Journal of Computer-Aided Molecular Design **18:** 549–562, 2004.
DOI 10.1007/s10822-004-4071-5

Validation tools for variable subset regression

Knut Baumann* & Nikolaus Stiefl
Department of Pharmacy and Food Chemistry, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Germany

Received 11 May 2004; accepted in revised form 23 August 2004

Key words: chance correlation, cross-validation, validation, variable selection

Summary

Variable selection is applied frequently in QSAR research. Since the selection process influences the characteristics of the finally chosen model, thorough validation of the selection technique is very important. Here, a validation protocol is presented briefly and two of the tools which are part of this protocol are introduced in more detail. The first tool, which is based on permutation testing, allows to assess the inflation of internal figures of merit (such as the cross-validated prediction error). The other tool, based on noise addition, can be used to determine the complexity and with it the stability of models generated by variable selection. The obtained statistical information is important in deciding whether or not to trust the predictive abilities of a specific model. The graphical output of the validation tools is easily accessible and provides a reliable impression of model performance. Among others, the tools were employed to study the influence of leave-one-out and leave-multiple-out cross-validation on model characteristics. Here, it was confirmed that leave-multiple-out cross-validation yields more stable models. To study the performance of the entire validation protocol, it was applied to eight different QSAR data sets with default settings. In all cases internal and external model performance was good, indicating that the protocol serves its purpose quite well.

Introduction

Fitting regression equations is ubiquitous in QSAR applications. Often, there is not only a single adequate regression model but a variety of alternatives owing to the abundance of available structure descriptors. These alternatives may result from different information used (e.g. different fields in CoMFA [1], different probes in Volsurf [2]), from differently complex models (factor selection in partial least squares regression (PLS) and principal component regression (PCR)), or from different data pre-processing steps (e.g. no scaling vs. scaling). All decisions that lead to the final model are referred to as model selection. Although often not explicitly stated, model selection has an impact on model validity. The more alternatives are tried, the larger is the probability of chance correlations [3–5]. An extreme case of model selection is variable selection, which is often applied to QSAR descriptors to ease the model interpretation. Variable selection evaluates thousands of models before arriving at the final one. As a result, the way the models are to be validated is affected.

While leave-one-out cross-validation (LOO-CV) typically works well as an objective function for selecting among few alternatives, e.g. factor selection in PCR and PLS [6], it yields heavily overfitted models in variable subset regression [7, 8]. Hence, choosing among many alternatives

*To whom correspondence should be addressed. Fax: +49-931-888-5494; E-mail: knut.baumann@mail.uni-wuerzburg.de

needs a more demanding validation criterion as objective function. Leave-multiple-out cross-validation (LMO-CV) is such a more stringent criterion. In LMO-CV a large portion (e.g. 40–60%) of the training data is set aside as validation data set [9, 10]. The remaining portion of the data is used as construction data set to build the model. The quality of this model is assessed with the validation data set. The random split into validation and construction data set is repeated many times to get an average cross-validated prediction error. The effect of LMO-CV is two-fold: First, it is more difficult to fit a good model with fewer construction data (in LOO-CV the construction data set is of size $n-1$ as opposed to LMO-CV, where it is often only of size $n/2$ with n being the number of objects in the training data set). Second, the model is assessed with a larger validation data set. Hence, the construction and the validation data set are less similar in each split than in LOO-CV. This allows to better estimate the predictive ability of the model. Both mechanisms prevent LMO-CV to learn the idiosyncrasies of the data set and as a result reduce the amount of overfitting in variable subset regression [10–12]. Instead of LMO-CV the respective bootstrap analogue can also be used to lower the tendency of overfitting [13, 14].

Apart from the necessity for more demanding validation criteria, the model selection process changes the characteristics of the final model. Two such characteristics are studied in this contribution. First, internal figures of merit are largely inflated [15]. Affected are not only measures that assess the fit of the model (e.g. R^2), which were already studied by various authors [3, 4, 15–17], but also cross-validated figures of merit [12, 18]. The degree of inflation depends on the modeling procedure and the validation criteria being used. A diagnostic plot revealing the degree of inflation and some relevant influence factors for the degree of inflation are pinpointed and discussed.

Second, models obtained by massively sifting through candidate models tend to be instable. Here, instable means that small changes in the data cause large changes in the model [19]. A simple experiment shows this instability. Using a fixed modeling protocol, remove a single molecule (object) from the training data set and run the entire modeling procedure from scratch. Repeat this for every object of the training data set. For each of the resulting final models predict the same test set and plot a measure of test set predictivity such as the coefficient of determination (R^2_{Test}). This experiment is shown in Figure 1 for the steroid data set [1, 20] ($n = 20$, $n_{\text{Test}} = 9$, $p = 170$; $p =$ number of variables) with the structure descriptors (MaP; see Methods) and the variable selection technique (Tabu-Search, LMO-CV; see Methods) used in ref. [21]. This data set will be referred to as STE. It can be seen that removing a single molecule from the training data set while keeping everything else fixed, causes large changes in test set predictivity. Fortunately, an ensemble average of the predicted values of all models largely smoothes out (i.e. stabilizes) the fluctuations of the individual models. Its value is shown as a horizontal line.

A method for the assessment of instability based on the idea of leverages is investigated. The same method can also be used to stabilize modeling procedures that are susceptible to overfitting. In classical regression theory [22] the leverage of the ith object (h_i) is defined as the change in the fitted value ($\hat{y}_i$) depending on the change in the observed value for this object (y_i), which can be expressed as $h_i = \partial\hat{y}_i/\partial y_i$. The larger the leverage of an object, the more will the fitted value for this object change, if its observed value changes (e.g. fluctuations around its expected value owing to random noise). Hence, the larger the sum of all leverage values for a given data set and a particular modeling procedure is, the more instable is the respective modeling procedure according to the definition of instability given above. A method to estimate the leverage values for an arbitrary modeling procedure was devised by Ye [23] and is used here to assess the stability. The information generated during these computations can also be used to stabilize the estimate of the test set prediction error by ensemble averaging at marginal additional computational cost [24]. Hence, this stabilized estimate of the prediction error for test sets will also be computed and compared to those obtained with a single model rather than an ensemble.

Methods

Permutation testing

The degree of inflation by model selection can be determined with the help of a permutation test. In

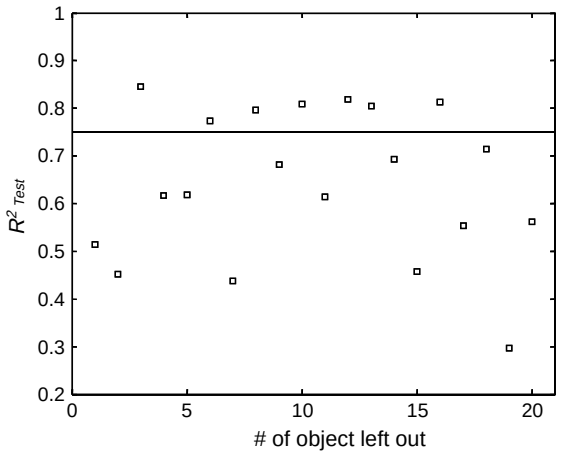

Figure 1. The instability of variable selection procedures illustrated with the steroid data set (STE). Removing a single molecule from the training data set (index on the x-axis) while keeping everything else of the variable selection procedure fixed, causes large changes in test set predictivity. The horizontal line gives the R^2_{Test}-value obtained with the ensemble average of the predicted values.

the permutation test used here, the relation between the independent variables (X-variables, e.g. structures descriptors) and the response variable (y-variable, e.g. biological activity) is deliberately destroyed by randomly permuting the response variable. In the QSAR literature this procedure is often called y-scrambling [e.g. 25–27] and is used to test the significance of the derived model. Since there is no relation between the X-variables and the response variable after y-scrambling, only bad models are expected. If a modeling procedure is, nonetheless, able to obtain good models, then this is an indicator for inflated internal figures of merit. Here, the degree of inflation is estimated as the median (i.e. the 50th percentile) of the permutation distribution and is referred to as R^2_{HO}. If the modeling procedure generates even better models than the original one (no permutation of the responses), then this is an indicator for a chance correlation and a reason for rejection of the respective model, since it is no better than chance. About 500 random permutations of the response variables are computed. For each of the permuted response vectors the entire modeling procedure is re-run from scratch.

The modeling procedure used for all experiments was a stepwise operating variable selection algorithm called reverse elimination Tabu search (REM-TS). REM-TS works like stepwise regression but cannot get trapped in local optima. It is a greedy search algorithm that either adds a variable to the model or removes a variable from the current model. At each step of the search the move which results in the largest improvement of the objective function is executed. In case REM-TS reaches a local optimum (i.e. no improvement is possible) it uses the path of the least detrimental step with respect to the objective function to escape from the local optimum. To avoid that REM-TS cycles between two solutions, a return to an already visited solution is prohibited (it is set 'tabu'). A detailed description of the method can be found in ref. [7]. REM-TS was used in combination with LOO-CV and LMO-CV as objective function. In case of LMO-CV, 50% of the data were set aside and used as validation data set (i.e. a leave-50%-out cross-validation was computed). The random split into validation data set and construction data set was repeated $3n$ times. Owing to averaging over such a large number of splits, an estimate of the cross-validated prediction error is obtained that only marginally depends on the random partitions into validation and construction data sets. REM-TS can be combined with any regression technique. It was used here in combination with either PCR or PLS. For the theory of PCR and PLS see ref. [28].

Two data sets were used for permutation testing. The Selwood data set ($n = 31$, $p = 53$) [29, 30], which is often used as a benchmark data set in variable selection, and a data set of endothelin A inhibitors (ET_A; $n = 36$, $p = 154$), which was also often used to compare 3D-QSAR methods (e.g. [31–33]). The descriptors used for the latter data set were those of ref. [34] (DiP: Distance Profiles). This data set is referred to as ET_A-DiP. REM-TS was run for 100 iterations in combination with PCR for the Selwood data set, and 50 iterations in combination with PLS for the ET_A-DiP data set. For the Selwood data set the influence of the number of variables in the final model was studied, whereas the ET_A-DiP data set was used to highlight the differences between using LOO-CV and LMO-CV as objective function.

Estimating the leverages and stabilization

The procedure to estimate the leverage values is based on intentionally perturbing the response variable [23]. For that reason, normally distributed

random noise is added to the original response values and the modeling procedure is re-run from scratch. Since the leverage for the ith object is defined as change in the fitted value depending on the change in the observed values for that object (see above), it can be estimated by determining the fitted values $\hat{y}_i^*$ after having perturbed the original response to $y_i^* = y_i + \delta_i$ (δ_i = random noise). Ye [23] shows that δ_i should be chosen as normally distributed noise with mean zero and standard deviation $\tau = 0.6\hat{\sigma}$, where $\hat{\sigma}$ is an estimator of the residual standard deviation. Since for biased regression techniques such as PCR and PLS the estimator of the residual standard deviation (i.e. the root mean squared error of calibration) is biased downwards, we used the usual LOO-CV root mean squared error of prediction ($\text{RMSEP}_{\text{CV}-1}$) as the respective multiplier. Once more, to be able to estimate the leverages, the responses need to be perturbed many times. Here $T = n$ repetitions were used. For each repetition $t = 1, \ldots, T$, the values $\hat{y}_{i,t}^*$ and the corresponding $\delta_{i,t}$ are stored for all i objects. Using these values the leverages can be estimated as the slope $\hat{h}_i$ using the perturbation terms $\delta_{i,t}$ ($t = 1, \ldots, T$) as independent variables and the values $\hat{y}_{i,t}^*$ ($t = 1, \ldots, T$) as dependent variables of a simple linear regression with intercept term.

In multiple linear regression (MLR) the degrees of freedom equal the number of variables in the model (p) [23]. It is instructive to note the coincidence in linear models that p also equals the sum of all leverage values. Since $\hat{h}_i$, the estimated leverage value for object i, and thus $\sum \hat{h}_i^*$ ($i = 1, \ldots, n$), the sum of the leverage values, can be computed for every modeling procedure M, the term $\hat{D}(\text{M}) \sum \hat{h}_i$ ($i = 1, \ldots, n$) is called generalized degrees of freedom of modeling procedure M [23]. As mentioned earlier, the larger $\hat{D}(\text{M})$ is, the less stable and the more complex is the respective modeling procedure M. In Table 1 the pseudo-code for determining the leverages and the generalized degrees of freedom is given for a modeling procedure that is linear in the model parameters **b** (e.g. multivariate regression techniques). However, it straightforwardly generalizes to any modeling technique since the fitted values of the perturbed models ($\hat{y}_{i,t}^*$) can be obtained with any modeling procedure.

The T models obtained with the perturbed responses can also be used for two different purposes. First, a variable selection statistics can be assembled showing the selection frequency of each variable. If the model is stable, it is anticipated that the same variables that are part of the unperturbed model are repeatedly selected. Hence, apart from the generalized degrees of freedom the variable selection statistics can be used to assess the stability of the selection procedure. Second, the perturbed models can be used to obtain a stabilized estimate of the prediction error for an

Table 1. Pseudo-code for determining the leverages and the generalized degrees of freedom.

1	Read data **X**, **y**
2	Compute τ as 0.6 $RMSEP_{\text{CV}-1}$ (i.e. the leave-one-out cross-validated root mean squared error of prediction of the full PCR or PLS model).
2	For $t = 1, \ldots, T$ ($T \geq n$)
3	Generate normally distributed random numbers $\Delta_t = [\delta_{1,t}, \ldots, \delta_{n,t}]^{\text{T}}$ according to N(0, τ).
4	Estimate the model parameters $\hat{\mathbf{b}}_t^*$ with training data $\mathbf{y}_t^* = \mathbf{y} + \Delta_t$ and **X** using modeling technique M. Obtain $\hat{\mathbf{y}}_t^* = \mathbf{X}\hat{\mathbf{b}}_t^*$.
5	Next t
6	For $i = 1, \ldots, n$
7	Estimate leverage h_i as the sensitivity $\hat{h}_i$ of the fitted values to perturbation as follows: $\hat{y}_{i,t}^* = \hat{h}_0 + \hat{h}_i\delta_{i,t} \quad t = 1, \ldots, T$
8	Next i
9	Estimate D(M) as $\hat{D}(\text{M}) = \sum_{i=1}^{n} \hat{h}_i$
10	End

Symbols and notation. **X**: matrix of independent variables consisting of n objects and p variables, **y**: response vector, Δ_t: tth vector of random noise with mean zero and standard deviation τ (see text), and **b**: model parameters (here: regression coefficients). Notation: Matrices: bold, uppercase symbols (e.g. **X**). Vectors: bold, lowercase symbols (e.g. **y**). Scalars: lowercase italicized symbols (e.g. n). All vectors are column vectors. The superscript T denotes the transpose of a vector or a matrix. Estimated quantities are wearing a hat ('^').

Table 2. Pseudo-code for the additional steps to obtain the stabilized estimate for an external test data set with independent variables X_{Test}.

4a	Predict $\hat{\mathbf{y}}^*_{Test,t} = \mathbf{X}_{\text{Test}}\hat{\mathbf{b}}^*_t$
9a	Compute stabilized estimate for the predictions as $\hat{\hat{\mathbf{y}}}^*_{\text{Test}} = T^{-1}\sum_{t=1}^{T}\hat{\mathbf{y}}^*_{\text{Test},t}$

Symbols and notation: see Table 1 and text.

external test set. This is achieved by predicting the test set objects for each of the T perturbed models and averaging the predicted values [24]. Like bagging [35] and boosting [36, 37], this technique has the potential to reduce the test set error by ensemble averaging. However, in contrast to bagging and boosting, the noise addition technique only changes the response variables to form perturbed training sets while the independent variables remain unchanged. The pseudo-code for the additional steps that are necessary to obtain the stabilized test set predictions for modeling procedures that are linear in the parameters is given in Table 2. These steps are to be added after the step with the corresponding number.

For illustrative experiments with the generalized degrees of freedom, two data sets were used: first the steroid data set ($n = 21$, $n_{\text{Test}} = 9$; $p = 45$) [1, 20] using the same descriptors (SESP: start-end shortest path vectors) as in ref. [38]. This data set is referred to as STE-SESP. The second data set comprises eye-irritating compounds and was used in combination with MaP-descriptors ($n = 25$, $n_{\text{Test}} = 13$, $p = 190$; EYES data set, see also below) [21]. The test set for the steroid data set consists of nine compounds only (i.e. the only fluorine substituted compound was removed).

Validation protocol

Since variable selection is unstable and tends to overfitting, rigorous validation is needed to result in a finally chosen model with satisfactory external predictivity. To safeguard against overfitting, a validation protocol evolved in our laboratory that is used by default. Clearly, the ideal settings depend on the class of structure descriptors that is used. However, the general structure of the protocol may also be interesting for QSAR scientists using variable selection. The protocol is given in Table 3. Some of the steps deserve further comments. For splitting the data set into training and test set, algorithms such as the Kennard–Stone CADEX algorithm have proven to be useful [39–41]. It is important to note that external prediction errors themselves show a large random fluctuation. The relative standard deviation of the root mean squared error of prediction is approximately $1/\sqrt{(2n_{\text{Test}})}$, which is of considerable size for small test sets [42]. Hence, it is also advisable to repeat this split with different partitions into training and test sets.

To avoid overfitting the search algorithm is executed with an early stopping rule (stop if improvement in RMSEP_{CV} in two consecutive steps is smaller than 3%). Although better internal solutions might be missed, it was found in many applications that it is advantageous with respect to robustness and external predictivity to stop early. The core of the validation protocol is the usage of LMO-CV (rather than LOO-CV) as objective function in combination with a rather high object-to-variable-ratio of at least 6 (see also [4, 18]). Once more, this is done to avoid overfitting that occurs with LOO-CV if the object-to-variable-ratio is not constrained [7, 8]. Next, the internal consistency of the model is checked with the usual tools (e.g. by inspecting internal figures of merit and residual plots). In addition to that the degree of inflation caused by model selection (R^2_{HO}) and model complexity/model stability expressed as generalized degrees of freedom is assessed. If the model passes all tests based on internal validation, external prediction is carried out to assess external predictivity. Before doing so, prediction outliers (i.e. unusual objects that are not covered by the calibration space) are removed [43]. Here, the closeness of an object to the calibration samples is measured as Mahalanobis distance [41, 43]. It should be noted that no information about the response variable of the test set is used at this step. In Table 3 some rules-of-thumb for cut-off values are given. However, these, as well as the entire protocol, should not be used in cookbook fashion. The characteristics of the data set under scrutiny have always to be taken into account. The protocol is neither complete, nor final. There is a steady evolution of novel validation tools (see e.g. [44] and those proposed here) which dynamically change the protocol. Moreover, there are important alternatives to this protocol that may also be employed [45]. In the work reported here, the protocol was applied to a broad range of data sets

Table 3. Standard validation protocol.

1	Split the available data into test data set (1/3) and training data set (2/3) using Kennard-Stone's CADEX algorithm. This split may also be done repeatedly at random.
2	Search algorithm: Tabu Search (REM-TS). Stop if the difference in the cross-validated root mean squared error of prediction ($RMSEP_{CV}$) of two consecutive steps is less than 3% or if it worsens.
3	Regression technique: PCR (default), or PLS.
4	Objective function: leave-50%-out cross-validation with $B = 3n$ splits of the training data set into construction and validation data sets are computed.
5	Constrain the maximum number of variables in the final model (p_{max}) to $n/p_{max} \geq 6$.
6	Check the internal consistency of the selected model by using internal figures of merit (e.g. cross-validated root mean squared error of prediction and the respective R^2_{CV}-value) and the usual regression diagnostics (residual plots etc.).
7	Determine the amount of inflation and the risk of chance correlation using a permutation test. The number of permutations ($n_{Permute}$) should be at least 200 (default: $n_{Permute} = 500$).
8	Determine the generalized degrees of freedom and inspect the selection statistics of the perturbed model.
9	Check the model: (1) Satisfactory internal consistency (e.g. leave-50%-out cross-validated coefficient of determination ($R^2_{CV-50\%}$) > 0.5, residual plots ok, etc.)? (2) Low amount of inflation (e.g. $R^2_{HO} < 0.2$, probability of chance correlation < 0.005). (3) Model complexity and stability ok (e.g. $D(M) < n/2$)
10	Check for prediction outliers [43], i.e. check for unusual objects in the test set that are remote from the calibration space. Do not use responses of the test sets here!
11	For those objects that are covered by the calibration space do test set prediction to assess the external predictivity of the finally chosen model (e.g. by the root mean squared error of test set prediction ($RMSEP_{Test}$) or the respective coefficient of determination (R^2_{Test})).

(see below) without changes of the default parameters given in Table 3.

Data sets used to evaluate the validation protocol

In total, eight data sets were processed using the aforementioned validation protocol. In all cases the MaP-descriptor (Mapping Property distributions of molecular surfaces), a translationally and rotationally invariant, surface-based 3D descriptor was used with default parameters [21]. In addition to that for two data sets (NIQ, PARP; see below) H-bond donor strength was differentiated [41], which is an extension of the default descriptor. The first three data sets were the steroid data set (STE; $n = 20$, $n_{Test} = 9$, $p = 170$) [1, 20], a data set of eye-irritating compounds (EYES; $n = 25$, $n_{Test} = 13$; $p = 190$) [46], and allosteric modulators of the muscarinic M_2 receptor (W84; $n = 29$, $n_{Test} = 15$, $p = 435$). These data sets are described in detail in ref. [21] (for STE one training and one test set compound was removed). The fourth data set was the aforementioned endothelin A inhibitors data set [31], this time using MaP descriptors (ET_A; $n = 24$; $n_{Test} = 12$, $p = 270$) [47, 48]. The fifth data set was also previously published [41] and consists of antimalarially active naphthylisoquinoline alkaloids (NIQ; $n = 29$, $n_{Test} = 15$, $p = 420$). The remaining data sets are part of a forthcoming manuscript and deal with acetylcholinesterase-inhibitors (AChE; $n = 32$, $n_{Test} = 16$, $p = 360$; training set molecule **3r** according to the enumeration scheme of the original reference was removed) [49], non-nucleoside HIV-1 RT inhibitors (NNI; $n = 29$, $n_{Test} = 15$, $p = 225$) [50], and poly(ADP-ribose)polymerase inhibitors (PARP; $n = 30$, $n_{Test} = 16$, $p = 399$) [51]. All data sets were partitioned into training and test set using

Kennard-Stone's CADEX algorithm on the structure descriptors. An exception to this was the steroid data set where the previously published training and test set partition was adopted [20, 21]. For all data sets PCR was used as regression technique. Differences between PCR and PLS tend to be small when these regression techniques are used in combination with variable selection [7]. The descriptor data were mean centred but not scaled. The column means were re-estimated for each split in cross-validation.

Figures of merit

The leave-multiple-out cross-validated root mean squared error was computed as follows:

$$\mathrm{RMSEP}_{\mathrm{CV}-k} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\frac{1}{k}\sum_{i=1}^{k}(y_{b,i,\mathrm{obs}} - y_{b,i,\mathrm{pred}})^2},$$

where B is the number of cross-validation runs (default: $B = 3n$), k the number of objects left out (default: nearest integer to $0.5n$), $y_{b,i,\mathrm{obs}}$ is the observed property value of the ith object that was left out in the bth cross-validation run, $y_{b,i,\mathrm{pred}}$ is the corresponding predicted property value of this object, and the subscript $-k$ indicates the number (or the percentage) of objects left out. The acronym RMSEP used in this contribution is also sometimes referred to as SDEP in the QSAR literature. From the $\mathrm{RMSEP}_{\mathrm{CV}-k}$ value the respective cross-validated squared multiple correlation coefficient $R^2_{\mathrm{CV}-k}$ can be computed as follows:

$$R^2_{\mathrm{CV}-k} = 1 - \frac{((\mathrm{RMSEP}_{\mathrm{CV}-k})^2 n)}{\sum_{i=1}^{n}(y_{i,\mathrm{obs}} - \bar{y})^2},$$

where $y_{i,\mathrm{obs}}$ is the observed property value of the ith object that was left out and $\bar{y}$ is the mean of all property values. The denominator of the equation is also termed SYY. In case of the usual LOO-CV ($k = 1$, $B = n$), the corresponding R^2-value is $R^2_{\mathrm{CV}-1}$ which is often referred to as q^2. If $k \gg 1$ the estimate of the prediction error obtained by LMO-CV ($\mathrm{RMSEP}_{\mathrm{CV}-k}$) is biased upwards [52]. Consequently, results obtained by LMO-CV will always be worse than those obtained by LOO-CV. Figures of merit for test set prediction are computed in the same fashion. The subscript is changed to 'Test' to indicate that these are external predictions. The exact equations for these figures of merit can also be found in ref. [8].

Results and discussion

It was shown by several authors that the figures of merit characterizing the fit of the data are largely inflated in case of variable subset regression [3, 4, 15, 16]. Less well known is the fact that internally cross-validated figures of merit are also affected. Owing to the massive search for good objective function values (e.g. a small $\mathrm{RMSEP}_{\mathrm{CV}}$) the objective function is likely to be overoptimistic since the same data are used to select and to assess the model over and over again. This so-called selection bias [5] was well described by Mosteller and Tukey as follows [53] (cited in [54]):

Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimising process that chose it from among many possible procedures will have made the greatest use possible of any and all idiosyncrasies of those particular data. ... As a result, the procedure will likely work better for these data than for almost any other data that will arise in practice.

Evaluating the inflation caused by model selection was done by permutation testing. Since permuting the response vector destroys the relationship between responses and structure descriptors (independent variables), the permutation test evaluates what can be achieved by chance alone (simulation of the statistical null hypothesis). Two permutation distributions for the Selwood data set are shown in Figure 2. For generating both distributions the same search technique (REM-TS, 100 iterations, LOO-CV) was used. In the upper panel the REM-TS was allowed to select at most six variables ($p_{\max} = 6$), whereas in the lower panel REM-TS was restricted to $p_{\max} = 3$. The LOO-CV coefficient of determination ($R^2_{\mathrm{CV}-1} \equiv q^2$) for the untouched response vector is shown as a dotted arrow on the x-axis. The median of all $R^2_{\mathrm{CV}-1}$-values obtained with the permuted response values (abbrev.: R^2_{HO}) is displayed as a solid arrow. Figure 2 shows that no $R^2_{\mathrm{CV}-1}$-value using permuted responses ($R^2_{\mathrm{CV}-1,\mathrm{PT}}$) is larger than the real $R^2_{\mathrm{CV}-1}$-value where the responses were left in place. Hence, the probability of chance correlation is low ($p_{\mathrm{CC}} < 0.002 = 1/n_{\mathrm{Permute}}$).

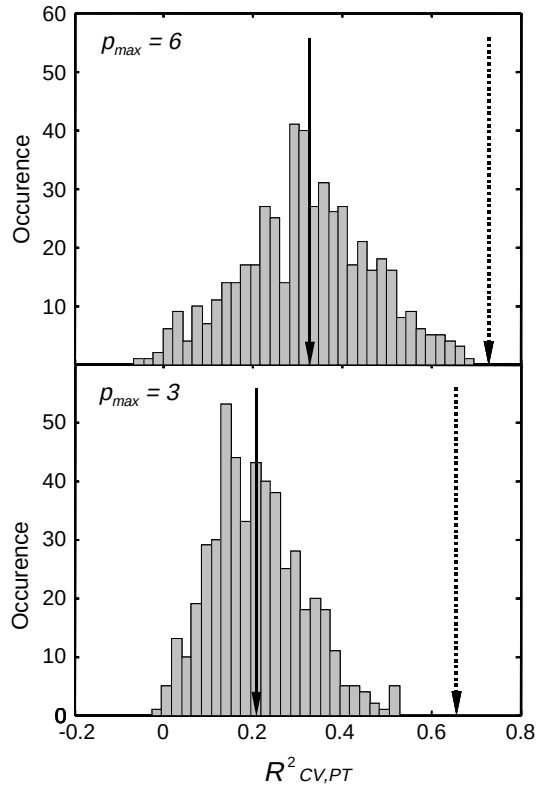

Figure 2. Permutation distributions for the Selwood data set: upper panel: $p_{\max} = 6$; lower panel: $p_{\max} = 3$. The location and the spread of the permutation distribution depend on the maximum number of variables allowed ($p_{\max}$) in the final model.

However, the degree of inflation is markedly different. The more variables are allowed to be selected, the larger is the degree of inflation (right shift of the permutation distribution on the x-axis from $p_{\max} = 3$ to $p_{\max} = 6$). Not surprisingly, the real R^2_{CV}-value is better for the six-variable model. However, the gap between the real R^2_{CV}-value (untouched response vector) and the median $R^2_{\mathrm{CV,PT}}$-value (R^2_{HO}) is larger if only three variables are allowed to be selected (the more stringent case from the mathematical modeling point of view). Put differently, for the three-variable model the permutation distribution is shifted farther away from the real value. That means that better models arisen due to chance are less likely in this case. Since chance and true model are separated more clearly for the three-variable model, the latter is considered to be superior. The differences between the real R^2_{CV}-value and the R^2_{HO}-value ($\Delta R^2_{HO} = R^2_{\mathrm{CV}} - R^2_{HO}$) can be used to characterize the performance of the modeling procedure. Here, the differences ΔR^2_{HO} for $p_{\max} = 6$ and $p_{\max} = 3$ are 0.40 and 0.45, respectively. This shows that the seemingly worse three-variable model is indeed bounded farther away from models that simply arise due to chance than the apparently better six-variable model.

In the second example ($\mathrm{ET_A}$-DiP data set), the influence of the objective function on the permutation distribution is studied. The object–variable ratio was restricted to six ($n/p_{\max} = 6$) for both cases, LOO-CV and LMO-CV. Figure 3 shows the results for LOO-CV (upper panel) and LMO-CV (lower panel; leave-50%-out, $B = 3n$). Once more, there is a marked shift in the permutation distribution. The permutation distribution using LMO-CV as objective function is shifted to the left on the x-axis. Thus, R^2_{HO} is shifted away from the real $R^2_{\mathrm{CV-50\%}}$-value with a difference between $R^2_{\mathrm{CV-50\%}}$ and R^2_{HO} equal to 0.58. For LOO-CV the respective difference is only $\Delta R^2_{HO} = 0.40$. This shows clearly that the degree of inflation does heavily depend on the objective function and it shows that LMO-CV is less prone to inflation. Summing up, internally cross-validated figures of merit are overoptimistic. As a consequence, assessing the validity of models generated by variable subset regression with an external test set is absolutely mandatory. The degree of inflation can be estimated by ΔR^2_{HO} (the larger, the less inflation). LOO-CV and models with a low object-to-variable-ratio are particularly prone to low ΔR^2_{HO}-values.

Leverages and generalized degrees of freedom

Apart from causing inflation of internal figures of merit, model selection adds to the instability and complexity of the chosen models. Instability and complexity can be expressed as leverage values. The larger the sum of all leverage values is, the more complex and the less stable is a particular model. In classical MLR the leverage values sum up to the number of parameters. In MLR (without subset selection) the number of parameters is equal to the number of degrees of freedom of that model (i.e. $df = p$). It is well known that more degrees of freedom (i.e. more parameters) in MLR indicate more complex and less stable models. The same line of reasoning is true for the generalized degrees of freedom $\hat{D}(\mathrm{M})$. Large values of $\hat{D}(\mathrm{M})$ indicate complex and instable models. However, in case of

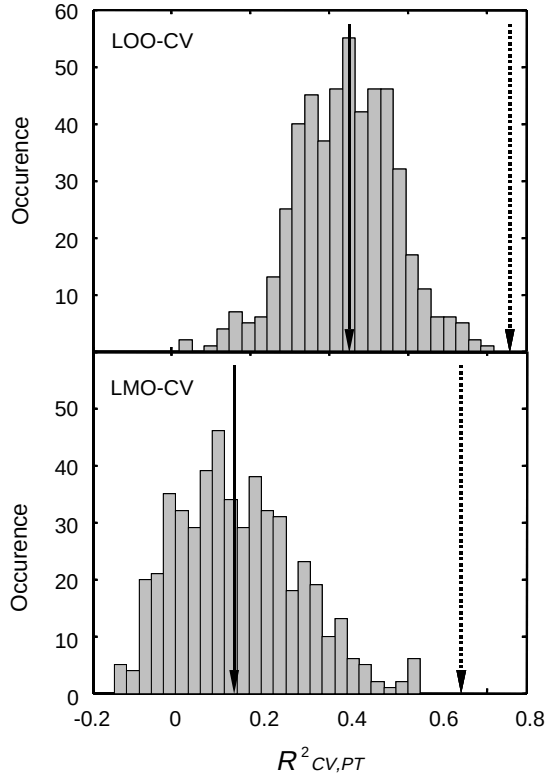

Figure 3. Permutation distributions for the ET_A-DiP data set: upper panel: LOO-CV; lower panel: LMO-CV. The object–variable ratio was restricted to six ($n/p_{max} = 6$). It can be seen that the location and the spread of the permutation distribution depend on the employed objective function.

model selection the sum of leverage values will generally be larger than the number of parameters in the model [18]. The difference between the nominal degrees of freedom ($\hat{D}_{\mathrm{nom}}$) and the generalized degrees of freedom ($\hat{D}(\mathrm{M})$) shows the impact of model selection. This is illustrated in the following example: For the steroid data set (STE-SESP), a model with eight variables using four principal components could be derived (REM-TS, LMO-CV: leave-50%-out, see [38]). The nominal degrees of freedom are five ($\hat{D}_{\mathrm{nom}} = 5$). One degree of freedom stems from estimating the column means (column mean centering) and the remaining four are due to the four principal components used for regression. Since PCR as well as PLS first projects the data to a lower dimensional subspace, the dimension of that subspace better approximates the nominal degrees of freedom than the number of variables included in the model (see e.g. [55]). In sharp contrast to the small number of nominal degrees of freedom, the generalized degrees of freedom amount to $\hat{D}(\mathrm{M}) = 10.7$. Hence, the nominal degrees of freedom do not reflect the true complexity and stability of models generated by variable subset regression. There is a 'cost' associated with model selection. This cost can be expressed with the help of the generalized degrees of freedom. The latter show that we pay for variable subset regression with added complexity ($\hat{D}(\mathrm{M}) > \hat{D}_{\mathrm{nom}}$) and thus with greater instability of the chosen models.

The large difference between the smaller nominal and the larger estimated degrees of freedom is the rule rather than the exception (see next subsection). This difference helps to explain the large fluctuation of models generated by variable subset regression (see e.g. introductory example). This large fluctuation, which is an indicator of instability, can also be highlighted by the selection statistics of all runs with perturbed response vector. Such a statistic is shown in Figure 4 for the steroid data set (STE-SESP). REM-TS was used in combination with leave-50%-out cross-validation as objective function. In this case, the selection function was not constrained to a particular object–variable ratio. As a result, the number of selected variables may differ between perturbation runs. The x-axis of the figure gives the variable index. On the y-axis the relative selection frequency is shown, i.e. the number of times a variable was selected in the final model divided by the number of perturbation runs. If a particular model was stable, it would be expected that the same variables of the original, unperturbed model were selected a large number of times even if the responses are slightly perturbed. Put differently, for stable models minor fluctuations in the responses (e.g. from measurement error) should cause only little variations of the selected variables. As a consequence, variables being truly important in the original, unperturbed model are expected to be selected more often than unimportant variables of that model. In Figure 4 two heuristically defined threshold values are given as horizontal lines. Variables that are selected in less than 30% (below the lower threshold) are considered weak and of little importance to the model. Variables that are selected more than 30% and less than 50% (the range in between the horizontal lines) are said to be of moderate strength and moderately important, whereas variables that are selected in more

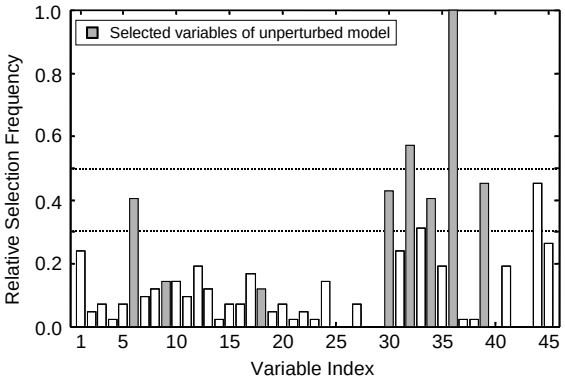

Figure 4. Variable selection statistic for the steroid data set (STE-SESP). The *x*-axis gives the variable index, the *y*-axis shows the relative selection frequency. The horizontal lines define the partition into weak, moderate and strong variables.

than 50% are considered to be strong and important. It can be seen that for the STE-SESP data set only two of the original variables are below the threshold. Hence, these were probably included by chance in the original, unperturbed model. This conclusion is supported by the fact that test set prediction improves from $R^2_{\text{Test}} = 0.80$ to $R^2_{\text{Test}} = 0.84$ when these two variables are removed. Moreover, the figure reveals that five of the remaining variables are quite reliably included in the perturbed models and that a single variable is found in all models. This variable encodes the presence of a ketone and a double bond in ring A of the steroids and is of utmost importance for good model quality. Summing up, the generalized degrees of freedom accompanied by the selection statistics allow a facile assessment of the stability of the model at hand.

The next example shall highlight the differences between the characteristics of LOO-CV and LMO-CV with respect to stability. A data set of eye-irritating compounds (EYES) was employed for this purpose ($n = 25$, $n_{\text{Test}} = 13$). For LOO-CV as well as LMO-CV (leave-50%-out, $B = 3n$) the object–variable ratio was constrained to six, which means that no more than four variables are to be selected by REM-TS. The estimated generalized degrees of freedom for the modeling techniques differing only in the objective function are $\hat{D}(\text{M}) = 10.6$ for LOO-CV and $\hat{D}(\text{M}) = 7.0$ for LMO-CV, respectively. This indicates that the model selected by LOO-CV is less stable than the one selected by LMO-CV, which can also be seen in the corresponding variable selection statistics shown in Figure 5. In contrast to LMO-CV (lower panel), for which three of the four variables of the original, unperturbed model were selected very frequently (≥50%), LOO-CV (upper panel) exhibits rather poor statistics, i.e. all variables of the original, unperturbed model are selected with less than 30% relative frequency. For LOO-CV the strongest variable in the selection statistic is not even part of the unperturbed model, as are two other moderately strong variables. Although test set prediction was comparable in both cases (LOO-CV: $R^2_{\text{Test}} = 0.68$, LMO-CV: $R^2_{\text{Test}} = 0.65$), the LMO-CV shows a far better stability in model selection, which is important for the interpretation of the model. It is rather undesirable to mechanistically interpret variables that simply showed up owing to the peculiarities of the training set composition. Summing up, models selected by LOO-CV are not only characterized by a larger degree of inflation (see previous subsection) but are also more complex and less stable than those selected by LMO-CV.

Results of the validation protocol

Eight data sets were processed using the proposed validation protocol with default settings (see Table 3). For all data sets a test set consisting of one third of the entire data set was used to assess external predictivity. In addition to this estimate of test set predictivity the 'stabilized' version was also computed (i.e. R^2_{Test} based on the average of the predicted values of all perturbation runs which is referred to as $R^2_{\text{Test,Stab}}$). Not only the validation protocol did not change but also the employed structure descriptor (MaP) was applied with default values to a broad range of problems. In two out of the eight cases an extension was used, though. The results are shown in Figures 6 and 7.

Figure 6 shows the different R^2-values. The degree of inflation (R^2_{HO}) is small in all cases ($R^2_{HO} < 0.15$). This was actually expected since LMO-CV (leave-50%-out) and the early stopping rule prevent an overdose of model selection. In most cases the early stopping rule renders REM-TS to forward selection and in some cases to stepwise regression. Since the early stopping rule allows no detrimental moves it always stops in, or even before the first local optimum. It can be seen that this simple search algorithm performs quite

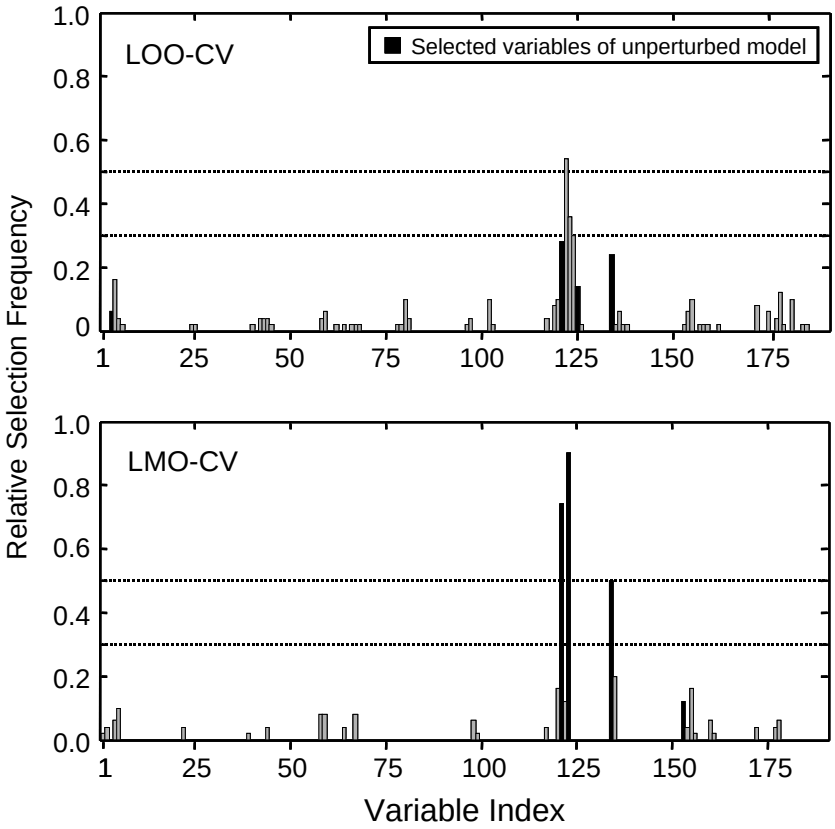

Figure 5. Variable selection statistics for the data set of eye-irritating compounds (EYES): upper panel: LOO-CV; lower panel: LMO-CV. The selection statistics depends on the chosen objective function. For LOO-CV a noisier, less reproducible selection statistics is observed.

well with respect to internal ($R^2_{CV-50\%}$) and external figures of merit ($R^2_{Test,}R^2_{Test,Stab}$). Basically, all of them are in a reasonable range. Moreover, internal and external predictivity is always comparable. The observed differences can be attributed to random fluctuations. Furthermore, it is interesting to note that the stabilized version of $R^2_{Test,}(R^2_{Test,Stab})$ does not yield significant improvements but that in some cases the opposite is true. Here, the $R^2_{Test,Stab}$-value even deteriorates but this difference can again be attributed to random fluctuations. Using ensemble averaging techniques such as bagging and boosting, improvement is not guaranteed [24, 35]. The same is true for the noise addition technique employed here. One possible explanation for this phenomenon is that the validation protocol already rigorously avoids overfitting, leaving little space for improvements by noise addition. Highly flexible modeling procedures and modeling procedures that tend to overfit benefit to a greater extent from stabilization [19, 24]. This could also be observed in an earlier study where stabilization was found to improve R^2_{Test}-values by up to 50% [18]. However, the largest improvement in that study was observed for REM-TS in combination with LOO-CV, which was the most flexible modeling procedure applied (i.e. it was most susceptible to overfitting).

Figure 7 shows the data set sizes, generalized degrees of freedom, and nominal degrees of freedom (number of principal components plus one). It can be seen that the generalized degrees of freedom that truly reflect model complexity and model stability are always by far larger than the nominal degrees of freedom. As a rule-of-thumb, a stable model should not use more than $(n/2) + 1$ generalized degrees of freedom, where the term '+1' is added because of sampling fluctuations of the statistic $\hat{D}(M)$. It can be seen that all models

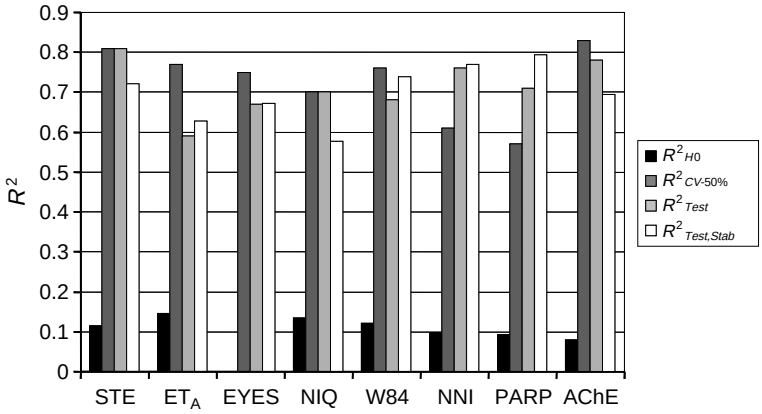

Figure 6. Results for eight data sets using the standard validation protocol and the MaP-descriptor: median of the permuted $R^2_{CV,PT}$-values (R^2_{HO}), original $R^2_{CV-50\%}$ value, R^2-values for the test set prediction (R^2_{Test}), and stabilized analog ($R^2_{Test,Stab}$).

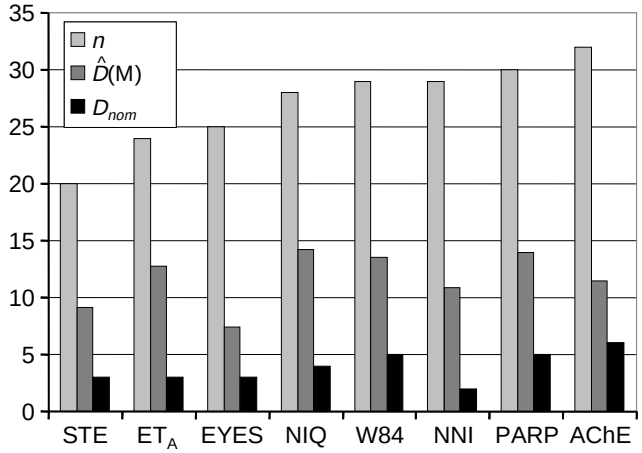

Figure 7. Results for eight data sets using the standard validation protocol and the MaP-descriptor: data set sizes (n), generalized degrees of freedom ($\hat{D}$(M)), and nominal degrees of freedom ($\hat{D}_{nom}$; number of principal components plus one).

are below this cut-off value. If a model showed generalized degrees of freedom larger than the cut-off, it is rejected as being overly complex and instable and is not used for predictions. The ET_A model and the NIQ model are close to this cut-off. Inspecting the variable selection statistics of the perturbation runs for both models shows that in both cases the two most important variables of the original model are selected reliably ($>45\%$) even though the response values were perturbed. Hence, the validity of these models is corroborated by the selection statistics. Further validation such as using multiple splits into training and test sets was used to assess their performance even more thoroughly. Once again, this validation step confirmed the validity of the models. However, these techniques are beyond the scope of this contribution.

Conclusions

Several diagnostic validation tools as well as a validation protocol for variable subset regression were presented. Using these tools the degree of inflation of the internal figures of merit, and the complexity and stability of models generated by variable subset regression can be assessed. The degree of inflation of internal figures of merit can be substantial. For a given data set it depends primarily on the maximum allowed number of variables in the final model (p_{max}) and the chosen

objective function. The larger the object–variable ratio (n/p_{max}) and the more rigorous the objective function is, the less will be the degree of inflation. However, with more demanding settings internal model performance also decreases. Put differently, the user has to trade the degree of inflation against internal model quality. Reasonable default settings are to use leave-50%-out cross-validation and $n/p_{max} \geq 6$. A possible indicator for the suitability of this trade-off is the difference between the original R^2_{CV}-value and the median of the $R^2_{CV,PT}$-values that can be obtained by chance alone (R^2_{HO}).

In addition to the objective function and n/p_{max}, the degree of inflation also depends on the applied mathematical modeling tool. For instance, it will be different for neural networks, regression trees, nearest neighbor regression, and multivariate regression techniques even if the same search algorithm is used. Moreover, the size of the data set plays an important role (the larger the data set, the smaller the degree of inflation) [12]. Since the degree of inflation depends on a number of different operational parameters, it should always be reported in addition to the probability of chance correlation. By doing so, the reported figures of merit can be better judged. In any case, reporting internal figures of merit for models generated by variable selection alone is not enough. External test set validation is mandatory, because the final model may be prone to a large selection bias.

Complexity and with it instability of models generated by variable selection was shown to be by far larger than the nominal degrees of freedom would suggest. Moreover, complexity and stability also depend on the objective function. LOO-CV was shown to be less stable than LMO-CV despite the fact that search algorithm and n/p_{max} were identical. The variable selection statistics obtained from the perturbation runs allows an easily accessible way to assess the stability of the generated model. By inspecting the relative selection frequency of a variable, its importance to the final model can be assessed.

The presented validation tools are part of a more sophisticated validation protocol, which was successfully applied to a broad range of real-world QSAR data with default settings. Hence, the protocol serves its purpose quite well. Owing to its rigorous nature it is expected to be applicable to other data sets and to safeguard against overfitting in variable selection.

Acknowledgements

This work was financially supported by the "Deutsche Forschungsgemeinschaft" (German Research Foundation); Grant: Sonderforschungsbereich 630, TP C5.

References

1. Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
2. Cruciani, G., Crivori, P., Carrupt, P.-A. and Testa, B., J. Mol. Struct., 503 (2000) 17.
3. Topliss, J.G. and Costello, R.J., J. Med. Chem., 15 (1972) 1066.
4. Topliss, J.G. and Edwards, R.P., J. Med. Chem., 22 (1979) 1238.
5. Zucchini, W., J. Math. Psychol., 44 (2000) 41.
6. Osten, D.W., J. Chemom., 2 (1988) 39.
7. Baumann, K., Albert, H. and von Korff, M., J. Chemom., 16 (2002) 339.
8. Baumann, K., von Korff, M. and Albert, H., J. Chemom., 16 (2002) 351.
9. Geisser, S., J. Am. Stat. Assoc., 70 (1975) 320.
10. Shao, J., J. Am. Stat. Assoc., 88 (1993) 486.
11. Cruciani, G., Baroni, M., Clementi, S., Costantino, G., Riganelli, D. and Skagerberg, B., J. Chemom., 6 (1992) 335.
12. Baumann, K., Trends Anal. Chem., 22 (2003) 395.
13. Shao, J., J. Am. Stat. Assoc., 91 (1996) 655.
14. Wehrens, R., Putter H. and Buydens, L.M.C., Chemom. Intell. Lab. Syst., 54 (2000) 35.
15. Rencher, A.C. and Pun, F.C., Technometrics, 22 (1980) 49.
16. Flack, V.F. and Chang, P.C., Am. Stat., 41 (1987) 84.
17. Hurvich, C.M. and Tsai, C.L., Am. Stat., 44 (1990) 214.
18. Baumann, K., Stiefl, N. and von Korff, M., In Ford, M., Livingstone, D., Dearden, J. and van de Waterbeemd, H. (Eds.), EuroQSAR 2002, Designing Drugs and Crop Protectants: Processes, Problems and Solutions, Blackwell Publishing, Oxford, UK, 2003, pp. 290–292.
19. Breiman, L., Ann. Stat., 24 (1996) 2350.
20. Coats, E.A., Perspect. Drug Discov. Des., 12–14 (1998) 199.
21. Stiefl, N. and Baumann, K., J. Med. Chem., 46 (2003) 1390.
22. Rao, R.C. and Toutenburg, H., Linear Models. 2nd edn., Springer, New York, 1999.
23. Ye, J., J. Am. Stat. Assoc., 93 (1998) 120.
24. Breiman L., Mach. Learning, 40 (2000) 229.
25. Klopman, G. and Kalos, A.N., J. Comput. Chem., 6 (1985) 492.
26. So, S.S. and Karplus, M., J. Med. Chem., 40 (1997) 4347.
27. Kubinyi, H., Hamprecht, F.A. and Mietzner, T., J. Med. Chem., 41 (1998) 2553.
28. Martens, H. and Naes, T., Multivariate Calibration, John Wiley & Sons, Chichester, UK, 1989.
29. Kubinyi, H., J. Chemom., 10 (1996) 119.
30. Selwood, D.L., Livingstone, D.J., Comley, J.C.W., O'Dowd, A.B., Hudson, A.T., Jackson, P., Jandu, K.S., Rose, V.S. and Stables, J.N., J. Med. Chem., 33 (1990) 136.

31. Krystek, S.R., Hunt, J.T., Stein, P.D. and Stouch, T.R., J. Med. Chem., 38 (1995) 659.
32. Robinson, D.D., Winn, P.J., Lyne, P.D. and Richards, W.G., J. Med. Chem., 42 (1999) 573.
33. Gancia, E., Bravi, G., Mascagni, P. and Zaliani, A., J. Comput.-Aided Mol. Des., 14 (2000) 293.
34. Baumann, K., Quant. Struct.-Act. Relat., 21 (2002) 507.
35. Breiman, L., Mach. Learning, 26 (1996) 123.
36. Freund, Y. and Schapire, R., In Saitta, L. (Ed.), Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufmann Publishers, San Francisco, CA, 1996, pp. 148–156.
37. Freund, Y. and Schapire, R., J. Comp. Syst. Sci., 55 (1997) 119.
38. Baumann, K., J. Chem. Inf. Comput. Sci., 42 (2002) 26.
39. Kennard, R.W. and Stone, L.A., Technometrics, 11 (1969) 137.
40. Wu, W., Walczak, B., Massart, D.L ., Heuerding, S., Erni, F., Last, I.R. and Prebble, K.A., Chemom. Intell. Lab. Syst., 33 (1996) 35.
41. Stiefl, N., Bringmann, G., Rummey, C. and Baumann., K., J. Comput.-Aided Mol. Des., 17 (2003) 347.
42. Faber, N.M., Chemom. Intell. Lab. Syst., 49 (1999) 79.
43. Jouan-Rimbaud, D., Bouveresse, E., Massart, D.L. and de Noord, O.E., Anal. Chim. Acta, 338 (1999) 283.
44. Golbraikh, A. and Tropsha, A., J. Mol. Graph. Mod., 20 (2002) 269.
45. Tropsha, A., Gramatica, P. and Gombar, V.K., QSAR Comb. Sci., 22 (2003) 69.
46. Kulkarni, A., Hopfinger, A.J., Osborne, R., Bruner, L.H. and Thompson, E.D., Toxicol. Sci., 59 (2001) 335.
47. Stiefl, N., Holzgrabe, U. and Baumann, K., In Ford, M., Livingstone, D., Dearden, J. and van de Waterbeemd, H. (Eds.), EuroQSAR 2002, Designing Drugs and Crop Protectants: Processes, Problems and Solutions, Blackwell Publishing, Oxford, UK, 2003, pp. 195–197.
48. Baumann, K. and Stiefl, N., In Ford, M., Livingstone, D., Dearden, J. and van de Waterbeemd, H. (Eds.), EuroQSAR 2002, Designing Drugs and Crop Protectants: Processes, Problems and Solutions, Blackwell Publishing, Oxford, UK, 2003, pp. 153–157.
49. Sippl, W., Contreras, J.M., Parrot, I., Rival, Y.M. and Wermuth, C.G., J. Comput.-Aided Mol. Des., 15 (2001) 395.
50. Barreca, M.L., Carotti, A., Carrieri, A., Chimirri, A., Monforte, A.M., Pellegrini Calace, M. and Rao, A., Bioorg. Med. Chem., 7 (1999) 2283.
51. Costantino, G., Macchiarulo, A., Camaioni, E. and Pellicciari, R., J. Med. Chem., 44 (2001) 3786.
52. Burman, P., Biometrika, 76 (1989) 503.
53. Mosteller, F. and Tukey, J.W., Data Analysis and Regression, Addison-Wesley, Reading, MA, 1977.
54. Picard, R.P. and Cook, R.D., J. Am. Stat. Assoc., 79 (1984) 575.
55. Kubinyi, H. and Abraham, U., In Kubinyi, H. (Ed.), 3D QSAR in Drug Design – Theory Methods and Applications, ESCOM Science Publishers, Leiden, The Netherlands, 1993, pp. 717–728.