Journal of Computer-Aided Molecular Design **17:** 265–275, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics

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

MS received 22 November 2002; accepted for publication 10 December 2002

Key words: cross-validation; dissimilarity selection; molecular diversity; OptiSim; PLS; projection onto latent structures; representativeness; boosted LMO

Summary

It is becoming increasingly common in quantitative structure/activity relationship (QSAR) analyses to use external test sets to evaluate the likely stability and predictivity of the models obtained. In some cases, such as those involving variable selection, an internal test set – *i.e.*, a cross-validation set – is also used. Care is sometimes taken to ensure that the subsets used exhibit response and/or property distributions similar to those of the data set as a whole, but more often the individual observations are simply assigned 'at random.' In the special case of MLR without variable selection, it can be analytically demonstrated that this strategy is inferior to others. Most particularly, D-optimal design performs better if the form of the regression equation is known and the variables involved are well behaved. This report introduces an alternative, non-parametric approach termed 'boosted leave-many-out' (boosted LMO) cross-validation. In this method, relatively small training sets are chosen by applying optimizable k-dissimilarity selection (OptiSim) using a small subsample size ($k = 4$, in this case), with the unselected observations being reserved as a test set for the corresponding reduced model. Predictive errors for the full model are then estimated by aggregating results over several such analyses. The countervailing effects of training and test set size, diversity, and representativeness on PLS model statistics are described for CoMFA analysis of a large data set of COX2 inhibitors.

Introduction

It is rare that a regression model is constructed solely to concisely summarize the behavior of a data set of interest, with no desire or intention of predicting the results for observations falling outside the data set; generally there is interest in making interpolated or (cautiously) extrapolated predictions as well. If one is lucky enough to have many observations and a relatively small number of mutually independent variables descriptors *and* the error in the response variable(s) is normal and independently and identically distributed (IID) across the responses, the standard error of prediction (SDEP) can be directly calculated from the standard error (SE) of multiple linear regression (MLR). In fact, if the cited conditions are met, that SDEP value is a BLUE statistic – a best linear unbiased estimator. Moreover, the analytical expressions involved can be expanded to accommodate inhomogeneous variance in errors, so long as the dependence of error on the descriptors used is well-behaved and known. Unfortunately, few or none of these conditions are realized for most of the data sets upon which quantitative structure-activity relationship (QSAR) analyses are based. In particular, population sampling in structural space is usually heavily biased in practice, so that descriptors are not independent of one another. In that case, the residual sum of squares upon which calculation of SE is based will be consistently underestimated, leading to an unjustifiably inflated confidence in the model obtained [1].

Alternative least-squares techniques developed to cope with various limitations on the applicability of MLR to real data sets include supervised and un-

*Correspondence. E-mail: bclark@tripos.com supervised variable selection [*e.g.*, 2,3], principal component regression (PCR [4]), and projections to latent structures (PLS [5, 6]). None of these techniques allow for direct and unbiased estimation of predictive error (or, equivalently, confidence intervals) when descriptors are not independent [7], though some attempts have been made to do so ([8, 9, 10]). An analogous dilemma arises for models based on artificial neural nets [*e.g.*, 11].

Hence empirical cross-validation approaches have needed to be used to estimate SDEP for such regression methods, with predictive errors being estimated based on errors in prediction for a subset of observations omitted during model construction [12]. The most conservative such approach is to select a *training set* of *n* observations from which a model is to be derived, and to set aside the remaining N – n observations as a *test set* whose values are predicted using the model obtained. The sum of squared deviations of those predictions from the actual values (PRESS) can then be used to calculate an external standard error of prediction, SDEP_X.

$$SDEP_X = \sqrt{\frac{PRESS}{N - n}}$$

This has become a 'gold standard' of sorts for QSAR, particularly for supervised variable selection methods involving iterative cross-validation [13].

When model building and evaluation are 'cheap,' it is simple to reduce the risk of accidentally getting a 'bad' test or training set by generating many pairs of sets by random assignment of observations to one or the other, then consolidating the predictive errors obtained. In *leave-some-out* (LSO) cross-validation, each observation appears in one test set (*i.e.*, in one *cross-validation group*). If the number of observations is not evenly divisible by the number of cross-validation groups desired, uneven groupings must be used or multiple instances need to be allowed. Alternatively, cross-validation groups can be obtained by drawing several random samples independently but without replacement, in which case some observations are likely to not make it into any test set. Drawing multiple random samples *with* replacement (*bootstrapping*) [11, 14] will produce somewhat different results for most data sets.

Use of external test sets and LSO cross-validation both cause some concern because of their random aspects. Moreover, both methods entail generating predictivity statistics from significantly reduced models, which are necessarily less statistically powerful *on average* than is the model based on the full data set. Both concerns are often finessed by employing *leave-one-out* (LOO) cross-validation. In this approach, each of the N observation in the data set is set aside in turn, and the value of the omitted response is predicted using a reduced model generated from the other $N - 1$ observations. An LOO cross-validated standard error (SE_CV) is calculated from the residual predictive sum of N squared deviations (PRESS) obtained. By convention, the number of degrees of freedom for this sum of squares is taken as $N - c - 1$, where c is the number of components in the reduced models [15]. Hence:

$$SE_{CV} = \sqrt{\frac{PRESS}{N - c - 1}}$$

This statistic is nominally deterministic, in that its value is defined exactly for any given training set, and is based on minimally reduced models. Unfortunately, comparison of internal predictive error (SE_CV) with that found for completely external test sets indicates that this measure confounds predictivity of the model with redundancy in the data set being considered [13, 16]. The same holds true for the associated predictivity, q^2:

$$q^2 = 1 - \frac{(N - c - 1)SE_{CV}^2}{(N - 1)SD^2}$$

where SD denotes the standard deviation of the training set (the implied model for the null hypothesis is the mean of the training set). This statistic is widely quoted in the literature but is not well behaved with respect to sampling even for quite large sample sizes. Hence this report will focus on standard errors instead.

The overly optimistic statistics obtained using LOO can be made more reasonable by using LSO cross-validation. It is not at all obvious, however, how many cross-validation groups should be used in any given case, and confounding of predictivity with redundancy remains a problem for larger data sets. As a result, several other approaches have been developed to address concerns about the statistical power and indeterminacy of randomly constructed test and training sets. If enough is known about the population and the functional form relating the response to the descriptors, it can be shown analytically that experimental design [17, 18] (typically D-optimal design for least squares regression) can be used to identify optimal training sets. Unfortunately, the 'true' functional is not adequately known or determinable for most systems of interest to QSAR analysts. The most prominent alternative strategies employed for creating external test sets include balanced sampling across the response range and/or structural classes [2, 19, 20] and maximizing training set diversity by some pre-set criterion [21, 22]. The former emphasizes making both test and training sets as *representative* as possible. The latter favors assignment of the 'most unusual' compounds to the training set to increase the statistical power of the models obtained. This has the potential drawback that some of the compounds of greatest interest – those 'at the edge' of the data set – will be excluded from the SDEP calculation, though they often exemplify exactly the kind of modest extrapolation critical in real-world applications.

Here optimizable *k*-dissimilarity selection (OptiSim) [23, 24] is used to systematically vary the balance between representativeness and diversity of training and test sets for comparative molecular field analysis (CoMFA [25]) of a literature COX2 data set [20]. The results obtained lead naturally to a non-parametric method for identifying series of training and test set pairs that, taken together, provide an efficient and robust measure of model predictivity for full models. Because of its non-parametric nature, such a 'boosted leave-many-out' approach should be applicable not only for other PLS analyses, but in almost any regression context where external predictivity is of interest.

The COX2 data set

Carrying out systematic cross-validation studies places unusually severe demands on a data set. It must be large enough to allow meaningful sampling statistics to be obtained; structurally heterogeneous enough that redundancy – especially uneven redundancy – will not distort the statistics obtained; and be comprised of responses evenly distributed over a fairly wide range of values. For analyses involving 3D QSAR, there needs to be a clear alignment rule as well. Within the limits of those constraints, the data set must be reasonably representative of QSAR data sets likely to be encountered in practice.

COX2 inhibitors constitute one such data set. Chavatte *et al.* [20] have compiled an extensive list of structures and IC50s for 305 compounds falling into five major structural classes (Figure 1): most (114) are *N*-aryl imidazoles (**1**), with the bulk of the remainder comprised of *N*-substituted pyrroles (**2**), pyrazoles (**3**) and cyclopentenes (**4**), and *o*-diaryl benzenes (**5**). Six compounds in the data set – two cyclopentadienes, two isoxazoles, an *N*-cyclohexyl pyrrole, and a thiophene (DuP-697) – fall outside of these five classes. A 3-(1-hydroxy-2,2,2-trifluoroethyl)-pyrrole analog was omitted because of its chirality. In most cases, one of the aryl substituents (denoted by Ar in Figure 1) is a 4-methanesulfonyl- or a 4-amidosulfonylphenyl moiety. A largely overlapping data set of 314 inhibitors compiled by Kauffman and Jurs [2] is likely to be useful in future cross-validation studies, but many of the 'extra' compounds contained therein only have limiting values for the IC50 (*e.g.*, > 5000 μM) associated with them.

The molecular alignments used for the current study will be described in detail so as to facilitate their potential use in future studies and by other groups.

3D structures were constructed in SYBYL 6.8.1 [26] based on the structures given in the supporting data provided in [20], then regularized using CONCORD [27]. This produced conformers in which the vicinal aryl substituents were both perpendicular to the central ring. The positioning of *ortho* and *meta* substituents with respect to the central ring varied, however, depending upon the exact substitution pattern, as did positioning of the sulfonyl group with respect to the other rings. Hence inter-ring torsions were modified by 180° rotations where necessary so as to position the 2′- and 3′-substituents on the same side of the central ring as much as possible.

Imidazole template **1a** was oriented along its principal axes [28] to define a coordinate system, then each analog was aligned to that template by rotation and translation so as to minimize the RMSD between four atoms in the template (circled in Figure 1) and the corresponding atoms in the analog. This was accomplished using the DATABASE ALIGN command in SYBYL. The alignments produced are shown in Figure 2A, where the view is down along one of the XY diagonals. This differs somewhat from the alignments employed by Chavatte *et al.*, who matched the torsions between the peripheral Ar and Ar' rings and the central ring to the respective 161 and 101° torsions found in the crystal structure for SC-558 (**3a**) bound to the mouse enzyme (1CX2 [29]).

Atomic partial charges were calculated using the GAST_HUCK option in SYBYL, which utilizes the Gasteiger-Marsili method [30] for σ charges and an extension thereof for distributing charges across π bonds. Steric and electrostatic fields were calculated using the default parameters: a cationic sp^3 carbon atom bearing a unit charge as probe across a rectilin-

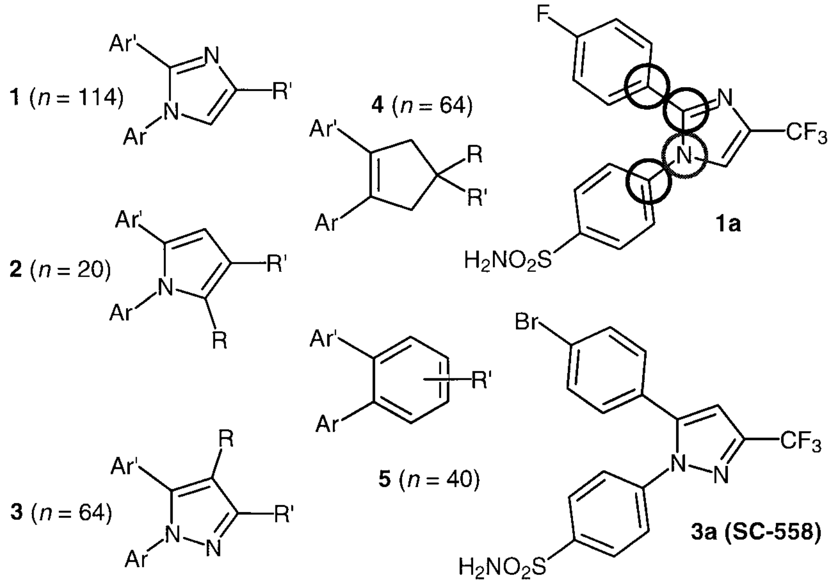

Figure 1. Generic structures for compounds in the data set, where Ar is characteristically an otherwise unsubstituted 4-amidosulfonyl- or 4-methanesulfonylphenyl moiety; see the supplemental material for the original reference [20] for a full list of structures. Two pyrrole (**2**) and four pyrazole (**3**) examples bear 4-halo- or 4-methoxyphenyl groups at this position instead. Ar' represents one of a broad range of variously substituted phenyl, pyridyl and other heteroaromatic groups; one pyrrole (**2**) example bears a cyclohexyl group at this position. The substituent labeling shown corresponds to the alignment rule used; see text for details.

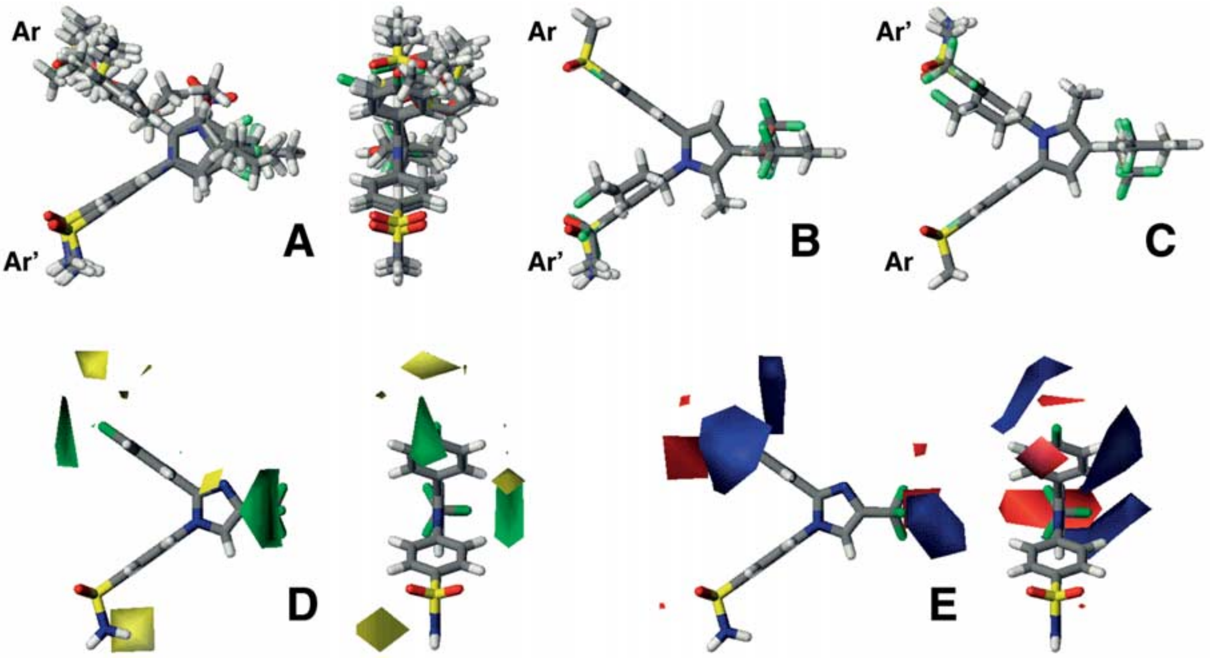

Figure 2. Alignment rules considered and CoMFA field contours obtained for the full data set. (**A**) Overlay of examples taken from structural classes **1**, **3**, **4** and **5** showing the torsions applied to get conformers such that bulky substituents on the top (Ar') ring lie to the right of the central ring orientation shown. (**B**) Orientation used for pyrrole (**2**). (**C**) Alternative orientation considered for pyrroles but rejected. (**D**) Steric field contours for the full CoMFA model, with yellow indicating the areas where steric bulk is relatively disfavored and green indicating areas where steric bulk is favored. The contours are drawn at the 20 and 80% levels of the STDEV*COEFFICIENT levels. (**E**) Electrostatic field contours for the full CoMFA model. Blue contours enclose areas where negative charge on the ligand is disfavored (*i.e.*, low energy for the sp^3-C^+ probe atom is associated with high activity) and red contours enclose areas where negative charge is favored.

Table 1. CoMFA statistics for models by structural class and considered *en masse.*

Class (structure)	N	mean ± SD	c	SE_{CV}	q^2
imidazoles (**1**)	114	6.56±0.74	7[a]	0.504	0.559
pyrroles (**2**)	21	6.60±0.76	1	*0.816*	*−0.088*
pyrazoles (**3**)	70	7.08±0.76	5	0.694	0.233
cyclopentenes (**4**)	62	7.87±0.88	7	0.417	0.800
benzenes (**5**)	40	7.74±0.66	6	0.482	0.547
full model	304	7.11±0.93	5	0.601	0.591
			6	0.594	0.602
			7	0.589	0.609

[a] Number of component latent variables included in each model.

ear 2Å lattice. Hydrophobic fields were utilized in the analyses of Chavatte *et al.* but were not included here.

The five analog classes afforded CoMFA models of varying quality (Table 1). The model statistics cited here are in good qualitative agreement with the CoMFA results reported by Chavatte *et al.* [20], with models based on imidazoles (**1**), cyclopentenes (**4**) and o-diarylbenzenes (**5**) giving the best LOO cross-validation statistics (Table 1) and pyrroles (**2**) giving a very weakly predictive model (Table 1). The statistics for the full model were consistent with the statistics for those restricted by class. Perhaps more importantly, the models were consistent across classes. In particular, the cycolpentene model ($n = 62$) was predictive of pyrrole activity using the alignment shown in Figure 2B, where the pyrrole nitrogen is matched to the imidazole nitrogen in **1a**. The cyclopentene model was not predictive, however, for pyrroles oriented using the alternative alignment rule based on placement of the sulfonyl groups (Figure 2C). The former gave a slope of 1.075 between the predicted and observed pIC50s, with a residual SE of ±0.665 [31]. In contrast, predictions obtained for the alternative alignment were actually *negatively* correlated with observed values (slope $= -0.515$). The model based on imidazole analogs gave qualitatively similar results. Hence the pyrrole orientation shown in Figure 2B was used for subsequent perturbation studies of the consolidated model.

Figures 2D and 2E show the steric and electrostatic contours obtained for the full model. These are rather less fragmented than those reported by Chavatte *et al.* [20], possibly as a result of differences in alignment and the greater simplicity of the model considered here (two fields instead of three and eight components instead of seven).

Selection of training and test Sets

Optimizable k-dissimilarity selection (OptiSim [23, 32]) provides a useful non-parametric means of selecting training sets varying in representativeness and diversity. The method involves drawing a series of random subsamples of size k from the population of interest, then selecting from each subsample a 'best' candidate for inclusion in the selection set. When diversity is used to determine the best candidate, the selection sets obtained become progressively less representative and more diverse a s k is increased. The limiting cases of $k = 1$ and $k = N$ correspond to random and maximal dissimilarity selection, respectively [23].

The criterion for 'best' used in the work described here was structural diversity with respect to those compounds selected from preceding subsamples in terms of Tanimoto similarity with respect to UNITY substructural fingerprints [23, 33, 34, 35]. The similarity between a candidate and the set of compounds already selected was taken as the similarity to the candidate's nearest neighbor in the set [35, 36]. The initial selection was a single representative drawn at random from the full population.

In other applications, it is advantageous to filter out redundant candidates from the subsamples using an exclusionary, minimum *dis*similarity threshold r [23]. That option was not employed here, so as to make the extreme case of $k = 1$ synonymous with purely random sampling. Other than that, standard methodology was used. In particular, sampling was done without replacement, and in such a way as to ensure that each individual was considered for inclusion in a subsample once before any candidate was considered a second time.

The effect of k on the test set is indirect but basically complementary to that on the training set, since the former is comprised of all compounds in the population *not* selected. Hence increasing k will reduce the diversity of the test set and increase its representativeness. This effect is modest for small values of k and n. It can become significant, however, as the product of the number considered per subsample (k) and the number selected (n) approaches N, the size of the data set as a whole. When the product of the two parameters equals the size of the data set, each candidate will have been considered exactly once when selection is complete.

Results: predictivity vs diversity

The differences among predictivity statistics shown in Table 1 for the full model using five, six or seven components are small. Moreover, the COX2 data set employed here, though more diverse than many QSAR data sets its size, is still redundant to some degree. Such redundancy can inflate predictivity statistics and lead to over-specification (*i.e.*, inclusion of too many latent variable components in the model). This expectation was substantiated by the results obtained from progressive scrambling stability analysis [16], which indicated that the models obtained at all three levels of complexity were comparably robust. A conservative approach was therefore taken for subsequent analyses, with the complexity of reduced models restricted to five components. This helped stabilize the statistics obtained with smaller training sets, making the results clearer and easier to interpret. Results obtained with six and seven component models were qualitatively similar, however.

Five training sets were selected at a series of progressively larger sizes n using OptiSim with subsample size set to $k = 1$, 2, 3, 4, 5 or 6, each training set having been started with a different random number seed. The internal cross-validation statistics obtained are plotted in Figure 3 as the root mean square across the five training sets for each combination of k and n. The abscissa is linear in $\sqrt{n}$, since this is how the variances and information content are expected to scale with sample size [37]. As expected, the internal predictivity determined by LOO cross-validation improves as the size of the training set n increases regardless of the diversity of the training set selected, resulting in a drop in cross-validated standard error ($\mathrm{SE_{CV}}$; Figure 4). Increasing training set diversity by increasing k consistently yields higher $\mathrm{SE_{CV}}$ values; given that some QSAR does indeed exist, this, too, is expected. The more diverse a training set is, the more likely each single individual it contains is to carry unique information making it difficult to predict using the information borne by its fellows.

The magnitude of this effect exhibits a striking and somewhat unexpected dependency on both diversity and training set size. $\mathrm{SE_{CV}}$ jumps sharply between $k =$ 3 and $k = 4$ for small ($n \leqslant 50$) training sets, whereas all non-random training sets ($k > 1$) exhibit elevated errors for larger training sets ($n \geqslant 75$). It is apparent that the internal predictive error plateaus even when the diversification of the training set is relatively small (k = 2 to 4) (Figure 3). Not surprisingly, the data is relatively noisy for small training sets, but the trend is clear. The inset shows the corresponding curves for q^2 (Figure 5). The effect on this statistic is distorted somewhat by the fact that the variance of the training set – which is factored into q^2 but not into $\mathrm{SE_{CV}}$ – generally also increases with diversity.

This behavior stands in sharp contrast to that of the standard error of *external* prediction ($\mathrm{SDEP_X}$) and of its corresponding predictivity statistic, Q^2 (Figure 4). Predictions for compounds not included in the training set are *most* accurate (*i.e.*, $\mathrm{SDEP_X}$ is smallest) at the same subsample size ($k = 4$) at which internal predictivity is poorest. The inverse relation between internal and external predictivity evident when Figure 4 is compared to Figure 3 has been noted by others, particularly in connection with using q^2-guided variable selection [13, 38, 39].

The trend is especially clear for external predictivity because it is more stable for any given k and n than is $\mathrm{SE_{CV}}$. This is in part because the number of data points going into the $\mathrm{SDEP_X}$ calculation (*i.e.*, the size of the test set) is largest for models based on the smallest training set, which is when the models are least stable. For this data, the standard error of the mean for $\mathrm{SDEP_X}$ ranged from 0.004 to 0.031 across the 60 combinations of k and n examined, but with no clear pattern of variation. The grand root mean square was ± 0.021, which is the size of the exemplary error bar indicated in Figure 4 for $k = 4$ and $n = 50$.

The broadly general inverse correlation between internal and external predictivity is also evident in Figure 5, which shows $\mathrm{SE_{CV}}$ values for individual training sets as a function of $\mathrm{SDEP_X}$ (in the interests of clarity, only models with $65 \leqslant n \leqslant 100$ are shown). The randomly selected training sets indicated with open symbols are evenly distributed above and below the 1:1 line that defines an unbiased estimator. In contrast, the internal cross-validation error found for the more diverse training sets lie above that line, indicating that $\mathrm{SE_{CV}}$ from such models consistently overestimates the 'real' predictivity estimated by $\mathrm{SDEP_X}$. The edge of the distribution is formed by models based on training sets created using $k = 4$ (denoted by the filled circles in Figure 5); these points form a rather sharp band with minimal error in external predictivity despite a wide range of internal predictivity.

The situation is brought into sharper relief when the respective root mean square values across replicates are plotted for the full spectrum of training set sizes and levels of diversity examined (Figure 6). Here the plot is dominated by the effect of training set size,

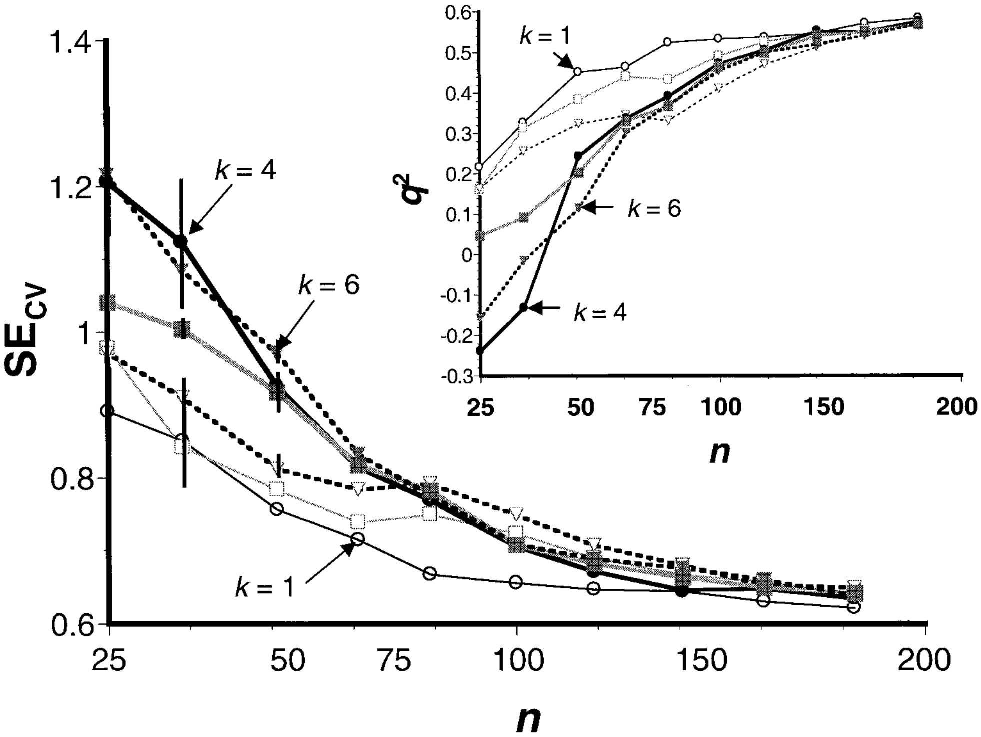

Figure 3. Plot of internal cross-validated standard error (SE_{CV}) as a function of training set size n and diversity. The OptiSim subsample sizes k used to select the training sets were 1 (○—○), 2 (□—□), 3 (▽– –▽), 4 (●—●), 5 (■—■) and 6 (▼– – ▼). Each point represents the root mean square error for five different training sets, each selected using a different random number seed. Superimposed vertical bars indicate the estimated standard error for each RMS. The inset shows the corresponding curves for the internal predictivity coefficient q^2.

especially for the more diverse training sets ($k > 2$). The data obtained at each particular level of training set diversity show the expected consistent increases in accuracy (decrease in $SDEP_X$ *and* SE_{CV}) with increasing N; the inverse correlation evident within each set of replicates is clearly overwhelmed by this trend when the replicates are considered in aggregate.

It is reassuring that despite the tendency for the internal and external predictivities of *individual* training sets to vary inversely (Figure 5 and [13, 38, 39]), the root mean squares for the random ($k = 1$) case fall reasonably close to the 1:1 SE_{CV}:$SDEP_X$ line that defines an unbiased estimator. Note, too, that though the curves shown in Figure 6 are purely empirical quadratic fits intended primarily to aid visualization rather than for any theoretical significance, the underlying relationships follow them quite well and are surprising linear, at least for this data set. This is particularly evident for the more diverse data sets ($k \geqslant 3$), where the curves are steeper than for the more representative training sets, with the points from the $k = 4$ case forming the limiting curve.

Discussion

The results presented here suggest a straightforward alternative approach to ‘classical’ LOO and LSO cross-validation that has some of the best attributes of both. Such an approach entails using OptiSim to create a series of t training sets, then calculating the residual sums of squared deviations for each complementary training set with respect to the corresponding reduced PLS model. Each OptiSim run takes the same values of k and n as arguments, but utilizes a different random number seed. The cumulative PRESS with $t(N - n)$ degrees of freedom [40] obtained by summing across all test sets yields an SDEP for the full model based on all N observations.

Such a ‘boosted leave-many-out’ (boosted LMO) approach can be seen as a nonparametric variation of the stratified sampling technique used for sampling populations in which some individuals exert a disproportionately large influence on the particular statistic of interest [1]. Here, the bias is in favor of including ‘more diverse’ compounds in the training set, since these have greater leverage in the calculations and are

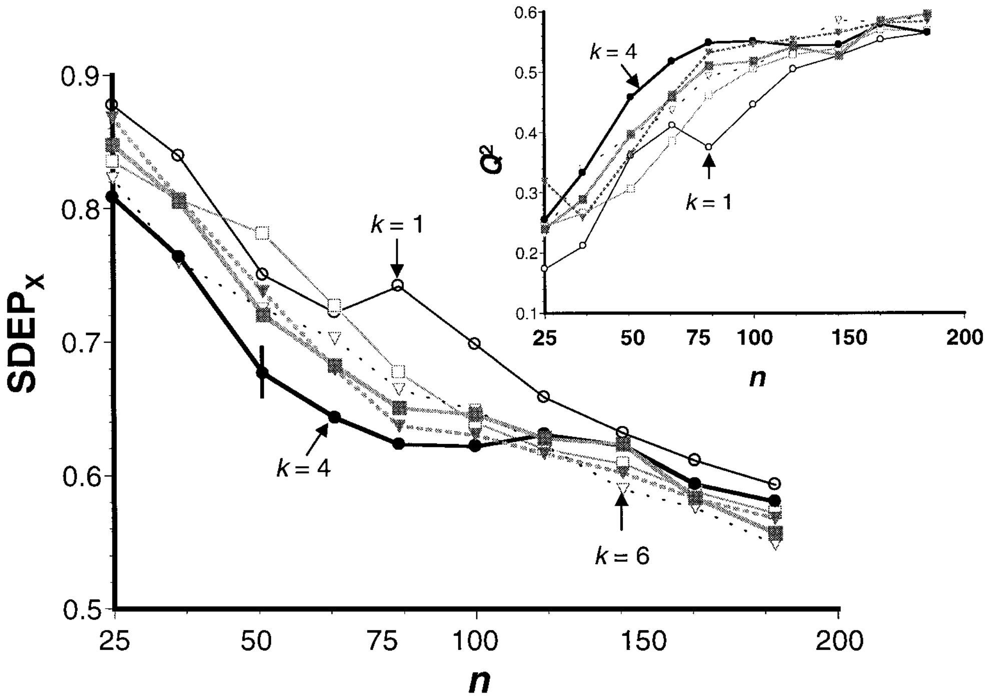

Figure 4. Plot of external standard error of prediction (SDEP$_X$) as a function of training set size and diversity. The symbols used match those specified for Figure 3. The single vertical bar shown indicates the SEM corresponding to the grand mean for the variation in RMSE across all pairs of k and n; it is applicable to every data point shown. The inset shows the corresponding values obtained for the external predictivity coefficient Q^2.

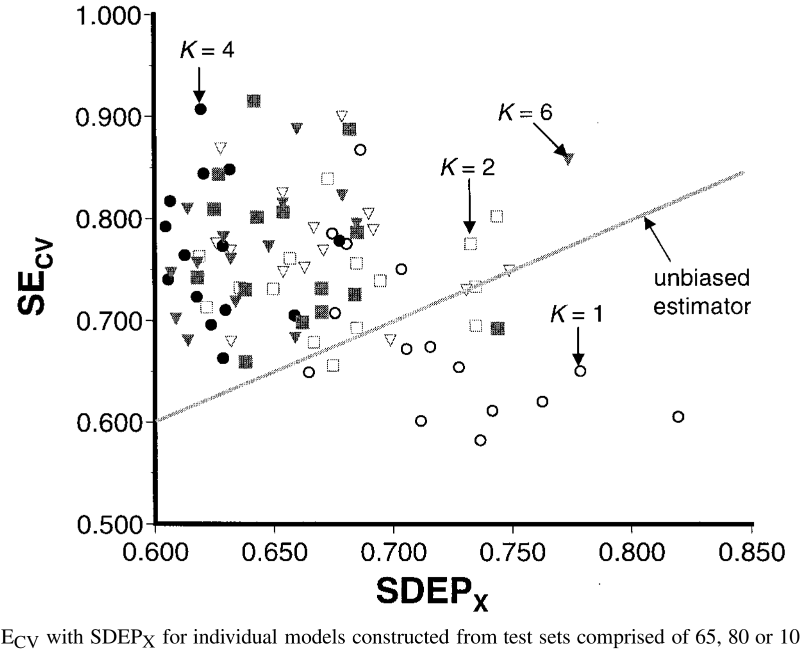

Figure 5. Variation in SE$_{CV}$ with SDEP$_X$ for individual models constructed from test sets comprised of 65, 80 or 100 observations. The line indicates the unbiased estimator line along which the two statistics are equal. See Figure 3 for an explanation of the symbols used.

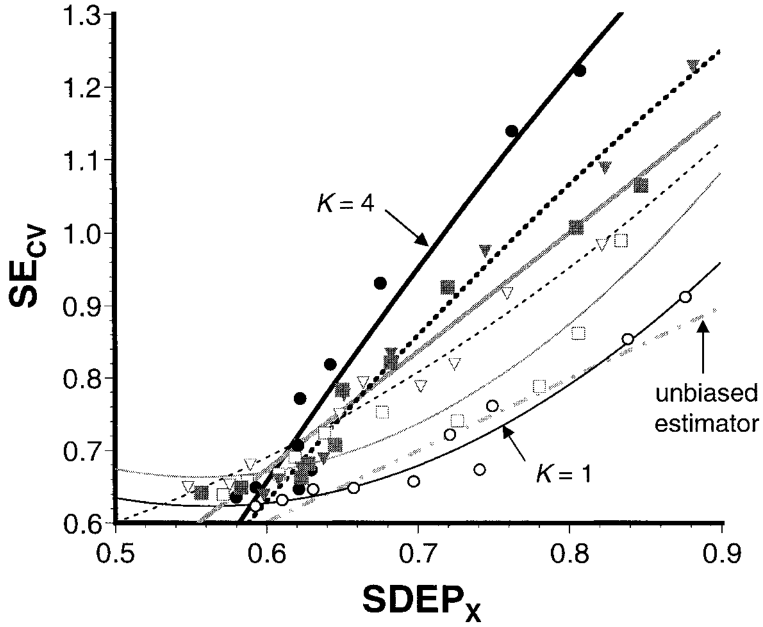

Figure 6. Variation in SE_{CV} with $SDEP_X$ for the root mean square statistics evaluated across the five models obtained from the OptiSim-selected training sets using each pair of *k* and *n* values. See Figure 3 for an explanation of the symbols used. The doubly broken line (– · – · –) indicates the unbiased estimator line along which the two statistics are equal. The curves shown are least squares quadratic fits to the results obtained at each OptiSim subsample size *k*.

more likely to unduly inflate the estimate of predictive error for the full model if they are placed in the test set instead (as indicated by comparing SE_{CV} to $SDEP_X$ for diverse training sets; Figure 6). Moreover, to the degree that the data set is a biased sampling of structural space, such compounds tend to be especially informative about compounds of potential interest falling at the 'edges' of the data set as a whole.

The results presented here indicate that a little diversity goes a long way in this regard, in that models constructed from the most diverse training sets ($k = 5$ and 6) were less reliably predictive than were those based on OptiSim selections using $k = 4$. Evidently representativeness is comparable in importance to diversity in this context – studies with other data sets have shown that applying OptiSim using values of k from 3 to 5 generally increases selection set diversity with little loss in representativeness [23, 24]. A greater tendency to over-train may account for some of the rather unexpected penalty for having too much diversity in the training set, and breakdown in the adequacy of a linear model may account for another portion [41]. Whatever its source, it is fortunate in that the use of a relatively small k in selecting training sets results in more diverse test sets. As noted above, the $SDEP_X$ values obtained from such test sets are likely to better reflect expectations for 'new' compounds of the greatest interest. Indeed, the optimality of $SDEP_X$ for $k = 4$ is, if anything, underestimated with respect to the more diverse training sets due to the removal of harder-to-predict compounds from the test sets for the latter.

What values of k and n should be used? The choice of $k = 4$ will probably be good in general, though not necessarily optimal. Given that choice, setting $n = N/4$ is reasonable, since that will result in each individual in the data set being visited once; it is unlikely to be accidental that $SDEP_X$ and Q^2 level off above this point ($n = 76$) for the COX2 data set. In any event, considering each individual in the population at least once ensures that isolated outliers are very likely to go into the training set. This seems prudent, since they are most likely to contribute to the applicable scope of the final model and unduly deflate predictivity estimates.

A smaller training set may prove more suitable for very large N. It may also be desirable to introduce an exclusion threshold (r) [23] where that is the case, since very large data sets tend to be particularly redundant. A smaller value of k will likely be more appropriate for small data sets.

Note that the internal predictivity reflected in the SE_{CV} and q^2 statistics is ignored in such a cross-validation analysis. The degree of fit reflected in the 'classical' MLR SE and r^2 statistics probably should be taken into consideration in the general case, however, since they establish a secondary limit on the predictive error. They are not considered further here, mainly because the goodness-of-fit statistics obtained from PLS analysis are inherently much more optimistic than is the external predictive error [6].

Although the COX2 data set is unusual in several ways, its distinctive attributes are precisely those that make it a particularly suitable subject for the analysis at hand. The descriptor and similarity measure used to assess diversity – Tanimoto similarity with respect to UNITY substructural fingerprints – differs from that used to derive the QSAR (CoMFA). This is partly a matter of operational convenience, but also reflects a deliberate effort to minimize the potential for incidental confounding interactions between the two parts of the analysis. The same descriptors could certainly be used for both, but maintaining a modest degree of independence is probably a good idea. It is difficult to imagine a QSAR problem in which fingerprint similarity would not adequately reflect structural diversity for purposes of cross-validation.

Finally, it should be noted that boosted LMO is likely to be very broadly applicable. It is nonparametric in that it does not require any assumptions about the distribution of error, descriptor independence *or* the form of the underlying function(s) involved, and so should prove useful for other regression methods as well. Since it is able to deliver more precise error estimates with smaller training sets than does random partition (about half as many observations are required to reach a given level of $\mathrm{SDEP_X}$ for $k = 4$ as for $k = 1$ in Figure 4), methods which scale badly with *N* may benefit substantially from this approach. In addition, the potential for sharply reducing the number of validation runs required to reach the same precision as LOO cross-validation (from 304 to 5 for this data set) without sacrificing model scope should make it appealing for computationally intensive applications. It may even find application in PCR and other QSAR techniques in which principal components analysis is used to 'de-correlate' descriptors for a particular data set, since the principal components so obtained are not always statistically independent enough to support accurate direct estimates of predictive error [42].

References

1. Snedecor, G.W and Cochran, W.G., Statistical Methods, Eighth Ed., Iowa State University Press, Iowa City, 1989.
2. Kauffman, G.W. and Jurs, P.C., J. Chem. Inf. Comput. Sci. 41 (2001) 1553.
3. Baumann, K., von Korff, M. and Albert, H., J. Chemometrics 16 (2002) 351.
4. Næs, T. and Martens, H., J. Chemometrics 2 (1988) 155.
5. Wold, S., Ruhe, A., Wold, H. and Dunn, W.J. III, SIAM J. Sci. Statist. Comput. 5 (1984) 735.
6. Wold, S., in: van de Waterbeemd (Ed.) Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 195-218.
7. Morsing, T. and Ekman, C., J. Chemometrics 12 (1998) 295.
8. Kleinknecht, R.E., J. Chemometrics 10 (1996) 687.
9. Denham, M.C., J. Chemoetrics 11 (1997) 39.
10. Faber, K. and Kowalski, B.R., J. Chemometrics 11 (1997) 181.
11. Agrafiotis, D.K., Cedeño, W. and Lobanov, V.S., J. Chem. Inf. Comput. Sci. 42 (2002) 903.
12. Wold, S. and Eriksson, L., in: van de Waterbeemd (Ed.) Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309-318.
13. Golbraikh, A. and Tropsha, A., J. Molec. Graphics Modell. 20 (2002) 269.
14. Wehrens, R. and van der Linden, W.E., J. Chemometrics 11 (1997) 157.
15. Wold, S., Johansson, E. and Cocchi, M., in: Kubinyi, H. (Ed.) 3D QSAR in Drug Design, ESCOM, Leiden, 1993, pp. 523-550.
16. 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, 2001, pp. 475-485.
17. Höskuldsson, A., J. Chemometrics 10 (1996) 637.
18. van de Waterbeemd, H., in: van de Waterbeemd, H. (Ed.) Structure-Property Correlations in Drug Research, .G. Landes, Austin, 1996, pp. 55-80.
19. Oprea, T.I., Waller, C.L. and Marshall, G.R., J. Med. Chem. 37 (1994) 2206.
20. Chavatte, P., Yous, S, Marot, C., Baurin, N. and Lesiur, D., J. Med. Chem. 44 (2001) 3223.
21. Matter, H., Defossa, E., Heinelt, U., Blohm, P.-M., Schneider, D., Müller, A., Herok, S., Schreuder, H., Liesum, A., Brachvogel, V., Lönze, P., Walser, A., Al-Obeidi, F. and Wildgoose, P., J. Med. hem. 45 (2002) 2749.
22. Golbraikh, A. and Tropsha, A., J. Comput.-Aided Molec. Design 16 (2002), 357.
23. Clark, R.D., J. Chem. Inf. Comput. Sci. 37 (1997) 1181.
24. Clark, R.D. and Langton, W.J., J. Chem. Inf. Comput. Sci. 38 (1987) 1079.
25. Cramer, R. D., Patterson, D. E., Bunce, J. D., J. Amer. Chem. Assoc., 110 (1998) 5959.
26. SYBYL and UNITY are available from Tripos, Inc., 1699 S. Hanley Rd., St. Louis MO 63144 USA.
27. CONCORD was developed by R.S. Pearlman, A. Rusinko, J.M. Skell and R. Balducci at the University of Texas, Austin TX and is available exclusively from Tripos, Inc., 1699 S. Hanley Road, St. Louis MO 63144 U.S.A.
28. Clark, R.D., Ferguson, A.M. and Cramer, R.D., in: Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.), 3D QSAR in Drug Design, Vol. 2: Ligand-Protein Interactions and Molecular Similarity, Kluwer/ESCOM, Dordrecht, 1998, 213-224.

29. Kurumbail, R.G., Stevens, A.M., Gierse, J.K., McDonald, J.J., Stegeman, R.A., Pak, J.Y., Gildehaus, D., Miyashiro, J.M., Penning, T.D., Seibert, K., Isakson, P.C. and Stallings, W.C., Nature 385 (1997) 555.
30. Gasteiger, J. and Marsili, M., Tetrahedron, 36, (1980) 3219.
31. The predicted class average activity for pyrrole analogs was 6.00 *vs* an observed class mean pIC50 of 6.60.
32. US Patent 6,535,819 (2003). OptiSim is available as an option in the Selector and HiVol modules of SYBYL and in ChemEnlighten. OptiSim, HiVol, Selector, SYBYL and ChemEnlighten are trademarks of Tripos, Inc., 1699 S. Hanley Road, St. Louis MO 63144 (http://www.tripos.com).
33. Wilett, P. and Winterman, V.A., Quant. Struct.-Activity Relat. 5 (1986) 18.
34. Cheng, C., Maggiora, G., Lajiness, M. and Johnson, M., J. Chem. Inf. Comput. Sci. 36 (1996) 909.
35. Clark, R.D., in: Ghose, A.K. and Viswanadhan, V.N. (Eds.) Combinatorial Library Design and Evaluation, Marcel Dekker, Inc., New York, 2001, pp. 337-362.
36. Holliday, J.D. and Willett, P., J. Biomolec. Screening, 1 (1996) 145.
37. Wold, S., Berglund, A. and Kettaneh, N., J. Chemometrics 16 (2002) 377.
38. Baumann, K., Albert, H. and von Korff, M., J. Chemometrics 16 (2002) 339.
39. Bauman, K., von Korff, M. and Albert, H., 16 (2002) 351.
40. Using this value for the degrees of freedom is not strictly correct, since the samples are not in fact independently drawn. Such an assumption is generally justified by noting that it is likely to hold true to a good approximation for most data sets and for most regression techniques.
41. Such an effect is not limited to linear models; *any* assumed functional model is more likely to break down for a more diverse training set.
42. Suppose, for example, X_1, X_2, X_1^2 and $X_1 \cdot X_2$ are being used as descriptors. Four orthogonal linear combinations can be derived for any given training set, but those ‘decorrelated’ variables will not be statistically independent when the broader descriptor space is sampled.