Model with dual indices and complete graphs. The heterogeneous description of the dipole moments and polarizabilities

Lionello Pogliani

Dipartimento di Chimica, Università della Calabria, 87030, Rende (CS), Italy.
E-mail: lionp@unical.it

Received (in Montpellier, France) 23rd October 2002, Accepted 23rd January 2003
First published as an Advance Article on the web 6th May 2003

A preliminary study was carried out in an attempt to predict the induced dipole moments of (A) alcohols, amines, and ethers, (B) aldehydes, ketones, and esters, and (C) sulfides, and phosphines, as well as the molecular polarizabilities of halogenated organic derivatives. The description of these two properties was based on graph-theoretical molecular connectivity and pseudoconnectivity descriptors, and served to check the model ability of (i) the dual molecular connectivity and pseudoconnectivity basis indices as well as (ii) the descriptive power of molecular connectivity indices derived from regular complete graphs. These last types of graphs were used to encode the inner-core electrons of atoms with principal quantum number $n \geq 2$. The descriptions obtained in this way not only emphasize the importance of the dual basis indices but especially the importance of encoding the inner-core electrons for heteroatoms with odd complete graphs. Another relevant feature of the attempted descriptions is the importance assumed by the zeroth-order valence molecular connectivity basis index, ${}^0\chi^v$, which allows a direct physical interpretation of the model.

Introduction

Among the computational methods that are used with increasing success in the field of prediction of properties and/or activities, the mathematical methods based on topological concepts occupy an eminent place. They are normally known under the acronym QSPR/QSAR, that is , quantitative-structure-property or structure-activity relationships. These mathematical methods are mainly, but not always, based on non-empirical structure-related graph-theoretical descriptors, where molecules are seen as a set of vertices attached to each other by a set of non-metrical connections. These computational procedures allow one to forecast properties or activities of specific chemical compounds once the properties or activities of a set of those compounds has been modeled.[1–7] The range of applicability of these methods goes well beyond the QSPR field.[8–11]

A successful mathematical topological method, employed in QSPR/QSAR studies, is the molecular connectivity method that allows one to derive, by a series of rules, a set of graph-theoretical indices of wide applicability.[1,2,12–15] This method has recently undergone interesting developments with the introduction of the electrotopological state indices,[16] indices derived from line-graphs (edge-connectivity indices),[17–23] pseudoconnectivity indices, molecular connectivity, pseudoconnectivity terms, and mixed terms, dual indices,[21–25] and variable descriptors.[26–28] The recent evolution of the molecular connectivity theory, if on the one hand has greatly improved its descriptive power, has, on the other hand, deepened its 'non-absolute' character. 'Non-absolute theories' are theories that can quantitatively predict a certain property of a system from its fundamental parameters without the help of 'adjustable external' parameters. In fact, a subset of molecular connectivity indices, the valence molecular connectivity indices for higher-level atoms with $n \geq 3$ (n = principal quantum number) are defined by the aid of atomic concepts like the atomic number, the number of valence electrons, and the principal quantum number.[2,16] Thus, for atoms with $n \geq 3$ the valence molecular connectivity indices are no longer pure graph-theoretical descriptors but mixed 'graph-quantum' descriptors. Recently, this problem has been solved with the introduction of an algorithm based on complete graphs.[29] The present work will further check the validity of the complete graph conjecture for the inner-core electrons.[25] This conjecture will be used to model two properties of two classes of organic compounds that have been studied with MM methods by Ma, Lii and Allinger:[30] the molecular polarizability and the induced dipole moment.

Methodology

I. The structure-property relation

Two types of structure (S)-property (P) relations will be used throughout this work: the linear type, $P = c_1S + c_0U_0$ and the multi-linear type relation, $P = \Sigma c_iS_i$. Here P is the modeled property, c_i are the regression coefficients, and c_0 is the regression coefficient of the unitary index, $U_0 \equiv 1$. The structural descriptors, S, can either be (i) a basis index, β, (ii) a molecular connectivity, $X = \mathrm{f}(\chi)$, (iii) a pseudoconnectivity term, $Y = \mathrm{f}(\psi)$, or (iv) a mixed molecular connectivity-pseudoconnectivity term, $Z = \mathrm{f}(X,Y)$, $\mathrm{f}(Z,\beta)$ (for details about β see the next paragraph). The multiple linear relation can normally be written as a dot product: $P = \mathbf{C}\cdot S$, where $\mathbf{C} = (c_1, c_2,\ldots, c_0)$, and $\boldsymbol{S} = (\beta_1, \beta_2,\ldots,U_0)$. To avoid negative calculated P values, with no biological or physical meaning, it is advantageous to use the modulus equation: $P = |\Sigma c_iS_i|$, where bars stand for absolute value. The best basis indices to include in a linear combination of basis indices (LCBI) are chosen by the aid of a search procedure that is performed on the total combinatorial space described by the same basis indices. The procedure used to construct the dominant molecular connectivity terms, $S = X,\ Y$, is a trial-and-error procedure that chooses the best basis indices, β, and optimizes them with the several adjustable parameters.[22] The mathematical expression for the X and Y terms looks like the rational function of eqn. (1), where β is a basis index and $S = X$ or Y for $\beta = \chi$ or

View Online

$\beta = \psi$, respectively. Parameters, *a–d*, *m–p*, and *q,r* are the adjustable parameters and they can either be negative, zero or one, and in this case the rational function can be reduced into a much simpler form.

$$S = [a(\beta_1)^m + b(\beta_2)^n]^q / [c(\beta_3)^o + d(\beta_4)^p]^r \qquad (1)$$

The use of terms for modeling could be loosely called configuration interaction of graph-type basis indices (CI-GTBI) for its vague resemblance with the quantum chemistry method of configuration interaction of molecular orbitals made up of antisymmetrized basis functions. The construction of the $Z = \mathrm{f}(X,Y)$ terms is performed by the aid of a search procedure, which consists in trying the different ways to combine X and Y together (sometimes a β index is included). The mixed Z terms allow one to short-circuit the huge combinatorial problem due to the high number of basis indices.

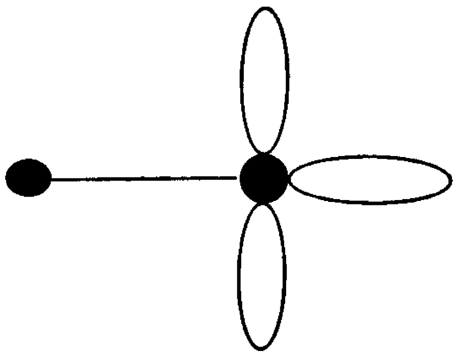

Fig. 1 The hydrogen-suppressed pseudograph of CH_3–F. The inner-core electrons of the C and F atoms are encoded by a K_1 complete graph.

II. The main basis indices

The set of basis indices, $\{\beta\}$, used in this study, is made up of three subsets of basis indices: a medium-sized subset of eight molecular connectivities, $\{\chi\} = \{D, {}^0\chi, {}^1\chi, \chi_t, D^v, {}^0\chi^v, {}^1\chi^v, \chi_t^v\}$; a medium-sized subset of eight molecular pseudoconnectivity indices, $\{\psi\} = \{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^T\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\}$; and of a medium-sized subset of twelve dual indices, $\{\beta_d\} = \{{}^0\chi_d, {}^1\chi_d, {}^1\chi_s, {}^0\chi_d^v, {}^1\chi_d^v, {}^1\chi_s^v, {}^0\psi_{Id}, {}^1\psi_{Id}, {}^1\psi_{Is}, {}^0\psi_{Ed}, {}^1\psi_{Ed}, {}^1\psi_{Es}\}$, that is:

$$\{\beta\} = \{\{\chi\},\{\psi\}, \{\beta_d\}\} \qquad (2)$$

The subset of χ indices is directly based on the vertex degrees δ and δ^v of a hydrogen-suppressed graph and pseudograph, respectively. The degree of a vertex is the number of edges incident with it, a loop at a vertex contributes twice to its degree.[31,32] If the molecule does not contain any higher row atoms, that is atoms with $n \geq 3$, then δ and δ^v values can be derived from the corresponding chemical graph and pseudograph of a molecule, respectively.[22] The δ^v values of the valence connectivity χ^v indices of higher-row atoms with $n \geq 3$ (here, Si, P, S, Cl, Br, and I) are calculated with the following $\delta^v(Z,Z^v)$ algorithm:[2]

$$\delta^v = [Z^v - h]/[Z - Z^v - 1] \qquad (3)$$

where Z^v is the number of valence electrons, Z is the atomic number, and h is the number of suppressed hydrogen atoms. For $n = 2$, $\delta^v = [Z^v - h]$, as $Z - Z^v - 1 = 1$, and δ^v can be derived from the hydrogen-suppressed pseudograph of a molecule. Thus, second-row atoms can be encoded by the aid of graph-theoretical concepts only. We remember that the hydrogen-suppressed graph of a molecule allows only single connections while the corresponding pseudograph allows for multiple connections that mimic multiple bonds and loops, that is self-connections that mimic non-bonding electrons.[22] Thus, while the hydrogen-suppressed graph of CH_3F can be represented as two connected points, •–•, where each vertex has $\delta = 1$, in the CH_3F pseudograph, instead, the C encoding vertex has $\delta^v = 1$ and the F encoding vertex has $\delta^v = 7$ (see Fig. 1, where loops count twice, plus the single connection).

Basis ψ indices are indirectly related to δ and δ^v numbers through the I-state (ψ_I subset) and S-state (ψ_E subset) atom level indices,[16] which are defined in eqns. (4) and (5):

$$I_i = [(2/n)^2 \delta_i^v + 1]/\delta_i \qquad (4)$$

$$S_i = I_i + \Sigma_j \Delta I_{ij} \qquad (5)$$

Here, the δ_i^v of the I-state index equals the δ^v that can be derived from the pseudograph (ps) of a molecule, that is, $\delta_i^v \equiv \delta_i^v(\mathrm{ps})$. The inner-core electron contribution for heteroatoms is here encoded by the $(2/n)^2$ factor, that is, $\delta^v(\mathrm{F}) = \delta^v(\mathrm{ps})$, $\delta^v(\mathrm{Cl}) = (2/3)^2\delta^v(\mathrm{ps})$, $\delta^v(\mathrm{Si}) = (2/3)^2\delta^v(\mathrm{ps})$, and so on. In practice, we can rewrite δ_i^v as $(2/n)^2\delta_i^v(\mathrm{ps})$. In eqn. (5) the factor ΔI_{ij} equals $(I_i - I_j)/r_{ij}^2$, where r_{ij} counts the atoms in the minimum path length separating two atoms i and j, which is equal to the usual graph distance, $d_{ij} + 1$. The factor $\Sigma_j \Delta I_{ij}$ incorporates the information about the influence of the remainder of the molecular environment. These two atom-level indices encode simultaneously the graph and pseudograph representation of a molecule, as they are directly (I) and indirectly (S) related to δ and δ^v numbers of a graph and a pseudograph, respectively. From what has been said, it is evident that χ^v and ψ indices are based on different δ^v values for $n > 2$. This means that a description based on χ^v and ψ indices, for $n > 2$, has a heterogeneous character. As the electrotopological state choice for $\delta_i^v = (2/n)^2\delta_i^v(\mathrm{ps})$ seems optimal,[16] we are left with the possibility to test the χ^v indices derived by the aid of this relation.

Basis χ and ψ indices are formally similar, as can be seen from the following relations:

$$D = \Sigma_i \delta_i \qquad (6a)$$

$${}^S\psi_I = \Sigma_i I_i \qquad (6b)$$

$${}^0\chi = \Sigma_i (\delta_i)^{-0.5} \qquad (7a)$$

$${}^0\psi_I = \Sigma_i (I_i)^{-0.5} \qquad (7b)$$

$${}^1\chi = \Sigma(\delta_i\delta_j)^{-0.5} \qquad (8a)$$

$${}^1\psi_I = \Sigma(I_iI_j)^{-0.5} \qquad (8b)$$

$$\chi_t = (\Pi\delta_i)^{-0.5} \qquad (9a)$$

$${}^T\psi_I = (\Pi I_i)^{-0.5} \qquad (9b)$$

χ_t (and χ_t^v) is the total molecular connectivity index and it has its ψ counterpart in the total molecular pseudoconnectivity index, ${}^T\psi_I$ (and ${}^T\psi_E$). Sums in eqns. (6) and (7), as well as the products Π in eqns. (9), are taken over all vertices of the hydrogen-suppressed chemical graph. Sums in the vertex-connectivity index of eqns. (8) are over all edges of the chemical graph (σ bonds in a molecule). Upon replacing δ with δ^v the subset of valence χ^v indices $\{D^v, {}^0\chi^v, {}^1\chi^v, \chi_t^v\}$ for a hydrogen-suppressed chemical pseudograph is obtained. By replacing I_i with S_i the ψ_E subset $\{{}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\}$ is obtained. Superscripts S and T stand for sum index and total index (we have adopted the name total for this index, following the definition of the χ_t index[12]), while the other sub- and superscripts follow the established denomination for χ indices.[2]

The dual basis indices introduced recently are based on a Boolean-like algorithm used in a rather unconventional way, as can be seen from the following definitions, where subscript d stands for dual and s for soft dual.[25] Actually, these indices can be considered as sorts of contra-indices, where the contra-expression is obtained by interchanging sums and products and zeros and ones (for more details on this topic see ref. 25).

$$^0\chi_d = (-0.5)^N \Pi_i(\delta_i) \tag{10a}$$

$$^0\psi_{Id} = (-0.5)^N \Pi_i(I_i) \tag{10b}$$

$$^1\chi_d = (-0.5)^{(N+\mu-1)} \Pi(\delta_i + \delta_j) \tag{11a}$$

$$\psi_{Id} = (-0.5)^{(N+\mu-1)} \Pi\ (I_i + I_j) \tag{11b}$$

$$^1\chi_s = \Pi(\delta_i + \delta_j)^{-0.5} \tag{12a}$$

$$^1\psi_{Is} = \Pi\ (I_i + I_j)^{-0.5} \tag{12b}$$

If, in these expressions δ is replaced by δ^v and I_i by S_i the corresponding χ^v valence dual and ψ_E dual indices are obtained. The exponent, μ, in eqns. (11) is the cyclomatic number. The cyclomatic number, $\mu = q - N + 1$, of a graph (q = number of edges, N = number of vertices), indicates the number of cycles of a chemical graph and it is equal to the minimum number of edges necessary to be removed in order to convert a (poly)-cyclic graph to an acyclic graph.[5] For acyclic molecules $\mu = 0$, for monocyclic compounds $\mu = 1$, and for bicyclic compounds $\mu = 2$. The total number of basis indices is twenty-eight, this means that the combinatorial space of these basis indices amounts to almost billions of combinations.[22] The easiest way to avoid this huge combinatorial problem is to use the dual basis indices to improve, whenever possible, the model quality of the X, Y and Z terms, giving rise to X', Y' and Z' terms.[25]

A result from the I_S concept[16] is that $\Sigma_i S_i = \Sigma_i I_i$, with the consequence that $^S\psi_I = {}^S\psi_E$, such that the ψ subset will consist of seven indices only. Now, as S_i can be negative, which gives rise to imaginary ψ_E values, it has been necessary to rescale every S_i value of the class of compounds whose atoms show negative S_i to the Si value in SiF_4: $S(Si) = -6.611$. Inevitably, this rescaling procedure invalidates the cited result of the I_S concept, with the consequence that, now, $^S\psi_I \neq {}^S\psi_E$.[23,24]

The statistical performance of the graph-structural molecular connectivity invariant, S, is controlled by a quality factor, $Q = r/s$, and by the Fischer ratio F. Here r is the correlation coefficient and s the standard deviation of the estimates. Parameter Q is only able to compare the descriptive power of different descriptors for the same property (more about Q in the web address given in ref. 6). The F ratio tells us, even if Q improves, which additional descriptor endangers the statistical quality of the combination. For every invariant, X, Y, Z, β_i, and U_0, the fractional utility, $u_i = |c_i/s_i|$, where s_i is the confidence interval of c_i, and the average fractional utility $<u> = \Sigma u_i/(v+1)$ will be given. If the relation is linear, then $<u> = (u_1 + u_0)/2$, where u_1 and u_0 are the utilities of the invariant and of the unitary index, $U_0 \equiv 1$ (the constant parameter) of the linear regression, respectively. The utility statistics check descriptors that give rise to unreliable coefficient values (c_i), whenever they have a high deviation interval (s_i). Recently,[25] the critical importance of the standard deviation of the estimate s has been underlined; thus, it will be advantageous to know how much this statistic is 'squeezed' by the next best descriptor. To achieve this goal the ratio $s_R = s_0/s_i$ (where R means ratio) is here introduced, where s_0 is the s value of the best single-index description and s_i refers to the s values of the next best descriptions. Thus, halving s_i will double s_R, thus allowing a direct measure of the progress of s along a series of sequential descriptions. It should be stressed that, now (i) all statistical parameters will grow with improving model, (ii) every description is under the control of all these statistics, and (iii) an improved Q is not a sufficient sign of an improved model. To avoid bothering the reader with the dimensional problem of the model equation every property P should be read as $P/P°$ where $P°$ is the unitary value of the property, so that this choice allows P to be read as a pure numerical value.[33]

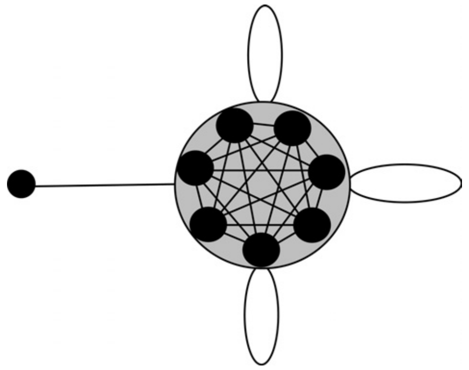

Fig. 2 The hydrogen-suppressed pseudograph-complete graph of CH_3–I. The first vertex at the left mimics the carbon atom. The second vertex mimics the iodine atom. The blow-up of the iodine vertex shows a K_7 complete graph, which encodes the inner-core electrons of I. The loops around this K_7 vertex (pseudograph characteristics) mimic the nonbonding electrons of the valence shell.

III. The complete graphs

In this study we wish to see how the $\{D^v$, $^0\chi^v$, $^1\chi^v$, χ_t^v, $^0\chi_d^v$, $^1\chi_d^v$, $^1\chi_s^v\}$ subset for atoms with $n > 2$ works when (i) $\delta_i^v = (2/n)^2\delta_i^v(\text{ps})$ ('old' homogeneous description) and when (ii) δ_i^v are derived by the aid of an algorithm based on regular complete graphs, K_p ('new' or just heterogeneous description), which are used to encode the inner-core electrons of heteroatoms. A graph G is complete if every pair of its vertices is adjacent. A complete graph of order p is denoted by K_p, $(p - 1 = r)$ and is r-regular, where in general for a graph r denotes its regularity, if it has all vertices with the same degree r (not to be confounded with the correlation coefficient r). In a previous study good results[29] have been obtained with odd complete graphs and with the following algorithm:

$$\delta^v = \delta^v(\text{ps})/[p{\cdot}r + 1] \tag{13}$$

where $p = 1, 3, 5, 7$, that is, $\delta^v(\text{F}) = 7$, $\delta^v(\text{Cl}) = 7/7$, $\delta^v(\text{Br}) = 7/21$, $\delta^v(\text{I}) = 7/43$ (see Figs. 1 and 2) The given algorithm based on the K_1 graphs (K_1 is just a vertex) for any second-row atom allows their graph representation (a vertex) and their δ^v values to be preserved. The product $p{\cdot}r$ in the algorithm involving the complete graph K_p is a well-known parameter in graph theory. In fact, from the '*handshaking theorem*' it equals twice the number of connections.[22,32] For every graph and pseudograph it is possible to write its adjacency A matrix.[2,22] The adjacency matrix for a hydrogen-suppressed chemical pseudograph of a triatomic system that includes the contribution of the odd regular complete graph for the inner-core electrons is given in Fig. 3 together with two applications. Here, $g_{i,j}$ can either be 0 or 1; it is one if vertices i and j are connected, otherwise it is zero (graph characteristics); ps_i is the sum of the self-connections (they count twice) and multiple connections of vertex i (pseudograph characteristics). The factor $(p{\cdot}r + 1)$ encodes the complete graph characteristics and depends on the p value of the complete graph; this renders the adjacency matrix asymmetric as can be seen from the 2×2 and 3×3 A matrixes for hydrogen-suppressed CH_3–Br (K_5 for Br and K_1 for C) and CH_3–S–CH_3 (K_1 for C and K_3 for S) at the bottom of Fig. 3. The term $1/1 = 1$ has been written to allow an easier decoding of the formalism.

Results and discussion

I. Induced dipole moments

A. Alcohols, amines and ethers. In Table 1 are collected the experimental induced dipole moment values, together with the corresponding calculated ones, and the corresponding residual

$$A = (p \cdot r + 1)^{-1}_{Kp} \begin{pmatrix} ps_{1,1} & g_{1,2} & g_{1,3} \\ g_{2,1} & ps_{2,2} & g_{2,3} \\ g_{3,1} & g_{3,2} & ps_{3,3} \end{pmatrix}$$

$$A(C-Br) = \begin{pmatrix} 0 & 1/1 \\ 1/21 & 6/21 \end{pmatrix},\ A(C-S-C) = \begin{pmatrix} 0 & 1/1 & 0 \\ 1/7 & 4/7 & 1/7 \\ 0 & 1/1 & 0 \end{pmatrix}$$

Fig. 3 Top: The adjacency$A$ matrix for a hydrogen-suppressed pseudograph-complete graph of a triatomic system. Bottom: the adjacency matrices for CH_3Br and $S(CH_3)_2$.

modulus $|\Delta\mu| = |\mu(\mathrm{E}) - \mu(\mathrm{C})|$ for these and the following two classes of compounds. For this class of compounds and for the next one the odd complete graph algorithm gives rise to the same δ^{v} values, which can be derived with eqn. (3) and with the $(2/n)^2\delta^{\mathrm{v}}(\mathrm{ps})$ algorithm, as for the second-row atoms the three algorithms give rise to equal values. Thus, there is only one type of description here. The best single-, two-index and three-index descriptions are:

$$\{{}^{\mathrm{S}}\psi_{\mathrm{I}}\}\text{: } Q = 2.037,\ F = 22,\ r = 0.723,$$
$$s_0 = 0.35,\ s_{\mathrm{R}} = 1,\ n = 22$$

$$\{{}^0\chi^{\mathrm{v}}, {}^{\mathrm{S}}\psi_{\mathrm{E}}\}\text{: } Q = 6.001,\ F = 95,\ r = 0.953,\ s_{\mathrm{R}} = 2.2,$$
$$\langle u\rangle = 12,\ \mathbf{u} = (13,\ 14,\ 8.1),\ n = 22$$

$$\{{}^0\chi,\ D^{\mathrm{v}}, {}^0\chi^{\mathrm{v}}\}\text{: } Q = 7.919,\ F = 110,\ r = 0.974,\ s = 0.12,$$
$$s_{\mathrm{R}} = 3.5,\ n = 22,\ \langle u\rangle = 7.1,\ \mathbf{u} = (5.0,\ 6.9,\ 9.2,\ 7.5),$$
$$(\mathbf{C} = 0.84028,\ 0.10678,\ -1.17700,\ 0.99778)$$

Observe the consistent improvement at every statistical level of the second description. From now on the attentive reader should keep an eye to the ${}^0\chi^{\mathrm{v}}$ index. The correlation vector, **C**, of the last description will be used to derive the calculated $\mu(\mathrm{C})$ values and the corresponding residual modulus $|\Delta\mu|$ for this class of compounds (see Table 1). Thanks to the dual index ${}^0\psi_{\mathrm{Ed}}$ an interesting Z'_μ-type term can be detected:

$$Z'_\mu = [Z - 610^{-10} \cdot ({}^0\psi_{\mathrm{Ed}})]\text{: } Q = 7.144,\ F = 269,\ r = 0.965,$$
$$s_{\mathrm{R}} = 2.6,\ n = 22,\ \mathbf{u} = (16, 16)$$

Here, $Z = [X^{0.9} + 0.0002 \cdot Y^{0.9}]^{0.01}$: $Q = 6.78$, $F = 242$, $s_{\mathrm{R}} = 2.5$; $X = [\chi_{\mathrm{t}}^{\mathrm{v}} \cdot {}^1\chi^{\mathrm{v}} - 0.009 \cdot {}^0\chi]^{-0.01}$: $Q = 5.15$, $F = 140$, $s_{\mathrm{R}} = 1.9$; $Y = [{}^{\mathrm{S}}\psi_{\mathrm{I}} - 1.9 \cdot {}^0\psi_{\mathrm{I}}]^{1.8}$: $Q = 3.20$, $F = 54$, $s_{\mathrm{R}} = 1.3$.

If the experimental values for these compounds are regressed *vs.* the MM3(2000) dipole moment calculated values, taken from ref. 30, $r = 0.978$ and $s = 0.11$ are obtained. These results emphasize the good quality of the present study (ref. 30 assigns only the s values and for a slightly different set of compounds).

B. Aldehydes, ketones, acids and esters. It is rather difficult to achieve a satisfactory description for the dipole moment of this class of compounds. The best single-basis index is quite poor and is improved only by a Z'-type term, where the dual index, ${}^0\psi_{\mathrm{Id}}$, introduces an interesting improvement in the Z term:

$$\{{}^{\mathrm{T}}\psi_{\mathrm{I}}\}\text{: } Q = 0.959,\ F = 7.8,\ r = 0.539,$$
$$s_0 = 0.56,\ s_{\mathrm{R}} = 1,\ n = 21$$

$$Z'_\mu = [Z - 12 \cdot {}^0\psi_{\mathrm{Id}}]^{0.01}\text{: } Q = 2.987,\ F = 75,\ r = 0.894,$$
$$s = 0.30,\ s_{\mathrm{R}} = 1.9,\ n = 21,\ \mathbf{u} = (8.7,\ 8.8),$$
$$\mathbf{C} = (-141.83,\ 152.659)$$

Here, $Z = [X^{1.1} + Y]^{1.7}$: $Q = 2.08$, $F = 36$, $s_{\mathrm{R}} = 1.4$; $X = [D^{\mathrm{v}} - 1.4 \cdot D]^{0.7}({}^0\chi)^{1.1}$: $Q = 2.03$, $F = 35$, $s_{\mathrm{R}} = 1.4$; $Y = [({}^{\mathrm{T}}\psi_{\mathrm{I}} \cdot {}^{\mathrm{S}}\psi_{\mathrm{E}}) - 1.5 \cdot ({}^1\psi_{\mathrm{E}})^{0.4}]^{0.5}$: $Q = 1.64$, $F = 23$, $s_{\mathrm{R}} = 1.2$.

The calculated dipole moment values given in Table 1, together with the residual modulus, have been obtained with the given correlation vector, **C**, of the Z'_μ term. The obtained residual values, and the comparison with the MM3(2000) calculated values,[30] underline the non-optimal quality of our description; in fact $r(\mathrm{MM3}) = 0.979$ and $s(\mathrm{MM3}) = 0.13$.

C. Sulfides and phosphines. Due to the sulfur and phosphorus atoms the δ^{v} (K_p), $\delta^{\mathrm{v}}(Z, Z^{\mathrm{v}})$, and $(2/n)^2\delta^{\mathrm{v}}(\mathrm{ps})$ algorithms give rise to different δ^{v} values for sulfides and phosphines, and, consequently, to different $\{\chi^{\mathrm{v}}\}$ values.

i. Heterogeneous description. The heterogeneous description uses $\delta^{\mathrm{v}} = (Z^{\mathrm{v}} - h)/(Z - Z^{\mathrm{v}} - 1)$ [eqn. (3)], where $\delta^{\mathrm{v}}(\text{–SH}) = 0.56$, $\delta^{\mathrm{v}}(\text{–S–}) = 0.67$, $\delta^{\mathrm{v}}(\text{–PH}_2) = 0.33$, $\delta^{\mathrm{v}}(\text{–PH–}) = 0.44$, $\delta^{\mathrm{v}}(\text{–P}<) = 0.56$. The best single-$\beta$-basis and two-$\beta$-basis descriptors are:

$$\{\chi_{\mathrm{t}}^{\mathrm{v}}\}\text{: } Q = 4.113;\ F = 8;\ r = 0.636;\ s_0 = 0.15,\ s_{\mathrm{R}} = 1,\ n = 14$$

$$\{{}^0\chi, {}^0\chi^{\mathrm{v}}\}\text{: } Q = 13.88;\ F = 47;\ r = 0.946;\ s_{\mathrm{R}} = 2.1,$$
$$\langle u\rangle = 13,\ \mathbf{u} = (9.6,\ 9.6,\ 20),\ n = 14$$

The improvement from the single-basis descriptor to the two-basis descriptor is impressive. The ψ indices based on $\delta^{\mathrm{v}} = (2/n)^2\delta^{\mathrm{v}}(\mathrm{ps})$ are not good descriptors in this case. The following X term shows an interesting improvement at every statistical level:

$$X = [({}^0\chi^{\mathrm{v}} - {}^0\chi)^{2.4} + 0.08 \cdot \chi_{\mathrm{t}}^{\mathrm{v}}]/({}^1\chi^{\mathrm{v}})^{0.3}\text{: } Q = 21.02,$$
$$F = 213,\ r = 0.973,\ s_{\mathrm{R}} = 3.3,\ \langle u\rangle = 39,\ \mathbf{u} = (15,\ 64),\ n = 14$$

A Z' term improves the model a bit, thanks to a poor Y term: $Y = [{}^{\mathrm{S}}\psi_{\mathrm{I}} - 1.3\ {}^0\psi_{\mathrm{E}}]^{0.1}$: $Q = 4.053$, $F = 7.9$, $r = 0.631$, $s_{\mathrm{R}} = 0.9$, and a connectivity dual index, ${}^0\chi_{\mathrm{d}}$,

$$Z'_\mu = [Z + 0.07 \cdot {}^0\chi_{\mathrm{d}}]\text{: } Q = 23.54,\ F = 268,\ r = 0.978,$$
$$s_{\mathrm{R}} = 3.6,\ \langle u\rangle = 42,\ \mathbf{u} = (16,\ 67),\ \mathbf{C} = (-1.04727,\ 1.77842)$$

Here, $Z = [X \cdot Y]^{0.9}$: $Q = 21.77$, $F = 229$, $s_{\mathrm{R}} = 3.3$.

ii. Homogeneous description. The homogeneous description uses $\delta^{\mathrm{v}} = (2/n)^2\delta^{\mathrm{v}}(\mathrm{ps})$, where $\delta^{\mathrm{v}}(\text{–SH}) = 2.22$, $\delta^{\mathrm{v}}(\text{–S–}) = 2.67$, $\delta^{\mathrm{v}}(\text{–PH}_2) = 1.33$, $\delta^{\mathrm{v}}(\text{–PH–}) = 1.77$, $\delta^{\mathrm{v}}(\text{–P}<) = 2.22$. This description is deceiving; in fact, the best basis descriptors are ($s_0 = 0.15$): $\{\chi_{\mathrm{t}}^{\mathrm{v}}\}$: $Q = 3.970$, $F = 7.6$, $r = 0.623$, $s_{\mathrm{R}} = 0.9$; and $\{{}^0\chi, D^{\mathrm{v}}\}$: $Q = 8.043$, $F = 16$, $r = 0.860$, $s_{\mathrm{R}} = 1.4$, $n = 14$. While the ψ indices are not good descriptors, the found terms are deceptive:

$$Z_\mu = [X \cdot (Y)^{1.9}]\text{: } Q = 7.592,\ F = 28,\ r = 0.836,$$
$$s = 0.11,\ s_{\mathrm{R}} = 1.4,\ n = 14$$

Here, $X = [{}^0\chi \cdot (\chi_{\mathrm{t}}^{\mathrm{v}})^{1.4}]$: $Q = 6.970$, $F = 23$, $r = 0.813$, $s_{\mathrm{R}} = 1.3$ (see previous Y based on the same δ^{v})

iii. Heterogeneous $\boldsymbol{K_p} - (\boldsymbol{p} - odd)$ *description.* Here $\delta^{\mathrm{v}} = \delta^{\mathrm{v}}(\mathrm{ps})/[p \cdot r + 1]$, where $\delta^{\mathrm{v}}(\mathrm{SH}) = 5/7$, $\delta^{\mathrm{v}}(\mathrm{S}) = 6/7$, $\delta^{\mathrm{v}}(\mathrm{PH}_2) = 3/7$, $\delta^{\mathrm{v}}(\mathrm{PH}) = 4/7$, $\delta^{\mathrm{v}}(\mathrm{P}) = 5/7$. While the single- and the two-basis index combinations are somewhat worse than in the heterogeneous description, for the two-basis index there is a noticeable improvement over the homogeneous description:

$$\{\chi_{\mathrm{t}}^{\mathrm{v}}\}\text{: } Q = 4.025,\ F = 7.8,\ r = 0.628,\ s_{\mathrm{R}} = 0.9,\ n = 14$$

$$\{{}^0\chi,\ D^{\mathrm{v}}\}\text{: } Q = 10.8,\ F = 28,\ r = 0.915,\ s_{\mathrm{R}} = 1.9,\ n = 14$$

The ψ indices are, even here, not better descriptors. Concerning the X, Z and Z' terms there is an interesting improvement over the heterogeneous description (note the importance of the ${}^0\chi^{\mathrm{v}}$ index):

$$X = [({}^0\chi^{\mathrm{v}} - {}^0\chi)^{2.4} + 0.06 \cdot \chi_{\mathrm{t}}^{\mathrm{v}}]/[({}^1\chi^{\mathrm{v}})^{0.5} - 0.1{}^0\chi^{\mathrm{v}}]^{1.2}\text{: } Q = 23.72,$$
$$F = 272,\ r = 0.979,\ s_{\mathrm{R}} = 3.8,\ \langle u\rangle = 45,\ \mathbf{u} = (16,\ 73),\ n = 14$$

Table 1 Experimental (E) and calculated (C) dipole moments, μ, and residual modulus, $|\Delta\mu| = |\mu(\text{E}) - \mu(\text{C})|$, for (A) alcohols, amines, ethers, (B) aldehydes, ketones, acids, esters, and (C) sulfides and phosphines.[a]

Compound[b]	μ(E)/D	μ(C)/D	$\|\Delta\mu\|$/D	Compound	μ(E)/D	μ(C)/D	$\|\Delta\mu\|$/D
Methanol	1.700	1.616	0.084	3-Pentanone	2.720	2.623	0.097
Ethanol	1.680	1.591	0.088	Cyclopentanone	3.250	2.917	0.333
Propanol	1.550	1.567	0.017	Cyclohexanone	3.250	3.563	0.313
2-Propanol	1.580	1.512	0.068	Formic acid (*t*)	3.790	3.224	0.566
tert-Butanol	1.670	1.414	0.256	Acetic acid	1.700	1.737	0.037
1,2-Ethanediol (*g*)	2.410	2.644	0.234	Propionic acid	1.550	2.233	0.683
1,2-Propanediol (*g*)	2.568	2.565	0.003	2-Methylpropionic acid	1.790	1.575	0.215
Dimethyl ether	1.310	1.292	0.018	2,2-Dimethylpropionic acid	1.700	1.755	0.055
Ethyl methyl ether	1.174	1.268	0.094	Methyl formate	1.770	1.842	0.072
Diethyl ether	1.061	1.243	0.182	Ethyl formate (*t*)	1.980	2.096	0.116
Methyl propyl ether (*t*-*t*)	1.107	1.243	0.136	Methyl acetate	1.706	2.062	0.356
Diisopropyl ether	1.130	1.084	0.046	Ethyl acetate	1.780	1.600	0.180
THF	1.750	1.654	0.096	Methyl propionate	1.750	1.600	0.150
THP	1.740	1.629	0.111	Ethyl propionate	1.810	1.798	0.012
2-Methoxyethanol	2.360	2.321	0.039	Methane thiol	1.520	1.510	0.010
Methylamine	1.296	1.249	0.047	Ethane thiol (*t*)	1.580	1.603	0.023
Ethylamine	1.220	1.224	0.004	1-Propane thiol (*t*)	1.598	1.622	0.024
Dimethylamine	1.010	0.971	0.039	Dimethyl sulfide	1.500	1.487	0.013
n-Propylamine	1.180	1.200	0.020	Ethyl methyl sulfide (*t*)	1.560	1.502	0.058
i-Propylamine	1.190	1.145	0.045	2-Methyl-2-propane thiol	1.660	1.630	0.029
Trimethylamine	0.612	0.801	0.189	Diethyl sulfide (*t*)	1.520	1.557	0.037
N-Methylaminoethane	0.880	0.946	0.066	Methyl phosphine	1.100	1.112	0.012
Formaldehyde	2.360	2.659	0.299	Ethyl phosphine (*t*)	1.226	1.205	0.021
Acetaldehyde	2.730	3.011	0.281	Dimethyl phosphine	1.230	1.187	0.043
Propanal	2.520	2.455	0.075	*i*-Propyl phosphine (*t*)	1.230	1.196	0.033
Butanal	2.740	2.730	0.010	Trimethyl phosphine	1.192	1.2647	0.072
2,2-Dimethylpropanal	2.660	2.640	0.020	*tert*-Butyl phosphine	1.170	1.211	0.041
Acetone	2.930	2.412	0.518	Ethyldimethyl phosphine	1.310	1.310	0.0001
2-Butanone	2.775	2.736	0.039				

[a] Class A includes methanol to *N*-Me-aminoEt; class B includes formaldehyde to Et-Propionate; class C includes methane thiol to Et-diMe phosphine. [b] g = gauche; *t* = *trans*; *t*-*t* = *trans*–*trans*, THF = tetrahydrofuran; THP = tetrahydropyran.

$$Z'_\mu = [Z + 0.03\cdot{}^0\chi_d]:\ Q = 25.76,\ F = 321,\ r = 0.982,\ n = 14,\ s_R = 3.8,\ \langle u\rangle = 48,\ \mathbf{u} = (18, 79),\ \mathbf{C} = (-1.7911, 1.73691)$$

Here, $Z = [X\cdot Y]$: $Q = 24.21$, $F = 283$, $s_R = 3.8$. The Y term is the same as for the heterogeneous model. The model for sulfides and phosphines, whose δ^v values have been derived with the $K_p - (p - \text{odd})$ algorithm for sulfur and phosphorus atoms, seems optimal. With the vector **C** of the Z'_μ term have been obtained the calculated μ(C) and the residual modulus values given in Table 1 and shown in Fig. 4 for the first and third classes of these compounds (here are shown the algebraic $\Delta\mu$ values). MM3(2000) results[30] have $r = 0.843$ and $s = 0.11$.

II. Molecular polarizability

The polarizability measures the distortion of a molecule in an electric field, E, thus non-spherical molecules have anisotropic polarizabilities, α_i, and are rotationally Raman active. Other interesting characteristics of polarizability have recently been studied.[34–38] The experimental (E) polarizability of fifty-four organic compounds, including many halogenated compounds, as well as the molecular principal polarizabilities, α_1, α_2 and α_3, of forty organic compounds are given in Table 2. Some $\langle\alpha(\text{E})\rangle$ values, whenever the α_i(E) values are absent, are the result of quantum mechanical calculations. In this table are also shown (i) the fifty-four calculated polarizability values, $\langle\alpha(\text{C})\rangle$, derived from the direct description of $\langle\alpha(\text{E})\rangle$; (ii) the calculated α_i(C) values, which include fourteen inferred α_i(C) values, derived from the direct description of the forty α_i(E) values; and (iii) the average values $A[\alpha_i(\text{C})] = [\alpha_1(\text{C}) + \alpha_2(\text{C}) + \alpha_3(\text{C})]/3$ (Ave column in Table 2). This last set of values has fourteen values that arise from the fourteen inferred calculated α_i(C) values belonging to unknown α_i(E) values. The $A[\alpha_i(\text{C})]$ values include, thus, fourteen inferred values. The $A[\alpha_i(\text{C})]$ values can then be considered the result of a leave-14-out method. Further, the model of the single α_i(E) will not be optimized, as this will be done with the optimal descriptor for $\langle\alpha(\text{E})\rangle$.

A. Heterogeneous description.

$\delta^v = (Z^v - h)/(Z - Z^v - 1)$ for $\{\chi^v\}$ and $\{\psi\} = \text{f}(\delta^v = (2/n)^2\delta^v(\text{ps}))$. The best index for the different molecular polarizabilities is ${}^0\chi^v$ with:

P	n	$\{\beta\}$	Q	F	r	s_0	$\langle u\rangle$	**u**
$\langle\alpha\rangle$	54	$\{{}^0\chi^v\}$	0.829	545	0.955	1.15	12	(23, 0.9)
α_1	40	$\{{}^0\chi^v\}$	0.534	200	0.917	1.71	8.3	(14,2.4)
α_2	40	$\{{}^0\chi^v\}$	0.803	498	0.964	1.20	12	(22, 1.2)
α_3	40	$\{{}^0\chi^v\}$	0.800	498	0.956	1.20	11	(20, 1.7)

The low utility value (u_0) of the unitary index (U_0) of the constant parameter of the regression is due to the fact that the value of this regression parameter is nearly zero, with the consequence that even small errors result in a low utility. The best linear combination for $\langle\alpha\rangle$ is a four-index LCBI:

P	n	$\{\beta_1, \beta_2, \beta_3, \beta_4\}$	Q	F	r	s_R	$\langle u\rangle$	**U**
$\langle\alpha\rangle$	54	$\{{}^0\chi^v, {}^S\psi_I, {}^0\psi_I, {}^S\psi_E\}$	1.694	569	0.989	2.0	7.8	(25, 5.9, 2.0, 5.0, 0.9)
α_1	40	$\{{}^0\chi^v, {}^S\psi_I, {}^0\psi_I, {}^S\psi_E\}$	0.739	96	0.957	1.3	3.9	(8.7, 3.5, 1.5, 3.0, 2.9)
α_2	40	$\{{}^0\chi^v, {}^S\psi_I, {}^0\psi_I, {}^S\psi_E\}$	1.316	334	0.987	1.6	5.8	(18, 4.2, 1.5, 3.6, 1.6)
α_3	40	$\{{}^0\chi^v, {}^S\psi_I, {}^0\psi_I, {}^S\psi_E\}$	0.973	151	0.972	1.2	3.2	(13, 0.2, 1.4, 0.12, 1.8)

The following Z term, where $X = [2{}^0\chi^v + {}^1\chi]^{1.1}$ (for $\langle\alpha\rangle$: $Q = 1.221$, $F = 1183$, $r = 0.979$, $s_R = 1.3$, $\langle u\rangle = 18$), and

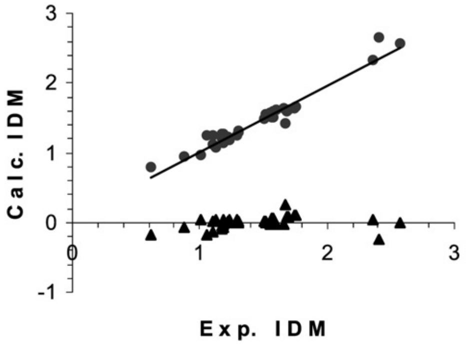

Fig. 4 Plot of the calculated *versus* the experimental induced dipole moments in Debye units, and plot of the algebraic residual values for compound classes A and C of Table 1.

$Y = [{}^0\psi_I - 0.9{\cdot}{}^0\psi_E]^{0.6}$ (for $\langle\alpha\rangle$: $Q = 0.359$, $F = 102$, $r = 0.814$, $s_R = 0.5$), shows an improvement in the model quality:

$$Z = [1.3{\cdot}X + 5.2{\cdot}Y - 0.6{\cdot}{}^0\chi - 3.1{\cdot}{}^1\psi_E]^{1.1}$$

P	n	Q	F	r	s_R	$\langle u\rangle$	**U**
$\langle\alpha\rangle$	54	1.430	1621	0.984	1.7	20	(40, 0.5)
α_1	40	0.684	327	0.947	1.2	10	(18, 2.8)
α_2	40	1.153	1026	0.982	1.4	17	(32, 1.7)
α_3	40	1.011	653	0.972	1.3	14	(26, 2.1)

The model improves further with Z' term, thanks to two dual indices,

$$Z' = [Z + 0.4({}^1\chi_s)^{3.7} + 0.001{\cdot}{}^1\psi_{Id}]$$

P	n	Q	F	r	s_R	$\langle u\rangle$	**U**	**C**
$\langle\alpha\rangle$	54	1.675	2225	0.989	1.9	24	(47, 0.34)	(0.33596, 0.063998)
α_1	40	0.707	350	0.950	1.3	11	(19, 2.8)	(0.32843, 1.25587)
α_2	40	1.369	1447	0.987	1.7	20	(38, 2.2)	(0.35851, –0.52276)
α_3	40	0.979	612	0.970	1.2	13	(25, 2.0)	(0.32046, –0.6798)

From the c_0 value of the correlation vector (–0.064, rounding) for $\langle\alpha\rangle$ it is evident that this value (i) is practically zero and (ii) that even a small deviation around it gives rise to a very large error, as can be seen from the small utility u_0 value (0.34). Nearly the same is true for α_i. The only meaningful utility value is the impressive u_1 utility value. Furthermore, the two calculated parameters, $\langle\alpha(C)\rangle$ and $A[\alpha_i(C)]$, even if similar in their statistical performance, differ, nonetheless, by a constant: $\langle\alpha(C)\rangle = A[\alpha_i(C)] + 0.88$. MM3(2000) calculations show a somewhat poorer quality: $r = 0.982$, $s = 0.75$, and $s_R = 1.5$, assuming $s_0 = 1.15$.

B. Homogeneous description. $\delta^v = (2/n)^2\ \delta^v(\text{ps})$ for $\{\chi^v\}$ and $\{\psi\}$. This is a deceiving case. Note that the best descriptors for $\langle\alpha\rangle$ are again $\{{}^0\chi^v\}$, $\{{}^0\chi^v, {}^0\psi_I\}$, and $\{{}^0\chi^v, {}^S\psi_I, {}^0\psi_I, {}^S\psi_E\}$. No good terms could be detected here.

C. Heterogeneous K$_p$-(p-odd) description. $\delta^v = \delta^v(\text{ps})/[p{\cdot}r + 1]$, $p = 1, 3, 5, 7$ for $\{\chi^v\}$ only. Let us first emphasize that the model due to a χ combination with $\chi^v = f[\delta^v = (Z^v - h)/(Z - Z^v - 1)]$ is:

$$\{{}^1\chi,\ D^v,\ {}^0\chi^v\}\text{: } Q = 1.472,\ \ F = 573,\ \ r = 0.986,\ \ s_R = 1.7,$$
$$\mathbf{u} = (8.5,\ 5.0,\ 19,\ 2.4),\ \langle u\rangle = 8.9,\ n = 54$$

The following results found with the odd complete graph algorithm are somewhat unexpected: index ${}^0\chi^v$ becomes the dominant index and the valence molecular connectivity indices become the overall dominant indices, excluding the pseudoindices from the model. The best linear combination is the four-χ-index combination The importance of the ${}^0\chi^v$ index allows to use the forward combination procedure[22] to search for the optimal combination ($s_0 = 1.15$).

P	n	$\{\beta\}$	Q	F	r	s_R	$\langle u\rangle$	**u**
$\langle\alpha\rangle$	54	$\{{}^0\chi^v\}$	1.045	867	0.971	1.3	15	(29, 0.4)
α_1	40	$\{{}^0\chi^v\}$	0.576	232	0.927	1.1	8.8	(15, 2.4)
α_2	40	$\{{}^0\chi^v\}$	0.888	609	0.970	1.1	13	(25, 1.4)
α_3	40	$\{{}^0\chi^v\}$	0.996	633	0.971	1.2	14	(25, 2.5)

P	n	$\{\beta_1, \beta_2, \beta_3, \beta_4\}$	Q	F	r	s_R	$\langle u\rangle$	**U**
$\langle\alpha\rangle$	54	$\{{}^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	2.086	863	0.993	2.5	6.5	(12, 9.5, 7.7, 2.7, 3.4)
α_1	40	$\{{}^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	0.799	112	0.963	1.5	3.1	(2.3, 5.3, 4.5, 2.4, 1.0)
α_2	40	$\{{}^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	1.414	386	0.989	1.8	5.2	(7.4, 6.5, 5.2, 2.3, 4.7)
α_3	40	$\{{}^0\chi^v, {}^1\chi, D^v, \chi_t^v\}$	1.093	191	0.978	1.4	3.0	(8.1, 1.6, 1.6, 0.7, 3.3)

Now, let us see how the Z and Z' terms behave:

$$Z = [2.5{\cdot}X + 6{\cdot}Y - 0.6{\cdot}{}^0\chi - 3.3{\cdot}{}^1\psi_E]^{1.1}$$

P	n	Q	F	r	s_R	$\langle u\rangle$	**U**
$\langle\alpha\rangle$	54	1.569	1952	0.987	1.8	23	(44, 0.9)
α_1	40	0.671	315	0.945	1.2	10	(18, 2.2)
α_2	40	1.120	968	0.981	1.4	17	(31, 2.4)
α_3	40	1.128	813	0.977	1.4	16	(29, 3.2)

The term $X = [3{\cdot}{}^0\chi^v + {}^1\chi]$, for $\langle\alpha\rangle$ has: $Q = 1.414$, $F = 1587$, $r = 0.984$, $s_R = 1.6$, $\langle u\rangle = 21$; the Y term does not change, as $\psi = \psi[\delta^v = (2/n)^2\delta^v(\text{ps})]$.

For

$$Z' = [Z + 0.7({}^1\chi_s)^{3.7} + 0.002{\cdot}{}^1\psi_{Id}]$$

P	n	Q	F	r	s_R	$\langle u\rangle$	**U**	**C**
$\langle\alpha\rangle$	54	1.738	2397	0.989	2.0	25	(49, 1.1)	(0.16733, –0.193)
α_1	40	0.680	324	0.946	1.2	10	(18, 2.2)	(0.16225, 1.06788)
α_2	40	1.228	1165	0.984	1.5	18	(38, 2.2)	(0.17727, –0.73533)
α_3	40	1.040	690	0.974	1.3	15	(26, 2.9)	(0.15948, –0.91766)

The improvement brought about by the K_p conjecture is evident. With the correlation vectors, $\mathbf{C} = (c_1,\ c_0)$ of vector $\mathbf{S} = (Z',\ U_0)$ the different calculated values for the different types of polarizabilities have been obtained; they are shown in Table 2. A less evident but important result is that, now, $\langle\alpha(C)\rangle$ values overlap with $A[\alpha_i(C)]$ values, that is, the calculated $\alpha_i(C)$ values, which include inferred $\alpha_i(C)$ values calculated with the aid of the correlation vectors, give rise to $A[\alpha_i(C)]$ values (Ave column in Table 2) that have the same quality of the $\langle\alpha(C)\rangle$ values, with nearly no shifting between them: $\langle\alpha(C)\rangle = 1.006{\cdot}A[\alpha_i(C)] + 0.003 \cong A[\alpha_i(C)]$. In Fig. 5 are plotted the $\langle\alpha(C)\rangle$ values *versus* the experimental $\langle\alpha(E)\rangle$ ones. Here the calculated values have been obtained with the correlation vector of the four-index combination, $\chi = ({}^0\chi^v,\ {}^1\chi,\ D^v,\ \chi_t^v,\ U_0)$, that is, $\mathbf{C} = (1.50028,\ 2.47483,\ -0.13929,\ 0.23527,\ -0.55904)$. This combination shows optimal r and s values for this property. The impressive figure confirms the good performance of the odd complete, K_p, graphs. The performance of the MM3(2000) calculations is: $r = 0.982$, $s = 0.75$, and $s_R = 1.5$.

To avoid unreliable linear combinations due to collinearity among the basis indices, while maintaining the descriptive quality, it is advantageous to orthogonalize the corresponding basis indices. To obtain the correlation coefficients of the correlation vector, $\mathbf{C}(\Omega)$, of the orthogonal descriptors there is no need to orthogonalize the basis indices, as they can be obtained

Table 2 Experimental (E) and computed (C) molecular polarizabilities of organic compounds $\langle\alpha\rangle/\text{Å}^3$; $\langle\alpha_{\text{loo}}\rangle$ is the predicted value based on the leave-one-out method[a]

Compound	$\langle\alpha(\text{E})\rangle$	$\langle\alpha(\text{C})\rangle$	$\langle\alpha_{\text{loo}}\rangle$	Ave.[b]	$\alpha_1(\text{E})$	$\alpha_2(\text{E})$	$\alpha_3(\text{E})$	$\alpha_1(\text{C})$	$\alpha_2(\text{C})$	$\alpha_3(\text{C})$
Ethane	4.48	4.65	4.65	4.61	5.49	3.98	3.98	5.76	4.39	3.69
Propane	6.38	6.46	6.46	6.42	7.66	5.74	5.74	7.52	6.32	5.43
Neopentane	10.20	10.54	10.55	10.48	10.20	10.20	10.20	11.48	10.64	9.31
Cyclopropane	5.50	5.53	5.73	5.50	5.74	5.74	5.04	6.62	5.33	4.54
Cyclopentane	9.15	9.22	9.22	9.16	9.68	9.17	8.40	10.20	9.24	8.05
Cyclohexane	11.00	11.08	11.09	11.01	11.81	11.81	9.28	12.00	11.21	9.83
Ethylene	4.12	3.34	3.30	3.32	4.82	3.71	3.25	4.49	3.01	2.45
Propene	6.26	5.48	5.46	5.44				6.57	5.28	4.49
2-Methylpropene	8.29	7.60	7.59	7.55				8.62	7.52	6.51
trans-2-Butene	8.49	8.47	8.47	8.41				9.47	8.44	7.36
Cyclohexene	10.70	10.47	10.47	10.40				11.41	10.56	9.25
Butadiene	7.87	6.37	6.34	6.33	11.93	6.14	5.54	7.44	6.22	5.34
Benzene	9.92	9.19	9.17	9.13	11.20	11.20	7.36	10.16	9.20	8.02
Toluene	12.30	11.30	11.27	11.23				12.21	11.44	10.04
Hexamethylbenzene	22.63	22.77	22.84	22.63	22.63	22.63	22.63	23.33	23.59	20.96
Acetylene	3.50	2.73	2.69	2.71	4.79	2.85	2.85	3.91	2.36	1.87
Propyne	4.68	4.96	4.97	4.93	6.14	3.94	3.94	6.07	4.73	4.00
$C(CCH)_4$	12.19	12.66	12.69	12.58	12.19	12.19	12.19	13.53	12.88	11.33
Allene	5.00	4.63	4.62	4.60	8.97	4.43	4.43	5.75	4.38	3.86
Methanol	3.32	3.37	3.37	3.35	4.09	3.23	2.65	4.52	3.04	2.48
Ethanol	5.11	5.18	5.19	5.15	5.76	4.98	4.50	6.28	4.96	4.21
2-Propanol	6.97	7.22	7.23	7.17				8.26	7.12	6.15
Cyclohexanol	11.56	11.80	11.82	11.73				12.70	11.97	10.52
Dimethylether	5.24	5.70	5.71	5.66	6.38	4.94	4.39	6.78	5.51	4.70
p-Dioxane	8.60	9.60	9.63	9.54	9.40	9.40	7.00	10.57	9.64	8.42
Methylamine	3.59	3.69	3.69	3.66	3.94	3.40	3.38	4.83	3.37	2.78
Formaldehyde	2.45	2.63	2.63	2.61	2.76	1.70	1.83	3.80	2.25	1.77
Acetaldehyde	4.59	4.76	4.77	4.73				5.87	4.52	3.81
Acetone	6.39	7.01	7.03	6.97	7.37	7.37	4.42	8.05	6.90	5.95
Fluoromethane	2.62	3.19	3.22	3.17	3.18	2.34	2.34	4.35	2.85	2.31
Trifluoromethane	2.81	3.94	3.98	3.91	2.87	2.87	2.69	5.07	3.64	3.02
Tetrafluoromethane	2.92	4.11	4.16	4.08	2.92	2.92	2.92	5.24	3.82	3.18
Chloromethane	4.55	4.46	4.46	4.43	5.68	3.99	3.98	5.58	4.20	3.52
Dichloromethane	6.82	6.14	6.13	6.10	8.81	6.30	5.36	7.21	5.98	5.12
Trichloromethane	8.53	8.00	7.99	7.95	9.42	9.42	6.74	9.02	7.95	6.89
Tetrachloromethane	10.51	9.94	9.93	9.88	10.51	10.51	10.51	10.89	10.00	8.74
Bromomethane	5.61	5.96	5.97	5.92	6.91	4.96	4.96	7.03	5.78	4.95
Dibromomethane	8.68	9.21	9.23	9.15				10.19	9.23	8.05
Tribromomethane	11.84	12.73	12.78	12.65	13.00	13.00	9.53	13.60	12.95	11.40
Iodomethane	7.59	8.01	8.02	7.96	9.02	6.87	6.87	9.02	7.96	6.90
Diiodomethane	12.90	12.36	12.33	12.35				13.31	12.64	11.11
Triiodomethane	18.04	17.65	17.59	17.55	18.69	18.69	16.74	18.37	18.17	16.09
$CH_2{=}CCl_2$	7.83	7.29	7.29	7.24	8.96	8.79	5.75	8.32	7.19	6.21
cis-CHCl=CHCl	7.78	7.38	7.38	7.34	9.46	7.80	6.08	8.42	7.29	6.30
Disilane	11.10	11.42	11.44	11.35				12.33	11.57	10.15
Formamide	4.08	3.76	3.75	3.74				4.91	3.46	2.85
Acetamide	5.67	5.89	5.90	5.86				6.97	5.71	4.88
Acetonitrile	4.48	4.67	4.67	4.63	5.74	3.85	3.85	5.78	4.41	3.71
Propionitrile	6.24	6.53	6.54	6.49				7.59	6.39	5.50
tert-Butylcyanide	9.59	10.62	10.66	10.57	10.71	9.03	9.03	11.55	10.72	9.39
Benzylcyanide	11.97	11.57	11.56	11.50	16.16	11.60	8.15	12.47	11.73	10.29
Trichloroacetonitrile	10.42	10.13	10.12	10.07	10.70	10.29	10.29	11.08	10.20	8.92
Pyridine	9.92	10.07	10.08	10.01	10.72	10.43	6.45	11.02	10.14	8.87
Thiophene	9.00	8.44	8.43	8.34	10.15	10.14	6.70	9.44	8.41	7.31

[a] α_1, α_2 and α_3 are the principal molecular polarizabilities; $\langle\alpha(\text{E})\rangle = \Sigma_i\alpha_i(\text{E})/3$; $\langle\alpha(\text{E})\rangle$ values given without $\alpha_i(\text{E})$ values were computed with quantum methods. [b] Ave. = $A[\alpha_i(\text{C})] = \Sigma_i\alpha_i(\text{C})/3$ (see text).

by the aid of sequential stepwise regressions.[22,39,40] The orthogonal correlation vector for $\mathbf{S} = ({}^0\chi^{\text{v}}, {}^1\chi, D^{\text{v}}, \chi_{\text{t}}^{\text{v}}, U_0) \rightarrow \mathbf{S}(\Omega) = ({}^1\Omega, {}^2\Omega, {}^3\Omega, {}^4\Omega, U_0)$ is $\mathbf{C}(\Omega) = (2.36719, 0.83838, -0.12524, 0.23527, 0.11242)$.

The attentive reader should not miss the importance of the fact that we have not optimized the model of the single $\alpha_1(\text{E})$, $\alpha_2(\text{E})$, and $\alpha_3(\text{E})$ properties. These three properties have just been modeled with the best descriptor for the $\langle\alpha(\text{E})\rangle$ property; nevertheless, their description is very satisfactory. Practically, even if not theoretically, $\alpha_1(\text{E})$, $\alpha_2(\text{E})$, and $\alpha_3(\text{E})$ have been externally validated. The leave-one-out calculated values for the $\langle\alpha(\text{E})\rangle$ property are shown in Table 2, throughout the fourth column, $\langle\alpha_{\text{loo}}\rangle$. These values underline the good quality of the model; in fact the standard error of estimates due to this leave-one method is $s_{\text{loo}} = 0.58$ (average error), and consequently: $s_{\text{R}} = s_0/s_{\text{loo}} = 2.0$ ($s_0 = 1.15$).

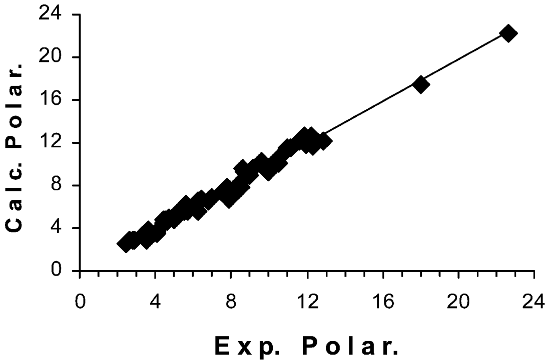

Fig. 5 Plot of the calculated *versus* the experimental polarizabilities in units of Å^3 for the fifty-four organic compounds in Table 2.

Conclusions

Four points can be emphasized from this study: (i) the quality of the achieved descriptions, (ii) the importance of the descriptions that are based on the odd complete graphs for the inner-core electrons, (iii) the improving characteristics of the dual indices, and (iv) the descriptive quality of the $^0\chi^v$ index. While the model the induced dipole moments can be considered reasonable, the model of the molecular polarizabilities is very good. The model of sulfides and phosphines just confirms the ability of odd complete graphs to encode the inner-core electrons for the higher-row ($n > 2$) atoms. The homogeneous model, which uses χ^v indices based on the δ^v for the I- and S-state indices, gives rise to deceiving results. The dual indices confirm their ability to produce improved terms, whenever they can be found. The presence in many descriptions of the atom-based valence index $^0\chi^v$ emphasizes the importance of the pseudograph representation of a molecule and the importance of the odd complete graphs throughout the model of the molecular polarizability. The $^0\chi^v$ index correlates closely with the number and kinds of atoms in a molecule. The importance of this index throughout the descriptions of both properties is not a random result; induced dipole moments and polarizabilities are, in fact, connected.[41] For those who are interested in deepening the physical meaning of the molecular connectivity indices perusal of refs. 42–44 is mandatory.

In model studies it is not always easy to derive a physical visualization for the found algorithms. Regarding the topic of 'physical visualization' an interesting work has underlined[45] that an easy physical visualization is not only difficult with the more accurate quantum mechanical calculations but also nearly impossible with some pH calculations. This seems to hold for our terms also. Property calculations with these terms seem to resemble PVT calculations of fluid properties more than resembling quantum chemistry calculations.[46,47] Empirical equations of state are normally used for prediction and correlation of the volumetric properties of fluids in any density region. Scores of such equations have been proposed and they range from the relatively simple van der Waals equation to complex equations suitable only for computerized calculations and involving as many as twenty or more parameters.

Acknowledgements

I would like to remember an interesting discussion with Prof. Milan Randić of Drake University, at the last Pacifichem 2000, about the possibility of defining a statistical parameter based on the standard deviation of the estimates. I would like to thank also Prof. R. Bruce King of the University of Georgia, Athens, for a helpful suggestion about complete graphs. I wish also to thank the referees for their valuable suggestions.

References

1 M. Randić, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
2 L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York, 1986.
3 *Topological Indices and Related Descriptors in QSAR and QSPR*, eds. J. Devillers and A. T. Balaban, Gordon and Breach, Amsterdam, 1999.
4 G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h and R. Buelow, *Nature Biotech.*, 1998, **16**, 748.
5 M. Reinhard and A. Drefahl, *Handbook for Estimating Physicochemical Properties of Organic Compounds*, Wiley, New York, 1999.
6 R. Todeschini and V. Consonni, *The Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000. For a discussion about *Q* see: http://www.disat.unimib.it/chm/CHMnews.htm.
7 *QSPR/QSAR Studies by Molecular Descriptors*, ed. M. V. Diudea, Nova Science Pub. Inc., New York, 2000.
8 N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, 1992.
9 O. Temkin, A. V. Zeigarnik and D. Bonchev, *Chemical Reaction Networks*, CRC Press, New York, 1996.
10 M. Randić, Topological Indices, in *The Encyclopedia of Computational Chemistry*, ed. P. V. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. R. Scheiner, Wiley, New York, 1998, pp. 3018–3032.
11 *Topology in Chemistry*, ed. R. B. King and D. Rouvray, Horwood, New York, 2001.
12 D. E. Needham, I. C. Wei and P. G. Seybold, *J. Am. Chem. Soc.*, 1988, **110**, 4186.
13 D. Rouvray, *J. Mol. Struct. (THEOCHEM)*, 1989, **185**, 187.
14 J. Galvez, R. Garcia-Domenech, V. De Julian-Ortiz and R. Soler, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1198; J. Galvez, R. Garcia-Domenech, V. De Julian-Ortiz and R. Soler, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 272.
15 M. Kuanar and B. K. Mishra, *Bull. Chem. Soc. Jpn.*, 1998, **71**, 191.
16 L. B. Kier and L. H. Hall, *Molecular Structure Description. The Electrotopological State*, Academic Press, New York, 1999.
17 I. Gutman and E. Estrada, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 541.
18 S. C. Basak, S. Nikolić, N. Trinajstić, D. Amić and D. Beslo, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 927.
19 E. Estrada and L. Rodriguez, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1037.
20 S. Nikolić, N. Trianjstić and S. Ivanis, *Croat. Chim. Acta*, 1999, **72**, 875.
21 L. Pogliani, *J. Phys. Chem. A*, 1999, **103**, 1598.
22 L. Pogliani, *Chem. Rev.*, 2000, **100**, 3827.
23 L. Pogliani, *J. Phys. Chem. A*, 2000, **104**, 9029.
24 L. Pogliani, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 836.
25 L. Pogliani, *J. Mol. Struct. (THEOCHEM)*, 2002, **581**, 87.
26 M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 614.
27 M. Randić, D. Mills and S. C. Basak, *Int. J. Quantum Chem.*, 2000, **80**, 1199.
28 M. Randić and M. Pompe, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 575; M. Randić and M. Pompe, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 581.
29 L. Pogliani, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1028.
30 B. Ma, J. H. Lii and N. L. Allinger, *J. Comput. Chem.*, 2000, **21**, 813.
31 K. H. Rosen, *Discrete Mathematics and its Applications*, McGraw-Hill, New York, 1995.
32 E. Estrada and E. Uriarte, *Curr. Med. Chem.*, 2001, **8**, 1537.
33 M. N. Berberan-Santos and L. Pogliani, *J. Math. Chem.*, 1999, **26**, 255.
34 P. G. Chattaraj, P. Fuentealba and A. Toro-Labbe, *J. Phys. Chem. A*, 1999, **103**, 9307.
35 P. G. Chattaraj and S. Sengupta, *J. Phys. Chem.*, 1996, **100**, 16 126.
36 T. K. Ghanty and S. K. Ghosh, *J. Phys. Chem.*, 1996, **100**, 12 295.
37 P. G. Chattaraj and A. Poddar, *J. Phys. Chem. A*, 1998, **102**, 9944 1999, **103**, 1274.
38 R. G. Parr and R. G. Pearson, *J. Am. Chem. Soc.*, 1983, **105**, 7512.
39 M. Randić, *New J. Chem.*, 1991, **15**, 517.
40 M. Randić, *Croat. Chem. Acta*, 1991, **64**, 43.
41 P. W. Atkins, *Physical Chemistry*, Oxford University Press, Oxford, 4th edn., 1990, ch. 22, p. 190.

42 L. B. Kier and L. H. Hall, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 792.
43 E. Estrada, *J. Phys. Chem. A*, 2002, **106**, 9085.
44 J. Galvez, *J. Mol. Struct. (THEOCHEM)*, 1998, **429**, 255.
45 H. N. Po and N. M. Senozan, *J. Chem. Educ.*, 2001, **78**, 1499.
46 K. S. Pitzer and L. Brewer, *Thermodynamics*, McGraw-Hill, New York, 1961 (revision of Lewis and Randall).
47 M. N. Berberan-Santos, E. N. Bodunov and L. Pogliani, *Am. J. Phys.*, 2002, **70**, 438.