

# From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors

Lionello Pogliani\*

*Dipartimento di Chimica, Università della Calabria, 87030 Rende, Italy*

*Received January 24, 2000*

## Contents

I. Introduction	3827	I. Modeling and Cis/Trans Isomerism	3854
A. Background on Molecular Connectivity	3827	J. Orthogonal Descriptors	3855
B. Challenges of Molecular Connectivity	3829	IV. Recent and Alternative Elaborations	3856
C. Some Recent Trends in Molecular Connectivity Modeling	3829	V. Conclusion	3856
II. Mathematical Tools and Algorithms	3830	VI. Acknowledgments	3857
A. The Molecular Connectivity Concept	3830	VII. Glossary	3857
1. The Molecular Connectivity Index $\chi$	3830	VIII. References	3857
2. The Medium-Sized Set of Molecular Connectivity Indices	3830		
B. The Graph Mass Index	3831		
C. Dimensionality Problem	3831		
D. The Cis/Trans Isomerism	3831		
E. Linear Combination of Connectivity Indices (LCCI)	3832		
F. Correlation Problem and Randić's Orthogonalization Procedure	3834		
G. Special Molecular Connectivity Indices	3835		
III. Modeling Properties of Different Classes of Compounds	3836		
A. Amino Acids	3837		
1. Side-Chain Molecular Volume	3837		
2. Isoelectric Point	3838		
3. Crystal Density	3838		
4. Specific Rotation	3839		
5. Solubility	3840		
B. Purine and Pyrimidine Bases	3841		
1. Solubility	3841		
2. Singlet Excitation Energy, Oscillator Strength, and Molar Absorption Coefficient	3842		
C. Solubility of the Mixed Class of [AA + PP]	3843		
D. Alkanes	3843		
1. Melting Points	3844		
2. Motor Octane Number	3844		
E. Four Properties of Organophosphorus Compounds	3845		
F. Lattice Enthalpy of Inorganic Salts	3847		
G. Unfrozen Water Content of the Mixed Class of Amino Acids and Metal Chlorides	3848		
H. Random Organic Solvents	3849		
1. Boiling Point	3849		
2. Refractive Index	3851		
3. Density	3852		
4. Cutoff UV Values	3852		
5. Dipole Moment	3854		

## I. Introduction

### A. Background on Molecular Connectivity

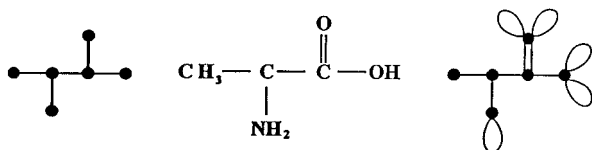
Many structure–property studies use graph theoretical indices that are based on the topological properties of a molecule viewed as a graph. The main goal of topology is always toward the general, i.e., toward relations and theorems that apply to any space, without reference to measurements or any kind of metrics. Thus, atoms embedded in a graph will no longer be Euclidean points but any unspecified thing to which we can apply these relationships meaningfully. This kind of generalization is natural to mathematics; six pairs are a dozen, whether loaves, or atoms, or days. A graph, in a topological context becomes, thus, the abstracted essence of the properties of traversing and joining, and conversely, a molecule is a concrete manifestation of an abstracted graph where the Euclidean metric together with the notions of congruence (see Glossary) and similarity (see Glossary) go by the board. A graph  $G$  can be defined as a set of  $V$  vertexes with a set of  $E$  edges that connect these vertexes, i.e.,  $G = (V, E)$ . Thus, a graph is determined by the set of vertexes and by the set of edges joining the vertexes and not by the particular appearance of the configuration. A chemical graph is a graph where atoms and bonds are represented by vertexes and edges, respectively. Clearly, double bonds or lone-pair electrons cannot be fitted by a graph; for this reason, pseudographs are also used to represent organic molecules. A pseudograph  $G = (V, E)$  is the most general type of graph, since it may contain multiple edges between pairs of vertexes and loops, which are edges from a vertex to itself.<sup>1</sup> Every graph is, thus, a pseudograph, but not every pseudograph is a simple graph. Some mathematicians, in fact, reserve the term 'simple graph' for a graph with no multiple edges and loops.

\* Fax: +39-(0)984-492044. E-mail: lionp@unical.it.

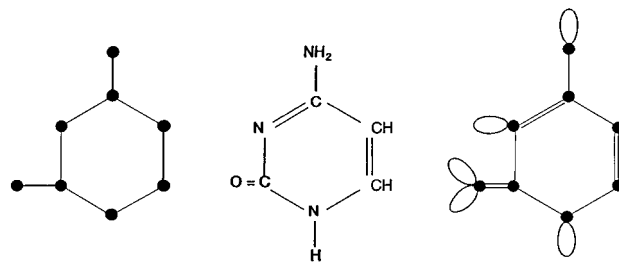


Lionello Pogliani graduated in Chemistry in 1969 at the University of Firenze. He received his postdoctoral training at the Department of Molecular Biology of the C.E.A. (Centre d'Etudes Atomiques) of Saclay, France, at the Physical Chemistry Institute of the Technical and Free University of Berlin, and at the Pharmaceutical Department of the University of California, San Francisco. During this last stage he was awarded the GM Neural Trauma Research Award. He joined the Faculty of the University of Calabria and was appointed Associate Professor in 1987. He was visiting professor at the Centro de Química-Física Molecular of the Technical University of Lisbon. He contributed more than 100 papers and made more than 40 symposium presentations in different fields of physical chemistry, notably, in chemical and medical applications of NMR, chemical applications of graph theory, and educational chemistry (mainly in chemical kinetics and thermodynamics). He has also co-authored contributions in quantum chemistry, history and meaning of numbers, electrochemistry, dimensional analysis, and theory of gases.

An important characteristic of graphs and pseudographs is the degrees of their vertexes, i.e., the number of edges incident with each vertex, where loops are to be considered self-incident edges. The degree of a vertex suggests one of the chemical concepts of valence, and in fact, in chemical graph theory it is often used with this meaning (see refs 2–8 and references therein). However, while the degree of a vertex in a simple chemical graph denotes the connections of the chemical vertexes, the degree of a vertex in a chemical pseudograph is directly related to the chemical concept of valence, with loops and multiple edges simulating lone pairs and  $\pi$  bonds, respectively. From what has been said, it is evident that organic molecules are well suited to be represented by chemical graph or pseudographs whose mathematical properties can be used in QSAR/QSPR studies. The most widely considered chemical graphs and pseudographs are hydrogen-suppressed graphs and pseudographs, and from now on, graphs and pseudographs will be assumed to be hydrogen-suppressed graphs and pseudographs. In Figures 1 and 2 are reported the molecules of the amino acid alanine and of the base cytosine and their corresponding hydrogen-suppressed graph (left) and pseudograph (right), respectively.



**Figure 1.** Molecule of the amino acid Ala and its corresponding hydrogen-suppressed graph (left) and pseudograph (right).



**Figure 2.** Molecule of the base cytosine and its corresponding hydrogen-suppressed graph (left) and pseudograph (right).

A common way to represent chemical graphs is to use adjacency matrixes.<sup>3</sup> An adjacency matrix of a graph is a  $n \times n$  0,1 matrix with 1 as its  $(i,j)$ th entry when vertex  $v_i$  and vertex  $v_j$  are connected and 0 as its  $(i,j)$ th entry when they are not connected, i.e.,  $a_{ij} = 1$  if  $\{v_i, v_j\}$  is an edge of the graph and  $a_{ij} = 0$  otherwise. The symmetric square 0,1 matrix of eq 1 is the connection matrix representing the hydrogen-suppressed chemical graph of the amino acid Ala of Figure 1. This and the matrix of eq 2 were built using the following ordering of the atoms: C<sub>α</sub>, C<sub>β</sub>, N, O (of the OH group), O (of the C=O group), and C (of the CH<sub>3</sub> group). As self-adjacencies are not allowed in normal chemical graphs, all entries along the main diagonal are zero.

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (1)$$

The symmetric square adjacency matrix for the hydrogen-suppressed chemical pseudograph of Ala is given by matrix of eq 2

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 5 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (2)$$

Relative to matrix 1, some entries along the diagonal are seen to be different from zero, since in pseudographs self-connections (loops), which mimic the presence of lone-pair electrons, contribute twice to the degree of a vertex and multiple connections, which mimic the presence of  $\pi$  bonds, contribute once for each multiplicity to the degree of a vertex.<sup>1</sup> Thus, the single  $\pi$  bond of the carboxyl carbon of the C=O group (first row) contributes  $a(1,1) = 1$  in the diagonal entry; the nitrogen of the amino group with a lone-pair electron or self-connection has entry  $a(3,3) = 2$ ; the oxygen of the C–OH group with its two lone pairs has entry  $a(4,4) = 4$ ; the oxygen of the C=O group with two lone pairs and a  $\pi$  bond has entry  $a(5,5) = 5$ . There are other ways to write the adjacency matrix of a pseudograph,<sup>1</sup> but the one given is the most useful for computational purposes,

as matrix 1 can be obtained from matrix 2 with an algorithm that zeros the diagonal.

The vertex degree or valence  $\delta_i$  and the pseudover-  
tex degree or valence  $\delta^v_i$  of an atom can now be  
computed from matrix 2 in the following way:  $\delta_i$  is  
equal to the sum of the elements in row  $i$  (or column  
 $i$ ) in the diagonal-suppressed adjacency matrix (or  
equal to the sum of the elements in adjacency matrix  
1);  $\delta^v_i$  is the sum of all the elements of row  $i$  of matrix  
2. It can be shown that the number of nonzero entries  
in matrixes 1 and 2 is twice the number of bonds or  
connections in the corresponding graph and pseudo-  
graph. Thus, let  $G = (V, E)$  be a graph (pseudograph)  
with  $E$  edges and  $V$  vertexes, then  $2e = \sum \delta_i$  ( $2e = \sum \delta^v_i$   
in pseudographs) where  $e$  is the number of edges.  
Since an edge is incident with exactly two vertexes,  
it contributes twice to the sum of the degrees of the  
vertexes. This result is sometimes referred to as  
*hand-shaking theorem*<sup>1</sup> because of the analogy be-  
tween an edge having two end points and a hand-  
shake involving two hands. Such a peculiarity of  
graphs is encoded in the symmetric form of matrixes  
1 and 2. For pairs of enantiomers such as in amino  
acids or sugars, the given sum can be seen either as  
the sum of the vertex degrees of the L form or as the  
sum of the vertex degrees of the D form, i.e.,  $2e_D =$   
 $2e_L$  as  $e_L = e_D$ . In fact, invariants derived from  $\delta_i$  and  
 $\delta^v$  numbers are unable to distinguish between D and  
L forms of enantiomeric pairs as  $\delta_L = \delta_D$ , and  $\delta^v_L =$   
 $\delta^v_D$ . We will come back to this topic and discuss some  
consequences of this theorem later on. From numbers  
 $\delta$  and  $\delta^v$ , following certain rules laid down by the  
molecular connectivity theory, it is possible to derive  
a whole set of molecular connectivity (MC) indices.  
For an amino acid like Ala, e.g., it is possible to derive  
up to 20 MC indices.

## B. Challenges of Molecular Connectivity

The type of quantitative structure–property rela-  
tionships (QSPR) in which we are interested is based  
on molecular connectivity invariants (MCI). Thus, a  
molecular connectivity modeling of properties is that  
modeling which successfully relates these invariants  
or indices to specific properties of a class of com-  
pounds; the more the properties and classes of  
molecules we are able to model with molecular  
connectivity indices, the more these indices assume  
the character of property meters. This result makes  
us confident that our conclusions with respect to  
these invariants are quite general and that the  
equation  $P = f(\text{MCI})$  is valid for all properties which  
occur in the world. We need this generalization so  
that we can use it for the prediction of properties of  
molecules which have not yet been determined. We  
cannot, of course, be absolutely certain that every  
prediction will be correct, but confidence in gener-  
alization grows with every successful prediction.

A great deal of successful QSPR and QSAR studies  
are based on hydrogen-suppressed chemical graphs  
for which graph theoretical indices, the molecular  
connectivity  $\chi$  indices, have been defined and further  
refined for the last 20 years<sup>9–29</sup> into a self-consistent  
theoretical frame known as the molecular connectiv-  
ity theory (MCT) or molecular connectivity model

(MCM). Throughout these years, interesting contri-  
butions to this theory have been made by many  
groups scattered around the world<sup>29–53</sup> (for a detailed  
bibliography before 1986, see references in ref 3). The  
references cited are certainly not exhaustive but  
nevertheless indicate the rich development under-  
gone by molecular connectivity and related topologi-  
cal concepts. Reference 8 is an interesting contribu-  
tion about the story of chemical graph theory. The  
main challenge of MC modeling can be phrased into  
the following way: *All predictions can be reached  
using nothing more than pencil and paper.*<sup>54</sup> Obvi-  
ously, more than pencil and paper are needed to  
model physicochemical properties of compounds, but  
what these words are stressing is the fact that the  
modeling of properties can be done by the aid of a  
few elementary, direct, and easily understandable  
mathematical tools.

## C. Some Recent Trends in Molecular Connectivity Modeling

With the recent introduction of four new molecular  
connectivity indices, the sum-delta, the valence sum-  
delta,<sup>55</sup> and the total and valence total<sup>25</sup> indices,  
it has been possible to define a medium-sized set of  
eight molecular connectivity indices  $\{\chi\}$  which seems  
able to offer both a satisfactory model of many  
physicochemical properties of many classes of com-  
pounds and reduce the dimension of the otherwise  
severe combinatorial problem required to derive  
optimal linear combinations of molecular connectivity  
indices, LCCI.<sup>55–71</sup> With this medium-sized set of  
molecular connectivity indices it has been possible  
to model the properties of natural amino acids, purine  
and pyrimidine bases, alkanes, organic phospho-  
derivatives, unsaturated organic compounds, inor-  
ganic salts, mixed classes of amino acids plus pep-  
tides, amino acids plus inorganic salts, amino acids  
plus purine and pyrimidine bases, and so on.<sup>55–64</sup> The  
modeled properties include the pH at the isoelectric  
point, the longitudinal relaxation time, the side-chain  
molecular volume, the specific rotation, the solubility,  
the crystalline density, the melting points, the motor  
octane number, the retention index for paper chro-  
matography, enthalpy values and hydration proper-  
ties, etc. It was also attempted to define an index for  
the cis/trans isomerism in unsaturated compounds.<sup>58</sup>

Subsequently, the need to model with fewer indices  
and to extend the applicability of the method to  
highly heterogeneous classes of compounds has led  
to the development of nonlinear higher-level molec-  
ular connectivity terms,  $X = f(\chi)$ , and semiempirical  
terms,  $X = f(\chi, P_{\text{exp}})$ , where  $P_{\text{exp}}$  is an experimental  
property different from the modeled property.<sup>65–70</sup>  
Further, it has been seen that  $D$  and  $D^v$  indices,  
which can be derived from the hand-shaking theorem  
(see section A), led to the development of the concept  
of graph mass, a concept with its own identity and  
is not redundant with the concept of molar mass.<sup>71</sup>

Another interesting procedure of notable impor-  
tance in molecular connectivity modeling is the  
orthogonalization procedure to derive from normal  
MCIs the corresponding orthogonal molecular con-  
nectivity indices. It is a rather general procedure



which allows the construction of stable relationships, the derivation of dominant descriptors and enhancement of reliability of the relationship.<sup>26–29</sup> Using this approach, ordered orthogonalized connectivity bases have been proposed and successfully tested with amino acids.<sup>44</sup> The recent introduction of the inverse imaging procedure, which consists of building molecules from modeling equations, allowed the path from graph to the modeling equation to be inverted.<sup>36–37</sup>

## II. Mathematical Tools and Algorithms

### A. The Molecular Connectivity Concept



#### 1. The Molecular Connectivity Index $\chi$

The term molecular connectivity was adopted by Kier and Hall in 1975<sup>10,11</sup> in connection with a mathematical algorithm proposed by Randić<sup>9</sup> 25 years ago. This algorithm can be considered the first concept of the MC theory. It has long been known that branched-chain alcohols and hydrocarbons generally have lower boiling points and higher solubility than the corresponding straight-chain isomers. Randić not only suggested a simple computational method for correlating a physicochemical data with such topological characteristics as branching, but the proposed method did show a quantitative character that up to then had failed in previous works, like, e.g., in the works of Hosoyea and Smolenski.<sup>72,73</sup> Randić's branching index, which is generally known as the path-one molecular connectivity index or the first-order molecular connectivity index ( ${}^1\chi$ ), was defined as  $\chi = \Sigma(m \cdot n)^{-0.5}$ , where the summation includes one term for each edge in the hydrogen-suppressed chemical graph and the variables  $m$  and  $n$  are the valencies of the adjacent points joined by each edge.

Let us show in few words how Randić did attack the problem to find, by the aid of topological considerations, a numerical solution for the inequalities in the boiling points of alkane isomers. Consider the five hexane isomers *n*-hexane (6), 3M-pentane (3M5), 2M-pentane (2M5), 2,3MM-butane (23MM4), and 2,2MM-butane (22MM4) and the inequalities that follow their boiling points:

$$\begin{aligned} 2(1,2) + 3(2,2) &> 2(1,2) + (1,3) + 2(2,3) > (1,2) + \\ 2(1,3) + (2,2) + (2,3) &> 4(1,3) + (3,3) > (1,2) + \\ &3(1,4) + (2,4) \end{aligned}$$

Here  $i(m,n)$  represent the number  $i$  of bond type where the end vertexes have valences  $m$  and  $n$ , respectively. Randić solved the problem to find numerical values for  $(m,n)$  that would represent a solution of given inequalities, with the suggestion that the contribution for bond type  $(m,n)$  is  $1/\sqrt{(m \cdot n)}$ , which, when summed over all bonds, defined the connectivity index. The connectivity index values for the given hexane isomers were thus found to be  $2.9142 > 2.8081 > 2.7701 > 2.6427 > 2.5607$ , i.e.,  $\chi(6) > \chi(3M5) > \chi(2M5) > \chi(23MM4) > \chi(22MM4)$ . It is interesting to point out how the connectivity indices of the different isomers parallel the boiling

point of the five alkanes. Randić's original paper also alludes to the concept of extended connectivity, which acknowledges the presence of more distant neighbors. The natural extension of this view to define additional indices for subgraphs corresponding to paths with lengths greater than one (including the zeroth-order  $\chi$  index), to clusters, to path-clusters, to cycles, and to define indices for pseudographs which are able to mimic the presence of multiple bonds and lone-pair electrons was undertaken by Kier and Hall.<sup>10–16</sup> Thus, e.g., the second-order molecular connectivity index,  ${}^2\chi$ , stretches over three linear-contiguous vertexes,  $\bullet-\bullet-\bullet$ , and the third-order molecular connectivity index,  ${}^3\chi$ , stretches over four linear-contiguous vertexes, i.e.,  $\bullet-\bullet-\bullet-\bullet$ , and are defined as  $\Sigma(mnp)^{-0.5}$  and  $\Sigma(mnpq)^{-0.5}$ , respectively. The summation here is over the overall number of subgraphs into which the graph can be partitioned. The third- and fourth-order cluster and path-cluster indices stretch instead over  and  subgraphs, respectively, and so on. In this way, it can easily be understood that even for a small molecule like Ala, up to 20  $\chi$  and  $\chi^v$  indices can be defined, giving rise to a huge combinatorial problem as we shall see later on. Nearly in the same period Balaban proposed an analogous index to  ${}^1\chi$  but based on the distance matrix (see Glossary) of a graph, the so-called  $J$  index.<sup>74</sup>

Concerning the first-order molecular connectivity index, it has recently been demonstrated<sup>49</sup> that a relationship exists between molecular orbital theory and molecular topology and that the  ${}^1\chi$  index reproduces the values obtained for the  $\pi$  electronic energies as well as for the resonance energies calculated with the HMO method for conjugated alternant hydrocarbons. This parallelism between molecular connectivity and Hückel theory of conjugated molecules is quite interesting as both are based on the topology of a molecular framework ( $\pi$  network only for HMO) rather than its geometry; further, they share the same simplicity and limited computational effort.

#### 2. The Medium-Sized Set of Molecular Connectivity Indices

In the remainder of this article we will mainly be interested in a limited set of molecular  ${}^i\chi$  and valence  ${}^i\chi^v$  connectivity indices of hydrogen-suppressed chemical graphs that can easily be computed by the aid of adjacency matrixes 1 and 2. This medium-sized set of eight molecular connectivity indices has been systematically and successfully used during the past years both in deriving powerful linear combinations and even more powerful molecular connectivity terms, where powerful means the good quality of the achieved modeling. As already explained, the basic parameter for these and other molecular connectivity indices is the connectivity degree or valence of a vertex  $\delta$  in a molecular chemical graph and  $\delta^v$  in a pseudograph. From the first row of matrixes 1 and 2, respectively, it is possible, for example, to obtain for the carboxyl carbon of Ala  $\delta(C_o) = 3$  and  $\delta^v(C_o) = 4$ . For higher row atoms, e.g., for the third and further quantum level atoms, like Na, Mg, Cl, Br, I, P, S, ..., the delta values,  $\delta$ , are computed from their chemical graphs

while the  $\delta^v$  values are computed by the aid of eq 3, as, in this case, it has been proposed to take into account the core electrons also.<sup>13</sup> Clearly, here the concept of pseudograph is of no help.

$$\delta^v = Z^v / (Z - Z^v - 1) \quad (3)$$

Here,  $Z$  is the atomic number and  $Z^v$  is the number of valence electrons. Thus, e.g., for the following atoms, the following  $\delta^v$  values are normally used

$$\begin{pmatrix} \text{Li} & \text{Na} & \text{K} & \text{Rb} & \text{Cs} & \text{Be} & \text{Mg} & \text{Ca} & \text{Sr} & \text{Ba} & \text{Cu} & \text{Cl} & \text{Br} & \text{I} & \text{S} \\ 1 & 1/9 & 1/17 & 1/35 & 1/53 & 2 & 2/9 & 2/17 & 2/35 & 2/53 & 2/26 & 7/9 & 7/27 & 7/45 & 5/9 \end{pmatrix}$$

For the amino acid Cys, a  $\delta^v(\text{S}) = 0.56$  has been chosen, while for the organophosphorous compounds a  $\delta^v(\text{P}) = 2.22$  has been chosen.<sup>13</sup>

The eight indices, which will be used throughout this review, are now presented and succinctly discussed. The zeroth- and first-order molecular connectivity indices<sup>13</sup>

$${}^0\chi = \sum(\delta_i)^{-0.5} \quad (4)$$

and

$${}^1\chi = \sum(\delta_i \delta_j)^{-0.5} \quad (5)$$

Their dimensions are  $[\delta^{-0.5}]$  and  $[\delta^{-1}]$ , respectively. The sum in eqs 4 and 5 run over all  $N$  vertexes (atoms) and all edges ( $\sigma$  bonds) of the molecular graph, respectively. Replacing, in these and following equations,  $\delta$  with valence  $\delta^v$  of the corresponding pseudograph, we obtain the corresponding valence molecular connectivity indices,  $\chi^v$ , where the sum is to be taken again over all vertexes and over all single edges, i.e., multiple edges are taken only once.

In 1988 Needham et al.<sup>25</sup> introduced a quite useful molecular connectivity index, the total structure molecular connectivity index of a chemical graph over all  $N$  vertexes,  $\chi_t$ ,

$$\chi_t = (\delta_1 \delta_2 \dots \delta_N)^{-0.5} \quad (6)$$

with dimension  $[\delta^{-N/2}]$ . Replacing  $\delta$  with  $\delta^v$ , the corresponding total valence molecular connectivity index,  $\chi_t^v$ , for the pseudograph is retrieved.

Recently, the following sum-delta (and valence sum-delta,  $D^v$ ) molecular connectivity index has been introduced,<sup>55</sup> where the sum runs over the vertexes of the chemical graph (or pseudograph)

$$D = \sum \delta_i \quad (7)$$

Indices  $D$  and  $D^v$  are strictly related to the *hand-shaking* theorem, which was discussed in a preceding section. These eight indices, the  $\chi$  and  $\chi^v$  indices, build the following medium-sized set of molecular connectivity indices

$$\{\chi\} = \{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t^v\} \quad (8)$$

## B. The Graph Mass Index

The  $D$  and  $D^v$  indices, which have as their basis the *hand-shaking* theorem, a theorem which concerns a property of the graph or pseudograph as a whole

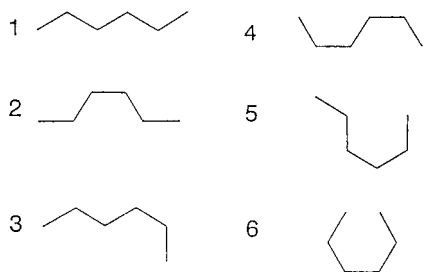
entity, have recently been called graph and pseudo-graph mass indices, respectively.<sup>71</sup> This, does not mean that they are good descriptors of the molar  $M$  mass of molecules. On the contrary, it has been demonstrated that they are autonomous indices which can sometimes and only sometimes offer a good description of the molar masses of certain classes of molecules. Other indices are more effective descriptors of the molar mass of organic and nonorganic compounds. In fact, the simulation of the molar mass of many classes of compounds has shown that normally  ${}^0\chi$  and  ${}^1\chi$  are the best descriptors for  $M$ , while graph,  $D$ , and pseudograph,  $D^v$ , mass descriptors, are normally poor descriptors of  $M$ . It seems that as soon as the graph and pseudograph do not superpose anymore, as it does in alkanes, both  $D$  and  $D^v$  acquire an autonomous descriptive dimension, with minimal superposition with  $M$ . Theoretical graph mass parameters  $D$  and  $D^v$  cannot, thus, be exchanged with  $M$  or be considered redundant with it.

## C. Dimensionality Problem

Graphs have been considered two-dimensional (2-D) objects<sup>75-78</sup> even if very important information on 3-D structure is implicit in the set of connections contained in the chemical graph and encoded in the molecular connectivity indices. In fact, graphs containing only  $-\text{CH}_2-$  and graphs containing quaternary,  $>\text{C}<$ , and tertiary,  $-\text{CH}<$ , carbons are evidently designing a cyclic molecule and a molecule with consistent steric crowding, respectively. These differences, as well as differences in bonding (see pseudographs), are detected by the different molecular connectivity indices. However, even if graphs encode some 3-D information implicitly, it is not unworthy to elucidate again what a graph is. Fundamentally, the term graph refers to a mathematical object that represents the structure of the various interconnections of a molecule. Being a collection consisting of two sets, the set of  $V$  vertexes and the set of  $E$  edges that connect these vertexes, it is essentially a statement of objects and their relations. A graph is only determined by the set of vertexes and by the set of edges joining these vertexes, and it is not a Cartesian representation of a structure; thus, the spatial dimension of a graph is a quite elusive entity, as dimension in graph theory does not have the same meaning as the concept of dimension in physics. In light of these considerations, the claim that molecular graphs do not encode 3-D structures has a minor value. A more rigorous definition of graph states that the term graph and one-dimensional complex are synonymous and that they are a set of zero-dimensional objects or vertexes and a set of one-dimensional objects or connections together with a rule which assigns to each connection two distinct vertexes.<sup>1</sup>

## D. The Cis/Trans Isomerism

The algorithm to encode this isomerism,<sup>58</sup> which normal graph theory does not allow to be encoded, is based on the first-order  ${}^1\chi$  index. The procedure to derive it starts by considering the different hexatriene



**Figure 3.** Six conformers of hexatrienes.

conformational isomers of Figure 3 as embedded on an idealized graphite grid as already suggested for three-dimensional structures.<sup>79,80</sup> Now, we notice that if we increase, in chemical graphs 2–6, the  $\delta$  values of the two cis points by 1, connecting them by an edge, they can form four-membered rings and, precisely, one four-membered ring for graphs 2 and 3, two for graphs 4 and 5, and three for graph 6. The six-membered rings of the graphite grid can be thought of as stable embedding forms. The newly formed rings, a sort of virtual rings with the raised  $\delta^r$  valence of the newly connected cis points, can be considered virtual ring fragments. The  $\delta$  vector of the virtual ring which includes the connected cis points can be used to define the following cis  ${}^n\chi_c$  connectivity index, where  $c$  stands for cis and (here) the embedded ring form has  $n = 4$

$${}^n\chi_c = \Sigma(\delta^r_1 \delta^r_2 \dots \delta^r_n)^{-6/n} \quad (9)$$

Exponent 6 is contributed by the number of edges of the embedding rings of the graphite grid. The numbering at the different delta values denotes the different vertexes of the virtual ring fragment, and the limits of the summation are the number of four-membered-ring fragments that can be formed inside the embedding rings. If  $\delta = \delta^r = 2$ , eq 9 has a constant solution for every  $n$  ( ${}^n\chi_c = 0.01563$ ). This new index, which is reminiscent of the  ${}^n\chi_{ch}$  index of Kier and Hall,<sup>13</sup> (i) has no meaning for trans structures, where no cis points are present, and furthermore (ii) cannot encode branching. To short circuit these limitations, a global molecular  $\chi_{ct}$  connectivity index has been defined in a way that its maximum value should correspond to the all-trans isomer while it decreases the more 'cis-rich' and the more branched the chain gets; i.e., the new index should include the characteristics of the  ${}^1\chi$  index, which decreases with increasing branching, and of the  ${}^n\chi_c$  index, which increases with increasing cis-rich compounds; it should look like the index of eq 10

$$\chi_{ct} = {}^1\chi - {}^n\chi_c \quad (10)$$

Here,  $n = 4$ , and  ${}^1\chi$  has been defined in the preceding section. Then for all-trans graphs for which  ${}^n\chi_c = 0$ , the new index simplifies into  ${}^1\chi$

$$\chi_{ct} = {}^1\chi = \Sigma(\delta_i \delta_j)^{-1/2} \quad (11)$$

The given definition of  $\chi_{ct}$  allows us to encode not only different cis and trans olefins, but also different types of chemical graphs and pseudographs such as al-

kanes and alkenes. The  ${}^n\chi_c$  index alone, instead, can only be used among specific sets of olefins such as the given trienes but is useless, for example, to differentiate between *cis*-2-butene and *cis*-2-octene. This new index,  $\chi_{ct}$ , proved to be very useful in modeling some physicochemical properties of olefins, such as the boiling points, the refractive index, the density, and the molar refractivity.<sup>58</sup>

## E. Linear Combination of Connectivity Indices (LCCI)

The method of linear combination of connectivity indices is a powerful tool used to model different properties of different classes of compounds with no recourse to empirical, quantum mechanical, or other kind of 'external' parameters, which are in some cases used in connection with  $\chi$  indices to improve the modeling. The validity of this powerful method had already been recognized and proved for the simulation of properties of alcohols, alkanes, quaternary salts, polyaromatic hydrocarbons, and octane isomers.<sup>13,81</sup> This method starts with the choice of a set of optimal molecular connectivity indices, among which, through a combinatorial technique and following statistical criteria that will be introduced later on, the best modeling indices are sorted out. The rationale for the choice of only eight molecular connectivity indices (see eq 8) is to keep under control the combinatorial problem that arises in the choice of the best modeling combination of  $\chi$  indices. Two standard combinatorial techniques can be used for the search of the best combination of indices belonging to the  $\{\chi\}$  set:<sup>62,65,67</sup> the *forward selection* technique and the *full combinatorial* technique. The forward selection technique, also known as the '*greedy algorithm*', is a sequential method for index selection based on the notion that connectivity indices should be inserted one at a time until an optimal LCCI is obtained. This method spans a subspace of the full combinatorial space. The procedure is as follows: (a) choose the best single  $\chi$  index, (b) then choose the next best  $\chi$  index of the  $\{\chi\}$  set that further enhances the description of the property, in the presence of the previous index, (c) and so on until the description starts to worsen with the introduction of the next  $\chi$  index of the set. One of the advantages of this algorithm is that it gives an *ordered* list of descriptors that can be used in deriving orthogonalized descriptors. The more elaborate and precise full combinatorial technique, instead, searches the full combinatorial space spanned by the indices of the set to extract the optimal LCCI. Experience has shown that the first method offers an adequate alternative to the more precise but more time-consuming full combinatorial method. A drawback of this second combinatorial algorithm results from the fact that in a stepwise regression it produces results in which at different successive steps, in addition to a new descriptor, old descriptors can be altered and a completely different combination of descriptors may emerge.<sup>82</sup> The difference in terms of searched combinations between the two combinatorial techniques is shown in Table 1, where the overall number of possible combinations for both procedures with grow-



**Table 1. Number of Possible Combinations for  $m$  Indices with the Forward Selection (fs) and Full Combinatorial (fc) Technique**

$m$	fs	fc
2	3	3
3	6	7
4	10	15
5	15	31
6	21	63
7	28	127
8	36	255
9	45	511
10	55	1023
20	210	1 048 575
25	325	33 554 431
30	465	1 073 741 823

ing number of indices is summarized. From this table it can be seen that the complete combinatorial space practically explodes with a growing number of descriptors while the forward selection technique looks quite tractable up to nearly 10 indices. It should be added that 30 normal plus valence indices per molecule (like Phe) is no extraordinary number of indices, keeping in mind that  $C_{7,8}$  alkanes, which do not have valence-type indices, can have up to 17 molecular connectivity indices per molecule.<sup>83</sup> Thus, our set of eight indices will give rise to 36 forward selection combinations and 255 full combinations, which should be searched to find the best LCCI, a task that a normal PC or a programmable pocket calculator can handle.

Assuming that the relationship between properties,  $P$ , and molecular connectivity indices is linear, then the modeling equation is given by the following dot product modulus

$$P = |\mathbf{C} \cdot \chi| \quad (12)$$

Here,  $P$  is the calculated property, the row vector,  $\mathbf{C} = (c_D, c_{Dv}, c_0, c_{0v}, c_1, c_{1v}, c_t, c_{tv}, c_U)$ , is the vector of the coefficients determined by a least-squares procedure, and the column vector,  $\chi = (D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t^v, U_0)$  is the vector of the connectivity descriptors. The aforementioned selection techniques in choosing the best descriptors will determine which coefficients,  $c_k$ , of vector  $\mathbf{C}$  are zero. The multivariate regression can be regarded as a linear combination of connectivity indices where the constant term can be considered to multiply the unitary index,  $U_0 = \chi^0 \equiv 1$ . Even if  $P$  is not always a linear function of  $\chi$ , it is nevertheless a linear function of the  $c_k$  coefficients. If  $\chi$  is a  $m \cdot n$  matrix (where,  $n$  = number of compounds), then  $\mathbf{P}$  is a column vector of the entire class of compounds. The bars in eq 12 stand for the absolute value to get rid of negative  $P$  values with no physical meaning and simultaneously enhance the description of the property. As  $\chi$  indices are dimensionless numbers holding no unit, they are, strictly speaking, able to describe only dimensionless parameters. To avoid this pitfall, the scalar property  $P$  should be read as  $P/P^\circ$ , where  $P^\circ$  has the same units as  $P$  but unitary value, following a well-known algorithm of quantity calculus which allows treatment of  $P/P^\circ$  as a dimensionless quantity.<sup>84,85</sup> The reader should be reminded that every time he reads  $P$ ,  $P/P^\circ$  is intended, even with the legends of the figures.

The different combinations of indices are controlled and sorted by the aid of two statistical parameters: (i) the quality factor,  $Q = r/s$ , where  $r$  = correlation coefficient and  $s$  = standard deviation of estimates, and (ii) the variance  $F$  (Fischer) ratio,  $F = fr^2/[(1 - r^2)m]$ , where  $f$  = number of degrees of freedom,  $f = n - m - 1$ ,  $m$  = number of variables, and  $n$  = number of data points. For every optimal combination,  $r$  and  $s$  will be given also. It should be noticed that  $Q$  and  $F$  values have been derived with original calculated  $r$  and  $s$  with five digits. The modeling was taken to be optimal when  $Q$  reached a maximum together with  $F$ , even if slightly nonoptimal  $F$  values have normally been accepted. Instead, a significant decrease in  $F$  with the introduction of one additional variable, with increasing  $Q$ , due to a decreasing  $s$ , could mean that the new descriptor has endangered the statistical quality of the combination, which nevertheless can again improve with the ulterior introduction of a more convincing descriptor. For every index of a LCCI equation, the fractional utility (i.e., the inverse of the fractional error),  $u_k = |c_k/s_k|$  as well as the average fractional utility  $\langle u \rangle = \sum u_k/m$  will be given. The statistical parameter,  $u_k$ , will allow detection of the paradoxical situation of a LCCI with a good predictive power but with a poor utility at the level of some or all of its coefficients.<sup>41</sup> This paradox can, in part, be removed with the introduction of orthogonal molecular connectivity indices. It should be noticed that  $Q$ ,  $r$ , and  $s$  values as well as  $\langle u \rangle$  and  $u_k$  values, even if they seem redundant, offer a more direct view of the statistical behavior of a modeling and can also be used as a check for eventual printing errors.

Sometimes some compounds have a property with a negative value, such as the specific rotation, SR, of some amino acids (see the  $SR_L$  column in Table 2). In this case the modeling equation loses its bars as it must be able to model the negative value also. Let us see, e.g., how the modeling equation should be formulated for the specific rotations of amino acids. In this case, eq 12 should be recast into the more general form given by eq 13, as SR cannot only assume negative values, but can also assume antithetical values for the L- and D-forms.

$$\mathbf{P}_{LUD} = \mathbf{C}_{LUD} \cdot \chi \quad (13)$$

with

$$\mathbf{C}_D = -\mathbf{C}_L \quad (14)$$

where  $\cup$  stands for the logical sign 'or'. The automatic extension of the modeling to the other form may be regarded as a kind of pseudoexternal validation test as the model is applied to the prediction of SR of a set of amino acids which are similar to the training set but are not involved in the development of the model.

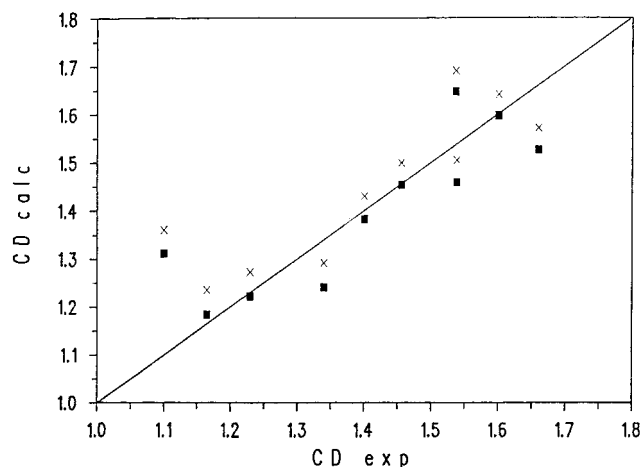
Particular care should be taken with multilinear relations where even apparently small rounding effects are magnified to consistent errors in predicted values. Coefficients returned by the regression procedure should always be examined in a critical manner and verified before using them in a predictive

**Table 2. Experimental Values of the Solubility,  $S$ , for 20 Amino Acids (AA, at 25 °C in units of g/kg of water), pH at the Isoelectric Point, pI of 21 AA, Crystal Densities, CD, of 10 AA, Side-Chain Molecular Volume,  $V$  (in Å<sup>3</sup>), of 18 AA, Specific Rotations in Angular Degrees,  $SR_L$ , of 16 L-AA in Water (in parentheses,  $\pm 1$  °C, when  $T \neq 20$  °C), Unfrozen Water Content UWC (g of H<sub>2</sub>O/g of AA) of 8 AA, and Solubility,  $S$  (at the indicated  $T$  (°C), in units of g/100 mL of water) of 23 Purines and Pyrimidines (PP)<sup>87-92</sup>**

AA	$S$	pI	CD	$V$	$SR_L$	UWC	PP <sup>(a)</sup>	$S(T, \text{°C})$
Gly	251	5.97	1.601	36.3			7I8MTp	0.63 (20)
Ala	167	6	1.401	52.6	2.7 (22)		7B8MTp	0.45 (20)
Cys		5.07					7ITp	2.7 (20)
Ser	422	5.68	1.537	54.9	-6.83	0.48	7BTp	0.37 (30)
Val	58	5.96	1.230	85.1	6.42		1BTb	0.56 (30)
Thr	97	5.60		71.2	28.4 (26)	0.72	7PTp	23.11 (30)
Met	56	5.74	1.340		-8.11 (25)		1PTb	1.38 (30)
Pro	1622	6.30		73.6	-85 (23)	1.07	7ETp	3.66 (30)
Leu	23	5.98	1.165	102	-10.8 (25)		1ETb	3.98 (30)
Ile	34	6.02		102	11.29		Cf	2.58 (30)
Asn	25	5.41		72.4			Tp	0.81 (30)
Asp	5	2.77	1.660	68.4	4.7 (18)		Tb	0.054 (30)
Lys	6	9.74		105.1	14.6	0.93	UA	0.002 (20)
Hyp	361	5.8			-75.2 (23)	0.70	OA	0.18 (18)
Gln	42	5.65		92.7			X	0.05 (20)
Glu	8.6	3.22	1.538	84.7	11.5 (18)	0.97	IsoG	0.006 (25)
His	43	7.59		91.1	-39.01 (25)	0.66	G	0.004 (40)
Arg	181	10.76	1.100	109.1	12.5	0.46	HypoX	0.07 (19)
Phe	29	5.48		113.9	-35.14		A	0.09 (25)
Tyr	0.5	5.66	1.456	116.2			T	0.40 (25)
Trp	12	5.89		135.4	-31.5 (23)		5MC	0.45 (25)
							U	0.36 (25)
							C	0.77 (25)

<sup>a</sup> For the meaning of these names, see the footnote for Table 4.

manner. In ref 59 the case of the simulation of a property with two and with five decimal figures is reported, where the difference can even be detected at the very small scale of a figure in a paper. To underline this aspect, in Figure 4 the calculated vs experimental plot of the crystal densities of 10 amino acids is shown. The modeling has here been done with a single molecular connectivity term (see paragraph on specific rotations of amino acids). In this figure the solid squares (■) have been obtained with the three decimal places and the times (×) signs with just one decimal place. The difference can be enhanced further if linear combinations with more descriptors are used, and in this case even the second decimal place might not be enough to obtain an



**Figure 4.** Plot of the calculated (calcd) versus the experimental (exp) crystal density, CD, of 10 amino acids. ■ and × have been obtained with correlation parameters with three decimal figures and with one decimal figure, respectively.

optimal plot, as shown in ref 59. The best way to avoid this kind of error consists of checking the accuracy of the prediction with calculated vs experimental plot methods. Plot methods are not always taken into consideration in modeling studies. They can illustrate and detect violation of assumptions; i.e., values should show random fluctuations around the main diagonal of the figure. This is equivalent to saying that residuals should show random fluctuation around a value of zero. Clusters of positive and negative values might suggest that a curvilinear trend in the data should be investigated. In a set of values obtained in sequence, there should not be long runs of values on the same side of the main diagonal of the figure; i.e., there should not be systematic trends in the sequence of residuals. Unfortunately, it is difficult to quantify what constitutes a 'long' run. Furthermore, by employing plotting methods it is easier to detect the presence of outliers in the data set, which lead to an inflated standard deviation, and in some cases, this allows a strategy to be outlined for their treatment.

## F. Correlation Problem and Randić's Orthogonalization Procedure

In many cases LCCIs have a series of drawbacks, as molecular descriptors are normally interrelated, and the type of correlation is  $n$ -dependent ( $n$ , number of points). The interrelation is measured by the correlation matrix of the indices of a LCCI obtained regressing every index as a function of every other index of the LCCI and measuring the corresponding regression coefficient,  $r$ . An interesting collinearity criterion was proposed by Mihalić et al.<sup>35</sup> They in fact suggested that those with  $r \geq 0.98$  should be considered as being strongly interrelated indices. A direct



test for the overall collinearity among indices of our  $\{\chi\}$  set for a specific property,  $P$ , is given by the mean interrelation matrix value,  $\langle r_{IM}(P:\{\chi\}) \rangle$ .<sup>59</sup> One should be reminded that even strongly interrelated indices can further enhance the quality of a description as the small fraction of an index which is not reproduced by its strongly interrelated companion can provide positive contributions to the modeling.

Thus, inclusion or exclusion of an index from an LCCI on the exclusive basis of its collinearity can be misleading.<sup>27–29,86</sup> Now, the mutual relatedness among the different indices can result in highly unstable regression coefficients of vector  $\mathbf{C}$ . These coefficients are, in fact, not stable under addition or deletion of a single index into the regression equation. Further, this mutual relatedness (i) may render the values predicted for compounds not in the original ‘training’ data set unreliable, (ii) may render an analysis of the relative importance of an index in a modeling an useless task, and (iii) may underestimate the utilities of regression coefficients with consequent loss of validity of the LCCI. A way to short circuit these drawbacks is to construct orthogonal indices,  $\Omega$ , by the aid of an orthogonalization procedure outlined by Randić.<sup>26–29</sup> These orthogonal indices (i) can render the regression equation stable with the inclusion or deletion of a new index, i.e., the regression parameters,  $c_k$ , are constant, (ii) can give information about the importance of the indices in the regression equation, thus detecting dominant descriptors, and (iii) can improve the utilities of the coefficients of the dominant descriptors. It should be noticed that the statistical performance of the best LCOCI equals the statistical performance of the best LCCI from which it was derived, as orthogonal indices cannot expand the information content of the original indices from which they are derived.

It should be kept in mind that the ordering in which the subsequent ordinary descriptors are added in the normal LCCI fixes the orthogonalization procedure. The drawback of doing further calculations to derive the orthogonal indices, a drawback that can become quite heavy if the number of indices to be orthogonalized is large and the number of compounds to be modeled is not held constant, can be avoided. It is, in fact, possible to obtain an orthogonalized regression equation without recourse to the orthogonal indices. All that is needed is to derive stepwise LCCI regressions and then use the diagonal coefficients as those of the sought-after regression equation. The  $c_U$  value of the unitary term is given by the single- $\chi$  linear regression. The orthogonalization procedure starts with the first best connectivity index, which is chosen as the first  ${}^1\Omega$  index; the second orthogonal index,  ${}^2\Omega^1 \equiv {}^2\Omega$ , is obtained by subtracting from the second index that part which can be reproduced by  ${}^1\Omega$ . Such a process goes on obtaining from a  ${}^i\chi$  index the corresponding  ${}^i\Omega^{i-1} \equiv {}^i\Omega$  index, which is orthogonal to every  ${}^{i-1}\Omega$  index.

## G. Special Molecular Connectivity Indices

It is not rare to have a case of modeling where some outliers dramatically influence the description with

the consequence that they have to be left out of the modeling. Now, if the outliers do not represent any form of experimental error, they should clearly be included in the modeling. Otherwise, some thought should be given to what respect they may differ from the rest of the set, if that is possible. As the concept of outlier has a meaning in the context of a model, knowledge of the reasons that give rise to them should always be used to improve the model. The unusual value of the property of some compound can, in some cases, be better grasped supposing the existence of association phenomena, either self-association or association with other types of molecules through noncovalent interactions. Such association phenomena can be modeled with the introduction of the following medium-sized set of supra-connectivity indices or supraindices:  $\{a\chi\} = \{aD, aD^v, a^0\chi, a^0\chi^v, a^1\chi, a^1\chi^v, \chi_t/a, \chi_t^v/a\}$ . Here the connectivity indices are either multiplied or divided by an association factor  $a$  that can also be a noninteger factor.<sup>56,60–63,65</sup> This parallels the method to give outliers different weight on some kind of subjective basis as this turns out to be equivalent to the subjective assertion that the model is correct but the data need to be adjusted. Now in some cases a critical analysis of these outliers can reveal the existence of self-association or association phenomena with the solvent. In other cases, instead, owing to the lack of more detailed information about the behavior of compounds in solution, these phenomena have to be inferred from an anomalous value of the property. In the given set, the total connectivity indices,  $\chi_t$  and  $\chi_t^v$ , are divided by  $a$ . This choice resides in their definition (see eq 6), as their values decrease with increasing complexity of the chemical graph. Further, the given set is only the most simple form of a set of supraindices, whose most general form can be defined in the following way, where  $p$  can be any rational number

$$\{(a\chi)^p\} = \{(aD)^p, (aD^v)^p, (a^0\chi)^p, (a^0\chi^v)^p, (a^1\chi)^p, (a^1\chi^v)^p, (\chi_t/a)^p, (\chi_t^v/a)^p\} \quad (15)$$

Throughout our studies it has been shown that for  $p = -1$  and 2, i.e., with reciprocal and squared molecular connectivity supraindices,<sup>60–65</sup> an optimal modeling for some properties can be reached.

One of the main drawbacks of the LCCI method is that in some cases an excessive number of indices are necessary to achieve a satisfactory modeling of a limited set of compounds, with the consequence of a loss of meaning of the corresponding LCCI. This loss of meaning can be detected either from a decrease in  $F$  under inclusion or deletion of a new index and/or from the corresponding deterioration of some utilities. The best solution would be to derive a modeling equation with just a single descriptor, but with  $\chi$  indices this is hardly possible. This task, instead, can be achieved with the introduction of molecular connectivity terms,  $X = f(\chi)$ ,<sup>66–70</sup> which are higher-order indices derived by a trial-and-error composition procedure. This last procedure is performed either on indices of  $\{\chi\}$  set (eq 8) or on indices of the set of eq 15 and sometimes on subsets of these

two sets. In this last case, the subset of best molecular connectivity indices derived by the aid of a full combinatorial procedure is chosen. Clearly, the fewer indices to be tried, the easier the trial-and-error search. In many cases it is possible to derive dead-end terms, i.e., single dominant descriptors that do not give rise to any improved linear combination with other connectivity descriptors. This fact renders the orthogonalization procedure useless. In other cases the *greedy* algorithm alone is able to find an optimal linear combination made up of connectivity terms and other connectivity descriptors. Experience has shown that the most general form for molecular connectivity terms is

$$X_P(\chi_i, \chi_j, \chi_k, \chi_l) = \frac{(\chi_i + b\chi_j)^P}{(c\chi_k + d\chi_l)^Q} \quad (16)$$

The optimization parameters  $b$ ,  $c$ ,  $d$ ,  $p$ , and  $q$  can either be positive, negative or even zero. If  $p$  and  $q$  are integers, eq 16 represents rational functions, i.e., functions in which both the numerator and the denominator are polynomials. Subscript  $P$  stands for the acronym of the modeled property. Chi indices  $\chi_i$ ,  $\chi_j$ ,  $\chi_k$ , and  $\chi_l$  are normally taken from set of eq 8, but sometimes they can be taken from the set of eq 15, and the result is a quite powerful but rather convoluted term. The rather tedious trial-and-error search technique for the best term, if it works, is rather straightforward and consists of the following steps: (i) start with the best index, (ii) add to it another index, (iii) optimize it, (iv) back-optimize the previous index, (v) check if the introduced index should not be optimized again (normally it does not), (vi) introduce a new index and restart from step ii to step v with an additional optimization step for the new index, (vii) if further addition of a new index gives no improvement, then construct the fraction and restart from step ii to step v with additional optimization steps, (viii) introduce and optimize the different coefficients  $b$ ,  $c$ ,  $d$ ,  $p$ , and  $q$ . This procedure can be schematized for the case of four parameters by the aid of the following symbolism, where  $I$  stands for introduction,  $O$  stands for optimization and  $C$  for check operations, and the fraction can be constructed at every level:

$$O(1)$$

$$I(2), O(2|1), O(1|2), C(2|1)$$

$$I(3), O(3|1, 2), O(2|3, 1), O(1|2, 3), C(3|1, 2)$$

$$I(4), O(4|1, 2, 3), O(3|4, 1, 2), O(2|3, 4, 1),$$

$$O(1|2, 3, 4), C(4|1, 2, 3)$$

Usually, this procedure either converges pretty rapidly or does not work at all. Normally, when molecular connectivity terms are used the optimal modeling equation, eq 12, quenches into the simple linear relation

$$P = c_1 X + c_U U_0 \quad (17)$$

Here  $U_0 = X^0 \equiv 1$  is the unitary connectivity term.

Already Kier and Hall<sup>13</sup> had suggested that the composition of  $\chi$  indices into a single descriptor could give rise to improved descriptors.

Molecular connectivity terms are, practically, theoretically optimized higher-order graph descriptors, and even if they constitute a powerful tool for modeling, they can nevertheless have problems in modeling some classes of compounds. Problems normally arise with classes of highly 'heterogeneous' compounds, e.g., classes made up of saturated, unsaturated, nonsubstituted, substituted, highly substituted, nonpolar, slightly polar, and highly polar compounds. These are classes of compounds which differ among them in the different level of noncovalent interactions, including hydrogen-bond interactions. To model the properties of these compounds, semiempirical molecular connectivity terms have been introduced quite recently, i.e., either terms (i) whose molecular connectivity indices are multiplied by an empirical parameter or (ii) that just include in their expression one or more empirical parameters.<sup>70</sup> The empirical parameters are the dielectric constants,  $\epsilon$ , 'ad hoc'  $\epsilon$ -related parameters which can describe hydrogen bonds, and the molar masses,  $M$ . Introduction of  $M$ ,  $\epsilon$ , and  $\epsilon$ -related parameters also allows analysis of some general characteristics of the optimal descriptors. The dielectric constant has been selected to improve the modeling as: (i) it is related to the noncovalent character of a compound, (ii) a wide wealth of values for this property are known, and (iii) one can follow what is normally done in molecular dynamics simulations, where the solvent is normally mimicked by using its dielectric constant. For a class of highly heterogeneous solvents, it has been found that the best 'ad hoc'  $\epsilon$ -related parameters are  $a_w \approx \epsilon/15$ ,  $a_{OH}$ , and  $a_\epsilon$ . Parameter  $a_w \approx \epsilon/15$  is truncated at the first figure, and for  $\epsilon/15 < 1$ ,  $a_w = 1$  is assumed. The number 15 has been chosen as it represents the molar mass of a  $CH_3$  radical. Hydrogen bonds in alcohols and acids contribute  $a_w = 2$  whatever the value of  $\epsilon/15$  is, but for compounds with medium dielectric constant, like ethylenecarbonate,  $a_w = 3$  is preferred. Compounds with a very high  $\epsilon$  value, like formamide, have  $a_w = 7$  and the contribution due to the hydrogen bond is neglected, while for compounds with quite low  $\epsilon$  value, like morpholine,  $a_w = 1$  is preferred. The second  $\epsilon$ -related parameter is  $a_{OH} = 2 + \epsilon/15$ , truncated at the second figure. When the number of alcohols in the data set is rather low, then  $a_\epsilon = a_w = \epsilon/15$  is used instead of  $a_w = 2$ .

### III. Modeling Properties of Different Classes of Compounds

The experimental values of modeled properties and the calculated values of the corresponding molecular connectivity indices of the different classes of compounds are collected in Tables 2–13. The original experimental values of the studied properties can be found in the works of the authors cited in the References section and also in refs 25 and 87–97. Let us now see how the molecular connectivity indices and connectivity terms "earn their keep" by helping us to model the greatest number of properties of as

many different classes of compounds as possible and possibly to uncover regularities in these descriptions.

## A. Amino Acids

### 1. Side-Chain Molecular Volume

Let us start modeling the side-chain molecular volume,  $V_{AA}$ , of  $n = 18$  amino acids (AA) shown in Table 2 (no Met, Cys, and Hyp), while in Table 3 are reported the  $\chi$  and  $M$  values of AA. This is one of the oldest and most successful modelings of an amino acid property by molecular connectivity indices. The first study on this property<sup>87</sup> noticed that the  ${}^1\chi^v$  index of the side chain could model  $V_{AA}$  in a remarkable way. More recently<sup>59,65</sup> it has been noticed that the single index  $\{{}^0\chi^v\}$  of the whole amino acid could reach the same statistical score:  $Q = 0.25$ ,  $F = 691$ ,  $r = 0.989$ ,  $s = 3.95$ ,  $\langle u \rangle = 14.8$ . For a comparison,  $M$  describes this property with  $Q = 0.069$  and  $F = 83$ . Both the forward selection and the full combinatorial procedures choose the same LCCI, made up of two  $\chi$  indices,  $\{D^v, {}^0\chi^v\}$ . This is the overall best LCCI for this property<sup>65</sup> with  $Q = 0.40$ ,  $F = 887$ ,  $r = 0.996$ ,  $s = 2.48$ ,  $\langle u \rangle = 12$ . This result is even better than the result obtained with a LCCI consisting of more descriptors. Here, the increasing  $F$  value guarantees that the introduction index  $D^v$  does not endanger the description. The utility vector for each component of the connectivity vector  $(D^v, {}^0\chi^v, U_0)$  is quite satisfactory, especially for the second index,  ${}^0\chi^v$ , with (5.1, 29, 4.0), where the last value is the utility of the unitary index. The presence in the best LCCI of valence indices only underlines the importance of a pseudograph description for the side-chain molecular volume of AA. In fact, many amino acid side chains have lone pairs and/or double bonds: Thr, Ser, Gln, Glu, Asn, Asp, Phe, Tyr, Trp, His, Lys, Pro, and Arg. The correlation between the two LCCI indices,  $r(D^v, {}^0\chi^v) = 0.826$ , following the criteria of Mihalovic et al., is not excessive. It is even less than the average value of the interrelation matrix for this property:  $\langle r_{IM}(V: \{\chi\}) \rangle = 0.883$ . A rapid trial-and-error search for a higher-order index, based on the two best indices

found, finds the following  $X$  term,  ${}^{3.5}X_V = [(D^v)^{3.5}/{}^0\chi^v]$ , which by itself is a rather poor descriptor ( $Q = 0.031$  and  $F = 11$ ) but together with  ${}^0\chi^v$  again shows an improved modeling at every statistical level

$$\{{}^{3.5}X_V, {}^0\chi^v\}: Q = 0.424, F = 989, r = 0.996, \\ s = 2.35, \mathbf{u} = (5.3, 34, 7.4)$$

Index  ${}^0\chi^v$  thus confirms its dominant character in the description of  $V_{AA}$ , as not only is it a component of the term found but it is also a component of this last LCCI. The interrelation between the term and index is better than the preceding interrelation between  $D^v$  and  ${}^0\chi^v$ , in fact  $r({}^{3.5}X_V, {}^0\chi^v) = 0.725$ . With the new term we are, thus, decreasing the interrelation. Nevertheless, the desirable solution remains a description of  $V_{AA}$  with just one molecular connectivity descriptor, which at this level of the problem can only be an improved sort of molecular connectivity term. A more ambitious trial-and-error search with all the indices of the set of eq 8 discovers the following molecular connectivity term

$$X_V = \frac{(D^v)^{1.3} + ({}^0\chi^v)^{2.1}}{D^v - 0.7 \cdot D} \quad (18)$$

This term performs  $Q = 0.438$ ,  $F = 2109$ ,  $r = 0.996$ ,  $s = 2.3$ ,  $\mathbf{u} = (46, 17)$ . The improvement is impressive,  $F$  more than doubles, and the utility of the unitary term also. With this term, which is a dominant 'dead-end' term, as it does not allow a better LCCI, the orthogonality problem has been bypassed. The correlation vector for vector  $(X_V, U_0)$  is  $\mathbf{C} = (18.1182, -52.5871)$ . Thus, the final modeling equation is  $\mathbf{V}_{AA} = \mathbf{18.12X}_V - \mathbf{52.59}$ . Rounding here to the first two decimal figures is justified by the simplicity of the equation. Anyway, when modeling with LCCIs, the full values of  $c_k$  of the correlation vector  $\mathbf{C}$  should be consulted and used.

We notice now that the side-chain molecular volume is no longer dependent on  ${}^0\chi^v$ , but through  $D^v$ , the other index of the first LCCI, continues to be

**Table 3. Molecular Connectivity Indices for 21 Amino Acids and Their Molar  $M$  Masses**

AA( $M$ )	$D$	$D^v$	${}^0\chi$	${}^0\chi^v$	${}^1\chi$	${}^1\chi^v$	$\chi_t$	$\chi_t^v$
Gly (75)	8	20	4.28446	2.63992	2.27006	1.18953	0.40825	0.03727
Ala (89)	10	22	5.15470	3.51016	2.64273	1.62709	0.33333	0.03043
Cys (121)	12	23.56	5.86181	4.55358	3.18074	2.40290	0.23570	0.02875
Ser (105)	12	28	5.86181	3.66448	3.18074	1.77422	0.23570	0.00962
Val (117)	14	26	6.73205	5.08751	3.55342	2.53777	0.19245	0.01757
Thr (119)	14	30	6.73205	4.53473	3.55342	2.21862	0.19245	0.00786
Met (149)	16	26.67	7.27602	6.14607	4.18074	4.04355	0.11785	0.01859
Pro (115)	16	28	5.98313	4.55413	3.80453	2.76688	0.08333	0.00932
Leu (131)	16	28	7.43916	5.79462	4.03658	3.02094	0.13608	0.01242
Ile (131)	16	28	7.43916	4.79462	4.09142	3.07578	0.13608	0.01242
Asn (132)	16	36	7.43916	4.70278	4.03658	2.30434	0.13608	0.00254
Asp (133)	16	38	7.43916	4.57273	4.03658	2.23927	0.13608	0.00196
Lys (146)	18	32	7.98313	5.91594	4.68074	3.36624	0.08333	0.00439
Hyp (132)	18	34	6.85337	4.87159	4.19838	2.84158	0.06804	0.00340
Gln(146)	18	38	8.14627	5.40997	4.53658	2.80434	0.09623	0.00179
Glu (147)	18	40	8.14627	5.27984	4.53658	2.73927	0.09623	0.00139
His (155)	22	42	8.26758	5.81918	5.19838	3.15529	0.03402	0.00080
Arg (174)	22	42	9.56048	6.70883	5.53658	3.60022	0.04811	0.00078
Phe (165)	24	42	8.97469	6.60402	5.69838	3.72222	0.02406	0.00069
Tyr (181)	26	48	9.84493	6.97388	6.09222	3.85651	0.01964	0.00027
Trp (204)	32	54	10.83650	8.10402	7.18154	4.71624	0.00567	0.00009



determined by valence types of indices even if not exclusively by them. Clearly we might wonder if somewhere an even better term does not hide. The characteristics of every trial-and-error procedure is that it is an improvement–open search procedure.

## 2. Isoelectric Point

It was during the simulation of the pH at the isoelectric point, pI, of  $n = 21$  amino acids that indices  $D$  and  $D^v$  were introduced, together with the concept of fragmentary molecular connectivity indices, i.e., indices which were mainly determined by the characteristics of the secondary functional groups in amino acids.<sup>55</sup> In fact, as this property is highly dependent on the type of side chain an amino acid has, the normal connectivity indices of set eight achieve a totally unsatisfactory modeling. The construction of the first fragmentary molecular connectivity indices in the cited paper was rather awkward. Recently, an entire new and sound set of fragmentary molecular connectivity terms has been proposed, which were derived with a rather easy trial-and-error procedure.<sup>65,66,68</sup> These terms are defined in the following way

$$X_{pI} = \frac{\chi}{\chi^v} \left( 1 + \frac{\Delta n}{n_T} \right) \quad (19)$$

where  $\Delta n = n_A - n_B$ ,  $n_A = \text{no. of acidic groups}$  (two for Asp and Glu, one for all others),  $n_B = \text{no. of basic groups}$  (two for Lys and His, three for Arg, and one for all others), and  $n_T = 3$  (total number of functional groups); notice that for  $n_T = 2$ ,  $\Delta n = 0$ . Clearly, there are eight such terms following the type of index which enters in numerator  $\chi$ . The nomenclature for such terms can be defined in the following way for  $\chi = D^v \rightarrow X \equiv {}^D X^v$  etc.. The best single descriptor for pI is  ${}^0 X^v$ , with  $Q = 2.12$ ,  $F = 267$ ,  $r = 0.966$ ,  $s = 0.46$ ,  $\mathbf{u} = (16, 28)$ . These statistics, especially the utility statistics, seem quite satisfactory. Now,  $Q$  statistics can be improved, at the expenses of  $F$  and  $u$  statistics, with the following LCXCI (linear combination of  $X$  terms made up of connectivity indices), which can be derived by the aid of both forward and full combinatorial techniques

$$\{ {}^D X^v, {}^0 X^v, {}^0 X^v, {}^1 X^v \}: Q = 2.53, F = 95, \\ r = 0.980, s = 0.39, \mathbf{u} = (3.1, 2.8, 4.7, 2.8, 26)$$

Average  $\langle u \rangle$  drops from 22.4 to 7.9, the utility of  ${}^0 X^v$  drops dramatically, and only the unitary index maintains a good utility.

To improve these utilities and detect possibly dominant descriptors, use is made of the following vector of orthogonalized terms:  $\Omega = ({}^1\Omega, {}^2\Omega, {}^3\Omega, {}^4\Omega, U_0)$ , where  ${}^1\Omega \equiv {}^0 X^v$ ,  ${}^2\Omega \leftarrow {}^D X^v$ ,  ${}^3\Omega \leftarrow {}^1 X^v$ ,  ${}^4\Omega \leftarrow {}^0 X^v$ . The orthogonalized vector shows the following utilities:  $u = (19, 1.3, 1.0, 2.8, 33)$ . This utility vector tells us that only the first  ${}^1\Omega \equiv {}^0 X^v$  and the last  $U_0 \equiv \Omega^0 \equiv 1$  parameters are important descriptors. We are thus back to the single-term description but with an enhanced utility for  ${}^1\Omega$  and  $U_0$ : 19 and 33 instead of 16 and 28. It should be noticed that the statistical score of the molar masses for pI is  $Q = 0.002$  and  $F$

$= 0.14$ . An inspection of the interrelation between the eight terms confirms their small interrelation as  $\langle r_{IM}(\text{pI}; \{X\}) \rangle = 0.560$ ,  $r_w({}^D X^v, X^v) = 0.004$  and  $r_s({}^D X^v, {}^1 X^v) = 0.975$ , where  $r_w$  and  $r_s$  stand for the weakest and strongest interrelation, respectively.

A critical analysis of the term  ${}^0 X^v$  lets us notice that this term is rather trivial, as it is nothing other than  $(1 + \Delta n/n_T)$ . Now, as the best description is given by a relation consisting of this term only, this means that molecular connectivity indices are not needed to simulate this property. Let us resort to a deeper trial-and-error search. This time we discover the following not at all trivial term

$$X'_{pI} = \frac{({}^1\chi^v)^{0.5} - 180\chi_t^v}{D} \left( 0.04 \cdot \chi_t^v + \frac{\Delta n}{n_T} \right) \quad (20)$$

The modeling power of this dominant term is quite remarkable:  $Q = 3.41$ ,  $F = 693$ ,  $r = 0.987$ ,  $s = 0.29$ ,  $\langle u \rangle = 58$ ,  $\mathbf{u} = (26, 90)$ , and the correlation vector is  $\mathbf{C} = (77.99429, 5.75382)$ . Thus, the final modeling equation can be written as  $\text{pI} = 77.99X'_{pI} + 5.75$ . Not only is the improvement in  $F$  and  $u$  more than expected, but further this term is a highly dominant 'dead-end' term, as it does not allow any better combination with any other index or term. This term, like the preceding  ${}^0 X^v$  term, is mainly based on valence-type molecular connectivity indices, an expected result as side-chain functional groups in amino acids are rich in double bonds and lone-pair electrons.

## 3. Crystal Density

The crystal density, CD, of 10 amino acids can, with a seemingly satisfactory  $Q$ ,  $F$ , and  $u$  statistics, be modeled with a LCCI consisting of the following molecular connectivity indices,<sup>59,68</sup>

$$\{ D, D^v, {}^0\chi^v, {}^0\chi^v \}: Q = 32.6, F = 87, r = 0.993, \\ s = 0.03, \mathbf{u} = (5.2, 11, 11, 8.1, 24), \langle u \rangle = 12$$

It should here be noticed that in ref 59, owing to a wrong value of  $\chi$  for Thr, a somewhat different but also quite satisfactory result is obtained at the level of the statistical parameters. The found LCCI shows the best  $Q$ ,  $F$ , and  $u$  values. Linear combinations with more or less indices show worse  $Q$ ,  $F$ , and  $u$  statistics. The LCCI with five indices has  $Q = 29.2$  and  $F = 36$ . The best single index is  ${}^0\chi^v$  index with  $Q = 3.44$ ,  $F = 3.9$ ,  $r = 0.570$ , which is a bad descriptor for this property. Using four indices to model 10 property values is a dubious choice that strongly suggests using a trial-and-error procedure to look for a single-term description. Two dominant but poor molecular connectivity terms<sup>68</sup> are thus discovered and finally the following more satisfactory dominant term

$$X_{CD} = \frac{({}^0\chi^v)^{1.2} + 1.8 \cdot {}^0\chi^v}{({}^1\chi)^{0.8} - 1.3(\chi)^{2.1}} \quad (21)$$

The statistical factors of this rather convoluted term are  $Q = 7.9$ ,  $F = 20$ ,  $r = 0.848$ ,  $s = 0.1$ ,  $\langle u \rangle = 5.4$ ,  $\mathbf{u} = (4.5, 6.4)$ . The modeling equation is  $\text{CD} = -0.51X_{CD}$

+ **4.18**, while the correlation vector written in full form, i.e., to five decimal points, is  $\mathbf{C} = (-0.50967, 4.81717)$ . The less-than-optimal plot of the modeled versus the experimental values for this property has already been shown in Figure 4, section II.E. Relative to the precedent LCCI made up of four  $\chi$  indices, the only improvement is that now the crystal density can be modeled with only one descriptor. Analyzing both the LCCI and this term, we can conclude that the crystal density is modeled by a mixed contribution of valence and normal molecular connectivity indices and that the  ${}^0\chi$  and  ${}^0\chi^v$  indices seem to be 'focal' for this property. To have an idea of the validity of the modeling of CD by the  $X_{CD}$  term, it should be added that the molar masses rate only  $Q = 1.9$  and  $F = 1.1$ , i.e., connectivity indices and terms are, by far, better descriptors than  $M$ , a fact which underlines that these descriptors are poorly related to  $M$ .

#### 4. Specific Rotation

The specific rotation of  $n = 16$  L-amino acids,  $SR_L$  (in angular degrees, normally given as  $[\alpha]_D^{25}$ ), is a property that can be modeled by the aid of eq 13 only, as some of its values are negative. Further, the modeling of  $SR_L$  can automatically be extended, with eq 14, to the modeling of  $SR_D$  of D-amino acids, which are just opposite in sign relative to  $SR_L$ . Practically the modeling of  $SR_L$  is completely equivalent to the modeling of  $SR_D$ , and in describing this property, we will drop the subscripts L and D and just define this property as SR. A first modeling of the 16 SR values of Table 2 done with a subset of six indices,  $\{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v\}$ ,<sup>57,59</sup> and with a full combinatorial procedure is rather deceiving but, nevertheless, much better than the modeling achieved by the molar masses; in fact, the best two index  $\{{}^0\chi, {}^1\chi\}$ -LCCI rates  $Q = 0.053$ ,  $F = 22$ ,  $r = 0.88$ ,  $s = 17$ ,  $\langle u \rangle = 5.9$  while the molar masses rate  $Q = 0.0003$ ,  $F = 0.34$ .

Running the medium-sized set of eq 11 with a 'fc' procedure<sup>68,69</sup> allows us to detect (i) the best single-index LCCI,  $\{\chi_t\}$ , with quite poor statistics,  $Q = 0.014$ ,  $F = 3.2$ ,  $r = 0.43$ ,  $s = 30$ ,  $\langle u \rangle = 2.1$ , and (ii) the following optimal 3- $\chi$ -LCCI, where no valence molecular connectivity indices contribute to the modeling

$$\{D, {}^0\chi, \chi_t\}: Q = 0.088, F = 41; r = 0.955, \\ s = 11, \mathbf{u} = (6.8, 9.8, 4.3, 7.9), \langle u \rangle = 7.2$$

Combinations with more than three indices show an increasingly unsatisfactory modeling. The interrelation among these indices is not excessively bad as  $r(D, {}^0\chi) = 0.88$ ,  $r(D, \chi_t) = 0.86$ , and  $r({}^0\chi, \chi_t) = 0.80$ . The introduction and use of the following set of reciprocal molecular connectivity indices  $\{R\} = \{{}^D R, {}^D R^v, {}^0 R, {}^0 R^v, {}^1 R, {}^1 R^v, R_t, R_t^v\}$ , where, e.g.,  ${}^0 R^v = ({}^0\chi^v)^{-1}$ , allows detection of an even better description for SR. The full combinatorial search procedure extracts the following optimal reciprocal two-index LCRCI

$$\{{}^D R, {}^0 R\}: Q = 0.089, F = 62, r = 0.952, \\ s = 11, \mathbf{u} = (11, 11, 5.8), \langle u \rangle = 9.2$$

The improvement in  $F$  and utility over the preceding

3- $\chi$ -LCCI combination is noteworthy. The higher interrelation of these indices relative to the best  $\chi$  indices,  $r({}^D R, {}^0 R) = 0.92$ , underlines the fact that a good description does not always need poorly interrelated indices. Here we notice again that reciprocal valence indices do not bring any contribution to the modeling of SR and also that no combination with more reciprocal indices shows a better statistical rating than this two-reciprocal index combination. Before testing a trial-and-error procedure to discover a more effective  $X$  term, let us point out that the best single-index LCRCI,  $\{{}^D R\}$ , rates very badly, even poorer than  $\chi_t$ , with  $Q = 0.009$ ,  $F = 1.4$ ,  $r = 0.3$ ,  $s = 32$ ,  $\langle u \rangle = 1.4$ .

The trial-and-error procedure with subset  $\{D, {}^0\chi, \chi_t\}$  discovers the following general optimal term for SR

$$X_{SR} = {}^0\chi / (D + a\chi_t) \quad (23)$$

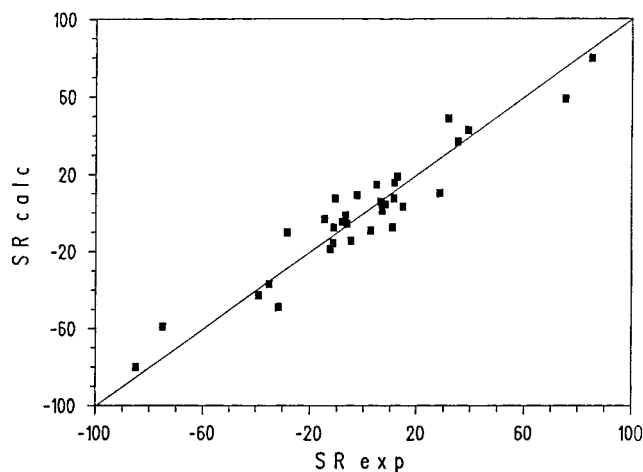
For  $a = 1$  and 7, two nearly equivalent optimal terms are obtained whose descriptive power is much better than  $\chi_t$  and  ${}^D R$ , i.e.,  $\{X_{SR}(1)\}$   $Q = 0.035$ ,  $F = 20$ ,  $r = 0.77$ ,  $s = 22$ ,  $\langle u \rangle = 4.8$  and  $\{X_{SR}(7)\}$   $Q = 0.044$ ,  $F = 30$ ,  $r = 0.83$ ,  $s = 19$ ,  $\langle u \rangle = 5.6$ . Both terms allow the search procedure for the best combination to the forward selection procedure to be restricted. This combinatorial procedure done with the expanded set,  $\{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t^v, X_{SR}\}$ , allows the following optimal combinations to be discovered

$$\{{}^1\chi, X_{SR}(1)\}: Q = 0.100, F = 80, r = 0.961, \\ s = 9.6, \mathbf{u} = (7.6, 12, 12), \langle u \rangle = 11$$

Here  $r({}^1\chi, X_{SR}(1)) = 0.76$ , a rather poor interrelation indeed. Notice the good utility of the  $X_{SR}(1)$  term and the improved utility of the unitary term. The optimal combination with the  $X_{SR}(7)$  term, instead, is made up of three descriptors,  $\{{}^1\chi, \chi_t, X_{SR}(7)\}$ :  $Q = 0.097$ ,  $F = 50$ ,  $\langle u \rangle = 8.1$ ; i.e., the best single term with the lowest interrelation with  ${}^1\chi$ ,  $r({}^1\chi, X_{SR}(7)) = 0.58$ , offers the worse two- and three-index description. The other correlations are rather satisfactory with  $r({}^1\chi, \chi_t) = 0.87$  and  $r(\chi_t, X_{SR}(7)) = 0.57$ . Analysis of the type of  $\chi$  indices involved in the modeling of SR bring us to question ourselves if optical activity is really contributed only by a chemical graph representation of a molecule, i.e., by the  $\sigma$ -electron framework of the amino acids only. On the other hand, the possibility to model SR with just one optimal 'dead-end' term is too tempting not to be tried. A new trial-and-error search on the entire set of eq 8 unveils the following quite satisfactory 'dead-end' term, where valence molecular connectivity indices also play a role and  ${}^1\chi$  index does not appear anymore

$$X_{SR} = \frac{{}^0\chi - (\chi_t^v)^{0.3}}{D^{0.8} + 0.2(\chi_t)^{0.02}} \quad (24)$$

This term shows the following statistics and regression vectors,  $\mathbf{C}_D$  and  $\mathbf{C}_L$ , for the L- and D-forms of 32 [L + D] amino acids:  $Q = 0.084$ ,  $F = 112$ ,  $r = 0.943$ ,  $s = 11$ ,  $\langle u \rangle = 11$ ,  $\mathbf{u} = (11, 11)$ ,  $\mathbf{C}_L = (573.114, -430.56)$ , and  $\mathbf{C}_D = (-573.114, 430.56)$  (see eqs 13 and 14). The



**Figure 5.** Plot of the calculated (calcd) versus the experimental (exp) specific rotation, SR, for 32 amino acids.

final modeling equation for the L-form can be written as  $SR_L = 573 \cdot X'_{SR} - 431$ , and in Figure 5 the calculated versus the experimental plot of this property for the 32  $SR_L$  plus  $SR_D$  values is shown. The connectivity term of eq 24 shows a tremendous improvement relative to the single index description  $\{\chi_t\}$ :  $Q = 0.014$ ,  $F = 3.2$ , and an improvement in  $F$  relative to the  $\{^1\chi, X_{SR}(1)\}$  combination. Thus, 32 SR ( $SR_L + SR_D$ ) values can satisfactorily be described by a single descriptor which combines information from the chemical graph and pseudograph of a molecule.

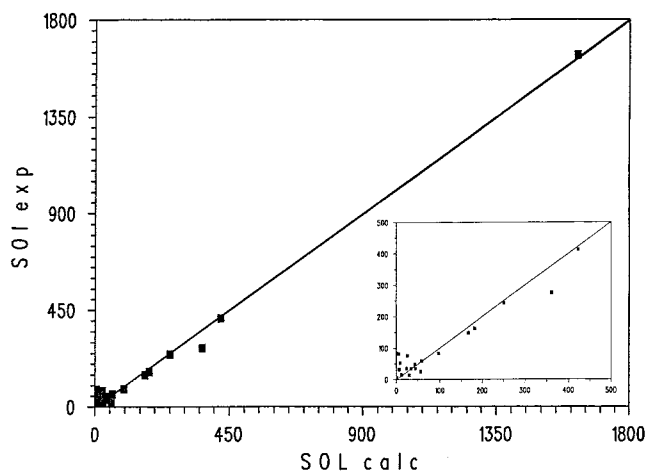
To justify with graph theory the differences in D- and L-amino acids, i.e., that  $C_D = -C_L$ , we might resort to the introduction of digraphs (and pseudodigraphs), i.e., of directed graphs in which the direction associated with the edges can be drawn with inverted arrows for D- and L-forms of Ala, as shown in Figure 6.



**Figure 6.** Digraphs of the amino acid L-Ala (left) and D-Ala.

### 5. Solubility

The solubility,  $S$ , of amino acids, whose experimental values are shown in Table 2, was and is, together with the solubility of purines and pyrimidine bases, the most problematic property studied with LCCI and with molecular connectivity terms.<sup>57,59,60,63,65–69</sup> It was the need to model the solubility of the entire class of amino acids, including the strong outliers Arg, Ser, Hyp, and Pro, that obliged us to introduce the supraconnectivity indices of eq 15, with  $a(\text{Pro}) = 8$ ,  $a(\text{Ser, Hyp, Arg}) = 2$ , and  $a(\text{others}) = 1$  and  $p = 1$ . This parameter,  $a$ , can either have the statistical meaning of a weighting factor or the physical meaning of an association parameter due to inter- and intramolecular association phenomena in solution. The association parameter for Pro, Ser, Hyp, and Arg can, to date, be inferred only from their anomalous solubility value rather than from experimental evidence. Thus, at this time, it is better to look at  $a$  as



**Figure 7.** Plot of the calculated (calcd) versus the experimental (exp) solubility, Sol, for 20 amino acids. Insert: zoom of the solubility region 0–50. Calculated values have been obtained with the modulus equation (see text).

a statistical weighting factor. Let us start the modeling noticing that the molar masses are quite bad descriptors of the solubility of amino acids with  $Q = 0.0009$  and  $F = 2.1$ . While the introduction of supraconnectivity indices did not achieve the expected improvement, the introduction, instead, of the following set of suprareciprocal connectivity indices improved the modeling in a consistent way (see eq 15 with  $p = -1$  and  $a$  outside the parentheses)

$$\{aR\} = \{a^D R, a^D R^v, a^O R, a^O R^v, a^1 R, a^1 R^v, R/a, R_t^v/a\} \quad (25)$$

Here,  $R = 1/\chi$ , e.g., for  $\chi = {}^0\chi^v$  we have  $R = {}^0R^v$  etc. Now, the best single suprareciprocal index regression made up of  $\{a^O R^v\}$  index alone shows an astounding modeling power relative to  $M$

$$\{a^O R^v\}: Q = 0.0287, F = 2052, r = 0.996, s = 35, \mathbf{u} = (45, 13), \langle u \rangle = 29$$

The regression vector for this descriptor is  $\mathbf{C} = (1010.789, -139.389)$ . The resulting modeling equation is thus the following modulus equation:  $S_{AA} = |1011(a^O R^v) - 139|$ . Using the absolute value to remove a meaningless negative solubility increases  $F(S_{calcd}/S_{exp})$  from 2052 to 2127. The calculated versus the experimental solubility values are shown in Figure 7, where four outliers are clearly visible in the insert. As the full combinatorial procedure does not uncover any better LCRCI, we can conclude that the given suprareciprocal valence index  $a^O R^v = a^O \chi^v$  is a highly dominant 'dead-end' descriptor.

Now, descriptor  $\{a^O R^v\}$  does not seem very robust since after excluding from the modeling the outliers Arg, Hyp, Pro, and Ser and, then, modeling the 16 remaining solubility points with the  $\{a^O R^v\}$  index again, the following statistics are obtained  $Q = 0.0294$ ,  $F = 59$ ,  $r = 0.899$ , and  $s = 31$ , with an evident decrease in  $r$  and  $F$  values. For these 16 points, the  $\{a^O R^v\}$  reciprocal index is the best single descriptor with  $Q = 0.038$ ,  $F = 97$ ,  $r = 0.935$ ,  $s = 25$ , and  $\langle u \rangle = 8.7$ . Both full and forward combinatorial procedures uncover, for these same 16 points, the  $\{a^O R, R/a\}$



combination, which has an improved  $Q$  and a not too deteriorated  $F$ :  $Q = 0.049$ ,  $F = 81$ ,  $r = 0.962$ ,  $s = 20$ , and  $\langle u \rangle = 8.2$ . It should be noted that for the given 16 points, and only for them,  $a = 1$ . The most interesting aspect of  $\{a^0R\}$  and  $\{a^0R, R_t/a\}$  descriptors resides in their stability, as both seem to be quite robust descriptors. Modeling, in fact, the 20 solubility points with their help we obtain the following very interesting results

$$\{a^0R\}: Q = 0.023, F = 1358, r = 0.993, \\ s = 43, \mathbf{u} = (37, 9.6), \langle u \rangle = 23$$

$$\{a^0R, R_t/a\}: Q = 0.023, F = 672, r = 0.994, \\ s = 43, \mathbf{u} = (36, 0.9, 7.4), \langle u \rangle = 15$$

The bad utility of the  $R_t/a$  index is the source for the decreasing  $\langle u \rangle$  value in second combination. This descriptor,  $R_t/a$ , is then practically useless for the modeling, and we can conclude that a good and robust descriptor for the solubility of amino acids is the single  $a^0R$  suprareciprocal index.

## B. Purine and Pyrimidine Bases

### 1. Solubility

The modeling of solubility for purine and pyrimidine (PP) bases of Table 2 (see Table 4 for  $\chi$  and  $M$  values) was first achieved in 1996 by the aid of squared supramolecular connectivity indices of eq 15, with  $p = 2.62^{65}$ . A preceding study<sup>61</sup> on a smaller set of bases for which there was experimental evidence of association phenomena in solution was, however, seminal for the introduction of supramolecular indices. The association values used are  $a(7PTp) = 4$ ,  $a(1ETb, 7ETp, Cf) = 2$ , and  $a(7ITp) = 1.5$ . Molar masses are again very bad descriptors for this

property of PP, with  $Q = 0.006$  and  $F = 1.9$ . The best single-descriptor regression has the following descriptor and statistics (notice we are modeling in g/1000 mL of water in accordance with amino acids)

$$\{(a^1\chi)^2\}: Q = 0.176, F = 1553, r = 0.993, \\ s = 5.7, \langle u \rangle = 22$$

Both forward and full combinatorial procedures uncover the following best two-descriptor LCSCI (S stands for squared)

$$\{(a^1\chi)^2, (\chi_t/a)^2\}: Q = 0.240, F = 1445, \\ r = 0.997, s = 4.2, \langle u \rangle = 22$$

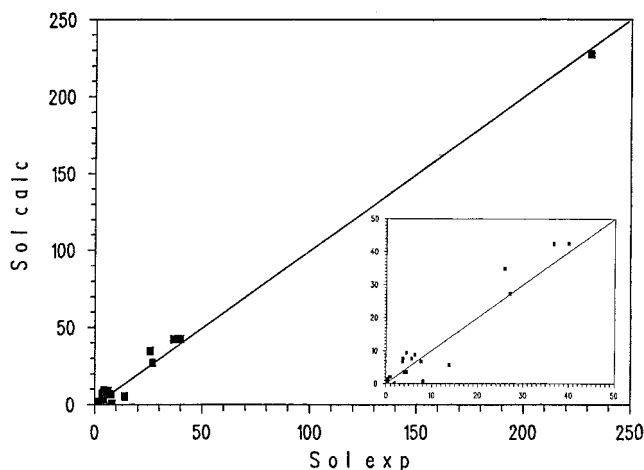
Linear combinations with more than two descriptors show (i) an unsatisfactory statistical behavior of the combination chosen with the forward selection combinatorial procedure and (ii) a dramatically deteriorating utility vector of combinations chosen by the aid of both search procedures. The choice between the single-index and the two-index combination is not obvious, but we prefer the two-index LCSCI. The modeling equation then is  $\mathbf{S}_{PP} = |0.2582(\mathbf{a}^1\chi)^2 + 1779(\chi_t/a)^2 - 9.275|$ , while the full regression and the utility vectors are  $\mathbf{C} = (0.25815, 1778.7, -9.2746)$ ,  $\mathbf{u} = (53, 4.3, 8.0)$ . In Figure 8 are shown the calculated versus the experimental solubility values with the modulus equation to get rid of meaningless negative values. The use of the modulus equation improves  $Q(S_{\text{calcd}}/S_{\text{exp}})$  from 2.46 to 2.59 and  $F(S_{\text{calcd}}/S_{\text{exp}})$  from 3037 to 3367. The insert in Figure 8 shows the evident presence of some outliers.

Regarding association phenomena in solution, it should be emphasized that they have been studied for no more than four compounds (see ref 61 and references therein), for which the following association values have been proposed: (i) for 7PTp,  $a = 4$

**Table 4. Calculated  $\chi$  Values for 23 Purine and Pyrimidine Bases<sup>a</sup> and Their Molar  $M$  Masses**

PP ( $M$ )	$D$	$D^v$	${}^0\chi$	${}^0\chi^v$	${}^1\chi$	${}^1\chi^v$	$\chi_t$	$\chi_t^v$
7I8MTp (250.3)	38	62	13.61036	11.38981	8.34111	5.97071	0.003564	8.51E-05
7B8MTp (250.3)	38	62	13.44723	11.22667	8.48527	6.11486	0.003086	7.37E-05
7ITp (236.3)	36	60	12.74012	10.46716	7.93043	5.53989	0.004365	9.82E-05
7BTp (236.3)	36	60	12.57699	10.30402	8.07459	5.68405	0.00378	8.51E-05
1BTb (236.3)	36	60	12.57699	10.30402	8.07459	5.68405	0.00378	8.51E-05
7PTp (222.2)	34	58	11.86988	9.59691	7.57459	5.18405	0.005346	0.00012
1PTb (222.2)	34	58	11.86988	9.59692	7.57459	5.18405	0.005346	0.00012
7ETp (208.2)	32	56	11.16277	8.88981	7.07459	4.68405	0.00756	0.00017
1ETb (208.2)	32	56	11.16277	8.88981	7.07459	4.68405	0.00756	0.00017
Cf (194.2)	30	54	10.45567	8.1827	6.53658	4.10793	0.01069	0.00024
Tp (180.2)	28	52	9.58542	7.23549	6.1259	3.71758	0.013095	0.000269
Tb (180.2)	28	52	9.58542	7.23549	6.10906	3.7135	0.013095	0.000269
UA (168.1)	26	54	8.71518	5.72474	5.6647	3.11237	0.01604	0.00013
OA (156.1)	22	50	8.43072	5.24931	5.09222	2.66333	0.03928	0.00027
X (152.1)	24	48	7.84493	5.34106	5.27086	2.92873	0.01964	0.00034
IsoG (151.1)	24	46	7.84493	5.45738	5.27086	2.96049	0.01964	0.00043
G (151.1)	24	46	7.84493	5.45738	5.27086	2.96049	0.01964	0.00043
HypoX (136.1)	22	42	6.97469	4.95738	4.87701	2.74509	0.02406	0.00085
A (135.1)	22	40	6.97469	5.07369	4.87701	2.77277	0.02406	0.00108
T (126.1)	18	36	6.85337	4.89385	4.19838	2.4856	0.06804	0.00301
5MC (125.1)	18	34	6.85337	5.01016	4.19838	2.51736	0.06804	0.0038
U (112.1)	16	34	5.98313	3.9712	3.78769	2.06893	0.08333	0.00347
C (111.1)	16	32	5.98313	4.08751	3.78769	2.1007	0.08333	0.00439

<sup>a</sup> A = Adenine, G = Guanine, U = Uracil, T = Thymine, C = Cytosine, OA = orotic acid, UA = uric acid, X = Xanthine, M = methyl, P = propyl, B = butyl, I = isobutyl, Cf = Caffein = 137MMMX = 7MTp, Tb = Theobromine = 37MMX, Tp = theophylline = 13MMX.



**Figure 8.** Plot of the calculated (calcd) versus the experimental (exp) solubility, Sol, for 23 purine and pyrimidine bases. Insert: zoom of the solubility region 0–50. Calculated values have been obtained with the modulus equation (see text).

and 2 in unknown proportions, (ii) for Cf,  $a = 1, 2$ , and 4 in unknown proportions, (iii) for 7ETp and 1Etb,  $a = 2$ . The value  $a = 1.5$  for 7Itp has been inferred assuming an equimolar mixture of monomer and dimer. Quite probably also other bases are expected to undergo some degree of self-association or association with the solvent in solution. An indirect answer to this topic can be given from the following leaving one (or more) out procedure: (i) excluding from the modeling the inferred value for 7Itp and modeling the  $n = 22$  compounds with the same LCSCI, we obtain a satisfactory statistical result:  $Q = 0.234$ ,  $F = 1372$ ,  $r = 0.997$ ; (ii) excluding 7PTp also,  $r$  starts to decrease consistently:  $Q = 0.297$ ,  $F = 120$ , and  $r = 0.964$ ; (iii) excluding 7Itp, 7Ptp, Cf, 7ETp, and 1Etb, a poor modeling of the remaining  $n = 18$  solubility points is obtained:  $Q = 0.204$ ,  $F = 4.8$ ,  $r = 0.624$ . If we, instead, model the 12 compounds from Tb to C (see Table 2) using the same LCSCI, we obtain  $Q = 0.896$ ,  $F = 26$ ,  $r = 0.922$ . The modeling of these 12 compounds can be improved further if index  $(\chi/a)^2$  alone is used; in fact, in this last case we obtain  $Q = 0.944$ ,  $F = 57$ ,  $r = 0.944$ . Such an erratic behavior for the modeling of PP might be explained if it is assumed that more than five purines and pyrimidines undergo, to some extent, association phenomena in solution.

## 2. Singlet Excitation Energy, Oscillator Strength, and Molar Absorption Coefficient

The five properties of DNA–RNA bases, the first and second singlet excitation energies,  $\Delta E_1$  and  $\Delta E_2$ , the first and second oscillator strengths of the first singlet excitation energy,  $f_1$  and  $f_2$ , and the molar absorption coefficient,  $\epsilon_{260}$ , at 260 nm and pH = 7 (see Table 5) have been thoroughly analyzed in recent work.<sup>62,66,68,69</sup> The simulation of the molar absorption coefficient  $\epsilon_{260, \text{exp}}$  at 260 nm and pH = 7 of nucleotides UMP, TMP, AMP, GMP, and CMP is done using the connectivity indices of U, T, A, G, and C only, because the only uncommon part of these nucleotides are these bases. Different kinds of optimal terms have

**Table 5. Experimental (exp) Molar Absorption Coefficient  $\epsilon_{260, \text{exp}}$  at 260 nm and pH = 7.0, First ( $\Delta E_1$ ) and Second ( $\Delta E_2$ ) Singlet Excitation Energies in eV, and First ( $f_1$ ) and Second ( $f_2$ ) Oscillator Strength Values (of the first singlet excitation energies) of the Nucleotide DNA–RNA Bases<sup>93</sup>**

bases	$\epsilon_{260}/1000$	$\Delta E_1$	$\Delta E_2$	$f_1$	$f_2$
A	15.4	4.75	5.99	0.28	0.54
G	11.7	4.49	5.03	0.20	0.27
U	9.9	4.81	6.11	0.18	0.30
T	9.2	4.67	5.94	0.18	0.37
C	7.5	4.61	6.26	0.13	0.72

been discovered, and all of them are better descriptors than the corresponding molar masses,  $M$ . Up to now the best terms for the first and second oscillator strengths,  $f_1$  and  $f_2$ , and the molar absorption coefficient  $\epsilon_{260}$  at 260 nm and pH = 7 are

$$X_{f_1} = \frac{1\chi^v}{(0\chi + 0.6\cdot 0\chi^v)} \quad (26)$$

$$X_{f_2} = \frac{\chi_t^v}{(\chi_t - 16\cdot 0\chi_t^v)} \quad (27)$$

$$X_{\epsilon} = \frac{1\chi^v}{(0\chi + 2\cdot 0\chi^v)} \quad (28)$$

The resulting description is reasonable with  $f_1$  and  $\epsilon_{260}$  described by the same kind of term and  $f_2$  described by a term which is a function of total connectivity indices alone

$$f_1: Q = 31, F = 12, r = 0.89, s = 0.03, \\ \langle u \rangle = 3.1; Q(M) = 11, F(M) = 1.4$$

$$f_2: Q = 11, F = 17, r = 0.92, s = 0.08, \\ \langle u \rangle = 3.3; Q(M) = 2.3, F(M) = 0.7$$

$$\epsilon_{260}: Q = 0.8, F = 21, r = 0.936, s = 1.2, \\ \langle u \rangle = 4.2; Q(M) = 0.2, F(M) = 1.9$$

The modeling of the first and second singlet excitation energies,  $\Delta E_1$ , and  $\Delta E_2$ , can be achieved with a unique term. The following term is in fact a rather efficient descriptor for both properties, especially for the second one

$$X_{\Delta E} = \left( \frac{0\chi}{\chi_t + 10^3 \cdot 0\chi_t^v} \right)^5 \quad (29)$$

$$\Delta E_1: Q = 8.9, F = 5.0, r = 0.79, s = 0.1, \\ \langle u \rangle = 55; Q(M) = 4.9, F(M) = 1.5$$

$$\Delta E_2: Q = 6.9, F = 44, r = 0.97, s = 0.1, \\ \langle u \rangle = 46; Q(M) = 3.7, F(M) = 12$$

Practically, the five properties are described by a formally similar molecular connectivity term while the two energies are modeled by the same term, which is highly dependent on total molecular connectivity indices. Further, every term is a mixing of

$\chi$  and  $\chi^v$  types of indices. Every property is thus described by terms which are  $\delta$  and  $\delta^v$  dependent, a result that seems in keeping with quantum chemistry calculations.<sup>93</sup>

### C. Solubility of the Mixed Class of [AA + PP]

Modeling of the solubility of the mixed class of amino acids plus purines and pyrimidines for a total of  $n = 43$  compounds has been attempted and further refined since the introduction of supramolecular connectivity indices.<sup>66–69</sup> The modeling of the solubility of this special mixed set requires the introduction of the following new set of supraindices

$$\{aD \cdot \chi_t^v, aD^v \cdot \chi_t^v, a^0 \chi_t \cdot \chi_t^v, a^0 \chi_t^v \cdot \chi_t^v, a^1 \chi_t \cdot \chi_t^v, a^1 \chi_t^v \cdot \chi_t^v, \chi_t \cdot a^{-1}, \chi_t^v \cdot a^{-1}\} \quad (30)$$

Here,  $a = 8$  for Pro,  $a = 2$  for Ser, Hyp, and Arg, and  $a = 1$  for the other amino acids. For purines and pyrimidines, instead, we have  $a = 4$  for 7PTp,  $a = 2$  for 7Etb, ETp, and Cf,  $a = 1.5$  for 7Itp, and  $a = 1$  for the remaining bases. To simplify things, set 30 will be renamed as

$$\{^D S, ^D S^v, ^0 S, ^0 S^v, ^1 S, ^1 S^v, S_t, S_t^v\} \quad (31)$$

The trial-and-error search for an optimal descriptor for the 43 solubility points (no Cys but with Hyp included) discovers the three dominant 'dead-end' terms of eqs 32–34 which can achieve a quite satisfactory modeling

$$X_{S1} = \frac{D_S^v - ^D S}{(\chi_t + 350 \cdot \chi_t^v)^{0.7}} \quad (32)$$

$$Q = 0.020, F = 1079, r = 0.982, s = 50, \langle u \rangle = 24, \mathbf{u} = (33, 15)$$

$$X_{S2} = \frac{^1 S^v - (^0 S)^{1.1}}{(S_t - 0.0002)^{1.2}} \quad (33)$$

$$Q = 0.010, F = 297, r = 0.937, s = 91, \langle u \rangle = 9.9, \mathbf{u} = (17, 2.6)$$

$$X_{S3} = \frac{(^D S^v)^{0.3} - 0.9(^D S)^{0.3}}{(S_t^v)^{0.3} + 0.9(S_t)^{0.3}} \quad (34)$$

$$Q = 0.0096, F = 260, r = 0.929, s = 97, \langle u \rangle = 14, \mathbf{u} = (16, 13)$$

Term  $X_{S1}$  seems to be the best term to model the solubility of this mixed class of compounds. Its statistical  $Q$ ,  $F$ ,  $r$ ,  $s$ , and  $u$  values are very good. However, let us see which of these terms is also a good descriptor for the subclasses composed of AA and of PP alone

$$X_{S1} \rightarrow \text{AA}: Q = 0.020, F = 1007, r = 0.991, s = 49, \langle u \rangle = 21, \mathbf{u} = (32, 11)$$

$$X_{S1} \rightarrow \text{PP}: Q = 0.026, F = 35, r = 0.791, s = 30, \langle u \rangle = 5.1, \mathbf{u} = (5.9, 4.2)$$

$$X_{S2} \rightarrow \text{AA}: Q = 0.008, F = 155, r = 0.946, s = 120, \langle u \rangle = 7.7, \mathbf{u} = (12, 3.1)$$

$$X_{S2} \rightarrow \text{PP}: Q = 0.282, F = 4005, r = 0.997, s = 3.5, \langle u \rangle = 33, \mathbf{u} = (64, 3.1)$$

$$X_{S3} \rightarrow \text{AA}: Q = 0.029, F = 2070, r = 0.996, s = 35, \langle u \rangle = 38, \mathbf{u} = (46, 31)$$

$$X_{S3} \rightarrow \text{PP}: Q = 0.042, F = 88, r = 0.899, s = 22, \langle u \rangle = 8.7, \mathbf{u} = (9.4, 8.1)$$

Descriptor  $X_{S1}$  is a quite good descriptor for  $S(\text{AA})$  but an inadequate descriptor for  $S(\text{PP})$ , with a rather low  $r$  value. Descriptor  $X_{S2}$  is the overall best descriptor for  $S(\text{PP})$  and a rather good descriptor for  $S(\text{AA})$  but with quite poor  $s$  statistics, which deteriorates  $Q(\text{AA})$  consistently; further, the  $U_0$  utility for both  $S(\text{AA})$  and  $S(\text{PP})$  is not brilliant. Finally, descriptor  $X_{S3}$  is up to now the best descriptor for  $S(\text{AA})$  and a less satisfactory descriptor for  $S(\text{PP})$ . Considering, as already said, that association phenomena in solution of a large majority of these compounds is quite far from being understood, we think that  $X_{S3}$  should, for the moment, be considered as the best descriptor for this mixed class of compounds. Notice that if the solubility is modeled with g/100 mL of  $\text{H}_2\text{O}$ , the statistical results at the level of  $s$  and  $Q$  change by a factor of 10. Notice also the amazing formal symmetry of the  $X_{S3}$  term.

To emphasize the ability of the trial-and-error method to discover satisfactory terms, we show here the dominant term of eq 35, whose statistical values for the mixed class [AA+PP] are  $Q = 0.021$ ,  $F = 1199$ ,  $r = 0.983$ ,  $s = 47$ ,  $\langle u \rangle = 23$ ,  $u = (35, 11)$ .

$$X_{S4} = \frac{(^D S^v)^{1.1} - (^D S)^{1.1}}{(\chi_t + 10^3 \cdot \chi_t^v)^{0.7}} \quad (35)$$

This term fails at the level of the single subclasses, as for  $S(\text{AA})$  it has  $Q = 0.018$ ,  $F = 805$ ,  $r = 0.989$  and for  $S(\text{PP})$  things are even worse with  $Q = 0.021$ ,  $F = 22$ ,  $r = 0.742$ . This term together with index  $D$  shows somewhat improved  $Q$  statistics, due to a better  $r$  and  $s$

$$\{X_{S4}, D\}: Q = 0.024, F = 779, r = 0.987, s = 42, \langle u \rangle = 17, \mathbf{u} = (38, 3.6, 7.8)$$

For comparison, the molar masses rate  $Q = 0.001$ ,  $F = 2.9$ ,  $r = 0.26$ .

### D. Alkanes

The melting points, MP, and the motor octane numbers, MON, of alkanes, especially this last property, have been thoroughly examined by different authors on many occasions and with different descriptors<sup>25,34,61,65,68,69</sup> and even with quantum theoretically based indices<sup>94</sup> (MON only).



**Table 6. Molar Masses,  $M$ , Experimental Melting Points, MP (K), Motor Octane Numbers, MON, and Calculated Molecular Connectivity Indices for 17 and 39 Alkanes<sup>a 25,94</sup>**

alkanes	$M$	MP	MON	$D$	${}^0\chi$	${}^1\chi$	$\chi_t$
4	58.1		90.1	6	3.41421	1.91421	0.5000
2M3	58.1		97.6	8	4.28445	2.27005	0.4082
2M4	72.2		90.3	6	3.57735	1.73205	0.5774
2M5	86.2		73.5	10	4.99156	2.77005	0.2887
24MM6	114.2	135.65	69.9	14	6.56981	3.66390	0.1667
33MM5	100.2	138.69	86.6	12	5.91421	3.12132	0.2500
5	72.2		61.9	8	4.12132	2.41421	0.3536
23MM4	86.2	144.61	94.4	10	5.15470	2.64273	0.3333
33MM6	114.2	147.05	83.4	14	6.62132	3.62132	0.1768
22MM5	100.2	149.34	95.6	12	5.91421	3.06066	0.2500
22MM6	114.2	151.97	77.4	14	6.62132	3.56066	0.1768
4M7	114.2		39	14	6.40577	3.80806	0.1443
3M7	114.2		35	14	6.40577	3.80806	0.1443
3M6	100.2		55.0	12	5.69867	3.30806	0.2041
24MM5	100.2	153.91	83.5	12	5.86180	3.12589	0.2357
23MM5	114.2	154.05	88.5	12	5.86180	3.18073	0.2357
3E5	100.2		65.0	12	6.69867	3.34606	0.2041
2M6	100.2		46.4	12	5.69867	3.27005	0.2041
3M5	86.2		74.3	10	4.99156	2.80806	0.2887
23ME5	114.2	158.19	88.1	14	6.56891	3.71874	0.1667
223MMM5	114.2		99.9	14	6.78445	3.48138	0.2041
234MMM5	114.2		95.9	14	6.73205	3.55341	0.1925
2M7	114.2		23.8	14	6.40577	3.77005	0.1443
224MMM5	114.2		100.0	14	6.78445	3.41650	0.2041
233MMM5	114.2		99.4	14	6.78445	3.50403	0.2041
22MM4	86.2	173.28	93.4	10	5.20710	2.56066	0.3536
6	86.2		26.0	10	4.82842	2.91421	0.2500
25MM6	114.2	181.95	55.7	14	6.56981	3.62589	0.1667
7	100.2		0.0	12	5.53553	3.41421	0.1768
23MM7	128.3	157.15		16	7.27602	4.18073	0.1179
22MM7	128.3	160.15		16	7.32842	4.06066	0.1250
26MM7	128.3	170.25		16	7.27602	4.12589	0.1179
33ME5	114.2	182.28		14	6.62132	3.68198	0.1768
33EE5	128.3	240.04		16	7.32842	4.24264	0.1250
22MM3	72.2	256.60	80.2	8	5.50000	2.00000	0.5000

<sup>a</sup> 2 = ethane, 3 = propane, etc.; M = methyl, E = Ethyl; e.g., 34ME6 = 3-methyl-4-ethylhexane.

### 1. Melting Points

The melting point constitutes up to now one of the properties that resists any attempt of a satisfactory modeling by molecular connectivity indices and/or terms. To model the MP of 56 alkanes, a strategy named 'double-sieve' was adopted, which consist of (i) first performing a general modeling of the class of compounds, (ii) detecting patterns in this modeling which (iii) allow sorting of subclasses, which are apt to be modeled in a more satisfactory way. This sorting procedure is not a random procedure, as normally subclasses contain compounds with common characteristics, such as the subclass of 17 melting points of Table 6. This subclass, [MMi + MEi + EEi], is made up of alkanes with evident common features; in fact, i stands for the main chain, and M and E stand for methyl and ethyl substitutes along the main chain, respectively. The reader should not forget that as alkanes can be represented by simple graphs only, they do not have any  $\chi^v$ -type indices. The best single descriptor for this subclass of 17 alkanes is the term of eq 36, which performs  $Q = 0.033$ ,  $F = 19$ ,  $r = 0.749$ ,  $s = 23$ ,  $\langle u \rangle = 5.7$ ,  $u = (4.4, 7.0)$ .

$$X_{MP} = \frac{(D - {}^0\chi)^2}{({}^1\chi - 3.9 \cdot {}^0\chi)^{0.6}} \quad (36)$$

This term is far from being an optimal descriptor. A better but not optimal description for this property can be achieved by a normal LCCI composed of the following two indices whose correlation value is  $r({}^1\chi, \chi_t) = 0.97$

$$\{{}^1\chi, \chi_t\}: Q = 0.043, F = 16, r = 0.834, s = 20, \\ \langle u \rangle = 5.0, \mathbf{u} = (5.2, 5.6, 4.3)$$

This (and preceding) modeling is nevertheless much better than the modeling obtained with the molar mass as a descriptor, which achieves  $Q = 0.005$ ,  $F = 0.38$ , and  $r = 0.16$ .

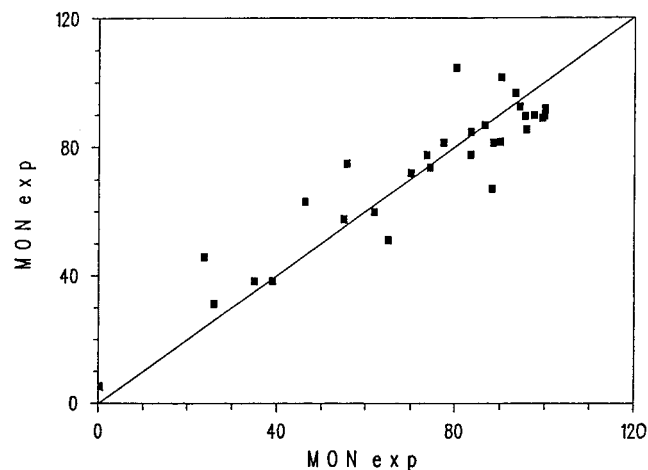
### 2. Motor Octane Number

The motor octane number of the 30 alkanes of Table 6 can rather satisfactorily be modeled at the single index level by a molecular connectivity term found by the aid of a trial-and-error search within the  $\{D, {}^0\chi, {}^1\chi, \chi_t\}$  set. While molar masses are again quite bad descriptors with  $Q = 0.006$ ,  $F = 0.79$ ,  $r = 0.17$ , the obtained  $X_{MON}$  of eq 37 shows the following statistics

$$Q = 0.085, F = 146, r = 0.916, s = 11, \\ \langle u \rangle = 20, \mathbf{u} = (12, 27)$$

$$X_{MON} = \frac{({}^0\chi \cdot \chi_t)^{0.1} + (D)^{1.3}}{({}^0\chi - 1.5 \cdot {}^1\chi)^{1.2}} \quad (37)$$

The correlation vector for this term is  $\mathbf{C} = (-1.6714, 121.02)$ , and a simplified modeling equation could be  $\text{MON} = |-1.67X_{MON} + 121|$ . In Figure 9 the calculated MON values (with modulus equation) versus the corresponding experimental ones are plotted. This dominant term offers the possibility, through a forward combinatorial search, to find a mixed linear combination of a molecular connectivity term and



**Figure 9.** Plot of the calculated (calcd) versus the experimental (exp) solubility, MON, for 30 alkanes. Calculated values have been obtained with the modulus equation (see text).

**Table 7. Molar Mass,  $M$ , Observed Molar Refractivity,  $MR_D$ , Refractivity Index,  $n_D^{20}$ , Density,  $d_4^{20}$  and Calculated Molecular Connectivity Values of 17 MPO(OR')<sub>2</sub> Neutral Organophosphorus Compounds<sup>a,94</sup>**

R'	$M$	$MR_D$	$n_D^{20}$	$d_4^{20}$	$D$	${}^0\chi$	${}^1\chi$	${}^1\chi^v$	$\chi_t$
Bu	208.2	54.42	1.4259	0.9638	24	10.15685	6.12132	5.48471	0.03125
isoBu	208.2	54.86	1.4226	0.9653	24	10.48313	5.83300	5.19640	0.04167
secBu	208.2	54.81	1.4222	0.9657	24	10.48313	5.90901	5.34997	0.04167
terBu	208.2	54.46			24	10.91421	5.41421	4.90140	0.06250
n-Pe	236.3	64.19			28	11.57107	7.12132	6.48471	0.01563
isoPe	236.3	63.61	1.5264	0.9529	28	11.89734	6.83300	6.19640	0.02083
22MMP	236.3	64.27			28	12.32843	6.41421	5.77761	0.03125
n-H	264.3	73.45	1.4353	0.9401	32	12.98528	8.12132	7.48471	0.00781
c-H	260.3	69.23			36	12.13998	8.15660	7.59755	0.00391
n-Hep	292.4	82.87	1.4401	0.9303	36	14.39949	9.12132	8.48471	0.00391
Octyl	320.5	91.24	1.44	0.9257	40	15.81371	10.12132	9.48471	0.00195
1Mhep	320.5	92.02	1.4381	0.9146	40	16.13998	9.90901	9.34997	0.00260
2EH	320.5	91.18	1.4414	0.9289	40	16.13998	9.98502	9.34842	0.00260
Nonyl	348.5	101.1	1.4445	0.9164	44	17.22792	11.12132	10.48471	0.00098
Decyl	376.6	109.7	1.4427	0.9093	48	18.64213	12.12132	11.48471	0.00049
Undec	404.6	119.8	1.4498	0.9077	52	20.05635	13.12132	12.48471	0.00024
Dodec	432.7	129.3	1.4512	0.9012	56	21.47056	14.12132	13.48471	0.00012

<sup>a</sup> M = methyl, E = ethyl, P = propyl, Bu = butyl, Pe = pentyl, H = hexyl, Hep = heptyl, Undec = undecyl, Dodec = dodecyl, c = cyclo.

**Table 8. Observed Retention Index  $R_f$  for Paper Chromatography, Calculated Molecular Connectivity Values, and Molar Masses of 14 RPO(OR')<sub>2</sub> Neutral Organophosphorus Compounds (see abbreviations for Table 7, Oc = octyl)<sup>94</sup>**

R	R'	$M$	$R_f$	${}^0\chi$	${}^1\chi$	${}^1\chi^v$	$\chi_t$
M	E	152.1	0.80	7.32843	4.12132	3.48471	0.125
M	P	180.2	0.71	8.74264	5.12132	4.48471	0.06250
M	Bu	208.2	0.62	10.15685	6.12132	5.48471	0.03125
E	Bu	222.3	0.59	10.86396	6.68198	5.99524	0.02210
P	Bu	236.3	0.53	11.57107	7.18198	6.49524	0.01563
M	Pe	236.3	0.48	11.57107	7.12132	6.48471	0.01563
Bu	Bu	250.3	0.46	12.27817	7.68198	6.99524	0.01105
M	H	264.3	0.38	12.98528	8.12132	7.48471	0.00781
Pe	Bu	264.3	0.38	12.98528	8.18198	7.49524	0.00781
H	Bu	278.4	0.34	13.69239	8.68198	7.99524	0.00276
Hep	Bu	292.4	0.26	14.39949	9.18198	8.49524	0.00138
M	Hep	292.4	0.24	14.39949	9.12132	8.48471	0.00391
Oc	Bu	306.4	0.22	15.10660	9.68198	8.99524	0.00069
M	Oc	320.5	0.15	15.81371	10.12132	9.48471	0.00195

three molecular connectivity indices, with better  $Q$ ,  $r$ , and  $s$  values

$$\{X_{MON}, D, {}^0\chi, {}^1\chi\}: Q = 0.129, F = 85, r = 0.965, \\ s = 7.9, \langle u \rangle = 7.5, \mathbf{u} = (4.8, 5.6, 5.4, 5.4, 4.1)$$

For comparison, the best LCCI is the following linear combination, which consists of the three  $\chi$  indices of the prior mixed combination

$$\{D, {}^0\chi, {}^1\chi\}: Q = 0.092, F = 57, r = 0.932, s = 10, \\ \langle u \rangle = 5.7, \mathbf{u} = (6.1, 7.1, 4.3, 5.3)$$

### E. Four Properties of Organophosphorus Compounds

The modeling of the four experimental properties of organophosphorus compounds of Tables 7 and 8, which was already satisfactorily achieved by quantum theoretical based indices,<sup>94</sup> has been achieved here with two different minimal basis sets. For the  $n = 17$  molar refractivities,  $MR_D$ , the  $n = 14$  refractivity indices,  $n_D^{20}$ , and the  $n = 14$  density  $d_4^{20}$  values, a minimal basis set of five molecular con-

nectivity indices,  $\{D, {}^0\chi, {}^1\chi, {}^1\chi^v, \chi_t\}$ , was used. For the  $n = 14$  retention indices for paper chromatography,  $R_f$ , the minimal basis set of four indices,  $\{{}^0\chi, {}^1\chi, {}^1\chi^v, \chi_t\}$ , was instead used. As indices  $D$  and  $D^v$  as well as the  ${}^0\chi$  and  ${}^0\chi^v$  indices for all these compounds differ from each other by a constant term, we can then choose to neglect  $D^v$  and  ${}^0\chi^v$  throughout the given set of compounds as well as  $\chi_t^v$  as  $r(\chi_t, \chi_t^v) = 1$  and retain  $\chi_t$  only. The further reduction of the basis set for  $R_f$  is based on the same reason, i.e.,  $r(D, D^v) = 1$ , which means that one of the indices is redundant. It has been remarked that the molar refractivity,  $MR_D$ , and the retention index,  $R_f$ , are properties dependent more on the size of the molecule while the density,  $d_4^{20}$ , and the refractive index,  $n_D^{20}$ , are more shape-dependent.<sup>94</sup> This different quality of these two sets of properties is reflected by the different modeling of the molar masses

$$MR_D: Q(M) = 1.3, F(M) = 16\,299, r(M) = 0.9994$$

$$R_f: Q(M) = 45, F(M) = 970, r(M) = 0.994$$

$$n_D^{20}: Q(M) = 370, F = 163, r(M) = 0.965$$

$$d_4^{20}: Q(M) = 140, F(M) = 148, r(M) = 0.962$$

While the first two (more) size-dependent properties are extraordinarily well modeled by  $M$ , the last two (more) shape-dependent properties are less adequately modeled by  $M$ . It will be, nevertheless, interesting to see (i) if a better description than  $M$  can be found, (ii) which of the connectivity indices is more size- or shape-dependent, (iii) which index is, indirectly, the best descriptor for  $M$ , and (iv) as  $r({}^1\chi, {}^1\chi^v) = 0.99990$ , it is then possible to test the importance of  ${}^1\chi^v$  along with the modeling of the four properties, whose value is contributed by  $\delta^v(P)$  only. Thus, the modeling of these two types of physical properties becomes an interesting objective. To calculate the  ${}^1\chi^v$  values, a  $\delta^v(P) = 2.22$  value for the phosphorus atom has been used.<sup>13</sup> The descriptions of these properties by  $\chi$  indices is excellent and better

than the description achieved by quantum theoretically based indices.<sup>94</sup> The best single index and the best LCCI chosen with a full combinatorial approach for these four properties are

$$\mathbf{MR}_D (n = 17)$$

$$\{^0\chi\}: Q = 0.554, F = 2831, r = 0.9974, s = 1.8, \\ \langle u \rangle = 31, \mathbf{u} = (53, 8.4)$$

$$\{^0\chi, ^1\chi\}: Q = 1.980, F = 18116, r = 0.9998, \\ s = 0.5, \langle u \rangle = 14, \mathbf{u} = (18, 13, 11)$$

$$\{(^0\chi \cdot ^1\chi)^{0.55}\}: Q = 1.420, F = 18579, \\ r = 0.9996, s = 0.7, \langle u \rangle = 70, \mathbf{u} = (136, 3.5)$$

Here only the second and third descriptions are improved compared to the description of the molar masses. The description starts to worsen if LCCI with three or more molecular connectivity indices is performed. The description achieved by term  $\{(^0\chi \cdot ^1\chi)^{0.55}\}$ , which is composed of the two main indices of the best LCCI, is spoiled only by the low utility of the unitary term of the regression,  $c_U = 3.5$ ; all other statistics are excellent. We choose, in fact, this term to model  $\mathbf{MR}_D$ , as modeling 17 points with just one descriptor is a better choice than modeling them with two descriptors. The correlation vector for the found term is  $\mathbf{C} = (5.48211, 2.12168)$ . The resulting modeling equation can be written as  $\mathbf{MR}_D = 5.48 \cdot (^0\chi \cdot ^1\chi)^{0.55} + 2.12$ . In Figure 10 the calculated versus the experimental  $\mathbf{MR}_D$  values are plotted.

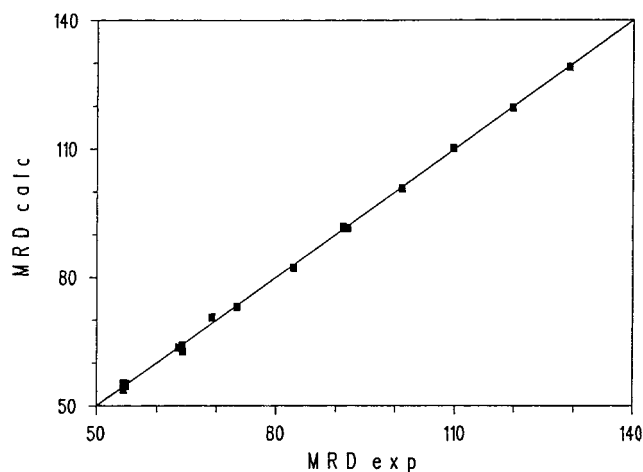
$$\mathbf{R}_f (n = 14)$$

$$\{^0\chi\}: Q = 44.5, F = 970, r = 0.9939, \\ s = 0.022, \langle u \rangle = 38, \mathbf{u} = (31, 44)$$

$$\{^0\chi, ^1\chi^v\}: Q = 57.5, F = 809, r = 0.9966, \\ s = 0.017, \langle u \rangle = 4.5, \mathbf{u} = (3.1, 3.7, 6.8)$$

$$\{(^0\chi \cdot ^1\chi^v)^{0.75}\}: Q = 55.2, F = 1494, r = 0.9960, \\ s = 0.018, \langle u \rangle = 52, \mathbf{u} = (40, 64)$$

Here the two-index LCCI is no more optimal as (i)  $F(^0\chi, ^1\chi^v) < F(^0\chi)$ , and (ii) an unsatisfactory  $\langle u \rangle$  renders this combination even more suspect relative to the single-index combination, even if this two-index combination outperforms  $M$  in its  $r$  and  $Q$  value. The description achieved by the third descriptor, a term composed by the two indices of the best LCCI, seems optimal under every insight. Notice the similar  $\chi$  modeling of  $\mathbf{MR}_D$  and  $\mathbf{R}_f$ . The correlation vector and modeling equation are  $\mathbf{C} = (-0.02117, 1.04713)$ ,  $\mathbf{R}_f = -0.02 \cdot (^0\chi \cdot ^1\chi^v)^{0.75} + 1.05$ . The best size-dependent descriptor and indirectly the best descriptor for the molar masses seems then to be index,  $^0\chi$ , which can thus be considered the molar mass molecular connectivity index, as shown elsewhere.<sup>71</sup> For both properties the modeling achieved with LCCI with more than two indices is no more reliable. Let us now inquire into the modeling behavior of the two (more)



**Figure 10.** Plot of the calculated (calcd) versus the experimental (exp) molar refractivities,  $\mathbf{MR}_D$ , for 17 phosphoderivatives.

shape-dependent properties.

$$\mathbf{n}^{20}_D (n = 14)$$

$$\{^1\chi\}: Q = 405, F = 195, r = 0.971, s = 0.0023, \\ \langle u \rangle = 306, \mathbf{u} = (14, 598)$$

$$\{^1\chi, \chi_t\}: Q = 579, F = 200, r = 0.987, s = 0.0017, \\ \langle u \rangle = 134, \mathbf{u} = (7.5, 3.6, 390)$$

$$\{(^1\chi \cdot \chi_t)^{0.3}\}: Q = 592, F = 418, r = 0.986, \\ s = 0.0016, \langle u \rangle = 650, \mathbf{u} = (20, 1280)$$

The description by the single- and two-index LCCI are more satisfactory than the description given by the molar masses. The best LCCI which has only two indices shows better  $Q$  and  $F$  values than the single-index description but an unsatisfactory utility  $u_2$  value (3.6). The third descriptor, a term composed again by the indices of the best LCCI, seems perfect for the description of this property. We can then choose the following modeling equation:  $\mathbf{n}^{20}_D = -0.05(^1\chi \cdot \chi_t)^{0.3} + 1.46$ , whose correlation vector is  $\mathbf{C} = (-0.0538, 1.45793)$ . The four digits for the  $s$  value are given here to aid in the comparative analysis.

$$\mathbf{d}_4^{20} (n = 14)$$

$$\{^0\chi\}: Q = 143, F = 155, r = 0.964, s = 0.007, \\ \langle u \rangle = 73, \mathbf{u} = (13, 134)$$

$$\{(^0\chi \cdot ^1\chi)^{0.01}\}: Q = 171, F = 222, r = 0.974, \\ s = 0.006, \langle u \rangle = 17, \mathbf{u} = (15, 18)$$

Both descriptions are better than the description offered by the molar masses. A two-index LCCI, description,  $\{^0\chi, \chi_t\}$ , is no more optimal with  $Q = 156$ ,  $F = 92$ , and  $r = 0.971$ . Further, multi-index descriptions show unsatisfactory statistical values. A description offered by the term composed of the first and second best single indices is the most impressive. The correlation vector for this term is  $\mathbf{C} = (-4.0719, 5.20933)$ , and the regression equation is  $\mathbf{d}_4^{20} = -4.07(^0\chi \cdot ^1\chi)^{0.01} + 5.21$ . The collinearity among the



indices that belong to the best LCCI for the three properties is  $r(^0\chi, ^1\chi) = 0.986$  for  $MR_D$ ,  $r(^0\chi, ^1\chi^v) = 0.997$  for  $R_f$  and  $r(^1\chi, \chi_t) = 0.844$  for  $n_4^{20}$ . While the collinearity among the best indices for the modeling of  $R_f$  is quite strong, the collinearity among the best indices for  $n_4^{20}$  is relatively weak, and nevertheless, the  $R_f$  modeling is much better than  $n_4^{20}$  modeling. Throughout these four simulations we can notice (i) that notwithstanding the mentioned high correlation,  $r(^1\chi, ^1\chi^v) = 0.99990$ , index,  $^1\chi$  is critical for the modeling of  $MR_D$ ,  $r^{20}_D$ , and  $d_4^{20}$ , while a similar role is played by index  $^1\chi^v$  for the modeling of  $R_f$ ; (ii) the good description of the four properties by a single descriptor, which has the features of a dominant descriptor, and (iii) that terms found for these properties are normally composed of the best indices of the corresponding best LCCI.

All in all,  $^0\chi$  index seems the best descriptor of size-dependent properties, while  $^1\chi$  seems to be the best descriptor of shape-dependent properties. In fact, while  $^1\chi$  for  $d_4^{20}$  performs  $Q = 134$ ,  $F = 136$ , and  $r = 0.959$  and similar values are obtained by  $^0\chi$  index, the modeling of  $r^{20}_D$  by  $^0\chi$  performs  $Q = 340$ ,  $F = 138$ , and  $r = 0.959$ . These last values are consistently worse than the values obtained by  $^1\chi$ . Further, the fact that the  $D$  index is not a good descriptor of any of these properties shows that it shares no attributes in common with the molar masses, i.e., it describes a graph property of its own, and for this reason it that has been named graph mass index.<sup>71</sup> It is evident that the quality of the modeling of the first two size-dependent properties is better than the quality of the modeling of the last two shape-dependent properties, a fact that is clearly reflected by the value of the  $F$  parameter. Notice that  $\delta^v(P)$  is not based on a pseudograph representation of the phosphoorganic molecules; this 'external' definition of a valence delta value will become important in the next paragraphs.

## F. Lattice Enthalpy of Inorganic Salts

Conceptually the definition of a graph and, even more, of a pseudograph for inorganic salts is quite problematic and can, in principle, be solved only with a graph representation of solid ionic structures, a representation which up to now remains to be satisfactorily answered. The problem will nevertheless be strongly simplified by the aid of pure 'a posteriori' considerations, i.e., the given inorganic compounds will be modeled using eq 3 for  $\delta^v$  and their graph representation for  $\delta$ , e.g., the graph for  $CaCl_2$ , will be written as  $G(CaCl_2) = \bullet-\bullet-\bullet$ . The use of a graph representation for the given metal halides is equivalent to considering these compounds in their gaseous state. The quality of the achieved modeling will tell us how far our approximation has been 'crude'. This procedure, which sharply simplifies the modeling, shows no minor practical advantages.<sup>64,65</sup> Further, it should be remembered that there is no such thing as 'purely covalent' molecule or 'purely ionic' crystal; these two categories represent limiting cases. Thus, the practical application of graph concepts to ionic compounds, even if theoretically ques-

**Table 9. Lattice Enthalpies  $\Delta H_L^\ominus$  at 298.15°K (kJ mol<sup>-1</sup>) of 20 Metal Halides (MeX) and Their Corresponding Molar Masses and Molecular Connectivity Values<sup>95</sup>**

MeX	<i>M</i>	$\Delta H_L^\ominus$	$D^v$	$^0\chi^v$	$^1\chi^v$	$D^Z$
LiF	25.9	1037	8	1.37796	0.37796	4
NaF	42.0	926	7.11111	3.37796	1.13389	3.83333
KF	50.1	821	7.05882	4.50107	1.55839	3.75000
RbF	104.5	789	7.02857	6.29404	2.23607	3.70000
CsF	151.9	750	7.01887	7.65807	2.75162	3.66667
LiCl	42.4	852	1.77778	2.13389	1.13389	2.83333
NaCl	58.4	786	0.88889	4.13389	3.40168	2.66667
KCl	74.6	717	0.83660	5.25700	4.67516	2.58333
RbCl	120.9	695	0.80635	7.04997	6.70820	2.53333
CsCl	168.4	678	0.79665	8.41400	8.25487	2.50000
LiBr	86.8	815	1.25926	2.96396	1.96396	2.25000
NaBr	102.9	752	0.37037	4.96396	5.89188	2.41667
KBr	119.0	689	0.31808	6.08707	7.09762	2.00000
RbBr	165.4	668	0.28783	7.88004	11.6190	1.95000
CsBr	212.8	654	0.27813	9.24407	14.2979	1.91667
LiI	133.8	761	1.15556	3.53546	2.53546	1.90000
NaI	149.9	705	0.26667	5.53546	7.60639	1.73333
KI	166.0	649	0.21438	6.65857	10.4540	1.65000
RbI	212.4	632	0.18413	8.45154	15.0000	1.60000
CsI	259.8	620	0.17442	9.81557	18.4585	1.56667

tionable, has some loose theoretical basis. The transition between the covalent and ionic bonding type is not an abrupt transition, and between the two extreme bonding types, NaF and diamond, there is a wide region of intermediate cases<sup>98-100</sup> to which most of our metal halides (MeX) belong. Further, it should not be forgotten that the electrostatic ionic model for salts is just a model even if, up to date, a very successful one. The lattice enthalpies,  $\Delta H_L^\ominus$ , of 20 metal halides at 298.15 K of Table 9 can optimally be modeled only with the introduction of a new index to the subset of valence molecular connectivity indices  $\{D^v, ^0\chi^v, ^1\chi^v\}$ . The triviality of the graph for these metal halides ( $\bullet-\bullet$ ) renders  $\chi$  indices useless as (i)  $D = ^0\chi, ^1\chi = \chi_t$  and (ii)  $\chi(\text{MeX}_i) = \chi(\text{MeX}_j)$ , with  $i \neq j = 1-20$ . Thus, the only valuable indices are the  $\chi^v$  indices ( $^1\chi^v = \chi_t^v$ ), whose basic parameter,  $\delta^v$ , has been 'pragmatically' defined (see eq 3). Thus, practically (i) we do not need a graph representation, similar to the one used for organic molecules, to model these compounds, (ii) this modeling becomes a benchmark for the proposed definition of  $\delta^v$ , and (iii) this modeling also becomes a benchmark for the following new index

$$D^Z = \sum \delta^z_i = \sum_i (Z^v / n_i) \quad (38)$$

Here  $Z^v$  is the number of valence electrons and  $n$  is the principal quantum number. With this new index, the set of connectivity indices used to model the given property of the inorganic halides becomes  $\{\chi\} = \{D^v, ^0\chi^v, ^1\chi^v, D^Z\}$ . Notice that for these compounds  $^1\chi^v = \chi_t^v$ . The  $\delta^z$  values of the atoms of some inorganic compounds are (the corresponding  $\delta^v$  values are given in section A.2)

$$\begin{pmatrix} \text{Li} & \text{Na} & \text{K} & \text{Rb} & \text{Cs} & \text{F} & \text{Cl} & \text{Br} & \text{I} \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & 7/2 & 7/3 & 7/4 & 7/5 \end{pmatrix}$$

The average interrelation matrix value for this property with this  $\{\chi\}$  set is  $\langle r(\Delta H_L^\ominus; \{\chi\}) \rangle = 0.68$ , while the strongest and weakest interrelated indices



**Table 10. Unfrozen Water Content UWC (g of H<sub>2</sub>O/g of MeCl) for Five Metal Chlorides and Their Corresponding Molecular Connectivity Index Values<sup>92</sup>**

MeCl	UWC	D	<sup>0</sup> χ	<sup>1</sup> χ	D <sup>v</sup>	<sup>0</sup> χ <sup>v</sup>	<sup>1</sup> χ <sup>v</sup>	χ <sub>t</sub>	χ <sub>t</sub> <sup>v</sup>
LiCl	6.5	2	2	1	1.7778	2.1339	1.1339	1	1.1339
NaCl	3.0	2	2	1	0.8889	4.1339	3.4016	1	3.4017
KCl	1.8	2	2	1	0.8366	5.2570	4.6752	1	4.6752
CaCl <sub>2</sub>	4.0	4	2.7071	1.4142	1.6732	5.1832	6.6117	0.70711	3.7485
CuCl <sub>2</sub>	4.0	4	2.7071	1.4142	1.6325	5.8733	8.1777	0.70711	4.6357

four index LCCI:

$$Q = 2.79, F = 328, r = 0.984, s = 0.3_5,$$

$$\langle u \rangle = 13, \text{ and } \mathbf{u} = (18, 7.6)$$

$$X_{\text{UWC}} = \frac{{}^1\chi^v}{D^v - {}^0\chi^v} \quad (40)$$

A somewhat better modeling can be achieved if the modulus of  $X_{\text{UWC}}$  is considered, i.e.,  $|X_{\text{UWC}}|$ , as some  $X_{\text{UWC}}$  values are negative. Now, with  $|X_{\text{UWC}}|$ , we obtain the following improved statistics  $Q = 3.14$ ,  $F = 417$ ,  $r = 0.987$ ,  $s = 0.3$ ,  $\langle u \rangle = 13$ ,  $\mathbf{u} = (20, 5.0)$ . The correlation vector and the modeling equations are  $\mathbf{C} = (1.83423, 0.55209)$  and  $\mathbf{UWC} = 1.83 \cdot |X_{\text{UWC}}| + 0.55$ . Further, while  $X_{\text{UWC}}$  in combination with  $\chi_{t^v}$  allows for improved  $Q$  statistics with  $Q = 3.11$ ,  $|X_{\text{UWC}}|$  term, instead, is a strict 'dead-end' term allowing no improved combinations.

## H. Random Organic Solvents

Until now, with the exception of [AA + PP] and [AA + MeCl] mixed classes, modeling was normally restricted to rather homogeneous classes of compounds. The fact that mixed classes of compounds could satisfactorily be simulated with special types of indices constitutes a good suggestion to model more complex classes of compounds, especially classes of compounds that are characterized by different and strong intra- and intermolecular interactions. Lately,<sup>70</sup> modeling of 11 physicochemical properties of a highly heterogeneous class of compounds has been attempted and for many of them a successful modeling has been achieved with the introduction of semiempirical molecular connectivity terms, i.e., terms, which include in their definition empirical parameters. The modeled heterogeneous class of organic solvents was composed of saturated, unsaturated, substituted, unsubstituted, nonpolar, slightly, and highly polar compounds. These solvents, together with some of their properties and their molecular connectivity indices, are collected in Tables 11 and 12, respectively. None of these properties can satisfactorily be modeled by theoretical  $\chi$  indices or  $X$  terms, as their properties are highly dependent on noncovalent interactions.

Table 11 also shows the values of the molar masses and of the dielectric constants of the different solvents, two central properties in this study. These two properties, either directly or indirectly, together with ad hoc  $\epsilon$ -related parameters, which will be used to describe hydrogen bonds in alcohols and acids, will be used here to overcome the inherent limitation of the molecular connectivity indices, which do not encode van der Waals and hydrogen-bond interac-

tions. As already said in section II.G (last portion), the  $\epsilon$ -related parameters that are derived from the dielectric constant are (i)  $a_w \approx \epsilon/15$ , truncated at the first figure and rounded to 0.5 or 0.0, (ii)  $a_{\text{OH}}$ , and (iii)  $a_\epsilon$ . Parameter  $a_{\text{OH}}$  will be defined in the next paragraph, and for parameter  $a_\epsilon$ , the reader is referred to ref 70. It will be assumed that for  $\epsilon/15 < 1$ ,  $a_w = 1$ . The values for  $a_w \neq 1$  are collected in Table 11, first column, in parentheses. Hydrogen bonds in alcohols and acetic acid contribute  $a_w = 2$ , whatever the value of  $\epsilon/15$  is; but for ethylenecarbonate,  $a_w = 3$  has been preferred. This ad hoc noncovalent parameter,  $a_w$ , shows the following descriptive power for the dielectric constant  $Q = 0.105$ ,  $F = 250.3$ ,  $r = 0.895$ ,  $s = 8.49$ , and  $n = 64$ , which means that it is not redundant with the dielectric constant.

### 1. Boiling Point

Let us start with the modeling of the boiling points,  $T_b$ , of  $n = 63$  solvents. The modeling due to the molar  $M$  masses, the dielectric  $\epsilon$  constant, and  $a_w$  are quite misleading with (in ref 70 second and third  $Q$  values are wrongly reported)

$$M: Q = 0.006, F = 7.9; \epsilon: Q = 0.009, F = 16.4; a_w: Q = 0.008, F = 14.2$$

If these ad hoc descriptors have to play a role in the modeling of  $T_b$ , then these values tell us that they should mainly be  $\epsilon$  and/or  $a_w$ . The best molecular connectivity model is achieved by  $\{{}^0\chi^v\}$  with  $Q = 0.013$ ,  $F = 33.9$ ,  $r = 0.598$ , and  $s = 36.4$ . The use of linear combination of molecular connectivity indices does not improve the modeling. Introduction, instead, of the following set of semiempirical supramolecular connectivity indices improves the description substantially:

$$\{a_w \cdot \chi\} = \{a_w \cdot D, a_w \cdot D^v, a_w \cdot {}^0\chi, a_w \cdot {}^0\chi^v, a_w \cdot {}^1\chi, a_w \cdot {}^1\chi^v, \chi_t/a_w, \chi_t^v/a_w\} \quad (41)$$

The rationale for dividing the two last total indices by  $a_w$  has been explained with the set of eq 15. The improvement in the modeling of  $T_b$  is noticeable both at the level of the single supraindex as well as at the level of a two-supraindex description ( $u$  values will be given only for the ultimate best descriptions)

$$\{a_w \cdot D\}: Q = 0.030, F = 183, r = 0.866, s = 29.0$$

$$\{a_w \cdot D, \chi_t^v/a_w\}: Q = 0.031, F = 98, r = 0.875, s = 28.3$$

This description can be improved further with the introduction of the following semiempirical molecular



**Table 11. Properties of Organic Solvents<sup>a</sup>**

solvent	<i>M</i>	<i>T<sub>b</sub></i>	RI	<i>d</i>	ε	μ	UV
acetone (1.5)	58.1	56	1.3590	0.791	20.7	2.88	330
acetonitrile (2.5)	41.05	82	1.3440	0.786	37.5	3.92	190
benzene	78.1	80	1.5010	0.874	2.3	0	280
benzonitrile (1.5)	103.1	188	1.528	1.010	25.2		
1-butanol (2)	74.1	117.7	1.3990	0.810	17.1		215
2-butanone	72.1	80	1.3790	0.805	18.5		330
butyl acetate	116.2	125	1.3940	0.882	5.01		254
CS <sub>2</sub>	76.1	46	1.6270	1.266	2.6	0	380
CCl <sub>4</sub>	153.8	77	1.4595	1.594	2.2	0	263
Cl-benzene	112.6	132	1.5240	1.107	5.62		287
1Cl-butane	92.6	77.5	1.4024	0.886	7.39		225
CHCl <sub>3</sub>	119.4	61	1.4460	1.492	4.8	1.01	245
cyclohexane	84.2	80.9	1.426	0.779	2.0	0	200
cyclopentane	70.1	50	1.4000	0.751	2.0		200
1,2-dCl-benzene	147.0	179.5	1.5510	1.306	9.9	2.50	295
1,2-dCl-ethane	98.95	83	1.4438	1.256	10.37	1.75	225
dCl-methane	84.9	39.9	1.4240	1.325	7.5	1.60	235
<i>N,N</i> -dM-acetamd (2.5)	87.1	165.2	1.4380	0.937	37.8	3.8	268
<i>N,N</i> -dM-formamd (2.5)	73.1	153	1.431	0.944	36.7	3.86	268
1,4-dioxane	88.1	101	1.4220	1.034	2.2	0.45	215
ether	74.1	34.6	1.3530	0.708	4.3	1.15	215
ethyl acetate	88.1	77	1.3720	0.902	6.0	1.8	260
ethyl alcohol (2)	46.1	78	1.3600	0.785	24.5	1.69	210
heptane	100.2	98	1.3870	0.684	1.92		200
hexane	86.2	69	1.3750	0.659	1.89		200
2-methoxyethanol (2)	76.1	124.5	1.4020	0.965	16.0		220
methyl alcohol (2)	32.0	64.6	1.3290	0.791	32.7	1.70	205
2-methylbutane	72.15	30	1.3540	0.620	1.843		
4-M-2-pentanone	100.2	117.5	1.3960	0.800	13.1		334
2-M-1-propanol (2)	74.1	108	1.3960	0.803	17.7		
2-M-2-propanol (2)	74.1	83	1.3870	0.786	10.9	1.66	
DMSO (3)	78.1	189	1.4790	1.101	46.7	3.96	268
nitromethane (2.5)	61.0	100.9	1.3820	1.127	35.9	3.46	380
1-octanol (2)	130.2	196	1.4290	0.827	10.34		
pentane	72.15	35.5	1.3580	0.626	1.844		200
3-pentanone	86.1	102	1.3920	0.853	17.0		
1-propanol (2)	60.1	97	1.3840	0.804	20.1		210
2-propanol (2)	60.1	82.4	1.3770	0.785	18.3		210
pyridine	79.1	115	1.5100	0.978	12.4	2.2	305
ttCl-ethylene	165.8	121	1.5056	1.623	2.3		
tt-hydrofuran	72.1	67	1.4070	0.886	7.6	1.75	215
toluene	92.1	111	1.4960	0.867	2.4	0.36	285
1,1,2tClFethane	187.4	47.5	1.3578	1.575	2.41		230
2,2,4-tM-pentane	114.2	98.5	1.3910	0.692	1.94		215
<i>o</i> -xylene	106.2	144	1.5050	0.870	2.568		
<i>p</i> -xylene	106.2	138	1.4950	0.866	2.374		
acetic acid (2)	60.05	117.9	1.3719	1.049	6.1	1.2	
decaline	138.2	191.7	1.4758	0.879	2.20		
dBr-methane	173.8	97.0	2.4970	1.542	7.5	1.43	
1,2-dCl-E-en(Z)	96.9	60.6	1.4490	1.284	9.2	1.90	
1,2-dCl-E-en(E)	96.9	47.7	1.4462	1.255	2.1	0	
1,1-dCl-E-en	96.9	31.6	1.4247	1.213	4.6	1.34	
dmethoxymethane	76.1	42.3	1.3563	0.866	2.6		
dimethyl ether	46.1	-24			5.02		
E-encarbonate (3)	88.1	238	1.4250	1.321	89.6	4.91	
formamide (7)	45.0	210.5	1.4475	1.133	109	3.73	
methyl chloride	50.5	-24.1	1.3389	0.916	12.6	1.87	
morpholine	87.1	128.9	1.4573	1.005	7.4		
quinoline	129.2	237.1	1.6293	1.098	9.0	2.2	
SO <sub>2</sub>	64.1	-10.0		1.434	17.6	1.6	
2,2-ttCl-ethane	167.8	146.2	1.4868	1.578	8.2	1.3	
ttM-urea (1.5)	116.2	176.5	1.4493	0.969	23.1	3.47	
tCl-E-en	131.4	87.2	1.4800	1.476	3.4		

<sup>a</sup> Molar mass *M* (g·mol<sup>-1</sup>), *T<sub>b</sub>* = boiling points (°C), RI = refractive index (20 °C), *d* = density (at 20 ± 5 °C relative to water at 4 °C), ε = dielectric constant, dipole moments, μ, in Debye (1D = 10<sup>-18</sup> esu cm = 3.3356 × 10<sup>-3</sup> C m), UV = UV cutoff (nm, wavelength at which absorbance is 1 Å for a good LC-grade solvent).<sup>96,97</sup> Abbreviations: amd = amide, d = di, E-en = ethylen, M = methyl, t = tri, tt = tetra. *T<sub>b</sub>* has been modeled in Kelvins. For values in parentheses, see text.

connectivity term found by a trial-and-error procedure

$$X_{BP} = \frac{(a_w \cdot D)^{0.7}}{\chi_t \cdot (a_w)^{-1} + 1.5} \quad (42)$$

which shows  $Q = 0.031$ ,  $F = 198$ ,  $r = 0.87$ ,  $s = 28.1$ , and  $\langle u \rangle = 22$ . The improvement relative to the single supraindex is not impressive, but now the following enhanced description with linear combinations of  $X_{BP}$  and  $\{a_w \cdot \chi\}$  supraindices can be obtained. Here  $X_{BP}$  is a dominant descriptor, which allows the reduction of the total combinatorial search into a forward selection search.

$$\{X_{BP}, \chi_t^v/a_w\}: Q = 0.034, F = 119, r = 0.894, \\ s = 26, \langle u \rangle = 13$$

$$\{X_{BP}, \chi_t^v/a_w, \chi_t^v/a_w\}: Q = 0.035, F = 85, \\ r = 0.902, s = 25, \langle u \rangle = 6.6$$

$$\{X_{BP}, \chi_t^v/a_w, \chi_t^v/a_w, \epsilon, M\}: Q = 0.039, F = 63, \\ r = 0.920, s = 23, \langle u \rangle = 4.6$$

The last description, even if its  $F$  and  $\langle u \rangle$  have worsened throughout the modeling, can be used to model the boiling points of the given solvents. The utility vector of the last combination  $\mathbf{u} = (4.9, 1.8, 2.7, 3.3, 2.8, 12.3)$  is somewhat misleading but can further be enhanced, up to  $\langle u \rangle = 11$  and  $u_1 = 17$ , with the introduction of orthogonalized descriptors.<sup>70</sup> The kind of descriptors involved in the modeling allow us to understand the basis of this property. While  $X_{BP}$  is made up of nonvalence molecular connectivity indices, which are shape-dependent, the improvement caused with the introduction of the total valence supraindex  $\chi_t^v/a_w$  underlines the importance of a pseudograph representation for these molecules. At the same time, the indirect influence of  $a_w$  and further of  $\epsilon$  shows that subtle electrostatic intermolecular interactions are also important for  $T_b$ . The additional improvement caused by the molar mass  $M$ , as without  $M$  we have  $Q = 0.037$ ,  $F = 69$ , and  $r = 0.909$ , seems to tell us that bulk factors contribute in some minor way to  $T_b$ . Previous studies on boiling points<sup>101,102</sup> of some nonpolar or slightly polar compounds have underlined the importance of polarizability and further of shape and size (molar volume) on  $T_b$ . Molar mass can indirectly help, through density, to model the size factor, which determines the boiling points. Clearly, better accuracy can be achieved and has been achieved with more homogeneous classes of compounds and with different molecular structure indices, like the recent modeling of  $C_2$ – $C_{10}$  alkenes and cycloalkenes,  $C_1$ – $C_2$  and  $C_1$ – $C_4$  chlorofluoroalkenes, and  $C_2$  chlorofluoroalkenes,<sup>40,103,104</sup> while modeling based on a linear combination of more sophisticated nontopological indices recently achieved a very interesting description of a wide set of boiling points and of critical transition temperatures.<sup>52</sup>

## 2. Refractive Index

The refractive index, RI or  $n_r$ , is related to the molar refractivity,  $R_m$ , through the equation  $R_m =$

$M(n_r^2 - 1)/d(n_r^2 + 1)$ , where  $M$  stands for molar mass and  $d$  for density. The best descriptor for the  $n = 61$  points of this property is  $\{\chi_t^v\}$  which rates  $Q = 6.00$ ,  $F = 50$ ,  $r = 0.676$ ,  $s = 0.11$ , while  $M$ ,  $\epsilon$ , and  $a_w$  show the following descriptive power

$$Q(M) = 3.52, F(M) = 17; Q(\epsilon) = 0.47, \\ F(\epsilon) = 0.31; Q(a_w) = 0.68, F(a_w) = 0.63$$

From these values and from the relation which relates  $R_m$  to RI, it should be expected that  $M$  should play some role in the description of RI. The following LCCL, where  $\chi_t^v$  is a dominant index, offers a better description for RI than the aforementioned empirical parameters

$$\{\chi_t^v, D^v\}: Q = 10.2, F = 72, r = 0.844, s = 0.08$$

$$\{\chi_t^v, \chi_t, D^v, {}^0\chi, a_w\}: Q = 14.8, F = 60, \\ r = 0.919, s = 0.06$$

The following dominant  $X_{RI}$  molecular connectivity term alone can explain most of the modeling with  $Q = 13.6$ ,  $F = 256$ ,  $r = 0.902$ ,  $s = 0.07$ ,  $\langle u \rangle = 90$

$$X_{RI} = (\chi_t^v)^3/(\chi_t)^{2.5} \quad (43)$$

This modeling can be improved with the following linear combinations, where the five-index combination includes the empirical parameters  $M$  and  $a_w$

$$\{X_{RI}, D\}: Q = 16.7, F = 198, r = 0.932, \\ s = 0.06, \langle u \rangle = 33$$

$$\{X_{RI}, D, {}^0\chi, M, a_w\}: Q = 19.9, F = 109, \\ r = 0.953, s = 0.05, \langle u \rangle = 11$$

Eliminating from the modeling the two strong outliers,  $CS_2$  and decaline, i.e., working with 59 points, the single- and multi-index descriptions improve to

$$\{X_{RI}\}: Q = 14.7, F = 290, r = 0.914, \\ s = 0.06, \langle u \rangle = 94$$

$$\{X_{RI}, D\}: Q = 20.8, F = 289, r = 0.955, \\ s = 0.05, \langle u \rangle = 38$$

$$\{X_{RI}, D, {}^0\chi, a_w, M\}: Q = 30.0, F = 242, \\ r = 0.979, s = 0.03, \langle u \rangle = 16$$

Relative to the significant improvement in  $Q$ , the small decrease in  $F$  throughout this series can be neglected while the overall utility continues to be meaningful. The correlation and utility vectors of the last combination are  $\mathbf{C} = (0.02135, 0.02478, -0.06271, 0.02591, 0.00103, 1.32902)$  and  $\mathbf{u} = (24, 10, 6.7, 5.0, 4.8, 60)$ . The modeling equation, to simplify matters, can be written in the following algebraic form:  $\mathbf{RI} = \mathbf{a} \cdot \mathbf{X}_{RI} + \mathbf{b} \cdot \mathbf{D} + \mathbf{c} \cdot {}^0\chi + \mathbf{d} \cdot \mathbf{a}_w + \mathbf{e} \cdot \mathbf{M} + \mathbf{f}$  where regression parameters  $\mathbf{a}$ – $\mathbf{f}$  are to be taken from the given correlation vector  $\mathbf{C}$ .

The modeling of RI, thus, requires (i) an  $a_w$  parameter which is not embedded in a term, (ii) a  $X_{RI}$  term made up of two total molecular connectivity

indices, (iii) two nonvalence  $D$  and  ${}^0\chi$  indices, and (iv)  $M$  to help to further improve the description, which without  $M$  lowers to  $Q = 25$ ,  $F = 212$ ,  $r = 0.970$ . Intermolecular interactions seem to play a lesser role here than in the  $T_b$  modeling. Notice that up to the rather good description with combination  $\{X_{RI}, D, {}^0\chi\}$ ,  $Q = 23.3$ ,  $F = 242$ ,  $r = 0.964$  for  $n = 59$  and  $Q = 17.9$ ,  $F = 147$ ,  $r = 0.941$ ,  $s = 0.05$  for  $n = 61$ , no intermolecular  $a_w$  or bulk  $M$  descriptors are required.

### 3. Density

Both single index and LCCI show no interesting modeling for the  $n = 62$  points of this property. The best single index for the density is  $\{\chi_t^v\}$  with  $Q = 1.64$ ,  $F = 12$ ,  $r = 0.41$ ,  $s = 0.25$ . Descriptors  $M$ ,  $\epsilon$ , and  $a_w$  perform as follows

$$Q(M) = 2.75, F(M) = 34; Q(\epsilon) = 0.21, \\ F(\epsilon) = 0.20; Q(a_w) = 0.16, F(a_w) = 0.11$$

That is, descriptor  $M$  rates better than  $\chi_t^v$  and should be expected to play an interesting role in the description of this property. The importance of  $M$  is not unexpected, as densities are strictly related to molar masses. In fact, the following linear combination of a molecular connectivity index and  $M$  show a remarkable modeling

$$\{{}^0\chi^v, M\}: Q = 9.71, F = 213, r = 0.937, s = 0.10$$

An improved description can be obtained with the following set of molar mass-based  $\chi$  indices, where the total indices are, instead, multiplied as they describe an inverted domain relative to the other indices

$$\{\chi \cdot M^{-1}\} = \{D/M, D^v/M, {}^0\chi/M, {}^0\chi^v/M, {}^1\chi/M, \\ {}^1\chi^v/M, \chi_t \cdot M, \chi_t^v \cdot M\} \quad (44)$$

These indices offer a significant single-index and multi-index modeling, where the single index is a dominant descriptor which transforms the full combinatorial search into a forward selection search

$$\{{}^0\chi^v/M\}: Q = 8.88, F = 357, r = 0.925, s = 0.10$$

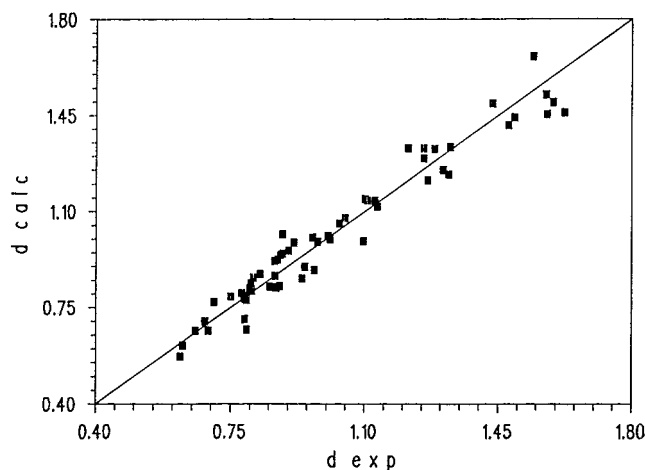
$$\{{}^0\chi^v/M, {}^1\chi/M\}: Q = 10.9, F = 270, r = 0.949, \\ s = 0.09$$

$$\{{}^0\chi^v/M, {}^1\chi/M, D/M\}: Q = 15.9, F = 380, \\ r = 0.975, s = 0.06$$

There is no improvement with more descriptors. The best overall description is achieved with the following semiempirical molecular connectivity term, derived by a trial-and-error procedure and centered around  ${}^0\chi^v$  and  $M$ . This term is a dead-end descriptor, as no further improvement can be achieved with a linear combination of this and other indices of any set

$$X_d = {}^0\chi^v \cdot ({}^1\chi + \chi_t)^{0.4} / M^{1.4} \quad (45)$$

Its statistics are  $Q = 15.7$ ,  $F = 1122$ ,  $r = 0.974$ ,  $s = 0.06$ ,  $\langle u \rangle = 48$ , a quite good score for a single



**Figure 12.** Plot of the calculated (calcd) versus the experimental (exp) densities,  $d$ , for 62 random organic solvents.

descriptor of 62 points. Its correlation, utility vectors and modeling equations are  $\mathbf{C} = (-96.3939, 2.10663)$ ,  $\mathbf{u} = (34, 63)$ , and  $\mathbf{d} = -96.39 \cdot X_d + 2.11$ . In Figure 12 the calculated versus the experimental density values are plotted. We can notice from the utility vector that the single utility of the  $X_d$  term and of the unitary term,  $U_0$  are excellent. The term given by eq 45 is a special case of a more convoluted semiempirical term given by eq 46, in which the dielectric constant also plays a direct role

$$X_d = \frac{{}^0\chi^v \cdot ({}^1\chi + \chi_t)^{0.4}}{M^{1.4} + (0.12\epsilon)^{1.2}} \quad (46)$$

This term rates even better than term 45, with  $Q = 16.2$ ,  $F = 1182$ , and  $r = 0.976$ . The small loss in quality of term 45 is more than compensated by its higher simplicity. Before closing this modeling, let us notice that around  ${}^0\chi^v$  is also centered a term which perfectly models the side-chain volume of amino acids, a property that surely has something to do with density.

### 4. Cutoff UV Values

The simulation of  $n = 37$  UV cutoff values starts very badly with index  $\{D^v\}$ , whose main statistical values are  $Q = 0.005$ ,  $F = 2.5$ ,  $r = 0.259$ . With normal  $\chi$  indices there is no way to improve this poor modeling, which seems to mimic the bad modeling that can be obtained with  $M$ ,  $\epsilon$ , and  $a_w$

$$Q(M) = 0.002, F(M) = 0.26; Q(\epsilon) = 0.003, \\ F(\epsilon) = 0.73; Q(a_w) = 0.0002, F(a_w) = 0.004$$

Even the introduction of the following type of semiempirical  $\epsilon/a_w$ -indices, which will later be used, does not improve the modeling

$$\{(\epsilon/a_w)\chi\} = \\ \{\epsilon \cdot D/a_w, \epsilon \cdot D^v/a_w, \epsilon \cdot {}^0\chi/a_w, \epsilon \cdot {}^0\chi^v/a_w, \epsilon \cdot {}^1\chi/a_w, \\ \epsilon \cdot {}^1\chi^v/a_w, \chi_t \cdot a_w/\epsilon, \chi_t^v \cdot a_w/\epsilon\} \quad (47)$$

In fact, the main statistics of the best single  $\epsilon/a_w$



**Table 12. Molecular Connectivity Indices for Compounds of Table 11**

solvent	D	D <sup>v</sup>	<sup>0</sup> χ	<sup>0</sup> χ <sup>v</sup>	<sup>1</sup> χ	<sup>1</sup> χ <sup>v</sup>	χ <sub>t</sub>	χ <sub>t</sub> <sup>v</sup>
acetone	6	12.0	3.57735	2.90825	1.73205	1.20412	0.57735	0.20412
acetonitrile	4	10.0	2.70711	1.94721	1.41421	0.72361	0.70711	0.22361
benzene	12	18	4.24264	3.46410	3	2	0.12500	0.03704
benzonitrile	16	28	5.81999	4.33397	3.93185	2.38429	0.07217	0.00717
1-butanol	8	12	4.12132	3.56853	2.41421	2.02333	0.35355	0.15811
2-butanone	8	14	4.28446	3.61536	2.27006	1.76478	0.40825	0.14434
butyl acetate	14	24	6.40578	5.43782	3.77006	2.90403	0.14434	0.02946
CS <sub>2</sub>	4	5.33333	2.70711	2.94949	1.41421	1.22474	0.70711	0.75000
CCl <sub>4</sub>	8	7.11112	4.50000	5.03557	2.00000	2.26778	0.50000	0.82653
Cl-benzene	14	19.7778	5.11288	4.52064	3.39385	2.47763	0.10206	0.03637
1Cl-butane	8	7.77778	4.12132	4.25521	2.41421	2.50889	0.35355	0.40089
CHCl <sub>3</sub>	6	5.33334	3.57735	3.97903	1.73205	1.96396	0.57735	0.84169
cyclohexane	12	12	4.24264	4.24264	3	3	0.12500	0.12500
cyclopentane	10	10	3.53553	3.53553	2	2	0.17678	0.17678
1,2-dCl-benzene	16	21.5556	5.98313	5.57718	3.80453	2.96124	0.08333	0.03571
1,2-dCl-ethane	6	5.55556	3.41421	3.68200	1.91421	2.10357	0.50000	0.64286
dCl-methane	4	3.55556	2.70711	2.97489	1.41421	1.60357	0.70711	0.90913
N,N-dM-acetamide	10	18	5.15470	4.35546	2.64273	1.82216	0.33333	0.09129
N,N-dM-formamide	8	16	4.28446	3.43281	2.27006	1.38833	0.40825	0.10541
1,4-dioxane	12	20	4.24264	3.64492	3	2.15470	0.12500	0.04167
ether	8	12	4.12132	3.82246	2.41421	1.99156	0.35355	0.20412
ethyl acetate	10	20	4.99156	4.02360	2.77006	1.90403	0.28868	0.05893
ethyl alcohol	4	8	2.70711	2.15432	1.41421	1.02333	0.70711	0.31623
heptane	12	12	5.53553	5.53553	3.41421	3.41421	0.17678	0.17678
hexane	10	10	4.82843	4.82843	2.91421	2.91421	0.25000	0.25000
2-methoxyethanol	8	16	4.12132	3.26968	2.41421	1.51315	0.35355	0.09129
methyl alcohol	2	6	2	1.44721	1	0.44721	1	0.44721
2-methylbutane	8	8	4.28446	4.28446	2.27006	2.27006	0.40825	0.40825
4-M-2-pentanone	12	18	5.86181	5.19271	3.12590	2.62063	0.23570	0.08333
2-M-1-propanol	8	12	4.28446	3.73167	2.27006	1.87918	0.40825	0.18257
2-M-2-propanol	8	12	4.5	3.94721	2	1.72361	0.5	0.22361
DMSO	6	8.66667	3.57735	3.63299	1.73205	2.94948	0.57735	0.5
nitromethane	6	18	3.57735	2.26371	1.73205	0.81236	0.57735	0.07454
1-octanol	16	20	6.94975	6.39696	4.41421	4.02333	0.08839	0.03953
pentane	8	8	4.12132	4.12132	2.41421	2.41421	0.35355	0.35355
3-pentanone	10	16	4.99156	4.32246	2.80806	2.32544	0.28868	0.10206
1-propanol	6	10	3.41421	2.86143	1.91421	1.52333	0.5	0.22361
2-propanol	6	10	3.57735	3.02456	1.73205	1.41290	0.57735	0.25820
pyridine	12	20	4.24264	3.33397	3	1.84973	0.12500	0.02869
ttCl-ethylene	10	11.1111	5.15470	5.53557	2.64273	2.51778	0.33333	0.41326
tt-hydrofuran	10	14	3.53553	3.23668	2.5	2.07735	0.17678	0.10206
toluene	14	20	5.11288	4.38675	3.39385	2.41068	0.10206	0.03208
1,1,2tClFethane	14	31.3333	7	5.53557	3.25	2.51778	0.25	0.01968
2,2,4-tM-pentane	14	14	6.78446	6.78446	3.41650	3.41650	0.20412	0.20412
<i>o</i> -xylene	16	22	5.98313	5.30940	3.80453	2.82735	0.08333	0.02778
<i>p</i> -xylene	16	22	5.98313	5.30940	3.78769	2.82137	0.08333	0.02778
acetic acid	6	16	3.57735	2.35546	1.73205	0.92773	0.57735	0.09129
decaline	22	22	6.81155	6.81155	4.96633	4.96633	0.02083	0.02083
dBr-methane	4	2.51852	2.70711	4.63502	1.41421	2.77746	0.70711	2.72740
1,2-dCl-E-enZ	6	7.55556	3.41421	3.42248	1.91421	1.64264	0.5	0.42857
1,2-dCl-E-enE	6	7.55556	3.41421	3.42248	1.91421	1.64264	0.5	0.42857
1,1-dCl-E-en	6	7.55556	3.57735	3.47489	1.73205	1.48745	0.57735	0.45457
dmethoxymethane	8	16	4.12132	3.52360	2.41421	1.39385	0.35355	0.11785
dmethyl ether	4	8	2.70711	2.40825	1.41421	0.81650	0.70711	0.40825
E-encarbonate	10	26	4.99156	3.04817	2.77006	1.27581	0.28868	0.01521
formamide	4	12	2.70711	1.56295	1.41421	0.56904	0.70711	0.13608
methyl chloride	2	1.77778	2	2.13389	1	1.13389	1	1.3389
morpholine	12	18	4.24264	3.73667	3	2.28446	0.125	0.05103
quinoline	22	34	6.81155	5.48867	4.96633	3.26450	0.02083	0.00239
SO <sub>2</sub>	4	13.6667	2.70711	1.59109	1.41421	0.63245	0.70711	0.12910
2,2-ttCl-ethane	10	9.11112	5.15470	5.69027	2.64273	2.95194	0.33333	0.55102
ttM-urea	14	24	6.73205	5.80268	3.55342	2.44019	0.19245	0.04082
tCl-E-en	8	9.33334	4.28446	4.47903	2.27006	2.07722	0.40825	0.42085

index,  $\{\epsilon \cdot D^v/a_w\}$ , are  $Q = 0.01$ ,  $F = 12$ ,  $r = 0.504$ . A closer look at this modeling lets us notice that leaving out 12 nonalcoholic solvents (with asterisks in Table 13) and introducing for alcohols the parameter  $a_{OH} = 2 + \epsilon/15$ , truncated at the second figure, instead of  $a_w$ , then the modeling of the remaining  $n = 25$  points with indices of set of eq 47, where  $a_{OH}$  replaces  $a_w$ , improves consistently. For  $n = 25$  points,  $M$ ,  $\epsilon$ ,

and  $a_{OH}$  continue to be poor descriptors with

$$Q(M) = 0.007, F(M) = 1.5; Q(\epsilon) = 0.007, \\ F(\epsilon) = 1.9; Q(a_{OH}) = 0.006, F(a_{OH}) = 1.4$$

Index  $\{D^v\}$  continues to be a bad descriptor with  $Q = 0.009$ ,  $F = 2.9$ , and  $r = 0.33$ . The best index,  $\{\epsilon \cdot D/a_{OH}\}$ , now guarantees a satisfactory model, while

**Table 13. Experimental (exp) and Calculated (calcd) UV Cutoff Values of  $n = 37$  Solvents**

solvent	UV <sub>exp</sub>	UV <sub>calcd</sub>	solvent	UV <sub>exp</sub>	UV <sub>calcd</sub>	solvent	UV <sub>exp</sub>	UV <sub>calcd</sub>
acetone*	330	269.9	1,2-dCl-benzene	295	310.6	methyl alcohol	205	215.2
acetonitrile*	190	247.5	1,2-dCl-ethane	225	212.2	4-M-2-pentanone	334	307.4
benzene*	280	213.7	dCl-methane	235	218.4	DMSO	268	250.9
1-butanol	215	219.8	<i>N,N</i> -dM-acetamd	268	331.0	nitromethane*	380	328.9
2-butanone	330	307.8	<i>N,N</i> -dM-formamd	268	303.1	pentane*	200	264.0
butyl acetate	254	247.8	1,4-dioxane	215	217.2	1-propanol	210	219.8
CS <sub>2</sub> *	380	299.2	ether	215	222.2	2-propanol	210	227.6
CCl <sub>4</sub> *	263	254.1	ethyl acetate	260	250.1	pyridine*	305	325.2
Cl-benzene*	287	250.6	ethyl alcohol	210	219.7	tt-hydrofuran*	215	247.4
1Cl-butane	225	216.5	heptane	200	228.0	toluene*	285	217.3
CHCl <sub>3</sub>	245	225.4	hexane*	200	242.1	1,1,2-tClthane	230	236.4
cyclohexane	200	220.9	2-methoxyethanol	220	228.3	2,2,4-tM-pentane	215	240.5
cyclopentane	200	229.9						

\*Twelve solvents not included in the optimal simulation are denoted by an asterisk.

linear combinations achieve no further improvement

$$\{\epsilon \cdot D/a_{\text{OH}}\}: Q = 0.063, F = 139, r = 0.926, s = 15$$

A trial-and-error search procedure discovers the brilliant semiempirical term of eq 48 centered around  $D^v$ ,  $a_{\text{OH}}$ , and  $\epsilon$  parameters. The full statistical values of this term that can be used to model the  $n = 25$  UV points are  $Q = 0.104$ ,  $F = 380$ ,  $r = 0.971$ ,  $s = 9.3$ ,  $\langle u \rangle = 42$

$$X_{\text{UV}} = \epsilon \cdot [(D^v)^{0.7} - 0.05 a_{\text{OH}}]/(a_{\text{OH}})^{1.5} \quad (48)$$

Linear combinations of this term with the supra-indices of eq 47 (with  $a_{\text{OH}}$  instead of  $a_w$ ) show interesting improvements in  $Q$ ,  $r$ , and  $s$

$$\{X_{\text{UV}}, \epsilon^1 \chi/a_{\text{OH}}, \epsilon^1 \chi^v/a_{\text{OH}}\}: Q = 0.127, F = 190, r = 0.982, s = 7.7, \langle u \rangle = 42$$

The following correlation vector and modeling equation based on the single term of eq 48 can be used to model the UV cutoff values,  $\mathbf{C} = (1.26155, 192.494)$ ,  $\mathbf{UV} = \mathbf{1.26} \cdot \mathbf{X}_{\text{UV}} + \mathbf{192}$ . The utility vector of the parameter of the linear regression,  $\mathbf{u} = (20, 64)$ , shows the very good utility of each parameter of the regression.

Now, let us reintroduce the 12 strong outliers, excluded from the previous modeling, and handle the  $n = 37$  solvents, with indices of set 47 and let us calculate their UV cutoff values with the optimal combination,  $\{\epsilon \cdot D/a_w, \epsilon \cdot D^v/a_w, \epsilon^1 \chi/a_w, \chi_t^v \cdot a_w/\epsilon\}$ , whose statistical values are  $Q = 0.019$ ,  $F = 16$ , and  $r = 0.70$ . This rather poor modeling is nevertheless able to predict satisfactory UV cutoff values for the  $n = 37$  points, as can be seen in Table 13. This is to say that even without extraordinary statistics, it is possible to obtain  $P_{\text{calcd}}$  values that are not quite absurd.

### 5. Dipole Moment

The modeling of the polarity of solvents started soon after the introduction of molecular connectivity indices,<sup>16</sup> and the proposed descriptor for their modeling was based on the  ${}^1\chi^v$  index. It is interesting to notice that also for the modeling of the dipole moment of the  $n = 35$  solvent molecules of Table 11, the  ${}^1\chi^v$  index is the best single index with  $Q = 0.17$ ,  $F = 1.8$ ,  $r = 0.23$ ,  $s = 1.3$ . This is evidently a rather poor

modeling and is similar to the modeling achieved by  $M$ ,  $\epsilon$ , and  $a_w$

$$Q(M) = 0.18, F(M) = 1.9; Q(\epsilon) = 0.74, F(\epsilon) = 33; Q(a_w) = 0.6, F(a_w) = 25$$

From these ratings we can notice that  $\epsilon$  should play a consistent role in the modeling of this property. Introduction and use of the following set of semiempirical descriptors improves the description consistently

$$\{\epsilon \cdot \chi\} = \{\epsilon \cdot D, \epsilon \cdot D^v, \epsilon^0 \chi, \epsilon^0 \chi^v, \epsilon^1 \chi, \epsilon^1 \chi^v, \chi_t/\epsilon, \chi_t^v/\epsilon\} \quad (49)$$

The following single- and a two-index combination are, in fact, interesting

$$\{\epsilon^0 \chi\}: Q = 1.06, F = 68, r = 0.82, s = 0.8$$

$$\{\epsilon^0 \chi, \chi_t/\epsilon\}: Q = 1.35, F = 55, r = 0.88, s = 0.7$$

With more descriptors the modeling starts to worsen. The only decisive improvement is obtained with the following semiempirical term, which is strongly dependent on  $\epsilon$  and on  $D^v$  and whose statistical values are noteworthy:  $Q = 2.12$ ,  $F = 272$ ,  $r = 0.94$ ,  $s = 0.4$ ,  $\langle u \rangle = 9.7$

$$X_\mu = \left( \frac{\epsilon D^v - 1.9D}{2.5D^v - \epsilon \chi_t} \right)^{0.45} \quad (50)$$

Its correlation and utility vectors are  $\mathbf{C} = (1.43421, -0.46761)$ ,  $\mathbf{u} = (17, 3)$ . The modeling equation can succinctly be written as  $\mu = \mathbf{1.43} \cdot \mathbf{X}_\mu - \mathbf{0.47}$ . The description of this property can further be improved at the  $Q$ ,  $r$ , and  $s$  level with the following combination

$$\{X_\mu, {}^1\chi^v\}: Q = 2.24, F = 151, r = 0.951, s = 0.4, \langle u \rangle = 7.4$$

### I. Modeling and Cis/Trans Isomerism

One of the major problems in chemical graph theory has been to differentiate between different conformers of a molecule such as cis and trans isomers. Two different solutions have been proposed: one<sup>79,80</sup> based on metric rather than on topological considerations and the other<sup>58</sup> based on

topological considerations. This last solution centers its attention on virtual ring fragments from which it is possible to derive a specific  $\chi_{ct}$  descriptor as is explained in section D of this review. For a better understanding of this cis/trans procedure, let us consider the hexatriene **1–6** graphs of Figure 3, where we notice that the delta vector of the central four atoms is  $\delta = (2, 2, 2, 2)$  while the corresponding  $\delta/\delta^r$  vector for graphs **2** and **3**, which can give rise to only one four-membered virtual ring fragment each, are  $\delta/\delta^r(\mathbf{2}) = (3, 2, 2, 3)$  and  $\delta/\delta^r(\mathbf{3}) = (3, 2, 2, 2)$ . Graph **4** can, instead, give rise to two identical four-membered virtual ring fragments whose  $\delta/\delta^r$  vector is  $\delta/\delta^r(\mathbf{4}) = (3, 3, 2, 2)$ . Here the second 3 is due to the adjacent virtual ring. Further, graph **5** can give rise to two four-membered virtual ring fragments whose  $\delta/\delta^r$  vectors are  $\delta/\delta^r(\mathbf{5}) = (3, 2, 2, 3)$  and  $\delta/\delta^r(\mathbf{5}) = (3, 2, 2, 2)$ , and finally, graph **6** can give rise to three four-membered virtual ring fragments whose  $\delta/\delta^r$  vectors are  $\delta/\delta^r(\mathbf{6}) = (3, 2, 2, 3)$  and  $\delta/\delta^r(\mathbf{6}) = (2, 2, 2, 3)$ , where this last vector is taken twice. We will briefly review the modeling power of the proposed index for two different physicochemical properties of a set of olefins: the boiling point,  $T_b$ , of 12 olefins and the molar refractivity,  $MR_D$ , of 8 olefins, whose experimental values together with the corresponding crucial molecular connectivity indices,  ${}^1\chi$  and  $\chi_{ct}$ , are given in Table 14. The modeling of  $T_b$  and  $MR_D$  with the normal  ${}^1\chi$  index and with the  $\chi_{ct}$  index for  $n = 12$  and 8 points, respectively, shows the following results

$$T_b$$

$$\{{}^1\chi\}: Q = 0.216, F = 1593, r = 0.99688, s = 4.6, \\ \langle u \rangle = 30, \mathbf{u} = (40, 19)$$

$$\{\chi_{ct}\}: Q = 0.217, F = 1600, r = 0.99689, s = 4.6, \\ \langle u \rangle = 30, \mathbf{u} = (40, 19)$$

$$MR_D$$

$$\{{}^1\chi\}: Q = 14.1, F = 68476, r = 0.99996, s = 0.07, \\ \langle u \rangle = 140, \mathbf{u} = (262, 18)$$

$$\{\chi_{ct}\} Q = 21.5, F = 160426, r = 0.99998, \\ s = 0.05, \langle u \rangle = 214, \mathbf{u} = (401, 28)$$

The small differences in statistical values between  ${}^1\chi$  and  $\chi_{ct}$  in modeling the boiling points can be ascribed to the very small and random  $\Delta_{ct}$  difference between the  $T_b$  values of the cis and trans isomer. In fact, for the boiling points,  $\Delta_{ct}$  ranges from  $-0.7$  to  $1.2$ , with an average of  $\langle \Delta_{ct} \rangle = -0.05$ , excluding the consistent difference between *trans*- and *cis*-butene. For the molar refractivity there is no such random variation and the  $MR_D$  value of the cis isomer is always smaller than the corresponding trans value with an average  $\langle \Delta_{ct} \rangle = 0.1225$ .

In Table 14 we notice that *trans*-2-octene, *trans*-3-octene, and *trans*-4-octene as well as *cis*-3-octene and *cis*-4-octene isomers show the same  $\chi_{ct}$  values. This fact means that the  $\chi_{ct}$  connectivity index is not able to distinguish between the different positional isomers of the given octenes. We will test the follow-

**Table 14. Calculated  ${}^1\chi$  and  $\chi_{ct}$  Values Together with  $n = 12$  Experimental Boiling Points,  $T_b$  ( $^{\circ}\text{C}$ ), and  $n = 8$  Molar Refractivity,  $MR_D$ , Points of Cis(c)/Trans(t) Olefins (taken from ref 88)**

olefins	${}^1\chi$	$\chi_{ct}$	$(1/p)^n$	$T_b$	$MR_D$
t-2-butene	1.91421	1.91421		0.88	
c-2-butene	1.91421	1.89859	0.0625	3.7	
t-2-pentene	2.41421	2.42421			
c-2-pentene	2.41421	2.40571	0.03125		
t-3-hexene	2.91421	2.91421		67.5	29.72
c-3-hexene	2.91421	2.90958	0.00137	66.85	29.61
t-2-octene	3.91421	3.91421		123.4	
c-2-octene	3.91421	3.90571	0.003906	124.6	
t-3-octene	3.91421	3.91421		122.4	39.04
c-3-octene	3.91421	3.90958	0.000152	122.3	38.91
t-4-octene	3.91421	3.91421		121.4	39.05
c-4-octene	3.91421	3.90958	0.000015	121.7	38.92
t-5-decene	4.91421	4.91421		170.2	48.34
c-5-decene	4.91421	4.90958	$10^{-7}$	169.5	48.22

ing positional cis/trans connectivity index  $\chi_{ct}(p)$ , defined in eq 51, which takes into account the different positions of the double bond in the given compounds. Such a test only has a wide orientating value as the properties of three different octene isomers only are known

$$\chi_{ct}(p) = \chi_{ct} - (1/p)^n \quad (51)$$

where  $p$  stands for position and  $n$  for the length of the chain, i.e., number of carbon atoms. In the third column of Table 14 are reported the  $(1/p)^n$  values. From eq 51 and from found  $(1/p)^n$  values, we notice that the more the double bond shifts toward the molecule midpoint, the more  $(1/p)^n$  decreases and the more  $\chi_{ct}(p)$  corresponds to  $\chi_{ct}$ . The statistical performance of the newly defined  $\chi_{ct}(p)$  index is

$$T_b: Q = 0.223, F = 1687, r = 0.997, \\ s = 4.5, \langle u \rangle = 30, \mathbf{u} = (41, 20)$$

$$MR_D: Q = 22.4, F = 173576, r = 0.99998, \\ s = 0.045, \langle u \rangle = 223, \mathbf{u} = (417, 29)$$

The new positional index,  $\chi_{ct}(p)$ , is thus able to improve in a small but noticeable way the modeling of both properties.

## J. Orthogonal Descriptors

In section H orthogonal descriptors were introduced; let us now, then, discuss an example of how these types of descriptors work. The property chosen is the lattice enthalpy of metal halides, treated in section F. The best descriptor of this property is a linear combination made up of a normal  $\chi$  index, a  $z$ -based index, and a molecular connectivity term, i.e.,  $\{\chi^v, D^z, {}^1R^v\}$ . Even if these indices are not highly correlated, with  $r(\chi^v, D^z) = 0.46$ ,  $r(\chi^v, {}^1R^v) = 0.69$ , and  $r(D^z, {}^1R^v) = 0.66$ , the introduction of the corresponding orthogonal indices brings some practical advantages, which, with higher correlated indices, are of paramount importance. The first step of the orthogonalization procedure is to rewrite the descriptor vector in a sequential order, i.e., the first best index followed by the second best index and then by the third best index. The reordered vector is  $\mathbf{X} = ({}^1R^v,$



${}^0\chi^v$ ,  $D^z$ ,  $U_0$ ), and its utility and correlation vectors are  $\mathbf{u} = (5.4, 7.7, 6.6, 30)$  and  $\mathbf{C} = (64.4373, -20.1665, 54.5942, 719.386)$ . Now, the first orthogonal index is equal to the first normal index,  ${}^1\Omega \equiv {}^1R^v$ , and Randić's orthogonalization procedure goes by orthogonalizing in a sequential way  ${}^0\chi^v \rightarrow {}^2\Omega$  and  $D^z \rightarrow {}^3\Omega$  (see section H). Even if the unitary term,  $U_0 \equiv 1$ , does not change, its regression parameter,  $u_U$ , nevertheless changes in the orthogonalized regression. The regression and utility vectors of the orthogonal vector,  $\Omega = ({}^1\Omega, {}^2\Omega, {}^3\Omega, U_0)$ , thus derived are  $\mathbf{C}(\Omega) = (160.628, -20.1725, 54.5942, 682.282)$  and  $\mathbf{u}(\Omega) = (22, 7.7, 6.6, 131)$ . It can readily be noticed that  $\langle u \rangle$  improves from 12 to 52 and that this improvement is mainly due to the first index and to the unitary index  $U_0$ . The first index is thus confirmed as the dominant descriptor, and the sequential ordering of the utilities show the importance of each descriptor, because now the orthogonal regression is stable, i.e.,  $\mathbf{C}(\Omega)$  vector values are constant under inclusion or deletion of a new index. The new regression equation has not only the same good quality of the parent regression, but also an enhanced utility and a total stability. Further, to derive the orthogonal regression equation, i.e., to derive the  $\mathbf{C}(\Omega)$  values, there is no need to calculate the values of the orthogonal indices because each  $C_i(\Omega)$  can be derived by the aid of the sequential regressions obtained by adding each time the next best  $\chi$  index and retaining the regression parameter of this index. Let us see in detail how this powerful and automatic stepwise method works with the  $\Delta H_L^\ominus$  case. The stepwise modeling regressions, starting with descriptor  ${}^1R^v$  (eq 52), adding next index  ${}^0\chi^v$  (eq 53), and finally index  $D^z$  (eq 54) and including explicitly the unitary term,  $U_0$ , to render things easier, are

$$160.628 \cdot {}^1R^v + 682.282 U_0 \quad (52)$$

$$106.517 \cdot {}^1R^v - 20.2175 \cdot {}^0\chi^v + 818.387 U_0 \quad (53)$$

$$64.4373 \cdot {}^1R^v - 20.1665 \cdot {}^0\chi^v + 54.5942 D^z + 719.386 U_0 \quad (54)$$

Now, if these regression parameters are compared with the aforementioned  $C_i(\Omega)$  values of the orthogonal correlation vector  $\mathbf{C}(\Omega) = (160.628, -20.1725, 54.5942, 682.282)$ , it is at once evident how the orthogonal correlation vector can be 'constructed' without even deriving the specific  ${}^i\Omega$  values (here  $i = 1-4$  and  ${}^4\Omega \equiv U_0$ ). From eqs 52-54 it is obvious that (i) in stepwise  $\chi$ - and/or X-regressions  $c_i(\chi)$  or  $c_i(X)$  changes with the inclusion of the next descriptor, (ii)  $c_i(\Omega)$ , instead, remains constant, and (iii) the parameter of the unitary index is the parameter,  $c_U$ , of the single- $\chi$  or -X regression made up of the best single descriptor plus  $U_0$ . This strategy is evident with the third descriptor, where the regression parameter both in the normal and orthogonal description are the same, because this is the last added descriptor. The importance of the interrelation between descriptors is underlined by the similarity of the second regression parameter in both normal and orthogonal representations. The low interrelation

value between  ${}^0\chi^v$  and  $D^z$  indices ( $r = 0.46$ ) renders the introduction of the third index,  $D^z$ , in eq 53 nearly unnoticeable for the correlation parameter of the second  ${}^0\chi^v$  index that remains practically constant. Instead, the limited interrelation between  ${}^1R^v$  and  ${}^0\chi^v$  ( $r = 0.69$ ) and between  ${}^1R^v$  and  $D^z$  (0.66) is enough to render the regression parameters of  ${}^1R^v$  and  $U_0$  seemingly 'random' with the addition of the next descriptor.

#### IV. Recent and Alternative Elaborations

Chemical graph theory and molecular connectivity have also developed along other lines. An already cited book,<sup>5</sup> together with other publications,<sup>86,105-109</sup> constitute a quite exhaustive review on QSPR/QSAR methodologies not included in the present review. Concerning molecular connectivity theory, some interesting aspects have been developed during the past years. One of the major concerns of molecular connectivity, the meaning of  $\chi$  indices, has recently been worked out around the concept of bimolecular accessibility,<sup>110</sup> while variable connectivity indices, introduced in 1992, have further been developed during recent years.<sup>111-115</sup> Furthermore, molecular connectivity descriptors based on line graphs and molecular connectivity edge indices with optimum exponent<sup>116-119</sup> are starting to be used as descriptors in QSAR/QSPR studies, where a line graph  $L(G)$  of a graph  $G$  is a graph derived from  $G$  in such a way that the edges in  $G$  are replaced by vertexes in  $L$ . It should here be noticed that the trial-and-error construction procedure of molecular connectivity terms indirectly provides descriptors with optimized exponents. Along a somewhat different line of reasoning, but always based on molecular connectivity chemical graph concepts, a new molecular structure descriptor, known as the electrotopological E state has recently been developed,<sup>120</sup> which seems quite powerful in predicting the activity and properties of drugs. In a different but very fruitful perspective new graph concepts have been developed which are able to describe chemical reactions and their intrinsic mechanisms.<sup>121</sup> Finally, a recent publication in *Nature*<sup>122</sup> cannot be forgotten, which uses topological graph descriptors for the rational design of immunosuppressive compounds. This last publication practically shows the usefulness of topological indices in filtering a huge mass of compounds (300 000) in order to design a compound whose activity is about 100 times the activity of the initial lead compound.

#### V. Conclusion

The procedure outlined throughout this review, which is practically focused on the passage from molecular connectivity indices to molecular connectivity terms, can be defined as *heuristics*, if for *heuristics* it is meant a procedure that provides aid or direction in the solution of a problem. This *heuristics* is mainly based on concepts derived from the chemical graph theory, which, like every theory, is set up on a set of rules that define boundaries and specify how to be successful at and within these boundaries. Success is measured by the problems

that can be solved using these rules. What we have put together throughout this work is a scheme with a set of rules. The scheme involves the construction of molecular connectivity indices or terms, their linear combinations, and eventual changes that should be done to meet the requirement to model properties of different classes of compounds. It is set up this way because we have no way to calculate in a direct, easy, and straightforward way the values of the properties. The remarkable thing about this scheme is its *generality*. It seems to apply equally well to amino acids and inorganic salts, to purines and pyrimidines, to alkanes, to mixed classes of compounds, and to a highly heterogeneous class of compounds. It is far easier than quantum mechanical methods, even if it does not have their physical charisma and their mathematical rigor, but it is extremely practical. Plainly, it just works. We can never be definitely sure that it will always work, since *heuristics* are rather problem-specific, but on the basis of the results obtained up to now, we are rather confident that with it one can try all sorts of modeling and that is its use.

## VI. Acknowledgments

I thank Professor Lemont B. Kier of the Virginia Commonwealth University for his suggestion and support for this review. Special thanks are due to Professor Milan Randić of Drake University and to two anonymous and attentive reviewers, who scrupulously read the entire article and suggested many interesting improvements. I also thank, during my long journey in the realms of chemical application of graph theory, S. Basak, of the University of Minnesota, L. H. Hall of the Eastern Nazarene College, and N. Trinajstić of Zagreb University for their help. I am also indebted to G. Trotta of the Dipartimento di Scienze della Terra of this University for precious technical help. The kind words of the European Editor are also acknowledged. The Italian Ministry for University and Scientific and Technological Research (MURST) is acknowledged for financial support.

## VII. Glossary

congruence	two or more geometric figures are congruent if they differ only in location in space
similarity	two geometric figures are similar if one is the enlargement of the other, i.e., two similar polygons have corresponding angles equal and proportional corresponding sides. A less rigorous definition of similarity well-suited for chemical applications can be found in ref 123
distance matrix	constructed with the set of topological distances, i.e., the number of connections in the shortest path between atoms <i>i</i> and <i>j</i> . Cyclic graphs present special problems since the distance between two points may be traversed along more than one path
AA	amino acids, normally natural amino acids
BP	boiling points
CD	crystal density
LCCI	linear combination of molecular connectivity indices

LCOCI	linear combination of orthogonal molecular connectivity indices
LCRCI	linear combination of reciprocal molecular connectivity indices
LCSCI	linear combination of squared molecular connectivity indices
LCXCI	linear combination of special molecular connectivity indices
MC	molecular connectivity
MCI	molecular connectivity indices
MeCl	metal chlorides
MP	melting points
MR <sub>D</sub>	molar refractivity
MeX	metal halides
MON	motor octane number
pI	pH at the isoelectric point
PP	purine and pyrimidine bases
R <sub>r</sub>	retention index for paper chromatography
RI	refractive index
S	solubility
SR	specific rotation
QSAR	quantitative structure–activity relationships
QSPR	quantitative structure–property relationships
UWC	unfrozen water content
V	side-chain molecular volume

## VIII. References

- (1) Rosen, K. H. *Discrete mathematics and its applications*, McGraw-Hill: New York, 1995. Bamberg, O.; Sternberg, S. *A Course in Mathematics for Students of Physics*, Cambridge University Press: Cambridge, 1990; Vol. 2. Harary, F. *Graph Theory*, 2nd printing; Addison-Wesley: Reading, MA, 1971.
- (2) *Chemical applications of Graph theory*, Balaban, A. T., Ed.; Academic Press: London, 1986; *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334.
- (3) Kier, L. B.; Hall L. H. *Molecular connectivity in structure–activity analysis*; Wiley: New York, 1986.
- (4) Trinajstić, N. *Chemical graph theory*, 2nd ed.; CRC Press: Boca Raton, 1992.
- (5) Reinhard, M.; Drefahl, A. *Handbook for Estimating Physico-chemical Properties of Organic Compounds*, Wiley: New York, 1999.
- (6) Rouvray, D. H. *J. Mol. Struct. (THEOCHEM)* **1989**, *185*, 187.
- (7) Turro, N. J. *Angew. Chem., Int. Ed. Engl.* **1986**, *25*, 882.
- (8) Randić, M.; Trinajstić, N. *Croat. Chem. Acta* **1994**, *67*, 1.
- (9) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- (10) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. *J. Pharm. Sci.* **1975**, *64*, 1971.
- (11) Kier, L. B.; Hall, L. H.; Murray, W. J. *J. Pharm. Sci.* **1975**, *64*, 1974.
- (12) Hall, L. H.; Kier, L. B. *J. Pharm. Sci.* **1976**, *65*, 1806.
- (13) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press: New York, 1976.
- (14) Hall, L. H.; Kier, L. B. *Tetrahedron* **1977**, *33*, 1953.
- (15) Kier, L. B.; Hall, L. H. *J. Pharm. Sci.* **1981**, *70*, 583.
- (16) Kier, L. B. *J. Pharm. Sci.* **1981**, *70*, 930.
- (17) Edward, J. T. *Can. J. Chem.* **1982**, *60*, 480.
- (18) Aravindakshan, P. *Mater. Chem. Phys.* **1983**, *8*, 171.
- (19) Mekenyan, O.; Bonchev, D.; Balaban, A. T. *Chem. Phys. Lett.* **1984**, *109*, 85.
- (20) Seybold, P. G.; May, M.; Bagal, A. U. *J. Chem. Educ.* **1987**, *64*, 575.
- (21) Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1988**, *15*, 201.
- (22) Randić, M.; Hansen, P. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 60.
- (23) Hansen, P. J.; Jurs, P. C. *J. Chem. Educ.* **1988**, *65*, 574.
- (24) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discr. Appl. Math.* **1988**, *19*, 17.
- (25) Needham, D. E.; Wei I.-C.; Seybold, P. G. *J. Am. Chem. Soc.* **1988**, *110*, 4186.
- (26) Randić, M. *N. J. Chem.* **1991**, *15*, 517.
- (27) Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
- (28) Randić, M. *Croat. Chem. Acta* **1991**, *64*, 43.
- (29) Randić, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.
- (30) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1991**, *7*, 243.
- (31) Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109.

- (32) Maier, B. J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 87.
- (33) Mihalić, Z.; Trinajstić, N. *J. Chem. Ed.* **1992**, *69*, 701.
- (34) Balaban, T. A.; Kier, L. B.; Joshi, N. *Match (Comm. Math. Chem.)* **1992**, *28*, 13.
- (35) Mihalić, Z.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28.
- (36) Kier, L. B.; Hall, L. H.; Frazer, J. W. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 3, 143.
- (37) Hall, L. H.; Kier, L. B.; Frazer, J. W. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 148.
- (38) Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1994**, *2*, 289.
- (39) Balaban, A. T.; Bertelsen, S. *Match (Comm. Math. Chem.)* **1994**, *30*, 55.
- (40) Balaban, A. T.; Basak, S. C.; Colburn T.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118.
- (41) Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1994**, *21*, 215.
- (42) Galvez, J.; Garcia, R.; Salbert, M. T.; Soler R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520.
- (43) Galvez, J.; Garcia-Domenech R.; de Julian-Ortiz, J. V.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272.
- (44) Lucić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. *Croat. Chem. Acta* **1995**, *68*, 435.
- (45) Basak, S. C.; Grunwald, G. D. *New J. Chem.* **1995**, *19*, 231.
- (46) Estrada, E. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31.
- (47) Galvez, J.; Garcia-Domenech, R.; De Gregorio Alapont, C.; de Julian-Ortiz, J. V.; Popa, L. *J. Mol. Graphics* **1996**, *14*, 272.
- (48) Garcia-Domenech, R.; de Gregorio Alapont C.; de Julian-Ortiz, J. V.; Galvez, J. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 567.
- (49) Galvez, J. *J. Mol. Struct. (THEOCHEM)* **1998**, *429*, 255.
- (50) Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23.
- (51) Kuanar, M.; Mishra, B. K. *Bull. Chem. Soc. Jpn.* **1998**, *71*, 191.
- (52) Katritzky, A. R.; Mu, L. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28, 293, and 300.
- (53) Basak, S. C.; Gute, B. D. *SAR QSAR Environ. Res.* **1999**, *10*, 1.
- (54) Trinajstić, N.; Mihalić, Z.; Harris, F. E. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1994**, *28*, 525.
- (55) Pogliani, L. *J. Pharm. Sci.* **1992**, *81*, 334 and 967.
- (56) Pogliani, L. *Comput. Chem.* **1993**, *17*, 283.
- (57) Pogliani, L. *J. Phys. Chem.* **1993**, *97*, 6731.
- (58) Pogliani, L. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 801.
- (59) Pogliani, L. *J. Phys. Chem.* **1994**, *98*, 1494.
- (60) Pogliani, L. *Amino Acids* **1995**, *9*, 217.
- (61) Pogliani, L. *J. Phys. Chem.* **1995**, *99*, 925.
- (62) Pogliani, L. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1082.
- (63) Pogliani, L. *Croat. Chem. Acta* **1996**, *69*, 95.
- (64) Pogliani, L. *Croat. Chem. Acta* **1997**, *70*, 803.
- (65) Pogliani, L. *J. Phys. Chem.* **1996**, *100*, 18065.
- (66) Pogliani, L. *Med. Chem. Res.* **1997**, *7*, 380.
- (67) Pogliani, L. *Amino Acids* **1997**, *13*, 237.
- (68) Pogliani, L. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 104.
- (69) Pogliani, L. *J. Mol. Struct. (THEOCHEM)* **1999**, *466*, 1.
- (70) Pogliani, L. *J. Phys. Chem.* **1999**, *103*, 1598.
- (71) Pogliani, L. The Concept of Graph Mass in Molecular Graph Theory. A Case in Data Reduction Analysis. In *QSAR studies by Molecular Descriptors*; Diudea, M., Ed.; Nova Science Publishers Inc.: New York, 2000.
- (72) Hosoyea, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
- (73) Smolenski, E. A. *Russ. J. Phys. Chem.* **1964**, *38*, 700.
- (74) Balaban A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
- (75) Brown, R. D.; Martin, Y. C. *SAR QSAR Environ. Res.* **1998**, *8*, 23.
- (76) Cumming, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750.
- (77) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909.
- (78) Zheng, W.; Cho, S. J.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 252 and 259.
- (79) Randić, M.; Jerman-Blazic, B.; Trinajstić, N. *Comput. Chem.* **1990**, *14*, 237.
- (80) Randić, M. *J. Quantum Chem.: Quantum Biol. Symp.* **1988**, *15*, 201.
- (81) Randić, M. *Croat. Chem. Acta* **1993**, *66*, 289.
- (82) Lucić, B.; Trinajstić, N. *SAR QSAR Environ. Res.* **1997**, *7*, 45.
- (83) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
- (84) Bridgman, P. W. *Dimensional Analysis*; Yale University Press: New Haven, 1931.
- (85) Berberan-Santos, M. N.; Pogliani, L. *J. Math. Chem.* **1999**, *26*, 255.
- (86) Randić, M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672.
- (87) Kier, L. B.; Hall, L. H. *J. Pharm. Sci.* **1981**, *70*, 583.
- (88) Chang, R. *Physical Chemistry with Applications to Biological Systems*; McMillan: New York, 1977; pp 357–361.
- (89) *CRC Handbook of chemistry and physics*, 65<sup>th</sup> ed; Weast, C. R., Ed.-in-Chief.; Boca Raton, FL, 1984–1985; pp C-724–C-727.
- (90) *CRC Handbook of chemistry and physics*, 72<sup>nd</sup> ed; Lide, R. D., Ed.-in-Chief.; Boca Raton, FL, 1991–1992; pp 7-1–7-3.
- (91) *CRC Handbook of Chemistry and Physics*, 74<sup>th</sup> ed.; Lide, R. D., Ed.-in-Chief.; Boca Raton, FL, 1993–1994; pp 6-148–6-155.
- (92) Nakashima, N.; Suzuki E.-I. *Appl. Spectrosc. Rev.* **1984**, *20*, 1.
- (93) Ladik, J.; Appel, K. *Theor. Chim. Acta* **1966**, *4*, 132.
- (94) Xu, L.; Wang, H. W.; Su, Q. *Comput. Chem.* **1992**, *16*, 187 and 195.
- (95) Atkins, P. W. *Physical Chemistry*; Oxford University Press: Oxford, 1990; p 933.
- (96) *Bruker Almanac*; Bruker Scientific Instruments: Rheinstetten, Germany, 1986; pp 106–107.
- (97) *Aldrich Solvents*; Aldrich Chemical Co., Inc.: Milwaukee, WI, 1995; p 9.
- (98) Gilman, J. J. *J. Chem. Educ.* **1999**, *76*, 1330.
- (99) Standard, J. M.; Clark, B. K. *J. Chem. Educ.* **1999**, *76*, 1363.
- (100) Laing, M. *Educ. Chem.* **1993**, *30*, 160.
- (101) Myers, R. T. *J. Phys. Chem.* **1979**, *83*, 294.
- (102) Rich, R. L. *Bull. Chem. Soc. Jpn.* **1993**, *66*, 1065.
- (103) Hansen, P. J.; Jurs, P. C. *Anal. Chem.* **1987**, *59*, 2322.
- (104) Carlton, T. S. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 158.
- (105) *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997.
- (106) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: UK, 1999.
- (107) Basak, S.; Balaban, T.; Grunwald, G. D.; Gute, B. D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891.
- (108) Gutman, I.; Popovic, Lj.; Estrada, E.; Bertz, S. H. *ACH-Models Chem.* **1998**, *135*, 147.
- (109) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley: New York, in press.
- (110) Kier, L. B.; Hall, L. H. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792.
- (111) Randić, M. *Chemom. Intel. Lab. Syst.* **1991**, *10*, 213.
- (112) Randić, M. *J. Comput. Chem.* **1991**, *12*, 970.
- (113) Randić, M.; Dobrowolski, J. Sz. *Int. J. Quantum Chem.* **1998**, *70*, 1209.
- (114) Randić, M. Topological Indices. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, 1998; pp 3018–3032.
- (115) Randić, M.; Basak, S. C.; Mills, D.; Pogliani, L. *New J. Chem.*, submitted for publication.
- (116) Amić, D.; Beslo, D.; Lucić, B.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 819.
- (117) Burden, F. R.; Winkler, D. A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 236.
- (118) Randić, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 26 1.
- (119) Basak, S.; Nikolic, S.; Trinajstić, N.; Amic, D.; Beslo, D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 927.
- (120) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: New York, 1999.
- (121) Temkin, O. N.; Zeigarnik, A. V.; Bonchev, D. *Chemical Reaction Networks*; CRC Press: New York, 1996.
- (122) Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floch R.; Buelow, R. *Nat. Biotechnol.* **1998**, *16*, 748.
- (123) Randić, M. Similarity Methods of Interest in Chemistry. In *Mathematical Methods in Contemporary Chemistry*; Kuchanov, I. S., Ed.; Gordon & Breach: Amsterdam, 1996; pp 1–100.

CR0004456