External Validation and Prediction Employing the Predictive Squared Correlation Coefficient − Test Set Activity Mean vs Training Set Activity Mean

Gerrit Schüürmann,*[,†,‡] Ralf-Uwe Ebert,[†] Jingwen Chen,[§] Bin Wang,[§] and Ralph Kühne[†]

Department of Ecological Chemistry, UFZ Helmholtz Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany, Institute for Organic Chemistry, Technical University Bergakademie Freiberg, Leipziger Strasse 29, 09596 Freiberg, Germany, and Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), Department of Environmental Science and Technology, Dalian University of Technology, Linggong Road 2, Dalian 116024, China

Received July 25, 2008

The external prediction capability of quantitative structure−activity relationship (QSAR) models is often quantified using the predictive squared correlation coefficient, q^2. This index relates the predictive residual sum of squares, *PRESS*, to the activity sum of squares, *SS*, without postprocessing of the model output, the latter of which is automatically done when calculating the conventional squared correlation coefficient, r^2. According to the current OECD guidelines, q^2 for external validation should be calculated with *SS* referring to the training set activity mean. Our present findings including a mathematical proof demonstrate that this approach yields a systematic overestimation of the prediction capability that is triggered by the difference between the training and test set activity means. Example calculations with three regression models and data sets taken from literature show further that for external test sets, q^2 based on the training set activity mean may become even larger than r^2. As a consequence, we suggest to always use the test set activity mean when quantifying the external prediction capability through q^2 and to revise the respective OECD guidance document accordingly. The discussion includes a comparison between r^2 and q^2 value ranges and the q^2 statistics for cross-validation.

INTRODUCTION

Qualitative and quantitative structure−activity relationships (QSARs) rely on empirically derived correlations between a target property (activity) of chemical compounds and one or more predictor variables associated with their molecular structures. For many years, it has been recognized that calibration statistics do not inform well about the prediction capability of QSAR models, the latter of which can be addressed through statistical procedures such as cross-validation, bootstrapping, and target value scrambling and with external test sets.[1−8]

More recently, it was argued that similarity to the training set compounds could be used as an indicator for the expected model reliability.[9] To this end, different methods such as kernel density estimation of the training set coverage of the descriptor space[10,11] and descriptor-based discrimination between high and low training set regression residuals have been developed.[12] Moreover, structural similarity, encoded through atom-centered fragments,[13,14] has been used to correct - without refitting the model - for local (structure-specific) prediction errors,[13] to serve as criterion for selecting best-suited models,[15] and to directly predict a target property through a *k* nearest neighbor approach.[16] Because QSARs - besides other alternative methods - are envisaged to support the toxicological and ecotoxicological evaluation of chemical substances under the new European chemical legislation REACH[17] and are already in such use in the U.S.A., there is a need for agreed procedures to address their application domain,[18] the latter of which represents those compounds for which the model can be applied with statistical confidence.

For the quantitative evaluation of the prediction capability of QSAR models, an index related to - but different from - the squared correlation coefficient, r^2, has become a popular measure. Initially termed prediction r^2 or accuracy of prediction,[1] it relates the predictive residual sum of squares, *PRESS*,[19] to the activity sum of squares, *SS*, and is now often called q^2. As opposed to r^2, q^2 does not involve postprocessing in terms of least-squares rescaling of the predicted values but compares the untuned output of the original model to the experimental data (for mathematical details see below).

While *SS* (that contributes to q^2) would usually refer to the arithmetic mean of the target values, it has been suggested (without explanation) to evaluate *SS* with respect to the training set activity mean when quantifying q^2 for external test sets.[20] In the meantime, this approach has become part of the OECD guidelines for the external validation of QSARs.[21] As shown below through a mathematical proof and calculation examples, however, use of the training set activity mean yields estimates of the prediction capability above or (at best) equal to the ones calculated when employing - in accord with the original introduction of q^2 - the respective test set activity mean. Moreover, employing the training set mean may even result in q^2 values larger than r^2 for external test sets. It follows that external prediction performance if evaluated according to the present OECD

* Corresponding author phone: +49-341-235-1262; fax: +49-341-235-1785; e-mail: gerrit.schuurmann@ufz.de.
† UFZ Helmholtz Centre for Environmental Research.
‡ Technical University Bergakademie Freiberg.
§ Dalian University of Technology.

10.1021/ci800253u CCC: $40.75
Published on Web 10/28/2008

guidance document tends to yield too optimistic results (within the framework of q^2 statistics). Thus, the present findings suggest using the test set activity mean when calculating q^2 for external test sets and revising the OECD guidance document accordingly.

MATERIALS AND METHODS

Regression Models and Data Sets. For comparative evaluations of the prediction performance according to two different formulas for the q^2 statistics as described in the next section, three regression models and data sets have been taken from the literature.

Two application examples deal with models to predict the sorption of organic compounds into soil organic matter, quantified through the respective logarithmic sorption constant normalized for organic carbon content, log K_{oc}.[22] First, the preliminary model calibrated using 80% of the data set (457 compounds) was employed, using the remaining 20% (114 compounds) as well as three associated subsets covering 2/3 of the experimental log K_{oc} range as external prediction sets. The subsets were generated such that one covers the lower 2/3 range of the experimental log K_{oc} values of the prediction set ("lower 2/3 subset"), one the respective center 2/3 range ("centre 2/3 subset"), and one the upper 2/3 range ("upper 2/3 subset"). In this way, the influence of both the reduction in target values range and of the shift in the subset-specific arithmetic mean of the target value (log K_{oc}) could be analyzed.

Second, the final log K_{oc} regression model calibrated on 571 compounds was applied to an external data set collected from two literature sources[23,24] as described earlier.[22] Again, the lower, center, and upper 2/3 subsets were also generated and included in the statistical analysis.

Third, the prediction performance of an increment method to predict the logarithmic water solubility at 25 °C, log S_w, from molecular structure[25] was analyzed using an external test set. The latter was selected from our in-house database[26] such that all test set compounds have water solubilities within the experimental value range of the original training set and belong to the structural domain of the original model as defined through second-order atom-centered fragments.[13−15] As with the other two examples, prediction subsets covering the lower, center, and upper 2/3 of the target value range (here: log S_w range) were included in the analysis.

Statistical Parameters. Calibration and prediction performance of the three regression models have been evaluated using the following statistical parameters: squared correlation coefficient (r^2), predictive squared correlation coefficient (q^2), predictive squared correlation coefficient using the training set mean as reference value (q^2_{tr}), root-mean-square error of prediction (*rms*), and systematic error (*bias*). The mathematical formulas for r^2, q^2, and q^2_{tr} are given in the next section.

MATHEMATICAL ANALYSIS OF THE Q^2 STATISTICS

Mimimum of the Sum of Squares. Consider the sum of squares

$$f(x_i, x_0) = \sum_{i=1}^{N} (x_i - x_0)^2 \quad (1)$$

In eq 1, x_i denotes the i-th observed value of the property of interest, N is their total number, and x_0 is some associated reference value. To minimize $f(x_i, x_0)$ with respect to x_0, the necessary condition is that its partial derivative is equal to zero:

$$\frac{\partial f}{\partial x_0} = \sum_{i=1}^{N} \frac{\partial}{\partial x_0}(x_i - x_0)^2 = \sum_{i=1}^{N} 2(x_i - x_0)(-1) = (-2)\left(\sum_{i=1}^{N} x_i - N \cdot x_0\right) \equiv 0 \quad (2)$$

Division by (−2) and rearrangement of the right-hand side (rhs) of eq 2 yields

$$\sum_{i=1}^{N} x_i = N \cdot x_0 \quad (3)$$

which implies that the necessary condition for x_0 to minimize $f(x_i, x_0)$ is

$$x_0 = \frac{1}{N}\sum_{i=1}^{N} x_i \equiv x_{mean} \quad (4)$$

It follows that introducing the arithmetic mean of all x_i data, x_{mean}, for x_0 leads to an extremum of the sum of squares $f(x_i, x_0)$. That this extremum is indeed a minimum (and not a maximum) can be seen from the fact that the second partial derivative of $f(x_i, x_0)$ is positive:

$$\frac{\partial^2}{\partial x_0^2}\left(\sum_{i=1}^{N} (x_i - x_0)^2\right) = \frac{\partial}{\partial x_0}\left(-2\sum_{i=1}^{N} (x_i - x_0)\right) = (-2)(-N) = 2N > 0 \quad (5)$$

In statistics, $f(x_i, x_{mean})$ is often called *SS*, the sum of squares of the differences between observed values (x_i) and their mean (x_{mean}):

$$SS = \sum_{i=1}^{N} (x_i - x_{mean})^2 \quad (6)$$

In the context of regression relationships between target properties y_i and independent variables x_i, *SS* evaluated for y_i (and their associated arithmetic mean, y_{mean}) contributes to the calibration and prediction performance in terms of r^2 and q^2 values as outlined in the next section.

Prediction Power in Terms of q^2. For a model to predict observations y_i of a target property Y through some regression relationship that yields the values y_i^{pred} as estimates of y_i, the predictive residual sum of squares, *PRESS*, is defined as sum of the squared differences between predicted and observed values:[19]

$$PRESS = \sum_{i=1}^{N} (y_i^{pred} - y_i)^2 \quad (7)$$

The predictive squared correlation coefficient, q^2, as measure of the prediction performance of the model[1,2] can then be written as

$$q^2 = 1 - \frac{\sum_{i=1}^{N} (y_i^{pred} - y_i)^2}{\sum_{i=1}^{N} (y_i - y_{mean})^2} \equiv 1 - \frac{PRESS}{SS} \quad (8)$$

In eq 8, *SS* and *PRESS* should of course refer to the same data set which is the one used for evaluating q^2, implying in

particular that y_{mean} is the associated arithmetic mean. It follows that when analyzing the model performance in terms of q^2 for an external test set with N experimental data y_i, y_{mean} should be their arithmetic mean and not the respective mean of the original training set that had been used to derive the model.

The latter, however, is currently proposed in the OECD guidance document for quantifying the external prediction capability of QSAR models,[21] following a respective suggestion from literature.[20] It corresponds to a slightly different definition of q^2, which we now call q^2_{tr} for the sake of clarity:

$$q^2_{tr} = 1 - \frac{\sum_{i=1}^{N} (y_i^{pred} - y_i)^2}{\sum_{i=1}^{N} (y_i - y_{mean}^{training})^2} \quad (9)$$

Because y_{mean} yields the minimum value for SS as shown above (see eqs 1−5), any other reference value $y_0 \neq y_{mean}$ (see eqs 1 and 6) such as $y_{mean}^{training}$ increases SS, decreases $PRESS/SS$, and thus artificially increases q^2. It follows that

$$q^2_{tr} \geq q^2 \quad (10)$$

With increasing the difference between y_{mean} and $y_{mean}^{training}$, q^2_{tr} becomes increasingly larger than q^2 as a general (but not strict) trend, provided there is still respective room between q^2 and 1 as upper boundary. In other words, q^2_{tr} would approach 1 faster than q^2 and would do so increasingly with an increasing difference between y_{mean} and $y_{mean}^{training}$. Thus, application of eq 9 - which would follow the current OECD guidelines[21] - tends to yield too optimistic estimates for the prediction power of regression models in terms of the q^2 statistics.

While a usual requirement for properly selected test sets is that they span both the target and descriptor value range of the training set and thus should have an activity mean at least similar to the one of the training set, it remains unclear why eq 9 should be preferred over eq 8, the latter of which we consider as a natural definition of q^2. Moreover, external test sets need not necessarily serve as comprehensive validation sets but could be deliberately selected to cover only a certain part of the structural domain or a certain part of the activity range (such as the higher or lower part) or both. A typical example would be a combined theoretical and experimental approach when exploring the activity potential of a certain compound class, selecting the compounds for (synthesizing and) testing based on predictions of an existing model. In such more targeted analyses, $y_{mean}^{training}$ is more likely to differ significantly from y_{mean}, resulting in correspondingly significant and thus misleading overestimations of the respective prediction capability when employing q^2_{tr} instead of q^2.

Note further that also from practical considerations, q^2 (eq 8) is preferred over q^2_{tr} (eq 9), because the latter requires having the training set available, which is not always the case. Eq 9 would thus - formally - make impossible the q^2 evaluation of models without published training sets, even if the model had been successfully implemented and would thus be ready for use. While it is true that the lack of knowledge of the training set makes it impossible to know exactly whether and to what the degree a given test set might in fact overlap with the unknown training set, comparative evaluation of q^2 and r^2 would still enable an indirect (though incomplete) assessment of the data situation. Only in the case of a model that shows no bias for the data set under investigation, r^2 and q^2 could become identical - if calibration and prediction power are identical for the data under investigation - even if the test set differs from the original training set. Otherwise, q^2 would be smaller than r^2, following the general relationship $q^2 \leq r^2$.

In this context, it is convenient to write r^2 in the form

$$r^2 = 1 - \frac{\sum_{i=1}^{N} (y_i^{fit} - y_i)^2}{\sum_{i=1}^{N} (y_i - y_{mean})^2} \equiv 1 - \frac{RSS}{SS} \quad (11)$$

where y_i^{fit} is the (possibly newly) fitted value of the i-th target property, as opposed to the (unmodified) predicted value of the model, y_i^{pred}, when calculating q^2 (eq 8). On the rhs of eq 11, RSS denotes the residual sum of squares. Clearly, y_{mean} (again) refers to the data set under investigation (as of course do both SS and RSS), which would be the training set when deriving the model or any test set when applying the already derived model to the respective test set compounds.

Value Ranges of r^2 and q^2 and Their Difference. An interesting difference between r^2 and q^2 concerns their value ranges. For the least-squares calibration, $0 \leq r^2 \leq 1$ holds true, with the actual value of r^2 quantifying the fraction of the activity variation that can be explained through the model (as is of course well-known). By contrast, q^2 may even become negative and has in fact no lower boundary but the value range $-\infty \leq q^2 \leq 1$. For r^2, the lower boundary results from the least-squares relationship between y_i^{fit} and y_{mean}. The latter implies $RSS \leq SS$, resulting in a respective compensation of possibly large calibration residuals by correspondingly large differences between the observed values and their mean.

With regard to q^2, y_i^{pred} (that is used instead of y_i^{fit}, see eqs 8 and 11) has no formal relationship with y_{mean}. Consequently, $PRESS$ may become larger than SS and in fact has no upper limit. The latter is easily seen when considering an arbitrarily bad model that yields arbitrarily large systematic errors, each of which might approach infinity. A general example can be constructed from any existing relationship between a target property Y and predictors X, $Y = f(X)$. Introduction of a slope a and intercept b (which happens automatically in a least-squares manner when calculating r^2 between Y and $f(X)$) yields

$$Y = a \cdot f(X) + b \quad (12)$$

Now, an arbitrary increase of slope a ($a \rightarrow \infty$) or b ($b \rightarrow \infty$) or both makes the predicted values of the target property approach infinity ($y_i^{pred} \rightarrow \infty$ or $y_i^{pred} \rightarrow -\infty$), leading to an arbitrarily bad model. Thus, $PRESS \rightarrow \infty$ is possible in principle, resulting in $q^2 \rightarrow -\infty$. When $q^2 = 0$, using y_{mean} to predict y_i is - on the statistical average - as good as using the model results y_i^{pred}, in which case $PRESS = SS$. For $q^2 < 0$, y_{mean} is even statistically preferred over y_i^{pred} to estimate y_i for the data set under investigation.

Eq 12 can also be used to illustrate the difference between r^2 and q^2 as measures of the calibration performance and prediction performance, respectively. If $f(X)$ would be a perfect regression model with respect to both precision

(trend) and accuracy (absolute values), its application to an external data set would result in $a = 1$ and $b = 0$ as well as $q^2 = 1$ and $r^2 = 1$. However, $r^2 = 1$ could also be achieved with a slope different from 1 and a nonzero intercept, in which case only the trend and thus the precision would be perfect, implying $q^2 < 1$. A simple example would be a model that systematically overestimates every y_i by, say, precisely 10. In this case, each individual prediction would be wrong by exactly 10 but could be easily corrected for through an intercept of $b = -10$.

When applying such a model to an external data set, calculation of r^2 implies an automatic recalibration of the original model predictions $y_i^{\text{pred}} = f(x_i)$ according to eq 12 to generate appropriately scaled $y_i^{\text{fit}} = a \cdot y_i^{\text{pred}} + b$ (which is the kind of postprocessing automatically performed when calculating r^2). By contrast, calculation of the respective q^2 value leaves y_i^{pred} unchanged and thus evaluates the model with respect to both precision and accuracy, which corresponds to setting $a = 1$ and $b = 0$ in eq 12. Consequently, increasingly large differences between r^2 and q^2 reflect an increasingly large inaccuracy of the model, allowing still a good (or even perfect) trend and thus precision of the model predictions. In other words, r^2 quantifies the model precision, and the difference between r^2 and q^2 is inversely related to the model accuracy (the smaller the difference, the greater the accuracy).

Cross-Validation. A special situation occurs when applying cross-validation instead of using an external test set in order to indirectly evaluate the prediction power of the model. In this case, subsets of the (initial) training set are employed to generate subset-calibrated models, which in turn are used to predict the target property for compounds left out in the respective subsets. Taking the prominent leave-1-out cross-validation as example, N subsets containing $N-1$ compounds would be generated, and the target property of each compound could then be predicted from the subset-calibrated model, whose associated subset did not contain that particular compound.

For the general case of leaving out k compounds and generating respective but different subsets, the cross-validated q^2 can be written as

$$q^2_{\text{cv}} = 1 - \frac{\sum_{i=1}^{N} \left(y_i^{\text{pred}(N-k)} - y_i\right)^2}{\sum_{i=1}^{N} \left(y_i - y_{\text{mean}}^{(N-k,i)}\right)^2} \qquad (13)$$

In eq 13, $y_i^{\text{pred }(N-k)}$ denotes the i-th predicted target value generated through application of a subset model trained on $N-k$ compounds where in particular the i-th compound was left out, and $y_{\text{mean}}^{(N-k,\,i)}$ is the arithmetic mean of that subset.

Thus, cross-validation would usually involve subset-specific means rather than the total training set mean. In practice, however, q^2_{cv} is usually evaluated using y_{mean} to simplify the calculation (and considering the fact that the subset means are usually at least close to each other and to y_{mean}), although this is formally not consistent with the procedure of cross-validation.[27] Interestingly, eqs 1−5 demonstrate that the use of y_{mean} - as compared to any other fixed reference value - yields the smallest *SS* and q^2_{cv} values possible (keeping in mind that *SS* evaluated with varying subset means could theoretically become even smaller). Note further that leave-k-out cross-validation tends to increasingly overestimate the prediction capability with decreasing k (because $q^2_{\text{cv}} \rightarrow r^2$ for $k \rightarrow 1$ and increasingly large N). Moreover, with some data sets a lack of correlation has been found between high leave-1-out q^2_{cv} and the external prediction capability,[28−30] and it was shown that the leave-1-out procedure does not yield the correct asymptotic behavior for large numbers of observations.[31] These findings suggest that for cross-validation, employing the total training set activity mean, $y_{\text{mean}}^{\text{training}}$, instead of the individual subset means appears to be also theoretically preferred. The reason is that $y_{\text{mean}}^{\text{training}}$ is the best single reference value possible that removes noise caused by ad hoc variations of the subset mean and likely (but not necessarily) tends to (slightly) reduce the numerical value of q^2_{cv} and thus provide a (slight) correction toward a more conservative estimation of the prediction capability. For large data sets and small k (and in particular for leave-1-out), however, the overall effect will usually be negligible, because leaving out one or few compounds will normally not affect the arithmetic mean significantly.

To avoid confusion, we emphasize that y_{mean} should refer to the original training set only when evaluating the calibration performance (r^2) or when performing cross-validation (q^2_{cv}); in both of these cases, the training set is indeed the data set under investigation. By contrast, any other test set used to evaluate the model performance in terms of q^2 would imply that the properly selected y_{mean} should refer to that test set, as was outlined in the previous section.

COMPARATIVE Q^2 AND Q^2_{TR} CALCULATION

As shown above, q^2_{tr} (eq 9) tends to overestimate the prediction power as compared to q^2 (eq 8), which in turn is triggered by the difference between the true data set mean, y_{mean}, and the training set mean, $y_{\text{mean}}^{\text{training}}$. In Table 1, the difference between q^2_{tr} and q^2 is illustrated for three examples with regression models and data sets taken from the literature.

The first example deals with a regression model to predict the sorption of organic compounds into soil organic matter as a logarithmic constant, log K_{oc}, from molecular structure.[22] Upon derivation of this model, the initial data set had been subdivided into an initial calibration set covering 80% of the compounds ($N = 457$, $y_{\text{mean}}^{\text{training}} = 2.67$) and an internal prediction set defined through the remaining 20% ($N = 114$, $y_{\text{mean}} = 2.79$). For the latter, the initial model yielded $r^2 = 0.85$ and $q^2 = 0.83$,[22] and application of eq 9 results in $q^2_{\text{tr}} = 0.84$ (see first data row in Table 1). To analyze the effect of varying $y_{\text{mean}}^{\text{training}}$ on the difference between q^2 and q^2_{tr}, respective statistics have also been evaluated for the subsets covering the lower and upper 2/3 of the log K_{oc} value range ("lower 2/3" and "upper 2/3" in Table 1), with associated y_{mean} (log K_{oc} arithmetic mean) values of 2.41 and 3.75, respectively (4th column in Table 1). Because the reduction in target value range reduces *SS* and thus generally lowers r^2 and q^2 (and correspondingly affects also q^2_{tr}) for a given *rms*, a third subset was analyzed that also covers 2/3 of the target value range but this time is centered around the arithmetic mean of the total prediction set ("center 2/3" in Table 1).

As can be seen from the table, for both the lower 2/3 and upper 2/3 subsets q^2_{tr} becomes even larger than r^2, which is

Table 1. Prediction Performance in Terms of r^2 (Eq 11), q^2 (Eq 8), and q^2_{tr} (Eq 9) Using Data Sets and Regression Models Taken from the Literature

prediction set[a]	N	minimum value	arithmetic mean	maximum value	r^2	q^2	q^2_{tr}	*rms*	*bias*
Initial Log K_{oc} Regression Model Calibrated on 80% of the Initial Data (457 Compounds), Leaving 20% As External Prediction Set;[22] $y^{training}_{mean} = 2.67$									
total	114	0.40	2.79	6.50	0.85	0.83	0.84	0.51	−0.07
center 2/3	103	1.45	2.77	5.37	0.80	0.80	0.80	0.46	−0.05
lower 2/3	99	0.40	2.41	4.14	0.72	0.71	0.73	0.44	0.03
upper 2/3	57	2.54	3.75	6.50	0.80	0.71	0.86	0.57	−0.28
Final Log K_{oc} Regression Model Calibrated on 571 Compounds, Using Two Additional Data Sets[23,24] as External Prediction Set;[22] $y^{training}_{mean} = 2.64$									
total	61	1.37	2.91	6.41	0.95	0.94	0.94	0.33	0.00
center 2/3	34	2.25	3.07	5.21	0.81	0.79	0.85	0.30	−0.03
lower 2/3	54	1.37	2.51	3.99	0.86	0.85	0.86	0.29	0.06
upper 2/3	23	3.10	4.23	6.41	0.92	0.88	0.95	0.42	−0.22
Log S_w [mol/L] Regression Model[25] Applied to an External Prediction Set;[b] $y^{training}_{mean} = -2.81$									
total	150	−9.42	−4.69	0.64	0.90	0.89	0.93	0.81	−0.15
center 2/3	119	−7.66	−4.22	−1.14	0.82	0.81	0.88	0.83	−0.13
lower 2/3	109	−9.42	−5.83	−2.72	0.87	0.86	0.96	0.71	0.09
upper 2/3	100	−6.02	−3.26	0.64	0.73	0.70	0.72	0.87	−0.27

[a] Besides the total prediction set, the following three subsets are used: center 2/3 = subset containing all compounds within the 2/3 interval of the target value range, centered around the arithmetic mean (leaving out the upper and lower 1/6 of the target value range); lower 2/3 = subset containing all compounds within the interval bracketing the lower 2/3 of the target value range (leaving out the upper 1/3 of the target value range); upper 2/3 = subset containing all compounds within the interval bracketing the upper 2/3 of the target value range (leaving out the lower 1/3 of the target value range). [b] From our in-house S_w database,[26] the prediction set compounds have been selected such that they belong to the application domain of the original model: First, their log S_w [mol/L] is within the target value range of the original training set (−11.6 − 2.0); second, their chemical structure belongs to the structural domain according to the 2nd-order atom-centered fragment approach.[13−15]

particularly pronounced for the upper subset that has the largest difference between $y^{training}_{mean}$ and y_{mean} (3.75 − 2.67 = 1.08 log K_{oc} units). Note further that except for the center 2/3 subset, q^2_{tr} is larger than q^2, with the largest difference of 0.15 being observed for the lower 2/3 subset. This latter subset also yields the largest difference between r^2 and q^2 (0.09), reflecting the fact that its bias is by far the largest (−0.28). Moreover, all subset r^2 and q^2 values are smaller than their total prediction set counterparts, which is caused by the above-mentioned effect of reducing the target values range.

In the second part of Table 1, a corresponding analysis is performed for the final log K_{oc} regression model calibrated on 571 compounds ($y^{training}_{mean} = 2.64$) and its application on an external prediction set collected from two literature sources[23,24] as described earlier.[22] Again, reduction of the target value range by 1/3 reduces both r^2 and q^2, the largest difference between r^2 and q^2 is observed for the largest bias, q^2_{tr} is larger than q^2 in three of the four cases, the by far largest difference between $y^{training}_{mean}$ and y_{mean} (upper 2/3 subset: 4.23 − 2.64 = 1.59) yields the largest difference between q^2_{tr} and q^2 (0.95 − 0.88 = 0.07), and in two cases q^2_{tr} is even larger than r^2.

The bottom part of Table 1 provides a third example, which is the application of a log S_w prediction model[25] on an external data set with 150 organic compounds. Here, q^2_{tr} is larger than q^2 in all four cases and larger than r^2 except for the upper 2/3 subset. As before, the largest difference between q^2 and r^2 is accompanied by the largest bias and the largest difference between q^2_{tr} and q^2 by the largest difference in the respective data set means.

To avoid misunderstanding, we emphasize that such large differences between $y^{training}_{mean}$ and y_{mean} would clearly indicate that the test set was not properly selected (see above). Nevertheless, the results demonstrate that q^2_{tr} is inherently flawed and that it does not appear to be preferred over q^2 in any reasonable situation. As a consequence, we suggest to refrain from using q^2_{tr} (eq 9), to employ q^2 (eq 8) instead, and to revise the respective OECD guidance document[21] accordingly.

CONCLUSIONS

For a given regression model and external test set, calculation of the q^2 statistics according to the current OECD guidelines tends to provide too optimistic estimates of the prediction capability. This overestimation is triggered by the difference between the arithmetic means of the training set and test set target values and generally increases (as far as possible with 1 as upper boundary for q^2) with an increasing respective difference. While judiciously selected validation sets would usually (but not necessarily) have activity means close to the training set activity mean, there is still no reason to use the latter for estimating q^2 for a given test set. Moreover, test sets deliberately selected to focus on a certain activity range would lead to systematically overestimated and thus misleading values of the q^2 statistics when employing the training set activity mean. Accordingly, we suggest to use, for a given test set, only the true arithmetic mean of its target values for the calculation of q^2 and to revise the OECD guidelines accordingly.

ACKNOWLEDGMENT

Financial support for the UFZ group by the European Union through the projects OSIRIS (contract No.: GOCE-CT-2007-037017) and CAESAR (contract No.: SSPI − 022674 - CAESAR) is gratefully acknowledged.

REFERENCES AND NOTES

(1) Cramer, R. D., III. BC(DEF) Parameters. 2. An Empirical Structure-Based Scheme for the Prediction of Some Physical Properties. *J. Am. Chem. Soc.* **1980**, *102*, 1849–1859.

(2) Cramer, R. D., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Linear Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.
(3) Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Statist. Assoc.* **1983**, *78*, 316–331.
(4) Efron, B. Better bootstrap confidence intervals. *J. Am. Statist. Assoc.* **1987**, *82*, 171–200.
(5) Klopman, G.; Kalos, A. N. Causality in Structure-Activity Studies. *J. Comput. Chem.* **1985**, *6*, 429–506.
(6) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.
(7) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
(8) Kolossov, E.; Stanforth, R. The quality of QSAR models: problems and solutions. *SAR QSAR Environ. Res.* **2007**, *18*, 89–100.
(9) Sheridan, R.; Feuston, R. P.; Maiorov, V. N.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
(10) Mekenyan, O.; Nikolova, N.; Schmieder, P.; Veith, G. COREPA-M: A multidimensional formulation of COREPA. *QSAR Comb. Sci.* **2004**, *23*, 5–18.
(11) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA* **2005**, *33*, 445–459.
(12) Guha, R.; Jurs, P. C. Determining the Validity of a QSAR Model - A Classification Approach. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 65–73.
(13) Kühne, R.; Kleint, F.; Ebert, R.-U.; Schüürmann, G. Calculation of Compound Properties Using Experimental Data From Sufficiently Similar Chemicals. In *Software Development in Chemistry 10, Proceedings of the 10th workshop "Computer in Chemistry", Hochfilzen, Austria, 1995*; Gasteiger, J., Ed.; PROserv Springer Produktionsgesellschaft: Berlin, Germany, 1996; pp 125−134.
(14) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, D.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.
(15) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Model selection based on structural similarity - Method description and application to water solubility prediction. *J. Chem. Inf. Model.* **2006**, *46*, 636–641.
(16) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Estimation of compartmental half-lives of organic compounds - structural similarity vs. EPI-suite. *QSAR Comb. Sci.* **2007**, *26*, 542–549.
(17) Regulation (EC) No 1907/2006 of the European Parliament and of the Council of December 18, 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing an European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official J. Eur. Union*, L 306, **2006**, *49* (December 30), 1−849.
(18) Tunkel, J.; Mayo, K.; Austin, C.; Hickerson, A.; Howard, P. Practical considerations on the use of predictive models for regulatory purposes. *Environ. Sci. Technol.* **2005**, *39*, 2188–2199.
(19) Allen, D. M. The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction. *Technometrics* **1974**, *16*, 125–127.
(20) 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.
(21) Organisation 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)2, 2007, pp 55 (paragraph no. 198) and 65 (Table 5.7).
(22) Schüürmann, G.; Ebert, R.-U.; Kühne, R. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. *Environ. Sci. Technol.* **2006**, *40*, 7005–7011.
(23) Sabljić, A. On the prediction of soil sorption coefficients of organic pollutants from molecular structure: Application of molecular topology model. *Environ. Sci. Technol.* **1987**, *21*, 358–366.
(24) Nguyen, T. H.; Goss, K. U.; Ball, P. W. Polyparameter linear free energy relationships for estimating the equilibrium partition of organic compounds between water and the natural organic matter in soils and sediments. *Environ. Sci. Technol.* **2005**, *39*, 913–924.
(25) Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* **1995**, *30*, 2061–2077.
(26) Schüürmann, G., Ebert, R.-U.; Nendza, M.; Dearden, J. C.; Paschke, A.; Kühne, R. *Prediction of Fate-Related Compound Properties. In Risk Assessment of Chemicals. An Introduction*, 2nd ed.; van Leeuwen, K.,Vermeire, T., Eds.; Springer Science: Dordrecht, The Netherlands, 2007; pp 375−426.
(27) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat* **1994**, *13*, 285–294.
(28) Novellino, E.; Fattorusso, C.; Greco, G. Use of comparative molecular field analysis and cluster analysis in series design. *Pharm. Acta Helv.* **1995**, *70*, 149–154.
(29) Norinder, U. Single and domain made variable selection in 3D QSAR applications. *J. Chemom.* **1996**, *10*, 95–105.
(30) Kubinyi, H.; Hamprecht, H.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSAR) from SEAL similarity matrices. *J. Med. Chem.* **1998**, *41*, 2553–2564.
(31) Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.

CI800253U