Support Vector Machine and the Heuristic Method to Predict the Solubility of Hydrocarbons in Electrolyte

Weiping Ma,[†] Xiaoyun Zhang,*[,†] Feng Luan,[†] Haixia Zhang,[†] Ruisheng Zhang,[‡] Mancang Liu,[†] Zhide Hu,[†] and B. T. Fan[§]

Department of Chemistry, Lanzhou University, Lanzhou 730000, China, 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: January 10, 2005; In Final Form: March 3, 2005

A new method support vector machine (SVM) and the heuristic method (HM) were used to develop nonlinear and linear models between the solubility in electrolyte containing sodium chloride and three molecular descriptors of 217 nonelectrolytes. The molecular descriptors representing the structural features of the compounds include two topological and one electrostatic descriptor. The three molecular descriptors selected by HM in CODESSA were used as inputs for SVM. The results obtained by HM and SVM both were satisfactory. The model of HM leads to a correlation coefficient (R) of 0.980 and root-mean-square error (RMS) of 0.219 for the test set. The same descriptors were also employed to build the model in pure water, and the prediction results were consistent with the experimental solubilities. Furthermore, a predictive correlation coefficient $R = 0.988$ and RMS error of 0.170 for the test set were obtained by SVM. The prediction results are in very good agreement with the experimental values. This paper provides a new and effective method for predicting the solubility in electrolyte and reveals some insight into the structural features that are related to the noneletrolytes.

1. Introduction

It is well-known that saturated hydrocarbons are important constituents of petroleum products. Anthropogenic activity associated with the use of these compounds in chemical industry and in energy generation releases hydrocarbons into the environment.[1] The aqueous solubility of these compounds is an important molecular property, playing a large role in the behavior of compounds in many areas of interest. In modeling the environmental impact of a contaminant, alone with the soil−water absorption coefficient, the solubility is a key term in the understanding of transport mechanisms and distribution in water. The petroleum and petrochemical industries require this information for estimating the partition of hydrocarbons between aqueous and organic phase[2,3] and for minimizing the presence of hazardous solutes in aqueous effluents.[4] Environmental chemistry and engineering also need the data for modeling of the transport and fate of hydrocarbon pollutants in the environment[5,6] and for the remediation of sites contaminated by petroleum spills.[7,8] The environmental risk for using these compounds should be assessed because these types of compounds are often the most long-lived of environmental contaminants due to their comparatively low level of biodegradability when compared to oxygen or nitrogen containing compounds. However, experimental solubility data are rather scarce for saturated hydrocarbons with 10 or more carbon atoms. Whereas a general equation would be of greatest use, the present study is limited to hydrocarbons which were expected to be advantageous in obtaining a significant correlation, as the elimination of compounds that will undergo specific interactions with water, such as hydrogen bonding, simplifies the nature of the interactions that must be accounted for. Property of hydrocarbons in water saturated with salt is useful upon its contact with seawater. Given the importance of solubility, a potential theoretical method for predicting the solubility is desired, as many compounds exist for which the solubility simply is not available.

Quantitative structure−property relationships (QSPR) studies have been demonstrated to be an effective computational tool in understanding the correlation between the structure of molecules and their properties.[9−11] In a QSPR study, one seeks to find a mathematical relationship between the property and one or more descriptors. Thus, this study can indicate which of the structural factors may play an important role in the determination of a property. Furthermore, its advantage over other methods lies in the fact that the descriptors used can be calculated from the structure alone and are not dependent on any experimental properties. However, the main problems encountered in this kind of research are still the description of the molecular structure using appropriate molecular descriptors and selection of suitable modeling methods. At present, many types of molecular descriptors such as constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors have been proposed to describe the structural features of molecules.[12−14] The same as the diversity of molecular descriptors many different chemometrics and chemoinformatics methods, such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), different types of artificial neural networks (ANN), and genetic

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

10.1021/jp0501446 CCC: $30.25
Published on Web 03/30/2005

algorithms (GA), can be employed to derive correlation models between the molecular structures and properties.

Recently, there is a growing interest in the use of SVM to chemical problems due to its remarkable generalization performance in modeling nonlinear problems. SVM is a new algorithm developed from the machine learning community and has attracted attention and gained extensive application, such as pattern recognition problems,[15−17] drug design,[18] prediction of protein structure,[19] identifying genes,[20] quantitative structure−activity relationship (QSAR),[21] and QSPR analysis.[22−24] Nevertheless, to the best of our knowledge, there is no prediction of solubility in electrolyte by the QSPR approach based on SVM.

In the present work, SVM was used for the prediction of solubility in water saturated with salt at a different temperature of 217 hydrocarbons using descriptors calculated by the software CODESSA.[25] The aim was to establish a QSPR model that could be used for the prediction of solubility of hydrocarbons from their molecular structure alone, to show the flexible modeling ability of SVM, and, at the same time, to seek the important structure features related to the solubility of hydrocarbons.

2. Experimental Section

2.1. Data Set. In our study, a set of 217 hydrocarbons collected from ref 26 is investigated. The solubilities in water without salt ($X = 0$) and in water saturated with salt (NaCl) [$X =$ 358 700 ppm (wt)] was measured at different temperature and represented as log S_w and log S, where S_w and S are the solubility (ppm). Solubilities of aliphatic, alicyclic, and aromatic hydrocarbons were measured. A complete list of the compounds' names and their corresponding experimental log S_w and log S is given in Table 1. The data set of log S was randomly divided into two subsets: training set and test set (174 and 43 chemicals, respectively). The training set was used to optimize the parameters of SVM and the test set was used to evaluate the prediction ability of SVM. Leave-one-out (LOO) cross-validation was employed on the training set to optimize the parameters of SVM.

2.2. Descriptors Calculation. All structures of the compounds were drawn with the Hyperchem program.[27] The final structural optimizations of compounds were performed using the AM1 parametrization within the semiempirical quantum-chemical program MOPAC 6.0.[28] The geometry optimization was performed without symmetry restrictions. In all cases, frequency calculations had been performed in order to ensure that all of the calculated geometries correspond to true minima. Thereafter, the CODESSA program was used to calculate five types of molecular descriptors: constitutional, topological, geometric, electrostatic, and quantum-chemical. 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. The geometrical descriptors describe the size of the molecule and require 3D-coordinates of the atoms in the given molecular. The electrostatic descriptors reflect characteristics of the charge distribution of the molecular. The quantum chemical descriptors offer information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels.

3. Methodology

3.1. Selection of Descriptors Based on the Heuristic Method. Successful QSPR depends on good descriptors selection. If molecular structures are represented by improper descriptors, they will not lead to reasonable predictions. In recent years, methodology for a general QSPR approach has been developed and coded as the CODESSA software package, which combines different ways of quantifying the structural information about the chemicals with advanced statistical analyses for the establishment of molecular structure−property relationships. To find the best QSPR model, the correlation analysis was carried out using HM which is based on the scale forward selection technique.

The HM provides collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for rapid selection of the best correlation, without testing all of the possible combinations of the available descriptors. HM of the descriptors selection proceeds with a pre-selection of the descriptors to ensure (1) those descriptors that are available for each structure, (2) those values having variation for all structures, (3) descriptors that provide an F-test's value below 1.0 on the one-parameter correlation, and (4) the descriptors whose t values are less than the user-specified value, etc.

Following the pre-selection of descriptors, multiple linear regression models are developed. The selection of best correlations proceeds as follows: (1) Beginning with the top descriptor from the pre-selected list of descriptors, the two-parameter correlations are calculated using the following pairs: the first descriptor with each of the remaining descriptors, second descriptor with each of the remaining descriptors, etc. This procedure is continued until for some nth descriptor no correlations with an F-test value above one-third of the maximum F-test value for a given set are found. (2) The best pairs of branching criteria (number of descriptors sets to select for next recursion level) with highest F-test values in the two-parameter correlations are selected and processed further as the working sets. (3) If not correlated over rsig (descriptors are considered to be noncollinear below the value of their pair correlation coefficient) with the descriptors already included, each of the remaining descriptors is added to the selected working set of descriptors. If the resulting correlation gives F-test value above F working $n/(n + 1)$ (where n is a number of descriptors in the working set plus one), i.e., if this correlation is more significant than the working correlation, then this extended set of descriptors is considered for further treatment. (4) After all descriptors have been applied one-by-one and if the maximum number of descriptors, allowed by the user, is not yet achieved, then best extended working sets, i.e., the sets with the highest F-values, are submitted to the procedure from step (3). Otherwise the procedure is completed and the maximum number of descriptors best correlations found. The goodness of the correlation is tested by the coefficient regression (R^2) and the F-test values (F).

The advantages of this method are as follows: it usually produces correlations 2−5 times faster than other methods and has 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.

3.2. Support Vector Machine. SVM is gaining popularity due to many attractive features and promising empirical performance. It originated from early concepts developed by Vapnik and Chervonenkis.[29−31] SVM represents a powerful technique for a general (nonlinear) classification, regression, and outlier detection with an intuitive model representation.

In comparison with other neural network regressions, SVM has three distinct characteristics in estimation of regression

TABLE 1: Predicted log S_w by HM and log S by HM and SVM

no.	compound	log S_w	log S_w (HM)	log S	log S (HM)	log S (SVM)	no.	compound	log S_w	log S_w (HM)	log S	log S (HM)	log S (SVM)
1	cyclopentane	2.204	2.746	2.048	2.594	2.096	72	nonane	−0.914	−0.807	−1.044	−0.939	−0.922
2	pentane	1.597	1.695	1.441	1.552	1.459	73[a]	2-methyloctane	−0.775	−0.75	−0.906	−0.885	−0.837
3[a]	2-methylbutane	1.679	1.853	1.549	1.707	1.586	74	3-methyloctane	−0.805	−0.709	−0.936	−0.844	−0.909
4	benzene	3.244	2.815	3.132	2.663	2.914	75	4-methyloctane	−0.939	−0.73	−1.070	−0.864	−0.890
5	cyclohexane	1.749	1.961	1.593	1.813	1.446	76	3-ethylheptane	−0.774	−0.856	−0.905	−0.988	−0.911
6	methylcyclopentane	1.621	1.884	1.465	1.735	1.697	77	4-ethylheptane	−0.728	−0.767	−0.859	−0.900	−0.939
7	hexane	0.976	1.064	0.846	0.924	0.822	78[a]	2,2-dimethylheptane	−0.511	−0.509	−0.641	−0.650	−0.648
8[a]	2-methylpentane	1.114	1.094	0.983	0.951	1.014	79	2,3-dimethylheptane	−0.710	−0.747	−0.841	−0.883	−0.813
9	3-methylpentane	1.253	0.956	1.122	0.815	0.926	80	2,4-dimethylheptane	−0.516	−0.661	−0.646	−0.798	−0.778
10	2,2-dimethylbutane	1.377	1.225	1.246	1.077	1.234	81	2,5-dimethylheptane	−0.595	−0.644	−0.726	−0.781	−0.787
11	2,3-dimethylbutane	1.281	1.081	1.150	0.936	1.091	82	2,6-dimethylheptane	−0.575	−0.584	−0.706	−0.723	−0.727
12	Toluene	2.734	2.91	2.604	2.755	2.673	83[a]	3,3-dimethylheptane	−0.621	−0.627	−0.752	−0.765	−0.754
13[a]	Ethylcyclopentane	0.980	1.116	0.825	0.971	1.012	84	3,4-dimethylheptane	−0.713	−0.791	−0.844	−0.926	−0.852
14	1,1-dimethylcyclopentane	1.353	1.222	1.197	1.072	1.265	85	3,5-dimethylheptane	−0.595	−0.691	−0.726	−0.826	−0.840
15	C-1,2-dimethylcyclopentane	1.076	1.136	0.920	0.989	1.120	86	4,4-dimethylheptane	−0.575	−0.725	−0.706	−0.863	−0.767
16	T-1,2-dimethylcyclopentane	1.259	1.181	1.103	1.034	1.125	87	3-ethyl-2-methylhexane	−0.647	−0.815	−0.778	−0.950	−0.843
17	C-1,3-dimethylcyclopentane	1.285	1.198	1.129	1.050	1.168	88[a]	4-ethyl-2-methylhexane	−0.539	−0.778	−0.670	−0.913	−0.823
18[a]	T-1,3-dimethylcyclopentane	1.262	1.17	1.107	1.023	1.161	89	3-ethyl-3-methylhexane	−0.713	−0.735	−0.844	−0.870	−0.832
19	methylcyclohexane	1.204	1.235	1.048	1.089	1.064	90	3-ethyl-4-methylhexane	−0.708	−0.866	−0.839	−0.999	−0.879
20	heptane	0.350	0.261	0.219	0.125	0.306	91	2,2,3-trimethylhexane	−0.534	−0.599	−0.664	−0.739	−0.680
21	2-methylhexane	0.405	0.336	0.274	0.197	0.406	92	2,2,4-trimethylhexane	−0.353	−0.615	−0.484	−0.757	−0.678
22	3-methylhexane	0.422	0.367	0.291	0.229	0.322	93[a]	2,2,5-trimethylhexane	−0.268	−0.452	−0.398	−0.596	−0.556
23[a]	3-ethylpentane	0.470	0.223	0.340	0.087	0.298	94	2,3,3-trimethylhexane	−0.638	−0.658	−0.769	−0.797	−0.716
24	2,2-dimethylpentane	0.643	0.559	0.513	0.414	0.590	95	2,3,4-trimethylhexane	−0.673	−0.769	−0.804	−0.906	−0.796
25	2,3-dimethylpentane	0.720	0.405	0.589	0.265	0.416	96	2,3,5-trimethylhexane	−0.476	−0.678	−0.607	−0.817	−0.733
26	2,4-dimethylpentane	0.644	0.432	0.514	0.289	0.499	97	2,4,4-trimethylhexane	−0.458	−0.629	−0.589	−0.770	−0.692
27	3,3-dimethylpentane	0.772	0.357	0.642	0.216	0.450	98[a]	3,3,4-trimethylhexane	−0.710	−0.692	−0.840	−0.830	−0.760
28[a]	2,2,3-trimethylbutane	0.759	0.558	0.628	0.412	0.606	99	3,3-diethylpentane	−0.854	−0.77	−0.985	−0.903	−0.904
29	ethylbenzene	2.218	2.482	2.045	2.329	2.197	100	3-ethyl-2,2-dimethylpentane	−0.540	−0.607	−0.671	−0.747	−0.703
30	M-xylene	2.241	2.566	2.025	2.411	2.083	101	3-ethyl-2,3-dimethylpentane	−0.818	−0.699	−0.948	−0.836	−0.781
31	O-xylene	2.344	2.434	2.113	2.280	2.058	102	3-ethyl-2,4-dimethylpentane	−0.613	−0.697	−0.744	−0.834	−0.774
32	P-xylene	2.305	2.452	2.045	2.298	2.043	103[a]	2,2,3,3-tetramethylpentane	−0.705	−0.48	−0.836	−0.624	−0.576
33[a]	propylcyclopentane	0.310	−0.079	0.154	−0.219	−0.262	104	2,2,3,4-tetramethylpentane	−0.519	−0.503	−0.650	−0.646	−0.591
34	ethylcyclohexane	0.465	0.472	0.309	0.330	0.374	105	2,2,4,4-tetramethylpentane	−0.244	−0.239	−0.375	−0.389	−0.306
35	1,1-dimethylcyclohexane	0.777	0.732	0.622	0.584	0.647	106	2,3,3,4-tetramethylpentane	−0.738	−0.58	−0.868	−0.721	−0.658
36	C-1,2-dimethylcyclohexane	0.778	0.555	0.622	0.410	0.463	107	butylbenzene	1.141	1.247	0.851	1.099	1.073
37	T-1,2-dimethylcyclohexane	0.675	0.518	0.520	0.374	0.466	108[a]	M-diethylbenzene	1.435	1.246	1.282	1.098	1.096
38[a]	C-1,3-dimethylcyclohexane	0.763	0.571	0.608	0.426	0.517	109	O-diethylbenzene	1.368	1.286	1.215	1.139	1.160
39	T-1,3-dimethylcyclohexane	0.653	0.58	0.497	0.435	0.518	110	P-dimethylbenzene	1.370	1.393	1.216	1.244	1.272
40	C-1,4-dimethylcyclohexane	0.654	0.55	0.498	0.405	0.510	111	1,2,3,4-tetramethylbenzene	0.866	1.298	0.713	1.148	0.960
41	T-1,4-dimethylcyclohexane	0.584	0.514	0.428	0.369	0.502	112	1,2,3,5-tetramethylbenzene	1.028	1.298	0.875	1.147	0.912
42	octane	−0.366	−0.199	−0.496	−0.334	−0.366	113[a]	1-cyclopentylpentane	−0.939	−0.622	−1.095	−0.760	−0.856
43[a]	2-methylheptane	−0.126	−0.215	−0.257	−0.352	−0.235	114	butylcyclohexane	−0.749	−0.586	−0.905	−0.724	−0.857
44	3-methylheptane	−0.101	−0.211	−0.232	−0.347	−0.303	115	decane	−1.284	−1.176	−1.415	−1.308	−1.559
45	4-methylheptane	−0.129	−0.24	−0.260	−0.376	−0.285	116	2-methylnonane	−1.374	−1.12	−1.504	−1.254	−1.401
46	3-ethylhexane	−0.149	−0.303	−0.280	−0.437	−0.330	117	3-methylnonane	−1.393	−1.17	−1.523	−1.302	−1.462
47	2,2-dimethylhexane	0.144	−0.01	0.013	−0.153	−0.054	118[a]	4-methylnonane	−1.342	−1.163	−1.473	−1.296	−1.458
48[a]	2,3-dimethylhexane	−0.075	−0.248	−0.206	−0.386	−0.214	119	5-methylnonane	−1.327	−1.272	−1.458	−1.404	−1.424
49	2,4-dimethylhexane	0.074	−0.196	−0.056	−0.335	−0.186	120	3-ethyloctane	−1.362	−1.271	−1.492	−1.402	−1.483
50	2,5-dimethylhexane	0.101	−0.049	−0.030	−0.190	−0.124	121	4-ethyloctane	−1.292	−1.282	−1.422	−1.413	−1.470
51	3,3-dimethylhexane	0.017	−0.056	−0.114	−0.196	−0.144	122	2,2-dimethyloctane	−1.126	−0.969	−1.256	−1.108	−1.167
52	3,4-dimethylhexane	−0.153	−0.272	−0.284	−0.408	−0.258	123[a]	2,3-dimethyloctane	−1.308	−1.198	−1.439	−1.332	−1.358
53[a]	3-ethyl-2-methylpentane	−0.076	−0.268	−0.207	−0.405	−0.248	124	2,4-dimethyloctane	−1.100	−1.121	−1.231	−1.257	−1.321
54	3-ethyl-3-methylpentane	−0.142	−0.21	−0.273	−0.348	−0.240	125	2,5-dimethyloctane	−1.165	−1.115	−1.296	−1.251	−1.323
55	2,2,3-trimethylpentane	0.064	−0.098	−0.067	−0.240	−0.093	126	2,6-dimethyloctane	−1.212	−1.131	−1.343	−1.266	−1.330
56	2,2,4-trimethylpentane	0.347	0.052	0.216	−0.094	0.017	127	2,7-dimethyloctane	−1.199	−1.026	−1.330	−1.164	−1.261
57	2,3,3-trimethylpentane	−0.053	−0.17	−0.184	−0.311	−0.139	128[a]	3,3-dimethyloctane	−1.232	−0.997	−1.363	−1.134	−1.291
58[a]	2,3,4-trimethylpentane	0.134	−0.199	0.003	−0.339	−0.162	129	3,4-dimethyloctane	−1.286	−1.233	−1.416	−1.366	−1.404
59	2,2,3,3-tetramethylbutane	0.154	−0.016	0.024	−0.164	−0.093	130	3,5-dimethyloctane	−1.187	−1.226	−1.318	−1.359	−1.374
60	propylbenzene	1.717	1.849	1.605	1.699	1.607	131	3,6-dimethyloctane	−1.222	−1.181	−1.353	−1.315	−1.387
61	cumene	1.699	1.831	1.628	1.679	1.414	132	4,4-dimethyloctane	−1.140	−1.149	−1.271	−1.285	−1.292
62	M-ethyltoluene	1.918	1.948	1.765	1.797	1.638	133[a]	4,5-dimethyloctane	−1.255	−1.272	−1.386	−1.405	−1.392
63[a]	O-ethyltoluene	1.969	1.911	1.816	1.760	1.634	134	4-propylheptane	−1.140	−1.293	−1.271	−1.424	−1.459
64	P-ethyltoluene	1.977	1.935	1.824	1.784	1.627	135	4-isopropylheptane	−1.175	−1.259	−1.306	−1.393	−1.385
65	1,2,3-trimethylbenzene	1.816	1.953	1.687	1.800	1.569	136	3-ethyl-2-methylheptane	−1.232	−1.221	−1.363	−1.354	−1.394
66	1,2,4-trimethylbenzene	1.756	1.916	1.598	1.763	1.497	137	4-ethyl-2-methylheptane	−1.108	−1.141	−1.239	−1.275	−1.370
67	mesitylene	1.683	1.994	1.496	1.840	1.558	138[a]	5-ethyl-2-methylheptane	−1.194	−1.259	−1.325	−1.393	−1.369
68[a]	butylcyclopentane	−0.366	−0.151	−0.521	−0.291	−0.257	139	3-ethyl-3-methylheptane	−1.296	−1.159	−1.427	−1.293	−1.386
69	propylcyclohexane	−0.169	−0.079	−0.325	−0.219	−0.260	140	4-ethyl-3-methylheptane	−1.256	−1.311	−1.387	−1.442	−1.426
70	C−c-1,3,5-trimethylcyclohexane	0.293	0.052	0.137	−0.093	−0.050	141	3-ethyl-5-methylheptane	−1.157	−1.234	−1.288	−1.366	−1.424
71	C−t-1,3,5-trimethylcyclohexane	0.241	0.015	0.085	−0.129	−0.065	142	3-ethyl-4-methylheptane	−1.276	−1.302	−1.406	−1.433	−1.428
143[a]	4-ethyl-4-methylheptane	−1.222	−1.121	−1.353	−1.255	−1.381	181	2,3,4,5-tetramethylhexane	−1.108	−1.21	−1.239	−1.347	−1.276
144	2,2,3-trimethylheptane	−1.143	−0.995	−1.273	−1.135	−1.167	182	3,3,4,4-tetramethylhexane	−1.446	−1.196	−1.577	−1.335	−1.238
145	2,2,4-trimethylheptane	−0.909	−0.987	−1.040	−1.128	−1.124	183[a]	2,4-dimethyl-3-isopropylpentane	−1.129	−1.084	−1.260	−1.223	−1.223
146	2,2,5-trimethylheptane	−0.972	−0.989	−1.103	−1.130	−1.132	184	3,3-diethyl-2-methylpentane	−1.439	−1.206	−1.570	−1.339	−1.390
147	2,2,6-trimethylheptane	−0.925	−0.948	−1.056	−1.091	−1.071	185	3-ethyl-2,2,3-trimethylpentane	−1.434	−1.206	−1.565	−1.339	−1.390
148[a]	2,3,3-trimethylheptane	−1.207	−1.055	−1.338	−1.193	−1.226	186	3-ethyl-2,2,4-trimethylpentane	−1.085	−1.004	−1.216	−1.145	−1.135
149	2,3,4-trimethylheptane	−1.200	−1.214	−1.331	−1.349	−1.327	187	3-ethyl-2,3,4-trimethylpentane	−1.432	−1.056	−1.563	−1.194	−1.241

TABLE 1. (Continued)

no.	compound	log S_w	log S_w (HM)	log S	log S (HM)	log S (SVM)	no.	compound	log S_w	log S_w (HM)	log S	log S (HM)	log S (SVM)
150	2,3,5-trimethylheptane	−1.220	−1.161	−1.350	−1.297	−1.299	188[a]	2,2,3,3,4-pentamethylpentane	−1.351	−0.932	−1.481	−1.076	−1.022
151	2,3,6-trimethylheptane	−1.103	−1.059	−1.234	−1.197	−1.238	189	2,2,3,4,4-pentamethylpentane	−1.184	−0.912	−1.315	−1.058	−0.960
152	2,4,4-trimethylheptane	−0.977	−0.979	−1.108	−1.119	−1.145	190	pentylcyclohexane	0.584	0.662	0.472	0.517	0.532
153[a]	2,4,5-trimethylheptane	−1.115	−1.182	−1.246	−1.319	−1.302	191	1-cyclopentylhexane	−1.420	−1.029	−1.576	−1.166	−1.384
154	2,4,6-trimethylheptane	−0.891	−1.134	−1.022	−1.272	−1.233	192	pentylcyclohexane	−1.233	−1.038	−1.389	−1.175	−1.386
155	2,5,5-trimethylheptane	−1.023	−1.048	−1.153	−1.188	−1.186	193[a]	undecane	−2.357	−1.644	−2.487	−1.774	−2.016
156	3,3,4-trimethylheptane	−1.095	−1.19	−1.379	−1.327	−1.292	194	hexylbenzene	0.008	0.186	−0.105	0.042	0.100
157	3,3,5-trimethylheptane	−1.095	−1.112	−1.226	−1.250	−1.244	195	1,2,3-triethylbenzene	0.590	0.365	0.436	0.221	0.410
158[a]	3,4,4-trimethylheptane	−1.229	−1.141	−1.360	−1.278	−1.285	196	1,2,4-triethylbenzene	0.590	0.435	0.436	0.290	0.508
159	3,4,5-trimethylheptane	−1.264	−1.337	−1.394	−1.471	−1.371	197	1,3,5-triethylbenzene	0.622	0.49	0.469	0.345	0.564
160	3-isopropyl-2-methylhexane	−1.367	−1.177	−1.497	−1.313	−1.307	198[a]	1-cyclopentylheptane	−1.790	−1.409	−1.946	−1.545	−1.823
161	3,3-diethylhexane	−1.357	−1.291	−1.487	−1.423	−1.435	199	1-cyclohexylhexane	−1.599	−1.391	−1.754	−1.527	−1.826
162	3,4-diethylhexane	−1.298	−1.331	−1.429	−1.462	−1.465	200	dodecane	−2.432	−1.927	−2.538	−2.057	−2.478
163[a]	3-ethyl-2,2-dimethylhexane	−1.105	−1.088	−1.236	−1.226	−1.233	201	1-phenyloctane	−0.324	−0.329	−0.437	−0.472	−0.340
164	4-ethyl-2,2-dimethylhexane	−0.876	−0.995	−1.007	−1.135	−1.167	202	1-cyclopentyloctane	−2.046	−1.768	−2.201	−1.904	−2.124
165	3-ethyl-2,3-dimethylhexane	−1.293	−1.175	−1.424	−1.311	−1.312	203[a]	1-cyclohexylheptane	−1.857	−1.795	−2.013	−1.931	−2.157
166	4-ethyl-2,3-dimethylhexane	−1.225	−1.252	−1.356	−1.386	−1.363	204	tridecane	−2.699	−2.322	−2.830	−2.451	−2.773
167	3-ethyl-2,4-dimethylhexane	−1.205	−1.174	−1.336	−1.309	−1.363	205	1-phenyloctane	−0.603	−0.727	−0.715	−0.868	−0.725
168[a]	4-ethyl-2,4-dimethylhexane	−1.229	−1.106	−1.360	−1.243	−1.265	206	1,2,3,4-tetraethylbenzene	−0.049	−0.543	−0.202	−0.683	−0.497
169	3-ethyl-2,5-dimethylhexane	−1.055	−1.179	−1.186	−1.315	−1.294	207	1,2,3,5-tetraethylbenzene	−0.041	−0.491	−0.194	−0.632	−0.448
170	4-ethyl-3,3-dimethylhexane	−1.273	−1.181	−1.404	−1.316	−1.325	208[a]	1,2,4,5-tetraethylbenzene	−0.033	−0.527	−0.186	−0.668	−0.477
171	3-ethyl-3,4-dimethylhexane	−1.254	−1.291	−1.385	−1.425	−1.364	209	1-cyclopentylnonane	−2.187	−2.086	−2.343	−2.221	−2.361
172	2,2,3,3-tetramethylhexane	−1.210	−0.948	−1.341	−1.090	−1.065	210	1-cyclohexyloctane	−1.996	−2.086	−2.151	−2.221	−2.388
173[a]	2,2,3,4-tetramethylhexane	−1.173	−1.08	−1.303	−1.220	−1.185	211	tetradecane	−2.658	−2.544	−2.770	−2.673	−3.050
174	2,2,3,5-tetramethylhexane	−0.912	−0.99	−1.042	−1.132	−1.090	212	1-phenylnonane	−0.793	−1.089	−0.906	−1.230	−1.082
175	2,2,4,4-tetramethylhexane	−1.048	−0.908	−1.178	−1.053	−0.985	213[a]	1-cyclopentyldecane	−2.229	−2.376	−2.385	−2.510	−2.458
176	2,2,4,5-tetramethylhexane	−0.898	−0.985	−1.029	−1.128	−1.078	214	1-cyclohexylnonane	−2.022	−2.439	−2.178	−2.573	−2.490
177	2,2,5,5-tetramethylhexane	−0.633	−0.761	−0.764	−0.909	−0.949	215	pentadecane	−2.959	−2.885	−3.092	−3.013	−3.133
178[a]	2,3,3,4-tetramethylhexane	−1.315	−1.117	−1.446	−1.256	−1.237	216	hexadecane	−3.046	−3.06	−3.398	−3.188	−3.047
179	2,3,3,5-tetramethylhexane	−1.030	−1.002	−1.161	−1.144	−1.118	217	heptadecane	−2.854	−3.359	−2.983	−3.488	−3.063
180	2,3,4,4-tetramethylhexane	−1.242	−1.14	−1.373	−1.279	−1.229							

[a] Test set; log S_w : solubility in water system; log S: solubility in water saturated with salt system.

function: (1) SVM estimates the regression using a hypothesis space of linear functions in a high dimensional feature space. (2) SVM carries out the regression estimation by risk minimization and the risk is measured by Vapnik's ϵ-insensitive loss function. Unlike artificial neural networks (ANN) that employ traditional empirical risk minimization (ERM) principle (Vapnik, 1998), SVM adopts the of structural risk minimization (SRM) principle.[32] This has been found to be superior to the ERM principle. So SVM is usually less vulnerable to overfitting problem. (3) The risk function of SVM is made up of the empirical error and a regularization term which is derived from the SRM principle.

The basic idea in SVR is to map the input data x into a higher dimensional feature space F via a nonlinear mapping Φ and then a linear regression problem is obtained and solved in the feature space. Therefore, the regression approximation addresses the problem of estimating a function based on a given data set $G = \{(\mathbf{x}_i,d_i)\}_{i=1}^{l}$ ($\mathbf{x}_i$ is input vector, d_i is the desired value, l corresponds to the size of the training data).

The generic SVR estimating function takes the form as eq 1

$$y = \sum_{i=1}^{l} w_i \cdot \Phi_i(\mathbf{x}) + b,\ w_i \subset R^n,\ b \subset \mathbf{R} \tag{1}$$

where $\{\Phi_i(\mathbf{x})\}_{i=1}^{l}$ denotes the features of inputs, $\{w_i\}_{i=1}^{l}$ and b are coefficients. The coefficients are estimated by minimizing the regularized risk function

$$R_{\mathrm{SVM}}(C) = C\frac{1}{l}\sum_{i=1}^{l} L\epsilon\,(d_i,y_i) + \frac{1}{2}||\mathbf{w}||^2 \tag{2}$$

where

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

In eq 2, the first term $C(1/l)\sum_{i=1}^{l} L_\epsilon\,(d_i,y_i)$ is the empirical error (risk). The ϵ-insensitive loss function given by eq 3 is used to measure them. This loss function provides the advantage of enabling one to use sparse data points to represent the decision function as eq 1. Also, the second term $^1/_2||\mathbf{w}||^2$ is the regularization term, where C is the regularized constant. C 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 to the regularization term to grow. ϵ is called the tube size and it corresponds to the approximation accuracy placed on the training data points. Both C and ϵ are user-prescribed parameters.

Then, by introduction of Lagrange multipliers (α_i,α_i^*) and satisfying the equality $\alpha_i \cdot \alpha_i^* = 0$, $\alpha_i \geq 0$, $\alpha_i^* \geq 0$, $i = 1, ..., l$, the decision function (1) becomes the following form:

$$f(\mathbf{x},\alpha_i,\alpha_i^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(\mathbf{x},\mathbf{x}_i) + b \tag{4}$$

In eq 4, the kernel function K is equivalent to $K(\mathbf{x},\mathbf{x}_i) = \Phi(\mathbf{x})\cdot\Phi(\mathbf{x}_i)$. All kernel functions must satisfy Mercer's condition (kernel function must be symmetric, and it must be positive definite) that corresponds to the inner product of some feature space. One has several possibilities for the choice of this kernel function, including linear, polynomial, spline, and radial basis function. 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 $\Phi(\mathbf{x})$ explicitly. In SVR, a commonly used kernel function is the Gaussian radial basis function.

3.3. SVM Implementation and Computation Environment. All calculation programs implementing SVM were written in

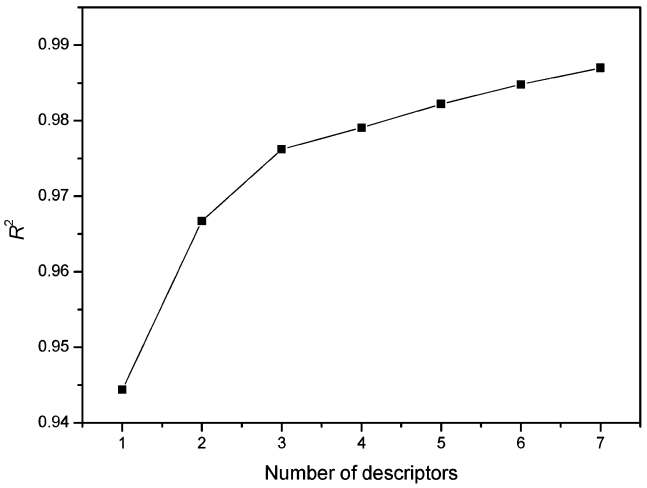

Figure 1. Influence of the number of descriptors on R^2.

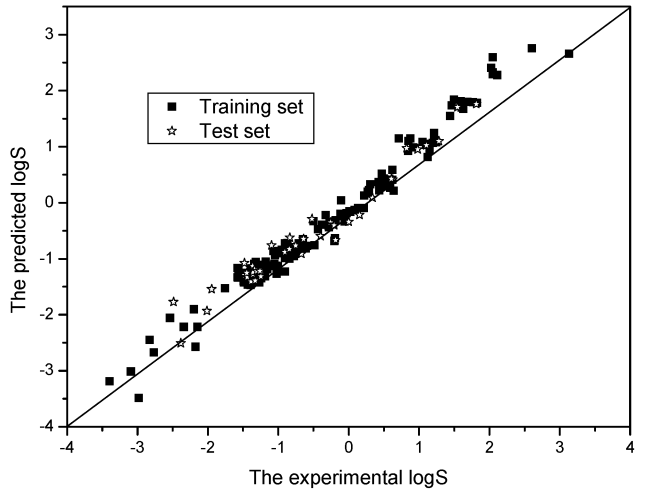

Figure 2. Experimental log *S* vs the calculated log *S* by HM.

R-file based on R script for SVM.[33] 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. Result of the Heuristic Method. The solubility of the training set has been correlated with three descriptors employing CODESSA software via HM. As shown in Figure 1, three descriptors appear to be sufficient for a successful regression model. For a three-parameter model, the squared correlation coefficient (R^2) is 0.976, and for a seven-parameter model, it improves to the value 0.987. As shown in Figure 1, the introduction of a new descriptor to the regression model does not significantly improve the value of R^2 and it was determined that the optimum subset size had been achieved. This model gave an RMS error of 0.187 for the training set, 0.219 for the prediction set, and 0.194 for the whole set, and the corresponding correlations (R) were 0.988, 0.980, and 0.986, respectively. The calculated and experimental values of log *S* were given in Table 1, the scatter plot was shown in Figure 2, and the statistical parameters of the model were shown in Table 2.

To compare with the solubility in pure water, the same descriptors were used to build the model and the prediction results were listed in Table 1. It can be seen that the selected descriptors can be used to predict the solubilities not only in water saturated with salt but also in pure water as shown in Table 3. The RMS error and correlation coefficient of solubility in pure water were 0.192 and 0.987.

By interpreting the descriptors in the regression model, it is possible to gain some insight into factors that are likely to govern the solubility in electrolyte. This model contains one electrostatic (PNSA-1 partial negative surface area [Zefirov's PC]) and two topological (average complementary information content (order 0), Kier&Hall index (order 2)) descriptors. These descriptors encoded different aspects of the molecular structure.

According to the t-test value (the ratio of the coefficient to the coefficient error) in Table 2, the most significant descriptor of the model is average complementary information content (order 0) ($^0\overline{\mathrm{CIC}}$). The average complementary 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 ρ is the radius of the molecular graph G. At the zero-order level, the atom set is partitioned solely on the basis on 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. The three topological indices, average complementary information content (order 0), average complementary information content (order 1) and average complementary information content (order 2), reflect the branching of the molecular and the diversity of the atoms of the branching. 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 hydrophobic and the steric property of the solute comprehensively. As the hydrophobic and steric interaction is the main interaction between the solute and the solvent, these three topological descriptors play an important role in the solubility. $^0\overline{\mathrm{CIC}}$ can be related to molecular shape and symmetry. In order for a solute to enter into aqueous solution, a cavity must be formed in the solvent for the solute molecular to occupy. Water as a solvent would much prefer to interact with itself or other hydrogen bonding or ionic species than with a nonpolar solute, so there is an increasing penalty (and thus lower solubility) for hydrocarbons with high-symmetry solutes. The negative regression coefficients for $^0\overline{\mathrm{CIC}}$ reflect the fact that hydrocarbons with higher symmetry have weaker coordination ability that leads to lower solubility.

The Kier&Hall index (order 2), KHI2, belongs to the well-known valence connectivity indices. The Kier&Hall indices $^mX^v$ defined by eqs 5 and 6 belong to the same "family" of descriptors

$$^mX^v = \sum_{i=1}^{Ns} \prod_{k=1}^{m+1} \left(\frac{1}{\delta_k^v}\right)^{1/2} \tag{5}$$

$$\delta_k^v = \frac{Z_k^v - H_k}{Z_k - Z_k^v - 1} \tag{6}$$

In eqs 5 and 6, Z_k is the total number of electrons in the kth atom, Z_k^v is the total number of valence electrons in the kth atom, H_k is the number of hydrocarbon atoms directly attached to the kth non-hydrocarbon atom, $m = 0$ is the atomic valence connectivity indices, $m = 1$ is the one bond path valence

TABLE 2: Linear Model in Water Saturated with Salt System[a]

no	descriptor	coefficient	standard error	t-test value
1	intercept	13.6540	0.2351	58.0858
2	average complementary information content (order 0)	−3.7800	0.0743	−50.8892
3	PNSA-1 partial negative surface area [Zefirov's PC]	−0.01344	0.0008	−15.0206
4	Kier&Hall index (order 2)	0.2458	0.0280	8.7660

[a] $R = 0.986$, $R^2 = 0.972$, $R_{cv}^2 = 0.971$, $F = 2514.7559$, RMS $= 0.194$, $N = 217$.

TABLE 3: Linear Model in Water System[a]

no	descriptor	coefficient	standard error	t-test value
1	intercept	13.8776	0.2310	60.0680
2	average complementary information content (order 0)	−3.8076	0.0730	−52.1570
3	PNSA-1 partial negative surface area [Zefirov's PC]	−0.0135	0.0009	−15.3598
4	Kier&Hall index (order 2)	0.2529	0.0276	9.1793

[a] $R = 0.987$, $R^2 = 0.974$, $R_{cv}^2 = 0.973$, $F = 2627.2671$, RMS $= 0.192$, $N = 217$.

connectivity indices, $m = 2$ is the two bond valence connectivity indices, and $m = 3$ is the three contiguous bond fragment valence connectivity indices, etc.

Kier and Hall have recently interpreted the molecular connectivity in terms of intermolecular accessibility starting from the interpretation of the bond contributions. Thus, they have concluded that[34] "the molecular connectivity index is the contribution of one molecule to the bimolecular interactions arising from encounters of bonds among two molecules". The significant positive coefficient of size related descriptors in the models indicate the higher probability of interaction leads to higher solubility.

The final descriptor, the partial negative surface area [Zefirov's PC] (PNSA1), is a sum of the negatively charged solvent-accessible atomic surface area

$$\mathrm{PNSA1} = \sum_A S_A \in\{\delta_A < 0\} \tag{7}$$

The PNSA1 encoded the distribution of the negative charge normalized by the total surface area of the molecular. The PNSA1 should be directly related to the hydrocarbon bond or Lewis basicity of the molecule. A large (in magnitude) value of PNSA1 should and does lead to lower log *S*.

From the above discussion, the three descriptors can account for the structural features responsible for the solubility of hydrocarbons in water saturated with salt.

4.2. Result of SVM. To obtain more accurate model, after the linear model was established, we built the nonlinear prediction model by SVM to further discuss the correlation between the molecular structure and the solubility based on the same subset of descriptors. Similar to other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters: capacity parameter C, ϵ of ϵ-insensitive loss function, the kernel type K, and its corresponding parameters. In this work, LOO cross-validation was performed for parameters optimization, which probably is the current best-performing approach to the SVM design problem. RMS was used as an error function which was defined as eq 8

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - o_i)^2}{n}} \tag{8}$$

C is a regulation 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. However, ref 35 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.

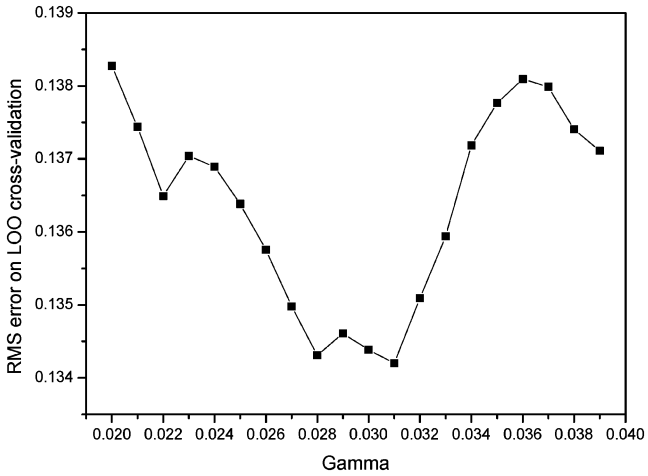

Figure 3. Gamma vs RMS error on LOO cross-validation ($C = 100$, $\epsilon = 0.1$).

For regression tasks, the Gaussian kernel shown as eq 9 is commonly used

$$\mathrm{K}(\mathbf{x}_i,\mathbf{x}) = \exp\{-\gamma|\mathbf{x} - \mathbf{x}_i|\}^2 \tag{9}$$

where γ is a constant, the parameter of the kernel, and $\mathbf{x}$ and $\mathbf{x}_i$ are two independent variables. 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 3), and the minimum was chosen as the optimal conditions. In this case, $\gamma = 0.031$.

The optimal value for ϵ depends on the type of the noise present in the data, which is usually unknown. Even if enough knowledge of noise is available to select an optimal value for ϵ, there is the practical consideration of the number of resulting support vectors. ϵ-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 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 4. The optimal ϵ was found as 0.04.

The last important parameter is the regularization parameter C, of which the effect on the RMS was shown in Figure 5. From Figure 5, the optimal C was found as 100.

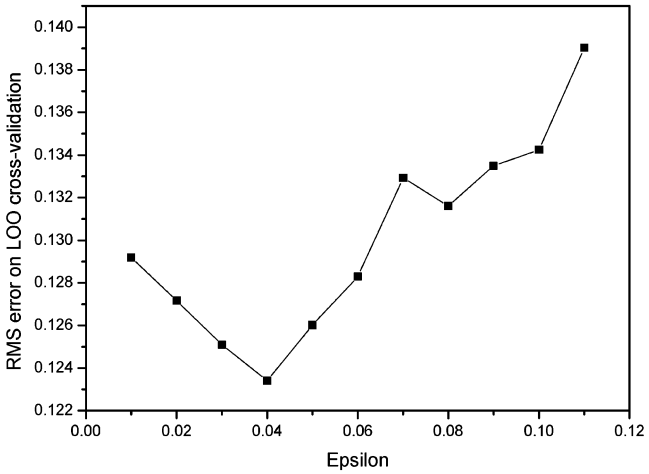

Figure 4. Epsilon vs RMS error on LOO cross-validation ($C = 100$, $\gamma = 0.031$).

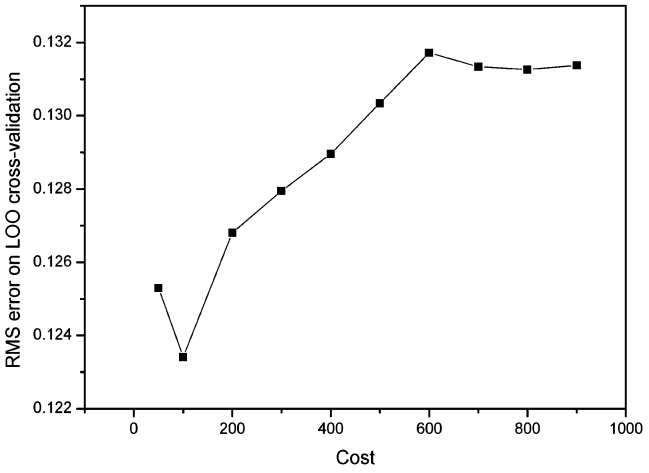

Figure 5. C vs RMS error on LOO cross-validation ($\gamma = 0.031$, $\epsilon = 0.04$).

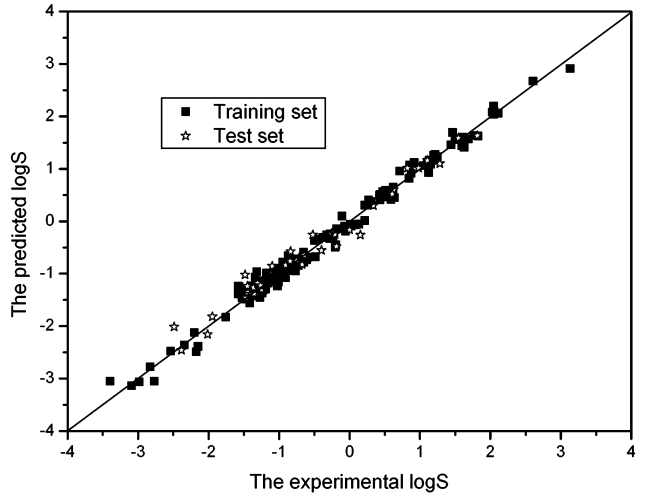

Figure 6. Predicted log S vs experimental log S by SVM.

Through the above process, γ, ϵ, and C were fixed to 0.031, 0.04, and 100, respectively, when the support vector number of the SVM model was 111, the predicted results of the optimal SVM were shown in Table 1 and Figure 6. The model gave an RMS of 0.123 for the training set, 0.170 for the test set, and 0.134 for the whole set, and the corresponding correlation coefficients (R) were 0.995, 0.988, and 0.994, respectively. The statistical parameters of different QSPR models were listed in Table 4, which indicate that the performance of SVM is a little better than that of HM.

TABLE 4: Comparison of Statistical Parameters of Different QSPR Model

	training set		test set		whole set			
approach	RMS	R	RMS	R	RMS	R	F-test	t-test
HM[a]	0.187	0.988	0.219	0.980	0.194	0.986	7615.0039	87.2640
SVM[b]	0.123	0.995	0.170	0.988	0.134	0.994	16263.1263	127.5270

[a] Model of HM. [b] Model of SVM.

5. Conclusion

Linear and nonlinear QSPR models of 217 hydrocarbons were built based on HM and SVM using the topological and electrostatic descriptors. Comparing with the linear and nonlinear models, it is proved that nonlinear SVM model gave better results than those of the linear model. It can be concluded that (1) The proposed models could identify and uncover that topological and electrostatic descriptors are related to the solubility of nonelectrolyte in the electrolyte from the molecular level. (2) The nonlinear model can describe the relationship between the structural parameters and the log S of the 217 hydrocarbons more accurately. (3) SVM proved to be a useful tool in the prediction of the solubility of nonelectrolyte in the electrolyte. It has some advantages over 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.

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) Tolls, J.; van Dijk, J.; Verbruggen, E. J. M.; Hermens, J. L. M.; Loeprecht, B.; Schuurmann, G. J. *Phys. Chem. A* **2002**, *106*, 2760−2765.
(2) Tsonopoulos, C. *Fluid Phase Equilib.* **1999**, *156*, 21−33.
(3) Tsonopoulos, C. *Fluid Phase Equilib.* **2001**, *186*, 185−206.
(4) Chen, W.; Kan, A. T.; Newell, C. J.; Moore, E.; Tomson, M. B. *Groundwater* **2002**, *40*, 153−164.
(5) Mackay, D. *Multimedia Environmental Models-The Fugacity Approach*; Lewis Publishers Inc.: Chelsea, MA, 1991.
(6) Turner, L. H.; Chiew, Y. C.; Ahlert, R. C.; Kosson, D. S. *AIChE J.* **1996**, *42*, 1772−1788.
(7) Hawthorne, S. B.; Yang Yu; Miller, D. J. *Anal. Chem.* **1994**, *66*, 2912−2920.
(8) Wilcock, R. J.; Corban, G. A.; Northcott, G. L.; Wilkins, A. L.; Langdon, A. G. *Environ. Toxicol. Chem.* **1996**, *15*, 670−676.
(9) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(10) Kubinyi, H. *Drug Discovery Today* **1997**, *2*, 457−467.
(11) Kubinyi, H. *Drug Discovery Today* **1997**, *2*, 538−546.
(12) Karelson, M. *Molecular Descriptors in QSAR/QSPR.*; John Wiley & Sons: New York, 2000.
(13) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(14) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers: Amsterdam, 1999.
(15) Pang, S. N.; Kim, D.; Bang, S. Y. *Pattern Recognit. Lett.* **2003**, *24*, 215−225.

(16) Liu, H. X.; Zhang, R. S.; Luan, F.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 900−907.
(17) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(18) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5−14.
(19) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. *Comput. Chem.* **2002**, *26*, 293−296.
(20) Bao, L.; Sun, Z. R. *FEBS Lett.* **2002**, *521*, 109−114.
(21) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288−1296.
(22) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *43*, 161−167.
(23) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 669−677.
(24) Xue, C. X.; Zhang, R. S.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 950−957.
(25) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual*; University of Florida: Gainesville, FL, 1994.
(26) Yaws, Carl L. *Chemical Properties Handbook*; McGraw-Hill: New York, 1999; pp 389−395.
(27) *HyperChem 4.0*, Hypercube, Inc., 1994.
(28) *MOPAC, v.6.0 Quantum Chemistry Program Exchange, Program 455*; Indiana University: Bloomington, IN.
(29) Cortes, C.; Vapnik, V. *Machine Learning* **1995**, *20*, 273−293.
(30) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
(31) Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons: New York, 1998.
(32) Gunn, S. R.; Brown, M.; Bossley, K. M. *Lecture Notes Comput. Sci.* **1997**, *1280*, 313−323.
(33) Venables, W. N. D.; Smith, M. *R manuals*; The R Development Core Team, 2003.
(34) Kier, L. B.; Hall, L. H. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792−795.
(35) Wang, W. J.; Xu, Z. B.; Lu, W. Z.; Zhang, X. Y. *Neurocomputing* **2003**, *55*, 643−663.