# Quantitative Prediction of log*k* of Peptides in High-Performance Liquid Chromatography Based on Molecular Descriptors by Using the Heuristic Method and Support Vector Machine

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

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

Received March 29, 2004

A new method support vector machine (SVM) and the heuristic method (HM) were used to develop the nonlinear and linear models between the capacity factor (log*k*) and seven molecular descriptors of 75 peptides for the first time. The molecular descriptors representing the structural features of the compounds only included the constitutional and topological descriptors, which can be obtained easily without optimizing the structure of the molecule. The seven molecular descriptors selected by the heuristic method in CODESSA were used as inputs for SVM. The results obtained by SVM were compared with those obtained by the heuristic method. The prediction result of the SVM model is better than that of heuristic method. For the test set, a predictive correlation coefficient $R = 0.9801$ and root-mean-square error of 0.1523 were obtained. The prediction results are in very good agreement with the experimental values. But the linear model of the heuristic method is easier to understand and ready to use for a chemist. This paper provided a new and effective method for predicting the chromatography retention of peptides and some insight into the structural features which are related to the capacity factor of peptides.

## 1. INTRODUCTION

Peptides belong to the most important biologically active substances. Acting as hormones, neurotransmitters, immunomodulators, coenzymes, enzyme substrates and inhibitors, receptor ligands, drugs, toxins, and antibiotics play a significant role in controlling and regulating many vitally important processes in living organisms. In addition, to understanding living cell functioning, a comprehensive investigation of the whole peptide set of a cell (peptidome) − peptidomics − will be necessary.[1,2] Consequently, separation and analysis of peptides are becoming more and more important.

Among the numerous separation techniques, chromatography has played a major role in understanding the mechanisms involved in the analysis and metabolism of proteins. Liquid chromatography can not only provide a rapid, sensitive, and very selective method of analysis of proteins but also permits the isolation or purification of most protein samples in the range of production most convenient for biochemistry or pharmaceutical studies, between a few nanograms and a few kilograms. It is very important because in vivo assays are often required to ascertain that the proper protein has been identified and analyzed. At the same time, an ideal separation method of biopolymer should be able to give high-purity material with a high production rate and a good recovery yield, which includes total conservation of their biological activity at any rate. High-performance liquid chromatography (HPLC) is the one of the separation methods which meets all these different requirements. It has been widely used for analytical separations. Its preparative and process applications are undergoing rapid development for the isolation and purification of peptides and proteins.[3]

Despite the ever increasing usage of HPLC for the separation and analysis of peptides and proteins, selection of chromatographic conditions for purification of a given peptide is still found by time-consuming trial and error methods. A priori knowledge of the retention time of a given peptide would simplify the selection of chromatographic conditions. At present, prediction of the retention behavior of peptides is mainly based on the amino acid composition.[4−8] However, using this method, some experiments for the standard samples must be performed in order to derive the group retention coefficients of the amino acid in the given conditions, which is still time-consuming and is difficult to generalize the calculated results.

Alternatively, quantitative structure−property relationship (QSPR) provides a promising method for the estimation of compounds' chromatographic behavior based on the descriptors derived solely from the molecular structure to fit experimental data. The advantage of this approach over other methods lies in the fact that it requires only the knowledge of chemical structure and is not dependent on any experiment properties.

The QSPR approach has become very useful in the prediction of physicochemical properties. This approach is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with changes in molecular features of the compounds termed descriptors.[9] The main steps involved in QSPR include the following: data

* Corresponding author phone: +86-931-891-2578; fax: +86-931-891-2582; e-mail: liuhx1003@yahoo.com.cn.
[†] Department of Chemistry, Lanzhou University.
[‡] Department of Computer Science, Lanzhou University.
[§] Université Paris 7-Denis Diderot.

10.1021/ci049891a CCC: $27.50 
Published on Web 08/31/2004

**Table 1.** Linear Model between the Structure and the log*k* of Peptides[a]

| descriptor | coefficient | error | t-test value |
|---|---|---|---|
| intercept | 3.8411 | 0.8732 | 4.3992 |
| average complementary information content (order 1) (ASIC1) | 0.3314 | 0.0421 | 7.8748 |
| relative number of single bonds | −6.5586 | 0.4850 | −13.5229 |
| relative number of s atoms | 27.3000 | 5.0266 | 5.4312 |
| average information content (order 2) (AIC2) | 0.9198 | 0.1055 | 8.7150 |
| relative number of N atoms | −11.3550 | 1.9517 | −5.8181 |
| number of rings | −0.0967 | 0.0228 | −4.2347 |
| average information content (order 0) (AIC0) | −0.9333 | 0.3961 | −2.3563 |

[a] $R^2 = 0.9051$, $R^2_{cv} = 0.8677$, F = 89.91, $s^2 = 0.0341$, N = 75.

collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development, and finally model performance evaluation.[10] This study can develop a method for the prediction of the property of new compounds that have not been synthesized or found. It can also identify and describe important structural features of the molecules that are relevant to variations in molecular properties, thus gaining some insight into the structural factors affecting the molecular properties. Although QSPR methods have been successfully used to predict many physicochemical properties, no research group has investigated the quantitative correlation between the structural parameters and the chromatographic retention of peptides, which might be due to the optimization of the structures of the peptides which is very time-consuming because in most of the cases, the size of the peptides is very large.

Artificial intelligence techniques have been applied to QSPR analysis since the late 1980s, mainly in response to increased accuracy demands.[11] Machine learning techniques have, in general, offered greater accuracy than have their statistical forebears, but there exist accompanying problems for the QSPR analyst to consider. Neural networks, for example, offer high accuracy in most cases but can suffer from the reproducibility of results, due largely to random initialization of the network and variation of stopping criteria and lack of information regarding the classification produced.[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 machine (SVM) is a new algorithm developed from the machine learning community. Due to its remarkable generalization performance, the SVM has gained much attention and extensive applications.[11,14−22]

Another important problem for the QSPR applications is the numerical representation (often called molecular descriptor) of the chemical structure. The built model performance and the accuracy of the results are strongly dependent on the way the structural representation was performed. Various numerical representations of organic compounds were proposed in QSPR studies: constitutional descriptors and topological descriptors; numerical code; quantum chemistry descriptors, etc. (Katritzky et al., 1995). The Software CODESSA[23,24] developed by the Katritzky group, which can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors, has been successfully used in various QSPR researches.

In the present investigation, for the first time, SVM and the heuristic method were used for the prediction of the capacity factor (log*k*) of 75 peptides using descriptors calculated by the software CODESSA as inputs. The aim was to explore the retention behavior of peptides in high-performance liquid chromatography, to establish a new quantitative structure-retention model, and to confirm the possibility of predicting retention behavior of peptides and, at the same time, to seek the structural factor affecting their retention behavior. The prediction results are very satisfactory in both training set and test set compounds, which proved SVM to be a useful tool in the prediction of the capacity factor (log*k*). Moreover, in this article, only the constitutional and topological descriptors were calculated without the need to optimize the structures of the peptides. This can avoid effectively the main problem among the application of the QSPR methods in the prediction of the properties of the peptides especially for those with a large size because the structural optimizing of peptides is very time-consuming as mentioned above.

## 2. EXPERIMENT

The sequences and the capacity factors of the studied peptides were collected from ref 25 and given in Table 1. The peptides were obtained by the enzymic degradation of calmodulin, Bene Jones proteins, lysozyme, and other proteins that had already been sequenced.

The column packed with carbonex porous microspherical carbon beads (were obtained from Biotech Research (Saitama, Japan), average particle size, 3.5 $\mu$m; specific surface area, >30 $m^2/g$; specific pore volume, 0.35 mL/g; apparent density, 0.57 g/mL; pore-size range, 10−700 Å) was used in this system. Sample peptides and proteins were applied to the column and eluted with a linear 30-min gradient from 10% to 70% acetonitrile in 0.1% aqueous trifluoroacetic acid at a flow-rate of 1.0 mL/min, with absorbance detection at 210 nm. The operating temperature was room temperature.

Most peptides were eluted under this condition. In this present investigation, the peptides which can be eluted are composed of standard amino acid and whose capacity factors are more than zero were selected as our research objects.

## 3. COMPUTATIONAL METHODS

**3.1. Calculation and Selection of the Descriptors.** To obtain a QSRR model, compounds are often represented by the molecular descriptors. The calculation process of the molecular descriptors is described as below: the three-dimensional structures of the peptides were drawn using the sequence editor of Hyperchem and saved as the hin files.

Then their hin files were transferred into software CODESSA, developed by the Katritzky group,[23,24] to calculate constitutional and topological descriptors, which has been successfully used in various QSPR/QSAR researches. The structural optimization of the large size peptides, which is a kind of biopolymer, is a challenging task, since the optimization is very time-consuming. In the present work, only the constitutional and topological descriptors were calculated because their calculation does not need the optimization of the molecular structure and is high-speed. Constitutional descriptors are related to the number of atoms and bonds in each molecule. Topological descriptors include valence and nonvalence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition, and the degree of branching of a molecule.

Once molecular descriptors are generated, the heuristic method in CODESSA was used to accomplish the preselection of the descriptors. Its advantages are the high-speed and no restrictions on the size of the data set. The heuristic method can either quickly give a good estimation about what quality of correlation to expect from the data or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which descriptors are insignificant (from the standpoint of a single-parameter correlation), and which descriptors are highly intercorrelated. This information will be helpful in reducing the number of descriptors involved in the search for the best QSPR model.

First of all, all descriptors are checked to ensure (a) that values of each descriptor are available for each structure and (b) that there is a variation in these values. Descriptors for which values are not available for every structure in the data in question are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested, and insignificant descriptors are removed. In the next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. All two-parameter regression models with remaining descriptors are subsequently developed and ranked by the regression correlation coefficient $R^2$. A stepwise addition of further descriptor scales is performed to find the best multiparameter regression models with the optimum values of statistical criteria (highest values of $R^2$, the cross-validated $R^2_{cv}$, and the $F$-value).

**3.2. Methodology.** After the descriptors are selected, the next step is to build the quantitative model by using some computational methods. In this work, the heuristic method and support vector machines were used to build the linear and nonlinear models for the prediction of *logk* of peptides, respectively. As the theory of the heuristic method has been well described in many monographs and articles, we only give a brief description on the theory of the SVM for regression.

**3.2.1. Theory of SVM for Regression.**[26,27] The support vector machine, developed by Vapnik,[28] as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. Compared to traditional neural networks, SVM possesses the following prominent advantages: (1) A strong theoretical background provides SVM with high generalization capability and can avoid local minima. (2) SVM always has a solution, which can be quickly obtained by a standard algorithm (quadratic programming). (3) SVM need not determine network topology in advance, which can be automatically obtained when the training process ends. (4) SVM builds a result based on a sparse subset of training samples, which reduce the workload. Originally, SVM are developed for pattern recognition problems, such as image recognition,[30] microarray gene expression classification,[11] protein folding recognition,[31] protein structural class prediction,[32] identification of protein cleavage sites, QSAR, and other pharmaceutical data analysis,[11,33] and now, with the introduction of $\epsilon$-insensitive loss function, SVM have been extended to solve nonlinear regression estimation and time-series prediction and excellent performances have been obtained.[29]

In SVM, the basic idea is to map the data $x$ into a higher-dimensional feature space F via a nonlinear mapping $\Phi$ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}_{i=1}^{l}$ ($x_i$ is input vector, $d_i$ is the desire $d$ value). SVM approximate the function in the following form

$$y = \sum_{i=1}^{l} w_i \Phi(x_i) + b \tag{1}$$

where $\{\Phi(x_i)\}_{i=1}^{l}$ are the features of inputs, and $\{\omega_i\}_{i=1}^{l}$ and $b$ are coefficients. They are estimated by minimizing the regularized risk function (2)

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\epsilon(d_i,y_i) + \frac{1}{2}|w|^2 \tag{2}$$

where

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

and $\epsilon$ is a prescribed parameter.

In eq 2, $C(1/N)\sum_{i=1}^{N} L_\epsilon(d_i,y_i)$ is the so-called empirical error (risk), which is measured by $\epsilon$-insensitive loss function $L_\epsilon(d,y)$, which indicates that it does not penalize errors below $\epsilon$. The second term, $1/|w|^2$, is used as a measurement of function flatness. $C$ is a regularization constant determining the tradeoff between the training error and the model flatness. Introduction of slack variables "$\xi$" leads eq 2 to the following constrained function:

$$\text{Max } R(w,\xi_i,\xi^*_i) = \frac{1}{2}|w|^2 + C^*\sum_{i=1}^{n}(\xi_i + \xi^*_i) \tag{4}$$

subject to

$$w\Phi(x_i) + b - d_i \leq \epsilon + \xi^*_i$$

$$d_i - w\Phi(x_i) - b \leq \epsilon + \xi_i$$

$$\xi_i, \xi_i^{**} \geq 0 \tag{5}$$

Thus, decision function (1) becomes the following form

$$f(x,\alpha_i,\alpha_i^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i,x_j) + b \qquad (6)$$

In eq 6, $\alpha_i,\alpha_i^*$ are the introduced Lagrange multipliers. They satisfy the equality $\alpha_i \cdot \alpha_i^* = 0, \alpha_i \geq 0, \alpha_i^* \geq 0$; $i = 1,\cdots,l$ and are obtained by maximizing the dual form of function (4); which has the following form:

$$\Phi(\alpha_i,\alpha_i^*) = \sum_{i=1}^{l} d_i(\alpha_i - \alpha_i^*) - \epsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\alpha_i,\alpha_j) \qquad (7)$$

subject to

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$

$$0 \leq \alpha_i \leq C,\ i = 1,\cdots,l$$

$$0 \leq \alpha_i^* \leq C,\ i = 1,\cdots,l \qquad (8)$$

Based on the Karush-Kuhn-Tucker (KKT) conditions of quadratic programming, only a number of coefficients ($\alpha_i - \alpha_i^*$) will assume nonzero values, and the data points associated with them could be referred to as support vectors.

In eq 6, $K(x_i,x_j)$ is the kernel function. The value is equal to the inner product of two vectors $x_i$ and $x_j$ in the feature space $\Phi(x_i)$ and $\Phi(x_j)$, that is, $K(x_i,x_j) = \Phi(x_i)\cdot\Phi(x_j)$. The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function. In support vector regression, the Gaussian kernel $K(u,v) = \exp(-|u - v|^2/\delta^2)$ is commonly used.

**3.2.2. SVM Implementation and Computation Environment.** All calculation programs implementing SVM were written in R-file based on R script for SVM. All scripts were compiled using R1.7.1 compiler running operating system on a Pentium IV with 256M RAM.

## 4. RESULTS AND DISCUSSION

**4.1. The Heuristic Method Model.** Through the heuristic method, the best linear model with seven descriptors was obtained, which was shown in Table 1. By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the chromatographic retention of peptide on a microspherical carbon column. It is generally agreed that three types of intermolecular interactions are the main factors influencing the retention of solute:[34] (a) polar interactions from permanent or induced dipoles between solute, stationary-phase, and mobile phase molecules; (b) dispersive interactions; (c) hydrogen bond interactions [Considering that the acetonitrile-trifluoroacetic acid mobile phase is liable to form a hydrogen bond with the oxygen and nitrogen atoms in the residues, the hydrogen bond must be taken into account]; (d) steric interactions between the solute and the stationary phase.

In the linear model, there are four constitutional descriptors and three topological descriptors. According to the t-test (Table 1), the most important descriptor affecting the retention of the peptides is a constitutional descriptor, the relative number of single bonds. The relative number of single bonds affects the density of the electron cloud of the molecule. The larger the relative number of single bonds is, the lower the density of the electron cloud of the molecule is and the weaker the polar interaction between the solute and mobile phase is. Thus, an increase in this descriptor leads to a decrease in the capacity factor of the compound. The relative number of single bonds is also related to the rigidity of the molecule. Generally, the flexibility of the molecule increases as the relative number of single bonds increases, and then resistance of the solute through the solvent will decrease, which leads to the decrease in the capacity factor. Therefore, the regression coefficient of this descriptor is negative.

The relative number of N atoms correlates with the ability of forming hydrogen bonds between the solute and mobile phase. As this value increases, the tendency of forming a hydrogen bond between the solute and mobile phase increases, leading to a decrease in a value of the capacity factor.

The positive coefficient for the number of sulfur atoms indicates an increase in the value of this descriptor which leads to an increase in the value of log*k*. The negligible difference of electronegativity between sulfur and carbon, about 0.03 units in the Pauling scale,[35] leads to relatively small bond dipoles disfavoring the solute−solvent interactions and therefore increasing the dispersive interactions with the stationary phase and consequently increasing the value of log*k*.

In this data set, the number of rings is the same as the number of aromatic rings. The special mobility of $\pi$ electrons will result in an enhanced polarizability and the interaction of unsaturated molecules with the mobile phase and therefore favors the elution process. Furthermore, the number of rings encodes the hydrophobicity of the compound, thus, an increase in this descriptor strengthens the hydrophobicity of the molecule, enhances the interaction between the solute and stationary phase, and then disfavors the elution process. Both these interactions can lead to a decrease in the value of log*k* on the whole.

The topological descriptors, the average complimentary information content, and average information content are defined on the basis of the Shannon information theory. They can be calculated for different orders of neighborhoods, $r$ ($r = 0, 1, 2, ...,\rho$), where $\rho$ is the radius of the molecular graph $G$. At the zero-order level, the atom set is partitioned solely on the basis of its chemical nature; at the level of the first-order topological neighborhood, the atoms are partitioned into disjoint subsets on the basis of their chemical nature and their first-order bonding topology. At the next level, the atom set is decomposed into equivalence classes using their chemical nature and bonding pattern up to the second-order bonded neighbors.[36] The three topological indexes average complimentary information content (order 1), average information content (order 0), and average information content (order 2) reflect the branching of the molecule and reflect how information rich the molecule is. "Information rich" describes how many different atoms there are in the molecule and how diverse the branching of these atoms is at zero to second valence level (coordination sphere). In other words,

they represent the difference between the maximum possible complexity of a graph and the realized topological information of the chemical species as defined by the information content. Therefore, they can describe the difference of the hydrophobicity and steric property of the solute comprehensively. As the hydrophobic and steric interaction is the main interaction between the solute and the stationary phase, these three topological descriptors play an important role in the elution process and have high correlation with the log*k*.

From the above discussion, these descriptors can account for the structural features responsible for the capacity factor of peptides in the certain condition. The calculated and experimental values of the logarithm of the capacity factor by the heuristic method were given in Table 2, and the scatter plot was shown in Figure 1. The root-mean-square error of this model is 0.1733, and the prediction correlation coefficient is 0.9513, respectively. From Table 1 and Figure 1, it can be seen that the model of MLR was not sufficiently accurate ($R^2 = 0.9051$, $R^2_{cv} = 0.8677$, $s^2 = 0.0341$), and for the separation system of high performance liquid chromatography, the factors influencing resolution were complex and not all of them were linear correlation with the chromatographic behavior. The nonlinear correlation model by SVM was used further to discuss the correlation between the molecular structure and the log*k*.

**4.2. Result of SVM. 4.2.1. 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 $C$, $\epsilon$ of the $\epsilon$-insensitive loss function, the kernel type K, and its 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. But, ref 29 indicated that the prediction error was scarcely influenced by C. To make the learning process stable, a large value should be set up for $C$ (e.g., $C = 100$).

The optimal value for $\epsilon$ 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 $\epsilon$, there is the practical consideration of the number of resulting support vectors. $\epsilon$-insensitivity 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 $\epsilon$ is critical from theory.

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

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

where $\gamma$ is a constant, the parameter of the kernel; $u$ and $v$ are two independent variables; and $\gamma$ controls the amplitude of the Gaussian function and, therefore, controls the generalization ability of SVM. We have to optimize $\gamma$ and find the optimal one.

To find the optimum values of two parameters ($\gamma$ and $\epsilon$) and prohibit the overfitting of the model, the data set was separated into a training set of 57 compounds and a test set of 18 compounds randomly, and the leave-one-out cross-validation of the whole training set was performed. The leave-one-out (LOO) procedure consists of removing one example from the training set, constructing the decision function on the basis only of the remaining training data, and then testing on the removed example. In this fashion one tests all examples of the training data and measures the fraction of errors over the total number of training examples. The root-mean-square error (RMS) was used as an error function which was defined as below

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(d_i - o_i)^2}{n}}$$

where $d_i$ are the teaching outputs (desired outputs) in the training set, $o_i$ are the actual outputs obtained from the leave-one-out cross-validation method, and $n$ is the number of the samples in the training set.

Detailed process of selecting the parameters and the effects of every parameter on generalization performance of the corresponding model were shown in Figures 2 and 3. To obtain the optimal $\gamma$, the support vector learning machines with different $\gamma$ were trained, the $\gamma$ varying from 0.002 to 0.02, every 0.001. We calculated the rms on different $\gamma$, according to the generalization ability of the model based on the LOO cross-validation for the training set in order to determine the optimal one. The curve of rms versus the gamma was shown in Figure 2. The optimal $\gamma$ was found as 0.011. To find an optimal $\epsilon$, the RMS on different $\epsilon$ was calculated. The curve of the RMS versus the epsilon was shown in Figure 3. From Figure 3, the optimal $\epsilon$ was found as 0.001.

**4.2.2. The Predicted Result of SVMs.** From the above discussion, the $\gamma$, $\epsilon$, and $C$ were fixed to 0.011, 0.001, and 100, respectively. The predicted results of the optimal SVM were shown in Table 2 and Figure 4. As can be seen from Figure 4, 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 listed in Table 2. The rms error of the training set, the test set, and the whole set is 0.1324, 0.1523, and 0.1374, and the prediction correlation coefficient is 0.9727, 0.9801, and 0.9714, respectively. It can be concluded that the predicted values are in very good agreement with the experimental values. By comparing results from the heuristic method and SVM, it can be seen that the SVM model has better predicting ability. At the same time, the complexity of the nonlinear model will not increase comparing with that of the linear model since the SVM method always can obtain the solution quickly by applying a standard optimization algorithm (quadratic programming) and only use a sparse subset of training samples.

Analysis of the results obtained indicates that the model we proposed can correctly represent structure-retention relationships of these compounds, and molecular descriptors calculated solely from structures could describe the structural features of the compounds responsible for their capacity factor. From the selected parameters, it can be seen that the capacity factor of peptides on the microspherical carbon column are mainly determined by several intermolecular

**Table 2.** Predicted Results Using SVM and Heuristic Method

| | | | results of SVM | | results of HM | |
|---|---|---|---|---|---|---|
| no. | peptide (sequence) | exp. | pred. | residue | pred. | residue |
| 1 | LI | −0.6990 | −0.6985 | 5E−4 | −0.1375 | 0.5615 |
| 2[a] | AS | −0.6990 | −0.8149 | −0.1159 | −0.8525 | −0.1535 |
| 3 | GV | −0.6021 | −0.7017 | −0.0996 | −0.8102 | −0.2081 |
| 4 | LG | −0.5229 | −0.5236 | −7E−4 | −0.4998 | 0.0231 |
| 5 | APK | −0.5229 | −0.3669 | 0.156 | −0.2474 | 0.2755 |
| 6 | PGK | −0.5229 | −0.5232 | −3E−4 | −0.3984 | 0.1245 |
| 7 | HK | −0.5229 | −0.1467 | 0.3762 | −0.0778 | 0.4451 |
| 8 | PSK | −0.4559 | −0.4407 | 0.0152 | −0.2976 | 0.1583 |
| 9 | QSNNK | −0.4559 | −0.3659 | 0.09 | −0.0573 | 0.3986 |
| 10[a] | AV | −0.4437 | −0.713 | −0.2693 | −0.7949 | −0.3512 |
| 11 | C | −0.1611 | −0.1619 | −8E−4 | −0.1216 | 0.0395 |
| 12 | PF | 0.2253 | 0.4144 | 0.1891 | 0.5491 | 0.3238 |
| 13[a] | MK | 0.4048 | −0.003 | −0.4078 | 0.2599 | −0.1449 |
| 14 | VDIK | 0.4654 | 0.6291 | 0.1637 | 0.4335 | −0.0319 |
| 15 | TPGSP | 0.5132 | −0.0242 | −0.5374 | 0.1509 | −0.3623 |
| 16[a] | HAVE | 0.5490 | 0.5279 | −0.0211 | 0.4676 | −0.0814 |
| 17 | ADSSPVK | 0.5922 | 0.4523 | −0.1399 | 0.5233 | −0.0689 |
| 18 | F | 0.6170 | 0.6163 | −7E−4 | 0.5874 | −0.0296 |
| 19 | AGVETTK | 0.6180 | 0.7328 | 0.1148 | 0.6021 | −0.0159 |
| 20[a] | KNSISPE | 0.6335 | 0.6101 | −0.0234 | 0.6776 | 0.0441 |
| 21 | ADGSPVK | 0.6415 | 0.5637 | −0.0778 | 0.6107 | −0.0308 |
| 22 | AGVETTTPSK | 0.6415 | 0.7316 | 0.0901 | 0.7811 | 0.1396 |
| 23[a] | VTALSQPK | 0.6794 | 0.769 | 0.0896 | 0.6878 | 0.0084 |
| 24 | DGDGTITTK | 0.6875 | 0.7719 | 0.0844 | 0.7501 | 0.0626 |
| 25 | HASLEKPKDE | 0.7059 | 0.8634 | 0.1575 | 0.9226 | 0.2167 |
| 26 | Y | 0.7101 | 0.7105 | 4E−4 | 0.6428 | −0.0673 |
| 27 | GF | 0.7292 | 0.7296 | 4E−4 | 0.6629 | −0.0663 |
| 28 | VFDK | 0.7679 | 0.944 | 0.1761 | 0.8363 | 0.0684 |
| 29 | TFKRD | 0.8129 | 0.8121 | −8E−4 | 0.7973 | −0.0156 |
| 30[a] | LTVLRQPK | 0.8274 | 0.9036 | 0.0762 | 0.7126 | −0.1148 |
| 31 | YINEHK | 0.8669 | 1.0939 | 0.227 | 1.1076 | 0.2407 |
| 32 | ADYEK | 0.8704 | 0.8712 | 8E−4 | 0.8848 | 0.0144 |
| 33 | LTVLGQPK | 0.8733 | 0.8729 | −4E−4 | 0.7348 | −0.1385 |
| 34 | VFGR | 0.8876 | 0.8873 | −3E−4 | 0.6867 | −0.2009 |
| 35 | EAFR | 0.8887 | 0.8698 | −0.0189 | 0.734 | −0.1547 |
| 36 | YVLNKHNE | 0.9106 | 1.0222 | 0.1116 | 1.0442 | 0.1336 |
| 37 | RVY | 0.9196 | 0.7844 | −0.1352 | 0.6231 | −0.2965 |
| 38 | LY | 0.9243 | 0.6805 | −0.2438 | 0.7743 | −0.15 |
| 39 | DGHAHSHLIQQHIEK | 0.9400 | 0.9404 | 4E−4 | 1.0319 | 0.0919 |
| 40 | ELGTVMR | 0.9415 | 0.9409 | −6E−4 | 1.0379 | 0.0964 |
| 41 | VDNALQSGNSQESVTEQDSK | 0.9479 | 0.9483 | 4E−4 | 1.1392 | 0.1913 |
| 42 | VPVVFVKKE | 0.9567 | 0.9559 | −8E−4 | 0.8788 | −0.0779 |
| 43 | VKDGHAHSHLIQQHIE | 0.9590 | 0.9794 | 0.0204 | 1.0471 | 0.0881 |
| 44[a] | HGLDNYR | 0.9881 | 0.9991 | 0.011 | 1.0891 | 0.101 |
| 45 | ISRGQHKYEPE | 0.9983 | 0.9991 | 8E−4 | 1.0891 | 0.0908 |
| 46 | HVLFGGGTK | 0.9987 | 1.098 | 0.0993 | 1.0358 | 0.0371 |
| 47 | KLSGHIYE | 1.0174 | 1.2728 | 0.2554 | 1.2562 | 0.2388 |
| 48[a] | TWGVTKAAELQ | 1.0183 | 1.1711 | 0.1528 | 1.152 | 0.1337 |
| 49[a] | VQWK | 1.0504 | 1.0799 | 0.0295 | 0.9761 | −0.0743 |
| 50[a] | AVRYINE | 1.0542 | 1.0984 | 0.0442 | 0.9635 | −0.0907 |
| 51 | DSTYSLSSTLTLSK | 1.0704 | 0.9913 | −0.0791 | 0.9981 | −0.0723 |
| 52 | DTDSEEEIR | 1.0730 | 1.0619 | −0.0111 | 1.0129 | −0.0601 |
| 53[a] | GQTLVVQFTVK | 1.0770 | 1.1258 | 0.0488 | 0.9499 | −0.1271 |
| 54 | W | 1.0874 | 1.0867 | −7E−4 | 1.2523 | 0.1649 |
| 55[a] | ANPTVTLFPPSSEELQANK | 1.0917 | 1.0333 | −0.0584 | 1.0632 | −0.0285 |
| 56 | ANPSVTLFPPSSEELQANK | 1.0924 | 1.002 | −0.0904 | 1.0434 | −0.049 |
| 57 | TWGVTKAAE | 1.0927 | 1.0629 | −0.0298 | 1.0818 | −0.0109 |
| 58 | VHVIFNYK | 1.1021 | 1.2279 | 0.1258 | 1.1452 | 0.0431 |
| 59 | GW | 1.1106 | 1.111 | 4E−4 | 1.0725 | −0.0381 |
| 60 | AKNWADD | 1.1193 | 0.8685 | −0.2508 | 0.9574 | −0.1619 |
| 61 | IHPF | 1.1238 | 1.1247 | 9E−4 | 1.0277 | −0.0961 |
| 62[a] | WKPRQIDNPE | 1.1483 | 1.0758 | −0.0725 | 1.0201 | −0.1282 |
| 63 | DPTVYFK | 1.1670 | 1.1662 | −8E−4 | 1.1998 | 0.0328 |
| 64[a] | NTDGSTDYGILQINSR | 1.1709 | 1.1593 | −0.0116 | 1.285 | 0.1141 |
| 65 | EAFSLFDKDGDGTITTK | 1.1709 | 1.2724 | 0.1015 | 1.3659 | 0.195 |
| 66[a] | DRVYIHPFHL | 1.1775 | 1.4758 | 0.2983 | 1.2082 | 0.0307 |
| 67 | HHQEHPYTAGE | 1.1844 | 1.1839 | −5E−4 | 1.0739 | −0.1105 |
| 68 | VKIDNSQVE | 1.1853 | 1.0637 | −0.1216 | 0.8629 | −0.3224 |
| 69 | NTDGSTDYGILQIN | 1.1878 | 1.0637 | −0.1241 | 0.8629 | −0.3249 |
| 70[a] | LLISDNYNRPSGVPARFSGSK | 1.2122 | 1.1099 | −0.1023 | 1.2006 | −0.0116 |
| 71 | GTDVQAWIR | 1.2460 | 1.2133 | −0.0327 | 1.161 | −0.085 |
| 72 | RTVAAPSVFIFPPSDEQLK | 1.2737 | 1.2745 | 8E−4 | 1.1645 | −0.1092 |
| 73[a] | DRVYIHPF | 1.2943 | 1.4118 | 0.1175 | 1.2365 | −0.0578 |
| 74 | VFDKDGDGYISAAELR | 1.3610 | 1.3607 | −3E−4 | 1.4076 | 0.0466 |
| 75 | RVYIHPE | 1.3705 | 1.361 | −0.0095 | 1.1594 | −0.2111 |

[a] The compounds in the test set for SVM model.

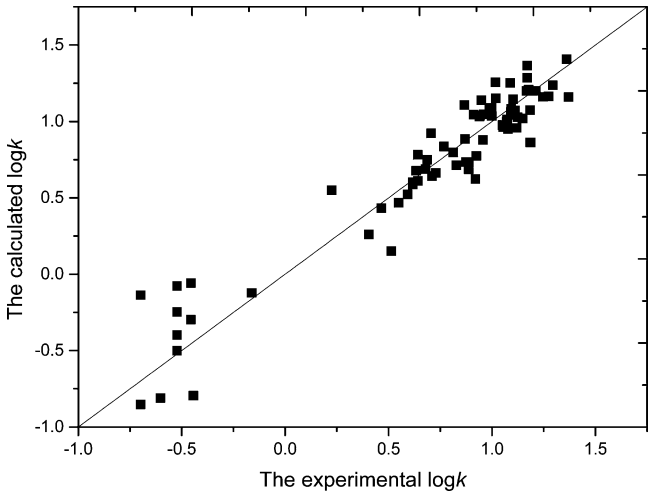


**Figure 1.** The experimental log*k* versus the calculated log*k* by heuristic method.

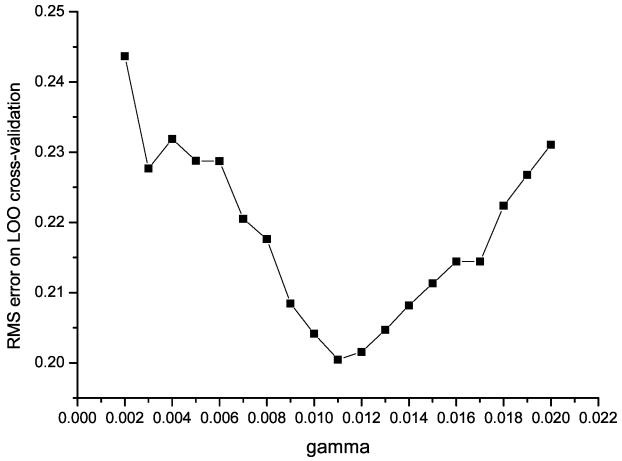


**Figure 2.** The gamma versus RMS error on LOO cross-validation ($C$=100, $\epsilon = 0.01$)

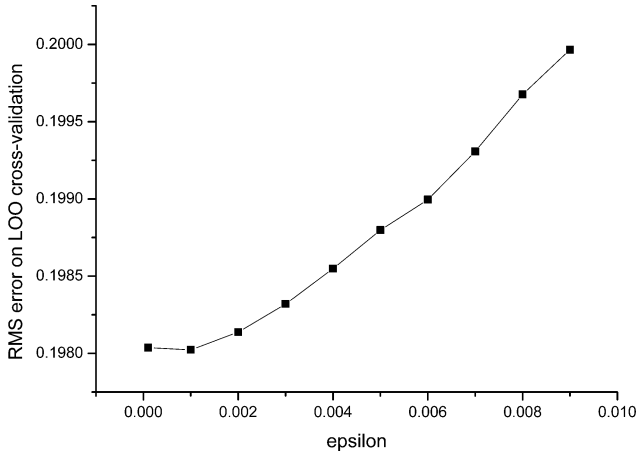


**Figure 3.** The epsilon versus RMS error on LOO cross-validation ($C$=100, $\gamma$=0.011).

interactions, such as hydrophobic and steric interactions between the solute and stationary phase and polar and hydrogen bond interactions between the solute and mobile phase. From the performance comparison of the nonlinear model and that of the heuristic method, it proved that nonlinear model can simulate the relationship between the structural descriptors and the chromatography retention of peptides more accurately.

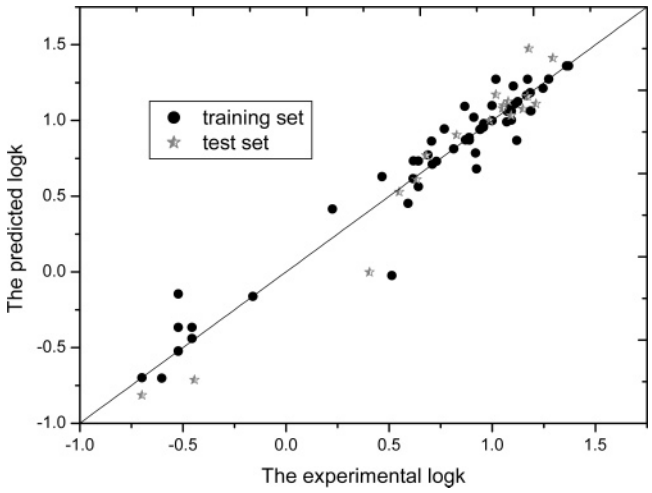


**Figure 4.** The predicted log*k* versus the experimental log*k* by SVM model.

## 5. CONCLUSION

Accurate linear and nonlinear QSRR models of 75 peptides were built based on the heuristic method and support vector machine, respectively, by using the constitutional and topological descriptors whose calculation is simple and fast. By comparing the linear and nonlinear models, it is proved that nonlinear SVM models gave better results with better predictive ability than a linear model. It can be concluded that (1) the proposed models could identify and provide some insight into structural features related to the mobility of peptides on the microspherical carbon column from the molecular level. (2) A nonlinear relationship can describe the relationship between the structural parameter and the log*k* of the 75 peptides more accurately. (3) SVM proved to be a useful tool in the prediction of the chromatography behavior of the peptides. It has some advantages over the other techniques, such as convergence to the global optimum and good generalization. Besides, because only support vectors (only a fraction of all data) are used in the generalization process, the SVM is suitable particularly to the problems with a great deal of data in cheminformatics. Furthermore, there are fewer free parameters to be adjusted in the SVM, and the model selecting process is easy to control. Therefore, the SVM is a very promising machine learning technique from many aspects and will gain more extensive application.

In summary, this investigation developed a new method to predict the chromatographic behavior of peptides and explained the factors affecting chromatographic behavior from the microcosmic perspective. It can also provide an idea for dealing with the QSAR/QSPR problem of biopolymers.

## 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) Wolters, D. A.; Washburn, M. P.; Yates, J. R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **2001**, *73*, 5683−5690.

(2) Kašička, V. Recent advances in capillary electrophoresis and capillary electrochromatography of peptides. *Electrophoresis* **2003**, *24*, 4013−4046.
(3) Huang, J. X.; Guiochon, G. Applications of preparative high-performance liquid chromatography to the separation and purification of peptides and proteins. *J. Chromatogr.* **1989**, *492*, 431−469.
(4) Guo, D.; Mant, C. T.; Taneja, A. K. and Hodges, R. S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography II correlation of observed and predicted peptide retention times and factors influencing the retention times of peptides. *J. Chromatogr.* **1986**, *359*, 519−532.
(5) Casal, V.; Martin-Alvarez, P. J.; Herraiz, Z. Comparative prediction of the retention behavior of small peptides in several reversed-phase high-performance liquid chromatography columns by using partial least squares and multiple linear regression. *Anal. Chim. Acta* **1996**, *326*, 77−84.
(6) Yoshida, T.; Okada, T. Prediction of peptide retention time in normal-phase liquid chromatography with only a single gradient run. *J. Chromatogr.* **1999**, *841*, 19−32.
(7) Palmblad, M.; Ramström, M.; Markides, K. E.; Håkansson, P.; Bergquist, J. Prediction of Chromatographic Retention and Protein Identification in Liquid Chromatography/Mass Spectrometry. *Anal. Chem.* **2002**, *74*, 5826−5830.
(8) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Paša-Tolić, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y. F.; Zhao, R. and Smith, R. D. Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. *Anal. Chem.* **2003**, *75*, 1039−1048.
(9) Yao, X. J.; Liu, M. C.; Zhang, X. Y.; Hu, Z. D.; Fan, B. T. Radial basis function network-based quantitative structure−property relationship for the prediction of Henry's law constant. *Anal. Chim. Acta* **2002**, *462*, 101−117.
(10) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(11) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S.; Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(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) Bao, L.; Sun, Z. R. Identifying genes related to drug anticancer mechanisms using support vector machine. *FEBS Lett.* **2002**, *521*, 109−114.
(15) Belousov, A. I.; Verzakov, S. A.; Von Frese J. A flexible classification approach with optimal generalization performance: support vector machines. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 15−25.
(16) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Prediction of protein structural classes by support vector machines. *Comput. Chem.* **2002**, *26*, 293−296.
(17) Morris, C. W.; Autret, A.; Boddy, L. Support vector machines for identifying organisms- a comparison with strongly partitioned radial basis function networks. *Ecological Modelling* **2001**, *146*, 57−67.
(18) Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; and Tugcu, N. Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1347−1357.
(19) 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.
(20) 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-$\kappa$B Mediated Gene Expression based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288−1296.
(21) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Prediction of Isoelectric Point of Amino Acid Based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 161−169.
(22) 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*.
(23) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Training Manual*. University of Florida, Gainesville, 1995.
(24) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual*. University of Florida, Gainesville, 1994.
(25) Yamaki, S.; Isobe, T.; Okuyama, T.; Shinoda, T. High-performance liquid chromatography of peptides on a microspherical carbon column. *J. Chromatogr. A* **1996**, *729*, 143−153.
(26) Tay, F. E. H.; Cao, L. J. Modified support vector machines in financial time series forecasting. *Neurocomputing* **2002**, *48*, 847−861.
(27) Burges, C. J. C.; A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2*, 1−47.
(28) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.
(29) Wang, W. J.; Xu, Z. B.; Lu, W. Z.; Zhang, X. Y. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing* **2003**, *55*, 643.
(30) Zhang, L.; Zhou, W. D.; Jiao, L. C. Support vector machine for 1-D image recognition. *J. Infrared Millimeter Waves* **2002**, *21*, 119−123.
(31) Ding, C. H. Q.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349−358.
(32) Karchin, R.; Karplus, K.; Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**, *18*, 147−159.
(33) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **2002**, *23*, 267−274.
(34) Xiang, Y. H.; Liu, M. C.; Zhang, X. Y.; Zhang, R. S.; Hu, Z. D. Quantitative Prediction of Liquid Chromatography Retention of N−Benzylideneanilines Based on Quantum Chemical Parameters and Radial Basis Function Neural Network. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 592−597.
(35) Delgado, E. J.; Alderete, J. B.; Jača, G. A. A Simple QSPR Model for Predicting Soil Sorption Coefficients of Polar and Nonpolar Organic Compounds from Molecular Formula. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1928−1932.
(36) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; and Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891−898.

CI049891A