EVALUATION OF DIFFERENT STATISTICAL APPROACHES FOR THE VALIDATION OF QUANTITATIVE STRUCTURE –ACTIVITY RELATIONSHIPS

Author:

Prof. Paola Gramatica
In collaboration with Dr. Pamela Pilutti

QSAR and Environmental Chemistry Research Unit
Department of Structural and Functional Biology (DBSF)
Via J.H. Dunant 3
University of Insubria
21100 Varese
Italy

E-mail: paola.gramatica@uninsubria.it
http://www.dipbsf.uninsubria.it
http://www.qsar.it

Sponsor:

The European Commission - Joint Research Centre
Institute for Health & Consumer Protection - ECVAM
21020 Ispra (VA)
Italy

Contact: Dr Andrew Worth
E-mail: andrew.worth@jrc.it
http://ecb.jrc.it/QSAR

JRC Contract ECVA-CCR.496576-Z

VERSION OF 4 NOVEMBER 2004
(with minor editorial modifications by JRC)

PREFACE

The validation of QSARs has been (and still is) a subject of intense debate within the academic regulatory and regulated communities. The questions discussed refer to: a) the validation principles that should be followed; b) the methods and approaches that can be used to apply the principles; c) the criteria for establishing scientific validity of models; and d) the flexibility and pragmatism that should be applied in the "real-world" use of QSARs for regulatory purposes.

The OECD Joint Meeting has now agreed to five principles that should be followed to establish the scientific validity of a QSAR, thereby facilitating its acceptance for regulatory purposes. One of these principles refers to the need to establish "appropriate measures of goodness-of-fit, robustness and predicitivity" for any model.

The present report was written by Professor Paola Gramatica (University of Insubria, Italy) with the support of a JRC contract. The report is based on a detailed investigation of the use of different statistical approaches for the validation of QSARs. It should therefore be a very valuable contribution to the development of the OECD Guidance Document on (Q)SAR Validation.

It should be noted that while the analyses are based on a selection of environmental, ecotoxicological and toxicological QSARs, the specific models were chosen to explore and illustrate the use of different statistical approaches. This study was not intended to be a formal validation study of the validities of the QSARs chosen.

DISCLAIMER

Any conclusion or opinion expressed in the present report reflects only the author's view and the European Community is not liable for any use that may be made of the information contained therein.

Andrew Worth
Coordinator of the JRC Activity on (Q)SARs
9 December 2004

TABLE OF CONTENTS

1. INTRODUCTION

Prof. Paola Gramatica (Insubria University, Italy) has, under the terms of a JRC contract (below), applied and evaluated different statistical techniques for the validation of some published QSAR models.

A more detailed explanation of the applied statistical techniques is reported below, and the availability of model statistics, including cross-validated statistics is discussed. Furthermore, this report draws some conclusions regarding to the need for statistical validation.

2. DESCRIPTION OF THE JRC CONTRACT WORK

In the contract work different statistical procedures for the statistical validation of QSARs were compared and evaluated.

Different kinds of internal validation: 1) cross validation by leave-one-out (LOO), 2) by leave-many-out (LMO), 3) boot-strapping, 4) response permutation (randomization testing or Y-scrambling) [a, b] were applied to the selected QSAR regression models and their relative performances compared (by the *MOBY-Digs* software).

Statistical external validation was performed by first splitting the original data sets by various techniques to ensure training and validation sets that were equally representative of the same chemical domain [b]. The splittings were performed by D-Optimal Experimental Design (*DOLPHIN* software) and by similarity analysis through Artificial Neural Network (Kohonen Maps, in *KOALA* software) or by random splitting.

The domain of applicability was verified by the leverage approach [a, b] and both the influential and the outlier chemicals were identified by the Williams graph (*SCAN* and *STATISTICA* packages).

The technical background of the proposed contract work has already been extensively written up by Gramatica in the cited background references [a ,b].

The verified QSAR models are related to the end-points required by the contract. The selection of models was made on the availability in the original papers of all the data needed for validation (chemical names, experimental data, descriptor values, model algorithm), thus allowing the development and validation of the same published QSAR model (note that unfortunately values of descriptors are not reported in most of the papers in the literature). The rationale for this selection was not deeper, as the aim of the contract, in contractor's idea, was not to comment on specific models but to apply and evaluate different statistical approaches for QSAR validation; in this case does not matter which the studied models are, because, since they were published, they can be considered "good" models!

The selected QSAR models on which we performed the statistical analyses are:
A physicochemical effect [1]:
the soil adsorption coefficient (Koc) of halogenated aromatics.

A parameter of environmental fate [2]:
the bioconcentration factor (BCF) of heterogeneous organic compounds.

c) An ecotoxicological endpoint [3]:
- acute toxicity to *P. promels* (LC_{50}); the data set includes heterogeneous chemicals collected into various groups (benzenes, alcohols, aldehydes, aliphatics, esters, ketones and phenols).

d) Some toxicological endpoints [4]:
-Toxicity and metabolic effects of alcohols on perfused rat liver; the activities of the following enzymes were assessed: glutamate-pyruvate-transaminase (GTP), lactate dehydrogenase (LDH), glutammate dehydrogenase (GLDH), and the amount of intracellular ATP [4].
-Toxicity of pyrimidines and bi-pyridines to mice (LD_{50}) [4].
- Lethality of halogenated hydrocarbons to *A. nidulans*: the lowest concentration inducing lethality in 63% of the cells of a sample of the mould *A. nidulans* (D_{37}) [4].

e) Mutagenicity of aromatic amines to *S. typhimurium* TA98 + S9 and TA100 + S9 microsomial preparation [5].

The majority of the software packages used were developed by the Milano Chemometrics Research Group of Prof. Todeschini, and have been widely applied by Prof. Gramatica's QSAR Research Unit that has published about 50 papers on QSAR in Environmental Research.

All the data sets used are reproduced in this report.
On the basis of the reported results (see below), suggestions are made regarding the appropriate choice of the statistical method for a given situation, particular attention being paid to the number of chemicals in the dataset.

References

Background references:

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

[b] 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, 69-76.

Studied QSAR models:

[1] Wei, D.B., Wu, C.D., Wang, L.S. and Hu, H.-Y. (2003). QSPR-based prediction of absorption of halogenated aromatics on yellow-brown soil. *SAR and QSAR in Environmental Research* 14 (3), 191-198.

[2] Gramatica, P. and Papa, E. (2003). QSAR modeling of bioconcentration factor by theoretical molecular descriptors. *QSAR & Combinatorial Science* 22 (3), 374-385.

[3] Kulkarni, S.A., Raje, D.V. and Chakrabarti, T. (2001). Quantitative structure-activity relationships based on functional and structural characteristics of organic compounds. *SAR and QSAR in Environmental Research* 12, 565-591.

[4] Cronin, M.T.D., Dearden, J.C., Duffy, J.C., Edwards, R.. Manga, N., Worth, A.P. and Worgan, A.D.P. (2002). The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. *SAR and QSAR in Environmental Research* 13 (1), 167-176.

[5] Gramatica, P., Consonni, V. and Pavan, M. (2003). Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. *SAR and QSAR in Environmental Research* 14 (4), 237-250.

Software packages:

Todeschini, R., Consonni, V. and Pavan, M. *Moby Digs* - Software for multilinear regression analysis and variable subset selection by Genetic Algorithm (2002). Version 1.2 for Windows, Talete srl, Milan, Italy.

Todeschini, R. and Mauri, A. *DOLPHIN* - Software for optimal distance-based experimental design (2002) Ver. 2.0 for Windows, Talete srl, Milan, Italy.

Todeschini, R. and Consonni, V.. *KOALA*-Software for Kohonen Artificial Neural Networks (2001). Rel. 1.0 for Windows, Talete srl, Milan, Italy.
SCAN - Software for Chemometric Analysis (1995). Rel. 1.1 for Windows, Minitab, USA.
STATISTICA (1997). Rel. 5.1 for Windows, StatSoft, USA.

3. TERMINOLOGY AND COMMENTS ON THE APPLIED STATISTICAL APPROACHES

Since the real utility of a QSAR model lies in its ability to accurately predict the modeled property for new chemicals, a realistic assessment of the model's true predictive power must be ascertained. Thus the two fundamental steps in QSAR modeling are:

1) Model validation, both internal and statistical external, which implies quantitative assessment of model robustness and its predictive power, and
2) Definition of the application domain of a model in the space of chemical descriptors used in deriving the model.

Before commenting on the results of model validation performed in this contract work, the different approaches applied are presented and commented on.

Boot-strapping

The basic premise of this method is that the data set is representative of the population from which it was drawn. Since there is only one data set, bootstrapping simulates what would happen if the population were resampled by randomly resampling the data set (Efron and Tibshirani, 1993; Wehrens et al., 2000).
In a typical bootstrap validation, K n-dimensional groups are generated by a randomly repeated selection of n-objects from the original data set. Each group of data is then always n-dimensional; some of these objects can in the same group more than once while others might never be inside. The model obtained on the first selected objects is used to predict the values for the excluded sample. From each bootstrap sample the statistical parameter of interest (here Q^2) is calculated. This yields an ensemble of estimates used to obtain a meta-estimate (for instance, we could generate 100 bootstrap samples and calculate the Q^2 for each sample). As in the case of LMO validation, a high average Q^2 in bootstrap validation is a demonstration of model robustness and internal predictivity.

Chemical Domain of Model applicability

As even a robust, significant and validated QSAR cannot be expected to reliably predict the modelled property for the entire universe of chemicals, its domain of application must be defined, and the predictions for only those chemicals that fall in this domain can be considered reliable. The chemical domain of applicability is a theoretical region in the space defined by the modeled response and the descriptors of the model, for which a given QSAR should make reliable predictions. This region is defined by the nature of the chemicals in the training set, and can be characterized in various ways. The Williams plot of the regression allows a graphical detection of both the outliers for the response and the structurally influential chemicals in a model.

Williams plot or Ordinary Least Squares (OLS) Outlier and Leverage Plot is the plot of jackknifed residuals versus leverages (hat diagonals) (see Leverage). In this plot the horizontal and vertical straight lines indicate the limits of normal values: the first for the outliers and the second for influential chemicals. The jackknifed residuals, also called Studentized residuals, are the standardized cross-validated residual. Each residual is divided by its standard deviation, which is calculated without the i^{th} observation. A simple formula for the jackknifed residual is:

$$r_i' = \frac{r_i}{s\sqrt{1-h_{ii}}}$$

It is important to note that, while the outliers for the response can be highlighted only for chemicals with known responses, the possibility of a chemical to be out of the structural applicability domain of a model, and thus the reliability of its predictions, can be verified for every new chemicals, the only knowledge needed being the molecular structure.

Collinearity:

Collinearity is a situation where there is an almost perfect linear relationship among some or all of the independent variables in a regression model. In practical terms, this means that there is some degree of redundancy or overlapping of the variables. Particular attention must be paid to the collinearity of the selected molecular descriptors in order to avoid multicollinearity without, or with, "apparent" prediction power (due to chance correlation). For all the studied models the *QUIK* rule (Q Under Influence of *K*) (Todeschini et al., 1999) was applied, verifying whether the models have a global correlation of [X+Y] block (K_{XY}) greater than the global correlation of the X block (K_{XX}) variable, X being the molecular descriptors and *y* the response variable. The influence of the value of the difference ΔK (K_{XY}- K_{XX}) on the quality of the studied models was verified. Collinearity among descriptors is not always dangerous, but it must be carefully controlled to have predictive models. Actually, in some cases intercorrelated descriptors can carry useful structural information in the parts where they do not correlate with other descriptors.

Correlation Coefficient of correlation (R):

The correlation coefficient (R) is a simple statistical measure of the relationship between a dependent variable *y* (e.g. an endpoint) and one or more independent variable(s) *x*. It is given a value from 0 (for no relationship) to 1 (for a perfect fit) (100% in percentage). In QSAR analysis, R can be used as a measure of the statistical fit of a regression-based model, but the preferred form is its squared value (coefficient of determination, see below), often adjusted for the degrees of freedom (R^2_{adj})

Coefficient of determination (R^2):

The total variation of any data set is made up of two parts, the part that can be explained by the regression equation and the part that cannot be explained by the regression equation. The coefficient of determination is the amount of dependent variable variance explained by a regression mode. It equals the square of the correlation coefficient R between the experimental response (the dependent variable *y*) and the predictors (the independent variables *x*). It represents the explained variance of the model, and is used as a measure of the goodness-of-fit of the model.

It is commonly believed that the closer the value to unity (or 100 in percentages), the better the model. However, it should be noted that R^2 is just a measure of the quality of the fit between model-calculated and experimental values, and it does not reflect the predictive power of the model at all. It is possible that a QSAR model with high R^2 could be a poor predictor.

Cross-validation:

Cross-validation refers to the use of one or more statistical techniques for internal validation in which different proportions of chemicals are omitted from the training set, with iterations, in order to verify the "internal predictivity" (e.g. LOO, LMO, bootstrapping). The QSAR is developed on the basis of the data for the training chemicals, and then used to make predictions for the chemicals that were omitted. This procedure is repeated a number of times (for instance hundreds or thousands, depending on the software), so that statistics can be derived from the comparison of predicted data with the known data. The final model is the model developed on all the chemicals: the repeated splittings are made to verify the "internal predictivity". Cross-validation techniques allow the assessment of internal predictivity, in addition to the robustness of the model (stability of QSAR model parameters). Nothing is known regarding the predictivity on new external chemicals (never included in the training set for model development), because in cross-validation each chemical can be included in training at least in one iteration, thus the final model takes into consideration it. (Tropsha et al, 2003). This type of internal validation provide only

a reasonable estimate of the internal predictive power of a QSAR model; it was demonstrated that the LOO Q^2 cannot indicate a significant external predictive power (known as Kubinyi paradox).

Cross-validation by the Leave-One-Out (LOO) procedure:

Cross-validation by LOO employs n training sets, and from each of these 1 object is excluded. A number of models are developed by using each training set of n-1 objects in the and predicting each excluded object in the test set. For each model, the excluded object is predicted and the cross-validated explained variance (Q^2) computed, as average value of a generally high number of validation runs.

Also the present work demonstrates that the LOO approach is not sufficient to assess robustness and predictivity, the estimated Q^2 being too similar to R^2. Although small Q^2 values in the LOO test indicate models with low robustness and low internal predictive ability, the opposite is not necessarily true (Shao, 1993; Golbraikh and Tropsha , 2002).

Cross-validated explained variance (Q^2_{LOO} or R^2_{cv}):

The cross-validated explained variance or cross-validated correlation coefficient (Q^2_{LOO} or R^2_{cv}) is widely used and abused as a measure of the goodness of internal predictivity.

The formula for the calculation is: 1- PRESS/TSS or

$Q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$ where y_i, $\hat{y}_i$, and $\bar{y}$ are, respectively, the measured, predicted, and averaged (over the entire data set) values of the dependent variable; the summations run over all compounds in the training set.

In contrast to the fitting parameter R^2, which always increases when more descriptors are added, the value of Q^2 increases only when the added predictors are useful, if they are not useful for prediction it decreases.

However, in reality, it has been shown that no correlation exists between cross-validated Q^2 (internal validation) and the correlation coefficient R^2 between the predicted and observed activities for an external validation set. Studies have indicated that while high Q^2 is a necessary condition for high predictive power in a model, it alone is not sufficient. Recent studies have systematically addressed the issue of Q^2 being an inadequate characteristic of a model's predictive power.

Cross-validation by the Leave-Many-Out (LMO) procedure:

Cross-validation by LMO employs smaller training sets than the LOO procedure. In a typical LMO validation, n objects of the data set are divided in G cancellation groups of equal size, m_j ($= n/G$). Based on the value of n, G is generally selected between 2 and 10. A large number of models are developed with

each of the n-m_j objects in the training set and m_j objects in the validation set. For each corresponding model, m_j objects are predicted and Q^2_{LMO} computed (as average value of a generally high number of validation runs). If, in LMO validation, a QSAR model has a high average Q^2 it is generally concluded that the model is robust and internally predictive.

When dealing with small data sets, LMO validation with too strong perturbation (up to 50% of data out of training, each run) often under-estimates predictivity because only a reduced part of the data is used each time for model calibration: this is a waste of valuable information and the model may not contain all the relevant structure information of the whole data set.

External Validation:

In typical situations, finding new experimentally tested compounds for external validation purposes is generally difficult. The new data should be in a statistically significant number (in fact, results on few data can give optimistic or unreliable information regarding to the model predictivity) and belonging to the same chemical domain of the compounds used for model development. This is the best way of external validation, performed after the model development and this is the way recommended by ECVAM for other alternative methods.

When additional data (in useful quantity and quality) are not available, statistical external validation can be helpful in defining the actual predictive power of the model more precisely. This is done by an adequate splitting of the available input data set into training (for model development) and validation (for model predictive assessment) sets, using experimental design and other procedures (see splitting training and validation sets). In this contract which is specifically devoted to evaluating statistical approaches to QSAR validation, only statistical external validation was applied.

Statistical external validation refers to a validation exercise in which the chemical structures selected for inclusion in the validation set are different from those included in the training set, but are representative of the same chemical domain. The QSAR model developed using only training set chemicals is then applied to the "unknown" validation set chemicals to verify, more reliably, the predictive ability of the model.

The formula for the calculation of Q^2 is :

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{valid} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{valid} (y_i - \bar{y}_{tr})^2}$$

where y_i and $\hat{y}_i$ are respectively the measured and predicted (over the validation set) values of the dependent variable, and $\bar{y}_{tr}$ the averaged value of the dependent variable for the training set; the summations cover all the compounds in the validation set.

Other useful parameters are R^2, calculated for the validation chemicals by applying the model developed on the training set, and MSE (Mean Squared Error) for the two sets.
QSAR models obtained with statistical external validation during model development are predictive for the "unknown" chemicals of the validation set, that are never included in model development, thus the predictive performance of such models can be considered more realistic than when verified by internal validation approaches.

Fitting power:
The ability of a model to reproduce the input data of the training chemicals.

Internal validation:
Internal validation is a validation in which the chemical structures selected for inclusion in the test set are iteratively subsets of those included in the training set. Internal validation results in one or more measures of robustness of model parameters (Q^2 LOO cross-validation, Y-scrambling) and internal predictivity (Q^2 LMO cross-validation or bootstrap).
Internal validation is an essential, but not sufficient, form of validation, which should ideally be supplemented by statistical external validation if QSAR models have to be used for predictive purposes. In fact, it is important to note that in internal validations the information related to each chemical in the data set is considered at least in one iteration of the validation process: these chemicals are never new chemicals in the model development and their information is taken into account in the modelling.
Internal validation provides only a reasonable first approximation of the predictive ability of a QSAR model.

Leverage
A simple measure of a chemical being too far from the applicability domain of the model is its leverage in the original variable space, h_i , (Atkinson, 1985) which is defined as:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \qquad (i=1,\ldots,n)$$

where x_i is the descriptor row-vector of the query compound, and X is the n x k-1 matrix of k model descriptor values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. The warning leverage h^* is generally fixed at $3k/n$, where n is the number of training compounds, and k the number of model parameters. A leverage greater than the warning leverage h^* means that the predicted response is the result of substantial extrapolation of the model and, therefore, may not be reliable, so the predicted value must be used with great care (Williams graph) Only predicted data for chemicals belonging to the chemical domain of the training set should be proposed and used.

It is important to note that the leverage approach can be applied *a posteriori* to new chemicals, only calculating the descriptors from the molecular structure, and then applying the model, developed on the training set: chemicals with high leverage could have unreliable predictions.

Mean Squared Error (MSE):

This is the medium quadratic error, and is calculated : $\sum_1^n (y_i - \hat{y}_i)^2 \Big/ n$

It is an useful parameter for the comparison of error difference in model calculations between training and validation set chemicals.

Outlier:

An outlier of a QSAR model refers to a data point (a chemical) that falls outside the confidence interval of the regression line. Typically, the outlier of a QSAR model has a cross-validated standardised residual greater than three standard deviation units.

The outliers of a QSAR model should always be identified by the Williams graph, and the reason for their outlying behaviour should be provided.

Overfitting

Overfitting is the use of models that include more descriptors than are necessary. While the fitting power of a model becomes greater by increasing the number of descriptors, the same is not true for the predictive power. It can generally be accepted that a regression model with *k* descriptors and *n* training set compounds could be acceptable for validation only if the following criterion is satisfied: $n > 4\text{-}5k$, where n is the object number and k the dependent variables X. (Hawkins, 2004)

Prediction set

See Validation set

Randomization testing: Y-scrambling or response permutation

Randomisation testing is a technique for checking the robustness of a QSAR model and the statistical significance of the estimated predicted power. In this test, the dependent variable vector, Y-vector, is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The process is repeated several times. It is expected that the resulting QSAR models will generally have low R^2 and low Q^2 LOO values.

If the new models developed from the data set with randomised responses have significantly lower R^2 and Q^2 than the original model, then this is strong evidence that the proposed model is well founded, and not just the result of chance correlation.

In contrast, if all the QSAR models obtained in the Y-randomisation test have relatively high r^2 and q^2 LOO, then it implies that, for the given data set, the current modelling method is unable to give an acceptable QSAR model. (Eriksson et al, 2003, Tropsha et al., 2003)

Robustness

A QSAR model can be considered robust when its performance remains satifactory and stable when heavy perturbation in the training composition are made (for instance leave-many-out or bootstrapping)

Splitting into Training and Validation Sets

In typical situations, finding new experimentally tested compounds for external validation purposes is generally difficult.

Therefore recourse is made to splitting the available data set into a training set, used to develop the QSAR model, and a validation set, used only for statistical external validation. At this point the underlying goal is to ensure that both the training and the validation sets span, separately, the whole descriptor space occupied by the entire data set, and that the chemical domain in the two data sets is not too dissimilar. (Sjostrom and Eriksson, 1995)

Therefore an ideal splitting leads to a validation set in which each of its members is close to at least one point of the training set. (Golbraikh et al, 2002, 2003)

The composition of the training and validation sets is of crucial importance. A representative selection of compounds spanning, to a good degree, the chemical domain of interest should be included in these sets. Based on our experience (Gramatica et al, 2004), we suggest that the training and validation sets must satisfy the following criteria: (i) representative points of both sets must be close to each other, and (ii) the training set chemicals must be diverse.

Developing a rational approach towards selecting training and validation sets is an active research area and here we have applied three training set selections: a D-optimal experimental design (Marengo and Todeschini, 1992) (by the software *DOLPHIN*) the Kohonen Artificial Neural Network (K-ANN) (Gasteiger and Zupan, 1993; Zupan, 1997) (by the software *KOALA*) and a random splitting.

The Marengo–Todeschini algorithm is an algorithm for optimal distance based experimental design that does not require any preliminary hypothesis about a regression model. The best set of compounds is defined through a fast exchange algorithm where, in each cycle, substitution provides the maximum increase in the minimum distance between the currently selected compounds. Such an algorithm provides

a final distribution of the most dissimilar compounds selected from the set of allowed candidates. Regression models are developed on the selected training set, and once the models are established, predictions are made for the remaining molecules under study (validation set).

The splitting of the data set, realized by Kohonen Artificial Neural Network (K-ANN), takes advantage of the clustering capabilities of K-ANN, allowing the selection of a meaningful training set and a representative validation set. The structural information represented by the X-variable (the molecular descriptors) and the Y-variable (the response) are used as variables to build a Kohonen map (defined neurons, defined epochs). At the end of the defined epochs of the net training, similar chemicals fall within the same neuron, i.e. they carry the same information. To select the training set of chemicals, it is assumed that the compound closest to each neuron centroid is the most representative of all the chemicals within the same neuron. Thus, the selection of the training set chemicals is performed by the minimal distance from the centroid of each cell in the top map. The remaining objects, close to the training set chemicals, are used for the validation set (or viceversa).

In the splitting procedure using D-optimal design, that selects the most dissimilar chemicals in the training set, the validation set chemicals are always inside the training set space. The splitting by K-ANN is better balanced: the chemical composition of the training and validation sets is more similar. Thus, it is obvious that splitting by D-optimal design gives a more optimistic idea of the external predictivity, as can be expected for a validation set lying completely within the descriptor space of the training chemicals. In fact, splitting with D-optimal design gives models with completely interpolated predictions whereas, in some cases, predictions from K-ANN splitting can be interpreted as extrapolation for some chemicals.

Splitting by a random selection of the chemicals into two sets, while useful in splitting for internal validation as applied iteratively, gives very variable results when applied in external validation, strongly depending on the dimension and representativity of the sets.

Standard Deviation Error in Calculation (SDEC)

The standard deviation error in calculation is similar to the standard error of the estimate. SDEC is the square root of the residual sum of squares (RSS) divided by the number n of objects in the training set. The standard error in calculation (SDEC) should be similar to the experimental variability of an endpoint.

$$SDEC = \sqrt{\frac{RSS}{n}}$$

Standard Deviation Error in Prediction (SDEP)

The standard deviation error in prediction is similar to SDEC, but the residuals are calculated by using the predicted value of the dependent variable when an observation is left out from the training set and put into

the test set. PRESS (Predictive Error Sum of Squares) is the sum of the squares of the differences (residuals) between the experimental and predicted responses when predictions are made for objects left out of the training sets, but included in the test set.

The standard error in prediction (SDEP) should be similar to the experimental variability of an endpoint.

$SDEP = \sqrt{\frac{PRESS}{n}}$ Where PRESS is calculated from: $PRESS = \sum_i (y_i - \hat{y}_{i/i})^2$

Standard error of the estimate (s):

The standard error of the estimate (s) is:

$$s = \sqrt{\frac{RSS}{n - p'}}$$

Where RSS is the sum of the squares of the differences (residuals) between the experimental and estimated responses when predictions are made for objects in the training set ($RSS = \sum_i (y_i - \hat{y}_i)^2$), p' is the number of model variables plus one, and n the number of objects used to calculate the model.

Test set:

It is a set of chemicals not present in the training set, and selected iteratively from the available complete data set to assess the predictive ability of a (Q)SAR. In internal validation, the test sets have different compositions depending on the applied method of cross-validation (LOO, LMO, bootstrap). The chemicals in test set must be representative of the same chemical domain of the training set.

Training set:

A training set is a set of chemicals used to develop a QSAR. The chemical domain of the training chemicals is the domain of applicability of the developed model. The dimension of the training set must be sufficiently high to allow the finding of a reasonable *X-y* relationship.

Total Sum of Squares (TSS):

The total sum of squares (TSS) is the sum of the squares of the differences between the experimental responses and the mean values.

Validation set:

The validation set is a set of chemicals selected to validate a QSAR model for statistical external predictivity. It can be named prediction set. The chemicals are completely new for the developed QSAR model.

For the purpose of (Q)SAR validation, it is important that the chemicals in the validation set belong to the chemical domain of the training set, used for model development. It must contain a sufficient number of chemical structures (at least 20% of the total data set is the suggested number for small data set of 25-30 chemicals, the suggested dimension for bigger data sets (70-240 in this exercise) is 40-50%).

The Statistical Diagnostics summarized above have been applied to the QSAR models planned for the contract.

According to the **SETUBAL PRINCIPLES**:
A QSAR should:

- be associated with a defined endpoint
- take the form of an unambiguous and easily applicable algorithm;
- ideally, have a mechanistic interpretation;
- be accompanied by a definition of domain of applicability
- be associated with a measure of goodness-of fit (internal validation);
- be assessed in terms of its predictive power by using data not used in the development of the model (external validation).

In this exercise, the principles of validation (both internal and statistical external) and of applicability domain were verified for all the selected models. All the models have an unambiguous and applicable algorithm, and a defined endpoint was modelled. The mechanistic basis is not considered, being beyond the contract aim.

4. STATISTICAL VALIDATION OF SOIL ADSORPTION COEFFICIENT (K_{OC}) MODEL (Wei et al.)

Wei, D.B., Wu, C.D., Wang, L.S. and Hu, H.-Y. (2003). QSPR-based prediction of absorption of halogenated aromatics on yellow-brown soil. ***SAR and QSAR in Environmental Research*** **14 (3), 191-198.**

The soil adsorption coefficient (Koc) of 28 halogenated aromatics (Table in Annex) is studied.
Two Ordinary Least Squares (OLS) models are proposed: one logKow-based (1) and one based on theoretical molecular descriptors (2). The published models and the related statistical parameters are:

(1) log Koc= 2.607 + 0.071 log Kow

n=28 r=0.195 r^2=0.039 SE=0.369 F=1.025 p=0.321

This model is completely unsatisfactory, already in the fitting. As stated, also by the authors, Log Kow is an inadequate descriptor for soil adsorption in this data set.
The OLS model by theoretical molecular descriptors is:

(2) log Koc= 18.372 + 1.551 β - 2.998 α + 0.024 Vcse -15.436O_v +1.818 ^{3}Xc

n=28 R^2adj=96 SE=0.074 F=1.025 p<0.0001

The regression line of this model reported in the paper reveals a perfect alignment of the observed and "predicted" (as written by the authors) Koc values. The validation performed in the paper is by Monte Carlo simulation and a modification of LOO (Jackknife): the authors conclude that "the model is so robust that it could be used to predict the absorption behaviour as well as to investigate the adsorption mechanism to a certain extent".

Actually, the regression coefficients and the intercept of the equation (2) are not reproducible by OLS. The authors, who we contacted, gave no reasonable explanation. The new OLS equation, recalculated by us on the molecular descriptors selected by the authors, is:

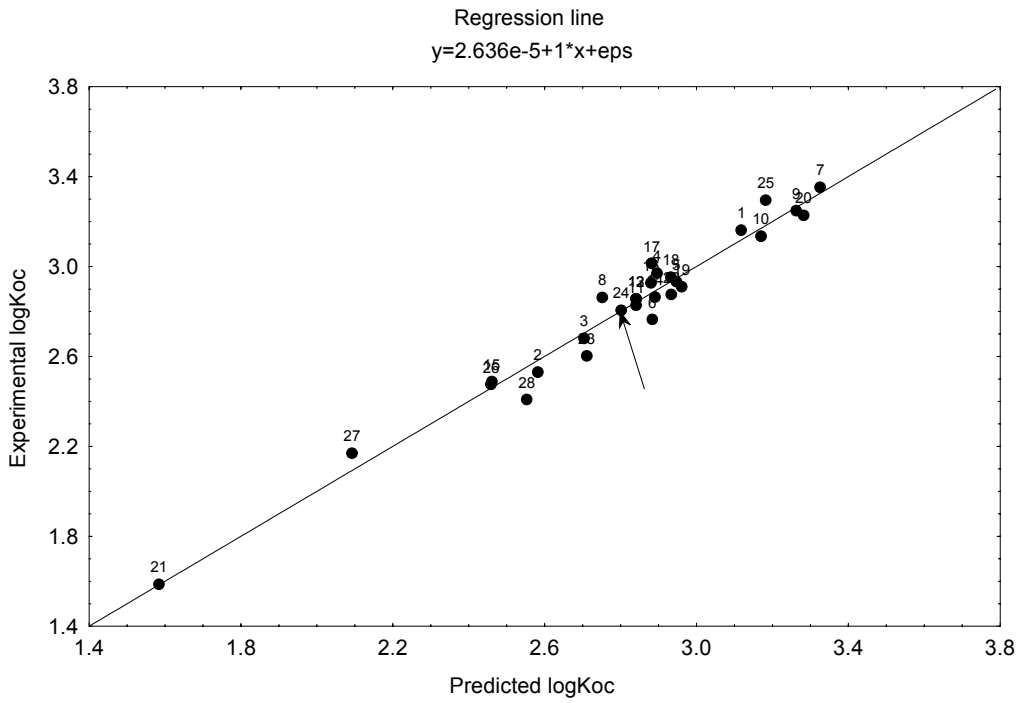
(3) log Koc= 13.22 + 1.15 β - 2.49 α + 0.02 Vcse -10.99 O_v + 1.80 ^{3}Xc

n=28 R^2=84.8 R^2adj=81.3 Q^2= 64.9 SDEC= 0.142 SDEP= 0.215 F=24.5

In this recalculated equation we report only two significant numbers for the coefficients in regression equations as this is better representative of the accuracy of the original data.

The regression line of our recalculated equation and the Williams plot are reported in Figure 1: 1 heavy outlier (24 , Pentachloro-phenol), not evidenced by the authors, is present. No highly influential chemicals are evidenced by the leverage approach. The Principal Component Analysis of the structural descriptors have been also performed in order to highlight the distribution of the chemicals in the structural space of the model descriptors and the eventual anomalous or isolated chemicals: the distribution in this case is substantially homogeneous.

As in the regression line of the published paper there were no outliers, we interpreted the difference (which we found) in the regression coefficients by a simple fitting of the data, performed by the authors.



It is important to note that the distribution of the responses in this model is not uniform: chemical 21 is isolated from other chemicals in that it has a significatively smaller Koc value. Note that an isolated datum has marked influence on the regression: the removal of such a chemical would completely change the model. Thus this model is unstable, being too determined by one single piece of chemical information.

VALIDATION:

The model (3) has been assessed, in the present contract work, by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Though small sized, this data set underwent statistical external validation by comparing different approaches for the preliminary splitting of the

chemicals into training (21 chemicals) and validation sets (7 chemicals) (D-optimal Distance, Kohonen-ANN; random). The small size of the data set did not allow a more drastic splitting: in order to maintain the necessary information for the modelling, the training chemicals must not be too few in number. Pentachlorophenol (24) was always included in the training set, but was always an outlier in each split model.

Table 1: Statistical Diagnostics of models

n tr	n valid	split	variables	Q^2	R^2	$Q^2_{LMO}50$	Q^2boot	R^2adj	Q^2ext	MSE tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
28			b a V_{CSE} O_v ^3X_C	64.9	84.8	54.4	59.5	81.3		0.020		0.215	0.142	24.5	0.160	50.4	52.3	1.9
21	7	K-ANN	b a V_{CSE} O_v ^3X_C	61.9	86.7	47.4	52.7	82.2	66.2	0.021	0.024	0.242	0.143	19.5	0.169	43.7	47.2	3.5
21	7	D-Optimal	b a V_{CSE} O_v ^3X_C	60.8	85.1	45.0	51.0	80.2	78.8	0.024	0.009	0.254	0.156	17.2	0.185	45.6	48.7	3.1
21	7	Random	b a V_{CSE} O_v ^3X_C	54.6	85.5	38.4	44.3	80.7	76.2	0.020	0.027	0.252	0.142	17.6	0.169	48.7	52.8	4.1

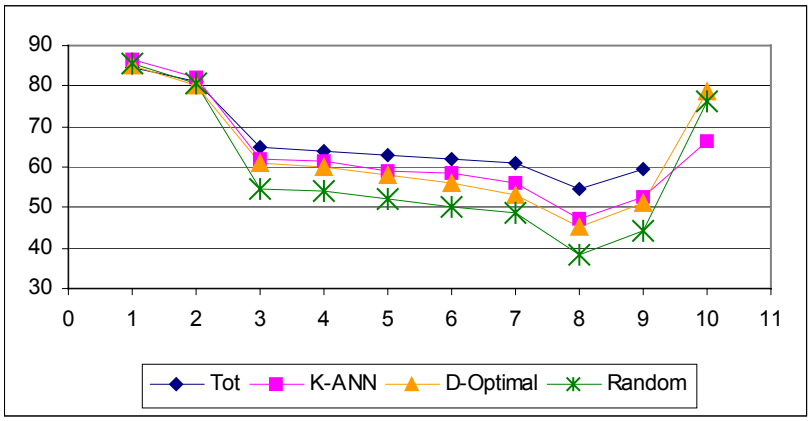
It is immediately evident from the Table analysis that the model, even with good fitting performances (high values of R^2 and R^2_{adj}), has unsatisfactory predictivity, verified by different internal validations. The LOO validation highlights that the model is unstable, overfitted, and poorly predictive, in fact the difference between R^2 and Q^2 is very high: 20%.

Regarding collinearity: while the authors stated that "it was proved that there was no serious multicollinearity among the parameters in Eq. 2", we verified that the descriptors are very correlated (medium Kxx: 47%) and most importantly the difference between the correlation among the block of the X variables and the response Y (Kxy) and the correlation among the X (Kxx) is very low (delta column, medium value: 3%) in comparison with other QSAR models and in our experience. This is a signal of multicollinearity without prediction power (in fact when we have applied the QUIK rule of Todeschini this model was excluded as predictive model). The model can be also considered in overfitting: actually 5 descriptors for 28 chemicals are probably too much, unnecessary terms are included to fit the data, but this is not useful for predictive purposes. The risk of chance correlation is verified also by the Y-scrambling procedure: in fact, some models obtained by Y-scrambling are of similar performances of the original model.

Table 1 bis: Statistical Diagnostics of models

		Total	K-ANN	D-Optimal	Random
1	R^2	84.8	86.7	85.1	85.5
2	R^2_{adj}	81.3	82.2	80.2	80.7
3	Q^2	64.9	61.9	60.8	54.6
4	Q^2_{LMO10}	64.1	61.3	59.8	54.2
5	Q^2_{LMO20}	63.1	59.0	57.9	52.0
6	Q^2_{LMO30}	61.9	58.5	56.0	50.1
7	Q^2_{LMO40}	61.0	56.1	53.2	48.7
8	Q^2_{LMO50}	54.4	47.4	45.0	38.4
9	Q^2_{boot}	59.5	52.7	51.0	44.3
10	Q^2ext		66.2	78.8	76.2

The following is the graphical representation of the parameters reported in the above table.

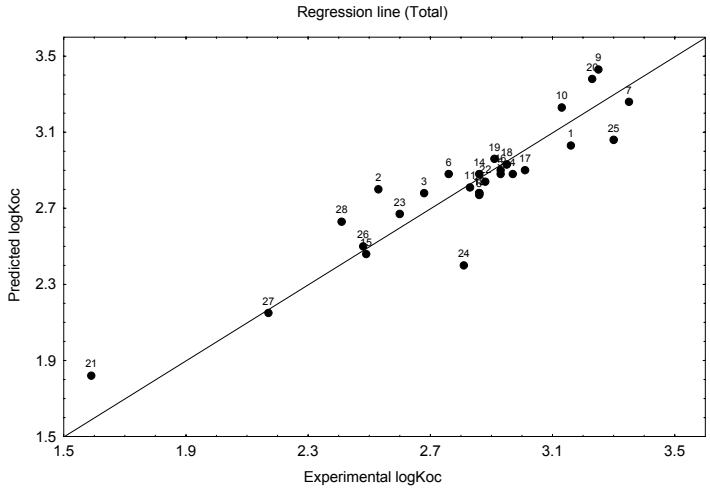


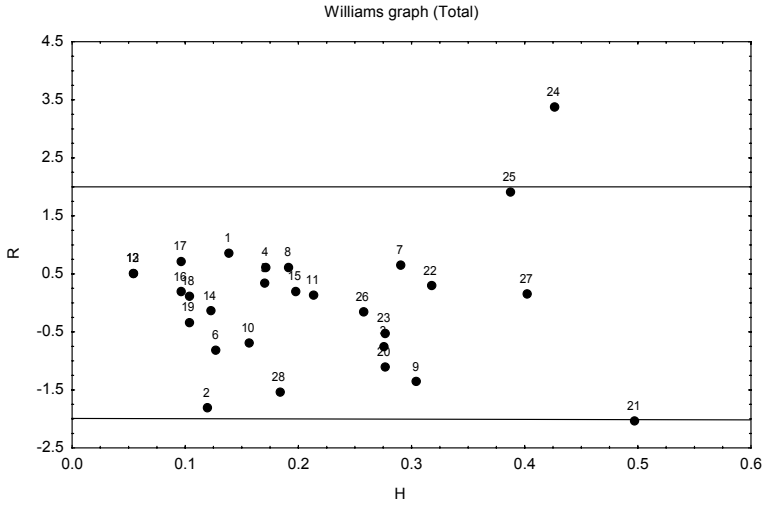
In the validation of this fitting model, the values of LOO and LMO (3-7) decrease slightly and regularly. Only when the perturbation is too high (50% relative to 14 chemicals in the test out of 28, point 8) the chemicals, on which the model is developed, give, in each run of perturbation, too little information regarding structural information useful for test chemicals. Thus the model appears less predictive (under-optimism), highlighted by the minimum in the graph.

In the case of a small data set (for all approaches), statistical external validation by preliminary splitting is not useful: the information obtained by Q^2ext is optimistic compared with the results of internal validation, and must be considered unreliable in its conclusions. The splitting is highly influential: in this case, the outlier 24 is always put into the training set, thus statistical external validation appears best in performance.

Validation by bootstrapping (9) gives intermediate and more realistic results regarding the internal predictivity of the model. This validation could be a good compromise for this kind of relatively small data set (20-30 chemicals).

FIGURE 1: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





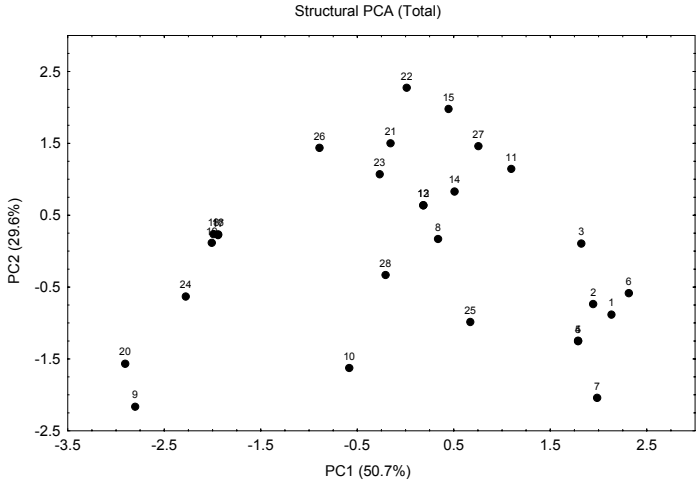
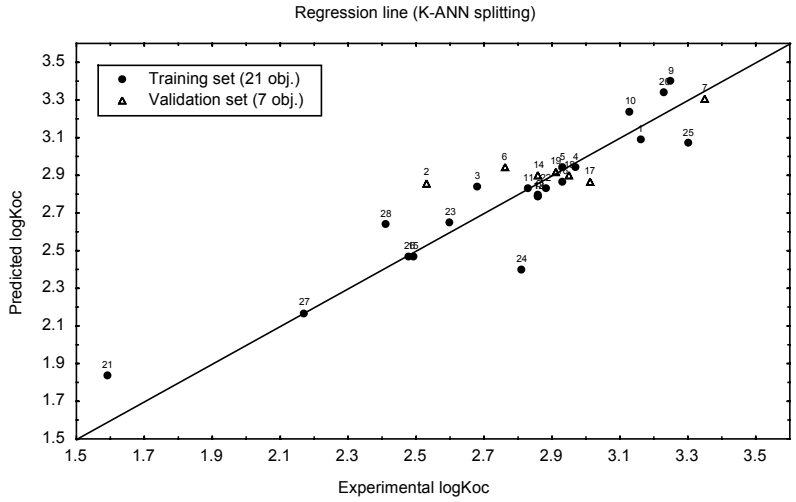
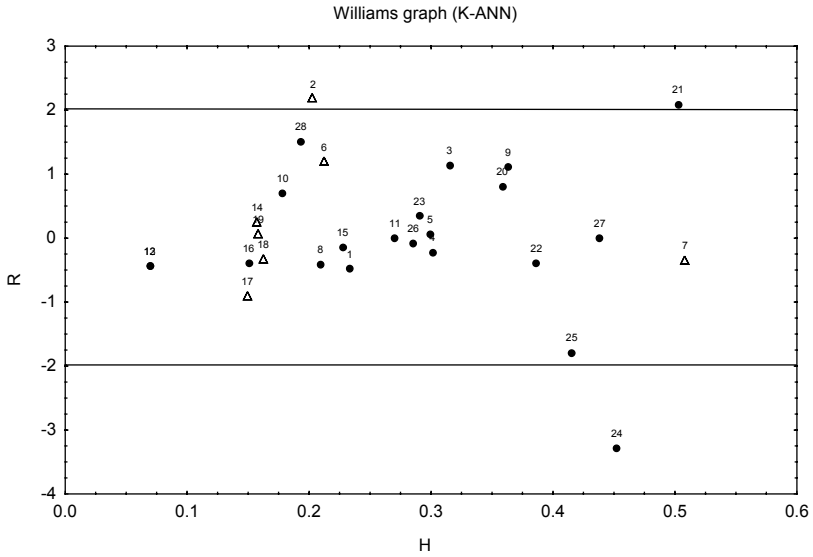


FIGURE 2: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





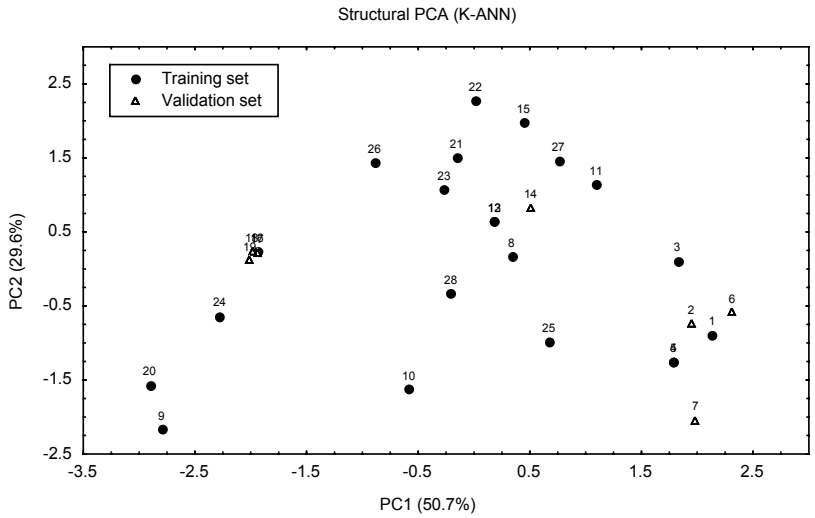
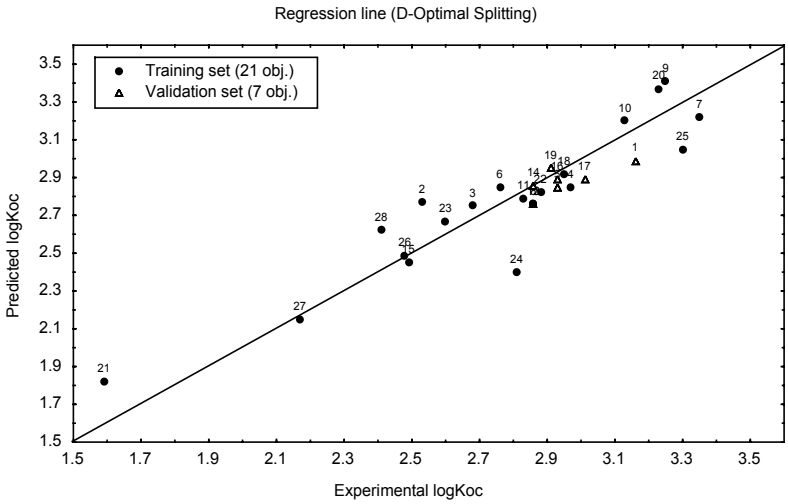
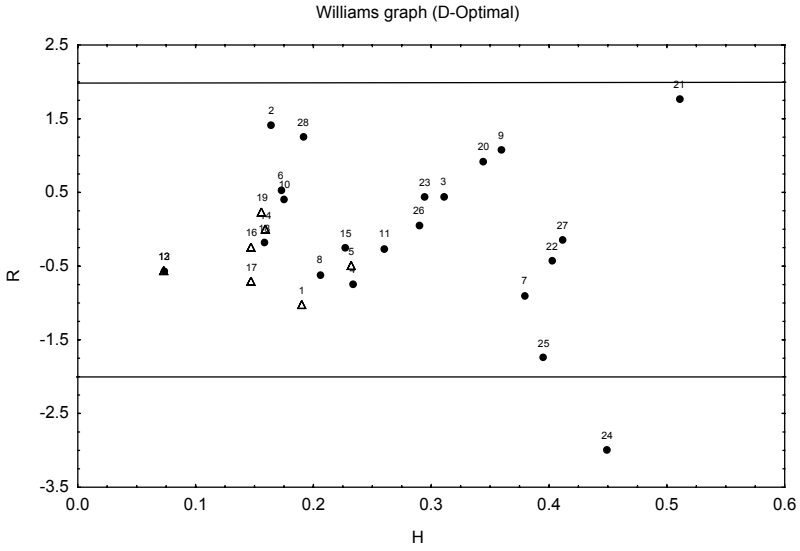


FIGURE 3: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





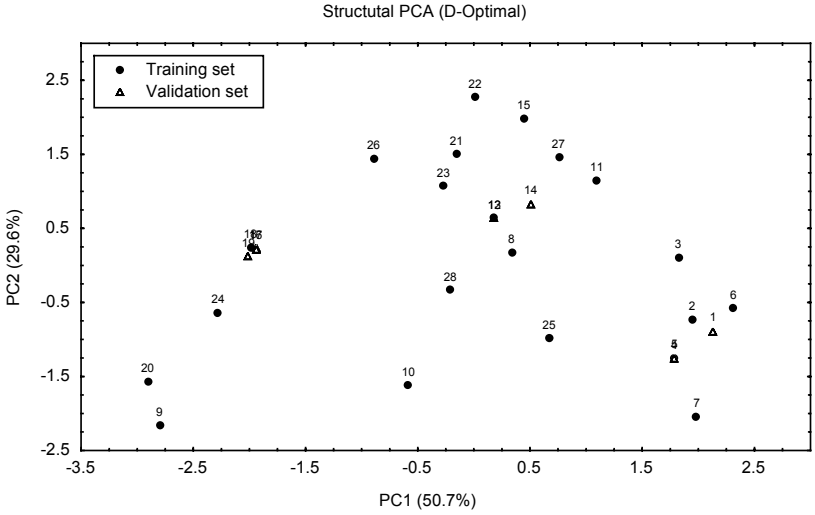
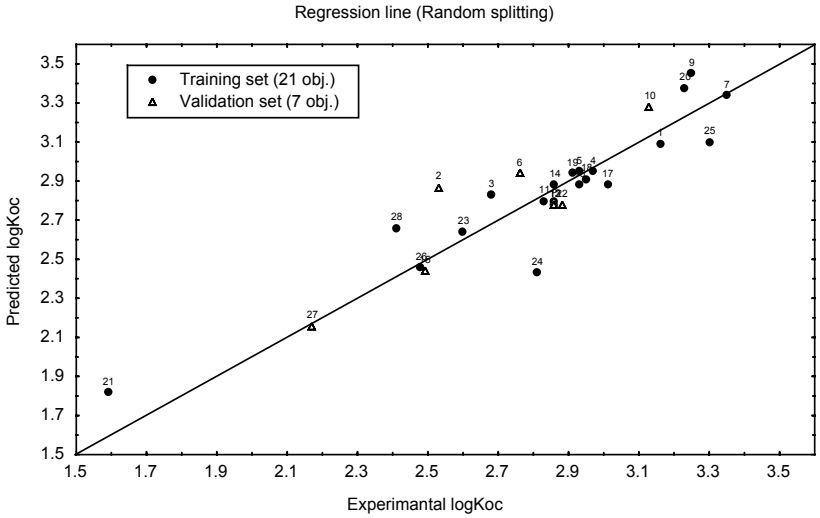
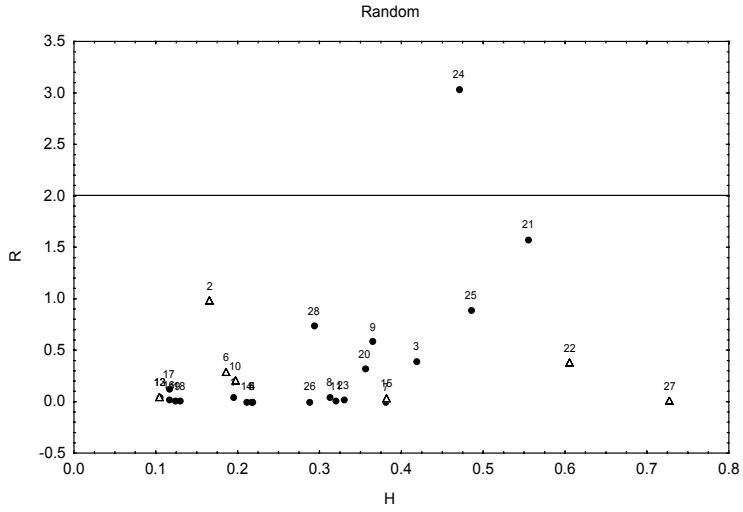
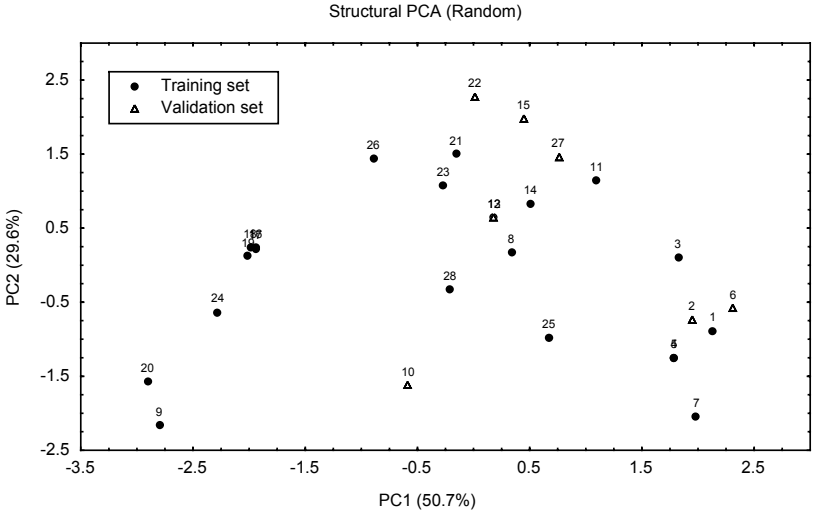


FIGURE 4: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







CONCLUSIONS for VALIDATION

The published model is not reproducible. The regression line in the published paper does not evidence a strong outlier, which is present in the real OLS model, redeveloped in this contract.

The new model obtained on the studied data set and by the descriptors selected by the authors has good fitting performances, but it is not robust, it is overfitting and without predictive power.

The internal validation parameters (Q^2_{LOO}, Q^2_{LMO} and Bootstrap) are significantly lower (>20%) than the fitting parameters.

The model is unstable, also because the response of one chemical (21) is too influential (isolated in the inferior part of the regression line): the model is "heavily driven" by this chemical.

In addition, it is based on multicollinear descriptors (Kxx) and without any significant increase in the correlation between the highly correlated block of descriptors and the response (Kxy-Kxx).

Also the Y-scrambling procedure highlights the risk of chance-correlation.

The applied statistical external validation by preliminary data splitting using D-optimal Design, Kohonen Maps-ANN and Random gives contradictory results, compared with the internal validation, and is too optimistic regarding apparent predictivity. In fact, with small data sets (20-30 chemicals), internal validation (LMO or bootstrap) is the only reasonable way of validation. Statistical external validation by splitting the available data set is too dependent on the splitting (here one outlier is always put into the training set), and is thus not useful. For completely new chemicals external predictivity can only be verified *a posteriori,* case-by-case.

As always, but particularly in this case, the inspection of the chemical domain of applicability is essential.

The presence of outliers must always be verified and their removal declared and possibly commented on.

Shortcomings of the model:

1. Not reproducibility
2. 1 strong outlier, not evidenced
3. Unbalanced data set (1 isolated response)
4. Overfitting
5. Too high correlation among the descriptors
6. Chance correlation
7. Not predictivity

5. STATISTICAL VALIDATION OF BIOCONCENTRATION FACTOR (BCF) MODEL (Gramatica and Papa)

Gramatica, P. and Papa, E. (2003). QSAR modeling of bioconcentration factor by theoretical molecular descriptors. ***QSAR & Combinatorial Science*** **22 (3), 374-385**

A data set of BCF in fish for 238 non-ionic organic compounds (Table in Annex), available in the literature, was studied. The entire data set and a reduced training set, obtained by Experimental Design (D-optimal distance) for statistical external validation of the model were modelled.
The published Ordinary Least Squares (OLS) models, by Genetic Algorithm variable subset selection, are:

$$\log BCF = - 18.87 + 1.68IDDM - 0.51nHAcc + 17.09MATS2m – 0.40GATS2e$$
$$n = 238 \quad R^2=81.6\ \% \quad Q^2 = 80.8\ \% \quad Q^2_{LMO}(25\%) = 80.7\ \% \quad Q^2_{LMO}(50\%) = 80.5\ \%$$
$$s=0.59 \quad F_{233} =257.84 \qquad SDEP= 0.60 \quad SDEC=0.59$$

$$\log BCF = -17.58 + 1.69IDDM - 0.45nHAcc + 15.65MATS2m - 0.36GATS2e - 1.64H6p$$
$$n = 179 \quad R^2=79.5\ \% \quad Q^2 = 78.0\ \% \quad Q^2_{LMO}(25\%) = 77.9\ \% \qquad Q^2_{LMO}(50\%) = 77.3\ \%$$
$$Q^2_{EXT} = 88.0\% \quad s=0.59 \quad F_{172} =134.52 \qquad SDEP= 0.60 \quad SDEC=0.58$$

The models were assessed by the authors, both internally (cross-validation by LOO, LMO up to 50%) and externally, on 59 selected new chemicals (D-optimal distance). The highest value of Q^2_{ext} was the parameter chosen by the authors for the selection of the best model.

VALIDATION:

In this contract work, the models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. In addition to the different internal validation approaches, statistical external validation was performed by comparing different approaches for the preliminary splitting of chemicals into training and validation sets (D-optimal Distance, Kohonen-ANN; random). The results are reported in the following Tables and graphs.
The regression lines and the Williams plot for the different splittings are reported below, with the PCA of structural descriptors to verify the distribution of the two sets regarding structural information. Some

common outliers, together with outliers present in a specific splitting, and some highly influential chemicals, depending on the applied splitting methodology, are highlighted.

Table 2: Statistical Diagnostics of BCF models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
238		Totale	IDDM MATS2m GATS2e H6p nHAcc	81.8	82.8	81.3	81.3	82.5	/	0.570		0.584	0.568	223.6	0.576	21.93	33.83	11.90
118	120	D-Optimal (a)	IDDM MATS2m GATS2e H6p nHAcc	77.8	80.2	76.5	76.6	79.3	83.8	0.345	0.332	0.623	0.588	90.9	0.603	25.36	35.34	9.98
179	59	D-Optimal (b)	IDDM MATS2m GATS2e H6p nHAcc	78.0	79.5	77.3	77.3	79.1	88.0	0.333	0.330	0.601	0.580	134.5	0.588	23.68	34.48	10.80
123	115	K-ANN (a)	IDDM MATS2m GATS2e H6p nHAcc	78.0	80.5	76.6	76.7	79.7	83.5	0.290	0.379	0.572	0.539	96.6	0.552	26.38	33.81	7.43
170	68	K-ANN (b)	IDDM MATS2m GATS2e H6p nHAcc	79.1	80.7	78.2	78.4	80.1	85.6	0.298	0.401	0.570	0.549	136.1	0.559	24.42	34.26	9.84
118	120	Rand. (a)	IDDM MATS2m GATS2e H6p nHAcc	83.4	85.3	82.2	82.5	84.6	80.4	0.323	0.356	0.603	0.568	129.0	0.583	19.44	32.76	13.32
179	59	Rand. (b)	IDDM MATS2m GATS2e H6p nHAcc	77.1	78.8	33.2	76.2	78.1	88.2	0.315	0.373	0.583	0.561	127.6	0.571	23.25	33.94	10.69

Regarding collinearity: the descriptors are not very correlated (medium Kxx=23) but, most importantly, the difference in the correlation between the block of X variables plus the response Y (Kxy) and that of X (Kxx) is sufficiently high (medium delta: 11) compared with other QSAR models and according to our experience.

All the models were also verified by Y-scrambling: the models on randomised response have all extremely low R^2 and Q^2 compared with the published models. This is a demonstration that the reported models are not obtained by chance correlation.

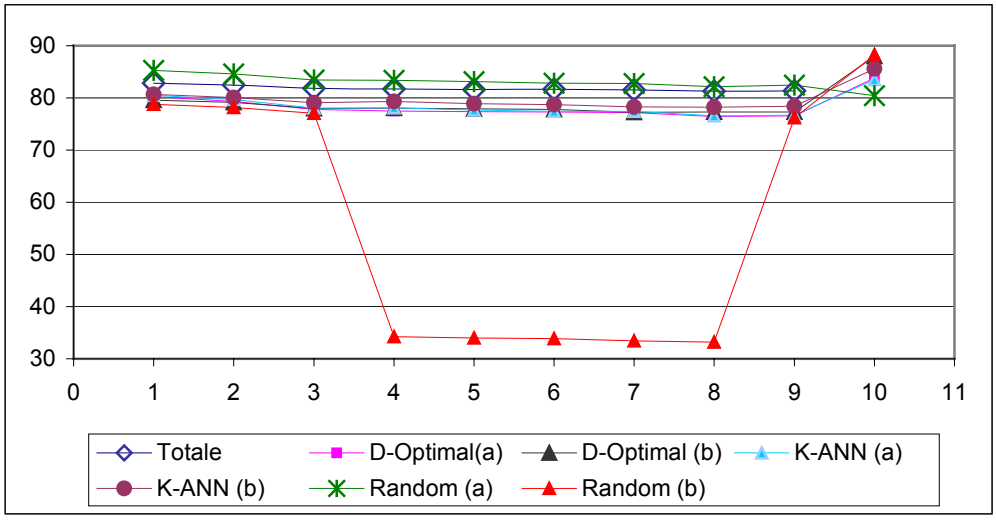
The MSE (mean squared error) values for the training and validation sets are similar, thus demonstrating that the models are able to predict BCF for chemicals not used in the model development (validation set) just as they do for chemicals used to find the relationship (training set).

An analysis was made of the results of the fitting and prediction parameters, on the data reported in the following table and plotted in the graph:

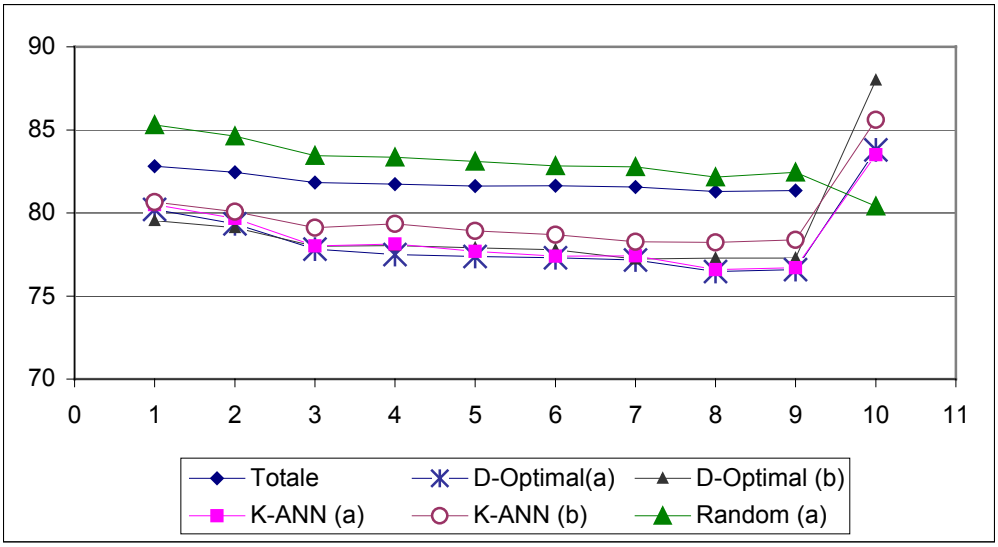
Table 2 bis: Statistical Diagnostics of BCF models

		Total	D-Optimal(a)	D-Optimal (b)	K-ANN (a)	K-ANN (b)	Random (a)	Random (b)
1	R^2	82.8	80.2	79.5	80.5	80.7	85.3	78.8
2	R^2_{adj}	82.5	79.3	79.1	79.7	80.1	84.6	78.1
3	Q^2	81.8	77.8	78.0	78.0	79.1	83.4	77.1
4	Q^2_{LMO10}	81.7	77.5	78.0	78.1	79.4	83.4	34.2
5	Q^2_{LMO20}	81.6	77.4	77.9	77.7	78.9	83.1	34.0
6	Q^2_{LMO30}	81.6	77.3	77.8	77.4	78.7	82.8	33.9
7	Q^2_{LMO40}	81.6	77.2	77.2	77.4	78.3	82.8	33.4
8	Q^2_{LMO50}	81.3	76.5	77.3	76.6	78.2	82.2	33.2
9	Q^2_{boot}	81.3	76.6	77.3	76.7	78.4	82.5	76.2
10	Q^2ext		83.8	88.0	83.5	85.6	80.4	88.2

The following is the graphical representation of the parameters reported in the above table.



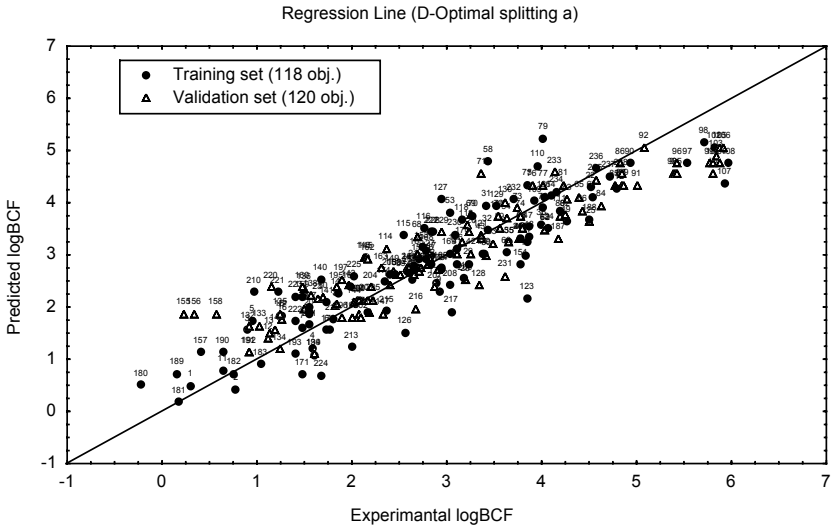
An extension of the upper part of the graph is reported:

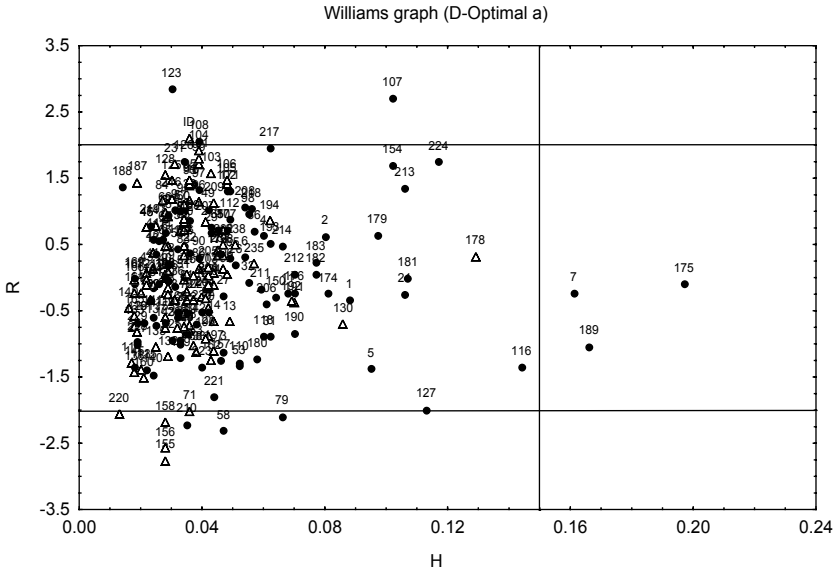


The models are robust: the difference between R^2 and Q^2 is small for D-optimal and K-ANN splittings (1-2%), in every splitting. The proposed models show satisfactory and stable performance both in internal and external validations, the only exception being one random splitting (b). In this splitting the distribution of the chemicals in the training and validation sets is so unbalanced that the internal validations by LMO are greatly influenced, while bootstrapping confirms the internal predictivity and stability of the model; as in other examples bootstrapping appears the best parameter for this internal validation.

Statistical external validation gives similar results for the model predictivity of validation chemicals in every splitting: in particular, the D-optimal splitting always gives the most optimistic idea of model predictivity, random splitting appears variable and K-ANN splitting falls in the medium.

FIGURE 5: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





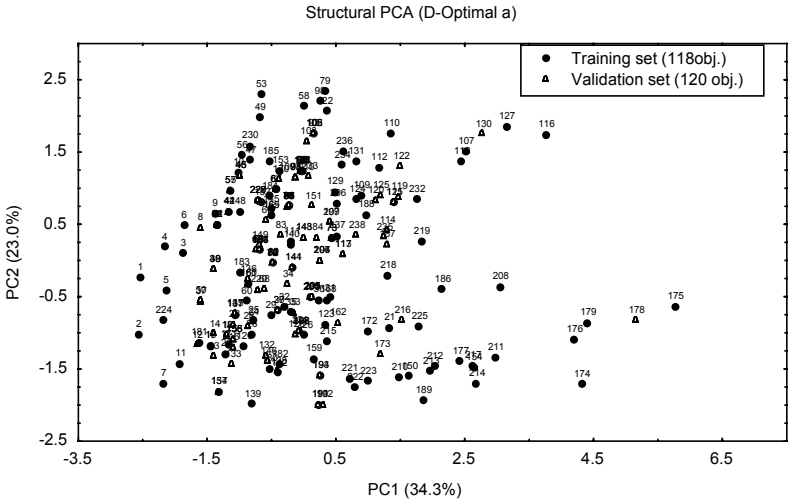
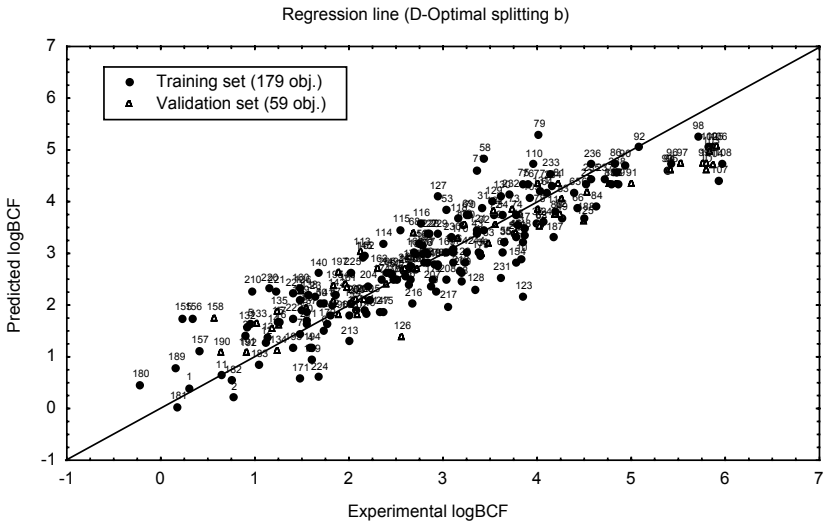
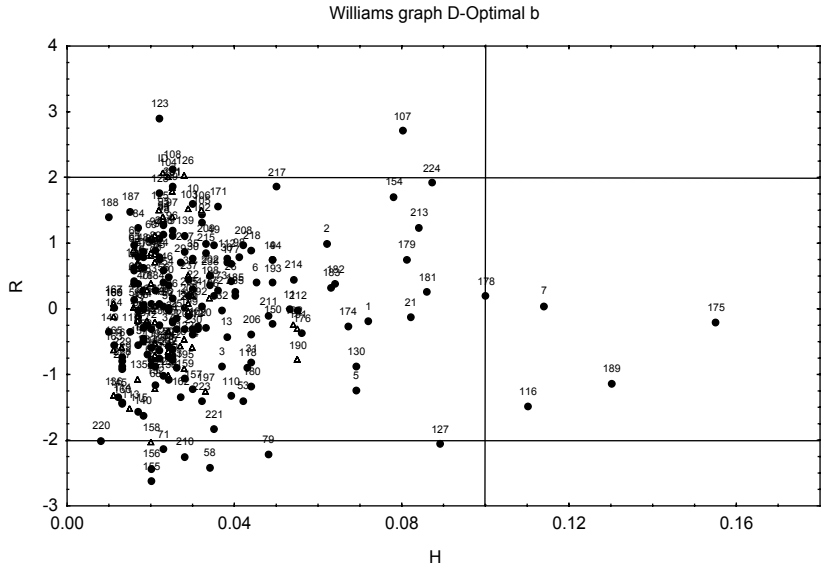


FIGURE 6: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





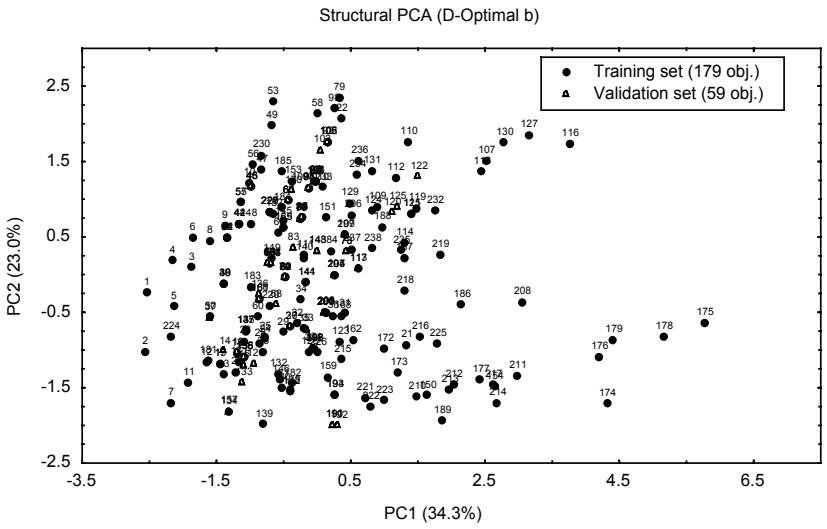
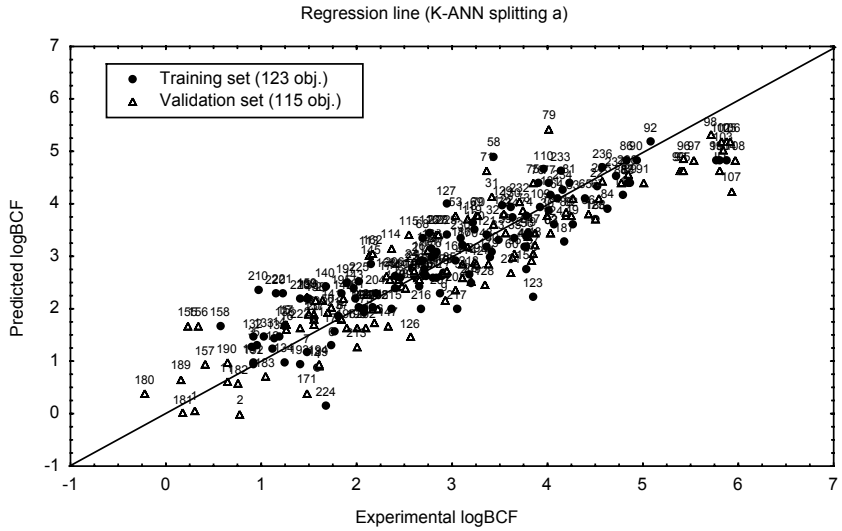
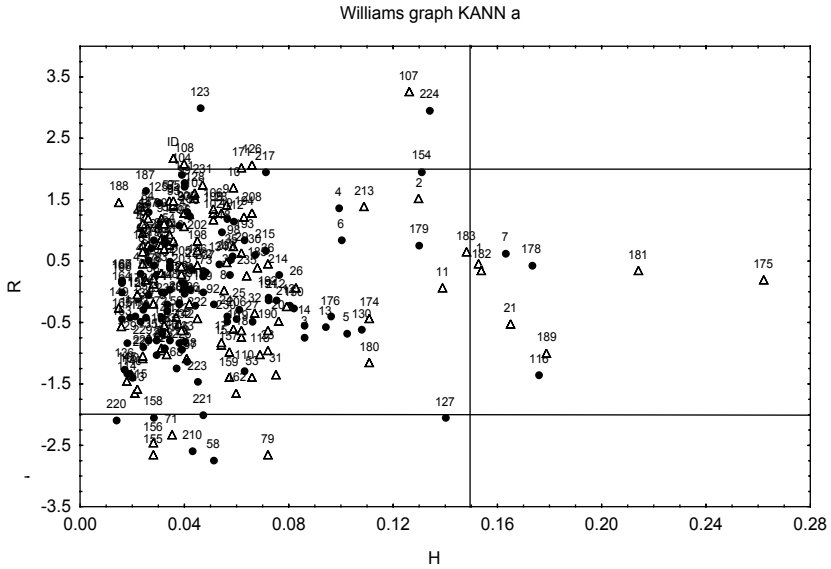


FIGURE 7: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





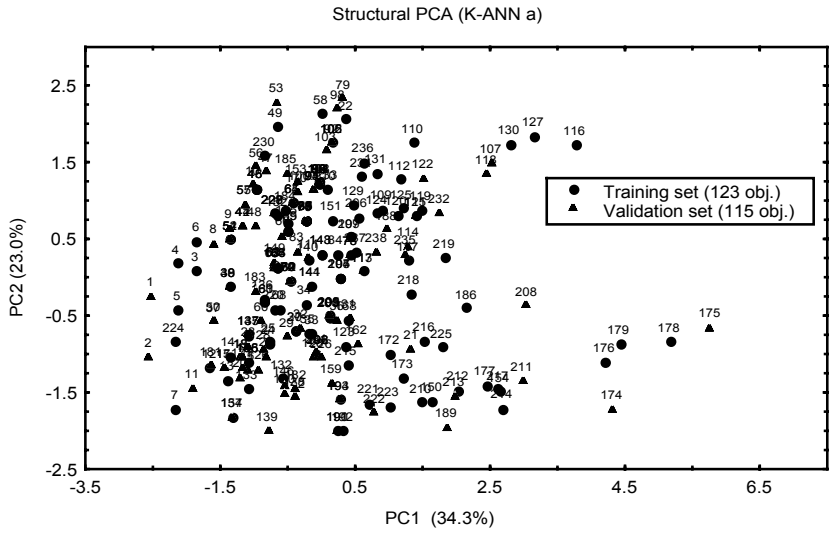
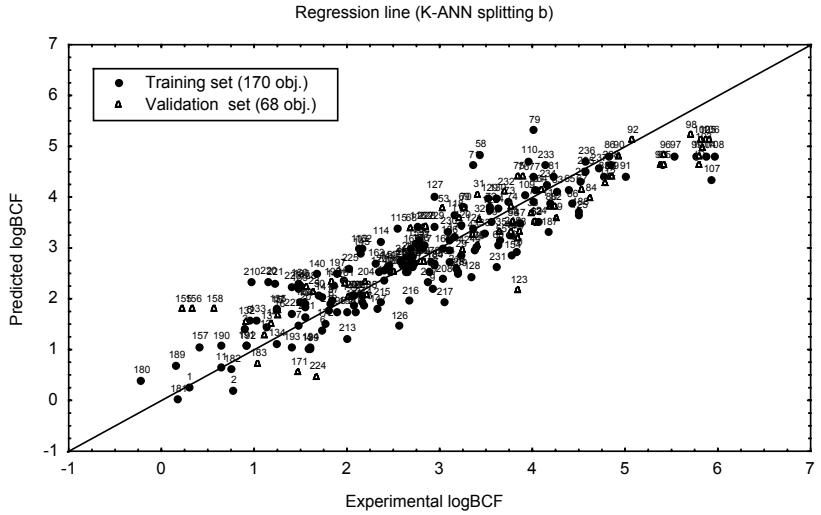
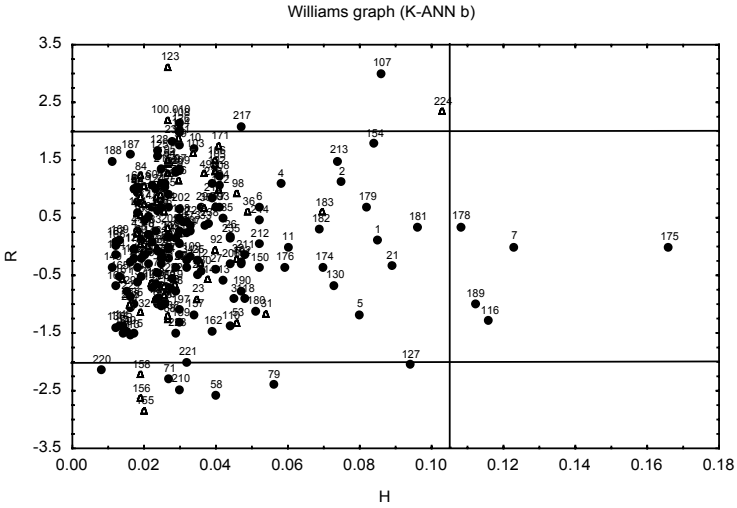


FIGURE 8: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





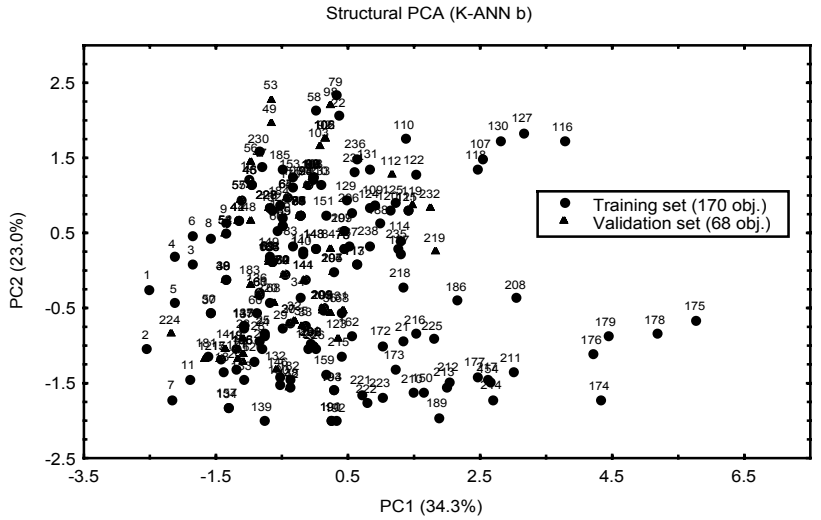
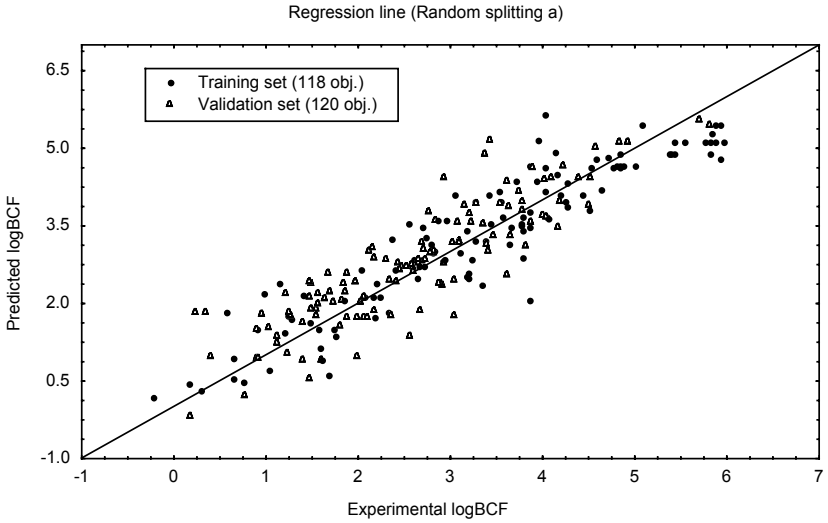
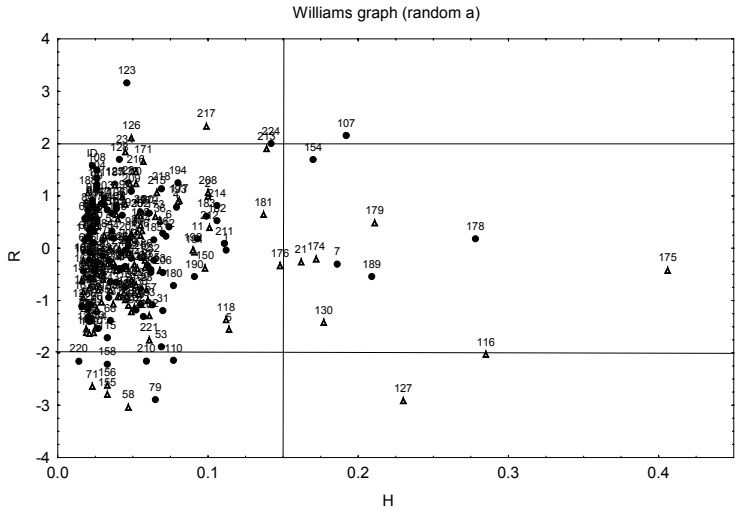


FIGURE 9: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





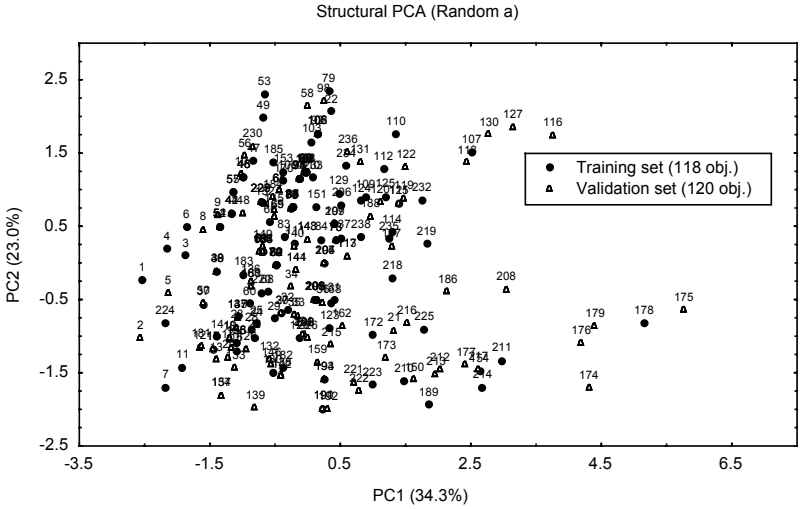
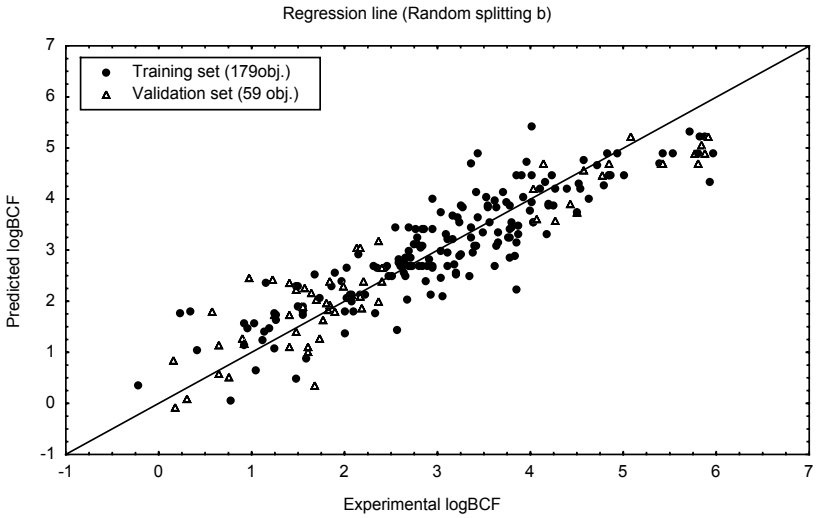
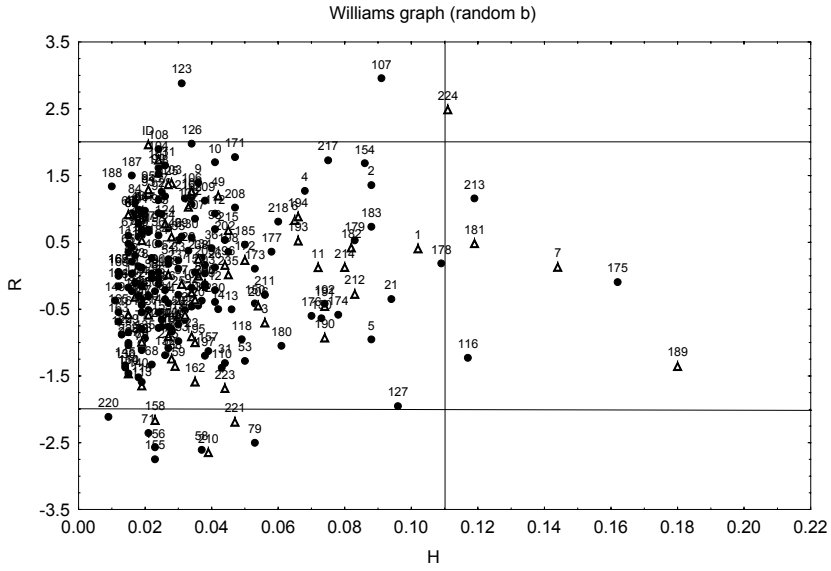
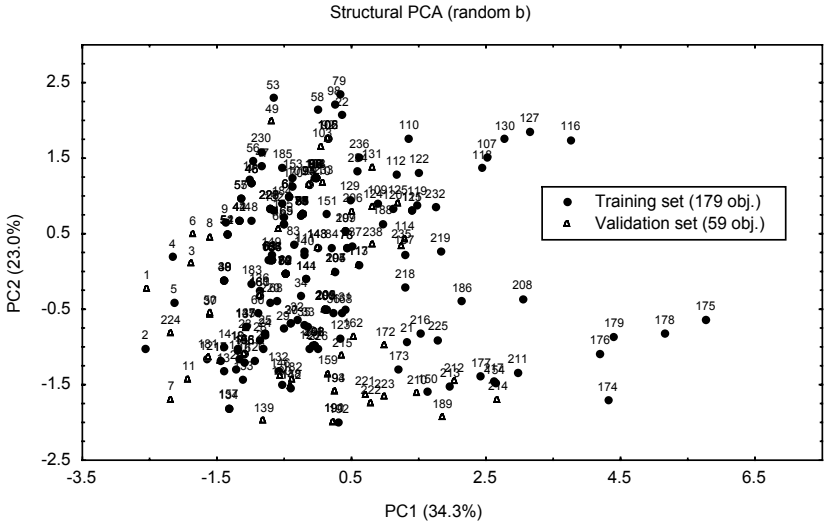


FIGURE 10: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







MAIN CONCLUSIONS FOR VALIDATION

The models published in this paper are stable, robust, with good fitting and predictive performance. They are predictive for the chemicals used in the model development (internal validation on training chemicals) and also for chemicals not used in the model development (statistical external validation on validation set chemicals). The available data set was split into training and validation sets by different splitting methodologies (D-optimal, K-ANN and random) and the predictivity of the models was verified in each splitting.

Internal validation by LMO (up to 50%) or, preferably, by **bootstrap** are the recommended approaches to verify the real internal predictivity for chemicals in the data set. Y-scrambling gives information regarding the exclusion of chance models. It is important to note that these approaches are internal validations as the information related to each chemical in the data set is considered in at least one run of the validation process: these chemicals are never new chemicals in the model development and their information is included. On the contrary, statistical external validation verifies the predictivity for chemicals not used in the model development, thus the information of these new chemicals is not included in the modelling.

In relation to this point, it must be noted that the real utility of this approach (splitting of the available data set and Q^2_{EXT} determination) particularly concerns model development.

As applied by the authors in this paper, the selection of the best models, from among one hundred possible models developed by Genetic Algorithm Variable Subset Selection applied to OLS modelling, was done by maximizing the **external predictivity** on the split validation set: it is very important to note that in the population of possible models, all with high values of R^2 and Q^2 (apparently good models), not all the models had high Q^2_{EXT} values. Some of the models with high fitting and internal predictivity (even higher than the reported models) had a low Q^2_{EXT} value (<50%).

This is a clear demonstration that internal validation is a necessary, but not sufficient validation procedure. The real predictivity of a model for chemicals not used for model development must be verified by statistical external validation. This is particularly true in model development by variable subset selection procedures.

Splittings: random splitting should be avoided for statistical external validation as the results of validation are strongly dependent on the training and validation set composition. A careful splitting into representative sets must be performed by suitable methodologies.

We have verified that D-optimal design and Kohonen Maps-Artificial Neural Networks are good splitting methodologies, giving different results in relation to the real external predictivity of the model: in general,

D-optimal distance is the more optimistic, the validation set being included in the training set, while K-ANN is the more balanced, the distribution of the chemicals in the training and validation sets being more balanced.

Shortcomings of the models:

No shortcomings found: the reason is that the statistical approaches, evaluated in this exercise, including the external validation for the choice of the best predictive model, were applied by the authors during the model development.

It is again verified that this is the best way to propose QSAR models with reliable predictions.

6. STATISTICAL VALIDATION OF ECOTOXICITY MODELS (Kulkarni et al.)

Kulkarni, S.A., Raje, D.V. and Chakrabarti, T. (2001). Quantitative structure-activity relationships based on functional and structural characteristics of organic compounds. *SAR and QSAR in Environmental Research* 12, 565-591.

The acute toxicity to *P. promelas* (LC_{50}) of a data set of 247 heterogeneous organic chemicals (Table in Annex), including priority pollutants, was studied. The chemicals are grouped in groups, as homogeneous as possible (benzenes, alcohols, aldehydes, aliphatics, esters, ketones and phenols), and modelled separately.

Five molecular descriptors (4 physico-chemical characteristic and 1 connectivity index) were selected as input descriptors by the authors and some of them used in the different models, checking the correlation with two approaches (condition index and K correlation index of Todeschini).

BENZENES:

The data set is composed of 38 benzenes, but 6 are outliers and removed from the modelling by the authors.

The proposed models and the reported statistical parameters are:

(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, in fact no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 0.088(7)\alpha - 0.507(7)\beta + 0.670(6)\log Kow + 0.522\pi^* - 1.280(7)$

$n= 32 \quad R^2= 91.6\% \quad R^2_{adj} = 90.35\% \quad S.E.= 0.2578$

(2) $\log(1/LC50) = -0.470(6)\beta + 0.665(5)\log Kow + 0.522\pi^* - 1.26(4)$

$n= 32 \quad R^2= 91.58\% \quad R^2_{adj} = 90.68\% \quad S.E.= 0.2534$

The statistical parameters reported by the authors are only related to the fitting performances that are very good.

The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figure 11 and 12. The authors did not point out that chemical 28 is an outlier, 32 is an influential

chemical in both the models and 16 only in model (1): these are evidenced in this contract work by the leverage approach.

The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors and any possible anomalous or isolated chemicals (32 is actually quite isolated from the other chemicals in the set.)

The authors performed external validation using two different test sets : a) 7 chemicals structurally similar to training chemicals; b) 11 chemicals with additional functional groups (for instance di- and tri-nitro).

The authors report only the predicted values in the paper, but we calculated the corresponding Q^2ext values and report them in the following table.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (7 chemicals) (D-optimal Distance, Kohonen-ANN; random) was also performed. The PCA of the structural descriptors to verify the distribution of the two sets regarding the structural information are reported below. The influential chemicals (16 and 32) are put into the training set in each following splitting.

Table 3: Statistical Diagnostics of models

n Tr	n valid	Split	Variables	Q^2	R^2	Q^2LMO50	Q^2boot	R^2adj	Q^2ext	Q^2ext (b)	MSE train	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
32	Test 7 or 11	Tot 1	π* β α logKow	87.9	91.5	76.8	80.3	90.3	a)89.0	-3.9	0.057		0.285	0.238	73.0	0.259	46.3	57.2	11.0
32	Test 7 or 11	Tot 2	π* β logKow	88.9	91.5	85.5	86.7	90.6	a)89.0	-3.9	0.057		0.273	0.238	100.8	0.254	36.9	56.2	19.3
25	7	K-ANN 1	π* β α logKow	85.8	92.5	68.9	73.4	91.0	86.1		0.050	0.095	0.308	0.224	61.7	0.250	48.9	59.5	10.6
25	7	K-ANN 2	π* β logKow	87.1	92.5	82.9	83.9	91.4	86.0		0.051	0.095	0.294	0.224	86.0	0.245	40.7	59.0	18.4
25	7	D-Opt. 1	π* β α logKow	89.5	93.4	76.7	80.4	92.0	44.1		0.054	0.075	0.293	0.233	70.4	0.261	46.4	57.3	10.9
25	7	D-Opt. 2	π* β logKow	90.6	93.4	85.7	88.2	92.4	44.0		0.054	0.075	0.277	0.233	98.3	0.254	37.9	57.0	19.1
25	7	Rand. 1	π* β α logKow	88.1	93.1	68.9	74.2	91.7	80.3		0.054	0.086	0.304	0.232	66.1	0.260	49.1	59.3	10.3
25	7	Rand.2	π* β logKow	89.9	93.0	86.4	87.6	92.0	80.4		0.054	0.083	0.280	0.233	91.5	0.260	40.3	58.7	18.4

The models demonstrate a satisfactory stability in internal validation.

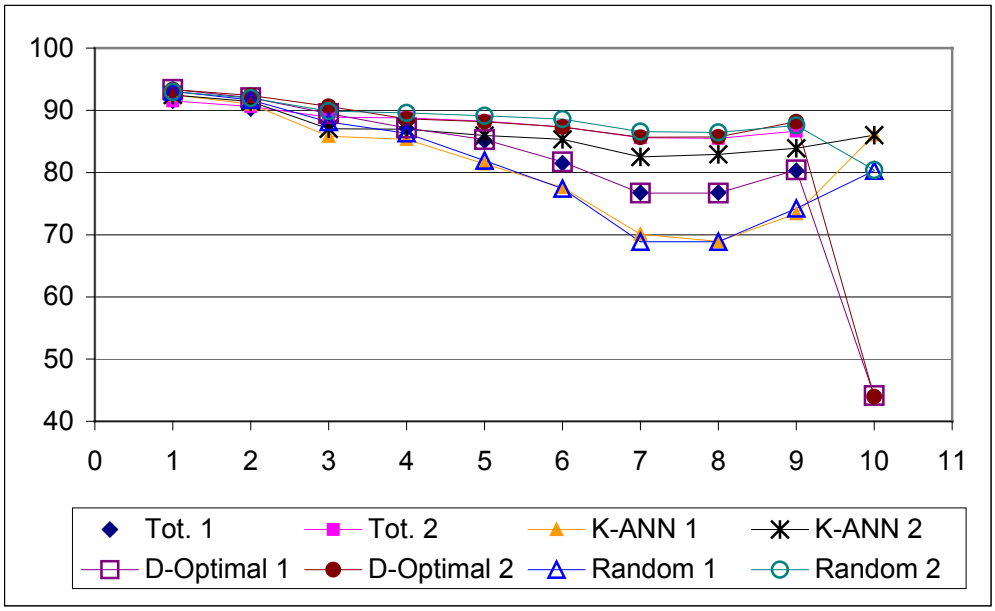
SDEP is similar to SDEC: the models have internal predictivity not too dissimilar from fitting power.
The models with 3 descriptors are always the more stable and internally predictive (less difference between Q^2_{LMO50} and Q^2_{LOO}), the addition of one descriptor does not improve the predictivity, only the fitting. Statistical external validations confirm the satisfactory prediction ability for chemicals included in the chemical domain of the training sets (set a). On the contrary , the models are not predictive for the 11 chemicals out of the domain (case b, with negative values of Q^2_{ext}) (see explanation in the Appendix on Leverage approach). The value of Q^2_{ext} = 44% for statistical external validation on the splitting by D-optimal is an apparent result, it is actually an under-estimation of the predictivity In this case, the MSE value must be considered. In fact, the MSE values for the training and validation sets are similar, thus demonstrating that the models are able to predict the response for chemicals not used in the model development (validation set) just as they do for chemicals used to find the relationship (training set).

The models demonstrate a satisfactory stability in internal validation.
Regarding collinearity: in general, the descriptors are very correlated (medium Kxx= 43) but, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy) and the correlation among the X (Kxx) is sufficiently high (medium delta= 15) compared with other QSAR models, and according to our experience.
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.
An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted :

Table 3 bis: Statistical Diagnostics of models

	Tot. 1	**Tot. 2**	**K-ANN 1**	**K-ANN 2**	**D-Optimal 1**	**D-Optimal 2**	**Random 1**	**Random 2**
R^2	91.5	91.5	92.5	92.5	93.4	93.4	93.1	93.0
R^2_{adj}	90.3	90.6	91.0	91.4	92.0	92.4	91.7	92.0
Q^2	87.9	88.9	85.8	87.1	89.5	90.6	88.1	89.9
Q^2_{LMO10}	87.1	88.8	85.3	87.0	87.2	88.6	86.3	89.6
Q^2_{LMO20}	85.1	88.2	81.4	86.0	85.3	88.2	81.9	89.1
Q^2_{LMO30}	81.5	87.3	77.6	85.3	81.7	87.3	77.4	88.6
Q^2_{LMO40}	76.8	85.6	70.1	82.5	76.7	85.7	68.9	86.6
Q^2_{LMO50}	76.8	85.5	68.9	82.9	76.7	85.7	68.9	86.4
Q^2_{boot}	80.3	86.7	73.4	83.9	80.4	88.2	74.2	87.6
Q^2ext			86.1	86.0	44.1	44.0	80.3	80.4

The following is the graphical representation of the parameters reported in the above table.



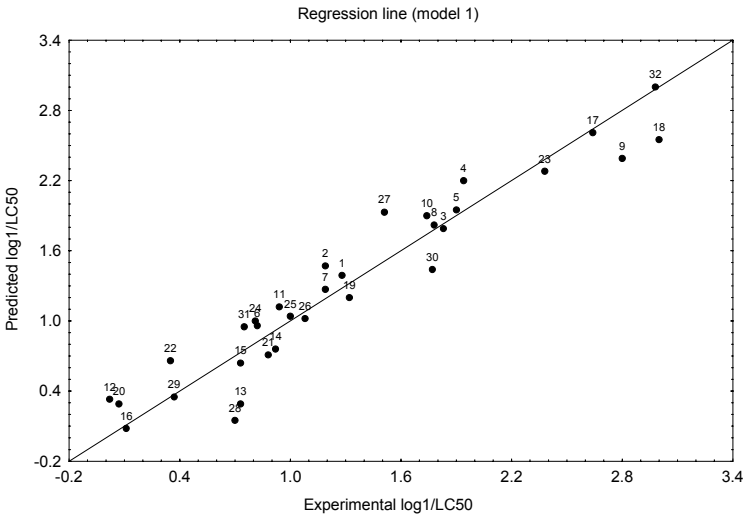
As already commented on above, the results of Q^2ext for the D-optimal splitting can be explained by the anomalous response distribution in this D-optimal splitting (all the validation chemicals with lower response value); on the contrary, this splitting worked well in relation to structure, as can be verified by the Structural PCA graph: the test chemicals are in the training chemicals, as usually happens in this splitting methodology. The much higher response variability in the training set compared with the validation set justifies the low TSS value in the formula for Q^2ext calculation and thus the apparent low predictivity.

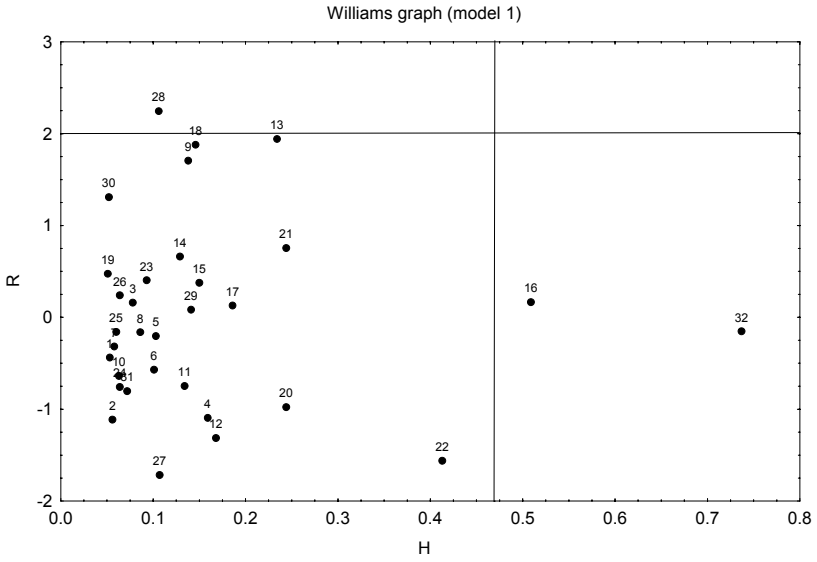
$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{valid} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{valid} (y_i - \bar{y}_{tr})^2} = \text{1-PRESS/TSS}$$

This is a demonstration that in some cases Q^2ext valuecan give apparent unreliable results: MSE and R^2 for validation set must be considered.

In Appendix the Leverage approach, applied to this data set.

FIGURE 11: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





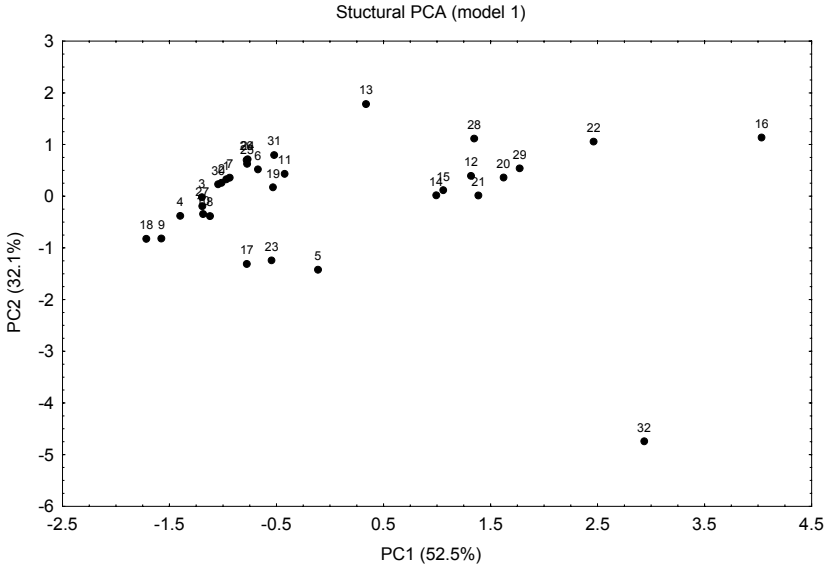
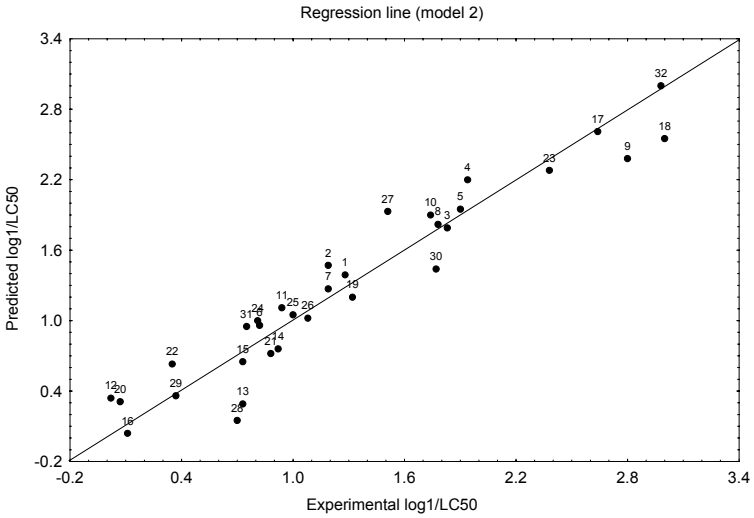
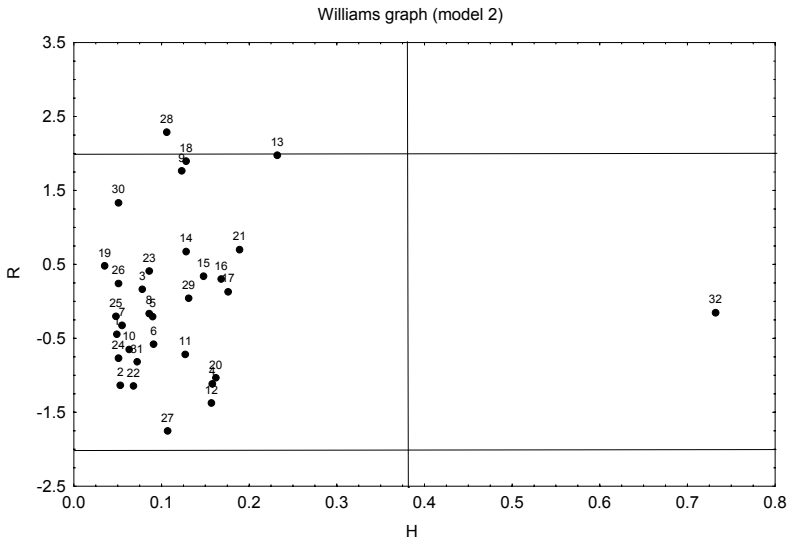


FIGURE 12: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





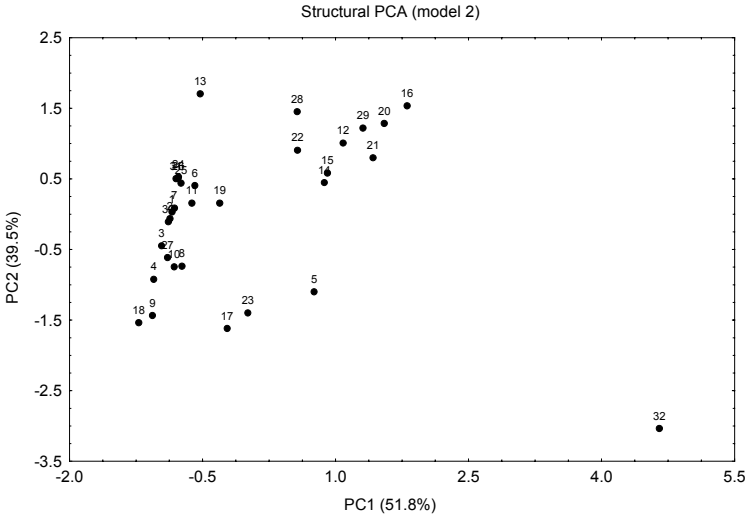
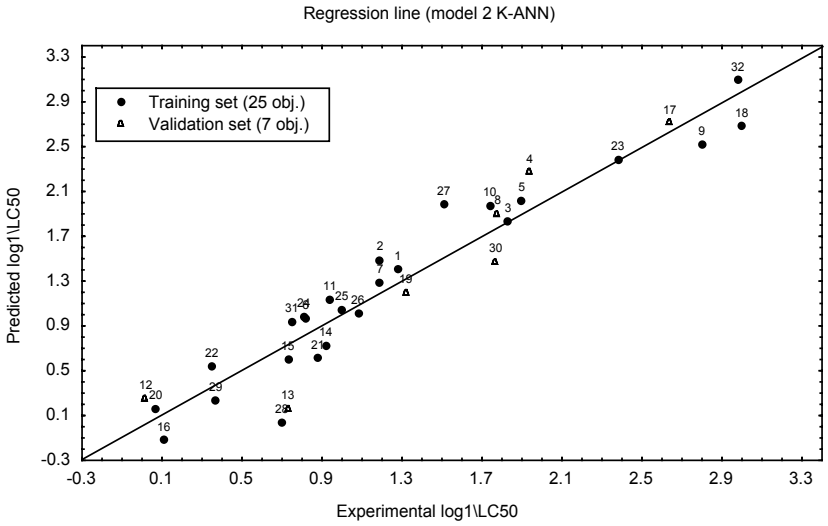
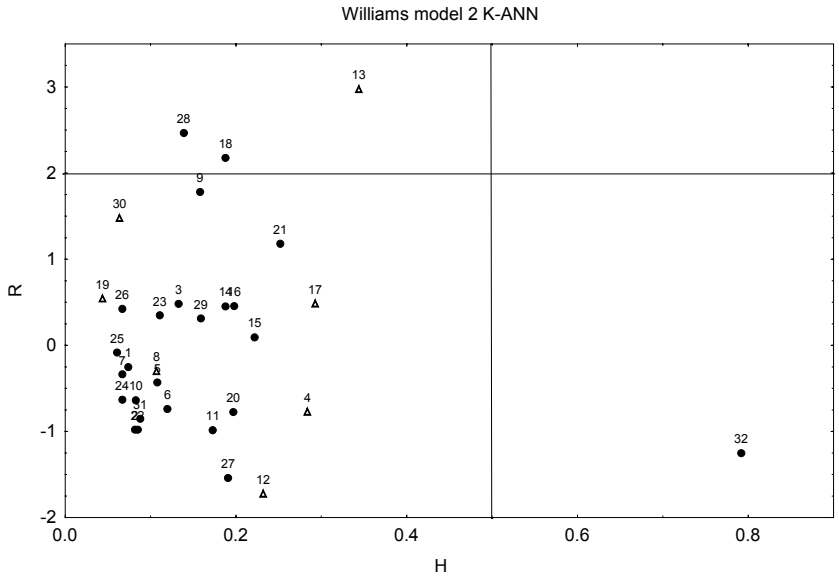


FIGURE 13: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





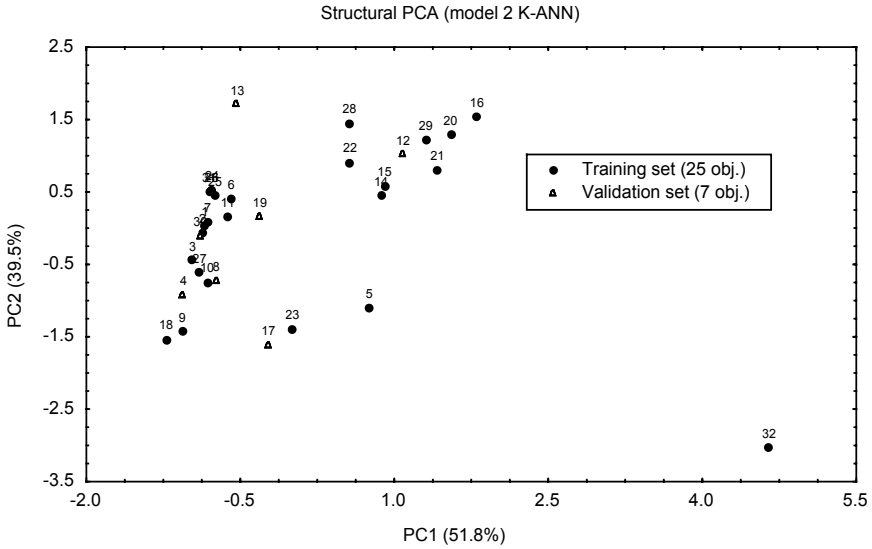
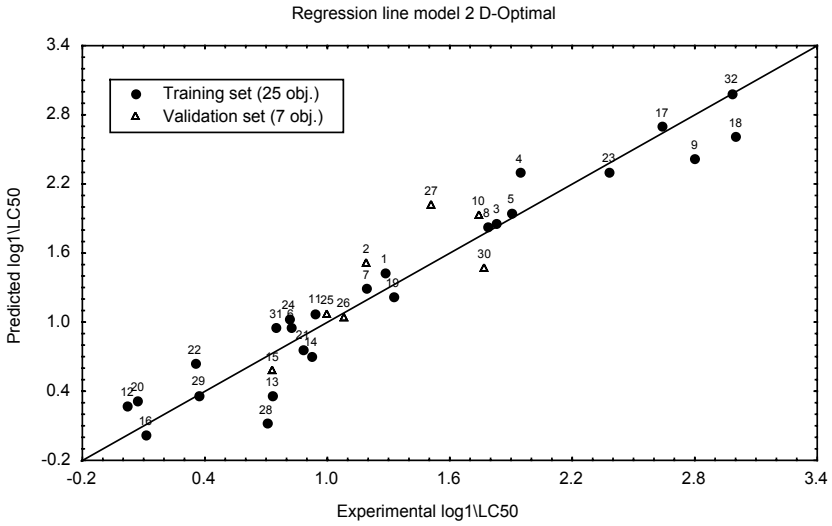
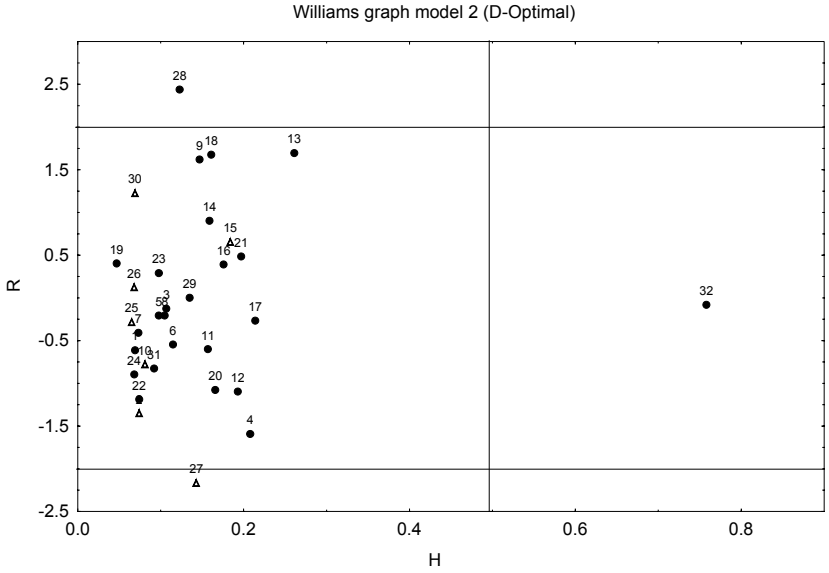


FIGURE 14: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





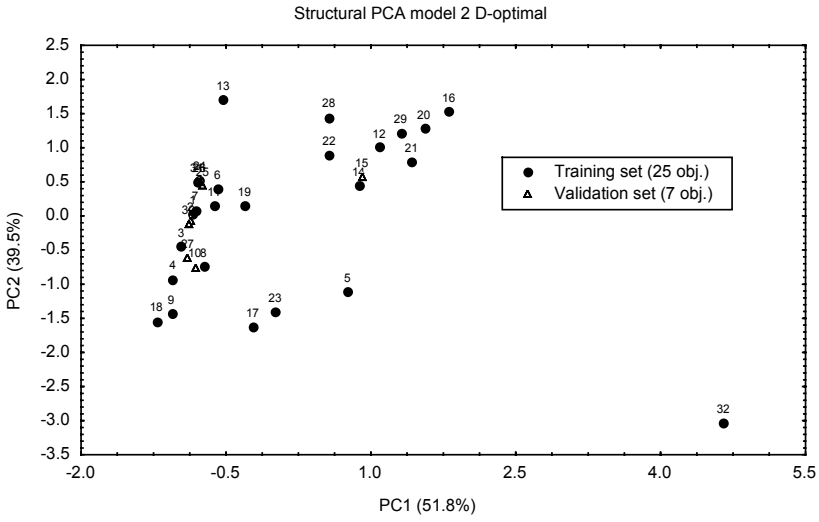
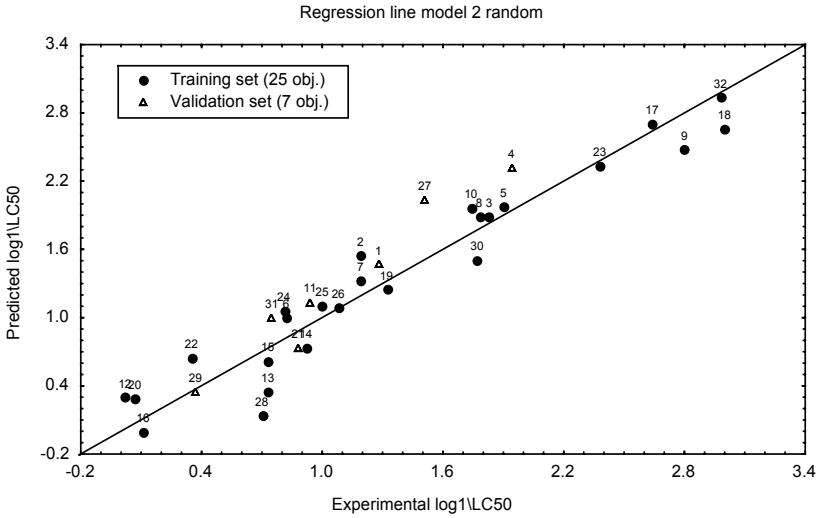
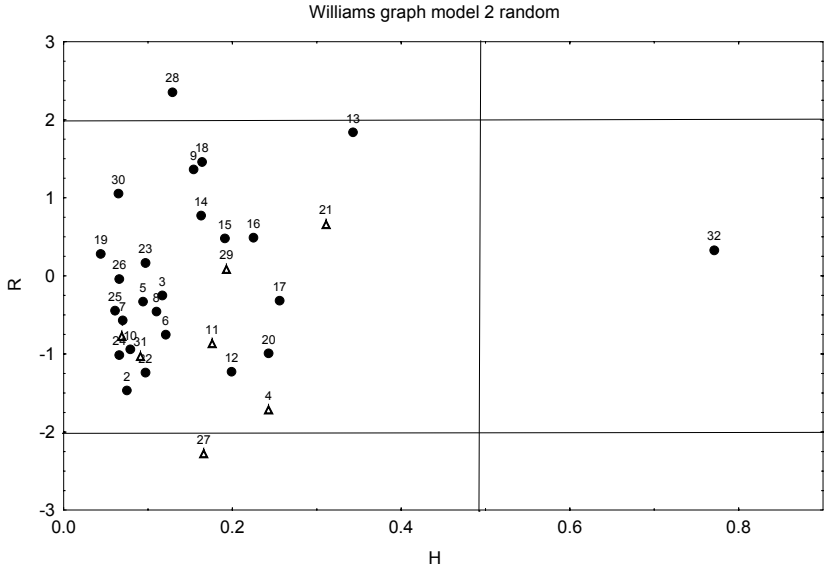
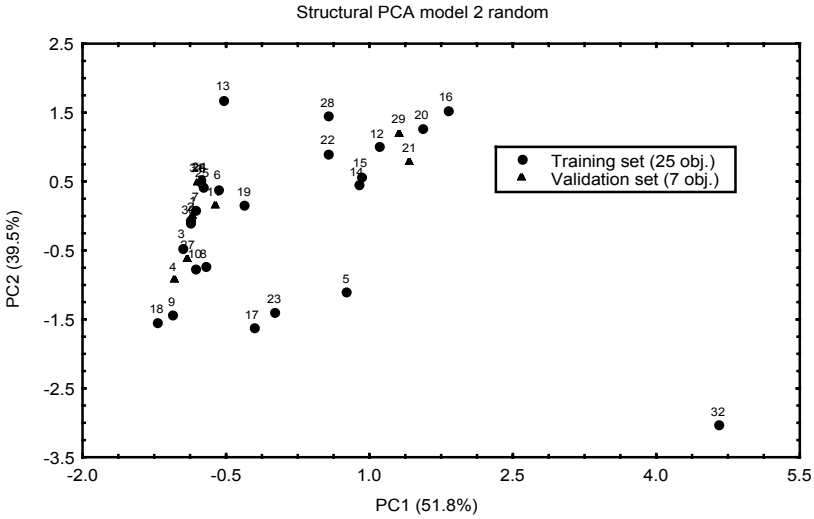


FIGURE 15: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







ALCOHOLS:

A data set of 34 alcohols was studied. 5 were evidenced as outliers by the authors and removed before the modelling.

The proposed models and the reported statistical parameters are:

(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 18.173(8)\ \alpha - 1.440(1)\ \beta + 0.809(8)\ \log Kow + 1.005\ \pi^* - 8.163(8)$

n= 29 R^2= 86.31% R^2 adj = 84.03% S.E.= 0.4746

(2) $\log(1/LC50) = 16.160(4)\ \alpha + 0.939\ \beta + 0.811\ \log Kow - 7.688(7)$

n= 29 R^2= 85.10% R^2adj = 83.31% S.E.= 0.4852

The statistical parameters reported by the authors are only related to the fitting performances, that are good.

The authors performed external validation using only one chemical (dodecyl alcohol, an alcohol with a chain longer than the alcohols in the training set: the predicted value is satisfactory, but it is impossible, basing only on this unique chemical to verify if by chance or by model quality)

The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figures 16 and 17. The authors did not point out that chemical 22 is an outlier, 14 an influential chemical in both the models and 29 only in model (1): these are evidenced in this contract work by the leverage approach.

The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors and any possible anomalous or isolated chemicals. The anomalous distribution of the chemicals in the structural space of the descriptors is immediately evident: the correlation among the variables results in a perfect alignment of some chemicals.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation was also performed by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (D-optimal

Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets regarding structural information is reported below. The influential chemicals are put into the training set in each following splitting.

Table 4: Statistical Diagnostics of models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2boot	R^2adj	Q^2ext	MSE tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
29	1	Tot.1	π* β α logKow	78.0	86.4	45.2	0.0	84.1		0.186		0.548	0.431	38.1	0.473	38.19	44.61	6.42
29	1	Tot.2	β α logKow	79.9	85.2	55.6	51.4	83.4		0.201		0.523	0.449	48.0	0.483	23.35	36.12	12.77
21	8	K-ANN 1	π* β α logKow	77.0	91.0	37.0	0.0	88.7	63.1	0.005	0.582	0.546	0.342	40.4	0.392	45.30	45.01	-0.29
21	8	K-ANN 2	β α logKow	86.5	91.0	69.2	23.1	89.4	63.1	0.118	0.582	0.418	0.342	57.2	0.380	30.65	38.93	8.28
21	8	D-Opt. 1	π* β α logKow	73.7	85.7	35.2	0.0	82.1	89.8	0.231	0.070	0.653	0.481	24.0	0.551	36.70	43.61	6.91
21	8	D-Opt. 2	β α logKow	77.2	84.0	50.1	42.5	81.2	92.0	0.258	0.055	0.608	0.509	29.8	0.565	22.55	35.23	12.68

It is immediately evident that the models, even with good fitting performances (high values of R^2 and R^2_{adj}), and satisfactory Q^2 values are unstable and not predictive if validated by LMO and bootstrapping (the validation by bootstrap is below the cut-off value of the software, resulting in 0 value).

SDEP are higher than SDEC: the models work slightly worse in prediction than in calculation.

Regarding collinearity: in general, the descriptors are very correlated (medium Kxx: 33) but, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy) and the correlation among the X (Kxx) in some model is too small and even negative (medium delta: 8). This is a signal of multicollinearity without prediction power (in fact when we applied the QUIK rule of Todeschini this model 1 was excluded as a predictive model). The model 1 can also be considered overfitting: unnecessary terms are included to fit the data, but these are not useful for predictive purposes.

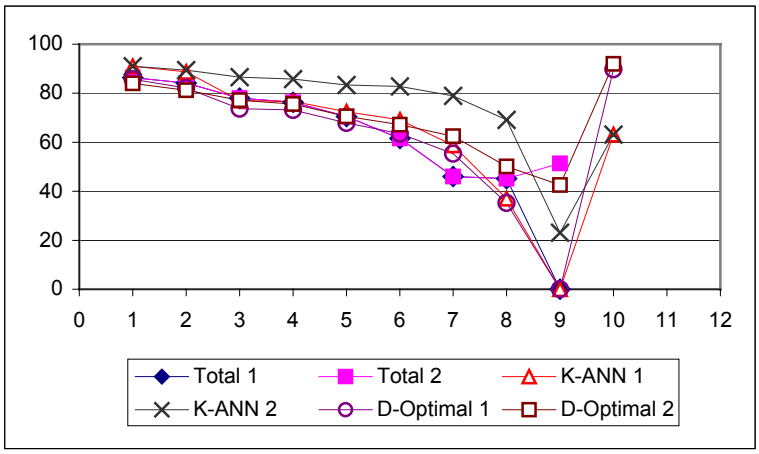
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted:

Table 4 bis: Statistical Diagnostics of models

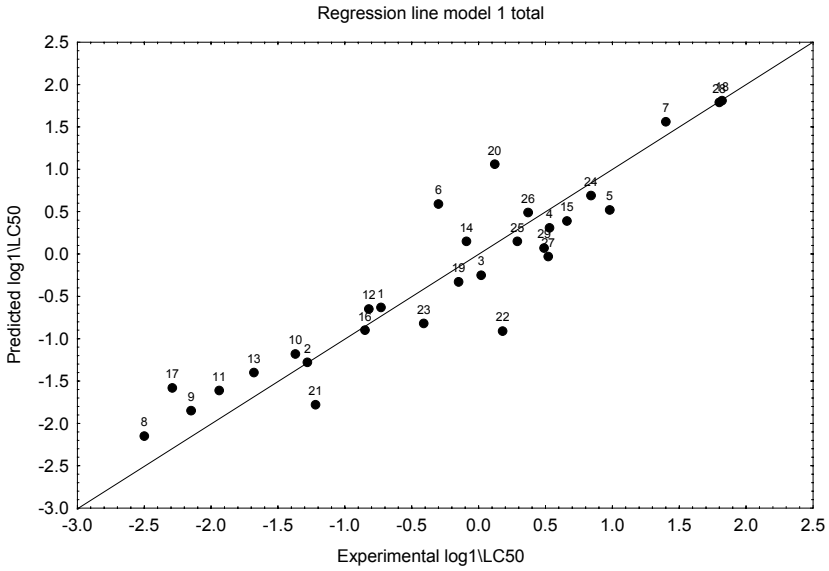
		Total 1	Total 2	K-ANN 1	K-ANN 2	D-Optimal 1	D-Optimal 2
1	R^2	86.4	86.4	91.0	91.0	85.7	84.0
2	R^2_{adj}	84.1	84.1	88.7	89.4	82.1	81.2
3	Q^2	78.0	78.0	77.0	86.5	73.7	77.2
4	Q^2_{LMO10}	76.3	76.3	76.6	85.8	73.1	75.6
5	Q^2_{LMO20}	70.5	70.5	72.4	83.4	67.9	70.5
6	Q^2_{LMO30}	61.5	61.5	69.1	82.8	63.5	67.2
7	Q^2_{LMO40}	46.1	46.1	58.7	78.9	55.4	62.5
8	Q^2_{LMO50}	45.2	45.2	37.0	69.2	35.2	50.1
9	Q^2_{boot}	0.0	51.4	0.0	23.1	0.0	42.5
10	Q^2_{ext}			63.1	63.1	89.8	92.0

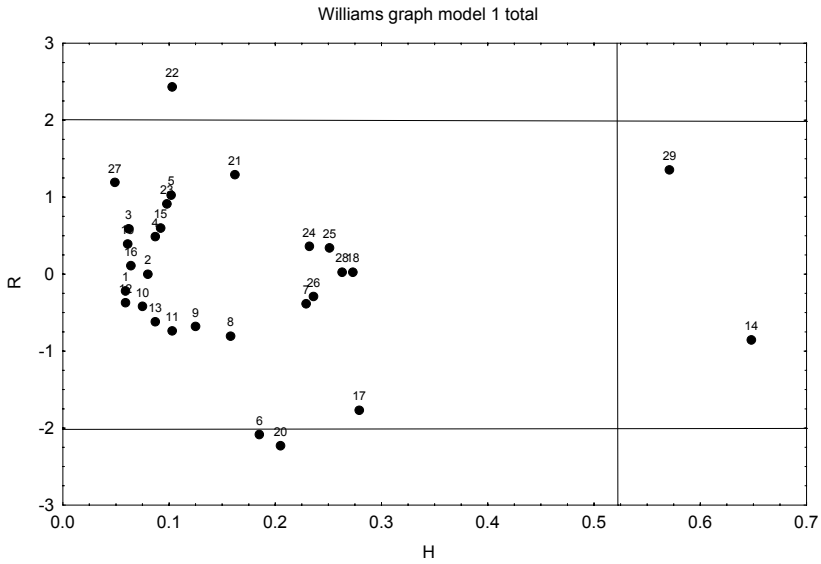
The following is the graphical representation of the parameters reported in the above table.



As in other cases, internal validation with more than 30% of the chemicals excluded from the data set in LMO cross-validation results in a strong under-optimism when applied to small data sets (20-30 chemicals). In fact the structural information in the training set of 12-15 chemicals (40-50%) is too reduced and and not sufficiently informative to be useful to demonstrate model predictivity. The models with higher predictivity are always those with fewer descriptors, even though with lower R^2 they have better prediction performance because the models with more descriptors are overfitted. The comparison of Q^2_{EXT} highlights that D-optimal splitting gives optimistic results compared with K-ANN splitting. The published models are apparently good models, they are fitting models, and, in particular, are overfitted with 4 descriptors, but their predictive ability, though verified by internal validation, must be considered insufficient.

FIGURE 16: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





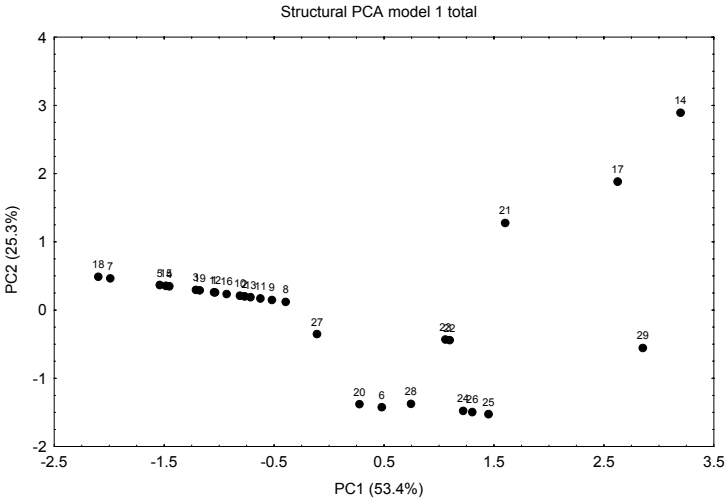
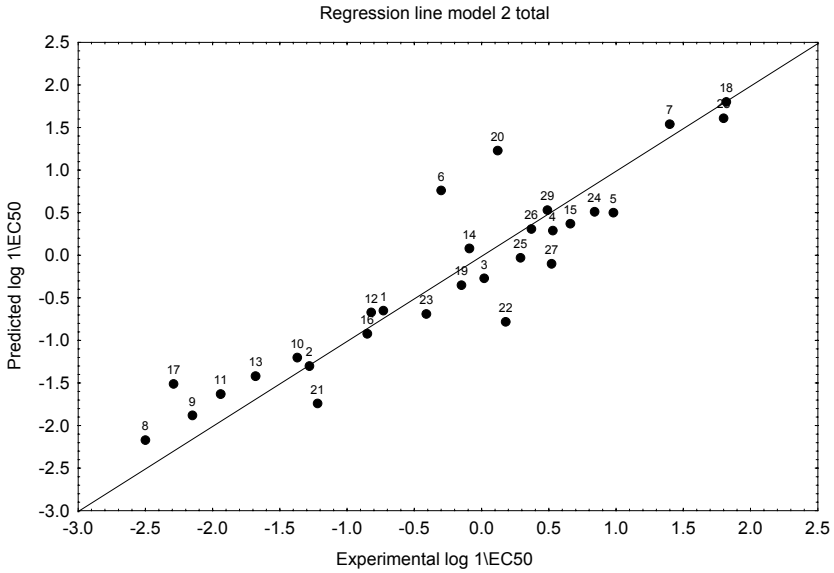
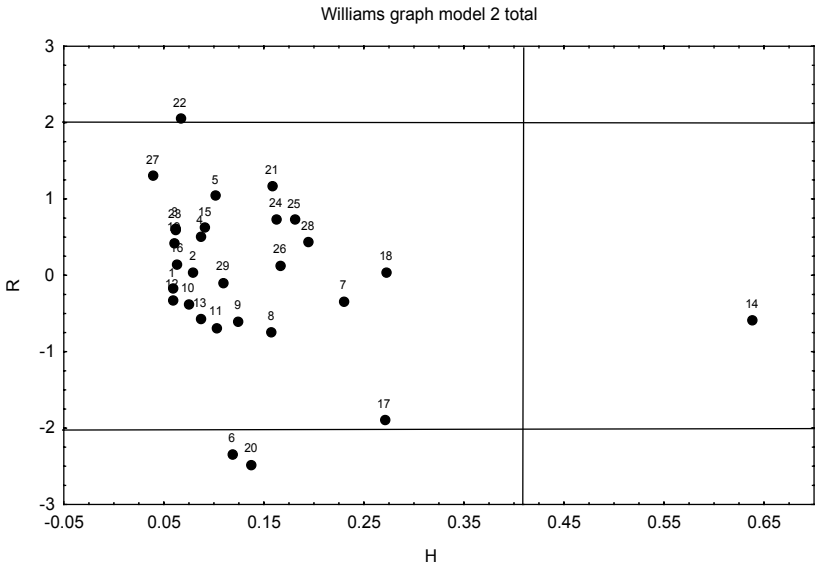


FIGURE 17: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





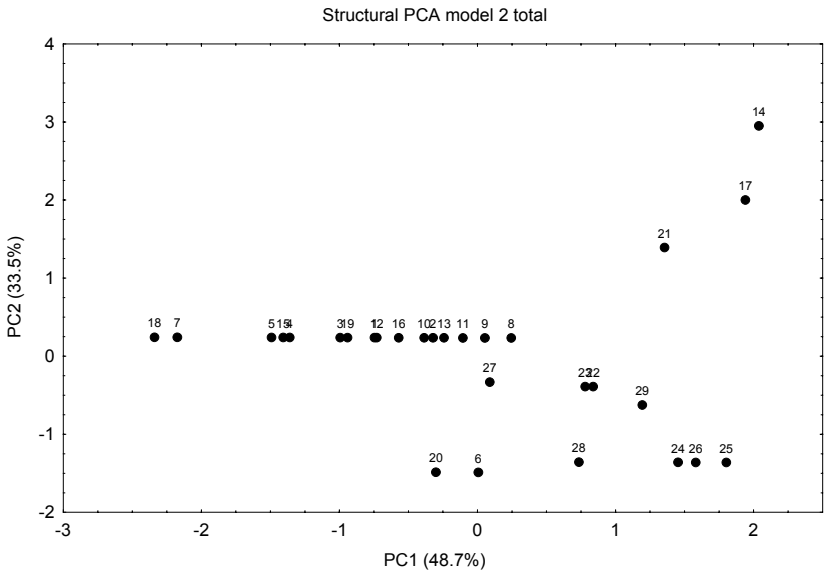
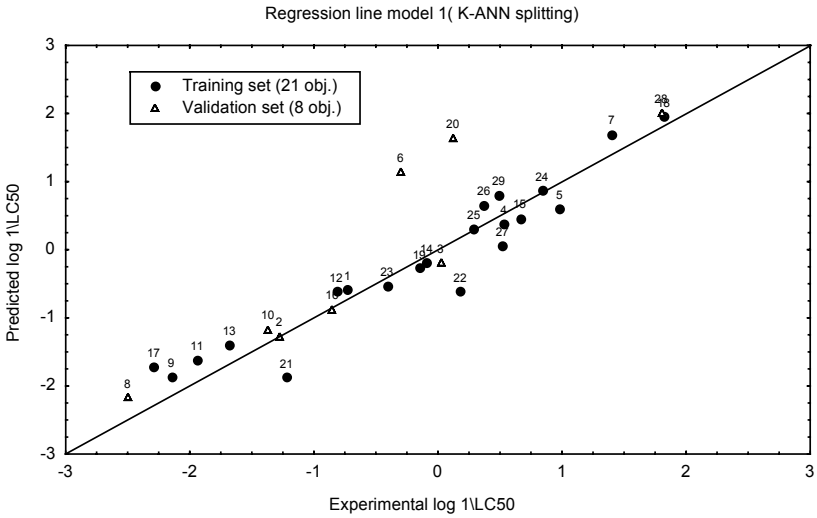
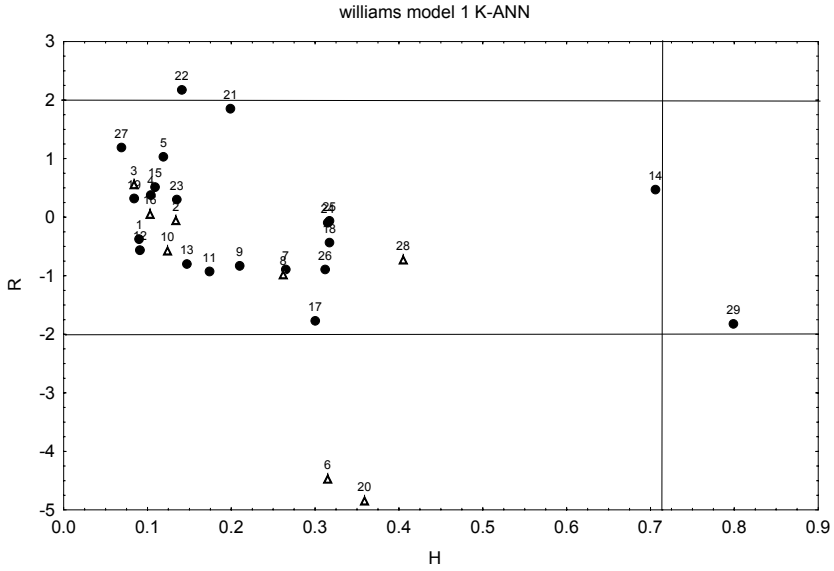


FIGURE 18: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





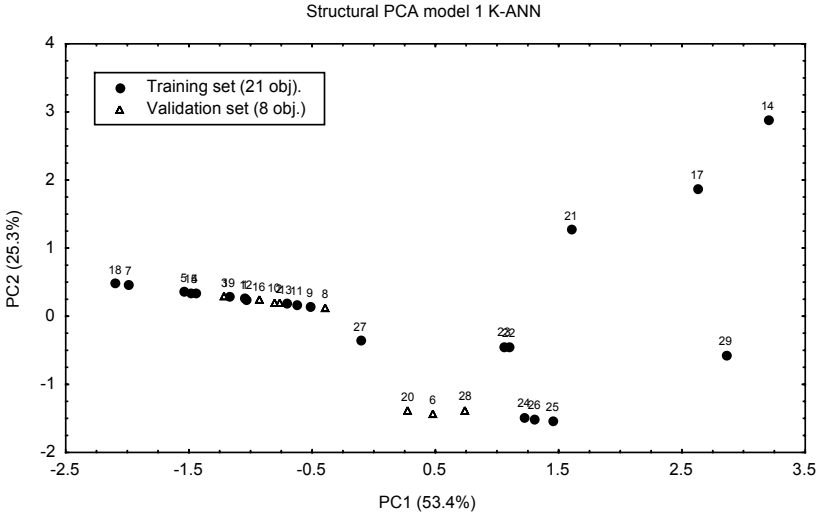
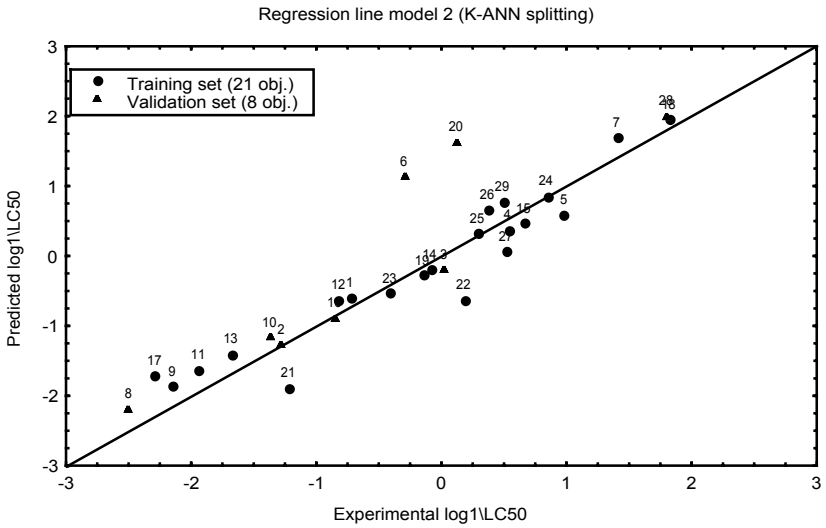
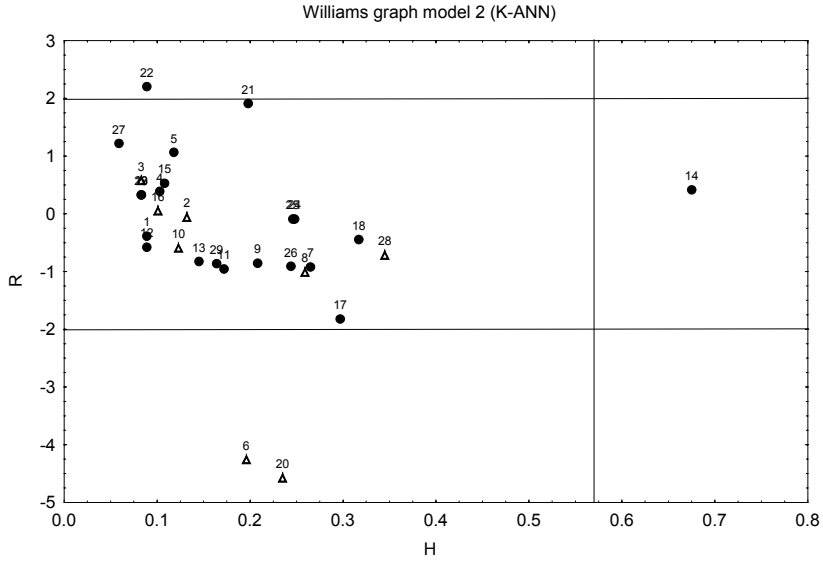


FIGURE 19: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





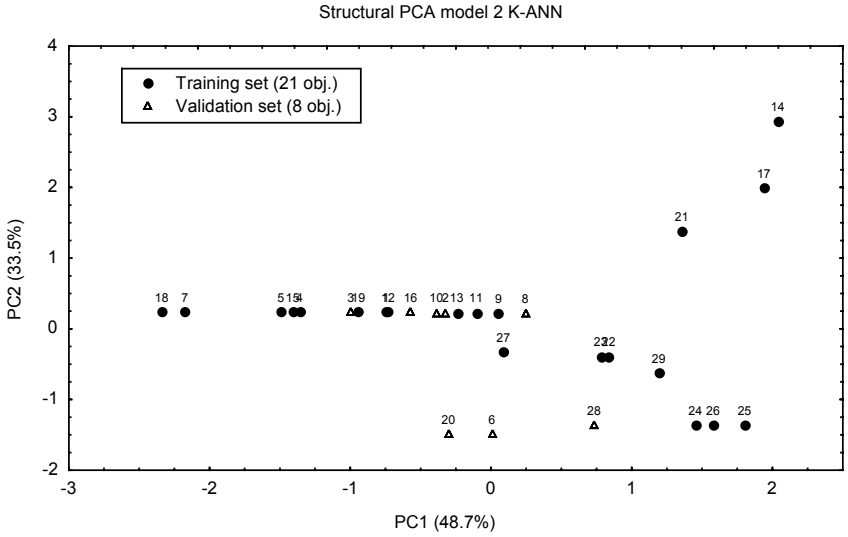
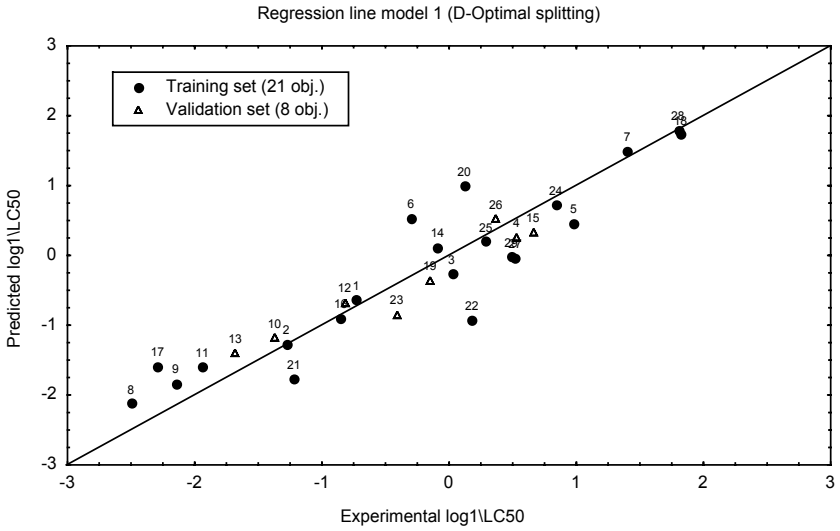
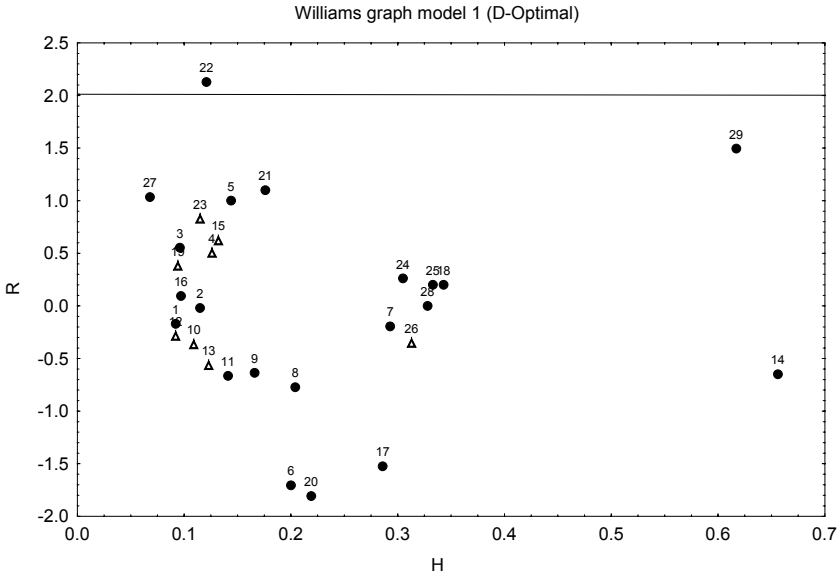


FIGURE 20: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





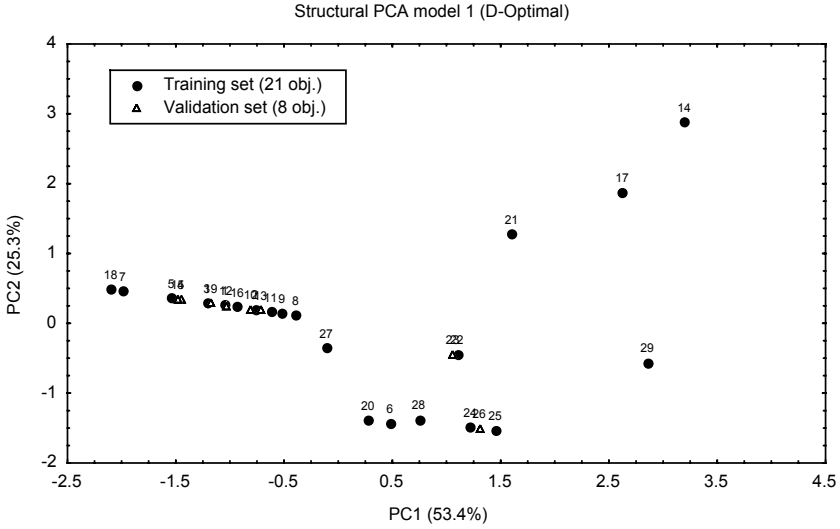
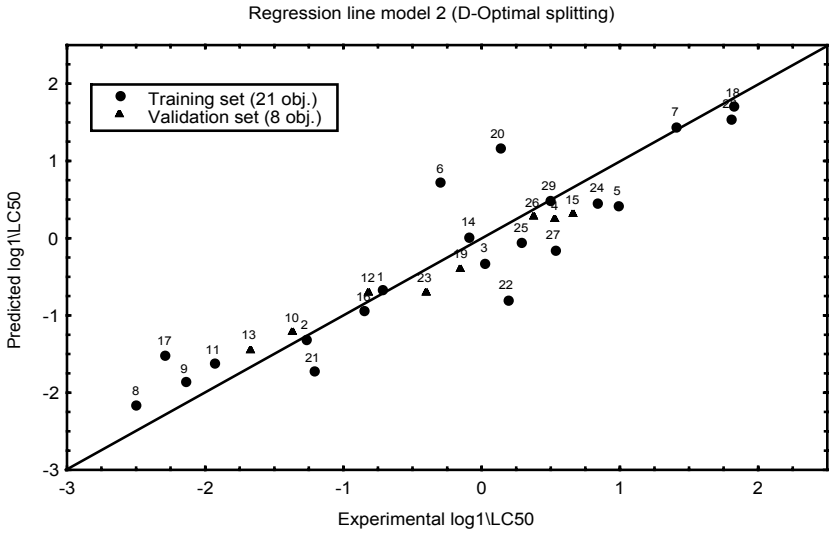
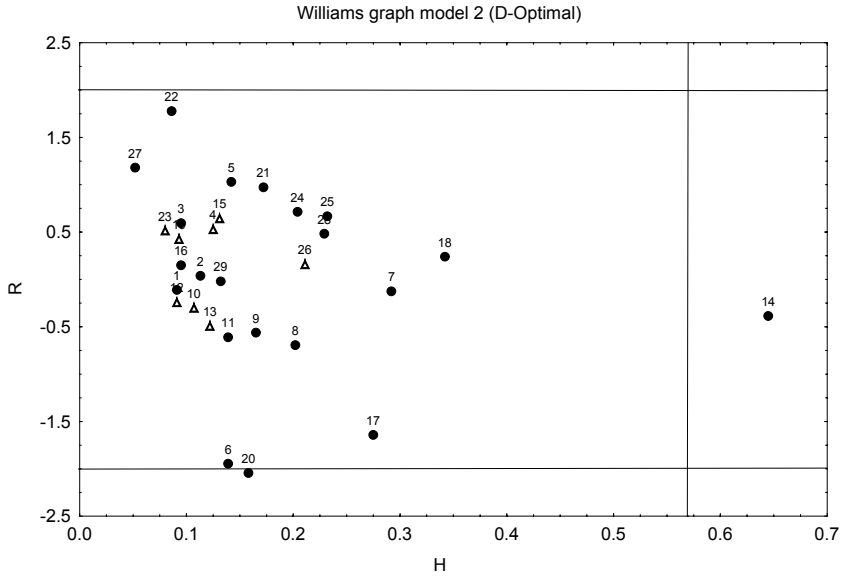
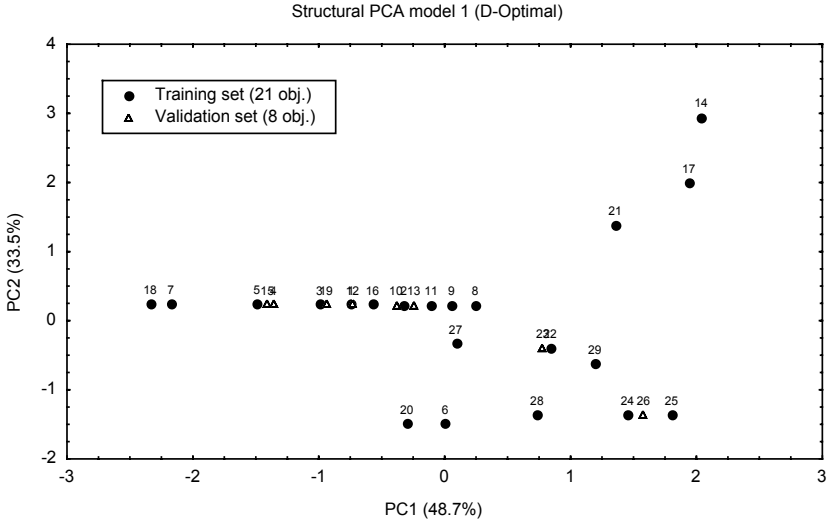


FIGURE 21: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







ALDEHYDES:

The data set is composed of 34 chemicals: 5 are outliers and removed before the modelling for the model (1), 4 outliers were removed for model (2).

The proposed models and the reported statistical parameters are:

(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 1.313(1)\,\alpha - 1.0118(9)\,\beta + 0.363(7)\,\log Kow + 0.743(2)\pi^* + 0.294(3)$

n= 29 R^2= 87.19% R^2 adj = 85.05% S.E.= 0.2742

(2) $\log(1/LC50) = 0.816(5)\,\alpha - 0.730(1)\beta + 0.49\,\log Kow + 0.583(8)$

n= 30 R^2= 79.47% R^2adj = 77.11% S.E.= 0.3347

The statistical parameter reported by the authors relate only to fitting performances, that are very good.

The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figures 22 and 23. The authors did not point out that chemicals 27 and 12 are outliers in Model (1), and that 30 is an outlier and an influential chemical in model (2).

The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors and any possible anomalous or isolated chemicals.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation was also performed by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (D-optimal Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets regarding structural information is reported below. The influential chemicals are put into the training set in each following splitting.

Table 5: Statistical Diagnostics of models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE train	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
29		Tot.1	π* β α logKow	83.3	87.3	78.1	80.2	85.1		0.062		0.285	0.249	41.1	0.273	44.26	49.88	5.62
30		Tot.2	β α logKow	67.7	79.6	62.9	64.4	77.2		0.096		0.391	0.311	33.7	0.334	25.91	40.48	14.57
22	7	K-ANN 1	π* β α logKow	86.4	91.1	77.5	81.9	89.0	75.7	0.044	0.121	0.259	0.209	43.2	0.238	36.56	45.25	8.69
24	6	K-ANN 2	β α logKow	68.7	81.9	62.5	64.6	79.2	63.7	0.097	0.099	0.409	0.311	30.0	0.341	20.24	38.04	17.80
22	7	D-Opt. 1	π* β α logKow	87.0	91.2	80.7	83.4	89.1	58.2	0.049	0.108	0.270	0.222	43.6	0.253	39.90	47.29	7.39
24	6	D-Opt. 2	β α logKow	71.1	83.4	66.6	68.4	81.0	47.9	0.088	0.131	0.393	0.298	33.4	0.326	20.78	38.69	17.91
22	7	Rand. 1	π* β α logKow	90.4	92.9	76.9	83.8	91.2	46.2	0.039	0.180	0.227	0.196	54.7	0.222	45.08	51.56	6.48
24	6	Rand. 2	β α logKow	47.9	71.2	37.8	40.9	66.8	90.4	0.103	0.092	0.432	0.322	16.4	0.352	30.20	45.07	14.87

The models demonstrate a satisfactory stability in internal validation. In this case, the models with 4 descriptors are always the more stable and internally predictive (less difference between Q^2_{LMO50} and Q^2_{LOO}), the addition of one descriptor is useful in improving not only the fitting but also the predictivity.

SDEP and SDEC are more similar in models 1 than in models 2: models 1 work better in internal prediction.

Statistical external validations confirm the lower prediction ability for chemicals in the validation set mainly for model 1. The MSE value must be considered, as Q^2_{ext} gives results depending on the splitting. In fact, the MSE values for training and validation set are similar for models 2, but the models 1 predict the response for chemicals not used in the model development (validation set) worse than for chemicals used to find the relationship (training set). The only exception is the model from the random 2 splitting which is more predictive externally than internally. This is further evidence of the strong influence of the splitting in verifying the model predictivity when small data sets(20-30 chemicals) are considered.

Regarding collinearity: in general, the descriptors are very correlated (medium Kxx: 33) but, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy) and the correlation among the X (Kxx) is sufficiently high, at least for some models (2), (medium delta:12) compared with other QSAR models, and according to our experience.

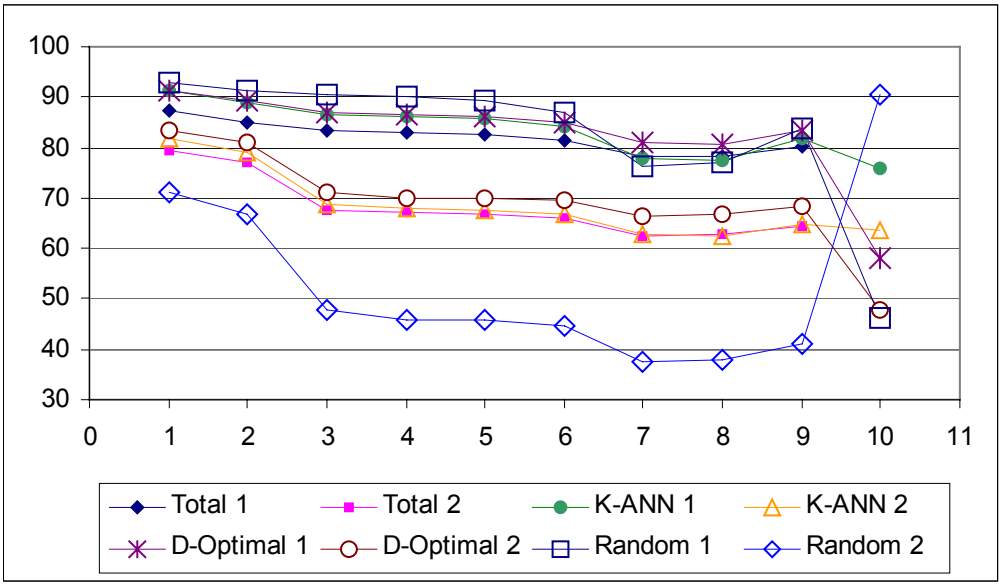
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted :

Table 5 bis: Statistical Diagnostics of models

		Total 1	Total 2	K-ANN 1	K-ANN 2	D-Optimal 1	D-Optimal 2	Random 1	Random 2
1	R^2	87.3	79.6	91.1	81.9	91.2	83.4	92.9	71.2
2	R^2_{adj}	85.1	77.2	89.0	79.2	89.1	81.0	91.2	66.8
3	Q^2	83.3	67.7	86.4	68.7	87.0	71.1	90.4	47.9
4	Q^2_{LMO10}	83.0	67.2	86.1	67.9	86.7	70.0	90.3	45.8
5	Q^2_{LMO20}	82.5	66.7	85.7	67.5	86.3	70.1	89.2	45.7
6	Q^2_{LMO30}	81.5	66.0	84.1	66.8	85.2	69.4	87.0	44.7
7	Q^2_{LMO40}	78.4	62.4	77.8	62.9	80.9	66.3	76.2	37.5
8	Q^2_{LMO50}	78.1	62.9	77.5	62.5	80.7	66.6	76.9	37.8
9	Q^2_{boot}	80.2	64.4	81.9	64.6	83.4	68.4	83.8	40.9
10	Q^2ext			75.7	63.7	58.2	47.9	46.2	90.4

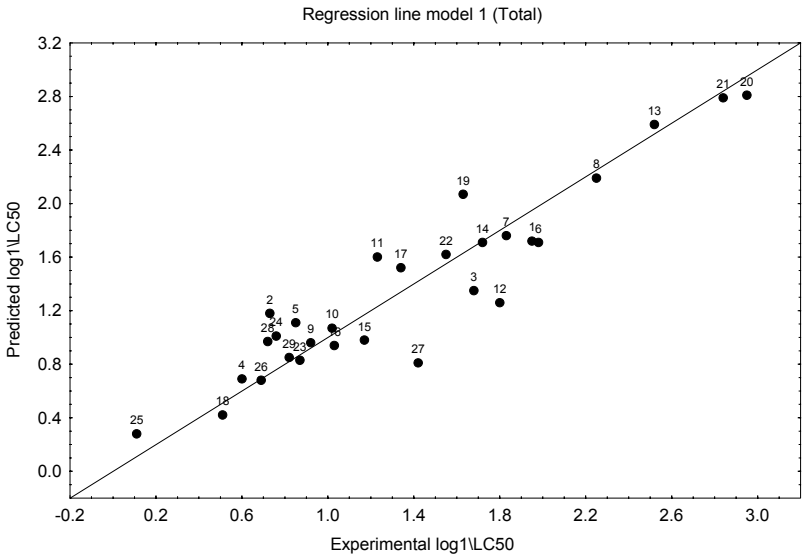
The following is the graphical representation of the parameters reported in the above table.

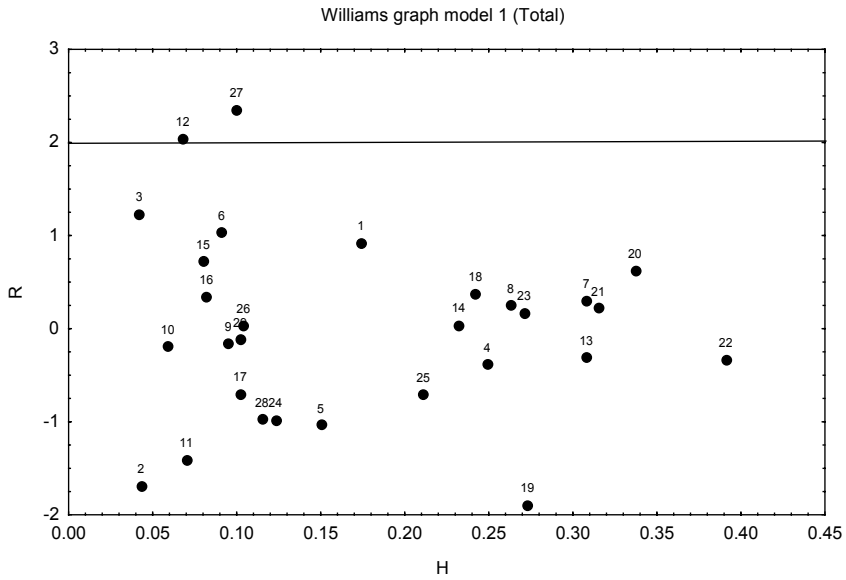


Internal validation results in the normal trend of similar validations: the exception is again the under-optimistic Q^2ext value in D-optimal splitting. The comparison of MSE value gives complementary information regarding predictive ability for new chemicals. The models 2, based on 3 descriptors, are the more predictive on new chemicals.

The random splitting 2 gives anomalous results: it actually highlights less predictivity by Q^2_{LOO}, the value of Q^2_{EXT} is over-optimistic, in fact MSE values confirm that the training and validation set values are calculated similarly. The anomalous result of random splitting is evident and we propose that the results derived from the analysis of models based on random splitting in small data sets cannot be considered reliable.

FIGURE 22: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





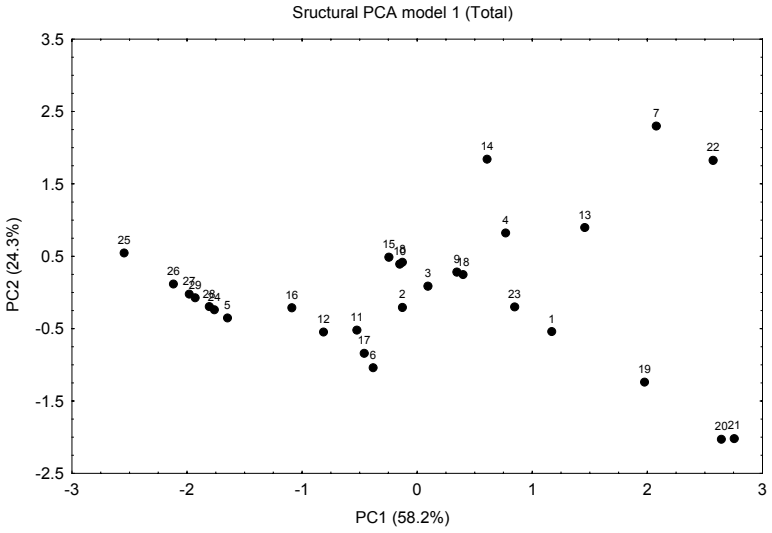
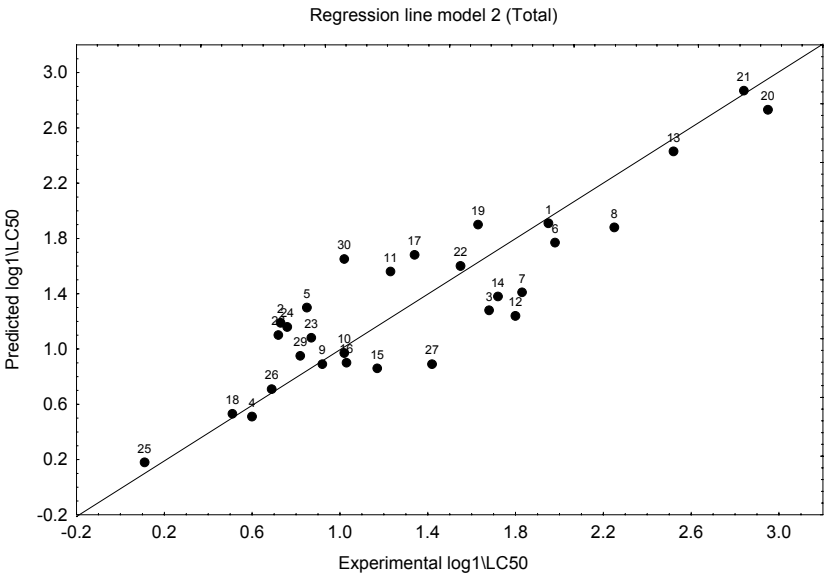
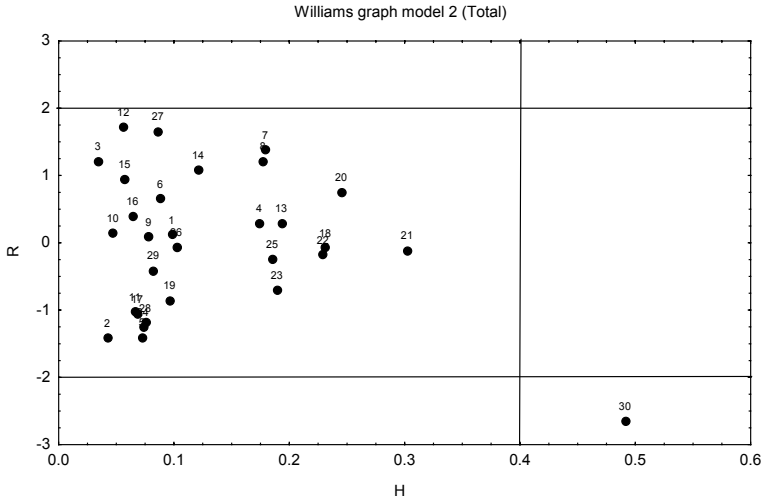


FIGURE 23: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





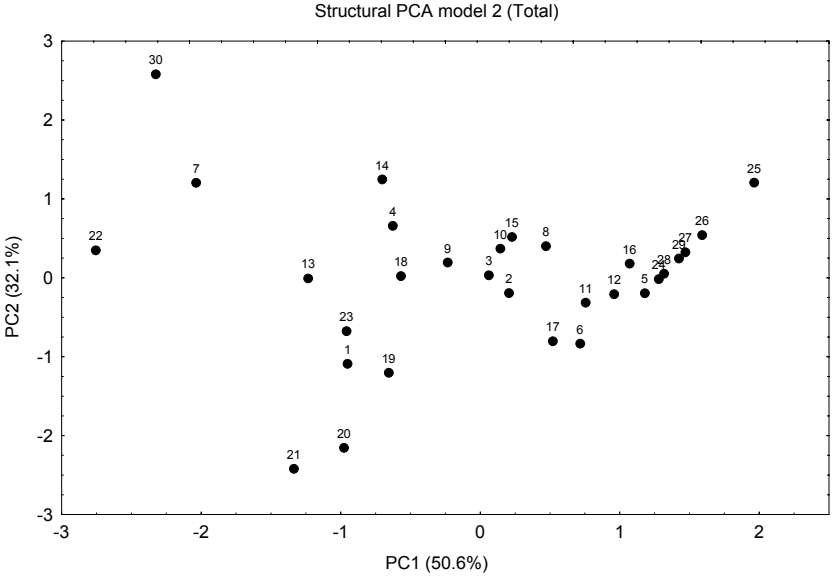
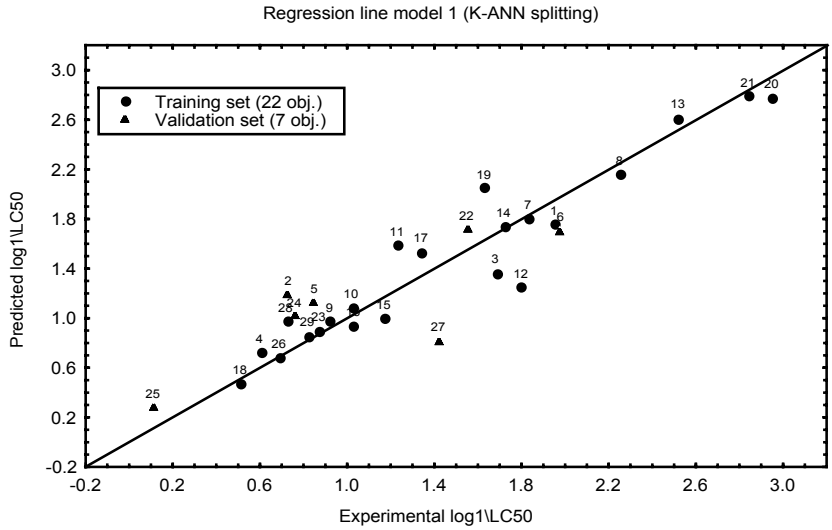
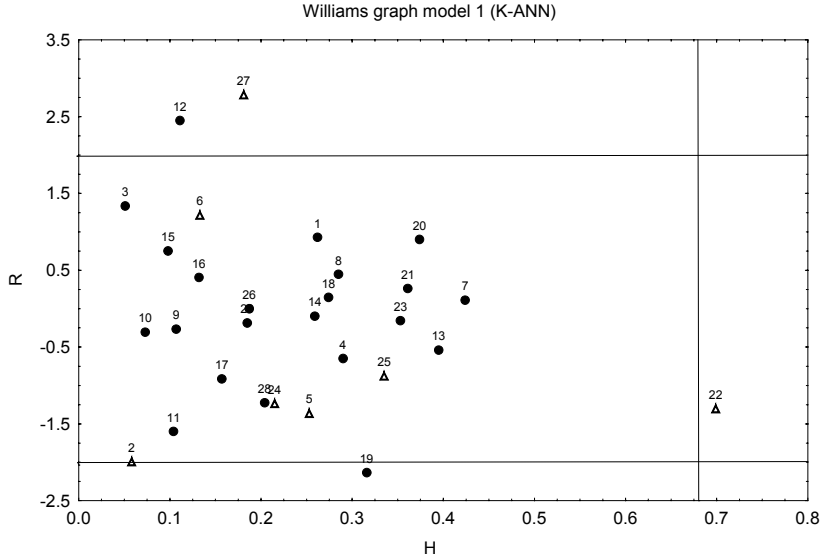


FIGURE 24: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





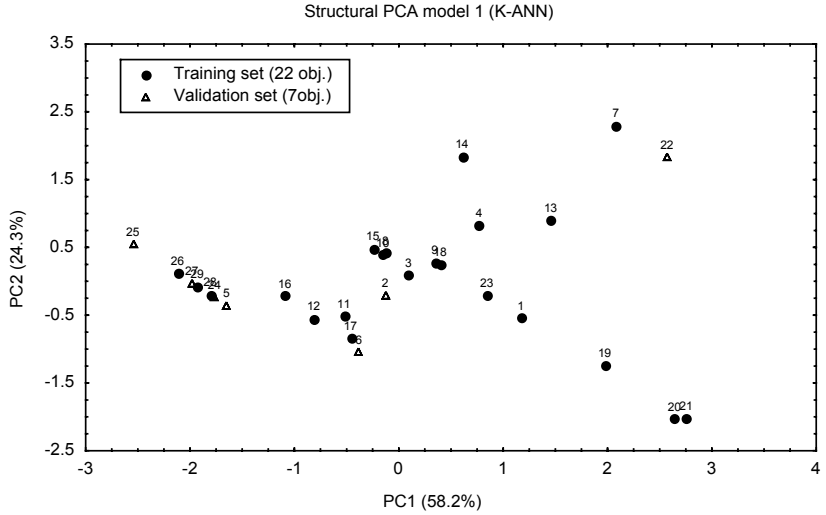
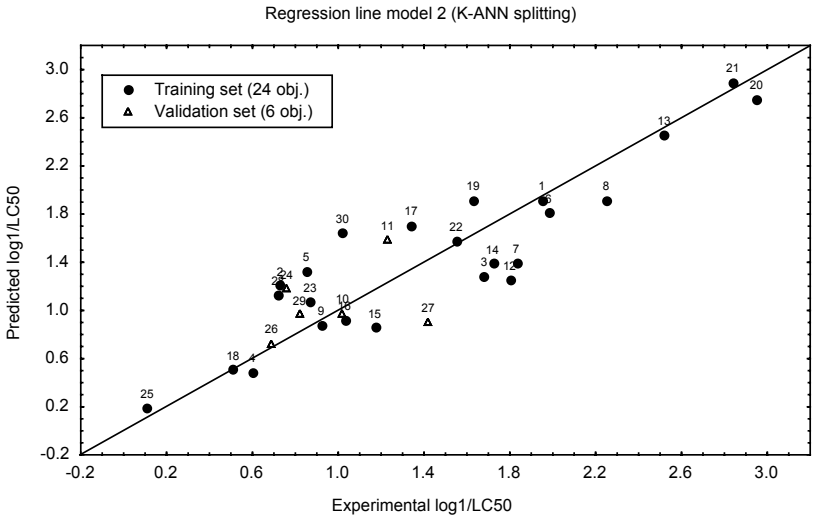
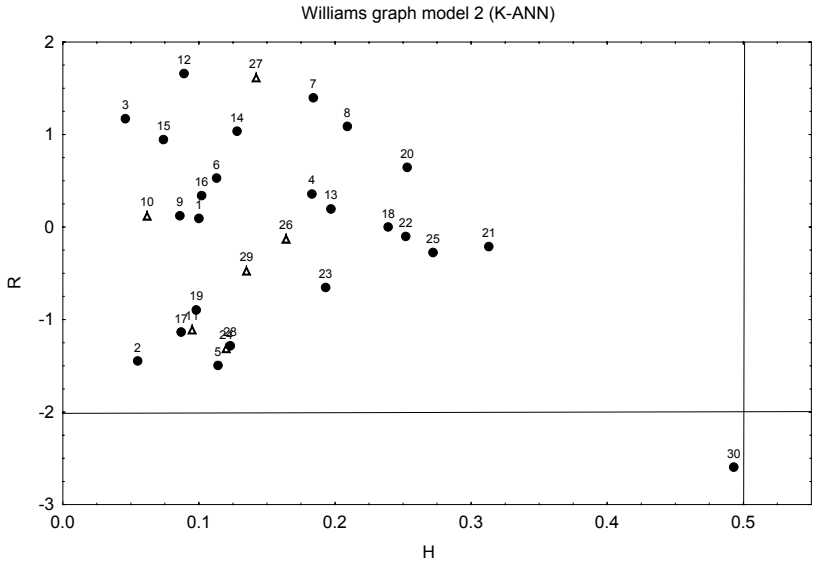


FIGURE 25: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





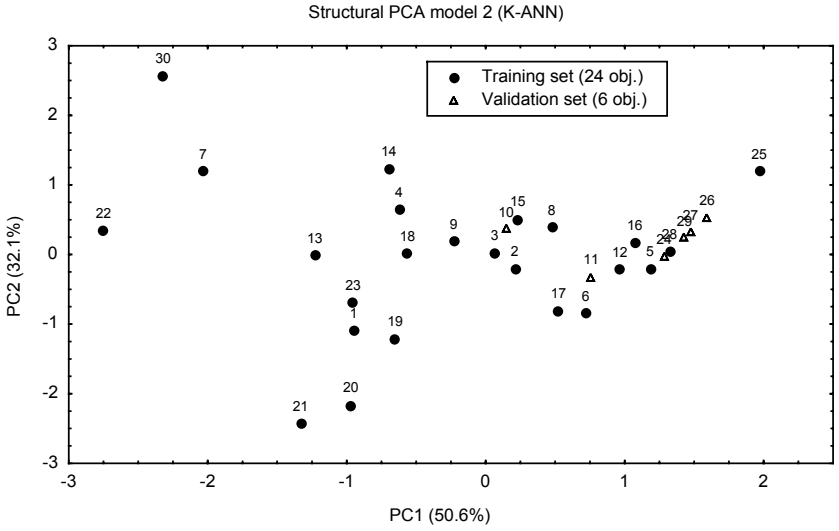
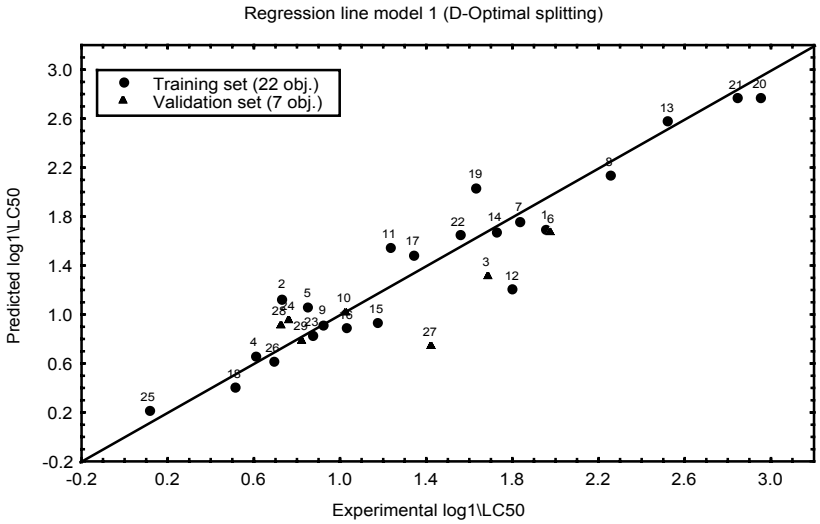
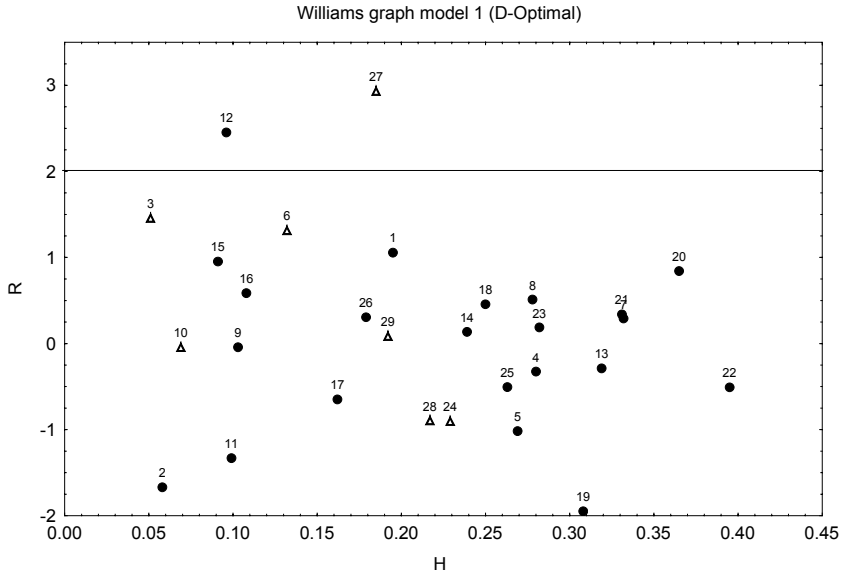


FIGURE 26: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





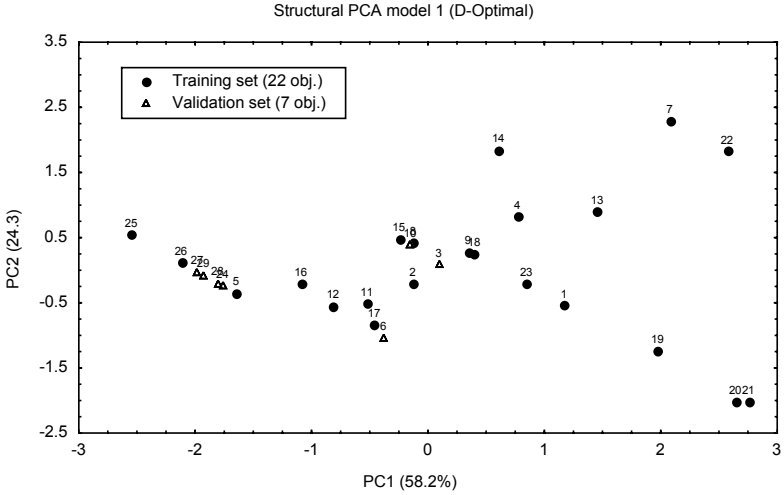
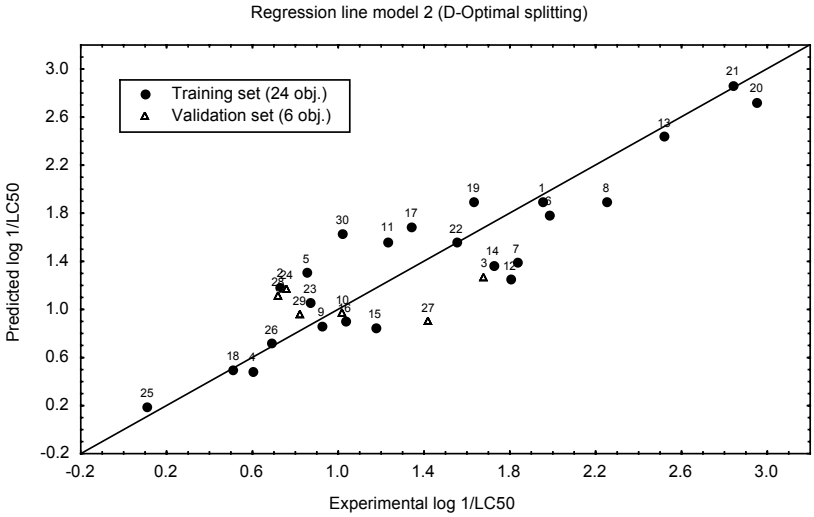
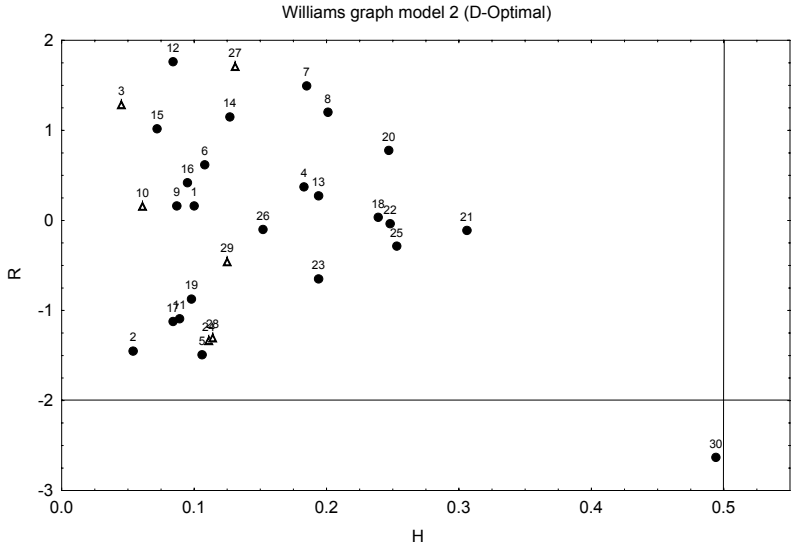


FIGURE 27: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





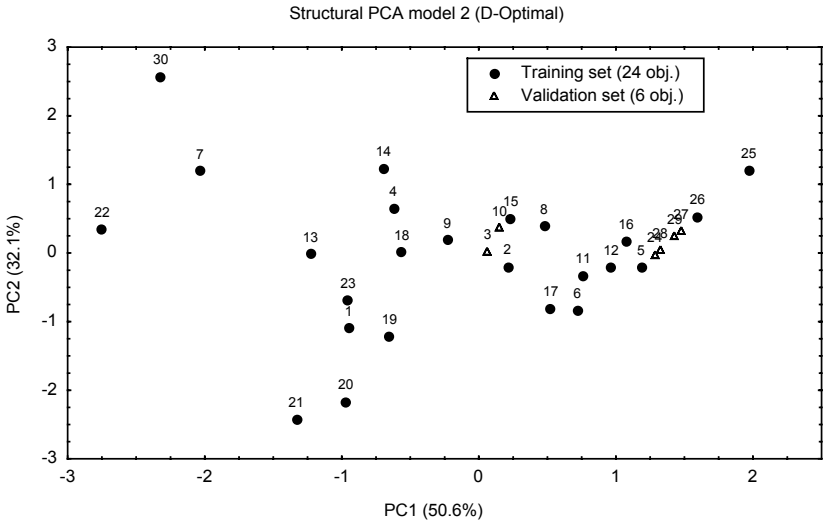
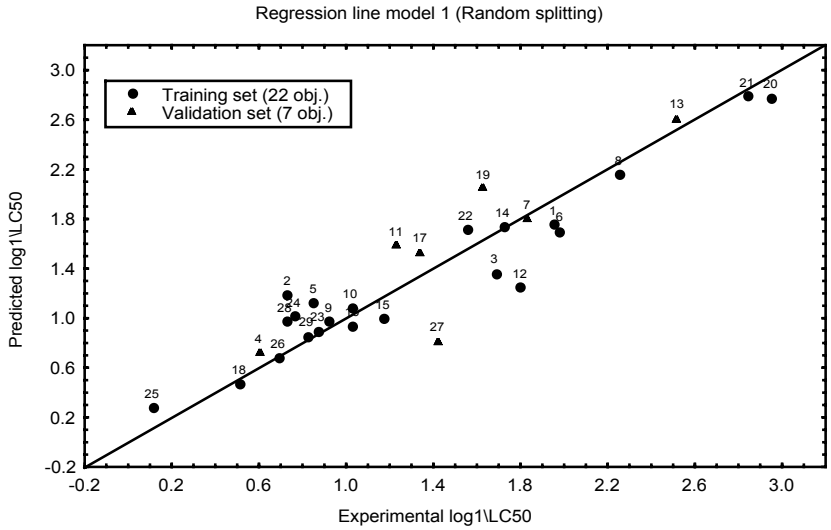
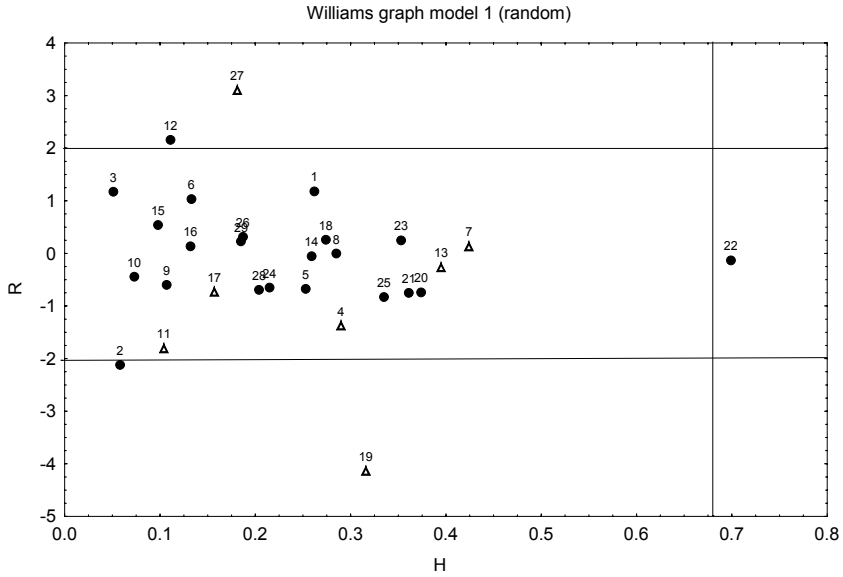


FIGURE 28: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





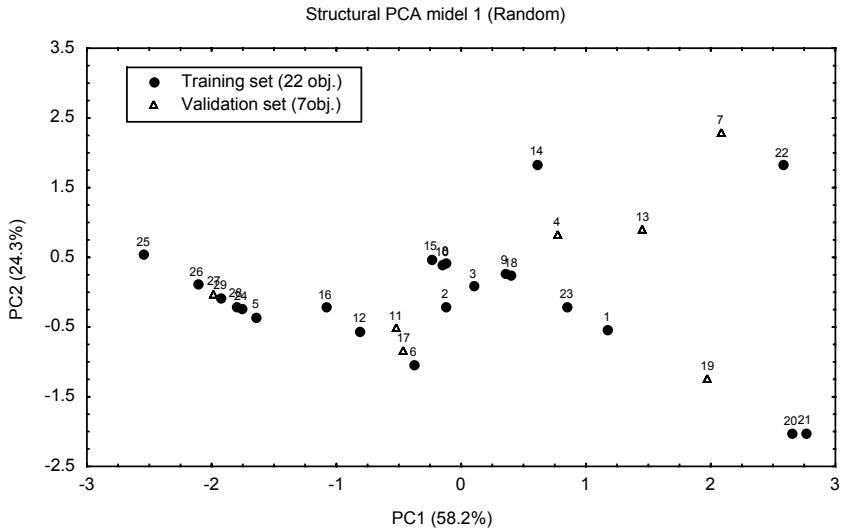
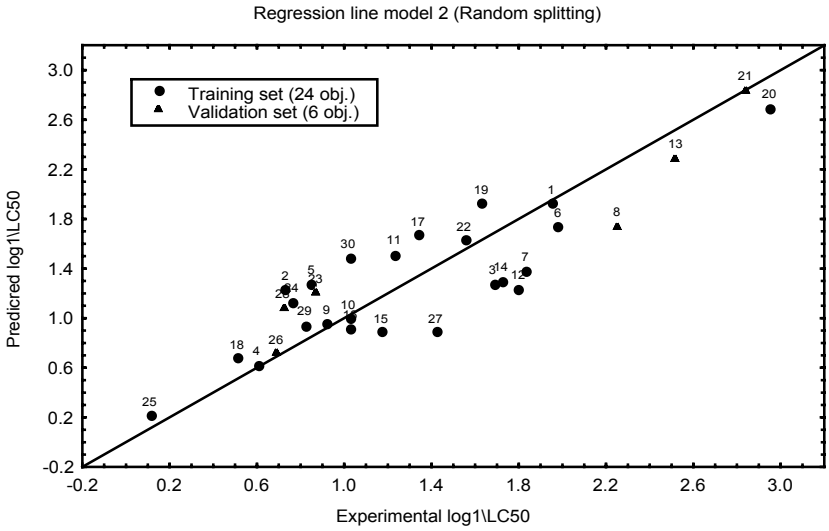
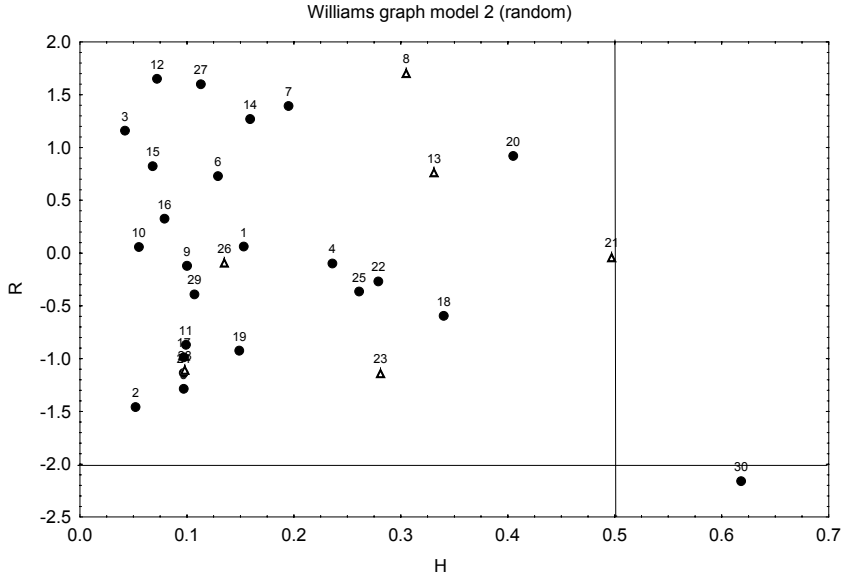
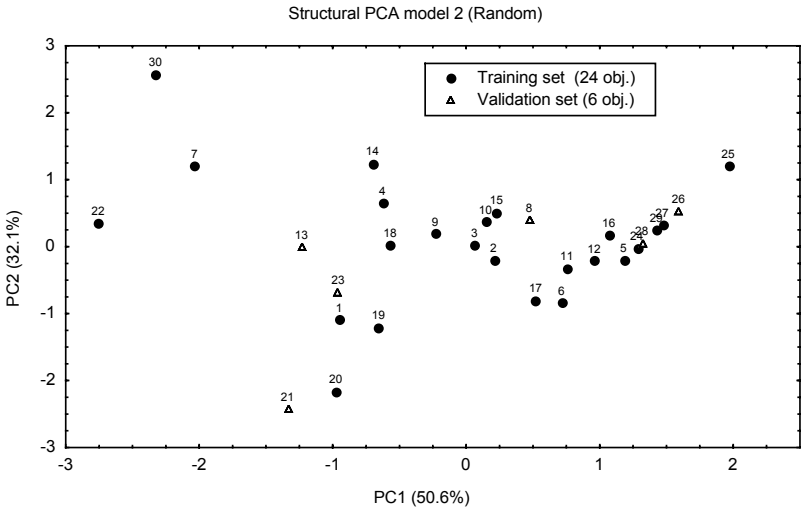


FIGURE 29: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







ALIPHATIC CHEMICALS:

A data set of 33 chemicals has been studied: 4 are outliers and removed by the authors before the modeling.

The proposed models and the reported statistical parameters are:

(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, in fact no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 1.522(8)\ \alpha - 0.009(2)\ \beta + 0.809(5)\ \log Kow + 0.003(4)\ \pi^* + 0.066(9)\ {}^1X^v - 1.861$

n= 29 R^2= 96.73% R^2 adj = 96.02% S.E.= 0.2337

(2) $\log(1/LC50) = 1.50\ \alpha + 0.809(5)\ \log Kow + 0.067(1)\ {}^1X^v - 1.862(9)$

n=29 R^2= 96.73% R^2adj = 96.34% S.E.= 0.2242

The statistical parameter reported by the authors are only related to the fitting performances, that are very good.

The authors performed external validation by using only one chemical (tetrachloromethane: the predicted value is satisfactory, but it is impossible to verify if by chance or by model quality, basing only on this unique observation)

The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figures 30 and 31. The authors did not point out that chemicals 3, 27 and 28 are outliers in both the models, while 5 is a strongly influential chemical only in model (1). The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors: the distribution of chemicals is regular.

VALIDATION:

The models were assessed in this contract work by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation was also performed by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (D-optimal Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets regarding to the structural information are reported below. The outliers are put in the training set in each following splitting, while the influential chemical 5 is put in training by D-optimal and Random and in validation by K-ANN splitting.

Table 6: Statistical Diagnostics of models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2boot	R^2adj	Q^2ext	MSE train	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
29	1	Tot.1	π^* β α $^1X^v$ logKow	-54.4	96.7	27.2	0.0	96.0		0.044		1.432	0.209	135.5	0.234	59.16	65.76	6.60
29	1	Tot.2	α $^1X^v$ logKow	94.7	96.7	93.9	94.1	96.3		0.044		0.266	0.209	245.5	0.225	39.05	56.76	17.71
24	5	K-ANN 1	π^* β α $^1X^v$ logKow	91.3	96.4	82.5	87.2	95.4	-0.3	0.048	2.041	0.338	0.217	93.5	0.251	64.07	69.59	5.52
24	5	K-ANN 2	α $^1X^v$ logKow	93.1	96.3	91.9	92.4	95.8	98.4	0.048	0.025	0.300	0.220	169.8	0.241	38.03	55.66	17.63
24	5	D-Opt. 1	π^* β α $^1X^v$ logKow	-105.7	96.9	26.3	0.0	96.0	95.1	0.047	0.031	1.753	0.216	107.8	0.250	59.21	65.73	6.52
24	5	D-Opt. 2	α $^1X^v$ logKow	94.5	96.9	93.5	93.8	96.4	95.4	0.047	0.029	0.286	0.217	198.7	0.238	39.00	56.43	17.43
24	5	Rand. 1	π^* β α $^1X^v$ logKow	-160.7	96.3	18.9	0.0	95.2	99.0	0.050	0.014	1.867	0.224	90.1	0.258	58.59	65.31	6.72
24	5	Rand. 2	α $^1X^v$ logKow	93.4	96.3	92.1	92.6	95.7	99.0	0.050	0.013	0.297	0.224	166.7	0.245	38.92	56.64	17.72

It is immediately evident that the model (1), even with excellent fitting performances (very high values of R^2 and R^2_{adj}), is overfitting and not predictive at all (Q^2 LOO with negative value, bootstrapping 0) SDEP>>SDEC

SDEP is similar to SDEC in models (2): the models have internal predictivity not too dissimilar from fitting power.

Regarding collinearity: in general, the descriptors are very correlated in models 1 (medium Kxx: 60) but, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy) and the correlation among the X (Kxx) in models with 5 descriptors is too small for model 1 (delta : 6). This is a signal of multicollinearity without prediction power (in fact when we applied the QUIK rule of Todeschini these models were excluded as predictive models). The model can also be considered overfitting: actually 5 descriptors for 29 chemicals are probably too many, thus unnecessary terms are included to fit the data, but these are clearly not useful for predictive purposes. On the contrary, the models with 3 descriptors have medium Kxx value of 39 and medium delta value of 18 and are better models.

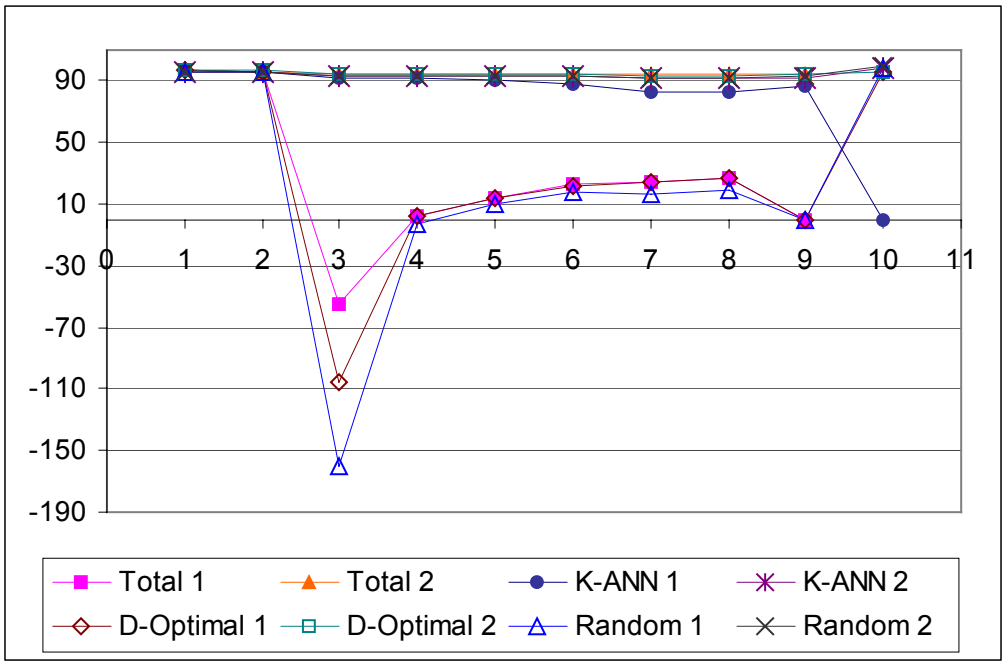
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted :

Table 6 bis: Statistical Diagnostics of models

	Split	Total 1	Total 2	K-ANN 1	K-ANN 2	D-Optimal 1	D-Optimal 2	Random 1	Random 2
1	R^2	96.7	96.7	96.4	96.3	96.9	96.9	96.3	96.3
2	R^2_{adj}	96.0	96.3	95.4	95.8	96.0	96.4	95.2	95.7
3	Q^2	-54.4	94.7	91.3	93.1	-105.7	94.5	-160.7	93.4
4	Q^2_{LMO10}	2.5	94.6	91.4	93.1	2.9	94.5	-2.5	93.3
5	Q^2_{LMO20}	14.2	94.5	90.6	92.9	13.8	94.5	10.0	93.1
6	Q^2_{LMO30}	22.8	94.3	88.4	92.7	21.8	94.1	17.4	92.8
7	Q^2_{LMO40}	24.5	93.9	82.4	91.9	24.0	93.5	15.9	92.2
8	Q^2_{LMO50}	27.2	93.9	82.5	91.9	26.3	93.5	18.9	92.1
9	Q^2_{boot}	0.0	94.1	87.2	92.4	0.0	93.8	0.0	92.6
10	Q^2ext			-0.3	98.4	95.1	95.4	99.0	99.0

The following is the graphical representation of the parameters reported in the above table.

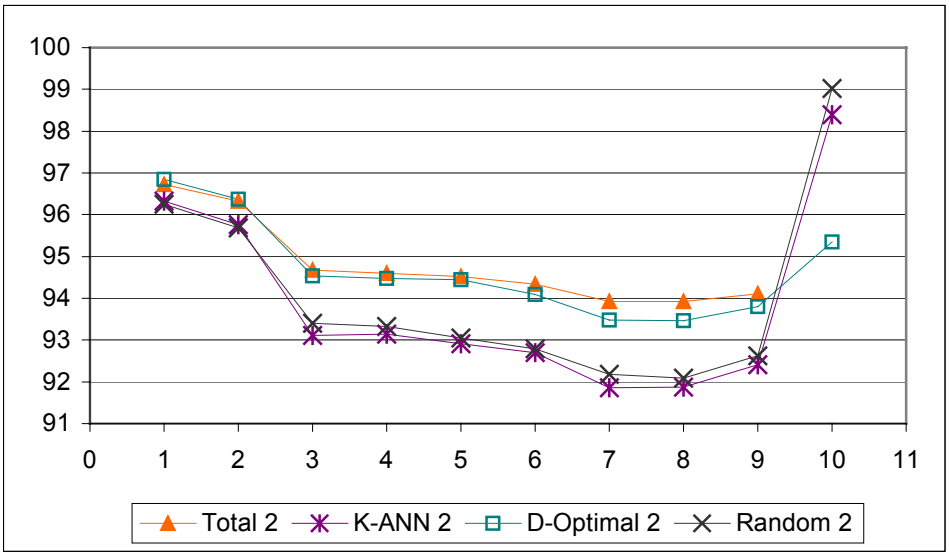


It is clearly evident that models with 5 descriptors (1) are unpredictive, while those with 3 descriptors (2) show high performance in both prediction and statistical external validation.

It is interesting to note that the model (1) verified by the K-ANN splitting appears internally predictive and completely unpredictive externally: in fact the validation set includes two chemicals that are outside

the chemical domain of the training set (n.5 and 26). This is a good demonstration that the chemical distribution between the training and validation sets strongly influences the statistical external validation results.

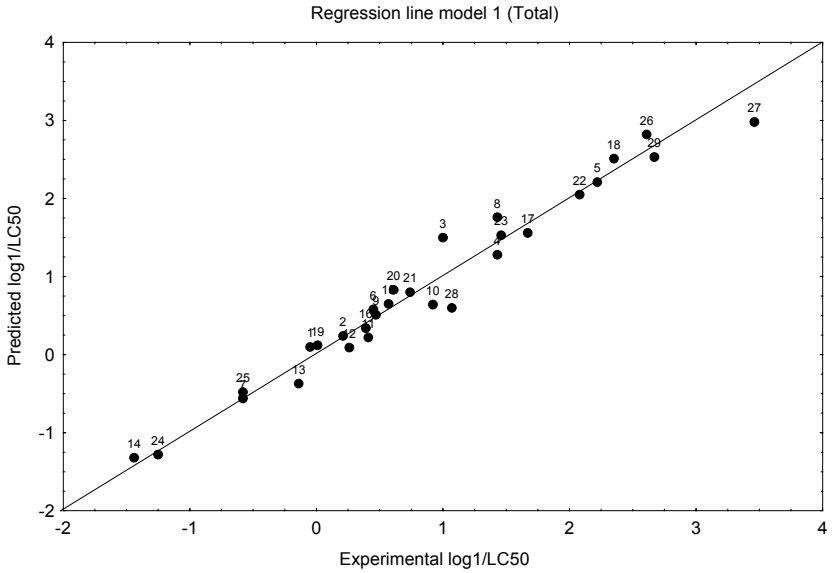
The following is an enlargement of the upper part of the previous graph, in which an over-optimism of the prediction power estimate based on Q^2_{EXT} is evident.

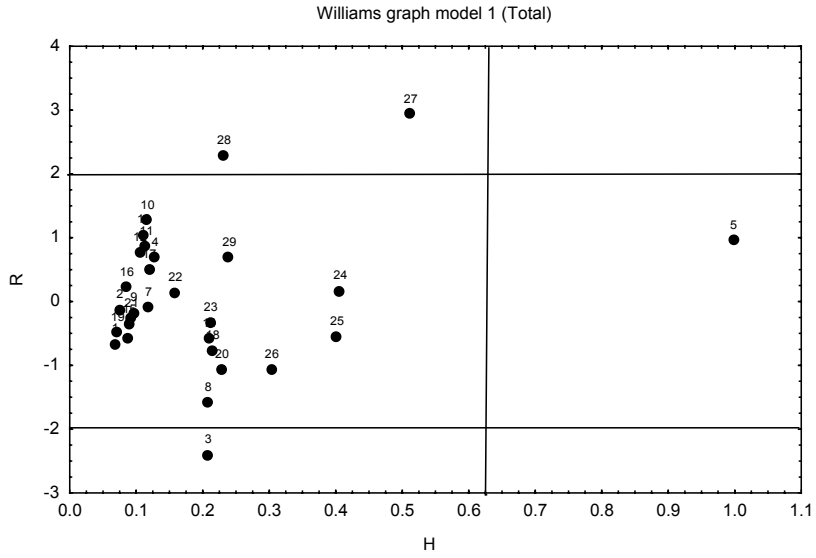


Model (2) can be considered a predictive model: internal validations give similar results, while statistical external validation (strongly influenced by the splitting) is over-optimistic in Q^2_{EXT} values. A more realistic idea of the quality of the predictivity can be obtained by comparing the MSE values (smaller for validation chemical than for training chemicals).

In Appendix the Leverage approach, applied to this data set.

FIGURE 30: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





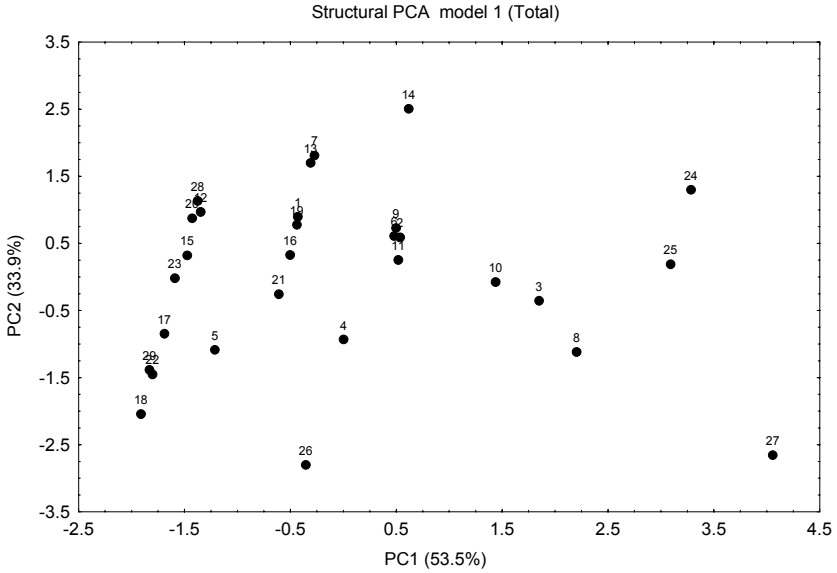
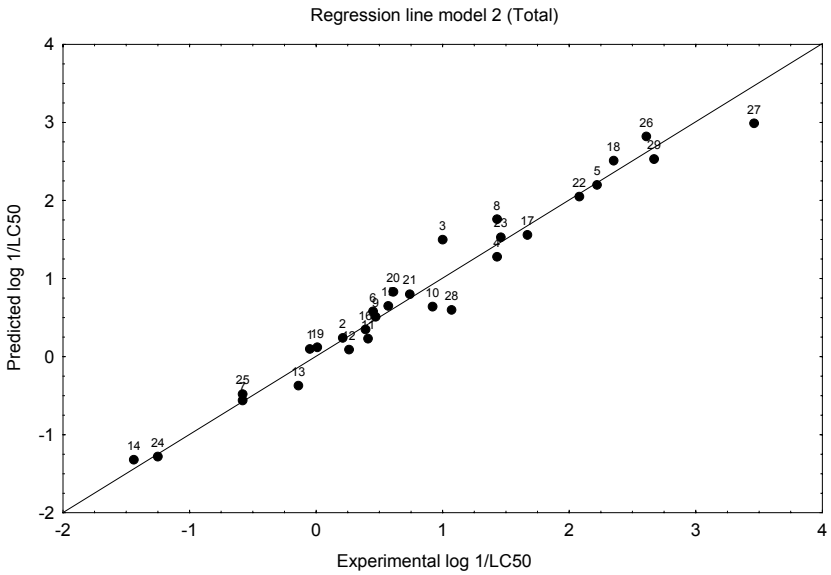
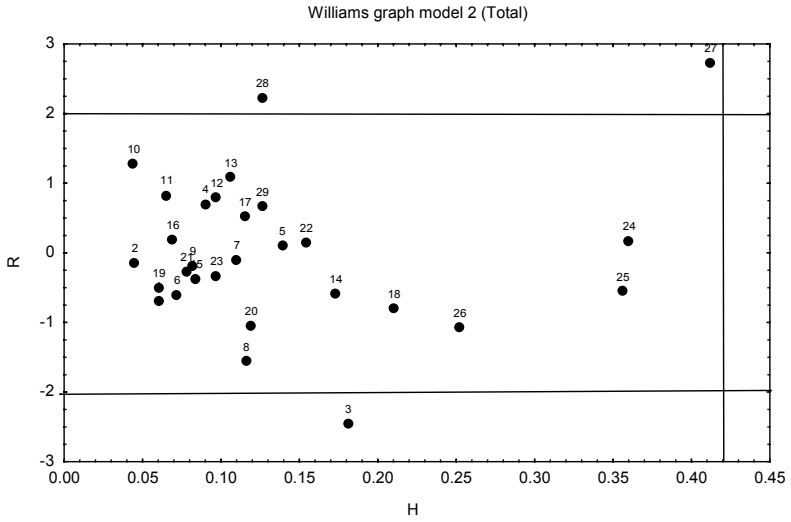


FIGURE 31: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





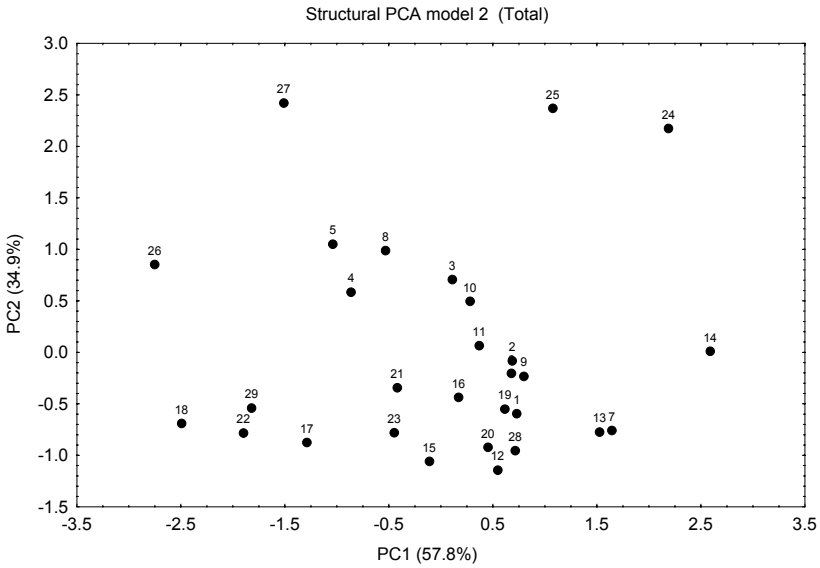
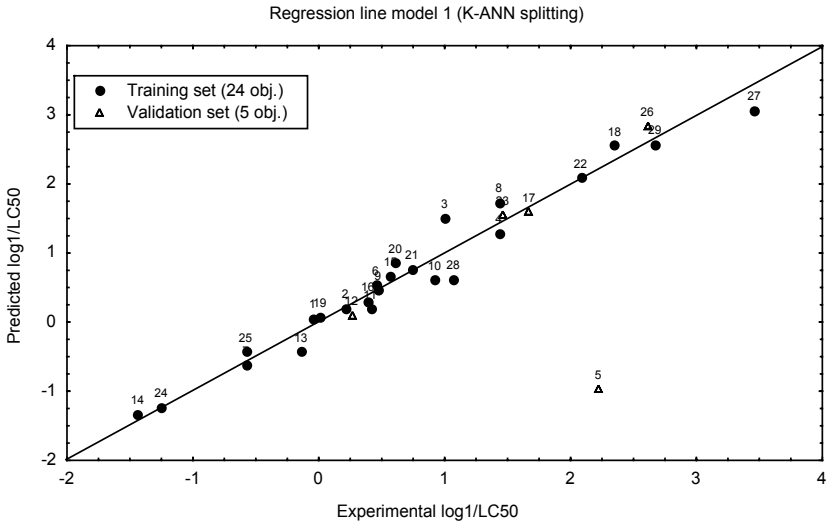
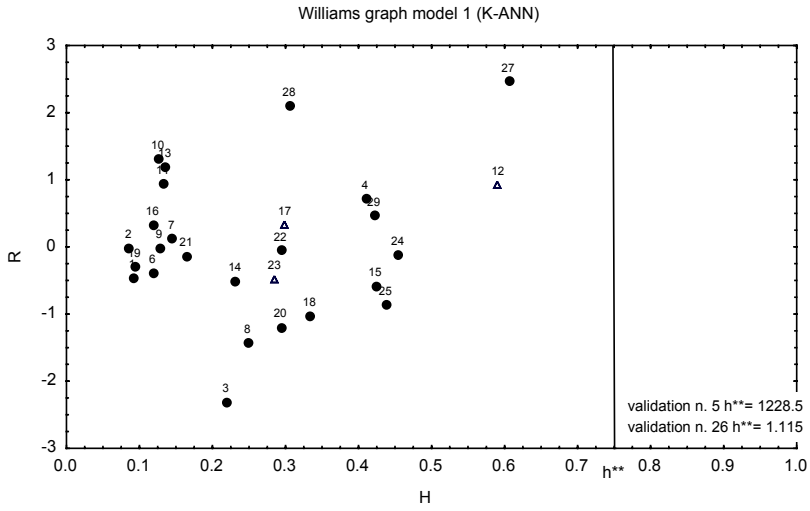


FIGURE 32: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





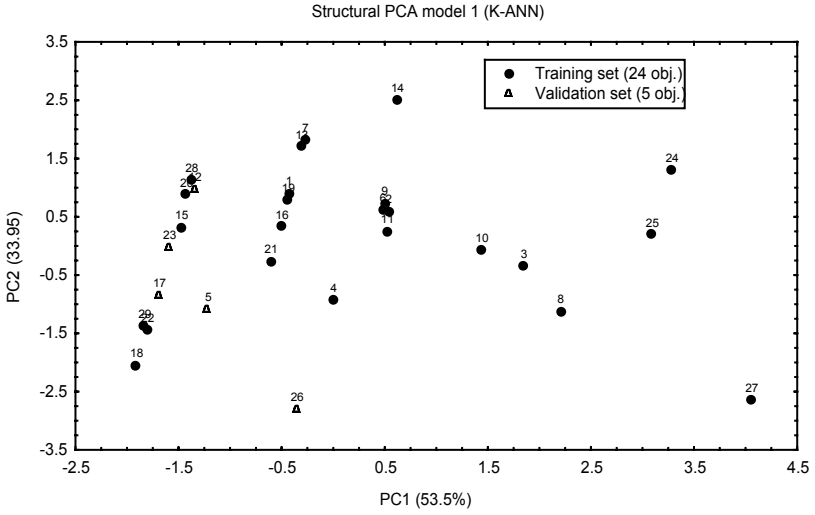
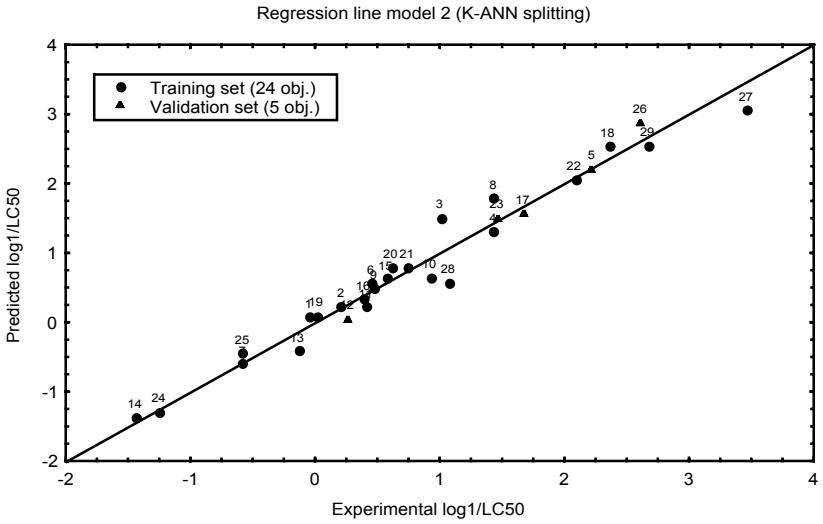
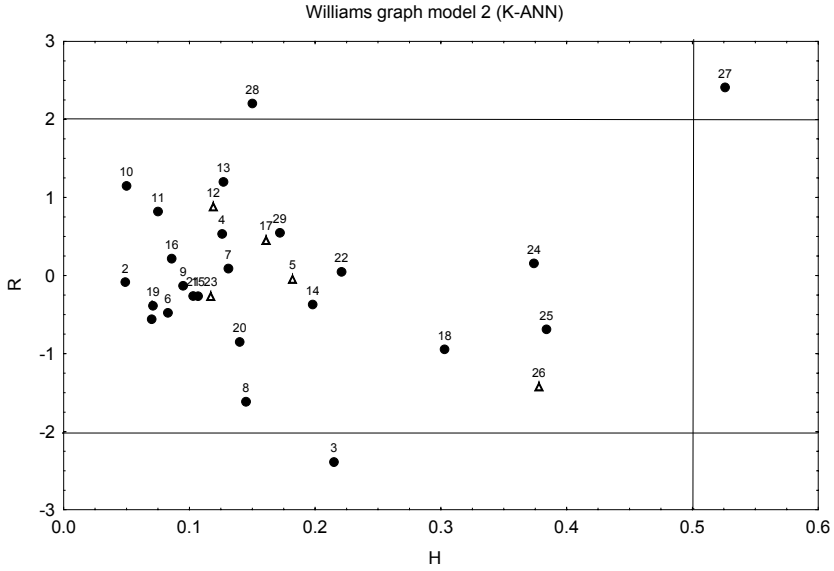


FIGURE 33: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





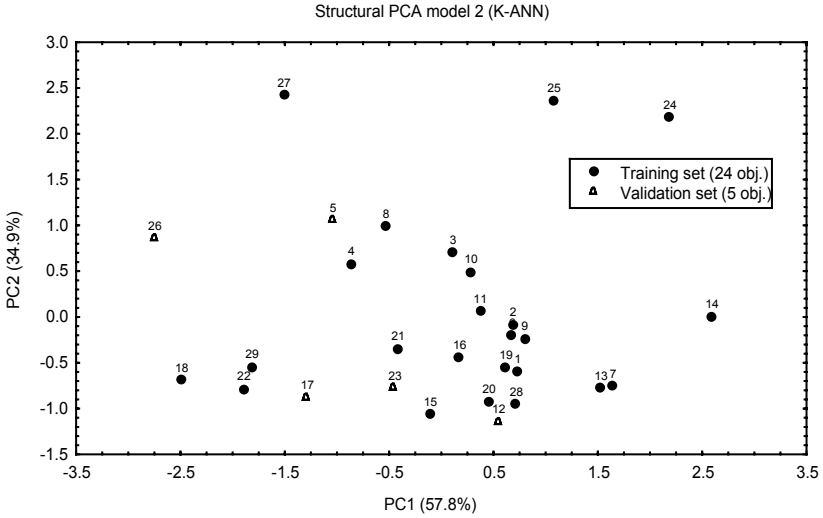
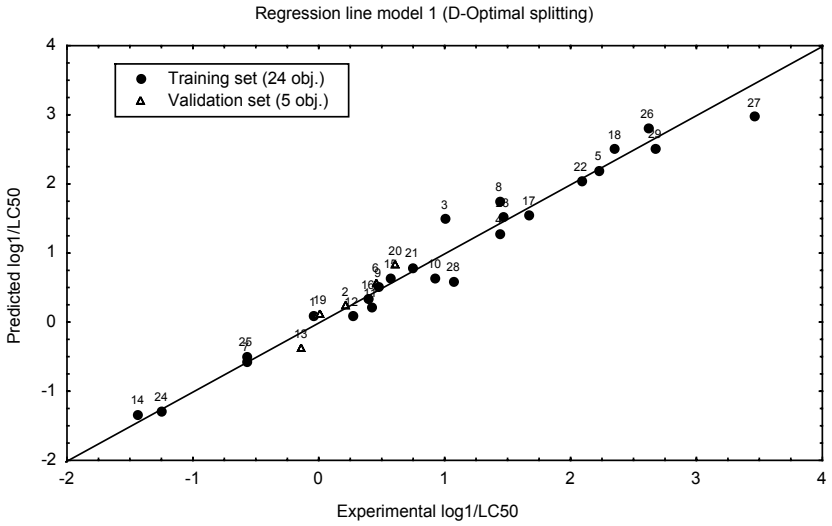
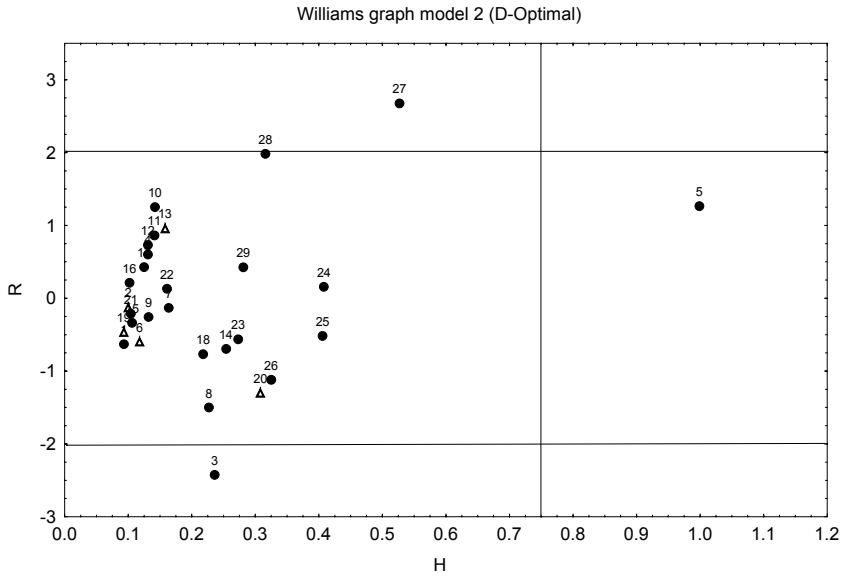


FIGURE 34: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





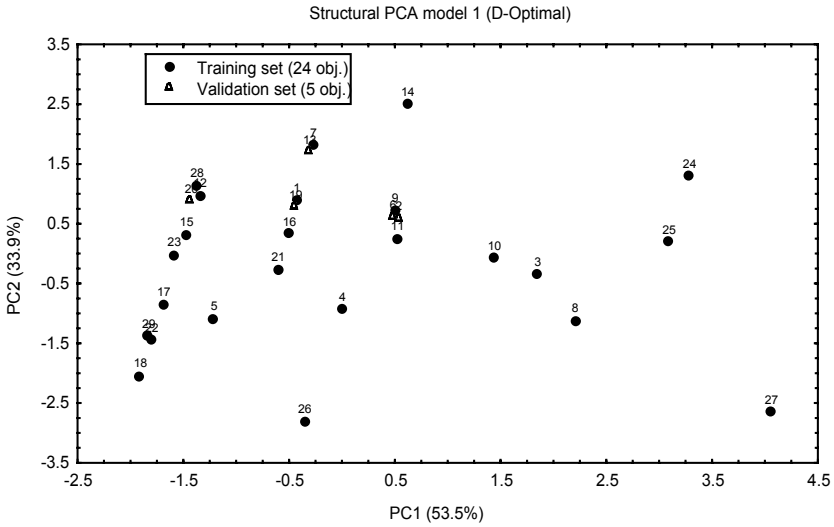
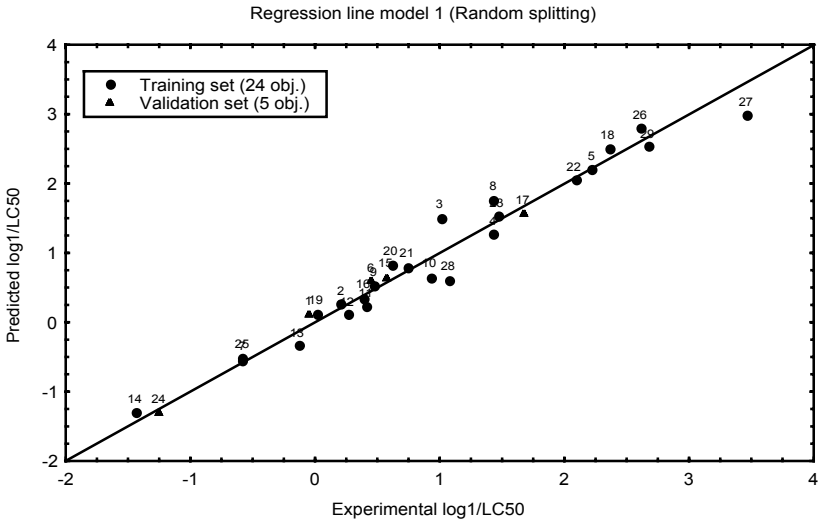
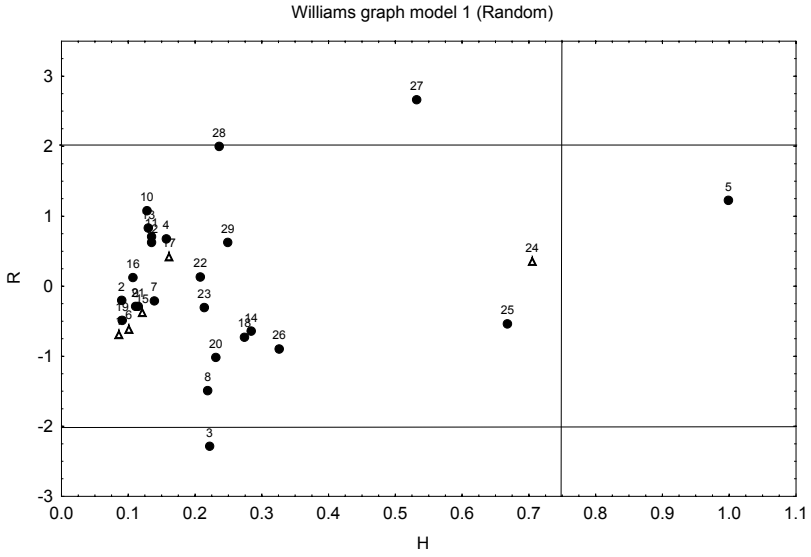


FIGURE 35: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





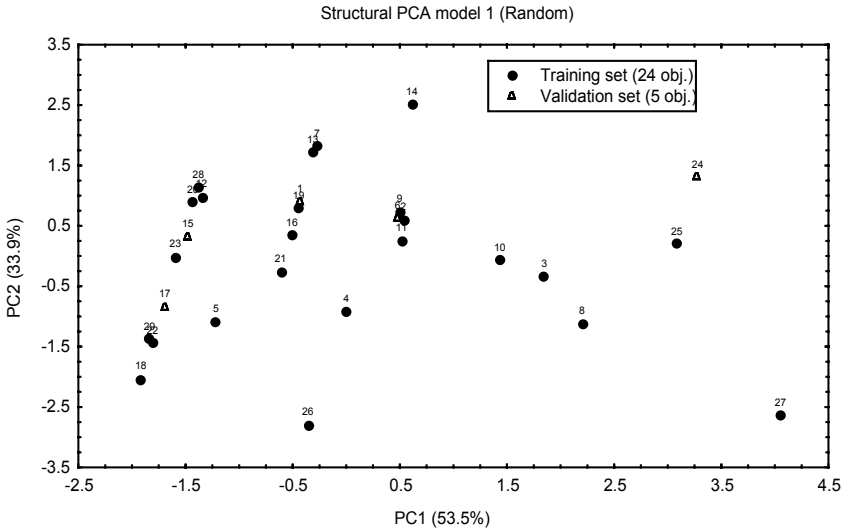
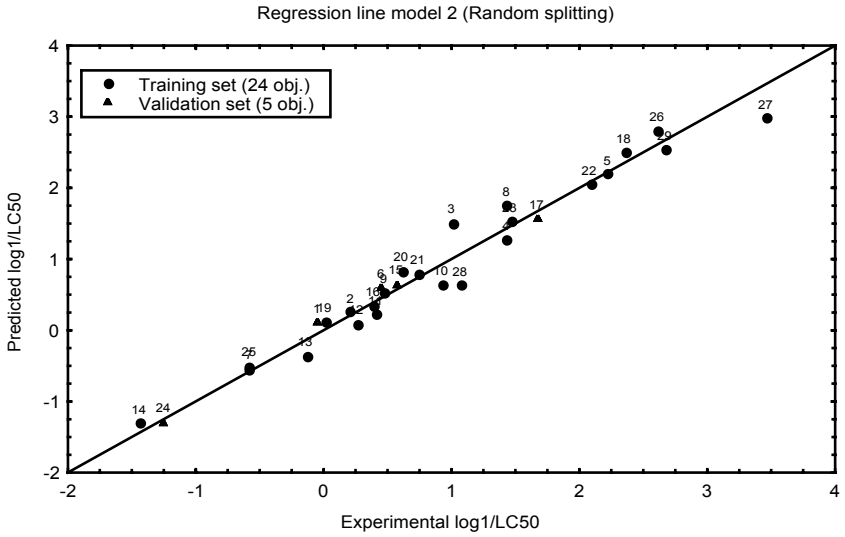
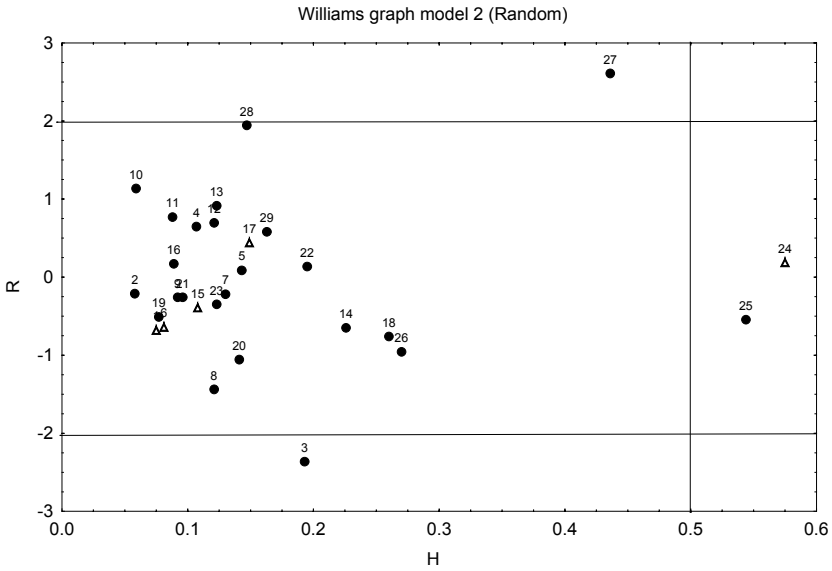
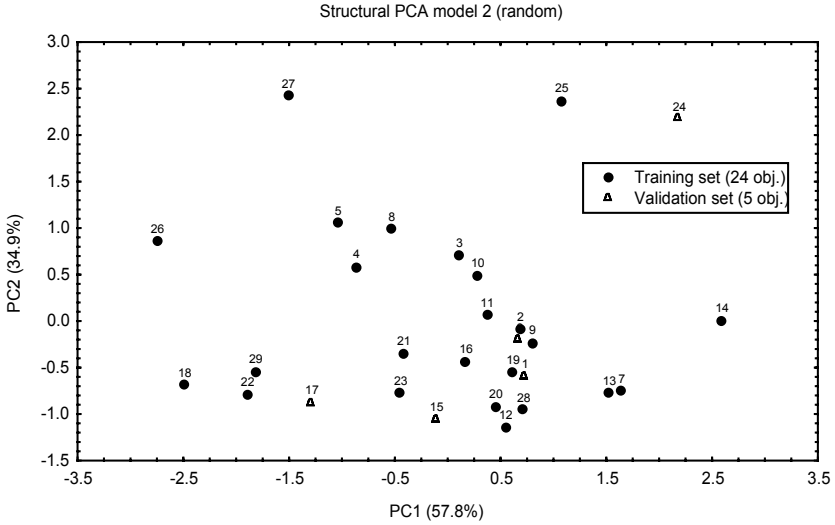


FIGURE 36: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







ESTERS:

A data set of 32 chemicals was studied (no outlier removed).

The proposed models and the reported statistical parameters are:

(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, in fact no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 0.133(4)\ \alpha - 1.298(4)\ \beta + 0.054(8)\ \log Kow + 1.491(8)\ \pi^* + 0.277(6)\ {}^1X^v - 0.542(9)$

n=32 R^2= 83.64% R^2 adj = 80.5% S.E.= 0.5229

(2) $\log(1/LC50) = -0.329\ \alpha + 1.184(7)\ \pi^* + 0.236(9)\ {}^1X^v - 0.895(6)$

n=32 R^2= 82.26% R^2adj = 80.36% S.E.= 0.5248

The statistical parameters reported by the authors are only related to the fitting performances, that are good.

The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figures 37 and 38. The authors did not point out that chemical 13 is an influential chemical in both models and that 23 is near the cut-off value in model (2). The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors. the anomalous distribution of the chemicals in the structural space of the descriptors is immediately evident: the degeneracy of some variables and variables correlation result in the perfect alignment of some chemicals.

VALIDATION:

The models were assessed in this contract work by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation was also performed by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (D-optimal Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets with regard to structural information are reported below.

Table 7: Statistical Diagnostics of models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
32		Tot.1	π* β α $^1X^v$ logKow	74.8	83.6	61.1	66.0	80.4		0.223		0.585	0.472	26.5	0.524	51.37	56.59	5.22
32		Tot.2	π* α $^1X^v$	77.0	82.2	69.7	72.4	80.3		0.242		0.559	0.492	43.1	0.525	34.05	52.69	18.64
26	6	K-ANN 1	π* β α $^1X^v$ logKow	73.2	85.2	50.4	72.3	81.5	79.0	0.182	0.423	0.573	0.426	22.8	0.49	51.26	56.62	5.36
26	6	K-ANN 2	π* α $^1X^v$	77.7	83.7	68.0	51.2	81.5	78.0	0.199	0.444	0.522	0.446	37.5	0.49	32.43	51.93	19.50
26	6	D-Opt. 1	π* β α $^1X^v$ logKow	74.5	85.6	50.9	58.5	82.0	75.9	0.179	0.463	0.561	0.422	23.6	0.48	51.20	56.66	5.46
26	6	D-Opt. 2	π* α $^1X^v$	78.3	83.8	67.0	71.5	81.6	75.8	0.200	0.464	0.518	0.448	37.6	0.49	33.12	52.40	19.28
26	6	Rand. 1	π* β α $^1X^v$ logKow	72.2	84.4	50.7	60.6	80.6	79.1	0.197	0.372	0.592	0.443	21.6	0.51	44.26	51.45	7.19
26	6	Rand. 2	π* α $^1X^v$	74.8	83.1	63.6	68.8	80.8	78.9	0.214	0.376	0.564	0.462	35.9	0.5	28.96	49.20	20.24

It is immediately evident that the model (1), even with better fitting performance (see values of R^2 and R^2_{adj}), is worse than model (2) in predictive ability (lower Q^2 LOO and higher SDEP). Also stronger internal validation highlights this point (anomalous result only for bootstrapping on K-ANN 2 splitting)

Regarding collinearity: in general, the descriptors are very correlated (medium Kxx: 41) but, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy), and the correlation among the X (Kxx) in models (1) is generally small (delta: 6), while in models (2) it is higher (delta: 19) . Model (1) is the more instable.

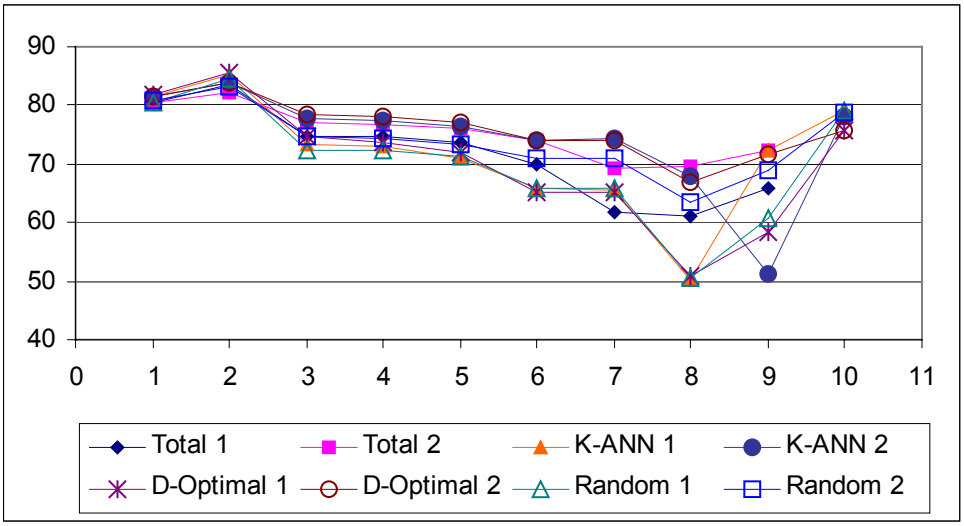
All the models were also verified by Y-scrambling: compared to the published models, the models on randomised response all have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and plotted in the graph:

Table 7 bis: Statistical Diagnostics of models

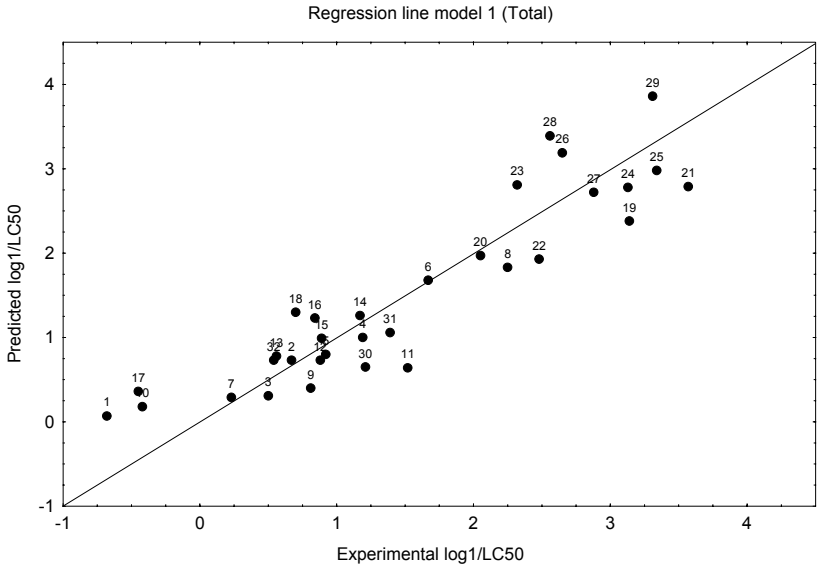
		Total 1	Total 2	K-ANN 1	K-ANN 2	D-Optimal 1	D-Optimal 2	Random 1	Random 2
1	R^2	80.4	80.3	81.5	81.5	82.0	81.6	80.6	80.8
2	R^2_{adj}	83.6	82.2	85.2	83.7	85.6	83.8	84.4	83.1
3	Q^2	74.8	77.0	73.2	77.7	74.5	78.3	72.2	74.8
4	Q^2_{LMO10}	74.6	76.8	73.0	77.5	73.7	78.2	72.5	74.3
5	Q^2_{LMO20}	73.7	76.1	71.1	76.5	72.1	77.2	71.3	73.5
6	Q^2_{LMO30}	69.8	74.1	65.8	74.2	65.3	74.1	65.8	71.1
7	Q^2_{LMO40}	61.6	69.3	65.4	74.3	65.2	74.1	65.7	71.1
8	Q^2_{LMO50}	61.1	69.7	50.4	68.0	50.9	67.0	50.7	63.6
9	Q^2_{boot}	66.0	72.4	72.3	51.2	58.5	71.5	60.6	68.8
10	Q^2ext			79.0	78.0	75.9	75.8	79.1	78.9

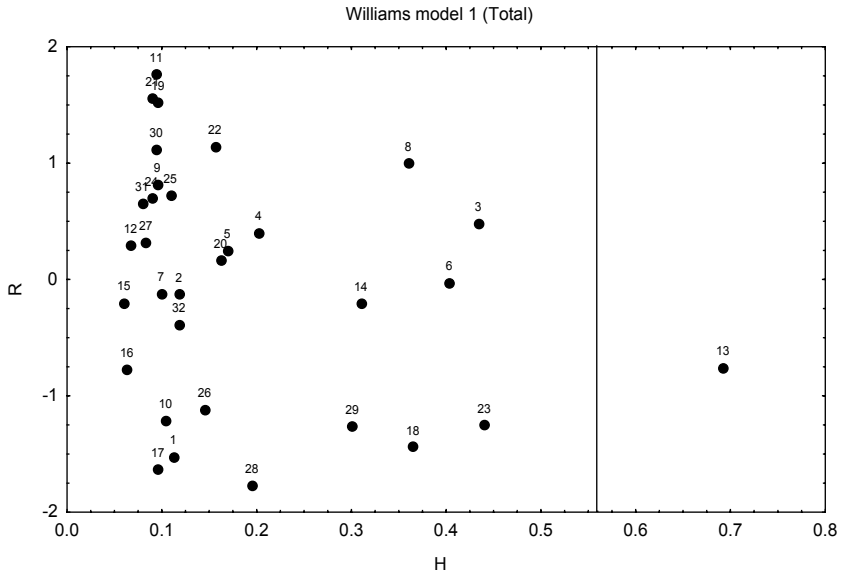
The following is the graphical representation of the parameters reported in the above table.



Only model (2) can be considered sufficiently stable and a sufficiently predictive model: internal validations give similar results, while statistical external validation (strongly influenced by the splitting) is over-optimistic in Q^2_{EXT} values. A more realistic idea of the quality of the predictivity can be obtained by comparing the MSE values: the validation chemicals are predicted two times worse than the training chemicals

FIGURE 37: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





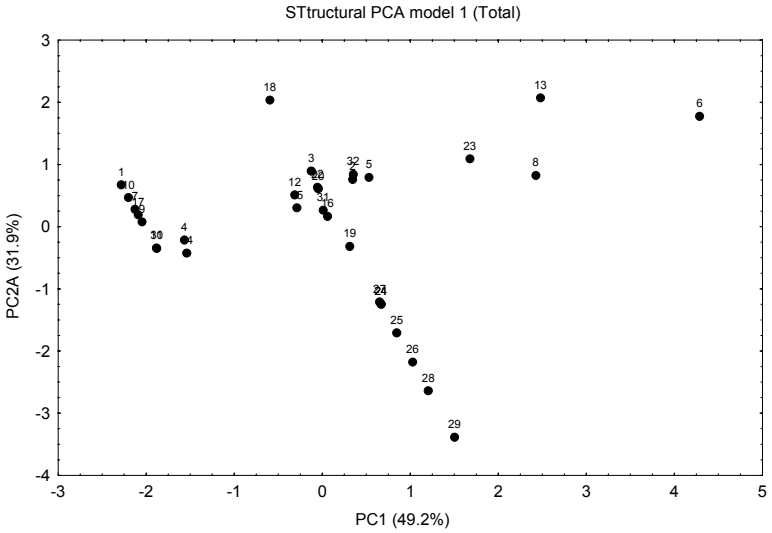
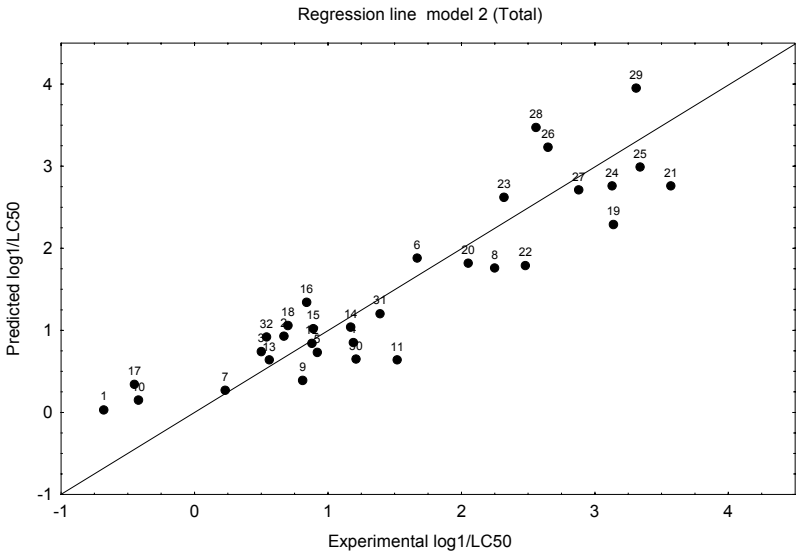
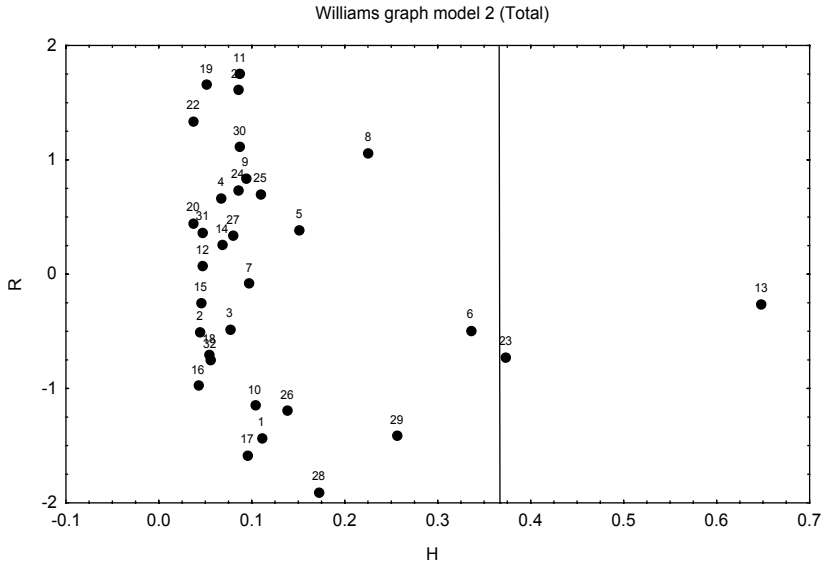


FIGURE 38: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





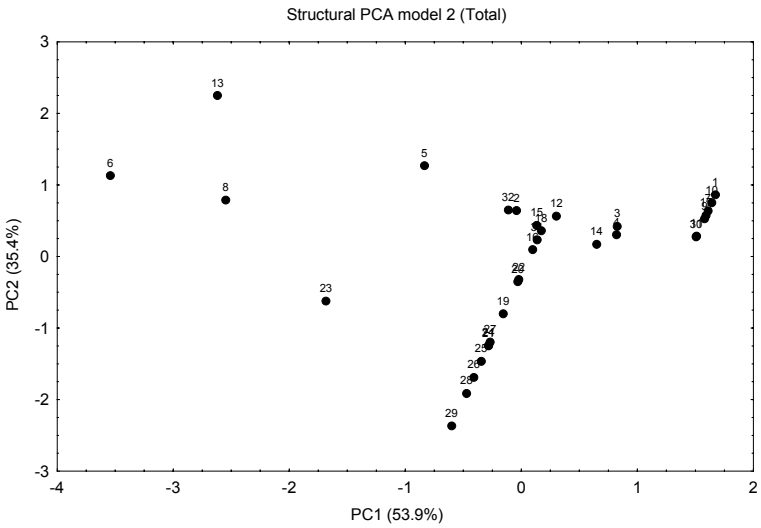
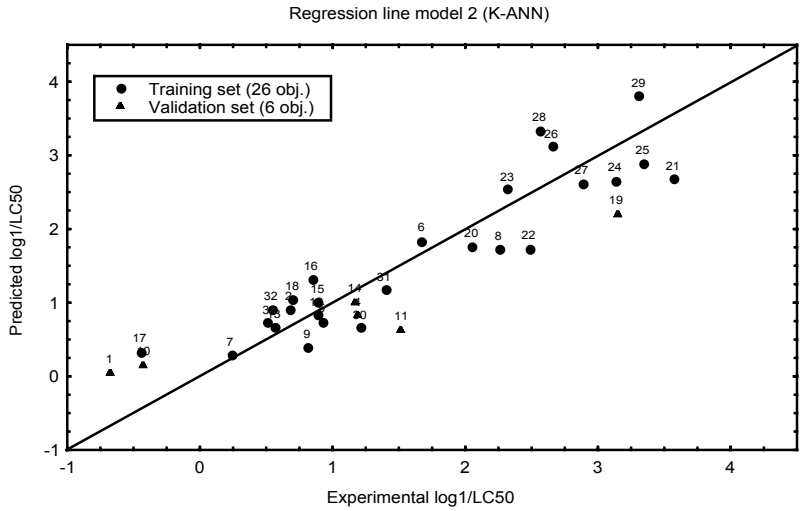
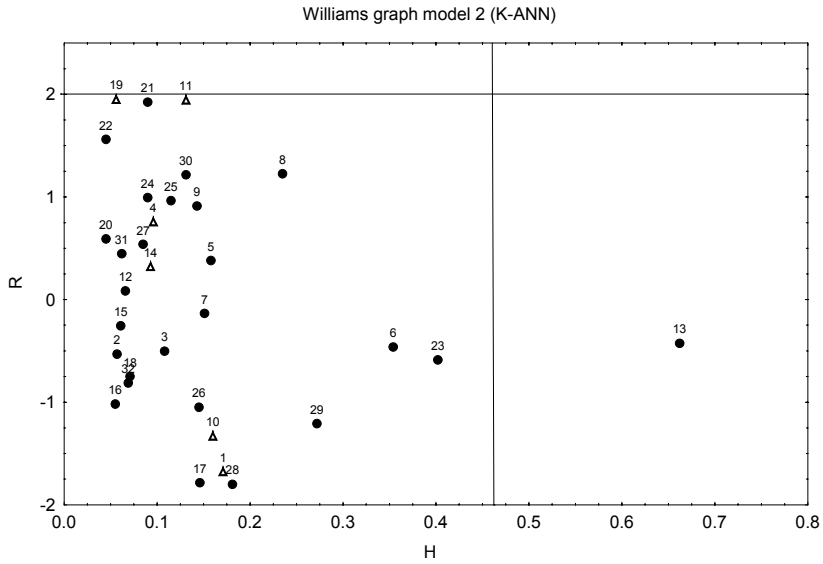


FIGURE 39: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





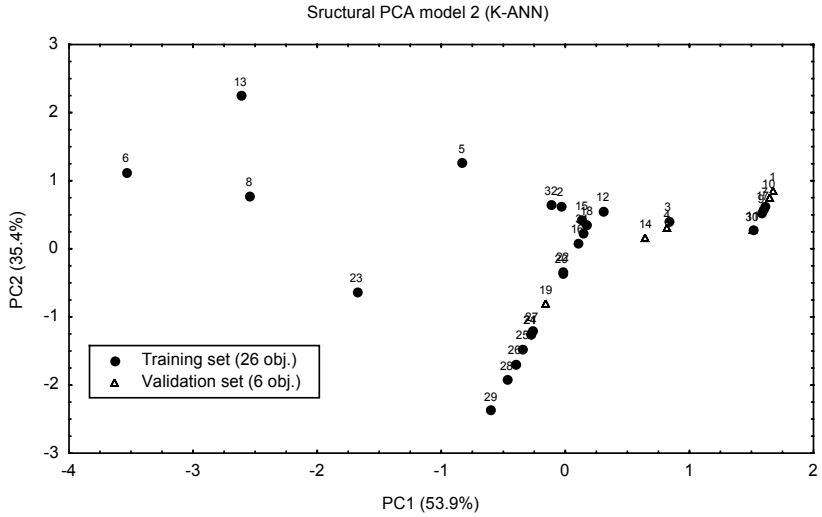
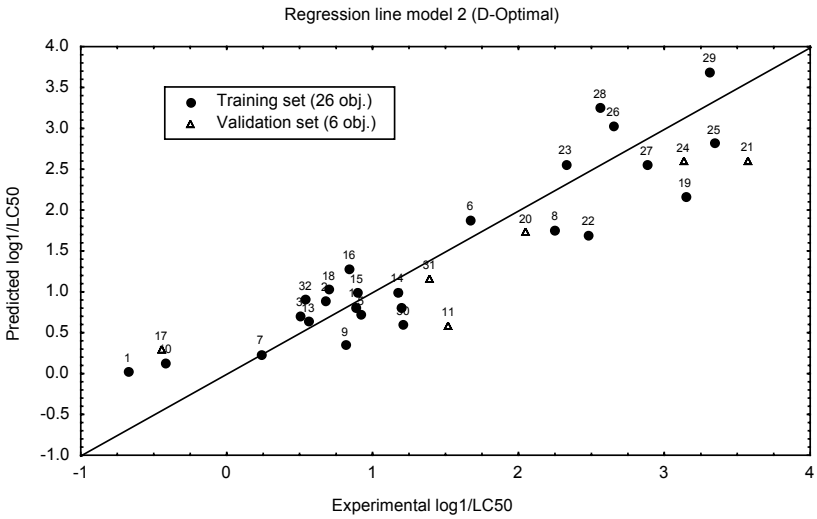
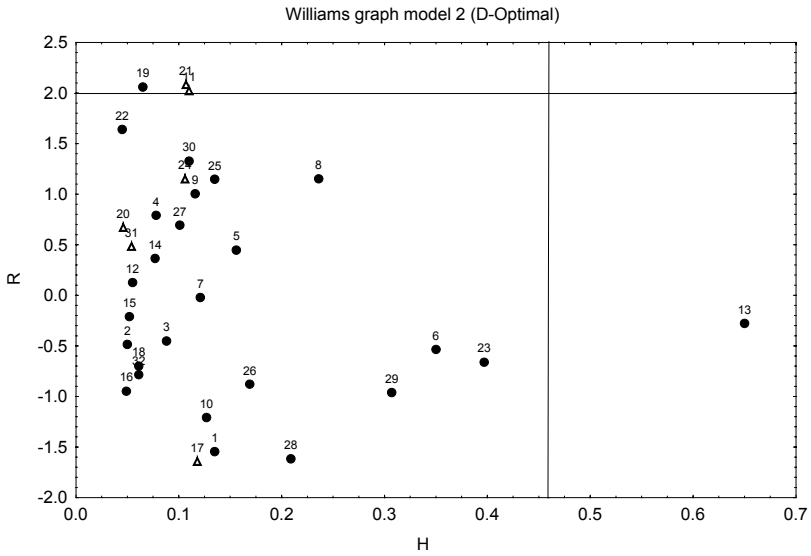


FIGURE 40: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





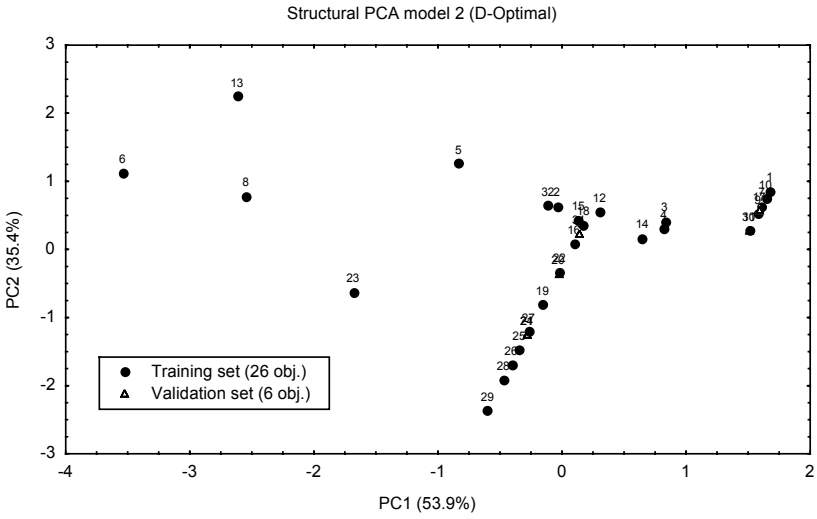
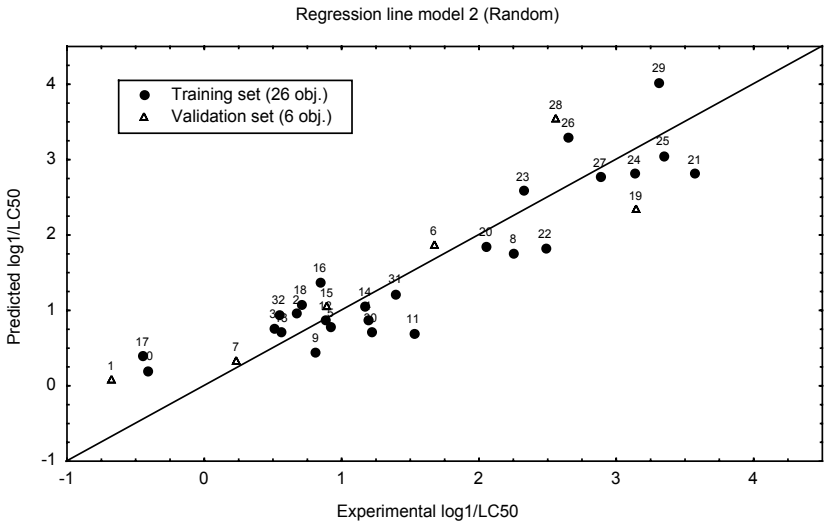
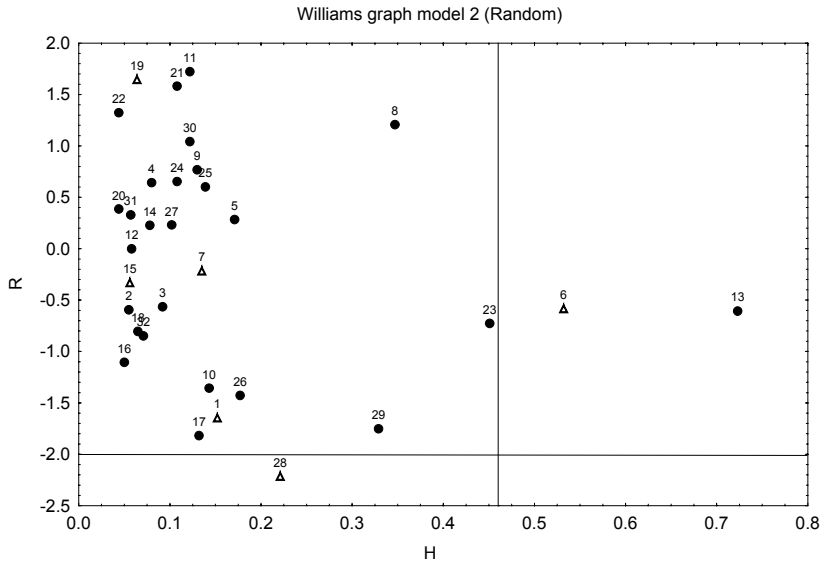
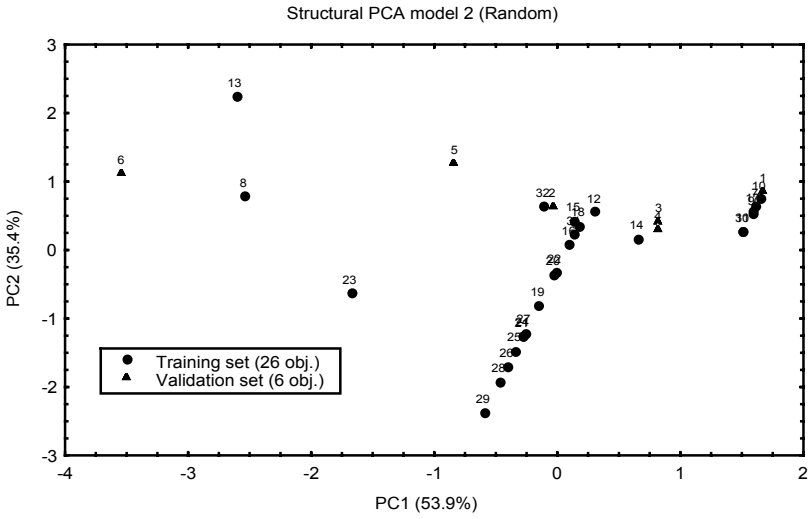


FIGURE 41: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.







KETONES:

The data set is composed of 34 chemicals: 6 are outliers and were removed before the modelling by the authors.

The proposed models and the reported statistical parameters are:

(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 0.842(9)\,\alpha - 0.882(6)\,\beta + 0.634(8)\,\log Kow + 0.417(3)\,\pi^* + 0.260(6)\,{}^1X^v - 1.914$

n=28 R^2= 95.32% R^2 adj = 94.26% S.E.= 0.2836

(2) $\log(1/LC50) = 0.514(7)\,\alpha + 0.804\,\log Kow + 0.360(6)\,\pi^* - 1.795(5)$

n=28 R^2= 92.39% R^2adj = 91.44% S.E.= 0.3465

The statistical parameter reported by the authors relate only to fitting performance that is very good.

The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figures 42 and 43. The authors did not point out that chemical 13 is an outlier in both models, while 22 is an influential chemical in all the models and 9 in Model (1). Depending on the splitting, some chemicals in the validation sets (10, 11, 17 and 22) are beyond the threshold level of H, thus their predicted values could be unreliable.

The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors and any possible anomalous or isolated chemicals. The anomalous distribution of the chemicals in the structural space of the descriptors is immediately evident: the degeneracy of some variables, and the correlation among the variables, results in the perfect alignment of some chemicals.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation was also performed by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (D-optimal Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets regarding structural information are reported below.

Table 8: Statistical Diagnostics of models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE train	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
28		Tot.1	π* β α $^1X^v$ logKow	90.5	95.3	68.0	76.5	94.3		0.063		0.359	0.251	87.7	0.284	52.1	60.2	8.1
28		Tot.2	π* α logKow	88.4	92.4	77.5	81.5	94.3		0.102		0.396	0.320	96.0	0.346	25.3	47.8	22.5
19	9	K-ANN 1	π* β α $^1X^v$ logKow	83.0	97.3	34.6	0.0	96.3	70.1	0.030	0.557	0.438	0.173	91.3	0.209	59.2	65.9	6.7
19	9	K-ANN 2	π* α logKow	95.7	97.0	63.9	74.9	96.3	80.5	0.034	0.362	0.220	0.185	154.0	0.208	31.9	51.7	19.8
19	9	D-Opt. 1	π* β α $^1X^v$ logKow	90.4	97.3	46.2	51.9	96.3	91.5	0.026	0.121	0.368	0.195	90.6	0.235	48.4	57.2	8.8
19	9	D-Opt. 2	π* α logKow	85.5	93.0	63.3	67.4	91.6	92.0	0.098	0.114	0.451	0.314	65.6	0.353	22.1	45.4	23.3
19	9	Rand. 1	π* β α $^1X^v$ logKow	91.0	98.6	29.2	0.0	98.1	55.9	0.026	0.179	0.403	0.159	170.0	0.192	53.6	61.8	8.2
19	9	Rand. 2	π* α logKow	991.1	97.4	61.0	37.6	96.9	33.0	0.046	0.272	0.399	0.214	183.4	0.241	26.7	49.6	22.8

It is immediately evident that Model (1), even with excellent fitting performance (very high values of R^2 and R^2_{adj}), is probably not predictive: in fact while the internal validation parameters demonstrate satisfactory predictive ability, statistical external validation highlights that the predictivity for new chemicals (different in different splittings) could be lower.

SDEP are very much higher than SDEC in Models (1): the models work in prediction in a decisively worse way than in calculation.

Model (2) is the best, being more stable and with predictivity verified by internal and statistical external validations. Again it is demonstrated that the addition of descriptors can improve the fitting, but not the predictivity, of a model. In this case, SDEP are not too much higher than SDEC: the models work slightly worse in prediction than in calculation.

Regarding collinearity: in general, the descriptors are very correlated (medium Kxx: 53) in model (1) and the difference in correlation between the block of X variables plus response Y (Kxy) and that of X (Kxx) in Model (1) is small (delta:8). This is a signal of multicollinearity without prediction power: unnecessary terms are included to fit the data, but these are clearly not useful for predictive purposes. The correlation among the descriptors in Model (2) is lower (medium Kxx: 53) and, most importantly, there is a significant increase in the correlation with response (medium delta: 22).

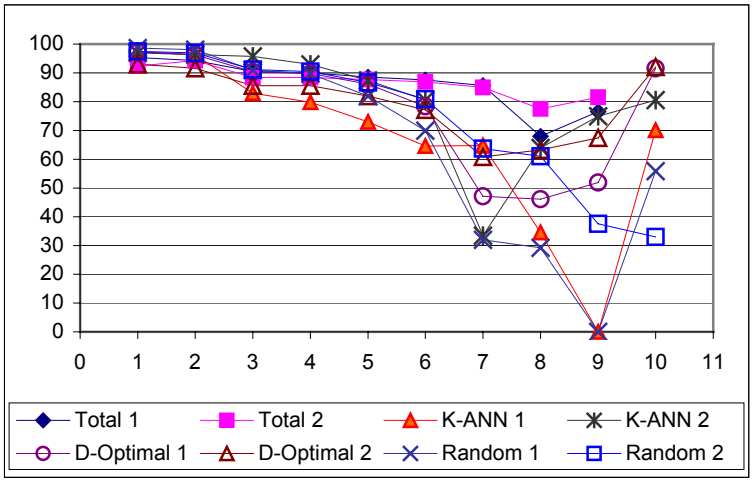
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted :

Table 8 bis: Statistical Diagnostics of models

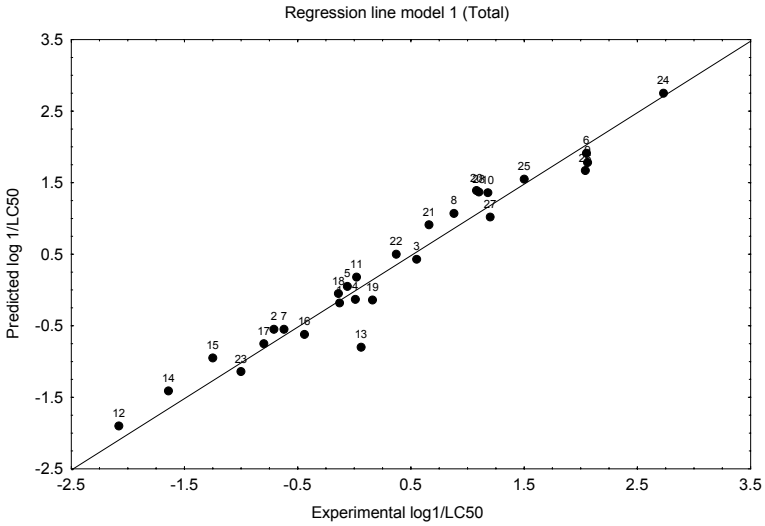
		Total 1	Total 2	K-ANN 1	K-ANN 2	D-Optimal 1	D-Optimal 2	Random 1	Random 2
1	R^2	95.3	92.4	97.3	97.0	97.3	93.0	98.6	97.4
2	R^2_{adj}	94.3	94.3	96.3	96.3	96.3	91.6	98.1	96.9
3	Q^2	90.5	88.4	83.0	95.7	90.4	85.5	91.0	91.1
4	Q^2_{LMO10}	90.0	88.4	79.8	92.9	89.8	85.5	89.7	90.4
5	Q^2_{LMO20}	88.6	87.6	73.0	87.4	86.1	81.9	81.9	86.7
6	Q^2_{LMO30}	87.6	86.9	64.6	80.5	78.4	77.1	69.9	80.9
7	Q^2_{LMO40}	85.6	84.9	64.7	33.5	47.2	60.8	32.0	63.7
8	Q^2_{LMO50}	68.0	77.5	34.6	63.9	46.2	63.3	29.2	61.0
9	Q^2_{boot}	76.5	81.5	0.0	74.9	51.9	67.4	0.0	37.6
10	Q^2ext			70.1	80.5	91.5	92.0	55.9	33.0

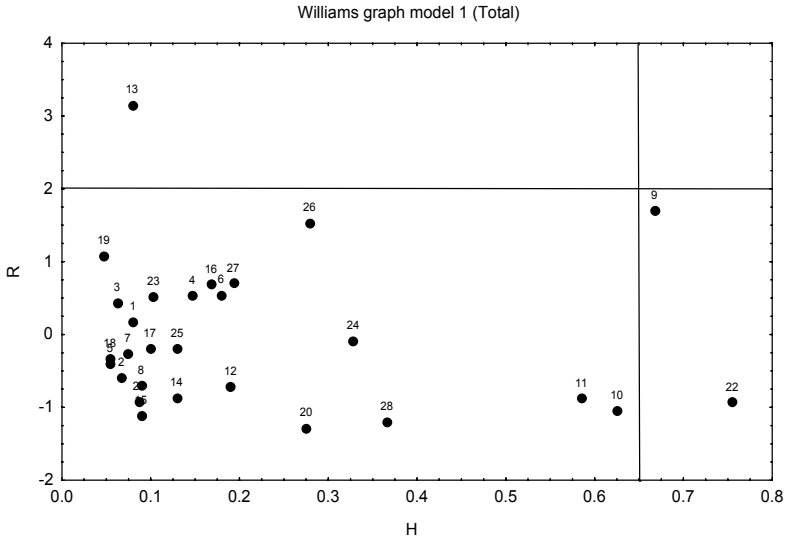
The following is the graphical representation of the parameters reported in the above table.



Models with 5 descriptors (1) are the worst and are, in general, scarcely predictive, while those with 3 descriptors (2) are, in general, more stable (exception Random 2) and can be considered sufficiently predictive. The Q^2_{EXT} values (again dependent on the splitting methodology) are over-optimistic and a more realistic idea of the quality of the predictivity can be obtained by comparing the MSE values: the validation chemicals are predicted worse than the training chemicals, but not in a dramatic way and again in an amount strongly dependent on the splitting.

FIGURE 42: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





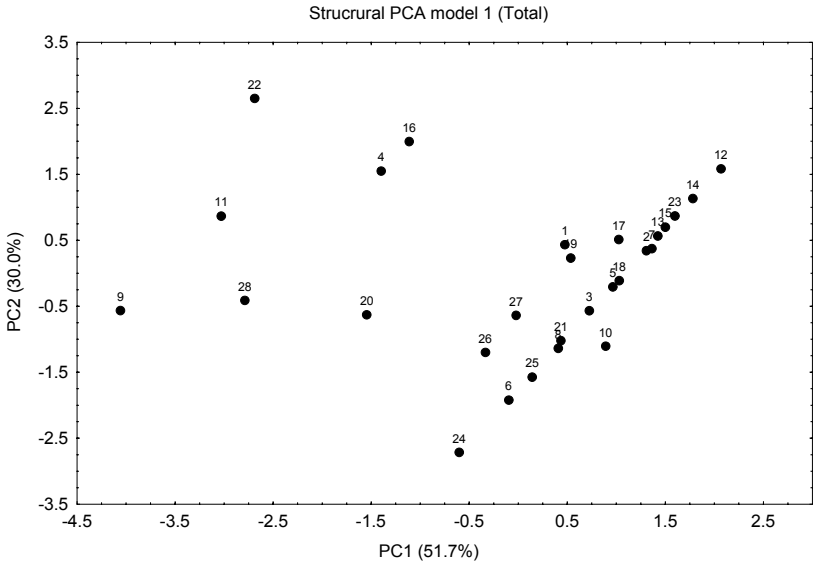
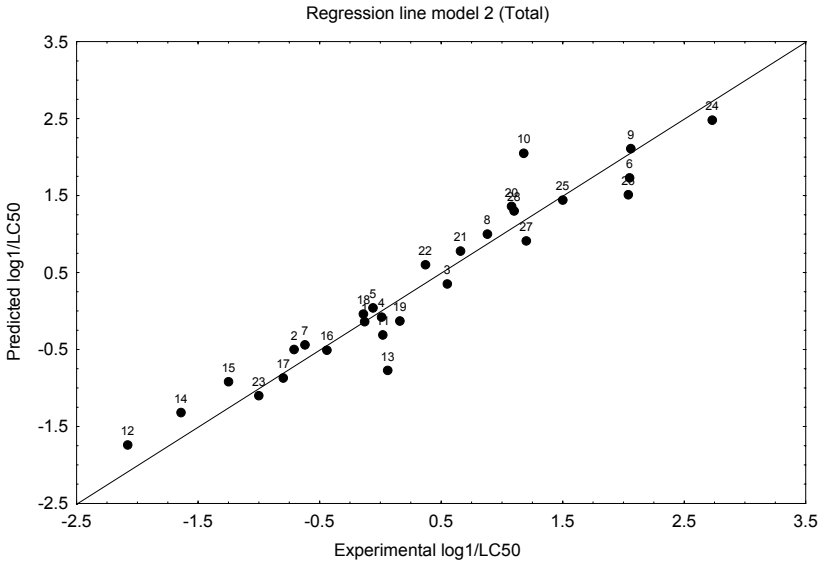
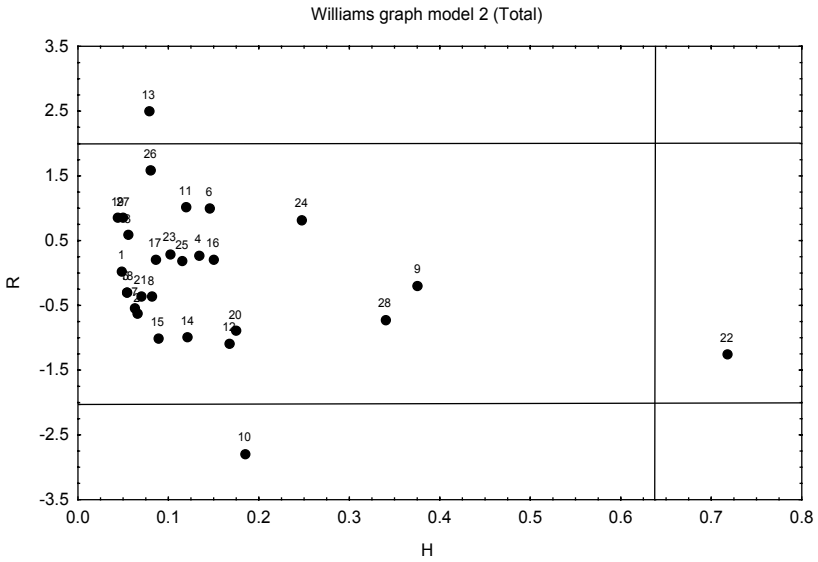


FIGURE 43: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





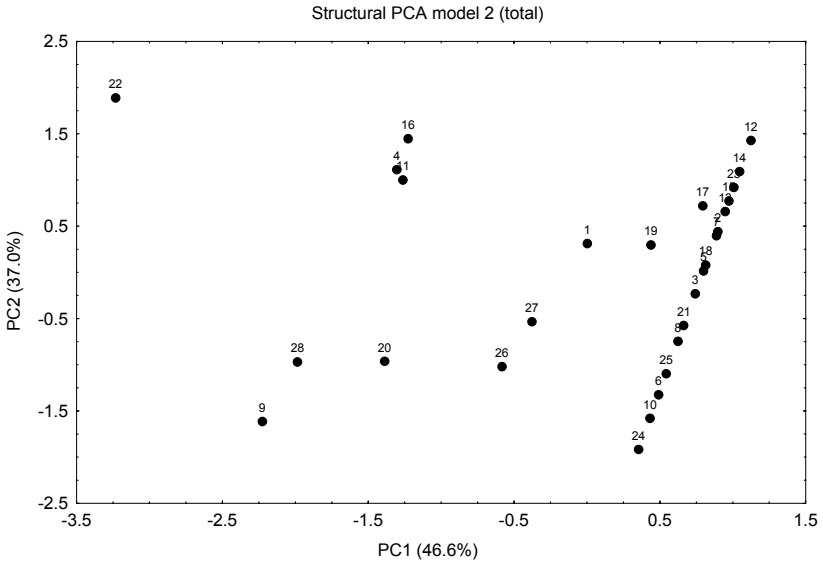
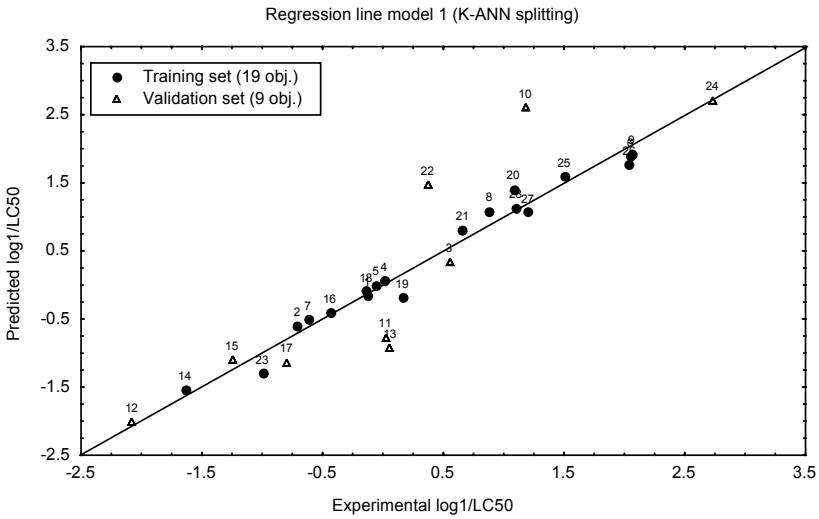
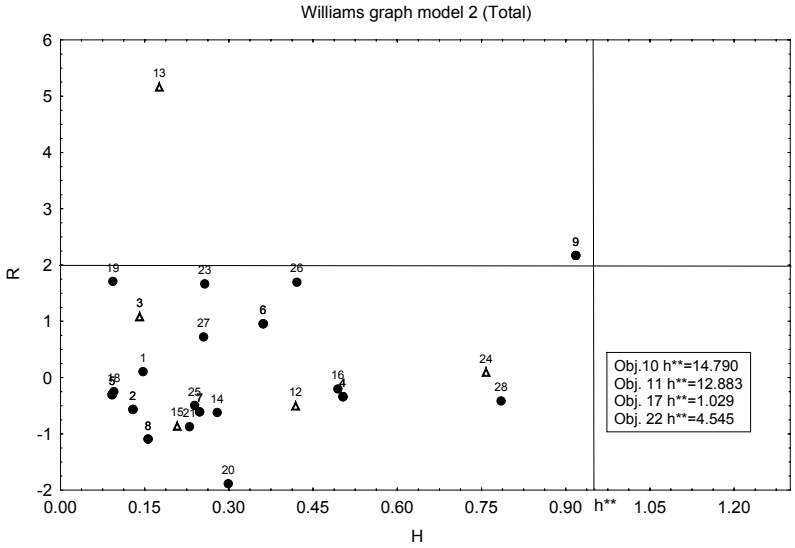


FIGURE 44: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





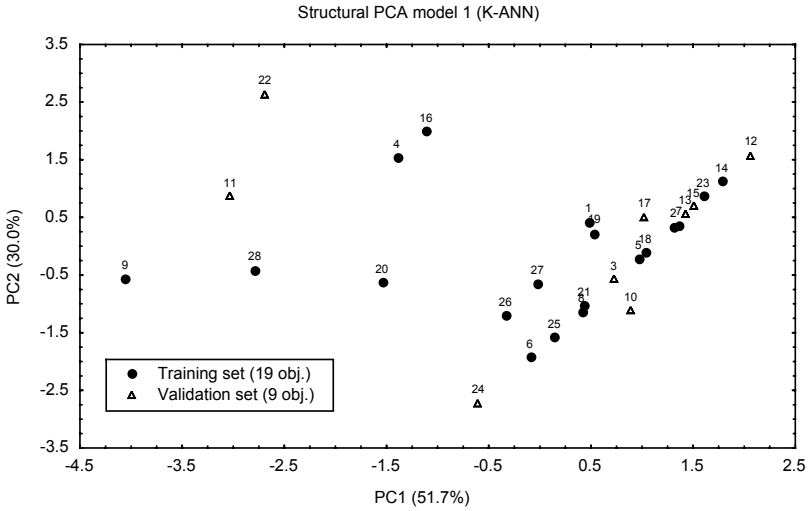
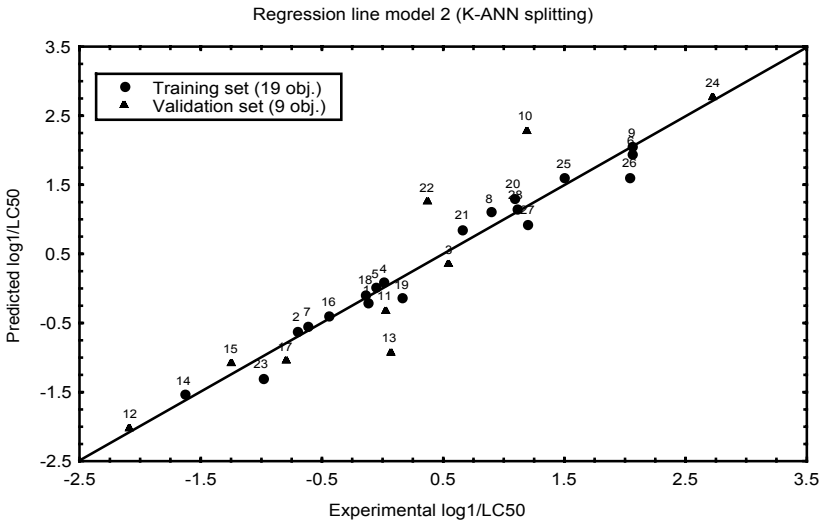
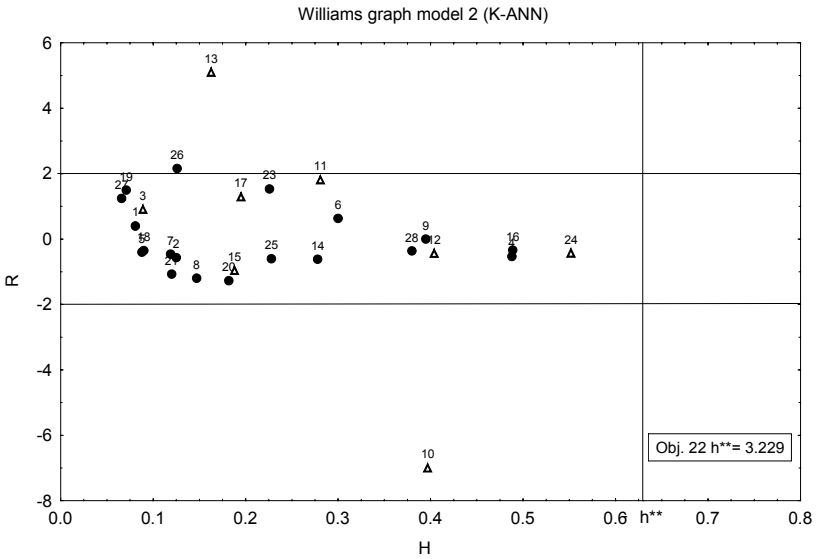


FIGURE 45: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





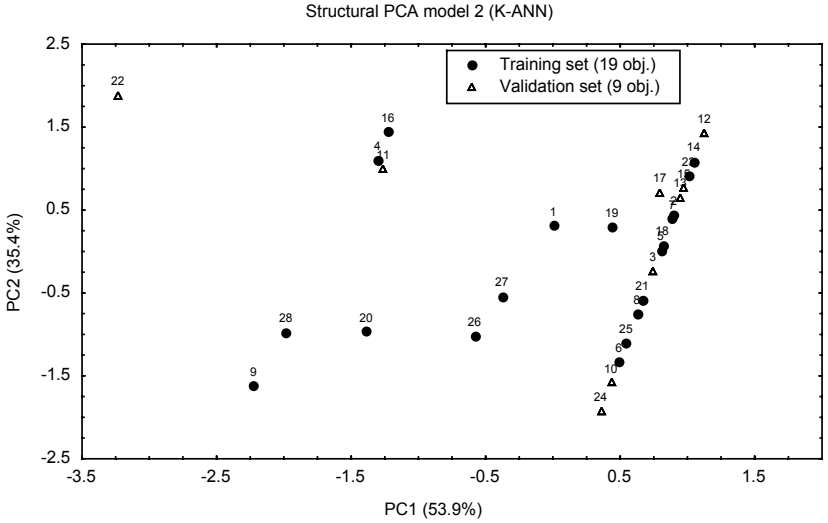
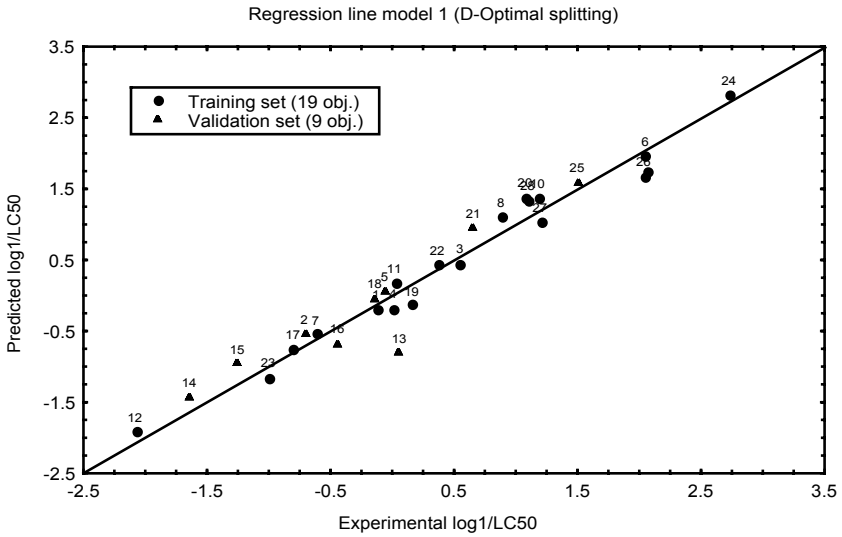
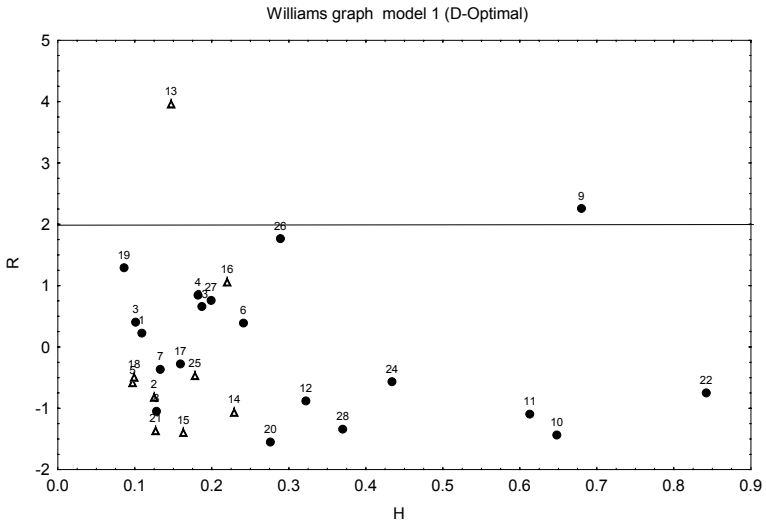


FIGURE 46: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





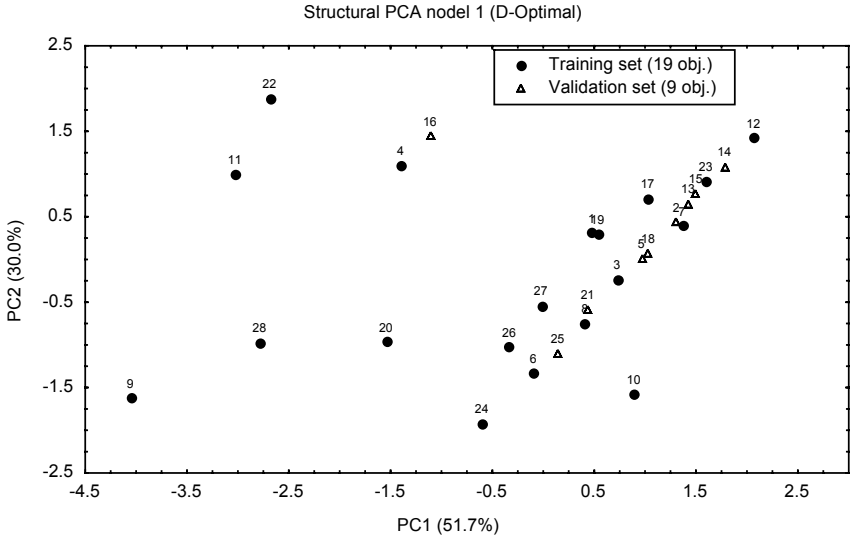
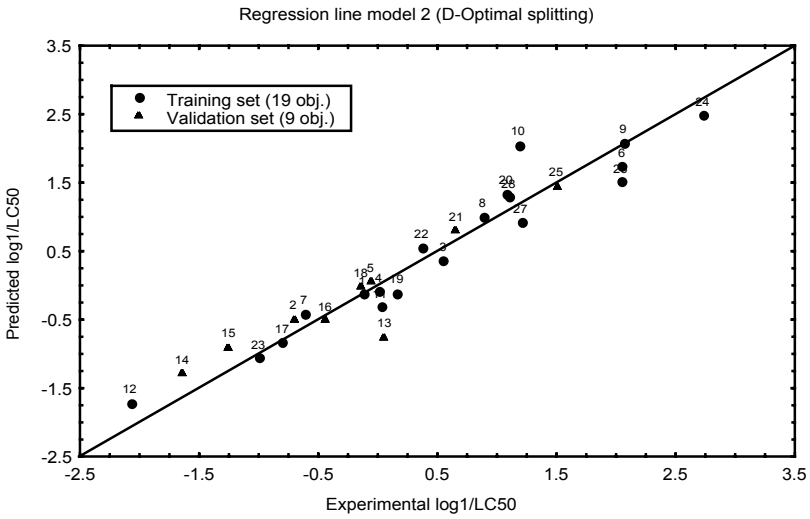
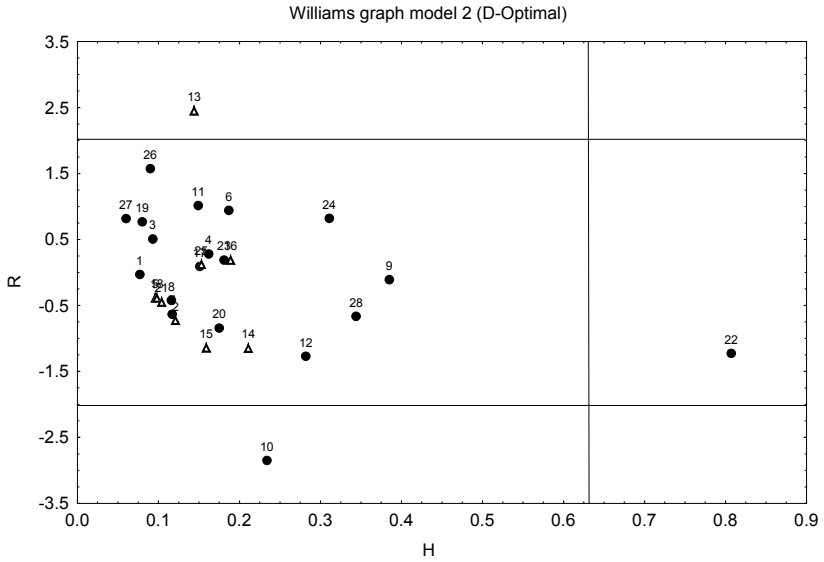


FIGURE 47: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





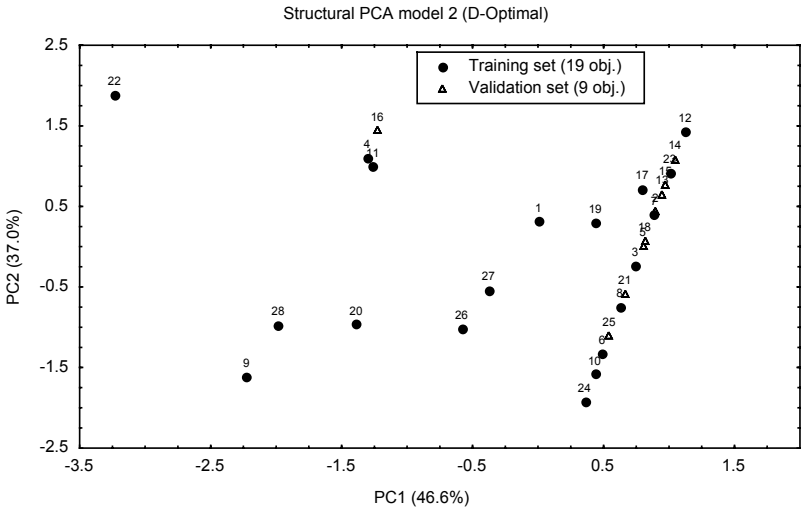
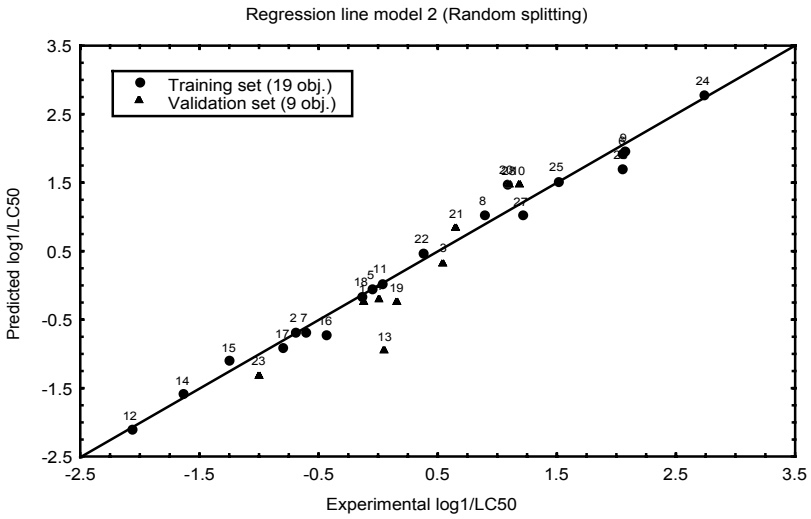
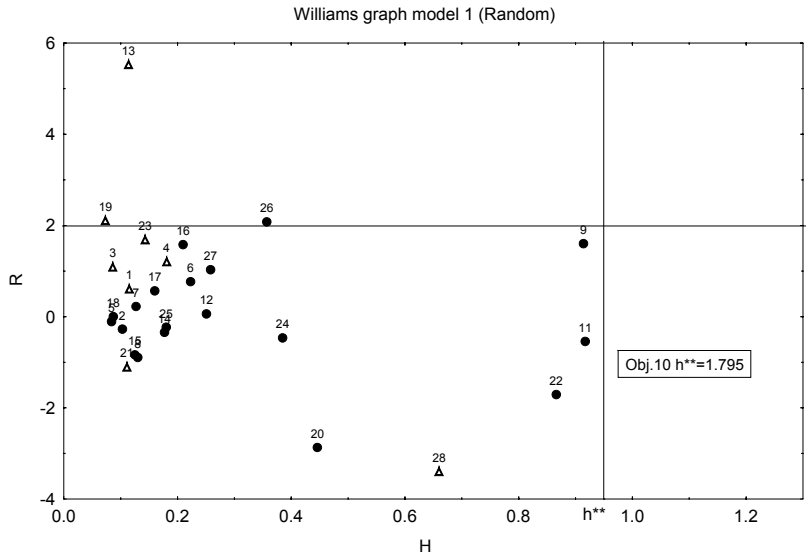


FIGURE 48: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors.





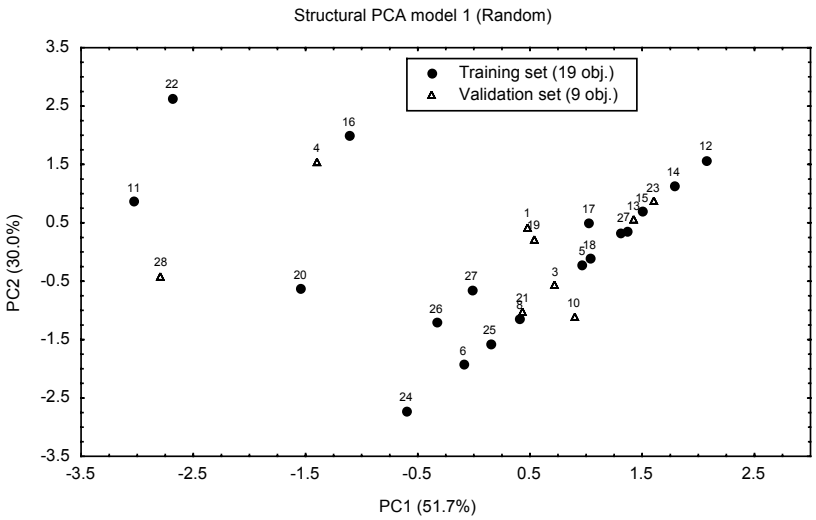
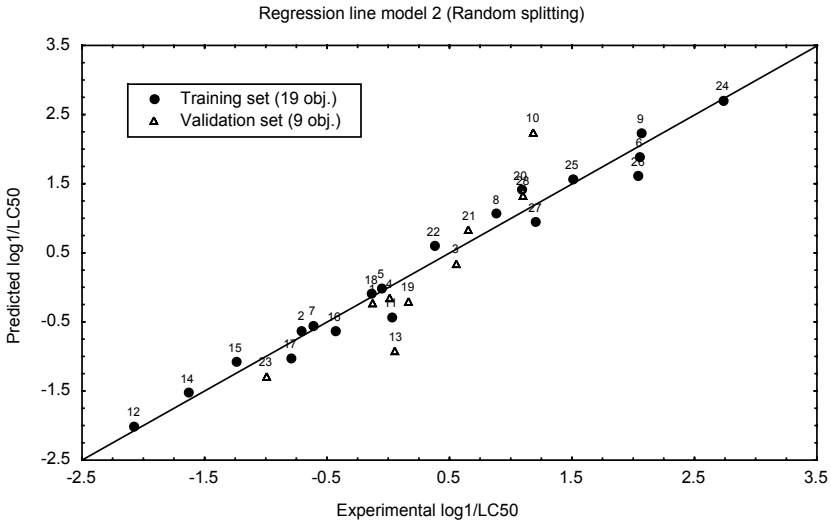
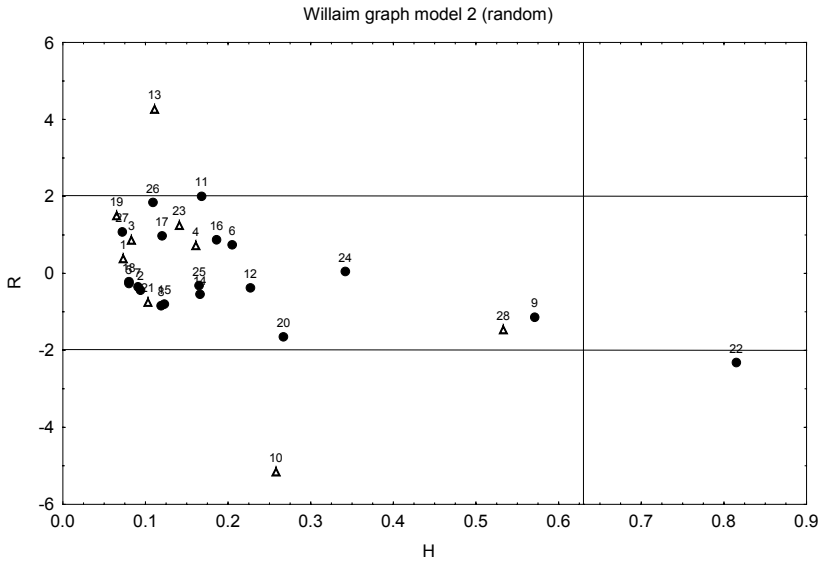
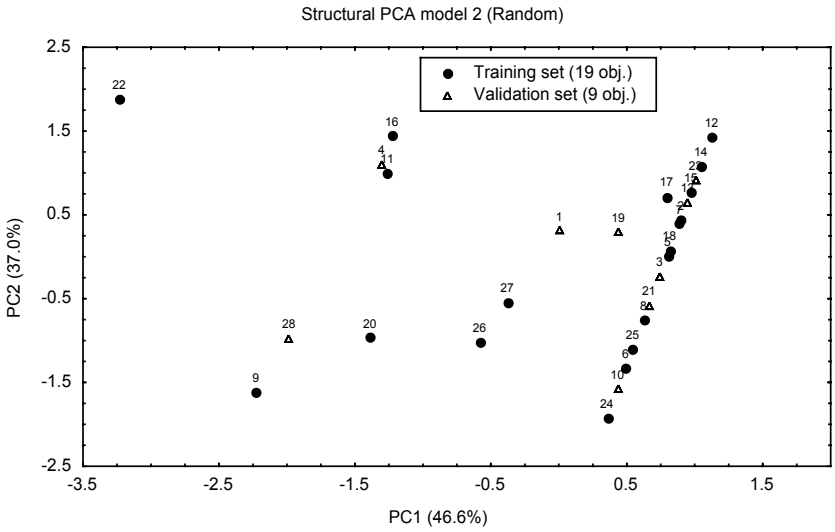


FIGURE 49: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors







PHENOLS:

The data set is composed of 42 phenols benzenes, but 5 are outliers and removed from the modelling.
The proposed models and the reported statistical parameters are:
(the last significant number of the regression coefficients, reported by the authors, is here put into brackets, in fact no more than three is the preferable number as this is representative of the accuracy of the original data)

(1) $\log(1/LC50) = 0.355(6)\ \alpha - 0.263(9)\ \beta + 0.406(8)\ \log Kow - 0.326(4)\ \pi^* + 0.200(6)\ {}^1X^v - 0.972(4)$

n= 37 R^2= 83.44% R^2 adj = 81.14% S.E.= 0.451

(2) $\log(1/LC50) = 1.621(5)\ \alpha + 0.710(3)\ \beta - 0.231\ \pi^* + 0.629\ {}^1X^v - 1.328(9)$

n= 37 R^2= 86.62% R^2adj = 84.95% S.E.= 0.3576

The statistical parameters reported by the authors are only related to fitting performance, that is good (but the statistical parameters of model (1) are not reproducible: recalculated parameters in Table 9)
The regression lines (not reported in the papers) and the corresponding Williams plot are reported in Figures 50 and 51. The authors did not point out that chemicals 2 and 15 are outliers. The Principal Component Analysis of the structural descriptors was also performed to highlight the distribution of the chemicals in the structural space of the model descriptors and any possible anomalous or isolated chemicals: also in this case, degeneracy of some variables and the correlation among the variables result in a perfect alignment of some chemicals.
The authors performed external validation by using 3 new phenols: in the paper they report only the predicted values, the corresponding Q^2ext was calculated by us and is reported in the following table.
The number of validation chemicals is too reduced to be significative for predictivity conclusions.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. Statistical external validation was also performed by comparing different approaches for the preliminary splitting of the chemicals into training and validation sets (5 chemicals) (D-optimal Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets regarding structural information are reported below. The outliers (2 and 15) are put in the training set in each following splitting.

Table 9: Statistical Diagnostics of models

n. Tr.	n valid	Split	Variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE train	MSE valid	SDEP	SDEC	F	s	x	Kxy	ΔK
37	3	Tot.1	π* β α $^1X^v$ logKow	82.5	87.4	78.1	78.0	85.3	83.4	0.104		0.381	0.323	42.6	0.353	47.5	56.2	8.7
37	3	Tot.2	π* β α $^1X^v$	82.7	86.6	76.4	79.8	85.0	48.1	0.111		0.378	0.333	51.4	0.358	38.0	48.5	10.5
32	5	K-ANN 1	π* β α $^1X^v$ logKow	79.3	85.7	68.3	72.2	83.0	98.0	0.119	0.017	0.415	0.344	31.1	0.381	47.0	55.6	8.7
32	5	K-ANN 2	π* β α $^1X^v$	79.8	84.8	72.1	75.2	82.6	98.1	0.108	0.015	0.410	0.355	37.5	0.386	38.0	48.1	10.2
32	5	D-Opt. 1	π* β α $^1X^v$ logKow	83.5	88.7	76.2	80.5	86.5	57.7	0.105	0.116	0.388	0.322	40.3	0.357	47.7	56.6	8.9
32	5	D-Opt. 2	π* β α $^1X^v$	83.6	87.7	78.2	78.8	85.8	62.9	0.113	0.102	0.388	0.336	47.6	0.366	38.5	49.1	10.5
32	5	Rand. 1	π* β α $^1X^v$ logKow	80.6	86.9	67.2	72.3	84.4	87.2	0.110	0.102	0.404	0.332	34.3	0.368	50.6	58.7	8.2
32	5	Rand. 2	π* β α $^1X^v$	80.0	85.4	71.7	74.0	83.2	95.0	0.124	0.020	0.410	0.351	39.1	0.382	42.6	51.3	8.7

The models demonstrate a satisfactory stability in internal validation.

SDEP is similar to SDEC: the models have internal predictivity not too dissimilar to fitting power.

The models with 4 or 5 descriptors are of similar stability and internal predictivity. Statistical external validations confirm the satisfactory prediction ability for chemicals included in the chemical domain of the training sets. In fact, the MSE values for training and validation sets are similar, thus demonstrating that the models are able to predict the response for chemicals not used in the model development (validation set) in similar way as for chemicals used to find the relationship (training set). The differences in Q^2_{ext} values for the different splitting must be attributed to the response distribution.

Regarding collinearity: in general, the descriptors are very correlated (medium Kxx: 53) but, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy) and the correlation among the X (Kxx) is sufficiently high (medium delta: 9) compared with other QSAR models, and according to our experience.

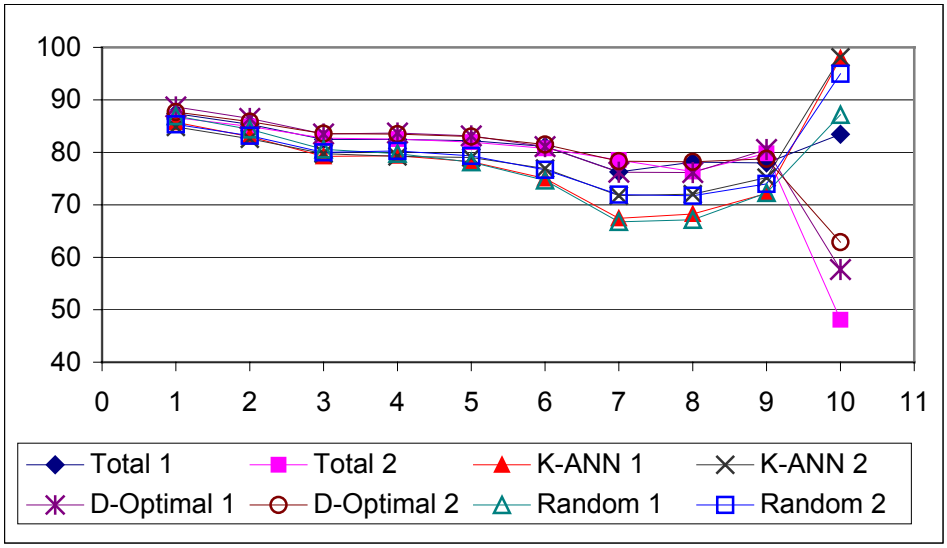
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted :

Table 9 bis: Statistical Diagnostics of models

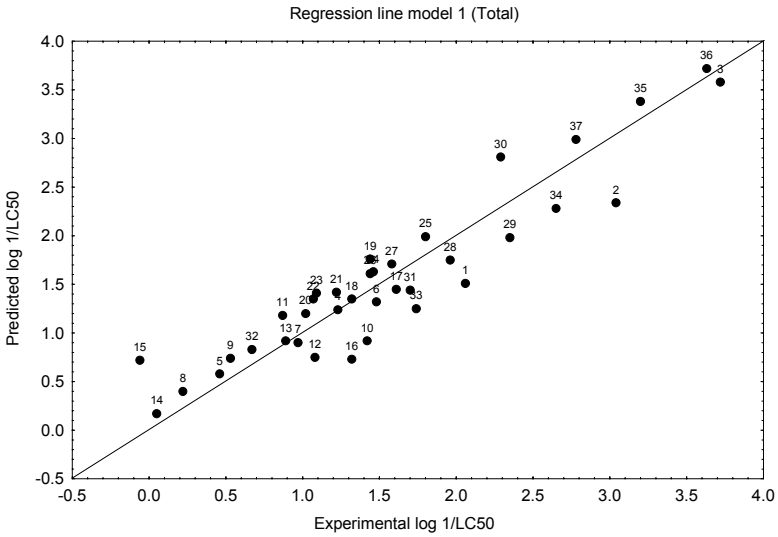
		Total 1	Total 2	K-ANN 1	K-ANN 2	D-Optimal 1	D-Optimal 2	Random 1	Random 2
1	R^2	87.4	86.6	85.7	84.8	88.7	87.7	86.9	85.4
2	R^2_{adj}	85.3	85.0	83.0	82.6	86.5	85.8	84.4	83.2
3	Q^2	82.5	82.7	79.3	79.8	83.5	83.6	80.6	80.0
4	Q^2_{LMO10}	82.5	82.5	79.4	79.2	83.7	83.5	79.8	80.3
5	Q^2_{LMO20}	82.2	82.0	78.2	79.0	83.1	83.0	78.2	79.3
6	Q^2_{LMO30}	81.1	80.8	75.0	76.9	81.1	81.5	74.7	76.7
7	Q^2_{LMO40}	76.3	78.6	67.4	71.8	76.2	78.3	66.8	72.0
8	Q^2_{LMO50}	78.1	76.4	68.3	72.1	76.2	78.2	67.2	71.7
9	Q^2_{boot}	78.0	79.8	72.2	75.2	80.5	78.8	72.3	74.0
10	Q^2ext	83.4	48.1	98.0	98.1	57.7	62.9	87.2	95.0

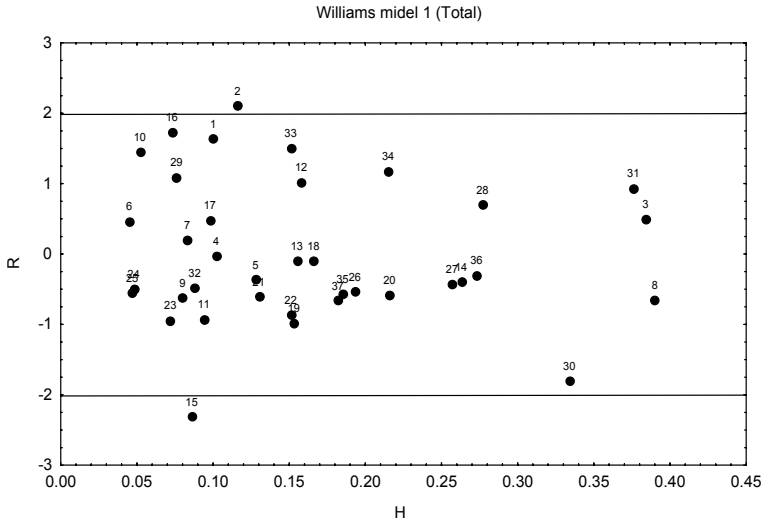
The following is the graphical representation of the parameters reported in the above table.



The models are stable and internally predictive. The statistical external validation, verified by Q^2_{EXT}, gives different results depending on the splitting methodology, this parameter being strongly influenced by response distribution. The comparison of MSE values (in Table) confirms the predictive ability of these models for the selected validation chemicals belonging to the chemical domain of the training.

FIGURE 50: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





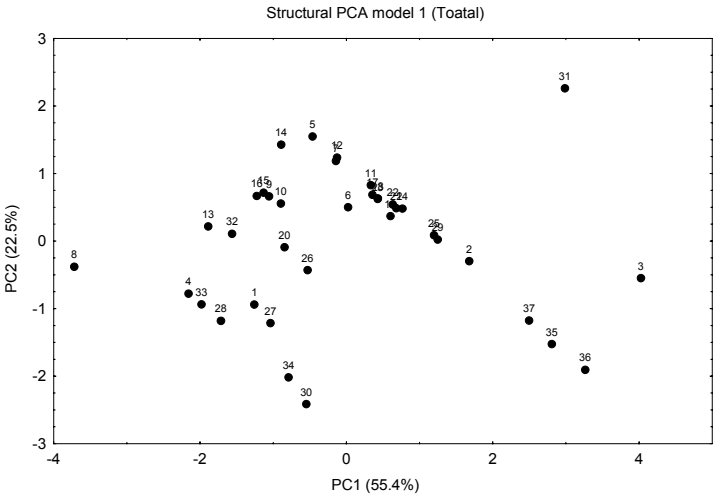
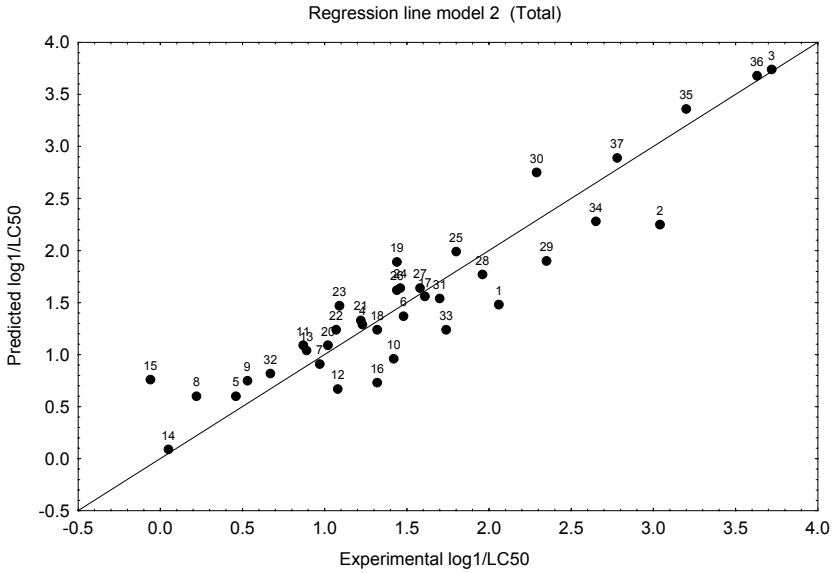
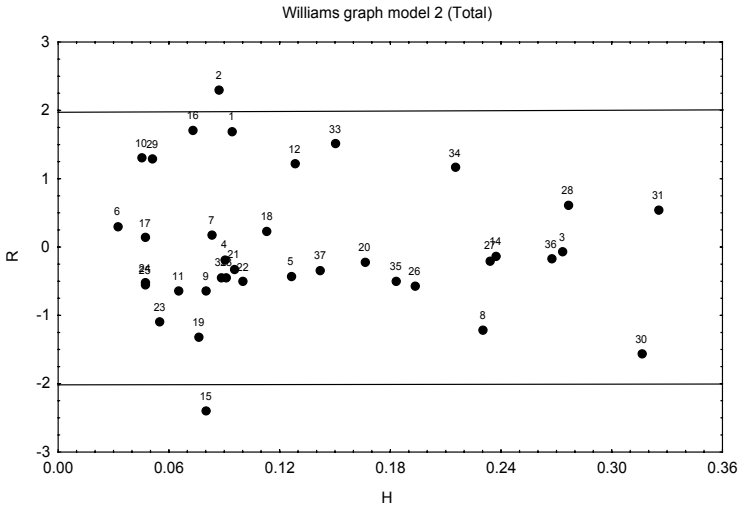


FIGURE 51: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





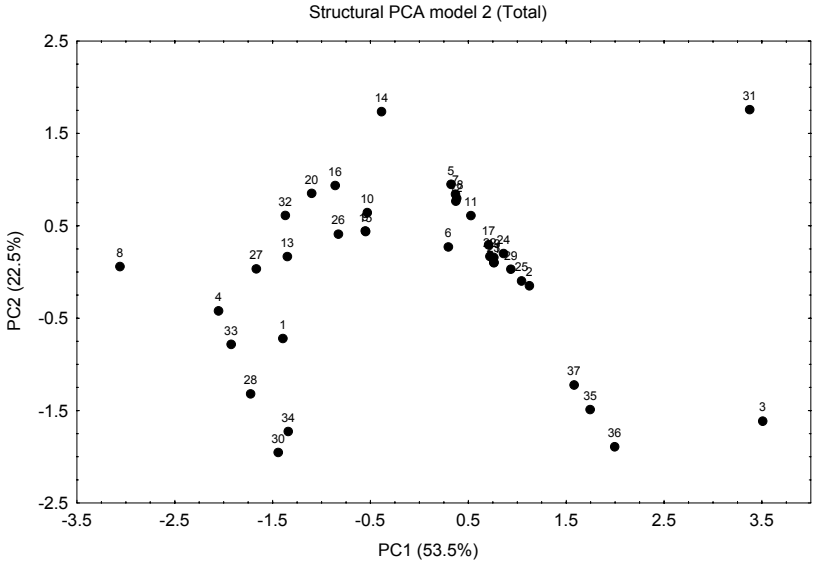
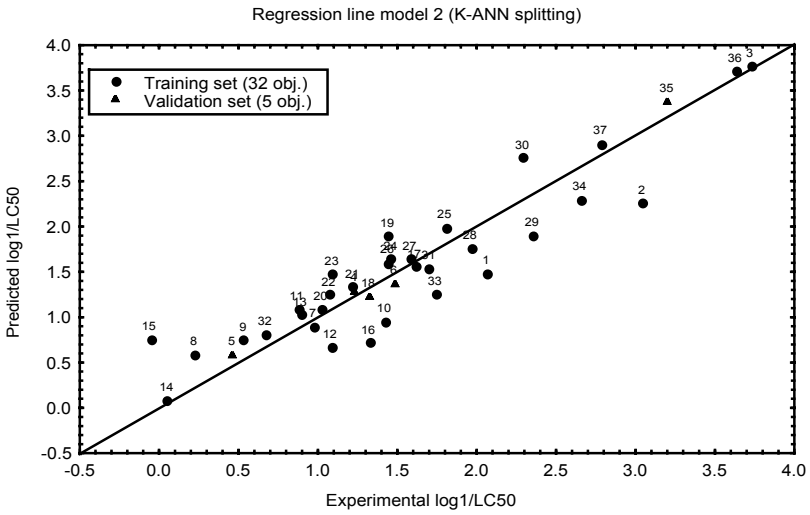
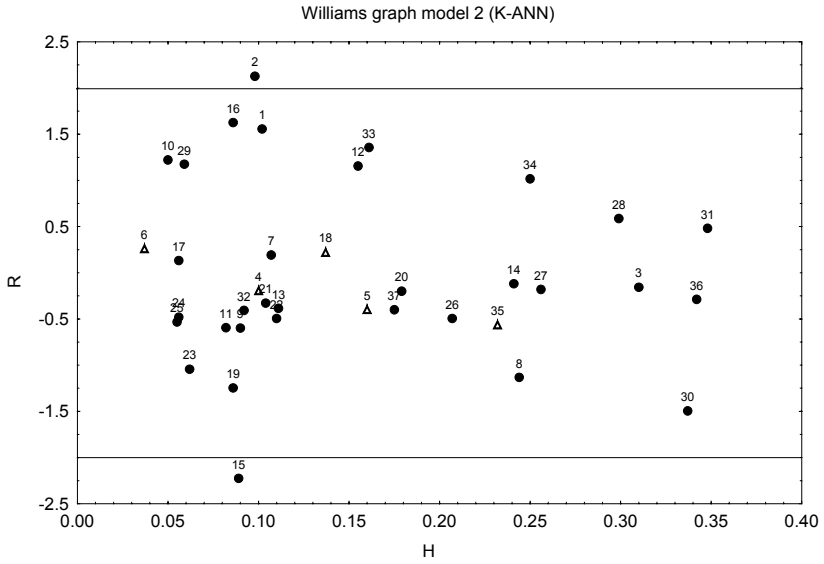


FIGURE 52: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





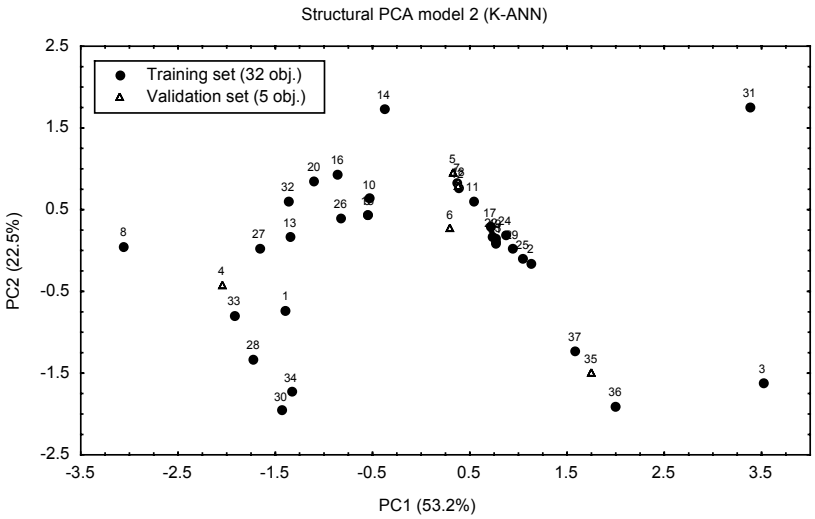
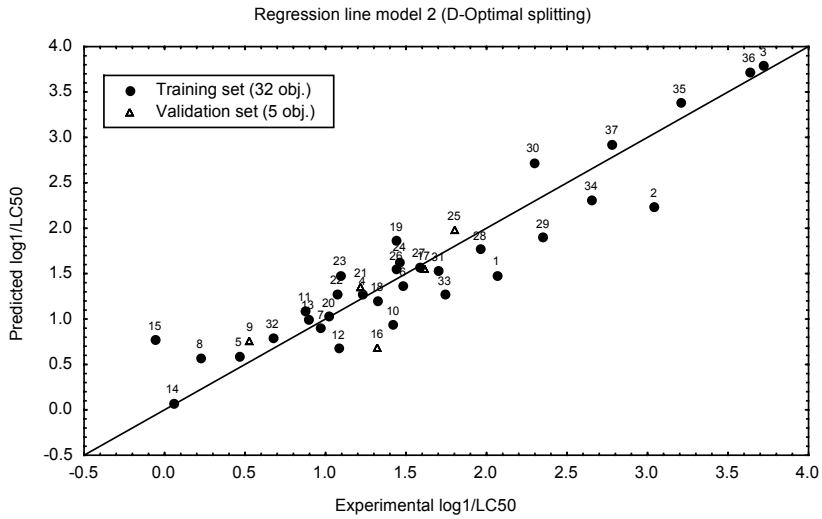
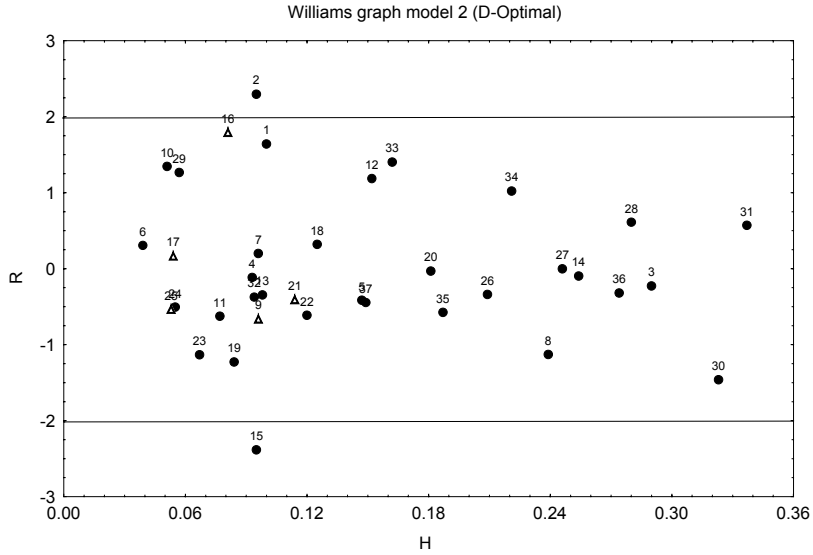


FIGURE 53: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





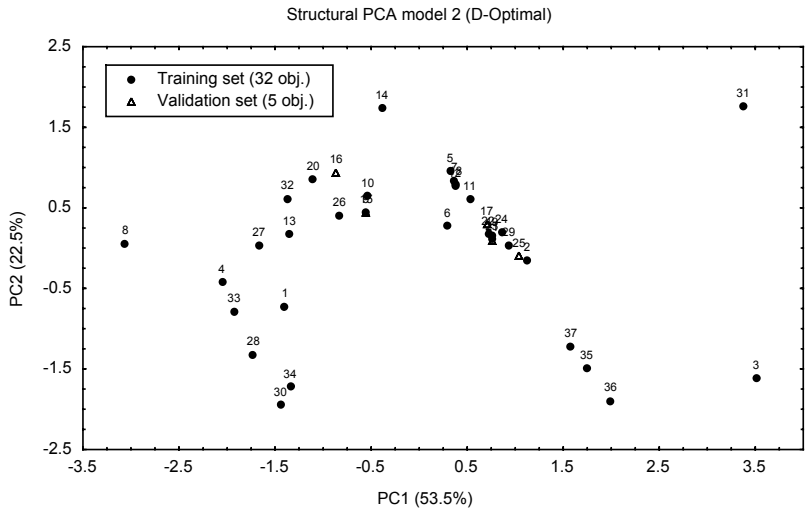
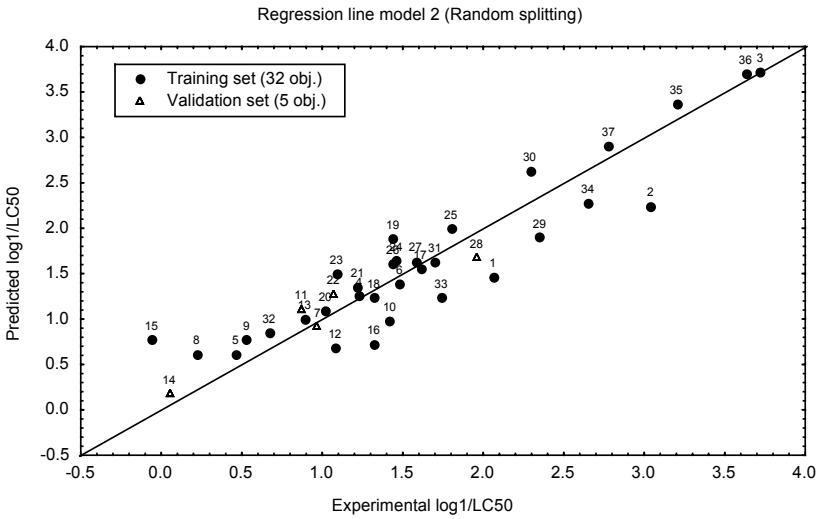
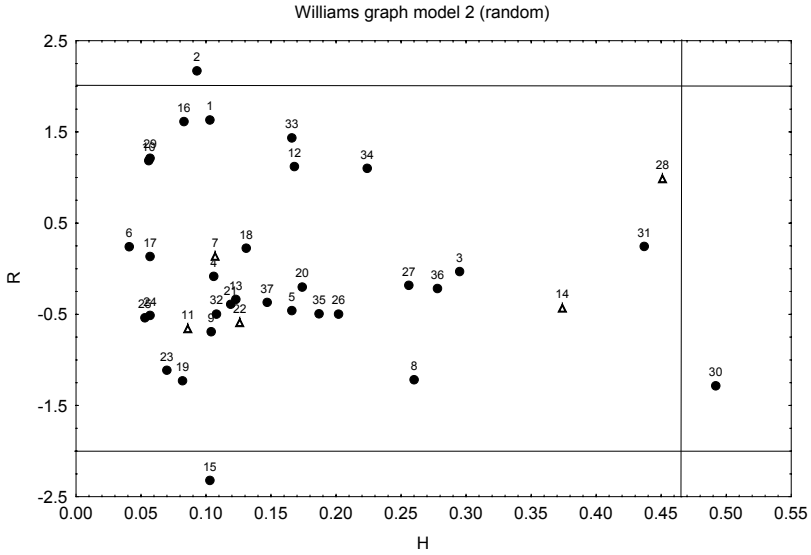
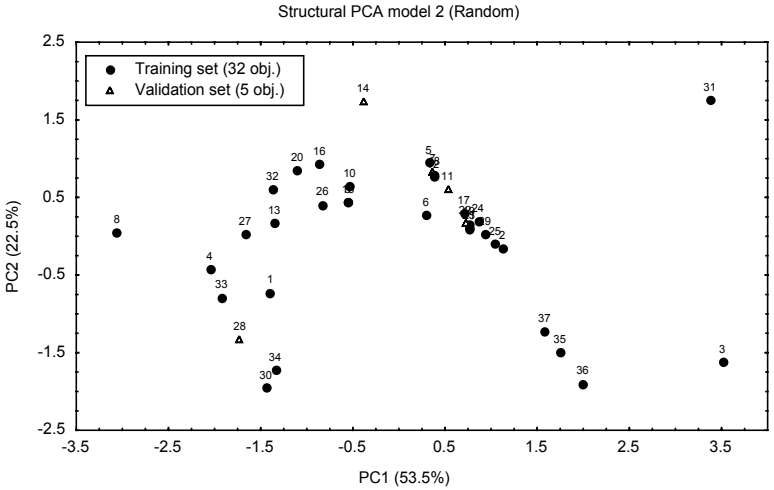


FIGURE 54: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors







MAIN CONCLUSIONS FOR VALIDATION

The models in this paper are often fitting models : their performances are excellent or good in fitting, but the fit is often too good compared with predictive performance, verified by different approaches, even only by LOO internal validation.

For each data set the authors propose two different models: they are nested models, and in each case it can be verified that the model with fewer descriptors is the more predictive. In fact the addition of additional descriptors, not useful for prediction, result in **overfitting** and not predictive models.

The best n/k ratio is 9 in all the studied data sets, the suggested ratio of >4 or >5 is too small.

The outliers and influential chemicals must always be highlighted in order to have the definition of the **chemical domain of applicability** of the models: these anomalous chemicals were not evidenced by the authors, we detected them by the Williams plot.

A comparison of the different approaches for **internal validation** allows the conclusions that:

LOO is too over-optimistic; this over-optimism can be counteracted by LMO but it is important that the perturbation is not too high so that enough information can be maintained in the training set (LMO 40-50% is too perturbative in a data set of 20-30 chemicals). Bootstrapping is the more balanced way of internal validation, and gives a more realistic idea of the real predictive ability of a model for chemicals used in model development.

Statistical external validation in the case of a small data set has not the significance it has in big data sets, but it is essential in order to verify if the model is able to predict response also for chemicals not used in model development. The results of this validation in small data sets is strongly influenced by the splitting methodology and, in addition to the Q^2_{EXT} parameter, MSE values on training and validation set must be checked.

It is interesting to note that one model of aliphatics, verified by the K-ANN splitting, which was internally predictive, appears completely unpredictive externally: in fact in the validation set two chemicals were included, which are out of the chemical domain of the training set. This is a strong demonstration that the distribution of chemicals between training and validation is strongly influent on the statistical external validation results. (see also Appendix of the Leverage approach).

The **distribution of the chemicals in the space of structural descriptors of the model** must always be verified: in the PCA of model descriptors some anomalous distribution of the chemicals (as in some of the studied models) could be indicative of an inadequate selection of modelling descriptors (degeneration, correlation). The **correlation among variables** must always be checked and verified, making a comparison with response correlation (QUIK rule).

Shortcomings of the studied models:

1. strong outliers and influential chemicals, not evidenced
2. overfitting
3. too high correlation among the descriptors
4. No, or less than apparent, predictivity

7. STATISTICAL VALIDATION OF TOXICITY MODELS (Cronin et al.)

Cronin, M.T.D., Dearden, J.C., Duffy, J.C., Edwards, R. Manga, N., Worth, A.P. and Worgan, A.D.P. (2002). The importance of hydrophobicity and electrophilicity descriptors in mechanistically-based QSARs for toxicological endpoints. ***SAR and QSAR in Environmental Research*** **13 (1), 167-176.**

Four data sets are studied in this paper.

1) The toxic and metabolic effects of 23 aliphatic alcohols (Table in Annex) on GPT (glutamate-pyruvate-transaminase), LDH (lactate-dehydrogenase), GLDH (glutamate-dehydrogenase) and the reduction of ATP concentration in rat liver are studied. The OLS regression coefficients of all the published models have been verified and confirmed, but the statistical parameters are slightly different (see Table).

Some outliers were removed from the data set by the authors, without any comment, before the modelling. The equations with (in the paper is written: without, probably due to a typing error) the removed outliers are not reported in the paper.

Our analysis of the OLS outlier and leverage plots (Williams plots) performed in this contract work allows the identification, in each model, of additional outliers not identified in the paper, whereas there were no influential chemicals (of high leverage value, H).

The published model for GPT is:

$$\text{Log GPT} = 0.576\ \text{logP} - 0.193\ E_{LUMO} - 0.494\ {}^{3}X_{PC} + 12.19$$

$$N=23 \quad s=0.183 \quad r^2=0.836 \quad r^2cv=0.801 \quad F=38.5$$

The regression line and the Williams plot are reported in Fig.55 : 2 outliers (20 and 22), not evidenced by the authors, are present.

The published model for LDH is:

$$\text{Log LDH} = 0.561\ \text{logP} - 0.297\ E_{LUMO} - 0.487\ {}^{3}X_{PC} + 1.57$$

$$N=22 \quad s=0.184 \quad r^2=0.848 \quad r^2cv=0.813 \quad F=40.1$$

1 outlier (3-methyl2-buten-1-ol, 22) was removed by the authors before the modelling.

The regression line and the Williams plot are reported in Fig. 56: 1 outlier (19), not evidenced by the authors, is again present.

The published model for GLDH is:

$$\text{Log GLDH} = 0.399\ \text{logP} - 0.037\ E_{LUMO} - 0.384\ {}^{3}X_{PC} + 0.579$$

$$N=19 \quad s=0.132 \quad r^2=0.846 \quad r^2cv=0.797 \quad F=34.1$$

4 outliers (2-methyl-1-butanol (11), 2-propyn-1-ol (18), 1-buten-3-ol (20) and 2-methyl-2-propen-1-ol (21)) have been removed by the authors before the modelling. The regression line and the Williams plot are reported in Fig. 57: no outliers are present.

The published model for ATP is:

$$\text{Log } 1/\text{ATP} = 0.393\ \text{logP} - 0.362\ E_{LUMO} - 0.263\ {}^{3}X_{PC} + 1.48$$

$$N=20 \quad s=0.159 \quad r^2=0.857 \quad r^2cv=0.789 \quad F=38.9$$

3 outliers (1- butanol (4), 2-methyl-1-butanol (11), 3-methyl-2-buten-1-ol (22)) have been removed by the authors before the modelling. The regression line and the Williams plot are reported in Fig. 58: 1 outlier (23), not evidenced by the authors, is again present.

VALIDATION:

The models have been assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap.

The limited size of the data set does not allow reasonable statistical external validation by preliminary splitting: in this case a splitting into training and validation sets would reduce the available information in the training chemicals too much, and the results of QSAR models on validation chemicals could be unreliable and even wrong.

Table 10: Statistical Diagnostics of models

	ntr	Split	variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	MSE	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
GPT	23	TOT	logP E_{Lumo} ^{3}Xpc	80.1	85.9	69.9	75.4	83.7	0.028	0.197	0.166	38.2	0.183	17.25	35.62	18.37
LDH	22	TOT	logP E_{Lumo} ^{3}Xpc	81.3	87.0	71.4	76.0	84.8	0.028	0.199	0.167	39.7	0.184	19.76	35.92	16.16
GLDH	19	TOT	logP E_{Lumo} ^{3}Xpc	79.6	87.2	60.6	64.6	84.6	0.014	0.148	0.117	33.7	0.132	6.91	32.69	25.78
1/ATP	20	TOT	logP E_{Lumo} ^{3}Xpc	79.4	88.2	69.6	74.2	86.0	0.019	0.185	0.140	39.6	0.157	20.30	34.31	14.01

Regarding collinearity: in general, the descriptors are not very correlated (Kxx) and, most importantly, the difference in correlation between the block of X variables plus response Y (Kxy) and the correlation among the X (Kxx) is quite high (delta column) compared with other QSAR models, and according to our experience.

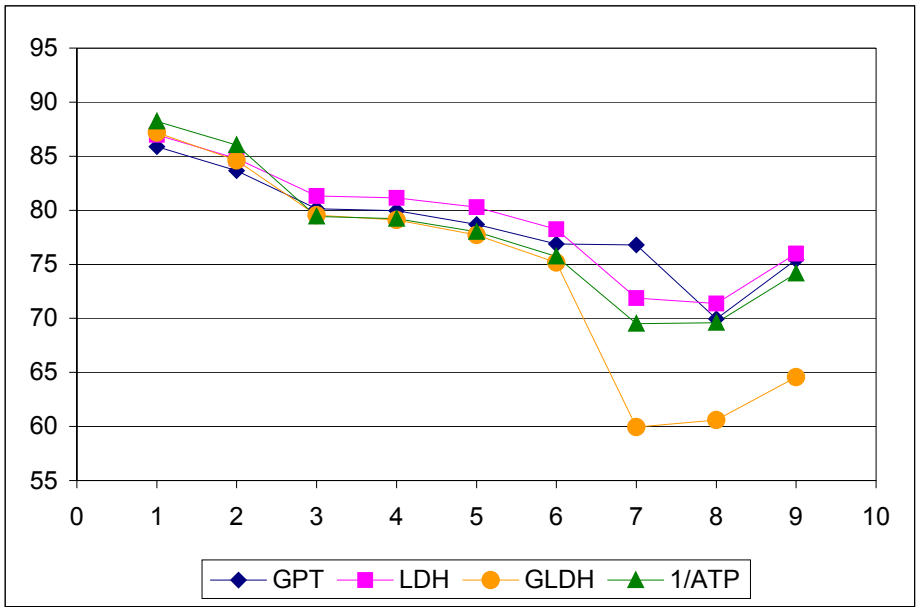
All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table, and a graph was plotted.

Table 10 bis: Statistical Diagnostics of models

		GPT	LDH	GLDH	1/ATP
1	R^2	85.9	87.0	87.2	88.2
2	R^2_{adj}	83.7	84.8	84.6	86.0
3	Q^2	80.1	81.3	79.6	79.4
4	Q^2_{LMO10}	80.0	81.2	79.1	79.2
5	Q^2_{LMO20}	78.7	80.3	77.7	78.0
6	Q^2_{LMO30}	76.9	78.2	75.2	75.7
7	Q^2_{LMO40}	76.8	71.9	60.0	69.5
8	Q^2_{LMO50}	69.9	71.4	60.6	69.6
9	Q^2_{boot}	75.4	76.0	64.6	74.2

The following is the graphical representation of the parameters reported in the above table.



Comparing the statistical parameters allows the following conclusions: the fitting parameters R^2 (1) and R^2_{adj} (2) give similar results and, obviously, R^2_{adj} is always a lower value than R^2. Internal validation by LOO and LMO gives results that decrease progressively (with regularity), and that are substantially similar up to the 30% of perturbation ($Q^2_{LMO30,}$ 6). When the perturbation is too high in small data sets (40-50% means 8-10 chemicals out of 20 put in the test set) then, in each run of perturbation, the chemicals on which the model is developed are not sufficiently informative with regard to the structural information useful for test chemicals. Thus, the models appear to have lower predictive performance (10-19% less than LOO). In general, validation by bootstrapping (9) gives intermediate results. As can be expected, the smallest data set on GLDH(19 chemicals) is the most sensitive to LMO and bootstrap validation.

By strong internal validations the published models reveal a smaller (5-19% less) internal predictivity than expected from the published R^2_{cv} values.

FIGURE 55: Regression line and residuals/leverage diagnostic (Williams graph) of model

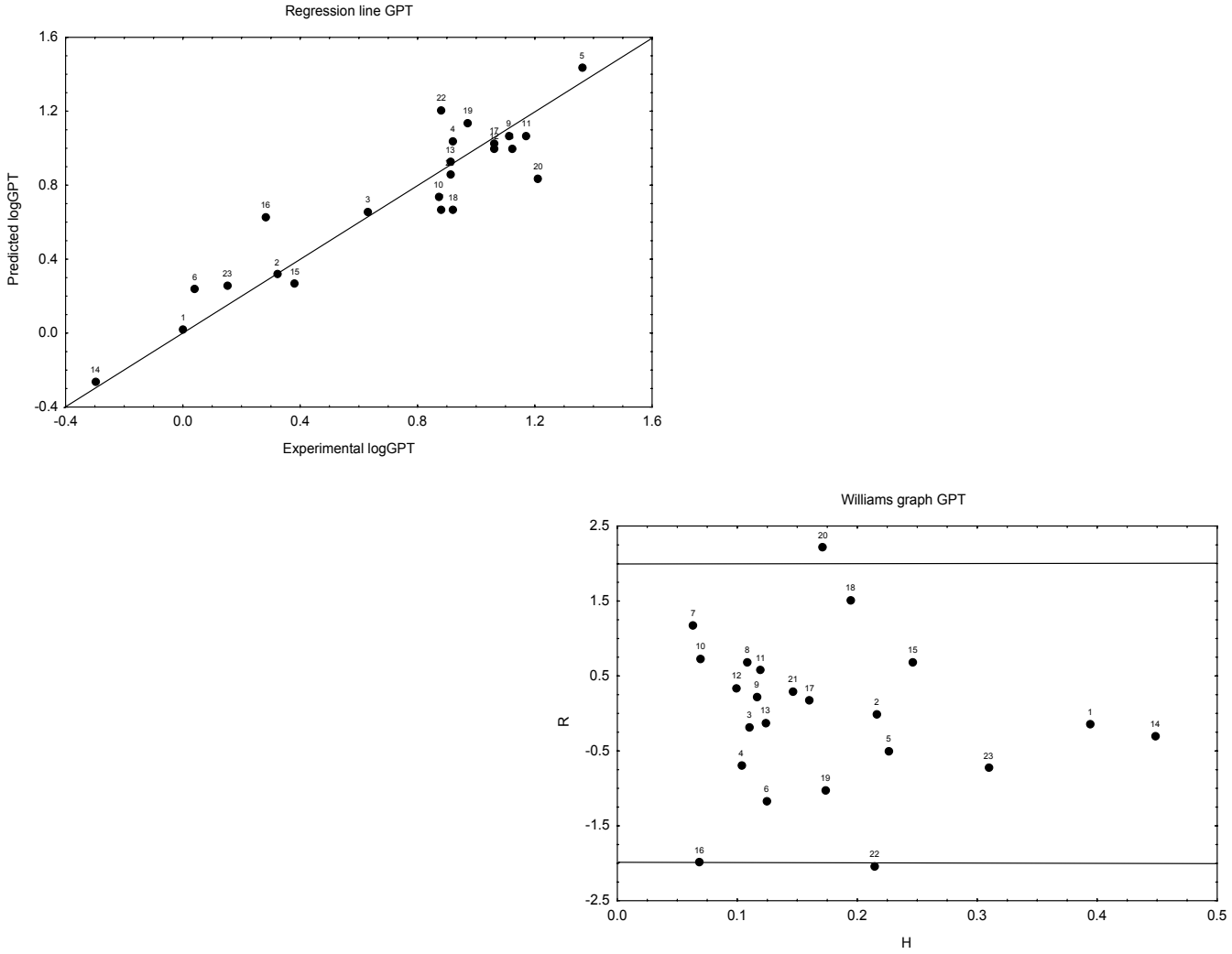


FIGURE 56: Regression line and residuals/leverage diagnostic (Williams graph) of model

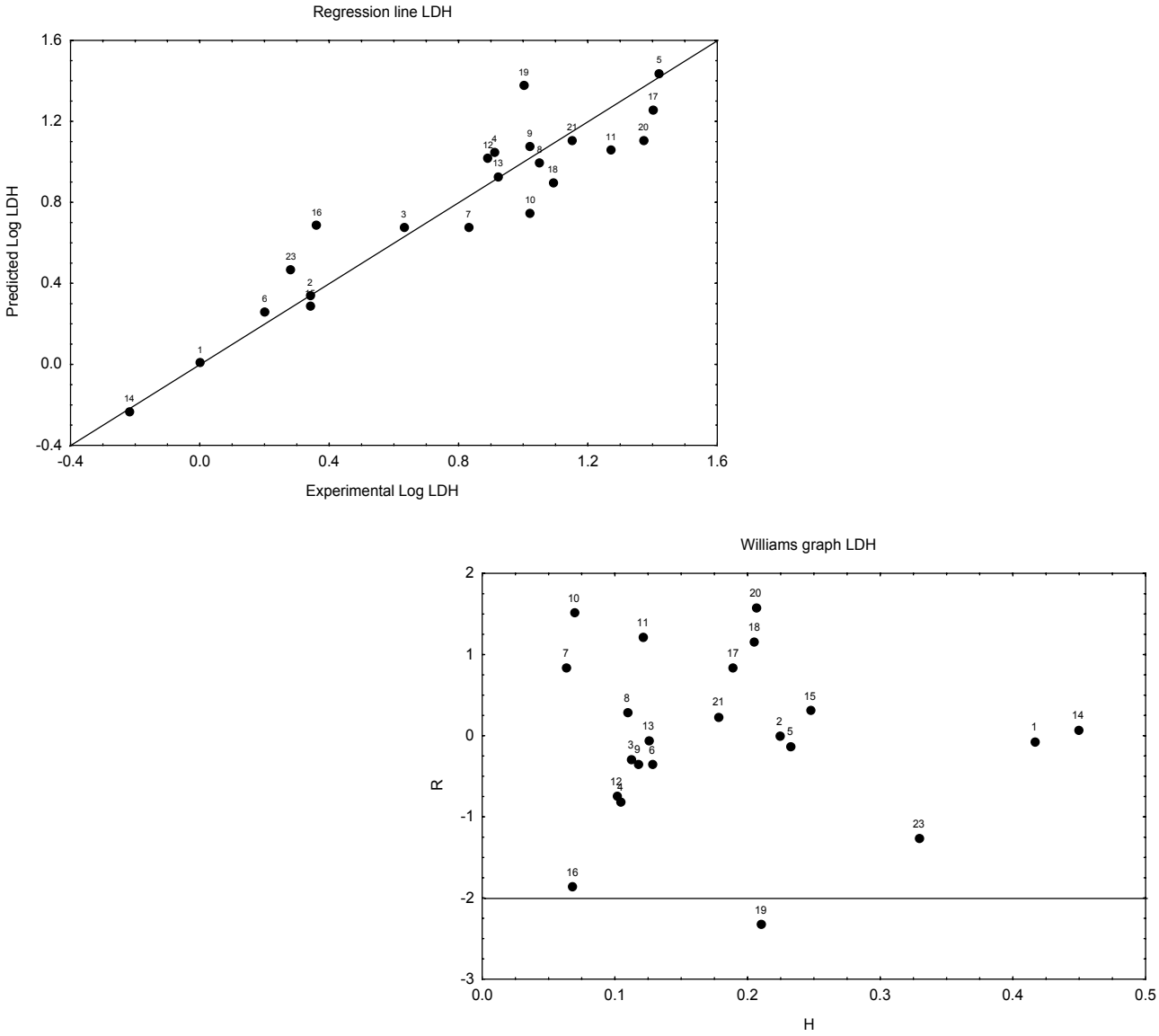
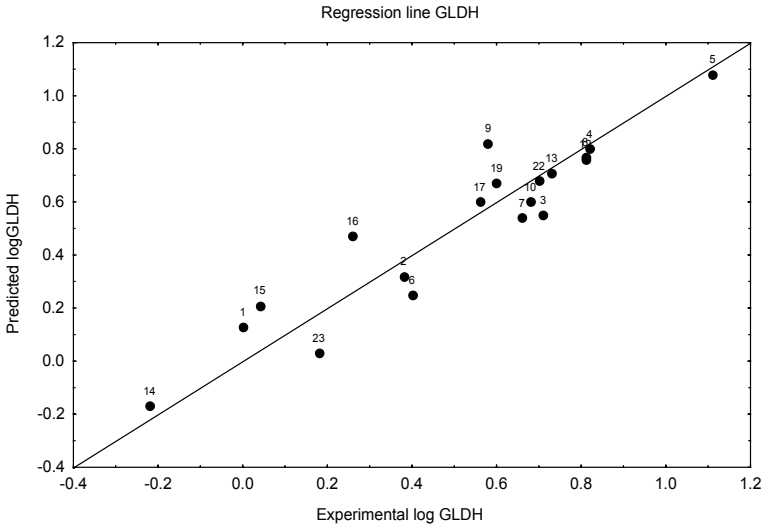


FIGURE 57: Regression line and residuals/leverage diagnostic (Williams graph) of model



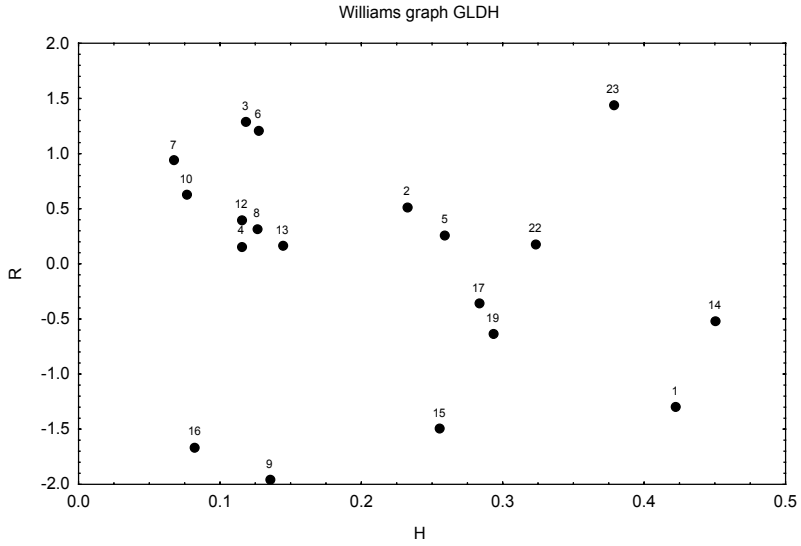
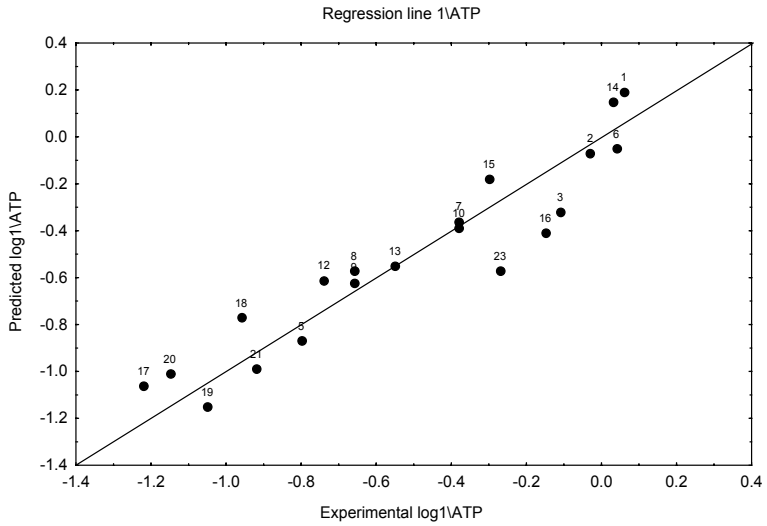
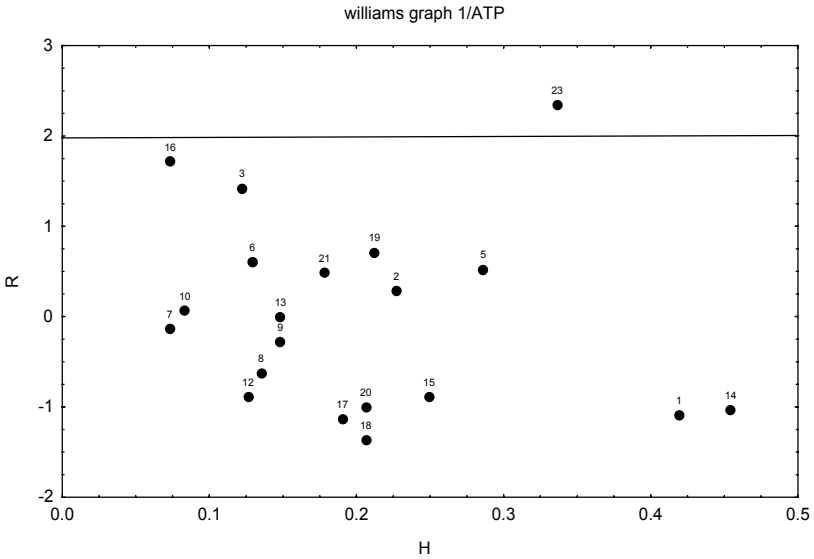


FIGURE 58: Regression line and residuals/leverage diagnostic (Williams graph) of model





2) The toxicity (LD_{50}) of 21 pyridines and by-pyridines (Table in Annex) to mice has been studied.

The published Ordinary Least Squares (OLS) model is:

$$\text{Log } 1/LD_{50} = 0.380 \text{ logP} - 0.660 \, E_{LUMO} + 1.81$$

$$N=21 \quad s=0.224 \quad r^2= 0.808 \quad r^2cv=0.766 \quad F= 43.1$$

1 outlier (3-hydroxy-N-oxide-pyridine, 16) was identified by the authors and removed, the new published model is:

$$\text{Log } 1/LD_{50} = 0.359 \text{ logP} - 0.682 \, E_{LUMO} + 1.85$$

$$N=20 \quad s=0.188 \quad r^2= 0.848 \quad r^2cv=0.819 \quad F= 54.1$$

The regression line and the Williams plots of these models, obtained in this work, are reported in Fig. 59 and 60.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap.

As before, the limited size of the data set does not allow reasonable statistical external validation by preliminary splitting.

Table 11: Statistical Diagnostics of models

	ntr	variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	MSE	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
$1/LD_{50}$	21	logP E_{LUMO}	76.5	82.7	71.7	72.7	80.8	0.043	0.242	0.208	42.74	0.224	28.46	46.03	17.57
$1/LD_{50}$	20	logP E_{LUMO}	80.9	86.3	76.3	78.0	84.7	0.030	0.205	0.174	53.32	0.188	26.65	46.90	20.25

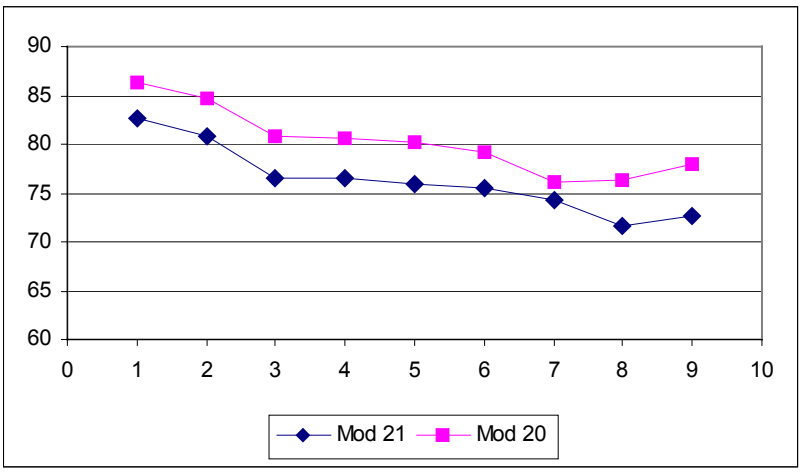
Regarding collinearity: the descriptors are quite correlated (Kxx) but, most importantly, the difference in the correlation between the block of X variables plus the response Y (Kxy) and the correlation among the X (Kxx) is quite high (delta column) compared with other QSAR models and according to our experience.

All the models were also verified by Y-scrambling: compared with the published models, the models on randomised response all have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

Table 11 bis: Statistical Diagnostics of models

		Mod 21	Mod 20
1	R^2	82.7	86.3
2	R^2_{adj}	80.8	84.7
3	Q^2	76.5	80.9
4	Q^2_{LMO10}	76.5	80.6
5	Q^2_{LMO20}	75.9	80.2
6	Q^2_{LMO30}	75.5	79.1
7	Q^2_{LMO40}	74.4	76.1
8	Q^2_{LMO50}	71.7	76.3
9	Q^2_{boot}	72.7	78.0

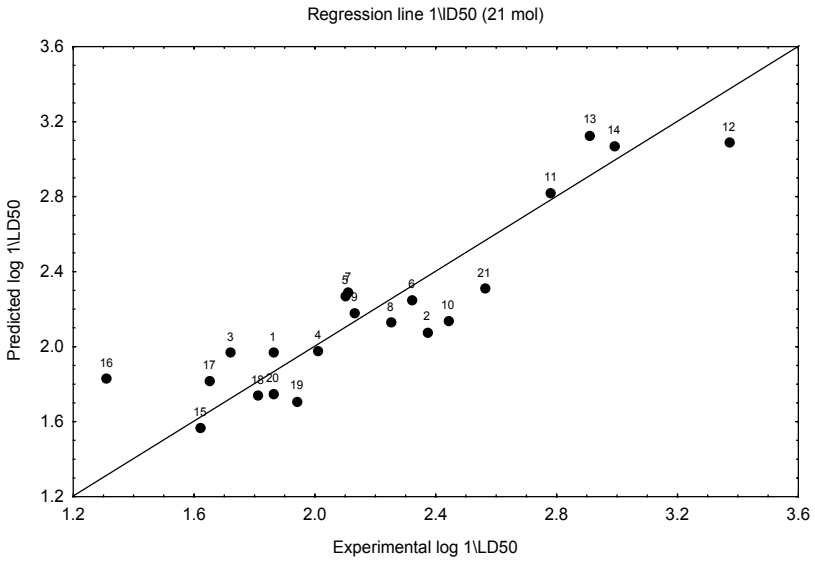
The following is the graphical representation of the parameters reported in the above table.



The size of this data set is similar to the previous one (20 chemicals), but in this case all the validation approaches applied give very similar information regarding the internal predictivity of the model. The reason is probably related to the greater structural homogeneity of this data set compared with the previous one: the information given here is more homogeneous in all the subset training/test in each run of LMO perturbation. The model without the outlier is obviously the best, and its performance is also higher in strong validations. The poorest performances are again verified by LMO 50% and bootstrapping.

In this case, the difference between Q^2_{LOO} and Q^2_{LMO50} is less than 5%: this highlights the stability and internal predictivity of the models, even though obtained on a small data set.

FIGURE 59: Regression line and residuals/leverage diagnostic (Williams graph) of model



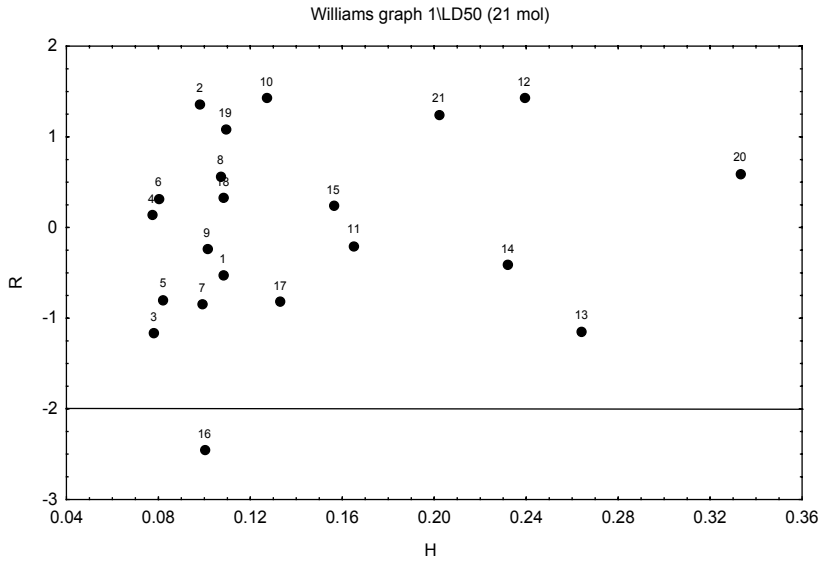
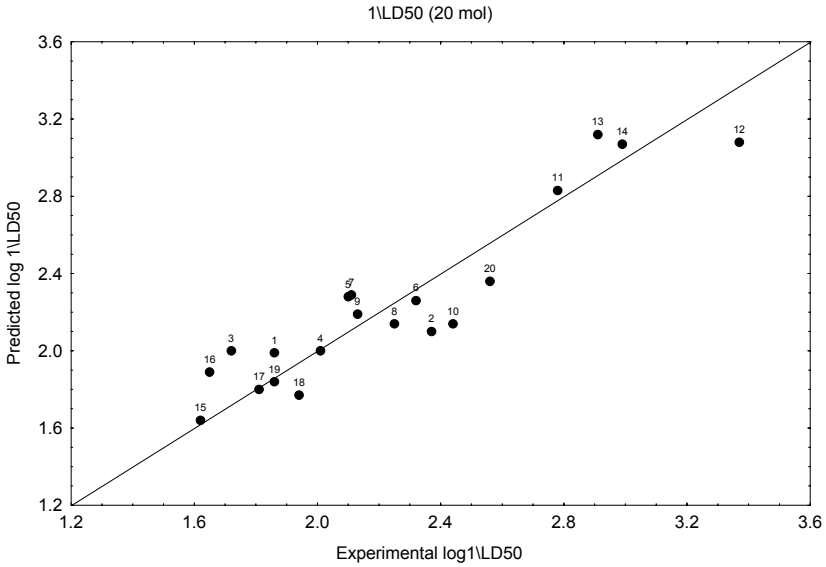
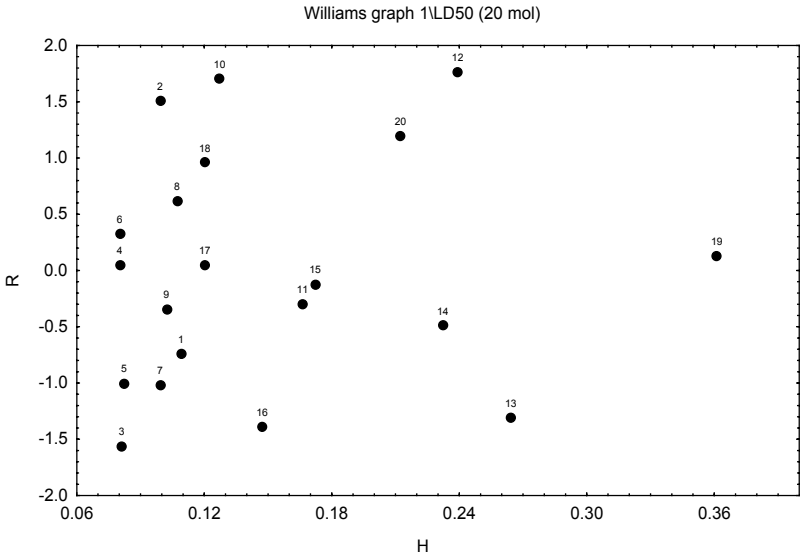


FIGURE 60: Regression line and residuals/leverage diagnostic (Williams graph) of model





3) The lethality effect ($1/D_{37}$) of 55 halogenated aliphatic hydrocarbons (Table in Annex) to A.nidulans was studied.

The published Ordinary Least Squares model is:

$$\text{Log } 1/D_{37} = 0.598 \text{ logP} - 0.331 \, E_{LUMO} - 2.17$$

$$N=55 \quad s=0.413 \quad r^2= 0.615 \quad r^2cv=0.567 \quad F= 44.2$$

The model does not have satisfactory performance, thus 6 outliers (2,3-Dichloropropene (30),tetrabromomethane (31), 1,1,2,2-tetrabromoethane (33), dichlorodibromomethane (51), 1,3-Dichloropropene(53), 3-chloro-2-chloromethyl-1-propene (54)) were removed by the authors (without any comments) and a new model was developed:

$$\text{Log } 1/D_{37} = 0.557 \text{ logP} - 0.202 \, E_{LUMO} - 2.21$$

$$N=49 \quad s=0.254 \quad r^2= 0.739 \quad r^2cv=0.703 \quad F= 65.1$$

The regression line and the Williams plot of this model, obtained in this contract work, are reported in Fig. 61: 11 and 28 are influential chemicals in this model (not highlighted in the paper). The need for applicability domain inspection is evident: the two highly influential chemicals are more distant from the centre of the model, they show some peculiarity in their structure and derived descriptor values, so their presence strongly influences the modelling. The structural descriptors also underwent Principal Component Analysis to highlight the distribution of the chemicals in the structural space of the model descriptors and any eventual anomalous or isolated chemicals (11 and 28 are actually more isolated from the other chemicals in the set.)

VALIDATION:

The second model was assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap.

On this data set, in addition to different internal validation approaches, the statistical external validation was performed by comparing different approaches for the preliminary splitting of chemicals into training (33 chemicals) and validation sets (16 chemicals) (D-optimal Distance, Kohonen-ANN; random). The PCA of structural descriptors to verify the distribution of the two sets regarding structural information are reported below.

Table 12: Statistical Diagnostics of models

ntr	nvalid	split	variables	Q^2	R^2	$Q^2_{LMO}50$	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
49		TOT	logP E_{LUMO}	70.2	73.9	68.3	68.8	72.7				0.263	0.247	64.750	0.254	5.16	44.82	39.66
33	16	D-opt	logP E_{LUMO}	72.8	78.3	70.1	72.8	76.82	59.1	0.063	0.063	0.280	0.250	53.770	0.263	7.86	46.96	39.10
33	16	K-ANN	logP E_{LUMO}	69.7	76.1	70.1	67.0	74.5	65.9	0.062	0.060	0.281	0.250	47.440	0.262	5.88	43.73	37.85
33	16	random	logP E_{LUMO}	73.3	78.4	70.7	71.0	76.9	51.3	0.062	0.060	0.277	0.249	54.150	0.261	4.42	45.86	41.44
33	16	random	logP E_{LUMO}	67.7	72.9	64.2	64.7	71.1	75.4	0.062	0.059	0.272	0.250	40.150	0.262	2.98	42.75	39.77

The models show satisfactory fitting performance (R^2 and R^2adj) but the validation parameters need commenting on. For the sake of clarity the MSE value was also considered. The MSE values for training and validation set are similar, thus demonstrating that the models are able to predict the response of chemicals not used in the model development (validation set) as they do for chemicals used to find the relationship (training set). Q^2_{EXT} gives anomalous results (see below).

SDEP is very similar to SDEC: the model has internal predictivity similar to fitting power, as highlighted by other internal validation parameters

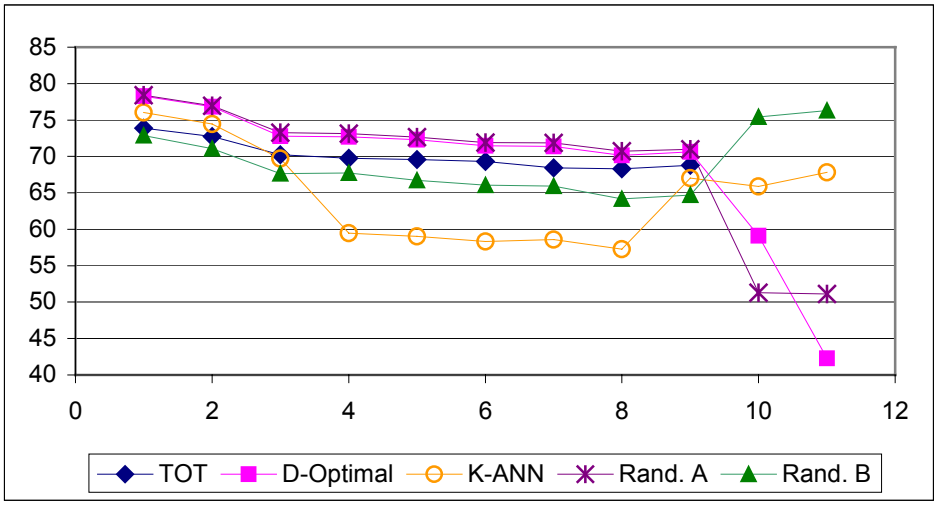
Regarding collinearity: the descriptors show little correlation (Kxx) and, most importantly, the difference in the correlation between the block of X variables plus the response Y (Kxy) and that of X (Kxx) is very high (delta column) compared with other QSAR models and according to our experience.

All the models were also verified by Y-scrambling: the models on randomised response all have low R^2 and Q^2 compared with the published models. This is a demonstration that the reported models are not obtained by chance correlation.

Table 12 bis: Statistical Diagnostics of models

	Split	TOT	D-Optimal	K-ANN	Rand. A	Rand. B
1	R^2	73.9	78.3	76.1	78.4	72.9
2	R^2_{adj}	72.7	76.8	74.5	76.9	71.1
3	Q^2	70.2	72.8	69.7	73.3	67.7
4	Q^2_{LMO10}	69.7	72.7	59.5	73.1	67.7
5	Q^2_{LMO20}	69.6	72.3	59.0	72.6	66.7
6	Q^2_{LMO30}	69.3	71.5	58.3	71.9	66.1
7	Q^2_{LMO40}	68.4	71.4	58.6	71.9	65.9
8	Q^2_{LMO50}	68.3	70.1	57.3	70.7	64.2
9	Q^2_{boot}	68.8	70.6	67.0	71.0	64.7
10	Q^2ext		59.1	65.9	51.3	75.4

The following is the graphical representation of the parameters reported in the above table.



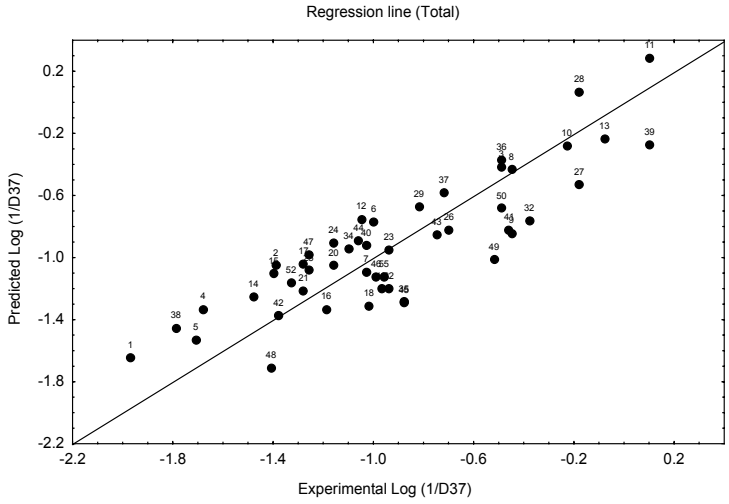
The models are substantially robust: the difference between R^2 and Q^2 is not too high (around 5%) for every splitting. The proposed models demonstrate progressively decreasing performance in internal validation, as normally happens, again highlighting that LOO cross-validation is over-optimistic. The internal validation in K-ANN splitting gives a probable under-estimation of the true predictive power. The bootstrapping approach gives similar results in all the models and confirms that it must be preferred as the internal validation parameter, data set composition influencing it the least. Statistical external validations verified by Q^2ext give extremely variable results: the distribution of chemicals between the training and validation sets is highly influential in Q^2ext calculation.

The checking of the MSE values confirms that the models are equally predictive for internal and external chemicals, while the Q^2ext value is under-optimistic or over-optimistic depending on the splitting methodology applied. This calls for a comment on the anomalous response distribution in D-optimal splitting (all the validation chemicals with lower response value); this splitting worked well in relation to structure, as can be verified by the Structural PCA graph: the validation chemicals are into the training chemicals, as usually happens in this splitting methodology. The much higher response variability in the training set compared with the validation set justifies the low TSS value in the formula for Q^2ext calculation and thus the apparent low predictivity.

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{valid} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{valid} (y_i - \bar{y}_{tr})^2} = \text{1-PRESS/TSS}$$

It is always advisable to combine the MSE or similarly RMS values calculation for the training and validation sets.

FIGURE 61 : Regression line and residuals/leverage diagnostic (Williams graph) of model



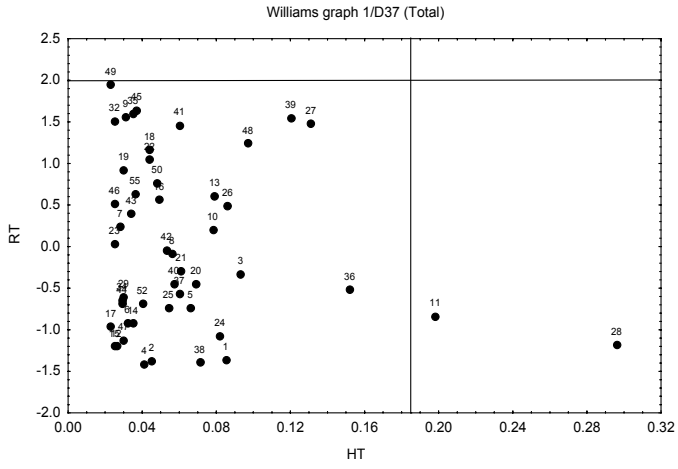
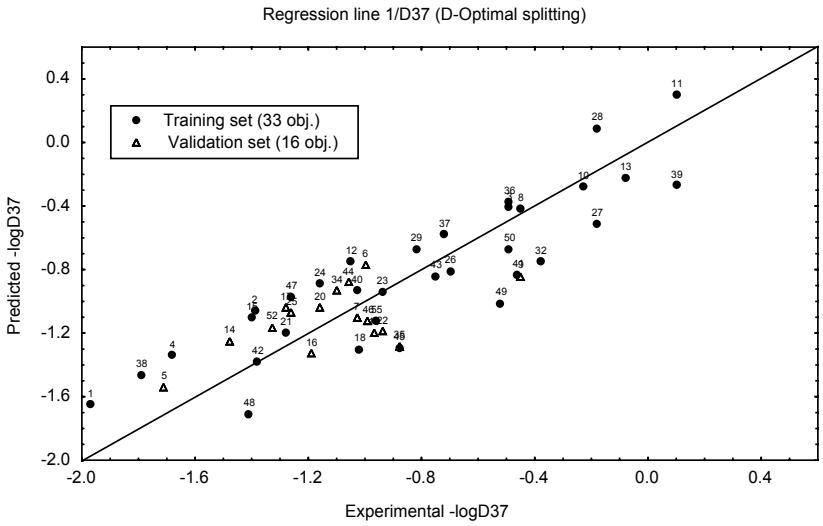
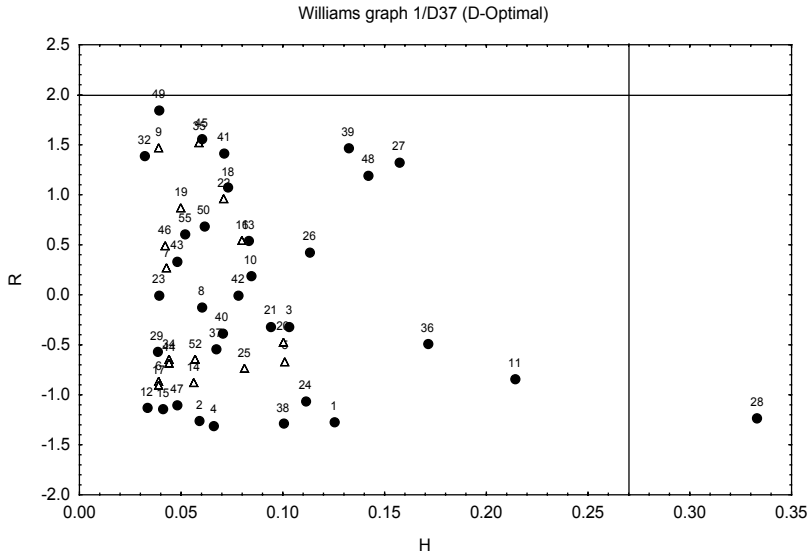


FIGURE 62: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





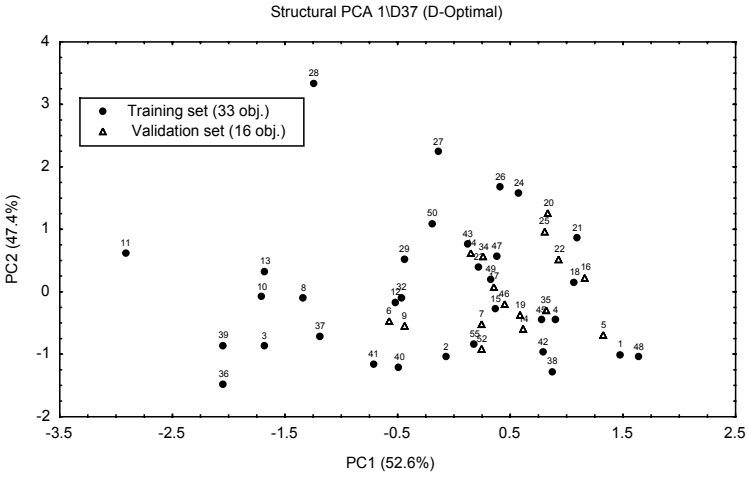
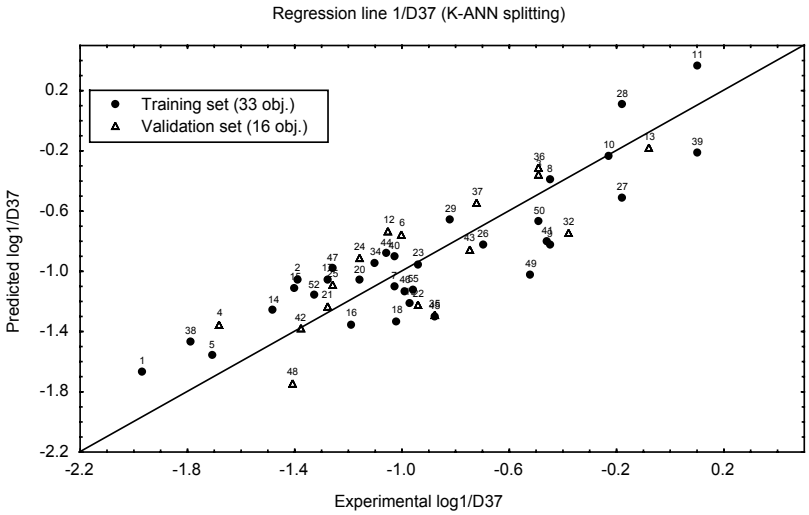
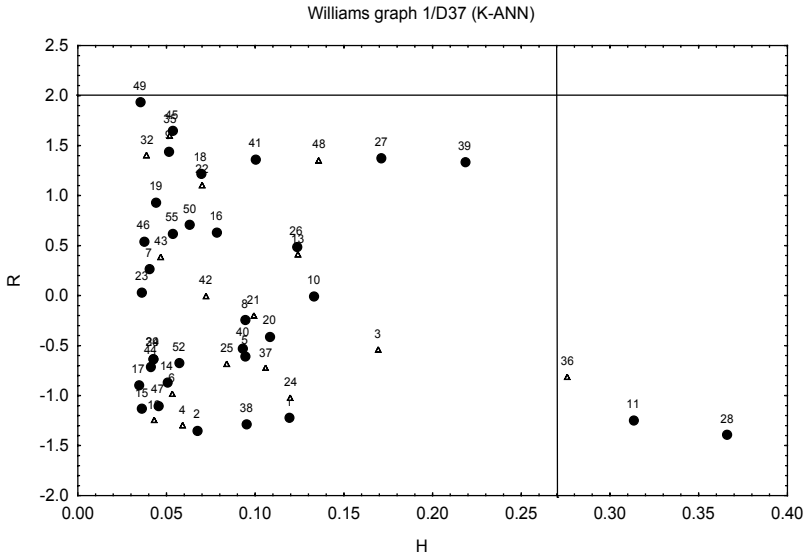


FIGURE 63: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





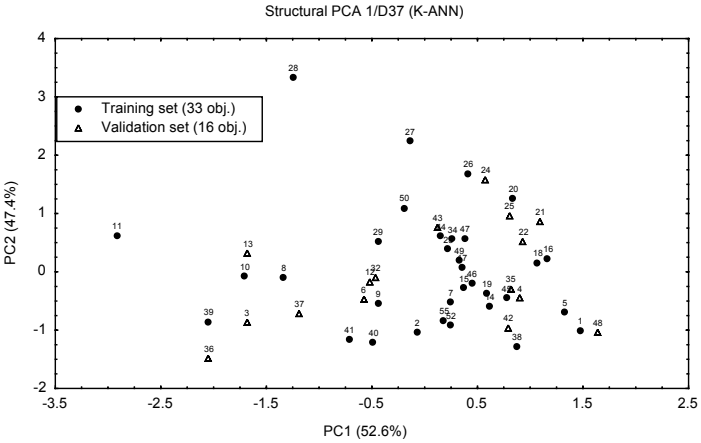
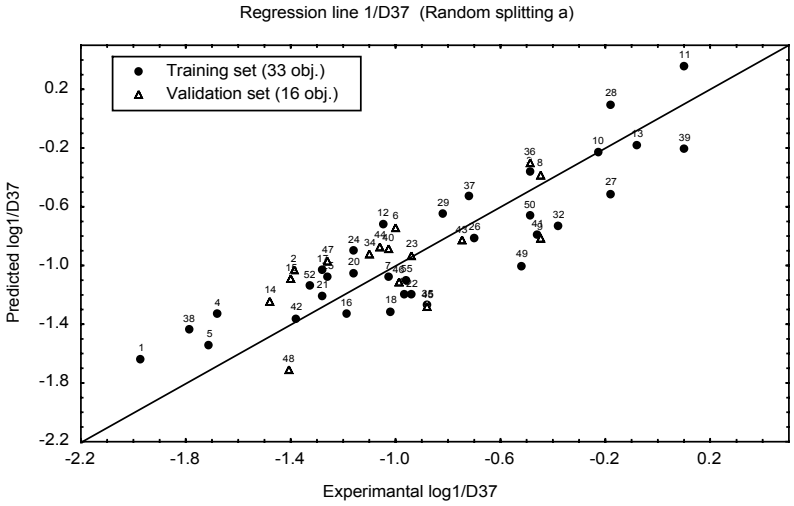
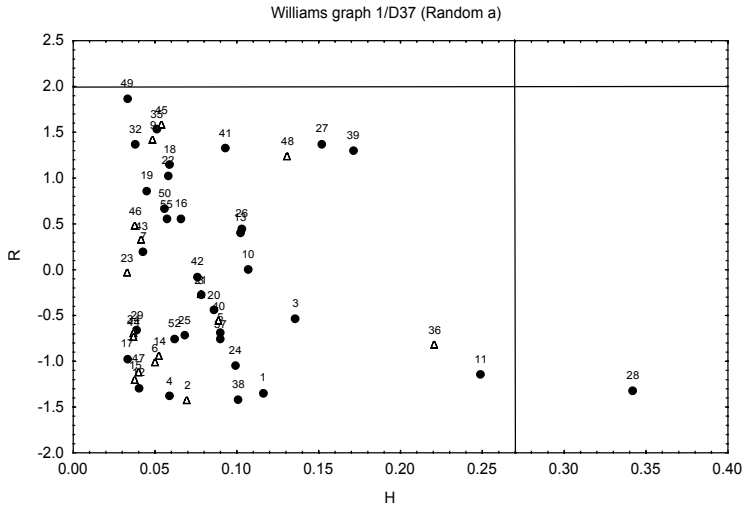


FIGURE 64: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





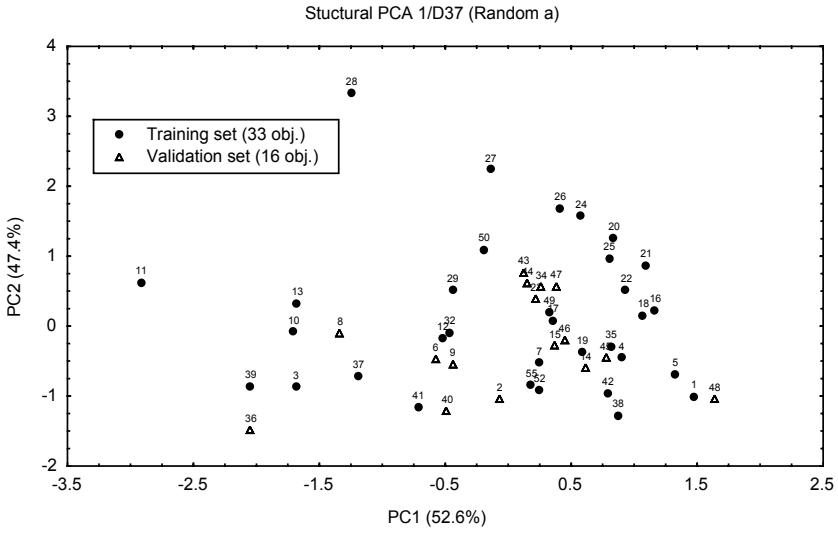
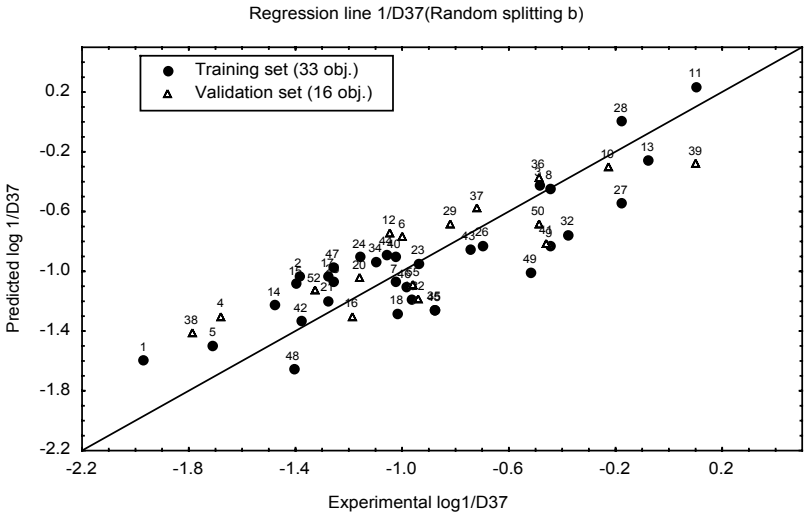
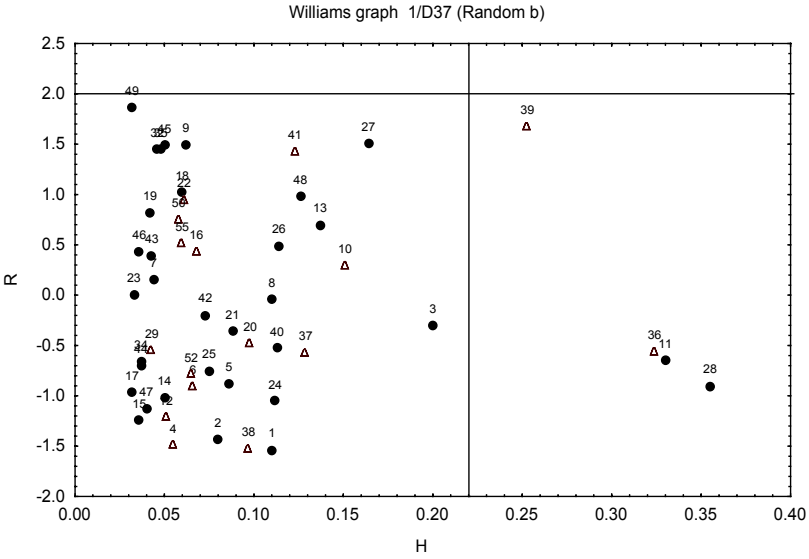
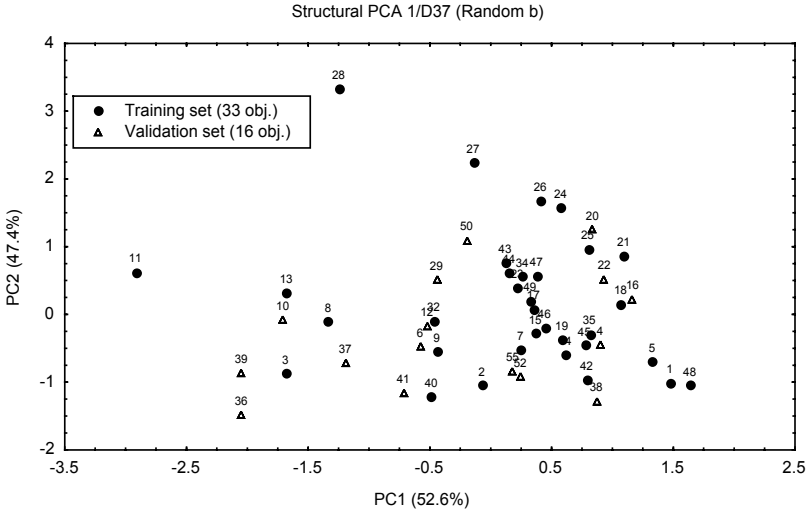


FIGURE 65: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors







4) The toxicity of 13 nitrobenzenes to C.vulgaris ($1/EC_{50}$) (Table in Annex) has been modelled.

The published Ordinary Least Squares model is:

$$\text{Log } 1/EC_{50} = 0.911 \text{ logP} - 1.55 \, E_{LUMO} - 3.88$$

$$N=13 \quad s=0.442 \quad r^2= 0.767 \quad r^2cv=0.701 \quad F= 20.8$$

1 outlier (4-chloronitrobenzene, 2) was identified by the authors and removed, the new model is:

$$\text{Log } 1/EC_{50} = 0.952 \text{ logP} - 1.68 \, E_{LUMO} - 4.24$$

$$N=12 \quad s=0.353 \quad r^2= 0.861 \quad r^2cv=0.813 \quad F= 35.2$$

The regression line and the Williams plot, performed during our work on these two models, are reported in Fig. 66 and 67: chemical 2 is an outlier, commented on by the authors in the first model, but chemical 9 (1,2-dinitrobenzene) is a new outlier in the second model, this second outlier was neither identified nor commented on by the authors. It is worth stating here that often the removal of outliers, and/or influential chemicals, results in models where new outliers and influential chemicals appear. In the present case the models show no influential chemicals to be present.

VALIDATION:

The models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap.

The limited size of the data set does not allow reasonable statistical external validation by preliminary splitting.

Table 13: Statistical Diagnostics of models

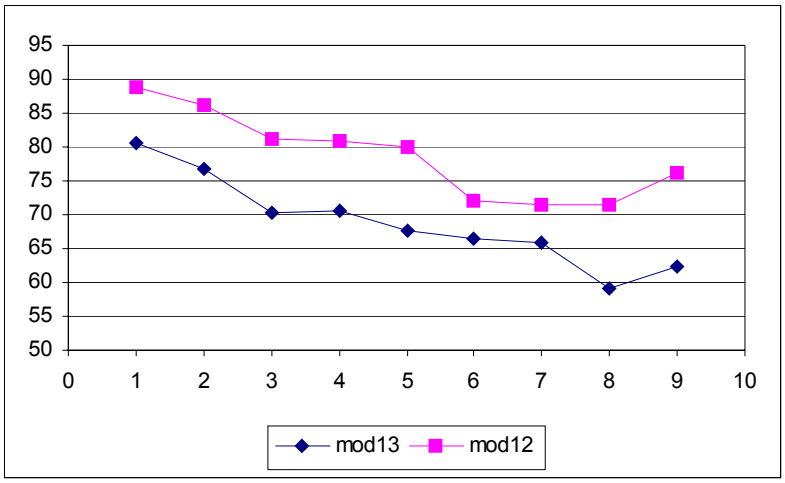
	ntr	variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	MSE	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
$1/EC_{50}$	13	logP E_{LUMO}	70.2	80.6	59.3	62.3	76.8	0.150	0.480	0.387	20.7	0.440	36.70	46.00	9.30
$1/EC_{50}$	12	logP E_{LUMO}	81.3	88.7	71.5	76.1	86.2	0.094	0.390	0.305	35.1	0.350	38.10	47.80	9.69

Regarding collinearity: the descriptors are correlated (Kxx) but, most importantly, the difference in the correlation between the block of X variables plus the response Y (Kxy) and the correlation among the X (Kxx) is sufficiently high (delta column) compared with other QSAR models and according to our experience. All the models were verified also by Y-scrambling: compared to the published models, the models on randomised response all have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

Table 13 bis: Statistical Diagnostics of models

		mod13	mod12
1	R^2	80.6	88.7
2	R^2_{adj}	76.8	86.2
3	Q^2	70.2	81.3
4	Q^2_{LMO10}	70.5	80.9
5	Q^2_{LMO20}	67.7	80.0
6	Q^2_{LMO30}	66.5	71.9
7	Q^2_{LMO40}	66.0	71.5
8	Q^2_{LMO50}	59.3	71.5
9	Q^2_{boot}	62.3	76.1

The following is the graphical representation of the parameters reported in the above table.



It is immediately evident that the two models, obtained on a small data set (only 13 chemicals), are not stable and probably useful mainly for fitting purposes, but scarsely for predictive aims (especially the first). The differences between R^2 and Q^2_{LOO} and between Q^2_{LOO} and Q^2_{LO50O} are nearly 10%. Again bootstrapping (9) gives the intermediate value for internal predictivity validation. Due to the small size of the data set there is no possibility of a reasonable splitting for statistical external validation, thus no conclusions can be drawn with regard to the predictive ability of these models for chemicals not used in the model development.

FIGURE 66: Regression line and residuals/leverage diagnostic (Williams graph) of model

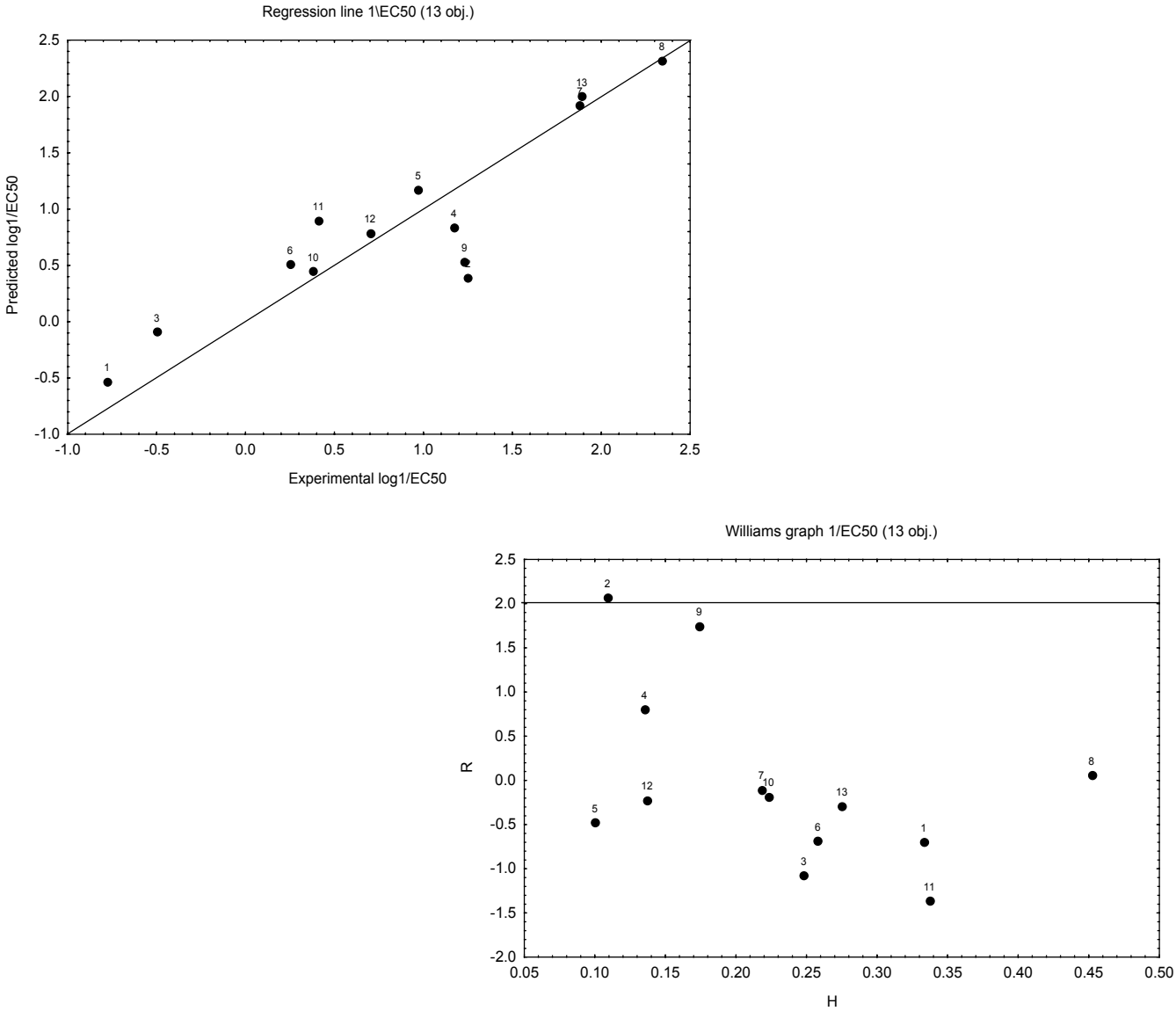
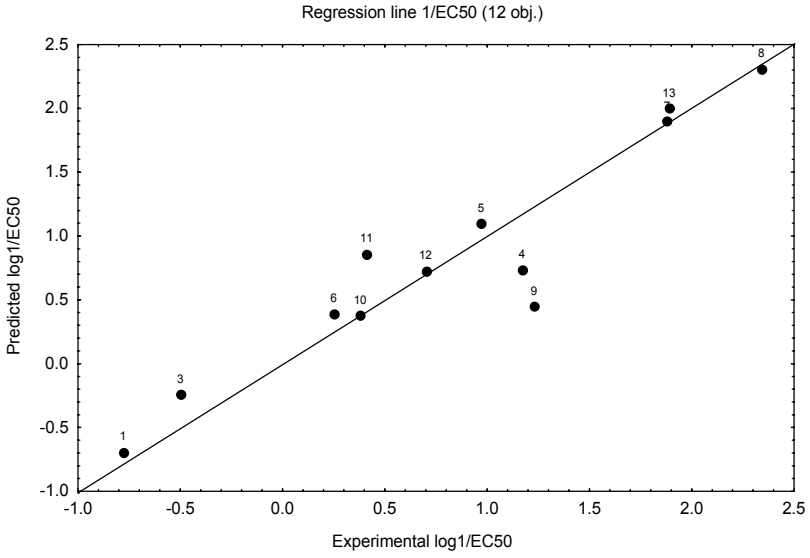
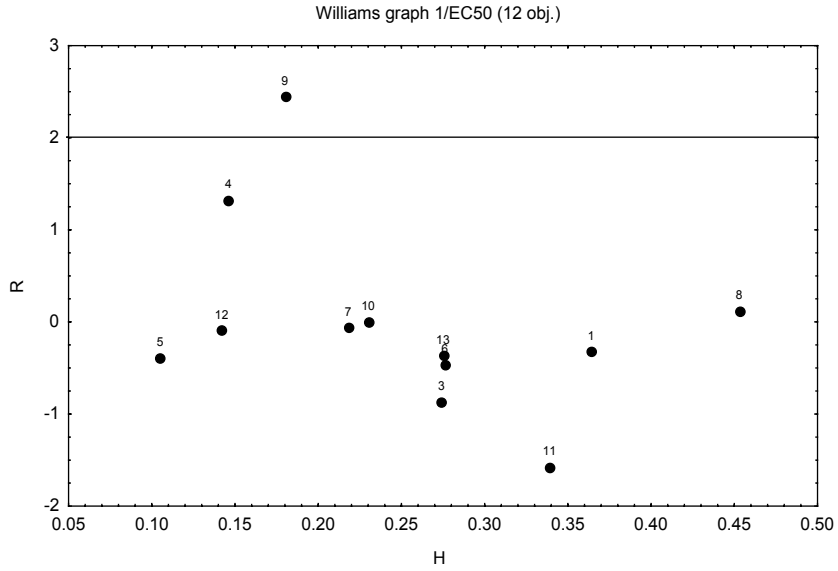


FIGURE 67: Regression line and residuals/leverage diagnostic (Williams graph) of model





MAIN CONCLUSIONS for VALIDATION

Internal validations

Models developed on small data sets (12-20 chemicals) can be useful for fitting models, their actual predictivity generally being lower than that presumed from LOO cross-validation. Stronger **internal validation** by LMO and bootstrapping demonstrate their real internal predictive power. Bootstrapping can be considered the best approach because the dimension of the training set in each validation run is the same as in the original data set, and information is not too scarce as can happen in 40-50% of data set perturbations (LMO). Note that with this kind of validation nothing can be concluded with regard to the model predictivity for new chemicals.

Internal cross-validation by LMO (up to 50% if the data set is not too small) and, preferably, by bootstrap are the recommended approaches to verify the real predictivity for data set chemicals used for the model development.

It is important to note that, as the information related to each chemical in the data set is considered in at least one run of the validation process, both approaches are internal validation: the chemicals are never new chemicals in the model development (they are included in some training of LMO or bootstrap procedure) and their information is considered and included in the model.

In conclusion, the models developed on small size data sets are generally of little use for the prediction of new data. Their principal utility, mainly in mechanistically-based models like those assessed in this work, is related to an understanding of some mechanism or some structural peculiarity of a chemical.

Chemical Domain of model applicability

In each model, both the outliers and the highly influential chemicals must be verified and, if possible, explained. In this contract work, we have verified the outliers by standardized cross-validated residuals (jackknifed), and the influential chemicals by the leverage approach through Hat diagonal values: the Williams plot of the regression is a graph that can allow the visual inspection of these anomalous chemicals. By the leverage approach it is possible to verify the Chemical Domain of model applicability. The predicted data for chemicals with leverage values higher than the H control (threshold value) must be considered with caution. Such data could be unreliable as these chemicals are at the border of the model space, thus their prediction could be the result of extrapolation not of interpolation; for chemicals with a leverage value below the threshold value the predictions are interpolated values.

In this contract work we always applied the leverage approach, highlighting chemicals that are highly influential in both the training set for model development and in the validation set.

Collinearity: collinearity among descriptors must be checked, and correlation between the X block and the Y response verified (we apply the QUIK rule of Todeschini based on K (see theoretical part in the introduction): K_{XY} must be significantly higher than K_{XX}). Collinearity among descriptors is not always dangerous, but it must be carefully controlled to have predictive models. Actually, in some cases, intercorrelated descriptors may carry useful structural information in the parts in which they don't correlate with the other descriptors.

Statistical external validation

Statistical external validation by the splitting of available data sets is possible only when the data sets are of a reasonable size, thus allowing the information in the training set to be maintained.

In this work an exercise of statistical external validation was performed only on the data set of A. nidulans toxicity (49 chemicals). Statistical external validation in small data sets is highly dependent on the applied splitting methodology: anomalous results (under- or over-optimistic) could be obtained by Q^2ext calculation. In this case it is better to verify the mean square error (MSE) of the models and compare the results of prediction for training and validation chemicals.

Shortcomings of some models:

1. Outliers and influential chemicals, not evidenced by the authors
2. Instability of GLDH model (less predictivity than apparent)
3. Only internal validation possible for small data sets (12 chemicals)

8. STATISTICAL VALIDATION OF MUTAGENICITY MODELS (Gramatica et al.)

Gramatica, P., Consonni, V. and Pavan, M. (2003). Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. ***SAR and QSAR in Environmental Research*** **14 (4), 237-250.**

The results of the two Ames test (TA98 e TA100) for a total of 146 aromatic amines (96 for TA98 and 76 for TA100) (Table in Annex) have been modelled by theoretical molecular descriptors.
The published Ordinary Least Squares (OLS) models, by Genetic Algorithm Variable Subset Selection, are:

$$\text{Log TA98} = -3.98 + 2.40\ \text{MWC07} + 0.56\ \text{MATS7m} + 2.44\ \text{Mor27u} + 1.12\ \text{Mor15m} \qquad (1)$$

$$N_{training} = 60 \quad R^2 = 80.3 \quad Q^2_{LOO} = 76.6 \quad Q^2_{LMO} = 75.9 \quad Q^2_{ext} = 68.9$$
$$K_{XX} = 27.9 \quad s = 0.827 \quad F_{(55)} = 55.87 \quad SDEC = 0.791 \quad SDEP = 0.861 \quad SDEP_{ext} = 0.991$$

$$\text{Log TA100} = -3.99 - 0.61\ \text{nHA} + 9.55\ \text{ATS5p} + 0.65\ \text{L2v} \qquad (2)$$

$$N_{training} = 46 \quad R^2 = 81.2 \quad Q^2_{LOO} = 78.0 \quad Q^2_{LMO} = 77.4 \quad Q^2_{ext} = 67.1 \quad K_{XX} = 17.1$$
$$s = 0.579 \quad F_{(42)} = 60.40 \quad SDEC = 0.553 \quad SDEP = 0.598 \quad SDEP_{ext} = 0.731$$

The models were assessed by the authors during the model development, both internally (cross-validation by LOO, LMO 20%) and externally by preliminary splitting with Experimental design (D-optimal distance). The highest value of Q^2_{ext} was the parameter chosen by the authors to select the best models.

VALIDATION:
In this contract work, the models were assessed by internal cross-validation (LOO and LMO up to 50% of perturbation), Y-scrambling, bootstrap. In addition to different internal validation approaches, statistical external validation was performed by comparing different approaches for the preliminary splitting of chemicals into training and validation sets (D-optimal Distance, Kohonen-ANN; random). The following tables and graphs show the results. Note that the value of Q^2ext is slightly different as a different formula was applied in the paper. The formula from the theoretical part of the Introduction was used here.
The regression lines and the Williams plot for the different splittings are reported below, together with the PCA of structural descriptors, to verify the distribution of the two sets regarding structural information. Some common outliers, some outliers present in a specific splitting, and no influential chemicals are

evidenced in the TA98 models, while in the TA100 models some outliers and some highly influential chemicals, depending on the applied splitting methodology, are highlighted.

TA98

Table 14: Statistical Diagnostics of TA98 models

n. tr	n. valid	split	variables	Q^2	R^2	Q^2_{LMO50}	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE Tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
60	38	D-Opt.	MWC07 MATS7m Mor27u Mor15m	76.6	80.3	74.0	74.4	78.8	72.2	0.626	1.008	0.863	0.791	55.9	0.827	28.0	40.9	12.9
61	37	K-ANN	MWC07 MATS7m Mor27u Mor15m	72.9	76.8	70.5	70.7	75.1	74.6	0.708	0.893	0.909	0.841	46.3	0.878	31.0	41.0	10.0
60	38	Rand. 1	MWC07 MATS7m Mor27u Mor15m	65.3	70.7	61.6	62.2	68.6	82.5	0.858	0.644	1.008	0.926	33.2	0.967	33.5	43.3	9.8
60	38	Rand. 2	MWC07 MATS7m Mor27u Mor15m	79.7	82.2	77.7	78.0	80.9	64.3	0.610	1.037	0.834	0.781	63.5	0.816	31.3	42.6	11.4

Regarding collinearity: the descriptors are quite correlated (medium Kxx: 31) but, most importantly, the difference in the correlation between the block of the X variables plus the response Y (Kxy) and the correlation among the X (Kxx) is sufficiently higher (medium delta: 11) compared with other QSAR models and according to our experience.

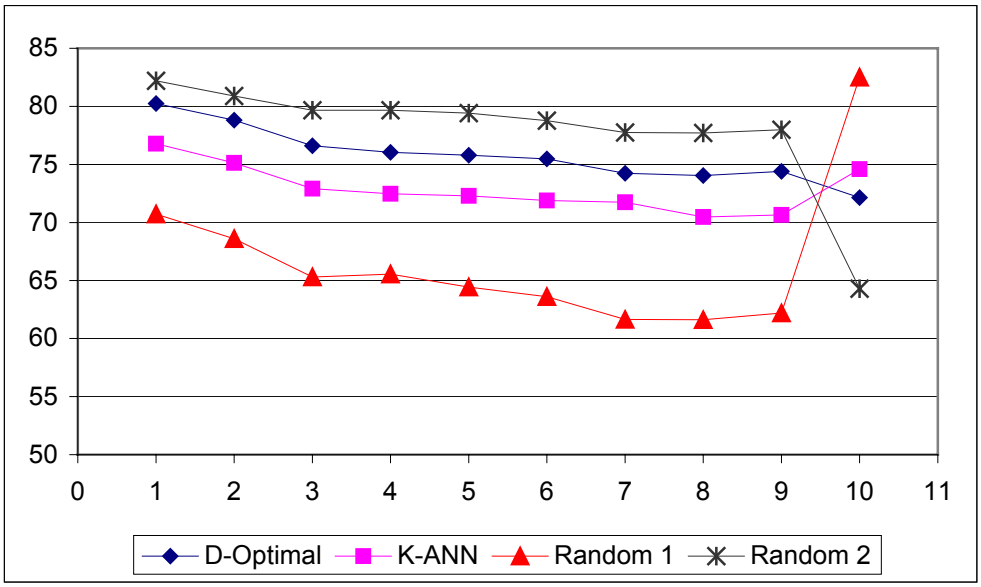
All the models were also verified by Y-scrambling: compared with published models the models on randomised response all have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table and plotted in the graph:

Table 14 bis: Statistical Diagnostics of TA98 models

		D-Optimal	K-ANN	Random 1	Random 2
1	R^2	80.3	76.8	70.7	82.2
2	R^2_{adj}	78.8	75.1	68.6	80.9
3	Q^2	76.6	72.9	65.3	79.7
4	Q^2_{LMO10}	76.1	72.5	65.6	79.7
5	Q^2_{LMO20}	75.8	72.3	64.4	79.4
6	Q^2_{LMO30}	75.5	71.9	63.6	78.8
7	Q^2_{LMO40}	74.2	71.8	61.6	77.8
8	Q^2_{LMO50}	74.0	70.5	61.6	77.7
9	Q^2_{boot}	74.4	70.7	62.2	78.0
10	Q^2ext	72.2	74.6	82.5	64.3

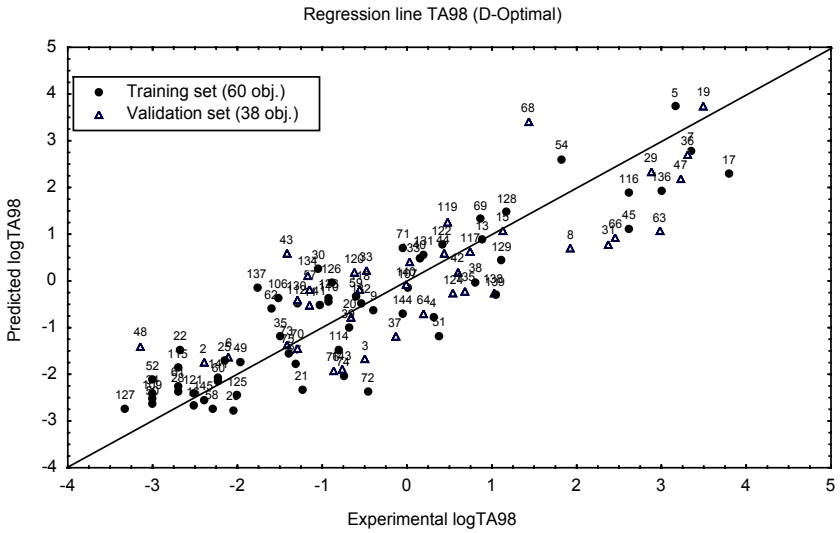
The following is the graphical representation of the parameters reported in the above table.

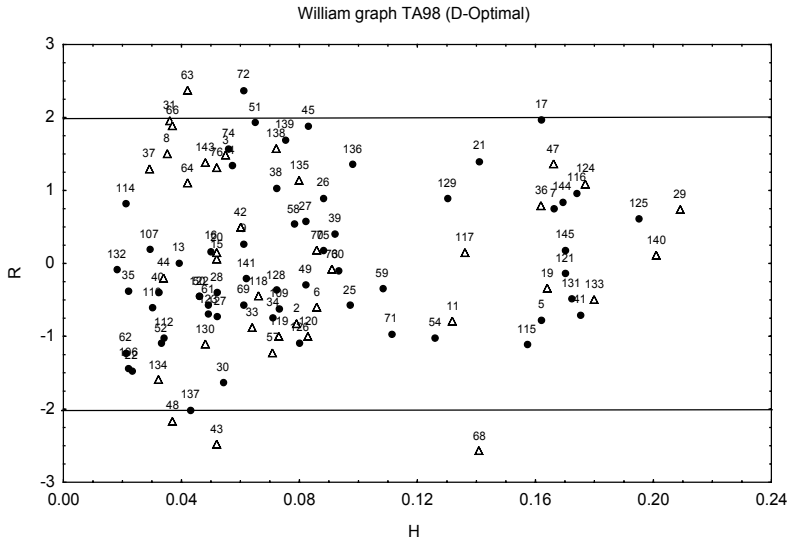


The proposed model has satisfactory performance in both the internal and statistical external validations. It is robust. The difference between R^2 and Q^2 is small for D-optimal and K-ANN splittings (about 4%), while random splittings give anomalous results very dependent on the splitting itself. The strong internal validations (LMO and Bootstrapping) highlight the predictivity of the model for chemicals in the training set: as in other examples the bootstrap approach appears the best compromise for this internal validation. Statistical external validation gives similar results for the model predictivity of validation chemicals in the D-optimal and K-ANN splitting approaches, while it must again be noticed that random splitting markedly influences the results of both internal and statistical external validation.

For this reason a careful choice of the splitting methodology is fundamental to avoid conflicting and deformed results: for instance, the model verified by Random 1 splitting appears less predictive internally because the outliers are in the training, but highly predictive externally, the contrary happens for the Random 2 splitting where the strongest outliers are in the validation set.

FIGURE 68: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





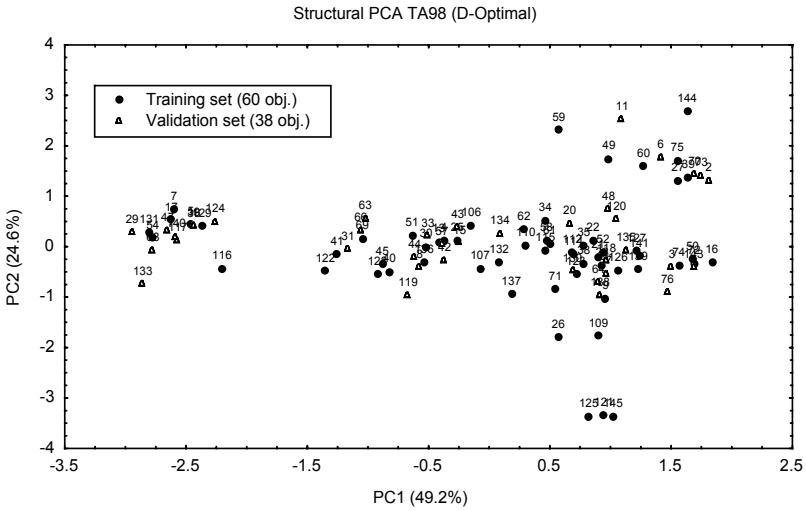
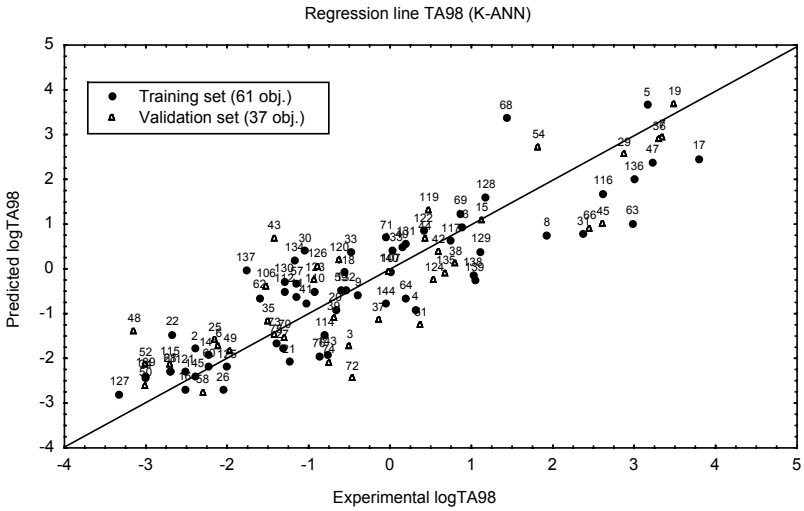
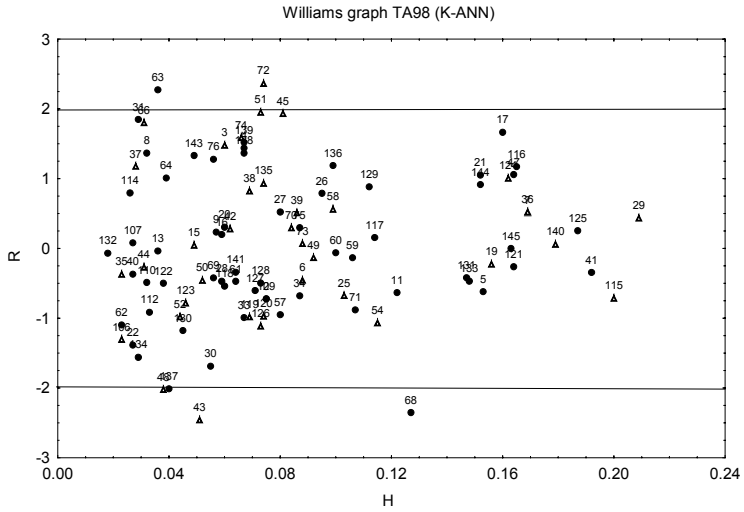


FIGURE 69: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





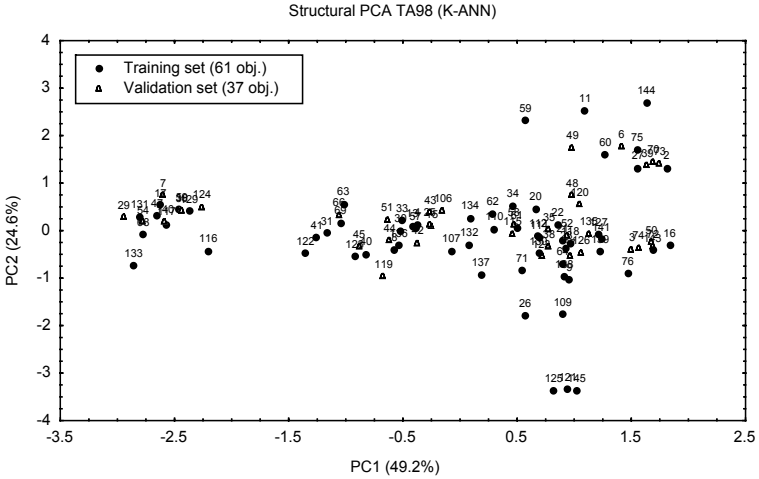
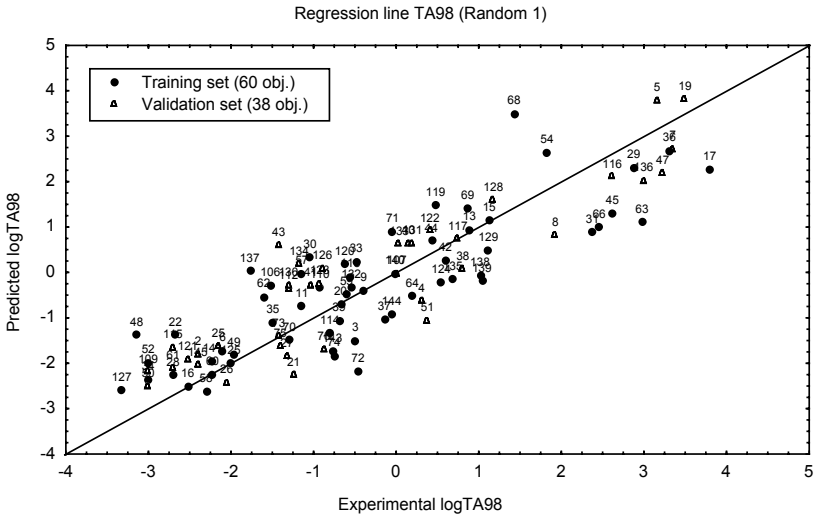
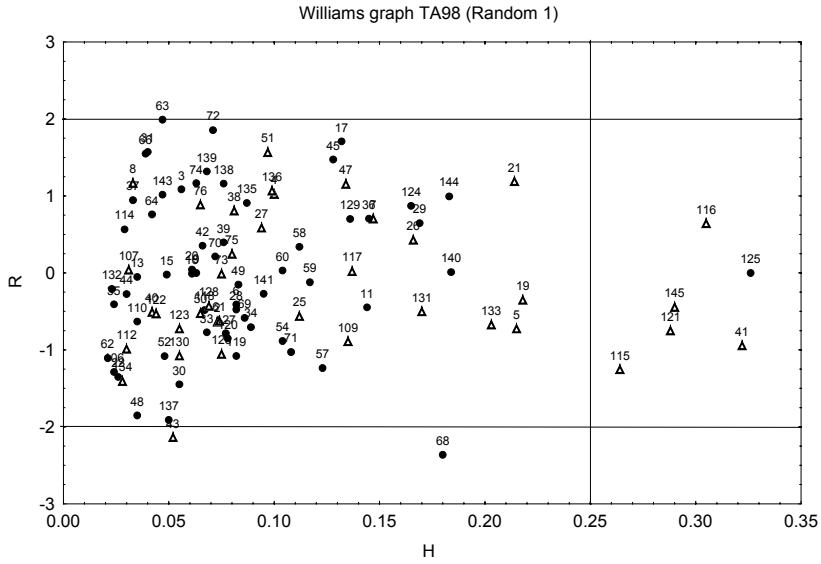


FIGURE 70: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





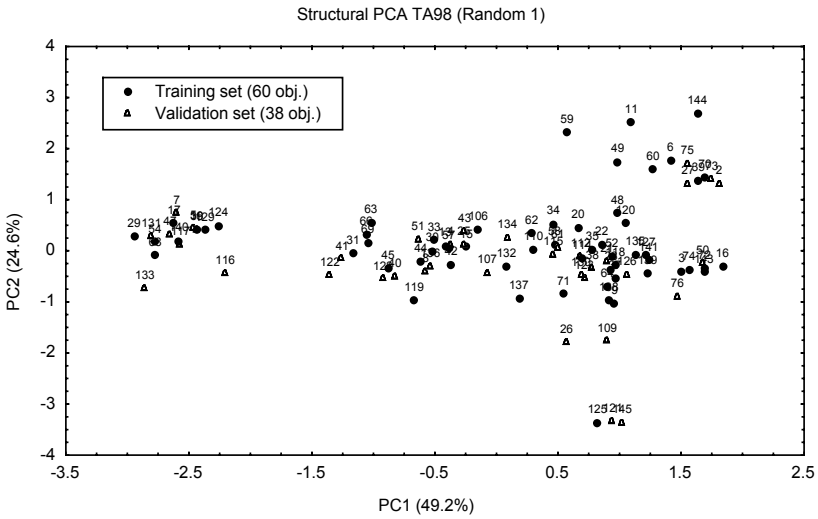
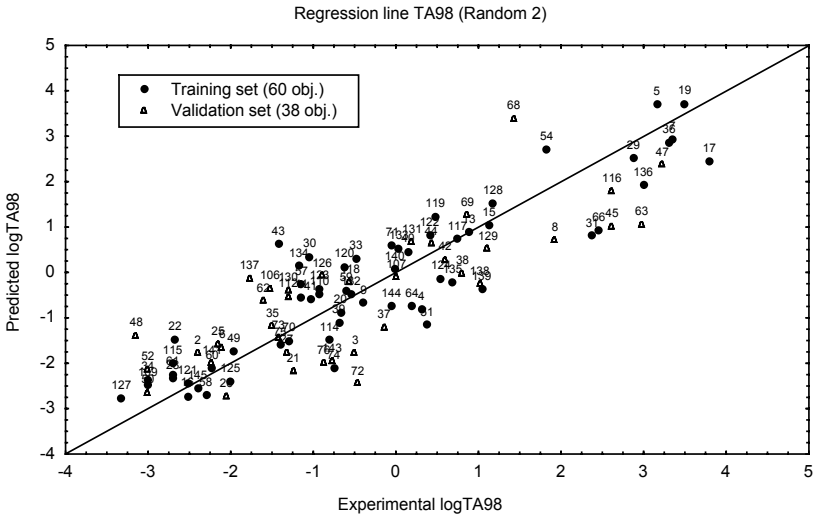
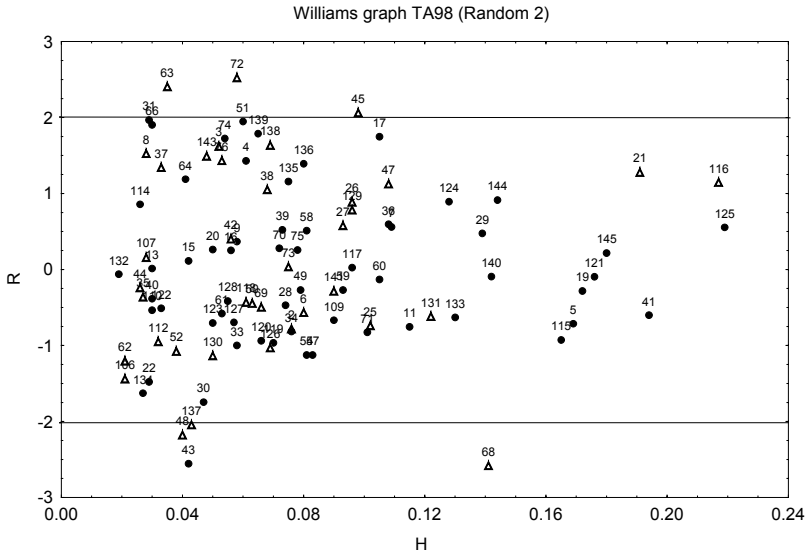
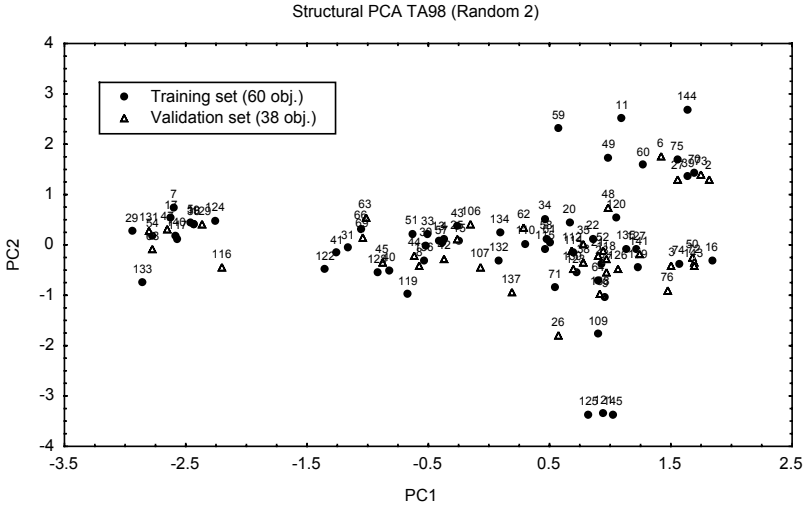


FIGURE 71: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors







TA100

Table 15: Statistical Diagnostics of TA100 models

n. tr	n. valid	split	variables	Q^2	R^2	$Q^2_{LMO}50$	Q^2_{boot}	R^2_{adj}	Q^2_{ext}	MSE Tr	MSE valid	SDEP	SDEC	F	s	Kxx	Kxy	ΔK
46	30	D-Optimal	nHA ATS5p L2v	78.0	81.2	75.5	76.1	79.8	78.0	0.306	0.653	0.598	0.553	60.4	0.579	17.11	39.25	22.14
48	28	K-ANN	nHA ATS5p L2v	73.1	77.1	70.1	70.7	75.6	80.6	0.423	0.452	0.705	0.651	50.2	0.679	19.64	39.90	20.26
46	30	Random	nHA ATS5p L2v	73.1	76.9	70.4	70.6	75.2	78.5	0.512	0.386	0.772	0.716	46.3	0.749	27.22	44.99	17.77

Regarding collinearity: the descriptors are not very correlated (medium Kxx: 21) but, most importantly, the difference in correlation between the block of X variables plus the response Y (Kxy) and the correlation among the X (Kxx) is very high (medium delta: 20) compared with other QSAR models and according to our experience.

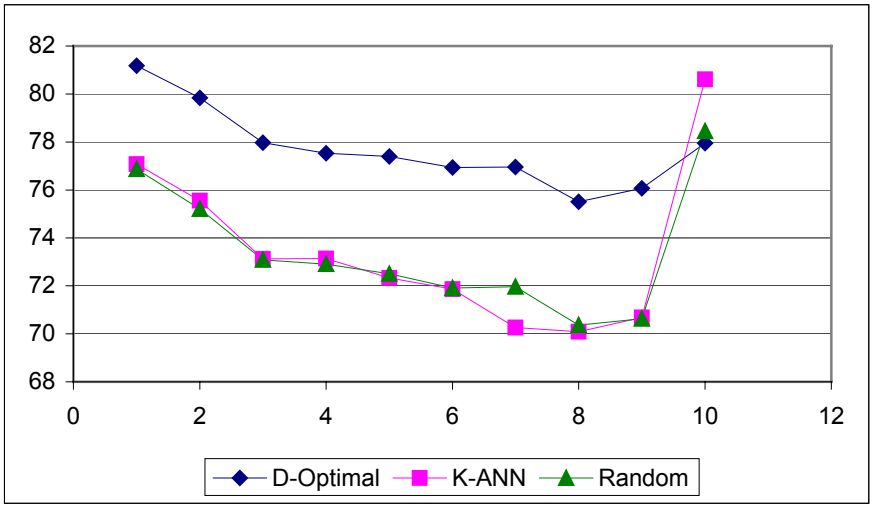
All the models were also verified by Y-scrambling: compared with published models the models on randomised response all have extremely low R^2 and Q^2. This is a demonstration that the reported models are not obtained by chance correlation.

An analysis of the results of fitting and prediction parameters was done on the data reported in the following table and plotted in the graph:

Table 15 bis: Statistical Diagnostics of TA100 models

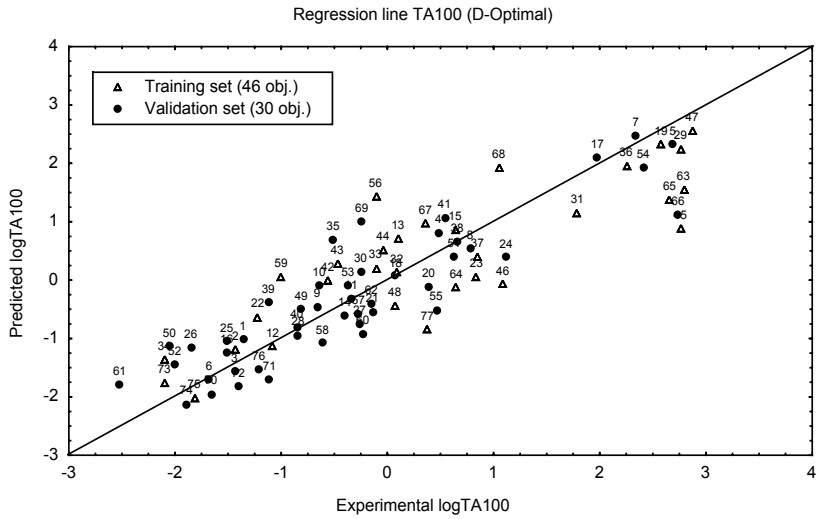
		D-Optimal	K-ANN	Random
1	R^2	81.2	77.1	76.9
2	R^2_{adj}	79.8	75.6	75.2
3	Q^2	78.0	73.1	73.1
4	Q^2_{LMO10}	77.5	73.1	72.9
5	Q^2_{LMO20}	77.4	72.3	72.5
6	Q^2_{LMO30}	76.9	71.9	71.9
7	Q^2_{LMO40}	77.0	70.3	72.0
8	Q^2_{LMO50}	75.5	70.1	70.4
9	Q^2_{boot}	76.1	70.7	70.6
10	Q^2ext	78.0	80.6	78.5

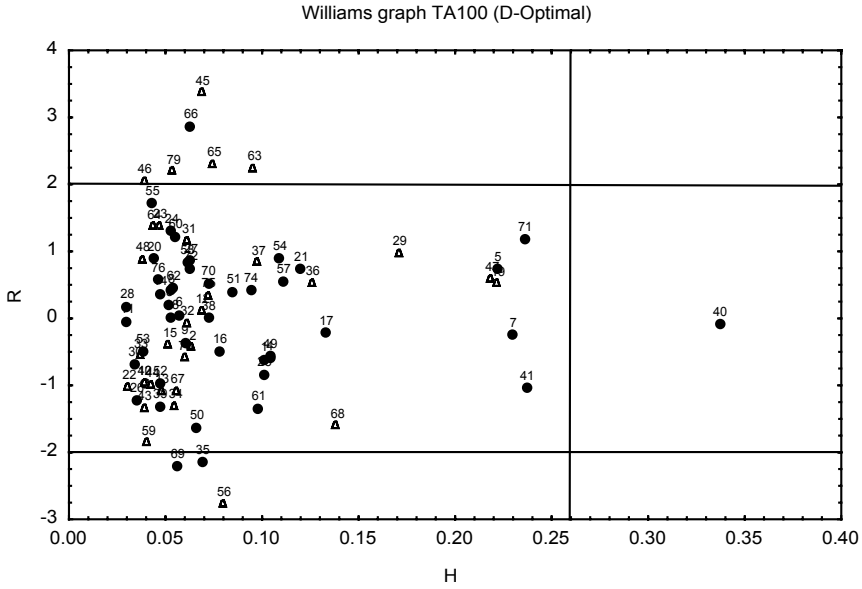
The following is the graphical representation of the parameters reported in the above table



The proposed model shows satisfactory performance in both internal and statistical external validation. It is robust. The difference between R^2 and Q^2 is small for each splitting (3-4%). Strong internal validation (LMO and Bootstrapping) highlights the predictivity of the model for chemicals in the training set: as in other examples. Bootstrapping appears the best compromise for this internal validation. In this case, splitting by D-optimal design give more optimistic results both in fitting and in internal validation because the outliers are mainly in the validation set (see the corresponding Williams plot): again the selection of chemicals in the training and validation sets is highly influential on the validation parameters. Statistical external validation gives similar results for the model predictivity of validation chemicals in the D-optimal, K-ANN and random splittings. Contrary to what happened in the TA 98 model, this random splitting was well balanced and more similar to K-ANN than to D-optimal splitting. It must be noted that in the random splitting there were 4 highly influential chemicals (3 of which were in the validation set): thus the prediction for these chemicals could be unreliable. One common high leverage chemical (40) is present in every splitting.

FIGURE 72: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





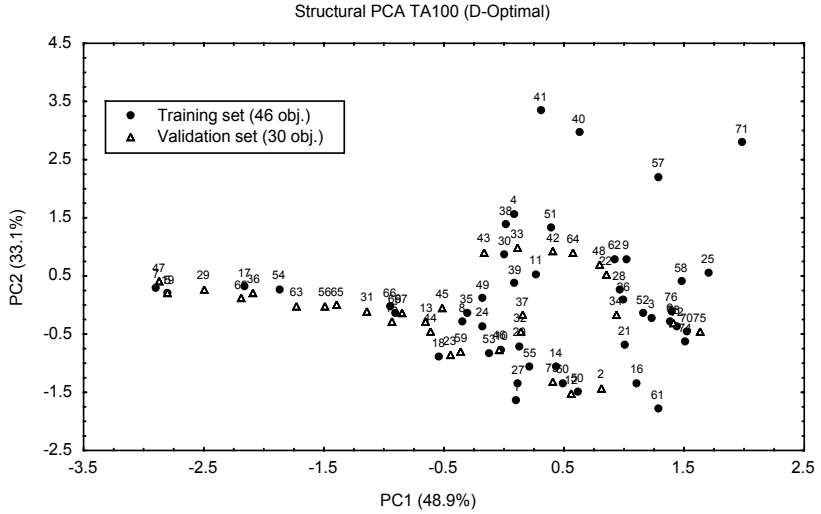
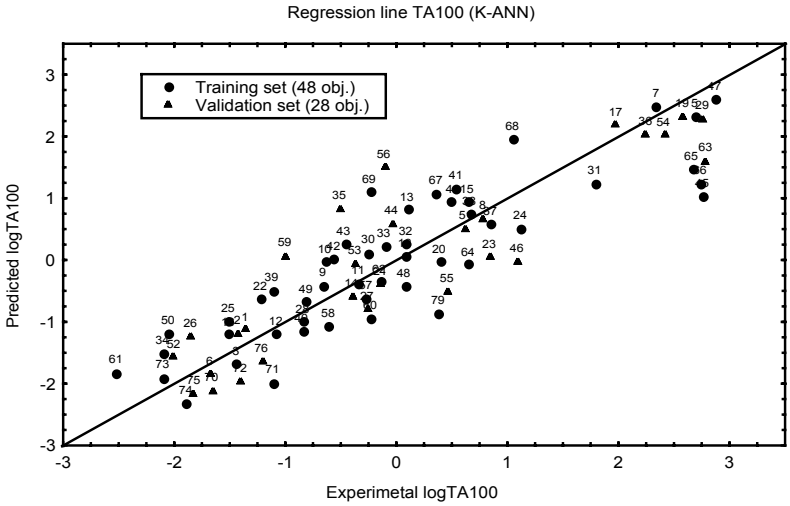
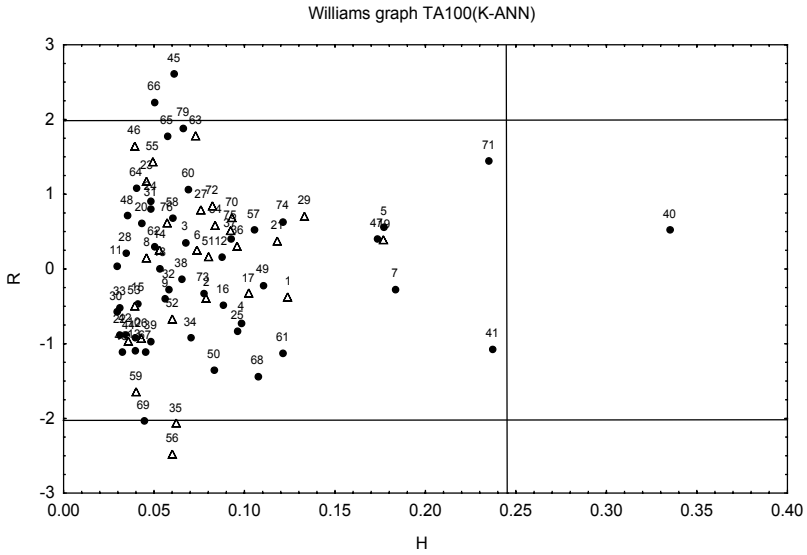


FIGURE 73: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors





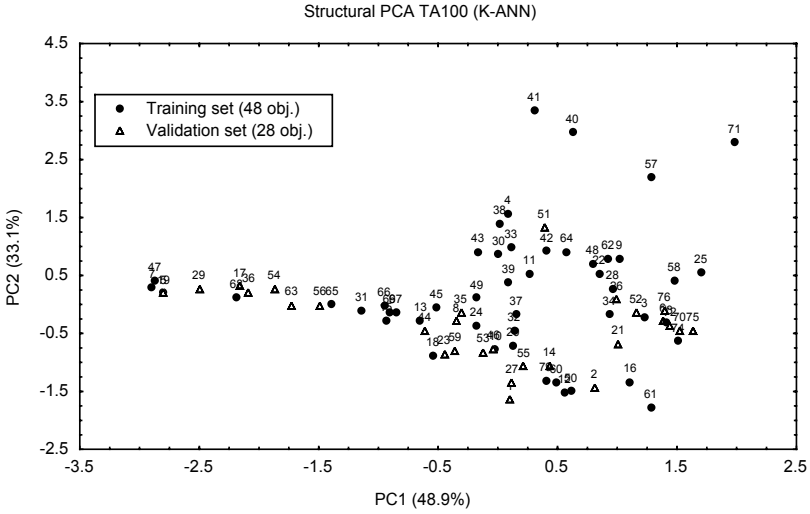
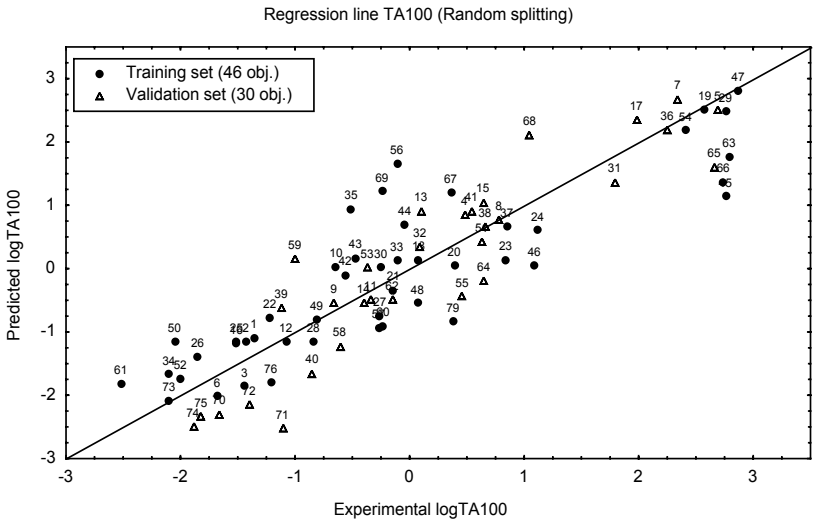
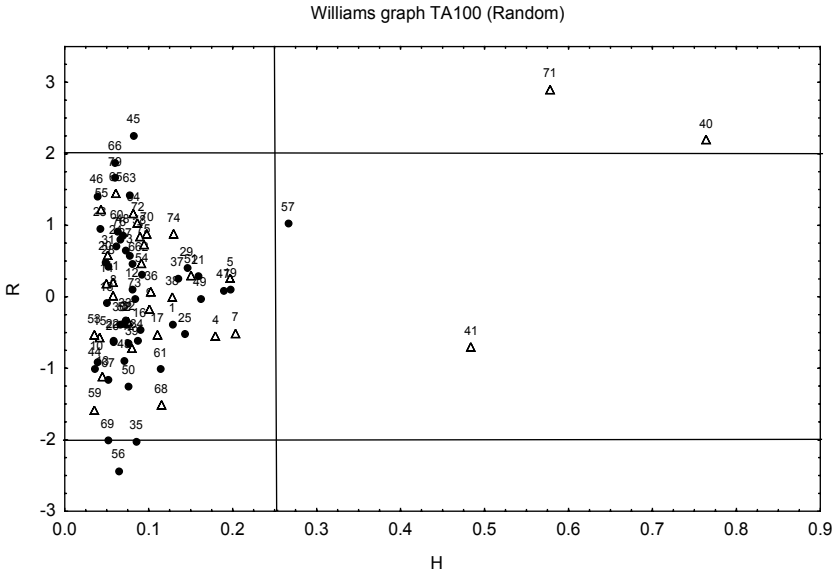
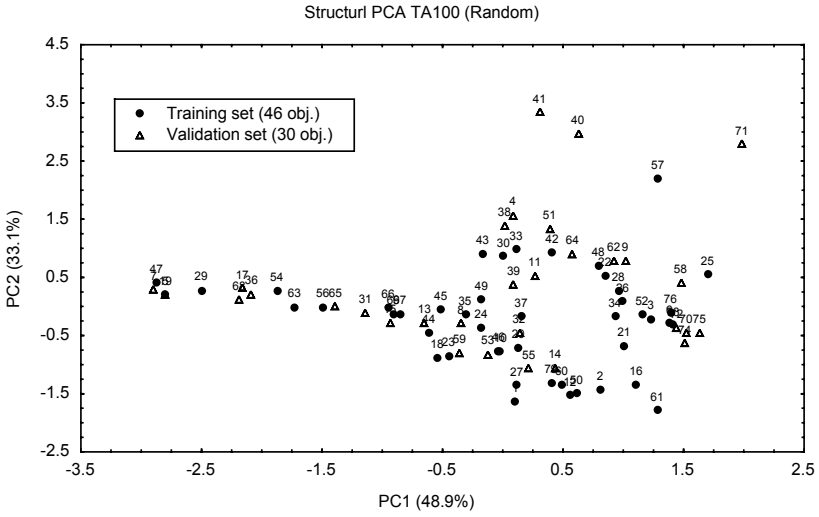


FIGURE 74: Regression line, residuals/leverage diagnostic (Williams graph) and structural analysis of the data set by Principal Component Analysis (PCA) of model descriptors







MAIN CONCLUSIONS for VALIDATION

The models published in this paper are stable and robust, with satisfactory fitting and predictive performances. They are predictive for the chemicals used in the model development (internal validation on training chemicals) and also for chemicals not used in the model development (statistical external validation on validation chemicals). The available data sets were split into training and validation sets by different splitting methodologies (D-optimal, K-ANN and random) and the predictivity of the models was verified in each splitting.

Internal validation by LMO (up to 50% if the data set is not too small) or, preferably, by bootstrapping are the recommended approaches to verify the real internal predictivity for chemicals in the data set.
It is important to note that both these approaches are internal validations as the information related to each chemical in the data set is considered at least in one run of the validation process: these chemicals are never new chemicals in the model development and their information is taken into account.

Statistical external validation, on the contrary, verifies the predictivity for chemicals not used in the model development, thus the information on these new chemicals is not included *a priori* in the modelling.
In relation to this point, it must be noted that the real utility of this approach: splitting of available chemicals and Q^2_{EXT} determination is during model development. The authors (we) selected the best models, from among a hundred possible models developed by Genetic Algorithm Variable Subset Selection applied to OLS modelling, by maximizing the external predictivity on the split validation sets. It is very important to note that in the population of possible models, all with high R^2 and Q^2 values (apparently good models), there were only a very few models with high Q^2_{EXT} values. The majority of the models with high fitting and internal predictivity (even higher than the reported models) had a very low Q^2_{EXT} value (31-52%). This is a clear demonstration that internal validation is a necessary, but not sufficient validation procedure. The real predictivity of a model for chemicals not used for model development must be verified by statistical external validation. This is particularly true in model development using variable subset selection procedures.

Splittings: random splitting must be avoided for statistical external validation as the results of the validation depend greatly on the training and validation set compositions. A careful splitting into

representative sets must be performed by suitable methodologies. We have verified that D-optimal design and Kohonen Maps-Artificial Neural Networks are good splitting methodologies.
The results of internal and statistical external validation highlight that such validations are sensitive to the distribution of response in the two split sets in relation to the presence of outliers, but are insensitive to the distribution of chemicals that are influential as a result of their structural descriptors (high leverage).

Chemical domain of applicability: the Williams plot of the OLS regression allows an easy check of outliers and influential chemicals.

Shortcomings of the models:
No shortcomings found: the reason is that the statistical approaches, evaluated in this exercise, including the external validation for the choice of the best predictive model, were applied by the authors during the model development.
It is again demonstrated that this is the best way to propose QSAR models with reliable predictions.

9. CONCLUSIONS ON THE STATISTICAL VALIDATION APPROACHES AND THEIR NEED FOR APPLICABILITY

Model validation is one of the most important aspects of QSAR analysis.
The assumption that training set data, used for model development, are representative for X-y relationships in a certain large population of chemicals is critical and too optimistic. In fact, a QSAR multivariate model gives the best results as an interpolation method, not an extrapolation method. Thus, QSAR models must be checked for predictivity and chemical domain of applicability.
Optimism is a well-known problem of "predictive" QSAR models.
In fact model performance for new chemicals is often worse than the performance expected on the basis of what is estimated from the development data set (training set) of the model, namely "apparent predictivity". This is particularly true in small data set (10-30 chemicals) modelling where, compared to predictive performance, the fit is too good. Fitting ability is improved by adding variables, however such addition could be absolutely not useful for predictivity improvement (**overfitting**).

Thus it is mandatory to first investigate the predictive relevance of a developed model.
The purpose of this contract work was to compare different methods of statistical validation for the estimation of the prediction ability of QSAR models.
In this work some literature models on diverse data sets, which differ in both the structure of the chemicals and in the response end-points, have been used as examples of QSAR modelling to verify the applicability and the derived information of different statistical validation techniques.

Such verification can be accomplished by **internal validation and statistical external validation** techniques resulting in information on the different levels of reliability.

First of all, the possibility that a model is only a fitting model, or that it is overfitting, must be checked (in this exercise some models, declared as predictive by the authors, are actually only fitting or overfitting models). In addition, careful inspection is needed to check for **collinearity** among the variables; collinearity among descriptors is not always dangerous, but it must be carefully controlled to have predictive models. In this contract work we verified the difference between K_{XY} and K_{XX} (delta) and found that for some models the correlation among the variables (K_{XX}) is too high (40-60 %), while that between variable block plus response and variable block ($\Delta K = Kxy - Kxx$) is too little (<6-8%).

Several approaches have been suggested to estimate model predictivity by internal validation. Cross-validation (CV) and bootstrapping techniques are internal validation techniques as only chemicals from the training set are used to estimate performance: the structural information of each chemical in the training set is taken into account in at least one validation run. Several runs are repeated (hundreds or thousand depending on the software) so that all the chemicals serve, at least once, to test the model. The average of the performance measures taken over several repetitions (iterations) is considered the performance estimate.

LOO cross-validation (LOO-CV) is the most commonly used parameter for the internal validation of a QSAR model, but it often over-estimates the true model prediction ability (over-optimism). When used for model selection, the selected model frequently has unnecessary variables, making the model larger than it should be and thus overfitting: this model often performs well in fitting the data used for model development but is poor in prediction. An improvement on this cross-validation is **LMO-CV**, where more than one sample at a time is left out for the validation. LMO is used as a way to counteract the slight over-optimism of LOO-CV. Methods to assess internal validity vary in the amount of data used to estimate the model and the amount of data kept out to test the model. With 50% of the data out, the performance of the full model was generally underestimated since only half the data was used to construct the model: this is a waste of valuable information and the model may not contain all the relevant structure information of the whole data set. Also the predictions were very unstable since half of the data set was used for validation.
In the case of small data sets (in this exercise about 20-30 chemicals), LMO-CVs (40-50%) are too strong in perturbation, and the stronger CV, which can give a more realistic idea of the real predictivity is LMO 30%.

Contrary to cross-validation, **bootstrap** methods are more efficient and stable as the entire data set is used for model development. Bootstrap can be seen as a smoothed version of CV. Predictivity estimates by bootstrapping are consistent with LMO-cross estimates but exhibit less variability. Thus, this method gives the most accurate estimates of model performance, as was also verified in the examples of this contract work; for this reason it is the suggested method for internal validation.

Y-scrambling highlights the presence of apparent models, obtained by chance correlation. In this exercise one verified model was probably a model by chance.

All the internal validation approaches verify model predictivity for chemicals that participated in model development, thus each chemical's information is already taken into account in the model (in terms of selected descriptors and regression coefficients). Nothing is known regarding the predictivity of the model for chemicals not participating in model development i.e. new or "unknown" chemicals.

The problem is the data availability. When a sufficiently large number of new (i.e. obtained after the model development) and reliable experimental data is available, the best proof of model accuracy is by testing the model performance on these additional data, at the same time checking the chemical domain of applicability. This is the best way of external validation, performed after the model development and this is the way recommended by ECVAM for other alternative methods.
But, in the absence of available additional data (in useful quantity and quality), statistical external validation can be usefully applied to more precisely define the actual predictive power of the model, which is done by adequately splitting the available input data set into training set (for model development) and validation set (for model predictive assessment) by experimental design and other procedures. In this contract, specifically devoted to the evaluation of statistical approaches to QSAR validation, only statistical external validation was applied.

Statistical external validation is needed to determine both the generalizability of QSAR models for new chemicals and the "realistic" predictive power of a QSPR model. The model must be tested on a sufficiently large number (not just one or two, at least 20% of the complete data set is recommended) of chemicals not used in the QSAR model development. It is the performance accuracy of the model on this validation set that determines the actual predictive power of a QSAR model. Without this kind of validation there can be no confidence in the model predictivity for new external chemicals.
But, to be really useful, a statistical external validation can be performed during the model development (Tropsha et al., 2003). This approach was applied *a priori* in Gramatica's models, studied in this exercise: this is the reason why the verified models of Gramatica do not contain the shortcomings found in the other published models, for which the authors did not verify the parameters which were applied in this exercise.
In this contract work we applied *a posteriori* this splitting on already developed models, thus this kind of "statistical external validation" is slightly different, being influenced by the nature of all the chemicals which have indeed participated in the original model development and the selection of descriptors.
"Statistical external validations" were made only when the dimension of the data sets was reasonably large.

In this context the **methodology of splitting** available data is crucial (here we applied D-optimal distance, Kohonen Map-ANN and random), in fact the distribution of chemicals into training and validation sets can give different apparent predictivity of the studied model.
In this case, the worst method is random solitting: highly variable and unreliable results were obtained in this exercise. In fact, splitting by a random selection of the chemicals into two sets, while useful in splitting for internal validation as applied iteratively, gives very variable results when applied in external validation , strongly depending on the dimension and representativity of the sets.
The best splitting must guarantee that the training and validation sets are scattered over the whole area occupied by representative points in the descriptor space (representativity) and that the training set is distributed over the area occupied by representative points for the whole data set (diversity). Experimental Design by D-optimal distance design, selecting the most dissimilar chemicals in the training set, would be the best to satisfy this diversity condition, but in this case predictive ability estimates are over-optimistic, the validation chemicals all being included in the training domain. Splitting by Kohonen Map-Artificial Neural Network, which guarantees the most balanced structural diversity in both sets, highlights, in general, model performances that are more realistic and similar to "true predictivity".
Accurate estimation of the validity of a predictive regression model is especially problematic when the data set is small: in fact, for large data sets the methods were revealed to differ only slightly, whereas when the different splittings were applied to small data sets the statistical external validation results were often contradictory and too dependent on the nature of chemicals put into the training or validation sets.
Obviously, when too few experimental data are available the only form of statistical validation is internal validation. In such a case different internal validation techniques must be combined and compared.

Chemical domain of applicability: No model can be expected to extrapolate successfully, yet it is not always obvious which predictions are extrapolations and which are interpolations. Simple single-variable range checks cannot do this reliably. Measures analogous to leverage, the technique applied here for Ordinary Least Squares (OLS) models, are important in multiple predictor problems.
Here we applied, as we normally do in our modelling, and we recommend it for general use, an easy and fast inspection of the chemical domain of the OLS model: the Williams plot of regression.
Williams plot or OLS Outlier and Leverage Plot is the plot of jackknifed residuals versus leverages (hat diagonals), it allows a graphical detection of both the outliers and the influential chemicals in a model, in fact, in the present plot the horizontal straight lines (s) and the vertical (H) indicate the limits of normal values for outliers and influential chemicals respectively.

Based on my experience and this exercise, my main conclusions are:

a) **Internal validation** is necessary to guarantee the predictivity of the model for chemicals used in model development (training set) and to exclude pure-fitting models

b) **Cross-validation by LOO** is too over-optimistic in predictive ability estimate.
Moreover, if the difference between Q^2_{LOO} and R^2 is high (more than 20-30%), the model is not stable or robust and is probably a fitting model, able only to fit the training data.

c) **Cross-validation by LMO** counteracts the above optimism and must support the CV-LOO to guarantee the stability and robustness of the model.
When dealing with small data sets no more than 30% of chemicals must be put into the internal test set to maintain the necessary information in the training.
CV-LMO 40-50% give under-optimistic results.

d) **Bootstrapping** is the more balanced and, probably, more "realistic" internal validation technique.

e) **Y-scrambling** verifies that the model is not by chance-correlation.

f) Comparison of the **SDEP** and **SDEC** values gives information regarding differing abilities to calculate the response of a chemical in the two different situations of use: for model development (SDEC) or for checking (by LOO) internal predictivity (SDEP).

g) **Collinearity** among the variables (Kxx) and between the variable block and the response (Kxy) must be verified: the difference (Δ) must be sufficiently high to have robust models.

h) **Internal validation is necessary but not sufficient** as, *per se,* it gives no guarantee with regard to generalizability. No information can be obtained in relation to predictivity for new chemicals ("unknown" during the model development) thus such validation gives no assurance of the predictive power of the model.

i) The main utility of **statistical external validation** is during the development of QSAR models. In fact, in a population of several possible models (as is normal in Variable Selection-based modelling),

some apparently stable and "predictive" models, when verified by internal validation parameters, can be externally less-, or even completely not-predictive, and thus not useful for model generalization.

j) **Statistical external validation** is possible by splitting the available input data set into a training set and a validation set. In this context the **methodology of splitting** available data is crucial, the distribution of chemicals into training and validation sets can give different apparent predictivity of the studied model (generally random splitting must be avoided, unless thousand of chemicals are available).

k) Although external validation is commonly needed, the results of statistical external validation, when applied to **small data sets and after model development**, are not always unambiguous and trustworthy and such results depend strongly on training/validation composition. Different parameters must be considered together (for instance, Q^2_{EXT}, MSE and R^2 for validation set). But, it is important to highlight that great care must be done in generalizing QSAR models developed on small data sets.

10. REFERENCES

Atkinson, A.C. *Plots, Transformations and Regression.* **1985,** Clarendon Press: Oxford.

Efron, B., Tibshirani, R.J., *An Introduction to the Bootstrap*, **1993**, Chapman & Hall, New York.

Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.; McDowell, R.M.; Gramatica, P. Methods for Reliability, Uncertainty Assessment, and Applicability Evaluations of Regression Based and Classification QSARs. *Environ. Health Perspect.* **2003**, *111* (10), 1361-1375.

Eriksson, L.; Johansson, E. Multivariate Design and Modeling in QSAR. Tutorial. *Chemom. Int. Lab. Syst.* **1996**, *34*, 1-19.

Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503-527.

Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational Selection of Training Sets for the Development of Validated QSAR Models. *J. Comput. Aided Mol. Des.* **2003**, *17*, 241-253.

Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269-276.

Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training Set Selection. *J. Comput. Aided Mol. Des.* **2002**, *16*, 357-369.

Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric Degradability: Splitting into Training-Test Set and Consensus Modeling, *J. Chem.Inf. Comput.Sci*., **2004,** in press.

Hawkins, D.M. The Problem of Overfitting, *J. Chem.Inf. Comput.Sci*., **2004**, *44*, 1-12.

Marengo, E.; Todeschini, R. A New Algorithm for Optimal Distance – Based Experimental Design. *Chemom. Int. Lab. Syst.* **1992**, *16*, 37-44.

Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486-494.

Sjostrom, M.; Eriksson, L. Applications of Statistical Experimental Design, in *Chemometric Methods in Molecular Design*. Van de Warerbeemd H., VCH, **1995**; Vol. 2, pp 63-90.

Tropsha, A.; Gramatica, P.; Gombar, V.K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Comb. Sci.* **2003**, *22*, 69-76.

Todeschini, R.;Maiocchi, A.; Consonni, V. The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemom. Int. Lab. Syst.* **1999**, *46*, 13-29.

Wehrens, R.; Putter, H.; Buydens, L.M.C., The Bootstrapp: A Tutorial. *Chemom. Int. Lab. Syst.* **2000***, 54*, 35-52.

Zupan, J.; Novic, M.; Ruisánchez, I. Kohonen and Counter propagation Artificial Neural Networks in Analytical Chemistry. *Chemom. Int. Lab. Syst.* **1997**, *38*, 1-23.

11. ANNEXES

Koc data set (Wei et al. 2003)- Chapter 4

ID	*Chemicals*	CAS	β	α	V_{CSE}	Ov	Xc	logKow	exp. (logKoc)	pred. (logKoc)
1	1,2,3-trichlorobenzene	87-61-6	0.04	0	113.18	1.25	0.66	4.13	3.161	3.118
2	1,3-dichloro-2-fluorobenzene	2268-05-5	0.02	0.08	113.01	1.25	0.66	3.78	2.531	2.582
3	1,3-dichloro-4-fluorobenzene	1435-48-9	0.02	0.08	103.02	1.25	0.76	3.45	2.681	2.704
4	2,3-dichloro-bromobenzene		0.03	0.1	118.94	1.25	0.66	4.64	2.97	2.896
5	2,6-dichloro-bromobenzene		0.03	0.1	118.88	1.25	0.66	4.12	2.933	2.947
6	4-choro-iodobenzene	637-87-6	0.07	0	111.53	1.25	0.58	4.12	2.764	2.884
7	1,2-dchloro-3-iodobenzene	2401-21-0	0.03	0	125.05	1.25	0.66	4.84	3.353	3.326
8	1-cloro-3-nitrobenzene	121-73-3	0.31	0.16	104.49	1.27	0.79	2.49	2.863	2.752
9	1-cloro-2,4-dinitrobenzene	97-00-7	0.5	0.32	124.27	1.3	1.21	2.18	3.249	3.263
10	3,4-Dichloronitrobenzene	99-54-7	0.27	0.16	118.42	1.28	0.97	3.29	3.134	3.17
11	2,3-dichloroaniline	608-27-5	0.45	0.26	105.86	1.24	0.66	2.74	2.828	2.841
12	2,4-dichloroaniline	554-00-7	0.45	0.26	106.14	1.26	0.76	2.76	2.856	2.841
13	2,5-dichloroaniline	95-82-9	0.45	0.26	106.14	1.26	0.76	2.75	2.856	2.841
14	3,4-dichloroaniline	95-76-1	0.45	0.26	106.04	1.25	0.76	2.55	2.864	2.891
15	2-chloro-4-fluoroaniline	2106-02-7	0.44	0.34	95.6	1.25	0.76	2.04	2.488	2.461
16	2-chloro-4-nitroaniline	121-87-9	0.68	0.42	110.98	1.28	0.97	2.14	2.927	2.88
17	2-chloro-5-nitroaniline	6283-25-6	0.68	0.42	111.07	1.28	0.97	2.23	3.014	2.882
18	4-chloro-2-nitroaniline	89-63-4	0.68	0.42	110.7	1.28	0.99	2.56	2.952	2.933
19	4-chloro-3-nitroaniline	635-22-3	0.68	0.42	111.95	1.28	0.99	2.09	2.91	2.961
20	2,6-dichloro-4-nitroaniline	99-30-9	0.65	0.42	124.99	1.29	1.16	2.84	3.228	3.282
21	2,4-dichlorophenol	120-83-2	0.3	0.6	103.77	1.25	0.76	2.9	1.587	1.584
22	2-amino-4-chlorophenol	95-85-2	0.71	0.32	96.39	1.25	0.76	1.79	2.876	2.934
23	4-chloro-2-nitrophenol	89-64-5	0.8	0.28	108.41	1.27	0.54	2.48	2.603	2.711
24	pentachlorophenol	87-86-5	0.27	0.6	114.91	1.28	1.16	5.04	2.805	2.802
25	4-chlorobenzaldehyde	104-88-1	0.53	0	118.42	1.28	0.49	2.16	3.295	3.183
26	2-chlorobenzamide	609-66-5	0.75	0.49	107.98	1.26	0.71	0.64	2.476	2.459
27	4-chlorobenzonitrile	623-03-0	0.47	0.22	98.52	1.27	0.49	2.24	2.169	2.093
28	3,4-dichlorobenzonitrile	6574-99-8	0.44	0.22	112.47	1.28	0.68	2.98	2.408	2.553

BCF data set (Gramatica and Papa 2003)– Chapter 5

ID	CAS	*Chemicals*	BCF Exp.	$^{V}I^{M}$	MATS2m	GATS2e	H6p	nHAcc
1	107-06-2	1,2-Dichlorethane	0.3	1.97	0.99	2.33	0	0
2	67-66-3	Trichloromethane	0.78	1.97	0.97	1.97	0	0
3	79-34-5	1,1,2,2-Tetrachlorethane	0.9	2.56	0.99	2.33	0	0
4	79-01-6	Trichloroethylene	1.59	2.29	1.00	2.15	0	0
5	71-55-6	1,1,1-Trichloroethane	0.95	2.29	1.00	0.75	0	0
6	127-18-4	Tetrachloroethylene	1.74	2.56	1.00	2.5	0	0
7	56-23-5	Tetrachloromethane	1.48	2.29	0.97	0	0	0
8	76-01-7	Pentachloroethane	1.83	2.78	1.00	2.14	0	0
9	67-72-1	Hexachloroethane	2.92	2.98	1.00	2.33	0	0
10	87-68-3	1,1,2,3,4,4-Hexachloro-1,3-Butadiene	3.83	3.29	1.01	2.5	0	0
11	71-43-2	Benzene	0.64	2.59	0.94	2.44	0	0
12	108-88-3	Toluene	1.12	2.8	0.95	2.03	0	0
13	100-42-5	Styrene	1.13	2.98	0.94	2.19	0.01	0
14	100-41-4	Ethylbenzene	1.19	2.98	0.95	2.17	0.01	0
15	95-47-6	o-Xylene	1.24	2.98	0.95	1.79	0	0
16	108-38-3	m-Xylene	1.27	2.98	0.95	1.79	0.05	0
17	106-42-3	p-Xylene	1.27	2.98	0.95	1.79	0	0
18	622-97-9	p-Methylstirene	1.5	3.14	0.95	1.9	0.01	0
19	100-80-1	m- Methylstirene	1.55	3.15	0.95	1.9	0.01	0
20	98-82-8	Isopropylbenzene	1.55	3.15	0.94	2.05	0.01	0
21	2719-61-1	2-Phenyldodecane	2.65	4.14	0.93	2.55	0.21	0
22	29082-74-4	Octachlorostyrene	4.52	3.97	1.03	2.19	0.08	0
23	91-20-3	Naphtalene	1.64	3.31	0.95	2.04	0	0
24	208-96-8	Acenaphtylene	2.58	3.57	0.95	1.76	0	0
25	83-32-9	Acenaphtalene	2.59	3.57	0.95	1.83	0	0
26	92-52-4	Biphenyl	2.64	3.56	0.94	2.14	0	0
27	120-12-7	Anthracene	2.83	3.79	0.95	1.88	0.02	0
28	91-57-6	2-Metylnaphtalene	3.2	3.44	0.95	1.8	0.01	0
29	86-73-7	Fluorene	3.23	3.68	0.95	1.85	0.02	0
30	85-01-8	Phenanthrene	3.42	3.79	0.95	1.88	0.02	0
31	50-32-8	Benzo[a]pyrene	3.42	4.3	0.95	1.65	0.05	0
32	129-00-0	Pyrene	3.43	3.99	0.95	1.69	0	0
33	2531-84-2	2-Methylphenanthrene	3.48	3.88	0.95	1.71	0.04	0
34	24423-11-8	2-Chlorophenanthrene	3.63	3.88	0.97	1.12	0.02	0
35	779-02-2	9-Methylanthracene	3.66	3.89	0.95	1.71	0.03	0
36	56-55-3	Benzo[a]anthracene	4	4.15	0.95	1.79	0.05	0
37	108-90-7	Chlorobenzene	1.85	2.8	0.98	1.22	0	0
38	95-50-1	1,2-Dichlorobenzene	2.48	2.98	0.99	1.29	0	0
39	106-46-7	1,4-Dichlorobenzene	2.52	2.98	0.99	1.29	0	0
40	541-73-1	1,3-Dichlorobenzene	2.65	2.98	0.99	1.29	0	0
41	87-61-6	1,2,3-Trichlorobenzene	3.11	3.15	1.01	1.44	0	0
42	120-82-1	1,2,4-Trichlorobenzene	3.26	3.15	1.01	1.44	0	0
43	634-90-2	1,2,3,5 -Tetrachlorobenzene	3.36	3.3	1.02	1.65	0	0
44	108-70-3	1,3,5-Trichlorobenzene	3.38	3.15	1.01	1.44	0	0
45	95-94-3	1,2,4,5-Tetrachlorobenzene	3.76	3.3	1.02	1.65	0	0
46	634-66-2	1,2,3,4-Tetrachlorobenzene	3.77	3.3	1.02	1.65	0	0
47	608-93-5	Pentachlorobenzene	3.86	3.44	1.02	1.97	0	0

48	6639-30-1	2,4,5-Trichlorotoluene	3.87	3.3	1.01	1.24	0	0
49	118-74-1	Hexachlorobenzene	4.26	3.56	1.03	2.44	0	0
50	108-86-1	Bromobenzene	1.7	2.8	0.98	1.25	0	0
51	108-36-1	1,3-Dibromobenzene	2.8	2.98	1.01	1.24	0	0
52	106-37-6	1,4-Dibromobenzene	2.83	2.98	1.01	1.24	0	0
53	87-82-1	Hexabromobenzene	3.04	3.56	1.04	2.44	0	0
54	583-53-9	1,2-Dibromobenzene	3.1	2.98	1.01	1.24	0	0
55	615-54-3	1,2,4-Tribromobenzene	3.66	3.15	1.02	1.36	0	0
56	636-28-2	1,2,4,5-Tetrabromobenzene	3.79	3.3	1.03	1.57	0	0
57	626-39-1	1,3,5-Tribromobenzene	3.85	3.15	1.02	1.36	0	0
58	2234-13-1	Octachloronaphtalene	3.44	4.15	1.03	2.04	0	0
59	1825-31-6	1,4-Dichloronaphtalene	3.56	3.57	0.99	1.15	0	0
60	91-58-7	2-Monochloronaphtalene	3.63	3.44	0.97	1.14	0	0
61	2050-74-0	1,8-Dichloronaphtalene	3.79	3.57	0.99	1.15	0	0
62	2050-75-1	2,3-Dichloronaphtalene	4.04	3.57	0.99	1.15	0	0
63	2198-77-8	2,7-Dichloronaphtalene	4.04	3.56	0.99	1.15	0	0
64	20020-02-4	1,2,3,4-Tetrachloronaphtalene	4.1	3.79	1.01	1.31	0	0
65	31604-28-1	1,3,5,8-Tetrachloronaphtalene	4.4	3.79	1.01	1.31	0	0
66	55720-37-1	1,3,7-Trichloronaphtalene	4.43	3.68	1.00	1.21	0	0
67	53555-64-9	1,3,5,7-Tetrachloronaphtalene	4.53	3.79	1.01	1.31	0	0
68	2051-62-9	4-Chlorobiphenyl	2.69	3.68	0.97	1.2	0	0
69	13029-08-8	2,2'-Dichlorobiphenyl	3.26	3.79	0.98	1.17	0	0
70	2050-68-2	4,4'-Dichlorobiphenyl	3.28	3.78	0.98	1.17	0	0
71	39485-83-1	2,2',4,4',6-Pentachlorobiphenyl	3.37	4.06	1.01	1.35	0	0
72	33284-50-3	2,4'-Dichlorobiphenyl	3.55	3.78	0.98	1.17	0	0
73	16606-02-3	2,4',5-Trichlorobiphenyl	3.75	3.88	0.99	1.21	0.11	0
74	34883-41-5	3,5-Dichlorobiphenyl	3.78	3.78	0.98	1.17	0	0
75	15968-05-5	2,2',6,6'-Tetrachlorobiphenyl	3.85	3.98	1.00	1.27	0	0
76	32598-13-3	3,3',4,4'-Tetrachlorobiphenyl	3.9	3.97	1.00	1.27	0	0
77	2437-79-8	2,2',4,4'-Tetrachlorobiphenyl	4.02	3.97	1.00	1.27	0	0
78	15862-07-4	2,4,5-Trichlorobiphenyl	4.02	3.88	0.99	1.21	0.11	0
79	2051-24-3	2,2',3,3',4,4',5,5',6,6'-Decachlorobiphenyl	4.02	4.44	1.03	2.14	0	0
80	34883-39-1	2,5-Dichlorobiphenyl	4.2	3.78	0.98	1.17	0	0
81	38444-93-8	2,2',3,3'-Tetrachlorobiphenyl	4.23	3.98	1.00	1.27	0	0
82	16605-91-7	2,3-Dichlorobiphenyl	4.25	3.79	0.98	1.17	0	0
83	37680-65-2	2,2',5-Trichlorobiphenyl	4.27	3.88	0.99	1.21	0	0
84	7012-37-5	2,4,4'-Trichlorobiphenyl	4.63	3.88	0.99	1.21	0.08	0
85	32598-11-1	2,3',4',5-Tetrachlorobiphenyl	4.77	3.98	1.00	1.27	0	0
86	35065-27-1	2,2',4,4',5,5'-Hexachlorobiphenyl	4.83	4.15	1.01	1.45	0	0
87	41464-39-5	2,2',3,5'-Tetrachlorobiphenyl	4.84	3.98	1.00	1.27	0	0
88	41464-40-8	2,2',4,5'-Tetrachlorobiphenyl	4.84	3.98	1.00	1.27	0	0
89	35693-99-3	2,2',5,5'-Tetrachlorobiphenyl	4.87	3.98	1.00	1.27	0	0
90	33979-03-2	2,2',4,4',6,6'-Hexachlorobiphenyl	4.93	4.14	1.01	1.45	0	0
91	70362-47-9	2,2',4,5-Tetrachlorobiphenyl	5	3.98	1.00	1.27	0	0
92	35694-08-7	2,2',3,3',4,4',5,5'-Octachlorobiphenyl	5.08	4.3	1.02	1.72	0	0
93	38380-02-8	2,2',3,4,5'-pentachlorobiphenyl	5.38	4.06	1.01	1.35	0	0
94	37680-73-2	2,2',4,5,5'-Pentachlorobiphenyl	5.4	4.06	1.01	1.35	0	0
95	41464-51-1	2,2',3',4,5-Pentachlorobiphenyl	5.43	4.06	1.01	1.35	0	0
96	38411-22-2	2,2',3,3',6,6'-Hexachlorobiphenyl	5.43	4.15	1.01	1.45	0	0
97	52744-13-5	2,2',3,5,5',6'-Hexachlorobiphenyl	5.54	4.15	1.01	1.45	0	0
98	40186-72-9	2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl	5.71	4.37	1.03	1.91	0	0
99	38380-07-3	2,2',3,3',4,4'-Hexachlorobiphenyl	5.77	4.15	1.01	1.45	0	0

100	57465-28-8	3,3',4,4',5-Pentachlorobiphenyl	5.81	4.06	1.01	1.35	0	0
101	52712-04-6	2,2',3,4,5,5'-Hexachlorobiphenyl	5.81	4.15	1.01	1.45	0	0
102	2136-99-4	2,2',3,3',5,5',6,6'-Octachlorobiphenyl	5.82	4.3	1.02	1.72	0	0
103	52663-69-1	2,2',3,4,4',5',6-Heptechlorobiphenyl	5.84	4.22	1.02	1.57	0	0
104	35694-06-5	2,2',3,4,4',5-Hexachlorobiphenyl	5.88	4.14	1.01	1.45	0	0
105	68194-17-2	2,2',3,3',4,5,5',6-Octachlorobiphenyl	5.88	4.3	1.02	1.72	0	0
106	52663-78-2	2,2',3,3',4,4',5,6-Octachlorobiphenyl	5.92	4.3	1.02	1.72	0	0
107	38411-25-5	2,2',3,4,5,5',6'-Heptachlorobiphenyl	5.93	4.23	1.02	1.57	0.34	0
108	32774-16-6	3,3',4,4',5,5'-Hexachlorobiphenyl	5.97	4.15	1.01	1.45	0	0
109	59080-33-0	2,4,6-Tribromobiphenyl	3.93	3.88	1.01	1.17	0.17	0
110	59261-08-4	2,2',4,4',6,6'-Hexabromobiphenyl	3.96	4.14	1.03	1.38	0.19	0
111	92-86-4	4,4'-Dibromobiphenyl	4.19	3.78	0.99	1.17	0.04	0
112	59080-37-4	2,2',5,5'-Tetrabromobiphenyl	4.8	3.98	1.02	1.22	0.19	0
113	33857-26-0	2,7-Dichlorodobenzo-p-dioxin	2.13	3.98	0.99	1.91	0.03	2
114	39227-58-2	1,2,4-Trichlorodibenzo-p-dioxin	2.36	4.07	1.00	1.92	0.11	2
115	30746-58-8	1,2,3,4-tetrachlorodibenzo-p-dioxin	2.55	4.15	1.01	1.97	0.11	2
116	3268-87-9	Octachlorodibenzo-p-dioxin	2.76	4.44	1.03	2.57	0.39	2
117	38964-22-6	2,8-Dichlorodibenzo-p-dioxin	2.82	3.98	0.99	1.91	0.03	2
118	35822-46-9	1,2,3,4,6,7,8-Heptachlorodibenzo-p-dioxin	3.16	4.37	1.02	2.35	0.22	2
119	39227-61-7	1,2,3,4,7-Pentachlorodibenzo-p-dioxin	3.21	4.22	1.01	2.06	0.11	2
120	67028-18-6	1,2,3,7-Tetrachlorodibenzo-p-dioxin	3.24	4.15	1.01	1.97	0.07	2
121	33423-92-6	1,3,6,8-tetrachlorodibenzo-p-dioxin	3.36	4.15	1.01	1.97	0.11	2
122	39227-28-6	1,2,3,4,7,8-Hexachlorodibenzo-p-dioxin	3.54	4.3	1.02	2.18	0.1	2
123	262-12-4	Dibenzo(1,4)dioxan	3.85	3.79	0.96	2.11	0.03	2
124	1746-01-6	2,3,7,8-Tetrachlorodibenzo-p-dioxin	4.06	4.15	1.01	1.97	0.03	2
125	40321-76-4	1,2,3,7,8-Pentachlorodibenzo-p-dioxin	4.5	4.22	1.01	2.06	0.07	2
126	271-89-6	Benzo[b]furan	2.56	3.16	0.95	2.24	0	1
127	39001-02-0	Octachlordibenzofuran	2.94	4.37	1.03	2.19	0.36	1
128	00132-64-9	Dibenzofuran	3.34	3.68	0.95	1.97	0.03	1
129	51207-31-9	2,3,7,8-Tetrachlorodibenzofuran	3.53	4.06	1.01	1.64	0.04	1
130	67562-39-4	1,2,3,4,6,7,8-Heptachlorodibenzofuran	3.62	4.3	1.03	1.99	0.32	1
131	57117-31-4	2,3,4,7,8-Pentachlorodibenzofuran	4.03	4.15	1.02	1.8	0.07	1
132	767-00-0	4-Cyanophenol	0.91	3.14	0.97	1.2	0	2
133	95-48-7	2-Methyl phenol	1.03	2.98	0.96	1.15	0	1
134	108-95-2	Phenol	1.24	2.8	0.95	1.25	0	1
135	108-43-0	3-Clorophenol	1.25	2.98	0.98	1.3	0	1
136	120-83-2	2,4-Dichlorophenol	1.5	3.15	0.99	1.43	0	1
137	106-41-2	4-Bromophenol	1.56	2.98	0.98	1.28	0	1
138	99-71-8	p-sec-Butyl phenol	1.57	3.43	0.95	1.11	0.03	1
139	123-31-9	Hydroquinone	1.6	2.98	0.95	1.33	0	2
140	1689-84-5	2,6-Dibromo-4-Cyanophenol	1.67	3.43	1.01	1.43	0	2
141	16766-31-7	4,6-Dichloroguaiacol	1.74	3.43	0.99	2.17	0.01	2
142	98-54-4	4-t-Butyl phenol	1.86	3.43	0.95	1.06	0.03	1
143	2668-24-8	4,5,6-Trichloroguaiacol	1.97	3.56	1.00	2.23	0.01	2
144	2460-49-3	4,5-Dichloroguaiacol	2.03	3.43	0.99	2.17	0.01	2
145	935-95-5	2,3,5,6-Tetrachlorophenol	2.15	3.44	1.01	1.84	0	1
146	105-67-9	2,4-Dimethylphenol	2.18	3.15	0.96	1.09	0.05	1
147	95-57-8	2-Chlorophenol	2.33	2.98	0.98	1.3	0	1
148	57057-83-7	3,4,5-Trichloroguaiacol	2.41	3.56	1.00	2.23	0.01	2
149	88-06-2	2,4,6-Trichlorophenol	2.43	3.3	1.00	1.6	0	1
150	104-40-5	p-Nonyl phenol	2.45	3.97	0.94	1.2	0.23	1
151	2539-17-5	Tetrachloroguaiacol	2.71	3.68	1.01	2.35	0.01	2

152	118-79-6	2,4,6-Tribromophenol	2.71	3.3	1.02	1.51	0	1
153	87-86-5	Pentachlorophenol	2.74	3.56	1.02	2.19	0	1
154	104-43-8	p-Dodecyl phenol	3.78	4.22	0.94	1.25	0.33	1
155	106-47-8	4-Chloroaniline	0.23	2.98	0.97	1.21	0	1
156	108-42-9	3-chloroaniline	0.34	2.98	0.97	1.21	0	1
157	062-53-3	Aniline	0.41	2.8	0.95	1.26	0	1
158	95-51-2	2-chloroaniline	0.57	2.98	0.97	1.21	0	1
159	122-39-4	Diphenylamine	1.48	3.68	0.94	1.97	0.07	1
160	95-76-1	3,4-Dichloroaniline	1.48	3.15	0.99	1.3	0	1
161	554-00-7	2,4-Dichloroaniline	1.98	3.15	0.99	1.3	0	1
162	135-88-6	N-phenyl-2-naphtylamine	2.17	4.06	0.95	1.88	0.06	1
163	634-67-3	2,3,4-Trichloroaniline	2.31	3.3	1.00	1.45	0	1
164	636-30-6	2,4,5-trichloroaniline	2.61	3.3	1.00	1.45	0	1
165	634-83-3	2,3,4,5-Tetrachloroaniline	2.69	3.44	1.01	1.64	0	1
166	634-91-3	3,4,5-trichloroaniline	2.7	3.3	1.00	1.45	0	1
167	634-93-5	2,4,6-trichloroaniline	2.73	3.3	1.00	1.45	0	1
168	91-94-1	3,3'-Dichlorobenzidine	2.79	3.97	0.98	1.18	0	2
169	3481-20-7	2,3,5,6-tetrachloroaniline	3.03	3.44	1.01	1.64	0	1
170	527-20-8	Pentachloroaniline	3.17	3.56	1.02	1.9	0	1
171	141-78-6	Ethyl acetate	1.48	2.55	0.98	1.6	0.01	2
172	131-11-3	Dimethyl phtalate	1.76	3.78	0.98	1.85	0.03	4
173	84-66-2	Diethyl phtalate	2.07	3.97	0.97	1.61	0.03	4
174	117-81-7	Bis(2-ethylhexyl)phtalate	2.34	4.77	0.95	1.41	0.34	4
175	52918-63-5	Deltamethrin	2.66	4.73	0.99	1.2	0.54	4
176	51630-58-1	Fenvalerate	2.79	4.83	0.97	1.26	0.31	4
177	85-68-7	Benzyl butyl phtalate	2.89	4.49	0.96	1.54	0.12	4
178	52315-07-8	Cypermethrin	2.91	4.78	0.98	1.35	0.45	4
179	52645-53-1	Permethrin	3.39	4.67	0.97	1.4	0.41	3
180	7580-85-0	2-t-Butoxy ethanol	-0.22	2.97	0.95	2.51	0.04	2
181	1634-04-4	t-Butyl methyl ether	0.18	2.55	0.95	3.14	0	1
182	17348-59-3	t-Butyl isopropyl ether	0.76	2.97	0.94	2.74	0.11	1
183	111-44-4	Bis(2-chloroethyl)ether	1.04	2.77	0.98	2.96	0.05	1
184	87-40-1	2,4,6-Trichloroanisole	2.94	3.43	1.01	2.27	0	1
185	607-99-8	2,4,6-Tribromoanisole	2.94	3.43	1.02	2.65	0	1
186	72-43-5	methoxychlor	3.1	4.36	0.97	1.94	0.19	2
187	52322-80-2	2,4,5-Trichlorodiphenyl ether	4.18	3.98	0.99	1.68	0.17	1
188	56348-72-2	3,3',4,4'-tetrachlorodiphenyl ether	4.51	4.06	1.00	1.71	0.11	1
189	534-52-1	2-Methyl-4,6-Dinitrophenol	0.16	3.78	0.99	1	0.01	7
190	100-01-6	4-Nitroaniline	0.64	3.3	0.97	0.84	0	4
191	88-74-4	2-Nitroaniline	0.91	3.3	0.97	0.84	0	4
192	99-09-2	3-Nitroaniline	0.92	3.3	0.97	0.84	0.01	4
193	554-84-7	3-Nitrophenol	1.4	3.3	0.98	1.03	0	4
194	88-75-5	2-Nitrophenol	1.6	3.3	0.98	1.03	0	4
195	89-69-0	2,4,5-Trichloronitrobenzene	1.84	3.56	1.01	1.46	0	3
196	121-73-3	3-Chloronitrobenzene	1.89	3.3	0.99	0.99	0	3
197	879-39-0	2,3,4,5-Tetrachloronitrobenzene	1.89	3.68	1.02	1.79	0	3
198	100-00-5	4-Chloronitrobenzene	2	3.3	0.99	0.99	0	3
199	89-61-2	2,5-Dichloronitrobenzene	2.05	3.44	1.00	1.2	0	3
200	611-06-3	2,4-Dichloronitrobenzene	2.07	3.43	1.00	1.2	0	3
201	99-54-7	3,4-Dichloronitrobenzene	2.07	3.43	1.00	1.2	0	3
202	88-73-3	2-chloronitrobenzene	2.1	3.3	0.99	0.99	0	3
203	3209-22-1	2,3-Dichloronitrobenzene	2.16	3.44	1.00	1.2	0	3

204	17700-09-3	2,3,4-Trichloronitrobenzene	2.2	3.56	1.01	1.46	0	3
205	618-62-2	3,5-Dichloronitrobenzene	2.23	3.44	1.00	1.2	0	3
206	82-68-8	Pentachloronitrobenzene	2.4	3.78	1.02	2.27	0	3
207	18708-70-8	2,4,6-Trichloronitrobenzene	2.88	3.56	1.01	1.46	0	3
208	1836-77-7	Chloronitrofen	3.04	4.22	1.00	1.46	0.24	4
209	117-18-0	2,3,5,6-Tetrachloronitrobenzene	3.2	3.68	1.02	1.79	0	3
210	13286-32-3	IBP	0.97	4.14	0.95	1.47	0.09	3
211	2597-03-7	Phenthoate	1.56	4.22	0.96	2.15	0.24	4
212	333-41-5	Diazinon	1.8	4.22	0.97	1.65	0.06	5
213	122-14-5	Fenitrothion	2	4.06	0.98	1.64	0.03	6
214	25311-71-1	Isofenphos	2.17	4.43	0.96	1.63	0.12	5
215	298-04-4	Disulfoton	2.37	3.77	0.95	2.34	0.04	2
216	55-38-9	Fenthion	2.68	3.97	0.97	2.09	0.12	3
217	2104-64-5	EPN	3.05	4.36	0.97	1.54	0.12	5
218	2921-88-2	Chloropyriphos	3.18	4.14	1.00	1.68	0.01	4
219	21609-90-5	Leptophos	3.78	4.3	0.99	2.02	0.15	2
220	10075-50-0	5-Bromoindole	1.15	3.31	0.98	1.64	0	1
221	63-25-2	carbaryl	1.22	3.88	0.96	0.98	0.02	3
222	2212-67-1	Molinate	1.41	3.56	0.95	1.25	0.13	2
223	3766-81-2	BPMC	1.41	3.88	0.96	0.93	0.06	3
224	107-13-1	Acrylonitrile	1.68	1.97	1.00	0.93	0	1
225	28249-77-6	Thiobencarb	2.03	3.97	0.97	1.27	0.2	2
226	260-94-6	Acridine	2.61	3.79	0.95	1.8	0.03	1
227	58-89-9	Lindane	2.84	3.56	1.00	2.01	0	0
228	319-85-7	b-BHC	2.86	3.56	1.00	2.01	0	0
229	319-84-6	a-BHC	2.95	3.56	1.00	2.01	0	0
230	77-47-4	Hexachlorocyclopentadiene	3.09	3.43	1.02	2.44	0	0
231	92-83-1	Xanthene	3.62	3.79	0.95	1.93	0.02	1
232	60-57-1	Dieldrin	3.71	4.23	1.01	1.34	0.19	1
233	76-44-8	Heptachlor	4.14	4.06	1.01	1.4	0.03	0
234	1024-57-3	Heptachlor epoxide	4.16	4.15	1.02	1.64	0.04	1
235	789-02-6	o-p'-DDT	4.57	4.22	0.99	0.87	0.17	0
236	57-74-9	Chlordane	4.58	4.14	1.02	1.44	0.09	0
237	72-55-9	p,p'-DDE	4.71	4.14	0.99	0.93	0.08	0
238	50-29-3	p,p'-DDT	4.84	4.22	0.99	0.87	0.11	0

Benzenes ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^{1}X^{v}$	t-obs	t-pre[1]	t-pre[2]
1	Isopropylbenzene	98-82-8	0.55	0.15	0.00	3.66	3.35	1.28	1.38	1.38
2	1,2,4-Trimethylbenzene	95-63-6	0.55	0.15	0.00	3.78	3.24	1.19	1.46	1.46
3	t-Butylbenzene	98-06-6	0.55	0.15	0.00	4.26	3.66	1.83	1.78	1.78
4	Amylbenzene	538-68-1	0.53	0.17	0.00	4.91	4.36	1.94	2.20	2.20
5	Biphenyl	92-52-4	1.20	0.28	0.00	4.09	4.06	1.90	1.95	1.95
6	Chlorobenzene	108-90-7	0.71	0.10	0.00	2.86	2.51	0.82	0.95	0.96
7	1,2-Dichlorobenzene	095-50-1	0.64	0.10	0.00	3.38	3.03	1.19	1.26	1.27
8	1,2,4-Trichlorobenzene	120-82-1	0.84	0.06	0.00	4.02	3.54	1.78	1.82	1.82
9	1,2,4,5-Tetrachlorobenzene	95-94-3	0.85	0.01	0.00	4.82	4.06	2.80	2.39	2.38
10	3,4-Dichlorotoluene	95-75-0	0.75	0.09	0.00	4.22	3.44	1.74	1.89	1.89
11	Bromobenzene	108-86-1	0.79	0.06	0.10	2.99	2.92	0.94	1.11	1.11
12	Nitrobenzene	98-95-3	1.01	0.34	0.16	1.85	2.45	0.02	0.33	0.33
13	3-Nitrotoluene	99-08-1	0.17	0.35	0.16	2.45	2.85	0.73	0.28	0.29
14	1-Chloro-3-nitrobenzene	121-73-3	1.06	0.29	0.16	2.41	2.96	0.92	0.75	0.75
15	1-Chloro-2-nitrobenzene	88-73-3	1.06	0.29	0.16	2.24	2.96	0.73	0.64	0.64
16	2-Phenyl-3-butyn-2-ol	127-66-2	0.88	0.58	0.73	1.68	3.42	0.11	0.07	0.04
17	Pentachloroanisole	1825-21-4	0.84	0.26	0.06	5.34	5.12	2.64	2.61	2.60
18	Pentachlorobenzene	608-93-5	0.74	0.04	0.00	5.17	4.58	3.00	2.55	2.54
19	Naphthalene	91-20-3	0.70	0.20	0.00	3.30	3.40	1.32	1.19	1.20
20	1,4-Dimethoxybenzene	150-78-7	0.79	0.58	0.12	2.15	3.05	0.07	0.28	0.30
21	2,6-Dimethoxytoluene	5673-07-4	0.77	0.59	0.12	2.80	3.46	0.88	0.71	0.72
22	Benzyl-*t*-butanol		0.70	0.39	0.60	2.57	3.84	0.35	0.66	0.62
23	a,a,2,6-Tetrachlorotoluene	81-19-6	1.04	0.19	0.06	4.64	4.63	2.38	2.28	2.27
24	1,2-Xylene	95-47-6	0.51	0.16	0.00	3.12	2.83	0.81	0.99	1.00
25	Ethylbenzene	100-41-4	0.55	0.15	0.00	3.15	2.96	1.00	1.04	1.04
26	1,4-Xylene	106-42-3	0.51	0.16	0.00	3.15	2.82	1.08	1.02	1.02
27	1,3-Diethylbenzene	141-93-5	0.55	0.18	0.00	4.50	3.93	1.51	1.93	1.93
28	1-Fluoro-4-nitrobenzene	350-46-9	0.72	0.34	0.24	1.80	2.25	0.70	0.15	0.15
29	2-Tolunitrile	529-19-1	0.77	0.53	0.22	2.21	3.04	0.37	0.35	0.36
30	1,2-Dibromobenzene	583-53-9	0.63	0.10	0.00	3.64	3.81	1.77	1.43	1.44
31	2-Fluorotoluene	95-52-3	0.60	0.10	0.08	2.93	2.21	0.75	0.95	0.95
32	a,a,a',a'-Tetrabromo-2-xylene	13209-15-9	2.31	0.82	0.20	5.17	6.98	2.98	2.99	2.99
33*	1,4-Dinitrobenzene	100-25-4	0.79	0.64	0.32	1.46	2.89	2.37		
34*	2,4-Dinitrotoluene	121-14-2	0.75	0.65	0.32	2.00	3.31	0.88		
35*	1,3-Dichloro-4,6-dinitrobenzene	3698-83-7	0.89	0.54	0.32	2.65	1.45	3.71		
36*	1,3,5-Trichloro-2,4-dinitrobenzene	6284-83-9	0.94	0.59	0.32	2.65	4.45	3.09		
37*	a,a'-Dichloro-4-xylene		1.25	0.48	0.12	3.27	3.37	3.65		
38*	Pentachloropyridine	2176-62-7	1.02	0.33	0.00	4.34	4.58	2.73		
39[a]	Acenaphthene	83-32-9	0.70	0.20	0.00	3.92	4.44	1.98	1.61	1.62
40[a]	Benzene	71-43-2	0.59	0.14	0.00	2.13	2.00	0.35	0.37	0.37
41[a]	2,6-Dichlorobenzonitrile	1194-65-6	0.89	0.41	0.22	2.99	3.42	1.46	1.00	0.99

42[a]	Benzyl alcohol	100-51-6	0.99	0.61	0.33	1.10	2.58	-0.62	-0.31	-0.30
43[a]	Toluene	108-88-3	0.59	0.14	0.00	2.69	2.41	0.41	0.75	0.76
45[a]	Diphenylether	101-84-8	1.28	0.50	0.06	3.20	4.23	1.63	1.28	1.29
46[a]	2-Methylnitrobenzene	88-72-2	0.69	0.39	0.16	2.37	2.85	0.57	0.48	0.48
47[b]	4-Methyl-1,3-dinitrobenzene		0.79	0.64	0.32	2.01	3.05	0.75	0.18	0.18
48[b]	5-Methyl-1,3-dinitrobenzene		0.79	0.64	0.32	2.06	3.31	0.91	0.21	0.22
49[b]	2-Methyl-1,3-dinitrobenzene	606-20-2	0.79	0.64	0.32	2.01	3.32	0.99	0.18	0.18
50[b]	1,3-Dinitrobenzene	99-65-0	0.79	0.64	0.32	1.49	2.89	1.38	-0.16	-0.16
51[b]	4-Methyl-1,3,5-trinitrobenzene		0.89	0.89	0.48	1.65	3.76	1.88	-0.12	-0.12
52[b]	3-Methyl-1,2-dinitrobenzene		0.79	0.64	0.32	1.98	3.32	2.01	0.16	0.16
53[b]	5-Methyl-1,2-dinitrobenzene		0.79	0.64	0.32	1.98	3.31	2.08	0.16	0.16
54[b]	3-Methyl-1,4-dinitrobenzene		0.79	0.64	0.32	1.98	3.31	2.15	0.15	0.16
55[b]	1,3,5-Trinitrobenzene	99-35-4	0.89	0.89	0.48	1.13	3.34	2.29	-0.46	-0.46
56[b]	1,2-Dinitrobenzene	528-29-0	0.79	0.64	0.32	1.46	2.90	2.48	-0.18	-0.18
57[b]	3-Methyl-1,2,4-trinitrobenzene		0.89	0.89	0.48	1.59	3.76	3.37	-0.15	-0.16

***** Outlier
[a] Test set
[b] Test set

Alcohols ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^{1}X^{v}$	t-obs	t-pre[1]	t-pre[2]
1	1-Pentanol	71-41-0	0.40	0.47	0.33	1.56	2.52	-0.73	-0.63	-0.65
2	2-Methyl-1-propanol	78-83-1	0.40	0.47	0.33	0.76	1.87	-1.28	-1.28	-1.29
3	1-Hexanol	111-27-3	0.40	0.47	0.33	2.03	3.02	0.02	-0.25	-0.26
4	1-Heptanol	111-70-6	0.40	0.47	0.33	2.72	3.52	0.53	0.31	0.29
5	1-Octanol	111-87-5	0.40	0.47	0.33	2.97	4.02	0.98	0.51	0.49
6	2,2,2-Trichloroethanol	115-20-8	0.75	0.37	0.43	1.42	2.47	-0.30	0.58	0.75
7	1-Nonanol	143-08-8	0.40	0.47	0.33	4.26	4.52	1.40	1.56	1.54
8	Ethanol	64-17-5	0.40	0.47	0.33	-0.31	1.02	-2.50	-2.14	-2.16
9	2-Propanol	67-63-0	0.40	0.47	0.33	0.05	1.41	-2.15	-1.85	-1.87
10	1-Butanol	71-36-3	0.40	0.47	0.33	0.88	2.02	-1.37	-1.18	-1.20
11	2-Methyl-2-propanol	75-65-0	0.40	0.47	0.33	0.35	1.72	-1.94	-1.61	-1.63
12	2-Methyl-3-pentanol	565-67-3	0.40	0.47	0.33	1.53	2.84	-0.82	-0.65	-0.67
13	2-Butanol	78-92-2	0.40	0.47	0.33	0.61	1.95	-1.68	-1.40	-1.41
14	2-Diisopropylaminoethanol	96-80-0	1.04	1.85	0.33	0.86	3.87	-0.09	0.15	0.08
15	2-Ethyl-1-hexanol	104-76-7	0.40	0.47	0.33	2.81	3.69	0.66	0.38	0.36
16	Cyclohexanol	108-93-0	0.40	0.47	0.33	1.23	3.07	-0.85	-0.90	-0.91
17	2-(2-Ethoxyethoxy)ethanol	111-90-0	0.94	1.37	0.33	-0.54	2.00	-2.29	-1.58	-1.50
18	1-Decanol	112-30-1	0.40	0.47	0.33	4.57	5.02	1.82	1.81	1.79
19	2.4-Dimethyl-3-pentanol	600-36-2	0.40	0.47	0.33	1.93	2.97	-0.15	-0.33	-0.34
20	1,1,1-Trichloro-2-methyl-2-propanol	57-15-8	0.75	0.37	0.43	2.00	2.27	0.12	1.07	1.25
21	2-Ethylaminoethanol	110-73-6	0.76	1.06	0.33	-0.46	2.22	-1.22	-1.78	-1.73

22	2-Chloroethanol	107-07-3	0.75	0.62	0.39	0.03	1.66	0.18	-0.91	-0.78
23	1-Chloro-2-propanol	127-00-4	0.75	0.62	0.39	0.14	2.09	-0.41	-0.82	-0.69
24	2-Butyn-1-ol	764-01-2	0.60	0.67	0.46	0.16	1.42	0.84	0.69	0.50
25	3-Butyn-1-ol	927-74-2	0.60	0.67	0.46	-0.50	1.45	0.29	0.15	-0.03
26	4-Pentyn-2-ol	2117-11-5	0.60	0.67	0.46	-0.08	1.88	0.37	0.49	0.30
27	1-Hexen-3-ol	4798-44-1	0.50	0.57	0.38	1.12	2.61	0.52	-0.03	-0.10
28	1-Heptyn-3-ol	7383-19-9	0.60	0.67	0.46	1.52	2.95	1.80	1.79	1.60
29	2,3-Dibromopropanol	96-13-9	1.26	0.81	0.43	0.63	2.00	0.49	0.06	0.53
30*	3-Methyl-1-pentyn-3-ol	77-75-8	0.60	0.67	0.46	0.86	2.32	-1.08		
31*	3-Methyl butynol	115-19-5	0.60	0.67	0.46	0.28	1.76	-1.60		
32*	3,4-Dimethyl-1-pentyn-3-ol	1482-15-1	0.60	0.67	0.46	1.26	2.70	-0.26		
33*	3,6-Dimethyl-1-heptyn-3-ol	19549-98-5	0.60	0.67	0.46	2.32	3.76	0.45		
34*	1-propyne-3-ol	107-19-7	0.60	0.67	0.46	-0.37	0.95	1.57		
35**	Dodecyl alcohol	112-53-8	0.40	0.47	0.33	4.94	6.02	2.26	2.11	2.09

* Outlier
** Test set

Aldehydes ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^{1}X^{v}$	t-obs	t-pre[1]	t-pre[2]
1	4-(Hexyloxy)-3-anisaldehyde	61096-84-2	1.12	1.00	0.12	3.99	6.11	1.95	1.72	
2	4-Ethoxybenzaldehyde	10031-82-0	1.02	0.78	0.06	2.31	3.54	0.73	1.17	1.19
3	2-Chloro-5-nitrobenzaldehyde	6361-21-3	1.07	0.76	0.16	2.28	3.15	1.68	1.35	1.27
4	2,4,5-Trimethoxybenzaldehyde	4460-86-0	1.22	1.22	0.18	1.38	4.02	0.60	0.69	0.51
5	2,3-Dimethyl pentanaldehyde	32749-94-3	0.65	0.41	0.00	2.07	3.17	0.85	1.11	1.29
6	2,4-Dichlorobenzaldehyde	874-42-0	1.02	0.46	0.00	3.11	3.46	1.98	1.71	1.77
7	4.6-Dimethoxy-2-hydroxybenzaldehyde		1.25	1.23	0.72	2.33	3.63	1.83	1.76	1.41
8	Pentafluorobenzaldehyde	653-37-2	1.07	0.31	0.40	2.45	1.47	2.25	2.19	1.88
9	2,4-Dimethoxybenzaldehyde	613-45-6	1.12	1.00	0.12	1.91	3.48	0.92	0.96	0.88
10	2-Nitrobenzaldehyde	552-89-6	1.02	0.81	0.16	1.74	2.88	1.02	1.07	0.97
11	2-Chloro-6-fluorobenzaldehyde	387-45-1	1.00	0.46	0.08	2.54	2.75	1.23	1.59	1.55
12	4-Chlorobenzaldehyde	104-88-1	0.97	0.51	0.00	2.10	2.95	1.80	1.26	1.24
13	3,5-Dibromosalicylaldehyde	90-59-5	1.13	0.71	0.60	3.83	4.38	2.52	2.59	2.43
14	Salicylylaldehyde	90-02-8	1.05	0.79	0.60	1.81	2.57	1.72	1.71	1.38
15	4-Nitrobenzaldehyde	555-16-8	1.02	0.81	0.16	1.50	2.22	1.17	0.98	0.86
16	Benzaldehyde	100-52-7	0.92	0.56	0.00	1.48	2.43	1.03	0.94	0.90
17	4-Isopropylbenzaldehyde	122-03-2	0.92	0.56	0.00	3.07	3.78	1.34	1.52	1.67
18	N,N-Dimethylamino-4-benzaldehyde	100-10-7	1.05	1.29	0.00	1.81	3.50	0.51	0.42	0.52
19	4-Phenoxybenzaldehyde	67-36-7	1.61	0.92	0.06	3.96	4.66	1.63	2.07	1.90
20	3-(3.4-Dichlorophenoxy)benzaldehyde	79124-76-8	1.71	0.82	0.06	5.49	5.69	2.95	2.80	2.72
21	3-(4-tert-Butylphenoxy)benzaldehyde	69770-23-6	1.61	0.92	0.06	5.93	6.32	2.84	2.78	2.86
22	4-(Diethylamino)salicylaldehyde	17754-90-4	1.18	1.52	0.60	3.34	4.75	1.55	1.62	1.60
23	4-(Diethylamino)benzaldehyde	120-21-8	1.05	1.29	0.00	2.94	4.38	0.87	0.83	1.08

24	Hexanal	66-25-1	0.65	0.41	0.00	1.78	2.85	0.76	1.01	1.15
25	Ethanal	75-07-0	0.65	0.41	0.00	-0.20	0.81	0.11	0.27	0.17
26	Butanal	123-72-8	0.65	0.41	0.00	0.88	1.85	0.69	0.67	0.71
27	Isovaleraldehyde	590-86-3	0.65	0.41	0.00	1.23	2.26	1.42	0.81	0.88
28	2-Methylvaleraldehyde	123-15-9	0.65	0.41	0.00	1.67	2.76	0.72	0.96	1.10
29	Valeraldehyde	110-62-3	0.65	0.41	0.00	1.36	2.35	0.82	0.85	0.95
30*	2.4-Dihydroxybenzaldehyde	95-01-2	1.18	1.02	1.20	1.71	2.71	1.02		1.65
31*	2-Fluorobenzaldehyde	446-52-6	0.95	0.51	0.08	1.76	2.24	1.95		
32*	a,a,a-Trifluoro-3-tolualdehyde	454-89-7	1.16	1.13	0.18	2.47	2.26	2.27		
33*	3-Ethoxy-4-hydroxybenzaldehyde	121-32-4	1.15	1.01	0.66	1.88	3.68	0.27		
34*	2-Tolualdehyde	529-20-4	0.92	0.56	0.00	2.26	2.85	0.35		

* Outlier

Aliphatic Chemicals ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^{1}X^{v}$	t-obs	t-pre[1]	t-pre[2]
1	1,2-Dichloropropane	78-87-5	0.70	0.30	0.12	1.99	2.53	-0.05	0.10	0.09
2	1,1,2-Trichloroethane	79-00-5	1.05	0.45	0.18	2.05	2.65	0.21	0.24	0.24
3	Tetrachloroethylene	127-18-4	1.50	0.70	0.29	3.40	2.65	1.00	1.50	1.50
4	1-Octyl cyanide	2243-27-8	0.65	0.44	0.22	3.12	4.27	1.43	1.28	1.27
5	Hexachloroethane	67-72-1	0.70	-0.30	0.30	4.14	3.86	2.22	2.21	2.19
6	1,1,1-Trichloroethane	71-55-6	1.05	0.45	0.18	2.49	2.31	0.45	0.57	0.57
7	Dichloromethane	75-09-2	0.70	0.30	0.12	1.25	1.70	-0.58	-0.55	-0.55
8	Pentachloroethane	76-01-7	1.75	0.75	0.30	3.63	3.48	1.43	1.75	1.75
9	Trichloroethylene	79-01-6	1.05	0.45	0.18	2.42	2.18	0.47	0.51	0.51
10	1,1,2,2-Tetrachloroethane	79-34-5	1.40	0.60	0.24	2.39	3.11	0.92	0.64	0.64
11	1,2,3-Trichloropropane	96-18-4	1.05	0.45	0.18	1.98	3.20	0.41	0.22	0.22
12	1-Bromopropane	106-94-5	0.43	0.17	0.05	2.10	2.61	0.26	0.08	0.08
13	1.2-Dichloroethane	107-06-2	0.70	0.30	0.12	1.48	1.70	-0.14	-0.37	-0.37
14	Propionitrile	107-12-0	0.65	0.44	0.22	0.16	1.27	-1.44	-1.31	-1.31
15	1-Bromobutane	109-65-9	0.43	0.17	0.05	2.75	3.11	0.57	0.64	0.64
16	1,4-Dichlorobutane	110-56-5	0.70	0.30	0.12	2.24	3.20	0.39	0.34	0.34
17	1-Bromohexane	0111-25-1	0.43	0.17	0.05	3.80	4.10	1.67	1.56	1.56
18	1-Bromooctane	111-83-1	0.43	0.17	0.05	4.89	5.10	2.35	2.51	2.51
19	1.3-Dichloropropane	142-28-9	0.70	0.30	0.12	2.00	2.70	0.01	0.11	0.11
20	2,4-Hexadiene	592-46-1	0.20	0.20	0.10	2.96	2.15	0.61	0.82	0.82
21	1,5-Dichloropentane	628-76-2	0.70	0.30	0.12	2.76	3.70	0.74	0.79	0.79
22	1-Bromoheptane	629-04-9	0.43	0.17	0.05	4.36	4.60	2.08	2.04	2.05
23	2,5-Dimethyl-2,4-hexadiene	764-13-6	0.20	0.20	0.10	3.76	2.91	1.46	1.52	1.52
24	1,4-Dicyanobutane	111-69-3	1.30	0.88	0.44	-0.32	2.65	-1.25	-1.28	-1.28
25	1,6-Dicyanobutane		1.30	0.88	0.44	0.59	3.65	-0.58	-0.47	-0.48
26	1-Undecylcyanide	2437-25-4	0.65	0.44	0.22	4.90	5.78	2.61	2.82	2.82
27	Hexachloro-1.3-butadiene	87-68-3	2.30	1.10	0.46	4.78	4.35	3.46	2.98	2.98

28	2,3-Dimethyl-1,3-butadiene	513-81-5	0.20	0.20	0.10	2.70	1.95	1.07	0.60	0.60
29	1,9-Decadiene	1647-16-1	0.20	0.20	0.10	4.90	4.13	2.67	2.53	2.53
30*	1,3-Dichloropropene	542-75-6	0.80	0.40	0.17	1.60	2.28	2.66		
31*	3,4-Dichloro-1-butene	760-23-6	0.80	0.40	0.17	1.97	2.69	1.18		
32*	2,3-Dichloro-2-methyl-1-propene		0.80	0.40	0.17	1.56	2.65	2.82		
33*	1,3-Dibromopropane	109-64-8	0.86	0.34	0.10	1.99	3.81	1.98		
34**	Tetrachloromethane	56-23-5	0.70	0.00	0.21	2.72	2.41	0.56	0.81	0.81

* Outlier
** Test set

Esters ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^{1}X^{v}$	t-obs	t-pre[1]	t-pre[2]
1	Methyl Acetate	79-20-9	0.55	0.45	0.12	0.18	1.32	-0.68	0.06	0.02
2	Ethyl 4-aminobenzoate	94-09-7	0.89	0.81	0.38	1.86	3.76	0.67	0.72	0.92
3	2-Ethoxyethyl acetate	111-15-9	0.82	0.9	0.12	0.65	2.95	0.5	0.31	0.73
4	Methyl 4-chlorobenzoate	1126-46-1	0.81	0.38	0.12	2.9	3.48	1.19	0.99	0.85
5	Ethyl Salicylate	118-61-6	0.89	0.66	0.72	3.14	3.43	0.92	0.8	0.73
6	Phenyl-4-aminosalicylate	133-11-9	1.61	1.18	0.98	3.15	5.02	1.67	1.67	1.87
7	Propylacetate	109-60-4	0.55	0.45	0.12	1.2	2.32	0.23	0.28	0.26
8	Phenyl Salicylate	118-55-8	1.48	0.8	0.72	4.12	4.82	2.25	1.83	1.76
9	Butyl acetate	123-86-4	0.55	0.45	0.12	1.73	2.82	0.81	0.39	0.38
10	Ethyl acetate	141-78-6	0.55	0.45	0.12	0.73	1.82	-0.42	0.17	0.14
11	Hexyl Acetate	142-92-7	0.55	0.45	0.12	2.79	3.9	1.52	0.64	0.64
12	Methyl 4-nitrobenzoate	619-50-1	0.86	0.68	0.28	2.02	3.42	0.88	0.73	0.84
13	Methyl 2,4-dihydroxybenzoate	2150-47-2	1.02	0.89	1.32	2.22	3.24	0.56	0.77	0.64
14	Methyl 2,5-dichlorobenzoate	2905-69-3	0.86	0.33	0.12	3.37	4.01	1.17	1.25	1.03
15	Methyl 4-chloro-2-nitrobenzoate	42087-80-9	0.91	0.63	0.28	2.49	3.93	0.89	0.98	1.02
16	Diethylphthalate	84-66-2	0.93	0.72	0.24	2.47	5.13	0.84	1.23	1.34
17	t-Butyl Acetate	540-88-5	0.55	0.45	0.12	1.38	2.61	-0.45	0.35	0.33
18	Dimethylphthalate	131-11-3	0.93	0.72	0.24	-4.8	3.95	0.7	1.3	1.06
19	Dihexyl phthalate	84-75-3	0.93	0.72	0.24	1.8	9.13	3.14	2.37	2.29
20	Dibutyl phthalate	84-74-2	0.93	0.72	0.24	-0.84	7.13	2.05	1.96	1.81
21	Dioctyl Phthalate	117-84-0	0.93	0.72	0.24	4.44	11.13	3.57	2.78	2.76
22	Diisobutyl Phthalate	84-69-5	0.93	0.72	0.24	-0.84	7	2.48	1.93	1.78
23	Butyl benzyl phthalate	85-68-7	1.52	0.86	0.24	-1.15	7.58	2.32	2.81	2.62
24	Diisooctyl Phthalate	27554-26-3	0.93	0.72	0.24	4.44	11.1	3.13	2.77	2.75
25	Diisononyl phthalate	28553-12-0	0.93	0.72	0.24	5.76	12.1	3.34	2.98	2.99
26	Diisodecylphthalate	26761-40-0	0.93	0.72	0.24	7.1	13.1	2.65	3.18	3.23
27	Di(2-ethylhexyl)phtalate	117-81-7	0.93	0.72	0.24	4.44	10.9	2.88	2.72	2.7
28	Diundecylphthalate	3648-20-2	0.93	0.72	0.24	8.4	14.1	2.56	3.39	3.46
29	Ditridecylphthalate	119-06-2	0.93	0.72	0.24	10.24	16.13	3.31	3.85	3.94
30	Ethyl hexanoate	123-66-0	0.55	0.45	0.12	2.79	3.95	1.21	0.65	0.65
31	Dimethyl aminoterephthalate	5372-81-6	0.93	0.72	0.24	2.45	4.52	1.39	1.06	1.19

32	Methyl 4-cyanobenzoate	1129-35-7	0.96	0.8	0.34	1.72	3.35	0.54	0.73	0.92

Ketones ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^{1}X^{v}$	t-obs	t-pre[1]	t-pre[2]
1	Acetophenone	98-86-2	0.98	0.53	0.06	1.58	2.86	-0.13	-0.17	-0.14
2	4-Methyl-2-pentanone	108-10-1	0.67	0.48	0.00	1.31	2.61	-0.71	-0.54	-0.50
3	2-Octanone	111-13-7	0.67	0.48	0.00	2.37	3.76	0.55	0.42	0.35
4	4'-Aminopropiophenone	70-69-9	1.11	0.91	0.32	1.43	3.62	0.01	-0.13	-0.08
5	2-Heptanone	110-43-0	0.67	0.48	0.00	1.98	3.26	-0.06	0.05	0.04
6	2-Undecanone	112-12-9	0.67	0.48	0.00	4.09	5.26	2.05	1.90	1.73
7	2-Hexanone	591-78-6	0.67	0.48	0.00	1.38	2.41	-0.62	-0.55	-0.44
8	2-Nonanone	821-55-6	0.67	0.48	0.00	3.18	4.26	0.88	1.07	1.00
9	2-Amino-4'-chlorobenzophenone	2894-51-1	1.94	1.31	0.06	3.95	5.70	2.06	1.78	2.11
10	2-Dodecanone	6175-49-1	0.67	0.48	0.00	4.49	2.18	1.18	1.36	2.05
11	2',3',4'-Trimethoxyacetophenone	13909-73-4	1.28	1.19	0.24	1.12	6.52	0.02	0.18	-0.31
12	Acetone	67-64-1	0.67	0.48	0.00	-0.24	1.20	-2.08	-1.89	-1.74
13	3,3-Dimethyl-2-butanone	75-97-8	0.67	0.48	0.00	0.97	2.45	0.06	-0.80	-0.77
14	2-Butanone	78-93-3	0.67	0.48	0.00	0.29	1.76	-1.64	-1.41	-1.32
15	3-Pentanone	96-22-0	0.67	0.48	0.00	0.79	2.32	-1.25	-0.95	-0.92
16	3'-Aminoacetophenone	99-03-6	1.11	0.91	0.32	0.90	3.06	-0.44	-0.61	-0.50
17	Cyclohexanone	108-94-1	0.76	0.52	0.00	0.81	3.05	-0.80	-0.74	-0.87
18	5-Methyl-2-hexanone	110-12-3	0.67	0.48	0.00	1.88	3.12	-0.14	-0.05	-0.04
19	6-Methyl-5-heptan-2-one		0.77	0.58	0.05	1.70	3.25	0.16	-0.13	-0.12
20	Benzophenone	119-61-9	1.57	0.67	0.06	3.18	4.52	1.08	1.39	1.35
21	5-Nonanone	502-56-7	0.67	0.48	0.00	2.91	4.32	0.66	0.91	0.78
22	2'-Hydroxy-4'-methoxyacetophenone	552-41-0	1.21	0.98	0.72	1.98	3.53	0.37	0.51	0.60
23	3-Methyl-2-butanone	563-80-4	0.67	0.48	0.00	0.56	2.15	-1.00	-1.14	-1.10
24	2-Tridecanone	593-08-8	0.67	0.48	0.00	5.02	6.25	2.73	2.75	2.48
25	2-Dodecanone	6175-49-1	0.67	0.48	0.00	3.73	4.75	1.50	1.55	1.44
26	2',3',4'-Trichloroacetophenone	13608-87-2	1.13	0.40	0.06	3.57	4.41	2.04	1.67	1.51
27	2,4'-Dichloroacetophenone	937-20-2	1.08	0.43	0.06	2.84	3.89	1.20	1.02	0.90
28	4-Bromophenyl-3-pyridylketone		1.89	0.92	0.06	2.97	5.27	1.10	1.37	1.30
29*	1-Benzoylacetone	93-91-4	1.28	0.88	0.06	1.05	4.11	2.16		
30*	2-Methyl-1.4-naphthaquinone		1.76	0.96	0.12	2.20	4.05	3.19		
31*	2,4-Pentanedione	123-54-6	0.60	0.70	0.00	-0.54	2.11	-0.13		
32*	a-Bromo-2',5'-dimethoxy acetophenone		1.61	1.14	0.23	2.39	5.17	3.58		
33*	5,5-Dimethyl-1,3-cyclohexanedione	126-81-8	1.52	1.04	0.00	0.51	3.53	-1.91		
34*	4'-Chloro-3'-nitroacetophenone	5465-65-6	1.13	0.73	0.22	1.96	3.67	1.52		

* Outlier

Phenols ecotoxicity data set (Kulkarni et al. 2001) – Chapter 6

ID	*Chemicals*	CAS	π*	β	α	logKow	$^1X^v$	t-obs	t-pre[1]	t-pre[2]
1	4,6-Dinitro-2-cresol	534-52-1	0.92	0.77	0.86	2.56	3.45	2.06	1.42	1.47
2	Pentachlorophenol	87-86-5	0.87	0.23	0.60	5.12	4.73	3.04	2.52	2.25
3	Pentabromophenol	608-71-9	0.33	0.03	0.60	5.74	6.67	3.72	3.39	3.74
4	2,4-Dinitrophenol	51-28-5	1.14	0.73	0.92	1.54	3.04	1.23	0.94	1.29
5	Phenol	108-95-2	0.72	0.33	0.61	1.46	2.13	0.46	0.55	0.59
6	1-Naphthol	90-15-3	0.83	0.43	0.60	2.51	3.54	1.48	1.18	1.37
7	2-Cchlorophenol	95-57-8	0.82	0.30	0.60	2.15	2.65	0.97	0.90	0.90
8	3-Methoxyphenol	150-19-6	1.27	1.05	0.80	-0.54	2.65	0.22	-0.26	0.59
9	4-Nitrophenol	100-02-7	0.82	0.59	0.66	1.58	2.57	0.53	0.65	0.74
10	4-Chloro-3-methylphenol	59-50-7	1.01	0.53	0.60	1.91	3.06	1.42	0.76	0.96
11	2,4-Dimethylphenol	105-67-9	0.77	0.32	0.60	3.10	2.95	0.87	1.35	1.09
12	Catechol	120-80-9	0.66	0.38	0.61	2.30	2.27	1.08	0.93	0.66
13	2-Cresol	95-48-7	1.07	0.52	0.85	0.88	2.55	0.89	0.56	1.03
14	4-Methoxyphenol	150-76-5	1.00	0.66	0.31	1.95	2.57	0.05	0.25	0.09
15	2-Nitrophenol	88-75-5	0.82	0.59	0.66	1.34	2.58	-0.06	0.56	0.75
16	4-Chlorophenol	106-48-9	1.11	0.50	0.62	1.85	2.64	1.32	0.65	0.72
17	2,4,6-Trichlorophenol	88-06-2	0.82	0.30	0.60	2.48	3.68	1.61	1.24	1.55
18	2,4-Dichlorophenol	120-83-2	0.97	0.23	0.60	3.69	3.16	1.32	1.59	1.24
19	2-Phenylphenol	90-43-7	0.92	0.26	0.60	2.92	4.21	1.44	1.50	1.89
20	2,4,6-Trimethylphenol	527-60-6	1.33	0.51	0.60	3.36	3.37	1.02	1.31	1.09
21	2,3,6-Trimethylphenol	2416-94-6	0.60	0.43	0.60	3.42	3.38	1.22	1.60	1.32
22	4-Ethylphenol	123-07-9	0.60	0.43	0.60	3.42	3.25	1.07	1.57	1.24
23	4-Propylphenol	645-56-7	0.68	0.39	0.60	2.58	3.60	1.09	1.28	1.47
24	4-tert-Butylphenol	98-54-4	0.78	0.29	0.60	3.31	3.79	1.46	1.61	1.64
25	4-tert-Pentylphenol	80-46-6	0.78	0.29	0.60	3.98	4.35	1.80	2.00	1.99
26	4-Phenylphenol	92-69-3	1.33	0.51	0.60	3.36	4.20	1.44	1.48	1.61
27	4-Phenoxyphenol	831-82-3	1.41	0.73	0.66	3.75	4.36	1.58	1.67	1.64
28	4-Chlorocatechol	2138-22-9	0.90	0.55	1.20	1.97	2.78	1.96	1.57	1.76
29	2,3,4,6-Tetrachlorophenol	58-90-2	0.82	0.28	0.60	4.45	4.21	2.35	2.15	1.90
30	Tetrachlorocatechol	1198-55-6	1.05	0.51	1.20	4.29	4.35	2.29	2.79	2.74
31	2,4,6-Tribromophenol	118-79-6	0.71	0.02	0.00	4.02	4.85	1.70	1.39	1.54
32	2,6-Dinitrophenol	573-56-8	1.16	0.70	0.62	1.91	3.04	0.67	0.69	0.82
33	2,5-Dinitrophenol	329-71-5	0.92	0.87	0.92	1.75	3.04	1.74	1.06	1.24
34	tert-Butyl-2,6-dinitrophenol	4097-49-8	0.88	0.89	0.92	3.36	4.70	2.65	2.06	2.28
35	4-Nonylphenol	104-40-5	0.68	0.39	0.60	6.36	6.60	3.20	3.42	3.36
36	2,4,6-Tri-t-butylphenol	732-26-3	0.60	0.43	0.60	6.95	7.13	3.63	3.79	3.68
37	2,6-Di-tert-Butyl-4-methylphenol	128-37-0	0.60	0.43	0.60	6.07	5.87	2.78	3.17	2.89
38*	3-Acetamidophenol	621-42-1	1.59	0.95	0.75	-0.54	3.25	-0.87		
39*	2-Acetamidophenol	614-80-2	1.43	1.05	0.82	-0.54	3.28	-0.73		
40*	2-Allylphenol	1745-81-9	0.68	0.39	0.60	3.18	4.22	0.95		

41*	Resorcinol	108-46-3	0.85	0.60	1.20	0.80	2.26	0.34		
42*	2,3,4,5-Tetrachlorophenol	4901-51-3	0.82	0.28	0.60	4.21	4.21	2.75		
43**	3,4-Xylenol	95-65-8	0.72	0.37	0.60	2.23	2.95	0.94	1.00	1.07
44**	2,6-Xylenol	576-26-1	0.72	0.37	0.60	2.36	2.96	0.65	1.06	1.07
45**	2,6-Dibromo-4-cyanophenol	1689-84-5	1.11	0.57	0.80	2.63	4.34	1.37	1.54	2.03

* Outlier
** Test set

Toxic and metabolic effects of alcohols on the perfused rat liver data set (Cronin et al. 2002) – Chapter 7

ID	Chemicals	CAS	logGPT	logLDH	logGLDH	log1/ATP	logP	E_{LUMO}	$^3X_{pc}$
1	Methanol	67-56-1	0.00	0.00	0.00	0.06	-0.77	3.778	0.000
2	Ethanol	64-17-5	0.32	0.34	0.38	-0.03	-0.31	3.565	0.000
3	1-Propanol	71-23-8	0.63	0.63	0.71	-0.11	0.25	3.489	0.000
4	1-Butanol	71-36-3	0.92	0.91	0.82	-0.15	0.88	3.424	0.000
5	1-Pentanol	71-41-0	1.36	1.42	1.11	-0.80	1.56	3.390	0.000
6	2-Propanol	67-63-0	0.04	0.20	0.40	0.04	0.05	3.576	0.577
7	2-Butanol	78-92-2	0.88	0.83	0.66	-0.38	0.61	3.476	0.408
8	2-Pentanol	6032-29-7	1.12	1.05	0.81	-0.66	1.19	3.517	0.408
9	3-Pentanol	584-02-1	1.11	1.02	0.58	-0.66	1.21	3.482	0.289
10	2-Methyl-1-propanol	78-83-1	0.87	1.02	0.68	-0.38	0.76	3.547	0.408
11	2-Methyl-1-butanol	137-32-6	1.17	1.27	1.35	-1.10	1.22	3.543	0.289
12	3-Methyl-1-propanol		1.06	0.89	0.81	-0.74	1.16	3.386	0.408
13	3-Methyl-2-butanol	598-75-4	0.91	0.92	0.73	-0.55	1.28	3.488	0.667
14	2-Methyl-2-propanol	75-65-0	-0.30	-0.22	-0.22	0.03	0.35	3.438	2.000
15	2-Methyl-2-butanol	75-85-4	0.38	0.34	0.04	-0.30	0.89	3.435	1.561
16	Cyclopropylmethanol	2516-33-8	0.28	0.36	0.26	-0.15	0.21	3.033	0.204
17	2-Propen-1-ol	107-18-6	1.06	1.40	0.56	-1.22	0.17	1.364	0.000
18	2-Propyn-1-ol	107-19-7	0.92	1.09	0.99	-0.96	-0.38	1.560	0.000
19	2-Buten-1-ol	6117-91-5	0.97	1.00	0.60	-1.05	0.34	1.280	0.000
20	1-Buten-3-ol	627-27-0	1.21	1.37	0.78	-1.15	0.12	1.134	0.408
21	2-Methyl-2-propen-1-ol	513-42-8	0.91	1.15	1.16	-0.92	0.21	1.299	0.408
22	3-Methyl-2-buten-1-ol	556-82-1	0.88	0.85	0.70	-0.36	0.74	1.089	0.408
23	2-Methyl-3-butyn-2-ol	115-19-5	0.15	0.28	0.18	-0.27	0.28	1.679	1.561

Toxicity of pyridines to mice data set (Cronin et al. 2002) – Chapter 7

ID	Chemicals	CAS	$log1/LD_{50}$	logP	E_{LUMO}
1	Pyridine	110-86-1	1.86	0.65	0.138
2	2-Hydroxypyridine	72762-00-6	2.37	0.93	0.121
3	3-Hydroxypyridine	109-00-2	1.72	0.47	0.021

4	4-Hydroxypyridine	626-64-2	2.01	0.47	0.017
5	2,3-Dihydroxypyridine	16867-04-2	2.10	1.29	0.044
6	2,6-Dihydroxypyridine	626-06-2	2.32	1.22	0.040
7	3-Hydroxy-6-methylpyridine	1121-78-4	2.11	1.43	0.101
8	2-Methylpyridine	109-06-8	2.25	1.11	0.150
9	3-Methylpyridine	108-99-6	2.13	1.20	0.129
10	4-Methylpyridine	108-89-4	2.44	1.22	0.207
11	2,2'-Dipyridine	366-18-7	2.78	1.73	-0.537
12	4,4'-Dimethyl-2,2'-bipyridine	1134-35-6	3.37	2.56	-0.460
13	5,5'-Dimethyl-2,2'-bipyridine	1762-34-1	2.91	2.56	-0.521
14	6,6'-Dimethyl-2,2'-bipyridine		2.99	2.56	-0.440
15	N-Oxide-pyridine		1.62	-1.20	-0.328
16	3-Hydroxy-N-Oxide-pyridine		1.31	-0.61	-0.385
17	4-Hydroxy-N-Oxide-pyridine		1.65	-0.84	-0.504
18	3-Methyl-N-Oxide-pyridine		1.81	-0.74	-0.318
19	4-Methyl-N-Oxide-pyridine		1.94	-0.74	-0.281
20	N,N'-Dioxide-2,2'-bipyridine		1.86	-1.68	-0.874
21	N-Oxide-2,2'-bipyridine		2.56	-0.08	-0.800

Lethality of halogenated hydrocarbons to *Aspergillus nidulans* data set (Cronin et al. 2002) Chapter 7

ID	**Chemicals**	**CAS**	**log1/D_{37}**	**logP**	**E_{LUMO}**
1	Dichloromethane	75-09-2	-1.97	1.25	0.594
2	Chloroform	67-66-3	-1.39	1.97	-0.304
3	Tetrachloromethane	56-23-5	-0.49	2.83	-1.116
4	1,1-Dichloroethane	75-34-3	-1.68	1.79	0.582
5	1,2-Dichloroethane	107-06-2	-1.71	1.47	0.686
6	1,1,1-Trichloroethane	71-55-6	-1	2.49	-0.266
7	1,1,2-Trichloroethane	79-00-5	-1.03	2.07	0.171
8	1,1,1,2-Tetrachloroethane	630-20-6	-0.45	3.03	-0.485
9	1,1,2,2-Tetrachloroethane	79-34-5	-0.45	2.39	-0.23
10	Pentachloroethane	76-01-7	-0.23	3.22	-0.681
11	Hexachloroethane	67-72-1	0.1	4.14	-0.968
12	1,1,2-Trichloroethylene	79-01-6	-1.05	2.61	-0.061
13	Tetrachloroethylene	127-18-4	-0.08	3.4	-0.437
14	1,2-Dichloroethylene	540-59-0	-1.48	1.86	0.34
15	1,1-Dichloroethylene	75-35-4	-1.4	2.13	0.379
16	1,2-Dichloropropane	78-87-5	-1.19	1.99	1.117
17	2,2-Dichloropropane	594-20-7	-1.28	2.31	0.575
18	1,3-Dichloropropane	142-28-9	-1.02	2	1.02
19	1,2,3-Trichloropropane	96-18-4	-0.97	1.98	0.449
20	1-Chlorobutane	109-69-3	-1.16	2.64	1.511

21	2-Chlorobutane	78-86-4	-1.28	2.33	1.434
22	1,3-Dichlorobutane	1190-22-3	-0.94	2.24	1.15
23	2,3-Dichlorobutane	7581-97-7	-0.94	2.52	0.671
24	1-Chloro-2-methylpropane	513-36-0	-1.16	2.92	1.544
25	2-Chloro-2-methyl propane	507-20-0	-1.26	2.52	1.325
26	1-Chloropentane	543-59-9	-0.7	3.05	1.509
27	1-Chlorohexane	544-10-5	-0.18	3.58	1.509
28	1-Chlorooctane	111-85-3	-0.18	4.64	1.508
29	1,1-Dichloropropene	563-58-6	-0.82	2.9	0.368
30	2,3-Dichloropropene	78-88-6	-0.43	2.04	0.491
31	1,3-Dichloropropene	542-75-6	-0.72	1.76	0.434
32	1,1,3-Trichloro-1-propene	2567-14-8	-0.38	2.62	0.012
33	3-Chloro-2-chloromethyl-1-propene	1871-57-4	-0.43	1.62	0.151
34	1-Chloro-2-methyl-1-propene	513-37-1	-1.1	2.58	0.789
35	3-Chloro-2-methyl-1-propene	563-47-3	-0.88	1.91	0.623
36	Chlorodibromofluoromethane	353-55-9	-0.49	2.71	-1.665
37	Bromoform	75-25-2	-0.72	2.67	-0.747
38	Bromochloromethane	74-97-5	-1.79	1.41	0.1
39	Bromotrichloromethane	75-62-7	0.1	3.01	-1.314
40	Bromodichloromethane	75-27-4	-1.03	2.09	-0.632
41	Chlorodibromomethane	124-48-1	-0.46	2.23	-0.728
42	1-Bromo-2-Chloroethane	107-04-0	-1.38	1.6	0.232
43	1-Bromobutane	109-65-9	-0.75	2.75	0.829
44	2-Bromobutane	78-76-2	-1.06	2.66	0.759
45	1-Bromo-3-chloropropane	109-70-6	-0.88	1.85	0.507
46	2-Bromo-1-chloropropane	3017-95-6	-0.99	2.13	0.466
47	1-Bromo-2-methylpropane	78-77-3	-1.26	2.53	0.862
48	2-Bromo-2-methylpropane	507-19-7	-1.41	1.15	0.661
49	1-Bromo-4-Chlorobutane	6940-78-9	-0.52	2.38	0.624
50	1-Bromo-3-methylbutane	107-82-4	-0.49	3.06	0.837
51	Dibromodichloromethane	594-18-3	0.7	3.15	-1.349
52	Dibromomethane	74-95-3	-1.33	1.88	-0.052
53	Tetrabromomethane	558-13-4	2	3.42	-1.227
54	1,1,2,2-Tetrabromoethane	79-27-6	0.4	3.2	-0.775
55	1,2-Dibromoethylene	624-61-3	-0.96	1.95	-0.041

Toxicity of nitrobenzenes to the alga *Chlorella vulgaris* data set (Cronin et al. 2002) – Chapter 7

ID	Chemicals	CAS	log1/EC_{50}	logP	E_{LUMO}
1	Nitrobenzene	98-95-3	-0.78	1.85	-1.068
2	4-Chloronitrobenzene	100-00-5	1.25	2.39	-1.343
3	3-Nitrotoluene	99-08-1	-0.50	2.42	-1.016

4	2-Methyl-3-chloronitrobenzene		1.17	3.09	-1.219
5	2,5-Dichloronitrobenzene	89-61-2	0.97	3.03	-1.467
6	2,4,6-Trimethylnitrobenzene	603-71-4	0.25	3.22	-0.935
7	2,4,5-Trichloronitrobenzene	89-69-0	1.88	3.47	-1.691
8	2,4,5,6-Tetrachloronitrobenzene	117-18-0	2.34	4.38	-1.42
9	1,2-Dinitrobenzene	528-29-0	1.23	1.69	-1.841
10	1,3-Dinitrobenzene	99-65-0	0.38	1.49	-1.912
11	1,4-Dinitrobenzene	100-25-4	0.41	1.47	-2.208
12	2,4-Dinitrotoluene	121-14-2	0.70	1.98	-1.841
13	2,4,6-Trichloro-1,3-dinitrobenzene		1.89	2.97	-2.037

TA98 and TA100 Aromatic Amines data set (Gramatica et al. 2003) – Chapter 8

ID	Compounds	T98exp	TA98est	TA100exp	TA100est	MWC07	MATS7m	Mor27u	Mor15m	nHA	ATS5p	L2v
1	2.3-Dimethylaniline	-	-1.90	-1.36	-0.98	0.5	0	0.164	0.434	1	0.22	2.291
2	2.5-Dimethylaniline	-2.40	-1.75	-1.43	-1.18	0.5	0.814	0.038	0.434	1	0.252	1.571
3	4-Chloro-1.2-phenylendiamine	-0.49	-1.68	-1.44	-1.57	0.5	0	0.146	0.666	2	0.266	1.715
4	4-Aminophenylsulfide	0.31	-0.77	0.48	0.84	0.8	0.038	-0.086	1.327	2	0.533	1.43
5	4-Aminopyrene	3.16	3.74	2.69	2.22	2.1	0.265	0.091	2.068	1	0.493	3.413
6	2-Amino-4-methylphenol	-2.10	-1.62	-1.68	-1.72	0.5	0.975	-0.033	0.621	2	0.255	1.639
7	1-Aminofluoranthene	3.35	2.78	2.34	2.51	2.2	0.26	-0.18	1.587	1	0.508	3.424
8	2-Aminofluorene	1.93	0.70	0.78	0.54	1.4	-0.12	0.04	1.152	1	0.422	1.738
9	Benzidine	-0.39	-0.61	-0.66	-0.45	0.9	-0.274	0.21	0.767	2	0.416	1.203
10	4-Methyl-2-bromoaniline	-	-1.39	-0.64	-0.07	0.5	0.302	-0.034	1.17	1	0.35	1.798
11	8-Aminoquinoline	-1.14	-0.53	-0.34	-0.31	0.8	1.36	-0.013	0.71	2	0.378	1.989
12	3.4-Dimethylaniline	-	-1.83	-1.08	-1.14	0.5	0.814	0.017	0.406	1	0.237	1.842
13	3-Aminofluorene	0.89	0.88	0.10	0.72	1.4	0.154	0.06	1.137	1	0.423	1.978
14	4-Methyl-2-chloroaniline	-	-2.12	-0.40	-0.62	0.5	0.705	-0.021	0.286	1	0.308	1.62
15	4-Aminofluorene	1.13	1.08	0.64	0.87	1.4	0.188	0.128	1.145	1	0.422	2.217
16	4-Chloroaniline	-2.52	-2.66	-1.51	-1.22	0.3	0	0.05	0.429	1	0.266	1.262
17	8-Aminofluoranthene	3.80	2.30	1.98	2.12	2.1	0.118	-0.234	1.558	1	0.511	2.81

18	2-Ethyl-4-chloroaniline	-	-2.38	0.08	0.07	0.5	0.652	-0.118	0.29	1	0.333	2.304
19	2-Aminopyrene	3.50	3.76	2.58	2.32	2.1	0.265	0.1	2.064	1	0.493	3.413
20	2-Aminonaphtalene	-0.67	-0.79	0.39	-0.13	0.7	0.351	0.069	1.029	1	0.36	1.631
21	4-Cyclohexylaniline	-1.24	-2.32	-0.14	-0.59	0.8	-0.102	-0.184	0.22	1	0.365	0.892
22	6-Aminoquinoline	-2.67	-1.47	-1.22	-0.63	0.7	0.154	-0.014	0.699	2	0.377	1.519
23	2-Amino-1-methylnaphtalene	-	-0.50	0.84	0.06	0.9	0.493	-0.013	0.966	1	0.338	2.205
24	4-Amino-3-methylbipheyl	-	-1.13	1.12	0.35	0.9	-0.129	0.003	0.678	1	0.411	1.649
25	4.4'-Ethylenbis(aniline)	-2.15	-1.70	-1.51	-1.00	0.9	-0.055	-0.29	0.774	2	0.381	0.805
26	2-Methoxy-5-methylaniline	-2.05	-2.76	-1.85	-1.12	0.5	-0.881	-0.101	0.687	2	0.31	1.704
27	2.4.5-Trimethylaniline	-1.32	-1.78	-0.26	-0.77	0.6	0.765	-0.038	0.379	1	0.263	2.083
28	2.4-Diamino-n-butylbenzene	-2.70	-2.38	-0.84	-0.95	0.6	-0.152	-0.117	0.476	2	0.338	1.605
29	7-Aminofluoranthene	2.88	2.33	2.76	2.25	2.2	-0.05	-0.286	1.57	1	0.506	3.112
30	1-Aminocarbazole	-1.04	0.27	-0.25	0.16	1.4	0.023	-0.072	0.941	2	0.426	1.981
31	1-Aminophenantrene	2.38	0.79	1.79	1.15	1.4	-0.038	-0.058	1.41	1	0.449	2.26
32	3-Amino-4-methylbiphenyl	-	-1.15	0.09	0.13	0.9	-0.217	0.061	0.581	1	0.397	1.457
33	3-Aminocarbazole	-0.48	0.22	-0.11	0.21	1.4	0.11	-0.097	0.905	2	0.444	1.814
34	4-Methoxy-2-methylaniline	-3.00	-2.41	-2.10	-1.36	0.5	0.188	-0.234	0.756	2	0.272	1.929
35	2-Aminobiphenyl	-1.49	-1.19	-0.51	0.78	0.8	0.118	0.009	0.701	1	0.445	1.603
36	3-Aminofluoranthene	3.31	2.71	2.25	1.97	2.2	0.128	-0.143	1.508	1	0.497	2.813
37	4-Aminobiphenyl	-0.14	-1.20	0.85	0.40	0.8	-0.109	0.081	0.653	1	0.439	1.25
38	3.3'-Dichlorobenzidine	0.81	-0.02	0.66	0.66	1.2	0.049	0.153	0.606	2	0.505	1.613
39	2.6-Dichloro-1.4-phenylendiamine	-0.69	-1.01	-1.12	-0.35	0.6	0.89	0.127	0.645	2	0.353	2.246
40	3.3'-Dimethoxybenzidine	0.15	0.47	-0.85	-0.79	1.3	-0.211	-0.005	1.311	4	0.407	2.667
41	4-Aminophenyldisulfide	-1.03	-0.50	0.54	1.23	0.8	-0.186	-0.14	1.804	3	0.629	1.362
42	2-Aminocarbazole	0.60	0.20	-0.56	-0.01	1.4	-0.082	-0.042	0.865	2	0.435	1.61
43	4-Aminocarbazole	-1.42	0.58	-0.47	0.29	1.4	0.267	0	0.941	2	0.434	2.086
44	1-Aminofluorene	0.43	0.59	-0.04	0.53	1.4	-0.048	-0.003	1.117	1	0.397	2.067
45	2-Aminoanthracene	2.62	1.12	2.76	0.88	1.3	-0.08	0.105	1.581	1	0.458	1.708

46	2-Amino-3-methylnaphthalene	-	-0.49	1.09	-0.07	0.9	0.22	0.027	1.019	1	0.352	1.802
47	2-Aminofluoranthene	3.23	2.20	2.87	2.57	2.1	0.005	-0.241	1.542	1	0.525	3.327
48	3-Aminoquinoline	-3.14	-1.39	0.07	-0.43	0.7	0.456	-0.028	0.649	2	0.402	1.449
49	3-Methoxy-4-methylaniline	-1.96	-1.73	-0.81	-0.46	0.5	0.888	-0.118	0.75	2	0.314	2.642
50	2-Chloroaniline	-3.00	-2.64	-2.05	-1.08	0.4	0	-0.005	0.357	1	0.242	1.776
51	4-Phenoxyaniline	0.38	-1.17	0.63	0.39	0.8	0.002	-0.216	1.271	2	0.497	1.345
52	2-Amino-4-chlorophenol	-3.00	-2.11	-2.00	-1.42	0.5	0	-0.051	0.711	2	0.277	1.723
53	1-Amino-2-methylnaphthalene	-	-0.73	-0.37	-0.07	0.9	0.455	-0.12	1.01	1	0.343	1.914
54	6-Aminocrysene	1.83	2.61	2.41	1.86	2.1	-0.032	-0.164	1.76	1	0.506	2.599
55	2-Methyl-4-bromoaniline	-	-1.47	0.46	-0.55	0.5	0	-0.043	1.274	1	0.306	1.807
56	4-Aminophenanthrene	-	0.39	-0.11	1.43	1.4	0.291	-0.205	1.206	1	0.463	2.483
57	4-Aminophenylether	-1.14	-0.17	-0.27	-0.61	0.9	0.059	0.01	1.432	3	0.459	1.338
58	4-Ethoxyaniline	-2.30	-2.73	-0.61	-1.11	0.4	-0.02	-0.234	0.79	2	0.359	1.088
59	1-Aminonaphthalene	-0.60	-0.33	-1.00	0.06	0.8	1.215	-0.034	1.01	1	0.346	2.093
60	2.4-Dimethylaniline	-2.22	-2.15	-0.23	-0.96	0.5	0.814	-0.147	0.478	1	0.265	1.771
61	2.4-Difluoroaniline	-2.70	-2.25	-2.52	-1.70	0.5	0	-0.158	0.828	1	0.199	1.427
62	4.4'-Methylenedianiline	-1.60	-0.60	-0.15	-0.42	0.9	0.248	-0.003	0.971	2	0.417	1.275
63	9-Aminophenanthrene	2.98	1.06	2.79	1.56	1.4	0.291	-0.031	1.429	1	0.463	2.676
64	3.4'-Diaminobiphenyl	0.20	-0.69	0.65	-0.12	0.9	-0.146	0.157	0.74	2	0.433	1.483
65	3-Aminophenanthrene		0.86	2.66	1.38	1.4	-0.105	0.002	1.377	1	0.464	2.394
66	2-Aminophenanthrene	2.46	0.93	2.74	1.02	1.4	0.166	-0.045	1.403	1	0.461	2.051
67	1-Aminoanthracene		0.99	0.36	0.97	1.4	-0.261	0.019	1.532	1	0.445	2.046
68	1-Aminopyrene	1.43	3.40	1.05	1.92	2.1	-0.056	0.017	2.082	1	0.481	2.969
69	9-Aminoanthracene	0.87	1.32	-0.24	1.08	1.4	0.139	0.049	1.559	1	0.445	2.089
70	2.4-Diaminotoluene	-1.29	-1.43	-1.66	-1.98	0.5	0.891	0.074	0.597	2	0.23	1.639
71	3.3'-Diaminobenzidine	-0.04	0.71	-1.11	-1.90	1.2	-0.132	0.308	1.013	4	0.38	1.681
72	1.3-Phenylendiamine	-0.46	-2.37	-1.40	-1.85	0.3	0	0.091	0.606	2	0.244	1.639
73	3.4-Diaminotoluene	-1.42	-1.36	-2.10	-1.76	0.5	0.891	0.1	0.606	2	0.25	1.636

74	1.2-Phenylendiamine	-0.75	-2.02	-1.89	-2.15	0.4	0	0.114	0.652	2	0.203	1.772
75	3-Amino-6-methylphenol	-1.40	-1.54	-1.82	-2.01	0.5	0.975	0.009	0.599	2	0.23	1.549
76	2.4-Diaminoethylbenzene	-0.87	-1.93	-1.21	-1.55	0.5	-0.254	0.13	0.607	2	0.283	1.503
77	3-Aminobiphenyl	-	-1.25	-	0.78	0.8	0.014	0.05	0.614	1	0.459	1.542
78	2.3-Diaminobiphenyl	-	-0.85	-	-0.07	1	0.024	-0.147	0.963	2	0.418	1.773
79	2-Methyl-4-chloroaniline	-	-2.38	0.38	-0.85	0.5	0	-0.015	0.395	1	0.269	1.825
80	2-Chloro-4-methylaniline	-	-2.10	-	-0.61	0.5	0.705	-0.02	0.3	1	0.308	1.62
81	4-Methoxyaniline	-	-3.16	-	-1.74	0.4	-1.064	-0.151	0.745	2	0.283	1.188
82	3-Methoxyaniline	-	-2.21	-	-1.49	0.4	0.975	-0.213	0.701	2	0.295	1.394
83	Aniline	-	-2.92	-	-1.75	0.2	0	0.014	0.492	1	0.206	1.365
84	3-Chloroaniline	-	-2.73	-	-0.54	0.3	0	0.006	0.468	1	0.315	1.622
85	3-Ethoxyaniline	-	-2.85	-	-0.79	0.4	-0.05	-0.247	0.722	2	0.369	1.39
86	2-Ethoxyaniline	-	-2.05	-	-0.57	0.5	0.495	-0.157	0.752	2	0.374	1.65
87	4-Aminophenol	-	-2.07	-	-2.33	0.3	1.11	-0.068	0.66	2	0.214	1.295
88	3-Aminophenol	-	-2.72	-	-1.98	0.3	0	-0.058	0.613	2	0.225	1.672
89	2.4.6-Trimethylaniline	-	-1.24	-	0.16	0.6	0.765	0.133	0.486	1	0.306	2.837
90	2.4.6-Tribromoaniline	-	-0.74	-	1.10	0.6	0	0.021	1.567	1	0.414	2.699
91	2.4.6-Trichloroaniline	-	-1.98	-	0.09	0.6	0	0.08	0.329	1	0.338	2.252
92	2.6-Diethylaniline	-	-2.17	-	-0.30	0.6	0.306	-0.156	0.517	1	0.32	1.928
93	3.5-Dimethylaniline	-	-2.82	-	0.24	0.4	0	-0.127	0.457	1	0.342	2.431
94	2.6-Dimethylaniline	-	-2.52	-	-0.96	0.5	0	-0.123	0.502	1	0.243	2.031
95	2.4-Dibromoaniline	-	-1.11	-	0.05	0.5	0	-0.028	1.561	1	0.357	1.918
96	2.4-Dichloroaniline	-	-2.36	-	-0.59	0.5	0	0.023	0.33	1	0.302	1.736
97	4-Iodoaniline	-	-	-	-0.11	0.3	0	-0.035	-0.828	1	0.391	1.176
98	2-Iodoaniline	-	-3.57	-	-0.40	0.4	0	-0.048	-0.387	1	0.318	1.791
99	2-Fluoroaniline	-	-2.45	-	-1.72	0.4	0	-0.07	0.671	1	0.204	1.447
100	2-Bromoaniline	-	-1.70	-	-0.91	0.4	0	-0.017	1.219	1	0.263	1.821
101	4-Ethylaniline	-	-2.72	-	-1.21	0.4	-0.278	-0.052	0.524	1	0.282	1.077

102	2-Ethylaniline	-	-1.93	-	-0.94	0.4	0.814	0.038	0.485	1	0.276	1.577
103	4-Methylaniline	-	-2.25	-	-1.57	0.3	0.883	-0.003	0.467	1	0.236	1.196
104	3-Methylaniline	-	-2.77	-	-0.99	0.3	0	-0.011	0.467	1	0.269	1.612
105	2-Methylaniline	-	-2.55	-	-1.51	0.4	0	-0.027	0.485	1	0.206	1.729
106	2.2'-Diaminobiphenyl	-1.52	-0.35	-	0.13	1	0.233	-0.052	1.101	2	0.431	1.886
107	3.3'-Dimethylbenzidine	0.01	-0.15	-	-0.22	1.2	-0.132	0.006	0.904	2	0.384	2.041
108	9-Aminofluorene	-	1.13	-	0.71	1.5	0.479	0	1.112	1	0.411	2.146
109	2.4-Diaminoisopropylbenzene	-3.00	-2.51	-	-1.19	0.6	-0.805	-0.032	0.509	2	0.312	1.615
110	2.4'-Diaminobiphenyl	-0.92	-0.43	-	-0.11	0.9	0.126	0.061	1.05	2	0.423	1.643
111	2-Aminophenol	-	-2.34	-	-2.15	0.4	0	-0.027	0.671	2	0.203	1.726
112	3.3'-Diaminobiphenyl	-1.30	-0.48	-	0.42	0.9	0.111	0.121	0.884	2	0.451	2.046
113	2-Methoxyaniline	-	-1.89	-	-1.15	0.4	0.975	-0.098	0.741	2	0.301	1.829
114	3-Trifluoromethylanilin e	-0.80	-1.47	-	-1.18	0.7	0	-0.015	0.777	1	0.248	1.632
115	4-Bromoaniline	-2.70	-1.86	-	-0.95	0.3	0	-0.014	1.285	1	0.3	1.218
116	2-Bromo-7-aminofluorene	2.62	1.89	-	1.27	1.5	-0.246	0.009	2.135	1	0.503	1.649
117	1.7-Diaminophenazine	0.75	0.63	-	-1.13	1.5	-0.185	-0.347	1.761	4	0.418	2.01
118	3-Amino-3'-nitrobiphenyl	-0.55	-0.19	-	-0.79	1.1	0.088	0.174	0.605	4	0.459	1.931
119	2.7-Diaminofluorene	0.48	1.28	-	-0.10	1.5	-0.332	0.161	1.305	2	0.421	1.682
120	2-Amino-4'-nitrobiphenyl	-0.62	0.17	-	-1.34	1.1	0.515	0.191	0.671	4	0.429	1.531
121	2-Amino-5-nitrophenol	-2.52	-2.42	-	-	0.6	-1.497	0.122	0.605	5	0.238	1.639
122	2-Hydroxy-7-aminofluorene	0.41	0.77	-	-0.10	1.5	-0.271	-0.085	1.356	2	0.419	1.713
123	4-Amino-2'-nitrobiphenyl	-0.92	-0.36	-	-1.16	1.1	-0.083	0.122	0.652	4	0.435	1.723
124	2-Aminophenazine	0.55	-0.27	-	-0.55	1.3	-0.061	-0.452	1.55	3	0.438	1.682
125	2.4-Dinitroaniline	-2.00	-2.46	-	-	0.8	-1.55	0.034	0.351	7	0.258	1.95
126	2-Amino-3'-nitrobiphenyl	-0.89	-0.03	-	-0.94	1.1	0.04	0.236	0.64	4	0.447	1.876
127	4-Fluoroaniline	-3.32	-2.74	-	-1.80	0.3	0	-0.077	0.639	1	0.202	1.345
128	2-Amino-7-acetamidofluorene	1.18	1.47	-	-0.53	1.7	-0.176	0.065	1.174	3	0.449	1.547

129	2.8-Diaminophenazine	1.12	0.43	-	-1.22	1.4	-0.049	-0.351	1.729	4	0.427	1.747
130	3-Amino-2'-nitrobiphenyl	-1.30	-0.41	-	-0.85	1.1	-0.067	0.104	0.642	4	0.451	1.963
131	1.6-Diaminophenazine	0.20	0.56	-	-	1.5	-0.161	-0.41	1.819	4	0.411	2.345
132	2-Bromo-4.6-dinitroaniline	-0.54	-0.47	-	-	1	-0.07	0.015	0.994	7	0.377	2.896
133	1.9-Diaminophenazine	0.04	0.41	-	-	1.5	-0.647	-0.364	1.833	4	0.411	2.406
134	2-Amino-1-nitronaphthalene	-1.17	0.13	-	-	1.2	0.233	0.032	0.915	4	0.351	2.185
135	3-Amino-4'-nitrobiphenyl	0.69	-0.21	-	-1.28	1.1	0.206	0.177	0.522	4	0.439	1.482
136	2-Amino-7-nitrofluorene	3.00	1.93	-	-1.25	1.7	0.041	0.197	1.188	4	0.428	1.688
137	1-Amino-7-nitronaphthalene	-1.77	-0.15	-	-	1.2	-0.323	0.086	0.825	4	0.359	2.965
138	4-Amino-3'-nitrobyphenil	1.02	-0.24	-	-1.17	1.1	-0.235	0.212	0.64	4	0.441	1.623
139	4-Amino-4'-nitrobyphenil	1.04	-0.31	-	-1.62	1	0.05	0.233	0.604	4	0.422	1.211
140	1-Aminophenazine	-0.01	-0.09	-	-0.41	1.4	-0.232	-0.471	1.62	3	0.428	2.039
141	4-Chloro-2-nitroaniline	-2.22	-2.06	-	-	0.7	0	-0.05	0.33	4	0.293	1.94
142	2.7-Diaminophenazine		0.47	-	-1.26	1.4	-0.049	-0.325	1.712	4	0.427	1.678
143	4-Chloro-1.3-phenylenediamine	-0.77	-1.89		-1.43	0.5	0	0.131	0.511	2	0.279	1.724
144	2-Nitro-1.4-phenylenediamine	-0.05	-0.69		-	0.7	1.51	0.052	0.566	5	0.246	1.929
145	4-Nitro-1.3-phenylenediamine	-2.4	-2.54		-	0.6	-1.509	0.11	0.526	5	0.244	1.777
146	4-Nitro-1.2-phenylenediamine		-2.05		-	0.6	-1.509	0.272	0.606	5	0.241	1.733

12. APPENDIX

The Structural Chemical Domain in Multiple Linear Regression: the Leverage from the Influence Matrix H

While the range of the independent variable X (descriptor) is useful to define the chemical domain of a univariate QSAR (based on one descriptor), in multivariate models the descriptor ranges are too limited to highlight those chemicals lying outside the domain.

In Multiple Linear Regression, where the data and residual distribution is generally normal, the predicted data Y are obtained from the experimental data Y through the Hat matrix of influence.

$$y = X(X^T X)^{-1} X^T y = H y$$

where X is the descriptor matrix.

The *ii* main diagonal entry of the Hat matrix (H, (*hii*) provides a measure of how far observation (ai) *i* is from the center of the X data (leverage).

The cut off value is defined by the formula $h^* = 3(p+1)/n$ where *p* are the variables and *n* the objects.

A chemical with high leverage in the training greatly influences the regression: the fitted regression line is forced near the observed value and the residuals are small. The chemical in the training set is not an outlier for the response fitting, but the predictions for high leverage chemicals in the test set could be extrapolated and unreliable.

A high leverage chemical is structurally anomalous in the chemical domain of the model, thus caution is needed for this chemical!

EXAMPLES:

A data set of 32 benzenes was modelled for ecotoxicity by Kulkarni et al (*SAR & QSAR Environ. Res*. 12, 565-591(2001), in this report Chapter 6 pag. 39): the published models have optimum fitting performances (R2 and R2 adj).

When applied by the authors to different test sets the results are opposite: one data set of seven 7 chemicals (set a) is well predicted (Q^2_{EXT}: 89%) while the set b of eleven 11 chemicals is not predicted (Q^2_{EXT} negative). The exercise gives evidence of model robustness and internal predictivity, as can be verified by the Table 3 parameters ($Q2_{LMO}$ and Bootstrap) and by the related graph. The statistical "external" validation by splitting the available data (by D-optimal design) highlighted the external unpredictivity of the model (Q2 ext: 44%)

It is interesting to verify the quality of the two test sets in the model descriptors space: the 7 test chemicals (a) are not high leverage chemicals and are well predicted by the model (as can be seen in the

corresponding regression line of Figure A) (the only high leverage chemical in the residual/leverage graph is n.32 of the training set, which is perfectly predicted, as normally happens for chemicals influential in training sets).

The 3 worst predicted chemicals of the 11 test chemicals (b) are high leverage chemicals. Other chemicals are badly predicted even if their structure is not beyond the model domain.

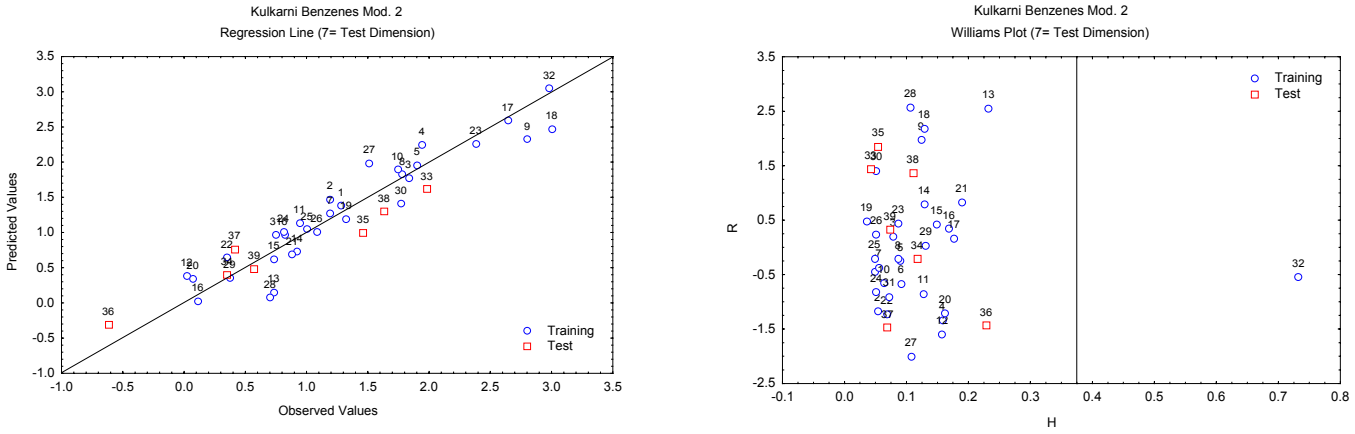


Figure A

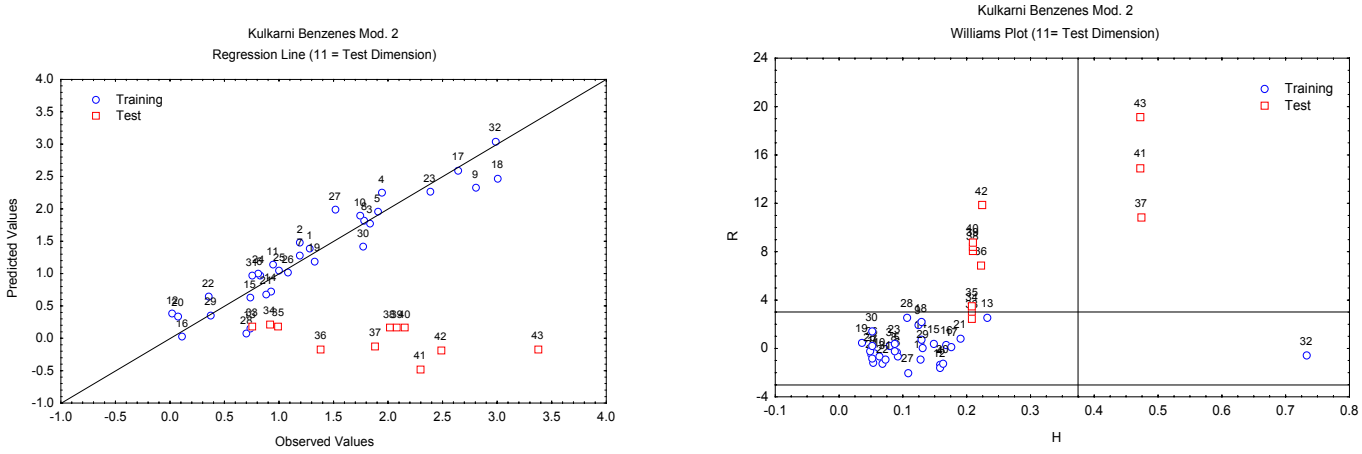


Figure B

Ecotoxicity of 29 aliphatic chemicals (hydrocarbons) was modelled by Kulkarni et al. (*SAR & QSAR Environ. Res*. 12, 565-591(2001)) by electronic and connectivity descriptors in addition to logKow (see Chapter 6 pag 68) 3 variables-models are robust and predictive as can be verified by all the validation parameters and the graph lines. 5-variables models are purely fitting, unpredictive models. Some unusual results are obtained by different splittings: the models appear highly sensitive to the presence of a peculiar chemical (n.5) put in the test (in this case the model is internally predictive, but not externally) or in the

training (in this case the model is internally unstable). This raised the need to verify the structural domain of model applicability. In the William graph of Figure 30 for the problematic model with 5 variables it is possible to verify that the chemical n.5 (hexachloroethane) is a high leverage chemical: it is very influential for the model (it is far from the majority of the chemicals) and perfectly fitted by the model (see the corresponding regression line), but its prediction when it is put in the text set hasn't the same reliability. This model is clearly an overfitting one (R^2: 96.7% but Q^2_{LOO} negative, influenced by the n.5). It is interesting to note that this high leverage chemical has only one descriptor with a minimum value. On the contrary, the same chemical is within the structural domain of the 3-variables model (Figure 31) and thus does not influence performance.

When splitting by K-ANN was performed, the model, apparently predictive by LOO CV, is externally unpredictive (R^2_{EXT} and Q^2_{EXT} negative). The regression line (Figure 32) highlights a strong outlier (n.5) and the residuals/leverage graph (Williams graph in Figure 32) two strong high leverage: n.5 in particular, but also n. 26. It is interesting to note that, while n.26 is a good leverage chemical (positively influencing regression), n.5 is a bad leverage chemical and also becomes an outlier for the response. Again we must observe that: "The predictions for high leverage chemicals could be unreliable" and for chemical n.5 they are.

The model developed without this peculiar chemical is stable and predictive for the data set (Q^2_{EXT} = 98.3%).