Inhibition of the Tyrosine Kinase, Syk, Analyzed by Stepwise Nonparametric Regression

T. John McNeany and Jonathan D. Hirst*

School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.

Received December 7, 2004

A set of 538 inhibitors of the tyrosine kinase, Syk, including purines, pyrimidines, indoles, imidazoles, pyrazoles, and quinazolines, has been analyzed using a stepwise nonparametric regression (SNPR) algorithm, which has been developed for QSAR studies of pharmacological data. The algorithm couples stepwise descriptor selection with flexible, nonparametric, kernel regression, to generate structure−activity relationships. A further 371 molecules have been used as a test set to evaluate the models generated. Descriptors were selected using an internal monitoring set, and models were assessed using 10% of the principal (538-compound) data set, selected randomly, as an external validation set. The best model had a Q^2 of 0.46 for the external validation set. Test set predictions were significantly less accurate, partly due to the higher mean activity of the test molecules. However at a more coarse-grain level the SNPR models classified active molecules accurately, giving good enrichments. The data sets are difficult to model accurately and SNPR performs better than multilinear regression and a neural network analysis. In the additive implementation of SNPR multidimensional models are considered as a sum of single dimensional regressions. This makes the resultant models easily interpretable. For example, in the most predictive SNPR models, there is a clear nonlinear relationship between hydrophobicity (*A*log*P*98*)* and inhibitory activity.

INTRODUCTION

A stepwise nonparametric regression (SNPR) method has been applied to a data set of 538 inhibitors of the tyrosine kinase, Syk. Kinases are a family of homologous proteins, involved in many cell signaling pathways, and are therefore potential drug targets for a range of diseases from inflammation responses and autoimmunity, to oncology and cardiovascular disease.[1] The kinase family of enzymes catalyzes phosphate transfer from ATP or GTP, to specific residues on a protein substrate. The tyrosine kinase, Syk,[2] is expressed in all haematopoietic cells, that is cells involved in the production of blood cells, and has been identified as essential for development of lymphocytes and for signal transduction in nonlymphoid cells.[3] Syk has two Src homology 2 (SH2) domains which bind to immunoreceptor tyrosine-based activation motifs (ITAM), and upon activation, phosphorylate the tyrosine residues of ITAM.[4] Syk may also be required in nonimmunological receptor pathways in the maintenance of vascular integrity.[5] It is therefore of medicinal interest and an interesting pharmaceutical target. Inhibitors of tyrosine kinases are typically competitive against ATP, due to the deep ATP-binding cleft,[6] but an additional binding pocket has been identified, which allows inhibitors to compete indirectly with ATP.[7,8] Efforts to identify compounds, which are active against such targets, are producing a wealth of data, and computational approaches are increasingly applied to the interpretation of this information.

The development of new quantitative structure−activity relationships (QSAR) methods for the generation of predictive models of drug activity is of particular interest to medicinal chemists.[9−13] When using QSAR to model pharmaceutical data, it is common to generate many physicochemical descriptors, thereby maximizing the probability that all relevant molecular properties will be available for the purpose of constructing models. The multiplicity of descriptors, however, introduces some problems, which hampers attempts to create accurate, unbiased, computationally cheap QSAR models. Foremost among these issues is the selection of a subset of representative descriptors, from the pool of available properties. Despite increasing computing power, it is infeasible to search through descriptor space exhaustively to select a useful subset, even for data sets with a modest number of molecular properties. In addition, indiscriminate searching increases the likelihood of identifying a correlation by chance.[14] Selected models are, consequently, less reliable and require more rigorous validation to establish their general applicability. Furthermore, in traditional parametric approaches, cross terms and functional forms of the available variables are often included explicitly.[15] This not only confounds the search problem but also introduces bias to the model, by constraining the data to fit a given functional form.

Determination of a descriptor set, in a multilinear regression problem, is often achieved using a forward selection procedure. The most predictive descriptors are added to a model one by one, until no further significant improvement in the model is seen. A related alternative is backward elimination of variables. In this case, a model is generated which uses all available descriptors. Descriptors are then removed from the model if they are not contributing significantly to the quality of the model. These competing processes can be combined in a stepwise procedure,[16] whereby descriptors are added if they are useful in the current model but removed if they no longer contribute valuable

* Corresponding author phone: +44 115 951 3478; fax: +44 115 951 3562; e-mail: jonathan.hirst@nottingham.ac.uk.

10.1021/ci049631t CCC: $30.25
Published on Web 03/25/2005

information as the algorithm progresses. This helps to avoid descriptor sets with redundant variables.

Nonparametric kernel regression is emerging as a useful tool for QSAR,[17−19] which does not require explicit inclusion of functional forms of variables because of the flexibility of the procedure. It does not bias the model to a specified form, allowing underlying trends to be identified and accurately represented. Nonparametric regression is similar in concept to k-nearest neighbors,[20] in that predicted activities are based on the activities of neighboring molecules, but the use of a bandwidth optimized kernel function gives a more general definition of a neighborhood. It is more computationally demanding than standard multilinear regression tools, but the recent increases in computing power are making nonparametric approaches more accessible. Other nonlinear fitting approaches have been considered in this context, most notably artificial neural networks,[21−24] which are capable of solving nonlinearly separable problems, and have become a standard tool in QSAR.

Neural network methods link input variables to response variables through a system of interconnected weights. Initial weights are determined randomly and then iteratively adjusted to reduce residual error. Problems can arise due to the interaction of weights, resulting in poor convergence. This can make neural networks slow to train, especially when many input variables are present. In addition, the random initial weights can cause neural networks to be unstable, with separate runs giving significantly different results for the same input data. To address the issue of instability, neural networks are often used in ensembles. This approach uses a number of networks, each independently solving a given problem. The final network predictions are then taken as the mean prediction of the network ensemble. Binary kernel discrimination,[25] support vector machines,[26,27] and smoothing splines[28−31] have also been applied to QSAR problems and are more predictive than multilinear regression in many cases.

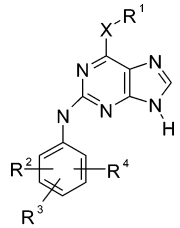
Numerous optimization approaches have been explored as descriptor selection routines. Stochastic search methods including simulated annealing,[32] genetic algorithms,[33] genetic programming,[34] and particle swarms[35] have been shown to be effective for neural network models and multilinear regression. Simulated annealing[36] uses a random walk to explore new descriptor combinations, accepting steps which improve an objective function and accepting detrimental steps, based on the Metropolis criterion.[37] The temperature, which is used to determine the Boltzmann factor in the Metropolis criterion, is lowered as the algorithm progresses, making acceptance of detrimental steps less likely. Given an infinitely slow cooling scheme, this method is guaranteed to find the global optimum. However, in practice, it performs poorly when the objective function surface is rough or has many similar local optima. Genetic algorithms[38] select good solutions from a population and then generate a new population of solutions, using evolutionary crossover and mutation operators. As new generations are created, the fitness of the solutions should increase, due to selective pressure, until a good solution is found. Particle swarms[35] use a population of individual search particles, which explore the objective function surface, and, through social interaction, are directed toward areas where the objective function is optimal.

In comparison, stepwise selection of variables is a deterministic, greedy algorithm, which is computationally cheap. It explores all alternative models in the immediate vicinity of a current solution, that is, models involving the addition or removal of a single descriptor, and accepts the solution that gives the most improvement in the objective function. It, therefore, proceeds through a series of local optima, until it can find no better solutions in the local neighborhood. This does not guarantee global optimality but rapidly converges on a good local optimum. In general a stepwise procedure leads to low dimensional models, identifying only a small number of relevant variables. Stepwise selection therefore samples a small number of possible descriptor combinations, making it a fast, efficient algorithm, and the resulting models are low-dimensional and easily interpreted. Given the discontinuous nature of descriptor space and the size of the combinatorial problem, we suggest that stepwise selection is a competitive method for determining an informative variable set for nonparametric regression.

METHODS

We present a deterministic, stepwise nonparametric regression (SNPR) procedure, which addresses several of the key problems encountered in the generation of unbiased, validated QSAR models. SNPR uses a greedy stepwise selection procedure, comprising forward selection and backward elimination of descriptors, to establish a meaningful subset of variables, for a nonparametric QSAR model. This is an algorithmically straightforward method, which is computationally cheap, due to the narrow focus of the search, sampling only a very small number of all possible descriptor combinations.

The data set contains two major structural classes: 92 purines and 316 pyrimidines, with remaining molecules split between indoles, imidazoles, pyrazoles, and quinazolines. Details of the purine derivatives have been described in a recent patent[39] and are based on the structure **1**. The R-groups are outlined in the patent and cover a wide range of structural features, ranging from simple alkyl and cycloalkyl substituents to haloalkyl and substituted amines. Substituted amines introduce further R-groups. The property and activity data for these compounds are available from the authors upon request, but the proprietary nature of the data set precludes further details of compound structures.



1

The molecular activity is given as a pIC_{50} value, and the molecules in the set span three log units. Molecules are considered to be highly active at concentrations of 50 nM and lower and inactive above 1 μM. Molecules between these limits are classified as moderately active. The data set comprises 14% highly active molecules, 43% moderately active, and 43% inactive molecules. A second data set of

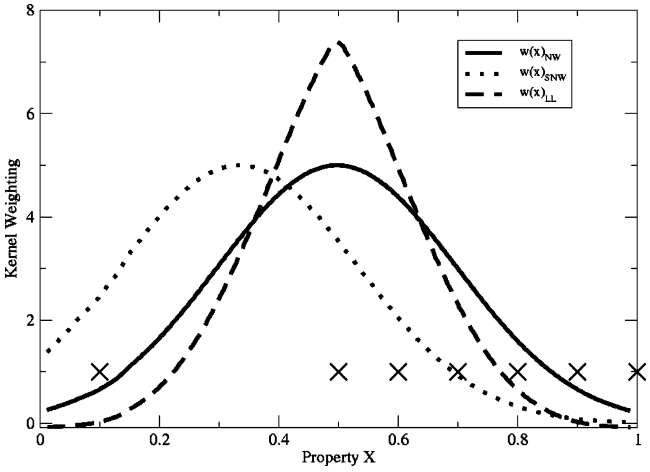


Figure 1. Nonparametric regression kernels. The local linear kernel has been scaled for comparison, and an arbitrary bandwidth has been chosen.

371 molecules, including 204 purines and 53 pyrimidines, has been retained as a test set to facilitate comparison with previous studies.[19] The molecules were synthesized and screened at a later stage in the development process; the percentage of actives (32%) is, thus by design, higher than that of the original set. The skew toward higher activity of the test set is reflected in the higher percentage of moderately active compounds (59%) and the lower fraction of inactives (9%). For both sets 74 physicochemical properties were generated using the Molecular Simulations Inc. Cerius2 package.[40] Kernel regression methods have been applied to this data set in previous work,[19] but descriptors were selected through an exhaustive search. Consequently, only low dimensional models were considered. The stepwise selection procedure, implemented here, allows higher dimensional models to be studied, and the rigorous validation procedure generates more reliable models.

Nonparametric regression as a QSAR tool has been presented in detail previously.[18] For a set of N molecules with associated activities y, the regression function m, corresponding to a predicted activity, at a point x in descriptor space is estimated as a distance-dependent, normalized, weighted sum of the activities of all N molecules:

$$\hat{m}(x) = \sum_{i=1}^{N} w_i(x) y_i \tag{1}$$

The locally fitted, distance based, weighting function $w(x)$ is determined by a normalized regression kernel, $K_h(x_i\text{-}x)$, a function of the distance of each datapoint x_i from the point of interest, x. Figure 1 shows a simple case of the three kernel functions considered in this study, centered on 0.5 and computed for the example datapoints shown.

The Nadaraya−Watson regressor[41,42] is implemented in a genuine multivariate fashion and is a Gaussian function

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{N} K_h(x_i - x) y_i}{\sum_{i=1}^{N} K_h(x_i - x)} \tag{2}$$

where

$$K_h(x_i - x) = h^{-1}(2\pi)^{-1/2} e^{-(x_i - x)^2/2h^2} \tag{3}$$

where h is the bandwidth of the Gaussian, which is selected using the block method.[43] As can be seen in Figure 1, the kernel weighting is highest around the point of interest and tails off toward zero as the distance increases. This means that molecules that lie close in property space have a greater contribution to the kernel estimate of an activity that those further away. The shifted Nadaraya−Watson kernel[44] is also a Gaussian but based on reverse mass recentering of the data to achieve a more robust estimate when data are sparse

$$\hat{m}_{SNW}(x) = \frac{\sum_{i=1}^{N} K_h(x_i - \xi^{-1}(x)) y_i}{\sum_{i=1}^{N} K_h(x_i - \xi^{-1}(x))} \tag{4}$$

where

$$\xi(x) = \frac{\sum_{i=1}^{N} K_h(x_i - x) x_i}{\sum_{i=1}^{N} K_h(x_i - x)} \tag{5}$$

and the reverse mass recentering function, $\xi^{-1}(x)$, is computed as a polynomial fit. The local linear regressor[45] uses a locally fitted polynomial correction to the Nadaraya−Watson kernel

$$\hat{m}_{LL}(x) = \frac{\sum_{i=1}^{N} \{\hat{s}_2(x) - \hat{s}_1(x)(x_i - x)\} K_h(x_i - x) y_i}{\sum_{i=1}^{N} \hat{s}_2(x)\hat{s}_0(x) - \hat{s}_1(x)^2} \tag{6}$$

where

$$\hat{s}_r(x) = \sum_{i=1}^{N} (x_i - x)^r K_h(x_i - x) \tag{7}$$

When data designs are uniform, the kernels show similar behavior, but when available data are nonuniform, as in the case illustrated in Figure 1, the kernels are significantly different. The shifted Nadaraya−Watson kernel is moved away from the center of mass of the datapoints, while the local linear kernel remains centered but shows a narrower distribution. The local linear regressor has a lower bias than the Nadaraya−Watson estimator, when data are irregularly distributed, but may perform poorly when data are sparse and discontinuities exist. The shifted Nadaraya−Watson kernel is more stable when discontinuities are present.

The shifted Nadaraya−Watson kernel and local linear kernel have both been applied in an additive fashion.[46,47] This separates a p-dimensional regression function into p, independent, one-dimensional functions as shown in eq 8, thereby overcoming the problem of data sparsity in high dimensional

spaces

$$m(x) = \alpha + \sum_{j=1}^{p} f_j(x_j) \tag{8}$$

where α is a constant. The independent nature of these functions allows high-dimensional models to be interpreted in terms of the individual contributions of each descriptor. This approach ignores the effect of cross terms, but the improved stability of kernel regression in multidimensional problems and the increased interpretability, which result from an additive approach, justify its implementation.

Stepwise selection of variables is a standard tool in multilinear regression[16] and has been used in many QSAR studies.[48−52] It is an iterative procedure with two components: a variable inclusion step (forward selection) and a descriptor removal step (backward elimination). Starting with a model where no variables are included and all activities are predicted to be the mean activity of the training set, forward selection is carried out, adding each unused variable in turn and comparing the new model with the initial model. The relative value of adding each variable is assessed using a partial *F*-test. The statistic, *F*, is defined as

$$F = \frac{SS_r - SS_f}{\hat{\sigma}^2} \tag{9}$$

where

$$\hat{\sigma}^2 = SS_f/df \tag{10}$$

SS_r and SS_f are the sum squared errors of the models with and without the particular variable, and the degrees of freedom, *df*, is the number of datapoints minus the number of variables in the model. If the highest *F*-value is above a user-defined threshold, that variable is accepted and added to the current model. This is continued until no further explanatory variables can be added to the model.

The inclusion of numerous descriptors introduces the possibility of redundancy, arising from combinations of descriptors, which provide essentially the same information as another single variable. Therefore, variables that initially had a significant contribution to the model may no longer be necessary. Backward elimination is a mechanism which allows the removal of redundant variables from a current model, thereby limiting models to the minimum number of explanatory variables. It is achieved by considering each possible reduced model, i.e. the current model with a single variable removed. The model with the lowest partial *F*-test is taken as the current model, if the *F*-value is below a threshold, and the superfluous descriptor is removed.

The partial *F*-test gives a quantitative comparison of two models of different dimensionality, which penalizes high-dimensional models, by inclusion of the $\hat{\sigma}^2$ term. In a stepwise selection procedure, the strictness of the penalty is determined by the user-defined thresholds chosen for the forward selection and backward elimination steps. To eliminate these parameters, SNPR has been implemented with a relaxed inclusion criterion for new descriptors. By removing the $\hat{\sigma}^2$ term from the partial *F*-test, eq 9 reduces to

$$F = SS_r - SS_f \tag{11}$$

This allows any model, which reduces the residual sum of squares, to be accepted. By having a relaxed selection criterion, models with a suitable number of descriptors can be selected with knowledge of the merits of models over a range of dimensionalities, rather than relying on thresholds, chosen a priori, for variable inclusion and elimination. Stepwise selection scales approximately linearly with increasing dimensionality, which compares very favorably with an exhaustive search.

A stepwise selection strategy, coupled with nonparametric regression, is a computationally cheap QSAR tool, but additional features are required to provide stringent validation of selected models. There are two main approaches to model assessment: cross-validation and holdout validation sets. Validation using holdout sets involves selecting a subset of compounds, which are not used in building the model, to establish the accuracy of the model. An *N*-fold cross-validation approach divides the data set into *N* subsets, each of which is excluded in turn. A model is built for each set, using the remaining data. Model quality is then assessed using the combined predictions of the *N* models. It has been argued that, particularly in the case of small data sets, cross-validation is less wasteful in terms of training data and also provides a more accurate measure of model quality.[53] However, when large data sets are available, holdout sets are as reliable as cross-validation[53] and less computationally demanding.

Training set performance is a generally poor objective function for descriptor selection, and a validated indicator is required to drive the selection. The cross-validated (or holdout set) performance is a more effective objective function to guide a search algorithm, but when used as an objective function, can no longer also guarantee model validity. We hierarchically partition the available data, to create a randomly selected external validation set, which is excluded from the process of variable selection. The remaining data are divided, using a Kennard-Stone set selection algorithm,[54] into training data for the generation of models and an internal monitoring set. The Kennard-Stone algorithm selects a diverse subset of molecules, within the descriptor space being considered. The objective function chosen to guide the stepwise search is the mean of the training set and monitoring set performances, which reduces the probability of overfitting, while avoiding the selection of internally inconsistent models. The absolute separation of validation data gives increased confidence in the quality of the selected model.

RESULTS

For model assessment, 10% of the molecules from the principal data set were selected randomly and withheld for validation of the models generated. The quality of the model fit is measured using R^2 and model predictivity in terms of Q^2

$$Q^2(R^2) = 1 - \frac{\sum_{i=1}^{N}(y_{i,obs} - y_{i,calc})^2}{\sum_{i=1}^{N}(y_{i,obs} - \bar{y})^2} \tag{12}$$

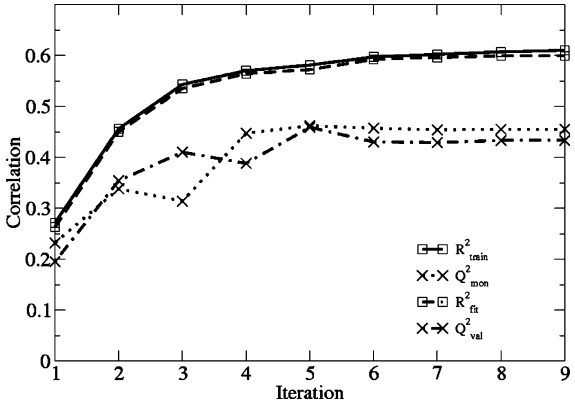


Figure 2. Quality of multivariate Nadaraya−Watson models selected by SNPR.

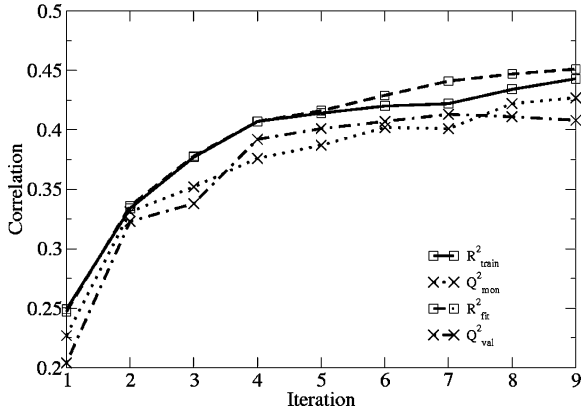


Figure 3. Quality of additive local linear models selected by SNPR.

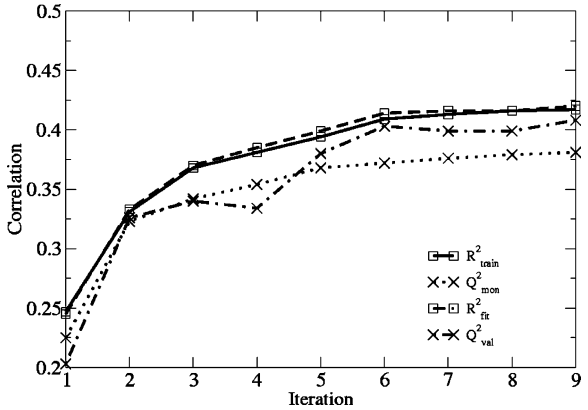


Figure 4. Quality of additive shifted Nadaraya−Watson models selected by SNPR.

where y_{obs} and y_{calc} are the observed and calculated activities respectively, and N is the number of molecules used for training. As this measure is dependent on the variance of the data, the standard deviation of the error of prediction (SDEP) and mean relative error (MRE) values are also included.

$$\mathrm{SDEP} = \sqrt{\frac{\sum_{i=1}^{N}(y_{i,calc} - y_{i,obs})^2}{N}} \tag{13}$$

$$\mathrm{MRE} = \frac{\sum_{i=1}^{N}\sqrt{\frac{(y_{i,obs} - y_{i,calc})^2}{y_{i,obs}}}}{N} \times 100\% \tag{14}$$

Because the SNPR approach to descriptor selection has a relaxed inclusion criterion, the addition of descriptors is only stopped when no higher dimensional models give a lower residual error, or when all available descriptors have been included. For computational efficiency, a cutoff has been added, limiting the models to 10 descriptors.

The progress of the selection for the Nadaraya−Watson kernel is summarized in Figure 2. R^2_{train} refers to the internal training data, Q^2_{mon} refers to the internal monitoring set, R^2_{fit} is the fit to the recombined internal data, and Q^2_{val} is the predictive accuracy for the external validation set. R^2_{train} increases with addition of descriptors, up to step eight. However, internal monitoring performance reaches a maximum at the fifth step, which corresponds to a five variable model. This model is also the most predictive, giving a Q^2_{val} of 0.46.

The results for the additive local linear kernel are shown in Figure 3. In this case training set performance continues to improve with the addition of variables, and the algorithm stops at the upper limit of 10 variables. There is also a noticeable improvement for steps five, six, and seven. Step six is an elimination step, which removes the first descriptor from the model. This is the only elimination step and gives a small reduction of the residual error. The following inclusion step, however, gives greater improvement. There is little justification for considering subsequent models, since the greater dimensionality increases the complexity of the model, with little reduction of the residual error for the internal validation set. While the overall error is higher for the local linear estimator than the Nadaraya−Watson, there is better agreement between the training and validation set performances. This suggests that the Nadaraya−Watson kernel is overfitting the training data or that, because it is not implemented additively, it is having difficulties with data sparsity in the regions of the validation molecules. This is consistent with a lower dimensional model giving the best performance and the early termination of the selection procedure.

The shifted Nadaraya−Watson kernel results (Figure 4) show good agreement between training and validation set performances, similar to the local linear kernel, but the models are less predictive. The choice of model is less clear in this case, but there is little improvement in the monitoring set performance beyond four descriptors. Performance statistics for the nonparametric approaches and the additional methods applied are summarized in Table 1.

The SDEP values for all models are around 0.5 log units, and the MRE values are approximately 6%. The data set is substantially larger than many medicinal chemistry data sets, and statistical measures of model quality are therefore more reliable.[53,55] The descriptors selected are shown in Table 1, and their identities are given in Table 2. The SNPR models all identify *A*log*P*98 as an important descriptor. It is a reparametrized *A*log*P* descriptor,[56] which is more easily computed than *C*log*P*. This is a measure of hydrophobicity, an important factor in transport through cell membranes. Other descriptors, which emerge as important, are CHI-2, Kappa-1-A, and SC-1. CHI-2 is a Kier and Hall connectivity

Table 1. SNPR, Multilinear Regression, and Neural Network Models

method	descriptors	R^2_{train}	Q^2_{mon}	R^2_{fit}	Q^2_{val}	SDEP	MRE%	Q^2_{test}
Nadaraya− Watson	40, 12, 13, 11, 41	0.58	0.46	0.57	0.46	0.47	6	0.10
additive local linear	13, 17, 38, 34, 26, 32	0.43	0.42	0.44	0.41	0.49	6	0.17
shifted Nadaraya−Watson	40, 13, 17, 26	0.38	0.35	0.39	0.33	0.52	7	0.10
multilinear regression	12, 15, 20, 10, 24, 38, 17	0.34	0.32	0.34	0.32	0.52	7	−0.07
neural network	30, 42, 73, 57, 20, 11, 70, 19, 35, 29	0.92	0.66	0.87	0.34	0.52	7	−0.15

Table 2. Descriptors Appearing in the Most Accurate Nonparametric QSARs

index	descriptor	index	descriptor
10	no. rotatable bonds	24	CHI-1
11	no. H-bond acceptors	26	CHI-2
12	no. H-bond donors	32	CHI−V-2
13	*A*log*P*98	34	CHI−V-3_C
15	Kappa-2	38	Wiener
17	Kappa-1-A	40	SC-1
20	Kappa-3-A	41	SC-2

index,[57] which is a representation of the degree of branching, and Kappa-1-A is a first-order shape index,[58] which indicates the degree of cyclicity of a molecule. SC-1 is a subgraph count[57] that measures the complexity of a molecule. Previous work[19] identified JX and PHI as important variables, where JX is a topological index and PHI is a flexibility index. These variables are not selected by any of the models in this study, but PHI is highly correlated with Kappa-1-A, and SC-1 is anticorrelated with JX. This suggests that similar information is presented by the new descriptor set.

To ensure that the models are not due to random correlations, the data set activities were randomized and the SNPR method was applied to the *y*-scrambled data. Five independent runs were carried out for each kernel resulting in no models with a positive Q^2_{val}. The mean Q^2_{test} for the additive local linear models was −0.50. For the Nadaraya−Watson and shifted Nadaraya−Watson models, the mean Q^2_{test} values were −0.53 and −0.56, respectively.

Using an additive approach, the contribution of each descriptor can be considered independently, improving the interpretability of the models. Figure 5 shows the separate contributions of each descriptor to the six-descriptor additive local linear model. The plots for each variable were for a fixed value for each of the remaining descriptors, which in most cases was suboptimal. The activity therefore only reflects the form of the activity relationship. All of the descriptors show a nonlinear dependence, demonstrating the need for a flexible fitting approach. The activity is predicted to be highest at *A*log*P*98 values around 2 and at Kappa-1-A values around 13. The separability of descriptor contributions makes additive nonparametric regressors more interpretable than multivariate approaches and neural networks, allowing regions of optimal activity to be identified for each descriptor.

For comparison, multilinear regression and neural network ensembles have also been applied to these data. The multilinear regression was implemented using the hierarchical partitioning scheme described previously. Q^2_{mon} reaches a maximum of 0.32, for a six-descriptor model. R^2_{train} for this model is 0.31, R^2_{fit} is 0.31, and Q^2_{val} is 0.32. Models give negative Q^2_{test} values up to eight variables and less than 0.05 for higher dimensional models.

The stepwise variable selection approach was used in conjunction with a cascade correlation neural network[59] ensemble. The ensemble consists of 10 networks, with predicted activities taken as the mean predicted activity of the 10 networks. The cascade correlation neural network ensemble is comparable to MLR in terms of model validation, but there is a large discrepancy between training fit and validation performance, suggesting that the neural network overfits the training data and generalizes poorly. The poor performance of the neural network may be a result of the variable selection procedure, and further investigation of descriptor selection procedures for neural networks would be required to assess a neural network approach more thoroughly. Results for *y*-scrambled data are similar to those for the nonparametric regression approaches.

The Q^2_{test} values for the SNPR models are low. The best is the local linear model, with a Q^2_{test} of 0.17, SDEP of 0.6, and MRE of 7%. The low Q^2_{test} may be due to the underrepresentation of active molecules in the training set, which has less than half the percentage of actives of the test set. The test set also has a significantly different distribution of structural classes. It is predominantly purine derivatives, compared with the training set which is dominated by pyrimidines. This may hinder the accurate prediction of the test set activities. However, the models perform well as a ranking method. The results of the ranking of the test compounds are presented in Figure 6, which shows the receiver operating characteristic (ROC) curves for the most predictive nonparametric regression models. ROC curves show the tradeoff between model sensitivity and model specificity.[60] Sensitivity is the ability of a model to identify positive results, defined as $TP/(TP+FN)$ where TP is the number of true positives and FN is the number of false negatives. The specificity of a model indicates its ability to identify negative results and is defined as $TN/(TN+FP)$ where TN and FP are true negatives and false positives, respectively. The associated area under a ROC curve, A_z, is a measure of model quality, with better models approaching unity. The additive models both have an A_z of 0.7 and the multivariate model 0.6.

The models can also be used as quantitative classifiers. If molecules are considered to be active above pIC_{50} of 7.3, all of the molecules predicted to be active by the local linear model are active, while the Nadaraya−Watson and shifted Nadaraya−Watson both have a hit rate of 50%. By lowering the prediction threshold below 7.3, the percentage of actives found rises rapidly. All the models identify the majority of actives, upon a reduction of the classification boundary by half a log unit, which is consistent with the SDEP for the models. The hit rate falls slowly and remains higher than the fraction of actives in the test set. The additive models both find half of the actives, with a hit rate greater than 50%, and the Nadaraya−Watson model finds half of the actives with a hit rate of approximately 47%. This is significantly better than the test set hit rate of 32%.

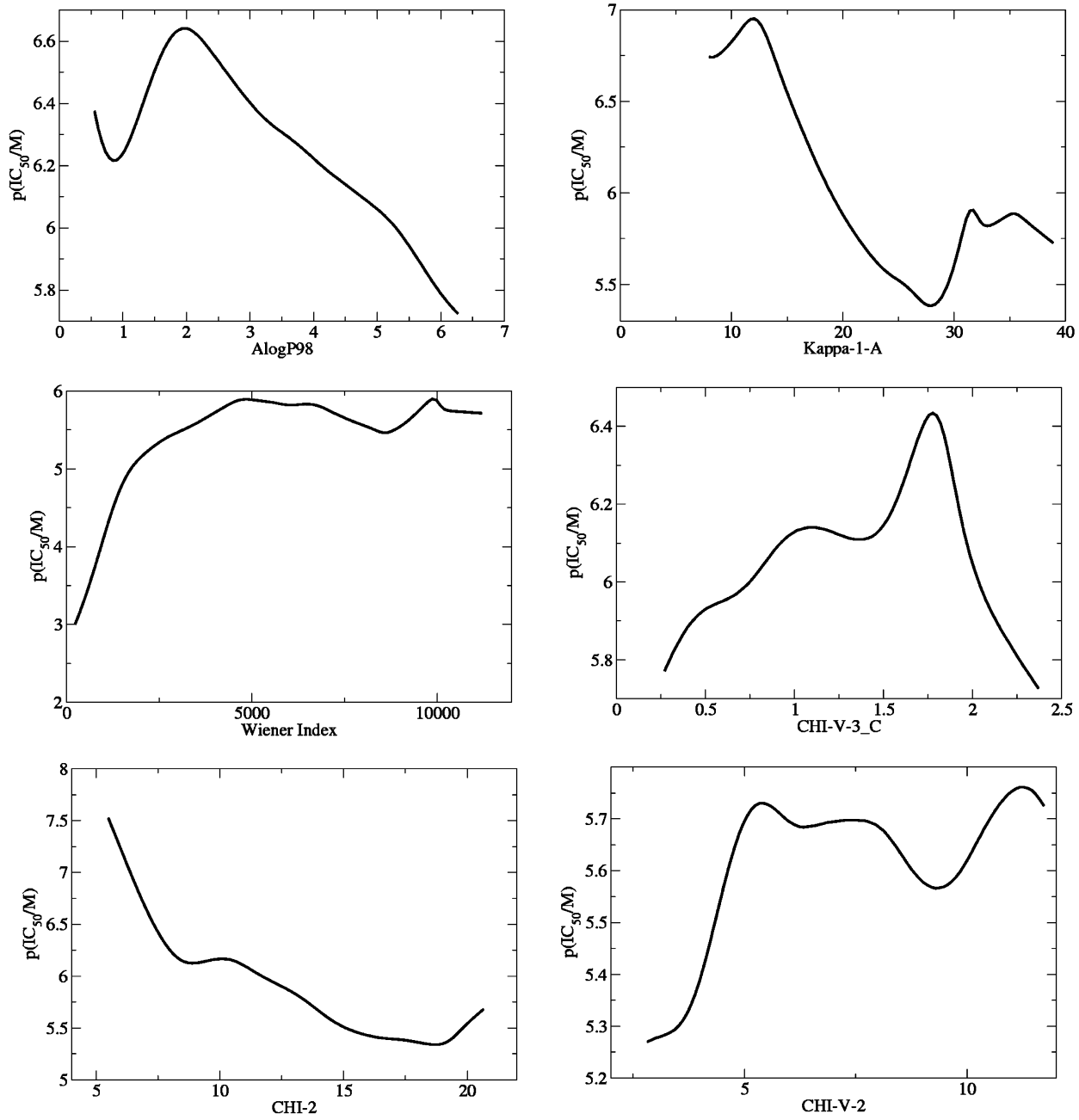


Figure 5. Individual contributions of each descriptor to the most predictive additive local linear model.

CONCLUSIONS

Analysis of the structure−activity relationships for inhibitors of the tyrosine kinase, Syk, has shown a nonlinear dependence on hydrophobicity and molecular shape. Activity has an almost parabolic dependence on *A*log*P*98, typical of drug molecules, which have to cross cell membranes. Low values for the Kappa-1-A shape index correspond to high activity, suggesting that more cyclic molecules have high activity. The best kernel regression model also highlights three connectivity indices as important, all of which show nonlinear behavior. The valence modified connectivity indices show a trend for increasing activity with increased branching, while the unmodified CHI-2 index suggests optimal activity at low values. It is possible that the Cerius2 descriptors are not the most appropriate for producing predictive models for these molecules. An alternative set of descriptors may prove more effective, in accurate modeling of these data and give more useful insight into the action of the molecules.

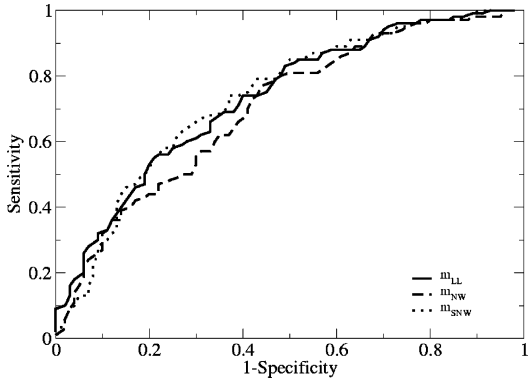


Figure 6. Receiver operating characteristic curves for the best kernel regression models.

The stepwise approach to descriptor selection has been shown to be effective, when coupled with nonparametric regression. It is a computationally cheap method, which makes the stepwise algorithm ideal for use with the nonparametric methods, which have large computational overheads. The use of an internal monitoring set prevents the inclusion of descriptors that produce overfitted models, and the external validation set gives a reliable measure of model quality.

The SNPR models generated have been validated using internal and external holdout sets and in terms of ability to classify a separate set of test compounds. There is good agreement between internal validation and external validation for the additive models, which are both effective as classifiers for the test set. The additive models have the further advantage of being readily interpretable. The genuine multivariate, Nadaraya−Watson kernel is less consistent, with large differences between internal and external validation, but with the highest internal and external validation Q^2. This suggests overfitting of training data or problems with data sparsity in high dimensions. The best Nadaraya−Watson model does still succeed as a classifier and finds the majority of active compounds with a high hit rate.

The kernel regression models are superior to multilinear regression, suggesting that underlying trends are nonlinear. This is confirmed by additive models, which demonstrate the nonlinearity of the structure activity relationships for the selected variables. In this case, nonparametric regression models are more predictive than a neural network approach and also more readily interpreted. The speed and accuracy of SNPR predictions of the activity of compounds, augurs well for its future as a QSAR tool for focusing high throughput screening efforts in drug discovery programs.

ACKNOWLEDGMENT

We thank EPSRC for a studentship and a JREI equipment grant (R/62052/01) for computers. We thank Novartis for financial support and the data sets. We also thank Peter Gedeck and Trevor Howe from Novartis and James Melville for helpful discussions.

REFERENCES AND NOTES

(1) Williams, D. H.; Mitchell, T. Latest developments in crystallography and structure-based design of protein kinase inhibitors as drug candidates. *Curr. Opin. Pharmacol.* **2002**, *2*, 567−573.

(2) Chu, D. H.; Morita, C. T.; Weiss, A. The Syk family of protein tyrosine kinases in T-cell activation and development. *Immunol. Rev.* **1998**, *165*, 167−180.

(3) Turner, M.; Schweighoffer, E.; Colucci, F.; Di Santo, J. P.; Tybulewicz, V. L. Tyrosine kinase SYK: essential functions for immunoreceptor signaling. *Immunol. Today* **2000**, *21*, 148−154.

(4) Niimi, T.; Orita, M.; Okazawa-Igarashi, M.; Sakashita, H.; Kikuchi, K.; Ball, E.; Ichikawa, A.; Yamagiwa, Y.; Sakamoto, S.; Tanaka, A.; Tsukamoto, S.; Fujita, S.; Tatsuta, K.; Maeda, Y.; Chikauchi, K. Design and synthesis of nonpeptidic inhibitors for the Syk C-terminal SH2 domain based on structure-based in-silico screening. *J. Med. Chem.* **2001**, *44*, 4737−4740.

(5) Watson, S. P.; Gibbins, J. Collagen receptor signaling in platelets: extending the role of the ITAM. *Immunol. Today* **1998**, *19*, 260−264.

(6) Hubbard, S. R. Protein tyrosine kinases: autoregulation and small-molecule inhibition. *Curr. Opin. Struct. Biol.* **2002**, *12*, 735−741.

(7) Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Structural mechanism for STI-571 inhibition of Abelson tyrosine kinase. *Science* **2000**, *289*, 1938−1942.

(8) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.

(9) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: Identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356−2364.

(10) Macchiarulo, A.; De Luca, L.; Costantino, G.; Barreca, M. L.; Gitto, R.; Pellicciari, R.; Chimirri, A. QSAR study of anticonvulsant negative allosteric modulators of the AMPA receptor. *J. Med. Chem.* **2004**, *47*, 1860−1863.

(11) Lapinsh, M.; Prusis, P.; Mutule, I.; Mutulis, F.; Wikberg, J. E. S. QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. *J. Med. Chem.* **2003**, *46*, 2572−2579.

(12) Shen, M.; Xiao, Y. D.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **2003**, *46*, 3013−3020.

(13) Austin, R. P.; Barton, P.; Bonnert, R. V.; Brown, R. C.; Cage, P. A.; Cheshire, D. R.; Davis, A. M.; Dougall, I. G.; Ince, F.; Pairaudeau, G.; Young, A. QSAR and the rational design of long-acting dual D-2-receptor/beta(2)-adrenoceptor agonists. *J. Med. Chem.* **2003**, *46*, 3210−3220.

(14) Topliss, J. G.; Edwards, R. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(15) Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Toward generating simpler QSAR models: Nonlinear multivariate regression versus several neural network ensembles and some related methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.

(16) Draper, N. R.; Smith, H. *Applied Regression Analysis*, 3rd ed.; Wiley: New York, 1998.

(17) Hirst, J. D. Nonlinear quantitative structure−activity relationship for the inhibition of dihydrofolate reductase by pyrimidines. *J. Med. Chem.* **1996**, *39*, 3526−3532.

(18) Constans, P.; Hirst, J. D. Nonparametric regression applied to quantitative structure−activity relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452−459.

(19) Hirst, J. D.; McNeany, T. J.; Howe, T.; Whitehead, L. Application of nonparametric regression to quantitative structure−activity relationships. *Bioorg. Med. Chem.* **2002**, *10*, 1037−1041.

(20) *Chemometric Methods in Drug Design*; van der Waterbeemd, H., Ed.; VCH: Weinheim, 1995.

(21) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative Structure−Activity-Relationships by Neural Networks and Inductive Logic Programming. 1. The Inhibition of Dihydrofolate-Reductase by Pyrimidines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 405−420.

(22) Aoyama, T.; Ichikawa, H. Neural Networks as Nonlinear Structure Activity Relationship Analyzers − Useful Functions of the Partial Derivative Method in Multilayer Neural Networks. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 492−500.

(23) Niculescu, S. P. Artificial neural networks and genetic algorithms in QSAR. *Theochem-J. Mol. Struct.* **2003**, *622*, 71−83.

(24) So, S. S.; Richards, W. G. Application of Neural Networks − Quantitative Structure-Activity-Relationships of the Derivatives of 2,4-Diamino-5- (Substituted-Benzyl)Pyrimidines as Dhfr Inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.

(25) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.

(26) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR study of ethyl 2-(3-methyl-2,5-dioxo(3-pyrrolinyl))amino -4-(trifluoromethyl) pyrimidine-5-carboxylate: An inhibitor of AP-I and NF-kappa B mediated gene expression based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288−1296.

(27) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048−2056.

(28) NguyenCong, V.; VanDang, G.; Rode, B. M. Using multivariate adaptive regression splines to QSAR studies of dihydroartemisinin derivatives. *Eur. J. Med. Chem.* **1996**, *31*, 797−803.

(29) Ren, S. J.; Kim, H. Comparative assessment of multiresponse regression methods for predicting the mechanisms of toxic action of phenols. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2106−2110.

(30) Ren, S. J. Modeling the toxicity of aromatic compounds to Tetrahymena pyriformis: The response surface methodology with nonlinear methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1679−1687.

(31) Lahsen, J.; Schmidhammer, H.; Rode, B. M. Structure−activity relationship study of nonpeptide delta- opioid receptor ligands. *Helv. Chim. Acta* **2001**, *84*, 3299−3305.

(32) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity-Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(33) Cho, S. J.; Hermsmeier, M. A. Genetic algorithm guided selection: Variable selection and subset selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927−936.
(34) Nicolotti, O.; Gillet, V. J.; Fleming, P. J.; Green, D. V. S. Multiobjective optimization in quantitative structure−activity relationships: Deriving accurate and interpretable QSARs. *J. Med. Chem.* **2002**, *45*, 5069−5080.
(35) Agrafiotis, D. K.; Cedeno, W. Feature selection for structure−activity correlation using binary particle swarms. *J. Med. Chem.* **2002**, *45*, 1098−1107.
(36) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.
(37) Metropolis, N.; Rosenbluth, A. W.; Teller, A. H.; Rosenbluth, M. N.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(38) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: New York, 1989.
(39) Collingwood, P. S.; Hayler, J.; LeGrand, D. M.; Mattes, H.; Menear, K. A.; Walker, C. V.; Cockroft, X.-L. *Purine Derivatives Inhibitors of Tyrosine Protein Kinase Syk*; Novartis AG, Basel: United States, 2003.
(40) *Cerius2*; Molecular Simulations Inc.: San Diego, CA.
(41) Nadaraya, E. A. *On Estimating Regression*. *Theory of Probability and Its Applications;* 1964; Vol. 10, pp 186−190.
(42) Watson, G. S. Smooth regression analysis. *Sankya-The Indian J. Statistics Ser. A* **1964**, *26*, 359−372.
(43) Hardle, W.; Marron, J. S. Fast and Simple Scatterplot Smoothing. *Computational Statistics Data Analysis* **1995**, *20*, 1−17.
(44) Mammen, E.; Marron, J. S. Mass recentered kernel smoothers. *Biometrika* **1997**, *84*, 765−777.
(45) Stone, C. J. Consistent Non-Parametric Regression. *Ann. Stat.* **1977**, *5*, 595.
(46) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman and Hall: New York, 1990.
(47) Buja, A.; Hastie, T.; Tibshirani, R. Linear Smoothers and Additive-Models. *Ann. Stat.* **1989**, *17*, 453−510.
(48) Churchwell, C. J.; Rintoul, M. D.; Martin, S.; Visco, D. P.; Kotu, A.; Larson, R. S.; Sillerud, L. O.; Brown, D. C.; Faulon, J. L. The signature molecular descriptor − 3. Inverse-quantitative structure−activity relationship of ICAM-1 inhibitory peptides. *J. Mol. Graph.* **2004**, *22*, 263−273.
(49) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. An accurate QSPR study of O−H bond dissociation energy in substituted phenols based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 669−677.
(50) Roy, K.; Leonard, J. T. QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino- 6-arylsulfonylbenzonitriles and congeners using molecular connectivity and E-state parameters. *Bioorg. Med. Chem.* **2004**, *12*, 745−754.
(51) Netzeva, T. I.; Dearden, J. C.; Edwards, R.; Worgan, A. D. P.; Cronin, M. T. D. QSAR analysis of the toxicity of aromatic compounds to Chlorella vulgaris in a novel short-term assay. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 258−265.
(52) Moon, T.; Song, J. S.; Lee, J. K.; Yoon, C. N. QSAR analysis of SH2-binding phosphopeptides: Using interaction energies and cross-correlation coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1570−1575.
(53) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(54) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.
(55) Faber, N. K. M. Estimating the uncertainty in estimates of root-mean-square error of prediction: application to determining the size of an adequate test set in multivariate calibration. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 79−89.
(56) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(57) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley & Sons: New York, 1986.
(58) Hall, L. H.; Kier, L. B. The Kappa Indices for Modeling Molecular Shape and Flexibility. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Reading U.K., 1999; pp 307−360.
(59) Fahlman, S. E.; Lebiere, C. The Cascade-Correlation Learning Architecture. *Advances in Neural Information Processing Systems*; Kaufmann Publishers: 1990; pp 524−532.
(60) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Applications*; Morgan Kaufmann: San Francisco, 1999.

CI049631T