Core electrons and hydrogen atoms in chemical graph theory

Lionello Pogliani

Dipartimento di Chimica, Università della Calabria, via P. Bucci, Cubo 14 C, 87036 Rende, Italy
E-mail: lionp@unical.it

General and complete graphs have recently been used to free chemical graph theory, and especially molecular connectivity theory, from spurious concepts, which belonged to quantum chemistry with no direct counterpart in graph theory. Both types of graph concepts allow the encoding of multiple bonds, non-bonding electrons, and core electrons. Furthermore, they allow the encoding of the bonded hydrogen atoms, which are normally suppressed in chemical graphs. This suppression could sometimes have nasty consequences, like the impossibility to differentiate between compounds, whose hydrogen-suppressed chemical graphs are completely equivalent, like for the CH_2F_2 and BHF_2 compounds. At the computational level the new graph concepts do not introduce any dramatic changes relatively to previous QSPR/QSAR studies. These concepts can nevertheless help in encoding the many electronic features of a molecule, achieving, as a bonus, an improved quality of the modeled properties, as it is here exemplified with a set of properties of different classes of compounds.

KEY WORDS: general graphs, complete graphs, hydrogen atoms, molecular connectivity theory

1. Introduction

Many QSPR/QSAR model studies are based on a chemical counterpart of Graph theory, which is a branch of mathematics that has to do with topology and combinatorics [1, 2], and which deals with the way objects are connected. Graph theory is centered on the concept of graph. A *graph G* can informally be defined as a set of *V* vertices (or points) with a set of E edges (also connections, or lines) that connect these vertices, that is, $G = (V, E)$. The degree of a vertex in a graph is the number of edges that occur with it. A *chemical graph* (also, *molecular graph, structural graphs, constitutional graphs*) is a graph where atoms and bonds are represented by vertices and edges, respectively [3–13]. The degree of a vertex in a chemical graph is normally called its *valence*. Double bonds and/or lone-pair electrons cannot be fitted by a graph. This task is accomplished by pseudographs, also named general graphs. A *pseudograph* (or general graph)
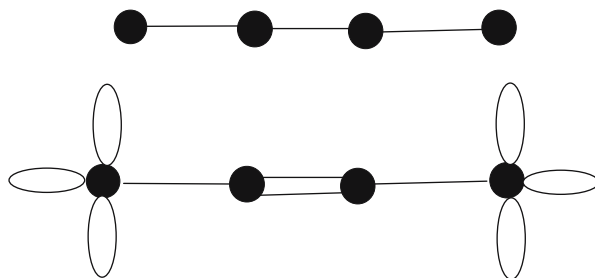
Figure 1. *Top*: The chemical graph of 1,2 difluoroethylene; *Bottom*: the chemical pseudograph of 1,2 difluoroethylene.

$G=(V, E)$ may contain multiple edges between pairs of vertices and self-connections (or loops), which are edges from a vertex to itself. The degree of a vertex in a pseudograph is the number of edges that occur with it, where self-connections contribute twice to its degree. Every graph is a pseudograph, but not every pseudograph is a graph. Some mathematicians reserve the term *simple graph* for a graph with no multiple edges and loops. A *chemical pseudograph* is a graph where vertices, edges, and loops encode atoms, bonds, and lone-pairs non-bonding electrons, respectively. The degree of a vertex in a chemical pseudo-graph is also called its valence. Normally chemical graphs either simple graphs or pseudographs are represented with depleted hydrogens (also called, hydrogen-suppressed graphs, or HS-graphs). In figure 1, *top*, and *bottom* are shown the simple chemical graph and the pseudograph of 1,2 difluoroethylene, where every pair of non-bonding electrons of the halogen atom (here F) is encoded by a loop.

It is evident that the chemical graph of figure 1, top, can encode a wide variety of molecules, like butane, dichloroethane, butadiene, and so on. The general graph of figure 1 bottom, instead, can only encode any 1,2 XCH=CHX haloethylenes, where X = F, Cl, Br, and I. A characteristic of chemical graphs and pseudographs is their impossibility to differentiate among different spatial isomers, and, especially, between the *cis* and *trans* isomers around a double bond. This last problem has partially been solved in molecular connectivity theory, with the concept of virtual rings [14, 15]. The degree of the vertices in the graph of figure 1, *top* is two for the mid-vertices and one for the extreme vertices, while for the corresponding pseudograph (*bottom*) is three for the mid-vertices and seven for the extreme vertices (every loop counts two). Another graph concept, quite useful in chemistry and which will be used in the foregoing study is the concept of complete graph, which can be used to encode the core electrons of an atom. A *complete graph*, K_p, of order p, is a graph where every pair of its vertices is adjacent. A complete graph is always r-regular, i.e., it has all of its vertices with the same degree r, where $r = p - 1$. Clearly, while a complete graph is always regular, the contrary is not true, i.e., pure (non-branched) cyclic graphs are regular. In figure 2 are shown four types of complete graphs: K_2, K_3, K_4,
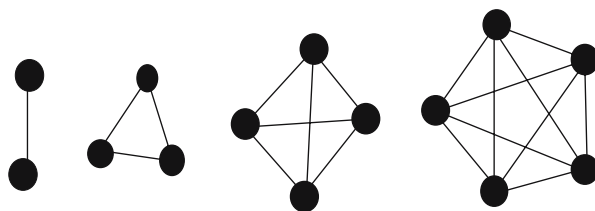
Figure 2. Five different types of complete graphs: K_2, K_3, K_4, and K_5.

and K_5. K_1, is just a vertex, this means that the single vertices of graphs can always be read as K_1 complete graphs.

A graph with no edges and vertices is a *null* (K_0) complete graph [16]. Null graphs could be used to encode the depleted hydrogens, i.e., the normal H-suppressed chemical graph is just a set of K_0, and K_1 graphs and a set of edges connecting the K_1 graphs.

2. Methods

For any kind of graphs it is possible to define an *adjacency matrix* and a *distance matrix* [14], which are both symmetric. An adjacency matrix has the element $g_{ij} = 1$ if vertices i and j are adjacent, otherwise $g_{ij} = 0$ (non-connected edges, or bonds). The adjacency matrix of a pseudograph has elements along the main diagonal, ps_{ii}, which count the multiple connections and the self-connections (they count twice each) in a general graph. The graph-pseudograph adjacency matrix can, thus, be written in a compact form, where the graph elements are $g_{i,j}$ and the pseudograph elements are ps_{ii}. A prototype of such a matrix for a four-vertex system is the 4×4 matrix M is shown below. This matrix could also be the prototype for a distance matrix. In this case, $ps_{i,i} = 0$, while the $g_{i,j}$ elements equal the number of edges in the shortest path between vertices i and j. Below the matrix M are given the examples of the adjacency matrix, M', for the chemical graph of 1,2 difluoroethylene (figure 1, top), of the adjacency matrix, M'', for the chemical pseudograph of 1,2 difluoroethylene (figure 1, bottom), and the distance matrix, D, of both chemical and general graph of 1,2 difluoroethylene.

$$M = \begin{pmatrix} ps_{1,1} & g_{1,2} & g_{1,3} & g_{1,4} \\ g_{2,1} & ps_{2,2} & g_{2,3} & g_{2,4} \\ g_{3,1} & g_{3,2} & ps_{3,3} & g_{3,4} \\ g_{4,1} & g_{4,2} & g_{4,3} & ps_{4,4} \end{pmatrix}, \tag{1}$$
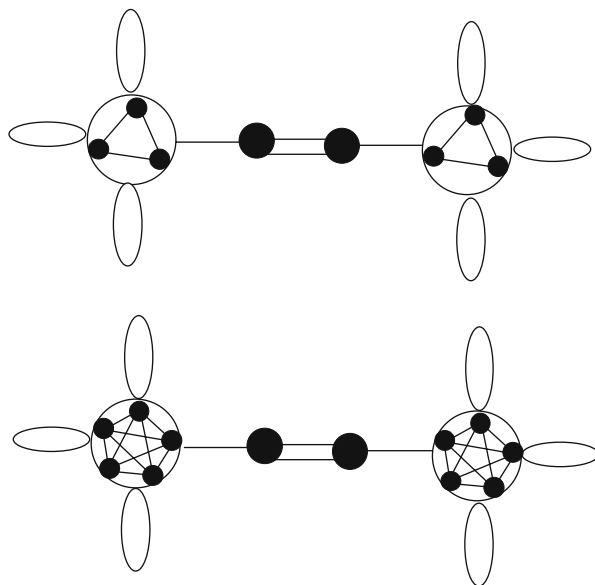
Figure 3. The chemical pseudograph plus complete graph of 1,2-dichoro- (top), and dibromoethylene.

$$M' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad M'' = \begin{pmatrix} 6 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 6 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}. \quad (2)$$

The sum of the elements either along a row or along a column in M', and M'' is the degree of a vertex of a chemical graph and pseudograph, respectively, normally called δ (delta) and δ^v (valence delta), which are used in molecular connectivity theory to derive the corresponding indices [3, 12]. These indices are referred as *graph invariants*, as isomorfic graphs possess identical topological indices.

2.1. The complete graphs and the core electron representation

As already told core electrons can be encoded with complete K_p graphs. The order p of these graphs mimics (broadly speaking) the principal quantum number, n. In figure 3 the chemical pseudograph plus complete graph of 1,2-dichoro- (top), and dibromoethylene (bottom) are shown.

Carbon atoms are always encoded by a K_1 complete graph, the Cl, and Br atoms atoms are, instead, encoded by K_3, and K_5 complete graphs, respectively, as up to now, with some exceptions, only odd complete graphs, with $p = 1, 3, 5, 7, \ldots$ showed an optimal model quality [17–20]. It should be under-

lined that the dimension of the halogen vertices in figure 3 has no metrical meaning, it is just a zoom of this vertex, whose dimensions are as meaningless as the dimensions of the K_1 vertex. The circle encircling the complete graph has no meaning it is just a frame for the complete graph. The proposed and successfully tested algorithm for the degree of a vertex of a chemical pseudograph-complete graph, δ^v, shown in equation (3) is able to quantify the contributions from multiple bonds, non-bonding electrons and core electrons,

$$\delta^v = q \cdot \delta^v(ps)/[p \cdot r + 1].\qquad(3)$$

Parameter $\delta^v(ps)$ is just the vertex degree number of an atom in a pseudograph, and it can be obtained with the adjacency matrix of the chemical pseudograph. The parameter $p \cdot r$ equals the sum of all vertex degrees in complete graphs. This sum equals twice the number of connections. Parameter q equals 1 or p, even if it could be used as an optimizing parameter. Four are the possible representations for the core electrons: for $q = 1$, and $p = $ odd a K_p-(p-odd) representation is obtained, for $q = 1$, and $p = 1, 2, 3, 4, \ldots$(i.e., sequential values), a K_p-(p-seq) representation is obtained, for $q = p$, and $p = $ odd, a K_p-(pp-odd) representation is obtained, and for $q = p$, and $p = $ seq, a K_p-(pp-seq) representation is obtained. Now, the full pseudograph-complete graph adjacency matrix, which is a product of two symmetric matrices, and only exceptionally will be a symmetric matrix, has the following general form, for a chemical graph with four vertices,

$$A = \begin{pmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 \\ 0 & 0 & k_3 & 0 \\ 0 & 0 & 0 & k_4 \end{pmatrix} \begin{pmatrix} ps_{1,1} & g_{1,2} & g_{1,3} & g_{1,4} \\ g_{2,1} & ps_{2,2} & g_{2,3} & g_{2,4} \\ g_{3,1} & g_{3,2} & ps_{3,3} & g_{3,4} \\ g_{1,4} & g_{2,4} & g_{3,4} & ps_{4,4} \end{pmatrix}.\qquad(4)$$

Here, $k_i = q/(pr + 1)_{Kpi}$, which depends on the type of complete graph chosen for the given vertex. For the HS-1Cl,2Br-Ethylene graph, i.e., Cl^1–C^2=C^3–Br^4, assuming, $q = 1$, and p odd, after multiplication we have, $ps_{1,1} = 6/7$, $g_{1,2} = 1/7$ (for Cl), and $ps_{4,4} = 6/21$, $g_{4,3} = 1/21$ (for Br), while for the remaining carbon atoms we have: $g_{2,1} = g_{2,2} = g_{2,3} = g_{3,2} = g_{3,3} = g_{3,4} = 1$, the other elements being zero. The final matrix for this compound is no more symmetric, i.e.,

$$A = \begin{pmatrix} 6/7 & 1/7 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1/21 & 6/21 \end{pmatrix}.\qquad(5)$$

The choice $q = 1$ or p has its rationale in the fact that for $q = p$ the δ^v values are rather similar to the $\delta^v = (2/n)^2\delta^v(ps)$ values of the electrotopological state, E_S, concept [21, 18–20]. For $q = 1$, instead, the δ^v values are rather similar to the $\delta^v = \delta^v(ps)/(Z - Z^v - 1)$ values, which were originally used to derive the valence

molecular connectivity indices for any kind of atom. Notice that for second-row atoms, $\delta^v = \delta^v(ps)$, always, as it should be.

2.2.   *The hydrogen perturbation*

Algorithm (3), as well as the previous algorithms of molecular connectivity, do not directly encode the bonded hydrogen atoms and do not allow to differentiate among p- and $\delta^v(ps)$-similar atoms, which differ only in the bonded number of hydrogen atoms. Let us take the first atom of the following compounds {LiF, BeHF, BH_2F, CH_3F}, they all have all the same δ^v, the same being true for {BeF_2, BHF_2, CH_2F_2}, and {BF_3, CHF_3} sets of compounds, and, correspondingly, for higher row atoms with $p > 1$. The depleted hydrogen atoms are responsible for this δ^v degeneracy. In fact, the HS-graphs and general graphs within these sets of compounds are equivalent. To circumvent this degeneracy a hydrogen perturbation parameter has recently been introduced and tested into δ^v [22]. The definition of this perturbation avoids the introduction of new graph concepts. It should be remarked that, normally, for alkanes equation (3) simplifies into $\delta^v = \delta^v(ps) = \delta$, as with alkanes the general chemical graph simplify into a simple chemical graph. Attempts to quantify hydrogen atoms have already been done by Kier and Hall [12, 21]. Recently, a modified connectivity index was proposed, where the contribution of the hydrogen atoms to the overall connectivity index was parametrized as $n_H/6$ [23]. The chosen guidelines to define the hydrogen perturbation parameter are: (i) the new δ^v should not contradict the δ^v of equation (3) for compounds with no hydrogens, (ii) the new values for δ^v should include a contribution from the bonded hydrogen atoms, which decreases with decreasing number of bonded hydrogen atoms (iii) due to the good results of algorithm (3) this contribution should be minimal, i.e., the resulting δ^v should not be affected in a macroscopic way by the hydrogen atoms, (iv) the hydrogen contribution should decrease in importance with growing p, i.e., the dependence on p should equal the one given in equation (3), and, last but not least, (v) the new algorithm should not introduce any new graph concept relatively to algorithm (3). These guidelines allow to define the following parameter,

$$f_\delta = [\delta^v_m(ps) - \delta^v(ps)]/\delta^v_m(ps) = 1 - \delta^v(ps)/\delta^v_m(ps) = n_H/\delta^v_m(ps). \qquad (6)$$

Here, $\delta^v_m(ps)$ is the maximal $\delta^v(ps)$ value a heteroatom (a vertex) can have in a chemical HS-pseudograph when all bonded hydrogens are substituted by heteroatoms, n_H equals the number of hydrogen atoms bonded to a heteroatom. Then, the 'hydrogen-perturbed' δ^v is,

$$\delta^v = \frac{(q + f_\delta^n)\delta^v(ps)}{(pr + 1)}. \qquad (7)$$

Notice (i) that for a completely substituted carbon atoms (or heteroatoms), $f_\delta = 0$ as $\delta_m^v(ps) = \delta^v(ps)$ (or $n_H = 0$), and equation (3) is retrieved, and (ii) that for hydrocarbons δ and δ^v are no more equal even if $\delta^v(ps) = \delta$, but, $\delta^v = (1 + f_\delta^n)\delta$ (here $p = 1$). For quaternary carbons $f_\delta = 0$ and $\delta^v = \delta$. Exponent n is no optimization parameter, it just quantifies the importance of the hydrogen atoms in a model study, i.e., the higher the n values the lower the importance of the perturbation introduced by the hydrogen atoms. For every property (or activity) of a class of compounds n is constant, i.e., the corresponding δ^v is constant. To avoid overlapping δ^v values for compounds with differing number of hydrogen atoms $n \geqslant 2$ should hold. The element $k_i = q/(pr + 1)_{Kpi}$ of matrix (4) has to be modified into $k_i = (q + f_\delta^n)/(pr + 1)_{Kpi}$. This algorithm implies to know two pseudograph adjacency matrices, A, for each compound: the normal adjacency ps-matrix for $\delta^v(ps)$, and the adjacency ps-matrix for the corresponding fully substituted compound for $\delta_m^v(ps)$. Thus, the adjacency ps-matrix for the HS-pseudograph of CH_3–NH_2 is used to derive $\delta^v(ps)$, while the adjacency ps-matrix for the pseudograph of CX_3–NX_2 is only be used to derive $\delta_m^v(ps)$, where X is a dummy monovalent heteroatom. Equation (7) with $n = 2$, i.e., with f_δ^2 has been plotted versus the atomic number of the first atom (A) (figure 4, left side) for the four sets of compounds for which $q = p = 1$: {LiF, BeHF, BH$_2$F, CH$_3$F}, {BeF$_2$, BHF$_2$, CH$_2$F$_2$}, {BF$_3$, CHF$_3$}, and {CF$_4$}. Figure 4, right side, shows the corresponding values for the sets (here: $q = p = 2$): {NaCl, MgHCl, AlH$_2$Cl, SiH$_3$Cl}, {MgCl$_2$, AlHCl$_2$, SiH$_2$Cl$_2$}, {AlCl$_3$, SiHCl$_3$}, and {SiCl$_4$}. Note that guideline (i) holds for compounds LiF, BeF$_2$, BF$_3$, CF$_4$, and for the corresponding second-row compounds, as it should be, while guarantee (ii) and (iii) hold for the other compounds. The $\delta^v(A)$ value of A in a set is never larger than the $\delta^v(A)$ value in the next set of compounds, and within a set it grows more and more slowly with growing number of bonded hydrogens, while for an atom A it grows with growing number of fluorine atoms. In fact, the difference, $\Delta\delta = \delta^v(A_{i+1}^s) - \delta^v(A_i^s)$, between adjacent $\delta^v(A)$ values within a set, s, of compounds, decreases (see the first and second set of compounds), this in keeping with the decreasing importance of the bonded hydrogen atoms. The difference between $\delta^v(A)$ values in different sets, $\Delta\delta' = \delta^v(A_i^{s+1}) - \delta^v(A_i^s)$, is nearly constant, in keeping with the constant contribution of the entering fluorine atom. Equation (7) guarantees that with growing p, $\delta^v(A)$ decreases in any case (guideline iv), as can be seen in figure 4. The $n \geqslant 2$ condition for equation (7) guarantees that no $\delta^v(A)$ of a set overlaps with a $\delta^v(A)$ of another set. In fact, for $n = 1$ the undesirable result $\delta^v[C, (CH_2F_2)] = \delta^v[B, (BF_3)] = 3$ is obtained.

All valence χ^v values obtained with δ^v of equation (7) are now able to differentiate among compounds with atoms where $\delta^v(ps) \neq \delta_m^v(ps)$. Regressing the $^1\chi^v(7)$ [see equation (14)] values obtained with δ^v of equation (7) versus the $^1\chi^v(3)$ values obtained with δ^v of equation (3), for the first four sets, it is possible to notice that they do not markedly diverge from each other, and that the
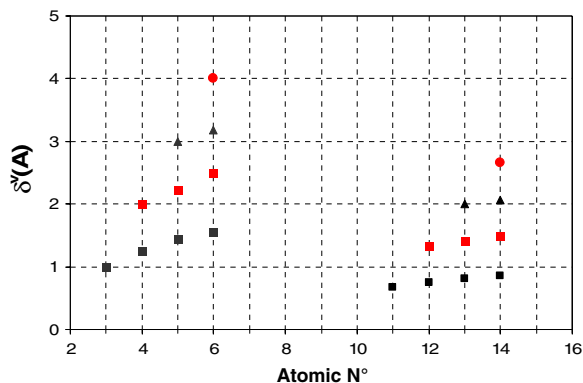
Figure 4. δ^v values [from equation (7)] versus the atomic number of the first atom (A) for eight sets of compounds. *Left side*: {LiF, BeHF, BH_2F, CH_3F}, (■); {BeF_2, BHF_2, CH_2F_2}, (■); {BF_3, CHF_3}, (▲); {CF_4}, (●). *Right side*: {NaCl, MgHCl, AlH_2Cl, SiH_3Cl}, (■); {$MgCl_2$, $AlHCl_2$, SiH_2Cl_2}, (■); {$AlCl_3$, $SiHCl_3$}, (▲); {$SiCl_4$}, (●).

linear regression is rather good,

$$^1\chi^v(7) = 1.121(\pm 0.091)\,^1\chi^v(3) - 0.091(\pm 0.033) : F = 340, r^2 = 0.977,$$
$$s = 0.03, N = 10. \tag{8}$$

The same calculations for the last four sets of compounds (here A is a third-row atom) shows that the divergence here decreases,

$$^1\chi^v(7) = 1.070(\pm 0.034)\,^1\chi^v(3) - 0.078(\pm 0.028) : F = 966, r^2 = 0.992,$$
$$s = 0.02, N = 10. \tag{9}$$

Let us see how equation (7) (with $n = 2$) and (3) value the two carbons of four HS-pseudographs of alkanes, where $q = p = 1$,

$C–C : \delta^v(3) = \delta^v(ps) = \delta = 1, \delta_m^v(ps) = 4, \delta^v(7) = 1.5625$,

$–C–C– : \delta^v(3) = \delta^v(ps) = \delta = 2, \delta_m^v(ps) = 4, \delta^v(7) = 2.5$ [with $n = 1$
  we have $\delta^v(7) = 3$],

$> C–C < : \delta^v(3) = \delta^v(ps) = \delta = 3, \delta_m^v(ps) = 4, \delta^v(7) = 3.1875$ [for $n = 1$
  we had $\delta^v(7) = 3.75$],

$\rightarrow C–C \leftarrow : \delta^v(3) = \delta^v(ps) = \delta = \delta_m^v(ps) = \delta^v(7) = 4$.

Here δ is the main molecular connectivity parameter of a simple HS-graph. Regressing for this four alkane carbons $\delta^v(7)$ against $\delta^v(3)$ from equation (3) confirms that the $\delta^v(3)$ values are not that inconsistent.

$$\delta^v(7) = 0.8(\pm 0.03)\delta^v(3) + 0.813(\pm 0.091) : F = 585, r^2 = 0.997,$$
$$s = 0.07, N = 4. \tag{10}$$

2.3. *The molecular connectivity indices*

To avoid a huge combinatorial problem only a restricted set of molecular connectivity basis indices $\{\beta\}$, will here be considered for model purposes. The $\{\beta\}$ set will be made of three subsets $\{\beta\} = \{\{\chi\}, \{\psi\}, \{\beta_d\}\}$, i.e., a subset of eight molecular connectivity basis indices $\{\chi\}$ [12], a subset of eight molecular pseudo-connectivity basis indices $\{\psi\}$ [24, 25], and a subset of 12 dual connectivity and pseudoconnectivity basis indices $\{\beta_d\}$ [26]. They all will be classified as molecular connectivity basis indices. It should be noted that a small molecule with different heteroatoms, such as acetic acid, can have more than forty molecular connectivity basis indices. The cited subsets of basis indices are,

$$\{\chi\} = \{D, {}^0\chi, {}^1\chi, \chi_t, D^v, {}^0\chi^v, {}^1\chi^v, \chi_t^v\},$$
$$\{\psi\} = \{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^T\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\},$$
$$\{\beta_d\} = \{{}^0\chi_d, {}^1\chi_d, {}^1\chi_s, {}^0\chi_d^v, {}^1\chi_d^v, {}^1\chi_s^v, {}^0\psi_{Id}, {}^1\psi_{Id}, {}^1\psi_{Is}, {}^0\psi_{Ed}, {}^1\psi_{Ed}, {}^1\psi_{Es}\}. \quad (11)$$

Following definitions show that connectivity and pseudoconnectivity indices are formally similar,

$$D = \Sigma_i \delta_i, \qquad {}^S\psi_I = \Sigma_i I_i, \qquad (12)$$

$$ {}^0\chi = \Sigma_i (\delta_i)^{-0.5}, \qquad {}^0\psi_I = \Sigma_i (I_i)^{-0.5}, \qquad (13)$$

$$ {}^1\chi = \Sigma (\delta_i \delta_j)^{-0.5}, \qquad {}^1\psi_I = \Sigma (I_i I_j)^{-0.5}, \qquad (14)$$

$$ \chi_t = (\Pi \delta_i)^{-0.5}, \qquad {}^T\psi_I = (\Pi I_i)^{-0.5}. \qquad (15)$$

Index 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 equations (12), and (13) as well as products (Π) in equations (15), are taken over all vertices of the hydrogen-suppressed chemical graph. Sums in equation (14) are over all edges of the chemical graph (σ bonds in a molecule). Replacing δ with δ^v the subset of valence χ^v indices $\{D^v, {}^0\chi^v, {}^1\chi^v, \chi_t^v\}$, is obtained. 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 and total, while the other sub- and superscripts follow the established denomination for χ indices. Basis ψ indices are indirectly related to δ^v through the I-State (ψ_I subset) and S-State (ψ_E subset) indices,

$$I = (\delta^v + 1)/\delta, \quad S = I + \Sigma \Delta I, \quad \text{with } \Delta I = (I_i - I_j)/r_{ij}^2, \qquad (16)$$

where, r_{ij} counts the atoms in the minimum path length separating atoms i and j, which equals the graph distance, $d_{ij} + 1$. Factor $\Sigma \Delta I$ incorporates the information about the influence of the remainder of the molecular environment, and, as it can be negative, S can also be negative. To avoid imaginary ψ_E values, every S value of many of our classes of compounds (as some atoms have $S < 0$), has

been rescaled. The rescaling procedure, which should be done with the intent to avoid either too small or too large S_i values, has a minor influence on the quality of the modeling [25]. The used dual basis indices are defined in the following way,

$$^0\chi_d = (-0.5)^N \Pi_i(\delta_i), \qquad ^0\psi_{Id} = (-0.5)^N \Pi_i(I_i), \qquad (17)$$

$$^1\chi_d = (-0.5)^{(N+\mu-1)} \Pi(\delta_i + \delta_j), \quad \psi_{Id} = (-0.5)^{(N+\mu-1)} \Pi(I_i + I_j), \quad (18)$$

$$^1\chi_s = \Pi(\delta_i + \delta_j)^{-0.5}, \qquad ^1\psi_{Is} = \Pi(I_i + I_j)^{-0.5}. \qquad (19)$$

Replacing, in these expressions δ with δ^v and I with S the corresponding χ^v valence dual and ψ_E dual indices are obtained. The exponent, μ, in equation (18) is the cyclomatic number, which indicates the number of cycles of a chemical graph and it is equal to the minimal number of edges necessary to be removed in order to convert the (poly)cyclic graph to an acyclic subgraph. Dual indices, excluding $^1\chi_s$, and $^1\psi_{Is}$, can be negative.

By the aid of a trial-and-error procedure, a series of higher-order descriptors, known as molecular connectivity terms [14, 24–26], $X = f(\chi)$, molecular pseudoconnectivity terms, $Y = f(\psi)$, and higher-order terms, $Z = f(X, Y, \beta_d)$, $Z' = f(Z, \beta_d)$ can be derived. X and Y terms can be as convoluted as the rational function of equation (19), even if, they are normally simpler,

$$T = [a(\beta_1)^m + b(\beta_2)^n]^q / [c(\beta_3)^o + d(\beta_1)^p]^r. \qquad (20)$$

Here, for $\beta = \chi$, $T = X$, and for $\beta = \psi$, $T = Y$. Parameters, $a-d$, and $m-r$ are parameters apt to improve the quality of a model and within a class of compounds are held constant. It seldom happens that all these parameters are different from either zero or one.

Used statistics are, the squared correlation coefficient, r^2, the standard deviation of the estimate, s, the F value, and the number of points, N. Plot methods, whose importance has recently been emphasized [27–29], and leave-more-out methods will constantly be used to check the quality of every model. The utility of each regression parameter, c_i (in parenthesis near c_i or collected under a vector form), will also be given for the fit-model equation, i.e. $|c_i/s_i|$, where s_i is the estimated error of c_i. A high utility underlines the importance of the corresponding index, even if a low utility should be expected when the corresponding regression parameter is nearly zero.

3.   Results and discussion

In the following we will review [20, 22, 30] the model of a series of properties following the decreasing hydrogen content of some classes of compounds (i.e., the ratio of the H atoms to the heteroatoms, n_H/n_{ht}). We will start with the Motor Octane Number (MON) of alkanes, which have the highest hydrogen
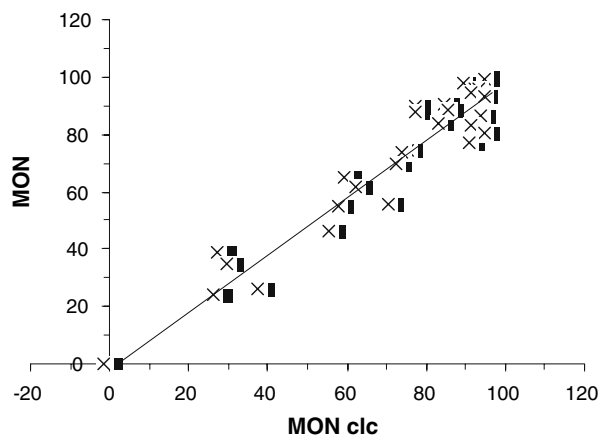
Figure 5. Experimental versus calculated (clc) MON, with a training set of 30 alkanes (■), and with a training and evaluation sets of 18 and 12 alkanes, respectively (×).

content, and going through the boiling points of a composite class of amines and alcohols, the molecular polarizability of a class of organic compounds, the partition coefficient of halocompounds, the molar refraction, and the boiling points of halomethanes, we will land on a class of compounds with no hydrogen atoms, the metal halides of which two properties will be studied, the lattice and the binding enthalpies.

3.1. The motor octane number, MON

Here the ratio n_H/n_{ht} equals 2.3. In table 1 are collected the MON values for 30 alkanes taken from [14]. Alkanes have no heteroatoms other than carbon atoms, which means that $\delta^v = (1 + f_\delta^n)\delta$, as $p = 1$, and as graph and general graph here coincide, $\delta^v(ps) \equiv \delta$, while in $f_\delta = (\delta_m - \delta)/\delta_m$, $\delta_m = 4$, and $\delta = 1, 2, 3, 4$. In f_δ^2 the δ^v of C_q shows a negative $S[C(CH_3)_4]$ value (-0.063), for this reason a rescaling factor, Resc $= -1$, was here introduced. Notice that for $n = 3$ (in f_δ^n) $S[C(CH_3)_4] > 0(= 0.078)$. The best descriptor, which has been obtained with $f_\delta^2 \neq 0$, is,

$$^{f2}Y(\text{MON}) = (^0\psi_E - 1.3^1\psi_E)^5 : F = 317, \quad r^2 = 0.920, \quad s = 7.7, \quad N = 30,$$
$$\text{MON} = -26.222^{f2}Y(18) + 94.707(50). \tag{21}$$

In parenthesis are the utility values of the regression coefficients. Equation (21) has been used to obtain the calculated values of figure 5 (full squares).

Twelve compounds chosen at random (with asterisk in table 1), but keeping $n_H/n_{ht} = 2.3$, were left-out from the model and the model was rerun. The new descriptor, $^{f2}Y'(\text{MON})$, equals the previous one and has the following

Table 1
Motor octane numbers, MON, for 30 alkanes.

Alkanes	MON	MON$_{Clc}$	Alkanes	MON	MON$_{Clc}$
*4	90.1	77.4	3M5	74.3	75.4
2M3	97.6	89.5	*23ME5	88.1	77.2
2M4	90.3	84.5	*223MMM5	99.9	94.9
*2M5	73.5	73.7	234MMM5	95.9	92.3
24MM6	69.9	72.3	2M7	23.8	26.2
*33MM5	86.6	93.8	224MMM5	100.0	94.6
5	61.9	62.3	233MMM5	99.4	94.9
*23MM4	94.4	91.1	*22MM4	93.4	94.6
33MM6	83.4	91.3	6	26.0	37.3
22MM5	95.6	93.4	*25MM6	55.7	70.2
*22MM6	77.4	90.8	7	0.0	−1.7
4M7	39	27.2	22MM3	80.2	94.9
*3M7	35	29.4			
*3M6	55.0	57.7			
24MM5	83.5	82.9			
*23MM5	88.5	85.4			
3E5	65.0	59.3			
2M6	46.4	55.4			

2 = ethane, 3 = propane, etc.: M = methyl, E = Ethyl; e.g., 34ME6 = 3-methyl-4-ethylhexane.
* Left-out values (see text).

$F/r^2/s/N$ values: 275/0.945/7.2/18. Notice that decreasing N favors the $f_\delta^2 \neq 0$ case over the $f_\delta = 0$ case. Now, the regression vector (utilities) for $^{f2}Y'(\text{MON})$ is: $\boldsymbol{C} = [-26.592(17), 94.971(41)]$.

The calculated values for the training ($N = 18$) and evaluation ($N = 12$) sets are in table 1 (MON$_{Clc}$) and are plotted in figure 5 (crosses).

3.2. Boiling points of amines and alcohols

For amines $n_{H}/n_{ht} = 2.2$, while for alcohols $n_{H}/n_{ht} = 2.0$. The experimental values collected in table 2 are taken from [26]. The model of each class of compounds shows no improvement with $f_\delta^n \neq 0$ (for any n value). The composite class amines plus alcohols [R-NH$_2$ + ROH], with $n_H/n_{ht} = 2.1$, shows, instead, with $f_\delta^8 \neq 0$ a clear improvement over the $f_\delta = 0$ case (first case).

$$X(\text{Am} + \text{Al}) = [(\chi_t)^{0.3} + 0.05^1\chi^v] : F = 383, \ r^2 = 0.895, \ s = 15.7, \ N = 48,$$
$$^{f8}X(\text{Am} + \text{Al}) = [(\chi_t^v)^{0.2} + 0.03^1\chi^v] : F = 1150, \ r^2 = 0.962, \ s = 9.4, \ N = 48.$$

The fit-model equation for the 48 points, whose plot is shown in figure 8, is,

$$\text{BP}(\text{Am} + \text{Al}) = -768.44^{f8}X(\text{Am} + \text{Al})(34) + 951.99(58). \tag{22}$$

To check if the given improvement is only due to the growing N (from 21 and 27 to 48) every second compound starting with the first one in both classes of
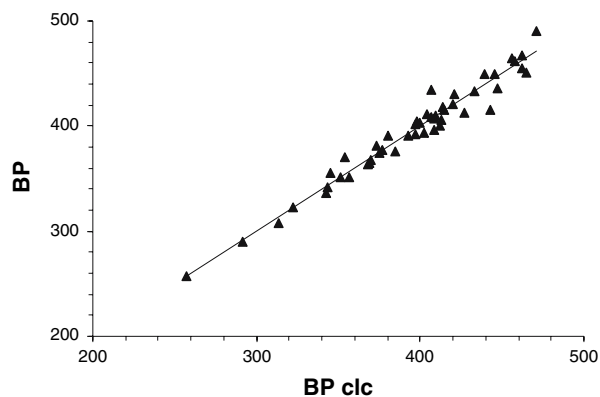
Figure 6. Model of the boiling points, BP, of 27 alcohols plus 21 amines.

amines and alcohols of table 2 (i.e., $1°$, $3°$, $5°$, ...) will be excluded from the model. Now, the number of training amines is 10 and the number of training alcohols is thirteen, while n_H/n_{ht} does not change. The best descriptors for $f_\delta \neq 0$, and $f_\delta = 0$ for the $N = 23$ training points are,

$$X'(\text{Am} + \text{Al}) = [(\chi_t)^{0.4} + 0.08^1\chi^v]^{0.6} : F = 192, \ r^2 = 0.901, \ s = 13.9,$$
$$^{f8}X'(\text{Am} + \text{Al}) = [(\chi_t^v)^{0.2} + 0.03^1\chi^v]^{1.4} : F = 693, \ r^2 = 0.970, \ s = 7.6.$$

The C vector (utilities) for $^{f8}X'(\text{Am} + \text{Al})$ is $[-599.92(26), \ 777.01(53)]$. The optimal term is only slightly different from the previous one with $N = 48$. Thus, here N changes do not to affect the importance and type of the hydrogen perturbation. The model of the 23 training points plus the 25 evaluating points with $^{f8}X'(\text{Am} + \text{Al})$ is hardly different from the model displayed in figure 6. In table 2 are given the residuals ($\text{res} = BP - BP_{\text{Clc}}$) for this last training case ($N = 23$).

3.3. Molecular polarizability of organic compounds, α

Here the ratio of the hydrogen atoms to the heteroatoms equals: $n_H/n_{ht} = 1.2$. The polarizabilty values, shown in table 3, were taken from [31]. To avoid some imaginary pseudoconnectivity values the S value of these organic compounds has been rescaled to the S value of Si in SiF_4, i.e., to $S[Si(SiF_4)] = -6.611$. The optimal model for α with molecular connectivity indices derived with a δ^v with $f_\delta = 0$, $q = 1$, and $p = \text{odd} = 1, 3, 5, \ldots$, is achieved by the following descriptor,

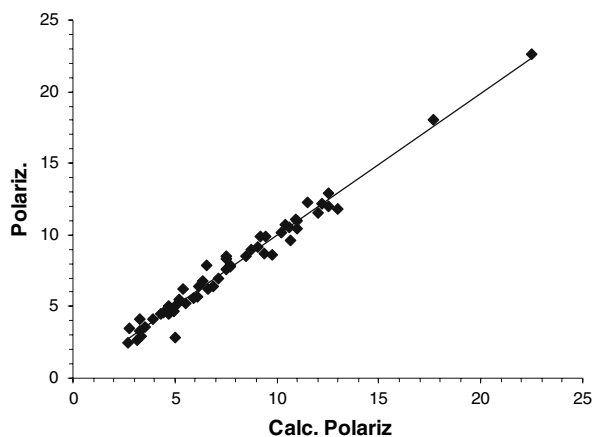$$X(\alpha) = (3^0\chi^v + 1.2^1\chi + 0.01^0\chi_d^v), \ F = 2013, \ r^2 = 0974, \ s = 0.6, \ N = 54.$$

Figure 7. Plot of the molecular polarizability (Polariz/Calc. Polariz) with a $N = 31$ training set.

Table 2

The boiling points (T_b in K) of primary amines and alcohols, and their calculated residuals (res).

Amines	T_b	res	Alcohols	T_b	res
CH_3-	256.6_5	11	$(CH_3)_2CH$-	355.5_5	9.2
CH_3CH_2-	290.1_5	−1.0	$CH_3CH_2CH_2$-	370.2_5	16
$(CH_3)_2CH$-	307.1_5	−7.2	$CH_3CH_2C(CH_3)_2$-	375.4_5	−9.6
$CH_3CH_2CH_2$-	322.1_5	−1.0	$CH_3CH(CH_3)CH_2$-	381.2_5	7.3
$CH_3CH_2CH(CH_3)$-	336.1_5	−6.9	$CH_3(CH_2)_3$-	390.7_5	10
$CH_3CH(CH_3)CH_2$-	341.1_5	−3.0	$CH_3CH_2CH_2CH(CH_3)$-	392.0_5	−4.3
$CH_3CH_2CH_2CH_2$-	350.9_5	−0.6	$CH_3C(CH_3)_2CH(CH_3)$-	393.5_5	−8.3
$CH_3CH_2C(CH_3)_2$-	351.1_5	−6.0	$CH_3(CH_2)_2C(CH_3)_2$-	396.1_5	−11
$(CH_3CH_2)_2CH$-	364.1_5	−4.4	$CH_3CH(CH_3)CH(CH_3CH_2)$-	400.6_5	−10
$CH_3CH_2CH_2CH(CH_3)$-	365.1_5	−4.1	$CH_3CH_2CH(CH_3)CH_2$-	402.0_5	5.0
$CH_3CH(CH_3)CH_2CH_2$-	368.1_5	−2.2	$CH_3CH(CH_3)CH_2CH_2$-	405.1_5	7.4
$CH_3C(CH_3)_2CH(CH_3)$-	375.1_5	−0.3	$CH_3CH(CH_3)CH_2CH(CH_3)$-	406.1_5	−5.9
$CH_3(CH_2)_4$-	377.5_5	0.7	$(CH_3CH_2)_2C(CH_3)$-	409.1_5	2.9
$CH_3(CH_2)_3CH(CH_3)$-	390.6_5	−1.9	$CH_3CH_2C(CH_3)_2CH_2$-	409.8_5	1.3
$CH_3(CH_2)_5$-	403.1_5	3.8	$CH_3(CH_2)_4$-	411.1_5	7.5
$CH_3CH_2CH(CH_3)CH_2CH(CH_3)$-	406.6_5	−0.5	$(CH_3CH(CH_3))_2CH$-	413.1_5	−12
$CH_3(CH_2)_4CH(CH_3)$-	415.1_5	2.0	$(CH_3CH_2)_3C$-	415.1_5	−24
$CH_3(CH_2)_6$-	430.0_5	11	$CH_3CH(CH_3)CH(CH_3)CH_2$-	418.1_5	5.8
$CH_3(CH_2)_7$-	449.1_5	12	$CH_3CH_2CH_2CH(CH_3)CH_2$-	421.1_5	3.0
$CH_3(CH_2)_8$-	465.1_5	13	$CH_3CH_2CH(CH_3)CH_2CH(CH_3)$-	432.9_5	2.3
$CH_3(CH_2)_9$-	490.1_5	25	$(CH_3CH_2)_2(CH_3)C$-	434.1_5	28
			$(CH_3(CH_2)_3)(CH_3CH_2)(CH_3)C$-	436.1_5	−7.3
			$CH_3(CH_2)_6$-	449.9_5	7.9
			$CH_3(CH_2)_5C(CH_3)_2$-	451.1_5	−8.7
			$(CH_3CH_2CH_2)_2(CH_3CH_2)C$-	455.1_5	−2.6
			$CH_3CH(CH_3)CH_2(CH_2)_4$-	461.1_5	7.2
			$CH_3(CH_2)_7$-	467.5_5	9.7

Here, a slight improvement is obtained with a $f_\delta^8 \neq 0$. The new term is formally equal to the previous unperturbed term,

$$^{f8}X'(\alpha) : F = 2061, \ r^2 = 0.976, \ s = 0.61, \ N = 54,$$
$$C = [0.6969(46), -0.4655(2.3)].$$

With the correlation vector (utilities), C, of $^{f8}X'(\alpha)$ the calculated values, α_{Clc}, shown in the corresponding column of table 3 have been obtained. Twenty-three compounds are now left-out (with asterisk in table 3), but preserving $n_H/n_{ht} = 1.2$. The optimal descriptors with the new training set of points for $f_\delta = 0$ and $f_\delta \neq 0$ are,

$$X(\alpha) = (3^0\chi^v + 1.4^1\chi + 0.01^0\chi_d^v)^{0.9} : F = 2157, \ r^2 = 0.986, \ s = 0.5, \ N = 31,$$
$$^{f8}X(\alpha) = (3^0\chi^v + 1.3^1\chi + 0.01^0\chi_d^v) : F = 2329, \ r^2 = 0.988, \ s = 0.5, \ N = 31,$$
$$C = [0.6937(48), -0.5501(2.6)].$$

Here the divergence between the $f_\delta = 0$ and $f_\delta \neq 0$ cases grows slightly favoring the $f_\delta \neq 0$ case. The number of data N is here important for the proton perturbation. The calculated values of the training ($N = 31$) plus evaluation sets ($N = 23$) obtained with $^{f8}X(\alpha)$ for $N = 31$, and with its regression vector, C, are very similar. Actually, the two descriptors are rather similar. Figure 7 shows the experimental versus calculated values, obtained with this last term.

3.4. Log P air partition coefficients at 37° C

Here, $n_H/n_{ht} = 0.56$. The partition coefficient, measured as Log P for 25 halocompounds (from [32]), is shown in table 4. For both $f_\delta = 0$, and $f_\delta^n \neq 0$ a K_p-(p-odd) representation for the core electrons gives rise to the optimal combination of indices for the saline, blood, and liver cases: $\{^0\chi, \, ^0\psi_I, D_F, D_{Cl}\}$. This is also a good descriptor for the other cases, i.e. $\{^0\chi, \, ^0\psi_I, D_F, D_{Cl}\}$, is practically a good descriptor for all the $N=150$ points. Here, $D_F = \Sigma_i \delta_F$, and $D_{Cl} = \Sigma_i \delta_{Cl}$, over all F and Cl atoms, respectively. The four basis indices of $\{^0\chi, \, ^0\psi_I, D_F, D_{Cl}\}$ are poorly correlated among them, as: $r^2(^0\chi, \, ^0\psi_I) = 0.88$, $r^2(^0\chi, D_F) = 0.14$, $r^2(^0\chi, D_{Cl}) = 0.37$, $r^2(^0\psi_I, D_F) = 0.01$, $r^2(^0\psi_I, D_{Cl}) = 0.55$, and $r^2(D_F, D_{Cl}) = 0.14$.

Let us model with a leave-5-out method (with asterisk in table 4, the first one chosen at random, and then every fifth but keeping $n_H/n_{ht} = 0.56$), and with a leave-10-out (with + and *) method but always keeping $n_H/n_{ht} = 0.56$. The training set will then be composed of 20 and 15 compounds, respectively. For these two training choices $\{^0\chi, \, ^0\psi_I, D_F, D_{Cl}\}$ is the best descriptor for the saline, blood and liver cases, for both the zero-H-perturbation ($f_\delta = 0$) and the non-zero-H-perturbation ($f_\delta^2 \neq 0$). The statistics are collected in table 5. Normally, the introduction of f_δ^2 improves the model, especially for the blood,

Table 3
Molecular polarizability, α, of organic compounds in units of $\mathring{A}^3$ and their computed (Clc) counterparts.

Compound	α	α_{Clc}	Compound	α	α_{Clc}
*Ethane	4.48	4.36	*Acetaldehyde	4.59	4.75
Propane	6.38	6.18	*Acetone	6.39	6.88
Neopentane	10.20	10.22	F-Methane	2.62	3.17
*Cyclopropane	5.50	5.21	*TriF-Methane	2.81	5.01
Cyclopentane	9.15	9.00	TetraF-Methane	2.92	3.32
Cyclohexane	11.00	10.90	Cl-Methane	4.55	4.46
*Ethylene	4.12	3.33	*DiCl-Methane	6.82	6.37
Propene	6.26	5.39	TriCl-Methane	8.53	8.46
*2MePropene	8.29	7.49	TetraCl-Methane	10.51	10.61
*Trans-2-Butene	8.49	7.54	*Br-Methane	5.61	5.99
Cyclohexene	10.70	10.37	*DiBr-Methane	8.68	9.43
Butadiene	7.87	6.52	TriBr-Methane	11.84	13.05
*Benzene	9.92	9.36	*I-Methane	7.59	7.55
Toluene	12.30	11.39	DiI-Methane	12.90	12.56
HexaMeBenzene	22.63	22.35	TriI-Methane	18.04	17.73
*Acetylene	3.50	2.80	*$CH_2=CCl_2$	7.83	7.69
Propyne	4.68	4.95	*cis*-CHCl=CHCl	7.78	7.73
$C(CCH)_4$	12.19	12.61	DiSilane	11.10	10.92
*Allene	5.00	4.70	*Formamide	4.08	3.94
Methanol	3.32	3.31	*Acetamide	5.67	6.12
Ethanol	5.11	5.11	Acetonitrile	4.48	4.67
*2-Propanol	6.97	7.12	*Propionitrile	6.24	6.60
Cyclohexanol	11.56	11.87	Tert-BuCyanide	9.59	10.69
Dimethylether	5.24	5.55	BenzylCyanide	11.97	12.41
*p-Dioxane	8.60	9.72	*TriCl-Acetonitrile	10.42	10.98
Methylamine	3.59	3.58	*Pyridine	9.92	9.15
Formaldehyde	2.45	2.72	Thiophene	9.00	8.70

* Left-out compounds (evaluation set).

oil, liver and fat cases. The different behavior of f_δ^2 in different media underlines the fact that the perturbation is property-dependent. Excluding the fat case where the $f_\delta^2 = 0$ and $\neq 0$ descriptions diverge as N gets smaller favoring the $f_\delta^2 \neq 0$ case, practically the other Log P cases show no consistent changes with N, an interesting result, actually an expected result with $n_H/n_{ht} = $ constant, but that only now shows up. In figure 8 is the model of 80 training points plus twenty external validated points (five left-out points for every set of Log P) obtained with the following vectors of the correlations coefficients for $^{f2}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ (utilities are in parenthesis). Notice that in multilinear relations coefficients with less than five figures are not always advantageous [14].

Table 4
Liquid and rat tissue, Log P air partition coefficients (37° C) for $N = 25$ halocompounds.

Molecule	Saline	Olive Oil	Blood	Liver	Liver-10-Out (Clc)	Muscle	Fat
CH_3Cl	−0.056	0.933	0.393	0.540	0.574	−0.013	1.130
*CH_2Cl_2	0.775	2.117	1.288	1.152	0.993	0.899	2.079
+$CHCl_3$	0.529	2.604	1.318	1.324	1.168	1.143	2.307
CCl_4	−0.456	2.573	0.655	1.152	1.257	0.657	2.555
$CH_2=CHCl$	−0.367	1.387	0.225	0.204	0.173	0.342	1.301
$CCl_2=CH_2$	−0.456	1.808	0.699	0.645	0.576	0.312	1.836
*$CHCl=CHCl(cis)$	0.512	2.444	1.334	1.185	0.965	0.785	2.356
+$CHCl=CHCl(tr)$	0.149	2.250	0.981	0.952	0.965	0.547	2.170
$CCl_2=CHCl$	−0.081	2.743	1.340	1.435	1.382	1.004	2.744
$CCl_2=CCl_2$	−0.102	3.329	1.276	1.847	1814	1.301	3.214
$CH_3–CH_2Cl$	0.037	1.590	0.611	0.558	0.480	0.508	1.587
*$CHCl_2–CH_3$	0.389	2.270	1.049	1.033	0.911	0.709	2.215
+$CH_2Cl–CH_2Cl$	1.057	2.563	1.483	1.553	1.477	1.369	2.537
$CCl_3–CH_3$	−0.125	2.470	0.760	0.934	1.058	0.498	2.420
$CHCl_2–CH_2Cl$	1.124	3.249	1.763	1.863	1.861	1.360	3.158
$CHCl_2–CHCl_2$	1.369	3.803	2.152	2.292	2.279	2.004	3.576
*$CCl_3–CH_2Cl$	0.548	3.429	1.620	1.945	2.008	1.597	3.332
+CH_2F_2	0.117	0.678	0.204	0.439	0.530	0.158	0.155
CH_2FCl	0.489	1.348	0.706	0.537	0.599	0.391	1.188
CH_2BrCl	0.937	2.558	1.618	1.465	1.356	1.045	2.512
CH_2Br_2	1.158	2.981	1.870	1.833	1.763	1.607	2.899
*$CF_3–CHClBr$	−0.301	2.297	0.721	0.882	0.600	0.649	2.260
+$CH_2=CHBr$	−0.357	1.748	0.607	0.522	0.249	0.354	1.692
$CH_2Br–CH_2Cl$	0.950	2.755	1.722	1.631	1.745	1.405	2.982
$CF_3–CH_2Cl$	−0.377	1.380	0.104	0.265	0.289	0.090	1.326

Throughout the sixth column are the calculated values for Liver with the leave-10-out method.
*+Left out compounds in two different leave-n-out methods.

Table 5
The $F - r^2 - s$ values due to $\{^0\chi, ^0\psi_I, D_F, D_{Cl}\}$ for $f_\delta^2 = 0$ and $\neq 0$, and $N = 20$ and 15.

f_δ^2	N	Saline	Blood	Liver	Oil	Muscle	Fat
0	20	21-0.846-0.3	32-0.895-0.2	36-0.906-0.2	75-0.953-0.2	29-0.885-0.2	138-0.974-0.2
$\neq 0$	20	15-0.799-0.3	44-0.922-0.2	57-0.939-0.2	164-0.978-0.1₅	36-0.904-0.2	257-0.986-0.1
0	15	11-0.819-0.3	19-0.885-0.3	27-0.916-0.2	47-0.951-0.2	21-0.891-0.2	88-0.972-0.2
$\neq 0$	15	9.1-0.785-0.4	25-0.910-0.2	48-0.951-0.2	96-0.974-0.1₅	29-0.922-0.2	281-0.992-0.1

$$C(Bl) = [-2.81203(8.2), 3.15516(9.9), 0.74932(4.7), 0.10050(1.2), 1.68496(5.8)],$$
$$C(Liv) = [-2.66129(8.7), 2.99186(11), 0.77664(5.5), 0.19568(2.7), 1.47851(5.7)],$$
$$C(Oil) = [-2.08605(8.3), 2.85394(12), 0.37152(3.2), 0.03685(0.6), 1.37054(6.5)],$$
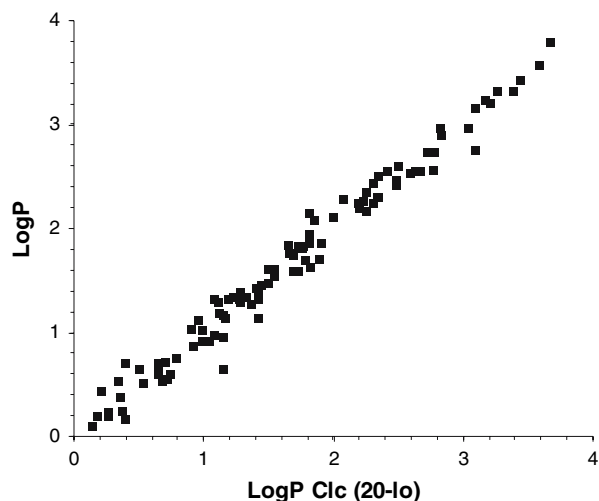$$C(Fat) = [-1.78911(8.6), 2.66236(14), 0.14340(1.5), -0.07069(1.4), 1.05050(6.0)].$$

Figure 8. Plot of Log P of $N = 80$ training plus $N = 20$ evaluated points with $\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$.

3.5. Two properties of halomethanes

The molar refraction, R_m, and boiling points, BP, of halomethanes (from [33, 34]) are collected in table 6. The S values of these halomethanes has been rescaled to the S value of the carbon atom in CF_4, *i.e.*, $S[C(CF_4)] = -5.5$. The model of these two properties has been achieved by the aid of a K_p-(p-odd) (i.e., with p odd-valued and $q = 1$) representation for the core electrons.

3.5.1. Molar refraction, R_m
Here $n_H/n_{ht} = 0.2$. For $f_\delta = 0$ the following optimal term has been detected

$$Z'(R_m) = (X + 1.3Y + 0.3^1\psi_{IS}) : F = 22805, \ r^2 = 0.9986, \ s = 0.3, \ N = 34.$$

Here, $X = (^0\chi^v + \chi_t)^{0.6}(F = 1092)$, and $Y = (0.97^0\psi_I - 1.15^0\psi_E)^{1.2}(F = 2959)$. With $f_\delta^8 \neq 0$ a Z term with improved model quality can be detected,

$$^{f8}Z(R_m) = (X + Y)^{0.9} : F = 52939, \ r^2 = 0.9994, \ s = 0.2, \ N = 34.$$

Where, $^{f8}X = [0.7^0\chi^v + 0.9(\chi_t)^{0.8}]^{0.5}(F = 1132)$, and $^{f8}Y_{podd} = |(0.87^0\psi_I - 1.1^0\psi_E)|^{1.2}(F = 3422)$. Bars here mean absolute value. The best fit-model equation for this property is,

$$R_m = 12.948 \cdot {}^{f8}Z(R_m)(52) - 11.954(84). \tag{23}$$

Let us leave-15-out (with an asterisk in table 6) but keeping $n_H/n_{ht} = 0.2$. For $N = 19$ the advantage of the ^{f8}Z term shrinks ($^{f8}F/^{f0}F = 1.05$, r^2 and s do not change),

Table 6
Molar refraction (cm^3 mol^{-1}) R_m, and boiling points, BP (K) of halomethanes, CH$_n$X$_{4-n}$.

CH$_n$X$_{4-n}$	R_m	BP	CH$_n$X$_{4-n}$	R_m	BP
*CH$_3$F	6.7	195	CBr$_4$	38.0	462
CH$_2$F$_2$	6.6	221	CF$_3$Br	14.3	214
CHF$_3$	6.5	189	CF$_2$Br$_2$	22.2	298
*CF$_4$	6.4	144	*CFBr$_3$	30.1	381
*CH$_3$Cl	11.7	249	CCl$_3$Br	29.3	378
CH$_2$Cl$_2$	16.6	313	*CCl$_2$Br$_2$	32.2	408
*CHCl$_3$	21.5	335	CClBr$_3$	35.1	433
*CCl$_4$	26.4	349	CH$_2$FBr	14.5	291
CFCl$_3$	21.4	297	*CH$_2$ClBr	19.5	342
CF$_2$Cl$_2$	16.4	243	CHFClBr	19.4	309
CHFCl$_2$	16.4	282	CHF$_2$Br	14.4	259
*CHF$_2$Cl	11.5	233	*CHFBr$_2$	22.3	338
CF$_3$Cl	11.4	192	CHCl$_2$Br	24.4	361
*CH$_2$FCl	11.6	264	CHClBr$_2$	27.3	394
CH$_3$Br	14.6	277	*CFCl$_2$Br	24.3	326
CH$_2$Br$_2$	22.4	370	*CFClBr$_2$	27.2	353
*CHBr$_3$	30.2	422	*CF$_2$ClBr	19.3	269

* Left-out values (see text).

$$^{f8}Z(R_m) = (X + 2Y) : F = 16723, \ r^2 = 0.9990, \ s = 0.3, \quad N = 19,$$
$$\boldsymbol{C} = [(6.241(129), -6.224(30)]$$

Here, $X = [0.7{}^0\chi^v + 0.9(\chi_t)^{0.8}]^{0.7}$, and $Y = |(0.83{}^0\psi_I - {}^0\psi_E)|$.

3.5.2. Boiling points, BP

Here $n_H/n_{ht} = 0.2$. The best description with no hydrogen perturbation and with a f_δ^6 perturbation are,

$$\{{}^0\chi, {}^0\psi_I, {}^S\psi_E\} : F = 4938, \ r^2 = 0.998, \ s = 3.6, \ N = 34,$$
$$^{f6}\{{}^0\chi, {}^0\psi_I, {}^S\psi_E\} : F = 5426, \ r^2 = 0.9982, \ s = 3.5, \ N = 34.$$

Leaving-15-out (with * in table 6) avoiding changes in n_H/n_{ht} renders the $f_\delta = 0$ and $f_\delta^6 \neq 0$ descriptions much closer together,

$$\{{}^0\chi, {}^0\psi_I, {}^S\psi_E\} : F = 1990, \ r^2 = 0.9974, \ s = 4.3, \ N = 19,$$
$$^{f6}\{{}^0\chi, {}^0\psi_I, {}^S\psi_E\} : F = 2046, \ r^2 = 0.9976, \ s = 4.2, \ N = 19,$$
$$\boldsymbol{C} = [-166.26(13), 189.92(23), 2.6726(4.6), 273.53(54)].$$

Table 7

Lattice enthalpies, ΔH_L°, binding enthalpies, ΔH_G° (at 298 K), calculated lattice enthalpy values, with K_p-(pp-odd) descriptors, $\Delta H_L^\circ(C)$, and calculated binding enthalpies with K_p-(p-odd) descriptors $\Delta H_G^\circ(C)$ (all energies are in kJ mol^{-1}).

MeX	ΔH_L°	$\Delta H_L^\circ(C)$	ΔH_G°	$\Delta H_G^\circ(C)$
LiF	1046	1042	769	766
NaF	928	910	687	673
*KF	826	830	588	614
RbF	792	784	563	572
CsF	756	752	566	545
*LiCl	861	891	641	654
NaCl	787	791	559	565
KCl	717	733	503	516
*RbCl	692	700	476	483
CsCl	668	678	475	463
LiBr	817	808	614	603
*NaBr	751	750	539	521
KBr	689	681	474	480
RbBr	665	656	453	454
*CsBr	649	638	440	437
LiI	761	765	570	571
NaI	703	692	505	493
*KI	648	653	449	455
RbI	629	632	421	432
CsI	610	617	418	417

3.6. The lattice, ΔH_L°, and binding, ΔH_G°, enthalpies, of metal halides

Alkali halides can be represented by deceptively simple graphs, i.e., •—•, in this case their non-valence χ indices are useless, as they are the same for all these compounds. Furthermore, some indices are redundant, i.e., $^1\chi^v \equiv \chi_t^v$, $^1\psi_I = {}^T\psi_I$, $^1\psi_E = {}^T\psi_E$, and so on [25]. The subsets of indices now are: $\{D^v, {}^0\chi^v, {}^1\chi^v\}$, $\{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E\}$, and $\{{}^0\chi_d^v, {}^0\psi_{Id}, {}^0\psi_{Ed}, {}^1\chi_s^v, 1\psi_{Is}, {}^1\psi_{Es}\}$. The two enthalpies of metal halides are collected in table 7, their values are taken from [34, 35] (see also [20]). Here $n_H/n_{ht} = 0$, this fact would allow to check the contribution of the electron-core description only. The S value was rescaled to $S[\text{Ba}(\text{BaF}_2)] = -3.083$.

The optimal description for the core electrons for the Lattice Enthalpy, ΔH_L°, of Metal Halides is a K_p-(pp-odd) description with p odd-valued, and $q = p$.

$$Z(L) = [(0.9X + 3Y)^{0.1} - 0.0002(^1\chi_s^v)^{0.4}] : F = 1705, \ r^2 = 0.990, \ s = 11,$$
$$N = 20, \ \mathbf{C} = [138086(41), -153703(41)],$$
$$X = (D^v)^{0.4}/[0.8D^v + 30(^0\chi^v)^{1.2}]^{0.8} : F = 1209, \ r^2 = 0.986, \ s = 14, \ N = 20,$$
$$Y = (^0\psi_I)^{0.01} : F = 220, \ r^2 = 0.925, \ s = 31, \ N = 20.$$

The $\Delta H_L^\circ(C)$ values calculated with the $Z(L)$ term, are in the third column of table 7. Compounds with an asterisk have now been left-out from the training set. With a training set of $N = 14$ things change only a bit, the best descriptor now is $[Z(L)]^{0.5}$ with $F = 2120$, $r^2 = 0.994$, and $s = 9.2$. Calculated values for both training ($N = 14$) and evaluation sets ($N = 6$) with this new term hardly change, only the last figure of $\Delta H_L^\circ(C)$ in table 7 is affected (± 1).

For the binding enthalpy, ΔH_G°, the optimal descriptor is a Z term belonging to the K_p-(p-odd) [$q = 1$, $p = $ odd] description for the core electrons. This *p-odd* term is also a good descriptor for both lattice and binding enthalpies together.

$$Z(B) = [X + 0.2Y + 0.02(^0\chi_d^v)]^{0.5:} F = 1058, \ r^2 = 0.982, \ s = 12, \ N = 20,$$
$$\mathbf{C} = [1342.54(33), -252.262(10)],$$
$$X = (D^v)^{0.5}/[D^v + 12(^0\chi^v)^{1.2}]^{0.3} : F = 797, \ r^2 = 0.978, \ s = 14, \ N = 20,$$
$$Y = (^0\psi_E)^3 : F = 104, \ r^2 = 0.854, \ s = 37, \ N = 20.$$

With this term the calculated binding enthalpies, $\Delta_G^\circ(C)$ of table 7 were obtained. With a training set of fourteen compounds only (those without asterisk in table 7) things hardly change.

Now, the K_p-(p-odd) $Z(B)$ term is a good descriptor for the lattice enthalpy as well. In fact,

$$\Delta H_L^\circ \text{ with } Z(B) : F = 1010, \ r^2 = 0.982, \ s = 15, \ N = 20,$$
$$\mathbf{C} = [1568.4(32), -170.55(5.9)].$$

With the $Z(B)$ term the calculated values of the plot of figure 9 for the lattice and binding enthalpies for a total of forty points have been obtained. In this figure is also shown the residual plot [Res = exp. − calc.], where spacing lines underline the $\pm s$, $\pm 2s$, and $\pm 3s$ intervals.

4. Conclusions

By the aid of three graph concepts, simple graph, general graph, and complete graph is possible to model properties and activities of the molecular world. Actually, with the last two concepts it is possible to encode every molecular structure in a specific way, inclusive of the bonded hydrogen atoms even if the
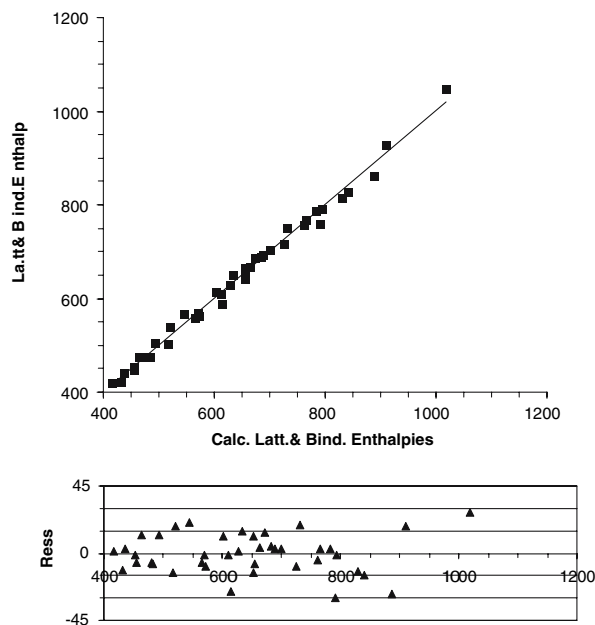
Figure 9. Plot of the the experimental versus calculated lattice and binding enthalpies, for a total of 40 points together with the corresponding residual plot for the lattice and binding enthalpies.

chemical graph is hydrogen depleted. In fact, the hydrogen contribution is encoded as a perturbation. The encoding of the core electrons by the aid of complete graphs, allow to retrieve an information, which is usually described, in atomic theory, by the principal quantum number, n. General graphs, instead, allow to encode any type of valence electrons, either σ-bonding, π-bonding, or non-bonding n-electrons. Parameter q of the δ^v algorithm [see equations (3) and (7)], which normally equals either 1 or p, might also be used as an optimization parameter, something like Randic's variable index [36, 37], but at another level.

The sets of properties here studied show that the hydrogen contribution can be property-dependent, as well as dependent on the number of compounds. The case of the composite boiling points of amines plus alcohols shows that composite classes might be considered a new sort of class, which in this case, seems independent of the number of compounds. The case of the lattice and bonding enthalpies of metal halides underlines again the intriguing importance of the odd complete graphs in encoding the core electrons.

References

[1] K.H. Rosen, *Discrete Mathematics and its Applications* (McGraw-Hill, New York, 1995).
[2] F. Harary, *Graph Theory* (2nd printing, Addison-Wesley, Reading, MA, 1971).
[3] M. Randić, J. Am. Chem. Soc. 97 (1975) 6609.

[4] M. Randić and N. Trinajstić, Croat. Chem. Acta 67 (1994) 1.

[5] R. Todeschini and V. Consonni, *The Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2000).

[6] D.G. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures* (Research Studies Press, Chichester, 1983).

[7] A.T. Balaban (ed.), *Chemical Applications of Graph Theory* (Academic Press, London, 1986),

[8] N.J. Turro, Angew. Chem. Int. Edn. Engl. 25 (1986) 882.

[9] R.B. King and D.H. Rouvray (Eds.), *Graph Theory and Topology in Chemistry* (Elsevier, Amsterdam, 1987).

[10] N. Trinajstić, *Chemical Graph Theory, 2nd ed.* (CRC Press, Boca Raton, 1992).

[11] J. Devillers and A.T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR/QSPR* (Gordon and Breach, Amsterdam, 1999).

[12] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis* (Wiley, New York, 1986).

[13] E. Estrada, Curr. Med. Chem. 8 (2001) 1573.

[14] L. Pogliani, Chem. Rev. 100 (2000) 3827.

[15] L. Pogliani, J. Chem. Inf. Comput. Sci. 34 (1994) 801.

[16] F. Harary and R. Read, *Proc. Graphs and Combinatorics Conference* (George Washington University; Springer: New York, 1977). Cited in J.D. Barrow, *The Book of Nothing* (Vintage Books: New York, 2000), p.155.

[17] L. Pogliani, J. Chem. Inf. Comput. Sci. 42 (2002) 1028.

[18] L. Pogliani, New J. Chem. 27 (2003) 919.

[19] L. Pogliani, J. Comput. Chem. 24 (2003) 1097.

[20] L. Pogliani, Int. J. Quant. Chem. 102 (2005) 38.

[21] L.B. Kier and L.H. Hall, *Molecular Structure Description. The Electrotopological State* (Academic press, New York, 1999).

[22] L. Pogliani, J. Comput. Chem. 27 (2006) 868.

[23] C. Yang and C. Zhong, J. Chem. Inf. Comput. Sci. 43 (2003) 1998.

[24] L. Pogliani, J. Phys. Chem. A 104 (2000) 9029.

[25] L. Pogliani, J. Chem. Inf. Comput. Sci. 41 (2001) 836.

[26] L. Pogliani, J. Mol. Struct. (Theochem) 581 (2002) 87.

[27] E. Besalu, J.V. de Juliàn-Ortiz, M. Iglesias and L. Pogliani J. Math. Chem. 39 (2006) 475.

[28] E. Besalu, J.V. de Julian-Ortiz and L. Pogliani, MATCH Commun. Math. Comput. Chem. 55 (2006) 281.

[29] L. Pogliani and J.V. de Julian-Ortiz, Chem. Phys. Lett. 393 (2004) 327.

[30] L. Pogliani, L. New J. Chem. 29 (2005) 1082.

[31] B. Ma, J.-H. Lii and N.L. Allinger, J. Comp. Chem 21 (2000) 813.

[32] M.L. Cargas, P.G. Seybold and M.E. Andersen, Toxicol. Lett. 43 (1988) 235.

[33] M. Laing, J. Chem. Educ. 78 (2001) 1544.

[34] R.J. Gillespie, J. Chem. Educ. 78 (2001) 1688.

[35] R.J. Schmid, J. Chem. Educ. 80 (2003) 931.

[36] M. Randić and S. Basak, J. Chem. Inf. Comput. Sci. 41 (2001) 614.

[37] M. Randić and M. Pompe, J. Chem. Inf. Comput. Sci. 41 (2001) 631.