# Assessing Model Fit by Cross-Validation

Douglas M. Hawkins,*,† Subhash C. Basak,‡ and Denise Mills‡

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, and Natural Resources Research Institute, University of Minnesota−Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received October 25, 2002

When QSAR models are fitted, it is important to validate any fitted model—to check that it is plausible that its predictions will carry over to fresh data not used in the model fitting exercise. There are two standard ways of doing this—using a separate hold-out test sample and the computationally much more burdensome leave-one-out cross-validation in which the entire pool of available compounds is used both to fit the model and to assess its validity. We show by theoretical argument and empiric study of a large QSAR data set that when the available sample size is small—in the dozens or scores rather than the hundreds, holding a portion of it back for testing is wasteful, and that it is much better to use cross-validation, but ensure that this is done properly.

## INTRODUCTION

When fitting any sort of predictive model to data, it is essential to verify that the fitted model can be generalized to future data of the same type. This is particularly true with fitting complex models. Social scientists ran into the problem of ensuring model generalizability some 50 years ago (Mosier[1]) and adopted the idea of splitting the available cases into what we would now call a "calibration" sample and a "test" sample. In what (somewhat confusingly for us), they called "cross-validation", the available subjects would be randomly split into two subsamples with one being used to fit and the other to test. This parallels a widespread current practice in QSAR modeling. We will call this method "sampling splitting" in the discussion below.

In a distinct approach, termed "validity generalization", the test sample would be from a different population than the calibration sample. This problem is also present in QSAR settings, but the conceptual problem of, for example, fitting a model to one chemical class of compounds and applying it to a different chemical class restricts this practice to the brave or trusting, and we will not try to discuss it.

In much more recent times, the method currently known as cross-validation, or more accurately "leave-one-out cross-validation", was developed. In this, a single sample of size $n$ is used. Each member of the sample in turn is removed, the full modeling method is applied to the remaining $n-1$ members, and the fitted model is applied to the hold-back member. An early (1968) application of this approach to classification is that of Lachenbruch and Mickey.[2] Allen[3] was perhaps the first application in multiple regression and Geisser[4] sketches other applications.

It is important to note that the much older sample splitting and the much newer cross-validation methods arose not only at very different times but also in very different subject matter contexts. In the typical psychometric test, test subjects are generally inexpensive and freely available. There is no particular problem with recruiting 500 "Introduction to Psychology" students and splitting them into two groups of size 250 for calibration and test purposes. But in the QSAR setting, finding and testing 500 potential carcinogens is a completely different matter, and the sample size is typically far smaller.

If the sample is already quite small, getting no more from half of it than a check on the other half seems a waste of valuable information. For this reason (if for no other) where sample splitting is used in QSAR, the social science tendency to use equal-sized samples for calibration and validation is not standard. Rather, in QSAR work the test sample is typically much smaller than the calibration sample. Add to this that some QSAR methods fit the model using both a "calibration" and a separate "validation" sample, and a large hold-out test sample becomes even less common.

Just as a large test sample is often unattractive in 21st century QSAR work, cross-validation was unthinkable in the 1950s since it involved repeating the entire analysis a total of $n+1$ times—once using the full data set, and again with each subject in turn deleted. But what would then have been a huge computational burden is now an insignificant part of the whole project. Thus there is seldom a computational obstacle to using cross-validation these days, and its technical merits are paramount.

Each of the two methodologies assesses the final model using test compounds that were not involved in the model fitting. The sample splitting method does so clearly, visibly, and unequivocally. Cross-validation does so less visibly, since the entire sample is fed into the computer together and the separation is done internally. But if cross-validation is to achieve the desired goal of providing a check of model fit that is independent of the model fitting procedure, it is vital when each compound is held out for prediction, that it not be used in any way in the model fitting applied to the remaining retained $n-1$ compounds. Doing so often requires vigilance to ensure that the hold-out compound is not still somehow reflected in some of the model choices but is

* Corresponding author phone: (612)624-4166; e-mail: doug@stat.umn.edu.

† School of Statistics, University of Minnesota.

‡ Natural Resources Research Institute, University of Minnesota−Duluth.

10.1021/ci025626i CCC: $25.00 
Published on Web 01/24/2003

essential if the results are to be trusted.

## STATISTICAL ISSUES

Write a generic QSAR model as

$$y_i = g(x_i) + e_i$$

where the subscript $i$ indicates the $i$th compound in the set, $y$ is the activity or property being predicted, $x$ is the collection of descriptors being used in the modeling, $g(x)$ is the "true" relationship carrying the descriptor information into a prediction of the activity, and $e$ is the "unexplained" portion of the activity. By definition, $e$ is independent of $x$ and hence of $g(x)$ or else there would be some additional explanatory power not yet captured in the true relationship.

Going over to the whole population, write $\mu$ for the grand mean of the $y$. This gives the variance decomposition

$$E[y-\mu]^2 = E[\{y-g(x)\}+\{g(x)-\mu\}]^2 = E[y-g(x)]^2 + E[g(x)-\mu]^2$$

or (writing $V_y$ for the variance of $y$) $V_y = V_e + V_g$.

The first term on the right side, $V_e$, is the variance of the true "error" terms $e$. It is a direct measure of how accurately $y$ can be predicted using the function $g$ of the predictors $x$ under the ideal circumstances that $g$ is known exactly and that no parameters need to be estimated. It is clearly the irreducible minimum amount of prediction variance.

The second term $V_g$ is the variance of the "true mean" function $g(x)$ across the population. Unlike $V_e$, it is not fixed, but is driven by the user's decision of whether to model a homogeneous or a diverse set of compounds. A more homogeneous library will make it small; a more diverse will make it large.

The coefficient of determination $R^2$ is defined by

$$R^2 = 1 - \frac{V_e}{V_y} = \frac{V_g}{V_g + V_e} = \frac{f}{f+1}$$

where

$$f = \frac{V_g}{V_e}$$

is the ratio of these two variances.

We will not dwell on it here, but note that since $V_g$ is driven by the user's decision of whether to study a homogeneous set of compounds or to broaden the net and include more diverse compounds, so too is $R^2$. If you choose to study a very homogeneous set of compounds, $R^2$ will almost inevitably be small even if the predictions are close to the truth. If you choose a highly diverse set, you may have a high $R^2$ even when the predictions, as measured by the standard deviation of prediction, are poor. $R^2$, in other words, does not exist in any absolute, context-free sense and generally should not be used as the primary basis for deciding on the adequacy of a model. This broad truth is well established in the statistical literature.

Leaving this fundamental concern aside, how do we estimate the $R^2$ of a particular modeling approach in a particular population of compounds? The obvious way is to take a random sample of the compounds, apply the modeling procedure to this random sample and use it to make predictions. Writing $\hat{y}_i$ for a generic estimate of $y_i$, the activity of the $i$th compound in the sample, and writing $\bar{y}$ for the sample mean of the $y_i$ gives the estimate (sums are over the set of compounds used to assess the fit)

$$^{\square}R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

We can distinguish three broad situations within this framework:

- The same data $y$ are used both for fitting the model to get the estimates $\hat{y}_i$. This leads to the so-called "resubstitution estimate" of $R^2$.
- One set of data ("training data") is used for fitting the model, and an entirely different set of data ("test data") is used for the estimation of $R^2$.
- Leave-one-out cross-validation is used. In this, all available data are used for both fitting and assessing. However, when $\hat{y}_i$ is being estimated, no use whatever is made of $y_i$. The estimate of $R^2$ obtained by leave-one-out cross-validation is often denoted $q^2$.

Another measure of fit sometimes seen is the square of the correlation between the $\hat{y}_i$ and the $y_i$. Where it differs from the formula above, this measure is generally invalid as it presupposes the ability to do a further post-fitting linear adjustment of the predictions to future data which, by definition, are unavailable when the model is being fitted.

We can quickly dispose of the resubstitution estimate. It has been known for 50 years to be overoptimistic about the ability of the fitted model to generalize to future compounds. Furthermore, the more flexible the modeling method, the more overoptimistic it is. Bad though it is for a regular multiple regression then, it is even worse for more flexible methods such as subset regression, neural nets, and $k$-nearest neighbors.

Thus the choice is between the test sample and the cross-validation approaches. Write $t$ for the number of compounds used for testing and $c$ for the number of compounds used for the calibration and recall that $n$ is the total number of compounds available. Looking at the formula, the sample estimate of $R^2$ involves two quantities:

$$\text{TSS} = \sum_{i-1}^{t}(y_i - \bar{y})^2, \quad \text{the total sum of squares}$$

and

$$\text{PSS} = \sum_{i=1}^{t}(y_i - \hat{y}_{i(i)})^2, \quad \text{the prediction sum of squares}$$

The subscript $i(i)$ on the estimate is to stress that in either of the two proposed methods the estimate of $y_i$ is made without any use of $y_i$ itself.

Of these two sums, TSS is not controversial; it is a regular un-normalized full-sample variance and for any given calibration data set is a fixed constant. Looking at PSS though, we note that

$$\text{PSS} = \sum_{i=1}^{t}(y_i - \hat{y}_{i(i)})^2 = \sum_{i=1}^{t}[\{y_i - g(x_i)\} - \{\hat{y}_{i(i)} - g(x_i)\}]^2 = \sum_{i=1}^{t}[y_i - g(x_i)]^2 + \sum_{i=1}^{t}[\hat{y}_{i(i)} - g(x_i)]^2 - C$$

where the correction term $C$ is twice the sum of cross products between the differences $\{y_i - g(x_i)\}$ and $\{\hat{y}_{i(i)} - g(x_i)\}$, a term that will be small in large samples because the true errors $e_i$ are independent of the $x_i$.

The first sum of squares in this decomposition estimates $tV_e$ and corresponds to the minimum possible prediction variance. The second (necessarily positive) sum of squares shows the impact on PSS of having to estimate the parameters in the model.

Because of the addition of the second term, the prediction mean square, PSS/$t$ overestimates the variance of prediction of future compounds. Notice that this is equally true of a hold-out method and cross-validation.

The expression for PSS shows that the best results come from making $t$ as large as possible. This is because, even though PSS/$t$ is estimating the same quantity whatever $t$ may be, larger $t$ (as usual in statistics) reduces random variability and ensures that the PSS seen from the data is a reliable picture of the population. Thus you want the test data set as large as possible. As usual in statistics though, there is a diminishing return to scale, so if $t$ is already "large"—for example into the hundreds—there will be little benefit in making it even larger.

At the same time, for most estimation methods, the quality of the estimates $\hat{y}_{i(i)}$ improves if the size of the calibration sample is increased. This argues for making the $c$ as large as possible. Once again, there is a diminishing return to scale, so that if $c$ is already large we do not gain much by making it larger.

These two objectives of large $t$ and large $c$ conflict in the hold-out test sample approach where $c + t = n$ and any compound kept for testing is a compound removed from calibration and vice versa. If the available sample size $n$ is large—for example several hundred—then we can have a large $c$ and a large $t$ in the hold-out test method, but where sample information is scarce and expensive, this limitation hurts. By contrast cross-validation uses the entire sample for validation and also uses the entire sample less one compound for each estimation, and so has $t = n$ and at the same time (for the cross-validation portion) $c = n - 1$. It thus maximizes the size of both calibration and test sample.

This suggests that, except where the sample is so abundantly available that you can create separate large data sets for calibration and testing, cross-validation should provide a much more reliable picture than hold-out sample validation.

This is just an heuristic argument, but a large body of statistical literature confirms the effectiveness and reliability of cross-validation. See for example ref 5 for a detailed discussion of sample reuse, while ref 6 is a key reference particularly because of the accompanying discussion. In the context of multiple regression, cross-validation's good properties are proved in refs 7 and 8, while ref 9 proves that as the sample size increases, cross-validation gives an increasingly accurate picture of the prediction error when the fitted model is applied to future compounds.

Cross-validation has one blind spot. If one of the compounds essentially determines one of the model parameters, then it can probably not be predicted well from the other compounds, and so its squared prediction error is likely to be large. So if for example only one compound has many chlorine atoms and chlorine is an important predictor, then the full data set may be able to calibrate the effect of chlorine and make good predictions, but the data set with the compound excluded will not, likely leading to a large prediction error on that compound. Note the direction of this deficiency; it is in the direction of the model truly being *better* (not worse) than cross-validation suggests it is. See for example refs 10 and 11 for further discussion and examples.

There is one other warning: cross-validation does not sufficiently penalize overfitting. If a model fits the data, and nonsense predictors are added to it, the cross-validation PSS will not necessarily increase as one would hope (see refs 12 and 9). This however is equally true of validation using a hold-out sample, so it is no particular indictment of cross-validation as against use of sample splitting as a method of model assessment.

In summary, a considerable body of evidence shows that cross-validation gives a reliable picture of the true performance of the prediction on future data, with a tendency toward conservatism. This means that if cross-validation suggests that a model generalizes well, then this can be believed. Further, a model may generalize well even if its cross-validation statistics do not look promising because, for example, of important molecular features found in only a handful of the compounds.

## EMPIRIC EVALUATION

The body of evidence validating the use of cross-validation is extensive but makes assumptions about the validity of the model being fitted. It is useful therefore to verify the statistical theory with some actual empirical QSAR modeling.

We explored these issues using a relatively large QSAR data set. This data set, described in refs 13 and 14, deals with the prediction of vapor pressure of a structurally diverse set of compounds using molecular descriptors. The full data set contains 469 compounds and 379 topological, geometrical, and quantum chemical descriptors. The analyses reported in ref 14 showed that there is a strong predictive relationship, with coefficients of determination as high as 90%. We used this data set as a test bed for methodology of checking model fit.

Preliminary thinning by Basak and Mills—removing predictors that were perfectly correlated with other predictors, or that showed no compound to compound variation over the set of compounds—reduced the descriptor pool from 379 to 268. Further, a descriptor that is seen in only a few compounds may be relevant in predicting a property, but there is no way to verify its relevance, and so it may be wiser to leave it out of a modeling exercise. We therefore removed another 37 descriptors that were all-but-constant across the 469 compounds, leaving 232 descriptors for our experiment.

The statistical method we used for the QSAR's was ridge regression—RR—(refs 15 and 16). This method was chosen because of its attractive statistical underpinnings—proven optimality properties if the linear model is correct (ref 17) and sturdy good performance of linear methods even if the linear model is violated in a number of ways (ref 18). We do not, however, have any reason to think that the broad conclusions would be different using other modeling approaches such as PLS, PCR, and neural nets. To outline the RR methodology, write $\mathbf{Y}$ for the $n$-component vector of

observed activities of the compounds and **X** for the matrix of descriptors, having $n$ rows and $p$ columns, one column for each of $p$ descriptors. Descriptors and dependent variables are often log-transformed to attain linearity; if this has been done then **Y** and/or **X** contains log transformed values. All variables are "autoscaled"—that is, their mean is subtracted and the resulting difference divided by the standard deviation. Then ridge regression fits a linear model with coefficient vector

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $k$ is a constant called the ridge constant. If $k = 0$, then ridge regression reduces to ordinary least squares regression.

The ridge constant $k$ needs to be selected. There are two conventional methods of doing this: cross-validation and generalized cross-validation. In each, various values of $k$ are tried in a line search, and the value giving the best criterion is used. In cross-validation, the criterion is the prediction sum of squares PRESS. This is found for each particular $k$ value by calculating the coefficient vector $b$ using all but one of the compounds and using it to predict the held-out compound. PRESS is then the sum of squared deviations between the observed activities and the prediction of the hold-out compound.

Generalized cross-validation (GCV) (ref 11) relies on a mathematical argument to simulate the PRESS of predicting future samples from the calibration sample itself. It thus uses just the full-sample fit for a given $k$ and avoids the $n$ separate fits that result in CV from omitting each case in turn and refitting.

To do a properly cross-validated ridge regression by these two criteria then involves the following steps:

**Cross-Validation.** Take each compound in turn. Remove it from the data set, leaving $n-1$ compounds for calibration. Fit a ridge regression to these $n-1$ compounds. This leads to the embedded problem of picking the appropriate $k$ for those $n-1$ compounds, which you do by cross-validation: varying $k$ and calculating the PRESS for each $k$ excluding each of the $n-1$ compounds in turn and retaining $n-2$ compounds. The $k$ value so selected is used with the $n-1$ compounds to get a $b$ vector and predict the held-out compound. Cross-validation therefore involves fitting a total of $(n-1)(n-2+1) = (n-1)^2$ regressions for each $k$ value considered.

**GCV.** In GCV, the inner cross-validation is avoided. The $k$ values are picked using just the ridge fit for the trial $k$ values to each of the samples created by holding out one compound. Thus the total number of regressions that need to be looked at is $(n+1)$ times the number of $k$ values tested.

Because of this roughly $n$-fold speed advantage, we used GCV in the experiments. Note though that even using GCV for the selection of $k$, this approach is a proper cross-validation since the fit is assessed predicting each $y_i$ using calculations in which compound $i$ played no part.

**First Experiment.** The experiment made 300 analyses. In each, we took a random 100 compounds from the 469 available and a random set of descriptors from the 232 available. We then fitted the ridge regression using these descriptors to the 100 calibration compounds. The size of the descriptor set was set to 5, 10, 20, and 50 descriptors, with 75 repetitions at each size. The resulting 300 regression fits covered a spectrum of quality. Where the random pick chose poor predictors, the resulting models had little or no predictive ability, whereas some 50-descriptor fits gave coefficients of determination approaching 90%.

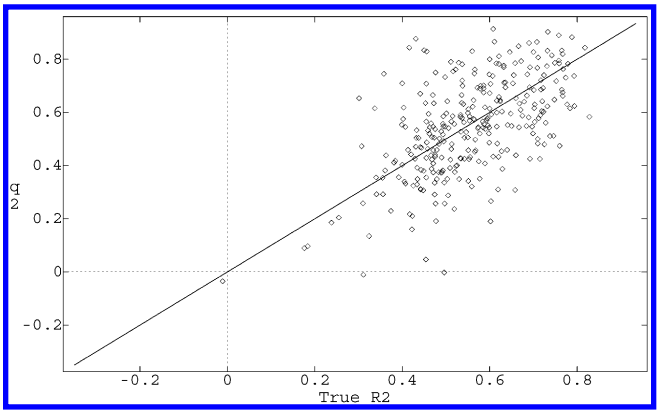


**Figure 1.** True R2 against q2.

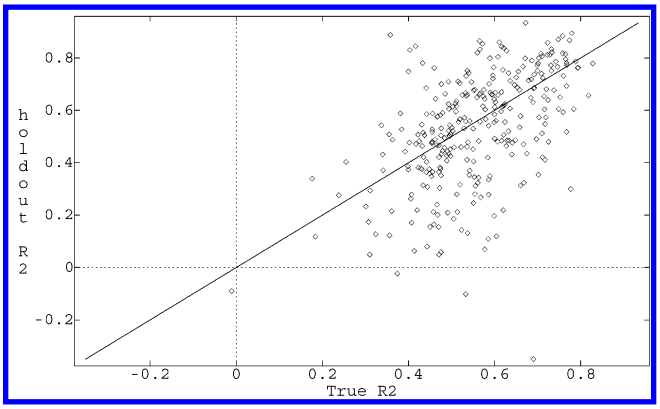


**Figure 2.** True R2 against 50-case holdout.

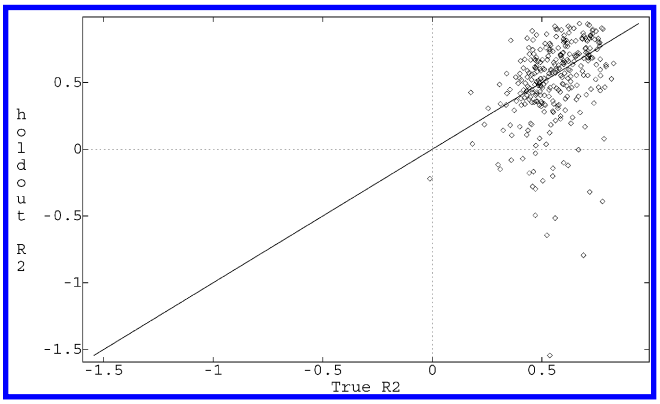


**Figure 3.** True R2 against 20-case holdout.

Of the 469 compounds, 100 were used for calibration, leaving 369 available for assessment. A random 50 of these were used to give "holdout" samples of size 10, 20, and 50, and the remaining 319 were used to calculate an $R^2$ which, in view of the much larger size of this set than the other sets, we regard as giving the "true" $R^2$.

Figures 1−4 show the relationship between the true $R^2$ and $q^2$ and between the true $R^2$ and the estimates given by the holdout test samples of size 50, 20, and 10. Each plot also shows the "$y = x$" line for reference.

(Figure 4 is truncated; four holdout estimates of $R^2$ between −3.3 and −1.5 are cropped.)

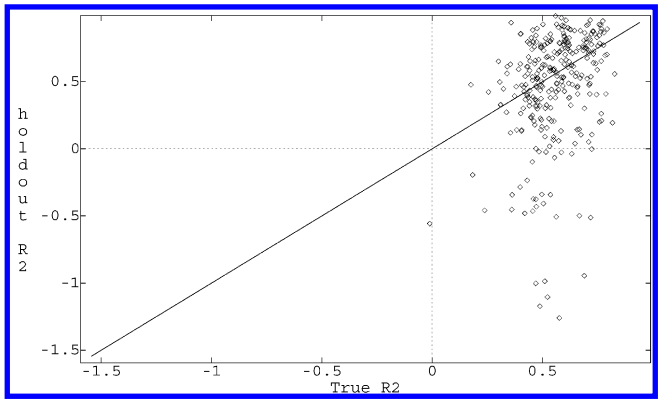


**Figure 4.** True R2 against 10-case holdout.

These figures make several clear points:

- The cross-validation $q^2$ seems to give a reliable picture of the true $R^2$. There is no apparent systematic over- or underestimation.
- The same can be said of the estimate of $R^2$ resulting from the 50-compound holdout. Visually though it does not seem quite as close as the $q^2$ estimate.
- The 20-compound holdout estimate is questionable. It shows much more variation from the true value than either the 50-compound holdout or $q^2$. It is clearly inferior to either.
- The 10-compound holdout is close to useless; there is enormous variability in the sample values to the point that a relationship with the $R^2$ is hard to see.

Going a stage further, it is helpful to look at the differences between the true $R^2$ and its sample estimates. The averages and standard deviations of the differences are as follows:

| | mean | standard deviation |
|---|---|---|
| true $R^2 - q^2$ | 0.010 | 0.149 |
| true $R^2$ − hold 50 | 0.028 | 0.184 |
| true $R^2$ − hold 20 | 0.055 | 0.305 |
| true $R^2$ − hold 10 | 0.123 | 0.504 |

These summary statistics confirm the visual impression of the plots; $q^2$ is the best of the estimates, having no perceptible bias and the smallest standard deviation. The estimate from the 50-compound holdout test sample is not as good, being also unbiased but quite a bit more variable. The 20-compound holdout sample is far inferior. It is not only substantially less precise but also seems biased. The 10-compound holdout is close to useless, as Figure 4 suggested. Some implications of this are immediate

1. If you have only 100 randomly selected compounds available, then you can use all 100 for calibration with the knowledge that the $q^2$ is accurate and precise.

2. If you have 120 compounds available, there is no point in splitting them into 100 for calibration and 20 for testing: the information in the test sample is far inferior to what you get from the calibration sample's $q^2$.

3. If you have more than 150 compounds available, you arrive at the possibility of using 100 for calibration and the remainder of more than 50 for testing. This test sample will give an estimate of $R^2$ fairly comparable with that given by $q^2$.

Taking the third of these points, if you have 150 or more compounds available, then you can certainly make a random split into 100 for calibration and 50 or more for testing. However it is hard to see why you would want to do this. The square root law of precision rules; using all 150 available compounds for calibration gives a 22% more precise estimate of the model coefficients than using 100 compounds, and the resulting $q^2$ will be even better than those illustrated by the experiment using 100 compounds.

The only motivation to rely on the holdout sample rather than cross-validation would be if there was reason to think the cross-validation not trustworthy—biased or highly variable. But neither theoretical results nor the empiric results sketched here give any reason to disbelieve the cross-validation results.

This conclusion is opposite to the recommendations of Golbraikh and Tropsha,[19] who argue that a holdout data set should always be used. However it is striking that they base their conclusion on holdout data sets as small as 10 compounds, at which, as Figure 4 clearly illustrates, the holdout estimate of $R^2$ is so imprecise that it has neither positive nor negative predictive value. We believe therefore that their advice is wrong.

**Second Experiment.** This raises the issue of how best one should use a modest-size pool of available compounds. Roughly matching sample sizes used in ref 19, we did a further experiment based on using a total available pool of 30 compounds. Either all 30 could be used for calibration with cross-validation for model checking, or 20 could be used for calibration and a hold-out sample of 10 used for assessment. We used a fixed set of nine descriptors that, in the full data, captured most of the descriptive power of the full set of descriptors and then made 300 random selections of the 30 compounds. With each selection, we either used all 30 for calibration or picked a random 20 for calibration and used the remaining 10 for checking. We used all remaining 439 compounds to get the true $R^2$ for the fitted model.

Some summary statistics from these 300 runs were as follows:

*Using all 30 for calibration and cross-validation*

| | first quartile | median | third quartile |
|---|---|---|---|
| true $R^2$ | 0.6130 | 0.7255 | 0.7868 |
| $q^2$ | 0.5116 | 0.7180 | 0.8433 |
| true $R^2 - q^2$ | −0.1370 | 0.0035 | 0.1422 |

*Using 20 to calibrate and a hold-out 10 to test*

| | first quartile | median | third quartile |
|---|---|---|---|
| true $R^2$ | 0.4447 | 0.6539 | 0.7369 |
| holdout $R^2$ | 0.1363 | 0.6426 | 0.8367 |
| true $R^2$-holdout | −0.1709 | −0.0267 | 0.2733 |

Noteworthy features of this experiment are as follows:

1. That the regressions fitted using 30 compounds are much better than those fitted using 20 compounds. This is shown conclusively by the true $R^2$ on the 439 compound population. The median 30-compound fit has an $R^2$ matching the upper quartile of the 20-compound fits. Thus the *average* model fitted using 30 compounds beats three-quarters of the models fitted using 20 compounds.

2. That the estimate of the true $R^2$ given by cross-validation of the 30-compound fits is much closer to the truth than is the estimate given by the holdout 10 compounds in the 20-compound fits. Its error of estimation has a central 50% range of −0.14 to +0.14; that of the 10-compound holdout was −0.17 to +0.27.

That you fit better models using 30 compounds than using 20 compounds should come as no surprise to anyone. That

the cross-validation estimate of the fit of the model is much closer to the truth than is that given by a hold-out sample of 10 may surprise some. In any event, the primary conclusion is inescapable: if the available sample is small, then you should use the whole sample for calibration and check the fit by cross-validation. This gives not only better fits but also a more reliable estimate of the fit than you get by splitting the compounds into a calibration and a test set.

**Third Experiment.** In the discussion so far, we covered only the situation in which the calibration and the test sample were independently and randomly sampled from the population of compounds of interest. In fact, this is not the best way to proceed. To have the widest range of validity and also to have the maximum precision for the estimates of fitted parameters, it is better to use a calibration sample that is chosen to be as diverse as possible. For example, the idea of "space-filling" designs (ref 20) looks for a calibration set of compounds each of which is as far as possible from its neighbors in descriptor space. A related but distinct idea is clustering a large set of compounds in descriptor space and then using one representative of each cluster in a calibration sample, as in for example ref 21. If regression models are fitted, then using "D-optimal" designs (ref 22) gives models whose precision is, in a certain technical sense, as high as possible given the set of available compounds and the total sample size available.

To investigate validating models in this framework, we repeated the sampling experiment but using a "smart" rather than a random set of compounds for the calibration. For this, we extracted a single common set of 100 compounds using the criterion of D-optimal design for a ridge regression with a modification of Galil and Kiefer's exchange algorithm, searching for the 100 compounds that maximized the determinant of the matrix $(\mathbf{X}^T\mathbf{X}+k\mathbf{I})$, with the ridge constant $k$ set to 0.1.

The resulting set of compounds has a distribution of descriptors appreciably different from a random pick. For example, in the data set, the descriptor *O* ranges from 0 to 1.792, but the distribution is left-skew, 40% of the readings have the value 1.389, and the mean and standard deviation are 1.23 and 0.40, respectively. The D-optimal design, seeking a better coverage of the scale, picks compounds whose *O* has a mean of 0.902 and a standard deviation of 0.56. The mean is substantially closer to the middle of the range. The standard deviation is 1.4 times that of a random sample. In fitting a regression, a large standard deviation of the predictor is good: it means that (all else being equal) the standard error of the slope of the regression on *O* will be smaller than that of a random sample of the same size by a factor of 1.4. To match this improvement in precision, a random sample would have to roughly double in size. As this illustrates, D-optimal design leads to a massive saving in sample size required for a given level of precision in the fitted model.

This choice of a "smart" rather than a random set of calibration compounds has implications on the cross-validated $q^2$ values. Because the design looks for compounds as widely spaced as possible, the compounds picked are more diverse than average. There tend to be many compounds out at the edge of descriptor space in some direction, and when cross-validating we will be verging on extrapolation much of the time. This brings us toward the "blind spot" of cross-validation, that its PSS overstates the error that will be incurred in predicting future compounds, and thereby understates the actual coefficient of determination.

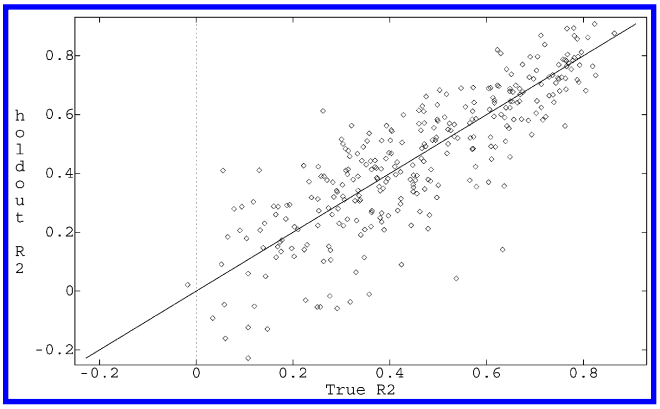


**Figure 5.** True R2 against 50-case holdout.

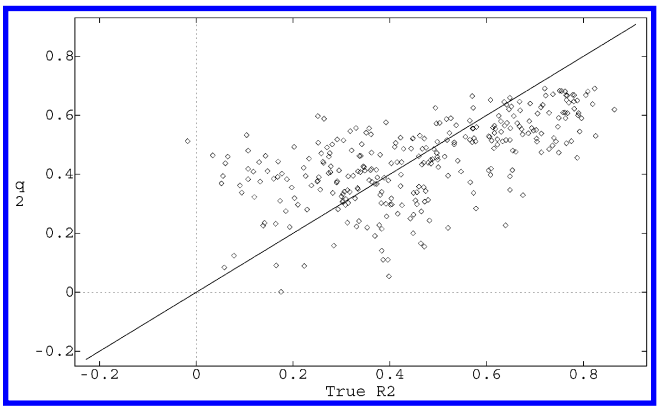


**Figure 6.** True R2 against q2.

Recall though the important fact that this deficiency is in the direction of conservatism; the sample $q^2$ will tend to underestimate the true $R^2$—a general feature that is particularly present in maximally diverse calibration samples.

To explore the impact of the choice of a smart calibration sample, we repeated the first experiment, again generating a total of 300 regressions using subsets of size 5, 10, 20, and 50 randomly chosen predictors. In this experiment though, instead of a random choice of 100 calibration samples, we used the fixed 100 compounds selected by the D-optimal design criterion. The remaining 369 compounds were used exactly as before to simulate hold-out test samples of size 10, 20, and 50, leaving 319 compounds to define for the true $R^2$. Figures 5 and 6 show the plot of the true $R^2$ against the 50-compound hold-out sample estimate and against $q^2$, respectively. The results for hold-out samples smaller than 50 parallel those for the random sample case in being substantially worse than the 50-compound results and so are not shown.

Figure 5 shows that, as one would expect, the hold-out sample's estimate of $R^2$ has no evidence of bias. The $q^2$ values are plotted in Figure 6. At the high end, they are considerably below the 45 degree line. This is exactly the phenomenon mentioned above that $q^2$ understates the true strength of the relationship. That good predictive models are even better than their $q^2$ suggests is hardly a cause for concern; modelers generally err on the side of caution, so having a relationship turn out to be stronger than advertised is seldom a bad thing.

There is perhaps more room for concern at the other end of the scale, where several relationships with low $q^2$ values have even lower $R^2$ values. The most extreme example is perhaps that with $q^2 = 0.51$, $R^2 = -0.02$. But this discrepancy turns out to be informative. For this particular set of predictors, the regression relationship is not linear, but curved, and underestimates the activity of the "silent majority" of compounds with more average chemistry than the extremist D-optimal design compounds. Since "average" compounds are underrepresented in the D-optimal design the curvature leads to underestimation of the average activity in the 319-compound population used for assessment. Some summary statistics paint the picture:

| | mean | SD |
|---|---|---|
| actual activity | 1.67 | 0.69 |
| prediction | 1.23 | 0.48 |
| difference | 0.44 | 0.53 |

The main reason for the small true $R^2$ of $-0.02$ is this miscentering. If we knew about the miscentering and knew to add 0.44 to all predictions from this model, we would have a true $R^2$ of 0.42, quite close to $q^2$, but the cross-validation does not give any warning of the offset except the indirect warning coming from the curvature in the model.

Detecting and correcting for this miscentering is a valid and possibly valuable use for an independent hold-out sample. Estimating this mean correction is a much easier problem, however, and needs a much smaller sample size than does attempting to estimate $R^2$ from a holdout sample. For example, a hold-out sample of size 10 is worthless for estimating $R^2$, but the size 10 holdout sample in this run had a mean residual of 0.37, enough to both demonstrate the nonzero mean error and estimate the correction needed to remove it.

## DISCUSSION AND CONCLUSION

Use of a test sample to check the fit of a model has a long tradition and has proved valuable in the social sciences where large test samples can frequently be had at little cost. In the QSAR setting, however, samples are more typically hard to get, expensive, or both, and the use of a large portion of the data merely for checking model fit seems a waste of valuable and often costly information. For this reason, hold-out test samples in QSAR work have tended to be small. We argue from theory and demonstrate by example that the multiple correlation as estimated from a modest-sized test sample is of little value in checking model fit.

The more recent approach is the use of cross-validation. This is computationally much more arduous than the use of a holdback test sample, since it involves carrying out the entire analysis as many times as one has sample. It appears to offer a free lunch, in that the entire sample is used in fitting the final model, but in addition, the entire sample is also used for model checking. This arouses the suspicion normally associated with offers of free lunch. Our trial using a population data set large enough to eliminate most uncertainty demonstrates that cross-validation does indeed provide a reliable picture of the fit of QSAR models calibrated using randomly selected samples. Hold-out samples of tolerable size, by contrast, do not match the cross-validation itself for reliability in assessing model fit and are hard to motivate.

There is a difference between calibration samples that are selected randomly and those selected more purposefully. Purposeful selection of compounds whose chemistry spans the population does not introduce any bias and is a good idea since it provides models that are both more precise and have a wider range of applicability than are given by random samples of the same size. Methods such as D-optimal designs, clustering in feature space, and space-filling designs are able to give substantial increases in efficiency over random choices. Here, there can be some value in using a small hold-out sample—not with the idea of using its computed $R^2$ to decide the fate of the model but for the more modest objective of checking whether there is evidence of a systematic misfit.

The bottom line is that in the typical QSAR setting where available sample sizes are modest, holding back compounds for model testing is ill-advised. This fragmentation of the sample harms the calibration and does not give a trustworthy assessment of fit anyway. It is better to use all data for the calibration step and check the fit by cross-validation, making sure that the cross-validation is carried out correctly.

Motivated by its well-understood properties and theoretical optimality, we used ridge regression as the fitting tool but believe that the broad conclusions hold for QSAR modeling methods in general.

## ACKNOWLEDGMENT

The research reported in this paper was supported in part by Grant F49620-01-0098 from the United States Air Force. We are grateful for the referees' suggested improvements. This is contribution number 329 from the Center for the Water and the Environment of the Natural Resources Research Institute.

## REFERENCES AND NOTES

(1) Mosier, C. I. Problems and designs of cross-validation. *Educ. Psychological Measurement* **1951**, *11*, 5−11.
(2) Lachenbruch, P. A.; Mickey, M. Estimation of error rates in discriminant analysis. *Technometrics* **1968**, *10*, 1−11.
(3) Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1968**, *16*, 125−127.
(4) Geisser, S. The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **1975**, *70*, 320−328.
(5) Efron, B. *The Jacknife, the Bootstrap and other Resampling Plans*; Society for Industrial and Applied Mathematics: 1982.
(6) Stone, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Royal Stat. Soc.* **1974**, *B36*, 111−147.
(7) Li, K.-C. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Annals Statistics* **1986**, *14*, 1101−1112.
(8) Droge, B. Asymptotic optimality of full cross-validation for selecting linear regression models. *Statistics Probability Lett.* **1999**, *44*, 351−357.
(9) Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486-494.
(10) Piepel, G. F.; Szychowski, J. M.; Loeppky J. L. Augmenting Scheffe linear mixture models with squared and/or crossproduct terms. *J. Quality Technol.* **2002,** *34*, *297*− 314.
(11) Golub, G. H.; Heath, M.; Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **1979**, *21*, 215−223.
(12) Stone, M. Asymptotics for and against cross-validation. *Biometrika* **1977,** *64*, 29−35.
(13) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical approach. *J. Chem. Inf. Comput. Sci*. **1997,** *37*, 651−655.

(14) Basak S. C.; Mills. D. Quantitative Structure−Property Relationships (QSPRs) for the estimation of vapor pressure: A hierarchical approach using mathematical structural descriptors. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 692−701.
(15) Hawkins, D. M.; Basak, S.; Shi, X. QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663−670.
(16) Hawkins, D. M.; Yin, X. A faster algorithm for ridge regression of reduced rank data. *Computational Statistics Data Anal.* **2002,** *40*, 253−262.
(17) Frank, I. E.; Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *3*5, 109−135.
(18) Li, K.-C.; Duan, N. Regression analysis under link violation. *Annals Statistics* **1989**, *17*, 1009−1052.
(19) Golbraikh, A.; Tropsha, A. Beware of $q^2$! *J. Mol. Graph. Modelling* **2002,** *20*, 269−276.
(20) Lam, R. L.; Welch, W. J.; Young, S. S. Uniform coverage designs for molecule selection. *Technometrics* **2002**, *44*, 99−109.
(21) Novellino, E.; Fattorusso, S.; Greco, G. Use of comparative molecular field analysis and cluster analysis in series design. *Pharm. Acta Helv.* **1995**, *70*, 149−154.
(22) Galil, Z.; Kiefer, J. Time- and space-saving computer methods, related to Mitchell's DETMAX, for finding D-optimum designs. *Technometrics* **1980,** *22*, 301−313.

CI025626I