Accurate Quantitative Structure−Property Relationship Model To Predict the Solubility of C_{60} in Various Solvents Based on a Novel Approach Using a Least-Squares Support Vector Machine

Huanxiang Liu,[†] Xiaojun Yao,*[,†] Ruisheng Zhang,[†,‡] Mancang Liu,[†] Zhide Hu,[†] and Botao Fan[§]

Department of Chemistry and Department of Computer Science, Lanzhou University, Lanzhou 730000, People's Republic of China, and Université Paris 7-Denis Diderot, ITODYS 1, rue Guy de la Brosse, 75005 Paris, France

Received: April 29, 2005; In Final Form: August 4, 2005

A least-squares support vector machine (LSSVM) was used for the first time as a novel machine-learning technique for the prediction of the solubility of C_{60} in a large number of diverse solvents using calculated molecular descriptors from the molecular structure alone and on the basis of the software CODESSA as inputs. The heuristic method of CODESSA was used to select the correlated descriptors and build the linear model. Both the linear and the nonlinear models can give very satisfactory prediction results: the square of the correlation coefficient R^2 was 0.892 and 0.903, and the root-mean-square error was 0.126 and 0.116, respectively, for the whole data set. The prediction result of the LSSVM model is better than that obtained by the heuristic method and the reference, which proved LSSVM was a useful tool in the prediction of the solubility of C_{60}. In addition, this paper provided a new and effective method for predicting the solubility of C_{60} from its structures and gave some insight into the structural features related to the solubility of C_{60} in different solvents.

1. Introduction

Since the discovery of C_{60} in 1985, there has been a burst of interest in identifying possible applications for this highly symmetrical molecule. This interest has resulted in extensive research in biochemistry, materials science, organic chemistry, and other fields.[1] One of the drawbacks to working with C_{60} is that it is not very soluble in most organic solvents[2] and there does not appear to be a specific physical property that one can use to determine what would constitute a good solvent.

Because there is no one property that accurately predicts C_{60} solubility, a quantitative structure−property relationship (QSPR) study would seem to be ideal in attempting to predict solubilities. The QSPR approach has become very useful in the prediction of many 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 the molecular features of the compounds termed descriptors.[3] The advantage of this approach over other methods lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on the experiment properties. The main steps involved in QSPR include data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development, and finally, model performance evaluation.[4] This study can develop a method for the prediction of the properties 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.

As a result of the advantages of this method, the QSPR method has already been used to study the solubility of C_{60} in various solvents. The theoretical linear solvation energy relationship approach was utilized to develop an equation taking into account the bulk, dipole moment, and hydrogen bonding ability of the compound.[5−7] These studies suffer from two drawbacks. First, they assume that the chosen solvent property contributes to solubility with the same weight for all solvents. This assumption may be valid for a family of solvents (e.g., for a homologous series) but not for widely different ones. Second, the solvent properties (especially the Hildebrand parameter) required for linear free energy relationship studies are not easily available; therefore, only a subset of the solvents were involved in the calculations.

A study by Murray et al.[8] also developed a model based on the electrostatic potential of a solvent and quantities related to the surface area of a solvent. This model is notable for the excellent fit to the data, showing a linear correlation coefficient of 0.954 and a standard deviation of 0.475 for the 22 organic solvents that were modeled. However, the predictive ability of this model was not tested.

A solubility study by Ruoff et al.[2] also showed that several properties can be potential indicators of solubility, but they do not always predict solubility. In that study it was shown that polarizability, polarity, molecular size, and cohesive energy density were correlated to solubility most of the time. It was also noted that C_{60} was much more soluble in aromatic solvents, but no predictive model was built.

The quantitative structure−solubility relationship models were suggested for individual data sets, such as alkanes, aklyl halides, alcohols, cycloalkanes, alkylbenzenes, and aryl halides, by

* Corresponding author. Tel.: +86-931-891-2578. Fax: +86-931-891-2582. E-mail address: xiaojunyao@yahoo.com.
[†] Department of Chemistry, Lanzhou University.
[‡] Department of Computer Science, Lanzhou University.
[§] Université Paris.

10.1021/jp052223n CCC: $30.25
Published on Web 10/12/2005

Sivaraman et al.[9] However, only several compounds for a class solvent were involved and no combined model for all classes was built; the built model is difficult to generalize.

In addition, to improve the accuracy of the model, artificial neural networks were applied to introduce nonlinearity in the numerical treatment.[1,10] Compared with the above models, the nonlinear models provided more accurate results. Although neural networks offer high accuracy, in most cases, however, they can suffer from the reproducibility of results, largely as a result of random initialization of the network and variation of the stopping criteria.[11]

The support vector machine (SVM) is a popular algorithm developed from the machine-learning community. Compared with traditional neural networks, SVM possesses prominent advantages: (1) A strong theoretical background provides the SVM with a high generalization capability so it can avoid local minima. (2) A SVM always has a solution that can be quickly obtained by a standard algorithm (quadratic programming). (3) The SVM need not determine the network topology in advance, which can be automatically obtained when the training process ends.[12] As a result of its advantages and remarkable generalization performance over other methods, the SVM has attracted attention and gained extensive applications.[13−22] The SVM shows outstanding performance because it can lead to global models that are often unique by embodying the structural risk minimization principle,[23,24] which has been shown to be superior to the traditional empirical risk minimization principle. Furthermore, as a result of their specific formulation, sparse solutions can be found and both linear and nonlinear regression can be performed. However, finding the final SVM model can be computationally very difficult because it requires the solution of a set of nonlinear equations (quadratic programming). As a simplification of the "traditional" SVM, Suykens and Vandewalle[25] have proposed the use of least-squares SVM (LSSVM). LSSVM encompasses advantages similar to those of SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple.

Besides the method to build the model, another important factor responsible for the quality of the QSPR model is the numerical representation (often called molecular descriptor) of the chemical structure. The performance and the accuracy of the results are strongly dependent on the way the structures are represented. The software CODESSA, developed by the Katritzky group, enables the calculation of a large number of quantitative descriptors solely on the basis of the molecular structural information and codes chemical information into mathematical form.[26,27] CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish molecular structure−property relationships. CODESSA has been applied successfully in a variety of QSPR analyses.[28,29]

In the present investigation, LSSVM was used for the first time as a novel machine-learning technique for the prediction of the solubility of C_{60} in a large number of diverse solvents using calculated molecular descriptors on the basis of the software CODESSA as inputs. The aim was to explore the solubility behavior of C_{60} in different solvents, to develop an accurate quantitative model correlating the structural descriptors of the solvents and the solubility of C_{60} in these solvents, and at the same time, to seek the structural factor affecting the solubility of C_{60} in the different solvents and the essential differences of the different solvents. The prediction results were very satisfactory in both the training set and the test set compounds, which proved LSSVM to be a powerful and useful tool.

2. Methodology

2.1. Data Preparation. The solubilities of C_{60} in 128 different solvents were taken from ref 10 and are listed in Table 1. The solubilities are not given in weight units (e.g., mg/mL) but in terms of logarithmic values of molar fractions (log *S*) because the log *S* values correspond to the free energy changes in the solvation process.

2.2. Calculation of the Descriptors. To obtain a QSPR/quantitative structure−activity relationship (QSAR) model, compounds are often represented by the molecular descriptors. The calculation process of the molecular descriptors is described as follows: All molecules were drawn into Hyperchem and preoptimized using an MM+ molecular mechanics force field.[30] A more precise optimization was done with a semi-empirical AM1 method in MOPAC.[31] All calculations were carried out at the restricted Hartree−Fock level with no configuration interaction. The molecular structures were optimized using the Polak−Ribiere algorithm until the root-mean-square (RMS) gradient was 0.01. The MOPAC output files were used by the CODESSA program to calculate five classes of descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.), topological (Wiener index, Randic indices, Kier−Hall shape indices, etc.), geometrical (moments of inertia, molecular volume, molecular surface area, etc.), electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.), and quantum chemical [reactivity indices, dipole moment, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, etc.].[26]

2.3. Selection of Descriptors on the Basis of the Heuristic Method (HM).[26] Once molecular descriptors were generated, the HM in CODESSA was used to accomplish the preselection of the descriptors and build the linear model. Its advantages are the high speed and no software restrictions on the size of the data set. The HM can either quickly give a good estimation about what quality of correlation to expect from the data or can 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 QSAR/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. As a 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 multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, the cross-validated R_{cv}^2, and the *F* value).

The HM usually produces correlations 2−5 times faster than other methods with comparable quality.[32] The rapidity of

TABLE 1: Experimental and Calculated log *S* of C_{60} in the Various Solvents[a]

no.	solvent	log S_{exp}	log S_{HM}	log S_{LSSVM}	no.	solvent	log S_{exp}	log S_{HM}	log S_{LSSVM}
1	pentane	−6.1	−5.7	−5.8	65[b]	1,4-dimethylbenzene	−3.3	−3.4	−3.3
2	hexane	−5.1	−5.4	−5.4	66	1,2,3-trimethylbenzene	−3.1	−2.9	−2.8
3	octane	−5.2	−5	−4.9	67	1,2,4-trimethylbenzene	−2.5	−3.1	−3
4	isooctane	−5.2	−5	−4.9	68	1,3,5-trimethylbenzene	−3.5	−3.1	−3.1
5	decane	−4.7	−4.7	−4.5	69[b]	1,2,3,4-tetramethylbenzene	−2.9	−2.6	−2.6
6	dodecane	−3.5	−4.4	−4.2	70	1,2,3,5-tetramethylbenzene	−2.4	−2.8	−2.7
7[b]	tetradecane	−4.3	−4.1	−3.8	71	tetralin	−2.5	−2.6	−2.6
8	**cyclopentane**	−6.6			72[b]	ethylbenzene	−3.4	−3.5	−3.4
9[b]	cyclohexane	−5.3	−4.9	−4.8	73	*n*-propylbenzene	−3.5	−3.4	−3.3
10	*cis*-decahydronaphthalene	−3.3	−3.6	−3.3	74	*iso*-propylbenzene	−3.6	−3.3	−3.2
11	*trans*-decahydronaphthalene	−3.5	−3.6	−3.3	75	*n*-butylbenzene	−3.4	−3.2	−3.2
12	cyclopentyl bromide	−4.2	−4.1	−3.9	76[b]	*sec*-butylbenzene	−3.6	−3	−3
13	cyclohexyl chloride	−4.1	−4.4	−4.3	77	*tert*-butylbenzene	−3.7	−3.2	−3.1
14	cyclohexyl bromide	−3.4	−4.1	−3.8	78	fluorobenzene	−4.1	−3.7	−3.7
15	cyclohexyl iodide	−2.8	−3.9	−3.6	79	chlorobenzene	−3	−3.5	−3.4
16	1,2-dibromocyclohexane	−2.6	−3.3	−3.1	80	bromobenzene	−3.3	−3.3	−3.2
17	cyclohexene	−3.8	−4.6	−4.5	81[b]	iodobenzene	−3.5	−3.1	−3.1
18[b]	1-methyl-1-cyclohexene	−3.8	−4.4	−4.3	82	1,2-dichlorobenzene	−2.4	−2.9	−2.9
19	methylcyclohexane	−4.5	−4.6	−4.5	83[b]	1,3-dichlorobenzene	−3.4	−3.2	−3
20[b]	*cis*-1,2-dimethylcyclohexane	−4.6	−4.2	−4	84[b]	1,2-dibromobenzene	−2.6	−2.2	−2.2
21	*trans*-1,2-dimethylcyclohexane	−4.6	−4.2	−4	85	1,3-dibromobenzene	−2.6	−2.4	−2.5
22[b]	ethylcyclohexane	−4.3	−4.4	−4.2	86	1-bromo-2-chloro-benzene	−2.4	−2.6	−2.6
23	dichloromethane	−4.6	−4.8	−4.8	87	1-bromo-3-chloro-benzene	−3	−2.7	−2.7
24[b]	chloroform	−4.8	−4.7	−4.2	88	1,2,4-trichlorobenzene	−2.8	−2.7	−2.6
25	carbon tetrachloride	−4.4	−3.9	−4.1	89	styrene	−3.2	−3.6	−3.5
26	dibromomethane	−4.5	−4.3	−4.1	90	***o*-cresol**	−5.7		
27	bromoform	−3.2	−3	−3.3	91	nitrobenzene	−3.9	−3.7	−3.6
28	iodomethane	−4.2	−4.7	−4.5	92	benzonitrile	−4.2	−3.4	−3.4
29	**diiodomethane**	−4.8			93	anisole	−3.1	−3.5	−3.4
30	bromochloromethane	−4.2	−4.4	−4.2	94	benzaldehyde	−4.2	−3.7	−3.7
31	bromoethane	−5.2	−5.2	−5.2	95	phenyl isocyanate	−3.4	−3.5	−3.4
32	iodoethane	−4.5	−4.5	−4.5	96[b]	2-nitrotoluene	−3.4	−3.3	−3.3
33	1,1,2,2-tetrachloroethane	−3.1	−3.1	−3.1	97	3-nitrotoluene	−3.4	−3.3	−3.3
34	**1,1,2-trichlorotrifluoroethane**	−5.6			98	thiophenol	−3	−3.5	−3.6
35	1,2-dichloroethane	−5	−4.8	−4.8	99[b]	benzyl chloride	−3.4	−3.1	−3.1
36[b]	1,2-dibromoethane	−4.2	−4	−4	100	benzyl bromide	−3.1	−3.1	−3
37	1,1,1-trichloroethane	−4.7	−5.6	−5.5	101	trichlorotoluene	−3	−2.5	−2.7
38	1-chloropropane	−5.6	−5.4	−5.5	102	1-methylnaphthalene	−2.2	−2.4	−2.5
39[b]	1-bromopropane	−5.2	−4.8	−4.8	103	dimethylnaphthalene	−2.1	−2	−2.2
40	1-iodopropane	−4.6	−4.5	−4.5	104	1-phenylnaphthalene	−1.9	−1.5	−1.7
41	2-chloropropane	−5.9	−5.8	−5.9	105[b]	1-chloronaphthalene	−2	−2.3	−2.3
42	2-bromopropane	−5.4	−5.1	−5.1	106[b]	1-bromo-2-methylnapthalene	−2.1	−1.7	−2
43	2-iodopropane	−4.8	−4.6	−4.7	107	ethanol	−7.1	−6.8	−6.7
44	1,2-dichloropropane	−4.9	−5	−4.9	108[b]	1-propanol	−6.4	−6.1	−6.1
45	1,3-dichloropropane	−4.8	−4.9	−4.9	109	1-butanol	−5.9	−5.7	−5.7
46	1,2-dibromopropane	−4.3	−3.9	−3.9	110	1-pentanol	−5.3	−5.4	−5.4
47[b]	1,3-dibromopropane	−4.2	−4	−4	111[b]	1-hexanol	−5.1	−5.2	−5.2
48	1,3-diiodopropane	−3.4	−3.3	−3.3	112[b]	1-octanol	−5	−4.9	−4.8
49	1,2,3-tribromopropane	−2.9	−3.1	−3.1	113	**nitroethane**	−6.7		
50	1,2,3-trichloropropane	−4	−4.2	−4	114	acetone	−7	−6.4	−6.6
51[b]	1-bromo-2-methylpropane	−4.9	−4.6	−4.6	115	***n*-butylamine**	−3.3		
52	1-chloro-2-methylpropane	−5.4	−5.2	−5.2	116[b]	acrylonitrile	−6.4	−6.1	−5.5
53	1-iodo-2-methylpropane	−4.3	−4.5	−4.4	117[b]	2-methoxyethyl ether	−5.2	−4.6	−4.5
54[b]	2-chloro-2-methylpropane	−5.7	−5.6	−5.8	118	*N*,*N*-dimethylformamide	−5.3	−5.4	−5.2
55	2-bromo-2-methylpropane	−5	−5.1	−5.2	119	tetrahydrothiophene	−5.4	−5	−4.9
56[b]	2-iodo-2-methylpropane	−4.4	−4.6	−4.6	120	thiophene	−4.4	−4.6	−4.4
57	1,2-dibromoethylene	−3.7	−3.4	−3.4	121	2-methylthiophene	−3	−3.1	−3.1
58[b]	trichloroethylene	−3.8	−4.4	−4.2	122	*N*-methyl-2-pyrrolidone	−3.9	−4.1	−4.1
59	tetrachloroethylene	−3.8	−4	−3.9	123	pyridine	−4	−4.4	−4.3
60	1-chloro-2-methylpropene	−4.5	−5	−5	124	quinoline	−2.9	−2.6	−2.7
61	benzene	−4	−4.1	−3.9	125	aniline	−3.9	−3.9	−3.9
62[b]	toluene	−3.4	−3.6	−3.6	126	*N*-methylaniline	−3.8	−3.6	−3.6
63	1,2-dimethylbenzene	−2.9	−3.2	−3.2	127	*N*,*N*-dimethylaniline	−3.2	−3.4	−3.4
64	1,3-dimethylbenzene	−3.3	−3.4	−3.3	128	1,5,9-cyclododecatriene	−2.7	−3.2	−3.1

[a] The outlier solvents in model 1 are boldfaced, which were not predicted by model 2 or the LSSVM model. [b] Compounds in the test set.

calculations from the HM renders it the first method of choice in practical research. Thus, in the present investigation, we used this method to select structural descriptors and build the linear model.

2.4. LSSVM. In recent years, the SVM, as a powerful new tool for data classification and function estimation, has been developed.[33] The SVM maps input data into a high-dimensional feature space where it may become linearly separable. Recently, SVM has been applied to a wide variety of domains such as pattern recognition and object detection,[23] function estimation,[34] and so forth. One reason that SVM often performs better than earlier methods is that SVM was designed to minimize structural

risk, whereas previous techniques are usually based on minimization of empirical risk. So, SVM is usually less vulnerable to the overfitting problem. Specifically, Suykens and Vandewalle[26] proposed a modified version of SVM called LSSVM, which resulted in a set of linear equations instead of a quadratic programming problem that can extend the application of the SVM.

There exist a number of excellent introductions for SVM, both printed[34−36] and electronically available.[37] The theory of LSSVM for classification and function estimation has also been described clearly by Suykens and Vandewalle.[26] For this reason, we only briefly described the differences between SVM and LSSVM for function estimation here.

In principle, LSSVM always fits a linear relation ($y = wx + b$) between the regressors (x) and the dependent variable (y). The best relation is the one that minimizes the cost function (Q) containing a penalized regression error term

$$Q_{\mathrm{LSSVM}} = \frac{1}{2} w^T w + \gamma \sum_{k=1}^{N} e_k^{\,2} \tag{1}$$

subject to

$$y_k = w^T \varphi(x_k) + b + e_k, \quad k = 1, ..., N$$

The first part of this cost function is a so-called L_2 norm on the regression weights. When this is used as the norm, weight values are penalized quadratically, and it aims at coefficients that are as small as possible. The second term takes into account the regression error (e_k) for all of the N training objects (the standard least-squares error approach). The relative weight of this part as compared to the first part is indicated by the parameter γ, which has to be optimized by the user. The third part gives the definition of the regression error to be the difference between the true and the predicted values, and this can be seen as a constraint. For comparison, note that the traditional SVM approach defines the regression error differently by neglecting all regression errors smaller than $\pm\epsilon$ (the ϵ-insensitive loss function). It is this difference in error definitions that makes the LSSVM optimization problem computationally much easier than the original SVM problem. Furthermore, the value of parameter ϵ does not have to be optimized for LSSVM, which is the case for SVMs.

Similar to SVM, the LSSVM also considers this optimization problem to be a constrained optimization problem and uses a Lagrange function to solve it. By solving the Lagrange style of eq 1, the weight coefficient (w) can be written as an expansion of the Lagrange multipliers with the corresponding training objects:

$$w = \sum_{k=1}^{N} \alpha_k x_k \quad \text{with} \quad \alpha_k = 2\gamma e_k \tag{2}$$

By substituting eq 2 into the original regression line ($y = wx + b$), the following result can be obtained:

$$y = \sum_{k=1}^{N} \alpha_k x_k^{\,T} x + b \tag{3}$$

It can be seen that the Lagrange multipliers can be defined as

$$\alpha_k = [x_k^{\,T} x + (2\gamma)^{-1}]^{-1} (y_k - b) \tag{4}$$

Finding these Lagrange multipliers is very simple as opposed to the SVM approach in which a more difficult relation has to be solved to obtain these values.

In addition, as a result of solving the optimization problem in terms of Lagrange multipliers such as SVM, LSSVM has the same advantage in that the final model can be written as a weighted linear combination of the inner product between the training points and a new test object. Therefore, it easily allows nonlinear regression as an extension of the linear approach by introducing the kernel function. This leads to the following nonlinear regression function:

$$f(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{5}$$

In eq 5, $K(x,x_k)$ is the kernel function. The value is equal to the inner product of two vectors x and x_k in the feature space $\Phi(x)$ and $\Phi(x_k)$, that is, $K(x,x_k) = \Phi(x)^T \cdot \Phi(x_k)$. The elegance of using the 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. The radial basis function (RBF) kernel $K(x,x_k) = \exp(-||x_k - x||^2/\sigma^2)$ is commonly used.

Note that, in contrast to the Lagrange multipliers, the choice of a kernel and its specific parameters together with γ do not follow from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik−Chervonenkis bounds, cross-validation, an independent optimization set, or Bayesian learning. In this paper, the Gaussian kernel was used as the kernel function and the 10-fold cross-validation was used to tune the optimized values of the two parameters γ and σ. The MSE was used as the error function, and it is computed according to the following equation:

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

where d_i are the teaching outputs (desired outputs), o_i are the actual outputs, and n is the number of samples.

In addition, the RMS error was used to evaluate the quality of the built model, which is defined as

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

All calculations of LSSVM were performed using the Matlab/C toolbox.[38]

3. Results and Discussion

3.1. Results of the HM. Through the HM of CODESSA, the best linear model with six parameters (model 1) was obtained, which is shown in Table 2. The statistical analysis results of the six-parameter model and the involved molecular descriptors as well as their corresponding physical-chemical meaning are summarized in Table 2.

The scatter plot between the calculated and experimental values of log S of C_{60} in the different solvents by model 1 is shown in Figure 1. From Figure 1 it can be seen that there are several obvious outliers in model 1, which are marked by the

TABLE 2: Linear Model (Model 1) Between the Structure and log *S* of C_{60} for All 128 Solvents[a]

descriptor	coefficient	error	*t* test value
intercept	2.09×10^{-1}	9.77×10^{-1}	0.2138
Randic index (order 3)	7.79×10^{-2}	8.87×10^{-3}	8.7796
relative molecular weight	9.96×10^{-1}	4.73×10^{-1}	2.1058
HOMO-1 energy	4.09×10^{-1}	7.18×10^{-2}	5.6931
RNCG	-8.31×10^{1}	3.91×10^{1}	−2.1273
ABIC1	5.34×10^{-1}	8.03×10^{-2}	6.6428
avg one-electron react. indext for a C atom	-2.37×10^{0}	5.76×10^{-1}	−4.1199

[a] $R^2 = 0.761$, $F = 64.29$, $s^2 = 0.322$, and $N = 128$.

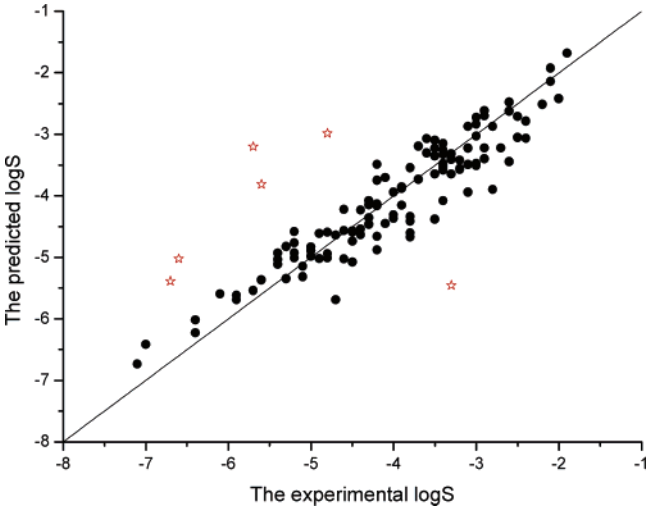

Figure 1. Predicted values of log *S* vs the experimental data by model 1 for 128 solvents.

TABLE 3: Model 2 Between the Structure and log *S* of C_{60}[a]

descriptor	coefficient	error	*t* test value
intercept	-7.91×10^{-1}	6.76×10^{-1}	−1.1701
Randic index (order 3)	4.78×10^{-1}	4.74×10^{-2}	10.0845
relative molecular weight	9.28×10^{-2}	6.14×10^{-3}	15.119
HOMO-1 energy	4.69×10^{-1}	5.57×10^{-2}	8.4204
RNCG	-2.50×10^{0}	3.79×10^{-1}	−6.6071
ABIC1	1.26×10^{0}	3.20×10^{-1}	3.9406
avg one-electron react. index for a C atom	-7.88×10^{1}	2.53×10^{1}	−3.1137

[a] $R^2 = 0.892$, $F = 159.05$, $s^2 = 0.134$, and $N = 122$.

red stars. They are the compounds cyclopentane, diiodomethane, 1,1,1-trichloroethane, *o*-cresol, nitroethane, and *n*-butylamine. The large solubility of some of the outliers can be related to chemical reactions either in the liquid or in the solid phase.[10] To build the proper model for most of solvents, the six outliers were removed and a new model for the remaining 122 solvents was built. The new six-parameter model (model 2) is given in Table 3, and the scatter plot between the predicted log *S*, which is based on this model, and the experimental value is shown in Figure 2. When Table 3 and Figure 2 and Table 2 and Figure 1 are compared, it can be seen that model 2 is much better than model 1 and that the predicted values are in good agreement with the experimental values.

When the descriptors in the regression model are interpreted, it is possible to gain some insight into factors that are likely to govern the solubility of C_{60} in the different solvents.

In the linear model, there are one constitutional descriptor, two topological descriptors, and three quantum chemical descriptors. The constitutional descriptor, *relative molecular weight*, accounts both for the atomic masses (volumes) and for their distribution within the molecular space and seems to

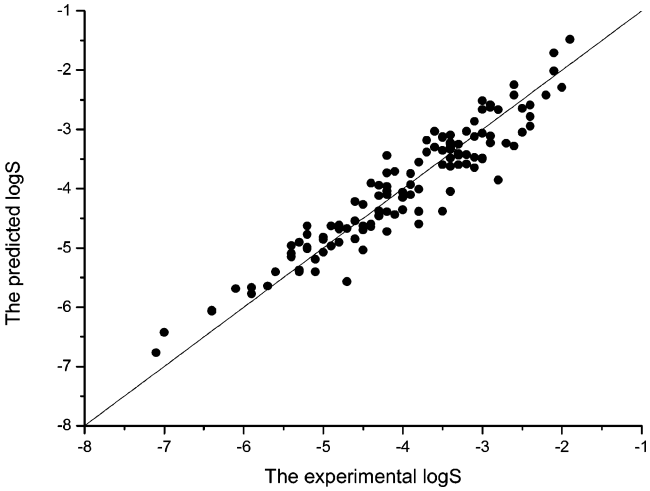

Figure 2. Predicted values of log *S* vs the experimental data by model 1 for 122 solvents.

quantify effectively the bulk cohesiveness of compounds arising from the dispersion and hydrophobic interactions. The topological descriptor, *Randic index* (*order 3*)[39] is calculated as a sum of atomic connectivities over molecular paths of a certain length ($n = 3$); thus, it reflects molecular size and branching. The second topological descriptor, *average bonding information content* (*order 1*; *ABIC1*), is defined on the basis of the Shannon information theory, reflecting the connectivity of atom−atom in the molecule at the first coordination sphere. The involved quantum chemical descriptors were *HOMO-1 energy*, *average one-electron reactivity index for a C atom*, and *RNCG* [*relative negative charge* (QMNEG/QTMINUS) *quantum chemical PC*]. *HOMO-1 energy* is the energy of the second highest occupied molecular orbital. One-electron reactivity indices for a given atomic species in the molecule are defined as[26]

$$R_A' = \sum_{i \in A} \sum_{j \in A} c_{i\mathrm{HOMO}} c_{j\mathrm{LUMO}}$$

where the summations are performed over all atomic orbitals *i* and *j* at the given atom, $c_{i\mathrm{HOMO}}$ and $c_{j\mathrm{LUMO}}$ denote the *i*th and *j*th atomic orbital coefficients on the HOMO and the LUMO, respectively. Here, the *average one-electron reactivity index for a C atom* was involved. Because the main weight atoms are C atoms, this descriptor can estimate the relative reactivity of the compounds.

The two quantum chemical descriptors combined with the constitutional descriptor and the two topological descriptors discussed above roundly reflect cavity forming and dispersion forces during the solubility process.

The last quantum chemical descriptor was *RNCG*, describing the negative partial charge distribution information in the molecule and accounting for the hydrogen-bonding formation ability of the solvents. The hydrogen atoms directly connected with the electronegative atom in the molecule are considered as possible hydrogen-bonding donors. The H-bonding character of the solvent is an obvious discriminating factor. C_{60} will have a lower solubility in solvents that organize themselves through polar or H-bond interactions because of the disruption in the solvent structure that would result from dissolution of C_{60}, which is nonpolar and does not participate in H bonds.

Model 2 gave a RMS error of 0.126, and the corresponding correlation coefficients (R^2) were 0.892 for the whole data set, confirming the good predictive capability of the model. Figure 2 showed a plot of the calculated versus experimental log *S*

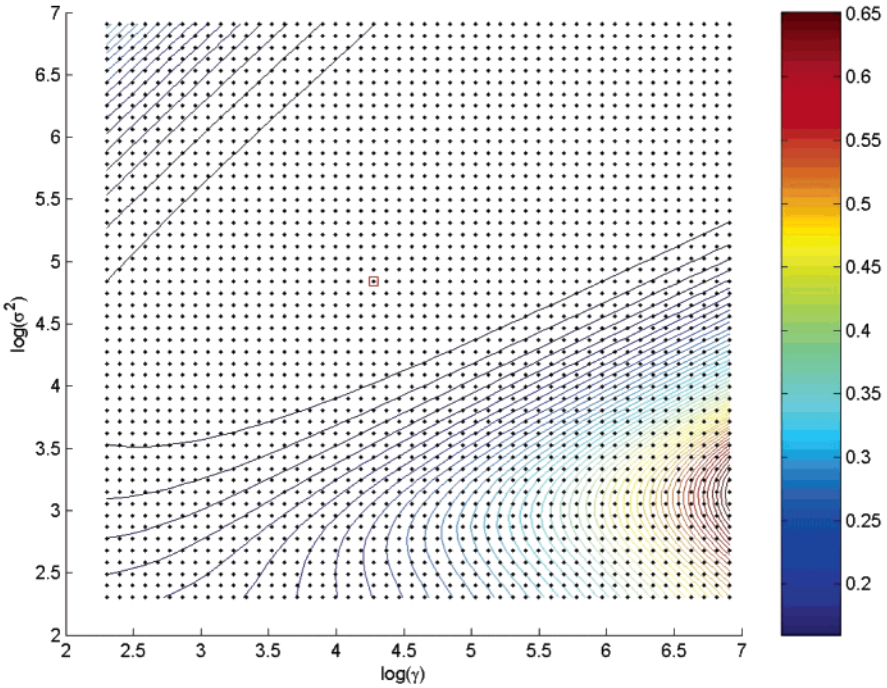

Figure 3. Contour plot of the optimization error for LSSVM when optimizing the parameters σ and γ in the regression problem. The red square indicates the selected optimal settings.

values for all of the 122 compounds studied. From Figure 2, it also can be seen that the predicted values are in good agreement with the experimental values, confirming the selected descriptors calculated solely from structures can describe the structural features of the solvents responsible for the solubility of C_{60} in these solvents. In addition, it was indicated in the introduction that the nonlinear model[1,10] provided much better results than the linear model when predicting the solubility of C_{60}. This could be a result of the complex factors influencing the solubility of C_{60} in the different solvents and that not all of them were in linear correlation with the solubility of C_{60}. So, we built the nonlinear prediction models by LSSVM to further discuss the correlation between the molecular structure and the solubility of C_{60} on the basis of the selected descriptors.

3.2. LSSVM Model. *3.2.1. Optimizing LSSVM.* As discussed in section 2.4, kernel function and its specific parameters together with γ have to be tuned by the user. In this paper, the RBF kernel was used as the kernel function. Thus, γ (the relative weight of the regression error) and σ (the kernel parameter of the RBF kernel) need to optimized. To find the optimized combination of the two parameters, the data set was separated into a training set of 92 compounds and a test set of 30 compounds randomly and a process of leave-one-out cross-validation of the whole training set was performed. Here, the optimal parameters are found by an intensive grid search. The result of this grid search is an error surface spanned by the model parameters. A robust model is obtained by selecting those parameters that give the lowest error in a smooth area.

The parameter (σ) of the RBF kernel in the style of σ^2 and the parameter γ were tuned simultaneously in a grid 50 × 50, ranging from 10 to 1000 and from 10 to 1000, respectively. In this way, parameter optimization was performed in the same orders of magnitude. Because the grid search has been performed over just two parameters, a contour plot of the optimization error can be visualized easily (Figure 3). This is an advantage of LSSVMs over SVMs in which three parameters have to be optimized. From Figure 3, the optimal parameter settings can now be selected from a smooth subarea with a low prediction error. The set of selected optimal values of γ and σ^2 was 71.97 and 126.49, respectively, which is marked by the red square in Figure 3.

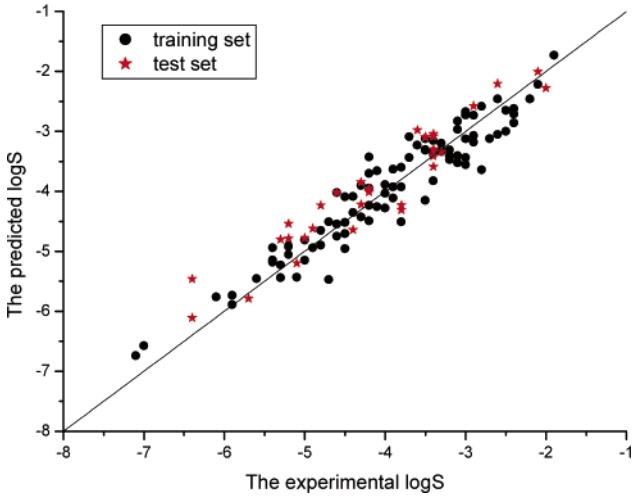

Figure 4. Predicted values of log S vs the experimental data by the LSSVM model for 122 solvents.

3.2.2. Predicted Results of LSSVM. The predicted results of the optimal LSSVM model ($\sigma^2 = 126.49$, $\gamma = 71.97$) are shown in Table 1 and Figure 4. The model gave RMS values of 0.104 for the training set, 0.153 for the prediction set, and 0.116 for the whole data set, and the corresponding correlation coefficients (R^2) were 0.910, 0.908, and 0.903, respectively. From the above results, it can be concluded that the predicted values are in good agreement with the experimental values. By comparing the results from the HM and the LSSVM, it can be seen that the performance of the LSSVM model is better than that of the HM model.

3.3. Comparison of the Results Obtained by Different QSPR Models. To test the suitability of the QSPR approach constructed by HM and SVM, we have compared the obtained results with those calculated in ref 10. Table 4 shows the statistical parameters of the results obtained from the three

TABLE 4: Comparison of Different QSPR Models To Predict log *S* of C_{60}

model	RMS	R^2	*F* test	*t* test	signal
ref 10	0.214	0.841	633.048	25.160	0.000
HM[a]	0.126	0.892	968.584	31.122	0.000
SVM[b]	0.116	0.903	1094.743	33.087	0.000

[a] Represents the HM model. [b] Represents the results by LSSVM.

studies for the same set of compounds. The RMS errors of the SVM model for the whole data set were much lower than that of the models proposed in ref 10 and the HM. The correlation coefficient (R^2) given by the SVM model was higher than that of the models in ref 10 and the HM. Through a regression analysis on the experimental and the calculated $-\log S$ obtained by different methods for the whole data set, the results of the *F* test and the *t* test were obtained and also are shown in Table 4. From Table 4 it can be seen that the SVM model gives the highest *F* and *t* values, so this model gives the most satisfactory results as compared with the results obtained from ref 6 and the HMs. Consequently, this SVM approach currently constitutes the most accurate method to predict the solubility of C_{60} in organic solvents.

4. Conclusion

The LSSVM was used to develop the nonlinear model for predicting the solubility of C_{60} in diverse organic solvents on the basis of calculated descriptors of compounds for the first time. Very satisfactory results were obtained with the proposed method. By analyzing the obtained results, the following can be concluded: (1) The proposed models could identify and provide some insight into what structural features are related to the solubility of C_{60} in the different solvents and help to improve the understanding for the C_{60} mechanism of C_{60} in organic solvents. (2) Additionally, nonlinear models using LSSVM produced slightly better models with good predictive ability than did the linear regression. LSSVM proved to be a powerful and useful tool in the prediction of the physical-chemical property of organic compounds. LSSVM can lead to global (and often unique) nonlinear models, and at the same time, LSSVM can be calculated easily and is easier to be controlled compared with SVM. Therefore, the LSSVM is a very promising machine-learning technique and will gain more extensive applications.

References and Notes

(1) Danauskas, S. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 419.
(2) Ruoff, R. S.; Tse, D. S.; Malhotra, R.; Lorents, D. C. *J. Phys. Chem.* **1993**, *97*, 3379.
(3) Yao, X. J.; Liu, M. C.; Zhang, X. Y.; Hu, Z. D.; Fan, B. T. *Anal. Chim. Acta* **2002**, *462*, 101.
(4) Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218.
(5) Marcus, Y. *J. Phys. Chem. B* **1997**, *101*, 8617.
(6) Makitra, R. G.; Pristanskii, R. E.; Flyunt, R. I. *Russ. J. Gen. Chem.* **2003**, *73*, 1227.
(7) Marcus, Y.; Smith, A. L.; Korobov, M. V.; Mirakyan, A. L.; Avramenko, N. V.; Stukalin, E. B. *J. Phys. Chem. B* **2001**, *105*, 2499.
(8) Murray, J.; Gagarin, S.; Politzer, P. *J. Phys. Chem.* **1995**, *99*, 12081.
(9) Sivaraman, N.; Srinivasan, T. G.; Vasudeva Rao, P. R. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1067.
(10) Kiss, I. Z.; Mándi, G.; Beck, M. T. *J. Phys. Chem. A* **2000**, *104*, 8081.
(11) Manallack, D. T.; Livingstone, D. J. *Eur. J. Med. Chem.* **1999**, *34*, 95.
(12) Gunn, S. R.; Brown, M.; Bossley, K. M. *Lect. Notes Comput. Sci.* **1997**, *1280*, 313.
(13) Belousov, A. I.; Verzakov, S. A.; Von Frese, J. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 15.
(14) Morris, C. W.; Autret, A.; Boddy, L. *Ecol. Modell.* **2001**, *146*, 57.
(15) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5.
(16) Tugcu, N.; Ladiwala, A.; Breneman, C. M.; Cramer, S. M. *Anal. Chem.* **2003**, *75*, 5806.
(17) Thissen, U.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. *Anal. Chem.* **2004**, *76*, 3099.
(18) Ma, W.; Zhang, X.; Luan, F.; Zhang, H.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *J. Phys. Chem. A* **2005**, *109*, 3485.
(19) Chen, N.; Lu, W.; Yang, J.; Li, G. *Support Vector Machine in Chemistry*; World Scientific: Singapore, 2004.
(20) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 389.
(21) Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257.
(22) Muller, K.-R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. *J. Chem. Inf. Model.* **2005**, *45*, 249.
(23) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048.
(24) Burges, C. J. C. *Data Min. Know. Discov.* **1998**, *2*, 1.
(25) Suykens, J. A. K.; Vandewalle, J. *Neural Process. Lett.* **1999**, *9*, 293.
(26) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Comprehensive Descriptors for Structural and Statistical Analysis*, version 2.0; Semichem, Inc.: Florida University, Gainesville, FL, 1994 (reference manual).
(27) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279.
(28) Oblak, M.; Randic, M.; Solmajer, T. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 994.
(29) Katritzky, A. R.; Tatham, D. B. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162.
(30) *HyperChem.*, version 4.0; Hypercube, Inc.: Gainesville, FL, 1994.
(31) Stewart, J. P. P. *MOPAC*, version 6.0; Quantum Chemistry Program Exchange, QCPE, No. 455; Indiana University: Bloomington, IN, 1989.
(32) Katritzky, A. R.; Petrukhin, R.; Jain, R.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1521.
(33) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273.
(34) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998.
(35) Schölkopf, B.; Burges, C.; Smola, A. *Advances in Kernel Methods - Support Vector Learning*; MIT Press: Cambridge, MA, 1999.
(36) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, U.K., 2000.
(37) URL: http://www.kernel-machines.org/(accessed Jan 2005).
(38) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, D.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. *LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines*; Internal Report 02-44, ESATSISTA; K. U. Leuven: Leuven, Belgium, 2002.
(39) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E., Jr. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794.