Novel Atomic-Level-Based AI Topological Descriptors: Application to QSPR/QSAR Modeling

Biye Ren*

Research Institute of Materials Science, South China University of Technology, Guangzhou 510640, P.R. China

Received January 2, 2002

Novel atomic level AI topological indexes based on the adjacency matrix and distance matrix of a graph is used to code the structural environment of each atomic type in a molecule. These AI indexes, along with Xu index, are successfully extended to compounds with heteroatoms in terms of novel vertex degree υ^m, which is derived from the valence connectivity δ^v of Kier−Hall to resolve the differentiation of heteroatoms in molecular graphs. The multiple linear regression (MLR) is used to develop the structure−property/activity models based on the modified Xu and AI indices. The efficiency of these indices is verified by high quality QSPR/QSAR models obtained for several representative physical properties and biological activities of several data sets of alcohols with a wide range of non-hydrogen atoms. The results indicate that the physical properties studied are dominated by molecular size, but other atomic types or groups have small influences dependent on the studied properties. Among all atomic types, −OH groups seem to be most important due to hydrogen-bonding interactions. On the contrary, −OH groups play a dominant role in biological activities studied, although molecular size is also an important factor. These results indicate that both Xu and AI indices are useful model parameters for QSPR/QSAR analysis of complex compounds.

1. INTRODUCTION

Many physical properties of molecules can be related to their composition and structures. This is very convenient when one needs to be able to predict the properties of any molecule and also from a theoretical viewpoint gives one more molecular understanding of the different forces present in any system. At an early stage, it was accepted that properties could be expressed as a sum of contributions from atoms, bonds, or larger structural subunits. A significant approach is based on various group contribution methods using atomic types, bonds, or molecular fragments as descriptors with respect to different properties or activities. At present, a great deal of quantitative structure−property/activity relationship (QSPR/QSAR) models have been developed by using various known physicochemical parameters such as the multiple solvatochromic parameters, octanol/water partition coefficient (*logP*), and other molecular descriptors such as geometric, electronic or electrostatic, polar, steric, topological ones, and particularly the recently developed three-dimensional (3D) descriptors,[1−4] which have been proven to be very effective in various QSPR/QSAR because they take into account the geometric conformation and the nature of the bonding of groups in a molecule.

In recent years, graph-theoretical topological indices have received considerable attention because they can be derived directly from the molecular structures without any experimental effort.[5,6] Hence, the topological index approach to QSPR/QSAR studies represents a simple and straightforward means from a viewpoint of molecular design. Among conventional indices, well-defined molecular connectivity index (χ),[7] Hosoya's index (Z),[8] Balaban's index (J),[9] Bonchev's index (I_D),[10] Schluze's index (MTI),[11] and Wiener's index (W)[12] are most popular in a variety of QSPR/QSAR studies. However, most of the existing topological indices have some demerits for lack of information about the consideration of multiple bonds and heteroatoms as well as molecular spatial conformation, which limits their field of application. In the first case, several different empirical or unempirical approaches have been introduced in the past few years, for example, Kier and Hall's concept of valence molecular connectivity[7,13] and Estrada' approach of edge weights using quantum-chemical parameters.[14] Further, to reflect molecular three-dimensional conformation, the 3D topological indexes have been developed based on 2D topological indices by means of the molecular mechanics or quantum chemical method,[15] such as Bogdanov's geometric distance matrixes,[16] Randic's approach of graph embedded into three-dimensional space,[17] and Estrada's electron charge density weighted graphs using semiempirical quantum chemical method PM3.[18] In the majority of the cases, the 3D topological indices outperform the corresponding 2D indices in the structure−property/activity correlations due to improvements in heteroatoms differentiation and 3D conformation of a molecule. It is worthwhile to note that the 3D and 2D indices are usually highly correlated with each other; in other words, they usually contain similar structural information to some extent.[19] In some cases, better results can be also obtained using 2D topological indices. For example, in a comparative study of molecular descriptors derived from the distance matrix, it was found that the 3D indices produced inferior structure-boiling point models to 2D ones.[20]

* Corresponding author phone/fax: +86-20-8711-2886; e-mail: renbiye@163.net.

10.1021/ci020362l CCC: $22.00
Published on Web 06/13/2002

On the other hand, we note that the role of each of individual atomic types or groups cannot be directly obtained from the structure−property/activity correlations using 2D and 3D topological indices because they conventionally characterize a molecule as a whole, i.e., molecular size or shape, and do not take into account the separate contributions of each of individual atomic types or groups to properties. Therefore, in many cases, attempts to correlate physical properties or biological activities of complex compounds, especially pharmaceutical molecules and polar protic compounds containing groups such as −OH, $-NH_2$, −COOH, etc., in terms of a single 2D or 3D index are not very successful in the end. This fact implies that atomic types or groups related to different fundamental intermolecular interactions may be important to physical properties or biological activities of a molecule. As is well known, there are many factors influencing the physical property or biological activity of a molecule, such as the molecular size, shape, polarity, and especially the ability of the molecule to participate in hydrogen bonding, which are related to various aspects of intermolecular interactions such as van der Waals forces and hydrogen bonding interactions. For example, alcohols have higher boiling points than ethers with the same skeleton. The higher boiling points for alcohols may be attributed to the contribution of hydrogen-bonding interactions formed among -OH within pure liquid molecules. Now one recognizes that the effects of atomic types and groups (molecular fragments) should be taken into account in QSPR/QSAR modeling.[21]

Recently, Kier and Hall[22] introduced a type of atom-type-based topological indexes called as the electrotopological state (E-state), which is based on the electronic state of each atomic type and its topological nature in a graph. The efficiency of E-state indexes has been verified in a variety of QSPR/QSAR studies of complex molecules.[23−27] However, the development of the atomic-level topological indices is not very advanced. This should be the primary stimulation to find new atomic level topological indices with respect to different physical properties and biological activities. In previous papers,[28−30] a type of new atom-type AI topological indices were derived from the adjacency matrix and distance matrix of a graph to model six properties of alkanes. Further, the high quality models were developed to correlate four physical properties of a small data of alcohols and three physical properties of a mixed set of compounds containing alkanes and alcohols with their chemical structures. The atom-type AI indices offer the possibility of understanding the role of individual groups in molecules.

The main aim of the present study is to further illustrate whether the novel AI indices can be applied to a wide range of physical properties and especially biological activities that depend on the strength of intermolecular interactions such as hydrogen bonding interactions of −OH moieties in molecules or not. Another goal is to use these indices as an aid in deciphering what structural features and groups would be important to the physical properties and biological activities studied. To achieve these goals, first the concept of novel degree of vertex, v^m, is further extended to modify these indices to complex compounds containing different types of multiple bonds and heteroatoms; second, the multiple linear models using the modified Xu and AI indices are developed to correlate the normal boiling points, water solubility and octanol/water partition, narcosis activities, and toxicities of several data sets of alcohols with a wide range of non-hydrogen atoms; third, the role of the structural features in molecules is also discussed in detail from the final regression models; and finally, the models obtained are tested by the cross-validation using a leave-one-out method.

2. METHOD

For a molecular graph $G = \{V, E\}$ with n vertexes, where V and E are the vertex set and edge set, respectively. The vertex-adjacency matrix, $\mathbf{A} = [a_{ij}]_{n\times n}$, is a square symmetric matrix. The elements a_{ij} of matrix **A** are 1 if vertices i and j are adjacent and 0 otherwise. The distance matrix, $\mathbf{D} = [d_{ij}]_{n\times n}$, is also a square symmetric matrix. The entries d_{ij} of matrix **D** are the length of the shortest path between the vertices i and j in a G, i.e., the number of C−C bonds. The sum over any row i (or column i) of matrix **A** yields local vertex-degree v_i; analogously, the sum over any row i (or column i) of matrix **D** yields distance sums s_i. The Xu index can be expressed below[31−34]

$$Xu = n^{1/2} \log\left(\sum_{i=1}^{n} v_i s_i^2 / \sum_{i=1}^{n} v_i s_i\right) \quad (1)$$

where the sum is over all i vertices in a graph.

For any atom i that belongs to jth atomic type in a molecular graph, the corresponding topological index value, $AI_i(j)$, is defined as[28]

$$AI_i(j) = 1 + \phi_i(j) \quad (2)$$

with

$$\phi_i(j) = v_i(j) s_i^2(j) / \sum_{i=1}^{n} v_i s_i \quad (3)$$

where parameter $\phi_i(j)$ is considered as a perturbing term of ith atom reflecting the effect of its structural environment on its $AI_i(j)$ value.

According to this definition, for any atom-type j in a molecular graph, the corresponding AI index of atom-type j, AI(j), is a sum of all $AI_i(j)$ values in a molecular graph

$$AI(j) = \sum_{i=1}^{m} AI_i(j) = m + \sum_{i=1}^{m} \phi_i(j) = m + \sum_{i=1}^{m} v_i(j) s_i^2(j) / \sum_{i=1}^{n} v_i s_i \quad (4)$$

where m is the count of the atoms or groups of the same type. Clearly, the AI value is equal to the count of atomic groups of the same type plus total perturbation terms.

As may be easily seen from above equations, both Xu and AI index combines the vertex degree (v_i) and the distance sums (s_i), and thus these indices are related to their structural environment and are expected to contain more structural information both at the molecular and at the atomic level. Clearly, these indices have the advantage that they can be easily adapted for the presence of multiple bonds and/or heteroatoms by multiplying a weighted parameter by the edge or vertex. To reflect the nature of multiple bonds and/or

heteroatoms, suitable weighted parameters may be the atomic number, the covalent radius, the bond order, and the electronegativity,[5] etc. The well-defined valence connectivity δ^v of Kier-Hall,[7,13] which contains much information on electric states and atomic orbitals of multiple bonds and heteroatoms in molecular structures, may be a suitable candidate for the present study. However, one notes that in many cases the δ^v value cannot be directly used to modify other indices expressed in different formula. Therefore, it is necessary to propose new vertex degree for heteroatoms in graphs to extend Xu and AI indices. This problem has been reasonably solved, as discussed in the following section.

As we know, for a molecular graph, the vertex degree v_i of the ith atom is the number (δ) of connections (edges) of that atom.

$$v_i = \delta \tag{5}$$

We consider that the newly proposed vertex degree should reflect the number of edges of that atom in a graph. Consequently, a novel degree of vertex, v^m, is derived from the valence connectivity δ^v of Kier−Hall[7,13] to take the place of the vertex degree v_i of heteroatoms or carbon atoms linked to multiple bonds in molecular graphs, as is presented as follows

$$v^m = \delta + k \tag{6}$$

with

$$k = 1/[(2/N)^2\delta^v + 1] \tag{7}$$

where parameter k is a perturbing term reflecting the effect of heteroatoms. N is the principal quantum number of valence shell. For heteroatoms in molecular graphs, the δ^v value of Kier-Hall[7,13] is expressed as

$$\delta^v = (Z^v - h)/(Z - Z^v - 1) \tag{8}$$

For carbon atoms with multiple bonds in molecular graphs, δ^v is expressed as

$$\delta^v = Z^v - h \tag{9}$$

where h is the number of hydrogen atoms connected to the heteroatom. Z and Z^v are the atomic number and the number of valence electrons for a heteroatom, respectively.

Clearly, for heteroatoms in molecular graphs, if only we use the new degree of vertex, v^m, instead of the original vertex degree v_i, and then both Xu and AI indices can be expressed with the same formula defined above (eqs 1 and 4).

For an easy comparison, the valence connectivity δ^v of Kier and Hall,[7,13] k parameters, vertex degree v^m, and van der Waals radius r_w of some representative heteroatoms are shown in Table 1. It can be readily observed that the value of k parameter is almost linear with van der Waals radius. This indicates that parameter k as the perturbing term of the connections of heteroatoms in graphs reflects the effect of atomic radius. For example, for −F, −Cl, and −Br groups in halogen-substituted hydrocarbons the number of edges δ is equal to 1, but the k values are 0.35, 0.80, and 0.95, and then the v^m values are 1.35, 1.80, and 1.95, respectively. Also, for the oxygen atom in alcohols and ethers the δ values are 1 and 2, and the k values are 0.167 and 0.143, and thus the v^m values are 1.167 and 2.143 for −OH and −O− groups, respectively. It is obvious that the novel vertex degree v^m contains the information about the number of connections and heteroatomic radius. Therefore the v^m parameter has explicit physical meanings.

Table 1. Valence Connectivity (δ^v),[7] k Parameter, van der Waals' Radius (r_w), and Novel Vertex Degree (v^m) for Representative Atom-Types

groups	δ^v	r_w	k	v^m	groups	δ^v	r_w	k	v^m
-CH_3	1		0	1	≡N	5		0.167	1.167
-CH_2−	2		0	2	−PH_2	0.333	1.90	0.871	1.871
-CH<	3		0	3	>PH	0.444		0.835	2.835
>C<	4		0	4	−P<	0.556		0.802	3.802
=CH_2	2		0.333	1.333	−OH	5	1.40	0.167	1.167
=CH−	3		0.250	2.250	−O−	6		0.143	2.143
=C<	4		0.200	3.200	=O	6		0.143	1.143
=C=	4		0.200	2.200	−SH	0.556	1.85	0.802	1.802
≡CH	3		0.250	1.250	−S−	0.667		0.771	2.771
≡C−	4		0.200	2.200	−F	7	1.35	0.125	1.125
−NH_2	3	1.50	0.250	1.250	−Cl	0.778	1.80	0.743	1.743
>NH	4		0.200	2.200	−Br	0.259	1.95	0.939	1.939
−N<	5		0.167	3.167	−I	0.149	2.15	0.977	1.977

3. MULTIPLE LINEAR REGRESSION ANALYSIS

For individual physical properties and biological activities the multiple linear regression (MLR) using the modified Xu and all atom-type AI indices present in molecular structures is used to develop models correlating physical properties and biological activities with chemical structures. The final model is obtained in the form of eq 10

$$\text{property (activity)} = a_0 + a_1X_u^m + b_1\text{AI}(1) + ... + b_j\text{AI(j)} \tag{10}$$

where X_u^m is the modified Xu index. a_0 is a constant, a_1 is the contribution coefficient of the modified Xu index, and b_j is the contribution coefficient of jth atom-type index, AI-(j). Each coefficient describes the sensitivity of a property to each of the individual indices, so the coefficients of these indices would measure the relative importance of each index. The significance of each index is evaluated by monitoring the statistical parameters such as t-test values (t-values) and Fisher ratios (F) so as to choose a high quality subset of indices.[35−37] The standard error (s) is used to evaluate the quality of model.

4. RESULTS AND DISCUSSION

In general, there are two different directions in which multiple regression analysis is usually used in QSPR/QSAR studies.[38,39] The first approach is based on a large number of compounds with a wide range of structural types; the other way only deals with a smaller group of structurally related compounds. Both approaches have their merits and their shortcomings, and they serve different purposes. The significant advantage of selecting a smaller set of structurally related compounds for QSPR/QSAR studies is the possibility of calculating the relative importance of all possible atomic types or groups to the properties studied.[21] In the present work, we select the second alternative to get some meaningful insights into the structural factors influencing the structure−property/activity relationships. The relative im-

portance of individual structural features is estimated by the contribution of each index, which is estimated by multiplying the coefficients by the mean index values; fraction contributions are calculated by multiplying the absolute values by the coefficient of determination (r^2), i.e., the square of the multiple correlation coefficient r, and dividing by the sum of the absolute values.[40]

Alcohols represent an especially attractive class of polar protic compounds for QSPR/QSAR studies considering the influence of hydrogen-bonding interactions formed among −OH groups. To illustrate the potential of these indexes in QSPR/QSAR studies, two series of examples of applications are analyzed. First, we deal with the relationships between physical properties of alcohols and these indices. Several representative properties such as the normal boiling points (*BP*), water solubility, and octanol/water partitions of alcohols with a wide range of non-hydrogen atoms are selected for this case. The other examples are related to biological activities and toxicities of alcohols.

Correlation to Boiling Points. As a starting point, we select a data set of 138 alcohols up to 17 non-hydrogen atoms to develop the structure-boiling point model. The experimental data of *BP* for 138 alcohols with up to 16 non-hydrogen atoms are listed in Table 2. The majority of the data are taken from refs 41 and 42 and some are from refs 43 and 44. The precision of most data is within 1 °C, but in several cases when boiling points differing by 1−4 °C were reported, their arithmetic mean value is utilized so that in these cases the imprecision is about twice as large. In the majority of cases, the data among the four sources agree fairly well although there are some discrepancies between the four sources. Most of the discrepancies between the four sources are about 1 °C or less. Thus, these discrepancies are ignored. A boiling point model for 138 compounds is then generated using X_u^m and all five AI indices. As an illustration, we give the best five-parameter model (eq 11) below:

$$BP = -12.0683\ (\pm 3.1252) + 41.8616\ (\pm 1.2467)\ X_u^m + 6.7930\ (\pm 0.5718)\ \mathrm{AI}(-\mathrm{OH}) - 0.9839(\pm 0.1046)\mathrm{AI}(>\mathrm{CH_2}) - 2.6927\ (\pm 0.2254)\ \mathrm{AI}(>\mathrm{CH}-) - 5.2437\ (\pm 0.3308)\mathrm{AI}(>\mathrm{C}<) \quad (11)$$

$$r = 0.9957;\ s = 3.576\ ^\circ\mathrm{C};\ F = 3091;\ P < 0.0001;\ N = 138$$

The *t*-values are 3.862, 33.58, 11.88, −9.410, −11.95, and −15.85, respectively. The indices in the final model are not highly correlated with each other, and each coefficient is clearly highly significant. This model explains more than 99% of the variances in the experimental values of *BP* for these compounds. However, for the same 138 compounds, the simple linear regression with X_u^m index leads to a poor correlation ($r = 0.9702$ and $s = 9.277$ °C). Clearly, a single X_u^m index cannot give a simple and accurate correlation. As is well-known, the boiling point is a physical property that strongly depends on the intermolecular interactions and is very much influenced by the molecular weight, while the intermolecular interactions in complex compounds, as pointed out by Pitzer et al.[45] in 1955, are not only interactions between molecular center but also a sum of interactions between various parts of the molecules. According to this point of view, both the molecular size and individual atomic types or groups related to different fundamental interactions make the separate contributions to physical properties of a molecule. For alcohols the hydrogen-bonding interactions formed among −OH groups may be the second important factor. As we expect, inclusion of the AI (−OH) index produces a significantly improved model ($r = 0.9866$ and $s = 6.268$ °C). The improvement in the statistical quality is more than 30% relative to the linear model with X_u^m alone, which indicates that the AI (−OH) index contains information about the hydrogen-bonding interaction of −OH moieties. For the final model, the improvement is more than 60%.

The calculated values and residuals for 138 compounds are shown in Table 2. A comparison of calculated and observed data is shown in Figure 1. One can observe that the agreement between correlation and data is quite good. So the final model (eq 11) represents an excellent QSPR model considering the data from different sources with different experimental accuracy and also judging from the standard error and the plot in Figure 1.

In fact, the boiling point of alcohols is a property not correlated satisfactorily with a single or a small number of descriptors due to the influence of the hydrogen-bonding interaction of −OH moieties in molecules. As reported by Xu et al.,[46] a model using the EAΣ and EAmax indices (EAΣ is the sum of the absolute eigenvalues of expanded adjacency matrix EA. EAmax is the maximum of the absolute eigenvalues of EA.) gave $r = 0.9837$ and $s = 6.35$, but the molecular connectivity $^1\chi$ and $^1\chi^v$ indices provided a slightly superior model to EAΣ and EAmax indices. Analogously, the multiple linear model found by Xu et al.[47] using A_{x1}, A_{x2}, and A_{x3} (A_{x1}, A_{x2}, and A_{x3} indices are half the largest eigenvalues of the corresponding matrices $Z_1 = G_1{*}G'_1$, $Z_2 = G_2{*}G'_2$, and $Z_3 = G_3{*}G'_3$, respectively, where G is the expanded path matrix and G′ is the transpose matrix of matrix G) gave $r = 0.978$ and $s = 7.4988$ for the same series of 37 alcohols. Recently, Galvez et al.[48] reported that the three-parameter model with N (the number of vertices) and two charge indices G_1 and J_2 gave $r = 0.979$ and $s = 3.63$ for 29 alcohols. Kier and Hall[49] also developed a QSPR model to describe *BP* of 28 alcohols using E-state indices and bimolecular encounter parameters ($s = 5.8$ °C). Analogously, Estrada et al.[21] developed a model for the same series of 28 alcohols using a set of novel molecular descriptors based on local spectral moments of the bond matrix. The model accounts for more than 98% of variance ($s = 4.2$ °C). However, in the present study, this problem has been satisfactorily solved in terms of a combined use of X_u^m and AI indexes although our main aim is not to search for the best predictive models.

On the other hand, an understanding of what structural features of a molecule are important to the properties studied could be obtained from the relative or fraction contributions of individual indexes. The contributions of each index in the expression (eq 11) are then calculated according to the aforementioned procedure. The results indicate that X_u^m index makes a major contribution (75%), while other AI indices have smaller contributions, which decrease in the order of AI (−OH) (10%) > AI (>CH_2) (6.1%) > AI

Table 2. Calculated and Experimental Boiling Points (*BP*) of 138 Alcohols

		BP (°C)					*BP* (°C)		
no.	compounds	exp	calcd	res	no.	compounds	exp	calcd	res
1	1-butanol	117.6	108.3	9.3	70	3,4-dimethyl-2-hexanol	165.5	168.4	−2.9
2	2-methyl-1-propanol	107.9	100.6	7.3	71	2,5-dimethyl-2-hexanol	154.5	156.5	−2.0
3	2-butanol	99.5	98.0	1.5	72	4-methyl-4-heptanol	161.0	163.2	−2.2
4	2-methyl-2-propanol	82.4	87.2	−4.8	73	2,4,4-trimethyl-1-pentanol	168.5	164.0	4.5
5	1-pentanol	137.5	132.5	5.0	74	3-ethyl-3-hexanol	160.5	167.1	−6.6
6	3-methyl-1-butanol	131.0	125.3	5.7	75	2,3-dimethyl-2-hexanol	160.0	157.7	2.3
7	2-pentanol	119.3	121.0	−1.7	76	3,5-dimethyl-3-hexanol	158.0	156.0	2.0
8	2-methyl-1-butanol	128.0	123.1	4.9	77	2,3-dimethyl-3-hexanol	158.1	156.6	1.5
9	3-pentanol	116.2	118.9	−2.7	78	2-methyl-3-ethyl-2-pentanol	156.0	156.5	−0.5
10	3-methyl-2-butanol	112.9	113.7	−0.8	79	2,4,4-trimethyl-2-pentanol	147.5	144.7	2.8
11	2,2-dimethyl-1-propanol	113.1	112.6	0.5	80	2,2,4-trimethyl-3-pentanol	150.5	149.4	1.1
12	2-methyl-2-butanol	102.3	108.2	−5.9	81	2,2-dimethyl-3-hexanol	156.0	156.5	−0.5
13	1-hexanol	157.0	155.2	1.8	82	2,5-dimethyl-3-hexanol	157.5	161.3	−3.8
14	4-methyl-1-pentanol	151.9	148.0	3.9	83	4,4-dimethyl-3-hexanol	160.4	157.6	2.8
15	2-hexanol	140.0	142.6	−2.6	84	3,4-dimethyl-2-hexanol	165.5	165.1	0.4
16	3-methyl-1-pentanol	153.0	146.7	6.3	85	6-methyl-2-heptanol	174.0	173.4	0.6
17	2-methyl-1-pentanol	148.0	144.8	3.2	86	3-methyl-1-heptanol	186.0	186.4	−0.4
18	3-hexanol	135.0	139.7	−4.7	87	2-methyl-3-ethyl-3-pentanol	158.0	155.2	2.8
19	2-ethyl-1-butanol	146.5	142.6	3.9	88	2,3,4-trimethyl-3-pentanol	156.5	149.8	6.7
20	4-methyl-2-pentanol	132.0	135.3	−3.3	89	1-nonanol	213.3	214.9	−1.6
21	3,3-dimethyl-1-butanol	143.0	141.8	1.2	90	7-methyl-1-octanol	206.0	206.3	−0.3
22	2,3-dimethyl-1-butanol	144.5	137.8	6.7	91	2-nonanol	198.5	199.0	−0.5
23	2-methyl-2-pentanol	121.5	128.5	−7.0	92	3-nonanol	195.0	195.4	−0.4
24	3-methyl-2-pentanol	134.3	134.2	0.1	93	4-nonanol	192.5	193.4	−0.9
25	2-methyl-3-pentanol	129.5	132.5	−3.0	94	5-nonanol	193.0	192.8	0.2
26	2,2-dimethyl-1-butanol	136.5	132.5	4.0	95	2-methyl-2-octanol	178.0	179.8	−1.8
27	3-methyl-3-pentanol	123.0	127.2	−4.2	96	2,6-dimethyl-2-heptanol	173.0	172.3	0.7
28	3,3-dimethyl-2-butanol	120.4	123.4	−3.0	97	2,6-dimethyl-3-heptanol	175.0	177.3	−2.3
29	2,3-dimethyl-2-butanol	118.4	121.7	−3.3	98	2,6-dimethyl-4-heptanol	174.5	176.3	−1.8
30	1-heptanol	176.4	176.4	0	99	3,6-dimethyl-3-heptanol	173.0	171.3	1.7
31	5-methyl-1-hexanol	170.0	169.0	1.0	100	2,2,3-trimethyl-3-hexanol	156.0	160.7	−4.7
32	2-heptanol	160.4	162.8	−2.4	101	3,5-dimethyl-4-heptanol	171.0	172.4	−1.4
33	4-methyl-1-hexanol	173.0	168.5	4.5	102	2,3-dimethyl-3-heptanol	173.0	172.8	0.2
34	2-methyl-1-hexanol	164.0	165.2	−1.2	103	2,4-dimethyl-4-heptanol	171.0	171.8	−0.8
35	3-heptanol	157.0	159.5	−2.5	104	2-methyl-3-ethyl-3-hexanol	177.5	171.4	6.1
36	3-methyl-1-hexanol	169.0	167.2	1.8	105	2-methyl-3-ethyl-1-hexanol	193.0	191.8	1.2
37	4-heptanol	156.0	158.4	−2.4	106	5-methyl-3-ethyl-3-hexanol	172.0	170.7	1.3
38	5-methyl-2-hexanol	151.0	155.2	−4.2	107	2,4,4-trimethyl-3-hexanol	170.0	165.8	4.2
39	2-methyl-3-hexanol	145.5	151.1	−5.6	108	3,4,4-trimethyl-3-hexanol	165.5	161.4	4.1
40	2-methyl-2-hexanol	143.0	147.3	−4.3	109	4-methy-4-octanol	180.0	179.4	0.6
41	2,4-dimethyl-1-pentanol	159.0	157.9	1.1	110	4-ethyl-4-heptanol	182.0	177.9	4.1
42	5-methyl-3-hexanol	148.0	151.9	−3.9	111	2-methy-2-octanol	178.0	179.8	−1.8
43	3-methyl-3-hexanol	143.0	146.0	−3.0	112	1-decanol	231.1	232.3	−1.2
44	2,4-dimethyl-2-pentanol	133.1	140.0	−6.9	113	8-methyl-1-nonanol	219.9	222.8	−2.9
45	2,4-dimethyl-3-pentanol	140.0	144.1	−4.1	114	2-decanol	211.0	215.2	−4.2
46	3-ethyl-3-pentanol	142.0	144.4	−2.4	115	4-decanol	210.5	209.3	1.2
47	2,3-dimethyl-2-pentanol	139.7	140.1	−0.4	116	3,7-dimethyl-1-octanol	212.5	212.0	0.5
48	2,3-dimethyl-3-pentanol	139.0	139.2	−0.2	117	2,7-dimethyl-3-octanol	193.5	191.6	1.9
49	2,3,3-trimethyl-2-butanol	130.5	129.4	1.1	118	2,6-dimethyl-4-octanol	195.0	191.5	3.5
50	3-methyl-2-hexanol	151.0	153.7	−2.7	119	2,3-dimethyl-3-octanol	189.0	187.7	1.3
51	1-octanol	195.2	196.3	−1.1	120	5-methyl-5-nonanol	202.0	194.5	7.5
52	6-methyl-1-heptanol	188.6	188.4	0.2	121	4-methyl-1-nonanol	216.0	223.1	−7.1
53	2-octanol	180.0	181.6	−1.6	122	2-methyl-3-nonanol	200.0	201.1	−1.1
54	3-octanol	175.0	178.1	−3.1	123	2,2,5,5-tetramethyl-3-hexanol	170.0	174.1	−4.1
55	4-methyl-1-heptanol	188.0	187.9	0.1	124	4-propyl-4-heptanol	191.0	192.7	−1.7
56	4-octanol	176.3	176.4	−0.1	125	2,4,6-trimethyl-4-heptanol	181.0	178.4	2.6
57	2-ethyl-1-hexanol	184.6	180.8	3.8	126	3-ethyl-3-octanol	199.0	193.7	5.3
58	2-methyl-2-heptanol	156.0	164.4	−8.4	127	3-ethyl-2-methyl-3-heptanol	193.0	186.7	6.3
59	2,5-dimethyl-1-hexanol	179.5	176.4	3.1	128	1-undecanol	245.0	248.5	−3.5
60	5-methyl-2-heptanol	172.0	173.7	−1.7	129	2-undecanol	228.0	230.1	−2.1
61	6-methyl-3-heptanol	174.0	170.0	4.0	130	3-undecanol	229.0	226.5	2.5
62	3,5-dimethyl-1-hexanol	182.5	179.0	3.5	131	5-undecanol	229.0	222.7	6.3
63	3-methyl-2-heptanol	166.1	168.9	−2.8	132	6-undecanol	228.0	222.3	5.7
64	2-methyl-3-heptanol	167.5	165.3	2.2	133	1-dodecanol	261.9	263.5	−1.6
65	2-methyl-4-heptanol	164.0	168.6	−4.6	134	2-dodecanol	246.0	243.9	2.1
66	5-methyl-3-heptanol	172.0	170.1	1.9	135	1-tridecanol	276.0	277.5	−1.5
67	3-methyl-3-heptanol	163.0	163.5	−0.5	136	1-tetradecanol	289.0	290.4	−1.4
68	4-methyl-3-heptanol	170.0	169.8	0.2	137	1-pentadecanol	304.9	302.3	2.6
69	3-methyl-4-heptanol	162.0	168.6	−6.6	138	1-hexadecanol	312.0	313.1	−1.1

(>CH-) (4.8%) > AI (>C<) (3.5%). In a previous paper,[31] X_u index was interpreted as a parameter characterizing molecular size. Therefore, the fact indicates that the boiling points of alcohols are dominated by molecular size, but the hydrogen bonding interaction formed among the −OH group is also an important factor. This is in accordance with the polar character of alcohols. It is worth noting that AI (−OH) in primary alcohols makes a greater contribution to *BP* than

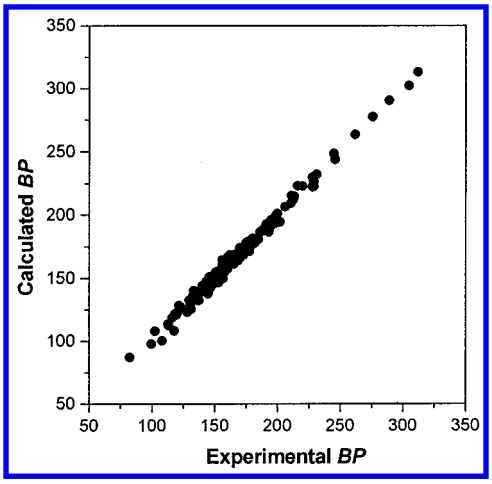

Figure 1. A plot of calculated versus experimental *BP* for 138 alcohols.

that in secondary or tertiary alcohols. The relative contribution of the AI (−OH) index decreases from 29.3 in primary alcohols to 18.8 in tertiary alcohols. This can be readily understood because in tertiary alcohols the $-CH_3$ groups located on a quaternary carbon (to which the −OH group is directly attached) prevent close contact with neighboring molecules and alter the average intermolecular distances. Since the intermolecular interaction energies fall sharply with the average intermolecular distances, the hydrogen bonding strength of −OH groups is reduced. The net result is that secondary or tertiary alcohols have a lower boiling point than primary alcohols with the same non-hydrogen atoms as observed in Table 2. AI (−OH) seems to simulate this effect and may be used as a simple estimation of the hydrogen-bonding strengths of −OH moieties in alcohols with different chemical structures. Our results seem to indicate that both X_u^m and AI indices are very sensible for describing *BP* of alcohols and especially the AI (−OH) index which contains information about the hydrogen-bonding interaction, which is important to explaining some processes such as aqueous solubility, octanol/water partition, and other biological processes.

Correlation to Water Solubilities and Octanol/Water Partition. As an extension of the above study, we select two data sets of alcohols with their aqueous solubility and octanol/water partition to develop the structure−property models. Aqueous solubility is a particularly important property of organic compounds and is widely applied in the field of pharmaceutical chemistry, biological chemistry, and environmental science. It is also valuable in understanding drug transport and environment impact. The experimental water solubilities as *log*(1/*S*), where *S* is the solubility in moles per liter, are listed in Table 3 for 63 alcohols. These data, which have been critically evaluated to be reliable, are taken from ref 50.

A model is then developed using X_u^m and all AI indexes. We give the best three-parameter model (eq 12) below:

$$log(1/S) = -2.4551\ (\pm 0.09403) + 0.9675\ (\pm 0.0313)X_u^m + 0.1518\ (\pm 0.03768)\mathrm{AI}(-\mathrm{OH}) - 0.04703\ (\pm 0.0123)\mathrm{AI}(>\mathrm{C}<) \quad (12)$$

$$r = 0.9874;\ s = 0.1669;\ F = 764;\ P < 0.0001;\ N = 63$$

The *t*-values are −26.11, 30.91, 4.028, and −3.824, respectively. Clearly, each coefficient in this equation is highly significant. This model explains more than 97% of the variances in the experimental values of *log*(1/*S*) for these

Table 3. Calculated and Experimental Water Solubility *log*(1/*S*) of 63 Alcohols

			log(1/*S*)				*log*(1/*S*)		
no.	compounds	exp	calcd	res	no.	compounds	exp	calcd	res
1	ethanol	−1.10	−1.40	0.30	33	5-methyl-2-hexanol	1.38	1.47	−0.09
2	1-propanol	−0.62	−0.75	0.13	34	2-methyl-3-hexanol	1.32	1.34	−0.02
3	1-butanol	−0.03	−0.08	0.05	35	2-methyl-2-hexanol	1.07	1.16	−0.09
4	2-methyl-1-propanol	−0.10	−0.22	0.12	36	2,2-dimethyl-1-pentanol	1.52	1.21	0.31
5	2-butanol	−0.47	−0.28	−0.19	37	4,4-dimethyl-1-pentanol	1.55	1.37	0.18
6	1-pentanol	0.59	0.57	0.02	38	2,4-dimethyl-1-pentanol	1.60	1.47	0.13
7	3-methyl-1-butanol	0.51	0.45	0.06	39	3-methyl-3-hexanol	0.98	1.08	−0.10
8	2-pentanol	0.28	0.35	−0.07	40	2,4-dimethyl-2-pentanol	0.93	1.05	−0.12
9	2-methyl-1-butanol	0.46	0.37	0.09	41	2,4-dimethyl-3-pentanol	1.22	1.18	0.04
10	3-pentanol	0.21	0.28	−0.07	42	3-ethyl-3-pentanol	0.83	1.01	−0.18
11	3-methyl-2-butanol	0.18	0.21	−0.03	43	2,3-dimethyl-2-pentanol	0.87	0.99	−0.12
12	2-methyl-2-butanol	−0.15	0.01	−0.16	44	2,3-dimethyl-3-pentanol	0.84	0.96	−0.12
13	1-hexanol	1.21	1.19	0.02	45	2,2-dimethyl-3-pentanol	1.15	0.99	0.16
14	4-methyl-1-pentanol	1.14	1.09	0.05	46	2,2,3-trimethyl-3-butanol	1.27	1.19	0.08
15	2-hexanol	0.87	0.97	−0.10	47	2,3,3-trimethyl-2-butanol	0.71	0.72	−0.01
16	2-methyl-1-pentanol	1.11	0.98	0.13	48	1-octanol	2.35	2.39	−0.04
17	3-hexanol	0.80	0.86	−0.06	49	2-octanol	2.09	2.16	−0.07
18	2-ethyl-1-butanol	1.01	0.90	0.11	50	2-ethyl-1-hexanol	2.11	2.02	0.09
19	4-methyl-2-pentanol	0.79	0.85	−0.06	51	2-methyl-2-heptanol	1.72	1.72	0
20	3,3-dimethyl-1-butanol	0.50	0.81	−0.31	52	3-methyl-3-heptanol	1.60	1.63	−0.03
21	2,3-dimethyl-1-butanol	0.37	0.86	−0.49	53	1-nonanol	3.01	2.97	0.04
22	2-methyl-2-pentanol	0.49	0.59	−0.10	54	7-methyl-1-octanol	2.49	2.90	−0.41
23	3-methyl-2-pentanol	0.71	0.78	−0.07	55	2-nonanol	2.74	2.74	0
24	2-methyl-3-pentanol	0.70	0.74	−0.04	56	3-nonanol	2.66	2.59	0.07
25	2,2-dimethyl-1-butanol	0.91	0.65	0.26	57	4-nonanol	2.59	2.51	0.08
26	3-methyl-3-pentanol	0.36	0.53	−0.17	58	5-nonanol	2.49	2.48	0.01
27	3,3-dimethyl-2-butanol	0.61	0.51	0.10	59	2,6-dimethyl-4-heptanol	2.16	2.31	−0.15
28	2,3-dimethyl-2-butanol	0.37	0.47	−0.10	60	3,5-dimethyl-4-heptanol	2.51	2.00	0.51
29	1-heptanol	1.81	1.80	0.01	61	2,2-diethyl-1-pentanol	2.42	2.10	0.32
30	2-heptanol	1.55	1.57	−0.02	62	1-decanol	3.63	3.53	0.10
31	3-heptanol	1.44	1.45	−0.01	63	1-dodecanol	4.67	4.63	0.04
32	4-heptanol	1.40	1.41	−0.01					

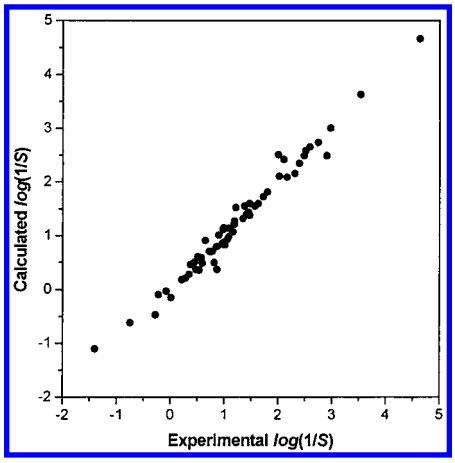

Figure 2. A plot of calculated versus experimental *log*(1/*S*) for 62 alcohols.

compounds. It should be mentioned that a single X_u^m index yields a slightly poor correlation ($r = 0.9772$ and $s = 0.220$). Here the results indicate again that a single X_u^m index related to molecular size cannot model aqueous solubility satisfactorily. The two-variable regression based on a combined use of X_u^m and AI (−OH) produces an obviously improved model with $r = 0.9842$ and $s = 0.1848$. Here, the role of −OH groups is evidenced again. Finally, the best correlation is obtained in terms of up to three indices (eq 12). The calculated values and residuals for 63 alcohols are shown in Table 3. A plot of calculated versus experimental data is shown in Figure 2.

The octanol/water partition (*logP*) is also a particularly important pharmaceutical property of organic compounds and has often been used to represent molecular lipophilicity, which seems to be a key factor related to the transport process through cell membranes and to many other biological events.[51] In particular, *logP* is a crucial parameter in QSAR/QSPR studies and drug design. The compounds and the experimental data of *logP* are shown in Table 4 for 62 alcohols. The majority of experimental data in this data set are taken from ref 3 and only several are from ref 51. The best two-parameter model is obtained, as shown below:

$$\begin{aligned} logP = &-1.4141\ (\pm 0.04999) + \\ &0.9130\ (\pm 0.02852) X_u^m + \\ &0.1288\ (\pm 0.03079) \mathrm{AI}\ (-\mathrm{OH}) \quad (13) \end{aligned}$$

$$r = 0.9958;\ s = 0.1519;\ F = 3488;\ P < 0.0001;\ N = 62$$

The *t*-values are −28.28, 32.02, and −4.182, respectively. This model explains more than 99% of the variances in the experimental data of *logP* for these compounds. On the other hand, X_u^m yields a correlation ($r = 0.9946$; $s = 0.1715$) slightly inferior to that obtained by a combined use of X_u^m and AI (−OH). The role of −OH groups in molecules is evidenced by AI (−OH) index again. Predicted values and residuals for 62 compounds are shown in Table 4. A plot of calculated versus experimental data is shown in Figure 3. One can observe from Tables 3 and 4 that calculated values are very close to experimental data. Hence, both eqs 12 and 13 represent excellent QSPR models judging from the standard error and plots in Figures 2 and 3.

Table 4. Calculated and Experimental Octanol/Water Partition (*logP*) of 62 Alcohols

		logP					*logP*		
no.	compounds	exp	calcd	res	no.	compounds	exp	calcd	res
1[a]	ethanol	−0.31	−0.50	0.19	32	3-heptanol	2.31	2.23	−0.08
2[a]	1-propanol	0.34	0.16	0.18	33	4-heptanol	2.31	2.19	0.12
3[a]	2-propanol	0.05	0.02	0.03	34	5-methyl-2-hexanol	2.19	2.24	−0.05
4	1-butanol	0.84	0.79	0.05	35	2-methyl-3-hexanol	2.19	2.13	0.06
5[a]	2-methyl-1-propanol	0.65	0.66	−0.01	36	2-methyl-2-hexanol	1.84	2.13	−0.29
6	2-butanol	0.61	0.60	0.01	37	2,2-dimethyl-1-pentanol	2.39	2.14	0.25
7[a]	2-methyl-2-propanol	0.37	0.46	−0.09	38	4,4-dimethyl-1-pentanol	2.39	2.31	0.08
8[a]	1-pentanol	1.40	1.39	0.01	39	2,4-dimethyl-1-pentanol	2.19	2.24	−0.05
9[a]	3-methyl-1-butanol	1.42	1.28	0.14	40	3-methyl-3-hexanol	1.87	2.03	−0.16
10	2-pentanol	1.14	1.19	−0.05	41	2,4-dimethyl-2-pentanol	1.67	2.02	−0.35
11	2-methyl-1-butanol	1.14	1.21	−0.07	42	2,4-dimethyl-3-pentanol	2.31	1.98	0.33
12	3-pentanol	1.14	1.13	0.01	43	3-ethyl-3-pentanol	1.87	1.95	−0.08
13	3-methyl-2-butanol	1.14	1.06	0.08	44	2,3-dimethyl-2-pentanol	2.27	1.96	0.31
14	2,2-dimethyl-1-propanol	1.36	1.08	0.28	45	2,3-dimethyl-3-pentanol	1.67	1.91	−0.24
15	2-methyl-2-butanol	0.89	1.00	−0.11	46	2,2-dimethyl-3-pentanol	2.27	1.96	0.31
16[a]	1-hexanol	2.03	1.97	0.06	47	1-octanol	3.15	3.09	0.06
17	4-methyl-1-pentanol	1.78	1.88	−0.10	48	2-octanol	2.84	2.89	−0.05
18	2-hexanol	1.61	1.77	−0.16	49	2-ethyl-1-hexanol	2.84	2.76	0.08
19	2-methyl-1-pentanol	1.78	1.78	0	50	1-nonanol	3.57	3.63	−0.06
20	3-hexanol	1.61	1.68	−0.07	51	2-nonanol	3.36	3.43	−0.07
21	2-ethyl-1-butanol	1.78	1.71	0.07	52	3-nonanol	3.36	3.30	0.06
22	4-methyl-2-pentanol	1.67	1.66	0.01	53	4-nonanol	3.36	3.23	0.13
23	3,3-dimethyl-1-butanol	1.57	1.71	−0.14	54	5-nonanol	3.36	3.20	0.16
24	2-methyl-2-pentanol	1.39	1.57	−0.18	55	2,6-dimethyl-4-heptanol	3.13	3.04	0.09
25	3-methyl-2-pentanol	1.67	1.60	0.07	56	1-decanol	4.01	4.16	−0.15
26	2-methyl-3-pentanol	1.67	1.56	0.11	57[a]	2-undecanol	4.42	4.47	−0.05
27	2,2-dimethyl-1-butanol	1.57	1.60	−0.03	58	1-dodecanol	5.13	5.18	−0.05
28	3-methyl-3-pentanol	1.39	1.50	−0.11	59	1-tetradecanol	6.11	6.17	−0.06
29	3,3-dimethyl-2-butanol	1.19	1.47	−0.28	60	1-pentadecanol	6.64	6.65	−0.01
30	2,3-dimethyl-2-butanol	1.17	1.44	−0.27	61	1-hexadecanol	7.17	7.13	0.04
31	1-heptanol	2.34	2.54	−0.20	62	1-octadecanol	8.22	8.06	0.16

[a] From ref 51.

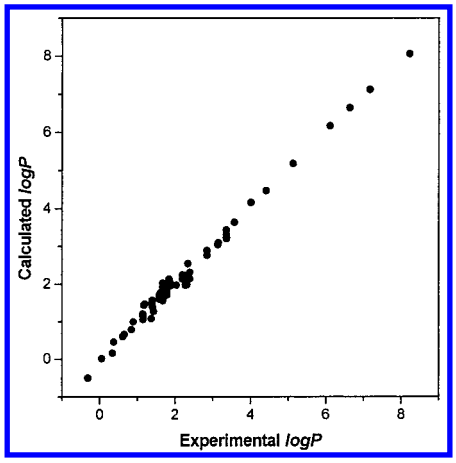

Figure 3. A plot of calculated versus experimental *logP* for 62 alcohols.

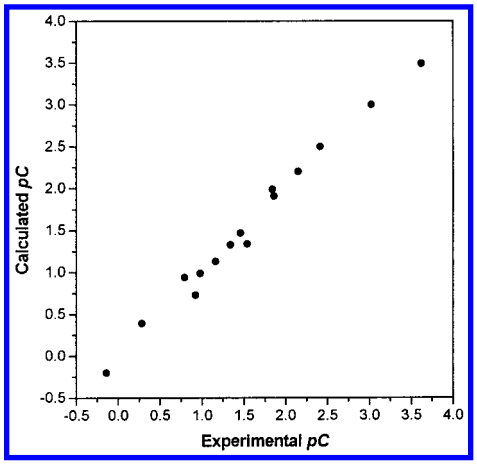

Figure 4. A comparison of calculated and observed narcosis activities (*pC*) of 15 alcohols on barnacle larvae.

Table 5. Narcosis Activities (*pC*) of 15 Alcohols on Barnacle Larvaes

				pC		
no.	compounds	X_u^m	AI(−OH)	exp	calcd	res
1	methanol	0	1.5385	−0.14	−0.20	0.06
2	ethanol	0.7243	2.0002	0.28	0.39	−0.11
3	1-propanol	1.3798	2.4486	0.79	0.94	−0.15
4	2-propanol	1.2619	2.1748	0.92	0.73	0.19
5	1-butanol	2.0022	2.8923	1.46	1.47	−0.01
6	2-methyl-1-propanol	1.8767	2.7668	1.54	1.34	0.20
7	2-butanol	1.8722	2.4003	1.16	1.13	0.03
8	2-methyl-2-propanol	1.7298	2.266	0.98	0.99	−0.01
9	1-pentanol	2.5996	3.3339	1.84	1.99	−0.15
10	3-methyl-1-butanol	2.4843	3.2796	1.86	1.91	−0.05
11	2-methyl-2-butanol	2.3083	2.3948	1.34	1.33	0.01
12	1-hexanol	3.1763	3.7743	2.41	2.50	−0.09
13	1-heptanol	3.7350	4.2139	3.02	3.00	0.02
14	1-octanol	4.2779	4.6531	3.62	3.49	0.13
15	benzyl alcohol	3.3931	3.0798	2.15	2.20	−0.05

The fact that in both cases a good correlation can be obtained with the addition of AI (−OH) into the equation indicates that AI (−OH) index reflects the character of polar protic alcohols. This may be readily understood because the −OH group would participate in the hydrogen bonding interactions with water, which would influence the magnitude of water solubility and octanol/water partition. This may be further illustrated by analyzing contributions of each index from the expressions (eqs 12 and 13). The fraction contributions to water solubility are 84%, 12%, and 1.5% for X_u^m, AI (−OH), and AI (>C<), respectively; the contributions to octanol/water partition are 87% and 12% for X_u^m and AI (−OH), respectively. The results clearly show that although X_u^m makes a major contribution to both aqueous solubility and octanol/water partition, which indicates the additive behavior of the two properties, while other atomic groups, especially −OH groups, are also important factors influencing the values of the two properties. The results clearly reveals why a single X_u^m index cannot account for the difference between observed values of alcohol isomers, which have been satisfactorily probed by AI indexes, especially the AI (−OH) index.

Correlations to Biological Activities. In this section, we will provide other examples of applications of these novel topological indexes with the aim to further verify their applicability to biological activities and toxicities. Both narcosis activities of alcohols on barnacle larvae and toxicities of alcohol steams on tomatoes and spiders are used in this study.

First, let us focus attention on the narcosis activity of alcohols on barnacle larvae. The data of narcosis activities of 15 alcohols are directly taken from ref 13 and listed in Table 5. The narcosis activities are recorded in terms of *pC* values, where $pC = log(1/C)$ and C is the molar concentration that elicits a constant biological response. A model for narcosis activities of 15 alcohols is then generated. The best two-parameter model is obtained as shown as follows:

$$pC = -1.0671\ (\pm 0.1561) + 0.4499\ (\pm 0.07796) X_u^m + 0.5662\ (\pm 0.1043)\ \mathrm{AI}(-\mathrm{OH}) \quad (14)$$

$$r = 0.9939;\ s = 0.117;\ F = 491;\ P < 0.0001;\ \text{and}\ N = 15$$

The *t*-values are −6.835, 5.771, and 5.427, respectively. This model explains more than 98% of the variances in the experimental *pC* values of narcosis activities for 15 alcohols; particularly the standard error is only 0.117. This model is favorably compared with that obtained by using the molecular connectivity $^1\chi$ and $^1\chi^v$ ($r = 0.987$; $s = 0.163$). The calculated *pC* values from eq 14 and residuals are shown in Table 5. The agreement between correlation and data is quite good. A comparison of calculated and observed data is shown in Figure 4. It is worth noting that in contrast to the physical properties studied, AI (−OH) index makes a major contribution to narcosis activities of alcohols because the fraction contributions to narcosis activity are 37% and 62% for X_u^m and AI (−OH), respectively. As is well known, the interactions of −OH groups present in small molecules with biological macromolecules are carried through the hydrogen bonding formed between them. The strength of hydrogen bonding interactions is mainly controlled by the position of −OH group in a molecule, which have been reasonably differentiated and characterized in terms of AI (−OH) index. The larger contributions of AI (−OH) index reveal the influence of the hydrogen bonding interaction formed between alcohols and biological macromolecules and suggest that the hydrogen bonding interaction plays an important role in these biological processes. This may be further illustrated by the following examples.

Table 6. Toxicities (*PC*) of 14 Alcohols on Tomatos and Spiders

				pC^a			pC^b		
no.	compounds	X_u^m	AI (−OH)	exp	calcd	res	exp	calcd	res
1	methanol	0	1.5385	2.60	2.52	0.08	2.80	2.67	0.13
2	ethanol	0.7243	2.0002	2.76	2.92	−0.16	3.00	3.06	−0.06
3	1-propanol	1.3798	2.4486	3.33	3.30	0.03	3.32	3.42	−0.10
4	2-propanol	1.2619	2.1748	3.18	3.16	0.02	3.26	3.29	−0.03
5	1-butanol	2.0022	2.8923	3.69	3.66	0.03	3.77	3.76	0.01
6	2-methyl-1-propanol	1.8767	2.7668	3.57	3.57	0	3.72	3.68	0.04
7	2-butanol	1.8722	2.4003	3.46	3.44	0.02	3.62	3.56	0.06
8	2-methyl-2-propanol	1.7298	2.2660	3.41	3.35	0.06	3.28	3.48	−0.20
9	1-pentanol	2.5996	3.3339	4.05	4.01	0.04	4.09	4.10	−0.01
10	3-methyl-1-butanol	2.4843	3.2796	3.95	3.95	0	4.09	4.04	0.05
11	2-pentanol	2.4789	2.6804	3.77	3.74	0.03	3.90	3.86	0.04
12	2-methyl-1-butanol	2.4456	3.0508	3.77	3.86	−0.09	3.96	3.96	0
13	3-pentanol	2.4404	2.4734	3.69	3.66	0.03	3.81	3.78	0.03
14	2-methyl-2-butanol	2.3083	2.3948	3.51	3.59	−0.08	3.75	3.71	0.04

[a] Toxicities on tomatoes. [b] Toxicities on spiders.

Toxicity of organic compounds is one of the particular interesting biological activities in the scientific community due to its impact on environment and human health. The toxicities of 14 alcohols on tomatoes and spiders are directly taken from ref 13 and shown in Table 6, where the toxicities (*pC*) is 50% inhibitory growth impairment concentration ($-log LC_{50}$). First, the model for describing toxicities on tomatoes is generated. We give the best two-parameter model below:

$$pC = 1.9972\ (\pm 0.1275) + 0.3330\ (\pm 0.04908) X_u^m + 0.3433\ (\pm 0.07559)\ \mathrm{AI}\ (-\mathrm{OH}) \quad (15)$$

$$r = 0.9872\ s = 0.0716;\ F = 211;\ P < 0.0001;\ \text{and}\ N = 14$$

The *t*-values are 15.67, 6.784, and 4.541, respectively. Each coefficient in eq 15 is clearly highly significant. This model explains more than 97% of the variances in the experimental values of *pC* for 14 alcohols with a fit error of only 2.1%.

For toxicity on spiders, we obtain the following two-variable model:

$$pC = 2.1873\ (\pm 0.1516) + 0.3334\ (\pm 0.05837) X_u^m + 0.3140\ (\pm 0.08989)\ \mathrm{AI}\ (-\mathrm{OH}) \quad (16)$$

$$r = 0.9809\ s = 0.0852;\ F = 140;\ P < 0.0001;\ \text{and}\ N = 14$$

The *t*-values are 14.43, 5.712, and 3.494, respectively. Each coefficient is also highly significant. This model explains more than 96% of the variances in the experimental values of *pC* for 14 alcohols with a fit error of only 2.37%. The calculated *pC* values and residuals are shown in Table 5. A comparison of calculated and observed toxicities is shown in Figure 5.

As we expect, in both cases a good correlation is obtained when the AI (−OH) index is utilized as the second parameter, especially the AI (−OH) index has slightly larger contributions than X_u^m index. For example, the contributions of X_u^m and AI (−OH) to toxicities on tomatoes are 40% and 57%, respectively; the contributions of X_u^m and AI (−OH) to toxicities on spiders are 41.6% and 54.6%, respectively.

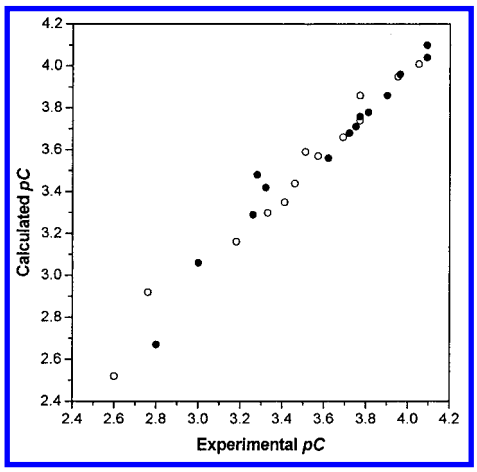

Figure 5. A comparison of calculated and observed toxicities (*pC*) of 14 alcohols on tomatoes (○) and spiders (●).

These examples further support our above belief that both narcosis activities and toxicities of alcohols tend to be more sensitive to molecular fragments or atomic groups related to hydrogen-bonding interactions. The present study further verifies the high potential of these indices for application to biological activities or toxicities.

Model Validation. The cross-validation is a practical and reliable method for testing the significance of a model. Hence, the cross-validation using a leave-one-out method is used to determine the validity of all models obtained. In principle, the performance of the model (its predictive ability) can be given by the standard error of prediction (*SEP*) defined as[36]

$$SEP = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N}} \quad (17)$$

where y_i and $\hat{y}_i$ is the experimental and predictive value, respectively. *N* is the number of samples used for model building.

The predictive ability of the model is also quantified in terms of the corresponding leave-one-out cross-validated

Table 7. Statistics of MLR and Leave-One-Out Cross-Validation for the Six Final Models

properties	r	s	SEP	r_{cv}	s_{cv}	SEP_{cv}
BP	0.9957	3.576	3.499	0.9953	3.742	3.660
$log(1/S)$	0.9874	0.167	0.160	0.9849	0.180	0.175
logP	0.9958	0.152	0.148	0.9954	0.159	0.155
pC^a	0.9939	0.117	0.106	0.9911	0.142	0.127
pC^b	0.9872	0.0716	0.0633	0.9753	0.0992	0.0880
pC^c	0.9809	0.0852	0.0772	0.9619	0.120	0.106

[a] Narcosis activities on barnacle larvae. [b] Toxicities on tomatoes. [c] Toxicities on spiders.

parameters, r_{cv}^2 and s_{cv} values, which are defined as[52]

$$r_{cv}^2 = 1.0 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y)^2} \tag{18}$$

where $\bar{y}$ is the mean value of y_i and

$$s_{cv} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N - M - 1}} \tag{19}$$

where M is the numbers of descriptors.

As a quantitative evaluation of the cross-validated results, the statistical parameters (r, s, and SEP) of the final models and the corresponding cross-validated statistical parameters (r_{cv}, s_{cv}, and SEP_{cv}) of the jack-knifed approach are also listed in Table 7. It is expected that the cross-validated SEP_{cv} of the jack-knifed approach should be slightly larger than SEP of the final models. One can see that for each property or activity studied the r_{cv} and s_{cv} values of the cross-validated approach are very close to the r and s values of the final models, and both s_{cv} and SEP_{cv} of the jack-knifed approach are only slightly larger than s and SEP of the final models. This cross-validation demonstrates the final models to be statistically reliable. On the other hand, plots of calculated versus observed values and plots of residuals versus calculated values can also be used as evidence of the validity of a model. Plots of residuals versus calculated values show that the residuals are randomly distributed. Plots of calculated versus observed data show no obviously observable pattern (Figures 1−5). Therefore, the final models are statistically significant and successfully validated.

Finally, it should be mentioned that both Xu and AI indices show an excellent discrimination power of isomers for all compounds in the database, indicating that the υ^m values are adequate to extension of Xu and AI indices to compounds with heteroatoms.

5. CONCLUSION

Novel atom-type AI topological indexes based on the adjacency matrix and distance matrix of molecular graphs is used to code the structural environment of each atom-type in a molecular graph. Both AI and Xu indexes are successfully extended to compounds with heteroatoms in terms of novel vertex degree, υ^m, which is derived from the valence connectivity δ^v of Kier−Hall to resolve heteroatoms differentiation in a graph. The results indicate that the proposed approach is adequate to modification of these indices.

The efficiency of these indices is demonstrated through several QSPR/QSAR examples. The high quality QSPR models have been obtained by a combined use of Xu and AI indices for the properties and activities of several data sets of alcohols. These models not only offer a better understanding of the role of each structural feature in a molecule but also illustrate the high potential of these indices for application to a wide range of physical properties and biological activities. Furthermore, the results indicate that the physical properties studied are dominated by molecular size, but atomic types or groups have an important effect, especially the −OH group seems to be most important due to hydrogen-bonding interactions. In contrast to physical properties studied, the biological activities studied are more sensitive to hydrogen-bonding interactions formed among −OH groups. Therefore, it is expected that both Xu and AI indices can be used as QSPR/QSAR model parameters in the future.

NOTE ADDED AFTER ASAP POSTING

References 9 and 30 have been modified. The correct version was posted 7/22/2002.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(2) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995.

(3) Xu, L.; Hu, C. *Applied Chemical Graph Theory*; Scientific Press: Beijing, 2000.

(4) Puri, S.; Chickos, J. S.; Welsh, W. J. Three-dimensional Quantitative Structure−Property Relationship (3D-QSPR) Models for Prediction of Thermodynamic Properties of Polychlorinated biphenyls (PCBs): Enthalpy of Sublimation. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 109−116

(5) Balaban, A. T. Chemical Graphs: Looking Back and Glimpsing Ahead. *J. Chem. Inf. Comput., Sci.* **1995**, *35*, 339−350

(6) Trinajstic, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, 1992.

(7) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(8) Hosoya, H. Topological Index. A Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339

(9) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(10) Bonchev, D.; Trinajstic, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.

(11) Schultz, H. P. Topological Organic Chemistry, 1. Graph Theory and Topological Indices. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227−228.

(12) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(13) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Studies*; Research Studies Press: Letchworth, 1986.

(14) Estrada, E. Edge Adjacy Relationships in Molecular Graphs Containing Heteroatoms: A New Topological Index Related to Molar Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 701−707.

(15) Diudea, M. V.; Horvath, D.; Graovac, A. Molecular Topology. 15. 3D Distance Matrices and Related Topological Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 129−135.

(16) Bogdanov, B.; Nikolic, S.; Trinajstic, N. On the Three-Dimensional Wiener Number. *J. Math. Chem.* **1989**, *3*, 299−309.

(17) Randic, M.; Jerman-Blazic, B.; Trinajstic, N. Development of 3-Dimensional Molecular Descriptors. *Comput. Chem.* **1990**, *14*, 237−246.

(18) Estrada, E. Three-dimensional Molecular Descriptors Based on Electron Charge Density Weighted Graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 708−713.
(19) Toropov, A.; Toropova, A.; Ismailov, T.; Bonchev, D. 3D Weighting of Molecular Descriptors for QSPR/QSAR by the Method of Ideal Symmetry (MIS): 1. Application to Boiling Points of Alkanes. *J. Mol. Struct. (THEOCHEM)* **1998**, *424*, 237−247.
(20) Mihalic Z.; Nikolic, S.; Trinajstic, N. Comparative Study of Molecular Descriptors Derived from the Distance Matrix. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 28−37.
(21) Estrada, E.; Molina, E. Novel Local (fragment-based) Topological Molecular Descriptors for QSPR/QSAR and Molecular Design. *J. Mol. Grap. Mod.* **2001**, *20*, 54−64.
(22) Kier, L. B.; Hall, L. H.; Frazer, J. W. An index of Electrotopological State for Atoms in Molecules. *J. Math. Chem*. **1991**, *7*, 229−241.
(23) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structural Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(24) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat*. **1991**, *10*, 43−51.
(25) Hall, L. H.; Kier, L. B. Binding of Saliccylamides: QSAR Analysis with Electrotopological State indices. *Med. Res. Rev*. **1992**, *2*, 497−502.
(26) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 1074−1080.
(27) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(28) Ren, B. Application of Novel Atom-Type AI Topological Indices to QSPR Studies of Alkanes. *Comput. Chem.* **2002**, *26*, 357.
(29) Ren, B. Novel Atom-Type AI Topological Indices for QSPR Studies of Alkanols. *Comput. Chem.* **2002**, *26*, 223.
(30) Ren, B. Application of Novel Atom-Type AI Topological Indices in the Structure−Property Correlations. *J. Mol. Struct. (THEOCHEM)* **2002**, *586*, 137−148.
(31) Ren, B. A New Topological Index for QSPR of Alkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139−143.
(32) Ren, B.; Chen, G.; Xu, Y. A Novel Topological Index for QSPR/QSAR Study of Organic Compounds. *Acta Chimi. Sinica (in Chinese)* **1999**, 57, 563−571.
(33) Ren, B.; Xu, Y.; Chen, G. Estimation of Heat Capacity of Complex Organic Compounds by a Novel Topological Index. *J. Chem. Eng. China (in Chinese)* **1999**, *50*, 280−286.
(34) Ren, B.; Luo, B.; Zhang, Y. QSPR Studies of Solubilities and Octanol/Water Partition Coefficients of Organic Compounds. *J. S. China University Technol. (in Chinese)* **1999**, *27* (5), 89−95.
(35) Lucic, B.; Trinajstic, N. Multivarite Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.
(36) Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.
(37) Lucic, B.; Amic, D.; Trinajstic, N. Nonlinear Multivarite Regression Outperforms Several Concisely Designed Neural Networks in QSPR Modeling. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403−413.
(38) Firpo, M.; Gavernet, L.; Castro, E. A.; Toropov, A. A. Maximum Topological Distances Based Indices as Molecular Descriptors for QSPR. Part 1. Application to Alkyl Benzenes Boiling Points. *J. Mol. Struct. (THEOCHEM)* **2000**, *501/502*, 419−425.
(39) Castro, E. A.; Tueros, M.; Toropov, A. A. Maximum Topological Distances Based Indices as Molecular Descriptors for QSPR. 2−Application to Aromatic Hydrocarbons. *Comput. Chem.* **2000**, *24*, 571−576.
(40) Needham, D. E.; Wei, I.-C.; Seybold, P. G. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194.
(41) Weast, R. *CRC Handbook of Chemistry and Physics*, 70th ed.; CRC Press: Boca Raton, FL, 1989−1990.
(42) Lide, D. R.; Milne, G. W. A. *Handbook of Data on Common Organic Compounds*; CRC Press: Boca Raton, FL, 1992.
(43) Dean, J. A. *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: Beijing, 1999.
(44) Yaws, C. L. *Chemical Properties Handbook*; McGraw-Hill: Beijing, 1999.
(45) Pitzer, K. S.; Lippmann, D. Z.; Curl, R. F.; Huggins, C. M.; Petersen, D. E. The Volumetric and Thermodynamic Properties of Fluids. II. Compressibility Factor, Vapor Pressure and Entropy of Vaporization. *J. Am. Chem. Soc.* **1955**, *77*, 3433−3400.
(46) Yang, Y.-Q.; Xu, L.; Hu, C.-Y. Extended Adjacency Matrix Indices and Their Applications. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1140−1145.
(47) Yao, Y.-Y.; Xu, L.; Yang, Y.-Q.; Yuan, X. S. Study on Structure−Activity Relationships of Organic Compounds. 3. New Topological Indices and Their Applications. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 590−595.
(48) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520−525.
(49) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: New York, 1999.
(50) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 601−609.
(51) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated *logP* Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(52) Xu, L. *Chemometrical Method (in Chinese)*. Scientific Press of China: Beijing, 1996.

CI020362L