Principles of QSAR models validation: internal and external

Paola Gramatica

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, via Dunant 3, 21100 Varese, Italy; E-mail: paola.gramatica@uninsubria.it

Keywords: External validation, Internal validation, OECD principles, Predictivity, Quantitative structure–activity relationships (QSAR)

Received: 7 November 2006; Accepted: 5 February 2007

DOI: 10.1002/qsar.200610151

Abstract
The recent REACH Policy of the European Union has led to scientists and regulators to focus their attention on establishing general validation principles for QSAR models in the context of chemical regulation (previously known as the Setubal, nowadays, the OECD principles). This paper gives a brief analysis of some principles: unambiguous algorithm, Applicability Domain (AD), and statistical validation. Some concerns related to QSAR algorithm reproducibility and an example of a fast check of the applicability domain for MLR models are presented. Common myths and misconceptions related to popular techniques for verifying internal predictivity, particularly for MLR models (for instance cross-validation, bootstrap), are commented on and compared with commonly used statistical techniques for external validation. The differences in the two validating approaches are highlighted, and evidence is presented that only models that have been validated externally, after their internal validation, can be considered reliable and applicable for both external prediction and regulatory purposes.
("Validation is one of those words...that is constantly used and seldom defined" as stated by A. R. Feinstein in the book *Multivariate Analysis: An Introduction*, Yale University Press, New Haven, 1996).

1 Introduction

The New Chemicals Policy of the European Commission (REACH: Registration, Evaluation and Authorisation of Chemicals) [1] explicitly states that at chemical registration level the registrant "should include information from alternative sources (*e.g.*, from Quantitative Structure–Activity Relationships (QSARs), *etc.*) which may assist in identifying the presence or absence of hazardous properties of the substance, and which can in certain cases replace the results of animal tests. Obviously, for the purposes of the REACH legislation, it is essential to use QSAR models that produce reliable estimates, *i.e.*, validated QSAR models [2]. Model validation has been the subject of much recent debate in the scientific and regulatory communities. It was considered important to develop an internationally recognized set of principles for QSAR validation, to provide regulatory bodies with a scientific basis for making decisions on the acceptability of QSAR estimates of regulatory endpoints, and to promote the mutual acceptance of QSAR models.

Several principles for assessing the validity of QSARs were proposed in 2002, as the "Setubal Principles", at an international workshop held in Setubal (Portugal) [3]; these were then modified in 2004, by the OECD Work Programme on QSARs, as the OECD Principles [4] for QSAR validation.

To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity; (5) a mechanistic interpretation, if possible.

The most substantial revision of the Setubal principles in the new OECD principles was the unification of the previous Principles 5 (internal validation) and 6 (external validation) into the single Principle 4. However, it is important to note that, at the OECD meeting in September 2004, some experts requested this Principle be reworded as two separate Principles, like the original Setubal version, on the basis that the new approach does not sufficiently emphasize the need for external validation. Other participants felt that the single Principle was more appropriate to allow flexibility in regulatory acceptance. The author, because of her approach to QSAR modeling [5–14], is biased towards the clear separation of these two aspects of QSAR validation. In fact, as 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 in the most rigorous and realistic way possible [14].

The present paper examines the OECD Principles 2–4, paying particular attention to the differences in internal and external validations, as many authors frequently do not realize the importance of the difference in these two validation approaches. The theoretical constructs are illustrated with examples taken from the literature as well as from the recent report for ECVAM on "Evaluation of different statistical approaches to the validation of QSARs" [15].

2 Discussion

2.1 Principle 2: An Unambiguous Algorithm

According to OECD principles, for a QSAR model to be acceptable in chemical regulations it must be clearly defined, easily and continuously applicable in such a way that the calculations for the prediction of the endpoint can be reproduced by everyone, also for new chemicals. Thus, the unambiguous algorithm is characterized not only by the mathematical method of calculation used, but also by the specific molecular descriptors required in the model mathematical equation. Thus, the exact procedure used to calculate the descriptors, including compound pretreatment (*e.g.*, energy minimization, partial charge calculation, *etc*), the software employed, and the variable selection method for QSAR model development should be considered as integrative parts of the overall definition of an unambiguous algorithm.

Although all QSAR models (linear and nonlinear) are based on algorithms, the most common and transparent method, where models are described by clearly expressed mathematical equations, is Multiple Linear Regression (MLR), applied by the author in her QSAR studies (for instance in [5–13]). The reproducibility of a proposed MLR model must be guaranteed by the possibility of its future application in the defined regression equation for new chemical prediction and, thus, by the continuous availability of reproducible descriptors included in the proposed equation.

To exemplify this point, the following QSAR model for log BCF validated prediction was published in 2003 [8]:

$$\text{Log BCF} = -17.58 + 1.69\ {}^{\text{V}}\text{I}^{\text{M}}_{\text{D,deg}} - 0.45\ \text{nHAcc} + 15.65\ \text{MATS2m} - 0.36\ \text{GATS2e} - 1.64\ \text{H6p} \qquad (1)$$

R^2: 0.79; Q^2_{LOO}: 0.78; $Q^2_{\text{LMO(50\%)}}$: 0.77; Q^2_{EXT}: 0.88

To guarantee the reproducibility of the unambiguous algorithm (applying the above equation with exactly the same intercept value and coefficients, also to new chemicals), the calculated descriptors must be exactly reproducible. Unfortunately, we verified that this condition was no longer applicable, as a new version of the software Dragon calculates some of the selected descriptors differently. Thus, to have a model reproducible with the descriptors actually available (calculated by the updated version of Dragon), we redeveloped [13] the model and proposed the following new updated Eq. 2. Genetic Algorithms (GA) applied for variable subset selection again selected similar and mechanistically interpretable descriptors from the new input descriptors.

$$\text{Log BCF} = -0.74 + 2.55\ {}^{\text{V}}\text{I}^{\text{M}}_{\text{D,deg}} - 1.09\ \text{HIC} - 0.42\ \text{nHAcc} - 1.22\ \text{GATS1e} - 1.55\ \text{MATS1p} \qquad (2)$$

R^2: 0.81; Q^2_{LOO}: 0.79; Q^2_{boot}: 0.79; Q^2_{EXT}: 0.88

The same crucial problem of descriptor reproducibility can also be found in the more general and widely applied QSAR models, based on log K_{ow} as the molecular descriptor, for instance, for nonpolar narcosis in fish [16]

$$\text{Log}\ (\text{LC}_{50}) = -0.846\ \log K_{\text{ow}} - 1.39$$

This equation could be correctly applied for the prediction of new toxicity data only if the descriptor value is homogeneously calculated, but it is well known that, in apparent contradiction to its wide use, log K_{ow} is not a universal descriptor and its value varies, depending on the experimental procedure or the calculation method applied [17, 18]. Thus, the predicted data will differ if log K_{ow} used in the equation comes from a different experimental method or different software [18]. This critical point of variability in toxicity prediction, resulting from the use of differently calculated log K_{ow} values, was recently demonstrated by the author group in modeling Fathead Minnow toxicity [12].

In general terms, principle 2 can be applied differently to different modeling methods. The more portable models, such as MLR models, have surely an unambiguous algorithm (an explicit function relating the dependent variable to each predictor) and its variable coefficients allow it to be applied to any future dataset. A crucial point is the use of the same software version for the variable calculations.

2.2 Principle 3: A Defined Domain of Applicability

Another crucial problem is the definition of the Applicability Domain (AD) of a QSAR model. Not even a robust, significant, and validated QSAR model can be expected to reliably predict the modeled property for the entire universe of chemicals. In fact, only the predictions for chemicals falling within this domain can be considered reliable and not model extrapolations.

It is important to note that the AD of a model cannot be verified by studying only a few chemicals (even less than five) [19], as in such cases it could happen that extrapolated predictions are good, but probably only by chance, so it is impossible to rely on the possibility of obtaining general conclusions.

The AD is a theoretical region in chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors.

This topic was dealt with at a recent Workshop where several different approaches were proposed [20], in relation to the different model typologies. One of the simplest, and immediately applicable to multiple predictor problems with normally distributed data such as in Ordinary Least Squares (OLS) models and Partial Least Squares models (PLS), is based on distance-based measures, analogous to leverage [20–24]. It is important to note that in multiple predictor models, simple single-variable range checks are not sufficient to verify AD.

Through the leverage approach [21] (shown below), it is possible to verify whether a new chemical will lie within the structural model domain (in this case predicted data can be considered as interpolated and with reduced uncertainty, hence reliable) or outside the domain (so predicted data are extrapolated by the model and must be considered to have increased uncertainty, hence unreliable). If it is outside the model domain, a warning must be given to the users. In fact, leverage used as a quantitative measure of the model applicability domain is suitable for evaluating the degree of extrapolation, which represents a sort of compound "distance" from the model experimental space (the structural centroid of the training set) and is a measure of the influence a particular chemical's structure has on the model: chemicals close to the centroid are less influential in model building than the extreme points. A compound with high leverage in a QSAR model would reinforce the model if the compound is in the training set (good leverage), but such a compound in the test set could have unreliable predicted data, the result of substantial extrapolation of the model (bad leverage).

Prediction should be considered unreliable for compounds of high leverage value ($h > h^*$, the critical value being $h^* = 3p'/n$, where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). When the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and actual values is as high as that for the training set chemicals. Conversely, a high leverage chemical in the test set is structurally distant from the training chemicals, thus it can be considered outside the AD of the model.

To clarify recent doubts [25], it is important to point out here that each QSAR model has its own specific AD based on the training set chemicals, not just on the kind of included chemicals but also on the values of the specific descriptors used in the model itself; such descriptors are dependent on the typology of the training chemicals.

For example, a population of MLR models of similar good quality, developed by variable selection performed with GA [26], can include a hundred different models developed on the same training set but based on different molecular descriptors: even if the models are developed on the same chemicals, the AD for new chemicals can differ from model to model, depending on the specific descriptors.

To visualize the AD of a QSAR model, the plot of standardized crossvalidated residuals (R) *versus* leverage (Hat diagonal) values (h) (the Williams plot) can be used for an immediate and simple graphical detection of both the response outliers (*i.e.*, compounds with crossvalidated standardized residuals greater than three standard deviation units, $>3\sigma$) and structurally influential chemicals in a model ($h > h^*$).

Figure 1 shows the Williams plot of a model for polar narcotics in *Pimephales promelas* [12] as an example: here chemical 347 is wrongly predicted ($>3\sigma$); it is a test chemical completely outside the AD of the model, as defined by the Hat vertical line (high h leverage value). Thus, it is both a response outlier and a high leverage chemical. Two other chemicals (squares at 0.35 h) slightly exceed the critical hat value (vertical line) but are close to three chemicals of the training set (rhombus), slightly influential in the model development: the predictions for these test chemicals can be considered as reliable as those of the training chemicals. Chemical 283 is wrongly predicted ($>3\sigma$), but in this case it belongs to the model AD, being within the cutoff value of Hat. This erroneous prediction could probably be attributed to wrong experimental data rather than to molecular structure.

2.3 Principle 4: Appropriate Measures of Goodness-of-fit, Robustness and Predictivity

The R^2 [1 – (RSS/TSS), where RSS is the Residual Sum of Squares and TSS is the Total Sum of Squares] is the most widely used measure of the ability of a QSAR model to reproduce the data in the training (goodness of fit), but nothing is known of its robustness and predictivity. Several

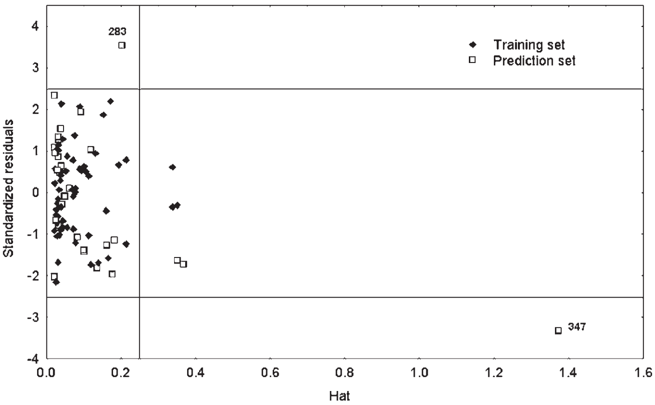

Figure 1. Williams Plot for an externally validated, recently published [12] model for Polar Narcotics. Cut-off value: 2.5 h*.

approaches to estimate model predictivity by internal validation have been suggested. Cross-Validations (CV), the most commonly used techniques for internal validation, are statistical techniques in which different proportions of chemicals are iteratively held-out from the training set used for model development (an optimal parameters' selection step) and "predicted" as new by the developed model in order to verify internal "predictivity" {*e.g.* by Q^2_{LOO} [Leave-One-Out; 1 – PRESS/TSS where PRESS is the Predictive Error Sum of Squares], Q^2_{LMO} (Leave-Many-Out), Q^2_{boot} (bootstrapping [27]).

The statistical parameters should be as high as possible (maximized) in the best model. However, it is important to remember that, in contrast to the fitting parameter R^2 which increases as more and more descriptors are added (until there is dangerous overfitting), the value of Q^2 generally increases only when the added predictors are useful in predicting left out compounds. Although LOO-CV is the sole technique that uses all the information available (very relevant mainly in small datasets), thus using the data more economically, it often overestimates the true model prediction ability. In fact, recent studies have systematically addressed the issue of Q^2_{LOO} being an incomplete measure of a model's predictive power [14, 28–33]; while it is essential at the model development step, it is still inadequate for a reliable estimate of model predictivity for completely new chemicals.

A stronger CV is LMO-CV where more than one chemical at a time is left out for the validation (groups of CV). LMO is used to counteract the slight overoptimism of LOO-CV. LMO methods vary in the data amount held-out to assess internal validity in relation to the dataset dimension, its structural heterogenicity, and the modeled response. For big datasets, it is suggested that the highest splitting possible be applied: It can be arbitrarily high as long as the remaining percentage of the data is sufficient to obtain a meaningful model. For medium and small datasets ($n < 50$), if too many chemicals are held-out (for instance 50%), the performance of the full model is generally underestimated, since only half the data are used to construct all the reduced parallel models for validation: the reduced models may not contain all the relevant structural information of the whole dataset. In fact, it has been verified on different datasets [15] that the strongest CV, able to give a more realistic idea of the true internal predictivity in small datasets (20–30 chemicals), is LMO 30%; even if it is evident that the more robust models of easily modeled endpoints (*e.g.* log K_{ow}, solubility, *etc.*) would allow a higher perturbation percentage.

Contrary to LOO- and LMO-CV, bootstrap methods [27] are more efficient and stable: they can be seen as a smoothed version of CV. Thus, this method generally gives the most accurate estimates of model performance in terms of "internal predictivity".

Although small Q^2 values, measured by the above mentioned CV (in my opinion those with $Q^2_{\mathrm{LOO}} < 0.7$) indicate models with low robustness and low internal predictive ability, the opposite is not necessarily true. In fact, while high Q^2 for each internal validation is a necessary condition for robustness and possible high predictive power in a model, it alone is not sufficient and could give an overoptimistic estimate of model predictivity for really external chemicals.

In the QSAR community, there are discordant opinions on the different outcomes of internal and external validations on QSAR models: most of the modelers still apply only internal validation approaches, convinced that this kind of validation gives an adequate and reliable idea of model predictivity. However, a relatively restricted modelers' group always externally validates models, previously internally validated by CV at the model development step; such action is based on the modelers' conviction that only models which have been additionally externally validated can be considered predictive for new chemicals ("Kubinyi paradox" [30]) [5–14, 29–43].

The supporters of using CV alone, for instance Hawkins [44, 45], Helma [46, personal communication], Asikainen [47], and Niemelå [48], point out the advantages of using many different test sets iteratively put out of the training; such test sets do not participate in the selection of variables for model development, in the turn in which they are left out. In this economical way of validation, the randomly selected data in each of the test sets are, in a sense, similar to the real life situation of unknown new chemicals. But, the author wishes to again highlight that CV techniques are essential at the variable selection step (for instance in our work [5–13, 49, 50], we always use Q^2_{LOO} as the optimization parameter for GA-based variable selection) and also to verify model robustness (stability of CV-Q^2), but provide a reasonable estimate only of the internal predictive power of a QSAR model. However, nothing can be concluded concerning the predictivity of new totally external chemicals (never included in the training set for model development, so never participating in variable selection, not even in one run), because any really new chemical has never been verified [14, 29, 31, 32].

To explain this point, it is important to take into account that CV, even if introduced at the proper stage [44, 45], is performed by iteratively developing the model on the same descriptors, selected only on the reduced training sets. These parallel models are, in turn, used to make predictions for chemicals that are iteratively omitted from each run; such chemicals being considered test chemicals and thus iteratively independent of model selection. Statistical parameters are then derived from the comparison of predicted and known experimental data. This is repeated time and time again (for tens or thousands of runs, depending on the software) and the average of Q^2 values over these runs is used. It is important to highlight that, in this way, the structural information of each chemical in the training set (the features represented by the selected descriptors) is taken into account in at least one validation run (no chemical is "new" at the end of this process). Baumann too pointed out [31, 32] that although the data used for validation are independent of the model-building process in each single split of the CV process, the resulting internal estimate of the prediction error is overoptimistic since the same data are repeatedly used to build and to assess the model. The average of the performance measures of these parallel models, taken over several iterations, is considered as the performance estimate of the final proposed model (the full model), whose equation is developed on all the available chemicals (this is the unambiguous algorithm of Principle 2). The parallel models, developed on reduced training sets for CV, are different (in terms of variable coefficients) from the final full model.

The limiting problem for efficient model validation is, obviously, data availability. When, after model development, a sufficiently large number of really new and reliable experimental data are available, the best proof of already developed model accuracy is to test the model performance on these additional data and at the same time check the chemical AD. This external validation can be considered the best. However, it is generally difficult to have data available (in useful quantity and quality) for external validation purposes for new experimentally tested compounds. Thus, in the absence of additional statistical data, external validation can be usefully applied by the QSAR modeler before the model proposal, to more precisely define the actual predictive power of the proposed model. This can be done by adequately splitting, before the model development, the available input data set into training set (for model development) and prediction set (for model predictive assessment) by different methods. At this point the underlying goal is to ensure that both the training and the prediction sets span, separately, the whole descriptor space occupied by the entire dataset, and that the chemical domain in the two datasets is not too dissimilar [33, 34, 51].

The composition of the two sets is of crucial importance. The best splitting must guarantee that the training and prediction 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 entire area occupied by representative points for the whole dataset (diversity). The more widely applied splitting methodologies are based on similarity analysis (for instance, D-optimal distance [51, 52], Kohonen Map-Artificial Neural Network (K-ANN) or Self Organizing Map (SOM) [33, 34, 53, 54]) or on random selection through activity sampling. The random splitting, while useful if applied iteratively in splitting for CV internal validation and more similar to real situations, gives very variable results when applied to statistical external validation, depending greatly on set dimension and representativity. In addition, there is a greater probability of having chemicals outside the model AD in the prediction set.

Statistical external validation should be applied at the model development step, in order to determine both the generalizability of QSAR models for new chemicals, that must obviously belong to the model AD, and the "realistic" predictive power of the model. The model must be tested on a sufficiently large number of chemicals not used in the QSAR model development (at least 20% of the complete dataset is recommended, but the most stable and robust models can also be checked on prediction set larger than the training set [39, 50]). In fact, there are in the literature examples of highly predictive models obtained by using significantly reduced training sets and larger test set. For instance, Kahn *et al.* [39] used about the 20% of the available data for model development and a big prediction set (about 80% of the original data) for model performance inspection in the soil sorption modeling. Very recently, the author group modeled the same property [50] by using even less than 15% of the original dataset in training. Moreover, it is important to highlight that, as explained above for AD, a model cannot be verified for its predictivity by checking only a few chemicals (even less than five) [19], as in such cases the results could be obtained by chance and it is impossible to obtain general conclusions.

The dimension of the available dataset is of crucial importance to all kinds of validation processes (internal or external). In dealing with small datasets, Hawkins [23, 44, 45] stated that CV, if done properly, provides a reliable picture of the fit of QSAR models. When the available sample size is small (less than 50 chemicals), he suggests to assess the model fit only by LOO, as stronger validations (LMO and external validation on a split prediction set) result in a waste of valuable information and give more variable estimates. For small sample sets (in my opinion this should be the case of less than 25 compounds) this is obviously correct, but consideration must be given to the fact that the main consequence in such cases is that the obtained models are not incontrovertibly generalizable models and great care must be taken when applying them for the prediction of new data.

The above cited papers [23, 44, 45] that support CV in any case, explain the proper stage for its application and come to the conclusion that CV is the preferable approach for estimating the predictive ability rather than external validation on split data (as used for instance in [5–14, 33–43, 49, 50]); however, there has been a crucial misunderstanding, evident in the last paper [45].

I am in full agreement with an anonymous reviewer who has stated that it is time for published papers to be cited correctly. Thus, it is very important to clarify here again that most of the supporters of external validation do not suggest external validation as an alternative to CV, but as a necessary additional validation step to be taken, CV being considered internal validation as it is based on several iterations (as explained before). Our basic idea is that a QSAR modeler must provide models that have been verified in a validation step for reliable predictive performance for future applications, mainly in the field of human and environmental protection and in regulation contexts [1]. Many of our papers [5 – 13, 49, 50] were aimed at verifying this predictive ability before model proposal. In the absence of new additional data, we assume that there is less data than is actually available: this is the reason for splitting the data in a reasonable way (commented on above) into a training set and a prediction set of "momentarily forgotten chemicals". First of all, in this approach the best model is selected by maximizing all the CV internal validation parameters, by applying CV in the proper way and step. Then, only the good models ($Q^2_{\mathrm{LOO}} > 0.7$), stable and internal predictive (with similar values of all the different CV-Q^2), are subjected to external validation on the split prediction set.

The invaluable quality of the proper CV for the selection of the best model must be fully recognized, and CV is surely necessary for model validation; however, it is not sufficient to demonstrate the external predictive ability for chemicals that never participated in the variable selection. This can be demonstrated by additional examples.

In a validation exercise of some literature models commissioned by ECVAM [15], it was verified that different QSAR models published as predictive, reporting only good fitting parameters and satisfactory CV-LOO results, are decidedly less predictive than reported (indeed, quite unpredictive) and thus not generalizable when checked by LMO (different levels of perturbation from 10 to 50%), bootstrap, and, mainly, external validation.

Additionally, it is not unusual that models with high internal predictivity, verified by different internal validation methods (CV-LOO, CV-LMO, and bootstrap) but externally less predictive or even absolutely unpredictive, are present in the population of models developed by the evolutionary techniques.

An example of this crucial point is highlighted in Table 1, which lists the first 30 models of a GA-population of PAH mutagenicity models (TA100 on 48 PAHs) [54]. Some models (in bold) appear stable and predictive by internal validation parameters (Q^2 and Q^2boot), but are less predictive (or even unpredictive: $Q^2_{\mathrm{EXT}} = 0$) when applied to external chemicals that were really never presented to the GA during model development. It is also important to note that the less predictive models (in bold) are based on different kinds of molecular descriptors, thus model instability cannot be attributed to a particular descriptor. The best combination of modeling variables must be chosen in this GA population from among the models, guaranteeing, first of all, a stable and internally predictive model (verified by CV) and, additionally, externally predictive ability (verified on the "momentarily forgotten chemicals"). From the model population of Table 1 we selected the variables of model 2 as the most stable of all the validation parameters and for three different kinds of splittings, regardless of the composition of the three training sets [54].

Y-randomization, randomly scrambling the responses, is another validation approach that must be used in parallel with CV, and must always be applied to test the significance of the derived QSAR model, highlighting the presence of apparent models, obtained only by chance correlation [14, 22, 32].

Other useful parameters to be considered are the RMSEs (Root Mean Squared Errors) calculated on different sets: on training sets (also called SDEC), CV (also called SDEP) and external prediction set. The R^2 and Q^2 values are good tests for evenly distributed data, but they are not always reliable for unevenly distributed datasets; instead RMSEs provide a more reliable indication of the fitness of the model, independently of the applied splitting. It is important to note that RMSE values must not only be low but also as similar as possible for the training, CV and external prediction sets: this suggests that the proposed model has both predictive ability (low values) as well as sufficient generalizability (similar values) [35].

A final point to highlight is that the validation of a defined QSAR model must be performed by applying the

Table 1. GA Population of models for 48 Nitro-PAH mutagenicity [49] (31 in training and 17 in prediction set), fitting (R^2), CV (Q^2_{LOO} and Q^2_{boot}) and external validation (Q^2ext) parameters.

ID	Model descriptors	R^2	Q^2_{LOO}	Q^2_{boot}	Q^2_{ext}
1	PW2 SIC1	85.7	82.44	82.36	72.27
2	PW2 CIC1	84.88	80.78	80.71	75.34
3	X1A MATS1e	82.42	79.32	79	85.75
4	Mv MATS2e	83.37	79.04	79.25	84.27
5	Mv MATS1e	81.76	78.47	78.42	74.86
6	Mv GATS2m	81.57	77.87	78.1	69.13
7	GATS1e VED2	81.07	77.64	77.68	88.06
8	Xt nPyr	80.25	77.48	77.41	81.71
9	Mv PW2	80.95	77.39	77.97	71.85
10	**PW2 IC1**	**80.89**	**77.04**	**77.32**	**60.07**
11	JGI3 VED2	80.27	76.76	76.91	66.67
12	Mp LUMO	80.78	76.54	76.55	70.13
13	Mv LUMO	80.26	76.15	76.11	63.74
14	**BELe8 HATS4u**	**80.53**	**76.1**	**76.17**	**47.59**
15	IC1 VED2	80.17	76.09	76.55	80.94
16	Xt MATS1e	80.23	76.08	75.96	86.79
17	PW2 HIC	80.14	75.99	76.16	69.62
18	SIC1 VED2	79.92	75.78	76.11	81.65
19	VED2 Hy	79.55	75.52	75.63	86.98
20	**VED2 R6u+**	**79.27**	**75.52**	**75.5**	**27.18**
21	**HATS3u R3v**	**79.55**	**75.52**	**75.23**	**0**
22	Mv MATS2m	79.25	75.37	75.64	69.21
23	Xt BELm2	79.89	75.35	75.4	69.54
24	GGI3 VED2	79.1	75.34	75.58	63.5
25	**BELe8 R4u+**	**80.06**	**75.32**	**75.3**	**50.23**
26	**SIC2 BEHm8**	**79.14**	**75.13**	**75.48**	**61.48**
27	VED2 RTe	78.65	75.13	75.32	69.76
28	CIC2 VED2	79.49	75.06	75.08	77.75
29	**SIC2 BELv5**	**79.4**	**75.02**	**75.36**	**58.31**
30	X1A LUMO	79.13	74.96	74.91	78.98

Models with reduced predictive performance in external validation in comparison to internal are shown in bold.

same equation (descriptors and coefficients) of the model developed on the training set to the new chemicals, and not by redeveloping a new equation (even if based on the same descriptors) and fitting it on the new chemicals, as has sometimes been published [55, 56]. For instance, a full model for log BCF, developed on 539 chemicals, was validated [55] "by removing 25% of the chemicals from the training set, redeveloping the QSAR and using it to predict the logBCF values of the 135 chemicals removed". The conclusion of the authors was that, as the observed and predicted values for the 135 chemicals correlated well (even with a relatively low $r^2=0.637$), the full model on 539 chemicals had good predictive ability. This conclusion should not be drawn, nor verified in this way as the two models (the full one and that redeveloped on the reduced dataset) are different models: the equation is changed as the intercept and the coefficients of the molecular descriptors are surely modified by the mutation in the training set composition. Such a change in the regression coefficients and intercept in the equations of two models (one developed on training chemicals, the other redeveloped on external validation chemicals) was observed in an evaluation of QSARs for *Tetrahymena* toxicity [56], but the conclusions were that the external validation of the model was performed and the robustness of the models was demonstrated by the similarity of the coefficients.

A comment could be: Which is the externally predictive model? Which is the unambiguous algorithm that must be applied on new chemicals for data prediction? We must always be very aware that a model redeveloped on different chemicals results in a new model (different chemicals influence the model), and this new model cannot verify the applicability for new chemical prediction of the previous model (developed on training).

Therefore, the inclusion of the term unambiguous is not redundant in the OECD Principle 2 for specifying algorithm quality.

3 Conclusions

In the past decade, QSAR model validation issues have been the subject of wide debate, from the Svante Wold school [57] to the fundamental papers of Kubinyi [30], Tropsha *et al.* ([14] and references cited therein) and Baumann [31, 32]. The aim of this paper has been to comment on three crucial OECD principles for QSAR validation.

Some interesting conclusions can be derived from this analysis.

Principle 2: The reproducibility of the molecular descriptors and the application of the coefficients of the unambiguous algorithm of Principle 1 must be guaranteed.

Principle 3: A fast and simple way to verify the AD of an MLR model is the Williams plot, a plot of standardized cross-validated residuals versus leverages (derived from the Hat diagonal) values.

Principle 4: External validation on a significant and representative number of chemicals must always supplement the necessary, but not sufficient, internal validation for predictive QSAR models. This can be done through statistical external validation, by properly splitting *a priori* the available data.

Again my suggestion is, after "The importance of being Earnest" [14], to externally validate, always and rigorously, models which have first been verified, by CV techniques, to be stable and internally predictive. This will avoid the proposal of overoptimistic, erroneously called, "predictive" QSAR models.

Acknowledgements

I wish to thank Dr Pamela Pilutti, Dr. Elisa Giani and, particularly, Ester Papa (PhD) for their precious help in the QSAR model validations and useful discussions on this crucial topic. Anonymous reviewers are gratefully acknowledged for their helpful suggestions that have led to improving the paper.

References

[1] Regulation (EC) No 1907/2006 of the European Parliament and of the Council (18/12/2006) concerning REACH; http://eur-lex.europa.eu /LexUriServ/site/en/oj/2006/l_396/l_39620061230en00010849.pdf (accessed January 22, 2007).

[2] http://ecb.jrc.it/qsar/ (accessed January 22, 2007).

[3] J. Jaworska, M. Comber, C. Auer, C. J. van Leeuwen, *Environ. Health Perspect.* **2003**, *111*, 1358–1360.

[4] http://www.oecd.org/dataoecd/33/37/37849783.pdf (accessed January 22, 2007).

[5] P. Gramatica, P. Pilutti, E. Papa, *SAR QSAR Environ. Res.* **2002**, *13*, 743–753.

[6] P. Gramatica, P. Pilutti, E. Papa, *QSAR Comb. Sci.* **2003**, *22*, 364–373.

[7] P. Gramatica, P. Pilutti, E. Papa, *Atmos. Environ.* **2003**, *37*, 3115–3124.

[8] P. Gramatica, E. Papa, *QSAR Comb. Sci.* **2003**, *22*, 374–385.

[9] P. Gramatica, V. Consonni, M. Pavan, *SAR QSAR Environ. Res.* **2003**, *14*, 237–250.

[10] P. Gramatica, P. Pilutti, E. Papa, *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1794–1802.

[11] P. Gramatica, P. Pilutti, E. Papa, *Atmos. Environ.* **2004**, *38*, 6167–6175.

[12] E. Papa, F. Villa, P. Gramatica, *J. Chem. Inf. Model.* **2005**, *45*, 1256–1266.

[13] P. Gramatica, E. Papa, *QSAR Comb. Sci.* **2005**, *24*, 953–960.

[14] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69–76.

[15] P. Gramatica, *Evaluation of different statistical approaches to the validation of Quantitative Structure–Activity Relationships*, **2004.** Available online at: http://ecb.jrc.it/DOCUMENTS/QSAR/Report_on_QSAR_validation_methods.pdf (accessed January 22, 2007).

[16] European Commission – QSAR for Predicting Fate and Effects of Chemicals in the Environment, Final report of DG XII contract No. EV5V-CT92-0211, **1995**.
[17] R. Renner, *Environ. Sci. Technol.* **2002**, *36*, 410 – 413.
[18] E. Benfenati, G. Gini, N. Piclin, A. Roncaglioni, M. R. Varý, *Chemosphere* **2003**, *53*, 1155 – 1164.
[19] E. M. Hulzebos, E. Posthumus, *SAR QSAR Environ. Res.* **2003**, *14*, 285 – 316.
[20] T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. D. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. W. Roberts, T. W. Schultz, D. T. Stanton, J. J. M. van de Sandt, W. Tong, G. Veith, C. Yang, *ATLA* **2005**, *33*, 155 – 173.
[21] A. C. Atkinson, *Plots, Transformations and Regression,* Clarendon Press, Oxford, **1985**.
[22] L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* **2003**, *111*, 1361 – 1375.
[23] D. M. Hawkins, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1 – 12.
[24] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, *ATLA* **2005**, *33*, 445 – 459.
[25] J. Tunkel, K. Mayo, C. Austin, A. Hickerson, P. Howard, *Environ. Sci. Technol.* **2005**, *39*, 2188 – 2199.
[26] R. Leardi, R. Boggia, M. Terrile, *J. Chemom.* **1992**, *6*, 267 – 281.
[27] R. Wehrens, H. Putter, L. M. C. Buydens, *Chemom. Int. Lab. Syst.* **2000**, *54*, 35 – 52.
[28] J. Shao, *J. Am. Stat. Assoc.* **1993**, *88*, 486 – 494.
[29] A. Golbraikh, A. Tropsha. *J. Mol. Graph. Model.* **2002**, *20*, 269 – 276.
[30] H. Kubinyi, *Quant. Struct-Act. Relat.* **2002**, *21*, 348 – 356.
[31] K. Baumann, *TrAC* **2003**, *22*, 395 – 406.
[32] K. Baumann, N. Stiefl, *J. Comput-Aided Mol. Design* **2004**, *18*, 549 – 562.
[33] A. Golbraikh, A. Tropsha, *J. Comput-Aided Mol. Design*. **2002**, *16*, 357 – 369.
[34] A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K. Lee, A. Tropsha, *J. Comput-Aided Mol. Design*, **2003**, *17*, 241 – 253.
[35] R. Guha, J. R. Serra, P. C. Jurs, *J. Mol. Graph. Model.* **2004**, *23*, 1 – 14.
[36] L. He, P. C. Jurs. *J. Mol. Graph. Model.* **2005**, *23*, 503 – 523.
[37] R. Perkins, H. Fang, W. D. Tong, W. J. Welsh, *Environ. Tox. Chem.* **2003**, *22*, 1666 – 1679.
[38] W. Tong, H. Fang, H. Hong, Q. Xie, R. Perkins, D. M. Sheehan, in: *Predicting Chemical Toxicity and Fate*, M. C. Cronin, D. J. Livingstone (Eds.), CRC Press, Boca Raton, **2004**, pp. 285 – 314.
[39] I. Kahn, D. Fara, M. Karelson, U. Maran, P. L. Andersson, *J. Chem. Inf. Model.* **2005**, *45*, 94 – 105.
[40] T. Oberg, *Chem Res. Toxicol.* **2004**, *17*, 1630 – 1637.
[41] T. Oberg, *Atmos. Environ*. **2005**, *39*, 2189 – 2200.
[42] S. Ren, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1679 – 1687.
[43] N. S. Zefirov, V. A. Palyulin, *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 1022 – 1027.
[44] D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 579 – 586.
[45] J. J. Kraker, D. M. Hawkins, S. C. Basak, R. Nataralan, D. Mills, *Chemom. Intell. Lab. Syst.* **2006**, online, doi:10.1016/j.chemolab.2006.03.001.
[46] C. Helma, *SAR QSAR Environ. Res.* **2004**, *15*, 367 – 383.
[47] A. H. Asikainen, J. Ruuskanen, K. A. Tuppurainen, *SAR QSAR Environ. Res.* **2004**, *15*, 19 – 32.
[48] ENV/JM/MONO (2004) 24. Available online at: http://appli1.oecd.org/olis/2004doc.nsf/linkto/env-jm-mono (2004) 24, p. 122 (accessed January 22, 2007).
[49] P. Gramatica, P. Pilutti, E. Papa, *SAR QSAR Environ. Res.* **2007**, *18*, 169 – 178.
[50] P. Gramatica, E. Giani, E. Papa, *J. Mol. Graph. Model.* **2007**, *25*, 755 – 766.
[51] M. Sjostrom, L. Eriksson, in: *Chemometric Methods in Molecular Design*, H. van de Waterbeend (Ed.), VCH, NY, **1995**, Vol. 2, p. 63.
[52] E. Marengo, R. Todeschini, *Chemom. Int. Lab. Syst.* **1992**, *16*, 37 – 44.
[53] J. Gasteiger, J. Zupan, *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503 – 527.
[54] J. Zupan, M. Novic, I. Ruisánchez, *Chemom. Int. Lab. Syst.* **1997**, *38*, 1 – 23.
[55] J. C. Dearden, N. M. Shinnawei, *SAR QSAR Environ. Res.* **2004**, *15*, 449 – 455.
[56] T. W. Schultz, T. I. Netzeva, in: *Predicting Chemical Toxicity and Fate*, M. C. Cronin, D. J. Livingstone (Eds.), CRC Press, Boca Raton, **2004**, pp. 265 – 284.
[57] S. Wold, L. Eriksson, in: *Chemometric Methods in Molecular Design*, H. van de Waterbeend (Ed.), VCH, NY, **1995**, pp. 309 – 318.