# Application of novel atom-type AI topological indices in the structure-property correlations

**ARTICLE** *in* JOURNAL OF MOLECULAR STRUCTURE THEOCHEM · JUNE 2002

Impact Factor: 1.37 · DOI: 10.1016/S0166-1280(02)00080-5

Biye Ren
South China University of Technology

ELSEVIER

Journal of Molecular Structure (Theochem) 586 (2002) 137–148

THEO CHEM

www.elsevier.com/locate/theochem

# Application of novel atom-type AI topological indices in the structure–property correlations

Biye Ren*

*Research Institute of Materials Science, South China University of Technology, Guangzhou 510640, People's Republic of China*

Received 29 November 2001; accepted 22 February 2002

**Abstract**

Novel atom-type AI topological indices are generated as new parameters in quantitative structure–property/activity relationships (QSPR/QSAR) models to encode the structural environment of each atom-type in a molecule. These AI indices, along with previously proposed Xu index, are extended to complex compounds with heteroatoms by using the novel vertex degree $v^m$, which is derived on the basis of the valence connectivity $\delta^v$ of Kier–Hall. The efficiency of the approach is demonstrated through three high quality QSPR models of the molar volumes (MV), molar refractions (MR), and molecular total surface areas (TSA) for three data sets of compounds consisting of alkanes and alcohols. The results indicate that combination of the atomic-based AI indices and Xu index can produce a significant improvement in the statistical quality of the models obtained for the three properties. The significant improvement indicates the high potential of these indices for application to various physical properties and structural types, especially complex compounds with special functional groups. For the final multiple linear regression models, the correlation coefficients $r$ are 0.9965, 0.9993, and 0.9990, and the standard errors $s$ are 2.603, 0.3223, and 3.393 for MV, MR, and TSA, respectively. In addition, the results indicate that three properties are dominated by molecular size but other atomic groups are also important although their contributions are much smaller than that of the molecular size. The cross-validation using the more general leave-*n*-out method demonstrates the final models to be highly statistically reliable.

*Keywords*: Molar volumes; Molar refractions; Molecular total surface areas; Multiple linear regression; Topological index; Quantitative structure–property/activity relationships

## 1. Introduction

The quantitative structure–property/activity relationships (QSPR/QSAR) studies of organic compounds have been a focus of great attention by the scientific community for a long time [1,2]. A great deal of QSPR/QSAR models have been developed using various model parameters to describe and predict the physical properties and biological activities of organic compounds from their molecular structures as well as for molecular design. Among these model parameters, the graph-theoretical topological indexes are particularly of interest because they can be derived directly from the molecular structures without any experimental effort.

As we know, there are two types of topological indices. A number of early-proposed topological indices conventionally characterize a molecule as a whole, i.e. molecular size or shape. These conventional indices, such as well-known molecular connectivity index ($\chi$) [3], Hosoya's index ($Z$) [4], Balaban's

* Tel./fax: +86-20-8711-2886.
*E-mail address:* renbiye@163.net (B. Ren).

0166-1280/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0166-1280(02)00080-5

index ($J$) [5], Bonchev's index ($I_D$) [6], Schluze's index (MTI) [7], and Wiener's index ($W$) [8], are most popular in developing estimation models for a variety of QSPR/QSAR studies. However, most of the existing topological indices largely limit their field of application for lack of information on multiple bonds and/or heteroatoms in molecular graphs. In order to differentiate the multiple bonds and heteroatoms in molecular graphs, several different empirical and unempirical approaches have been introduced in the past few years. These are Kier and Hall's concept of valence molecular connectivity [3,9], Bogdanov's topographic distance matrices [10], Randic's approach of graph embedded into three-dimensional [11], Estrada' approach of edge weights using quantum-chemical parameters [12], and recently developed 3D topological indices [13,14], etc. Further development is in progress but these conventional indices have some shortcomings. The major problem is that these conventional indices do not take into account the separate contributions of individual groups to properties, especially heteroatoms. As we know, for complex molecules, polarity, and especially the ability of the molecule to participate in hydrogen bonding caused by polar groups may be very important factors influencing physical properties, which depend directly on the strength of intermolecular forces. However, the conventional indices tend to obscure this fact and cannot directly reflect what structural features would be important.

The second type of topological indexes is the atom-type topological indexes, which further describe the structural information of a molecule at the atomic level. One of the most interesting indices of this type is the electrotopological state (E-state) index developed by Kier and Hall [15]. The E-state indices have been successfully used in a variety of QSPR/QSAR studies of complex molecules [15–20]. The development of atom-type indexes provides a new possibility for understanding the role of individual groups in molecules. The development of the atomic-level topological indices is not very advanced but progress can be anticipated. It is expected that the atom-type topological index could help make a break-through in QSPR/QSAR studies of complex compounds. This stimulates us to find new atomic level topological indices to describe different physical properties and biological activities. In the previous papers [21,22], we proposed a set of new atomic-based AI topological indices different from E-state indices. These atom-type AI indices and recently proposed Xu index were further extended to hetero-atom-containing compounds using the novel vertex degree ($v^m$) based on the valence connectivity $\delta^v$ of Kier–Hall [3].

The main goal of the present paper is to further verify the high potential of these indices for application to various properties and different structural types. We select three physical properties (molar volumes (MV), molar refractions (MR), and molecular total surface areas (TSA) of three mixed sets of representative organic compounds including non-polar alkanes and polar alcohols for this study. We also wish that this study could help in understanding what structural features or groups are likely to be important to the three physical properties.

## 2. Method

In a previous paper [23], Xu index was generated based on the vertex-adjacency matrix and the distance matrix of a graph. To facilitate the calculation of Xu index, herein the method to derive this index is briefly introduced. For a simple molecular graph $G = \{V, E\}$ with $n$ vertices, 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 $\mathbf{A}$ are 1 if vertices $i$ and $j$ are adjacent and 0 otherwise, where $n$ is the number of vertices. The distance matrix, $\mathbf{D} = [d_{ij}]_{n\times n}$, is also a square symmetric matrix. The entries $d_{ij}$ of matrix $\mathbf{D}$ are the length of the shortest path between the vertices $i$ and $j$ in a $G$, i.e. the number of C–C bond for alkanes. The over row or column $i$ of matrix $\mathbf{A}$ yields local vertex-degree $v_i$; the analogous sum for matrix $\mathbf{D}$ yields distance sums $s_i$. For alkanes the local vertex-degree $v_i$ is the sum of edges adjacent to vertex $i$. For a simple molecular graph Xu index can be expressed as below [23–26]:

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

where the sum is over all $i$ vertices in a graph. $n$ is the number of vertices.

For any atom $i$ belonging to $j$th atom-type in a molecular graph, the topological index, $\mathrm{AI}_i(j)$, is defined as follows [21,22]:

$$\mathrm{AI}_i(j) = 1 + \phi_i(j) \quad (2)$$

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

where parameter $\phi_i$ is considered as a perturbing term of $i$th atom reflecting the effect of its structural environment on its $\mathrm{AI}_i(j)$ value. The sum is over all $i$ vertices in a graph.

According to this definition, for any atom-type $j$ the atom-type AI index, $\mathrm{AI}(j)$, is defined as a sum of $\mathrm{AI}_i(j)$ values of all atoms or groups of the same atom-type in a molecular graph:

$$\mathrm{AI}(j) = \sum_{i=1}^{m} \mathrm{AI}_i(j) = m + \sum_{i=1}^{m} \phi_i(j)$$

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

where $m$ is the count of the atoms or groups of the same atom-type $j$. The AI value is clearly related to the count of the atoms or groups and its structural environment.

The differentiation of heteroatoms and multiple bonds has been successfully resolved by adding a perturbing term to the number of connections (edges) of that atom as the new degree of vertex, $v^m$ [19,20]. The novel degree of vertex for heteroatom $i$, $v^m$, is defined as follows:

$$v^m = \delta + k \quad (5)$$

where $\delta$ is the number of connections (edges) of that atom, and the $k$ parameter is a perturbing term of the number of edges. The number of connections (edges) is calculated as

$$\delta = \sigma - h \quad (6)$$

where $\sigma$ is the number of electrons in orbital, and $h$ is the number of hydrogen atoms connected to the atom. The $k$ parameter is calculated using the expression

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

where $N$ is the principal quantum number of valence shell. $Z$ and $Z^v$ are the atomic number and the number of valence electrons for heteroatom, respectively. $\delta^v$ is the valence connectivity of Kier and Hall [9], for multiple bonds in molecular graphs the $\delta^v$ value is expressed as

$$\delta^v = Z^v - h \quad (8)$$

For heteroatoms in molecular graphs, $\delta^v$ is expressed as

$$\delta^v = (Z^v - h)/(Z - Z^v - 1) \quad (9)$$

Consequently, if only we use the new degree of vertex, $v^m$, instead of the original vertex degree $v_i$, and then both Xu and atom-type AI indices can easily be extended to systems with multiple bonds and/or heteroatoms using the same formula defined for a simple molecular graph.

As an illustration, Table 1 shows the valence connectivity $\delta^v$ of Kier and Hall [3], numbers of connections $\delta$, $k$ parameters, and proposed $v^m$ values for some representative heteroatoms. For example, for oxygen in alcohols and ethers the $\delta$ 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.

## 3. Data set and regression analysis

### 3.1. Data set

MV are calculated as MW/$d$, where MW is the molecular weight and $d$ is the density (g/cm$^3$) at 20 °C. MR at 20 °C are calculated according to the lorentz–lorenz expression:

$$\mathrm{MR} = \frac{n_0^2 - 1}{n_0^2 + 2}\mathrm{MV} \quad (10)$$

where $n_0$ is the index of refraction at 20 °C. The Experimental values of MV and MR for alkanes at 20 °C are taken from Ref. [27]. The experimental values of $n_0$ and $d$ used to calculate MV and MR of alcohols are taken from Refs. [28,29]. For alcohols, in the majority of the cases, the data between the two sources agree fairly well although there are some discrepancies between the two sources. The molecular TSA at 25 °C, defined as the cavity dimension of the solute when placed in the water media, are taken from Refs. [30,31].

### 3.2. Regression analysis

For individual properties the multiple linear regression using the modified Xu (represented as $X_u^m$) and all

Table 1
The valence connectivity ($\delta^v$) of Kier–Hall [9], number of connections ($\delta$), $k$ parameters, and novel vertex degree ($v^m$) for some heteroatoms and multiple bonds

| Groups | $\delta^v$ | $\delta$ | $k$ | $v^m$ | Groups | $\delta^v$ | $\delta$ | $k$ | $v^m$ |
|---|---|---|---|---|---|---|---|---|---|
| $-CH_3$ | 1 | 1 | 0 | 1 | $\equiv N$ | 5 | 1 | 0.167 | 1.167 |
| $-CH_2-$ | 2 | 2 | 0 | 2 | $-PH_2$ | 0.333 | 1 | 0.871 | 1.871 |
| $-CH<$ | 3 | 3 | 0 | 3 | $>PH$ | 0.444 | 2 | 0.835 | 2.835 |
| $>C<$ | 4 | 4 | 0 | 4 | $-P<$ | 0.556 | 3 | 0.802 | 3.802 |
| $=CH_2$ | 2 | 1 | 0.333 | 1.333 | $-OH$ | 5 | 1 | 0.167 | 1.167 |
| $=CH-$ | 3 | 2 | 0.250 | 2.250 | $-O-$ | 6 | 2 | 0.143 | 2.143 |
| $=C<$ | 4 | 3 | 0.200 | 3.200 | $=O$ | 6 | 1 | 0.143 | 1.143 |
| $=C=$ | 4 | 2 | 0.200 | 2.200 | $-SH$ | 0.556 | 1 | 0.802 | 1.802 |
| $\equiv CH$ | 3 | 1 | 0.250 | 1.250 | $-S-$ | 0.667 | 2 | 0.771 | 2.771 |
| $\equiv C-$ | 4 | 2 | 0.200 | 2.200 | $-F$ | 7 | 1 | 0.125 | 1.125 |
| $-NH_2$ | 3 | 1 | 0.250 | 1.250 | $-Cl$ | 0.778 | 1 | 0.743 | 1.743 |
| $>NH$ | 4 | 2 | 0.200 | 2.200 | $-Br$ | 0.259 | 1 | 0.939 | 1.939 |
| $-N<$ | 5 | 3 | 0.167 | 3.167 | $-I$ | 0.149 | 1 | 0.977 | 1.977 |

AI indices present in molecular structures is used to develop the final model to correlate the physical property with chemical structures. The final model is obtained in the form of Eq. (11).

$$\text{property} = a_0 + a_1 X_u^m + b_1 \mathrm{AI}(1) + \cdots + b_j \mathrm{AI}(j) \quad (11)$$

where $a_0$ is a constant, and $a_1$ is the contribution coefficient of the modified Xu index, and $b_j$ is the contribution coefficient of $j$th AI index of atom-type $j$. Each coefficient describes the sensitivity of a property to each of the individual indices, so the coefficients of these parameters would measure the relative important of each index. As indices are added and removed, the changes in the statistics from model and model can be monitored. Therefore, the significance of each index is evaluated by monitoring the statistics ($t$ and $F$ values) so as to choose a high quality subset of indices [32–34]. The standard error ($s$) is used to evaluate the quality of model. In order to avoid colineality, intercorrelations between indexes are also examined.

## 4. Results and discussion

In general, there are two different directions in which multiple regression analysis is usually used in QSPR/QSAR studies [35,36]. 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 shortcomings, and they serve to somewhat different purposes. The main advantage of selecting a smaller set of structurally related compounds for QSPR/QSAR studies is the possibility of calculating the contributions of molecular size and all possible atom-types or groups to the properties studied. In present work, we select the second alternative so as to demonstrate what structural features or groups are likely to be important to the three properties studied.

### 4.1. Correlations to MV

As a starting point for illustrating the applicability of these indices, first we will consider a mixed set of 110 compounds containing 68 alkanes and 42 alcohols. The experimental values of MV for 110 compounds are listed in Table 2. A model for 110 compounds is then generated. As an illustration, we below give the final four-parameter model (Eq. (12)).

$$\begin{aligned} \mathrm{MV} = {} & 48.2504(\pm 1.2195) + 21.2665(\pm 1.3221) X_u^m \\ & - 3.5353(\pm 0.2938)\mathrm{AI}(\text{–OH}) \\ & + 2.8670(\pm 0.2893)\mathrm{AI}(\text{–CH}_3) \\ & + 0.7535(\pm 0.1004)\mathrm{AI}(>\mathrm{CH}_2) \qquad (12) \end{aligned}$$

$r = 0.9965$; $s = 2.603$; $F = 3800$; $P < 0.0001$; and $N = 110$

The $t$-values are 39.57, 16.09, $-12.03$, 9.911, and

Table 2
The calculated and experimental values of MV and MR for alkanes and alcohols

| Number | Compounds | MV (cm$^3$/mol) | | | MR (cm$^3$/mol) | | |
|---|---|---|---|---|---|---|---|
| | | Exp. | Calcd | Res. | Exp. | Calcd | Res. |
| 1 | 5 | 115.205 | 111.590 | 3.615 | 25.2656 | 24.9246 | 0.3410 |
| 2 | 2M4 | 116.426 | 109.899 | 6.527 | 25.2923 | 24.8402 | 0.4521 |
| 3 | 6 | 130.688 | 129.391 | 1.297 | 29.9066 | 29.8260 | 0.0806 |
| 4 | 2M5 | 131.933 | 127.681 | 4.252 | 29.9459 | 29.6956 | 0.2503 |
| 5 | 3M5 | 129.717 | 126.805 | 2.912 | 29.8016 | 29.5584 | 0.2432 |
| 6 | 23MM4 | 130.240 | 125.679 | 4.561 | 29.8104 | 29.5356 | 0.2748 |
| 7 | 22MM4 | 132.744 | 126.195 | 6.549 | 29.9347 | 29.6176 | 0.3171 |
| 8 | 7 | 146.540 | 147.272 | −0.732 | 34.5504 | 34.7249 | −0.1745 |
| 9 | 2M6 | 147.656 | 145.623 | 2.033 | 34.5908 | 34.5531 | 0.0377 |
| 10 | 3M6 | 145.821 | 144.415 | 1.406 | 34.4597 | 34.3785 | 0.0812 |
| 11 | 3E5 | 143.517 | 141.816 | 1.701 | 34.2827 | 33.6724 | 0.6103 |
| 12 | 22MM5 | 148.695 | 144.198 | 4.497 | 34.6166 | 34.4389 | 0.1777 |
| 13 | 23MM5 | 144.153 | 142.472 | 1.681 | 34.3237 | 34.1897 | 0.1340 |
| 14 | 24MM5 | 148.949 | 143.725 | 5.224 | 34.6192 | 34.3548 | 0.2645 |
| 15 | 33MM5 | 144.530 | 142.579 | 1.951 | 34.3323 | 34.1898 | 0.1425 |
| 16 | 223MMM4 | 145.191 | 141.847 | 3.344 | 34.3736 | 34.2095 | 0.1641 |
| 17 | 8 | 162.592 | 165.295 | −2.703 | 39.1922 | 39.6376 | −0.4454 |
| 18 | 2M7 | 163.663 | 163.716 | −0.053 | 39.2316 | 39.4188 | −0.1872 |
| 19 | 3M7 | 161.832 | 162.354 | −0.522 | 39.1001 | 39.2356 | −0.1355 |
| 20 | 4M7 | 162.105 | 161.953 | 0.152 | 39.1174 | 39.1864 | −0.0690 |
| 21 | 22MM6 | 164.285 | 162.579 | 1.706 | 39.2525 | 39.2854 | −0.0330 |
| 22 | 23MM6 | 160.395 | 160.164 | 0.231 | 38.9808 | 38.9835 | −0.0027 |
| 23 | 24MM6 | 163.093 | 160.506 | 2.587 | 39.1300 | 38.9977 | 0.1323 |
| 24 | 25MM6 | 164.697 | 161.929 | 2.768 | 39.2596 | 39.1758 | 0.0838 |
| 25 | 33MM6 | 160.879 | 160.689 | 0.190 | 39.0087 | 39.0778 | −0.0691 |
| 26 | 34MM6 | 158.814 | 159.036 | −0.222 | 38.8453 | 38.8010 | 0.0443 |
| 27 | 3E6 | 160.072 | 160.291 | −0.219 | 38.9441 | 38.8771 | 0.0670 |
| 28 | 223MMM5 | 159.526 | 158.760 | 0.766 | 38.9249 | 38.8235 | 0.1014 |
| 29 | 224MMM5 | 165.083 | 160.407 | 4.676 | 39.2617 | 39.0151 | 0.2466 |
| 30 | 233MMM5 | 157.292 | 158.030 | −0.738 | 38.7617 | 38.7220 | 0.0397 |
| 31 | 234MMM5 | 158.852 | 158.107 | 0.745 | 38.8681 | 38.7424 | 0.1257 |
| 32 | 23ME5 | 158.794 | 158.405 | 0.389 | 38.8362 | 38.6692 | 0.1670 |
| 33 | 33ME5 | 158.852 | 158.764 | 0.088 | 38.7171 | 38.7447 | −0.0276 |
| 34 | 9 | 178.713 | 183.520 | −4.807 | 43.8423 | 44.5798 | −0.7375 |
| 35 | 2M8 | 179.773 | 181.986 | −2.213 | 43.8795 | 44.3026 | −0.4231 |
| 36 | 3M8 | 177.952 | 180.542 | −2.590 | 43.7296 | 44.1210 | −0.3914 |
| 37 | 4M8 | 178.150 | 179.921 | −1.771 | 43.7687 | 44.0518 | −0.2831 |
| 38 | 3E7 | 176.410 | 178.154 | −1.744 | 43.6420 | 43.7169 | −0.0749 |
| 39 | 4E7 | 175.685 | 177.380 | −1.695 | 43.4907 | 43.5897 | −0.0990 |
| 40 | 22MM7 | 180.507 | 181.211 | −0.704 | 43.9138 | 44.1429 | −0.2291 |
| 41 | 23MM7 | 176.653 | 178.277 | −1.624 | 43.6269 | 43.8267 | −0.1998 |
| 42 | 24MM7 | 179.120 | 178.187 | 0.933 | 43.7393 | 43.7839 | −0.0446 |
| 43 | 25MM7 | 179.371 | 178.785 | 0.586 | 43.8484 | 43.8276 | 0.0208 |
| 44 | 26MM7 | 180.914 | 180.292 | 0.622 | 43.9258 | 44.0066 | −0.0808 |
| 45 | 33MM7 | 176.897 | 178.504 | −1.607 | 43.6870 | 43.8063 | −0.1193 |
| 46 | 34MM7 | 175.349 | 175.709 | −0.360 | 43.5473 | 43.2503 | 0.2970 |
| 47 | 35MM7 | 177.386 | 177.240 | 0.146 | 43.6378 | 43.6367 | 0.0010 |
| 48 | 44MM7 | 176.897 | 177.640 | −0.743 | 43.6022 | 43.7024 | −0.1002 |
| 49 | 23ME6 | 175.445 | 175.653 | −0.208 | 43.6550 | 43.3793 | 0.2757 |
| 50 | 24ME6 | 177.386 | 176.320 | 1.066 | 43.6472 | 43.4416 | 0.2056 |
| 51 | 33ME6 | 173.077 | 175.593 | −2.516 | 43.2680 | 43.3310 | −0.0630 |

Table 2 (*continued*)

| Number | Compounds | MV (cm$^3$/mol) | | | MR (cm$^3$/mol) | | |
|---|---|---|---|---|---|---|---|
| | | Exp. | Calcd | Res. | Exp. | Calcd | Res. |
| 52 | 34ME6 | 172.844 | 174.700 | −1.856 | 43.3746 | 43.2227 | 0.1519 |
| 53 | 223MMM6 | 175.878 | 176.793 | −0.915 | 43.6226 | 43.6092 | 0.0134 |
| 54 | 224MMM6 | 179.220 | 177.492 | 1.728 | 43.7638 | 43.6353 | 0.1285 |
| 55 | 225MMM6 | 181.346 | 179.196 | 2.150 | 43.9356 | 43.8230 | 0.1126 |
| 56 | 233MMM6 | 173.780 | 175.665 | −1.885 | 43.4347 | 43.4701 | −0.0354 |
| 57 | 234MMM6 | 173.498 | 174.685 | −1.187 | 43.3917 | 43.3230 | 0.0687 |
| 58 | 235MMM6 | 177.656 | 176.325 | 1.331 | 43.6471 | 43.5271 | 0.1200 |
| 59 | 244MMM6 | 177.187 | 176.405 | 0.782 | 43.6598 | 43.5044 | 0.1554 |
| 60 | 334MMM6 | 172.055 | 174.631 | −2.576 | 43.3407 | 43.2871 | 0.0536 |
| 61 | 33EE5 | 170.185 | 173.454 | −3.269 | 43.1134 | 42.9352 | 0.1782 |
| 62 | 223MME5 | 174.537 | 174.664 | −0.127 | 43.4571 | 43.2412 | 0.2159 |
| 63 | 233MME5 | 170.093 | 173.487 | −3.394 | 42.9542 | 43.0830 | −0.1288 |
| 64 | 234MEM5 | 173.804 | 173.701 | 0.103 | 43.4037 | 43.1366 | 0.2671 |
| 65 | 2233MMMM5 | 169.495 | 174.231 | −4.736 | 43.2147 | 43.2719 | −0.0572 |
| 66 | 2234MMMM5 | 173.557 | 174.532 | −0.975 | 43.4359 | 43.3110 | 0.1249 |
| 67 | 2244MMMM5 | 178.256 | 177.387 | 0.869 | 43.8747 | 43.6002 | 0.2745 |
| 68 | 2334MMMM5 | 169.928 | 173.406 | −3.478 | 43.2016 | 43.1910 | 0.0106 |
| 69 | Ethanol | 58.368 | 63.234 | 4.866 | 12.9267 | 12.9411 | −0.0144 |
| 70 | 1-Propanol | 74.798 | 78.533 | 3.735 | 17.5651 | 17.5425 | 0.0226 |
| 71 | 2-Propanol | 76.561 | 78.904 | 2.343 | 17.6135 | 17.8095 | −0.1960 |
| 72 | 1-Butanol | 91.529 | 93.656 | 2.127 | 22.1447 | 22.0692 | 0.0755 |
| 73 | 2-Methyl-1-propanol | 92.338 | 92.740 | 0.402 | 22.1820 | 22.0995 | 0.0825 |
| 74 | 2-Butanol | 91.903 | 94.879 | −2.976 | 22.1437 | 22.4159 | −0.2722 |
| 75 | 2-Methyl-2-propanol | 94.216 | 94.958 | −0.742 | 22.0334 | 22.6257 | −0.5923 |
| 76 | 1-Pentanol | 108.160 | 108.775 | −0.615 | 26.7978 | 26.5695 | 0.2283 |
| 77 | 3-Methyl-1-butanol | 108.559 | 107.422 | 1.137 | 26.7697 | 26.4935 | 0.2762 |
| 78 | 2-Pentanol | 108.962 | 110.952 | −1.990 | 26.7237 | 27.0175 | −0.2938 |
| 79 | 2-Methyl-1-butanol | 108.027 | 107.899 | 0.128 | 26.7535 | 26.5547 | 0.1988 |
| 80 | 3-Pentanol | 107.265 | 111.305 | −4.040 | 26.5647 | 27.0613 | −0.4966 |
| 81 | 3-Methyl-2-butanol | 107.631 | 109.409 | −1.778 | 26.6383 | 26.9254 | −0.2871 |
| 82 | 2-Methyl-2-butanol | 108.962 | 111.168 | −2.206 | 26.7179 | 27.1921 | −0.4742 |
| 83 | 2,2-Dimethyl-1-propanol | 108.559 | 107.610 | 0.949 | | | |
| 84 | 1-Hexanol | 125.590 | 123.981 | 1.609 | 31.6361 | 31.0686 | 0.5675 |
| 85 | 2-Methyl-1-pentanol | 123.795 | 123.684 | 0.111 | 31.2623 | 31.1022 | 0.1601 |
| 86 | 2-Ethyl-1-butanol | 122.401 | 123.119 | −0.718 | 31.1300 | 30.9724 | 0.1576 |
| 87 | 4-Methyl-2-pentanol | 126.774 | 125.335 | 1.439 | 31.4975 | 31.4309 | 0.0666 |
| 88 | 2,3-Dimethyl-2-butanol | 124.065 | 125.949 | −1.884 | 31.2388 | 31.6552 | −0.4164 |
| 89 | 3,3-Dimethyl-1-butanol | 124.005 | 121.954 | 2.051 | 31.2237 | 31.5680 | −0.3443 |
| 90 | 3,3-Dimethyl-2-butanol | 124.838 | 124.508 | 0.330 | 31.2682 | 31.4405 | −0.1723 |
| 91 | 3-Hexanol | 124.716 | 127.912 | −3.196 | 31.2971 | 31.7325 | −0.4354 |
| 92 | 3-Methyl-3-pentanol | 123.391 | 127.203 | −3.812 | 31.1343 | 31.7116 | −0.5773 |
| 93 | 1-Heptanol | 141.345 | 139.332 | 2.013 | 36.0153 | 35.5818 | 0.4335 |
| 94 | 2-Heptanol | 142.176 | 142.981 | −0.805 | 36.0768 | 36.1447 | −0.0679 |
| 95 | 3-Heptanol | 141.535 | 144.501 | −2.966 | 35.9814 | 36.3893 | −0.4079 |
| 96 | 4-Heptanol | 142.002 | 145.014 | −3.012 | 35.9278 | 36.4759 | −0.5481 |
| 97 | 2,4-Dimethyl-3-pentanol | 140.101 | 141.640 | −1.539 | 35.7942 | 36.1038 | −0.3096 |
| 98 | 1-Octanol | 157.473 | 154.875 | 2.598 | 40.6792 | 40.1218 | 0.5574 |
| 99 | 2-Octanol | 158.720 | 159.062 | −0.342 | 40.6681 | 40.7020 | −0.0339 |
| 100 | 4-Octanol | 158.972 | 162.111 | −3.139 | 40.6490 | 41.2119 | −0.5629 |
| 101 | 2-Ethyl-1-hexanol | 156.357 | 155.699 | 0.658 | 40.5140 | 40.1805 | 0.3335 |
| 102 | 2,2,4-Trimethyl-1-pentanol | 155.221 | 153.964 | 1.257 | 40.0974 | 39.9761 | 0.1213 |
| 103 | 3,5-Dimethyl-1-hexanol | 156.960 | 151.798 | 5.162 | 40.1345 | 39.6663 | 0.4682 |

Table 2 (*continued*)

| Number | Compounds | MV ($cm^3$/mol) | | | MR ($cm^3$/mol) | | |
|---|---|---|---|---|---|---|---|
| | | Exp. | Calcd | Res. | Exp. | Calcd | Res. |
| 104 | 1-Nonanol | 174.417 | 170.643 | 3.774 | 45.2664 | 44.6971 | 0.5693 |
| 105 | 2,6-Dimethyl-4-heptanol | 177.638 | 176.600 | 1.038 | 45.2441 | 45.4493 | −0.2052 |
| 106 | 5-Nonanol | 172.642 | 179.711 | −7.069 | 44.5890 | 46.0313 | −1.4423 |
| 107 | 1-Decanol | 190.252 | 186.672 | 3.580 | 49.7344 | 49.3186 | 0.4158 |
| 108 | 1-Undecanol | 207.652 | 202.983 | 4.669 | 54.6404 | 53.9898 | 0.6506 |
| 109 | 2,6,8-Trimethyl-4-nonanol | 227.438 | 228.100 | −0.662 | 59.2889 | 59.4354 | −0.1465 |
| 110 | 1-Tridecanol | 236.965 | 236.527 | 0.438 | 63.3751 | 63.5041 | −0.1290 |

7.507, respectively. The number in parentheses is the standard deviation associated with the coefficient. The uncertainties of the regression coefficients in the model correspond to 95% confidence intervals. The indices in the final models 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 MV for these compounds with a fit error of only 1.71%. The results are quite good considering the data from different sources with different experimental accuracy.

On the other hand, if we divide the mixed data set of 110 compounds into alkane and alcohol two subsets, which are run individually, the quality of the model can be further improved. For alkane subset the best three-parameter model leads to $r = 0.9964$ and $s = 1.368$ and for alcohol subset we obtain the final four-parameter model with $r = 0.9997$ and $s = 0.9387$. The above models are also comparable to that reported by Needham et al. [27] for the same series of alkanes. The best single-parameter model using the zero-order connectivity $^0\chi$ index gave $r = 0.962$ and $s = 4.8$, but the molecular connectivity $^0\chi$ and $^1\chi$ indices provided a slightly superior model and produced $r = 0.982$ and $s = 3.3$. In fact, the MV is a property not correlated successfully with other topological indices. This may be further illustrated by a single-parameter model obtained by Estrada [37] using $\varepsilon$ index for the same series of alkanes ($r = 0.9931$ and $s = 2.032$).

The calculated values and residuals for 110 compounds used to generate the final model are shown in Table 2. A comparison of calculated and experimental data for MV is shown in Fig. 1. One can observe that the agreement between correlation and data is quite good although there is a compound (5-nonanol) with a slightly large residual. In general, the quality of the QSPR models can be conveniently measured by the correlation coefficient $r$ and the standard deviation $s$. Mihalic and Trinajstic [38] suggested that a good QSPR model must have $r > 0.995$, and $s$ depends on the property studied. According to this statement, the final model (Eq. (12)) represents an excellent QSPR model judging from the standard error and the plot in Fig. 1.

### 4.2. Correlations to MR

The MR is related to the bulk and polarizability of a molecule and is also a useful physical parameter in the field of chemical, biological and pharmaceutical sciences. The experimental values of MR for a mixed set of 109 compounds containing 68 alkanes and 41 alcohols are shown in Table 2.

The final model for 109 compounds is then developed using the modified Xu index and all five AI indices. The best six-parameter model is shown as follows (Eq. (13)):

$$\begin{aligned}\mathrm{MR} = {} & 5.7912(\pm 0.3032) + 5.4878(\pm 0.2034)X_u^m \\ & + 0.2306(\pm 0.07719)\mathrm{AI}(\text{–OH}) \\ & + 1.3066(\pm 0.08166)\mathrm{AI}(\text{–CH}_3) \\ & + 0.1631(\pm 0.01444)\mathrm{AI}(\text{>CH}_2) \\ & - 0.3124(\pm 0.04477)\mathrm{AI}(\text{>CH–}) \\ & - 0.5598(\pm 0.08560)\mathrm{AI}(\text{>C<}) \qquad (13)\end{aligned}$$

$r = 0.9993$; $s = 0.3223$; $F = 11\,749$; $P < 0.0001$; and $N = 109$

The $t$-values are 19.10, 26.98, 2.987, 16.00, 11.30, −6.978, and −6.540, respectively. The model accounts

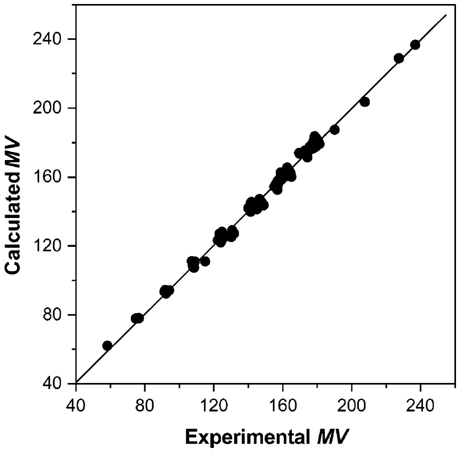

Fig. 1. A plot of calculated versus experimental MV for a mixed set of alkanes and alcohols.

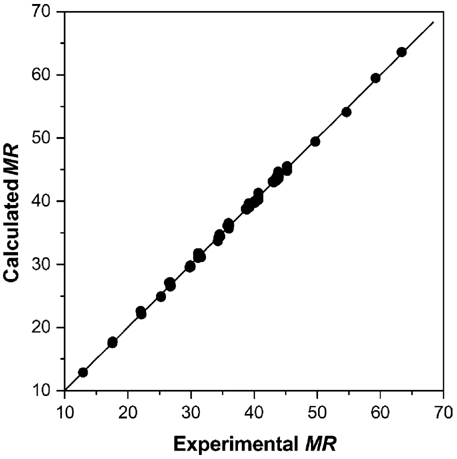

Fig. 2. A plot of calculated versus experimental MR for a mixed set of alkanes and alcohols.

for 99.9% of the variances in the experimental values of MR for 109 compounds with a fit error of only 0.86%. The calculated values and residuals for 109 compounds used to generate the final model are shown in Table 2. A comparison of calculated and experimental data is shown in Fig. 2 for 109 compounds. The agreement between correlation and experimental data is quite good. Therefore, the final model (Eq. (13)) represents an excellent QSPR model judging from the standard error and the plot in Fig. 2 according to above discussion. We observe that there is only 5-nonanol with a slight large residual out of 109 compounds. This compound is the same as one in correlation of MV. It is, therefore, possible that its MR is in error.

On the other hand, when the 109 compounds are divided into two subsets containing only alkanes or alcohols, which are run individually, the models obtained produce a significant improvement in the statistical quality, especially the reduction in the standard errors. For alkane subset we obtain the final four-parameter model with $r = 0.9996$ and $s = 0.1379$. This model is comparable to that obtained by Kier and Hall [3] by using four molecular connectivity indexes for the same series of 46 alkanes ($r = 0.9999$ and $s = 0.0738$) [3]. For alcohol subset the final four-parameter model with $r = 0.9998$ and $s = 0.1871$ can be obtained. This model is also comparable to that obtained by Kier and Hall [3] by using four molecular connectivity indexes for the same series of 30 alcohols ($r = 0.9998$ and $s = 0.153$).

### 4.3. Correlations to TSA

TSA is a practically valuable property in estimation of aqueous solubility of organic compounds. It is worth noting that although computer computations have been carried out to some extent, the molecular surface areas are still not easily available. The compounds that are used to generate a TSA model consist of 16 alkanes and 52 alcohols. The data of TSA in the literature [30,31] for 68 compounds are shown in Table 3.

The final model for 68 compounds is then developed using $X_u^m$ and five AI indices. The best six-parameter model is shown as follows (Eq. (14)):

$$\begin{aligned}\mathrm{TSA} = {} & 204.412(\pm 4.0015) + 47.8657(\pm 1.7286)X_u^m \\ & - 8.3149(\pm 1.0071)\mathrm{AI({-}OH)} \\ & - 4.4526(\pm 1.0998)\mathrm{AI({-}CH_3)} \\ & + 1.6547(\pm 0.09795)\mathrm{AI({>}CH_2)} \\ & + 3.7438(\pm 0.6933)\mathrm{AI({>}CH{-})} \\ & + 4.4809(\pm 1.3368)\mathrm{AI({>}C{<})}\end{aligned} \tag{14}$$

$$r = 0.9990;\ s = 3.393;\ F = 5279;\ P < 0.0001;\ \text{and}\ N = 68$$

The $t$-values are 51.08, 27.69, −8.256, −4.049, 16.89, 5.400, and 3.352, respectively. The model explains

Table 3
The molecular total surface areas (TSA) of alkanes and alcohols

| Number | Compounds | TSA ($\text{Å}^2$) | | | Number | Compounds | TSA ($\text{Å}^2$) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Lit. | Calcd | Res. | | | Lit. | Calcd | Res. |
| 1 | Ethane | 191.5 | 191.1 | 0.42 | 35 | 3,3-Dimethyl-1-butanol | 307.5 | 301.3 | 6.2 |
| 2 | Propane | 223.4 | 224.9 | −1.55 | 36 | 2,3-Dimethyl-2-butanol | 301.2 | 296.9 | 4.3 |
| 3 | Butane | 255.2 | 256.9 | −1.7 | 37 | 3,3-Dimethyl-2-butanol | 296.7 | 295.9 | 0.8 |
| 4 | 2-Methyl propane | 249.1 | 245.3 | 3.8 | 38 | 4-Methyl-1-pentanol | 323.0 | 324.0 | −1.0 |
| 5 | Pentane | 287.0 | 288.5 | −1.5 | 39 | 4-Methyl-2-pentanol | 314.9 | 314.5 | 0.4 |
| 6 | 2-Methyl butane | 274.6 | 275.7 | −1.1 | 40 | 1-Heptanol | 367.5 | 366.1 | 1.4 |
| 7 | 2,2-Dimethyl propane | 270.1 | 260.3 | 9.8 | 41 | 3-Heptanol | 357.1 | 355.8 | 1.3 |
| 8 | Hexane | 319.0 | 320.1 | −1.1 | 42 | 4-Heptanol | 357.1 | 355.1 | 2.0 |
| 9 | 2-Methyl pentane | 312.4 | 307.5 | 4.9 | 43 | 2-Methyl-2-hexanol | 346.1 | 341.9 | 4.2 |
| 10 | 3-Methyl pentane | 300.1 | 304.6 | −4.5 | 44 | 3-Methyl-3-hexanol | 337.7 | 336.8 | 0.9 |
| 11 | 2,2-Dimethyl butane | 290.8 | 289.4 | 1.4 | 45 | 2,4-Dimethyl-2-pentanol | 328.6 | 328.9 | −0.3 |
| 12 | Heptane | 351.0 | 352.0 | −1.0 | 46 | 2,4-Dimethyl-3-pentanol | 331.7 | 330.1 | 1.6 |
| 13 | 2,4-Dimethyl pentane | 324.7 | 326.7 | −2.0 | 47 | 3-Ethyl-3-pentanol | 324.4 | 331.9 | −7.5 |
| 14 | Octane | 383.0 | 384.1 | −1.1 | 48 | 2,3-Dimethyl-2-pentanol | 323.8 | 324.7 | −0.9 |
| 15 | 2,2,4-Trimethyl pentane | 338.9 | 340.5 | −1.6 | 49 | 2,2-Dimethyl-3-pentanol | 326.1 | 324.7 | 1.4 |
| 16 | 2,2,5-Trimethyl hexane | 373.0 | 374.3 | −1.3 | 50 | 2,3-Dimethyl-3-pentanol | 321.8 | 323.7 | −1.9 |
| 17 | 1-Butanol | 272.1 | 276.7 | −4.6 | 51 | 1-Octanol | 399.4 | 396.6 | 2.8 |
| 18 | 2-Methyl-1-propanol | 263.8 | 264.6 | −0.8 | 52 | 2-Octanol | 391.0 | 389.4 | 1.6 |
| 19 | 2-Butanol | 264.1 | 266.0 | −1.9 | 53 | 2-Ethyl-1-hexanol | 371.3 | 377.9 | −6.6 |
| 20 | 1-Pentanol | 303.9 | 306.3 | −2.4 | 54 | 2,2,3-Trimethyl-3-pentanol | 335.2 | 336.0 | −0.8 |
| 21 | 3-Methyl-1-butanol | 291.4 | 294.0 | −2.6 | 55 | 1-Nonanol | 431.2 | 427.5 | 3.7 |
| 22 | 2-Pentanol | 295.9 | 296.4 | −0.5 | 56 | 2-Nonanol | 423.2 | 421.0 | 2.2 |
| 23 | 2-Methyl-1-butanol | 289.4 | 292.1 | −2.7 | 57 | 3-Nonanol | 420.8 | 419.3 | 1.5 |
| 24 | 3-Pentanol | 293.5 | 294.4 | −0.9 | 58 | 4-Nonanol | 420.8 | 418.4 | 2.4 |
| 25 | 3-Methyl-2-butanol | 284.3 | 283.4 | 0.9 | 59 | 7-Methyl-1-octanol | 418.7 | 416.5 | 2.2 |
| 26 | 2-Methyl-2-butanol | 282.5 | 279.7 | 2.8 | 60 | 5-Nonanol | 420.8 | 418.1 | 2.7 |
| 27 | 1-Hexanol | 335.7 | 336.0 | −0.3 | 61 | 2,6-Dimethyl-4-heptanol | 394.0 | 394.9 | −0.9 |
| 28 | 2-Hexanol | 327.7 | 327.2 | 0.5 | 62 | 3,5-Dimethyl-4-heptanol | 379.3 | 379.3 | 0 |
| 29 | 3-Hexanol | 325.3 | 324.8 | 0.5 | 63 | 3,5,5-Trimethyl-1-hexanol | 376.6 | 381.1 | −4.5 |
| 30 | 2-Methyl-2-pentanol | 314.3 | 310.5 | 3.8 | 64 | 2,2-Diethyl-1-pentanol | 372.5 | 380.2 | −7.7 |
| 31 | 3-Methyl-2-pentanol | 311.3 | 311.1 | 0.2 | 65 | 1-Decanol | 463.0 | 459.1 | 3.9 |
| 32 | 2-Methyl-3-pentanol | 314.3 | 312.0 | 2.3 | 66 | 1-Dodecanol | 527.0 | 524.0 | 3.0 |
| 33 | 3-Methyl-3-pentanol | 305.8 | 306.7 | −0.9 | 67 | 1-Tetradecanol | 591.0 | 591.6 | −0.6 |
| 34 | 2-Ethyl-1-butanol | 308.6 | 318.1 | −9.5 | 68 | 1-Pentadecanol | 623.0 | 626.6 | −3.6 |

99.8% of the variances in the experimental values of TSA for 68 compounds with a fit error of only 1.0%. The calculated values and residuals for 61 compounds used to generate the final model are shown in Table 3. A plot of calculated versus data in the literature is shown in Fig. 3. The agreement of correlation and data is quite good. Therefore, the final model represents an excellent QSPR model judging from the standard error and the plot in Fig. 3 according to above discussion. Hence, it may be of interest to use the final model (Eq. (14)) as a simple approach to estimation of TSA from molecular structures although our main aim is not to search for the best predictive models. It should be mentioned that there are only 2 compounds with slightly large residuals (>9.0). However, it is difficult to propose a reason for why these compounds have slightly large residuals, other than to say that the data used may be erroneous, because the data values of TSA cannot directly obtained by experiment up to now.

On the other hand, when we divide the mixed data set of 68 compounds into two subsets, which are run individually, we can obtain slightly improved models. For 16 alkanes we obtain the final two-parameter

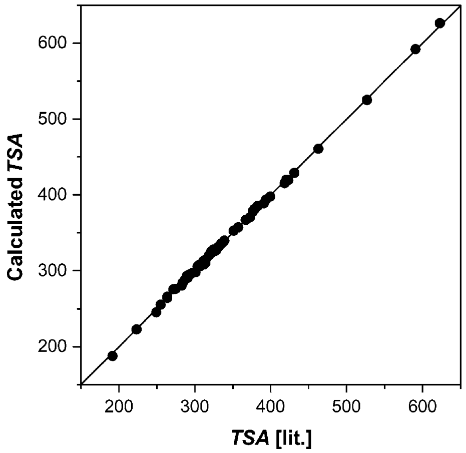

Fig. 3. A plot of calculated TSA versus data in literatures TSA [lit.] for a mixed set of alkanes and alcohols.

model with $r = 0.9985$ and $s = 3.088$, and for 52 alcohols the final six-parameter model with $r = 0.9991$ and $s = 3.348$ can be obtained.

Another aim of this study is to use these indices as an aid in deciphering what structural features and groups would be important to the three properties. As mentioned above, the present study provides a possibility for calculating the contributions of molecular size and all possible atomic groups to the studied properties from the models obtained. The relative contribution of each index can be simple estimated by multiplying the coefficients in final model by the mean index values; fraction contributions are obtained by multiplying the absolute values by the coefficients determination ($r^2$) and dividing by the sum [27]. The results are shown in Table 4.

One can see from Table 4 that the contributions of the modified Xu index to the three properties are within the range of 53–64%, while the other atom-type AI indices have smaller contributions and cover a wide range dependent on the investigated properties. The fact that the modified Xu index makes a major contribution to the three properties suggests that these properties are dominated by the dispersion because Xu index was physically interpreted as a parameter characterizing molecular size. This is particularly true because MV and molar refractivity are considered as molecular bulk or steric parameters. Our results are in accord with that reported by Needham et al. [27] in a comparative study of eight physical properties of alkanes based on ad hoc descriptors. On the other hand, our results reasonably explain why contributions of these AI indexes cannot be ignored in developing high quality QSPR models although their contributions are much smaller than that of the molecular size.

However, as it is well known, physical properties can be quite sensitive to some interactions and yet insensitive to others. We can see from Table 4 that the contributions of AI ($-CH_3$) index to MV and MR are 27.93 and 12.77%, respectively, indicating that the branching plays an important role in determining MV and MR because AI ($-CH_3$) index is clearly related to the counts of $-CH_3$ groups in a molecule, which is a crude measure of branching [27]. Obviously, branching prevents close contact with neighboring molecules in space. As a result, both MV and MR increase with the degree of branching for molecules with the same number of non-hydrogen atoms. We note that MV displays negative dependence on AI (–OH) index, suggesting that the hydrogen-bonding interaction plays an important role in determining MV, which is a property depending strongly on the intermolecular interactions. As we know, for heteroatom-containing compounds, polarity and/or hydrogen-bonding interactions lead to a decrease in MV as a result of the closer packing in the pure liquids. The hydrogen-bonding interaction forming among –OH groups will obviously result in a reduction in MV. On the other hand, –OH group present in the side chain likes a pseudobranch structure, which will prevent close contact with neighboring molecules in space

Table 4
The contributions of individual $X_u^m$ and AI indices to three properties

| Properties | $X_u^m$ | AI(–OH) | AI($-CH_3$) | AI($>CH_2$) | AI(>CH–) | AI(>C<) |
|---|---|---|---|---|---|---|
| MV | 72.71(0.64) | −4.436(0.066) | 27.93(0.25) | 7.410(0.065) | | |
| MR | 18.82(0.53) | 0.286(0.008) | 12.77(0.36) | 1.615(0.046) | 1.131(0.032) | 0.634(0.018) |
| TSA | 157.71(0.64) | −21.56(0.087) | −33.54(0.14) | 19.97(0.081) | 9.392(0.038) | 4.576(0.019) |

Table 5
The cross-validation results for all three properties

| Properties | $r$ | $s$ | $s$ of residuals | $r_{\text{press}}$ | $s$ of jackknifed residuals |
|---|---|---|---|---|---|
| MV | 0.9965 | 2.603 | 2.555 | 0.9943 | 3.200 |
| MR | 0.9993 | 0.3223 | 0.3133 | 0.9992 | 0.3396 |
| TSA | 0.9990 | 3.393 | 3.244 | 0.9983 | 4.511 |

and thus cause an increase in MV. The fact that the net results cause MV to display smaller negative dependence on AI (–OH) index indicates that the hydrogen-bonding interaction of –OH groups moieties in alcohols play an important role in determining the MV. It is particularly worth noting that MR displays positive dependence on AI (–OH) index. According to the above discussion, it may be possible that the increase in MR caused by the pseudobranch of –OH group present in the side chain compensate the reduction in MR caused by the hydrogen bonding interaction. The net results cause MR to display smaller positive dependence on AI (–OH) index. The fact suggests that the MR is less sensitive to the hydrogen-bonding interaction than the MV.

In the case of TSA, we see from Table 3 that the molecular surface areas have smaller values for alcohols that are small or highly branched, and have larger values for alcohols that are large and linear. The fact implies that the branching causes the reduction in TSA, in other words, the branching seems to be a very important factor determining the molecular surface areas. This is in accordance with our results that TSA displays negative dependence on AI ($-CH_3$) and AI (–OH) indices and positive dependence on AI ($>CH_2$), AI (>CH–), and AI (>C<) indexes. According to the definition of the molecular surface areas, the peripheral or terminal atomic groups (e.g. $-CH_3$ and –OH groups) should make a greater contribution to TSA than the 'hidden' or 'inside' groups (e.g. $-CH_2-$, –CH<, and >C< groups). In the present study, the contributions to TSA for each AI index decrease in the order of AI ($-CH_3$) > AI (–OH) > AI ($>CH_2$) > AI (>CH–) > AI (>C<). Therefore, the contributions for each AI index simply reflect the role of individual groups in molecules. These results seem to indicate that $X_u^m$ and AI indices, especially AI (–OH) index, are very sensible for describing the three physical properties of compounds investigated. The fact further indicates that the novel vertex degree $v^m$ proposed to replace the original vertex degree of heteroatom in molecular graphs is adequate to extension of Xu and AI indices to complex molecules with heteroatoms.

### 4.4. Model validation

Finally, the final models generated individually for three properties are validated by the cross-validation using the more general leave-*n*-out method since jackknifing of these data sets would be an extremely tedious process. For MR and MV, a leave-5-out method is used, and for TSA a leave-3-out method is used. As a quantitative evaluation of the results of the cross-validation, the PRESS statistical parameters ($r_{\text{press}}$ and $s_{\text{press}}$) and the standard error of the jackknifed residuals are listed in Table 5. The statistical parameters ($r$ and $s$) and the standard error of the residuals of the final models are also shown in Table 5. It is expected that the standard error of the jackknifed residuals should be larger than that of the residuals of the final models for the three properties. One can see that for each property the PRESS statistics ($r_{\text{press}}$ and $s_{\text{press}}$) obtained from the remaining compounds are very close to the statistics ($r$ and $s$) of the final models and the standard error of the jackknifed residuals is only slightly larger than the standard error of the residuals of the final models. This cross-validation demonstrates the outstanding predictive power of the final models. 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 observable pattern (Figs. 1–3). Therefore, the three final models are highly statistically reliable and successfully validated.

## 5. Conclusion

The multiple linear regression using the modified Xu and AI indices is used to develop high quality QSPR models to describe the three physical properties of three mixed data sets of organic compounds including alkanes and alcohols with a wide range of non-hydrogen atoms. For all three physical properties, the correlation coefficients $r$ are larger than 0.995 and particularly the standard errors are significantly reduced (within the range of 79–86%) as compared with the linear models with a single $X_u^m$ index. The results indicate that the novel vertex degree $v^m$ is adequate to the modification of Xu and AI indices and also indicates the high potential of these indices for application to various physical properties and different structural types, especially complex compounds with hydrogen bonding interactions. The cross-validation demonstrates the final models to be highly statistically reliable. The role of the molecular size and individual groups in molecules are simply illustrated. The results indicate that the three physical properties are dominated by molecular size directly associated with intermolecular dispersion forces. Other atomic groups, especially –OH group containing information about the hydrogen-bonding interaction, are also important although their contributions are much smaller than that of the molecular size. The contributions of individual indexes in final models satisfactorily account for the role of molecular size and each atom-type in molecules. This study can help in understanding what structural features and groups of a compound are important to physical properties. It is possible to use the modified Xu and AI indices to develop other high quality QSPR/QSAR models in the future.

## References

[1] A.T. Balaban, J. Chem. Inf. Comput. Sci. 35 (1995) 339.

[2] N. Trinajstic, Chemical Graph Theory, 2nd ed., CRC Press, Boca Raton, 1992.

[3] L.B. Kier, L.H. Hall, Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.

[4] H. Hosoya, Bull. Chem. Soc. Jpn 44 (1971) 2332.

[5] A.T. Balaban, Chem. Phys. Lett. 89 (1982) 399.

[6] D. Bonchev, N. Trinajstic, J. Chem. Phys. 67 (1977) 4517.

[7] H.P. Schultz, J. Chem. Inf. Comput. Sci. 29 (1989) 227.

[8] H. Wiener, J. Am. Chem. Soc. 69 (1947) 17.

[9] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–Activity Studies, Research Studies Press, Letchworth, 1986.

[10] B. Bogdanov, S. Nikolic, N. Trinajstic, J. Math. Chem. 3 (1989) 299.

[11] M. Randic, B. Jerman-Blazic, N. Trinajstic, Comput. Chem. 14 (1990) 237.

[12] E. Estrada, J. Chem. Inf. Comput. Sci. 35 (1995) 701.

[13] A. Toropov, A. Toporova, T. Ismailov, D. Bonchev, J. Mol. Struct. (Theochem.) 424 (1998) 237.

[14] M.V. Diudea, D. Horvath, A. Graovac, J. Chem. Inf. Comput. Sci. 35 (1995) 129.

[15] L.B. Kier, L.H. Hall, J.W. Frazer, J. Math. Chem. 7 (1991) 229.

[16] L.H. Hall, B. Mohney, L.B. Kier, J. Chem. Inf. Comput. Sci. 31 (1991) 76.

[17] L.H. Hall, B. Mohney, L.B. Kier, Quant. Struct.–Act. Relat. 10 (1991) 43.

[18] L.H. Hall, L.B. Kier, Med. Res. Rev. 2 (1992) 497.

[19] L.H. Hall, L.B. Kier, B.B. Brown, J. Chem. Inf. Comput. Sci. 35 (1995) 1074.

[20] L.H. Hall, L.B. Kier, J. Chem. Inf. Comput. Sci. 35 (1995) 1039.

[21] B. Ren, Comput. Chem. (2002) in press.

[22] B. Ren, Comput. Chem. 26 (2002) 223.

[23] B. Ren, J. Chem. Inf. Comput. Sci. 39 (1999) 139.

[24] B. Ren, G. Chen, Y. Xu, Acta Chimica Sinica 57 (1999) 563 in Chinese.

[25] B. Ren, Y. Xu, G. Chen, J. Chem. Engng China 50 (1999) 280 in Chinese.

[26] B. Ren, B. Luo, Y. Zhang, J. S. China Univ. Technol. 27 (1999) 89 in Chinese.

[27] D.E. Needham, I.-C. Wei, P.G. Seybold, J. Am. Chem. Soc. 110 (1988) 4186.

[28] F. Huang, X. Liu, Alcohols, Encyclopedia of Chemical Industry, vol. 2, Chemical Industry Press, Beijing, 1991 in Chinese.

[29] C.L. Yaws, Chemical Properties Handbook, McGraw-Hill, Beijing, 1999.

[30] L.B. Kier, L.H. Hall, W.J. Murray, M. Randic, J. Pharm. Sci. 64 (1975) 1971.

[31] G.L. Amidon, H. Yalkowsky, S.J. Leung, J. Pharm. Sci. 63 (1974) 3225.

[32] B. Lucic, N. Trinajstic, J. Chem. Inf. Comput. Sci. 39 (1999) 121.

[33] B. Lucic, N. Trinajstic, S. Sild, M. Karelson, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 39 (1999) 610.

[34] B. Lucic, D. Amic, N. Trinajstic, J. Chem. Inf. Comput. Sci. 40 (2000) 403.

[35] M. Firpo, L. Gavernet, E.A. Castro, A.A. Toropov, J. Mol. Struct. (Theochem.) 501–502 (2000) 419.

[36] E.A. Castro, M. Tueros, A.A. Toropov, Comput. Chem. 24 (2000) 571.

[37] E. Estrada, J. Chem. Inf. Comput. Sci. 35 (1995) 31.

[38] Z. Mihalic, N. Trinajstic, J. Chem. Edu. 69 (1992) 701.