An Accurate QSPR Study of O−H Bond Dissociation Energy in Substituted Phenols Based on Support Vector Machines

C. X. Xue,[†] R. S. Zhang,[†,‡] H. X. Liu,[†] X. J. Yao,[†,||] M. C. Liu,[†] Z. D. Hu,*[,§,†] and B. T. Fan[||]

Department of Chemistry, Lanzhou University, Lanzhou 730000, China, Department of Computer Science, Lanzhou University, Lanzhou 730000, China, State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou 730000, China, and Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France

Received October 30, 2003

The support vector machine (SVM), as a novel type of learning machine, was used to develop a Quantitative Structure−Property Relationship (QSPR) model of the O−H bond dissociation energy (BDE) of 78 substituted phenols. The six descriptors calculated solely from the molecular structures of compounds selected by forward stepwise regression were used as inputs for the SVM model. The root-mean-square (rms) errors in BDE predictions for the training, test, and overall data sets were 3.808, 3.320, and 3.713 BDE units (kJ mol^{-1}), respectively. The results obtained by Gaussian-kernel SVM were much better than those obtained by multiple linear regression, radial basis function neural networks, linear-kernel SVM, and other QSPR approaches.

1. INTRODUCTION

Phenolic compounds play a major role in the chemistry of living organisms and life-supporting substances. Phenols are widely used as synthetic organic materials and also as antioxidants in living organisms.[1] The enormous interest in their antioxidant activity in vitro and in vivo has been demonstrated by a wealth of research in recent years.[2] The properties of the O−H bond appear to be essential to understanding the chemical and biochemical behavior of phenolic compounds, for this is the bond which must be broken to generate the truly active species: the phenoxy radical. Phenols are also of special interest in organic chemistry, since their acid−base equilibria have often been used as reference values in establishing linear free energy relationships.[3] Consequently, much effort has been put into understanding the factors governing the O−H bond dissociation energies, O−H BDE, both in the solution and the gas phase.[4−6] There are many experimental studies for the determination of the O−H BDE of substituted phenols.[4−6] These studies were generally carried out in solutions (such as water, DMSO, etc.), and subsequently gas-phase O−H BDE values were determined under some assumptions.[5,6] Unfortunately, the O−H BDE values obtained from different experimental studies vary in a wide range. For example, different experimental studies suggested the O−H BDE for phenol was from 83.3 kcal/mol to 89.6 kcal/mol.[4] This discrepancy clearly shows that there remains a lot of uncertainties even in the experimental determined O−H BDE values.

The computational approaches used to predict O−H BDE can be classified into two categories. The rigorous approaches involve quantum mechanical methods including density functional methods,[7] ab initio,[8] and semiempirical methods.[9] The accurate estimation of BDE from theoretical calculations is a challenging task, since high levels of calculations are necessary for taking into account the effect of both the dynamical and the nondynamical part of electron correlation. High level ab initio calculations are thus prohibitive for the large size of the substituted phenol molecules. In most of the cases the calculated BDE values differ widely from the experimental value.[7]

The other methods are more empirically based quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSPR) approaches. In a recent review,[10] quantitative structure−activity relationships have been established for both biological and nonbiological activities and properties of substituted phenols with electronic properties of the substituents. More recently, Bosque et al. have applied the QSPR approach to predict the BDE of phenols.[11] Despite all efforts, the present knowledge on the calculation of O−H BDE is still unsatisfactory. Even for the simplest of those molecules—phenol itself—the published values for the O−H BDE vary over a wide range (∼30 kJ mol^{-1}).

Of those previous studies that were aimed at predicting the O−H BDE, the most promising method is to use QSPR, which uses descriptors derived from the molecular structure alone representing the character of the molecule. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from structure alone and are not dependent on any experimental properties. Once the structure of a compound is known, any descriptor can be calculated no matter whether they are synthesized of not. So once a reliable model is established, we can use this method to predict the property of compounds. This study can also tell us which of the structural factors may play an important role in the determination of a property. QSPR methods have been successfully used to predict many physicochemical properties. Nevertheless, to the best of our knowledge there is only one calculation of bond energies by the QSPR approach and the maximum absolute relative

* Corresponding author phone: +86-931-891-2578; fax: +86-931-891-2582; e-mail address:huzd@lzu.edu.cn.
[†] Department of Chemistry, Lanzhou University.
[‡] Department of Computer Science, Lanzhou University.
[§] State Key Laboratory of Applied Organic Chemistry, Lanzhou University.
[||] Université Paris 7-Denis Diderot.

10.1021/ci034248u CCC: $27.50
Published on Web 02/04/2004

error is 5.24% which showed the obtained results were not satisfactory.[11]

Machine learning techniques such as neural networks, genetic algorithm, etc., have been applied to QSPR analysis since the late 1980s, mainly in response to increased accuracy demands. The most popular neural network model is the back-propagation (BP) neural network due to its simple architecture yet powerful problem-solving ability. However, the BP neural network suffers from a number of weaknesses which include the need for a large number of controlling parameters, difficulty in obtaining a stable solution, and the danger of over-fitting. The over-fitting problem is a critical issue that usually leads to poor generalization because the neural network has too large a capacity which causes it to capture not only the useful information contained in the training data but also unwanted noises. As a result, it will end up only memorizing the training data and generalizing poorly to the out-of-sample data.[12] Genetic algorithms can suffer in a similar manner. The stochastic nature of both population initialization and the genetic operators used during training can make results hard to reproduce.[13] Owing to the reasons outlined above, there is a continuing need for the application of more accurate and informative techniques to QSPR analysis.

The support vector machines (SVM) is a new algorithm developed from the machine learning community. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application, such as pattern recognition problems,[14] classification and diagnosis of breast cancer,[15] drug design,[16] and other QSAR analysis.[17]

In the present investigation, for the first time, SVM was used for the prediction of O−H BDE in phenols using descriptors calculated by the software CODESSA.[18] Multiple linear regression (MLR) and radial basis function neural networks (RBFNNs) methods were also utilized to establish quantitative linear and nonlinear relationships to compare the results with that obtained by SVM. The aim was to establish an accurate quantitative structure−property relationship model, to confirm the possibility of predicting O−H BDE of phenols, and, at the same time, to seek for the important structural features related to the O−H BDE. The prediction results were very satisfactory in both the training set and the test set, which proved SVM was a useful tool in the prediction of the O−H BDE.

2. EXPERIMENTAL SECTION

2.1. Data Preparation. The values of bond dissociation energy of the substituted phenols studied were taken from the review published by R. M. Borges dos Santos and J. A. Martinho Simoes.[4] Table 1 contains the BDE of the data set, in kJ mol^{-1}. The compounds contain 25 kinds of substituents with different electronic effects. Some of these substituents are electron-donating such as Me, *t*-Bu, NH_2, NMe_2, OMe, while others are electron-withdrawing, such as F, Cl, Br, I, NO_2, CF_3, COR, SO_2Me. These substituents occupy the *ortho*-, *meta*-, and *para*-positions of the aromatic ring; 40 of the phenols studied have at least one substituent in the *ortho*-position. The bond dissociation energy in data set fall in the range of 321.3 kJ mol^{-1} for phenol substituted with *p*-NMe_2 to 397.3 kJ mol^{-1} for *o*-$(NO_2)_2$, respectively, with a mean value of 362.3 kJ mol^{-1}. To compare the results with the literature, the separation of the substituted phenols in the training and test sets is identical with that in ref 11. The training set of 62 compounds was used to adjust the parameters of the model, and the test set of 16 compounds was used to evaluate its prediction ability. Leave-one-out cross-validation was performed to evaluate the modeling ability of models.

2.2. Descriptor Calculation. All structures of the molecules were drawn with the HyperChem program and exported in a file format suitable for MOPAC.[19] The final geometries were obtained with the semiempirical AM1 method in the MOPAC 6.0 program.[20] All the geometries have been fully optimized without symmetry restrictions. In all cases frequency calculations have been performed in order to ensure that all the calculated geometries correspond to true minima. The resulted geometry was transferred into software CODESSA that can calculate constitutional, topological, geometrical, electrostatic, and quantum-chemical descriptors.

3. METHODOLOGY

3.1. Feature Selection and Regression Analysis. Once descriptors were generated, in this work, the correlation analysis of descriptors was performed first. In the process of correlation analysis, either parameter which the correlation coefficient is more than 0.85 was discarded. After the correlation analysis of the descriptors, descriptor-screening methods were used to select the most relevant descriptor to establish the models for prediction of the molecular property. Here, the forward stepwise regression method was used to choose the subset of the molecular descriptors. Forward stepwise regression starts with no model terms, and at each step it adds the most statistically significant term (the one with the highest *F*-statistic or lowest *P*-value) until there are none left.

After the descriptor was selected, multiple linear regression was used to develop the linear model of the property of interest, which takes the form below:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

In this equation, Y is the property, that is, the dependent variable, X_1-X_n represents the specific descriptor, while b_1-b_n represents the coefficients of those descriptors, and b_0 *is* the intercept of this equation.

3.2. Theory of Radial Basis Function Neural Networks. The theory of RBFNNs has been extensively described in the paper of Yao et al.[21] Here, only a brief description of the RBFNNs principle was given. The RBFNNs consists of three layers: the input layer, the hidden layer, and the output layer. The input layer does not process the information; it only distributes the input vectors to the hidden layer. Each neuron on the hidden layer employs a radial basis function (RBF) as a nonlinear transfer function to operate on the input data. The most often used RBF is a Gaussian function that is characterized by a center (c_j) and width (r_j). In this study, the Gaussian was selected as a radial basis function. The operation of the output layer is linear, which is given as below

$$y_k(x) = \sum w_{kj}h_j(x) + bk$$

where y_k is the kth output unit for the input vector x, w_{kj} is

Table 1. Experimental and Calculated BDE (kJ mol^{-1}) for the Training and Test Sets

no.	compd[a]	exptl	ref 11[b]	res[c]	MLR	RBFNNs	L−K SVM[d]	G−K SVM[e]	res[c]
1	PhOH	371.3	372.4	0.30	367.4	367.4	367.6	369.4	−0.51
2	2-Me	362.3	358.5	−1.05	360.9	360.9	363.4	365.8	0.95
3	4-Me	363.3	365.1	0.50	361.4	361.0	362.2	362.0	−0.37
4	4-t-Bu	364.3	358.3	−1.65	363.3	363.9	362.9	361.6	−0.74
5	4-Ph	359.3	364.6	1.48	363.4	364.1	365.3	363.4	1.15
6	2-OH	341.3	347.7	1.88	343.9	343.0	349.7	346.0	1.38
7	3-OH	372.3	364.8	−2.01	359.4	361.1	358.4	370.3	−0.54
8	4-OMe	349.3	353.7	1.26	353.5	352.6	352.2	356.7	2.11
9	2-CH_2OH	361.3	352.4	−2.46	355.9	365.2	354.0	360.1	−0.32
10	2-COMe	365.3	362.3	−0.82	368.6	368.6	369.2	368.5	0.87
11	4-COMe	380.3	377.8	−0.66	379.5	380.0	377.7	378.0	−0.59
12	3-COEt	380.3	375.3	−1.31	379.7	380.8	379.6	379.0	−0.34
13	4-COPh	382.3	387.4	1.33	386.4	386.2	388.0	387.3	1.30
14	4-OCOMe	360.3	366.9	1.83	369.4	364.1	370.2	369.6	2.59
15	3-COOEt	378.3	377.3	−0.26	378.9	380.0	377.6	376.6	−0.45
16	2-NH_2	340.3	342.8	0.73	343.6	341.5	344.2	340.0	−0.08
17	4-NH_2	331.3	337.8	1.96	333.2	325.5	332.2	331.6	0.09
18	3-NMe_2	363.3	344.3	−5.23	351.7	354.6	351.0	361.1	−0.62
19	4-NMe_2	321.3	335.6	4.45	342.4	341.9	342.6	322.8	0.46
20	2-CN	371.3	377.1	1.56	378.7	378.0	382.1	371.7	0.11
21	3-CN	384.3	382.7	−0.42	382.9	384.1	383.6	386.2	0.49
22	4-CN	389.3	381.7	−1.95	382.7	384.2	383.0	383.7	−1.43
23	2-NO_2	366.3	372.8	1.77	383.3	384.5	382.0	371.3	1.36
24	3-NO_2	390.3	385.0	−1.36	387.1	387.6	385.9	386.0	−1.09
25	4-NO_2	396.3	390.3	−1.51	395.7	396.6	394.9	397.4	0.27
26	2-F	363.3	365.0	0.47	364.2	363.0	365.8	357.4	−1.61
27	3-F	377.3	376.5	−0.21	377.9	379.6	377.9	377.3	0.00
28	4-Cl	370.3	371.8	0.41	370.5	370.4	371.9	370.3	0.00
29	2-Br	364.3	372.4	2.22	368.9	368.8	374.3	372.1	2.15
30	4-Br	373.3	375.9	0.70	376.0	376.9	378.6	378.6	1.43
31	4-I	370.3	379.4	2.46	358.7	357.9	360.0	369.8	−0.13
32	4-CF_3	388.3	386.6	−0.44	393.0	388.9	392.6	396.1	2.01
33	3-SO_2Me	382.3	380.5	−0.47	379.9	380.0	381.0	383.8	0.40
34	4-SO_2Me	393.3	396.0	0.69	388.3	388.4	388.6	390.4	−0.73
35	2,6-Me_2	357.3	355.6	−0.48	357.9	358.4	356.5	355.4	−0.53
36	3,5-Me_2	368.3	363.1	−1.41	357.7	356.6	358.8	361.9	−1.73
37	2,4-t-Bu_2	349.3	346.5	−0.80	351.4	349.8	354.7	354.3	1.42
38	2,6-t-Bu_2	345.3	347.7	0.70	351.9	350.3	349.7	347.2	0.56
39	3,5-t-Bu_2	365.3	349.5	−4.33	359.9	358.9	360.9	362.0	−0.90
40	2,6-Ph_2	360.3	359.8	−0.14	357.8	358.0	354.3	352.0	−2.30
41	2,6-$(OMe)_2$	350.3	354.5	1.20	349.8	354.4	343.9	347.5	−0.79
42	3,5-$(OMe)_2$	364.3	364.6	0.08	366.6	364.2	369.7	363.3	−0.29
43	2,6-$(NO_2)_2$	397.3	398.8	0.38	398.6	388.6	394.9	391.3	−1.52
44	3,5-Cl_2	385.3	379.1	−1.61	382.5	385.0	384.8	388.0	0.71
45	2,4,6-Me_3	348.3	346.3	−0.57	347.2	348.3	345.3	346.9	−0.41
46	2,6-t-Bu_2−4-Me	339.3	339.2	−0.03	340.5	341.0	339.8	339.6	0.08
47	2,4,6-t-Bu_3	334.3	329.4	−1.47	335.6	337.0	336.1	336.8	0.74
48	2,4,6-Ph_3	348.3	352.6	1.23	343.0	347.2	341.3	350.6	0.67
49	2,6-t-Bu_2−4-OtBu	322.3	312.8	−2.95	320.4	323.1	312.6	320.0	−0.72
50	2,6-t-Bu_2−4-COMe	345.3	353.1	2.26	350.0	350.2	351.0	347.6	0.66
51	2,6-t-Bu_2−4-OCOMe	345.3	342.9	−0.70	340.3	346.0	337.4	342.9	−0.70
52	2,6-t-Bu_2−4-CH_2NMe	333.3	333.9	0.18	329.9	330.8	337.9	329.6	−1.12
53	2,4,6-$(OMe)_3$	336.3	342.0	1.69	348.5	338.4	347.5	339.3	0.90
54	2,6-Me_2−4-NO_2	371.3	375.2	1.05	362.7	374.0	359.3	369.9	−0.39
55	2,6-t-Bu_2−4-NO_2	352.3	361.1	2.50	358.8	350.1	361.3	349.2	−0.87
56	2,6-t-Bu_2−4-CHNOH	328.3	339.6	3.44	331.7	328.9	341.7	329.4	0.32
57	3,4,5-Cl_3	384.3	379.1	−1.35	372.9	375.2	373.3	383.9	−0.09
58	2,6-(NO_2)2−4-Cl	396.3	394.2	−0.53	393.8	399.7	392.2	399.8	0.88
59	2,3,6-Me_3−4-OMe	333.3	333.6	0.09	341.1	337.8	339.7	339.1	1.73
60	2,3,5,6-Me_4−4-OMe	344.3	341.5	−0.81	333.4	333.2	329.7	338.2	−1.76
61	2,3,5,6-F_4	385.3	375.7	−2.49	371.4	371.8	358.7	382.3	−0.79
62	2,3,4,5,6-F_5	360.3	369.5	2.55	373.8	371.5	389.2	360.5	0.06
63[f]	2,4,6-Cl_3	369.3	369.0	−0.08	366.8	374.0	373.4	375.6	1.70
64[f]	2,6-Cl_2	370.3	372.7	0.65	369.0	368.1	378.0	375.6	1.43
65[f]	2,6-Me_2−4-OMe	329.3	336.9	2.31	343.0	342.8	340.7	330.0	0.22
66[f]	2,6-t-Bu_2−4-Et	328.3	335.9	2.31	334.3	335.9	333.9	334.8	1.98
67[f]	2-CH═CH_2	361.3	360.1	−0.33	360.1	360.9	363.1	358.0	−0.91
68[f]	2-Cl	368.3	369.7	0.38	361.3	362.3	367.7	368.1	−0.04
69[f]	2-OMe	354.3	357.3	0.85	350.5	357.6	352.4	353.5	−0.23
70[f]	3-COMe	376.3	377.8	0.40	376.6	376.9	374.8	376.0	−0.08
71[f]	3-Cl	376.3	374.2	−0.56	374.0	374.3	374.8	374.5	−0.49
72[f]	3-Me	368.3	368.3	0	362.8	361.7	363.6	365.1	−0.88
73[f]	3-NH_2	366.3	355.0	−3.08	359.5	361.3	360.8	365.5	−0.21
74[f]	3-OMe	371.3	362.1	−2.48	363.1	367.9	363.8	366.5	−1.29
75[f]	4-F	367.3	367.1	−0.05	374.6	375.6	373.4	371.4	1.12
76[f]	4-OH	344.3	348.4	1.19	347.5	343.4	351.1	343.9	−0.13
77[f]	2,6-t-Bu_2-4-CHO	343.3	353.2	2.88	355.2	355.4	352.2	343.4	0.04
78[f]	3-CF_3	384.3	382.5	−0.47	378.3	387.0	374.4	384.3	0

[a] Substituent on phenol. [b] Predicted BDE in ref 11. [c] Relative error: value of 100[(calc−exp)/exp]. [d] Predected BDE by linear-kernel SVM. [e] Predicted BDE by Gaussian-kernel SVM. [f] Test set.

the weight connection between the kth output unit and the jth hidden layer unit, and h_j is the notation for the output of the jth RBF unit.

The training procedure when using RBF involves selecting centers, width and weights. In this paper, the forward subset selection routine was used to select the centers from training set samples.[22,23] The adjustment of the connection weight between the hidden layer and the output layer was performed using a least-squares solution after the selection of centers and width of radial basis functions.

3.3. Support Vector Machines. 3.3.1. Structural Risk Minimization.[24,25] Previous approaches to statistical learning have tended to be based on finding functions to map vector-encoded data to their respective classes. The conventional minimization of the empirical risk over the training data does not, however, imply a good generalization to the novel test data. Indeed, there could be a number of different functions which all give a good approximation to a training set. It is nevertheless difficult to determine a function which best captures the true underlying structure of the data distribution. Structural risk minimization (SRM) aims to address this problem and provides a well-defined quantitative measure for the capacity of a learned function to generalize over unknown test data. Due to its relative simplicity, the Vapnik-Chervonenkis (VC) dimension (Vapnik, 1995) in particular has been adopted as one of the more popular measures for such a capacity. By choosing a function with a low VC dimension and minimizing its empirical error to a training data set, SRM can offer a guaranteed minimal bound on the test error.

3.3.2. Theory of SVM for Regression.[26] The SVM method was proposed by Vapnik[27] on the basis of the Structural Risk Minimization Principle.[25] It was initially designed to solve pattern recognition problems,[28] but it was later applied to function estimation problems.[29] The estimated function is a linear expansion in terms of functions defined on a certain subset of the data (support vectors), and the final number of coefficients in such an expansion does not depend on the dimensionality of the space of input variables. These two properties make SVM an especially useful technique for dealing with very large data sets in a high-dimensional space.

Compared to other neural network regressors, there are three distinct characteristics when SVM are used to estimate the regression function. First of all, SVM estimate the regression using a set of linear functions that are defined in a high dimensional space. Second, SVM carry out the regression estimation by risk minimization where the risk is measured using Vapnik's ϵ-insensitive loss function. Third, SVM use a risk function consisting of the empirical error and a regularization term which is derived from the structural risk minimization principle.

Figure 1 contains a graphical overview over the different steps in the regression stage of SVM. Given a set of data points $G = \{(x_i,d_i)\}_i^n$ (x_i is the input vector, d_i is the desired value, and n is the total number of data patterns), SVM approximate the function using the following equation

$$y = f(x) = w\Phi(x) + b \tag{1}$$

where $\Phi(x)$ is the high dimensional feature space which is nonlinearly mapped from the input space x. The coefficients w and b are estimated by minimizing

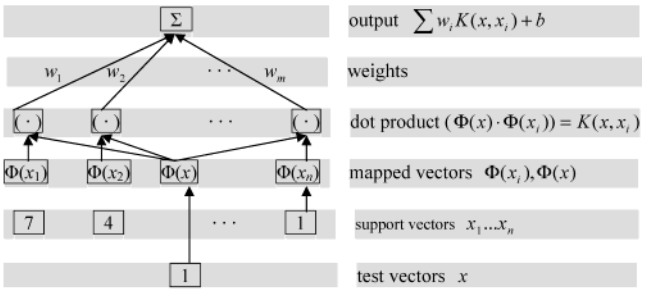

Figure 1. Architecture of a regression machine constructed by the SV algorithm.

$$R_{\mathrm{SVMs}}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\epsilon(d_i,y_i) + \frac{1}{2}\|w\|^2 \tag{2}$$

$$L_\epsilon(d,y) = \begin{cases} |d-y| & -\epsilon|d-y| \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In the regularized risk function given by eq 2, the first term $C(1/n)\sum_{i=1}^{n} L_\epsilon(d_i,y_i)$ is the empirical error (risk). They are measured by the ϵ-insensitive loss function given by eq 3. This loss function provides the advantage of enabling one to use sparse data points to represent the decision function given by eq 1. The second term $1/2\,\|w\|^2$, on the other hand, is the regularization term. C is referred to as the regularized constant, and it determines the tradeoff between the empirical risk and the regularization term. Increasing the value of C will result in the relative importance of the empirical risk with respect to the regularization term to grow. ϵ is called the tube size, and it is equivalent to the approximation accuracy placed on the training data points. Both C and ϵ are user-prescribed parameters.

To obtain the estimations of w and b, eq 2 is transformed to the primal function given by eq 4 by introducing the positive slack variables ξ_i and ξ_i^* as follows:

$$\text{minimize } R_{\mathrm{SVMs}}(w,\xi^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{subject to } \begin{cases} d_i - w\Phi(x_i) - b_i \leq \epsilon + \xi_i \\ w\Phi(x_i) + b_i - d_i \leq \epsilon + \xi_i^* \\ \xi_i,\xi_i^* \geq 0 \end{cases} \tag{4}$$

Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function given by eq 1 has the following explicit form

$$f(x,a_i,a_i^*) = \sum(a_i - a_i^*)K(x,x_i) + b \tag{5}$$

where the kernel function K corresponds to $K(x,x_i) = \phi(x)^T\phi(x_i)$. One has several possibilities for the choice of this kernel function, including linear, polynomial, splines, and radial basis function (RBF). In the support vector regression, the Gaussian Radial Basis Function kernel $K(x,y) = \exp(-(x-y)^2/\delta^2)$ is commonly used.

Table 2. Six-Parameter Correlation Equation for the Linear Model[a]

descriptor	chemical meaning	coefficient	SE	beta	*t*-test	sig
intercept	intercept	−124.548	55.361		−2.250	0.028
HOMO	HOMO energy	−26.375	2.301	−0.759	−11.465	0.000
YZS	YZ shadow	−1.020	0.117	−0.472	−8.728	0.000
PP/SD	polarity parameter/square distance	55.465	9.375	0.329	5.916	0.000
Qmin	min partial charge (Qmin)	−1481.575	262.085	−0.446	−5.653	0.000
WNSA-3	WNSA-3 weighted PNSA (PNSA3*TMSA/1000) [Zefirov's PC]	−4.490	0.944	−0.400	−4.756	0.000
RNCS	RNCS relative negative charged SA (SAMNEG*RNCG) [quantum-chemical PC]	−1.360	0.415	−0.186	−3.278	0.002

[a] $R = 0.937$; $R^2 = 0.879$; SE of the estimate = 7.205; rms = 6.786; $n = 62$; $F = 66.354$.

Table 3. Correlation Matrix of the 6 Descriptors Used in This Work[a]

	HOMO	YZS	PP/SD	Qmin	WNSA-3	RNCS
HOMO	1.000	0.190	−0.103	−0.587	0.594	−0.311
YZS		1.000	−0.024	−0.105	0.000	−0.405
PP/SD			1.000	0.154	0.171	0.246
Qmin				1.000	−0.736	0.092
WNSA-3					1.000	−0.128
RNCS						1.000

[a] The definitions of the descriptors were given in Table 2.

The overall performances of RBFNNs and SVM were evaluated in terms of the rms error which was defined as below

$$\mathrm{rms} = \sqrt{\frac{\sum_{i=1}^{n_s} (y_k - \hat{y}_k)^2}{n_s}} \qquad (6)$$

where y_k is the desired output and $\hat{y}_k$ is the actual output of the model, and n_s is the number of compounds in the analyzed set.

3.3.3. RBFNNs and SVM Implementation and Computation Environment. All calculation programs implementing RBFNNs were written in M-file based on a basis MATLAB script for RBFNNs. All calculation programs implementing SVM were written in R-file based on an R script for SVM and compiled using a R1.7.1 compiler. The scripts were run on a Pentium IV PC with 256M RAM.

4. RESULTS AND DISCUSSION

4.1. Results of MLR. About 600 descriptors were calculated by the CODESSA program. After the correlation analysis of the descriptors, the pool of descriptors was reduced to 241. A stepwise regression routine was used to develop the linear model for the prediction of the O−H BDE of substituted phenols using calculated structural descriptors. The best linear model contains 6 molecular descriptors. The regression coefficients of the descriptors and their physical-chemical meaning were listed in Table 2, and the correlation matrix of these descriptors was shown in Table 3. The linear correlation coefficient value of each of the two descriptors is < 0.85, which means the descriptors were independent in this MLR analysis. Of them, one is geometrical (YZS), two are quantum-chemical (HOMO, RNCS), and three are electrostatic (PP/SD, Qmin, and WNSA-3) descriptors. This model gave an rms error of 6.786 BDE units for the training set, 6.556 for the test set, and 6.739 for the whole set, and the corresponding correlation coefficients (R) were 0.937, 0.930, and 0.934, respectively. Figure 2 showed these predicted versus experimental O−H BDE.

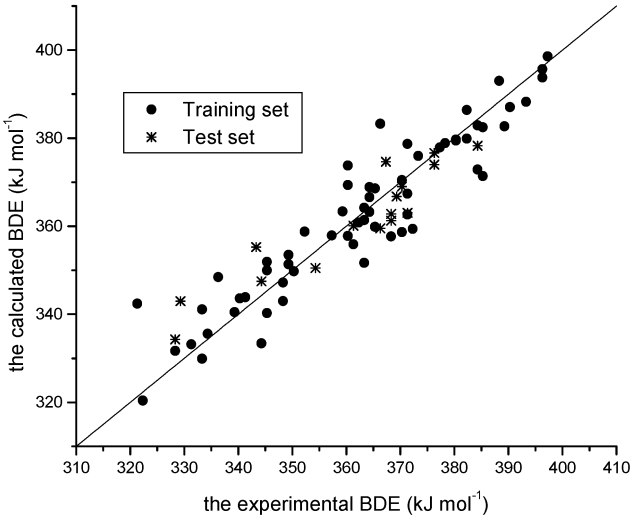

Figure 2. Plot of calculated vs experimental BDE (kJ mol^{-1}) (MLR).

By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to relate to the O−H BDE of phenols. The geometrical descriptors describe the size of the molecules and are derived from the three-dimensional coordinates of the atomic nuclei and the atomic masses and the atomic radii in the molecule. The only descriptor contained in the model that belongs to this group is the YZ shadow (YZS). It receives a negative coefficient in the regression; this indicates that the O−H BDE decreases with the increasing of the YZ shadow. The two quantum-chemical descriptors are E_{HOMO} and RNCS. The HOMO energy (E_{HOMO}) is the energy of the highest occupied molecular orbital. RNCS is defined as a relative negative charged surface area (RNCS). The electrostatic descriptors reflect the characteristic of the charge distribution of the molecule. The three electrostatic descriptors are as follows: PP/SD, Qmin, and WNSA-3. The polarity parameter/square distance (PP/SD) is defined as the polarity parameter factorized by the division with the square of the distance between the atoms bearing minimum and maximum partial charges. Qmin is the minimum partial charge in the molecule. WNSA-3 belongs to the charged partial surface area descriptors (CPSA), in terms of the whole surface area of the molecule or its fragments and the charge distribution in the molecule. WNSA-3 is defined as surface weighted CPSA (WNSA-3). The strength of the O−H bonds is predominantly an electrostatic interaction, and we reasoned that the atomic

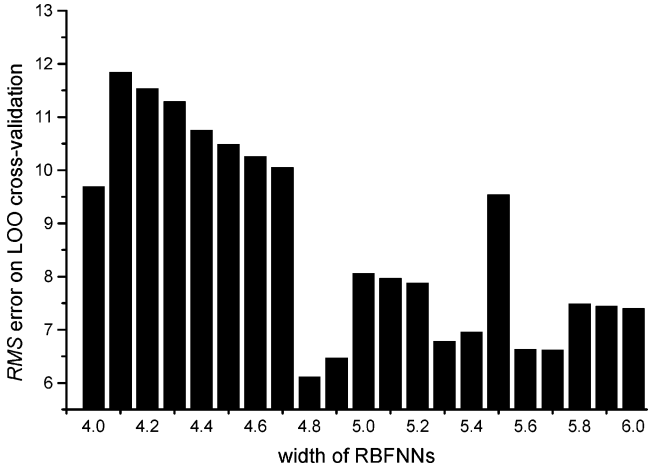

Figure 3. The width of RBFNNs vs rms error on LOO cross-validation.

charge gap between oxygen and hydrogen of O−H bond should be a valid measure of the strength of the O−H bond. The larger the gap is, the higher the O−H BDE is. The gap can be described by one quantum-chemical descriptor, RNCS, and three electrostatic descriptors: PP/SD, Qmin, WNSA-3. Of these four descriptors, RNCS, Qmin, and WNSA-3 receive negative coefficients in the regression, and this indicates that increasing the values of these three descriptors of the molecule leads to a low O−H BDE. While PP/SD receives a positive coefficient in the regression, this indicates that increasing the PP/SD leads to a high BDE. Moreover, the strength of the O−H bond also has a covalent or charge-transfer component, which can be described by E_{HOMO}. It receives a negative coefficient in the regression, which indicates that increasing the HOMO energy leads to a low BDE.

According to the beta values (Table 2), the more relevant descriptors are the E_{HOMO}, YZS, and Qmin.

4.2. Result of RBFNNs. From Table 2, it can be seen that the model of MLR was not sufficiently accurate (rms = 6.786, SE = 7.205) and showed the factors influencing O−H BDE were complex and not all of them were a linear correlation with the BDE. So, we built the nonlinear prediction models by RBFNNs and SVM to further discuss the correlation between the molecular structure and the BDE based on the same descriptor set.

After the establishment of a linear model, RBFNNs were used to develop a nonlinear model based on the same subset of descriptors. Each minimum error on the LOO cross-validation was plotted versus the width (Figure 3), and the minimum was chosen as the optimal conditions.

Through the above process, the optimum width and the best number of hidden layer units were selected as 4.8 and 16, respectively. From the best network, the inputs in the test set were presented with it, and the results with RBFNNs were obtained. They were shown in Table 1 and Figure 4. The network gave an rms error of 6.112 for the training set, 6.238 for the test set, and 6.138 BDE units for the whole set.

4.3. Results of SVM. 4.3.1. SVM Parameters Optimization. Similar to other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are capacity parameter C, ϵ of ϵ-insensitive loss function, the kernel type K, and its

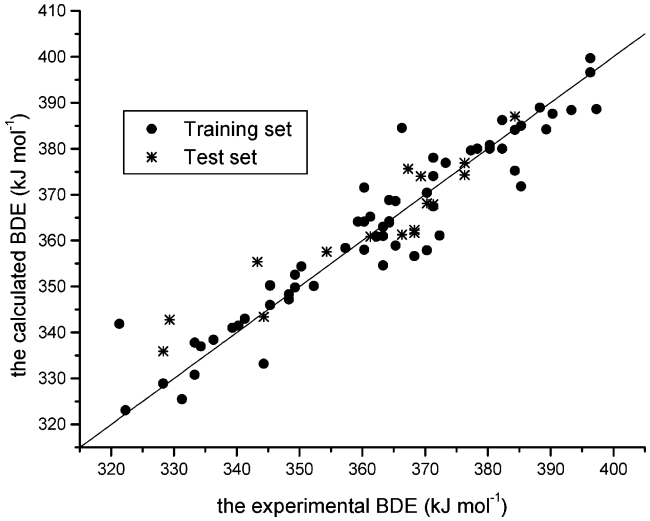

Figure 4. Plot of calculated vs experimental BDE (kJ mol^{-1}) (RBFNNs).

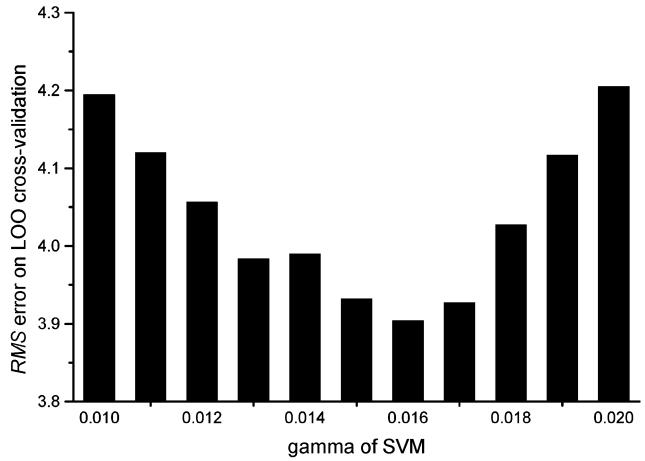

Figure 5. The gamma versus rms error on LOO cross-validation ($C = 100$, $\epsilon = 0.1$)

corresponding parameters. C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. To make the learning process stable, a large value should be set up for C (e.g., $C = 100$).

The kernel type is another important parameter. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function is as follows

$$exp(-\gamma^*|u - v|^2)$$

where γ is a constant, the parameter of the kernel; u and v are two independent variables; and γ controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. Each rms error on the LOO cross-validation was plotted versus γ (Figure 5), and the minimum was chosen as the optimal conditions. In this case: $\gamma = 0.016$.

The optimal value for ϵ depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ϵ, there is the practical consideration of the number of resulting support vectors. ϵ-insensitivity

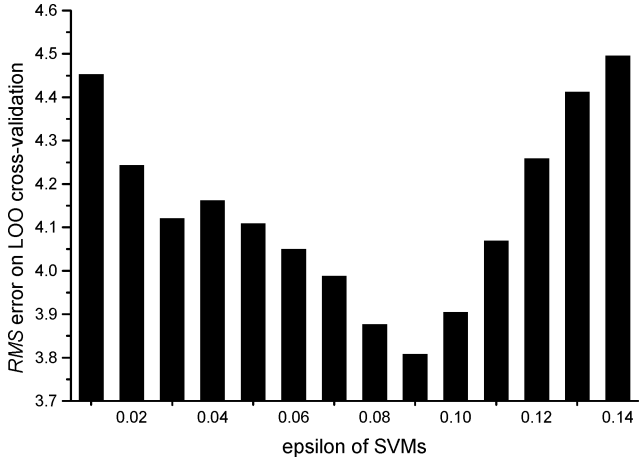

Figure 6. The epsilon vs rms error on LOO cross-validation ($C = 100$, $\gamma = 0.016$).

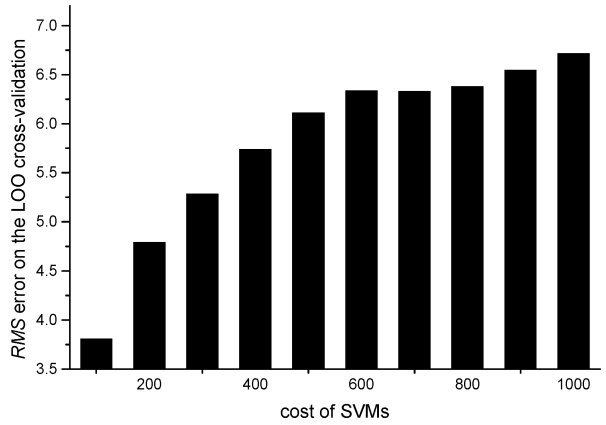

Figure 7. The cost versus rms error on LOO cross-validation ($\gamma = 0.016$, $\epsilon = 0.09$).

prevents the entire training set meeting boundary conditions and so allows for the possibility of sparsity in the dual formulation's solution. So, choosing the appropriate value of ϵ is critical from theory. To find an optimal ϵ, the rms on LOO cross-validation on different ϵ was calculated. The curve of rms versus the epsilon was shown in Figure 6. The optimal ϵ was found as 0.09.

The last important parameter is the regularization parameter *C*, of which the effect on the rms was shown in Figure 7. From Figure 7, the optimal *C* was found as 100.

To compare the results obtained by MLR, the linear kernel function was also used to build another SVM model. In the linear kernel function, there is one parameter to be adjusted, i.e., *C*, of which the effect on the rms error was shown in Figure 8. From Figure 8, the optimal *C* was found as 400.

4.3.2. The Predicted Results of SVM. From the above discussion, in the Gaussian-kernel SVM, the γ, ϵ, and *C* were fixed to 0.016, 0.09, and 100, respectively, when the support vector number of the SVM model was 44, the predicted results of the optimal SVM were shown in Table 1 and Figure 9. As can be seen from Table 1 and Figure 9, the proposed models were statistically stable and fitted the data well. The experimental and predicting values of the test set by the SVM model were also listed in Table 1. The model gave an rms of 3.713 for the whole set, 3.808 for the training set, and 3.320 for the prediction set, and the corresponding correlation coefficients (*R*) were 0.980, 0.981, and 0.979. It

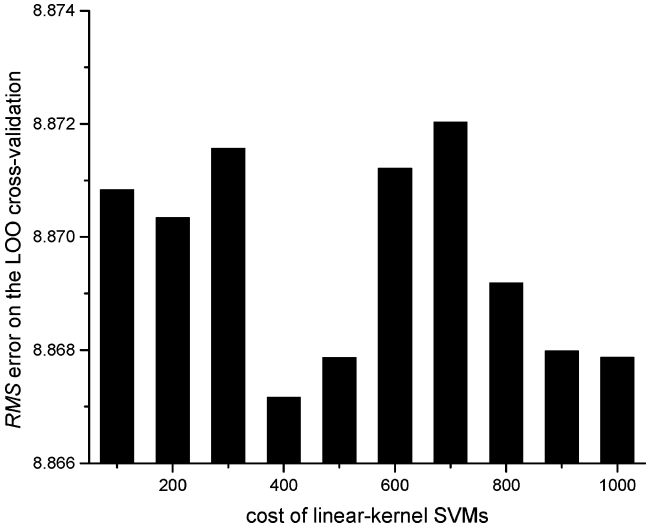

Figure 8. The cost vs rms error on LOO cross-validation (linear-kernel SVM).

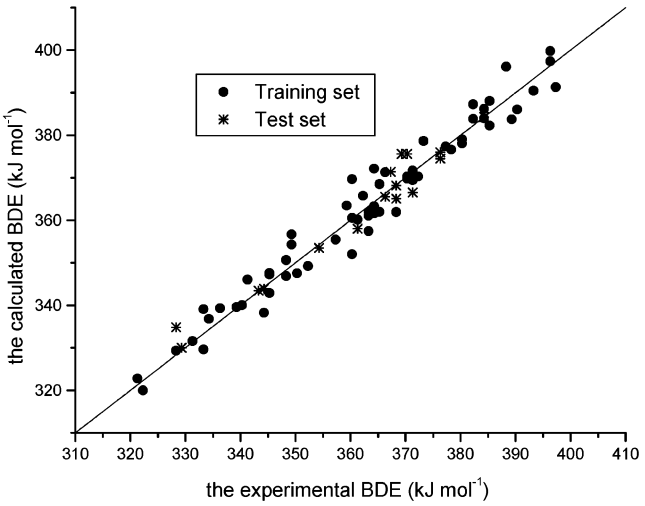

Figure 9. Plot of calculated vs experimental BDE (kJ mol^{-1}) (Gaussian-kernel SVM).

can be concluded that the predicted values obtained by the Gaussian-kernel SVM are in very good agreement with the experimental values.

When the linear kernel function was used to build another SVM model, *C* was fixed to 400, the support vector number of the model was 47, and the predicted results of the linear-kernel SVM were shown in Table 1 and Figure 10.

4.4. Compare the Results Obtained by Gaussian-Kernel SVM with Those by Other Methods. To test the suitability of the QSPR approach that was proposed by us, we have compared the obtained BDE with those calculated in ref 11. The predicted BDE and the relative error obtained in this reference were also listed in Table 1. From Table 1, it can be seen that the overall data set in the above reference gave the maximum absolute relative error of 5.24% and the average absolute relative error of 1.32%, while our Gaussian-kernel SVM model gave 2.59% and 0.82%, respectively. The maximum absolute relative error and the average absolute relative error of our work were much lower. Table 4 showed the statistical parameters of the results obtained from the two studies for the same set of compounds. The model proposed in the reference contained seven descriptors which were

Table 4. Statistics of Different QSPR Methods of O−H BDE Prediction

		training set		test set		whole set					
method	X_n[a]	rms	R	rms	R	rms	R	F-test	Sig	t-test	Sig
ref 11	7	6.225	0.948	5.425	0.953	6.070	0.947	659.512	0.000	25.681	0.000
MLR	6	6.786	0.937	6.556	0.930	6.739	0.934	520.551	0.000	22.816	0.000
RBFNNs	6	6.112	0.950	6.238	0.934	6.138	0.946	642.858	0.000	25.355	0.000
L−K SVM[b]	6	8.867	0.893	6.218	0.937	8.392	0.897	311.444	0.000	17.648	0.000
G−K SVM[c]	6	3.808	0.981	3.320	0.979	3.713	0.981	1893.832	0.000	43.518	0.000

[a] Number of descriptors used in the model. [b] Linear-kernel SVM. [c] Gaussian-kernel SVM.

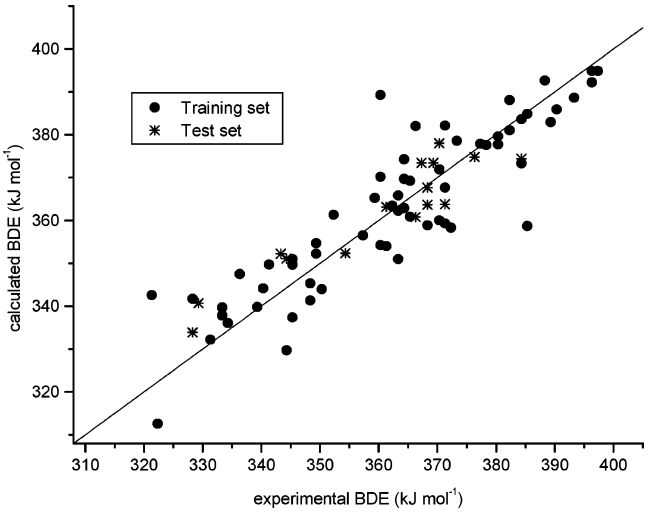

Figure 10. Plot of calculated vs experimental BDE (kJ mol^{-1}) (linear-kernel SVM).

selected by the heuristic multilinear regression procedure in the CODESSA program, while our model contained six descriptors which were selected by the stepwise regression procedures in the SPSS program. Through a regression analysis on the experimental BDE and the calculated BDE obtained by the different methods for the whole data set, the results of the F-test and the t-test were obtained and also shown in Table 4. From Table 4, it can be seen that the Gaussian-kernel SVM model gives the lowest rms error and the highest F and t values, so this model gives the most satisfactory results, compared with the results obtained from ref 11, MLR, RBFNNs, and linear-kernel SVM methods. Consequently, this SVM approach currently constitutes the most accurate method to predict the BDE of phenols.

5. CONCLUSION

A new method for the prediction of the O−H bond dissociation energy for a set of 78 substituted phenols based on support vector machines using descriptors calculated from the molecular structure alone were developed. Very satisfactory results were obtained with the proposed method. The models proposed could also provide some insight into what structural features are related to the O−H BDE of substituted phenols. Additionally, using Gaussian-kernel SVM produced even better nonlinear models with good predictive ability than another QSPR approach such as MLR, RBFNNs, linear-kernel SVM, and the method proposed by Bosque et al. This study of the QSPR model shows that the SVM proved to be a very promising tool in the prediction of O−H BDE. The training procedure is also simple when using SVM because there are fewer parameters having to be optimized, and only support vectors (only a fraction of all data) are used in the generalization process. Besides, the SVM exhibits the better whole performance due to embodying the Structural Risk Minimization principle and some advantages over the other techniques of converging to the global optimum and not to a local optimum. Furthermore, the proposed approach can also be extended in another QSPR or QSAR investigation.

ACKNOWLEDGMENT

The authors thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Program PRA SI 02-03). The authors also thank the R Development Core Team for affording the free R1.7.1 software.

REFERENCES AND NOTES

(1) Denisov, E. T.; Khudyakov, I. V. Mechanisms of action and reactivities of the free radical of inhibitors. *Chem. Rev.* **1987**, *87*, 1313−1357.
(2) Halliwell, B.; Gutteridge, J. M. C. *Free Radicals in Biology and Medicine*, 2nd ed.; Oxford University Press: Oxford, 1989.
(3) Hine, J. *Structural Effects on Equilibria in Organic Chemistry*; Wiley-Interscience: New York, 1975.
(4) Borges dos Santos, R. M.; Martinho Simóes, J. A. Energetics of the O−H bond in phenol and substituted phenols: a critical evaluation of literature data. *J. Phys. Chem. Ref. Data* **1998**, *27*, 707−739, and references therein.
(5) Lind, J.; Shen, X.; Eriksen, T. E.; Merényi, G. The one-electron reduction potential of 4-substituted phenoxyl radicals in water. *J. Am. Chem. Soc.* **1990**, *112*, 479−482.
(6) Bordwell, F. G.; Cheng, J.-P. Substituent effects on the stabilities of phenoxyl radicals and the acidities of phenoxyl radical cations. *J. Am. Chem. Soc.* **1991**, *113*, 1736−1743.
(7) Chandra, A. K.; Uchimaru, T. The O−H bond dissociation energies of substituted phenols and Proton affinities of substituted phenols ions: a DFT study. *Int. J. Mol. Sci.* **2002**, *3*, 407−422.
(8) Wright, J. S.; Carpenter, D. J.; McKay, D. J.; Ingold, K. U. Theoretical calculation of substituent effects on the O−H bond strength of phenolic antioxidants related to vitamin E. *J. Am. Chem. Soc.* **1997**, *119*, 4245−4252.
(9) Haandel, M. J. H.; Rietjens, I. M. C. M.; Soffers, A. E. M. F.; Veeger, C.; Vervoot, J.; Modi, S.; Mondal, M. S.; Patel, P. K.; Behere, D. V. Computer calculation-based quantitative structure−activity relationships for the oxidation of phenol derivatives horseradish peroxidase compound II. *J. Biol. Inorg. Chem.* **1996**, *1*, 460−467.
(10) Hansch, C.; Gao, H. Comparative QSAR: Radical Reactions of Benzene Derivatives in Chemistry and Biology. *Chem. Rev.* **1997**, *97*, 2995−3059.
(11) Bosque, R.; Sales, J. A QSPR Study of O−H Bond Dissociation Energy in Phenols. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 637−642.
(12) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery: have they lived up to their promise? *Eur. J. Med. Chem.* **1999**, *34*, 95−208.
(13) Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
(14) Ng, J.; Gong, S. G. Composite support vector machines for detection of faces across views and pose estimation. *Image Vision Comput.* **2002**, *20*, 359−368.
(15) Liu, H. X.; Zhang, R. S.; Luan, F.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Diagnosing breast cancer based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 900−907.

(16) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine leaning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(17) 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-1 and NF−KB Mediated Gene Expression Based on Support Vector Machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288−1296.
(18) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Training Manual;* University of Florida, Gainesville, FL, 1995.
(19) HyperChem 4.0, Hypercube, Inc. 1994.
(20) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange;* QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
(21) Yao, X. J.; Wang, Y. W.; Zhang, X. Y.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217−225.
(22) Orr, M. J. L. *Introduction to Radial basis function networks, center for cognitive science*; Edinburgh University: 1996.
(23) Orr, M. J. L. *MATLAB routines for subset selection and ridge regression in linear neural networks, Center for cognitive science*; Edinburgh University: 1996.
(24) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2*(2), 1−47.
(25) Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer: Berlin, 1982.
(26) Smola, A. J.; Schölkopf, B. *A tutorial on support* V*ector regression*; NeuroCOL2 Technical report series, NC2-TR-1998-030; October, 1998.
(27) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(28) Burges, C. J. C. A tutorial of support vector machines for pattern recognition. http://svm.research.bell-labs.com/SVMdoc.html, 1998.
(29) Vapnik, V.; Golowich, S.; Smola, A. Support Vector Method for function approximation, regression estimation, and signal processing. *Adv. Neural Inform. Process. Systems 9* **1997**, 281−287.

CI034248U