Unsupervised Forward Selection: A Method for Eliminating Redundant Variables

D. C. Whitley,*,† M. G. Ford,† and D. J. Livingstone†,‡

Centre for Molecular Design, Institute of Biomedical and Biomolecular Science, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, PO1 2DY, U.K., and ChemQuest, Delamere House, 1, Royal Crescent, Sandown, Isle of Wight, PO36 8LZ, U.K.

Received March 10, 2000

An unsupervised learning method is proposed for variable selection and its performance assessed using three typical QSAR data sets. The aims of this procedure are to generate a subset of descriptors from any given data set in which the resultant variables are relevant, redundancy is eliminated, and multicollinearity is reduced. Continuum regression, an algorithm encompassing ordinary least squares regression, regression on principal components, and partial least squares regression, was used to construct models from the selected variables. The variable selection routine is shown to produce simple, robust, and easily interpreted models for the chosen data sets.

1. INTRODUCTION

Computational chemistry has considerable potential for drug design, but to assist in its rational use it would be helpful to support the molecular modeling studies with mathematical analyses of the relationships between the responses detected by bioassays and appropriate sets of molecular properties derived using computational methods. In the past this has proved difficult due to the amounts of redundancy and multicollinearity contained in typical data sets. This paper offers a procedure to overcome this and, as part of a structured approach to model building, produce statistical models with good predictive power based on a small number of relevant properties.

1.1. Relevance, Redundancy, and Multicollinearity. The ability of molecular modeling packages to generate large numbers of molecular descriptors and the development of 3-D QSAR procedures such as CoMFA and EVA has led to the frequent occurrence of data matrices with many more columns (descriptors) than rows (objects). This has resulted in a much wider choice of variables for possible inclusion in statistical models but has greatly increased the possibility of chance correlation[1] with data describing biological activity. Three main issues arise when developing predictive models for use in the design of new compounds or when investigating the relationship between biological data sets and chemical descriptors: relevance, redundancy, and multicollinearity. Relevance means simply that the variables included in the model should contain information pertinent to the response being modeled. Relevant descriptors have a statistically significant correlation with the response variable and do not have low variance; as variance tends to zero, so does the information content of a variable. Redundancy is an exact linear dependence between a subset of the columns in the data matrix, so that at least one column in this subset contributes no unique information. Redundancy implies that the data matrix has maximal rank. Multicollinearity is the existence of high multiple correlation between a subset of linearly independent columns which, nevertheless, still contribute some unique information. Multicollinearity implies that the data matrix has at least one small, nonzero eigenvalue. Redundancy occurs in the limiting case when this eigenvalue tends to zero, so that redundancy is the ultimate multicollinearity. Whereas redundancy is always to be avoided, multicollinearity that reflects the properties of the population rather than the sample can be an important feature of a successful model, for example reflecting characteristic features of a series of related chemical structures.

1.2. Preprocessing Data. Such considerations lead to the following preprocessing strategy for the derivation of models for use in structure−activity relationships (QSARs): 1. identify a subset of columns (variables) with significant correlation to the response; 2. remove columns (variables) with small variance; 3. remove columns (variables) with no unique information; 4. identify a subset of variables on which to construct a model; and 5. address the problem of chance correlation. Attention to these points will result in parsimonious QSAR models that are more likely to generalize successfully to new objects.

The increasing application of multivariate techniques[2,3] to the development of models for drug design has led to the widespread use of "over-square" data sets and thus, in the interests of minimizing chance correlation and improving the quality of the sets, various approaches have been proposed for data preprocessing. The identification and removal of variables with low or zero variance is a commonly used method. Indeed, this is almost a prerequisite in the analysis phase of 3-D QSAR studies using CoMFA or EVA which invariably result in over-square data sets with many variables of low or, particularly in the case of EVA, zero variance. Although useful, this approach normally removes only a small number of variables from typical data sets generated by molecular modeling packages and does nothing to address problems such as redundancy and multicollinearity.

A technique that does set out to remove redundancy, on the basis of pairwise correlation, is known as CORCHOP.[4] This procedure is unsupervised, in the sense that it depends only on the independent variables, and the response variable plays no role in the selection process. CORCHOP identifies

* Corresponding author phone: (023)9284 5080; e-mail: david.whitley@port.ac.uk.
† University of Portsmouth.
‡ ChemQuest.

10.1021/ci000384c CCC: $19.00
Published on Web 09/08/2000

variables whose correlation with one other variable is greater than some pre-set limit and suggests an appropriate member of the pair to remove. After the identification of such sets of variables the algorithm then identifies variables on the basis of the count of their pairwise correlation with others. The intention here is to remove the smallest number of variables while breaking the largest number of pairwise collinearities. Others have proposed similar procedures,[5−7] and neural network pruning[8,9] allows a nonlinear estimate of variable importance as recently reviewed.[10] However, for generalization, QSAR equations should be of low dimension with as few variables as possible.

1.3. Aims and Objectives. This paper presents an unsupervised forward selection (UFS) routine that reduces oversquare matrices to a size for which specification of robust models is possible. This algorithm was designed specifically to meet items 2 and 3 above but also deals partially with items 4 and 5. Rather than starting with all the variables and removing correlated columns in the manner of CORCHOP, UFS starts with the two variables which are least well correlated and selects additional variables on the basis of their multiple correlation with those already chosen, thus building a subset of variables that is as close to orthogonality as possible. Three examples are presented to demonstrate the utility of UFS as part of a QSAR model building procedure designed to address all the issues listed above.

2. THE UFS ALGORITHM

This section describes the unsupervised forward selection algorithm, applied to an $n \times p$ matrix $\boldsymbol{X} = (x_{ij})$, where x_{ij} is the value of the jth variable for the ith compound. Let $X_j = (x_{ij}, ..., x_{nj})^T$ denote the jth column of $\boldsymbol{X}$. The selection process halts when the R^2 value of each remaining variable with those already selected exceeds some preassigned limit $R^2_{\max} < 1$.

1. Mean-center the columns of $\boldsymbol{X}$

$$x_{ij} \rightarrow x_{ij} - \bar{x}_j$$

where

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

2. Reject columns with length

$$|X_j| = \sqrt{\sum_{i=1}^{n} x_{ij}^2} < \epsilon$$

for some small $\epsilon > 0$. These columns have small standard deviation and contribute no significant information.

3. Normalize the remaining columns to unit length:

$$x_{ij} \rightarrow x_{ij}/|X_j|$$

4. Calculate the correlation matrix $(r_{ij}) = \boldsymbol{X}^T\boldsymbol{X}$. Select as the first two columns those with the smallest squared correlation coefficient r_{ij}^2 and reject columns whose squared correlation coefficient with either exceeds $R^2_{\max}$.

5. Choose an orthonormal basis $\{c_1, c_2\}$ for the subspace of R^p spanned by the first two columns. (For example, follow the Gram-Schmidt procedure: if the first two columns are X_α and X_β, take $c_1 = X_\alpha$ and $c_2 = Y/|Y|$, where $Y = X_\alpha - (X_\beta \cdot X_\alpha)X_\alpha$.)

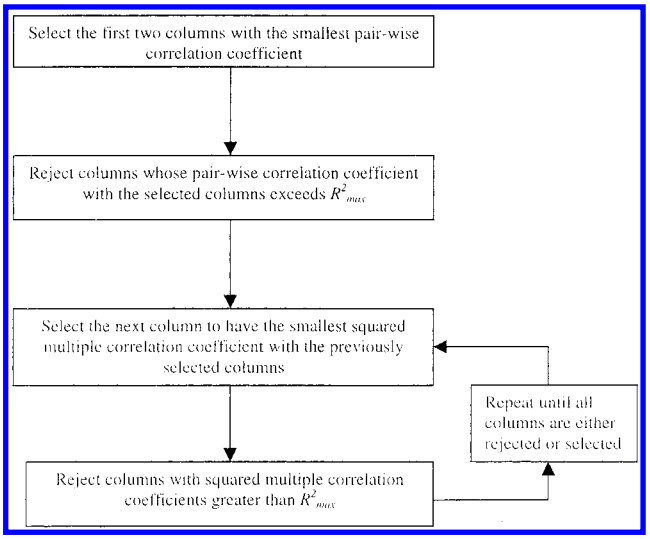

Figure 1. Unsupervised forward selection.

The remaining steps of the algorithm are repeated until each column is either selected or rejected. Suppose that $l \geq 2$ columns have been selected, and let $\{c_1, ..., c_l\}$ be an orthonormal basis for the subspace of R^P spanned by these.

6. For each remaining column X_j calculate its squared multiple correlation coefficient R_j^2 with the selected columns. This is the length of the orthogonal projection of X_j onto the subspace spanned by $\{c_1, ..., c_l\}$:

$$R_j^2 = |\sum_{k=1}^{l} (X_j \cdot c_k)c_k|$$

7. Reject columns X_j with $R_j^2 > R^2_{\max}$ and select from those remaining the column with the smallest R_j^2.

8. If any columns remain, choose an orthonormal basis for the subspace spanned by the selected columns and return to step 6. (For example, use $c_1, ..., c_l$ and $c_{l+1} = Y/|Y|$ where

$$Y = X_\omega - \sum_{k=1}^{l} (X_\omega \cdot c_k)c_k$$

and X_ω is the column selected at step 7.)

Figure 1 provides a flowchart for the major steps in this process. Source code implementing this algorithm is available from http://www.cmd.port.ac.uk. The algorithm as presented gives the user no control over the variables to be selected. Clearly it could be implemented as an interactive process, with the R^2 values for the unselected variables presented to the user at each stage, and allowing the user to over-ride the automatic choice of the variable with the smallest R^2, forcing the entry of favored variables (log P, etc.) into the data set.

3. APPLICATIONS OF THE UFS ALGORITHM TO DRUG DESIGN

To illustrate the use of the algorithm, we describe three applications to QSAR model building: a CoMFA data set used to model the relationship between a series of 21 steroid compounds and their testosterone binding globulin affinity;[11] a data set containing 70 descriptors used to model the biological activity of 19 pyrethroid insecticides;[12] and a data

set containing 53 descriptors used to model the biological activity of 31 antifilarial antimycin analogues.[13] The data sets used are available on the Centre for Molecular Design web site (http://www.cmd.port.ac.uk). The pyrethroid and Selwood data were mean-centered and normalized to unit length prior to analysis.

3.1. Model Specification Protocol. The modeling procedure adopted in each case was designed to address the issues of relevance, redundancy, and multicollinearity identified in the Introduction (section 1.1). First, variables whose correlation with the response variable was not significant at the 5% level were removed. Second, variables with small variance were removed. The UFS procedure was then applied repeatedly using values of R^2_{max} stepping from 0.1 to 0.9 with an increment of 0.1, together with $R^2_{max} = 0.99$. In each case models were built from the subset of variables identified by UFS using the Portsmouth formulation of Continuum Regression (CR),[14] a procedure in which the model selection criterion depends on a continuous parameter α in the range $0 \le \alpha \le 1.5$. CR is equivalent to Ordinary Least Squares (OLS) when $\alpha = 0$, Partial Least Squares (PLS) when $\alpha = 0.5$, and Principal Components Regression (PCR) when $\alpha = 1$. The CR calculations were performed with the in-house PARAGON[15] software using values of α stepping from 0 to 1.5 with an increment of 0.1. To address the issue of chance correlation, an optimal model was chosen to have values of R^2_{max} and α maximizing Q^2, the leave-one-out cross-validated R^2. At this stage the correlations between the residuals and the variables removed on the grounds of having an insignificant correlation with the response variable may be examined. Any variable found to have a significant correlation with the residuals may be added to the set of variables used to specify the model and CR repeated as above. This protects against committing a Type I error during the model specification procedure. As a final check against chance correlation, the optimal models were analyzed using (i) *n*-fold cross-validation for a range of values of *n*, where *n* is the number of cross-validation groups; and (ii) a randomization test that involved 1000 permutations of the y scores.

3.2. Steroid Data Set. The first application of UFS is to the data set of 21 steroid compounds used in the SYBYL CoMFA tutorial[11,16] to model their binding affinity to human testosterone binding globulin (TBG). A CoMFA column was calculated in SYBYL using the parameters recommended in the tutorial example, and the steric and electrostatic field values at each lattice point were extracted. After removing those lattice points whose field values exceeded the recommended 30 kcal/mol cutoff, this resulted in a dataset with 1248 columns. From this set, 858 columns not significantly correlated with the response variable TBG at the 5% level were removed, leaving a set of 390 columns. A further 367 columns with variance below 1.0 kcal/mol were removed as recommended,[16] leaving 23 columns. UFS and CR were then applied with the range of R^2_{max} and α values described above. For each value of R^2_{max} the value of α giving the largest Q^2 is shown in Table 1, along with the number of variables selected by UFS, the number of components in the CR model, and the fit R^2.

The number of variables selected by UFS always increases in a stepwise fashion with the value of R^2_{max}, used as a stopping criterion, rising in this instance to a maximum of 11 at $R^2_{max} = 0.99$. This reflects the increasing degree of multicollinearity observed as the number of selected variables increases. Table 1 shows that the number of selected variables is constant for $0.1 \le R^2_{max} \le 0.2$ and $0.3 \le R^2_{max} \le 0.5$. Over each of these ranges the same set of variables is provided to CR, and so identical models are generated. For this data set the number of components in the models found by CR tends to rise with R^2_{max}. The overall optimal model is a 3-variable, 1-component model with $Q^2 = 0.83$ and $R^2 = 0.85$ that is found over a range of values of R^2_{max}. The value of $\alpha = 0.3$ is determined from the plot of Q^2 versus α and R^2_{max} shown in Figure 2.

This produces the optimized QSAR model ($R^2_{max} = 0.3$, $\alpha = 0.3$):

$$TBG = 0.83\ C1 \quad (\pm 0.08)$$

Table 1. Optimal Models for the Steroid Data Set

R^2_{max}	α	variables	components	Q^2	R^2
0.1	1.0	2	1	0.7528	0.7617
0.2	1.0	2	1	0.7528	0.7617
0.3	0.3	3	1	0.8277	0.8535
0.4	0.3	3	1	0.8277	0.8535
0.5	0.3	3	1	0.8277	0.8535
0.6	0.0	4	1	0.7915	0.8949
0.7	0.3	5	2	0.7853	0.8891
0.8	0.3	6	2	0.7576	0.8802
0.9	1.0	7	4	0.8200	0.8911
0.99	0.7	11	3	0.8090	0.8860

1-Component Model ($R^2_{max} = \alpha = 0.3$)

Analysis of Variance

	DF	SS	MS	F-ratio	Prob>F
model	1	17.0700	17.0700	110.6943	0.0000
error	19	2.9300	0.1542		
total	20	20.0000			
S	0.3927				
R^2	0.8535				
adjusted R^2	0.8535				

Bootstrap 95.0% Confidence Limits (Based on 5000 Bootstraps)

	lower	median	upper	mean	stderr
E225	−0.6310	−0.4346	−0.2199	−0.4361	0.0985
E249	0.2955	0.5176	0.7076	0.5138	0.1075
S465	−0.6557	−0.4252	−0.2511	−0.4332	0.1010

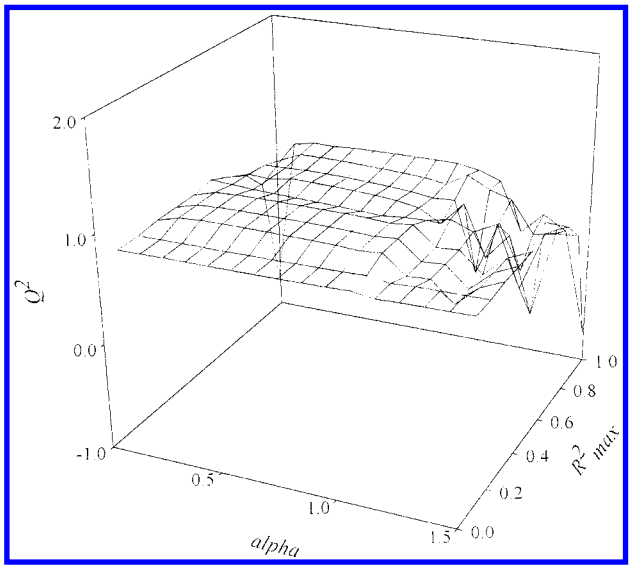

Figure 2. Maximizing Q^2 for the steroid data set.

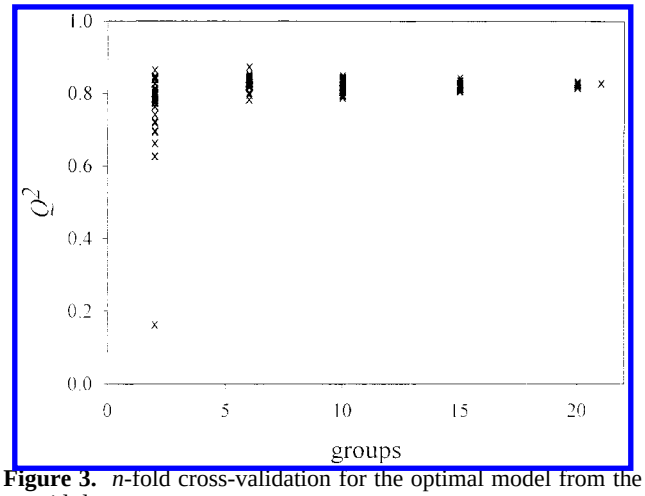

Figure 3. *n*-fold cross-validation for the optimal model from the steroid data set.

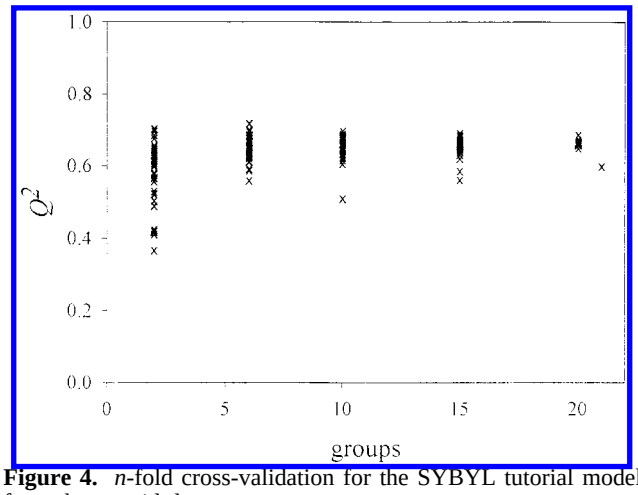

Figure 4. *n*-fold cross-validation for the SYBYL tutorial model from the steroid data set.

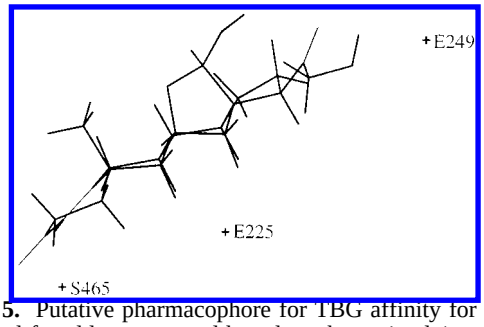

Figure 5. Putative pharmacophore for TBG affinity for steroids, illustrated for aldosterone and based on the optimal 1-component model, eq 3.

$$t = 10.5211,\ prob{>}t = 0.0000,\ R^2 = 0.85,\ Q^2 = 0.83 \quad (1)$$

where

$$C1 = -0.53\ E225 + 0.70\ E249 - 0.48\ S465$$

so that

$$TBG = \underset{(\pm 0.10)}{-0.44\ E225} + \underset{(\pm 0.11)}{0.58\ E249} - \underset{(\pm 0.10)}{0.39\ S465} \quad (2)$$

Here *E225* = electrostatic field at $x = -1.77$, $y = 0.19$, $z = -5.09$; *E249* = electrostatic field at $x = 6.23$, $y = 4.19$, $z = -5.09$; and *S465* = steric field at $x = -1.77$, $y = -5.81$, $z = 0.91$ are standardized variables ($\mu = 0$, $\sigma = 1$), and the standard errors in eq 2 are estimated by bootstrapping[17] using a sample size of 5000 (Table 1). It is worth noting that pairs of field points share common coordinates ($x = -1.77$ for *E225* and *S465*, and $z = -5.09$ for *E225* and *E249*). In terms of the original unstandardized variables eq 2 becomes

$$TBG = \underset{(\pm 0.08)}{8.96 - 0.35\ E225} + \underset{(\pm 0.07)}{0.35\ E249} - \underset{(\pm 0.04)}{0.15\ S465} \quad (3)$$

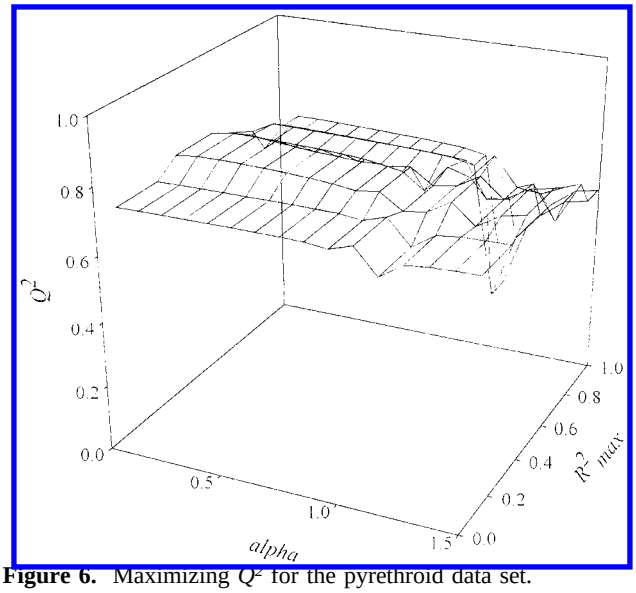

Figure 6. Maximizing Q^2 for the pyrethroid data set.

As a final check that chance correlation has been avoided and that the model is likely to generalize to new objects, *n*-fold cross-validation was carried out for $n = 2$, 6, 10, 15 and 20, followed by permutation of the y scores. The results of 40 cross-validations for each value of *n* are shown in Figure 3. Apart from a single low value of Q^2 when only 2 cross-validation groups were used, all the cross-validations produced values of $Q^2 > 0.6$, and when $n > 2$, $Q^2 > 0.75$. A randomization test for this model, using 1000 permutations of the response variable, produced tail probabilities less than 0.0001 for fit and 0.0012 for prediction.

By comparison, the SYBYL tutorial produces a 5 component model with $Q^2 = 0.6$ and $R^2 > 0.98$, an example of overfitting. The results of *n*-fold cross-validation of this model, shown in Figure 4, are generally weaker than those for the UFS/CR model in Figure 3, with a wider spread of lower values of Q^2 for each group size *n*.

Figure 5 shows the variables in eq 2 plotted with a representative structure (aldosterone). All three variables lie on the same side of the structure, and their common coordinates lead to them being approximately equidistant from the central plane of the structure. This illustrates how the modeling procedure followed here may lead to potential pharmacophores.

3.3. Pyrethroid Data Set. The second example is a data set consisting of 70 physicochemical descriptors used to model the killing activity (KA) of 19 pyrethroid insecticides.[12] Only 6 of these descriptors are significantly correlated with KA at the 5% level. In this case no variables were removed on the grounds of small variance. The results of the UFS/CR procedure are shown in Table 2. The optimal model, as chosen by maximal Q^2, is a 4-variable, 2-component model with $R^2 = 0.775$ and $Q^2 = 0.773$ obtained when $R^2_{max} = 0.7$ and $\alpha = 1.2$ (Figure 6):

Table 2. Optimal Models for the Pyrethroid Data Set

R^2_{max}	α	variables	components	Q^2	R^2
0.1	1.0	2	1	0.7196	0.7241
0.2	1.0	2	1	0.7196	0.7241
0.3	1.0	2	1	0.7196	0.7241
0.4	0.2	3	1	0.7563	0.8051
0.5	0.2	3	1	0.7563	0.8051
0.6	0.2	3	1	0.7563	0.8051
0.7	0.1	4	1	0.7172	0.7980
0.7	1.2	4	2	0.7733	0.7746
0.9	0.1	6	1	0.6823	0.7730
0.99	0.1	6	1	0.6823	0.7730

1-Component Model ($R^2_{max} = 0.4$, $\alpha = 0.2$)
Analysis of Variance

	DF	SS	MS	F-ratio	Prob>F
model	1	14.4913	14.4913	70.2119	0.0000
error	17	3.5087	0.2069		
total	18	18.0000			
S	0.4543				
R^2	0.8051				
adjusted R^2	0.8051				

Bootstrap 95.0% Confidence Limits
(Based on 5000 Bootstraps)

	lower	median	upper	mean	stderr
A5	0.3715	0.6974	0.9215	0.6860	0.1440
MIZ	0.0406	0.2914	0.4888	0.2846	0.1124
DVX	−0.5541	−0.3303	−0.0290	−0.3192	0.1332

2-Component Model ($R^2_{max} = 0.7$, $\alpha = 1.2$)
Analysis of Variance

	DF	SS	MS	F-ratio	Prob>F
model	2	13.9421	6.9710	27.4862	0.0000
error	16	4.0579	0.2563		
total	18	18.0000			
S	0.5063				
R^2	0.7746				
adjusted R^2	0.7613				

Bootstrap 95.0% Confidence Limits
(Based on 5000 Bootstraps)

	lower	median	upper	mean	stderr
A5	0.2531	0.7047	1.0101	0.6847	0.2018
A8	−0.2358	0.1232	0.5582	0.1521	0.2217
MIZ	−0.0673	0.2937	0.5876	0.2805	0.1684
DVX	−0.4967	−0.1321	0.2295	−0.1332	0.1919

$$KA = \underset{(\pm 0.15)}{-1.00\ C1} + \underset{(\pm 0.17)}{0.51\ C2} \quad (4)$$

where

$$C1 = -0.15\ A5 + 0.81\ A8 + 0.26\ MIZ + 0.50\ DVX$$

$$(n = 19,\ t = -6.7349,\ prob{>}t = 0.0000)$$

$$C2 = -0.7709\ A5 + 0.4162\ A8 - 0.3619\ MIZ + 0.3186\ DVX$$

$$(n = 19,\ t = 3.1006,\ prob{>}t = 0.0069)$$

so that

$$KA = \underset{(\pm 0.20)}{0.6951\ A5} + \underset{(\pm 0.22)}{0.0001\ A8} + \underset{(\pm 0.16)}{0.4941\ MIZ} - \underset{(\pm 0.19)}{0.0620\ DVX} \quad (5)$$

Here *A5*, *A8* = atomic charges, *MIZ* = z-component of moment of inertia, and *DVX* = x-component of dipole vector are standardized variables, and the standard errors in eq 5 are bootstrap estimates (Table 2). In terms of the original unstandardized variables eq 5 produces

$$KA = \underset{(\pm 2.80)}{-2.31} + \underset{(\pm 98)}{9.64\ A5} + \underset{(\pm 8.29\mathrm{E}-5)}{0.044\ A8} + 2.4\mathrm{E}{-4}\ MIZ - \underset{(\pm 0.11)}{0.037\ DVX} \quad (6)$$

A randomization test for this model using 1000 permutations of the response variable produced tail probabilities less than 0.0003 for fit and 0.0071 for prediction.

A very similar 3-variable, 1-component model occurs over the range $0.4 \le R^2_{max} \le 0.6$ with $\alpha = 0.2$:

$$KA = \underset{(\pm 0.10)}{0.81\ C1}$$

$$t = 8.3793,\ prob{>}t = 0.0000,\ R^2 = 0.81,\ Q^2 = 0.76 \quad (7)$$

where

$$C1 = 0.77\ A5 + 0.49\ MIZ - 0.41\ DVX$$

Thus, in terms of standardized variables

$$KA = \underset{(\pm 0.14)}{0.63\ A5} + \underset{(\pm 0.11)}{0.39\ MIZ} - \underset{(\pm 0.13)}{0.33\ DVX} \quad (8)$$

where the standard errors are bootstrap estimates (Table 2). In terms of the original unstandardized variables eq 8 becomes

$$KA = \underset{(\pm 2.00)}{-1.80} + \underset{(\pm 5.5\mathrm{E}-5)}{8.67\ A5} + 1.94\mathrm{E}{-4}\ MIZ - \underset{(\pm 0.08)}{0.20\ DVX} \quad (9)$$

A randomization test for this model using 1000 permutations of the response variable produced tail probabilities less than 0.0001 for fit and 0.0052 for prediction.

For both these models, *n*-fold cross-validation was carried out for $n = 2$, 6, 10, 15 and 18, with the results shown in Figures 7 and 8. The *n*-fold cross-validation indicates that the 1-component model is to be preferred over the 2-component model on grounds of parsimony, even though the latter has a marginally better value of Q^2. Equation 9 relates high insecticidal activity of pyrethroids to a low or negligible dipole component, a partial positive charge at the meta-position of the benzyl ring of the alcohol moiety, and a large moment of inertia in the *z*-direction. These features are consistent with a mode of action in which the order within a lipid bilayer or biological membrane is disrupted.

3.4. Selwood Data Set. The final example is the data set studied by Selwood et al.[13] in modeling the biological activity of 31 antifilarial antimycin analogues. Of the 53 descriptors in the data set only 12 are significantly correlated with the response at the 5% level. None of these variables was considered to have small variance, so all 12 were used in the UFS/CR modeling phase. The results, shown in Table 3 and Figure 9, are exceedingly poor. The best model, with $R^2 = 0.42$ and $Q^2 = 0.41$, was obtained when $R^2_{max} = 0.1$ and $\alpha = 1.0$:

$$\log \frac{1}{C} = \underset{(\pm 0.14)}{0.64\ C1}$$

$$t = 4.5683,\ prob{>}t = 0.0001,\ R^2 = 0.42,\ Q^2 = 0.41 \quad (10)$$

where

$$C1 = 0.71\ SUM\_F + 0.71\ MOFI\_Z$$

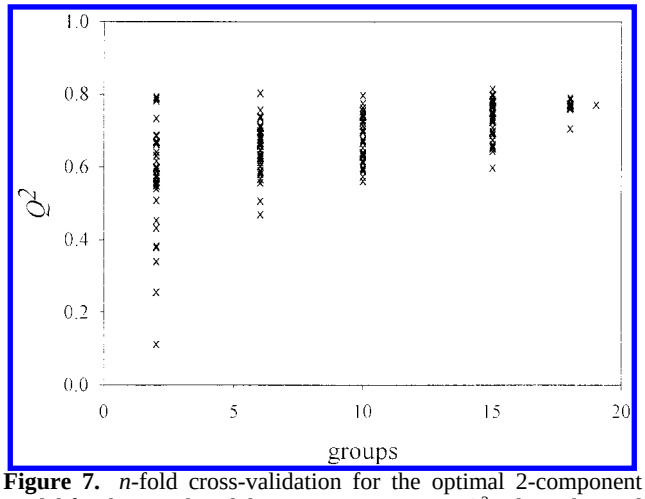

Figure 7. *n*-fold cross-validation for the optimal 2-component model for the pyrethroid data set. Four negative Q^2 values obtained when $n = 2$ are not shown.

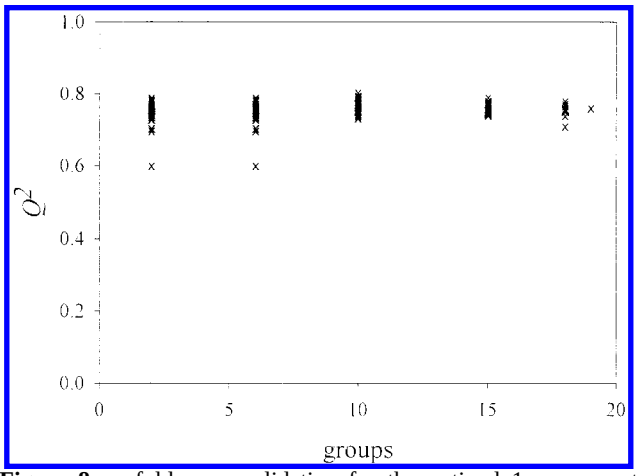

Figure 8. *n*-fold cross-validation for the optimal 1-component model for the pyrethroid data set.

Table 3. Optimal Models for the Selwood Data Set

R^2_{max}	α	variables	components	Q^2	R^2
0.1	1.0	2	1	0.4086	0.4240
0.2	1.0	2	1	0.4086	0.4240
0.3	1.2	3	1	0.1423	0.4435
0.4	1.2	3	1	0.1432	0.4435
0.5	1.0	4	1	0.0192	0.4150
0.6	1.0	4	1	0.0192	0.4150
0.7	1.0	5	1	0.1908	0.4478
0.8	1.0	7	1	0.3214	0.4776
0.9	1.0	8	1	0.3413	0.4743
0.99	0.8	12	1	0.3718	0.4475

1-Component Model ($R^2_{max} = 0.1$, $\alpha = 1.0$)

Analysis of Variance

	DF	SS	MS	F-ratio	Prob>F
model	1	12.3842	12.3842	20.8692	0.0001
error	28	16.6158	0.5934		
total	29	29.0000			
S	0.7703				
R^2	0.4270				
adjusted R^2	0.4270				

Bootstrap 95.0% Confidence Limits (Based on 5000 Bootstraps)

	lower	median	upper	mean	stderr
SUM_F	0.3287	0.4671	0.5945	0.4643	0.0667
MOFI_Z	0.3205	0.4670	0.5945	0.4593	0.0952

Thus

$$\log \frac{1}{C} = \underset{(\pm 0.07)}{0.45\ SUM\_F} + \underset{(\pm 0.09)}{0.45\ MOFI\_Z} \qquad (11)$$

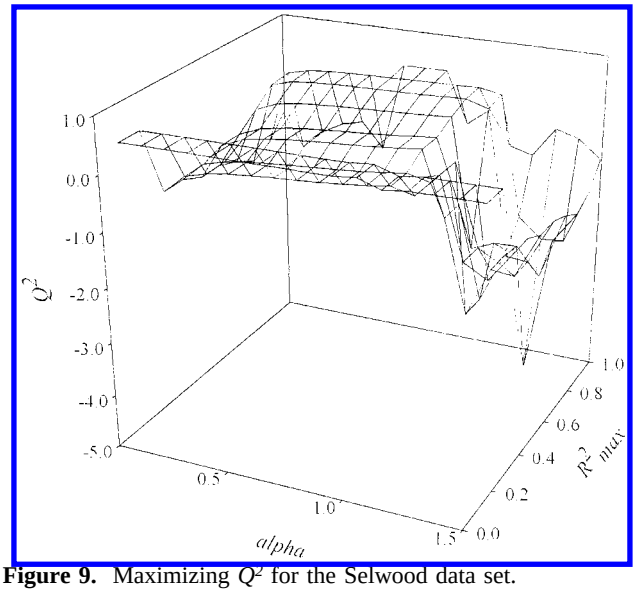

Figure 9. Maximizing Q^2 for the Selwood data set.

Table 4. Optimal Models for the Selwood Data Set with Compound M6 Removed

R^2_{max}	α	variables	components	Q^2	R^2
0.1	1.0	2	1	0.4113	0.4270
0.2	1.0	2	1	0.4113	0.4270
0.3	1.0	2	1	0.4113	0.4270
0.4	1.1	3	2	0.3286	0.4380
0.5	1.1	3	2	0.3286	0.4380
0.6	1.1	3	2	0.3286	0.4380
0.7	1.0	4	1	0.3793	0.4363
0.8	1.0	6	1	0.4458	0.4787
0.9	0.8	8	1	0.4866	0.5310
0.99	0.0	12	1	0.4964	0.8517

1-Component Model ($R^2_{max} = 0.99$, $\alpha = 0.0$)

Analysis of Variance

	DF	SS	MS	F-ratio	Prob>F
model	12	24.7004	2.0584	8.1386	0.0001
error	17	4.2996	0.2529		
total	29	20.3539			
S	0.5029				
R^2	0.8517				
adjusted R^2	0.7611				

Bootstrap 95.0% Confidence Limits (Based on 5000 Bootstraps)

	lower	median	upper	mean	stderr
ATCH8	−3.9277	0.0605	4.0518	0.0948	2.1537
ATCH9	−1.4681	−0.0129	1.2296	−0.0359	0.7986
ESDL5	−1.0053	−0.1880	0.9849	−0.1585	0.5902
ATCH7	−1.9265	0.2049	2.4041	0.2453	1.3341
SUM_F	0.0914	0.9037	1.8585	0.9185	0.4176
NSDL9	−0.9815	−0.0535	0.6861	−0.0768	0.4804
MOFI_Z	−0.1966	1.3536	2.6885	1.3390	0.7559
S8_1DX	−2.4146	−1.1739	0.0353	−1.1765	0.6543
SUM_R	−0.6086	0.4421	1.3414	0.4317	0.5071
ATCH4	−2.7713	0.0366	2.2178	0.0449	2.0804
S8_1DY	−2.0059	−0.8678	0.0353	−0.8980	0.5330
NSDL10	−0.1696	0.6742	1.5089	0.6700	0.4501

In terms of the original, unstandardized variables eq 11 is equivalent to

$$\log \frac{1}{C} = \underset{(\pm 0.29)}{-1.43 + 1.86\ SUM\_F} + \underset{(\pm 8.04\text{E}-6)}{3.94\text{E}-5\ MOFI\_Z} \qquad (12)$$

The bootstrap estimates for the standard errors suggest that both terms (*SUM_F* and *MOFI_Z*) are significant. A randomization test for this model using 1000 permutations of the response variable produced tail probabilities less than 0.0001 for fit and 0.0014 for prediction.

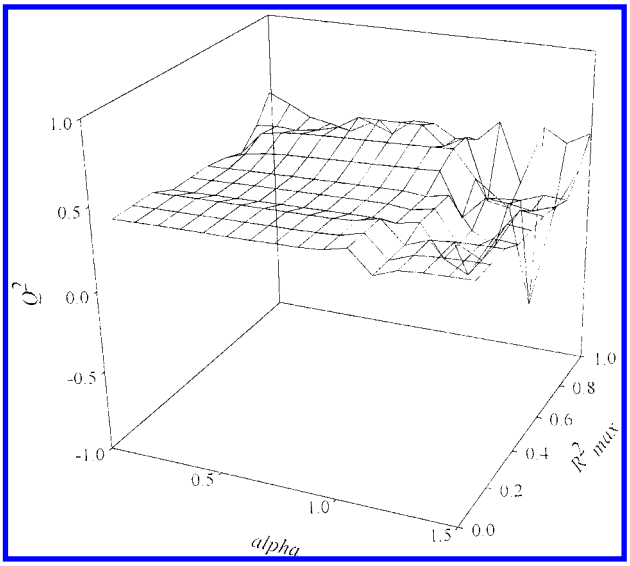

Figure 10. Maximizing Q^2 for the Selwood data set with compound M6 removed.

To understand why this result is so disappointing, scatter plots of the 12 variables against the response were examined, revealing that the data are very badly distributed: some variables have outlying values (*ATCH4*, *NSDL10*); others are clearly clustered into two groups (*ATCH8*, *SUM_R*). In fact none of these variables has a relationship with the response that is visible to the naked eye. In an attempt to obtain improved models, compound *M6* was removed from the data set. (This compound has a wildly outlying value of *NSDL10*: 0.9, while the remaining compounds have values between −0.2 and 0.1.) The UFS/CR modeling was then repeated with the results shown in Table 4 and Figure 10. These are an improvement over the earlier results, but are still poor, producing the optimized model with with $R^2 = 0.85$ and $Q^2 = 0.5$ when $R^2_{max} = 0.99$ and $\alpha = 0.0$:

$$\log \frac{1}{C} = \underset{(\pm 0.25)}{2.47\ C1}$$

$$t = 9.8824,\ prob{>}t = 0.0000,\ R^2 = 0.85,\ Q^2 = 0.5 \quad (13)$$

where

$$C1 = 0.037\ ATCH8 + 0.023\ ATCH9 - 0.043\ ESDL5 + 0.021\ ATCH7 + 0.39\ SUM\_F - 0.0091\ NSDL9 + 0.59\ MOFI\_Z - 0.48\ S8\_1DX + 0.24\ SUM\_R + 0.0047\ ATCH4 - 0.39\ S8\_1DY + 0.22\ NSDL10$$

so that, with standard errors estimated by bootstrapping,

$$\log \frac{1}{C} = \underset{(\pm 2.15)}{0.091\ ATCH8} + \underset{(\pm 0.80)}{0.056\ ATCH9} - \underset{(\pm 0.59)}{0.11\ ESDL5} + \underset{(\pm 1.33)}{0.051\ ATCH7} + \underset{(\pm 0.41)}{0.98\ SUM\_F} - \underset{(\pm 0.48)}{0.023\ NSDL9} + \underset{(\pm 0.76)}{1.47\ MOFI\_Z} - \underset{(\pm 0.65)}{1.2\ S8\_1DX} + \underset{(\pm 0.51)}{0.58\ SUM\_R} + \underset{(\pm 2.1)}{0.012\ ATCH4} - \underset{(\pm 0.53)}{0.96\ S8\_1DY} + \underset{(\pm 0.45)}{0.54\ NSDL10} \quad (14)$$

In terms of the original, unstandardized variables we have

$$\log \frac{1}{C} = -0.012 + \underset{(\pm 0.013)}{5.4\ ATCH8} + \underset{(\pm 46)}{3.2\ ATCH9} - \underset{(\pm 0.46)}{0.084\ ESDL5} + \underset{(\pm 60)}{2.3\ ATCH7} + \underset{(\pm 1.7)}{4.0\ SUM\_F} - \underset{(\pm 3.8)}{0.18\ NSDL9} + \underset{(\pm 6.6\text{E}-5)}{1.3\text{E}-4\ MOFI\_Z} - \underset{(\pm 0.27)}{0.49\ S8\_1DX} + \underset{(\pm 2.2)}{2.5\ SUM\_R} + \underset{(\pm 0.18)}{0.10\ ATCH4} - \underset{(\pm 0.13)}{0.23\ S8\_1DY} + \underset{(\pm 2.8)}{3.3\ NSDL10} \quad (15)$$

A randomization test for this model using 1000 permutations of the response variable produced tail probabilities less than 0.0008 for fit and 0.0328 for prediction.

The *n*-fold cross-validation results for $n = 2$, 6, 10, 15, 20, and 25 are shown in Figure 11 for the entire data set and in Figure 12 for the case where M6 was removed. In the latter case all but 2 of the cross-validation results for $n = 2$ gave values of $Q^2 < 0$ and are omitted from Figure 12. Although eq 13 produces the highest optimal Q^2, because of the degree of multicollinearity ($R^2_{max} = 0.99$), it contains only 2 significant terms (*ATCH8*, *SUM_F*) as judged by the z-test with standard errors calculated by bootstrapping. Moreover, the standard errors are highly variable and as a result inference is difficult. It is not clear whether the multicollinearity exhibited by the variables in eq 13 represents a feature of the underlying population, in which case the model may still have some predictive power.

4. DISCUSSION

The preprocessing procedure advocated here, identifying variables with a significant projection onto the response, eliminating irrelevant variables, and addressing redundancy and multicollinearity by UFS, has the aim of producing a subset of variables that meets the requirements of OLS. When these requirements are not met, a component-based construction is required for model building. Continuum regression provides a close to optimal construction covering OLS, PCR and PLS.[14] Leave-one-out cross-validation is employed to select a robust model, and the final *n*-fold cross-validation indicates the chances of the model generalizing to new objects. The issue of chance correlation is addressed by reducing the number of variables used during model specification, and by the selection of a robust model: good predictive properties are an indication that chance correlation has been avoided. Moreover, the selected variables are relevant, with unique information and minimal collinearity. In the examples studied here, this procedure leads to models with a small number of components (often only one) on a focused set of variables. Such models are far easier to interpret than models with several latent variables constructed from a large number of descriptors.

4.1. Achieving a Generalized Model. Many of the results obtained here illustrate the tradeoff between over-fitting and generalization. The results for the steroid data set in Table 1, for example, show that increasing the number of components produces a better fit, but at the expense of reducing Q^2. A 5-component model with $Q^2 = 0.6$ and $R^2 = 0.98$ has been reported for the same data set.[16] As noted earlier, the results of *n*-fold cross-validation for this model (Figure 4) are generally weaker than those for the UFS/CR model (Figure 3), with an increasing spread of values of Q^2 as the

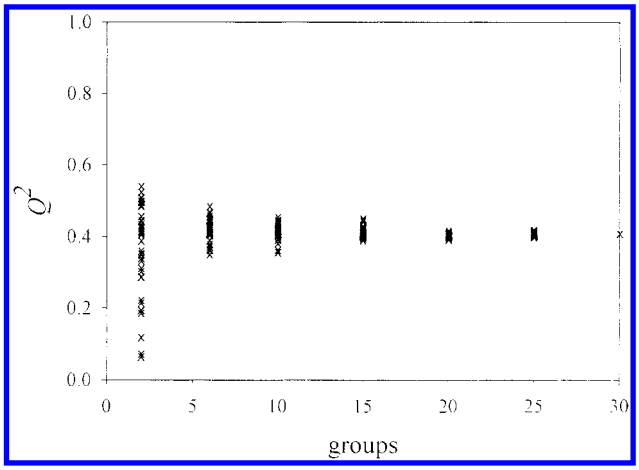

Figure 11. *n*-fold cross-validation for the optimal 2-variable model for the Selwood data set with compound M6 removed.

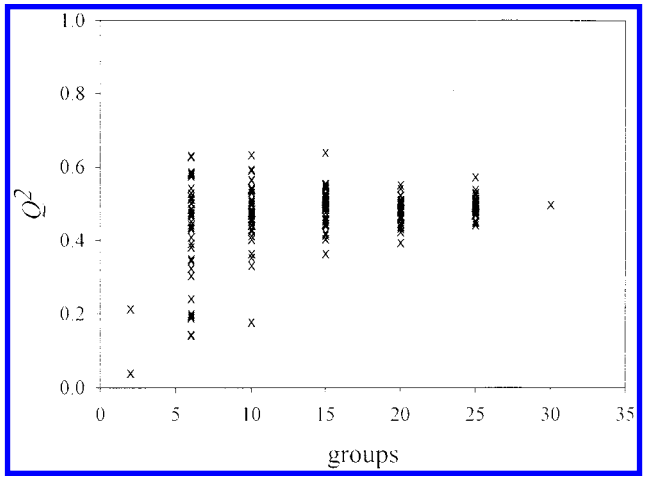

Figure 12. *n*-fold cross-validation for the optimal 12-variable model for the Selwood data set with compound M6 removed. 38 negative Q^2 values for $n = 2$ are not shown.

group size *n* decreases. The single value of Q^2 produced by leave-one-out cross-validation can be misleading since the results of *n*-fold cross-validation may have considerable variance while more closely representing the expected outcome when predicting the activities of a test set from a model constructed using an independent training set. Although use of independent training and test sets is the ultimate check, this is seldom practicable in the initial stages of a drug design program.

The results for the pyrethroid data set in Table 2 show that even maximizing Q^2 may not produce the "best" model. Here the *n*-fold cross-validation reveals that the 2-component model ($Q^2 = 0.77$) is inferior to the 1-component model ($Q^2 = 0.76$). In this case, choosing the more parsimonious eq 6 is supported by the increased precision indicated by the standard errors estimated by bootstrapping.

The models produced for the widely studied Selwood data set (Table 3) are much less successful, largely due to the fact that the data are so poorly distributed. Since this study was first published, and probably as a result of the full data being made available as Supplementary Material,[13] it has become accepted as a "standard" set for assessing new mathematical modeling techniques. Methods applied to this set have included joint eigenvector regression and alternating conditional expectations,[18] cluster significance analysis[19] and variants,[20] genetic function approximations,[21] Kohonen mapping,[22] evolutionary algorithms[23,24] back-propagation neural networks,[25] evolutionary programming,[26] systematic search,[27] genetic algorithms,[28] cascade-correlation neural networks,[9] genetic neural networks,[29] and neural networks with active neurons.[30] Despite this intense study, and with two notable exceptions,[28,31] few if any comments have been made on the quality of this data set. The poor data distributions may be due to a number of reasons such as incorrect structure representation in the original molecular models, imperfections in the algorithms[32] used to derive properties from the semiempirical calculation output or, perhaps, by the nature of the compounds themselves. Whatever the cause, and because the data set is available on the QSAR and Modeling Society Web site (http://www.pharma.ethz.ch/qsar/), perhaps it is time that this set was no longer regarded as a "standard" and is perhaps flagged as "difficult" if not flawed.

4.2. Reducing Dimensionality. The use of latent variables in regression attempts to address the problem of multicollinearity. A number of the close to optimal standardized models reported here have identical or very similar standard errors for each of the included terms (compare the one component models (2), (5), and (8)). As one more component is added, however, the standard errors increase in size and diverge (cf. the common terms in eqs 5 and 8). Thus, as more components are added the precision of the final prediction model decreases to result in less certainty of accurate prediction. Similarly, as more terms are used to construct a component, the standard errors obtained for the original variables by bootstrapping increase, leading to unstable beta estimates (the familiar problem of "bouncing betas"). For the Selwood data set, the optimal model obtained by removing an object (M6) has a maximum Q2 (0.50) but only 2 significant terms from a total of 12 variables used to construct this component (eq 13). These important results emphasize the requirement for using models of as low dimension as possible that are consistent with maximizing Q^2. Using low dimensional models will lead to smaller average distances for interpolation in the multivariate property space.

4.3. Feature Recognition. It is often argued that use of latent variables constructions on preprocessed data can lead to models that omit terms regarded by medicinal chemists as important explanatory variables. This problem can be addressed by calculating the component loadings (correlations of the original variables with a component) for all the variables in the over-square data matrix. The loading patterns can then be reviewed in order to identify those sets of variables highly associated with the component(s) included in the model. This can help to identify the features associated with the response.

This is illustrated for the pyrethroid set (1 component, 3 variable model). The component loadings significant at the 1% level are shown in Table 5. Based on this pattern of loadings, killing potency appears to be associated with the following features: (i) the atomic charges at the meta positions of the benzyl ring; (ii) the atomic charge of the ether linkage; (iii) the smallest component of the moment of inertia; (iv) the dipole along the long axis of the pyrethroid molecule; (v) the electrophilic superdelocalizability of the cyclopropane atom subtending the geminal dimethyl substituents; and (vi) the nucleophilic superdelocalizability of the vinyl carbon attached to the propane ring. These features

Table 5. Loadings for the 1-Component Pyrethroid Model with Tail Probability $p < 0.01$

variable	loading	variable	loading
A5	0.756	DVX	−0.603
A3	0.723	ES12	−0.584
A8	0.619	MIZ[a]	0.567
NS16	−0.605		

[a] Variable MIZ is included here because it occurs in the model, although its loading falls just below the value (0.575) required for significance at the 1% level.

may be useful for developing a putative pharmacophore for the killing action of pyrethroid insecticides.

ACKNOWLEDGMENT

The UFS and continuum regression algorithms used in this paper were developed as part of the BBSRC Cooperation with Industry Project "Improved Mathematical Methods for Drug Design" (BBSRC grant number 322/6284).

REFERENCES AND NOTES

(1) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−44.
(2) Livingstone, D. J. Pattern Recognition Methods for use in Rational Drug Design. In *Molecular Design and Modeling*: *Concepts and Applications*; Langone, J. J., Ed.; Academic Press: 1991; Vol. 203 of Methods in Enzymology, pp 613−638.
(3) Livingstone, D. J. *Data Analysis for Chemists*: *Application to QSAR and Chemical Product Design*; Oxford University Press: Oxford, 1995.
(4) Livingstone, D. J.; Rahr, E. Corchop − An Interactive Routine for the Dimension Reduction of Large QSAR Data Sets. *Quant. Struct-Act. Relat.* **1989**, *8*, 103−8.
(5) Kikuchi, O. Systematic QSAR Procedures with Quantum Chemical Descriptors. *Quant. Struct-Act. Relat.* **1987**, *6*, 179−84.
(6) Gute, B. D.; Basak, S. C. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131.
(7) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(8) Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.
(9) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Kholodovych, V. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 3. Variable Selection in the Cascade-Correlation Learning Architecture. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 651−9.
(10) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery: have they lived up to their promise? *Eur. J. Med. Chem.* **1999**, *34*, 195−208.
(11) Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis(CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(12) Ford M. G.; Greenwood R.; Turner C. H.; Hudson B.; Livingstone D. J. The Structure−Activity Relationships of Pyrethroid Insecticides. 1. A Novel Approach Based upon the use of Multivariate QSAR and Computational Chemistry. *Pestic. Sci.* **1989**, *27*, 305−326.
(13) Selwood, D. L.; Livingstone, D. J.; Comley J. C. W.; O'Dowd A. B.; Hudson A. T.; Jackson P.; Jandu K. S.; Rose V. S.; Stables J. N. Structure−Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.
(14) Malpass, J. A.; Salt, D. W.; Ford, M. G.; Wynn, E. W.; Livingstone, D. J. Continuum Regression: A New Algorithm for prediction of Biological Activity. In *Advanced Computer-Assisted techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim, New York, 1994; Vol. 3 of *Methods and Principles in Medicinal Chemistry*, pp 163−189.
(15) PARAGON drug design software, Centre for Molecular Design, University of Portsmouth (http://www.cmd.port.ac.uk/webdocs/paragon.html).
(16) *SYBYL 6.4 Ligand-Based Design Manual*; Tripos Inc.: St. Louis, MO, 1997; pp 41−50.
(17) Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Chapman and Hall, 1993.
(18) Forina, M.; Mosti, L. Joint Eigenvector Regression and Alternating Conditional Expectations. In *QSAR*: *Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier Science Publishers: Amsterdam, 1991; pp 181−4.
(19) McFarland, J. W.; Gans, D. J. On Identifying Likely Determinants of Biological Activity in High-Dimensional QSAR. *Quant. Struct.-Act. Relat.* **1994**, *13*, 11−17.
(20) Rose, V. S.; Wood, J.; MacFie, H. J. H. Generalized Single Class Discrimination (GSCD) − A New Method for the Analysis of Embedded Structure−Activity Relationships. *Quant. Struct.-Act. Relat.* **1992**, *11*, 492−504.
(21) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−66.
(22) Rose, V. S.; Croall, I. F.; MacFie, H. J. H. An Application of Unsupervised Neural Network Methodology (Kohonen Topology Preserving Mapping) to QSAR Analysis. *Quant. Struct.-Act. Relat.* **1991**, *10*, 6−15.
(23) Kubinyi, H. Variable Selection in QSAR Studies. 1. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−94.
(24) Kubinyi, H. Evolutionary Variable Selection in Regression and PLS Analyses. *J. Chemometrics* **1996**, *10*, 119−33.
(25) Wikel, J. H.; Dow, E. R. The Use of Neural Networks for Variable Selection in QSAR Studies. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−51.
(26) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−87.
(27) Kubinyi, H. Variable Selection In QSAR Studies. 2. A Highly Efficient Combination Of Systematic Search And Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.
(28) Leardi, R. Genetic Algorithms in Feature Selection. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, 1996; pp 67−86.
(29) So, S.-S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−30.
(30) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Ivakhnenko, A. G.; Livingstone, D. J. Application of Self-Organizing Neural Networks with Active Neurons for QSAR studies. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jørgensen, F. S., Eds.; Kluwer Academic/Plenum Publishers: New York, 2000; pp 444−5.
(31) Livingstone, D. J. The Trouble With Chemometrics. In *QSAR and Molecular Modelling*: *Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science Publishers: Barcelona, 1995; pp 18−26.
(32) Glen, R. C.; Rose, V. S. Computer program Suite for the Calculation, Storage and Manipulation of Molecular Property and Activity Descriptors. *J. Mol. Graph.* **1987**, *5*, 79−86.

CI000384C