Journal of Computer-Aided Molecular Design **18:** 563–576, 2004.
DOI 10.1007/s10822-004-4077-z

Statistical variation in progressive scrambling

Robert D. Clark* & Peter C. Fox
Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, USA

Received 20 July 2004; accepted in revised form 8 September 2004

Key words: cross-validation, PLS, progressive scrambling, redundancy, response randomization

Summary

The two methods most often used to evaluate the robustness and predictivity of partial least squares (PLS) models are cross-validation and response randomization. Both methods may be overly optimistic for data sets that contain redundant observations, however. The kinds of perturbation analysis widely used for evaluating model stability in the context of ordinary least squares regression are only applicable when the descriptors are independent of each other and errors are independent and normally distributed; neither assumption holds for QSAR in general and for PLS in particular. Progressive scrambling is a novel, non-parametric approach to perturbing models in the response space in a way that does not disturb the underlying covariance structure of the data. Here, we introduce adjustments for two of the characteristic values produced by a progressive scrambling analysis – the deprecated predictivity (Q_{s}^{*2}) and standard error of prediction (SDEP$_{\mathrm{s}}^{*}$) – that correct for the effect of introduced perturbation. We also explore the statistical behavior of the adjusted values (Q_0^{*2} and SDEP$_0^{*}$) and the sensitivity to perturbation ($\mathrm{d}q^2/\mathrm{d}r_{\mathrm{yy'}}{}^2$). It is shown that the three statistics are all robust for stable PLS models, in terms of the stochastic component of their determination and of their variation due to sampling effects involved in training set selection.

Introduction

Assessing the robustness of models involving quantitative structure–activity relationships (QSARs) is critical in medicinal chemistry applications, where being able to predict the biological activity of a new compound is the principal driver behind creating the models in the first place. In ‘classical’ modeling, descriptors (X_i) are assumed to be mutually independent variables and errors in the dependent variable (response Y) are assumed to represent a single normal distribution spread evenly across the range of responses. Under those conditions, estimating predictive error for ordinary least-squares regression entails a relatively straightforward partitioning of the variance in Y across the descriptors. The expected uncertainty in the prediction can then be calculated as a weighted sum of the variance of each individual descriptor (X_i) value about its training set mean plus a term representing the residual error of regression (SE) [1].

Unfortunately, the results of such calculations are reliable only when certain underlying assumptions are met – i.e. when the covariances between and among descriptors are negligible and the error in response is identically and independently distributed (IID) across Y. This is rarely the case for QSARs, which has led many workers in the field to turn to alternative approaches such as principal components regression (PCR [2]), partial least squares (PLS; also known as projection to latent structures [3]), or artificial neural nets [4]. Predictivity then has to be assessed empirically, because robust analytical measures do not exist. This is usually done by applying cross-validation, in which one or more of the observations are set

*To whom correspondence should be addressed. Fax: + 1-314-647-9241. E-mail: bclark@tripos.com

aside as a test set for which activities are predicted using a reduced model derived from the balance of the training set. This process is repeated, typically sampling without replacement until each observation has been predicted at least once. The test statistic q^2 is then calculated as

$$q^2 = 1 - \frac{\sum (y_j - \tilde{y}_j)^2}{\sum y_j^2} = 1 - \frac{\sum (y_j - \tilde{y}_j)^2}{(N-1)\mathrm{SD}_y^2} \quad (1)$$

The lower case (y) indicates a deviation from the mean response ($\bar{Y}$), $\tilde{y}_j$ denotes the predicted deviation of the response for the jth observation from that mean based on the reduced model from which that observation has been held back, and SD_y is the standard deviation for the full training set. Both summations are taken across all N observations in the training set.

The same equation is used to calculate the correlation coefficient R^2 for ordinary multiple linear regression (MLR) and to calculate the goodness-of-fit statistic r^2 for PLS and PCR models. In those cases, however, $\tilde{y}_j$ is the prediction for the jth observation using the full model, where *no* observations have been set aside.

The q^2 statistic was originally used in PLS analyses simply as a way to determine how many components can be included in a PLS model before over-fitting becomes a problem [5]. It has subsequently been used more broadly, however, as a general measure of model predictivity, including for the special case where the observations set aside are never folded back into the model – i.e. for external test sets. Here, that *external* predictivity statistic will be designated as q^2_{x} and the *internal* cross-validation statistic described above will be referred to as q^2_{cv}. In the former case, the summations in Equation 1 are only over the m observations in the test set, and the withheld test set observations do not contribute to either $\bar{Y}$ or SD_y.

Unfortunately, comparisons of q^2_{cv} to q^2_{x} have shown that good internal predictivity may or may not be a guarantee of good external predictivity [6–9], particularly when some form of variable selection or training set selection has been applied [10, 11]. This has led to reliance on a complementary approach (response randomization) wherein the response variables are scrambled across the data set and the analysis is rerun. If good q^2 values are still consistently obtained, it can be concluded that the QSAR method being used is simply fitting noise [5]. Unfortunately, the logical inverse is not true: getting low q^2 values for the scrambled data does not imply that the model is robust. In particular, the distribution of q^2 values obtained for a series of such scramblings cannot legitimately be used to calculate a probability that the q^2 observed for the intact responses arises by chance.

The problem is that most QSAR data sets encountered in medicinal chemistry are not random samples of structural space – not even of some limited space defined by one or more chemical series. Rather, they tend to be clumpy and more or less redundant due, in part, to the ease with which homologous subsets can be obtained synthetically. Literal redundancy – where a particular compound occurs more than once – is rare, which itself makes them non-random samples of structural space. Near redundancy (bromine vs. chlorine substitution, for example) is common, however. Consider the effect of response randomization for such pairs of observations, where activities and descriptor values will differ only slightly in most cases. Interchanging values within such a data set is likely to yield observations with y_j values that are substantially different – by $\sqrt{2}\mathrm{SD}_y$, on average. Injecting that level of anti-correlation into any data set will make it difficult or impossible to generate a model with nominally good predictivity statistics, regardless of how robust the original model may be, or how indiscriminate the modeling technique is.

The effect of redundancy on cross-validation statistics can be quite dramatic. Figure 1 shows the dependencies of the standard error of prediction ($\mathrm{SE}_{\mathrm{CV}}$) and q^2_{cv} on the complexity of HQSAR models [12] derived from 16 reverse transcriptase inhibitors. Plots are shown for the original data set as well as for data sets where half or all of the observations have been duplicated. Based on either statistic for the naïve data set (solid round symbols), the optimal model complexity is at $c = 1$ or 2 and the predictivity is modest ($q^2_{\mathrm{cv}} = 0.459$). The partially and fully redundant data sets contain the same information, yet their statistics imply that more complex models ($c = 3$ and 4, respectively) are justified, and that the models obtained are substantially more predictive (0.746 and 0.910, respectively).

These values are for runs in which only one observation is omitted at a time, a technique

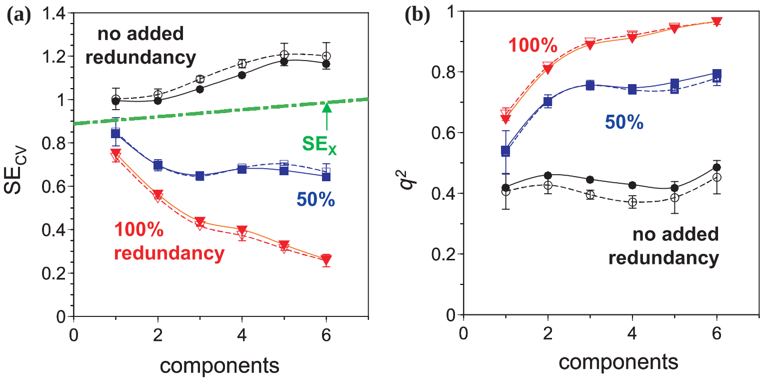


Figure 1. Dependence of cross-validation statistics on PLS model complexity for HEPT reverse transcriptase inhibitors where the training set has no added redundancy (●, ○; $N = 16$), 50% duplicated observations (■, □; $N = 1.5 \times 16 = 24$), or 100% duplicated observations (▼,▽; $N = 2 \times 16 = 32$). Open symbols represent LOO cross-validation statistics and closed symbols correspond to eightfold LSO cross-validation results. Fitted curves are simple splines. Descriptors were molecular holograms hashed to a length of 193. (a) Cross-validated standard error of prediction (SE$_{\mathrm{CV}}$); the green, doubly broken line indicates the curve for the external standard error of prediction against the remainder of the data set ($N = 85$). (b) q^2_{cv}.

known as 'leave one out' (LOO) cross-validation (filled symbols in Figure 1). Using larger cross-validation groups (leave some out) is often suggested as a remedy for the susceptibility of LOO to redundancy. The odds of leaving out *both* duplicates within any single run are small, however, so it has little effect on redundant data sets. The open symbols in Figure 1 correspond to an eightfold cross-validation analysis. The only obvious effect is to make the non-redundant model look worse (a result of having a smaller training set) and to make the redundant model look better.

Nor does response randomization behave as one might hope in this case. Figure 2 shows the distribution of q^2_{cv} values obtained in 100 replicate scramblings of the responses across the full range of the data set, with $c = 3$. The mean q^2_{cv} obtained is more negative for the original data set than for either of the adulterated versions (−0.778 versus −0.489 and −0.309), but it is not at all obvious by inspection that the distributions in Figure 2b, c are derived from seriously compromised models. In practice as in theory, a 'good' response randomization profile is no guarantee that a model is robust.

Several years ago our group developed an alternative approach – progressive scrambling – in which responses are scrambled only within quantiles rather than across the full range, thereby introducing a range of small perturbations into the data set [13]. The technique is closely akin to parametric perturbation analysis, in which model behavior is monitored as IID noise is added to the dependent (or independent) variables. Progressive scrambling is non-parametric, however, in that no assumptions are made about the normality or the

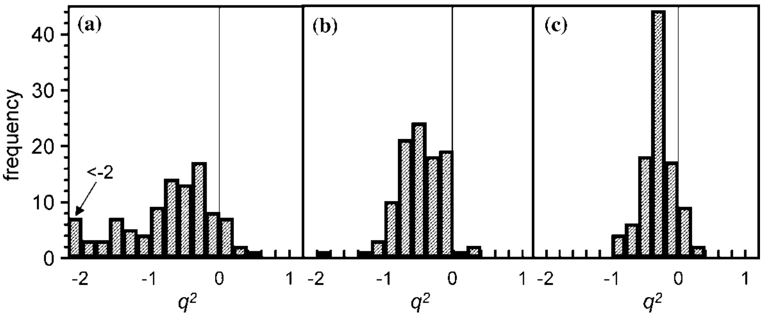


Figure 2. Effect of redundancy on full response randomization profiles for the HEPT inhibitor data set from Figure 1. (a) No added redundancy ($N = 16$). (b) Every other observation duplicated (50% redundancy; $N = 1.5 \times 16 = 24$). (c) Every observation duplicated (100% redundant; $N = 2 \times 16 = 32$). Vertical lines mark the bottom of the 0.0–0.2 bins.

correlation structure of the data set. In the original report, it was shown that models that exhibit good cross-validation statistics only because they are redundant are unstable to such perturbation, and several new indicators of model stability were introduced. Here, we propose adjustments that allow extrapolation of the values obtained back to the unperturbed model; explore the behavior of the indicators described in that report when the models in question are stable rather than pathological; and characterize them as statistics.

Methods

Data sets

Two artificial data sets were constructed by adding IID Gaussian random noise to linear and quadratic functions, with ordinates selected at random along a particular range of X values. A third, bimodal (clumpy) data set centered around two points placed near the ends of the linear data set's ranges was generated by adding Gaussian random noise independently to the X and Y coordinates of one or the other center. These three data sets were then modified by shifting the ranges and some points slightly so that the 'ordinary' and LOO cross-validated standard errors of regression (SE and $\mathrm{SE}_{\mathrm{CV}}$) of Y as a function of X were the same (2.00 and 2.12, respectively) for all three data sets.

Application of the method to such artificial data sets is informative, but application to a range of different 'real world' data sets is also desirable. Unfortunately, the number of available data sets that are both large enough and well-characterized enough to support the necessary sampling studies is small. We turned to a set of 101 reverse transcriptase inhibitors from the 1-hydroxyethoxymethyl-6-phenylthiothymine (HEPT) structural class [13–15]. Molecular holograms hashed to a length of 199 were used as descriptors for the data presented here [12], with default settings used for other HQSAR parameters [16]. The CoMFA models discussed in our earlier report [13] gave qualitatively similar results to the HQSAR analyses described here.

Experiments were also run on CoMFA models [17, 18] for subsamples drawn from a set of 304 cyclooxygenase 2 (COX-2) inhibitors built on a variety of azole, cyclopentenyl and phenyl scaffolds [8, 19] or on the 114 imidazoles included therein.

Potency expressed as pIC50 (i.e., −logIC50) served as the response variable in each case.

Scrambling

The mechanics of progressive scrambling are illustrated schematically in Figure 3. The first step is to re-order the observations from largest to smallest response. They are then blocked into $b_{\max}$ quantiles (here, 5), where $b_{\max}$ is the largest number of blocks such that each block contains at least three observations. There are 17 observations in this example, so they cannot be spread evenly across the five blocks. In such cases, block boundaries are adjusted so as to distribute the 'leftover' observations evenly among the centermost block(s) (Figure 3).

The responses within each block are then shuffled, and the modified pairings are submitted to PLS analysis to get the perturbed predictivity statistics $\mathrm{SE}^{*}_{\mathrm{cv}}, q^{*^2}$ and $r^2_{yy'}$ defined by

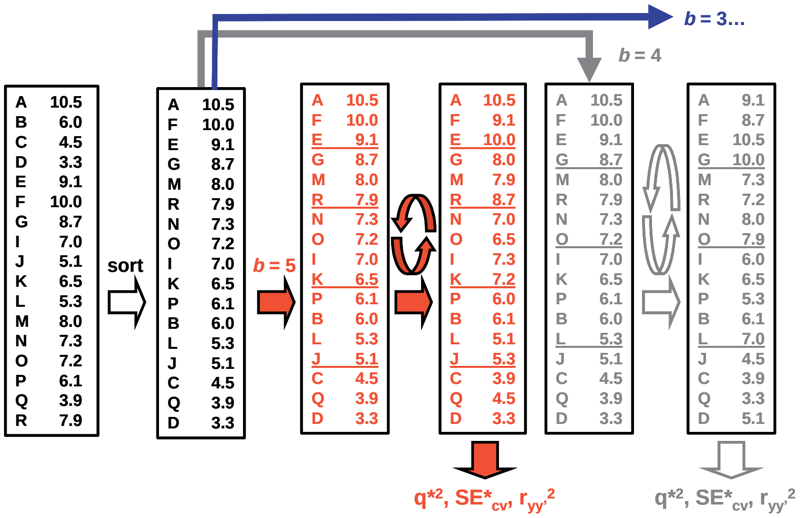
$$\mathrm{SE}^{*}_{\mathrm{cv}} = \sqrt{\frac{\sum(y'_j - \tilde{y}'_j)^2}{N - c - 1}} \tag{2}$$

$$q^{*^2} = 1 - \frac{\sum(y'_j - \tilde{y}'_j)^2}{\sum y'^2{}_j} = 1 - \frac{\sum(y'_j - \tilde{y}'_j)^2}{\sum y^2{}_j} \tag{3}$$

$$r^2_{yy'} = 1 - \frac{\sum(y'_j - y_j)^2}{\sum y_j^2} \tag{4}$$

Here, y' indicates the perturbed (scrambled) responses and c is the number of PLS components included in the model. A 'CV' subscript for q^{*^2} is implicit and so is omitted for the sake of simplicity. The statistic $r^2_{yy'}$ indicates the degree of correlation between the perturbed responses and the original ones.

Note that $\mathrm{SE}^{*}_{\mathrm{cv}}$ calculated using Equation 2 is deprecated by subtracting off the number of PLS components (c) in the denominator. This correction, which is rather *ad hoc* in nature [3], serves to penalize more complex models; it parallels the more directly justifiable correction for the number of descriptors or PLS components applied when calculating internal standard errors associated with goodness-of-fit. For regression methods other than PLS, some alternative measure of the degrees



A	10.5	A	10.5	A	10.5	A	10.5	A	10.5	A	9.1
B	6.0	F	10.0	F	10.0	F	9.1	F	10.0	F	8.7
C	4.5	E	9.1	E	9.1	E	10.0	E	9.1	E	10.5
D	3.3	G	8.7	G	8.7	G	8.0	G	8.7	G	10.0
E	9.1	M	8.0	M	8.0	M	7.9	M	8.0	M	7.3
F	10.0	R	7.9	R	7.9	R	8.7	R	7.9	R	7.2
G	8.7	N	7.3	N	7.3	N	7.0	N	7.3	N	8.0
I	7.0	O	7.2	O	7.2	O	6.5	O	7.2	O	7.9
J	5.1	I	7.0	I	7.0	I	7.3	I	7.0	I	6.0
K	6.5	K	6.5	K	6.5	K	7.2	K	6.5	K	6.5
L	5.3	P	6.1	P	6.1	P	6.0	P	6.1	P	5.3
M	8.0	B	6.0	B	6.0	B	6.1	B	6.0	B	6.1
N	7.3	L	5.3	L	5.3	L	5.1	L	5.3	L	7.0
O	7.2	J	5.1	J	5.1	J	5.3	J	5.1	J	4.5
P	6.1	C	4.5	C	4.5	C	3.9	C	4.5	C	3.9
Q	3.9	Q	3.9	Q	3.9	Q	4.5	Q	3.9	Q	3.3
R	7.9	D	3.3	D	3.3	D	3.3	D	3.3	D	5.1

Figure 3. Schematic representation of progressive scrambling.

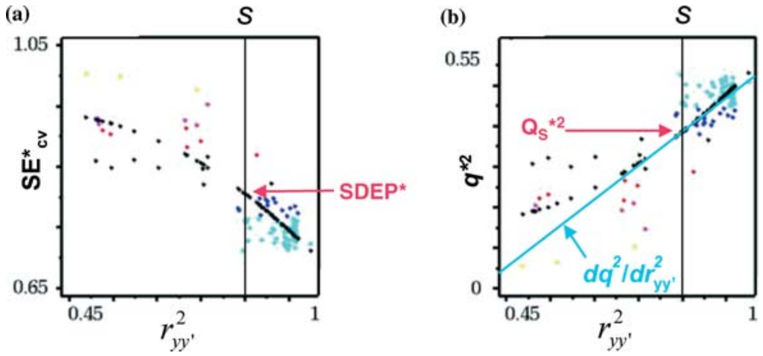


Figure 4. Progressive scrambling results as obtained in SYBYL 6.9.2 for CoMFA studies carried out on 100 COX-2 inhibitors. Scattered points are individual values colored by q^{*^2} value. Points along smooth curves are the corresponding fitted values for quadratic equations. (a) SE_{CV}; (b) q^{*^2}. Ordinate is the correlation between the original and the scrambled responses ($r^2_{yy'}$). The inset vertical line indicates the critical perturbation value (s).

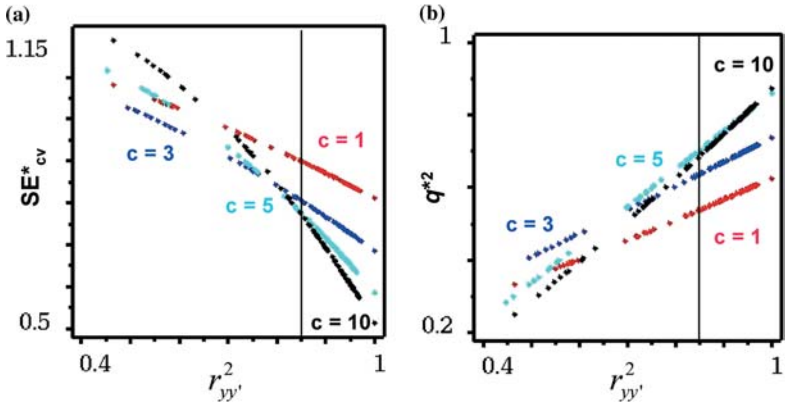


Figure 5. Effect of model complexity on scrambling profiles for (a) SE^*_{cv}, and (b) q^{*^2}. HQSAR analyses are for the full HEPT data set ($N = 101$) using molecular holograms as descriptors.

of freedom attributable to the model [20, 21] should be used instead.

The shuffling within blocks is repeated a prespecified number of times k (10, by default) at each binning level. The original sorted list of responses is then re-blocked into $b_{\max} - 1$ bins (4 in Figure 3), and the responses are randomized within each new block. Once the cross-validation statistics have been calculated for each perturbation, the responses are re-blocked into $b_{\max} - 2$ quantiles and the scrambling is repeated, with the process continuing down through $b_{\min}$ blocks.

Once scrambling is complete, the derived predictivity statistics ($\mathrm{SE}^*_{\mathrm{cv}}$ and q^{*^2}) are plotted against $r^2_{yy'}$ and fitted to quadratic or cubic functions, depending on the range of perturbations considered. The fitted curves are then used to estimate the standard error of prediction ($\mathrm{SDEP}^*_{\mathrm{s}}$) and predictivity ($Q^{*^2}_{\mathrm{s}}$) at some critical threshold level of perturbation s (SE and q^2 are used here to indicate point observations for a single model, whereas SDEP and Q^2 are used to indicate the corresponding population measures such as the summary statistics from a progressive scrambling run). Points corresponding to the individual scrambled data sets in Figure 4 are color-coded by q^{*^2} value in both cases. White points trace the predicted value along the fitted curves at each $r^2_{yy'}$ value.

The sensitivity to perturbation is calculated as the instantaneous slope of the predictivity with respect to the degree of perturbation ($\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$) evaluated at s.

In practice, the degree of the fitted polynomial makes little difference to the values obtained, except in the case of extremely unstable models. In SYBYL® 6.9.2, instantaneous slopes are calculated directly from the coefficients obtained from q^{*^2} plots. Indirect calculations based on coefficients from $\mathrm{SE}^*_{\mathrm{cv}}$ plots were used in our original report [13] because the regressions involved were expected to be better behaved in a statistical sense, given that q^2 involves a ratio of statistics whereas $\mathrm{SE}_{\mathrm{CV}}$ does not. The direct approach has proven itself marginally more satisfactory in most cases, however, and is less prone to errors in calculation [22].

Figure 5 illustrates the reason for choosing 0.85 as the default value of the critical point. The fitted quadratic curves for $\mathrm{SE}^*_{\mathrm{cv}}$ and q^{*^2} are shown for a range of model complexities. As the number of components increases, the slope of the corresponding $\mathrm{SE}^*_{\mathrm{cv}}$ curve increases and the slope of the q^{*^2} plot decreases. The vertical displacement of the curves also varies with complexity. The net result is that the curves for models above the complexity optimum (about $c = 5$ for this data set) cross near an $r^2_{yy'}$ value of 0.85 [13], making the number obtained using that critical value relatively insensitive to the choice of c. This degree of perturbation (15%) is very close to the amount of information removed (14%) in sevenfold LSO cross-validation.

Extrapolation

One problem with using SDEP^*_s and $Q^{*^2}_s$ as measures of model quality is that they are, by construction, very conservative. In particular, $Q^{*^2}_s$ has a maximum value of s. Besides being discouraging to the analyst, this significantly underrates the quality of the model under consideration. This could be dealt with by using randomization experiments to generate new probability tables and/or rules of thumb for significance, as was done for q^2_{cv} when PLS was first introduced for QSAR [23]. It is simpler to correct the statistics obtained for the noise added – extrapolating them, in a sense, back to the values expected for an unperturbed (but non-redundant) model.

Extrapolation for $Q^{*^2}_{\mathrm{s}}$ is simple – division by s yields the estimator $Q^{*^2}_0$, which is amenable to the same interpretation as is 'classical' q^2_{cv}. The corresponding correction for $\mathrm{SDEP}^*_{\mathrm{s}}$ is more complex:

$$\mathrm{SDEP}^*_0 = \sqrt{\frac{(2-s)(N-c-1)(\mathrm{SDEP}^*_s)^2 - (1-s)(N-1)\mathrm{SD}^2_y}{N-c-1}} \quad (5)$$

The form of the correction is complicated in part because the *model* sum of squares contribution to the response standard deviation (SD_y) should be reduced, but not the residual contribution reflected in $\mathrm{SDEP}^*_{\mathrm{s}}$. In addition, $Q^{*^2}_{\mathrm{s}}$ is a measure of the information in the model, whereas $\mathrm{SDEP}^*_{\mathrm{s}}$ is a measure of the noise not accounted for by the model, i.e. negative information. Hence the appropriate measure of the degree of perturbation in the latter case is 1− s rather than s itself.

Software environment

Regression and scrambling results presented here were obtained using the Advanced CoMFA module in SYBYL 6.9.2 [24], with LOO cross-validations carried out using SAMPLS [25]. All computations were carried out on SGI workstations operating in a unix environment.

Results

Artificial data sets

The effect of applying progressive scrambling to linear OLS models for three well-defined artificial data sets (Figure 6a) is illuminating. When Y is a linear function of X, plots for $\mathrm{SE}^*_{\mathrm{cv}}$ and q^{*^2} as functions of $r^2_{yy'}$ (Figures 6b and 6c) are both nearly linear. In principle, the loss of signal reflected in the reduction in q^{*^2} should be directly proportional to the applied perturbation $(1 - r^2_{yy'})$ when that perturbation is small. For this example, the relationship between q^{*^2} and $r^2_{yy'}$ is indeed linear, with a slope (0.9) that falls reasonably close to the theoretical value of 1.0. This agreement is particularly remarkable considering that the observed correlation between X and Y (r_{xy}) was 0.893 in this case.

In contrast, the scrambling plots for the data sets where the relationship between X and Y is quadratic or bimodal are decidedly curvilinear in both cases – positively so (concave down) for $\mathrm{SE}^*_{\mathrm{cv}}$ (Figure 6b) and negatively so (concave up) for q^{*^2} (Figure 6c). Slopes are much steeper for the bimodal data set for both statistics, with the slope for q^{*^2} in the absence of perturbation (i.e., at $r^2_{yy'} = 1$) being less than unity (0.8) for the quadratic case and greater than unity (1.5) for the bimodal case. This is a general effect: systematic error due to deviations from linearity will always be disproportionately 'smeared out' by scrambling, offsetting some of the perturbation effect.

The statistics obtained from such scrambling analyses – $Q^{*^2}_0 = 0.824 \pm 0.024$, 0.828 ± 0.001 and 0.813 ± 0.006 for the linear, quadratic and bimodal data sets, respectively – seem to better reflect the relative information content of the respective models than do the corresponding LOO q^2_{cv} values of 0.798, 0.773 and 0.890.

Precision

The randomization entailed in carrying out a progressive scrambling analysis is intrinsically non-deterministic. Stochastic effects show up at two different levels. One is sensitivity to the particular value used to seed the random number generator that drives the scrambling within quantiles, and the other is the effect of sampling variation when choosing a training set. Figure 7 shows the effect of using different random number seeds on predictivity curves for the full HEPT data set; although the particular q^{*^2} values obtained differ, the overall distribution and the fitted curves do not. This point is underscored by the results shown in Figure 8, where the fitted curves produced from four different random number seeds are overlaid. The mean (±SD) statistics obtained were $\mathrm{SDEP}^*_0 = 0.601 \pm 0.033$, $Q^{*^2}_0 = 0.823 \pm 0.004$, and $\mathrm{d}q^2/\,\mathrm{d}r^2_{yy'} = 1.064 \pm 0.033$.

The effect of structural variation in the training set is relatively greater, but not so much as to be problematic. To illustrate this, four 100 compound subsets were drawn at random from among the 304 making up the COX-2 data set, and the CoMFA models obtained were subjected to progressive scrambling. The plots obtained are shown in Figure 9. The curves for the replicate training sets diverge appreciably when the data has been substantially changed from its original state (i.e., at low values of $r^2_{yy'}$), but they are similar when the perturbation is small – particularly when close to the critical value of $s = 0.85$. The mean values obtained for this example were $\mathrm{SDEP}^*_0 = 0.780 \pm 0.014$, $Q^{*^2}_s = 0.455 \pm 0.026$, and $\mathrm{d}q^2/\mathrm{d}r^2_{yy'} = 0.997 \pm 0.144$.

Accuracy

The results presented above indicate that the statistics produced by progressive scrambling are *precise* enough to be useful, but they do not address the question of their *accuracy*. Considerable evidence has accumulated in recent years casting doubt on how well q^2_{cv} reflects predictivity on external test sets [6–8]; do SDEP^*_0 and $Q^{*^2}_s$ fare any better? One approach is to compare the internal measures with the corresponding statistics from external test sets across a population of training sets (Figures 10 and 11).

Training sets of 50 inhibitory imidazoles were drawn at random from the COX-2 data set. The complementary set of 64 imidazoles not selected served as the test set in each case. Each panel in Figures 10 represents a different level of model complexity (i.e. inclusion of a different number of

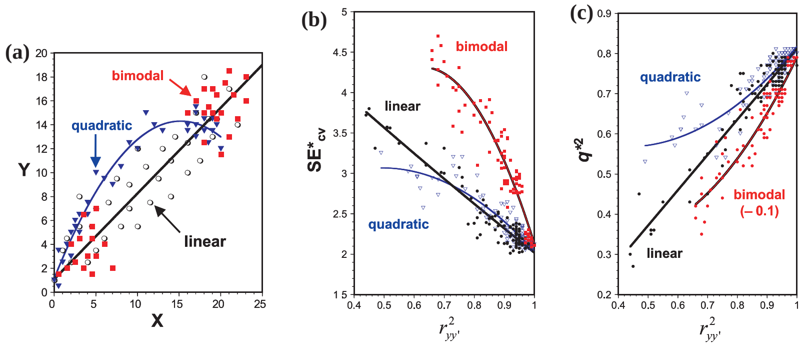


Figure 6. Artificially constructed data sets and their responses to progressive scrambling. Underlying relationships between X and Y were linear (○), quadratic (▼) or bimodal (■). (a) Original data. The bold straight line fits the linear and bimodal data sets. The curved line is quadratic. (b) $\mathrm{SE}^*_{\mathrm{cv}}$; all curves shown are quadratic fits. (c) q^{*^2}, with points corresponding to the bimodal data set offset to lower values by 0.1 units for clarity. Bimodal and quadratic data sets are fit to quadratic curves; the line shown for the linear data set is a linear fit.

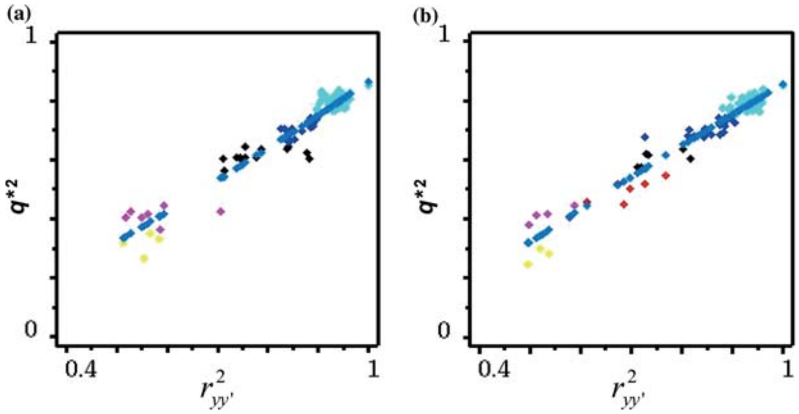


Figure 7. Replicate scrambling profiles for the full HEPT data set ($N = 101$) obtained using different random number seeds and $c = 8$. Scattered points correspond to individual perturbations and are color-coded by q^{*^2} value. The points along the smooth curve correspond to values as fit to a quadratic equation. Plots (a) and (b) were obtained using different random number seeds for scrambling.

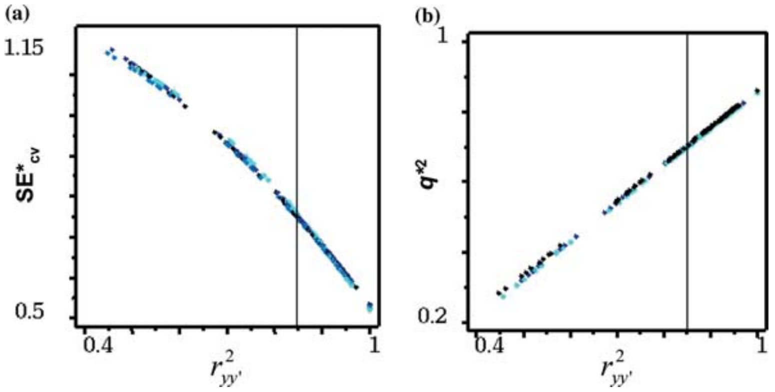


Figure 8. Variation fitted curves for progressive scrambling analyses with random number seed. The data set used was the same as for Figure 7. Fitted curves are shown for four different random number seeds, with points color-coded by seed. The inset vertical line indicates $s = 0.85$. (a) $\mathrm{SE}^*_{\mathrm{cv}}$; (b) q^{*^2}.

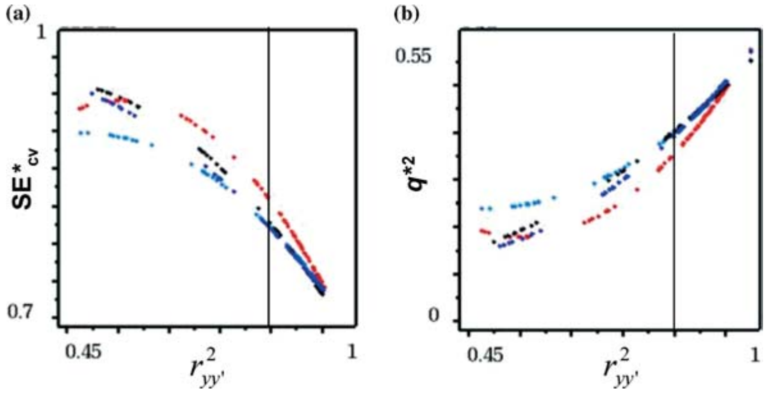


Figure 9. Variation of progressive scrambling profiles for 100 compound subsets drawn at random from the full COX-2 data set ($c = 5$). Each color represents a different training set, with only the fitted quadratic curves shown. The inset vertical line indicates $s = 0.85$. (a) $\mathrm{SE}^*_{\mathrm{CV}}$; (b) q^{*^2}.

PLS components in the model), with the abscissa indicating the standard error of prediction for the test set ($\mathrm{SE_x}$). The ordinate values reflect corresponding internal statistics for the training set – $\mathrm{SE_{CV}}$ (open symbols) and SDEP^*_0 (filled blue symbols).

Note that assigning a compound to a particular training set means that it does not appear in the corresponding test set. Moreover, any compound not assigned to the training set is automatically included in the test set. As a result, internal and external measures of predictivity are not statistically independent of each other [8, 10]. In fact, $\mathrm{SE_x}$ and $\mathrm{SE_{CV}}$ are inversely (though imperfectly) correlated, as shown by the regression lines in Figure 10. The slopes are all non-zero ($r^2 = 0.292$–0.503; $p \ll 0.01$ in all cases) but are also all less negative than −1 (−0.836 ± 0.052; range −0.766 to −0.902).

The negative correlation reflects natural variation in how much information individual observations carry. Including a distinctive molecule in

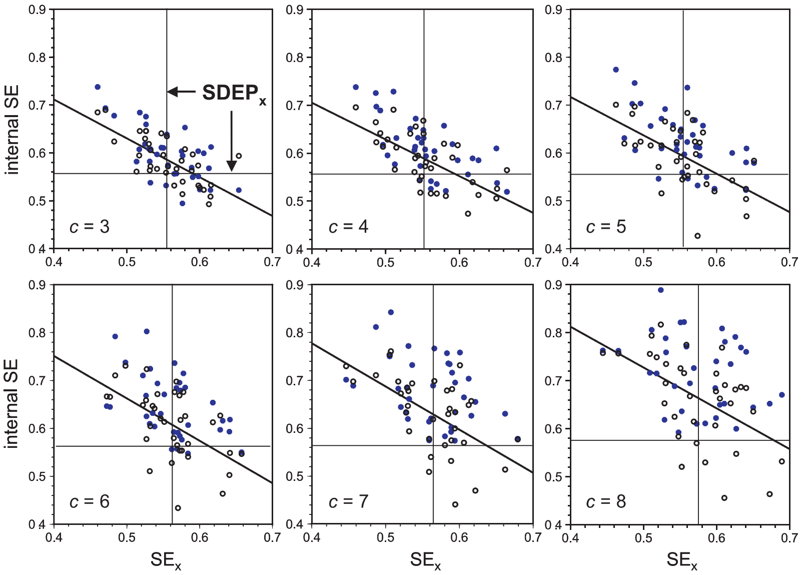


Figure 10. Variation of $\mathrm{SE_{CV}}$ (○) and SDEP^*_0 (●) across training sets as a function of external predictivity for the complementary test set ($\mathrm{SE_x}$). Each test set was comprised of 50 imidazoles drawn at random from the 114 in the COX-2 data set, with the remaining 64 held back as an external test set. The inset horizontal and vertical lines indicate the average value ($\mathrm{SDEP_x}$) for $\mathrm{SE_x}$ across all 35 test sets. Dark angled lines represent linear regressions of $\mathrm{SE_{CV}}$ on $\mathrm{SE_x}$.

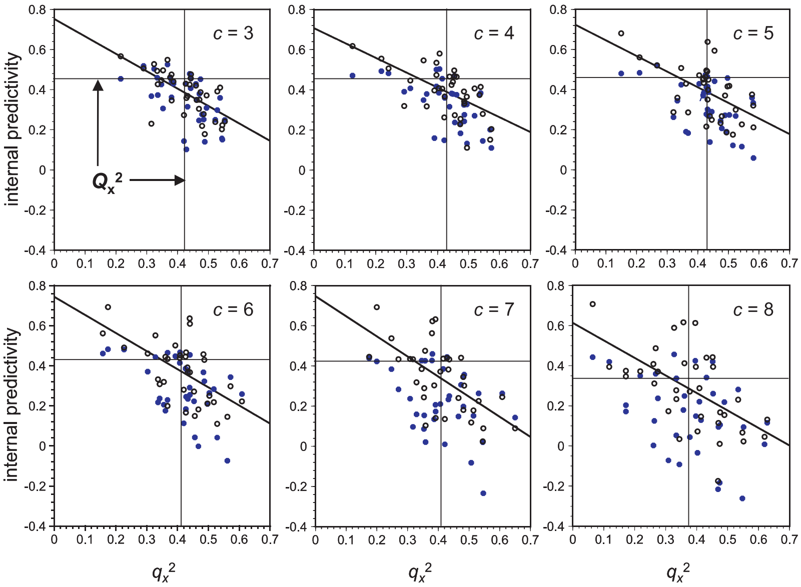


Figure 11. Variation of q^2_{cv} (○) and Q_0^* (●) across training sets as a function of external predictivity for the complementary test set (q^2_x). The same 50 compound training sets were used as for Figure 10. The inset horizontal and vertical lines indicate the average value (Q^2_x) for q^2_x across all 35 test sets. Dark angled lines represent linear regressions of q^2_{cv} on q^2_x.

the training set generally increases the external predictivity of the model produced but reduces internal predictivity. Placing such a molecule in the test set, on the other hand, increases the consistency of the training set and, hence, reduces internal error. Its properties will necessarily be difficult to predict, however, and the external error will generally increase.

Nonetheless, the *average* external predictivity – the average value across all training/test set combinations (SDEP_x – effectively the twofold SE_{cv} for the full data set) – should accurately reflect the properties of the population as a whole. The cross-hairs in each panel of Figure 10 represent the projection of SDEP_x onto the two plot axes. Each SE_{CV} regression line crosses the vertical SDEP_x axis above the horizontal one, with the offset increasing as model complexity increases. This shows that SE_{CV} is conservative, in that it is slightly biased towards overestimating external prediction error. Furthermore, that bias increases with increasing model complexity.

Of the four sectors defined by the cross-hairs, the lower left quadrant is the best of all worlds – external and internal errors are both small, which means that cross-validation (which is all the analyst would ordinarily have to go by [26]) correctly indicates that the model in hand is predictive. This quadrant is almost empty. Points in the lower right quadrant correspond to models where the internal cross-validation error is overly optimistic, which is clearly problematic. Points in the upper left quadrant represent a more subtle hazard; these are instances where the model is *more* predictive than indicated by the internal cross-validation statistics, potentially leading to rejection of a useful model. Such a situation typically arises when there is too *little* redundancy in the training set – e.g., when it is a result of applying a fractional factorial experimental design [27].

SDEP_0^* (filled blue symbols in Figure 10) is usually larger (more conservative) than SE_{CV} (Figure 10), a tendency that is more pronounced when the latter is overly optimistic (i.e., greater than SDEP_x). Conversely, cases where SDEP_0^* is smaller than SE_{CV} are decidedly more common to the left of the SDEP_x line in Figure 10. The net result is to reduce the number of training sets that put points into the undesirable lower right quadrant, particularly when the models in question are unjustifiably complex.

Figure 11 shows the corresponding plots for q^2_{cv} and Q_0^{*2} against q^2_x. Here the cross-hairs are defined by the average external predictivity Q^2_x, and 'the best of all worlds' lies in the upper right quadrant. Overly optimistic training sets fall in the upper left quadrant. The negative correlation between internal and external predictivity measures holds here as well, with regression slopes falling between −0.739 and −1.001 (mean −0.861 ± 0.093; r^2 0.323–0.437; $p \ll 0.01$ in all cases). Again, the internal cross-validation statistic is conservative, in that the regression line for q^2_{cv} crosses the vertical Q^2_{x} axis below the horizontal one.

As was seen for standard errors of prediction, scrambling consistently increases the accuracy of overly optimistic q^2_{cv} values, pulling them down towards the population value reflected in Q^2_{x}. Some of the overly pessimistic q^2_{cv} values falling in the lower right quadrant are improved as well. In most of the cases where $Q^{*^2}_0$ is less accurate than q^2_{cv}, the latter value was low because the response range for that particular training set was unusually small. A small value of SD_y reduces the q^2 (Equation 1). It is not involved in the calculation of standard errors, however (Equation 2), which explains why the effect is less apparent for SDEP^*_0 (Figure 10) than for $Q^{*^2}_0$ (Figure 11). It bears noting in passing that this dependence on SD_y is one reason predictive power is better characterized in terms of $\mathrm{SE_{CV}}$ than in terms of q^2.

Determining appropriate model complexity

One of the key uses of cross-validation is to determine the appropriate number of components to include in a PLS model. This is usually taken as the number of components above which q^2_{cv} first fails to increase significantly with an increase in complexity. More conservatively, it can be taken as the first point above which $\mathrm{SE_{CV}}$ fails to decrease.

For the artificial data sets described above (Figure 6), the response of q^2_{cv} to perturbation ($\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$) was qualitatively related to how well the complexity of the model matched that of the underlying relationship between X and Y; it was close to unity for the linear case ($\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$ for the perturbed model was 0.905 ± 0.055 when evaluated at $s = 0.85$); below unity for the quadratic case, where a second-order relation would be more appropriate; and greater than unity in the bimodal case, where a linear model is arguably too prescriptive. As noted above, a sensitivity near unity is expected to be optimal on theoretical grounds, which suggests that a number of PLS components yielding a $\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$ near 1 should be optimally robust. This hypothesis was examined by drawing random subsets of 75 COX-2 inhibitors as training sets and assessing the appropriate number of components to use by examining q^2_{cv}, q^2_{x} ($m =$ 229), or $\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$. For the instantaneous slope, that number of components was chosen for which $\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$ was maximal but less than 1.2 (the pooled SD was 0.100 across replicate scramblings, which is consistent with our general experience). The results are shown in Figure 12, where the training sets have been sorted in order of increasing q^2_{cv}. In 90% of the cases, scrambling indicates that a complexity between 4 and 5 components is appropriate, which matches results for the full data set and for the other two statistics. Internal and external predictivities are much more variable and often disagree, as expected, especially at either extreme; neither is as reliable an indicator of complexity as $\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$. Somewhat surprisingly, this is the case even when $Q^{*^2}_0$ is marginally significant.

Effect of introduced redundancy

The results presented above are intended to characterize the statistical properties of progressive scrambling analyses, and so focus on well-behaved, stable data sets. How does the method behave when applied to pathological cases? Figure 13a shows the dependence of $\mathrm{SDEP}^*_{\mathrm{s}}$ (open symbols) and SDEP^*_0 (filled symbols) on model complexity for HEPT training sets for which none, half or all of the observations have been entered in duplicate, along with that for the external predictive error, $\mathrm{SDEP_x}$ ($m = 85$). Figure 13b shows the plots for $\mathrm{d}q^2/\mathrm{d}r^2_{yy'}$, which indicate that no more than a single component can be justified for the original, non-redundant model. In contrast, two components would be appropriate for the case where half of the observations have been duplicated, and up to three components for the fully redundant model. The optimal complexity obtained across 13 independently drawn training sets averaged 2.00 (±0.71). The corresponding curves for $\mathrm{SE_{CV}}$ and q^2_{cv} are shown in Figure 1a and 1b; considerably more complex models are suggested by these statistics for both redundant models.

In this case the training set has too little redundancy, so simple cross-validation will consistently underestimate the predictive power of the model

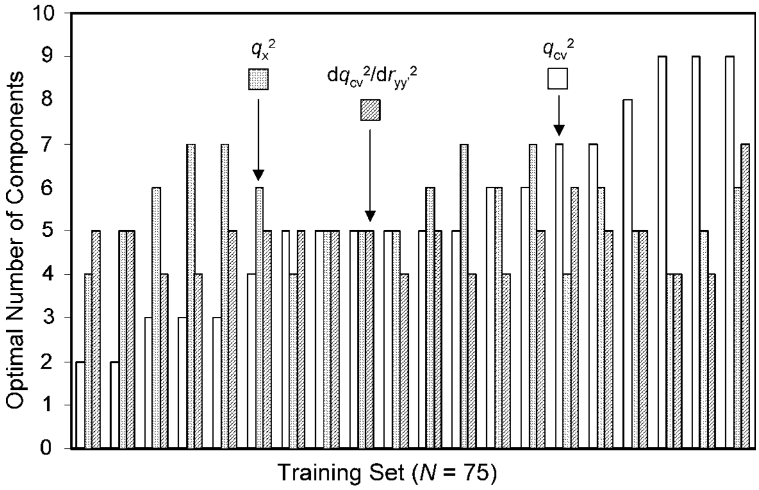


Figure 12. Variation across training sets in the optimal number of components as determined by q^2_{cv} (open bars), q^2_x (shaded bars) or $dq^2/dr^2_{yy'}$ (hashed bars). Training sets ($N = 75$) were taken from the full COX-2 data set; the 228 held back served as the test set in each case. Results are sorted in order of increasing complexity, as indicated by q^2_{cv} for clarity.

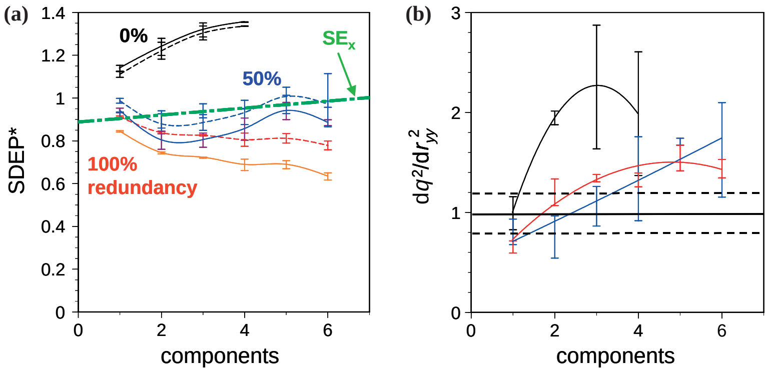


Figure 13. Dependence of SDEP estimates from progressive scrambling on model complexity for a 16 compound HEPT training set with no added redundancy (●, ○; $N = 16$, 0% redundancy); with alternate observations duplicated (■, □; $N = 1.5 \times 16 = 24$, 50% redundancy); or with every observation duplicated (▼, ▽; $N = 2 \times 16 = 32$, 100% redundancy). Error bars indicate standard deviations of the respective statistics for duplicate scramblings using different random number seeds. (A) Open symbols represent SDEP^*_s and closed symbols represent SDEP^*_0. The doubly broken line indicates the external predictivity ($\mathrm{SE_x}$) obtained against the balance of the HEPT data set ($N = 85$). (B) Perturbation sensitivity ($dq^2/dr^2_{yy'}$) as ordinate. The fitted curves shown are second order. The solid and dotted horizontal lines indicate the range (1.0 ± 0.2) within which the sensitivity corresponding to the optimum number of components is expected to fall.

and scrambling will only serve to exacerbate this excessive pessimism [8]. Systematically adding redundancy to the training set overcompensates, leading to a gross underestimate of predictive error (Figure 1a). It is interesting to note that subsequent deprecation by applying progressive scrambling yields SDEP* values quite close to $\mathrm{SDEP_x}$ when scrambling is used to determine an appropriate level of model complexity (Figure 13). This is reminiscent of the effect expected of applying LOO cross-validation to a fractional factorial [28] design. Further work will be needed to determine how useful such an approach might be in general.

Discussion

There is a clear need for statistically powerful ways to evaluate the stability and predictivity of regression models. Holding back randomly selected observations to create an external test set is effective and reliable, but the models evaluated are then necessarily less powerful than those that would result from including most or all of the available observations in the training set. This is not a trivial concern, since the test set must represent a large fraction – probably half – of the available observations if the statistics derived from it are themselves to be reliable. For relatively large data sets ($N \geq 100$), a stratified sampling validation technique such as boosted leave-many-out [8] can be used. For smaller data sets, LOO cross-validation has the advantage of representing a minimal perturbation to the full model whose predictivity is being sought. Both approaches may be seriously compromised, however, if the data set in question is significantly redundant.

In fact, there will be a problem with standard methods of cross-validation any time the error in the observed responses is not IID. Inclusion of near-duplicate observations is a common case in point, due in large part to filling out structural space in support of patent filings. Such redundancy may only involve a handful of pairs of similar structures, but it can also arise from excessive clumpiness due to having several 'tight' structural series within a data set, with activity varying much more from series to series than within a series. The bimodal data set represented by the red squares in Figure 6a is an extreme example of this; unfortunately, the problem is not always so self-evident when many descriptors are involved. Exhaustively (or nearly so) combinatorial designs entail a more subtle form of redundancy but can be equally problematic.

Unevenly distributed deviation from the functional form to which the data are being fit is another common source of correlation among errors. This is illustrated by the quadratic data set (blue triangles) in Figure 6a, where the systematic error is largest in the middle and at either end of the distribution, though the problem is by no means limited to linear regression models. One side effect of this is a large discrepancy between internal and external errors. A training set chosen mostly from the left or right side of the plot, for example, will be more predictive internally but more poorly predictive externally.

Minimizing differences in error correlation between the training and test set is the prime motivator for assigning observations to one or the other at random. In a sense, statistical experimental design of the training set takes a complementary approach: predictive error within the training set is *maximized* by including those observations whose statistical leverage is greatest, and internal predictivity is simply disregarded. The applicability domain of the model obtained is then reduced, however, because it encompasses only the molecules in the test set. Nor is the danger of non-IID error eliminated altogether, since the make-up of the training set is conditioned on the data set as a whole.

The existing alternatives to LOO cross-validation that are most often employed are leave-some-out cross-validation and (full) response randomization. Neither reliably identifies inflated predictivity statistics that are due to redundancy or other forms of error correlation, however. In contrast, the progressive scrambling technique characterized here can effectively overcome such distorting effects in many cases. Adding redundancy of any kind simply makes the high $r^2_{yy'}$ end of the scrambling plots more densely populated [see the plots for the bimodal (clumpy) data set in Figure 6b and 6c]. The sampling studies described here show that the statistics produced by such analyses are themselves robust and reliable when applied to stable and predictive models.

A potential drawback of progressive scrambling as originally formulated [13] is that the statistics produced reflect the properties of a perturbed model, and so are, by their nature, overly conservative. Here we have addressed this shortcoming by introducing adjustments to SDEP^*_s and Q_s – Equation 5 and re-scaling by s, respectively – and show that these adjustments yield statistics (SDEP^*_0 and Q^*_0) that are in good agreement with analyses of the full population. This should prompt a wider consideration of the technique, which is by no means limited in applicability to PLS. At a minimum, it represents a substantial improvement over the complete scrambling entailed in response randomization, where a failure to reject an hypothesis of stability is all too often interpreted as positive evidence that the hypothesis is true.

Acknowledgements

The authors would like to thank the reviewers of our original manuscript for their helpful suggestions for improvement, particularly as regards

broadening the discussion of the conditions under which 'ordinary' cross-validation can be expected to break down.

References

1. Snedecor, G.W. and Cochran, W.G., Statistical Methods, 10th Edition, Iowa State Press, Ames, IA, 1989.
2. Martens, H. and Næs, T., Multivariate Calibration, Wiley, Chichester, UK, 1989.
3. Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H. (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 523–550.
4. Zupan, J. and Gasteiger, J., Neural Networks in Chemistry and Drug Design, 2nd Edition, Wiley-VCH, Weinheim, Germany, 1999.
5. Wold, S. and Eriksson, L., In van de Waterbeemd, H. (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, Germany, 1995, pp. 309–318.
6. Tropsha, A., Grammatica, P. and Gombar, V.K., QSAR Comb. Sci., 22 (2003) 69.
7. Golbraikh, A. and Tropsha, A., J. Mol. Graph. Model., 20 (2002) 269.
8. Clark, R.D., J. Comput.-Aided Mol. Des., 17 (2003) 265.
9. Hawkins, D.M., Basak, S.C. and Mills, D., J. Chem. Inf. Comput. Sci., 43 (2003) 579.
10. Baumann, K., von Korff, M. and Albert, H., J. Chemom., 16 (2002) 351.
11. Hawkins, D.M., J. Chem. Inf. Comput. Sci., 44 (2004) 1.
12. Heritage, T.W. and Lowis, D.R., In Parrill, A.L. and Reddy, M.R. (Eds.), Rational Drug Design: Novel Methodology and Practical Applications, ACS Symposium Series 719, American Chemical Society, Washington, DC, 1999, pp. 212–225.
13. Clark, R.D., Sprous, D.G. and Leonard, J.M., In Höltje, H.-D. and Sippl, W. (Eds.), Rational Approaches to Drug Design, Prous Science, Barcelona, Spain, 2001, pp. 475–485.
14. Kireev, D.B., Chrétien, J.R., Grierson, D.S. and Monneret, C., J. Med. Chem., 40 (1997) 4257.
15. Luco, J.M. and Ferretti, F.H., J. Chem. Inf. Comput. Sci., 37 (1997) 392.
16. HQSAR™ is distributed by Tripos, Inc., St. Louis, MO; www.tripos.com
17. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
18. Cramer III, R.D., DePriest, S.A., Patterson, D.E. and Hecht, P., In Kubinyi, H. (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 443–485.
19. Chavatte, P., Yous, S., Marot, C., Baurin, N. and Lesiur, D., J. Med. Chem., 44 (2001) 3223.
20. van der Voet, H., J. Chemom., 13 (1999) 195.
21. Kalivas, J.H., Forrester, J.B. and Seipel, H.A., J. Comput.-Aided Mol. Design, 18 (2004) 537 (this issue).
22. In fact, Equation 5 in Ref. 13 includes a typographical error, with sSDEP′ substituted for *s*.
23. Clark, M., Cramer III, R.D., Jones, D.M., Patterson, D.E. and Simeroth, P.E., Tetrahedron Comput. Methodol., 3 (1990) 47.
24. Advanced CoMFA® and SYBYL® are distributed by Tripos, Inc., St. Louis, MO; www.tripos.com.
25. Bush, B.L. and Nachbar Jr., R.B., J. Comput.-Aided Mol. Design, 7 (1993) 587.
26. Given that the most statistically powerful model will always be the one based on all available observations [Refs. 9–11].
27. Otto, M., Chemometrics, Wiley-VCH, Weinheim, Germany, 1999.
28. A full factorial design includes *two* observations for each first-order factor, each of which is a partial replicate of its complement in the descriptor space (see Ref. 27).