JOURNAL OF PHYSICAL ORGANIC CHEMISTRY, VOL. 6, 574–582 (1993)

# USE OF THE MULTIFUNCTIONAL AUTOCORRELATION METHOD TO ESTIMATE MOLAR VOLUMES OF ALKANES AND OXYGENATED COMPOUNDS. COMPARISON BETWEEN COMPONENTS OF AUTOCORRELATION VECTORS AND TOPOLOGICAL INDICES

DRISS ZAKARYA
*Laboratoire de Chimie Organique, Faculté des Sciences, Université My Ismail, Meknès, Morocco*
FATHALLAH TIYAL
*Laboratoire de Chimie Organique, Faculté des Sciences, Université My Ismail, Meknès, Morocco, and Laboratoire de Chimie Organique Physique, URA 463, Université Claude Bernard, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France*
AND
MAURICE CHASTRETTE
*Laboratoire de Chimie Organique Physique, URA 463, Université Claude Bernard, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France*

**The Multifunctional Autocorrelation Method (MAM) derived from the autocorrelation method of Moreau *et al.* was used to calculate molar volumes for 300 linear or branched alkanes from $C_6$ to $C_{11}$ and 104 oxygenated compounds including 40 alcohols, 30 ethers and 34 ketones. Components $C_i$ of a MAM connectivity vector were compared with several topological descriptors. MAM descriptors were generally more efficient than topological descriptors but good regressions were obtained for alkanes only when more than two components $C_i$ were used. An excellent regression was obtained with four components ($C_0$ to $C_4$):**

$$V_m = 4\cdot45C_0 + 6\cdot44C_1 - 3\cdot33C_2 + 0\cdot46C_4 + 52\cdot9 \qquad (n = 300,\ r = 0\cdot999,\ s = 0\cdot81)$$

**Models obtained using descriptors based on connectivity were good enough for many practical purposes. However, to derive models adapted to estimate molar volumes for alkanes, an autocorrelation vector $V$ based on Van der Waals volumes (as given by Bondi) was considered. The best equation obtained again includes three components:**

$$V_m = 1\cdot750V_0 - 0\cdot123V_3 - 0\cdot037V_4 + 17\cdot97 \qquad (n = 300,\ r = 0\cdot999,\ s = 0\cdot85)$$

**To extend the use of MAM to compounds containing heteroatoms, a set of alcohols, ketones and ethers, where each functional group represents a different interaction type, was considered. Chemical families were first studied individually, then the total set was considered. The components of an autocorrelation vector $V$ based on Van der Waals volumes were used as molecular descriptors. Oxygenated compounds are correctly described by only one component $V_0$. The correlation coefficient is $0\cdot993$ for all the chemical families studied, and the standard error of the estimate is low ($s = 2\cdot56$). Comparison of standard errors of the estimate for Wiener's indices ($s = 17\cdot38$) and autocorrelation components ($s = 3\cdot95$), for all chemical families, shows that components of MAM are more efficient as molecular descriptors. The quality of regression equations was not significantly changed when 104 oxygenated compounds and 141 alkanes were studied in a global set.**

## INTRODUCTION

Molecular volume is an important parameter in quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) studies, particularly in drug design, in environmental toxicity problems and in the estimation of solubilities and partition coefficients. Estimation of molecular and molar volumes from chemical structure is therefore of great interest. Molecular volumes are estimated from bond

0894–3230/93/100574–09$09.50

*Received 3 March 1993*
*Revised 5 June 1993*

lengths and Van der Waals radii in Bondi's[1] and Meyer's[2,3] methods. In other methods, a molecule is put in a grid and its volume calculated. Molar volumes can be estimated from other physical properties such as the parachor[4] or from characteristic constants.[5] In other approaches, molar volumes are estimated from topological indices or other molecular descriptors.[5]

A very important step when establishing QSPR is a correct description of molecular structures. Among the methods used to describe molecular structures, many are based on topological information contained in molecular graphs, usually hydrogen-depleted graphs, which may be seen according to different points of view.

In approaches based on topological indices (TIs), the structural information contained in the distance matrices and connectivity matrices is condensed into descriptors, called topological indices, accounting for the topology and the connectivity of the molecule. Numerous topological indices have been proposed (for reviews, see Refs 6 and 7) and many used successfully in QSAR and QSPR studies.[7]

Some topological indices are only based on distances in the graph. The Wiener numbers $W$ and $P$ were introduced in 1947.[8] $W$ is the number of bonds between every pair of atoms in an acyclic molecule. Variants of the Wiener index such as $W_r$ (reduced Wiener index)[9] and $W_{exp}$ (expanded Wiener index)[7] were also introduced. $P$ was defined as the number of unique paths of three bonds lengths in the molecule and later associated with steric crowding. The Balaban index $J$ was introduced later.[10,11]

Other indices are based on distances and connectivities. The highly successful Randic connectivity index $\chi$ (also called the molecular connectivity index)[12] is defined as an additive quantity based on the connectivities of all non-hydrogen atoms in a molecule. Molecular connectivity indices of order $h$ (${}^h\chi$) have been proposed as extensions of the $\chi$ index.

These last two types of indices were found to be limited when electron distribution or hydrogen bonding interactions must be considered in QSPR. To improve the efficiency of topological indices, Kier and Hall[13] introduced the valence connectivity indices, including information not present in the graph such as the number of electrons and valence electrons in an atom. Balaban proposed[14] local vertex invariants (LOVI) based on the adjacency or distance matrices.

In the DARC system,[15] a molecular graph is considered as a structural environment around a focus which can be an atom, a group of atoms or a bond. In addition to this topological description, the DARC system introduces a progressive chromatic description of the molecule by considering several atomic properties. More recently a local description of the graph based on the same concept was introduced by elaborating FREL descriptors.[16]

With the autocorrelation method (AM), Moreau and Broto[17] introduced into the field of structure–activity relationships the mathematical autocorrelation function $F(\theta)$ of a function $f(x)$, defined as

$$F(\theta) = \sum_{AB} f(x)f(\theta + x)\,\mathrm{d}x \qquad (1)$$

where $\theta$ is a distance eventually determining periodicity.

Considering the hydrogen-depleted graph of the molecule as a topological space where a discrete autocorrelation function can be defined, the analogue of the distance $\theta$ is the topological distance $k$ between two atoms $i$ and $j$, taken as the smallest number of bonds connecting them. The analogue of the product $f(x)f(\theta + x)$ is the product $f(i)f(j)$, where $i$ and $j$ represent the nodes of the graph (atoms of the molecule) and $f(i)$ and $f(j)$ represent contributions or properties of the atoms $i$ and $j$.

The sum calculated for all pairs of atoms in the graph separated by a distance $d$ of $k$ bonds is considered as the component of order $k$ ($P_k$) of an autocorrelation vector $P$ corresponding to the specific property described by $f(i)$ and defined as

$$P_k = \sum_{j \geqslant i} f(i)f(j) \qquad d = k \qquad (2)$$

An autocorrelation vector can be calculated for every type of atomic property. Moreau *et al.*[18] used Pauling's electronegativity,[19] Van der Waals volume,[20] connectivity (number of non-hydrogen neighbours or vertex degree of the atom $i$), $\pi$ functionality, hydrogen bonding acceptor and donor ability. More recently, the octanol–water partition coefficient was also successfully used.[21] In several recent QSAR studies[22-24] the components of autocorrelation vectors were used as molecular descriptors.

However, it is very difficult to give a satisfactory physical interpretation of the components. For example, estimation of roughly additive molecular properties is difficult because the component zero of the autocorrelation vector is not the sum of atomic contributions. To reduce these drawbacks, Chastrette and Tiyal[25] modified the autocorrelation method in several important respects (see description and example under Experimental) to obtain the multifunctional autocorrelation method (${}^{0\cdot5}$MAM).

In this paper, the most widely used topological indices (excluding the DARC descriptors) are compared with autocorrelation descriptors for the estimation of molar volumes.

## EXPERIMENTAL

### Materials

Molar volumes at 25 °C were obtained from the API project[26] for a set including all alkanes from $C_6$ to $C_{11}$, except 2,2,3,3-tetramethylbutane, which is a solid at

25 °C. This set of 300 alkanes is called set 1. All values were consistent with those of a set of 69 alkanes from $C_5$ to $C_9$ (called set 2), used by Needham *et al.*[27] in a comparison of different descriptors in structure–property correlations. Volumes of 36 linear alkanes from $C_5$ to $C_{40}$ (set 3) were taken from the API project.[26] Molar volumes of 104 oxygenated compounds including 40 alcohols, 30 ethers and 34 ketones were calculated from data in the *Handbook of Chemistry and Physics*.[28]

### *Methods*

All equations relating molecular descriptors and molar volumes were established using a linear regression analysis program included in STATITCF software.[29] Components $C_i$ of autocorrelation connectivity vectors and components $V_i$ of a MAM volume vector were used as molecular descriptors.

*Multifunctional Autocorrelation Method (MAM).* The autocorrelation method AM was modified as follows:

(i) The notion of expanded atoms was introduced, as the hydrogen-suppressed graph where all carbon atoms are considered as equivalent was not suitable for properties such as molar volume. Using MAM, carbon atoms are coded as C, CH, $CH_2$ or $CH_3$ to take into account hydrogen atoms in the description of the molecule. Oxygen atoms are coded into several types depending on environment and hybridization state.

(ii) As it was very difficult to find correct atomic properties for atoms included in specific groups, such as oxygen or nitrogen in nitro group, some groups of atoms were considered as pseudo-atoms (Figure 1). In this study the only pseudo-atom is the carbonyl group.

(iii) The components $P_k$ of the autocorrelation vectors are calculated as

$$P_k = \sum_{j \geqslant i} f(p_i)f(p_j) \qquad d = k \qquad (3)$$

where $f(p_i)$ is a function of the property $p_i$ of the atom or group $i$. Any function of any property can be used and several known TIs can be calculated as particular cases of components of a vector $P$. In the particular case where $f(p_i) = (p_i)^m$, different versions of the MAM method will be denoted $^m$MAM.

To give a physical meaning to the vector components, $f(p_i)$ was taken as $(p_i)^{1/2}$. In this way, the first component $P_0$ is the sum of atomic properties. For a roughly additive molecular property, $P_0$ can be considered as a first approximation and the other components $P_i$ as measures of interactions between atoms and pseudo-atoms in the molecule. This corresponds to $^{0.5}$MAM. For the sake of comparison, another version of MAM called $^1$MAM in which the function $f(i)$ was the property itself as in the classical Moreau autocorrelation method was also used to study the influence of the function.

In this study, six components $C_i$ ($i = 0$–$5$) of a connectivity vector $C$ were calculated using classical connectivity. Six components $V_i$ ($i = 0$–$5$) of a volume vector $V$ were calculated using Bondi's contributions[1] for four types of expanded carbon atoms, two types of $sp^3$ oxygen atoms and the pseudo-atom carbonyl group. No attempt was made to optimize atomic or group contributions.

*Example.* 2-Methylhexanol (Figure 2) was chosen to illustrate the calculation of components of autocorrelation vectors, using the MAM method. The properties of the different groups present in the molecule (i.e. $CH_3$, $CH_2$, CH and OH) are given in Table 1. Some components of the autocorrelation vectors, calculated using equation (3), are shown in Table 2.

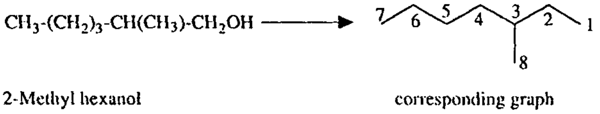

Figure 2. Structural formula and graph of 2-methyl hexanol

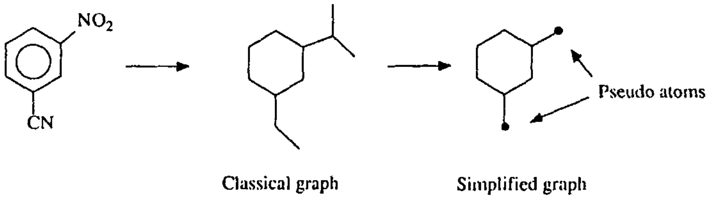

Figure 1. Classical and simplified graphs of 3-nitro benzonitrile

Table 1. Contributions of the atoms and groups to some molecular properties

| Property | No. of atom or expanded atom 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Connectivity, $C^a$ | 1 | 2 | 3 | 2 | 2 | 2 | 1 | 1 |
| Van der Waals volume, $V$ (cm$^3$mol$^{-1}$)$^b$ | 6·99 | 10·23 | 6·78 | 10·23 | 10·23 | 10·23 | 13·67 | 13·67 |
| Van der Waals surface, $S$ ( × $10^9$ cm$^2$mol$^{-1c}$ | 1·46 | 1·35 | 0·57 | 1·35 | 1·35 | 1·35 | 2·12 | 2·12 |

[a] Number of non-hydrogen neighbours of the atom or expanded atom.

[b,c] Values taken or estimated from Bondi's work[1].

Example: calculation of $V_0$ for 2-methylhexanol:

$$V_0 = \Sigma f(v_i)f(v_i) = \Sigma (v_i)^{1/2}(v_i)^{1/2} = \Sigma v_i.$$

$$V_0 = v_1 + v_2 + \cdots + v_8 = 6{\cdot}99 + 10{\cdot}23 + \cdots + 13{\cdot}67 = 82{\cdot}03 \text{ cm}^3\text{mol}^{-1}.$$

Table 2. Some components of three MAM vectors for 2-methylhexanol

| Property | Components $P_i$ $P_0$ | $P_1$ | $P_5$ | $P_6$ |
|---|---|---|---|---|
| Connectivity, $C$ | 14 | 13·46 | 3·83 | 1 |
| Van der Waals volume, $V$ (cm$^3$ mol$^{-1}$) | 82·03 | 67·03 | 33·95 | 9·78 |
| Van der Waals surface, $S$ ( × $10^9$ cm$^2$mol$^{-1}$) | 11·67 | 8·65 | 5·22 | 1·76 |

## RESULTS AND DISCUSSION

### Estimation of the molar volume of alkanes

Needham *et al.*[27] compared several structural parameters used in the molecular modelling of physical properties, including molar volumes for 69 alkanes. Here the components of $^{0\cdot5}$MAM and $^1$MAM vectors based on the same structural parameters are compared using Needham's *et al.* sample (set 2) and the larger set 1.

### *Structural information limited to connectivities and topological distances*

To compare MAM components with molecular descriptors discussed by Needham *et al.*,[27] components of autocorrelation vectors were calculated using only connectivity. To assess the performance of different descriptors, results are discussed first for separate sets of isomers and then for both sets 1 and 2.

### *Regressions for isomers*

When isomers were considered separately, good regressions using one descriptor were obtained with the Wiener number $p$ and component $C_3$ (connectivity vector). Both descriptors account for interactions between atoms (or groups) separated by three bonds. When two descriptors were used, all regressions were good except those using *ad hoc* descriptors (Table 3).

### *Regressions on set 1*

*Comparison between $^{0\cdot5}$MAM and $^1$MAM.* Regressions obtained with set 1 (300 alkanes) using one parameter [Table 4(a)], two parameters [Table 4(b)] and more than two parameters [Table 4(c)] are compared for components $C_i$ of the autocorrelation connectivity vectors calculated with both $^{0\cdot5}$MAM and $^1$MAM methods.

These results show that the components $C_i$ of the $^{0\cdot5}$MAM connectivity vector are successful in the estimation of the molar volume of this set of alkanes and present clear advantages over components of the $^1$MAM vector.

*Comparison between MAM descriptors and other descriptors.* Regressions obtained with sets 1 (300 alkanes) and 2 (69 alkanes) using Wiener indices, valence connectivity indices, *ad hoc* descriptors and components $C_i$ of the MAM autocorrelation connectivity vector were then compared (Table 5).

All regressions using one or two descriptors were significantly worse with sets 1 and 2 than with separate sets of isomers. $^{0\cdot5}$MAM descriptors were often more efficient than Wiener indices but good regressions were

Table 3. Comparison of correlation coefficients ($r$) and standard errors of the estimates ($s$) obtained in the estimation of molar volume of alkanes using MAM connectivity vector and volume vector components, topological indices and *ad hoc* descriptors

| Alkane | Descriptor | | | | |
|---|---|---|---|---|---|
| | *Ad hoc*: $T_m$, $T_3$,[a] $N^b$ | $\chi$: $^{0}\chi$, $^{1}\chi$, $\chi_t$ | Wiener: $W$, $W_r$, $p^c$ | MAM: $C_i$ ($i = 0$–$4$)[c] | MAM: $V_i$ ($i = 0$–$4$)[c] |
| $C_6$ | Not signif. | Not signif. | $W$, $p$ | Not signif. | $V_2$, $V_3$ |
| ($n = 5$) | | | 0·993 (0·23) | | 0·997 (**0·14**) |
| $C_7$ | Not signif. | $^{0}\chi$, $^{1}\chi$ | $W$, $p$ | Not signif. | $V_2$, $V_3$ |
| ($n = 9$) | | 0·887 (1·07) | 0·994 (**0·25**) | | 0·991 (0·30) |
| $C_8$ | $T_3$ | $^{0}\chi$, $^{1}\chi$ | $W$, $p$ | $C_3$, $C_4$ | $V_2$, $V_3$ |
| ($n = 18$) | 0·740 (1·81) | 0·923 (1·07) | 0·995 (**0·28**) | 0·949 (0·88) | 0·987 (0·45) |
| | | | | $C_3$, $C_4$, $C_5$ | |
| | | | | 0·995 (0·30) | |
| $C_9$ | $T_3$ | $^{1}\chi$, $\chi_t$ | $p$ | $C_3$, $C_4$ | $V_3$, $V_4$ |
| ($n = 35$) | 0·640 (2·54) | 0·971 (0·80) | 0·979 (0·68) | 0·978 (0·70) | 0·980 (0·66) |
| | | | | $C_1$, $C_3$, $C_4$ | |
| | | | | 0·982 (**0·64**) | |
| $C_{10}$ | $T_3$ | $^{1}\chi$, $\chi_t$ | $p$ | $C_1$, $C_2$ | $V_3$, $V_4$ |
| ($n = 75$) | 0·695 (2·80) | 0·949 (1·23) | 0·976 (0·84) | 0·961 (1·09) | 0·976 (0·86) |
| | | | | $C_1$, $C_2$, $C_4$ | |
| | | | | 0·980 (**0·79**) | |
| $C_{11}$ | $T_m$, $T_3$ | $^{0}\chi$, $^{1}\chi$ | $W$, $p$ | $C_1$, $C_2$ | |
| ($n = 159$) | 0·653 (3·30) | 0·904 (1·87) | 0·980 (0·86) | 0·966 (1·12) | |
| | | $^{0}\chi$, $^{1}\chi$, $\chi_t$ | | $C_1$, $C_2$, $C_4$ | $V_1$–$V_4$ |
| | | 0·960 (1·22) | | 0·981 (0·85) | 0·983 (**0·81**) |

[a] $T_m$ is the number of terminal methyl groups and $T_3$ is the number of terminal methyl groups separated by three bonds.

[b] $N_c$ is the number of carbon atoms of the compound.

[c] Values given are correlation coefficient and (in parentheses) standard error of the estimate in the regression (cm$^3$ mol$^{-1}$) (the best standard errors of the estimates are indicated in bold type).

obtained only when more than two components $C_i$ were used. An excellent regression ($n = 300$, $r = 0·999$, $s = 0·81$) was obtained with four components ($C_0$–$C_4$).

All regressions were improved by introducing descriptors such as $C_3$, $p$ or $T_3$ that account for butane-type interactions. With set 1 the contributions of $C_0$ and $C_3$ calculated according to Gore[30] are 78% and 22%, respectively. $C_0$ contributes positively and $C_3$ negatively, as expected for a rough measure of branching. Working on a more restricted sample, Needham *et al.*[27] obtained very significant models ($r = 0·988$, $s = 2·73$) by using *ad hoc* descriptors ($N_c$, $T_m$). Actually, $N_c$ is well correlated with $C_3$ ($r = 0·929$, $s = 1·41$).

Table 4. Quality of regressions using a single parameter (a), two parameters (b) or more than two parameters (c) in the estimation of the molar volume of 300 alkanes from the MAM connectivity vector $C$

| Parameters | Method | |
|---|---|---|
| | $^{0·5}$MAM | $^{1}$MAM |
| (a) One descriptor | $C_0$ | $C_0$ |
| $r(s)^a$ | 0·976 (3·92) | 0·814 (10·56) |
| (b) two descriptors | $C_0$, $C_3$ | $C_0$, $C_3$ |
| $r(s)^a$ | 0·995 (1·86) | 0·862 (9·22) |
| (c) >2 descriptors | $C_0$–$C_4$ | $C_0$–$C_4$ |
| $r(s)^a$ | 0·999 (0·81) | 0·979 (3·74) |

[a] $r$ = Correlation coefficient; $s$ = Standard error of the estimate.

*Autocorrelation descriptors based on volumes.* Generally, models obtained using descriptors based on connectivity are good enough for practical purposes. However, to derive more significant models to estimate molar volumes, it seemed logical to consider components of an autocorrelation vector based on Van der Waals volumes. In the following, we compare $^{0·5}$MAM or $^{1}$MAM descriptors and topological indices, first on Needham's *et al.* set of 69 alkanes and then on set 1 (300 alkanes). Results are reported in Tables 6 and 7.

### *$V_0$ ($^{0·5}$MAM and $^{1}$MAM) as a first approximation of molar volume*

*Linear alkanes (set 3).* For a set of 36 linear alkanes

Table 5. Comparison of correlation coefficients ($r$) and (in parentheses) standard errors of the estimates ($s$) (cm$^3$ mol$^{-1}$) obtained in the estimation of molar volume of alkanes in sets 1 and 2 using MAM connectivity vector components and some topological descriptors

| Set | No. of descriptors | *Ad hoc*: $T_m$, $T_3$, $N_c$ | $\chi$: ${}^0\chi$, ${}^1\chi$, $\chi_t$ | Wiener: $W$, $W_r$, $p$ | MAM: $C_i$[b] ($i$ = 0–4) |
|---|---|---|---|---|---|
| 1<br>$C_5$–$C_9$<br>($n$ = 69) | 1 | $N_c$[a]<br>0·987 (2·81) | ${}^0\chi$[a]<br>0·961 (4·80) | $W$[a]<br>0·972 (4·08) | $C_0$<br>0·987 (2·81) |
| 1<br>$C_5$–$C_9$<br>($n$ = 69) | 2 | $T_3$, $N_c$[a]<br>0·988 (2·73) | ${}^0\chi$, ${}^1\chi$[a]<br>0·982 (3·33) | $W$, $p$[a]<br>0·976 (3·78) | $C_0$, $C_3$<br>0·996 (1·57) |
| 2<br>$C_6$–$C_{11}$<br>($n$ = 300) | 1 | $N_c$[b]<br>0·976 (3·92) | ${}^0\chi$[b]<br>0·962 (4·80) | $W$[b]<br>0·962 (4·80) | $C_0$<br>0·987 (2·81) |
| 2<br>$C_6$–$C_{11}$<br>($n$ = 300) | 2 | $T_3$, $N_c$[b]<br>0·986 (3·05) | ${}^0\chi$, ${}^1\chi$[b]<br>0·973 (4·17) | $W$, $p$[b]<br>0·977 (3·92) | $C_0$, $C_3$<br>0·995 (1·86) |
| 2<br>$C_6$–$C_{11}$<br>($n$ = 300) | ≥3 | $T_m$, $T_3$, $N_c$[b]<br>0·987 (2·99) | ${}^0\chi$, ${}^1\chi$, $\chi_t$[b]<br>0·975 (4·04) | | $C_0$, $C_1$, $C_2$<br>0·998 (1·07)<br>$C_0$–$C_4$<br>0·999 (0·81) |

[a] Results from Ref. 27.
[b] This work.

($C_5$–$C_{40}$) an excellent correlation was obtained[31] between the molar volume and component $V_0$ of the MAM.

Volume vector:

$$V_m = 1\cdot605V_0 + 20\cdot911 \quad (n = 36,\ r = 0\cdot999,\ s = 0\cdot53) \qquad (4)$$

Analysis of the distribution of residuals suggested the introduction of a variable $D_{mp}$, defined as $D_{mp}$ = melting point $-25$ (°C). A two-variable equation was obtained:

$$V_m = 1\cdot615V_0 - 0\cdot023D_{mp} + 19\cdot003 \quad (n = 36,\ r = 0\cdot9999,\ s = 0\cdot14) \qquad (5)$$

This excellent equation shows that even in the very simple case of linear alkanes other molecular properties, not accounted for by descriptors, can play a role in the determination of molar volumes and limit the quality to be expected in simple QSPR.

Huggins[32] calculated the molar volume of $n$-alkanes ($C_5$–$C_{20}$) using a two-variable equation including $N_c$ and $1/N_c$. On our set 3 a similar equation gave good results ($n = 36$, $r = 0\cdot9999$, $s = 0\cdot06$) but no clear physico-chemical interpretation could be given to $1/N_c$.

*Sets 1 and 2.* The following equations were obtained:

$${}^{0\cdot5}\text{MAM:}\ V_m = 1\cdot467V_0 + 31\cdot765 \quad (n = 300,\ r = 0\cdot977,\ s = 3\cdot910) \qquad (6)$$

$${}^{1}\text{MAM:}\ V_m = 0\cdot117V_0 + 41\cdot196 \quad (n = 300,\ r = 0\cdot910,\ s = 7.325) \qquad (7)$$

$${}^{0\cdot5}\text{MAM:}\ V_m = 1\cdot444V_0 + 33\cdot011 \quad (n = 69,\ r = 0\cdot987,\ s = 2\cdot811) \qquad (8)$$

$${}^{1}\text{MAM:}\ V_m = 0\cdot120V_0 + 34\cdot158 \quad (n = 69,\ r = 0\cdot951,\ s = 5\cdot368) \qquad (9)$$

Equations using ${}^{0\cdot5}$MAM volume descriptors are very similar to those obtained using connectivity descriptors because of the high correlation between corresponding components. As ${}^{0\cdot5}$MAM vectors are clearly superior to ${}^1$MAM vectors in both sets, the latter will not be considered in the following discussion.

### *Comparison of MAM (using several $V_i$) and topological indices*

To obtain better regressions it was necessary to consider several components of a volume vector $V$. Results are presented first for separate families and then for mixed sets.

*Regressions on isomers.* Results in Table 6 show that MAM components and Wiener's indices led to comparable regressions while connectivity indices and *ad hoc* descriptors were less adequate.

*Regression on set 2* (*Table 6*). When higher components of the MAM vector were introduced to account

Table 6. Comparison of correlation coefficients ($r$) and (in parentheses) standard errors of the estimates ($s$) ($cm^3\ mol^{-1}$) obtained in the estimation of molar volume using MAM volume vector components or topological indices

| | Descriptors | | | |
|---|---|---|---|---|
| Alkanes | *Ad hoc*:<br>$T_m$, $T_3$, $N_c$ | $\chi$:<br>$^0\chi$, $^1\chi$, $\chi_t$ | Wiener<br>$W$, $W_r$, $p$ | MAM:<br>$V_i$ ($i = 0$–$4$)[a] |
| $C_5$–$C_9$<br>($n = 69$)[27] | $T_3$, $N_c$<br>0·988 (2·73) | $^0\chi$, $^1\chi$<br>0·982 (3·33) | $W$, $p$<br>0·976 (3·78) | $V_0$, $V_3$<br>0·998 (**0·97**) |
| $C_6$–$C_{11}$<br>($n = 300$) | | | $W$, $p$<br>0·977 (3·92) | $V_0$, $V_3$<br>0·998 (**1·23**) |
| | $T_m$, $T_3$, $N_c$<br>0·987 (2·99) | $^0\chi$, $^1\chi$, $\chi_t$<br>0·975 (4·04) | $W_r$, $p$<br>0·942 (6·09) | $V_0$ $V_3$, $V_4$<br>0·999 (**0·85**) |

[a] The best standard errors of the estimates are indicated in bold type.

for molecular branching, the best equation obtained with set 2 was

$$V_m = 1\cdot724V_0 - 0\cdot12V_3 - 0\cdot038V_4 + 19\cdot305 \quad (n = 69,\ r = 0\cdot999,\ s = 0.764) \qquad (10)$$

The contribution of each variable was estimated using Gore's method.[30] Molar volume appears as an approximately additive property, largely determined by $V_0$ (which has a contribution of 81%). However, the contributions of $V_3$ and $V_4$ were 15% and 4%, respectively, suggesting that branching plays a role, especially for atoms separated by three or four bonds. This corresponds to the well known butane and pentane interactions.[33] For this set other descriptors (including Wiener's indices) were clearly inferior in dual parameter regressions.

*Regression on set 1* (*Table 6*). The best equation obtained includes three components:

$$V_m = 1\cdot750V_0 - 0\cdot123V_3 - 0\cdot037V_4 + 17\cdot97 \quad (n = 300,\ r = 0\cdot999,\ s = 0\cdot85) \qquad (11)$$

MAM vector components based on only four different contributions of atoms or expanded atoms describing microscopic properties enabled us to correlate a macroscopic property such as the molar volume. This means that somehow the higher components of $V$ account for both molecular shapes and interactions in the liquid phase. Van der Waals volume, which is a specific atomic property, is more appropriate than connectivity (a non-specific atomic property) to derive molecular descriptors useful in the estimation of molar volumes.

## Estimation of molar volumes of oxygenated compounds

To extend the comparison between MAM and other approaches to compounds containing heteroatoms, we have considered[34] a set of 104 oxygenated compounds including 40 alcohols, 34 ketones and 30 ethers, where each functional group represents a different interaction type. Chemical families were first studied separately, then the total set was considered. Only the components of an autocorrelation vector based on Van der Waals volume were used as molecular descriptors. A summary of results obtained when comparing MAM descriptors with other topological descriptors is given in Table 7.

### *Alcohols*

A population of 40 alcohols (linear or branched) was used to establish a structure–volume relationship. For this set an excellent regression [Equation (12)] was obtained between the molar volume and components of the volume vector $V$. The contributions of components calculated according to Gore[30] were 88%, 7% and 5% for $V_0$, $V_1$ and $V_3$, respectively.

$$V_m = 1\cdot86V_0 - 9\cdot14V_1 - 0\cdot097V_3 + 3\cdot67 \quad (n = 40,\ r = 0\cdot999,\ s = 0\cdot40) \qquad (12)$$

### *Ethers*

A population of 30 ethers (linear or branched) was used. For this set higher components of the volume vector $V$ were not significant. The contributions calculated according to Gore[30] were 73%, 20% and 7% for $V_0$, $V_1$ and $V_2$, respectively.

$$V_m = 2\cdot52V_0 - 0\cdot64V_1 - 0\cdot26V_2 + 2\cdot95 \quad (n = 30,\ r = 0\cdot999,\ s = 1\cdot51) \qquad (13)$$

### *Ketones*

For this population of 34 ketones (linear or branched) an excellent regression was obtained using the same

Table 7. Comparison of correlation coefficients ($r$) and (in parentheses) standard errors of the estimates ($s$) ($cm^3\ mol^{-1}$) obtained in the estimation of molar volume of oxygenated compounds using MAM components or topological indices

| Compounds | No. of descriptors | $\chi^v$: $^0\chi^v$, $^1\chi^{v}$[a] | $\chi$: $^0\chi$, $^1\chi$ | Wiener: $W$, $W_r$, $p$ | MAM: $V_i$ ($i = 0$–$4$)[a] |
|---|---|---|---|---|---|
| Alcohols<br>($n = 40$) | 1 | $^1\chi^v$<br>0·994 (3·88) | $^0\chi$<br>0·994 (3·99) | $W$<br>0·907 (14·97 | $V_0$<br>0·999 (**1·05**) |
| | >1 | $^0\chi^v$, $^1\chi^v$<br>0·999 (1·59) | $^0\chi$, $^1\chi$<br>0·999 (1·44) | $W$, $p$<br>0·970 (8·81) | $V_0$, $V_3$<br>0·999 (**0·74**)<br>$V_0$, $V_1$, $V_3$<br>0·999 (**0·40**) |
| Ketones<br>($n = 34$) | 1 | $^1\chi^v$<br>0·990 (4·85) | $^0\chi$<br>0·994 (3·84) | $W$<br>0·953 (10·56) | $V_0$<br>0·999 (**1·06**) |
| | >1 | $^0\chi^v$, $^1\chi^v$<br>0·999 (1·43) | $^0\chi$, $^1\chi$<br>0·999 (1·17) | $W$, $p$<br>0·976 (7·75) | $V_0$, $V_5$<br>0·999 (**0·98**) |
| Ethers<br>($n = 30$) | 1 | $^1\chi^v$<br>0·996 (4·19) | $^0\chi$<br>0·997 (3·30) | $p$<br>0·995 (4·55) | $V_0$<br>0·999 (**1·61**) |
| | >1 | $^0\chi^v$, $^1\chi^v$<br>0·999 (**1·05**) | $^0\chi$, $^1\chi$<br>0·999 (1·59) | $W$, $p$<br>$W$ non-signif. | $V_0$, $V_1$, $V_2$<br>0·999 (1·52) |
| Alcohols<br>+ ketones<br>($n = 74$) | 1 | $^1\chi^v$<br>0·990 (5·20) | $^0\chi$<br>0·994 (4·11) | $W$<br>0·921 (14·46) | $V_0$<br>0·993 (**1·49**) |
| | >1 | $^0\chi^v$, $^1\chi^v$<br>0·998 (2·14) | $^0\chi$, $^1\chi$<br>0·999 (1·46) | $W$, $p$<br>0·967 (9·53) | $V_0$, $V_1$, $V_2$, $V_3$<br>0·999 (**1·12**) |
| Alkanes<br>+ alcohols<br>+ ketones | 1 | $^1\chi^v$<br>0·980 (6·67) | $^0\chi$<br>0·963 (9·02) | $W$<br>0·766 (21·59) | $V_0$<br>0·991 (**4·49**) |
| + ethers<br>($n = 245$) | 2 | $^0\chi^v$, $^1\chi^v$<br>0·992 (4·26) | $^0\chi$, $^1\chi$<br>0·968 (8·49) | $W$, $p$<br>0·856 (17·38) | $V_0$, $V_3$<br>0·993 (**3·95**) |

[a] The best standard errors of the estimate are indicated in bold type.

three components of the volume vector. The contributions calculated according to Gore[30] were 75%, 15% and 10% for $V_0$, $V_1$ and $V_2$, respectively.

$$V_m = 1{\cdot}22V_0 - 0{\cdot}24V_1 - 0{\cdot}14V_2 + 17{\cdot}22$$
$$(n = 34,\ r = 0{\cdot}9996,\ s = 0{\cdot}99) \qquad (14)$$

### *Alkanes and oxygenated compounds*

A set of 245 compounds including 141 alkanes ($C_6$–$C_{10}$), 40 alcohols, 30 ethers and 34 ketones (linear and branched) was finally used to establish a structure–volume relationship. A good correlation was obtained as above between the molar volume and components of the volume vector. Contributions of components $V_i$ calculated using Gore's method[30] were 60%, 20%, 13% and 7% for $V_0$, $V_1$, $V_2$ and $V_3$, respectively.

$$V_m = 3{\cdot}77V_0 - 1{\cdot}36V_1 - 0{\cdot}551V_2 - 0{\cdot}24V_3 - 22{\cdot}811$$
$$(n = 245,\ r = 0{\cdot}997,\ s = 2{\cdot}56) \qquad (15)$$

The results in Table 7 demonstrate that MAM component $V_0$ is sufficient to describe oxygenated compounds. Indeed, the correlation coefficient is higher than 0·993 for all chemical families and in addition the standard error is very low. The quality of the equation did not change when oxygenated compounds and alkanes were grouped into a global set.

Comparison of standard errors of the estimate for Wiener's indices ($s = 17{\cdot}38$) and autocorrelation components ($s = 3{\cdot}95$) for all chemical families shows that the components of MAM are more adequate as molecular descriptors. Connectivity indices appear as good descriptors, especially in the case of ethers. This could be ascribed to a good description of the polarity of the molecule by the atomic property $\delta^v$ used in the calculation of $^h\chi^v$. However, further study is needed to understand why these indices do not work so well with alcohols and ketones.

## CONCLUSION

On the whole, good results were obtained using MAM to estimate molar volumes of alkanes and oxygenated compounds.

Wiener indices ($W, p$) and their variants ($W_r$) turned out to be more efficient than MAM components when sets of isomeric alkanes were considered, while connectivity indices and *ad hoc* descriptors were less useful. However, for ethers (Table 7) MAM descriptors clearly became more adequate when alkanes were considered in sets 1 and 2. For example, with set 2 the standard deviation was 0·77 using MAM and 3·92 using Wiener indices.

It seems that the success of MAM is related to the amount of information brought by each component. Thus $V_0$, for example, was found to be highly correlated with the size of the molecule, whereas $V_3$ and $C_3$ (or $V_4$ and $C_4$) take into account interactions between atoms or groups separated by three (or four) bonds.

## REFERENCES

1. A. Bondi, *J. Phys. Chem.* **64**, 441–451 (1964).
2. A. Y. Meyer, *J. Chem. Soc., Perkin. Trans. 2*, 1161–1169 (1985).
3. A. Y. Meyer, *Chem. Soc. Rev.* **15**, 449–474 (1986).
4. W. J. Lyman, W. F. Reehl and D. H. Rosenblatt, *Handbook of Chemical Property Estimation Methods.* American Chemical Society, Washington, DC (1990).
5. J. C. McGowan and A. Mellors. *Molecular Volumes in Chemistry and Biology: Application Including Partitioning and Toxicity.* Ellis Horwood, Chichester (1986).
6. A. T. Balaban, *J. Mol. Struct. (Theochem)* **55**, 263–269 (1988).
7. S. S. Tratch, M. I. Stankevich and N. S. Zefirov, *J. Comput. Chem.* **11**, 899–908 (1990).
8. H. Wiener, *J. Chem. Phys.* **15**, 766–769 (1947).
9. H. Wiener, *J. Am. Chem. Soc.* **69**, 2636–2639 (1947).
10. A. T. Balaban, *J. Mol. Struct. (Theochem)* **55**, 263–269 (1988).
11. A. T. Balaban, *Theor. Chim. Acta* **165**, 243–253 (1988).
12. L. B. Kier, L. H. Hall, W. J. Murray and M. Randic, *J. Pharm. Sci.* **64**, 1971–1974 (1975).
13. L. B. Kier and L. H. Hall, *J. Pharm. Sci.* **65**, 1806–1809 (1976).
14. A. T. Balaban, *Pure Appl. Chem.* **55**, 199–202 (1983).
15. J. E. Dubois, D. Laurent and H. Viellard, *C. R. Acad. Sci.* **263**, 764–767 (1966).
16. J. E. Dubois, A. Panaye and R. Attias, *J. Chem. Inf. Comput. Sci.* **27**, 74–82 (1987).
17. G. Moreau and P. Broto, *Nouv. J. Chim.* **4**, 359–364 (1980).
18. G. Moreau, P. Broto and C. Vandycke, *Eur. J. Med. Chem. Chim. Theor.* **19**, 71–78 (1984).
19. L. Pauling, *The Nature of the Chemical Bond*, 3rd ed. Cornell University Press, Ithaca, NY (1960).
20. I. Moriguchi, Y. Kanada and K. Komatsu, *Chem. Pharm. Bull.* **24**, 1799–1806 (1976).
21. G. Moreau, P. Broto, M. Fortin and C. Turpin, *Eur. J. Med. Chem. Chim Theor.* **23**, 275–281 (1988).
22. J. Devillers, P. Chambon, D. Zakarya and M. Chastrette, *Chemosphere* **15**, 993–1002 (1986).
23. M. Chastrette, D. Zakarya and A. Elmouaffek, *Eur. J. Med. Chem. Chim. Theor.* **21**, 505–510 (1986).
24. J. Devillers, D. Zakarya, P. Chambon and M. Chastrette, *C. R. Acad. Sci., Sér. III* **304**, 195–198 (1987).
25. M. Chastrette and F. Tiyal, *C. R. Acad. Sci. Sér. II* **310**, 511–514 (1990).
26. *Research Project 44.* American Petroleum Institute, Thermodynamics Research Center, Texas A & M University, TX (1968).
27. D. E. Needham, I.-C. Wei and P. S. Seybold, *J. Am. Chem. Soc.* **110**, 4186–4194 (1988).
28. R. C. Weast and S. M. Shelby (Eds) *Handbook of Chemistry and Physics*, 70th ed. Chemical Rubber Company, Cleveland, OH (1989).
29. *STAT-ITCF.* Institut Technique des Céréales et des Fourrages, Boigneville, France (1988).
30. W. L. Gore, *Statistical Methods for Chemical Experimentation*, Interscience, New York (1952).
31. M. Chastrette, F. Tiyal, J. F. Peyraud and D. Zakarya, *C. R. Acad. Sci., Sér. II* **314**, 461–466 (1992).
32. M. Huggins, *J. Am. Chem. Soc.* **63**, 116–120 (1941).
33. Z. Mihalic, S. Nicolik, N. Trinajstic, *J. Chem. Inf. Comput. Sci.* **32**, 28–37 (1992).
34. F. Tiyal, Diplôme de Doctorat, No. 248, Université Lyon I, Lyon (1990).