Journal of Computer-Aided Molecular Design **17:** 607–620, 2003.

New atom-type-based AI topological indices: Application to QSPR studies of aldehydes and ketones

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

Received 28 December 2001; accepted 29 August 2003

Key words: aldehydes, graph theory, ketones, physical properties, QSPR, topological indices

Summary

Multiple linear regression (MLR) analysis based on a combined use of the modified Xu index and the atom-type based AI indices is performed to construct quantitative structure–property models on several data sets of organic compounds including aliphatic aldehydes and/or ketones. For each of the physical properties (the normal boiling points, molar refractions, gas heat capacities at 25 °C, water solubility at 25 °C, and *n*-octanol/water partition coefficient at 25 °C), high quality QSPR models are obtained, particularly the decrease in the standard error is within the range of 23.6–75.9% relative to the linear models with the modified Xu index alone. For individual subsets containing only aldehydes or ketones, in the majority of cases the quality of the model can be further improved. The significant improvement verifies the efficiency of the present approach and also indicates the usefulness of these indices for application to a wide range of physical properties. The results indicate that the physical properties studied are dominated by molecular size but atom types have smaller influences, especially the oxygen atom seems to be most important due to intermolecular polar interactions. The final models are validated to be statistically reliable using the leave-one-out cross-validation and/or an external test set.

Introduction

A close relationship often exists between chemical structures of organic compounds and many of their physical and chemical properties and biological activities. It is very convenient for practical use and molecular design if the physicochemical properties or biological activities of a molecule can be calculated and predicted from its molecular structure. The quantitative structure–property/activity relationships (QSPR/QSAR) provide an approach to this problem [1]. A large number of QSPR/QSAR models have been developed using various known physicochemical parameters such as the multiple solvatochromic parameters developed by Kamlet, Taft, and coworkers [2], octanol/water partition coefficient (*logP*) [3], and other molecular descriptors such as geometric, electronic or electrostatic, polar, steric, and topological ones [1, 3–5]. The geometric descriptors include molecular volume [6], molecular surface area [7, 8], charged partial surface area [9, 10], and dipole moments [11]. The electronic parameters include Hammett sigma (σ) constant [12], HOMO/LUMO energies [13], and partial atomic charges [14, 15]. Polar descriptors are usually described by molar refractivity [16]. The steric descriptors can be used as a measure of molecular bulkiness like the Taft constant (E_s) [17]. The topological descriptors include fragment and atom counts, substrate counts, and various conventional graph-theoretical indices, such as Kier-Hall's molecular connectivity index (χ) [18, 19], Hosoya's index (Z) [20], Balaban's index (J) [21], Bonchev's index (I_D) [22], Schluze's index (*MTI*) [23], Wiener's index (W) [24], and the Xu index recently introduced in our laboratory [25, 26].

Recently, the focus of attention has been turned to the use of topological indices because they can be readily derived directly from the molecular structures without any experimental effort [27]. However, most

*To whom correspondence should be addressed. E-mail: renbiye@163.net

of these conventional topological indices are being faced with some challenges because they tend to be simple, sensitive to various aspects of fundamental intermolecular interactions, and particularly they do not reflect the contributions of individual atomic groups to properties in QSPR/QSAR modeling. As a result, all attempts to model physical properties or biological activities of complex compounds, especially pharmaceutical molecules and polar protic compounds containing -OH, -NH_2, and -COOH, in terms of a single or few topological indices are not very successful in the end. The key problem is that heteroatoms in molecules may participate in polar interactions between molecules, especially hydrogen bonding interactions, because polarity and especially the ability of the molecule to participate in hydrogen bonding are very important factors determining physical properties except for molecular size and shape, and these factors are related to various aspects of intermolecular interactions.

In order to solve this problem, Kier and Hall introduced a type of atom-type topological indices called electrotopological state (E-state) indices [28], which further describe the structural information of a molecule at the atomic level. The E-state indices have been successfully used in a variety of QSPR/QSAR studies of complex molecules [29, 30], which indicates the efficiency of this approach. In addition, this approach also provides a new possibility for understanding the role of individual atom types in a molecule from a theoretical viewpoint. However, the development of the atomic level topological indices is not very advanced. Recently, we [31] proposed a new type of atom-type-based AI topological indices different from E-state indices. These AI indices have been successfully used as new model parameters in QSPR/QSAR studies [31–37]. However, the usefulness of these AI indices in QSPR/QSAR modeling has yet to be further verified in different systems.

The main aim of the present study is to further extend the application of these indices to complex compounds. Multiple linear regression (MLR) based on the AI and recently proposed Xu indices is performed to construct the structure–property models of several data sets of carbonyl compounds containing aldehydes and ketones. The physical properties include the normal boiling points, molar refractions, gas heat capacities at 25 °C, water solubility at 25 °C, and *n*-octanol/water partition coefficient at 25 °C. The following are our results.

Method

Let G = {V, E} be a molecular graph with n vertices, V (G) and $E(G)$ represent 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, 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 **D** are the length of the shortest path between the vertice i and j in a G. The sum over row or column i of matrix **A** yields local vertex-degree v_i; analogously, the sum over row i or column j of matrix **D** yields distance sums s_i. The Xu index can be expressed as [25]:

$$Xu = n^{1/2}\log\Big(\sum_{i=1}^{n} v_i s_i^2 \Big/ \sum_{i=1}^{n} v_i s_i\Big), \tag{1}$$

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

For any atom i that belongs to the jth atom-type in a graph, the topological index $\mathrm{AI}_i(j)$ is expressed as [31]

$$AI_i(j) = 1 + \phi_i(j) \tag{2}$$

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

where the parameter ϕ is considered as a perturbing term reflecting the effects of the structural environment of the ith atom on its topological index value.

According to this definition, for the jth atom-type in a graph, the corresponding AI index, AI (j), is a sum of all $\mathrm{AI}_i(j)$ values of the same atom-type:

$$\begin{aligned}\mathrm{AI}(j) &= \sum_{i=1}^{l} AI_i(j) = l + \sum_{i=1}^{l} \phi_i(j) \\ &= l + \sum_{i=1}^{l} v_i(j)s_i^2(j) \Big/ \sum_{i=1}^{n} v_i s_i,\end{aligned} \tag{4}$$

where l is the count of the same atom type. Clearly, the AI (j) value is equal to the count of the jth atom-type plus total perturbation terms and closely related to its structural environment.

In order to differentiate the multiple bonds and heteroatoms in a graph, a novel degree of vertex, v^m, was derived from the valence connectivity δ^v of Kier-Hall [18] and defined as [34]:

$$v^m = \delta + k \tag{5}$$

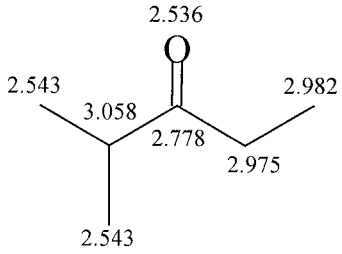

Figure 1. The molecular graph of 3-methyl-2-pentanone with the computed AI indices.

$$k = 1/[(2/N)^2\delta^v + 1], \tag{6}$$

where δ is the number of connections (edges) of that atom, and the parameter k is a perturbing term reflecting the effects of heteroatoms. N is the principal quantum number of valence shell. δ^v is the valence connectivity δ^v [18]. For heteroatoms in graphs, the δ^v value is expressed as [18, 19]:

$$\delta^v = (Z^v - h)/(Z - Z^v - 1). \tag{7}$$

The δ^v value for carbon with multiple bonds in a molecular graph is expressed as [18, 19]:

$$\delta^v = Z^v - h \tag{8}$$

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 heteroatom, respectively. It is obvious that v^m contains information on the number of connections and heteroatomic radius.

Consequently, for heteroatoms and multiple bonds in 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. 2 and 4). The v^m values of some heteroatoms and multiple bonds are shown in Table 1 [32, 34]. As an illustration, Figure 1 depicts a molecular graph of 3-methyl-2-pentanone with the computed AI indices.

Data set and regression analysis

Data set

The majority of the normal boiling points (*BP*) data are taken from refs. [38–42], but some values are taken from refs. [43, 44]. The precision of most data is within 1 °C, but for compounds with a boiling point range of 1–4 °C we quote their arithmetic mean *BP*. Molar refractions (*MR*) are calculated according to the lorentz-lorenz expression as described in ref. [36]. The experimental values of n_0 and d used to calculate *MR* are taken from refs. [43, 44]. The experimental values of gas heat capacity (C_p^G) are available in the literature [38]. The experimental values of water solubility (*logS*) and *n*-octanol/water partition coefficient (*logP*) are taken from ref. [1].

Regression analysis

For each data set MLR analysis using the modified Xu (represented as X_u^m) and all atom-type AI indices present in molecular structures is used to develop the structure–property models. The final model is obtained in the form of Eq. 9.

$$\begin{aligned} Property \;=\; & a_0 + a_1 X_u^m + b_1 AI(1) \\ & + \ldots + b_j AI(j), \end{aligned} \tag{9}$$

where a_0 is a constant, and a_1 is the contribution coefficient of X_u^m index, and b_j is the contribution coefficient of the jth atom type AI index, AI(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. The significance of each index is evaluated by monitoring the statistics (t and F values) to choose powerful indices. Generally, the number of the samples is six times descriptors at least to avoid overfitting of data. In order to avoid collinearity, intercorrelations between indexes are also examined. The variance inflation factor (VIF), defined as $(1-r^2)^{-1}$, is determined for each index as well. A VIF value larger than 10 is indicative of multicollinearity problems [45]. The standard error (s) is used to evaluate the quality of the model. Finally, the validity of the final models is tested using the leave-one-out cross-validation technique or an external test set (only for larger data set).

Model validation

In principle, cross-validation is a practical and reliable method for testing the significance of a model. Hence, to validate the final models generated individually for different physical properties, the leave-one-out method is used to do the cross-validation. The leave-one-out method consists of developing a number of models with one sample omitted at the time after developing each model. The omitted sample data is predicted and the differences between observed and predicted values are calculated. The predictive ability

Table 1. The valence connectivity (δ^v), k parameter, and proposed vertex degree (v^m) for some representative atom-types.

Groups	δ^v	k	v^m	Groups	δ^v	k	v^m
-CH_3	1	0	1	≡N	5	0.167	1.167
-CH_2-	2	0	2	-PH_2	0.333	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	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	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	0.125	1.125
-NH_2	3	0.250	1.250	-Cl	0.778	0.743	1.743
>NH	4	0.200	2.200	-Br	0.259	0.939	1.939
-N<	5	0.167	3.167	-I	0.149	0.977	1.977

of the model is quantified in terms of the corresponding leave-one-out cross-validated parameters, $r^2_{\rm cv}$ (or q^2) and $s_{\rm cv}$ values, which are defined as [46]:

$$r^2_{cv} = 1.0 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{10}$$

where y_i and $\hat{y}_i$ is the experimental and predictive value, respectively. $\bar{y}$ is the mean value of y_i.

$$s_{cv} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N - M - 1}}, \tag{11}$$

where N is the number of samples used for model building. M is the number of descriptors. For a reliable model, the $r^2_{\rm cv}$ or q^2 values should be > 0.6. The model is considered to be excellent if $r^2_{\rm cv}$ or $q^2 > 0.9$.

On the other hand, the actual prediction ability of the model is validated using an external prediction set not included in the training set. The performance of the model (its predictive ability) can be given by the standard error ($s_{\rm pred}$) of prediction, defined as [46]:

$$s_{pred} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}} \tag{12}$$

Correlations to physical properties

Correlations to boiling points (BP)

As a starting point, we consider a mixed data set of 28 aldehydes and 60 ketones with up to 16 non-hydrogen atoms to develop the structure-boiling point models. The observed *BP* values are listed in Table 2. A boiling point model is then generated using the X^m_u and all AI indices present in molecules. The best five-parameter model is obtained, as shown below (Eq. 10):

$$\begin{aligned} BP \;=\; & -31.6053(\pm 5.7197) + 43.2028 \\ & (\pm 0.6618)X^m_u + 30.3994 \\ & (\pm 4.2898)\mathrm{AI}(=\mathrm{O}) - 19.4295 \\ & (\pm 2.5606)\mathrm{AI}(>\mathrm{CH}=) - 13.2923 \\ & (\pm 1.8704)\mathrm{AI}(>\mathrm{C}=) - 2.4562 \\ & (\pm 0.2139)\mathrm{AI}(\mathrm{CH_3}) \end{aligned} \tag{13}$$

$$\begin{aligned} r^2 \;&=\; 0.9972;\ s = 3.28;\ F = 5841;\ P < 0.0001; \\ r^2_{cv} \;&=\; 0.9964;\ s_{cv} = 3.66;\ \text{and}\ N = 88. \end{aligned}$$

The t-values are −5.526, 65.28, 7.086, −7.107, −7.588, and −11.48, respectively. All indices in the model are statistically significant according to the t-values at the level of $p < 0.0001$. This model produces a standard error of only 3.28 °C and explains more than 99% of the variances in the experimental *BP* values for these compounds. In particular, the final model produces an improvement of 39.8% relative to the linear model with the X^m_u index ($s = 5.45$ °C).

Clearly, the significant improvement indicates the efficiency of a combined use of Xu and AI indices. The pairwise correlations between every pair of indices are performed also. Cross-correlation analysis shows that the indices in the model are not highly correlated with each other (the pairwise correlation coefficients $|r| \leq 0.75$). The largest VIF is 2.26 for the model, indicating the high internal stability of the model.

On the other hand, the model is further validated using the leave-one-out cross-validation and an external test set, respectively. The r^2_{cv} and s_{cv} are determined to be 0.9964 and 3.66 (°C), which are very close to the statistics of Eq. 13. The cross-validation demonstrates the model to be statistically significant. As a second strategy, the model (Eq. 13) is used to predict the *BP* values of 15 compounds not involved in regression analysis. The predictive standard error s_{pred} is 2.48 °C for the test set, indicating a good predictive power of the model.

In order to get some meaningful insights into the role of structural features influencing physical properties, the relative (ψ_{r}) or fraction contributions (ψ_{f}) of the topological indices to *BP* are analyzed to manifest the role of the corresponding structural features in molecules [47]:

$$\psi_r(i) = a_i \bar{T} I_i \tag{14}$$

$$\psi_f(i) = r^2 \, |\psi_r(i)| \, / \sum_i |\psi_r(i)| \times 100\% \tag{15}$$

where a_i and $\bar{T} I_i$ are the coefficient and the average value of the ith topological index in the model. $\psi_{\text{r}}(i)$ and $\psi_{\text{f}}(i)$ are the relative and fraction contributions of the ith topological indices in the model. The sum is over all indices involved in the model.

The results indicate that the ψ_{f} values decrease in the order of X^m_{u} (47.6%), AI(=O) (28.1%), AI(>C=) (10.4%), AI(-CH=) (8.2%), and AI(-CH_3) (5.4%), suggesting that the boiling points are dominated by molecular size (because the Xu index characterizes molecular size [25, 26]) but other atom types have smaller influences. This is similar to our previous results obtained for the physical properties of alkanes and alcohols [31, 32]. It is worth noting that *BP* display positive dependence on the AI(=O) index of O heteroatom, which is consistent with the polar character of aldehydes and ketones. This may be readily understood because the oxygen atom would participate in the polar interactions between molecules.

The calculated *BP* values and residuals for 88 compounds are shown in Table 2, and the predicted *BP* values for the 15 compounds are also given. According to Mihalić and Trinajstić's [48] comments on the quality of the model, Eq. 13 represents an excellent QSPR model judging from the statistics and is obviously comparable to that obtained by Li et al. [49] based on 12 distance-edge vectors (μ) for the same series of 72 compounds ($s = 3.93$ K).

On the other hand, when individual correlations are run for the two subsets containing respectively 28 aldehydes and 60 ketones, excellent results are obtained also:

The aldehyde subset:

$$\begin{aligned} BP \;=\; & -10.0214(\pm 1.8830) + 46.4733 \\ & (\pm 1.2480) X^m_{\text{u}} - 1.9120 \\ & (\pm 0.3839)\text{AI}(-CH_3) - 0.4317 \\ & (\pm 0.09412)\text{AI}(> CH_2) \end{aligned} \tag{16}$$

$$\begin{aligned} r^2 \;&=\; 0.9991;\; s = 2.41;\; F = 8792;\; P < 0.0001; \\ r^2_{cv} \;&=\; 0.9988;\; s_{cv} = 2.77;\; \text{and}\;\; N = 28. \end{aligned}$$

The t-values are −5.322, 37.24, −4.981, and −4.586, respectively. The largest VIF is 8.12, which is lower than the limiting value of VIF = 10. The standard error is only 2.41 °C. The improvement in the statistical quality is 26.5% relative to Eq. 13. As a comparison, calculated or predicted *BP* values and residuals for the aldehyde subset are shown in Table 2.

The ketone subset:

$$\begin{aligned} BP \;=\; & 9.1197(\pm 1.9695) + 46.2368 \\ & (\pm 1.1639) X^m_{\text{u}} - 3.2987 \\ & (\pm 0.2482)\text{AI}(-CH_3) - 0.3807 \\ & (\pm 0.07564)\text{AI}(> CH_2) \end{aligned} \tag{17}$$

$$\begin{aligned} r^2 \;&=\; 0.9980;\; s = 2.37;\; F = 9422;\; P < 0.0001; \\ r^2_{cv} \;&=\; 0.9976;\; s_{cv} = 2.63;\; \text{and}\;\; N = 60. \end{aligned}$$

The t-values are 4.630, 39.72, −13.29, and −5.034, respectively. The largest VIF is 6.26, indicating that the model is internally stable. The standard error is only 2.37 °C. The model produces an improvement of 27.7% relative to Eq. 10. Calculated or predicted *BP* values and residuals for the ketone subset are also shown in Table 2.

As an extension of the above study, we deal with other examples of applications of these topological indexes. *MR*, C^G_p, *logS* and *logP* of several data sets of

Table 2. Calculated and experimental boiling points for a mixed set of aldehydes and ketones.

No.	Compound	*BP* (°C)						
		Exp	Calcd[a]	Res	Calcd[b]	Res	Calcd[c]	Res
1	Acetaldehyde	20.8	18.4	2.4	19.4	1.4	16.7	4.1
2	Propionaldehyde	48.8	51.8	−3.0	48.7	0.1	52.2	−3.4
3	Butyl aldehyde	75.7	80.4	−4.7	75.8	−0.1	80.8	−5.1
4	2-Methyl propanal	64.4	72.2	−7.8	68.9	−4.5	72.7	−8.3
5	Pentaldehyde	103.0	106.3	−3.3	101.6	1.4	106.5	−3.5
6	2-Methyl butanal	92.5	96.5	−4.0	93.2	−1.2	96.6	−4.1
7	3-Methyl butanal	92.5	98.1	−5.6	95.7	−3.2	98.3	−5.8
8	2,2-Dimethylpropanal	77.5	80.6	−3.1	76.4	1.1	80.7	−3.2
9	Hexanal	128.0	130.2	−2.2	126.2	1.8	130.3	−2.3
10	2-Methylpentanal	117.0	119.8	−2.8	117.6	−0.6	119.9	−2.9
11	3-Methylpentanal	118.0	121.1	−3.1	118.4	2.6	121.2	−3.2
12	2-Ethylbutanal	117.0	117.5	−0.5	114.8	2.2	117.5	−0.5
13	Heptanal	152.8	152.6	0.2	149.6	3.2	152.6	0.2
14	3-Methylhexanal	143.0	142.9	0.1	141.4	1.6	142.9	0.1
15	2,2-Dimethylpentanal	126.5	132.0	−5.5	131.3	−4.8	132.3	−5.8
16	Octanal	171.0	173.8	−2.8	171.9	−0.9	174.0	−3.0
17	2-Ethylhexanal	160.0	159.8	0.2	159.7	3.3	159.8	0.2
18	2-Propylpentanal	160.0	158.6	1.4	158.3	1.7	158.6	1.4
19	Nonanal	191.0	194.1	−3.1	193.2	−2.2	194.3	−3.3
20	3,5,5-Trimethylhexanal	170.5	167.7	2.8	173.7	−0.7	167.3	3.2
21	Decanal	208.5	213.5	−5.0	213.6	−5.1	213.9	−5.4
22	Undecanal	233.0	232.2	0.8	233.2	−0.2	232.1	0.9
23	2-Methyldecanal	229.0	221.7	7.3	226.6	2.4	221.2	7.8
24	Dodecanal	254.0	250.3	3.7	251.9	2.1	249.9	4.1
25	2-Methylundecanal	246.0	239.7	6.3	245.7	0.3	239.2	6.8
26	Tridecanal	267.0	267.8	−0.8	269.9	−2.9	267.9	−0.9
27	Tetradecanal	287.0	284.8	2.2	287.1	−0.1	284.4	2.6
28	Pentadecanal	304.9	301.4	3.5	303.6	1.3	300.6	4.3
29	Acetone	56.2	49.2	7.0	54.3	1.9	48.5	7.7
30	2-Butanone	79.6	76.1	3.5	79.2	0.4	75.9	3.7
31	2-Pentanone	102	101.8	0.2	103.9	−1.9	101.8	0.2
32	3-Pentanone	101.7	99.1	2.6	101.6	0.1	99.0	2.7
33	3-Methyl-2-butanone	93.5	93.2	0.3	94.2	0.3	93.2	0.3
34	2-Hexanone	127.6	126.0	1.6	127.5	0.1	126.0	1.6
35	3-Hexanone	123.5	122.4	1.1	124.4	0.6	122.4	1.1
36	3-Methyl-2-pentanone	118.0	115.7	2.3	115.9	2.1	115.7	2.3
37	4-Methyl-2-pentanone	117.0	117.6	−0.6	118.2	−1.2	117.6	−0.6
38	2-Methyl-3-pentanone	115.5	113.8	1.7	115.2	−1.7	113.8	1.7
39	3,3-Dimethyl-2-butanone	106.0	107.8	−1.8	106.0	0	107.9	−1.9
40	2-Heptanone	151.4	148.9	2.5	150.0	1.4	148.8	2.6
41	3-Heptanone	147.0	145.2	1.8	146.7	0.3	145.1	1.9
42	4-Heptanone	144.0	143.8	0.2	145.6	−1.6	143.8	0.2
43	3-Methyl-2-hexanone	143.5	140.0	3.5	141.5	2.0	139.9	3.6
44	4-Methyl-2-hexanone	139.0	139.1	−0.1	138.9	3.1	139.1	−0.1
45	5-Methyl-2-hexanone	144.0	140.4	3.6	141.0	3.0	140.3	3.7
46	2-Methyl-3-hexanone	135.0	128.3	6.7	140.4	−5.4	126.7	8.3
47	4-Methyl-3-hexanone	134.5	134.8	−0.3	135.5	−1.0	134.8	−0.3

Table 2 (Continued).

No.	Compound	*BP* (°C)						
		Exp	Calcd[a]	Res	Calcd[b]	Res	Calcd[c]	Res
48	5-Methyl-3-hexanone	135.0	136.4	−1.4	137.6	−2.6	136.5	−1.5
49	2,2-Dimethyl-3-pentanone	125.6	126.9	−1.3	126.0	−0.4	126.9	−1.3
50	2,4-Dimethyl-3-pentanone	125.4	126.6	−1.2	127.6	−2.2	126.7	−1.3
51	4,4-Dimethyl-2-pentanone	126.4	131.5	−5.1	129.6	−3.6	131.9	−5.5
52	2-Octanone	172.5	170.6	1.9	171.4	1.1	170.6	1.9
53	3-Octanone	167.5	167.2	0.3	168.1	−0.6	167.2	0.3
54	4-Octanone	165.5	165.2	0.3	166.5	−3.5	165.2	0.3
55	2-Methyl-4-heptanone	154.0	156.4	−2.4	157.7	−3.7	156.4	−2.4
56	3-Methyl-4-heptanone	153.0	155.0	−2.0	155.8	−2.8	155.0	−2.0
57	3-Methyl-2-heptanone	164.0	159.8	4.2	159.7	4.3	159.7	4.3
58	6-Methyl-2-heptanone	167.0	161.9	5.1	162.6	4.4	161.7	5.3
59	2-Methyl-3-heptanone	158.0	156.8	1.2	158.1	−0.1	156.8	1.2
60	3,3-Dimethyl-2-hexanone	151.5	149.3	2.2	147.1	−0.1	149.2	2.3
61	2,2-Dimethyl-3-hexanone	146.0	147.2	−1.2	146.8	−0.8	147.3	−1.3
62	2,5-Dimethyl-3-hexanone	147.5	147.7	−0.2	149.0	−1.5	147.7	−0.2
63	4,4-Dimethyl-3-hexanone	148.0	146.2	1.8	144.7	3.3	146.1	1.9
64	2,2,4-Trimethyl-3-pentanone	135.1	138.3	−3.2	137.6	−2.5	138.6	−3.5
65	2-Nonanone	195.0	191.4	3.6	191.9	3.4	191.2	3.8
66	3-Nonanone	190.0	188.3	1.7	188.6	1.4	188.2	1.8
67	4-Nonanone	187.5	186.2	1.3	186.8	0.7	186.2	1.3
68	5-Nonanone	188.4	185.5	2.9	186.2	2.2	185.3	3.1
69	7-Methyl-3-octanone	182.5	179.0	3.5	179.7	2.8	178.9	3.6
70	3-Methyl-4-octanone	174.0	175.4	−1.4	176.0	−2.0	175.5	−1.5
71	7-Methyl-4-octanone	178.0	176.9	1.1	178.0	0	176.9	1.1
72	2,6-Dimethyl-4-heptanone	169.4	167.2	2.2	168.7	0.7	167.0	2.4
73	3,5-Dimethyl-4-heptanone	162.0	165.0	−3.0	165.3	−3.3	165.2	−3.2
74	2,2,4,4-Tetramethyl-3-pentanone	152.0	148.9	3.1	146.8	5.2	148.2	3.8
75	2-Decanone	210.0	211.2	−1.2	211.5	−1.5	211.2	−1.2
76	3-Decanone	211.0	208.6	2.4	208.3	−5.3	208.5	2.5
77	4-Decanone	206.5	206.6	−0.1	206.3	0.2	206.6	−0.1
78	2-Undecanone	231.5	230.2	1.3	230.2	1.3	230.0	1.5
79	3-Undecanone	227.0	228.2	−1.2	227.1	−0.1	228.3	−1.3
80	5-Undecanone	227.0	225.2	1.8	223.9	3.1	225.1	1.9
81	6-Undecanone	226.0	224.8	1.2	223.5	4.5	224.7	1.3
82	2-Dodecanone	246.5	248.5	−2.0	248.2	−1.7	248.8	−2.3
83	2-Tridecanone	263.0	266.2	−3.2	265.4	−2.4	266.9	−3.9
84	7-Tridecanone	261.0	262.2	−1.2	257.9	3.1	262.4	−1.4
85	2-Methyl-3-tridecanone	267.0	271.6	−4.6	268.2	−1.2	272.1	−5.1
86	7-Ethyl-2-methyl-4-undecanone	252.5	255.5	−3.0	252.0	0.5	256.1	−3.6
87	2-Pentadecanone	294.0	300.0	−6.0	297.8	−3.8	303.0	−9.0
88	8-Pentadecanone	291.0	298.0	−7.0	289.7	1.3	300.0	−9.0

Table 2 (Continued).

No.	Compound	Test set				
		Exp	Calcd[a]	Res	Calcd[b]	Res
T1	4-Methylpentanal	121.0	122.5	−1.5	120.4	0.6
T2	2-Methylhexanal	141.0	142.0	−1.0	140.8	0.2
T3	4-Methylhexanal	144.0	143.8	0.2	142.0	2.0
T4	5-Methylhexanal	143.5	144.7	−1.2	144.3	−0.8
T5	4-Methyl-2-heptanone	160.5	160.4	0.1	160.2	0.3
T6	5-Methyl-3-heptanone	161.0	157.0	4.0	157.5	3.5
T7	3,5-Dimethyl-2-hexanone	154.0	151.1	2.9	150.9	3.1
T8	4,5-Dimethyl-3-hexanone	152.0	147.6	4.4	147.7	4.3
T9	3-Ethyl-4-methyl-2-pentanone	154.0	148.0	6.0	146.9	7.1
T10	4-Methyl-2-octanone	184.0	181.0	3.0	180.7	3.3
T11	2,5-Dimethyl-3-heptanone	166.0	167.2	−1.2	168.0	−2.0
T12	3,3-Dimethyl-2-heptanone	173.5	170.1	3.4	168.1	5.4
T13	4,6-Dimethyl-3-heptanone	171.5	168.0	3.5	168.2	3.3
T14	2,3-Dimethyl-4-heptanone	167.5	166.6	0.9	167.2	0.3
T15	5-Decanone	204.0	205.5	−1.5	205.6	−1.6
T16	2,2,6,6-Tetramethyl-4-heptanone	185.0	186.6	−1.6	184.9	0.1

[a]From overall data set; [b]from subsets; [c]from the cross-validation.

compounds containing aldehydes and/or ketones are used in the following studies.

Correlations to molar refraction (MR)

Molar refraction (*MR*) is a particularly useful physical parameter in chemistry, biological chemistry, and pharmaceutical sciences because it is closely related to the bulkiness and polarizability of a molecule. The compounds used in this study contain 22 aldehydes and 24 ketones. The compounds and the corresponding *MR* values are listed in Table 3. The best model is shown as follows:

$$\begin{aligned} MR \;=\; & 3.8378(\pm 0.2257) + 5.6526 \\ & (\pm 0.1324)\mathrm{X}_{\mathrm{u}}^{m} + 1.2911 \\ & (\pm 0.08232)\mathrm{AI}(-\mathrm{CH_3}) + 0.1112 \\ & (\pm 0.008)\mathrm{AI}(>\mathrm{CH_2}) + 0.6778 \\ & (\pm 0.05386)\mathrm{AI}(-\mathrm{CH}=)0.3319 \\ & (\pm 0.048)\mathrm{AI}(>\mathrm{CH}-) - 0.5544 \\ & (\pm 0.09183)\mathrm{AI}(>\mathrm{C}<) \end{aligned} \tag{18}$$

$$r^2 \;=\; 0.9998;\; s = 0.1598;\; F = 39551;\; P < 0.0001;$$
$$r_{cv}^2 \;=\; 0.9996;\; s_{cv} = 0.1994;\; \text{and}\;\; N = 46.$$

The t-values are 17.00, 42.69, 12.58, 15.68, 13.89, −6.91, and −6.04, respectively, indicating that all variables in the model are statistically significant at the level of $p < 0.0001$. This model explains more than 99.9% of the variance in the experimental values of *MR* for 46 compounds; in particular, the decrease in the standard error is about 60% relative to the linear model with the $\mathrm{X}_{\mathrm{u}}^{m}$ index ($s = 0.8678$). The pairwise correlation coefficients ($|r| \leq 0.90$) show that the indices in the model are not highly correlated with each other. The largest VIF value is 5.21 for the model, indicating the high internal stability of the model. The model is validated to be statistically significant by the cross-validation. Moreover, the best model (Eq. 18) is used to predict the *MR* values of five carbonyl compounds as the test set. The s_{pred} is 0.2139 for the test set, indicating a good predictive power of the model. In addition, the ψ_{r} values decrease in the order of $\mathrm{X}_{\mathrm{u}}^{m}$ (62.3%), AI(-$\mathrm{CH_3}$) (26.8%), AI(>$\mathrm{CH_2}$) (4.5%), AI(-CH=) (4.2%), AI(>CH-) (1.7%), and AI(>C<) (0.4%), suggesting that molar refractions are dominated by the molecular size but other atom types, especially -$\mathrm{CH_3}$ groups have smaller influences. The calculated *MR* values and residuals are shown in Table 3. One observes that the agreement between correlation and data is quite good. Hence, this model represents an excellent model of *MR* judging from the statistics.

On the other hand, we perform two separate correlations on 22 aldehydes and 24 ketones, respectively. The results are shown below:

Table 3. Calculated and experimental molar refractions for 46 compounds.

No.	Compound	*MR* ($cm^3.mol^{-1}$)						
		Exp	Calcd[a]	Res	Calcd[b]	Res	Calcd[c]	Res
1	Acetaldehyde	11.5829	11.4714	0.1115	11.3677	0.2152	11.4174	0.1655
2	Propionaldehyde	16.1632	16.1152	0.0480	16.2100	−0.0468	16.1059	0.0573
3	Butyl aldehyde	20.8011	20.8496	−0.0485	21.0321	−0.2310	20.8565	−0.0554
4	2-Methyl propanal	20.8219	20.9118	−0.0899	20.8885	−0.0666	20.9200	−0.0981
5	Pentaldehyde	25.4983	25.5052	−0.0069	25.6948	−0.1965	25.5058	−0.0075
6	2-Methyl butanal	25.3943	25.4436	−0.0493	25.4634	−0.0691	25.4465	−0.0522
7	3-Methyl butanal	25.5327	25.4088	0.1239	25.5068	0.0259	25.3983	0.1344
8	Hexanal	30.0928	30.1479	−0.0551	30.2981	−0.2053	30.1550	−0.0622
9	2-Methylpentanal	29.8497	30.0052	−0.1555	29.9742	−0.1245	30.0164	−0.1667
10	2-Ethylbutanal	29.9981	29.8533	0.1448	29.8779	0.1202	29.8438	0.1543
11	2,3-Dimethylbutanal	30.0640	30.0555	0.0085	29.9632	0.1008	30.0546	0.0094
12	Heptanal	34.7004	34.7858	−0.0854	34.8651	−0.1647	34.7973	−0.0969
13	2,2-Dimethylpentanal	34.7537	34.6837	0.0700	34.6279	0.1258	34.6592	0.0945
14	Octanal	39.4396	39.4244	0.0152	39.4104	0.0292	39.4233	0.0163
15	2-Ethylhexanal	39.2395	39.0600	0.1795	39.2246	0.0149	39.0040	0.2355
16	2-Ethyl-3-methylpentanal	38.9423	38.7321	0.2102	38.5612	0.3811	38.6544	0.2879
17	Nonanal	44.2669	44.0717	0.1952	43.9488	0.3181	44.0434	0.2235
18	3,5,5-Trimethylhexanal	43.9887	43.8463	0.1424	43.8037	0.1850	43.6001	0.3886
19	Decanal	48.6737	48.7314	−0.0577	48.4882	0.1855	48.7433	−0.0696
20	2-Methyldecanal	53.0003	53.3295	−0.3292	53.6516	−0.6513	53.3959	−0.3956
21	Dodecanal	58.0913	58.1107	−0.0194	57.6060	0.4853	58.1240	−0.0327
22	2-Methylundecanal	57.9284	58.0073	−0.0789	58.3522	−0.4238	58.0355	−0.1071
23	Acetone	16.2963	16.0646	0.2317	16.2393	0.0570	16.0159	0.2804
24	2-Butanone	20.6039	20.6884	−0.0845	20.8044	−0.2005	20.6960	−0.0921
25	2-Pentanone	25.2926	25.3230	−0.0304	25.4096	−0.1170	25.3248	−0.0322
26	3-Pentanone	25.2487	25.2810	−0.0323	25.2885	−0.0398	25.2831	−0.0344
27	3-Methyl-2-butanone	25.2603	25.3436	−0.0833	25.2085	0.0518	25.3526	−0.0923
28	2-Hexanone	29.9308	29.9229	0.0079	29.9912	−0.0604	29.9223	0.0085
29	3-Hexanone	29.7251	29.9121	−0.1870	29.8696	−0.1445	29.9222	−0.1971
30	3-Methyl-2-pentanone	29.9453	29.8040	0.1413	29.5718	0.3735	29.7886	0.1567
31	4-Methyl-2-pentanone	29.9877	29.9402	0.0475	29.8770	0.1107	29.9374	0.0503
32	3,3-Dimethyl-2-butanone	29.6748	29.9865	−0.3117	29.5716	0.1032	30.1314	−0.4566
33	2-Heptanone	34.5463	34.4931	0.0532	34.5511	−0.0048	34.4892	0.0571
34	3-Heptanone	34.4230	34.5216	−0.0986	34.4588	−0.0358	34.5278	−0.1048
35	4-Heptanone	34.3083	34.5446	−0.2363	34.4389	−0.1306	34.5598	−0.2515
36	5-Methyl-2-hexanone	34.5773	34.5086	0.0687	34.5347	0.0426	34.5017	0.0756
37	2-Octanone	39.1959	39.0444	0.1515	39.1006	0.0953	39.0279	0.168
38	4-Octanone	39.0616	39.1650	−0.1034	39.0356	0.0260	39.1722	−0.1106
39	6-Methyl-3-heptanone	38.9478	39.2152	−0.2674	39.2223	−0.2745	39.2393	−0.2915
40	2-Nonanone	43.3542	43.5888	−0.2346	43.6529	−0.2987	43.6183	−0.2641
41	5-Nonanone	43.8710	43.8033	0.0677	43.6506	0.2204	43.7969	0.0741
42	2,6-Dimethyl-4-heptanone	43.8902	43.8246	0.0656	43.8518	0.0384	43.7953	0.0949
43	2-Decanone	48.5304	48.1341	0.3963	48.2172	0.3132	48.0771	0.4533
44	2-Undecanone	52.7129	52.6884	0.0245	52.8030	−0.0901	52.6848	0.0281
45	6-Undecanone	53.2109	53.0898	0.1211	52.9587	0.2522	53.0506	0.1603
46	2-Methyl-4-undecanone	57.7027	57.6747	0.0280	58.0037	−0.3010	57.6592	0.0435

Table 3 (Continued).

No.	Compound	Prediction set					
		Exp	Calcd[a]	Res	Calcd[b]	Res	Ref.
T1	2,2-Dimethypropanal	25.4202	24.5226	0.8976	24.0996	1.3206	43
T2	3-Methylhexanal	34.6622	34.7254	−0.0632	34.5532	0.1090	39
T3	2,4-Dimethyl-3-pentanone	34.1382	34.5659	−0.4277	34.2223	−0.0841	39
T4	2-Methyl-3-hexanone	34.5321	34.5960	−0.0639	34.5748	−0.0427	39
T5	6-Methyl-2-heptanone	39.4969	39.0535	0.4434	39.1826	0.3143	39

[a]From overall data set; [b]from subsets; [c]from the cross-validation.

Table 4. Calculated and experimental gas heat capacities for 18 compounds.

No.	Compound	C_p^G (J.mol^{-1}.K^{-1})						
		Exp	Calcd[a]	Res	Calcd[b]	Res	Calcd[c]	Res
1	Propionaldehyde	90.03	92.15	−2.12			92.72	−2.69
2	Pentaldehyde	144.07	147.20	−3.13			148.42	−4.35
3	2,2-Dimethylpropanal	132.42	125.75	6.67			124.22	8.20
4	Acetone	83.99	85.24	−1.25	84.99	−1.00	85.70	−1.71
5	2-Butanone	110.02	110.77	−0.75	110.56	−0.54	110.90	−0.88
6	2-Pentanone	136.23	136.65	−0.42	136.43	−0.20	136.68	−0.45
7	3-Pentanone	133.54	133.26	0.28	133.21	0.33	133.23	0.31
8	3-Methyl-2-butanone	131.09	130.95	0.14	130.75	0.34	130.94	0.15
9	2-Hexanone	161.50	162.19	−0.69	161.92	−0.42	162.24	−0.74
10	3-Hexanone	157.82	156.94	0.88	156.96	0.86	156.86	0.96
11	4-Methyl-2-pentanone	155.68	157.37	−1.69	157.09	−1.41	157.49	−1.81
12	3,3-Dimethyl-2-butanone	149.64	148.80	0.84	148.61	1.03	148.75	0.89
13	2-Heptanone	189.55	187.25	2.30	186.90	2.65	186.75	2.80
14	4-Heptanone	180.63	178.75	1.88	178.93	1.70	178.44	2.19
15	2-Methyl-3-hexanone	173.25	175.70	−2.45	175.97	−2.72	176.28	−3.03
16	2,4-Dimethyl-3-pentanone	171.98	169.56	2.42	169.71	2.27	169.20	2.78
17	2-Octanone	209.54	211.80	−2.26	211.35	−1.81	213.25	−3.71
18	5-Nonanone	221.36	222.01	−0.65	222.42	−1.06	222.38	−1.02

[a]From overall data set; [b]from ketone subset; [c]from the cross-validation.

For the aldehyde subset:

$$\begin{aligned} MR &= 5.2818(\pm 0.2655) + 6.6332 \\ &\quad (\pm 0.2180)X_u^m + 0.7355 \\ &\quad (\pm 0.05981)\mathrm{AI}(-CH_3) + 0.1271 \\ &\quad (\pm 0.01933)\mathrm{AI}(> CH_2) \end{aligned} \tag{19}$$

$$\begin{aligned} r^2 &= 0.9996;\ s = 0.2820;\ F = 14510;\ P < 0.0001; \\ r_{cv}^2 &= 0.9992;\ s_{cv} = 0.3925;\ \text{and}\ N = 22. \end{aligned}$$

The t-values are 19.90, 30.42, 12.30, and 6.577, respectively. The largest VIF is 5.88.

For the ketone subset:

$$\begin{aligned} MR &= 5.5891(\pm 0.2165) + 6.0373 \\ &\quad (\pm 0.1816)X_u^m + 0.7730 \\ &\quad (\pm 0.04466)\mathrm{AI}(-CH_3) + 0.1414 \\ &\quad (\pm 0.01441)\mathrm{AI}(> CH_2) \end{aligned} \tag{20}$$

$$\begin{aligned} r^2 &= 0.9997;\ s = 0.1941;\ F = 22816;\ P < 0.0001; \\ r_{cv}^2 &= 0.9994;\ s_{cv} = 0.2554;\ \text{and}\ N = 24. \end{aligned}$$

The t-values are 25.82, 33.26, 17.31, and 9.81, respectively. The largest VIF is 4.94.

The two models are only slightly inferior to that obtained for the whole data set (Eq. 18). As a comparison, calculated *MR* values and residuals from individual correlations are also shown in Table 3.

Table 5. Calculated and experimental water solubilities for 13 compounds.

No.	Compound	$-logS$ (mol.l^{-1})				
		Exp	Calcd[a]	Res	Calcd[b]	Res
1	2-Butanone	−0.68	−0.53	−0.15	−0.46	−0.22
2	2-Pentanone	0.17	0.14	0.03	0.14	0.03
3	3-Pentanone	0.23	0.11	0.12	0.09	0.14
4	3-Methyl-2-butanone	0.12	0.08	0.04	0.07	0.05
5	2-Hexanone	0.78	0.81	−0.03	0.81	−0.03
6	3-Hexanone	0.83	0.76	0.07	0.75	0.08
7	3-Methyl-2-pentanone	0.67	0.71	−0.04	0.71	−0.04
8	4-Methyl-2-pentanone	0.71	0.77	−0.06	0.77	−0.06
9	4-Methyl-4-pentanone	0.81	0.76	0.05	0.74	0.07
10	2-Heptanone	1.42	1.45	−0.03	1.46	−0.04
11	4-Heptanone	1.44	1.40	0.04	1.39	0.05
12	2,4-Dimethyl-3-pentanone	1.30	1.32	−0.02	1.34	−0.04
13	5-Octanone	2.57	2.60	−0.03	2.65	−0.08

[a]From overall data set; [b]from the cross-validation.

Correlations to gas heat capacity (C_p^G)

The gas heat capacity (C_p^G) is an important physical property of organic compounds for chemical engineering thermodynamics. Estimation methods applicable to C_p^G fall into four general categories: theoretical, group-contribution, corresponding-state, or Watson's thermodynamic cycle. However, reliable estimation procedures have not yet been developed for engineering use [50]. Here, we try to develop a QSPR model of C_p^G using these descriptors. Table 4 lists the experimental C_P^G data of 18 compounds, i.e., 3 aldehydes and 15 ketones. The best two-parameter model is presented below:

$$\begin{aligned} C_p^G &= 16.9047(\pm 4.6270) + 38.2859 \\ &\quad (\pm 0.8466)\mathrm{X}_\mathrm{u}^m + 9.5270 \\ &\quad (\pm 2.0669)\mathrm{AI}(=\mathrm{O}) \end{aligned} \tag{21}$$

$$\begin{aligned} r^2 &= 0.9960;\ s = 2.48;\ F = 1867;\ P < 0.0001; \\ r_{cv}^2 &= 0.9942;\ s_{cv} = 3.13;\ \text{and}\ \ N = 18 \end{aligned}$$

The t-values are 3.653, 45.22, and 4.609, respectively. The pairwise correlation r is 0.6147 and the VIF value is 1.61. The cross-validation indicates that the model is significant. This model explains more than 99% of the variances of C_p^G for 18 compounds with a fit error of only 1.63%. This is comparable to the traditional estimation methods [50]. As we expected, inclusion of AI (=O) index as the second parameter significantly improves the quality of the model. The improvement in the standard error is 33.5% relative to the linear model with the X_u^m index ($s = 3.73$). The significant improvement again indicates that the AI (=O) index contains information about the polar interactions between molecules, which may be important to explain some processes such as aqueous solubility, n-octanol/water partition, and other biological processes. In addition, by analyzing Eq. 21 we obtain the ψ_r values: 80.3% and 18.8% for X_u^m and AI(=O) indices, respectively. The results indicate that C_p^G are dominated by the dispersion force related to molecular size but the polar interaction between oxygen atoms in >CO or −CHO groups is important. The calculated values and residuals are also shown in Table 4.

In addition, the quality of the model can be significantly improved when only 15 ketones are used to generate a model using the same two indexes:

$$\begin{aligned} C_p^G &= 18.0856(\pm 3.4882) + 38.6729 \\ &\quad (\pm 0.7301)\mathrm{X}_\mathrm{u}^m + 8.6291 \\ &\quad (\pm 1.7316)\mathrm{AI}(=O) \end{aligned} \tag{22}$$

$$\begin{aligned} r^2 &= 0.9982;\ s = 1.64;\ F = 3378;\ P < 0.0001; \\ r_{cv}^2 &= 0.9966;\ s_{cv} = 2.26;\ \text{and}\ \ N = 15 \end{aligned}$$

The t-values are 5.185, 52.97, and 4.983, respectively. The standard error is only 1.64. The pairwise correlation r is 0.7242 and the VIF value is 2.10. As a comparison, the calculated values and residuals are also shown in Table 4.

Table 6. Calculated and experimental *logP* for 15 compounds.

No.	Compound	*logP*				
		Exp	Calcd[a]	Res	Calcd[b]	Res
1	Acetone	−0.21	−0.21	0	−0.22	0.01
2	2-Butanone	0.29	0.29	0	0.29	0
3	2-Pentanone	0.79	0.80	−0.01	0.80	−0.01
4	3-Pentanone	0.79	0.77	0.02	0.77	0.02
5	3-Methyl-2-butanone	0.59	0.61	−0.02	0.61	−0.02
6	2-Hexanone	1.29	1.31	−0.02	1.31	−0.02
7	3-Hexanone	1.29	1.27	0.02	1.27	0.02
8	3-Methyl-2-pentanone	1.09	1.08	0.01	1.08	0.01
9	4-Methyl-2-pentanone	1.09	1.11	−0.02	1.11	−0.02
10	4-Methyl-4-pentanone	1.09	1.08	0.01	1.08	0.01
11	2-Heptanone	1.79	1.81	−0.02	1.81	−0.02
12	4-Heptanone	1.79	1.78	0.01	1.78	0.01
13	2,4-Dimethyl-3-pentanone	1.39	1.37	0.02	1.35	0.04
14	5-Octanone	2.79	2.81	−0.02	2.82	−0.03
15	2-Nonanone	2.79	2.77	0.02	2.75	0.04

[a]From overall data set; [b]from the cross-validation.

Correlations to water solubility (logS)

Water solubility (*logS*) is a particularly important pharmaceutical property of organic compounds and has many uses in pharmaceutical chemistry, biological chemistry, and environmental science. It is also valuable in understanding drug transport and environmental impact [51]. The experimental *logS* data for 13 ketones are listed in Table 5. The best two-parameter model is shown as follows (Eq. 23):

$$\begin{aligned} -logS \;=\; & -2.7346(\pm 0.1056) + 1.0703 \\ & (\pm 0.03562)\mathrm{X}_{\mathrm{u}}^{m} + 0.04331 \\ & (\pm 0.01454)\mathrm{AI}(-\mathrm{CH_3}) \end{aligned} \tag{23}$$

$$\begin{aligned} r^2 \;&=\; 0.9928;\; s = 0.074;\; F = 687;\; P < 0.0001; \\ r_{cv}^2 \;&=\; 0.9864;\; s_{cv} = 0.100;\; \text{and}\;\; N = 13 \end{aligned}$$

The t-values are −25.90, 30.04, and 2.98, respectively. The pairwise correlation r and the VIF value are 0.5189 and 1.37, respectively. The cross-validation statistical parameters are very close to the statistics of Eq. 23, indicating that the model is significant. This model accounts for more than 99% of the variances in the experimental values of *logS* for the 13 ketones. The improvement in the standard error is 23.6% relative to the linear model using the $\mathrm{X}_{\mathrm{u}}^{m}$ index ($s = 0.09691$). The ψ_r values are 90.6% and 8.7% for $\mathrm{X}_{\mathrm{u}}^{m}$ and AI (-$\mathrm{CH_3}$), respectively, indicating that molecular size plays a major role in determining aqueous solubility but the $-\mathrm{CH_3}$ group in a molecule is also important. The calculated values and residuals are shown in Table 5.

Correlations to n-octanol/water partition (logP)

n-Octanol/water partition (*logP*) 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. In particular, *logP* is a crucial parameter in QSAR studies and drug design [52]. The experimental *logP* data for 15 ketones are listed in Table 6. The final two-variable model is shown below (24):

$$\begin{aligned} logP \;=\; & -1.1665(\pm 0.0233) + 0.7588 \\ & (\pm 0.0099)\mathrm{X}_{\mathrm{u}}^{m} + 0.01695 \\ & (\pm 1.17 \times 10^{-3})\mathrm{AI}(> \mathrm{CH_2}) \end{aligned} \tag{24}$$

$$\begin{aligned} r^2 \;&=\; 0.9996;\; s = 0.018;\; F = 14377;\; P < 0.0001; \\ r_{cv}^2 \;&=\; 0.9992;\; s_{cv} = 0.024;\; \text{and}\;\; N = 15 \end{aligned}$$

The t-values are −50.07, 76.67, and 14.54, respectively. The pairwise correlation r is 0.8499 and the VIF value is 3.60. This model is validated using the leave-one-out cross-validation to be statistically reliable. This model explains more than 99.9% of the variances in the experimental *logP* values for 15

ketones. The improvement in the standard quality is 75.9% relative to the linear model with the X_u^m index ($s = 0.07467$). Analogously, we obtain the ψ_r values of each index: 94.7% and 5.2% for X_u^m and AI(>CH_2), respectively. The calculated values and residuals are shown in Table 6.

Conclusions

MLR in terms of a combined use of the X_u^m and AI indices can provide high quality structure–property correlations for several data subsets of organic compounds containing aldehydes and/or ketones with a wide range of non-hydrogen atoms. For all data sets, the MLR models based on a combination of the X_u^m and AI indices can produce an improvement of 23.6–75.9% in the standard error relative to the linear model with the X_u^m index. For individual subsets, in the majority of cases the quality of the model can be further improved. The significant improvement verifies the efficiency of the present approach and also indicates the usefulness of these indices for application to various physical properties and different structural types, especially polar compounds. The results indicate that the physical properties are dominated by molecular size directly related to dispersion forces, but the role of atomic groups is important also. In general, the oxygen heteroatom (=O) seems to be most important due to intermolecular polar interactions. In general, the physical properties studied can be expressed as a linear combination of the individual indices related to molecular size and atom-type. The final models are validated to be statistically significant and reliable by the cross-validation using a leave-one-out method or an external test set.

References

1. Wang, L. and Han, S. (Eds.), Quantitative Structure-Activity Relationships of Organic Compounds (in Chinese), Environmental Scientific Press of China, Beijing, China, 1993.
2. Kamlet M., Abboud, J.L.M. and Taft R.W., Prog. Phys. Org. Chem., 13 (1981) 485, and references therein.
3. Hansch, C., Leo, A. and Hoekman, D., Exploring QSAR. Hydrophobic, Electronic and Steric Constants, American Chemical Society, Washington, DC, 1995.
4. Hansch, C. and Leo, A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC, 1995.
5. Xu, L. and Hu, C., Applied Chemical Graph Theory, Scientific Press, Beijing, China, 2000.
6. Bondi, A., J. Phys. Chem., 68 (1964) 441.
7. Hermann, R.B., J. Phys. Chem., 76 (1972) 2754.
8. Amidon, G.L., Yalkowsky, H. and Leung, S.J., J. Pharm. Sci., 63 (1974) 3225.
9. Stanton, D.T. and Jurs, P.C., Anal. Chem., 62 (1990) 2323.
10. Grigoras, S., J. Comput. Chem., 11 (1990) 493.
11. Kortvelyesi, T., Gorgenyi, M. and Heberger, K., Anal. Chim. Acta, 428 (2001) 73.
12. Reynolds, W.F., Prog. Phys. Org. Chem., 14 (1983) 165.
13. Franke, R., Theoretical Drug Design Methods, Elsevier, Amsterdam, The Netherlands, 1984.
14. Nelson, T.M. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 34 (1994) 601.
15. Dixon, S.L. and Jurs, P.C., J. Comput. Chem., 13 (1992) 492.
16. Grunenberg, J. and Herges, R., J. Chem. Inf. Comput. Sci., 35 (1995) 905.
17. Taft, R.W., J. Prog. Phys. Org. Chem., 14 (1983) 247.
18. Kier, L.B. and Hall, L.H., Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, 1976.
19. Kier, L.B. and Hall, L.H., Molecular Connectivity in Structure-Activity Studies, Research Studies Press, Letchworth, UK, 1986.
20. Hosoya, H., Bull. Chem. Soc. Jpn., 44 (1971) 2332.
21. Balaban, A.T., Chem. Phys. Lett., 89 (1982) 399.
22. Bonchev, D. and Trinajstić, N., J. Chem. Phys., 67 (1977) 4517.
23. Schultz, H.P., J. Chem. Inf. Comput. Sci., 29 (1989) 227.
24. Wiener, H., J. Am. Chem. Soc., 69 (1947) 17.
25. Ren, B., J. Chem. Inf. Comput. Sci., 39 (1999) 139.
26. Ren, B., Chen, G. and Xu, Y., Acta Chim. Sinica, 57 (1999) 563 (in Chinese).
27. Trinajstić, N., Chemical Graph Theory, 2nd ed., CRC Press, Boca Raton, FL, 1992.
28. Hall, L.H., Mohney, B. and Kier, L.B., J. Chem. Inf. Comput. Sci., 31 (1991) 76.
29. Maw, H.H. and Hall, L.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1248.
30. Rose, K., Hall, L.H. and Kier, L.B., J. Chem. Inf. Comput. Sci., 42 (2002) 651.
31. Ren, B., Comput. Chem., 26 (2002) 1121.
32. Ren, B., Comput. Chem., 26 (2002) 357.
33. Ren, B., J. Mol. Struct. (THEOCHEM), 586 (2002) 137.
34. Ren, B., J. Chem. Inf. Comput. Sci., 42 (2002) 858.
35. Ren, B., J. Chem. Inf. Comput. Sci., 43 (2003) 161.
36. Ren, B., J. Chem. Inf. Comput. Sci., 43 (2003) 1121.
37. Ren, B., Chemometrics Intel. Lab. Sys., 66 (2003) 29.
38. Weast, R., CRC Handbook of Chemistry and Physics, 70th ed., CRC Press, Boca Raton, FL, 1989–1990.
39. Lide, D.R. and Milne, G.W.A., Handbook of Data on Common Organic Compounds, CRC Press, Boca Raton, FL, 1992.
40. Dictionary of Organic Chemistry. 6th ed.; Chapman & Hall, London, 1996.
41. Dean, J.A., Lange's Handbook of Chemistry, 15th Ed., McGraw-Hill, Beijing, China, 1999.
42. Yaws, C.L., Chemical Properties Handbook, McGraw-Hill, Beijing, China, 1999.
43. Huang, F. and Liu, X., Aldehydes In Encyclopedia of Chemical Industry, Vol. 13. Chemical Industry Press, Beijing, China, 1997 (in Chinese).
44. Huang, F. and Liu, X., Ketones, In Encyclopedia of Chemical Industry, Vol. 16. Chemical Industry Press, Beijing, China, 1997 (in Chinese).
45. Wessel, M.D. and Jurs P.C., J. Chem. Inf. Comput. Sci., 35 (1995) 841.
46. Xu, L., Chemometrical Method (in Chinese), Scientific Press of China, Beijing, China, 1996.
47. Needham, D.E., Wei, I.-C. and Seybold, P.G., J. Am. Chem. Soc., 110 (1988) 4186.

48. Mihalić, Z. and Trinajstić, N. J. Chem. Educ., 69 (1992) 701.
49. Lin, Z., Xu, J., Liu, S., Zhen, X. and Li, Z., Acta Phys.-Chim. Sinica (in Chinese), 16 (2000) 153.
50. Reid, R.C., Prausnitz, J.M. and Poling, B.E., The Properties of Gases and Liquids, 4th ed., McGraw-Hill, New York, 1987.
51. Nelson, T.M. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 34(1994) 601.
52. Klopman, G., Li, J.-Y., Wang, S. and Dimayuga, M., J. Chem. Inf. Comput. Sci., 34(1994) 752.