Comments on the Definition of the Q^2 Parameter for QSAR Validation

Viviana Consonni,* Davide Ballabio, and Roberto Todeschini

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1 - 20126 Milano, Italy

Received March 30, 2009

This paper deals with the problem of evaluating the predictive ability of QSAR models and continues the discussion about proper estimates of the predictive ability from an external evaluation set reported in Schüürmann G., Ebert R.-U., et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140−2145. The two formulas for calculating the predictive squared correlation coefficient Q^2 previously discussed by Schüürmann et al. are one that adopted by the current OECD guidelines about QSAR validation and based on SS (sum of squares) of the external test set referring to the training set response mean and the other based on SS of the external test set referring to the test set response mean. In addition to these two formulas, another formula is evaluated here, based on SS referring to mean deviations of observed values from the training set mean over the training set instead of the external evaluation set.

INTRODUCTION

Quantitative structure−activity relationships (QSARs) are models relating molecular descriptors encoding information on the molecular structure to a target property of molecules. There are several diverse methods aimed at quantifying the relationship between the predictors, which are the selected molecular descriptors, and a dependent variable that is a property of molecules; among these the multiple linear regression (MLR) and partial least-squares regression (PLS) are the most popular methodologies useful for regression problems.

QSAR models can be used to find out how a molecular structure should be modified in order to achieve a desired property value or to predict property values when experimental measurements are lacking. In both cases, it is of primary importance that models were reliable or in other words that the effects of the selected molecular descriptors on the target property are not determined by chance. To this concern, model quality has to be properly evaluated by the aid of statistical parameters both related to the model fit and predictive ability toward future cases.

In recent years, an ever-growing interest in QSAR research has been shown by the scientific community addressing many efforts to the development of molecular descriptors and statistical methods for analysis of data related to molecules.[1]

Because QSAR models are in some cases recognized as one of the alternative methods to produce data on the environmental and toxicological behavior of chemicals, reliability of their predictions is of primary concern.[2−9] In addition to the classical internal validation procedure such as cross-validation and bootstrapping, validation by an external test set was suggested[10−13] and recommended by the current OECD guidelines concerning QSARs.[14] Validation of models by means of objects whose data have not taken part in the process of model development is usually referred to as external validation. Objects omitted from the model fitting are here called external test objects, and they should not be confused with test objects used for internal validation of models. The main difference between the two sets is that the former is used only for validation purposes while the latter is also used for model fitting. External test objects can be arbitrarily selected from the pool of available objects or are future cases whose response becomes available only later when the model has been already fitted. Several strategies, such as cluster analysis, optimal design, sphere exclusion algorithm, or other similarity/diversity tools,[10,15−18] were proposed to design an external test set that results sufficiently representative of the entire set of data in order to give reliable estimates of the model quality. However, most of these strategies suffer from a drawback; that is, external objects are selected by using information about chemical and/or activity space of the compounds used to train models, and this, in principle, negates the status of external objects: indeed, external objects must be completely independent of the training set, or, in other words, they should not have a priori similarity relationships with training compounds.

A really independent external validation set can be obtained by random selection, but obviously this selection should be repeated several times to derive reliable estimates of the model predictive quality. While repetition of random selection is a common practice for internal model validation, it is not feasible for external validation because external objects cannot be reused to fit the model. One might divide the external test set into a number of equal-sized subsets and use them as repetitive external test sets, allowing statistical inferences but bringing out another point that is the size of the external test set. This should be large enough to give reasonable estimates of model quality specifically when a random selection was carried out. However, a large

* Corresponding author. Tel: +39-0264482820; fax +39-0264482839; e-mail: viviana.consonni@unimib.it; URL: http://michem.disat.unimib.it/chm/.

10.1021/ci900115y CCC: $40.75 © 2009 American Chemical Society
Published on Web 06/15/2009

external set can be drawn only when the data set available is very large; in all the other cases, a small-sized external test set is useless because it will likely give poor estimates of predictive ability and more importantly the final result will be a random estimate. Then, in our opinion, bootstrapping, or alternatively Monte Carlo validation, gives rise in several cases to more reliable estimates of model predictive ability, especially when a small- or medium-sized data set is available.

A completely different argument concerns the evaluation of model quality by future cases. Models that are accepted and eventually adopted as the tools to produce data on chemicals can be further evaluated, as new experimental data are going to be produced. These objects are really external objects because they never encountered the model previously. In this case, the external test set may have any distribution and be very small, even composed of only one object, especially if generation of experimental data is a long and expensive process. However, information from these new data should always be accounted for to verify validity and quality of accepted models. Definitely, it is also relevant to point out that quantification of the predictive ability of a model should always be carried out only after having evaluated if test compounds effectively fall within the model applicability domain, which means evaluating whether the model estimated from the training set is also a proper model for future candidate objects.

In an interesting paper[19] published in this journal, two different expressions for the calculation of Q^2 from an external evaluation set were discussed. These two functions are both based on the sum of squares (SS) of the external test set referring the former to the training set mean and the latter to the test set mean; however, these functions, in some cases, have drawbacks yielding unreliable estimates of the external prediction capability of a model, as shown in this paper.

An alternative function will also be discussed and compared with the others by means of calculations on ad-hoc simulated data sets and two data sets taken from the literature. No position is taken in favor or against the use of an external data set for model validation purposes or concerning how to select a "good" external data set; the goal here is the comparison of some functions defined for evaluation of model external predictive quality, evaluation of their mathematical properties, and also their advantages and drawbacks in some extreme situations. It is also important to point out that validity of a model should always be assessed by considering not only the statistical parameter chosen for quantifying the predictive ability but also all the aspects concerning reliability, accuracy, and the applicability domain of the model. Moreover, the Q^2 value, calculated by any formula, is not sufficient to prove the goodness of a model, because different Q^2 values may arise from test sets with different sizes and object distributions. Thus, the Q^2 value should always be accompanied by descriptive statistics of the test set used to compute it.

ESTIMATES OF MODEL PREDICTIVE ABILITY

Once an external test set has been built, the correct way to evaluate the external predictive ability of a model is by comparison of observed values and model predictions, and this is properly quantified in terms of the root-mean-square error (rmse):

$$\mathrm{rmse} = \sqrt{\frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2}{n_{\mathrm{EXT}}}} \tag{1}$$

where $\hat{y}_i$ is the predicted value for the ith test compound and y_i its observed value; n_{EXT} is the total number of test compounds. This parameter depends only on the mean squared deviations between predictions and observed data and can always be calculated even when there is only one test compound.

A related quantity is the residual standard deviation RSD, which also accounts for the degrees of freedom correction $(n - p)$, where p is the number of model parameters, thus yielding more statistically reliable estimates of the predictive quality.

Unfortunately, both rmse and RSD suffer from a drawback when prediction capabilities of models for different target properties should be compared: their values depend on the measure scale of the property, and thus their comparison makes no sense. This is the reason why model predictive ability is often quantified in terms of the predictive squared correlation coefficient Q^2 calculated by methods based on some form of sample reuse, such as leave-one-out and leave-more-out cross-validation:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}} \tag{2}$$

where n is the total number of objects in the entire data set, TSS is the total sum of squares, that is, the sum of squared deviations from the data set mean, and PRESS is the sum of squares of the prediction errors.

This parameter is important and has become popular because it takes values in a standardized range (i.e., ≤ 1) thus allowing trivial interpretation of its values and easy comparison of different QSAR models and the different performance of fitting and predictive abilities of a model. While calculation of Q^2 by internal validation such as cross-validation is based on a well-known and accepted formula, its derivation from an external evaluation set is not trivial and different alternatives were discussed.

An erroneous way to quantify the predictive ability is the correlation coefficient between the predicted and observed values in the external evaluation set, that is, as:

$$r = \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - \hat{\bar{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - \hat{\bar{y}})^2} \sqrt{\sum_{i=1}^{n_{\mathrm{EXT}}} (y_i - \bar{y})^2}} \tag{3}$$

where $\hat{\bar{y}}$ and $\bar{y}$ are the average values of the predicted and observed values, respectively. This coefficient is not a suitable measure of the predictive ability because it is derived from statistical parameters related to external data distribution

as, for instance, the sum of squared deviations from the external mean; this means that values of this coefficient strongly depends on how the external test set was designed. Moreover, as it is well-known, if the predicted values were any linear combination of the observed ones, that is,

$$\hat{y}_i = a + by_i$$

the correlation coefficient would be expected to be equal to 1 regardless of the magnitude of the deviations between predicted and observed values.

In the paper[19] of Schüürmann et al., two different expressions for the calculation of external Q^2, that is, Q^2 based on predictions for external test compounds, were compared. These expressions are:

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2} = 1 - \frac{PRESS}{SS_{EXT}(\bar{y}_{TR})} \quad (4)$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2} = 1 - \frac{PRESS}{SS_{EXT}(\bar{y}_{EXT})} \quad (5)$$

where $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ indicate the response means of the training set and the external test set, respectively. PRESS is the predictive sum of squares, $SS_{EXT}(\bar{y}_{TR})$ and $SS_{EXT}(\bar{y}_{EXT})$ are the total sum of squares of the external set calculated by means of the training set mean and the external set mean, respectively.

The first function is discussed in the paper of Shi et al.[20] while the second one is discussed in the paper of Hawkins.[8] The two functions differ from each other only for the mean used in the denominator for the calculation of the sum of squares ($SS_{EXT}(\bar{y}_{TR})$ and $SS_{EXT}(\bar{y}_{EXT})$). The choice of using the training set mean instead of the test set mean mainly depends on the need to have a unique reference value independent of the composition of the test set. In effect, the training set mean may be assumed as the reference no-model or, in other words, as the more reliable prediction for all the objects in the test set when no significant relationship was established between response and independent variables. Then, in Q^2_{F1} squared deviations of model predictions from the experimental data are compared with squared deviations from the no-model predictions. This formula gives reasonable values when the test set spans the whole response domain of the model because in this case the test set mean approaches the training set mean. It was recommended in the OECD guidance document[14] for external prediction capability because in the same document selection of the external test set is suggested to follow specific criteria in order to reproduce as best as possible the training set.

However, formula Q^2_{F1} has two general drawbacks, as already pointed out by Schüürmann et al.:[19] it overestimates prediction capability when test compounds are on the boundary of the response domain while it underestimates prediction capability when test compounds are in the center of the response domain. This behavior is not correct and even opposite to the one commonly expected: predictions at the center of the model are expected to be better than those on the boundary of the model domain.

Function F1 appears in the OECD guidelines for validation purposes[14] and is criticized by Schüürmann et al.[19] Through a mathematical proof, they demonstrated that the use of the training set response mean yields estimates of the predictive ability above or equal to the ones calculated on the basis of the test set mean, thus stating that function F2 gives more reliable estimates of the external predictive ability and, accordingly, the OECD guidelines should be revised.

Schüürmann et al.[19] started their mathematical analysis highlighting that a sum of squared differences between a set of variable values and a reference value is minimum if the reference value is the arithmetic mean. Then, on the basis of this consideration, they argued that the arithmetic mean of the response values of the external set objects is the reference value which minimizes the denominator:

$$\min_{\bar{y}} \left[\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2 \right] = \bar{y}_{EXT} \Rightarrow Q^2_{F2} \leq Q^2_{F1}$$

deducing, correctly, that the Q^2 calculated using the external set mean is usually lower than that calculated using the training set mean. Finally, they concluded that formula F1 yields too optimistic predictive abilities and thus formula F2 is better. However, the mathematical relationship $Q^2_{F2} \leq Q^2_{F1}$ does not imply that Q^2_{F2} is the correct way to estimate the predictive ability of a model.

Schüürmann et al. reinforced their arguments in favor of function F2 using two other considerations:

(1) Q^2 has to be lower than R^2, but this relationship is not always fulfilled using function F1;

(2) "From practical considerations, Q^2_{F2} is preferred over Q^2_{F1} because the latter requires having the training set available, which is not always the case."[19]

The first point is true when model validation is carried out by an internal validation procedure that allows sample reuse for model fitting; however, it is well-known that for some selections of external objects, predictions may be very good yielding Q^2 values larger than R^2 values. The second point is very weak because a good practice is to keep trace of the training set information for any future model evaluation and also for assessing the degree of extrapolation for future cases on the basis of the model applicability domain.

Unlike function F2, function F1 quantifies the external predictive ability taking as the reference the training set no-model defined in terms of $\bar{y}_{TR}$. In function F2 no information about the reference model is accounted for because $\bar{y}_{EXT}$ only encodes information derived from the external set, and this information changes continuously on the basis of the objects belonging to the external data set.

In our opinion, to evaluate the predictive ability, predictions for all the test objects should be evaluated independently of test set composition which can be arbitrary or dependent on the size and distribution of the new data.

It should also be noted that function F2 has a relevant drawback if the test set would be composed of only one object: it cannot be calculated, the denominator being equal

to zero! Moreover, a downward bias in predictive capability is always obtained when test set objects have similar response values.

To derive an alternative function aimed at quantifying the external predictive ability of a model, we consider the common definition of the parameter R^2 used for assessing the model fit from the training set objects:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}/n_{\mathrm{TR}}}{\mathrm{TSS}/n_{\mathrm{TR}}} = 1 - \frac{\left[\sum_{i=1}^{n_{\mathrm{TR}}} (\hat{y}_i - y_i)^2\right]/n_{\mathrm{TR}}}{\left[\sum_{i=1}^{n_{\mathrm{TR}}} (y_i - \bar{y}_{\mathrm{TR}})^2\right]/n_{\mathrm{TR}}} \quad (6)$$

where RSS is the residual sum of squares, that is, the sum of the squared deviations between observed and calculated response values over the training set, and TSS the total sum of squares, which is related to the total response variance of the training cases. In the right term of the equation, both numerator and denominator have been divided by the number n_{TR} of objects belonging to the training set, highlighting that parameter R^2 can be equivalently calculated from the ratio of RSS over TSS, each expressed in units of samples, or, in other words, from the ratio of the average squared residual over the average squared deviation from the mean.

By analogy with this expression, the external predictive ability can be calculated as the following:

$$Q^2_{\mathrm{F3}} = 1 - \frac{\left[\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2\right]/n_{\mathrm{EXT}}}{\left[\sum_{i=1}^{n_{\mathrm{TR}}} (y_i - \bar{y}_{\mathrm{TR}})^2\right]/n_{\mathrm{TR}}} = 1 - \frac{\mathrm{PRESS}/n_{\mathrm{EXT}}}{\mathrm{TSS}/n_{\mathrm{TR}}} \quad (7)$$

where the summation in the numerator runs over the external test set while that in the denominator over the training set; the number n_{TR} of training set objects and the number n_{EXT} of external objects are usually different.

In our opinion, a fundamental property of a function for the assessment of model quality from the external evaluation set is that external test objects are independent of each other. This means that the Q^2 value derived from the whole external data set and the average of the Q^2 values obtained taking separately each external object one at a time should coincide (i.e., ergodic property). This assumption implies that functions for Q^2 from external data sets cannot contain any information derived from the whole set of external data, such as mean, range, standard deviation, and so on.

In order to verify this requirement, functions F1 and F3 are first applied to the case of a single *i*th external object resulting as in the following:

$$Q^2_{\mathrm{F1}}(i) = 1 - \frac{(\hat{y}_i - y_i)^2}{(y_i - \bar{y}_{\mathrm{TR}})^2}$$

$$Q^2_{\mathrm{F3}}(i) = 1 - \frac{(\hat{y}_i - y_i)^2/1}{\overline{\mathrm{TSS}}} \quad (8)$$

where $\overline{\mathrm{TSS}} = [\sum_{i}^{n_{\mathrm{TR}}}{}_{1}(y_i - \bar{y}_{\mathrm{TR}})^2]/n_{\mathrm{TR}}$ is a fixed constant for a given training set. Function F2 cannot be calculated for a single external object, and thus it cannot satisfy the ergodic property.

In the case of function F3, the average Q^2 value coincides with the Q^2 value calculated from the external objects considered all together as demonstrated in the following:

$$\bar{Q}^2_{\mathrm{F3}} = \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} Q^2_{\mathrm{F3}}(i)}{n_{\mathrm{EXT}}} = \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} \left(1 - \frac{(\hat{y}_i - y_i)^2/1}{\overline{\mathrm{TSS}}}\right)}{n_{\mathrm{EXT}}} = \frac{n_{\mathrm{EXT}} - \sum_{i=1}^{n_{\mathrm{EXT}}} \frac{(\hat{y}_i - y_i)^2}{\overline{\mathrm{TSS}}}}{n_{\mathrm{EXT}}} = \frac{n_{\mathrm{EXT}} - \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2}{\overline{\mathrm{TSS}}}}{n_{\mathrm{EXT}}} = 1 - \frac{\left[\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2\right]/n_{\mathrm{EXT}}}{\overline{\mathrm{TSS}}} = 1 - \frac{\mathrm{PRESS}/n_{\mathrm{EXT}}}{\mathrm{TSS}/n_{\mathrm{TR}}} = Q^2_{\mathrm{F3}} \quad (9)$$

For function F1 this equivalence is not true as shown below:

$$\bar{Q}^2_{\mathrm{F1}} = \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} Q^2_{\mathrm{F1}}(i)}{n_{\mathrm{EXT}}} = \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} \left(1 - \frac{(\hat{y}_i - y_i)^2}{(y_i - \bar{y}_{\mathrm{TR}})^2}\right)}{n_{\mathrm{EXT}}} = \frac{n_{\mathrm{EXT}} - \sum_{i=1}^{n_{\mathrm{EXT}}} \left[\frac{(\hat{y}_i - y_i)^2}{(y_i - \bar{y}_{\mathrm{TR}})^2}\right]}{n_{\mathrm{EXT}}} = 1 - \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} \left[\frac{(\hat{y}_i - y_i)^2}{(y_i - \bar{y}_{\mathrm{TR}})^2}\right]}{n_{\mathrm{EXT}}} \neq Q^2_{\mathrm{F1}} = 1 - \frac{\sum_{i=1}^{n_{\mathrm{EXT}}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{\mathrm{EXT}}} (y_i - \bar{y}_{\mathrm{TR}})^2} \quad (10)$$

Finally, as already pointed out, function F2 cannot be calculated for one single object and thus also the average Q^2 over the external test set cannot be estimated; in any case, this equivalence may not be true for function F2, the denominator being defined in terms of quantities that are dependent on the whole set of external objects.

RESULTS AND DISCUSSION

Simulated Data Sets. Characteristics and performance of the three different functions F1, F2, and F3 for model fit were first investigated by the use of some simulated data sets, whose main features are summarized in Table 1. The external data sets used here are not intended to be representative examples of true validation sets; most of them represent extreme situations aimed at analyzing the mathematical behavior of the different Q^2 functions. One should be aware that validation sets should be as representative of the data to which the model will be applied as possible in order to provide reliable estimates of model predictive ability.

Table 1. Descriptive Statistics of the Simulated Data Sets Used in This Study[a]

data set ID	n_{TR}	n_{EXT}	y_{TR}^{min}	y_{TR}^{max}	$\bar{y}_{TR}$	$\bar{y}_{EXT}$	TSS
D1	50	10	0.099	9.501	5.096	5.025	425.320
D2	50	10	0.099	9.501	5.096	4.962	425.320
D3	50	10	0.099	9.501	5.096	4.861	425.320
D4	50	10	0.099	9.501	5.096	−1.482	425.320
D5	50	10	0.099	9.501	5.096	5.397	425.320
D6	50	10	0.099	9.501	5.096	5.824	425.320
D7	50	10	0.099	9.501	5.096	4.008	425.320
D8	50	10	0.099	9.501	5.096	5.067	425.320
R	50	10	0.129	9.901	4.631	4.557	424.900

[a] n_{TR} and n_{EXT} are the number of data in the training and external evaluation set, respectively; y_{TR}^{min} and y_{TR}^{max} are the minimum and maximum values in the training set; $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ are the average value of the training and test set, respectively. TSS is the total sum of squares of the training set.

The first eight simulated data sets are indicated by the symbol D. A common training set of 50 objects was generated calculating 50 y response values uniformly distributed in the range 0 to 10; the predicted values $\hat{y}$ were simulated by adding random noise to the first series of y values. Then, 8 different external test sets (D1 - D8), each composed of 10 objects, were created with different distributions. In Figure 1, the plots of predicted vs observed values for all these data sets are shown; training and test objects are highlighted by different point marks. While the training set is fixed and, accordingly, the fitted model, which has a squared correlation coefficient of 0.990, the external test set was varied in such a way as to account for some extreme situations in order to verify the general validity of the functions under investigation.

The test sets D1 and D5 are randomly distributed in the response range of the training set, the test set D2 is concentrated at the center of the response range, the test set D3 is composed of two groups of similar values located at the extremes of the range; the test set D4 consists of very similar values out of the response range, having accordingly a mean value very different from the training set mean, the test set D6 is randomly distributed with large deviations between observed and predicted values, the test set D7 is randomly distributed with small deviations between observed and predicted values, and finally the test set D8 is concentrated at the training set center and is composed of values that lead to a value of root-mean-square error (rmse) equal to that of test set D1.

A random data set, indicated by R, was also created by taking both observed and predicted values randomly from a uniform distribution (between 0 and 10) for 50 training set objects. Random responses were created similarly for 10 external test objects.

For each data set, values of the quality functions under investigation are collected in Table 2 together with R^2 values for model fit and root-mean-square error over the external evaluation set (rmse).

From the results of Table 2 several considerations can be drawn.

D1, D5, D7. All the three functions F1, F2, and F3 give reasonable estimates of the model fit when test objects are uniformly distributed and cover the whole range of the training set. This is the case of test sets D1, D5, and D7, and with respect to these three test sets the largest value of Q^2 correctly corresponds to the smallest rmse (i.e., 0.315 of test set D5). However, while both functions F1 and F2 give similar Q^2 for test sets D1 and D7 (i.e., $Q^2(F1) = 0.868$, $Q^2(F2) = 0.867$ for D1 and $Q^2(F1) = 0.849$, $Q^2(F2) = 0.832$ for D7), function F3 gives a value of 0.967 for D1 and of 0.791 for D7, their deviation being in agreement with the deviation of the corresponding rmse values (i.e., 0.534 for D1 and 1.334 for D7).

D2, D8. When the test set is composed of very similar values located at the center of the training response domain as for D2 and D8, both functions F1 and F2 crash down regardless of the actual deviations between observed and predicted values, the former because the test object values are very near the training set mean and the latter because the test object values are very near the test set mean. It is noteworthy that if we consider test sets D1 and D8 that have the same root-mean-square error (i.e., 0.534) both functions F1 and F2 give opposite estimates of the model predictive ability, that is, a Q^2 very large from test set D1 and very small, even negative, from test set D8.

D3, D4. An upward bias in F1, which is based on the squared deviations from the training set mean, is apparent when test objects are far from the training set center as for test sets D3 and D4; in particular, it is noteworthy that the Q^2 value derived from test set D4 (i.e., 0.993) is even larger than that calculated from test set D5 (i.e., 0.983) although D5 gives a root-mean-square error of 0.315 against a value of 0.563 of D4. Also function F2 tends to overestimate the predictive ability from test set D3, the test objects having large deviations from their mean. In the case of test set D4 only function F2 gives a very small value; the problem here still is that test object values are very similar to each other.

D6. All the three functions behave similarly in the case of test set D6, giving very small values which account for the large deviations between observed and predicted values.

R. All the three functions give correct small values for the completely random data set: these agree with the small R^2 of the model.

Rank Correlation between Q^2 and rmse. In our opinion, the function chosen for assessing the model fit from an external test set should give values in agreement with those of the root-mean-square error over the test set, because this quantity is based only on the deviations between observed and predicted values, which are the model residuals. Then, the ranking of the predictive ability estimates from the eight test sets (D1−D8) on the basis of Q^2 (Table 2) should perfectly match the ranking given by the root-mean-square error (rmse). The degree of agreement between the two rankings was determined by the Spearman rank correlation between rmse and Q^2 values from each of the functions in analysis. Results of the rank correlation analysis are reported in Table 3. Only function F3 is able to perfectly reproduce the ranking of rmse.

PAH Data Set. A further investigation was based on a real data set taken from the literature. This is constituted by the octanol−water partition coefficients (logP) of 37 polyaromatic hydrocarbons (PAH).[21] Two different models (A and B) for prediction of logP, based on molecular descriptors as the independent variables, were considered: model A is a linear relationship of molecular weight and a topological index, while model B uses the total number of

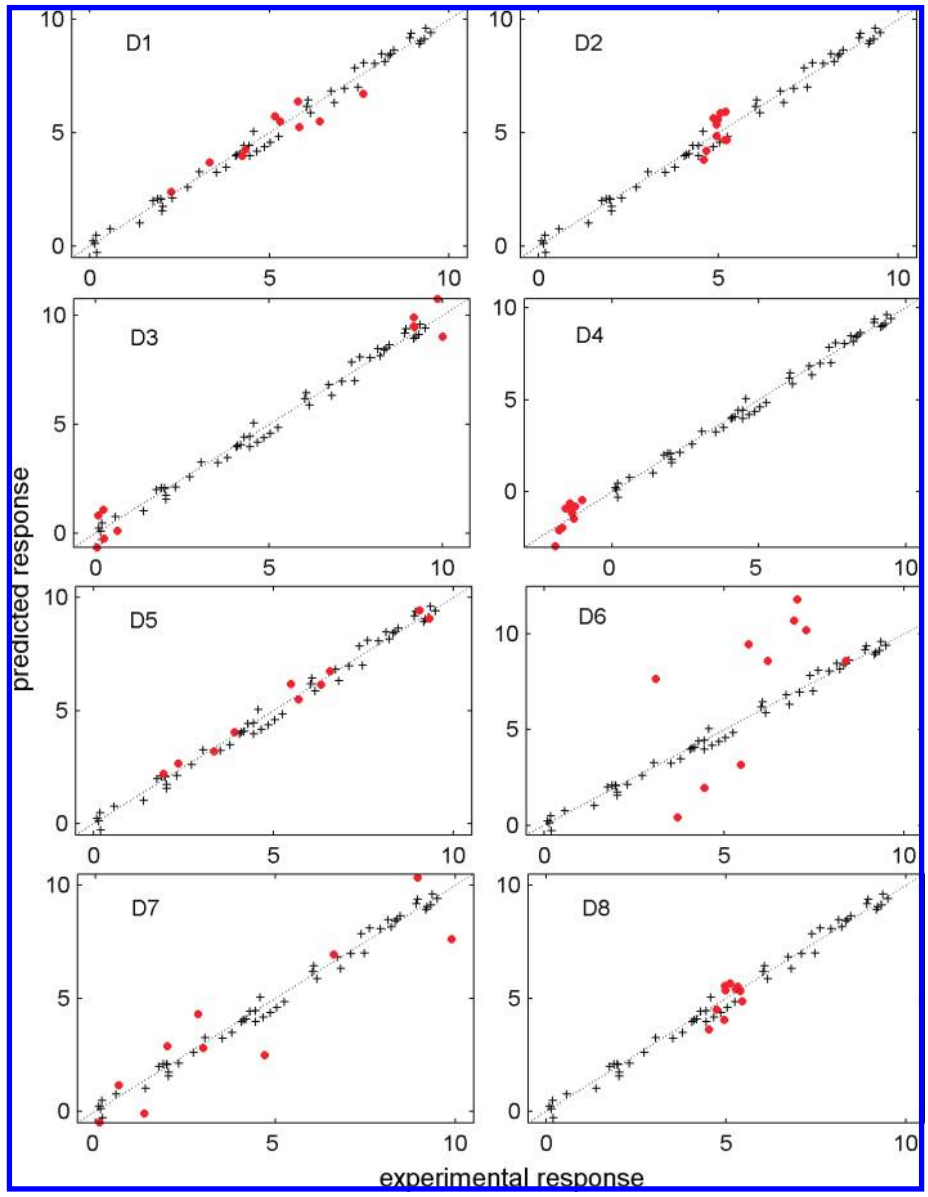

Figure 1. Predicted vs observed values for the simulated data sets D1−D8. The training set is fixed while the ten external test objects are distributed in different ways with respect to the model response domain. Training objects are indicated by crosses, while test objects are indicated by dot marks.

Table 2. Model Fit Estimates for Different Simulated Data Sets[a]

data set	R^2	Q^2(F1)	Q^2(F2)	Q^2(F3)	rmse
D1	0.990	0.868	0.867	0.967	0.534
D2	0.990	−5.060	−7.562	0.957	0.608
D3	0.990	0.977	0.977	0.943	0.696
D4	0.990	0.993	−3.485	0.963	0.563
D5	0.990	0.983	0.983	0.988	0.315
D6	0.990	−2.535	−3.269	−0.280	3.300
D7	0.990	0.849	0.832	0.791	1.334
D8	0.990	−2.507	−2.543	0.967	0.534
R	−0.493	−0.088	−0.089	−0.038	2.971

[a] R^2 is the squared correlation coefficient of the model assessed by the training set values. Q^2(F1), Q^2(F2), and Q^2(F3) are the model predictive ability from an external evaluation set on the basis of three different functions. rmse is the root-mean-square error over the external test set.

Table 3. Spearman Rank Correlation (absolute values) between Different Estimates of the Model Predictive Ability for Data Sets D1−D8

	Q^2(F1)	Q^2(F2)	Q^2(F3)	rmse
Q^2(F1)	1	0.50	0.43	0.43
Q^2(F2)		1	0.36	0.36
Q^2(F3)			1	1
rmse				1

bonds in molecules as the only independent variable. In both cases, a training/test splitting was carried out, selecting 20% (7) of the available compounds for the external test set and leaving the remaining ones (30) in the training set; this splitting was repeated three times, generating three different combinations of training and test sets. Descriptive statistics of the three PAH data sets are reported in Table 4.

The first training/test splitting was randomly carried out (PAH_R); the two models derived from the random training set are indicated by PAH_A_R and PAH_B_R, respectively. The second splitting was carried out by selecting as the seven

Table 4. Descriptive Statistics of the PAH Data Sets[a]

data set ID	n_{TR}	n_{EXT}	y_{TR}^{min}	y_{TR}^{max}	$\bar{y}_{TR}$	$\bar{y}_{EXT}$	TSS
PAH_R	30	7	2.920	6.900	4.962	5.393	30.375
PAH_Up	30	7	2.920	5.910	4.721	6.426	19.369
PAH_Centre	30	7	2.920	6.900	5.041	5.053	36.795

[a] n_{TR} and n_{EXT} are the number of data in the training and external evaluation set, respectively; y_{TR}^{min} and y_{TR}^{max} are the minimum and maximum values in the training set, and $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ are the average value of the training and test set, respectively. TSS is the total sum of squares of the training set.

Table 5. Model Fit Estimates for PAH Data Sets[a]

data set	R^2	Q^2(F1)	Q^2(F2)	Q^2(F3)	rmse
PAH_A_R	0.972	0.873	0.843	0.877	0.353
PAH_B_R	0.932	0.981	0.977	0.982	0.135
PAH_A_Up	0.963	0.961	0.254	0.816	0.345
PAH_B_Up	0.975	0.856	−1.766	0.319	0.663
PAH_A_Centre	0.962	0.324	0.321	0.986	0.132
PAH_B_Centre	0.945	−0.212	−0.218	0.974	0.177

[a] R^2 is the squared correlation coefficient of the model assessed by the training set values. Q^2(F1), Q^2(F2), and Q^2(F3) are the model predictive ability from an external evaluation set on the basis of three different functions. rmse is the root-mean-square error over the external test set.

test compounds those compounds with the largest response values (PAH_Up); the models derived from the remaining training compounds were referred to as PAH_A_Up and PAH_B_Up, respectively. In the third case, the external test set was composed of the seven compounds with values concentrated in the middle of the response interval (PAH_Centre); the two corresponding models were called PAH_A_Centre and PAH_B_Centre, respectively. Predicted vs observed values for the different PAH models are shown in Figure 2.

Model quality estimates from function F1, F2, and F3 are in Table 5.

Also from these results it follows that F3 provides reasonable estimates of the model predictive ability; it gives a very small value for the data set PAH_B_Up, but it is in agreement with the rmse value. Moreover, Q^2 values larger than R^2 are obtained for data sets PAH_B_R, PAH_A_Centre, and PAH_B_Centre, but these are also characterized by the smallest rmse values meaning that deviations between observed and predicted values for the test objects are very small. Both functions F1 and F2 confirm the same problems previously discussed:

• Function F1 has an upward bias when responses of the test objects are far from the training set mean as for the case PAH_B_Up where residuals of test objects are remarkably large and, accordingly, a scarce model quality would be expected. Conversely, it shows a downward bias when responses of test objects are all around the training set mean; for instance, in the case of PAH_A_Centre, Q^2 is 0.324 although responses are well estimated for all the test objects, the rmse being the smallest one (i.e., 0.132).

• Function F2 is not able to provide reliable estimates of model predictive ability when responses of all test objects are very similar to each other as for _Centre and _Up models. Moreover, in the case of the data set PAH_Up, function F2

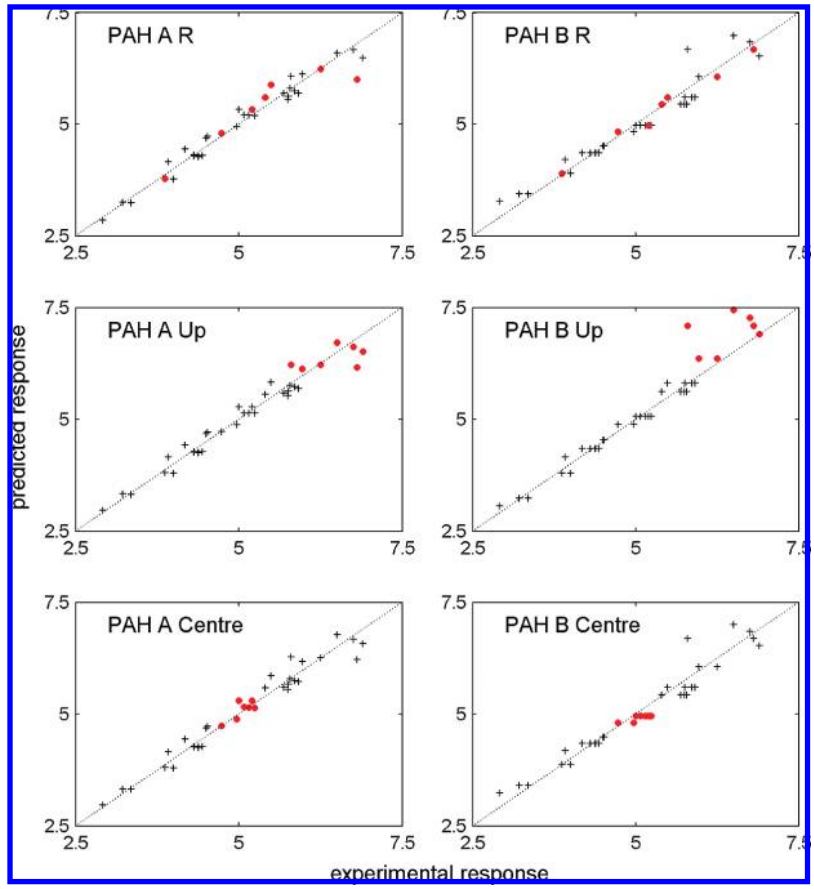

Figure 2. Predicted vs observed values for the PAH models. Training objects are indicated by crosses, while test objects are indicated by dot marks.

Table 6. Descriptive Statistics of the PCB Data Sets[a]

data set ID	n_{TR}	n_{EXT}	y_{TR}^{min}	y_{TR}^{max}	$\bar{y}_{TR}$	$\bar{y}_{EXT}$	TSS
PCB_R1	71	17	4.670	9.020	6.538	6.260	52.965
PCB_R2	71	17	4.670	9.020	6.538	6.494	52.965
PCB_R3	71	17	4.780	9.020	6.535	6.271	52.120
PCB_Up	71	17	4.670	7.500	6.227	7.559	29.940
PCB_Centre	71	17	4.670	9.020	6.419	6.757	58.494
PCB_Down	71	17	5.170	9.020	6.695	5.606	40.717

[a] n_{TR} and n_{EXT} are the number of data in the training and external evaluation set, respectively; y_{TR}^{min} and y_{TR}^{max} are the minimum and maximum values in the training set, and $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ are the average value of the training and test set, respectively. TSS is the total sum of squares of the training set.

Table 7. Model Fit Estimates for PCB Data Sets[a]

data set	R^2	Q^2(F1)	Q^2(F2)	Q^2(F3)	rmse
PCB_R1	0.872	0.773	0.736	0.832	0.354
PCB_R2	0.872	0.768	0.767	0.879	0.300
PCB_R3	0.876	0.766	0.735	0.809	0.375
PCB_Up	0.752	0.960	0.804	0.786	0.300
PCB_Centre	0.873	0.610	0.237	0.889	0.302
PCB_Down	0.825	0.930	0.656	0.819	0.322

[a] R^2 is the squared correlation coefficient of the model assessed by the training set values. Q^2(F1), Q^2(F2), and Q^2(F3) are the model predictive ability from an external evaluation set on the basis of three different functions. rmse is the root-mean-square error over the external test set.

provides nearly similar small values of Q^2 for both models A (i.e., 0.254) and B (i.e., −1.766), although predictions from model A (i.e., rmse = 0.345) are quite better than those from model B (i.e., rmse = 0.663).

PCB Data Set. Further results were obtained from another real data set constituted by the water solubility of 88 polychlorobiphenyls (PCB) taken from the literature.[21] QSAR models were generated on the basis of two topological indices that encode information on molecular structures. Also in this case, the external test set was generated by selecting 20% (17) of the available compounds and leaving the remaining ones in the training set (71). Six different training/test splitting were carried out yielding six different models because the training set changed each time. Descriptive statistics of the six PCB data sets are reported in Table 6.

The first three data splitting were randomly generated; the corresponding models were indicated by PCB_R1, PCB_R2, and PCB_R3. The fourth external test set was composed of the 17 compounds with the largest response values (PCB_Up), the fifth external set of the 17 compounds with central response values (PCB_Centre), and finally the sixth external set of the 17 compounds with the smallest response values (PCB_Down). Predicted vs observed values for training and test sets are shown in Figure 3, while R^2 and Q^2 values from the three functions for all the six PCB models are collected in Table 7. Also in this case function F3 is demonstrated to be the best one, giving values which better agree with rmse. Functions F1 and F2 still show a random variability mostly depending on the distribution of the test objects.

***Q*² Invariance to Random Sampling.** In order to better investigate the function behavior on test sets with the same rmse (as for D1 and D8 in the previous example), another simulated example was studied. A fixed training set of 500 objects was created, calculating y responses uniformly

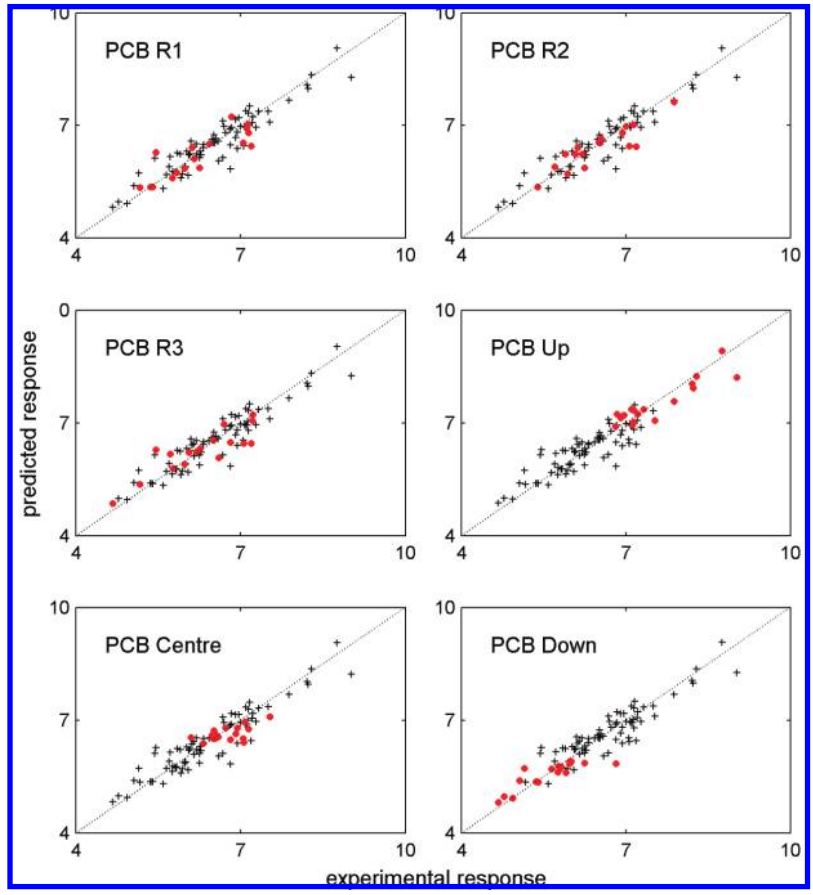

Figure 3. Predicted vs observed values for the PCB models. Training objects are indicated by crosses, while test objects are indicated by dot marks.

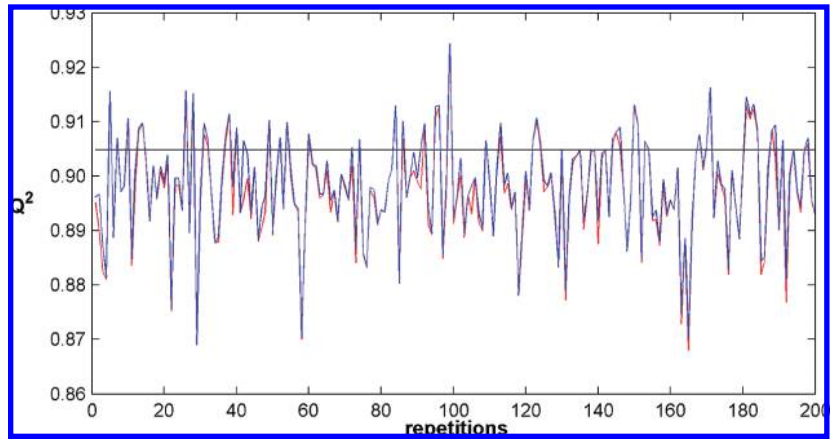

Figure 4. Q^2 values estimated by the functions F1 (blue), F2 (red), and F3 (black) from 200 test sets with the same rmse, each being composed of 100 test objects.

Table 8. Averages and Standard Deviations of the Q^2 from 1000 Random Test Sets of Various Sizes[a]

	Q^2(F1)		Q^2(F2)		Q^2(F3)		$\bar{y}_{TR}$		$\bar{y}_{EXT}$	
test set size	mean	std dev	mean	std dev	mean	std dev	mean	std dev	mean	std dev
2	0.637	3.560	−55.577	788.431	0.953	0.030	17.178	0.009	17.326	4.620
3	0.906	0.214	0.587	2.043	0.953	0.025	17.178	0.012	17.232	3.832
5	0.926	0.142	0.899	0.215	0.953	0.019	17.179	0.015	17.093	2.924
10	0.945	0.032	0.937	0.041	0.953	0.014	17.178	0.022	17.098	2.173
20	0.951	0.017	0.947	0.019	0.953	0.009	17.181	0.029	17.073	1.441
50	0.952	0.010	0.951	0.011	0.953	0.006	17.179	0.050	17.169	0.954
100	0.953	0.006	0.952	0.007	0.953	0.004	17.180	0.068	17.146	0.612
200	0.953	0.004	0.953	0.004	0.953	0.003	17.176	0.101	17.190	0.405
300	0.953	0.003	0.953	0.003	0.953	0.002	17.187	0.136	17.160	0.318
500	0.953	0.002	0.953	0.002	0.953	0.002	17.184	0.219	17.174	0.219

[a] The true model R^2 is 0.953 and the data mean from all the 1000 values is 17.1787. $\bar{y}_{TR}$ and $\bar{y}_{EXT}$ are the averages of the training set and test set means.

distributed in the range 0 − 10; the predicted values $\hat{y}$ were simulated by adding random noise, obtaining rmse equal to 0.570 and R^2 equal to 0.962. Then, test sets of 100 objects were created by randomly generating y responses uniformly distributed along the range 0.1−9.9. The predicted $\hat{y}$ values of the test objects were generated in such a way that the same rmse value equal to 0.898 was obtained each time. The test set random generation was repeated 200 times, and functions F1, F2, and F3 were calculated. In Figure 4, the three function values for each iteration are shown.

As expected, the function F3 gives always the same Q^2 value, because it is independent of the test set distribution, depending only on rmse. On the other side, both functions F1 and F2 reveal a nonconstant behavior, giving oscillating estimates of the model predictive ability Q^2 approximately in the range from 0.87 to 0.92.

As it is well-known, R^2 and rmse for training sets, having the same TSS, are correlated 1 (or, more precisely, −1); that is, the two indices give exactly the same information, although in two different scales. Therefore, it is quite natural that the same behavior was also required to functions estimating the model quality from validation sets: unlike functions F1 and F2, function F3 is correlated 1 to rmse, as also shown in Table 3.

***Q^2* Invariance to Sampling Size.** Following the same approach as that used by Hawkins,[8] the different mathematical behaviors of functions F1, F2, and F3 were further investigated. A simulated data set of 1000 objects was created on the basis of two independent variables x and one dependent variable y; a linear regression model with R^2 of 0.953 was estimated, with a training set mean equal to 17.1787. Then, a number of random test sets of various sizes were taken from the pool of 1000 objects to assess the model predictive ability. In this case, test objects are not really external objects, but this is not relevant to the main goal of this paper, that is, evaluating the differences among Q^2 values by functions F1, F2, and F3.

We used random test sets of size 2, 3, 5, 10, 20, 50, 100, 200, 300, and 500. For each size, 1000 random extractions were carried out and, for each of these, Q^2 values by functions F1, F2, and F3 were computed together with training set and test set means. The averages of all these quantities for the different sample sizes are collected in Table 8.

As it was already observed by Hawkins, the model predictive ability by F2 increases with the number of objects that are included in the external set and approaches the true R^2 when the test set is composed of about 20% of the available objects. Moreover, the results show that function F3 seems to be independent of the number of test objects, the Q^2 being constant and equal to the true R^2 of 0.953. This means that the model predictive ability is equally well estimated by using different numbers of test objects. On the contrary, the Q^2 value by function F2 increases with the number of test objects as a consequence of the fact that the test set mean approaches the true training set mean when the number of test objects, randomly extracted from the pool of 1000 objects, increases. When the test set size is increased, a convergence of function F2 toward function F1, which is based on the training set mean, is also noticed. Function F2 was demonstrated to have some drawbacks when the external test set is composed of objects with particular or extreme distributions, but it can work reasonably for repeated random selections.

Function F1 confirms a downward bias especially in the case of small-sized test sets. Moreover, it is noteworthy that, when 500 out of 1000 objects are used to validate the model, both functions F1 and F2 converge toward F3 because the number of test objects equals the number of training objects and the training and test set means nearly coincide.

Concluding Remarks. Finally, some general considerations about the evaluation of the predictive ability of a model naturally arise from the above discussion, in particular considering that commonly not much data are available for modeling. On one hand, the use of all the available information, that is, all the objects, to build a robust regression model seems to be a pressing temptation. On the other hand, an external data set could be useful to test the model predictive ability, but this should be composed of objects that have never encountered the training set objects previously: the optimal way to do this seems to be a random selection of the external objects! In effect, any deterministic algorithm for the selection of external objects such as optimal design, cluster analysis, sphere exclusion algorithm, or other similarity/diversity tools is in contrast to this requirement, the external objects being selected on the basis of some comparison with the training set objects. However, in the case of small- and medium-sized data sets, a random selection performed only once brings results that closely depend on the selection. The alternative is to repeat the random selection several thousands of times, and this is exactly what is done by validation techniques such as bootstrap, exhaustive leave-more-out cross-validation, or Monte Carlo validation. By these techniques, the test set is really built each time without any consideration about the training set, and the average predictive ability can be considered as a good estimate of the "true" predictive ability, still leaving the possibility to use all the objects to build the final regression model.

Moreover, it can also be noted that all the validation techniques based on a fixed number of test objects, as for instance cross-validation and Monte Carlo validation, are biased by the test set size, whereas this bias is not present in bootstrapping where the test set size can vary in each iteration.

CONCLUSIONS

Model validation is nowadays recognized as a mandatory stage in QSAR model development. However, different estimates of model predictive ability can be computed depending on the validation technique, the composition of test set, and the function used to quantify the deviations between predicted and observed values in the test set. This paper focused on this last point, comparing, from a mathematical point of view, the behavior of Q^2 values computed by three different functions over some data sets.

Functions F1 and F2 are based on the sum of squares SS of the external test set referring the former to the training set mean and the latter to the test set mean; function F3 is instead based on the mean squares of the training set in order to be independent of the distribution of test objects.

Functions F1 and F2 were demonstrated to suffer from some drawbacks when the external test objects are not uniformly distributed over the range of the training set. On the contrary, function F3 appeared independent of the external object distribution and satisfied the ergodic property, a condition that we consider a fundamental requirement. Nevertheless, despite the mathematical independence of size and distribution of external objects, statistical information about the validation set is always necessary to evaluate reliability of Q^2 provided by F3, because the larger and more representative the set of data, the more confident the final result. Moreover, this function is valid both for internal validation such as cross-validation or bootstrap and for external validation.

ACKNOWLEDGMENT

This study has been financed by funds PRIN 2007 (National Ministry of University and Research and University of Milano-Bicocca, code 2007R57KT7).

Note Added after ASAP Publication. This article was released ASAP on June 15, 2009, with minor errors in the text and Table 8. A new version was posted on June 24, 2009 with minor errors in equation 3. The final corrected version was published on July 1, 2009.

REFERENCES AND NOTES

(1) Todeschini R.; Consonni V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.
(2) Efron, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* **1983**, *78*, 316–331.
(3) Efron, B. Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.* **1987**, *82*, 171–200.
(4) Cramer, R. D., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.
(5) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.
(6) Eriksson, L.; Jonsson, J.; Berglind, R. External Validation of a QSAR for the Acute Toxicity of Halogenated Aliphatic Hydrocarbons. *Environ. Toxicol. Chem.* **1993**, *12*, 1185–1191.
(7) Wold S.; Eriksson L. Statistical Validation of QSAR Results. Validation Tools. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers: Weinheim, Germany, 1995; Vol. 2, pp 309−318.
(8) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
(9) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
(10) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241–253.
(11) 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*, 69–77.
(12) Schmuker, M.; Givehchi, A.; Schneider, G. Impact of different software implementations on the performance of the *Maxmin* method for diverse subset selection. *Mol. Div.* **2004**, *8*, 421–425.
(13) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
(14) Organization for Economic Co-operation and Development. Guidance document on the validation of (quantitative) structure-activity relationship ((Q)SAR) models. OECD Series on Testing and Assessment 69. OECD Document ENV/JM/MONO, 2007, pp 55−65.
(15) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* **1997**, *15*, 372–385.
(16) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training-Test Sets and Consensus Modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794–1802.
(17) Guha, R.; Serra, J. R.; Jurs, P. C. Generation of QSAR sets with a self-organizing map. *J. Mol. Graphics Modell.* **2004**, *23*, 1–14.
(18) He, L.; Jurs, P. C. Assessing the reliability of a QSAR model's predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503–523.
(19) Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140–2145.
(20) Shi, L. M.; Fang, H.; Tomg, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheenan, D. M. QSAR Models Using a Large Diverse Set of Estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186–195.
(21) Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. Part 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705.

CI900115Y