - \bar{y}')^2} \,. \tag{21}$$

Setting aside an external data set is not reasonable when the number of available data is small. If no similar data can be found in the QSAR literature, only internal predictivity can be assed using cross-validation or boostrap technique. But this overestimates the predictive power for new molecules.

Like any other kind of models, a QSAR model is never better than the series of measurements it was obtained from. As Aristotle was saying: "It is the mark of an instructed mind to rest easy with the degree of precision which the nature of the subject permits and not to seek an exactness where only an approximation of the truth is possible." This means that, whatever is the degree of validation of the model, the QSAR user has to keep in mind that it is never possible to predict the activity response of an unknown chemical with absolute certainty. This is then crucial to measure the uncertainty of the predictions provided by QSAR models and verify that this uncertainty is adequate for the purpose of the QSAR application.

7.2 Local external performance

Once the predictive power of the QSAR model has been assessed on an external test set, it can be used to predict the activity of a new compound. Only a small part of QSAR papers mention the uncertainty of such pointwise prediction. This subsection reviews this rare QSAR literature on prediction uncertainty for a given new compound.

For MLR, a 95% confidence interval for the expected activity value of the new compound $E[Y|\mathbf{x}]$ is centered around its estimated value $\mathbf{z}'\hat{\boldsymbol{\beta}}$ and is constructed as $\mathbf{z}'\hat{\boldsymbol{\beta}} \pm t_{N-K}^{0.975}\sqrt{s^2\mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}}$. A 95% prediction interval for the activity of the new compound $Y|\mathbf{x}$ is centered around its predicted value $\mathbf{z}'\hat{\boldsymbol{\beta}}$ and is constructed as $\mathbf{z}'\hat{\boldsymbol{\beta}} \pm t_{N-K}^{0.975}\sqrt{s^2(1+\mathbf{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z})}$.

In the report from the expert group on (Q)SAR, OECD (2004), the standard deviation of the residuals of the external set $s^{(ext)} = \sqrt{\frac{1}{N'-K}\sum_{i=1}^{N'}(y_i - \hat{y}_i)^2}$ is computed and used to construct a 95% prediction interval for the activity response y_0 of any new molecule as $\hat{y}_0 \pm z_{0.975} \cdot s^{(ext)}$. This seems not correct as the standard error of prediction is considered constant over the chemical domain although it is certainly greater at the edges than in the center of the experimental domain. But no other QSAR paper evokes the subject of prediction uncertainty of a new particular molecule.

8 Applicability domain

8.1 Definition of the applicability domain

Wold and Dunn (1983) have shown that QSAR models normally have local validity only. They can embrace only compounds with similar chemical and biological properties as the compounds in the training set. Whereas excellent fit to the training data may be attainable, often QSAR models fail to predict accurately chemicals that differ substantially from the training set molecules. In order yo state clearly this intrinsec limitation of any QSAR model, the OECD recommended, in the report from the expert group on (Q)SAR, OECD (2004), to define the applicability domain of the QSAR model. This way, any chemist that wants to use a QSAR model for the prediction of a new compound with descriptors $\mathbf{x_0}$ is able to first check if it is included in its applicability domain and the corresponding prediction is reliable.

The applicability domain is the group of substances for which the QSAR model is valid. Any new compound that lies in the chemical space beyond this boundary may possess a different structure-activity relationship than the molecules in the training set and the prediction provided by the QSAR model may be unreliable.

The applicability domain can be defined regarding the descriptors space and/or the response as well as the chemical class, the mechanism of action or the species considered. In practice, an applicability domain is defined for the descriptors space and another applicability domain is defined for the response. The QSAR model can then be applied for a new molecule if it is included in the two applicability domains.

Simple tools such as the descriptors ranges can be used or more complex tools such as the convex hull, the leverages and the William plot, some other distance measurements (Euclidean, Mahalanobis distance or DModX), the Hotteling T^2 or density measurements. Their exact definitions can often become complex due to the highly multivariate nature of the data.

The different definitions are given below and applied on a simple QSAR data set provided by Veith and Mekenyan (1993). A MLR was developed to model the $96h\ LC_{50}$ of organic chemicals to a fish species, the fathead minnow (*Pimephales promelas*). The $96h\ LC_{50}$ is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas* after an exposure of 96 hours. Two descriptors are used to model the $96h\ LC_{50}$: K_{OW}, the octanol-water partition coefficient and E_{LUMO}, the energy of the lowest unoccupied molecular orbital. The training set is composed of 114 observations and a MLR model was fitted on those data, transforming the response and the first descriptor with the logarithm function: $\hat{log}(LC_{50}) = -2.414 + 0.579log(K_{OW}) - 0.473E_{LUMO}$.

8.2 Applicability domain based on the ranges

The most simple definition for the applicability domain can be achieved using the ranges of the descriptors. The range of a variable is, by definition, the difference between the observed maximum and minimum.

Based on this idea, a new compound is considered to belong to the applicability domain if its computed descriptors $\mathbf{x_0}$ are within the minimum and the maximum of each descriptor. If the QSAR model uses only one descriptor, the applicability domain reduces to a simple interval. With two descriptors, it is a rectangle and with $K > 2$ descriptors, the applicability domain is a K-dimensional hyper-rectangle.

This definition of the applicability domain takes the descriptors into account only separately. The minimum and the maximum of each descriptor are listed and sometimes also an histogram of observed values is constructed for each descriptor. But the applicability domain should take into account the overall distribution of the data in the structure space. In addition, this definition relies on the assumption that the data are uniformly distributed. This is not always the case with QSAR data as observations are often clustered and the K-dimensional hyper-rectangle defined with the ranges contains often empty spaces, especially in the edges.

An alternative consists in applying first PCA and then defining the applicability domain on the ranges of the principal components scores. The resulting hyper-rectangle may also include empty spaces, but the volume enclosed will be most of the time less empty than in the original descriptors ranges case.

8.3 Applicability domain based on the convex hull

To insure no extrapolation, the convex hull of the descriptors in the training set can be computed. The convex hull is defined as the smallest convex polyhedra that contains the data. If a new molecule with descriptors $\mathbf{x_0}$ is outside this boundary, the prediction may be unreliable.

Efficient algorithms for convex hull calculation are available for dimension 2 or 3 (Graham, 1972; Jarvis, 1973). But the complexity increases rapidely in higher dimension. Moreover, this approach does not consider the global distribution of the data as the convex hull definition is only based on the boundary of the data set. This does not prevent from empty spaces in the convex hull.

8.4 Applicability domain based on the Euclidean, Mahalanobis or L_1 distance

The intuitive idea of distance-based applicability domain is that a compound too far from the compounds in the training set may be badly predicted. The distance of a new molecule to the training set may be computed as its distance to the mean of the elements of the training set, the averaged distance between this new molecule and all the molecules of the training set or the maximum distance between the new molecule and all the molecules of the training set. A definition of the distance between two substances is, for instance, already used with the kNN modelling technique. Each compound is represented by a K-vector of observed transformed descriptors $\mathbf{x}$. The distance between two substances with descriptors $\mathbf{x_1}$ and $\mathbf{x_2}$ can be defined using the Euclidean distance (d^E), the Mahalanobis distance (d^M) or the L_1-distance (d^{L_1}) defined as follows:

$$d^E(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{\sum_{k=1}^{K} (z_{1k} - z_{2k})^2}$$

$$d^M(\mathbf{x_1}, \mathbf{x_2}) = (\mathbf{x_1} - \mathbf{x_2}) S^{-1} (\mathbf{x_1} - \mathbf{x_2})'$$

$$d^{L_1}(\mathbf{x_1}, \mathbf{x_2}) = \sum_{k=1}^{K} |z_{1k} - z_{2k}|$$

As S denotes the covariance matrix of the observed transformed descriptors, the Mahalanobis distance takes into account the correlation between the transformed descriptors which is not the case with the two other definitions. For Euclidean and L_1 distance, it is necessary to standardise the descriptors ($\mathbf{z}$) otherwise the descriptors with wider scales would have higher weights in the distance calculation. For those two distances, it is also recommended to apply PCA and measure the distances in the principal components space where new variables are uncorrelated.

A new molecule with descriptors $\mathbf{x_0}$ is considered outside of the applicability domain of the QSAR model if its distance to the molecules of the training set exceeds a certain threshold fixed by the QSAR developer. This threshold can be defined as, for instance, the largest distance of any molecule in the training set to the rest of the molecules.

The Euclidean and Mahalanobis distances are specially adapted for observations that are normally distributed. The L_1 distance assumes that the observations are uniformely distributed. Any of those three distances take into account possible empty regions in the structure space.

8.5 Applicability domain based on leverages

The leverage is another measure of the distance between a new molecule and the molecules in the training set. As explained in Section 5, the leverage is originally defined in the MLR literature to detect outliers in the training set. This definition can be extended to any new molecule with the standardised descriptors contained in the K-vector $\mathbf{z_0}$ by $h_0 = \mathbf{z_0}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z_0}$.

The leverage of each molecule in the data set can be represented on a graph referred to as the William plot (report from the expert group on (Q)SAR, OECD, 2004). The William plot is a scatterplot representing the leverages and the standardised residuals in prediction as defined in Section 5. This graph allows to detect outliers and defines the boundary of the applicability domain with the limits 2 or 3 for the standardised residual in prediction and $2K/N$ or $3K/N$ for the leverage.

8.6 Applicability domain based on DModX

For PLSR or PCR, another tool is used to detect outliers and define the applicability domain: DModX, the distance from one observation to the model in the space of original explanatory variables space (SIMCA, 1998). More formally, for the i^{th} compound of the training set, DModX is computed as $DModX_i = \sqrt{\frac{1}{K-M}\sum_{k=1}^{K} e_{ik}^2}$ where e_{ik} is the element ik of the residual matrix $\mathbf{E} = \mathbf{X^c} - \mathbf{T_M}\mathbf{W_M}^{-1}$, $\mathbf{T_M}$ contains the M latent variables or PCs scores vectors of the model and $\mathbf{W}_M$ is the weight matrix used to construct the latent variables as linear combination of the original variables $\mathbf{X^c}$. This definition can be generalised to any new compound for which the descriptors can be computed as well as the scores in the new latent subspace.

A high value of $DModX$ indicates that the compound is quite different from the compounds in the training set. An applicability domain can be defined by normalising the distances and considering as extrapolating points observations having standardised distances greater than 3.

This definition of the applicability domain is not illustrated on the data of Veith and Mekenyan (1993) because PLSR or PCR has no sense as the two latent variables or the two principal components are significant.

8.7 Applicability domain based on the Hotelling's T^2

Another distance tool to detect strong outliers and define the applicability domain is given by the Hotelling's T^2 (Jackson, 1991). This statistic is a multivariate generalisation of Student's t-test, and provides a check for observations adhering to multivariate normality. For the i^{th} compound of the training set, the Hotelling's T^2 is given by $T_i^2 = (\mathbf{x_i} - \bar{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x_i} - \bar{\mathbf{x}})'$ where $\mathbf{x_i}$ is the observed K-vector of descriptors, $\bar{\mathbf{x}}$ is the K-vector of mean descriptors and $\mathbf{S}$ is the variance-covariance matrix of the observed descriptors in the training set. This definition can be generalised to any other new molecule using the corresponding observed descriptors $\mathbf{x_0}$. If the descriptors are centered, the Hotelling's T^2 is the same distance as the Mahalanobis one and the leverage.

A high value of T_i^2 denotes a compound that is far from the central molecule. The 95% significance limits of the Hotelling distribution (proportional to the Fisher) correspond to 95% of the data and can detect extreme compounds. Assuming normality of the data, the Hotelling's T^2 can be computed with any modelling technique but is mainly used with PLSR and PCR.

8.8 Applicability domain based on the density

Most of the tools used for the definition of applicability domain assume implicitly an underlying distribution of the data, either uniform or normal. Those assumptions are not realistic with QSAR data as such data are often clustered, with empty regions in the descriptors space. Nonparametric density estimation is an approach capable of identifying such empty regions (Jaworska *et al.*, 2005a,b).

The most well-known method for nonparametric density estimation is the kernel-based method. To simplify the multivariate density estimation, a PCA is first applied to the QSAR data to obtain orthogonal variables. Jaworska *et al.* (2005a,b) proposed to estimate each PC density using kernel estimators and take their product.

The applicability domain is then defined by high density regions of level $(1-\alpha)$, with $0 < \alpha < 1$. A $(1-\alpha)$ high density region is the smallest multidimensional region comprising $(1-\alpha) * 100$ percents of the probability mass. In practice, delimiting exactly this region may be quite complicated with high dimensional data.

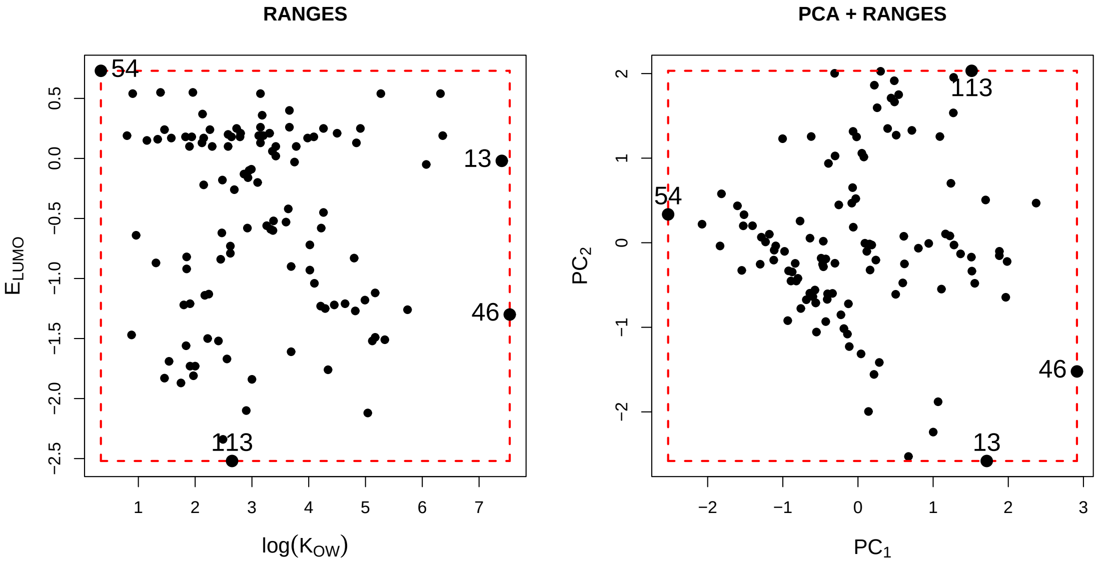

Figure 6: Applicability domain based on the ranges of the two observed descriptors (left) or on the ranges of the two principal components scores (left). Data of Veith and Mekenyan (1993).

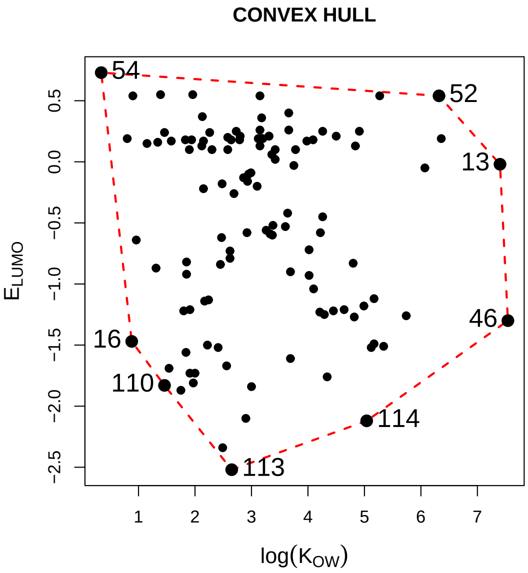

Figure 7: Applicability domain based on the convex hull of the 114 observed pairs of descriptors. Data of Veith and Mekenyan (1993).

8.9 Comparison of different applicability domains on a QSAR model example

Jaworska *et al.* (2005a,b) compared the different cited approaches and concluded that the range-based definition with a preliminary PCA rotation is the most adapted practical approach. In addition, only the PCA transformation and the ranges need to be stocked to be able to check if any new molecule is in the applicability domain of the QSAR model.

The report from the expert group on (Q)SAR, OECD (2004), recommend that further work is carried out to define more precisely the tools that can be used to define QSAR applicability domain. There is a need to reconcile the method used for the model development and the applicability domain definition. Whatever is

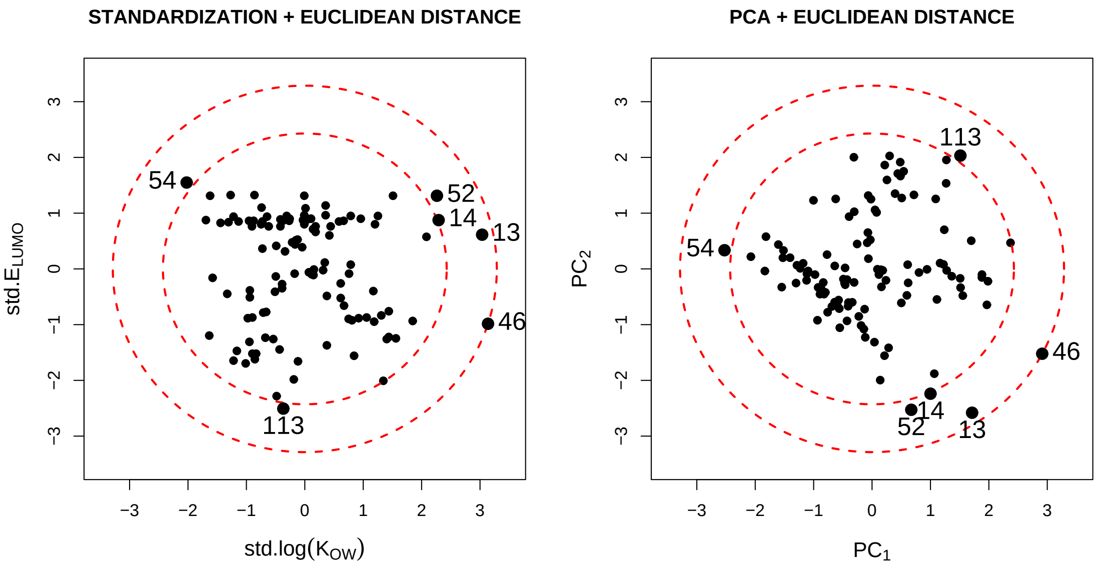

Figure 8: Applicability domain based on the Euclidean distance to the center of the dataset after a standardisation of the two descriptors (left) or after the transformation to the two PCs scores (right). The inner (resp. outer) dashed circle is the applicability domain considering a maximal Euclidean distance such that it contains 95% (resp. 100%) of the data.

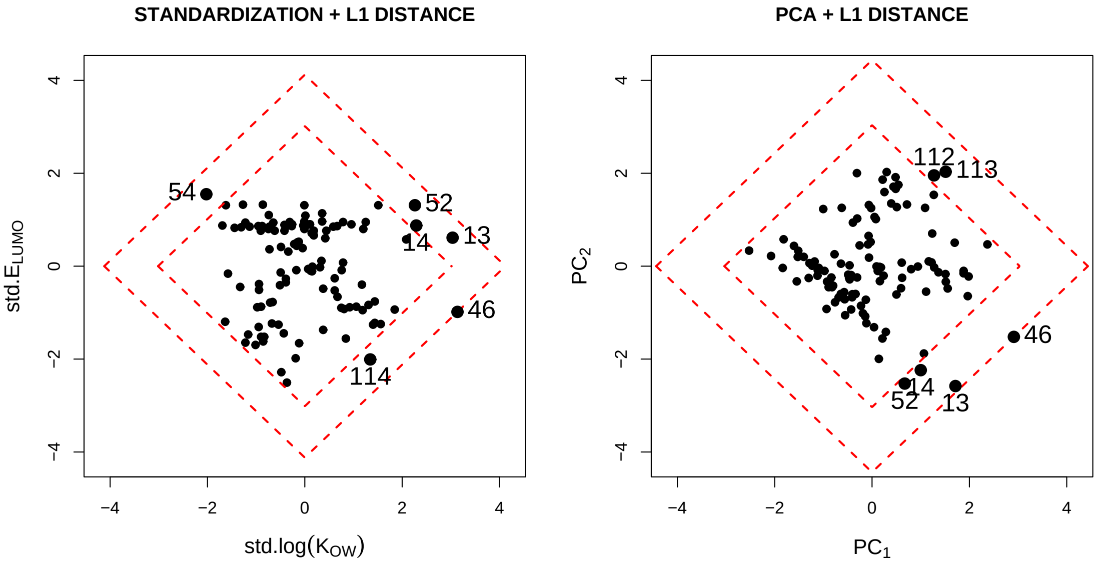

Figure 9: Applicability domain based on the L_1-distance to the center of the data after a standardisation of the two descriptors (left) or after the transformation to the two PCs scores (right). The inner (rep. outer) dashed circle is the applicability domain considering a maximal L_1-distance such that it contains 95% (resp. 100%) of the data. Data of Veith and Mekenyan (1993).

the statistical tool used, the exact definition of the applicability domain should be stated. If a new compound reveals to be outside the applicability domain, this information should be considered as a warning. Indeed, there is always the possibility that the QSAR model extrapolates correctly.

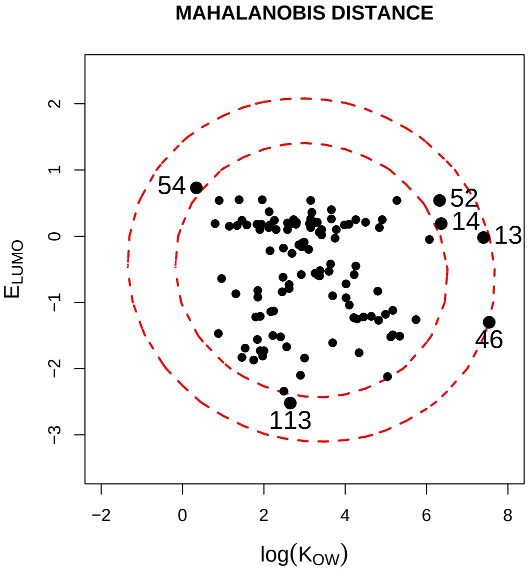

Figure 10: Applicability domain based on the Mahalanobis distance to the center of the data. The inner dashed circle is the applicability domain considering a maximal Mahalanobis distance such that it contains 95% of the data and the outer circle corresponds to 100% (the largest Mahalanobis distance). Data of Veith and Mekenyan (1993).

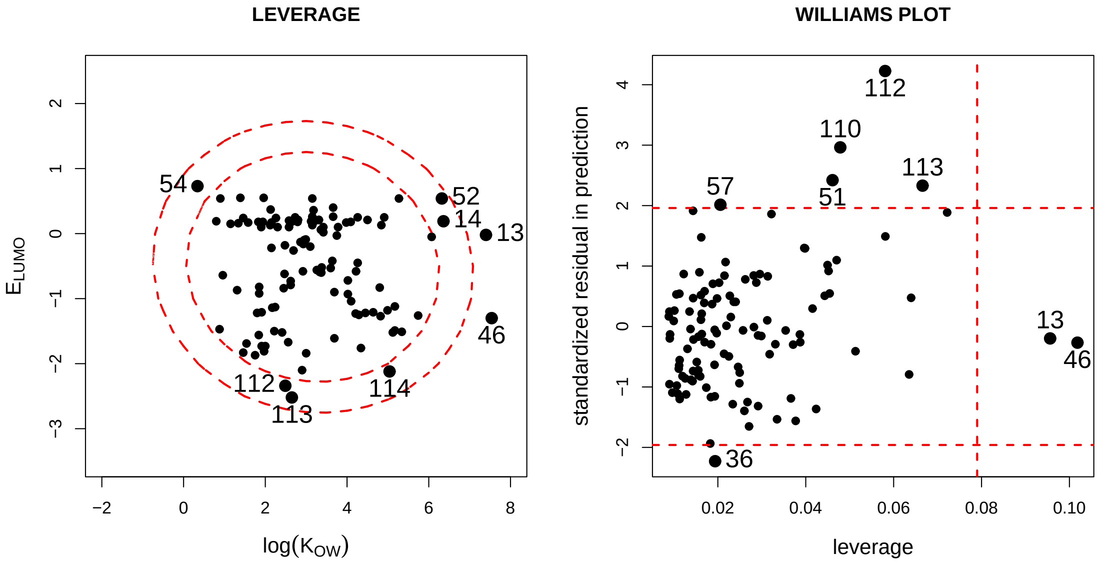

Figure 11: Left: Applicability domain based on the leverages. The inner dashed circle corresponds to the limit $2K/N$ and the outer circle to $3K/N$. Right: Williams plot. The vectical dashed line represent the limit $3K/N$ and the two horizontal dashed lines are the limits in ±2. Data of Veith and Mekenyan (1993).

Nearly all the presented methods are only applicable for quantitative descriptors. Jaworska *et al.* (2005a,b) suggest to divide the data set according to the levels of categorical descriptors and construct an applicability domain for each level. Only the $L1$-distance or the leverages can manage dummy variables for encoding the categorical descriptors.

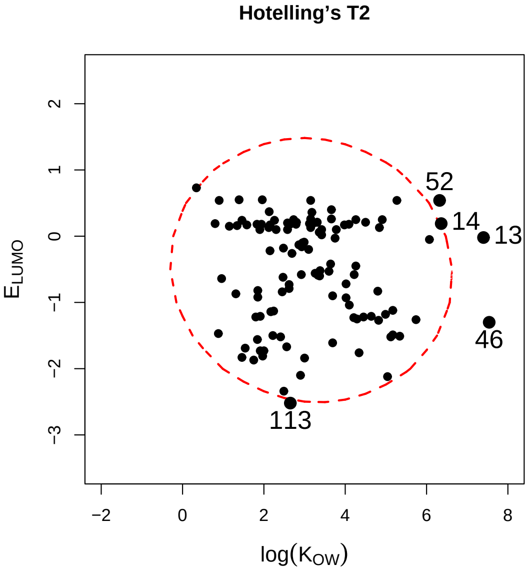

Figure 12: Applicability domain based on the Hotelling's T^2. Data of Veith and Mekenyan (1993).

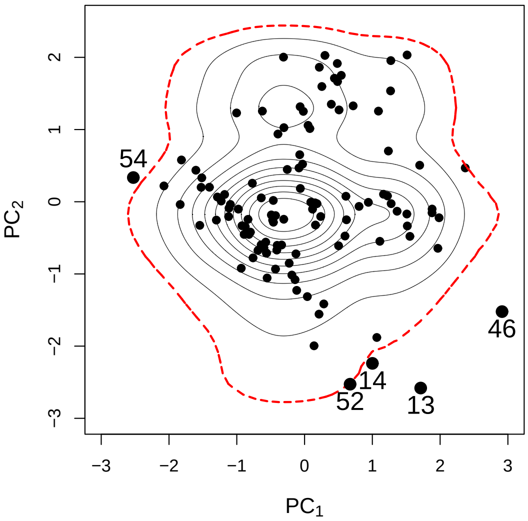

Figure 13: Applicability domain based on the joint density estimated using kernels after the transformation to the two principal components scores. The outer dashed line represents the high density region of level 0.95. The other inner lines represent different levels of density. Data of Veith and Mekenyan (1993).

9 Conclusion

QSAR modeling is an emerging tool in pharmaceutical industry. MLR, PLSR and RT are the three most often used classes of models. NN is becoming also more and more popular. In practice, it is recommended to fit different model families, to enter different subsets of descriptors, and select the "best" model. There is an increasing concern in statistically validating QSAR models and measuring their performance. The coefficient of determination (R^2) and its cross-validated version (Q^2) are the two most important statistical tools to measure respectively the goodness-of-fit and the predictive power of the model. Nevertheless, the importance of measuring the predictive power on an independent test set is not yet ingrained in mind of all practitioners.

The definition of the applicability domain of a QSAR model is encouraged to be able to check if the QSAR model is valid for new molecules and provides reliable predictions. However, in practice, a QSAR model is still often used for the prediction of new compounds without comparing them to the training set and without measuring the reliability of those local predictions.

References

De Aguiar, P.F., Bourguignon, B., Khots, M.S., Massart, D.L. and Phan-Than-Luu, R. (1995), "D-Optimal Designs", *Chemometrics and Intelligent Laboratory Systems*, 30, 199-210.

Andrea, T.A. and Kalayeh, H. (1991), "Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors", *Journal of Medicinal Chemistry*, 34, 2824-2836.

Atkinson, A.C. (1985), *Plots, transformations and regression*, Oxford:Clarendon Press.

Baroni, M., Costantino, G., Cruciani, G., Riganelli, D. , Valigi, R. and Clementi S. (1993), "Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems", *Quantitative Structure-Activity Relationships*, 12(1), 9-20.

Baurin, N., Mozziconacci, J.C., Arnoult, E., Chavatte, P., Marot, C. and Morin-Allory, L., (2002), "Insights into 2D-QSAR consensus prediction Virtual Screening of a 1992430 compounds database", available on the web at $http://www.univ-orleans.fr/SCIENCES/ICOA/communications/com2002/baurin.pdf$

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *Regression diagnostics: Identifying influencial data and sources of collinearity*, New-York, John Wiley and Sons.

Benigni, R. (2003), *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, Ed., CRC Press, Boca Raton, FL, USA.

Bishop, C.M. (1986), *Neural Networks for Pattern Recognition*, Oxford University Press.

Blockeel, H., Dzeroski, S., Kompare, B., Kramer, S., Pfahringer, B. and Van Laer W. (2004), "Experiments In Predicting Biodegradability", *Applied Artificial Intelligence*, 18(2), 157-181.

Box, G.E.P., Hunter, W.G., and Hunter, J.S., (1978), *Statistics for Experimenters*, John Wiley & Sons, Inc., New York.

Breiman, L., Freidman, J., Olshen, R. and Stone, C. (1984), *Classification and regression trees*, Belmont, CA: Wadsworth International Group.

Consonni, V. Mauri, A. and Pavan M. (2005), *DRAGON, version 5 (2005)*, Milano, Italy. Program for the calculation of molecular descriptors from HyperChem, Tripos, MDL file, SYBYL-molfile formats from ChemOffice and Tripos molecular design software. Download available at: $http://www.talete.mi.it/main_exp.htm.$

Crum-Brown, A. and Fraser, T. R. (1989), "On the connection between chemical constitution and physiological action. Part I. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia", *Transactions of the Royal Society of Edinburgh*, 25, 151.

Cronin, M.T.D. (2000), "Computational methods for the prediction of drug toxicity", *Current Opinion in Drug Discovery & Development*, 3, 292-297.

Cronin, M.T.D. (2002), "The current status and future applicability of quantitative structure-activity relationships (QSARs) inpredicting toxicity", *Alternatives to Laboratory Animals*, 30(2), 81-84.

Cronin, M.T.D. and Schultz, T.W. (2003), "Pitfalls in QSAR", *Journal of Molecular Structure*, 622, 39-51.

Cronin, M.T.D., Jaworska, J., Walker, J.D., Comber, M., Watts, C.D., and Worth, A.P. (2003), "Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances", *Environmental Health Perspectives*, 111, 1391-1401.

Cronin, M.T.D., Walker, J.D., Jaworska, J., Comber, M., Watts, C.D. and Worth, A.P. (2003), "Use of QSAR relationships in international decision-making frameworks to predict health effects of chemical substances", *Environmental Health Perspectives*, 111, 1376-1390.

Cronin, M.T.D. and Livingstone, D.J. (2004), *Predicting Chemical Toxicity and Fate*, Ed., CRC Press, Boca Raton, FL, USA.

Dearden, J.C., Barratt, M.D., Benigni, R., Bristol, D.W., Combes, R.D., Cronin, M.T.D. et al. (1997), "The development and validation of expert systems for predicting toxicity–the report and recommendations of an ECVAM/ECB workshop (ECVAM workshop 24)", *Alternatives to laboratory animals*, 25, 223-252.

Devillers, J. (1996), *Neural networks in QSAR and drug design*, London: Academic Press.

Doweyko, A.M. (2004), "3D-QSAR illusions", *Journal of computer-aided molecular design*, 18(7-9), 587-596.

Draper, N.R. and Smith, H. (1981), *Applied regression analysis*, 2nd edition. New York: John Wiley.

Duprat, A.F., Huynh, T. and Dreyfus, G. (1998), "Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of logP", *Journal of Chemical Information and Computer Sciences*, 38, 586-594.

Dzeroski, S. (2001), "Applications of symbolic machine learning to ecological modelling", *Ecological Modelling*, 146, 263-273.

Efron, B. (1982), *The jackknife, the bootstrap and other resampling planes*, Philadelphia : Society for Industrial and Applied Mathematics.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, London:Chapman & Hall.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (1996), "Multivariate Design and Modelling in QSAR", *Chemometrics and Intelligent Laboratory Systems*, 34, 1-19.

Eriksson, L., Johansson, E. and Wold, S. (1997), QSAR model validation. In: *Quantitative Structure-Activity Relationships in Environmental Sciences–VII*, (Chen, F., Schüürmann, G., eds), Pensacola, FL:SETAC Press, 381-397. Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., and Wold, S., (2000),

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C. and Wold, S. (2000), *Design of Experiments - principles and applications*, Umetrics AB.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (2001), *Multi- and megavariate data analysis - principles and applications*, Umetrics AB.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003), "Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs", *Environmental Health Perspectives*, 111(10), 1361-1375.

European Commission (1996), "Technical Guidance Document in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances and Commission Regulation (EC) No 1488/94 on Risk Assessment for Existing Substances", Luxembourg: European Commission, Office for Official Publications of the European Communities, http://ecb.jrc.it.

Fujita, T., Iwasa, J. and Hansch, C. (1964), "A new substituent constant, p, derived from partition coefficients", *Journal of the American Chemical Society*, 86, 5175-5180.

Golbraikh, A. and Tropsha, A. (2002), "Beware of q2!", *Journal of Molecular Graphics and Modelling*, 20(4), 269-276.

Graham, R. (1972), "An efficient algorithm for determining the convex hull of a finite point set", *Information Processing Letters*, 1, 132-133.

Gramatica, P., Navas, N. and Todeschini, R. (1998), "3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs)", *Chemometrics and Intelligent Laboratory Systems*, 40, 53-63.

Gramatica, P., Consonni, V. and Todeschini, R. (1999), "QSAR study of the tropospheric degradation of organic compounds", *Chemosphere*, 38, 1371-1378.

Gramatica P, Corradi M, Consonni V. (2000), "Modeling and prediction of soil sorption coefficients of non-ionic organic pesticides by different sets of molecular descriptors", *Chemosphere*, 41, 763-777.

Gramatica, P. and Papa, E. (2003), "QSAR modeling of bioconcentration factor by theoretical molecular descriptors", *Quantitative Structure-Activity Relationships*, 22, 374-385.

Hammett, L.P. (1937), "The effect of structure upon the reactions of organic compounds. Benzene derivatives", *Journal of the American Chemical Society*, 59, 96.

Hansch, C., Maloney, P.P., Fujita, T., and Muir, R.M. (1962), "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients", *Nature*, 194, 178-180.

Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, C.F. and Streich, M. (1963), "The Correlation of Biological Activity of Plant Growth-Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients", *Journal of the American Chemical Society*, 85, 2817-2824.

Hansch, C., and Fujita, T. (1964), "R-s-p analysis - A method for the correlation of biological activity and chemical structure", *Journal of the American Chemical Society*, 86, 1616-1626.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The elements of statistical learning. Data Mining, inference and predictions*, Springer.

Hawkins, D.M., Basak, S.C. and Shi, X. (2001), "QSAR with Few Compounds and Many Features", *Journal of Chemical Information and Computer Sciences*, 41, 663-670.

Hoffman, B. , Cho, S.J. and Zheng, W. (1999), "Quantitative structure-activity relationship modeling of dopamine D1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and k nearest neighbor methods", *Journal of Medicinal Chemistry*, 42, 3217-3226.

Hopfinger, A.J. and Tokarski, J.S. (1997), *Practical applications of computer-aided drug design. In: Practical Applications of Computer-Aided Design*, (Charifson PS, ed) New York:Marcel Dekker, 105-164.

Hulzebos, E.M., Schielen, P.C.J.I. and Wijkhuizen-Maslankiewicz, L. (1999), "(Q)SARs for human toxicological endpoints: a literature search", *RIVM Report 601516.00. A report by RIVM (Research for Man and Environment).* Bilthoven:RIVM.

Huuskonen, J.J., Livingstone, D.J. and Tetko, I.V. (2000), "Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices", *Journal of Chemical Information and Computer Sciences*, 40, 947-955.

Izrailev, S. and Agrafiotis, D. (2001), "A novel method of building regression tree models for QSAR based on artificial ants", *Journal of Chemical Information and Computer Sciences*, 41(1), 176-180.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, Wiley: New York.

Jarvis, R.A. (1973), "On the Identification of the Convex Hull of of a Finite Set of Points in the Plane", *Information Processing Letters*, 2, 18-21.

Jaworska, J.S., Comber, M., Auer, C. and Van Leeuwen, C.J. (2003), "Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints", *Environmental Health Perspectives*, 111(10), 1358-1360.

Jaworska, J.S., Aldenberg, T. and Nikolova, N. (2005), "Review of methods for assessing the applicability domains of SARs and QSARs", available on the web site of the European Chemicals Bureau at *http://ecb.jrc.it/home.php?CONTENU=/DOCUMENTS/QSAR/*

Jaworska, J.S., Nikolova, N. and Aldenberg, T. (2005), "QSAR applicability domain estimation by projection of the training set in descriptor space: a review", *Alternatives to Laboratory Animals*, 33, 445-459.

de Jong, S. (1993), "SIMPLS:An alternative approach to partial least squares regression", *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.

Jurs, P.C.(2003), *ADAPT–Automated Data Analysis and Pattern Recognition Toolkit.*, University Park, PA: Pennsylvania State University. Available: *http : //research.chem.psu.edu/pcjgroup/ADAPT.html* [accessed 23 April 2002].

Katritzky, A.R., Lobanov, V.S. and Karelson, M. (1994), *CODESSA, Reference Manual.*, Gainesville, FLUniversity of Florida. Available: *http : //www.semichem.com/codessarefs.html* [accessed 19 April 2002].

Kovesdi, I., Dominguez-Rodriguez, M.F. and Orfi, L. (1999), "Application of neural networks in structure-activity relationships", *Medicinal Research Reviews*, 19, 249-269.

Kubinyi, H. (1994), "Variable selection in QSAR studies. I. An evolutionary algorithm", *Quantitative Structure-Activity Relationships*, 13, 285-294.

Kubinyi, H. (1996), "Evolutionary variable selection in regression and PLS analyses", *Journal of Chemometrics*, 10, 119-133.

Kubinyi, H., Folkers, G. and Martin, Y.C. (1998), "3D QSAR in drug design–recent advances", *Perspectives in Drug Discovery and Design*, 12, R5-R7.

Linusson, A. (2000), *Efficient Library Selection in Combinatorial Chemistry*, Ph.D. Thesis, Umea University, Umea, Sweden

Livingstone, D. (1995), *Data Analysis for Chemists. Applications to QSAR and Chemical Product Design*, Oxford Science Publications, Oxford.

Livingstone, D.J. (2000), "The characterization of chemical structures using molecular properties. A survey.", *Journal of Chemical Information and Computer Sciences*, 40, 195-209.

Loehlin, J.C. (1998), *Latent variable models: an introduction to factor, path, and structural analysis*, Hillsdale, NJ: Lawrence Erlbaum Associates..

Manallack, D.T. and Livingstone D.J. (1999), "Neural networks in drug discovery: Have they lived up to their promise?", *European Journal of Medicinal Chemistry*, 34, 195-208.

Marengo, E and Todeschini, R. (1992), "A new algorithm for optimal, distance-based experimental design", *Chemometrics and Intelligent Laboratory Systems*, 16, 37-44.

Martens, H. (1985), *Multivariate Calibration*, Dr. techn. Thesis. Technical 25 University of Norway, Trondheim

Mekenyan, O. and Bonchev, D. (1986), "OASIS method for predicting biological activity of chemical compounds", *Acta Pharmaceutica Jugoslavica*, 36, 225-237.

Meyer, H. (1899), "Zur Theorie der Alkoholnarkose. 1. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung?", *Archiv fur experimentelle Pathologie und Pharmakologie*, 42, 109.

Myers, R.H. (1986), *Classical and modern regression with applications*, Boston: Duxbury Press.

Oprea,T.I., Mannhold, R., Kubinyi, H. and Folkers, G (2005), *Chemoinformatics in Drug Discovery*, eds. Hoboken, NJ:John Wiley & Sons, Inc.

Overton, E. (1897), "Ueber die Osmotischen Eigenschaften der Zelle in ihrer Bedeutung für die Toxicologie und Pharmakologie", *Zeitschrift für physik. Chemie* , 22, 189.

OECD series on testing and assessment No.49 (November 2004), "The report from the expert group on (Quantitative) Structure-Activity Relationship [(Q)SARs] on the principles for the validation of (Q)SARs", *ENV/JM/MOMNO(2004)24. Paris, France, OECD*

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

Richet, C. (1893), *Compt. Rend. Seances Soc. Biol.*, 9, 775.

Schultz, T.W. and Cronin, M.T.D. (2003), "Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships", *Environmental Toxicological and Chemistry*, 22, 599-607.

Schultz, T.W., Cronin, M.T.D., Walker, J.D. and Aptula, A.O. (2003), "Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective", *Journal of Molecular Structure*, 622, 1-22.

Selick, H.E., Beresford A.P. and Tarbit, M.H. (2002), "The emerging importance of predictive ADME simulation in drug discovery", *Drug Discovery Today*, 7(2), 7-109.

SIMCA 7.0. A new standard in multivariate data analysis, Manual, Umetrics AB, Umea, Sweden.

Sokal, R.R. and Rohlf, F.J. (1995), *Biometry: The Principles andPractice of Statistics in Biological Research*, Freeman, NewYork.

Stuper, A.J. and Jurs, P.C. (1976), "ADAPT: A computer system for automated data analysis using pattern recognition techniques", *Journal of Chemical Information and Computer Sciences*, 16, 99-105.

Taft, R.W.Jr. (1952), "Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters", *Journal of the American Chemical Society*, 74, 3120-3128.

Test guideline 401, 2001, *Trends in Pharmacological Sciences*, 22(2), 2375-2388.

Therneau, T.M. and Atkinson, E.J. (1997), "An introduction to recursive partitioning using the RPART routines", Technical Report #61. Mayo Clinic. Rochester, Minnesota.

Todeschini, R. and Consonni, V. (2000), *Handbook of molecular descriptors*, Wiley-VCH: Weinheim (Germany).

Todeschini, R. and Gramatica, P. (1997), "3D-modelling and prediction by WHIM descriptors. Part 6. Applications of WHIM descriptors in QSAR studies", *Quantitative Structure-Activity Relationships*, 16, 120-125.

Tong, W., Fang, H., Hong, H., Xie, Q., Perkins, R., Anson, J. and Sheehan, D. (2003), "Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective", *Pure and Applied Chemistry*, 75(11-12), 2375-2388.

Tong, W., Xie, Q., Hong, H., Shi, L., Fang, H. and Perkins, R. (2004), "Assessment of prediction confidence and domain extrapolation of two structure activity relationship models for predicting estrogen receptor binding activity", *Environmental Health Perspectives*, 112(12), 1249-1254.

Tropsha, A., Gramatica, P. and Gombar V.K. (2003), "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models", *QSAR & Combinatorial Science*, 22(1), 69-77.

Tutz, G. (1990), "Smoothed categorical regression based on direct kernel estimates", *Journal of Statistical Computation and Simulation*, 36, 139-156.

Veith, G.D. and Mekenyan, O.G. (1993), "A QSAR approach for estimating the aquatic toxicity of soft electrophiles", *Quantitative Structure-Activity Relationships*, 12, 349-356.

Vellman, P.F. and Welsch, R. (1981), "Efficient computing of regression diagnostics", *The American Statistician*, 35, 234-242.

Vinod, H.D. and Ullah, A. (1981), *Regression Methods*, New York: Marcel Dekker.

Walker, J.D., Rodford, R. and Patlewicz, G. (2002), "Quantitative structure-activity relationships for predicting percutaneous absorption rates", *Environmental toxicology and chemistry*, 22(8), 1870-1884.

Waller, C.L. and Bradley, M.P. (1999), "Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies", *Journal of Chemical Information and Computer Sciences*, 39, 345-355.

Weininger, D. (1988), "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules", *Journal of Chemical Information and Computer Sciences*, 28(1), 31-36.

Werbos, P. (1974), *Beyond regression: New tools for prediction and analysis in the behavioral sciences*, Ph.D. dissertation, Dept. of Applied Mathematics, Harvard University.

Whitley, D.C., Ford, M.G. and Livingstone, D.J. (2000), "Unsupervised forward selection: A method for eliminating redundant variables.", *Journal of Chemical Information and Computer Sciences*, 40, 1160-1168.

Worth, A. (2002), "ECVAM's Activities on Computer Modelling and Integrated Testing", *Alternatives to Laboratory Animals*, 30(2), 133-137.

Wold, H. (1966), *Estimation of principal components and related models by iterative least squares*, in *Multivariate analysis* Academic Press, New York.

Wold, S., Martens, H. and Wold H. (1982), *The multivariate calibration problem in chemistry solved by the PLS method*, Proceedings of the Conference on Matrix Pencils, March 1982, (A. Ruhe and B. Kagstrom, eds), Lecture Notes in Mathematics, 286-293. Heidelberg: Springer Verlag.

Wold, S. and Dunn, W.J. (1983), "Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for Their Applicability", *Journal of Chemical Information and Computer Sciences*, 23, 6-23.

Wold, S. and Dunn, W.J. (1983), "Multivariate Design", *Analytica Chimica Acta*, 191, 17-32.

Worth, A.P. and Cronin, M.T.D (2004), "Report of the worskop on the validation of QSARs and other computational prediction models", *Alternatives to Laboratory Animals*, 32(1), 703-706.

Xu, L., and Zhang, W.J. (2001), "Comparison of different methods for variable selection", *Analytica Chimica Acta*, 446, 477-483.

Wold, S. and Dunn, W.J. (1983), "Novel variable selection quantitative structure-property relationship approach based onthe k-nearest-neighbor principle", *Journal of Chemical Information and Computer Sciences*, 40, 185-194.