Nonparametric Regression Applied to Quantitative Structure−Activity Relationships

Pere Constans[†] and Jonathan D. Hirst*[,‡]

Department of Molecular Biology, TPC-6, The Scripps Research Institute, 10550 North Torrey Pines Road,
La Jolla, California 92037

Received August 23, 1999

Several nonparametric regressors have been applied to modeling quantitative structure−activity relationship (QSAR) data. The simplest regressor, the Nadaraya−Watson, was assessed in a genuine multivariate setting. Other regressors, the local linear and the shifted Nadaraya−Watson, were implemented within additive models—a computationally more expedient approach, better suited for low-density designs. Performances were benchmarked against the nonlinear method of smoothing splines. A linear reference point was provided by multilinear regression (MLR). Variable selection was explored using systematic combinations of different variables and combinations of principal components. For the data set examined, 47 inhibitors of dopamine β-hydroxylase, the additive nonparametric regressors have greater predictive accuracy (as measured by the mean absolute error of the predictions or the Pearson correlation in cross-validation trails) than MLR. The use of principal components did not improve the performance of the nonparametric regressors over use of the original descriptors, since the original descriptors are not strongly correlated. It remains to be seen if the nonparametric regressors can be successfully coupled with better variable selection and dimensionality reduction in the context of high-dimensional QSARs.

INTRODUCTION

Linear modeling of theoretical and experimental data includes a series of well-established and relatively simple techniques. These techniques have proven successful in a large variety of problems. In the development of quantitative structure−activity relationships (QSAR), methodologies based on ordinary multilinear regression, principal component analysis (PCA),[1] and partial least squares regression (PLS)[2] are well-known. Whenever QSARs are linear or nearly linear, and can be identified as such, the above methodologies, or variants, provide simple, interpretable, and robust modeling of data.

Parametric models are simple and consequently easy to understand and computationally straightforward to compute. But, if the wrong functional form is chosen, then the model will probably be highly biased, possibly to the point of uselessness. In contrast, a nonparametric model involves nearly no a priori assumptions concerning the underlying functional form. Such a model is much more flexible, due to its local nature, and is capable of capturing subtle and not so subtle nonlinear relationships. However, until recently nonparametric models have not been as widely applied as parametric models because of their associated computational cost and their greater statistical complexity. Ever-increasing computer power has started to change this. Moving beyond parametric linear and parabolic models to less restricted and potentially more powerful nonlinear QSARs is now an active area of research.[3−5] Neural networks have received much attention in this regard.[6−17] Other techniques have also been applied to QSAR, including machine learning methods,[18−21] interpolation,[22] genetic algorithm PLS (GAPLS),[23−25] and methods rooted in nonparametric statistics, such as smoothing splines.[26−28] It is probably fair to say that these methods are more difficult to apply to QSAR than their linear counterparts. Their statistical foundations are under continuous development, and difficulties arise from the need for large samples for a given level of confidence and from instabilities on irregular designs (data sets).

The present study is intended to contribute further analysis on the application of nonparametric techniques to QSAR. Several nonlinear methods, nonparametric kernel regression,[29−31] and smoothing splines[32] are comparatively applied to QSAR problems. Analyses using multilinear regression (MLR) provide linear reference points in our study.

NONPARAMETRIC KERNEL REGRESSORS

Nonparametric regressors model data without specific assumptions about the kind of functional dependencies underlying in the model. An overview of such methods is provided by several recent monographs,[29−31] and, in the chemical literature, by a succinct review of their application in analytical chemistry.[33] Statistically, the regression function m of a response variable Y in measuring variables X is defined as the conditional expectation of Y on X, at a given point x, by

$$\begin{aligned} m(x) &\equiv E(Y|X=x) \\ &= \int y f(y|x)\, \mathrm{d}y \\ &= \int y \frac{f(x,y)}{f_X(x)}\, \mathrm{d}y \end{aligned} \tag{1}$$

* To whom correspondence should be addressed. Telephone: (619) 784 9290. FAX: (619) 784 8688. E-mail: jhirst@scripps.edu.

[†] Present address: Department of Chemistry, MS-60, Rice University, Houston, TX 77005-1892.

[‡] Present address: School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom.

10.1021/ci990082e CCC: $19.00
Published on Web 03/09/2000

where $f(y|x)$, $f(y,x)$, and $f_X(x)$ are the conditional density of Y given X, the joint density of X and Y, and the marginal density of X, respectively.[31] Nadaraya[34] and Watson,[35] independently, applied the Parzen kernel estimate for probability density functions[36] to derive estimates of mean regression curves, $\hat{m}$. Accordingly, for a sample of N measures $\{x\}_{i=1}^{N}$ and its associated responses $\{y\}_{i=1}^{N}$, estimates of the joint density of X and Y, $\hat{f}(x,y)$, and the marginal density of X, $\hat{f}_X(x)$, are written as

$$\hat{f}(x,y) = \frac{1}{Nh_x h_y}\sum_{i=1}^{N} K_x\left(\frac{x - x_i}{h_x}\right) K_y\left(\frac{y - y_i}{h_y}\right) \tag{2}$$

$$\hat{f}_X(x) = \frac{1}{Nh_x}\sum_{i=1}^{N} K_x\left(\frac{x - x_i}{h_x}\right) \tag{3}$$

respectively. The smoothing kernel $K_h(u) = h^{-1}K(u/h)$, where h is a bandwidth, is often taken as a Gaussian density,

$$K(u/h) = (2\pi)^{-1/2}\mathrm{e}^{-u^2/2h^2} \tag{4}$$

Substituting eqs 2 and 3 into the definition of a regression curve (eq 1) yields, after integration, the *Nadaraya−Watson kernel estimator*,

$$\hat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^{N} K_x\left(\frac{x - x_i}{h_x}\right) y_i}{\sum_{i=1}^{N} K_x\left(\frac{x - x_i}{h_x}\right)} \tag{5}$$

The form of the above equation is simply a weighted sum of responses. In general, nonparametric estimates $\hat{m}$ of a function m appear as a linear combination of the measured responses y_i,

$$\hat{m}(x) = \sum_i w_i(x) y_i \tag{6}$$

with the weights $w_i(x)$ being positive and their sum equal to unity. In this vein, and in the context of QSAR, the simple and purely nonparametric estimate, with the weights defined

$$w_i(x) = \frac{d_i(x)^{-1}}{\sum_j d_j(x)^{-1}} \tag{7}$$

where d_i are Euclidean distances $\|x - x_i\|$, has recently been applied.[22] However, despite its appealing simplicity, this approach does not appear to be as accurate as the approaches we discuss in this paper, so we do not discuss it further.

In kernel regression, the parameter h, the *bandwidth* or *smoothing parameter*, plays an essential role, since it determines the degree of locality in the regression. Too low a value leads to highly rugged surfaces and thus sample dependence and a regression curve with high variance. In other words, an overly low bandwidth makes the smooth overly local and places too much emphasis on the idiosyncrasies of the sample. So, for a different set of measures, one would obtain a quite different regression curve estimate. On the other hand, an overly high bandwidth will introduce bias in the curve estimation and eliminate the fine structure of data. If one thinks of smoothing in terms of nearest neighbor methods, as computing the response Y at a given location X based on a weighted sum of the responses of neighbors, then the smoothing kernel is the function that governs the precise weight accorded to neighboring data points. Generally, the more distant the neighbor, the less weight it should be given.

For the nonparametric regressors discussed in this paper, the only explicit parameter in the methods is the bandwidth. However, while activity is modeled as a function of (perhaps relatively few) molecular descriptors, the functional form is based on N data points, the number of molecules in the training set. For a single descriptor, two parameters, the gradient and the intercept, would describe a linear model, and three parameters would describe a parabolic model. The maximum likelihood values for these parameters are usually estimated by a least-squares fitting to the training data, and the confidence intervals for these values depend on the number of parameters considered. In contrast, nonparametric regressors are built up using the N measurement values and the N associated responses. The maximum likelihood model, or curve, is established once an optimal bandwidth is determined. Such determination does not come from least-squares fitting, as discussed below, but from a tradeoff between “fitting” or bias minimization, and smoothing, or variance minimization. Consequently, in a nonparametric setting one should not think of a model as being “overfitted” using $2N$ parameters, since only the bandwidth is inferred, but not fitted. If the resulting model is simple enough, this might build confidence in the applicability of a given parametric approach. Thus, nonparametric regression is a valuable tool that can provide insight into the data and an unbiased procedure to determine the functional form of models.

The accuracy or closeness of the estimator $\hat{m}_h$ to the mean regression curve m is commonly measured as the mean squared error (MSE), or the expectation, E, of the squared error of the regressor,

$$\mathrm{MSE}(\hat{m}_h) = E(\hat{m}_h - m)^2 \tag{8}$$

As in classical parametric statistics, the MSE may be expressed in terms of the variance and squared bias

$$\mathrm{MSE}(\hat{m}_h) = \mathrm{Var}(\hat{m}_h) + (E(\hat{m}_h) - m)^2 \tag{9}$$

The influence of the bandwidth on the MSE and the rate of convergence of the estimator are readily derived with the assumption of a large sample size or *asymptotic approximation*. The asymptotic variance for the Nadaraya−Watson estimate is approximated, at a point x, by[37]

$$\mathrm{Var}(x)_{\mathrm{NW}} = h^{-1}N^{-1}\frac{\sigma^2(x)}{f_X(x)}\int K^2(u)\,\mathrm{d}u \tag{10}$$

where $\sigma^2(x)$ is the residual variance function, $E(Y^2|x) - m^2(x)$. The asymptotic bias, at a point x, is approximated by

$$E\hat{m}_{\mathrm{NW}}(x) - m(x) = \frac{h^2}{2}\left[m''(x) + 2\frac{m'(x)\,f_X'(x)}{f_X(x)}\right]\int u^2 K(u)\,\mathrm{d}u \tag{11}$$

These leading terms for the variance and bias quantify the influence of the bandwidth on the resulting model and may also be taken to determine asymptotic estimates for optimal bandwidths.

The pioneering work of Nadaraya and Watson remained virtually forgotten until the discovery of higher order nonparametric regressors. These regressors, by reducing the bias, improve the rate of convergence (with sample size) toward the mean curve. The higher order nonparametric regressors appear to be more stable in irregular designs (i.e., poorly distributed data) and the odd-degree higher order regressors also overcome boundary problems (i.e. problems arising at the ends of the range of the data). In light of this finding, the Nadaraya−Watson regressor can be seen as the zero order or local-constant estimate of a family of regressors based on local polynomial fitting. The first-order regressor, the local linear (LL),[38−41] behaves acceptably on irregular designs and includes, like all odd-degree local polynomials, boundary correction.[30] Approximating the regression function, locally, by a line, produces the following closed expression for the regressor:

$$\hat{m}_{\mathrm{LL}}(x) = N^{-1}\sum_{i=1}^{N}\frac{\{\hat{s}_2(x) - \hat{s}_1(x)(x_i - x)\}K_h(x_i - x)}{\hat{s}_2(x)\,\hat{s}_0(x) - \hat{s}_1(x)^2}\,y_i \qquad (12)$$

with

$$\hat{s}_r(x) = N^{-1}\sum_{i=1}^{N}(x_i - x)^r\,K_h(x_i - x) \qquad (13)$$

The variance appears to be the same as the Nadaraya−Watson smoother, while the bias is reduced to

$$E\hat{m}_{\mathrm{LL}}(x) - m(x) = {}^1\!/_2 h^2\,m''(x)\int u^2K(u)\,\mathrm{d}u \qquad (14)$$

thus adapting better to nonuniform designs, i.e., designs where $f_X'(x)/f_X(x)$ is large.

Recently, consideration of a mass recentering of the cloud of points led to the shifted Nadaraya−Watson (SNW) regressor, which exhibits the same good asymptotic behavior of LL (see eq 14) but also offers stability in sparse design.[42] The shifted Nadaraya−Watson estimate is given by

$$\hat{m}_{\mathrm{SNW}}(x) = \frac{\sum_{i=1}^{N}K_h(x_i - \xi^{-1}(x))y_i}{\sum_{i=1}^{N}K_h(x_i - \xi^{-1}(x))} \qquad (15)$$

with

$$\xi(x) = \frac{\sum_{i=1}^{N}K_h(x_i - x)x_i}{\sum_{i=1}^{N}K_h(x_i - x)} \qquad (16)$$

In our implementation of the SNW smoother, we compute the reverse mass-recentering function ξ^{-1} by a polynomial fit. This proves to be sufficiently accurate and greatly simplifies the evaluation of $\hat{m}_{\mathrm{SNW}}(x)$.

Practical application of nonparametric regression presents two critical challenges. As already pointed out, bandwidth selection plays an essential role regarding the performance and confidence of the resulting model. Therefore, there is a need for reliable data-driven bandwidth selectors. The second, and more profound, drawback has been called the *curse of dimensionality*,[43] which can be restated as "local" neighborhoods are almost certainly empty in high-dimensional spaces. An attempt to overcome it is provided by additive modeling.

Recently, excellent, optimal bandwidth selectors have been reported.[44,45] The Rupert−Sheather−Wand selector,[44] aside from its cost and complexity, is considered to be the reference one. It belongs to the family of well-established plug-in methods. Using the asymptotic expression for the MSE, or the integrated MSE (MISE) if considering a global bandwidth, one plugs into MSE expressions sample estimates of the unknown $\sigma^2(x)$, $f_X(x)$, and $m''(x)$ (see eqs 9−11). In an iterative procedure, the asymptotically optimal bandwidth is determined. Alternatively, the Hurvich−Simonoff−Tsai global bandwidth selector[45] uses an improved Akaike information criterion[46] (IAIC), which is based on a statistical information criterion instead of an asymptotic approach. It is also computationally demanding, but its applicability is more general, as it does not rely on asymptotic formulas (the assumption of an infinite sample). It also avoids instabilities in nearly linear designs, caused by the second derivative terms in the denominator of asymptotic formulas. In our study we consider the IAIC selector, due to its generality and formal simplicity. We also consider a simplified plug-in selector, called the block method selector,[47] which is fast and suited for additive modeling. First described for the Nadaraya−Watson regressor, its adaptation to LL and SNW is straightforward, by substituting eq 11 by eq 14. The block method selector divides data in contiguous blocks and computes, for each block, parametric estimates of the unknown quantities appearing in the expression of the asymptotically optimal bandwidth. The block method is the simplest and computationally cheapest and appears to be robust when dealing with small samples.[47]

The applicability of nonparametric regression in multidimensional settings may be quite limited, as mentioned above. Additive models attempt to circumvent the curse of dimensionality by generalizing ordinary, multiple linear regression, with linear terms replaced by nonparametric functions,[48,49]

$$m(\mathbf{x}) = \alpha + \sum_{j=1}^{d}f_j(x_j) \qquad (17)$$

where α is a constant. If the regression curve satisfies the additive constraint, one may determine the estimator $\hat{m}(\mathbf{x})$ with a convergence rate of the one-dimensional smoother.[50] In other words, if cross-terms are not important, then an *N*-dimensional problem may be reduced to *N* one-dimensional smooths rather than one *N*-dimensional smooth. A standard way to estimate each term $f_j(x_j)$ is the *back-fitting* algorithm.[48,49] This is an iterative procedure, where one uses the partial residuals in each step to get estimates of the function. The back-fitting often converges but may be time-consuming due to the necessity of smoothing at each step, and its performance is affected by highly correlated variables. Other procedures for generating additive models, such as the

Table 1. Molecules Used in This Study (Index Number, Substituents, Activity)

no.	R	pIC_{50}	no.	R	pIC_{50}	no.	R	pIC_{50}
1	2,6-$(CH_3)_2$	3.00	17	3-NO_2, 4-OCH_3	3.45	33	H	4.48
2	2,6-Cl_2	3.15	18	4-OCH_3	3.69	34	3-NO_2, 4-OH	4.51
3	2,6-$(OCH_3)_2$	3.30	19	3-OCH_3	3.80	35	3,4-Cl_2	4.55
4	2-Cl	3.45	20	3-OH	3.83	36	2,4-Cl_2	4.77
5	2-CH_3	3.47	21	3-CF_3, 4-OH	3.92	37	3-Br, 4-OH	4.92
6	3,4-$(OCH_3)_2$	3.47	22	2,4,6-Cl_3	3.99	38	3-Cl	4.92
7	4-CF_3	3.70	23	2,5-Cl_2	4.01	39	3-F	5.25
8	3-CF_3, 4-OCH_3	3.76	24	4-Cl	4.02	40	4-OH	5.59
9	2,6-Cl_2, 4-OCH_3	3.81	25	2,6-Cl_2, 4-OH	4.12	41	3,5-Cl_2	5.62
10	4-CH_3	3.83	26	2,3,5,6-F_4, 4-OH	4.21	42	3,4-$(OH)_2$	5.66
11	4-Br	3.94	27	4-NO_2	4.28	43	3-Cl, 4-OH	5.70
12	3-Br, 4-OCH_3	4.08	28	2,3-Cl_2	4.28	44	3-F, 4-OH	5.82
13	3-F, 4-OCH_3	4.13	29	3-CH_3, 4-OH	4.31	45	3,5-F_2	5.92
14	2-OCH_3	4.13	30	4-F	4.33	46	3,5-Cl_2, 4-OH	6.17
15	3-CH_3, 4-OCH_3	4.16	31	3,5-Cl_2, 4-OCH_3	4.33	47	3,5-F_2, 4-OH	7.13
16	2-OH	3.24	32	3,5-F_2, 4-OCH_3	4.44			

integration method,[51,52] are an active area of research. However, in the present study only the most widely known back-fitting algorithm has been considered.

METHODS

The kernel smoothers LL and SNW described above, in conjunction with the block bandwidth selector, in an additive modeling framework, are tested. In addition, results using a genuine multivariate regressor, the NW with a single bandwidth selected according to the IAIC, are presented. Despite the theoretical difficulties mentioned above, multivariate NW regressors are appealing for their simplicity and, thus, could have some applicability in coarse screenings of data. Smoothing splines (MARS)[32] have also been considered, since they are an established reference in nonparametric regression and there is already some experience in their application to QSAR.[26−28] This technique models data by introducing an undersmooth penalty in fitting splines, thus allowing a trade-off between bias and variance. An additional advantage is the possible inclusion of several degrees of interaction among variables.[32] We present results coming from purely additive splines and from single pair interactions.

In our analyses, variables are standardized with a mean of zero and a variance unity. The standardization was performed by estimating the mean and variance of the training data only, and thus there is no bias in the cross-validation trial (i.e., no information about the test set is introduced, even indirectly). All combinations of variables are systematically considered. As a technique for reduction of dimensionality, we only consider PCA, constructing models for all 31 possible combinations of the first five principal components. The potential importance of considering principal components with small eigenvalues has been remarked on.[53] While the study of the combinations of descriptors is a means for variable selection, it is not merely so. It also is a way to test these methodologies in a larger number of cases, thus providing more evidence about the performance and stability of the presented regressors. The problem of variable selection is a significant one, and there is interest in evaluating the robustness of methods on suboptimal selections of variables. Nonparametric approaches for selecting predictive regression exist, such as projection pursuit regression[54] (PPR) and sliced inverse regression,[55] and there is recent work applying PPR to QSAR.[27] However, consideration of these methods is beyond the scope of the present study. As a linear reference, multilinear regression (MLR) results are presented. The performance of MLR might be further enhanced by using an external selection of variables, as in the enhancement of PLS by GAPLS.[23−25]

The data set we examine is 47 1-(substituted benzyl)-imidazole-2(3*H*)-thiones with inhibitory activity against dopamine β-hydroxylase (shown in Table 1). These molecules can reduce blood pressure and are a treatment for cardiovascular disorders related to hypertension. The data have been used in QSAR studies based on linear free energy descriptors,[56] molecular shape analysis,[57] receptor surface models,[58] and genetic neural networks.[59] As such they represent a well-studied data set, comprising a reasonable number of compounds, and are well-suited to nonlinear analyses. We examine two previously employed molecular descriptors, those used in the molecular shape analysis study[57] and the energy terms used in the receptor surface model study.[58] The former descriptors are as follows: a common overlap steric volume, V_0, the composite charge density on carbon atoms 3, 4, and 5 of the substituted-benzyl ring, Q_{345}, the charge density on carbon atom 6, Q_6, the lipophilicty of the whole molecule (a CLOGP value), π_0, and the water/octanol fragment constant of the 4-substituent of the phenyl ring, π_4. The energy descriptors are as follows: a nonbonded interaction energy between the molecule and the receptor surface, E_{interact}, the internal strain energy of the molecule with respect to the surface, E_{inside}, the internal energy of the molecule after relaxation in the absence of the surface, E_{relax}, and E_{strain}, the difference between E_{inside} and E_{relax}. The origins of the descriptors have been discussed elsewhere.[58,60] Activities were reported as $-\log(IC_{50})$ values.

We assess the different nonlinear methods using several statistical measures. We use a cross-validated correlation coefficient, as employed in other recent studies,[59,61] which is defined as

$$q^2 = 1 - \frac{\sum_{i=1}^{N}(y_{i,obsd} - y_{i,pred})^2}{\sum_{i=1}^{N}(y_{i,obsd} - \bar{y}_{i,obsd})^2} \tag{18}$$

We also report mean absolute errors (MAE) and the Spearman rank correlation coefficient (SRCC).[62] The predictive performances of the methods have been assessed using

Table 2. Five Randomly Chosen Sets, with Molecules Referred to by Index Numbers Given in Table 1

set	molecules	set	molecules
1	37, 35, 17, 2, 47, 13, 38, 23, 27, 43	4	1, 36, 12, 22, 15, 26, 34, 40, 42
2	45, 4, 8, 19, 6, 33, 44, 16, 7, 24	5	3, 11, 9, 39, 25, 18, 28, 46, 21
3	31, 41, 14, 5, 30, 10, 32, 29, 20		

Table 3. Training and Test Sets for Cross-Validation Trials

trial	training sets	test set	trial	training sets	test set
1	2, 3, 4, 5	1	4	1, 2, 3, 5	4
2	1, 3, 4, 5	2	5	1, 2, 3, 4	5
3	1, 2, 4, 5	3			

Table 4. Performance of Nonparametric Methods and Multilinear Regression (MLR) on the Dopamine Inhibitors Data Set with Molecular Shape Descriptors

method	model	q^2	SRCC	MAE	r^2	SRCC (train)
multivariate NW	V_0	0.57	0.64	0.45	0.66	0.74
	Q_6, π_0, V_0	0.57	0.68	0.47	0.76	0.80
	$Q_{345}, Q_6, \pi_0, \pi_4, V_0$	0.55	0.76	0.49	0.80	0.89
local linear	Q_{345}, π_0, V_0	0.69	0.77	0.38	0.87	0.88
	Q_6, π_0, V_0	0.70	0.68	0.41	0.88	0.85
shifted NW	Q_{345}, π_0, V_0	0.60	0.79	0.43	0.85	0.88
	Q_6, π_0, V_0	0.75	0.77	0.36	0.86	0.85
MARS additive	Q_6, π_0, V_0	0.72	0.70	0.38	0.80	0.76
	$Q_{345}, \pi_0, \pi_4, V_0$	0.62	0.73	0.41	0.74	0.77
MARS interaction	Q_{345}, π_0	0.66	0.67	0.43	0.72	0.58
	Q_{345}, Q_6, π_0, V_0	0.64	0.72	0.40	0.74	0.60
	$Q_{345}, \pi_0, \pi_4, V_0$	0.63	0.75	0.41	0.84	0.79
MLR	$Q_{345}, \pi_0, \pi_4, V_0$	0.59	0.83	0.43	0.71	0.86

cross-validation trials. We have also in some cases used a leave-one-out (LOO) jackknife procedure, and comparable results to the cross-validation were obtained. As the LOO procedure can sometimes overestimate the true predictiveness, we only report the cross-validation results. In the cross-validation, the data were randomly divided into five independent sets (given in Table 2), and five trials were performed; each set was used as a test set in turn, and the remaining four sets were used as training data (see Table 3). The activity of every molecule is thus predicted once, and once only. Some techniques use cross-validation to determine meta−parameters; e.g. PLS uses cross-validation to determine the number of descriptors (or rank) that gives the best model. Hence, there is a perception that cross-validation is not a genuine test. In our case, however, cross-validation is not used to establish any meta−parameters and is exactly a training/test set protocol repeated over five trials. For completeness we also report some measures of training set performance, although they have no direct bearing on the predictiveness of the models.

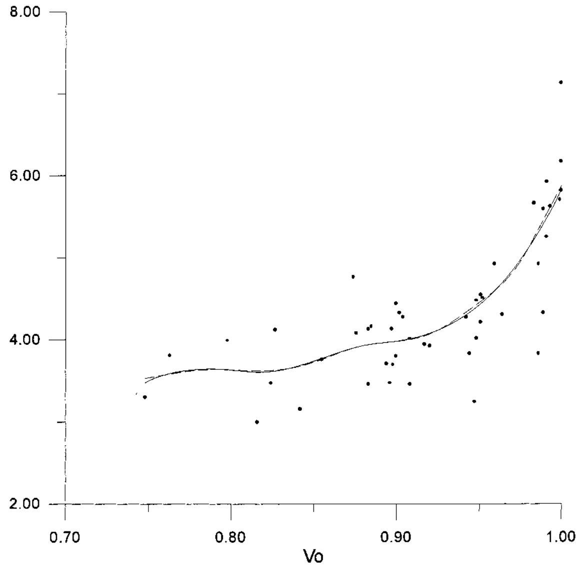

Figure 1. One-dimensional models of activity as a function of V_0, constructed using the local linear (solid line) and the shifted Nadaraya−Watson regressors.

Table 5. Performance of Nonparametric Methods and MLR on the Dopamine Inhibitors Data Set with Principal Components Derived from Molecular Shape Descriptors

method	model (principal components)	q^2	SRCC	MAE	r^2	SRCC (train)
multivariate NW	1, 4	0.59	0.74	0.43	0.79	0.86
local linear	1, 3, 4	0.73	0.80	0.38	0.82	0.84
shifted NW	1, 4	0.73	0.75	0.39	0.78	0.83
	1, 2, 3, 4	0.68	0.82	0.40	0.86	0.88
MARS additive	1, 2, 4	0.67	0.68	0.41	0.83	0.85
	1, 3, 4	0.67	0.70	0.42	0.77	0.77
MARS interaction	1, 2, 3,4	0.74	0.78	0.34	0.83	0.86
MLR	1, 3, 4	0.64	0.81	0.43	0.66	0.79
	1, 2, 3, 4	0.55	0.83	0.46	0.71	0.85

RESULTS

Table 4 shows the best models of the dopamine inhibitor data generated by each of the methods examined in this study. Up to three models are shown per method, with "best" defined with respect to each of the three predictive measures in turn, q^2, Spearman rank correlation coefficient, and the mean absolute error. In some cases, one model was best with respect to more than one of the measures. As described earlier, all possible combinations of the five descriptors were analyzed, i.e., five one-descriptor models, 10 two-descriptor models, 10 three-descriptor models, five four-descriptor models, and a model using all five descriptors, giving a total of 31 models. For illustrative purposes, the local linear and SNW models of activity as a function of V_0 are shown in Figure 1. The multivariate NW performs worse than the additive models or the smoothing splines. Aside from MNW, the nonparametric methods perform better than MLR with respect to q^2 and MAE, but MLR gives the best SRCC. Qualitatively, the robustness of the different methods may be reflected in the number of statistically significant models that a method is able to generate out of the possible 31 combinations of descriptors. A model may be deemed significant if $q^2 > 0.5$. In this sense, additive MARS (16 out of 31) is the most robust and LL (4 out of 31) is the least robust.

Using principal components instead of the original variables (see Table 5) does not change the overall qualitative picture. The multivariate NW remains poorer than the other nonparametric methods. Again, the nonparametric methods appear better than MLR with respect to q^2 and MAE. The difference in SRCC is narrowed, but MLR is still marginally better. Overall less models with $q^2 > 0.5$ are generated, perhaps reflecting a concentration of information resulting from the principal component analysis.

For the descriptors based on four energy terms, all 15 possible models were constructed. The best of these models, as defined previously, for each method are shown in Table 6. The local linear and SNW models of activity as a function

Table 6. Performance of Nonparametric Methods and MLR on the Dopamine Inhibitors Data Set with Energy Descriptors

method	model	q^2	SRCC	MAE	r^2	SRCC (train)
multivariate NW	E_{interact}	0.59	0.71	0.47	0.71	0.79
	E_{interact}, E_{relax}	0.58	0.68	0.46	0.73	0.79
local linear	E_{interact}	0.66	0.74	0.44	0.73	0.78
	E_{inside}, E_{interact}	0.71	0.71	0.39	0.83	0.87
shifted NW	E_{interact}, E_{relax}	0.68	0.72	0.41	0.70	0.80
	E_{interact}, E_{relax}, E_{strain}	0.66	0.76	0.42	0.80	0.85
MARS additive	E_{interact}	0.69	0.75	0.40	0.73	0.79
	E_{interact}, E_{relax}	0.70	0.72	0.39	0.77	0.78
MARS interaction	E_{interact}	0.69	0.75	0.40	0.73	0.79
	E_{interact}, E_{relax}	0.73	0.73	0.37	0.73	0.78
MLR	E_{inside}, E_{interact}, E_{relax}	0.45	0.72	0.54	0.55	0.75

of E_{interact} are shown in Figure 2, and a two-dimensional local linear model is shown in Figure 3. As for the other set of descriptors, q^2 and the MAE both show that the nonparametric methods offer a distinct improvement, and they are also a little better with respect to the SRCC. Multivariate NW again appears weaker than the other methods. A similar pattern to the one before is seen for the robustness of the methods, with additive MARS (9 out of 15) being the most robust and LL (4 out of 15) the least. Regression based on principal components (see Table 7) leads to some loss in predictive ability for the nonparametric methods. As a result of this, MLR performs relatively better as assessed by SRCC, but q^2 and MAE still favor the nonparametric methods.

DISCUSSION AND CONCLUSIONS

There are several studies in the literature on the dopamine inhibitors. A molecular shape analysis has been used to generate a regression equation involving linear and quadratic terms of the previously discussed properties V_0, Q_{345}, Q_6, π_0, and π_4.[57] So and Karplus[59] estimated the q^2 of this earlier regression equation to be 0.76. A receptor surface model based on the previously described energy terms E_{interact}, E_{inside}, E_{relax}, and E_{strain} has been generated[58] using splines and a variable selection technique called genetic function approximation. A q^2 of 0.67 (described as a fully cross-validated correlation coefficient) was reported. A higher value of 0.79 was reported for a regression-only cross-validated correlation coefficient. Presumably, the former is the more pertinent value and would seem to indicate the importance of genuine cross-validation. Molecular similarity matrices and genetic neural networks have been reported to give a model with a q^2 of 0.77.[59] However, this is apparently

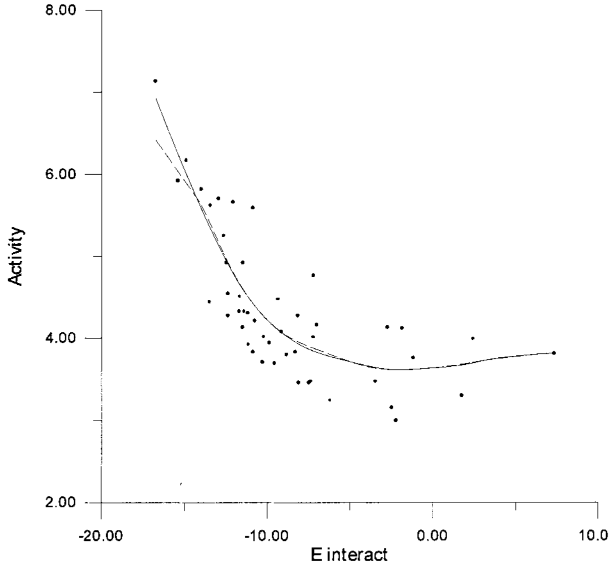

Figure 2. One-dimensional models of activity as a function E_{interact}, constructed using the local linear (solid line) and the shifted Nadaraya−Watson regressors.

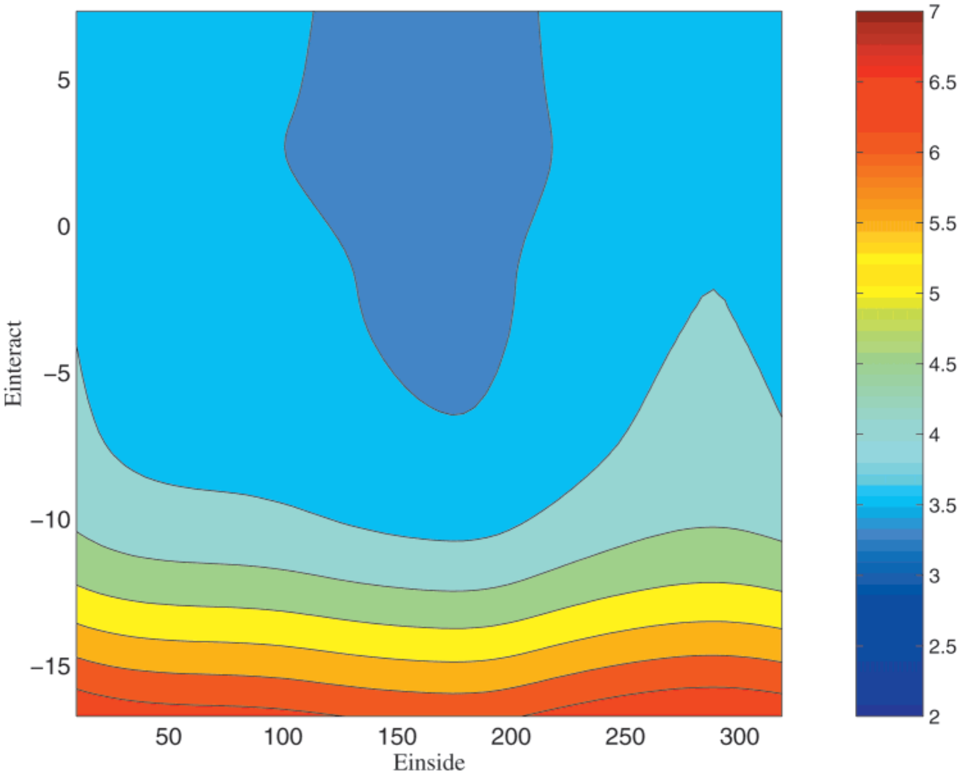

Figure 3. Two-dimensional local linear smooth of activity as a function of E_{inside} and E_{interact}.

Table 7. Performance of Nonparametric Methods and MLR on the Dopamine Inhibitors Data Set with Principal Components Derived from Energy Descriptors

method	model (principal components)	q^2	SRCC	MAE	r^2	SRCC (train)
multivariate NW	1	0.61	0.65	0.44	0.79	0.80
local linear	1	0.64	0.67	0.44	0.75	0.76
shifted NW	1	0.61	0.68	0.46	0.75	0.77
	1, 4	0.63	0.67	0.45	0.78	0.77
MARS additive	1, 4	0.68	0.62	0.42	0.72	0.66
MARS interaction	1, 3, 4	0.69	0.63	0.41	0.71	0.65
MLR	1, 2, 3	0.45	0.72	0.54	0.55	0.75

not a genuine q^2,[63] but an estimate used for computational expediency. While it is argued that the number is comparable to a genuine q^2, its reliability might be questionable. Despite significant differences in the above approaches, which preclude a direct comparison with our study, it appears that the nonparametric methods discussed in our study perform comparably with these other nonlinear methods, achieving models with the best q^2 values of 0.75.

In this study, we have attempted to illustrate the applicability of nonparametric methods to QSAR. We have assessed the methods on several data sets. Naturally, more extensive testing would be desirable, but is beyond the scope of the present work. The additive nonparametric methods perform better than MLR, as measured by q^2 and MAE. It is unclear why a similar improvement is not seen in the SRCC. One possibility is that in some cases the activity may be a monotonically increasing or decreasing function of a linear combination of descriptors. In such a case, a linear model would predict the relative ranks of the molecules well, even though the errors in predicted absolute activities could be quite large. Some caution should thus be exercised in the use of the SRCC as the sole measure of the predictive accuracy of QSAR methods.

As might be anticipated, there is no one method of choice. Depending on the data, LL, SNW, and MARS all perform well. The SNW extrapolates with a line, and thus is more stable than LL, particularly in the cross-validation trials, which involve a lot of extrapolations. The more direct multivariate NW appears to be less successful. The curse of dimensionality or the sparsity of data in higher dimensions probably leads to less predictive accuracy. We have explored the issue of variable selection for nonparametric methods, examining combinations of descriptors, principal components, and combinations of principal components. As mentioned in the Introduction, projection pursuit[54] and sliced inverse regression[55] are active lines of research in this area and have their analogues in parametric modeling. For example, the identification of appropriate variables, combinations of variables, and particular nonlinear functional forms has recently been shown to enhance the performance MLR.[64,65] We conclude with the general observation that nonlinear techniques (including nonparametric regression) may be considered to be more general and preferable to linear methods such as MLR and PLS, in that they can model nonlinear data more accurately without imposing assumptions about the explicit functional form of the nonlinearity or relying on assumptions about the error distributions in the measurements. Naturally, their applicability and utility depend on the nature of the particular data set in question, and, in particular, linear data should be treated by established and straightforward linear methods.

ACKNOWLEDGMENT

We thank Novartis Pharma for financial support. P.C. is grateful to the Spanish Ministry of Culture and Education for partial financial support. We thank our colleague Professor Angel Ortiz and Dr. Trevor Howe (Novartis) for useful discussions.

REFERENCES AND NOTES

(1) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.

(2) Cramer, R. D., III. Partial least squares (PLS): Its strengths and limitations. *Perspect. Drug. Discovery Des.* **1993**, *1*, 269−278.

(3) Hirst, J. D. Predicting ligand binding energies. *Curr. Opin. Drug Discovery Dev.* **1998**, *1*, 28−33.

(4) Kubinyi, H. QSAR and 3D QSAR in drug design Part I: Methodology. *Drug Des. Today* **1997**, *2*, 457−467.

(5) Kubinyi, H. QSAR and 3D QSAR in drug design Part 2: Applications and problems. *Drug Des. Today* **1997**, *2*, 538−546.

(6) Andrea, T. A.; Kalayeh, H. Applications of neural networks in quantitative structure−activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(7) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural networks applied to structure−activity relationships. *J. Med. Chem.* **1990**, *33*, 905−908.

(8) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural networks applied to quantitative structure−activity relationship analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.

(9) 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.

(10) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure−activity relationships by neural networks and inductive logic programming II. The inhibition of dihydrofolate reductase by triazines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 421−432.

(11) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure−activity relationships by neural networks and inductive logic programming I. The inhibition of dihydrofolate reductase by pyrimidines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 405−420.

(12) Ajay. On better generalization by combining two or more models: A quantitative structure−activity relationship example using neural networks. *Chem. Intell. Lab. Syst.* **1994**, *24*, 19−30.

(13) Ajay. A unified framework for using neural networks to build QSARs. *J. Med. Chem.* **1993**, *36*, 3565−3571.

(14) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758−3767.

(15) Duprat, A. F.; Huynh, T.; Dreyfus, G. Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of log *P*. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 586−594.

(16) Maddalena, D. J.; Johnston, G. A. R. Prediction of receptor properties and binding affinity of ligands to benzodiazepine/GABAA receptors using artificial neural networks. *J. Med. Chem.* **1995**, *38*, 715−724.

(17) Tetko, I. G.; Tanchuk, V. Y.; Chentsova, N. P.; Antonenko, S. V.; Poda, G. I.; Kukhar, V. P.; Luik, A. I. HIV-1 reverse transcriptase inhibitor design using artificial neural networks. *J. Med. Chem.* **1994**, *37*, 2520−2526.

(18) Bolis, G.; Pace, L. D.; Fabrocini, F. A machine learning approach to computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 617−628.

(19) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. E. Drug design by machine learning: The use of inductive logic programming to model the structure−activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 11322−11326.

(20) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M. J. E. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 438−442.

(21) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: A shape-based machine learning tool for drug design. *J. Comput.-Aided Drug Des.* **1994**, *8*, 635−652.

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

(23) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306−310.

(24) Hasegawa, K.; Funatsu, K. GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *J. Mol. Struct.* (*THEOCHEM*) **1998**, *425*, 255−262.

(25) Cho, S. J.; Zheng, W.; Tropsha, A. Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259−268.

(26) Nguyen-Cong, V.; van Dang, G.; Rode, B. M. Using multivariate adaptive regression splines to QSAR studies of dihydroartemisinin derivatives. *Eur. J. Med. Chem.* **1996**, *31*, 797−803.

(27) Nguyen-Cong, V.; Rode, B. M. Quantitative electronic structure−activity relationships of pyridinium cephalosporins using nonparametric regression methods. *Eur. J. Med. Chem.* **1996**, *31*, 479−484.

(28) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.

(29) Wand, M. P.; Jones, M. C. *Kernel Smoothing*; Chapman & Hall: New York, 1995.

(30) Fan, J.; Gijbels, I. *Local Polynomial Modeling and Its Application-Theory and Methodologies*; Chapman and Hall: New York, 1996.

(31) Simonoff, J. S. *Smoothing Methods in Statistics*; Springer-Verlag: Berlin, 1996.

(32) Friedman, J. H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1−141.

(33) Sekulic, S.; Seasholtz, M. B.; Kowalski, B. R.; Lee, S. E.; Holt, B. R. Nonlinear multivariate calibration methods in analytical chemistry. *Anal. Chem.* **1993**, *65*, 835A−845A.

(34) Nadaraya, E. A. On estimating regression. *Theory. Probability Its Appl.* **1964**, *10*, 186−190.

(35) Watson, G. S. Smooth regression analysis. *Sankhya, Ser. A* **1964**, *26*, 359−372.

(36) Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065−1076.

(37) Fan, J. Design-adaptive nonparametric regression. *J. Am. Stat. Assoc.* **1992**, *87*, 998−1004.

(38) Stone, C. J. Consistent nonparametric regression. *Ann. Stat.* **1977**, *5*, 595−645.

(39) Stone, C. J. Optimal rates of convergence for nonparametric estimators. *Ann. Stat.* **1980**, *8*, 1348−1360.

(40) Stone, C. J. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **1982**, *10*, 1040−1053.

(41) Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829−836.

(42) Mammen, E.; Marron, J. S. Mass recentered kernel smoother. *Biometrika* **1997**, *84*, 765−777.

(43) Bellman, R. E. *Adaptive Control Processes*; Princeton University Press: Princeton, NJ, 1961.

(44) Rupert, D.; Sheather, S. J.; Wand, M. P. An effective bandwidth selector for local least squares regression. *J. Am. Stat. Soc.* **1995**, *90*, 1257−1270.

(45) Hurvich, C. M.; Simonoff, J. S.; Tsai, C.-L. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *J. R. Stat. Soc., Ser. B* **1998**, *60*, 271−293.

(46) Akaike, H. *Information theory and an extension of the maximum likelihood principle*; Pterov, B. N., Csaki, F., Eds.; Akademia Kiado: Budapest, 1973; pp 267−281.

(47) Hardle, W.; Marron, J. S. Fast and simple scatterplot smoothing. *Comput. Stat. Data Anal.* **1995**, *20*, 1−17.

(48) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman and Hall: New York, 1990.

(49) Buja, A.; Hastie, T.; Tibshirani, R. Linear smoothers and the additive model. *Ann. Stat.* **1989**, *17*, 453−555.

(50) Stone, C. J. Additive regression and other nonparametric models. *Ann. Stat.* **1985**, *13*, 689−705.

(51) Linton, O. B.; Nielsen, J. P. Estimating structured nonparametric regression by the kernel method. *Biometrika* **1995**, *82*, 93−101.

(52) Linton, O. B. Efficient estimation of additive nonparametric regression models. *Biometrika* **1997**, *84*, 469−473.

(53) Jolliffe, I. T. A note on the use of principal components in regression. *Appl. Stat.* **1982**, *31*, 300−303.

(54) Friedman, J. H.; Stuetzle, W. Projection pursuit regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817−823.

(55) Li, K.-C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **1991**, *86*, 316−327.

(56) Kruse, L. I.; Kaiser, C.; DeWolf, W. E.; Frazee, J. S.; Ross, S. T.; Wawro, J.; Wise, M.; Flaim, K. E.; Sawyer, J. L.; Erickson, R. W.; Ezekiel, M.; Ohlstein, E. H.; Berkowitz, B. A. Multisubstrate inhibitors of dopamine β-hydroxylase. 2. Structure−activity relationships at the phenethylamine binding site. *J. Med. Chem.* **1987**, *30*, 486−494.

(57) Burke, B. J.; Hopfinger, A. J. 1-(Substituted-benzyl)imidazole-2(3*H*)-thione inhibitors of dopamine β-hydroxylase. *J. Med. Chem.* **1990**, *33*, 274−281.

(58) Hahn, M.; Rogers, D. Receptor surface models. 2. Application to quantitative structure−activity relationships studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.

(59) So, S.-S.; Karplus, M. Three-dimensional quantitative structure−activity relationships from molecular similarity matrixes and genetic neural networks. 2. Applications. *J. Med. Chem.* **1997**, *40*, 4360−4371.

(60) Hahn, M. Receptor surface models. 1. Definition and construction. *J. Med. Chem.* **1995**, *38*, 2080−2090.

(61) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of comparative molecular field analysis models: Effects of data scaling and variable selections using a set of human synovial fluid phospholipase A_2 inhibitors. *J. Med. Chem.* **1997**, *40*, 1136−1148.

(62) Press: W. H.; Teukolsky, S. A.; Vettering, W. T.; Flannery, B. P. *Numerical Recipes*; Cambridge University Press: Cambridge, U.K., 1992.

(63) So, S.-S.; Karplus, M. Three-dimensional quantitative structure−activity relationships from molecular similarity matrixes and genetic neural networks. 1. Method and validations. *J. Med. Chem.* **1997**, *40*, 4347−4359.

(64) Lucic, B.; Trinajsstic, N. Multivariate regression outperforms several robust architectures of neural networks in QSAR modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.

(65) Lucic, B.; Trinajstic, N.; Slid, S.; Karelson, M.; Katritzky, A. R. A new efficient approach for variable selection based on multiregression: prediction of gas chromatography retention times and response factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.

CI990082E