Comparison of Weighting Schemes for Molecular Graph Descriptors: Application in Quantitative Structure−Retention Relationship Models for Alkylphenols in Gas−Liquid Chromatography

Ovidiu Ivanciuc,*,† Teodora Ivanciuc,† Daniel Cabrol-Bass,*,‡ and Alexandru T. Balaban*,†

Department of Organic Chemistry, Faculty of Chemical Technology, University "Politehnica" of Bucharest, Oficiul 12 CP 243, 78100 Bucharest, Romania, and GRECFO-LARTIC, University of Nice−Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 2, France

Received September 16, 1999

Organic compounds containing heteroatoms or multiple bonds can be conveniently represented as vertex- and edge-weighted molecular graphs. These atom and bond parameters can be computed for any organic compound with two parameter sets that we have recently defined, namely, the relative electronegativity *X* and the relative covalent radius *Y* weighting schemes. Structural descriptors computed with these two weighting schemes and the previously defined atomic number *Z* parameter set are used to develop quantitative structure−retention relationship (QSRR) models for alkylphenols in gas−liquid chromatography. The QSRR models are generated with structural descriptors computed with several newly introduced graph operators, namely, the Wiener, hyper-Wiener, minimum eigenvalue, maximum eigenvalue, Ivanciuc−Balaban, and information on distance operators. These molecular graph operators were applied to the distance **D** and the reciprocal distance **RD** matrixes.

INTRODUCTION

Among the several hundreds of chemical descriptors and topological indices (TIs) that model the structure of organic compounds using the graph representation of molecules,[1−14] the large majority are defined only for simple graphs that represent alkanes and cycloalkanes. The first cause of this situation is that alkanes and cycloalkanes represent an ideal class of compounds for investigating the influence of chemical bonding, size, branching, cyclicity, and shape on the variation of molecular properties. A large number of mathematical relationships were discovered for the graph descriptors of alkanes and cycloalkanes, and thus chemists made important contributions to graph theory. However, apart from the mathematical beauty and interest of such theorems and relationships, the main chemical application of topological indices remains that of structural descriptors in structure−property and structure−biological activity models. Such studies require the computation of topological indices for molecular graphs containing heteroatoms and multiple bonds. Usually, an organic compound containing heteroatoms or multiple bonds can be represented as a vertex- and edge-weighted molecular graph. The absence of a rigorous mathematical theory of weighted graphs is the second cause that prevented the computation of a large number of molecular graph descriptors for organic compounds containing heteroatoms and multiple bonds.

Early applications of weighted molecular graphs are connected with the computation of polynomials and spectra of heteroconjugated compounds.[15−19] Employing graphs weighted with Hückel parameters, such methods can be used to compute Hückel molecular orbitals but the results are more general and can be applied to any weighted molecular graph. Several particular methods of computing TIs from molecular graphs containing heteroatoms or multiple bonds were proposed for the calculation of specific structural descriptors, giving valuable indices for quantitative structure−property relationship (QSPR) and quantitative structure−activity relationship (QSAR) models. Kier and Hall[1,2] used the valence atomic connectivity δ^v to define the connectivity indices ${}^m\chi_t^v$, the most successful class of topological indices, used in several hundreds of QSPR and QSAR studies. The same atomic invariant was employed in the formula of the electrotopological-state indices,[20−22] a group of descriptors that found important applications in establishing structure−property models. Basak applied information theory to compute the information content of the partitioning of atoms in equivalence classes; this ingenious method allows the computation of information theory indices for any organic compound without the need for special parameters for heteroatoms and multiple bonds.[23−26] Balaban extended the *J* index[27] by proposing special atomic parameters based on the electronegativity or atomic radius.[28−32] The importance of these four classes of molecular graph descriptors is emphasized by their use in more than 1000 QSPR and QSAR studies and by their incorporation in commercial molecular modeling software. Other methods for computing TIs from heteroatom-containing molecular graphs were proposed in the literature,[33−45] but their use is restricted to specific structural descriptors that did not find a wide spread use in molecular modeling.

Another important direction of research is represented by the development of general sets of atom and bond parameters

* Corresponding authors. E-mail: o_ivanciuc@chim.upb.ro; cabrol@unice.fr; and atbalaban@hotmail.com, respectively.
† University "Politehnica" of Bucharest.
‡ University of Nice−Sophia Antipolis.

10.1021/ci990129w CCC: $19.00
Published on Web 03/11/2000

that can be applied for the computation of all structural descriptors derived from vertex and edge contributions and molecular matrixes.[46−55] In the present paper we report a comparative study of three general weighting schemes, namely, the atomic number *Z*,[46] the relative electronegativity *X*, and the relative covalent radius *Y*[54] weighting schemes, in quantitative structure−retention relationship (QSRR) models for alkylphenols in gas−liquid chromatography. We have to mention that the *X* and *Y* weighting schemes are derived from the electronegativity and covalent radius parameters used in the computation of the Balaban index *J*.[28] The QSRR models are generated with structural descriptors computed with several newly introduced graph operators, namely, the Wiener, hyper-Wiener, minimum eigenvalue, maximum eigenvalue, Ivanciuc−Balaban, and information on distance operators.

MOLECULAR MATRICES AND STRUCTURAL DESCRIPTORS

The molecular graph operators were recently introduced as an extension of topological indices; a graph operator uses a mathematical equation to compute a family of related molecular graph descriptors with different molecular matrixes and various sets of parameters for atoms and bonds.[56] The use of molecular graph operators introduces a systematization of topological indices by putting together all descriptors computed with the same mathematical formula or algorithm. As a consequence, when new molecular matrixes are introduced, there is no need to invent new names and symbols for the topological indices derived from them; the notation of graph operators is simple and general and can accommodate new matrixes, weighting schemes, and any parameter used in the definition of a family of topological indices. In this section we present the weighting schemes and molecular graph operators that we use to compute the structural descriptors. Because the graph operators are newly introduced, we present several examples for the computation of the structural descriptors used in this study.

Weighting Schemes. Using graph theory, an organic compound containing heteroatoms or multiple bonds can be represented as a vertex- and edge-weighted molecular graph. A vertex- and edge-weighted (VEW) molecular graph $G = G(V,E,\mathrm{Sy},\mathrm{Bo},\mathrm{Vw},\mathrm{Ew},w)$ consists of a vertex set $V = V(G)$, an edge set $E = E(G)$, a set of chemical symbols of the vertexes $\mathrm{Sy} = \mathrm{Sy}(G)$, a set of topological bond orders of the edges $\mathrm{Bo} = \mathrm{Bo}(G)$, a vertex weight set $\mathrm{Vw}(w) = \mathrm{Vw}(w,G)$, and an edge weight set $\mathrm{Ew}(w) = \mathrm{Ew}(G)$. The elements of the vertex and edge weight sets are computed with the weighting scheme *w*. Usually, hydrogen atoms are not considered in the molecular graph, and in a VEW graph the weight of a vertex corresponding to a carbon atom is 0, while the weight of an edge corresponding to a carbon−carbon single bond is 1. Thus, the topological bond order Bo_{ij} of an edge e_{ij} takes the value 1 for single bonds, 2 for double bonds, 3 for triple bonds, and 1.5 for aromatic bonds.

A general approach of computing parameters for VEW graphs was developed by Trinajstić and co-workers by weighting the contributions of atoms and bonds with parameters based on the atomic number *Z* and the topological bond order;[46] a large variety of structural descriptors were computed with this method. In the atomic number weighting scheme *Z* the vertex parameter $\mathrm{Vw}(Z)_i$ of the vertex υ_i (representing atom *i* from a molecule) is defined as

$$\mathrm{Vw}(Z)_i = 1 - Z_C/Z_i = 1 - 6/Z_i \qquad (1)$$

where Z_i is the atomic number *Z* of the atom *i* and $Z_C = 6$ is the atomic number *Z* of carbon. The edge parameter $\mathrm{Ew}(Z)_{ij}$ that characterizes the bond between atoms *i* and *j* (represented in the molecular graph by the edge e_{ij} between vertexes υ_i and υ_j) is defined with the following equation:

$$\mathrm{Ew}(Z)_{ij} = Z_C Z_C/\mathrm{Bo}_{ij} Z_i Z_j = 6 \times 6/\mathrm{Bo}_{ij} Z_i Z_j \qquad (2)$$

where Bo_{ij} is the topological bond order of the edge between vertexes υ_i and υ_j.

In the relative electronegativity *X* weighting scheme the vertex parameter $\mathrm{Vw}(X)_i$ of the vertex υ_i is defined with the equation[54]

$$\mathrm{Vw}(X)_i = 1 - 1/X_i \qquad (3)$$

The edge parameter $\mathrm{Ew}(X)_{ij}$ that characterizes the bond between atoms *i* and *j* (represented in the molecular graph by the edge e_{ij} between vertexes υ_i and υ_j) is computed with the equation

$$\mathrm{Ew}(X)_{ij} = 1/\mathrm{Bo}_{ij} X_i X_j \qquad (4)$$

Values of the relative electronegativities *X* for some atoms are presented in Table 1, column 3.

In the relative covalent radius *Y* weighting scheme the parameter of the vertex υ_i, $\mathrm{Vw}(Y)_i$, is computed with the formula[54]

$$\mathrm{Vw}(Y)_i = 1 - 1/Y_i \qquad (5)$$

The edge parameter $\mathrm{Ew}(Y)_{ij}$ that characterizes the bond between atoms *i* and *j* is computed with the equation

$$\mathrm{Ew}(Y)_{ij} = 1/\mathrm{Bo}_{ij} Y_i Y_j \qquad (6)$$

Computed relative covalent radii *Y* for various elements are presented in Table 1, column 4.

A selected set of atomic properties used in the weighting schemes *X*, *Y*, and *Z* are presented in Table 1. These values can be used to compute the vertex and edge weights Vw and Ew for a large number of organic compounds. The atom and bond parameters computed with the relative electro-

Table 1. Selected Set of Atomic Properties Used with Different Weighting Schemes: Atomic Number *Z*, Relative Electronegativity *X*, and the Relative Covalent Radius *Y*

element	*Z*	*X*	*Y*
B	5	0.851	1.038
C	6	1.000	1.000
N	7	1.149	0.963
O	8	1.297	0.925
F	9	1.446	0.887
Si	14	0.937	1.128
P	15	1.086	1.091
S	16	1.235	1.053
Cl	17	1.384	1.015
As	33	0.946	1.379
Se	34	1.095	1.341
Br	35	1.244	1.303
Te	52	0.954	1.629
I	53	1.103	1.591

negativity X and the relative covalent radius Y have a periodic variation versus the atomic number Z, which is more chemically oriented when compared with the parameters derived from the Z weighting scheme.

Molecular Matrixes. The large majority of the topological indices proposed in the literature were derived from the adjacency and the distance matrixes. Recently, several new molecular matrixes were defined and used to compute new structural descriptors.[13] In the present paper the graph descriptors are computed from two molecular matrixes, namely, the distance **D** and reciprocal distance **RD** matrixes.

The distance matrix $\mathbf{D}(w) = \mathbf{D}(w,G)$ of a vertex- and edge-weighted molecular graph G with N vertexes is the symmetric square $N \times N$ matrix with real elements defined with the formula

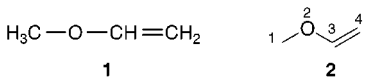
$$[\mathbf{D}(w)]_{ij} = \begin{cases} \mathrm{Vw}(w)_i & \text{if } i = j \\ d(w)_{ij} & \text{if } i \neq j \end{cases} \tag{7}$$

where $\mathrm{Vw}(w)_i$ is the weight of the vertex v_i, $d(w)_{ij}$ is the distance between vertexes v_i and v_j, and w is the weighting scheme used to compute the parameters Vw and Ew. The distance $d(w)_{ij}$ between a pair of vertexes v_i and v_j is equal to the length $l(p_{ij},w)$ of the shortest path p_{ij} connecting the two vertexes, where the length of the path p_{ij} is equal to the sum of the edge parameters $\mathrm{Ew}(w)_{ij}$ for all edges along the path.

The reciprocal distance matrix $\mathbf{RD}(w) = \mathbf{RD}(w,G)$ of a vertex- and edge-weighted molecular graph G with N vertexes is the square $N \times N$ symmetric matrix with real elements defined with the equation[57−60]

$$[\mathbf{RD}(w)]_{ij} = \begin{cases} [\mathbf{D}(w)]_{ij}^{-1} & \text{if } i \neq j \\ [\mathbf{D}(w)]_{ii} & \text{if } i = j \end{cases} \tag{8}$$

where $[\mathbf{D}(w)]_{ij}$ is the element corresponding to vertexes v_i and v_j from the distance matrix **D**, $[\mathbf{D}(w)]_{ii}$ is the diagonal element corresponding to vertex v_i, and w is the weighting scheme used to compute the parameters Vw and Ew. As an example of computing the reciprocal distance matrix consider methyl vinyl ether **1** and its corresponding molecular graph **2**.



Using the parameters from Table 1, one obtains the vertex- and edge-weighted distance matrix of graph **2** computed with the relative electronegativity weighting scheme X:

D(X,**2**)

	1	2	3	4
0	0.000	0.771	1.542	2.042
2	0.771	0.229	0.771	1.271
3	1.542	0.771	0.000	0.500
4	2.042	1.271	0.500	0.000

From the above distance matrix and eq 8 one obtains the reciprocal distance matrix of methyl vinyl ether modeled by graph **2**, **RD**(X,**2**):

RD(X,**2**)

	1	2	3	4
0	0.000	1.297	0.648	0.490
2	1.297	0.229	1.297	0.787
3	0.648	1.297	0.000	2.000
4	0.490	0.787	2.000	0.000

Vertex Sum Operator. Consider the vertex v_i from a graph G with N vertexes and a symmetric graph matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The vertex sum of vertex v_i, $\mathbf{VS}(\mathbf{M},w)_i = \mathbf{VS}(\mathbf{M},w,G)_i$, is defined as the sum of the elements in the column i, or row i, of the molecular matrix **M**:[12,55,56]

$$\mathbf{VS}(\mathbf{M},w,G)_i = \sum_{j=1}^{N} [\mathbf{M}(w)]_{ij} = \sum_{j=1}^{N} [\mathbf{M}(w)]_{ji} \tag{9}$$

For all vertexes in graph G, this vector is a local (atomic) invariant that is used to define several molecular graph descriptors. The vertex sums of the matrixes **D**(X,**2**) and **RD**(X,**2**) are

$$\mathbf{VS}(\mathbf{D},X,\mathbf{2}) = \{4.355,\ 3.042,\ 2.813,\ 3.813\}$$

$$\mathbf{VS}(\mathbf{RD},X,\mathbf{2}) = \{2.435,\ 3.610,\ 3.946,\ 3.276\}$$

Wiener Operator and Indices. Consider the VEW molecular graph G with N vertexes and its symmetric molecular matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The Wiener operator $\mathbf{Wi}(\mathbf{M},w) = \mathbf{Wi}(\mathbf{M},w,G)$ is[12,55,56]

$$\mathbf{Wi}(\mathbf{M},w,G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{M}(w)]_{ij} \tag{10}$$

The Wiener indices computed from the matrixes **D**(X,**2**) and **RD**(X,**2**) are $\mathbf{Wi}(\mathbf{D},X,\mathbf{2}) = 7.126$ and $\mathbf{Wi}(\mathbf{RD},X,\mathbf{2}) = 6.748$.

Hyper-Wiener Operator and Indices. Consider the vertex- and edge-weighted graph G with N vertexes and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The hyper-Wiener operator $\mathbf{HyWi}(\mathbf{M},w) = \mathbf{HyWi}(\mathbf{M},w,G)$ of the VEW graph G is:[12,55,56]

$$\mathbf{HyWi}(\mathbf{M},w,G) = {}^1/_2 \sum_{i=1}^{N} \sum_{j=i}^{N} ([\mathbf{M}(w)]_{ij}^2 + [\mathbf{M}(w)]_{ij}) \tag{11}$$

Topological indices computed with the hyper-Wiener operator were used to develop structure−property models for the boiling points of ethers, peroxides, acetals, and their sulfur analogues.[54] The hyper-Wiener indices obtained from the matrixes **D**(X,**2**) and **RD**(X,**2**) are $\mathbf{HyWi}(\mathbf{D},X,\mathbf{2}) = 8.390$ and $\mathbf{HyWi}(\mathbf{RD},X,\mathbf{2}) = 7.722$.

Matrix Spectrum Operator and Indices. Consider the VEW graph G with N vertexes and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w,G)$ computed with the weighting scheme w. The matrix spectrum operator $\mathbf{Sp}(\mathbf{M},w,G) = \{x_i,\ i = 1, 2, \ldots, N\}$ represents the eigenvalues of the matrix $\mathbf{M}(w)$ or the roots of the polynomial $\mathbf{Ch}(\mathbf{M},w,G,x)$, $\mathbf{Ch}(\mathbf{M},w,G,x) = 0$.[53] The $\mathbf{MinSp}(\mathbf{M},w,G)$ and $\mathbf{MaxSp}(\mathbf{M},w,G)$ operators are equal to the minimum and maximum values of $\mathbf{Sp}(\mathbf{M},w,G)$, respectively:[12,55,56]

$$\mathbf{MinSp}(\mathbf{M},w,G) = \mathbf{min}\{\mathbf{Sp}(\mathbf{M},w,G)\} \quad (12)$$

$$\mathbf{MaxSp}(\mathbf{M},w,G) = \mathbf{max}\{\mathbf{Sp}(\mathbf{M},w,G)\} \quad (13)$$

Molecular graph descriptors computed with these two spectral operators were used with success in QSPR studies to estimate the boiling points of acyclic compounds containing oxygen or sulfur atoms[54] and to model the boiling points, heat of vaporization, molar refraction, molar volume, critical pressure, critical temperature, and surface tension of alkanes.[60,61] The spectra of the distance and reciprocal distance matrixes of methyl vinyl ether computed with the weighting scheme X are

$$\mathbf{Sp}(\mathbf{D},X,\mathbf{2}) = \{3.591, -0.313, -0.627, -2.422\}$$

$$\mathbf{Sp}(\mathbf{RD},X,\mathbf{2}) = \{3.395, 0.128, -1.214, -2.080\}$$

One must note that the BCUT descriptors,[62−65] widely used as a molecular diversity measure and for the virtual screening of combinatorial libraries, are computed with the same formula as the **MinSp** and **MaxSp** operators, using molecular matrixes weighted with a different set of parameters.

Ivanciuc−Balaban Operator and Indices. The Ivanciuc−Balaban operator[54] of a VEW graph G, $\mathbf{IB}(\mathbf{M},w) = \mathbf{IB}(\mathbf{M},w,G)$, is computed with a formula modeled after Balaban's index J:[27]

$$\mathbf{IB}(\mathbf{M},w,G) = \frac{M}{\mu + 1}\sum_{E(G)}[\mathbf{VS}(\mathbf{M},w)_i\,\mathbf{VS}(\mathbf{M},w)_j]^{-1/2} \quad (14)$$

where $\mathbf{VS}(\mathbf{M},w)_i$ and $\mathbf{VS}(\mathbf{M},w)_j$ denote the vertex sums of the two adjacent vertexes v_i and v_j that are incident with an edge e_{ij} in the molecular graph G, M is the number of edges in the molecular graph, μ is the cyclomatic number (the number of cycles in the graph, $\mu = M - N + 1$, where N is the number of atoms of the molecular graph), w is the weighting scheme, and the summation goes over all edges from the edge set $E(G)$. The molecular graph **2** is acyclic and has four vertexes and three edges, giving the factor $M/(\mu + 1) = 3/(0 + 1) = 3$; using the vertex sums $\mathbf{VS}(\mathbf{D},X,\mathbf{2})$ and $\mathbf{VS}(\mathbf{RD},X,\mathbf{2})$, eq 14 gives the corresponding Ivanciuc−Balaban indices:

$$\mathbf{IB}(\mathbf{D},X,\mathbf{2}) = 3[(4.355 \times 3.042)^{-1/2} + (3.042 \times 2.813)^{-1/2} + (2.813 \times 3.813)^{-1/2}] = 2.766$$

$$\mathbf{IB}(\mathbf{RD},X,\mathbf{2}) = 3[(2.435 \times 3.610)^{-1/2} + (3.610 \times 3.946)^{-1/2} + (3.946 \times 3.276)^{-1/2}] = 2.641$$

Information on Distance Operators. The indices U, V, X, and Y using information on distances were defined for the distance matrix of simple graphs representing alkanes.[66−68] The extension of such indices to vertex- and edge-weighted graphs considers the possibility that the molecular matrix of a VEW graph may contain negative elements or elements with values between 0 and 1. The graph vertex operators $\mathbf{VUinf}(\mathbf{M},w,G)$, $\mathbf{VVinf}(\mathbf{M},w,G)$, $\mathbf{VXinf}(\mathbf{M},w,G)$, and $\mathbf{VYinf}(\mathbf{M},w,G)$ apply the information theory equations to the absolute values of the elements of the molecular matrix $\mathbf{M}(w,G)$.[69,70] All three weighting schemes have atomic weights Vw with negative values for certain elements: Vw(Z,B) = −0.200, Vw(X,B) = −0.175, Vw(X,Si) = −0.067, Vw(X,As) = −0.057, Vw(X,Te) = −0.048, Vw(Y,N) = −0.038, Vw(Y,O) = −0.081, and Vw(Y,F) = −0.127. Also, it is possible that with certain weighting schemes the edge parameters Ew have negative values. Because the logarithm is defined only for positive arguments, the four graph vertex operators are computed from the elements of a positive matrix $\mathbf{P}(w) = \mathbf{P}(w,G)$ whose element $[\mathbf{P}(w)]_{ij}$ is equal to the absolute value of the corresponding element from the $\mathbf{M}(w)$ matrix, element $[\mathbf{P}(w)]_{ij} = |[\mathbf{M}(w)]_{ij}|$. The graph vertex operators are defined with the following equations:[69]

$$\mathbf{VUinf}(\mathbf{M},w,G)_i = -\sum_{j=1}^{N}\frac{[\mathbf{P}(w)]_{ij}}{\mathbf{VS}(\mathbf{P},w)_i}\log_2\left[\frac{[\mathbf{P}(w)]_{ij}}{\mathbf{VS}(\mathbf{P},w)_i}\right] \quad (15)$$

$$\mathbf{VVinf}(\mathbf{M},w,G)_i = \mathbf{VS}(\mathbf{P},w)_i \log_2 \mathbf{VS}(\mathbf{P},w)_i - \mathbf{VUinf}(\mathbf{M},w)_i \quad (16)$$

$$\mathbf{VXinf}(\mathbf{M},w,G)_i = \mathbf{VS}(\mathbf{P},w)_i \log_2 \mathbf{VS}(\mathbf{P},w)_i - \mathbf{VYinf}(\mathbf{M},w)_i \quad (17)$$

$$\mathbf{VYinf}(\mathbf{M},w,G)_i = \sum_{j=1}^{N}[\mathbf{P}(w)]_{ij}\log_2[\mathbf{P}(w)]_{ij} \quad (18)$$

where $\mathbf{VS}(\mathbf{P},w)_i$ is the vertex sum of the vertex v_i computed from the matrix $\mathbf{P}$, w is the weighting scheme, and the summations in formulas (15) and (18) are done for the absolute values of the nonzero elements of the molecular matrix $\mathbf{P}$, $[\mathbf{P}(w)]_{ij} \neq 0$. For the notation of the four graph vertex operators $\mathbf{VUinf}(\mathbf{M},w,G)$, $\mathbf{VVinf}(\mathbf{M},w,G)$, $\mathbf{VXinf}(\mathbf{M},w,G)$, and $\mathbf{VYinf}(\mathbf{M},w,G)$, we have maintained the molecular matrix $\mathbf{M}$ to indicate the source of the invariants.

Because certain matrix elements $[\mathbf{M}]_{ij}$ may have values lower than 1, their logarithm gives negative values. In such conditions, some terms of the four graph vertex operators may have negative values, and the Randić-like formula used in the case of the topological indices U, V, X, and Y cannot be used. The bond contribution of the information on distances invariants is computed with the following equation:

$$f(x,y) = \begin{cases}(xy)^{-1/2} & \text{if } xy > 0\\ -(|xy|)^{-1/2} & \text{if } xy < 0\end{cases} \quad (19)$$

The information on matrix elements operators $\mathbf{U}(\mathbf{M},w)$, $\mathbf{V}(\mathbf{M},w)$, $\mathbf{X}(\mathbf{M},w)$, and $\mathbf{Y}(\mathbf{M},w)$ is computed with the equations

$$\mathbf{U}(\mathbf{M},w,G) = \frac{M}{\mu + 1}\sum_{E(G)} f(\mathbf{VUinf}(\mathbf{M},w)_i, \mathbf{VUinf}(\mathbf{M},w)_j) \quad (20)$$

$$\mathbf{V}(\mathbf{M},w,G) = \frac{M}{\mu + 1}\sum_{E(G)} f(\mathbf{VVinf}(\mathbf{M},w)_i, \mathbf{VVinf}(\mathbf{M},w)_j) \quad (21)$$

$$\mathbf{X}(\mathbf{M},w,G) = \frac{M}{\mu + 1}\sum_{E(G)} f(\mathbf{VXinf}(\mathbf{M},w)_i, \mathbf{VXinf}(\mathbf{M},w)_j) \quad (22)$$

$$\mathbf{Y}(\mathbf{M},w,G) = \frac{M}{\mu + 1}\sum_{E(G)} f(\mathbf{VYinf}(\mathbf{M},w)_i, \mathbf{VYinf}(\mathbf{M},w)_j) \quad (23)$$

The computation of the information on distances operators **U**(**M**,*w*), **V**(**M**,*w*), **X**(**M**,*w*), and **Y**(**M**,*w*) is presented for methyl vinyl ether **2**. From the distance matrix **D**(*X*,**2**) one obtains the four vertex vectors **VUinf**(**D**,*X*,**2**), **VVinf**(**D**,*X*,**2**), **VXinf**(**D**,*X*,**2**), and **VYinf**(**D**,*X*,**2**):

$$\mathbf{VUinf}(\mathbf{D},X,\mathbf{2}) = \{1.485, 1.811, 1.430, 1.395\}$$

$$\mathbf{VVinf}(\mathbf{D},X,\mathbf{2}) = \{7.759, 3.072, 2.767, 5.968\}$$

$$\mathbf{VXinf}(\mathbf{D},X,\mathbf{2}) = \{6.467, 5.508, 4.023, 5.320\}$$

$$\mathbf{VYinf}(\mathbf{D},X,\mathbf{2}) = \{2.777, -0.626, 0.174, 2.043\}$$

The **VYinf**(**D**,*X*,**2**) for vertex 2 in the molecular graph **2** is negative, and in this case the modified Randić formula (19) must be used to compute the corresponding **Y**(**D**,*X*,**2**) index. Using the formulas (20−23) one computes the four information theory indices: **U**(**D**,*X*,**2**) = 5.818, **V**(**D**,*X*,**2**) = 2.382, **X(D**,*X*,**2**) = 1.788, and **Y(D**,*X*,**2**) = −6.333.

The second example presents the computation of the descriptors **U**, **V**, **X**, and **Y** derived from the reciprocal distance matrix of methyl vinyl ether **2**. The elements of the reciprocal distance matrix **RD**(*X*,**2**) are used to compute the four vertex vectors **VUinf**(**RD**,*X*,**2**), **VVinf**(**RD**,*X*,**2**), **VXinf**(**RD**,*X*,**2**), and **VYinf**(**RD**,*X*,**2**):

$$\mathbf{VUinf}(\mathbf{RD},X,2) = \{1.458, 1.793, 1.453, 1.339\}$$

$$\mathbf{VVinf}(\mathbf{RD},X,2) = \{1.669, 4.892, 6.360, 4.271\}$$

$$\mathbf{VXinf}(\mathbf{RD},X,2) = \{3.550, 6.471, 5.732, 4.386\}$$

$$\mathbf{VYinf}(\mathbf{RD},X,2) = \{-0.423, 0.214, 2.081, 1.223\}$$

For the vertex 1 in the molecular graph **2** the atomic invariant **VYinf**(**RD**,*X*,2) is negative, and again the modified Randić formula (19) must be used to compute the **Y**(**RD**,*X*,**2**) index. The four information theory indices derived from the reciprocal distance matrix are **U**(**RD**,*X*,**2**) = 5.866, **V**(**RD**,*X*,**2**) = 2.163, **X**(**RD**,*X*,**2**) = 1.717, and **Y**(**RD**,*X*,**2**) = −3.596.

METHOD

Data. The identification of organic compounds from a mixture can be made with the method of chromatographic peak comparison with a standard sample of the each compound. Because samples of pure compounds are not always available, it is important to develop QSRR models that can efficiently predict retention parameters by using theoretical descriptors computed from the chemical structure. Chromatographic retention is a physical phenomenon that is primarily dependent on the interactions between the solute and the stationary phase. With the aid of QSRR the interactions associated with this phenomenon for any given stationary phase can be related to the constitutional, molecular graph (topological), geometric, electrostatic, and quantum descriptors of the molecules.[71−75] In the present study we will develop QSRR models for alkylphenols in gas−liquid chromatography with the aid of structural descriptors computed from the molecular graph. The structure of the alkylphenols and their experimental Kováts retention indices (RIs)[76] are presented in Table 2. The retention indices were determined on a column packed with 5% hexaphenyl ether on Chromatone N AW HMDS (0.16−0.20 mm) at 160 °C.

Structural Descriptors. The list of the 60 structural descriptors used in the QSPR study is presented as follows:

Table 2. Structure and Retention Indices (RI) of Alkylphenols and Calibration and Prediction Residuals Computed with the QSRR Model Represented by Equation 26 from Table 3

	RI		
compd	expt	res_{calc}[a]	res_{pr}[b]
phenol	1281	19	24
2-methylphenol	1354	6	7
3-methylphenol	1386	−8	−9
4-methylphenol	1385	−16	−18
2-ethylphenol	1430	14	15
3-ethylphenol	1483	15	16
4-ethylphenol	1473	−4	−4
2,3-dimethylphenol	1495	0	0
2,4-dimethylphenol	1456	−6	−6
2,5-dimethylphenol	1453	−10	−11
2,6-dimethylphenol	1416	9	10
3,5-dimethylphenol	1489	−12	−14
3,4-dimethylphenol	1530	2	2
4-isopropylphenol	1527	−32	−34
2-*n*-propylphenol	1502	12	13
3-*n*-propylphenol	1565	22	23
4-*n*-propylphenol	1563	10	10
2-ethyl-4-methylphenol	1523	−10	−10
2-ethyl-5-methylphenol	1529	−7	−7
2-ethyl-6-methylphenol	1485	−15	−17
3-ethyl-5-methylphenol	1581	3	3
4-ethyl-2-methylphenol	1539	−1	−1
4-ethyl-3-methylphenol	1608	10	10
2,3,4-trimethylphenol	1638	15	16
2,3,5-trimethylphenol	1593	8	9
2,3,6-trimethylphenol	1551	−7	−8
2,4,5-trimethylphenol	1593	2	3
3,4,5-trimethylphenol	1667	−4	−5
4-*sec*-butylphenol	1612	−44	−54
2-*n*-butylphenol	1600	15	16
3-*n*-butylphenol	1668	29	31
4-*n*-butylphenol	1661	11	12
2-methyl-4-*n*-propylphenol	1623	3	3
2-methyl-6-*n*-propylphenol	1553	−25	−28
3-methyl-6-*n*-propylphenol	1602	−13	−14
4-methyl-2-*n*-propylphenol	1593	−17	−18
2,4-diethylphenol	1602	−12	−13
2,5-diethylphenol	1624	5	5
3,4-diethylphenol	1682	−9	−9
2,3,4,5-tetramethylphenol	1782	33	41
2,3,4,6-tetramethylphenol	1690	9	11
2,3,5,6-tetramethylphenol	1683	2	2
2-ethyl-4,5-dimethylphenol	1656	−10	−10
2-*n*-pentylphenol	1700	11	13
4-*n*-pentylphenol	1765	10	11
4-*tert*-pentylphenol	1703	−3	−7
2-ethyl-5-*n*-propylphenol	1706	1	1
2-*n*-hexylphenol	1800	−3	−4
4-*n*-hexylphenol	1871	0	0
3-*n*-butyl-6-ethylphenol	1807	−5	−5

[a] The calibration residual computed with eq 26, $res_{calc} = RI_{expt} - RI_{calc}$. [b] The leave-one-out prediction residual of the MLR model represented by eq 26, $res_{pr} = RI_{expt} - RI_{pr}$.

(1) the molecular weight, **MW**; (2) the Kier and Hall's valence connectivity indices ${}^{0}\chi^{v}$, ${}^{1}\chi^{v}$, ${}^{2}\chi^{v}$, ${}^{3}\chi_{p}{}^{v}$, ${}^{3}\chi_{c}{}^{v}$;[1,2] (3) Wiener indices computed with the Wiener operator **Wi**(**M**,*w*); (4) hyper-Wiener indices computed with the hyper-Wiener operator **HyWi**(**M**,*w*); (5) the spectral operators **MinSp**(**M**,*w*) and **MaxSp**(**M**,*w*); (6) the Ivanciuc−Balaban operator **IB**(**M**,*w*); (5) the information-theory operators **U**(**M**,*w*), **V**(**M**,*w*), **X**(**M**,*w*), and **Y**(**M**,*w*).

QSPR Model. All studies that develop QSPR models from a large set of computed descriptors use a wide range of algorithms for selecting significant descriptors. Because the exhaustive test of all multilinear regression (MLR) equations

requires too large computational resources, we have used a heuristic method for descriptor selection. This heuristic algorithm starts from the set of 60 structural descriptors and develops QSPR models by applying the following steps:

(1) biparametric regression equations are computed with all possible pairs of descriptors. The most significant 200 pairs of molecular descriptors were used in the next step.

(2) To an MLR model containing n descriptors, a new descriptor is added to generate a model with $n + 1$ descriptors.

(3) The most significant 200 MLR models containing $n + 1$ descriptors are selected.

Steps 2 and 3 are repeated until MLR models with a certain maximum number of descriptors are obtained.

RESULTS AND DISCUSSION

In a previous QSRR study of this set of 50 phenols the following biparametric model was obtained:[77]

$$\mathrm{RI} = 1132.209 + 169.274\,{}^{3}\chi_{\mathrm{p}}{}^{v} + 0.9137Sz(P)^{24}$$
$$n = 50 \qquad r = 0.9528 \qquad s = 38 \qquad F = 232 \quad (24)$$

where ${}^{3}\chi_{\mathrm{p}}{}^{v}$ is a Kier and Hall connectivity index[1,2] and $Sz(P)$ is the Szeged index computed with the weighting scheme P,[55] computed from atomic polarizability values.[78] A slight improvement of this QSRR model was obtained with the use of the **X** information index computed from the three-dimensional molecular matrix **3D**:[70]

$$\mathrm{RI} = 1662.97 + 233.946\,{}^{3}\chi_{\mathrm{p}}{}^{v} - 767.255\mathbf{X}(\mathbf{3D})$$
$$n = 50 \qquad r = 0.9566 \qquad s = 37 \qquad F = 253 \quad (25)$$

With the 60 structural descriptors employed in this study we have obtained several QSRR models that show a significant improvement over these two equations. Several tests with QSRR models containing from two to five structural descriptors indicated that the highest Fisher test is obtained for the equations with four topological indices. In Table 3 we present the coefficients, confidence interval, structural descriptors, and statistical indices for the best ten QSRR models, eqs 26−35, with four independent variables that model the phenol retention indices.

It is well-known that correlational analysis develops structure−property models by proposing statistical relationships between molecular descriptors and a physical, chemical, or biological property; we have to point out here that a good statistical model does not imply a causal relationship between molecular descriptors and the investigated property. Correlations can be observed not only because a causal relationship exists between a set of descriptors and a property but also due to statistical bias resulting from experimental errors in measuring the property, or even due to chance alone. When a correlation appears due to errors or chance factors, the resulting QSPR model has no scientific content and offers misleading conclusions. Topliss et al. demonstrated that such a situation can easily occur if large numbers of structural descriptors are screened for potential inclusion into the final correlation equation.[79,80] Several model validation techniques were developed with the aim of distinguishing between true and random correlations and of estimating the predictive power of a QSPR model.[81] For the QSRR models represented

Table 3. Coefficients, Confidence Interval, Structural Descriptors $\mathbf{SD}_i$ ($i = 1-4$), Model Calibration Statistical Indices (r_{cal}, Correlation Coefficient; s_{cal}, Standard Deviation; F_{cal}, Fisher Test), and Statistical Indices for the Leave-One-Out Cross-Validation Prediction (r_{pr}, Correlation Coefficient; s_{pr}, Standard Deviation) for the Best 10 MLR Equations with Four Independent Variables That Model the Retention Indices of the 50 Alkylphenols from Table 2[a]

eq	a_0	a_1	$\mathbf{SD}_1$	a_2	$\mathbf{SD}_2$	a_3	$\mathbf{SD}_3$	a_4	$\mathbf{SD}_4$	r_{cal}	s_{cal}	F_{cal}	r_{pr}	s_{pr}
26	−2684.5 ± 353.5	63.573 ± 8.348	${}^{3}\chi_{\mathrm{p}}{}^{v}$	12.695 ± 1.667	**MaxSp(D,*Z*)**	−21304.1 ± 2797.6	**MinSp(RD,*Y*)**	19186.9 ± 2519.5	**MinSp(RD,*Z*)**	0.9931	15	811	0.9911	17
27	−2689.2 ± 353.2	63.535 ± 8.334	${}^{3}\chi_{\mathrm{p}}{}^{v}$	12.696 ± 1.667	**MaxSp(D,*X*)**	−21298.3 ± 2797.1	**MinSp(RD,*Y*)**	19179.1 ± 2518.8	**MinSp(RD,*Z*)**	0.9931	15	811	0.9911	17
28	−2757.7 ± 362.6	62.834 ± 8.261	${}^{3}\chi_{\mathrm{p}}{}^{v}$	12.718 ± 1.672	**MaxSp(D,*Y*)**	−21224.1 ± 2790.4	**MinSp(RD,*Y*)**	19075.5 ± 2507.9	**MinSp(RD,*Z*)**	0.9931	15	809	0.9911	17
29	−6058.8 ± 856.8	15.633 ± 2.211	**Y(D,*Z*)**	225.76 ± 31.93	**X(RD,*Z*)**	−25312.9 ± 3579.6	**MinSp(RD,*Y*)**	21658.6 ± 3062.9	**MinSp(RD,*Z*)**	0.9921	16	699	0.9901	17
30	−6040.8 ± 854.3	15.919 ± 2.251	**Y(D,*X*)**	225.96 ± 31.96	**X(RD,*Z*)**	−25302.6 ± 3578.5	**MinSp(RD,*Y*)**	21657.8 ± 3063.0	**MinSp(RD,*Z*)**	0.9921	16	699	0.9901	17
31	−6068.8 ± 858.7	15.659 ± 2.216	**Y(D,*Z*)**	225.85 ± 31.96	**X(RD,*X*)**	−25316.7 ± 3582.2	**MinSp(RD,*Y***	21658.3 ± 3064.6	**MinSp(RD,*Z*)**	0.9920	16	699	0.9901	17
32	−6060.8 ± 856.2	15.944 ± 2.256	**Y(D,*X*)**	226.05 ± 31.99	**X(RD,*X*)**	−25306.4 ± 3581.0	**MinSp(RD,*Y*)**	21657.5 ± 3064.7	**MinSp(RD,*Z*)**	0.9920	16	699	0.9901	17
33	−3764.5 ± 519.0	59.665 ± 8.225	${}^{3}\chi_{\mathrm{p}}{}^{v}$	164.20 ± 22.64	**X(RD,*Y*)**	−22069.8 ± 3042.5	**MinSp(RD,*Y*)**	19515.1 ± 2690.3	**MinSp(RD,*Z*)**	0.9924	16	736	0.9899	18
34	−3663.8 ± 508.2	59.275 ± 8.222	${}^{3}\chi_{\mathrm{p}}{}^{v}$	164.10 ± 22.76	**X(RD,*Z*)**	−21993.1 ± 3050.7	**MinSp(RD,*Y*)**	19481.4 ± 2702.3	**MinSp(RD,*Z*)**	0.9924	16	727	0.9898	18
35	−3668.3 ± 509.1	59.344 ± 8.236	${}^{3}\chi_{\mathrm{p}}{}^{v}$	164.09 ± 22.77	**X(RD,*X*)**	−21992.0 ±3052.2	**MinSp(RD,*Y*)**	19478.6 ± 2703.4	**MinSp(RD,*Z*)**	0.9923	16	726	0.9898	18

[a] The MLR equations have the general form $\mathrm{RI} = a_0 + a_1\mathbf{SD}_1 + a_2\mathbf{SD}_2 + a_3\mathbf{SD}_3 + a_4\mathbf{SD}_4$.

by the eqs 26−35 in Table 3 we have used the leave-one-out (LOO) cross-validation procedure; the statistical indices obtained for prediction (correlation coefficient r_{pr} and standard deviation s_{pr}) are reported in the last two columns of Table 3. The ten QSRR models from Table 3 are arranged in the ascending order of s_{pr}.

Compared with the previously obtained structure−retention models[70,77] for the set of alkylphenols from eqs 24 and 25, the QSRR models reported in this paper are significantly better, with a standard deviation 2.5 times smaller. Although eqs 26−35 have four descriptors each, the Fisher test F is several times higher, compared with the F indices from eqs 24 and 25 that have two descriptors each. The values of the statistical indices of eqs 26−35 show that the ten QSRR models, although obtained with different sets of structural descriptors, have close statistical quality. From the data set investigated in this paper, it is not possible to select only one QSRR model as the best one, because the differences in the statistical indices are small; this situation appears frequently whenever statistical models are developed from a large pool of structural descriptors. The analysis of the topological indices selected in eqs 26−35 offers the possibility of studing the distribution of weighting schemes, molecular matrixes, graph operators, and structural descriptors.

Using various atomic properties, the weighting schemes employed in this study offer the atom and bond parameters for the computation of the topological indices. The correlational ability of the structural descriptors depends heavily on the weighting scheme; those computed with atomic number Z were selected in the majority of cases: Z, 16 times; Y, 12 times; X, 6 times. The conclusion of these results is that all three weighting schemes are useful and provide structural invariants with a good correlational power, but further experiments are required to determine if this frequency order is a particular behavior or represents a more general trend.

An analysis of the presence of the molecular matrixes in the structural descriptors selected in eqs 26−35 shows that the reciprocal distance matrix **RD** was selected 27 times, while the distance matrix **D** was selected only 7 times. Until recently, the molecular graph descriptors were mainly calculated from the adjacency and distance matrixes. Our finding indicates that the structural descriptors computed from the recently introduced reciprocal distance matrix may be more suitable for developing relevant structure−property models.

From the large set of structural descriptors tested in the QSRR models, several were not selected in eqs 26−35: **Wi**(**M**,w), **HyWi**(**M**,w), **IB**(**M**,w), **U**(**M**,w), and **V**(**M**,w). These results indicate that in our particular QSRR model the above descriptors are not relevant; however, this finding does not rule out their utility in other QSPR or QSAR models. From the remaining graph operators, the structural descriptors computed with the spectral operator **MinSp** were selected with a higher frequency: **MinSp**(**RD**,w), 20 times; **X**(**RD**,w), 7 times; **Y**(**D**,w), 4 times; **MaxSp**(**D**,w), 3 times. This analysis shows again the importance of the reciprocal distance matrix, because the selected **MinSp** and **X** descriptors were computed only with the **RD** matrix.

Finally, we direct our attention to the frequency of individual structural descriptors. Two descriptors are found in each QSRR model from eqs 26−35 namely, **MinSp**(**RD**,Z), and **MinSp**(**RD**,Y). The **MinSp** operator is a measure of molecular shape and branching, relatively independent of the molecular size. The connectivity index ${}^3\chi_p{}^v$ appears in six QSRR models, and this is the single descriptor from its class selected in eqs 26−35. The index ${}^3\chi_p{}^v$ represents the weighted contribution of subgraphs, containing butane-like subgraphs and it is a measure of molecular size and shape.

Table 4. Intercorrelation Matrix of the Structural Descriptors in Equation 26 from Table 3 and Correlation Coefficient between Each Descriptor and the Experimental Retention Index

		1	2	3	4
${}^3\chi_p{}^v$	1	1.000	0.721	−0.486	−0.504
MaxSp(**D**,Z)	2	0.721	1.000	0.096	0.111
MinSp(**RD**,Z)	3	−0.486	0.096	1.000	0.980
MinSp(**RD**,Y)	4	−0.504	0.111	0.980	1.000
RI	5	0.907	0.857	−0.250	−0.303

An examination of the QSRR models from Table 3 reveals their high similarity in what concerns the structural descriptors involved. This situation indicates that the selection algorithm converged to a stable set of statistically good equations that have only small differences. Also, a comparison of the statistical indices from Table 3 shows that the statistical differences of the QSAR models in eqs 26−35 are not high. We will now examine in detail eq 26 because it gives the best leave-one-out prediction results. In Table 2 we give the residuals in calibration (res_{cal}) and prediction (res_{pr}) computed with eq 26. These results show that the predictions are as good as the calibration data, indicating that the model is stable and can be used for predicting the retention indices of new alkylphenols.

The strong intercorrelation between structural descriptors from a MLR equation may lead to misinterpretation of the corresponding structure−activity model. The algorithm used in this work does not test the intercorrelation of the structural descriptors selected in a MLR equation, and the final QSRR models can contain highly intercorrelated independent variables. In Table 4 we give the intercorrelation matrix of the four structural descriptors in eq 26 together with the individual correlation coefficients of the four descriptors with the retention indices of alkylphenols (i.e. in monoparametric correlations). From this matrix one can see that **MinSp**(**RD**,Z) and **MinSp**(**RD**,Y) have a very high intercorrelation coefficient, equal to 0.980. Considering this special situation and the fact that this QSRR model gives the best results, we will investigate in detail a method of avoiding this problem. Also, we will determine if both descriptors must be present in the QSRR model in order to obtain a high correlation or if one of them can be deleted without greatly influencing the correlational power of eq 26.

Several techniques can be applied to overcome the problem of highly intercorrelated descriptors: PCA, PLS, or Randić's sequential orthogonalization. We have selected the recently defined sequential orthogonalization[82] that was applied with success in numerous QSPR and QSAR studies.[83−95] In practice, the sequential orthogonalization of descriptors can be used to simplify QSPR and QSAR models that contain many intercorrelated descriptors, by removing the variables with a small contribution. Recently it was found that this statistical method can give better results than the neural

networks[94] or the heuristic algorithm implemented in CODESSA.[95]

In the sequential orthogonalization algorithm, a descriptor from the set of intercorrelated structural descriptors can be made orthogonal by removing the part of its information content that it shares with the other descriptors in the set. The order in which descriptors are orthogonalized is important, because it strongly affects the information content of the thus obtained orthogonal descriptors. We apply the orthogonalization algorithm to the four descriptors in eq 26 considered in the order from Table 3. The scope is to orthogonalize the set of structural descriptors to obtain the orthogonalized set of descriptors $\Omega({}^3\chi_p{}^v)$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$. For the set of four intercorrelated structural descriptors orthogonalized in the order ${}^3\chi_p{}^v$, **MaxSp**(**D**,*Z*), **MinSp**(**RD**,*Z*), and **MinSp**(**RD**,*Y*), the construction of the orthogonal descriptors follows the following steps:

(1) The first orthogonal descriptor $\Omega({}^3\chi_p{}^v)$ is identical with the original descriptor ${}^3\chi_p{}^v$:

$$\Omega({}^3\chi_p{}^v) = {}^3\chi_p{}^v \tag{36}$$

(2) The linear regression equation between the second descriptor **MaxSp**(**D**,*Z*) and orthogonal descriptor $\Omega({}^3\chi_p{}^v)$ is computed:

$$\mathbf{MaxSp}(\mathbf{D},Z) = a_{2,1} + b_{2,1}\Omega({}^3\chi_p{}^v) \tag{37}$$

The second orthogonal descriptor $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$ is the residual of the above equation, i.e., the difference between its real value **MaxSp**(**D**,*Z*) and that computed with eq 37:

$$\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) = \mathbf{MaxSp}(\mathbf{D},Z) - a_{2,1} - b_{2,1}\Omega({}^3\chi_p{}^v) \tag{38}$$

(3) The orthogonalization of the third descriptor begins with the computation of the linear regression equation between descriptor **MinSp**(**RD**,*Z*) and orthogonal descriptor $\Omega({}^3\chi_p{}^v)$:

$$\mathbf{MinSp}(\mathbf{RD},Z) = a_{3,1} + b_{3,1}\Omega({}^3\chi_p{}^v) \tag{39}$$

The residual of eq 39 gives $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))_1$, the part of **MinSp**(**RD**,*Z*) that is orthogonal to $\Omega({}^3\chi_p{}^v)$:

$$\Omega(\mathbf{MinSp}(\mathbf{RD},Z))_1 = \mathbf{MinSp}(\mathbf{RD},Z) - a_{3,1} - b_{3,1}\Omega({}^3\chi_p{}^v) \tag{40}$$

The vector $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))_1$ is then orthogonalized against $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$ by computing the linear regression equation between these two descriptors:

$$\Omega(\mathbf{MinSp}(\mathbf{RD},Z))_1 = a_{3,2} + b_{3,2}\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) \tag{41}$$

Finally, the third orthogonal descriptor $\Omega(X_3)$ is the residual of eq 41:

$$\Omega(\mathbf{MinSp}(\mathbf{RD},Z)) = \Omega(\mathbf{MinSp}(\mathbf{RD},Z))_1 - a_{3,2} - b_{3,2}\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) \tag{42}$$

(4) Using a similar procedure, the fourth descriptor is orthogonalized against the first three orthogonal descriptors, to give $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$.

The structure−property model from eq 26 is computed with the orthogonal descriptors $\Omega({}^3\chi_p{}^v)$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$:

$$\begin{aligned}\mathrm{RI} = {} & (1087.2 \pm 142.8) + (285.12 \pm 37.44)\Omega({}^3\chi_p{}^v) + \\ & (7.0617 \pm 0.9273)\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) + \\ & (36.517 \pm 4.795)\Omega(\mathbf{MinSp}(\mathbf{RD},Z)) - \\ & (21304.1 \pm 2797.6)\Omega(\mathbf{MinSp}(\mathbf{RD},Y)) \\ n = 50 \quad & r = 0.9931 \quad s = 15 \quad F = 811\end{aligned} \tag{43}$$

We have to point out that eqs 26 and 43 have identical statistical indices (correlation coefficient *r*, standard deviation *s*, and Fisher test *F*). The leave-one-out cross-validation statistical indices, $r_{pr} = 0.9911$ and $s_{pr} = 17$, are equal to those obtained with eq 26, which has nonorthogonal descriptors. We will now turn our attention to the possibility of generating from eq 26 a simpler QSRR model by deleting one of the highly intercorrelated descriptors, either **MinSp**(**RD**,*Z*) or **MinSp**(**RD**,*Y*). Of course, to be practical, such a reduction must not degrade the statistical indices. After deleting **MinSp**(**RD**,*Y*) we obtain the QSRR model:

$$\begin{aligned}\mathrm{RI} = {} & (1178.6 \pm 325.9) + (189.92 \pm 52.52){}^3\chi_p{}^v + \\ & (7.021 \pm 1.942)\mathbf{MaxSp}(\mathbf{D},Z) + \\ & (36.52 \pm 10.10)\mathbf{MinSp}(\mathbf{RD},Z) \\ n = 50 \quad & r = 0.9528 \quad s = 38 \quad F = 151\end{aligned} \tag{44}$$

It is clear that this equation that contains three structural descriptors, with statistical indices comparable with those obtained with the QSRR models with only two descriptors from eqs 24 and 25, is not an interesting alternative to eq 26. The second possibility is to delete **MinSp**(**RD**,*Z*), in which case we obtain a QSRR equation with much lower statistical indices than those from eq 26:

$$\begin{aligned}\mathrm{RI} = {} & -(3375.2 \pm 887.4) + (123.40 \pm 32.44){}^3\chi_p{}^v + \\ & (9.885 \pm 2.599)\mathbf{MaxSp}(\mathbf{D},Z) - \\ & (2234.3 \pm 587.5)\mathbf{MinSp}(\mathbf{RD},Y) \\ n = 50 \quad & r = 0.9571 \quad s = 37 \quad F = 167\end{aligned} \tag{45}$$

The results obtained with the QSRR models from eqs 44 and 45 clearly indicate that it is not possible to simplify eq 26 without losing a significant part of its modeling power. A possibility of obtaining a simpler QSRR model is suggested by the results of the orthogonalization of eq 26 The partial correlation coefficients of the four orthogonal descriptors, i.e., $\Omega({}^3\chi_p{}^v)$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$, with the retention indices of alkylphenols are 0.9065, 0.2936, 0.0015, and −0.2800. The correlation coefficient of the third orthogonal descriptor, $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$, with RI is very small, suggesting that the contribution of this descriptor to the overall correlation can be neglected. Indeed, if we delete this orthogonal descriptor from the QSRR model, we obtain a regression equation with

three independent variables that has the same statistical indices (r and s) as those of eqs 26 and 43:

$$\begin{aligned}\mathrm{RI} = (1087.2 \pm 111.2) + (285.12 \pm 29.16)\Omega({}^{3}\chi_{\mathrm{p}}{}^{\upsilon}) + \\ (7.0617 \pm 0.7222)\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) - \\ (21304.1 \pm 2178.6)\Omega(\mathbf{MinSp}(\mathbf{RD},Y)) \\ n = 50 \quad r = 0.9931 \quad s = 15 \quad F = 1105 \end{aligned} \tag{46}$$

Obviously, the F test is larger because eq 46 has only three descriptors, but the coefficients are identical with those from eq 43. The leave-one-out cross-validation statistical indices, $r_{\mathrm{pr}} = 0.9911$ and $s_{\mathrm{pr}} = 17$, are equal to those obtained with eqs 26 and 43, which have four structural descriptors. We have to mention that in eq 46 the orthogonal descriptor $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$ is orthogonalized against three descriptors, namely, $\Omega({}^{3}\chi_{\mathrm{p}}{}^{\upsilon})$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$. In this way, although the orthogonal descriptor $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$ is not explicitly present in eq 46, its values were used to compute $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$. The generation of the QSRR model from eq 46 was possible only because the sequential orthogonalization of descriptors suggested simpler structure−property models by removing the orthogonal descriptors with a small contribution to the overall correlation.

As already pointed before, the highly intercorrelated descriptors $\mathbf{MinSp}(\mathbf{RD},Z)$ or $\mathbf{MinSp}(\mathbf{RD},Y)$ contain similar structural information; however, the results obtained with eqs 44 and 45 indicate that it is not possible to simplify eq 26 by deleting one of the above two descriptors. The conclusion is that although highly intercorrelated, both $\mathbf{MinSp}(\mathbf{RD},Z)$ and $\mathbf{MinSp}(\mathbf{RD},Y)$ contain some important structural information for the modeling of the retention indices of phenols. Equation 46 shows that by using the sequential orthogonalization of descriptors, it is possible to eliminate the orthogonal descriptor $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$ and to conserve the good statistical indices of the QSRR model. We will now consider a slightly different orthogonalization order of the descriptors, by interchanging the last two highly intercorrelated descriptors, i.e., $\Omega({}^{3}\chi_{\mathrm{p}}\upsilon)$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$. The corresponding QSRR model has the same modeling power as the models from eqs 26 and 43:

$$\begin{aligned}\mathrm{RI} = (1087.2 \pm 142.8) + (285.12 \pm 37.44)\Omega({}^{3}\chi_{\mathrm{p}}{}^{\upsilon}) + \\ (7.0617 \pm 0.9273)\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) - \\ (2234.3 \pm 293.4)\Omega(\mathbf{MinSp}(\mathbf{RD},Y)) + \\ (19186.9 \pm 2519.5)\Omega(\mathbf{MinSp}(\mathbf{RD},Z)) \\ n = 50 \quad r = 0.9931 \quad s = 15 \quad F = 811 \end{aligned} \tag{47}$$

The partial correlation coefficients of the four orthogonal descriptors, i.e., $\Omega({}^{3}\chi_{\mathrm{p}}{}^{\upsilon})$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$, with RI are 0.9065, 0.2936, −0.0900, and 0.2652. The correlation coefficient between RI and the third orthogonal descriptor, $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$, is small, indicating that the contribution of this descriptor to the overall correlation can be ignored. By deleting this orthogonal descriptor from the QSRR model, we obtain a regression equation with three independent variables that has statistical indices slightly lower than those of eqs 26 and 43:

$$\begin{aligned}\mathrm{RI} = (1087.2 \pm 140.8) + (285.12 \pm 36.93)\Omega({}^{3}\chi_{\mathrm{p}}{}^{\upsilon}) + \\ (7.0617 \pm 0.9147)\Omega(\mathbf{MaxSp}(\mathbf{D},Z)) + \\ (19186.9 \pm 2485.3)\Omega(\mathbf{MinSp}(\mathbf{RD},Z)) \\ n = 50 \quad r = 0.9891 \quad s = 19 \quad F = 689 \end{aligned} \tag{48}$$

A comparison of eqs 46 and 48 shows that the best QSRR model is offered by the former equation, indicating that the orthogonalization order of the descriptors is essential for the generation of significant structure−property models.

CONCLUDING REMARKS

Molecular graph descriptors represent valuable structural descriptors that reflect the molecular size, shape, branching, and structural influence of multiple bonds and heteroatoms. They are extensively used to develop QSPR and QSAR, alone or together with other classes of structural descriptors, such as constitutional, geometrical, electrostatic, and quantum descriptors. We have to mention that graph descriptors are intended to complement (and not to substitute) the structural information encoded into other classes of descriptors. The interest of developing new molecular graph descriptors was stimulated in recent years by their use in database mining, virtual screening of combinatorial libraries, and similarity and diversity assessment. In the present investigation we made a comparative study of three general weighting schemes, namely, the atomic number Z, the relative electronegativity X, and the relative covalent radius Y weighting schemes. Recently we have defined several molecular graph operators as a convenient and efficient method to compute from a unique mathematical equation a family of related molecular graph descriptors; such an operator can be applied to all molecular matrixes and sets of parameters for atoms and bonds. Using the three weighting schemes, 60 structural descriptors were computed with several graph operators, namely, the Wiener, hyper-Wiener, minimum eigenvalue, maximum eigenvalue, Ivanciuc−Balaban, and information on distance operators. These descriptors were used to generate quantitative structure−retention relationship models for the retention indices of 50 alkylphenols in gas−liquid chromatography.

The best QSRR model was obtained with four structural descriptors, namely, ${}^{3}\chi_{\mathrm{p}}{}^{\upsilon}$, $\mathbf{MaxSp}(\mathbf{D},Z)$, $\mathbf{MinSp}(\mathbf{RD},Z)$, and $\mathbf{MinSp}(\mathbf{RD},Y)$, with good calibration ($r_{\mathrm{cal}} = 0.9931$, $s_{\mathrm{cal}} = 15$) and prediction ($r_{\mathrm{pr}} = 0.9911$, $s_{\mathrm{pr}} = 17$) results. Since the last two descriptors are highly intercorrelated, we have applied to the above four parameters the sequential orthogonalization of descriptors. Because the third orthogonal descriptor, $\Omega(\mathbf{MinSp}(\mathbf{RD},Z))$, has a small contribution to the overall QSRR model, we have obtained a simplified model with the remaining three orthogonal descriptors, namely, $\Omega({}^{3}\chi_{\mathrm{p}}{}^{\upsilon})$, $\Omega(\mathbf{MaxSp}(\mathbf{D},Z))$, and $\Omega(\mathbf{MinSp}(\mathbf{RD},Y))$. The QSRR model with these three orthogonal descriptors has the same statistical indices (correlation coefficient and standard deviation) as the QSRR model containing the four nonorthogonal descriptors. In this way, the sequential orthogonalization algorithm can be used to simplify a MLR model, by removing the variables with a small contribution.

In the present study, all structural descriptors computed from graph operators were computed from two molecular matrixes, namely, the distance **D** and reciprocal distance **RD**

matrixes. Several other molecular matrixes were proposed in the literature to translate into a numerical form the structure of a molecular graph:[8,10−13] chi χ,[96] Burden **B**,[97] edge adjacency **EA**,[98−100] laplacian **L**,[101−103] distance path $\mathbf{D_p}$,[61,104,105] distance delta $\mathbf{D_\Delta}$,[61,104,105] reciprocal distance-path $\mathbf{RD_p}$,[60,61,104,105] resistance distance matrix Ω,[106,107] detour Δ,[108] detour distance Δ-**D**,[108] distance-detour quotient $\mathbf{D}/\Delta$,[109] edge Wiener $\mathbf{W_e}$,[110,111] path Wiener $\mathbf{W_p}$,[110,111] edge Szeged $\mathbf{Sz_e}$,[112−114] path Szeged $\mathbf{Sz_p}$,[112−114] reciprocal Szeged $\mathbf{RSz_p}$,[112−114] edge Cluj $\mathbf{Cj_e}$,[113−117] and path Cluj $\mathbf{Cj_p}$.[113−117] For various reasons, the above matrixes were not used to derive structural descriptors. The main reason is the absence of published algorithms and parameters that enable the computation of the above matrixes for vertex- and edge-weighted graphs, because these matrixes were defined only for the molecular graphs of alkanes and cycloalkanes. The edge and path Wiener matrixes are defined only for simple, acyclic graphs representing alkanes. The computation of the detour, detour-distance, distance-detour quotient, edge Cluj, and path Cluj matrixes require the enumeration of all paths in the molecular graph, a very time-consuming algorithm for polycyclic compounds.

Using various atomic properties, the three weighting schemes employed in this study give the atom and bond parameters for the computation of the structural descriptors. All weighting schemes are useful and provide structural invariants with a good correlational power, but those computed with *Z* (computed from the atomic number) were selected in the majority of cases, followed by those obtained with *Y* (obtained from the relative covalent radius) and *X* (generated from the relative covalent radius). In other QSPR or QSAR studies this order may change, but the conclusion is that all three weighting schemes give structural descriptors with good correlational power in structure−property models. A set of compounds such as the present one, containing just a single heteroatom, cannot discriminate well the three scales encoding information about heteroatoms; therefore, the decision must await future studies involving databases containing more than one species of heteroatoms.

ACKNOWLEDGMENT

O.I. thanks the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche of France for a PAST grant. T.I. and O.I. acknowledge the kind hospitality of the LARTIC group during their stay in Nice. We acknowledge the partial financial support of this research by the Romanian Ministry of National Education under Grant 33084 T94.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, U.K., 1986.
(3) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(4) Balaban, A. T. From Chemical Graphs to 3D Molecular Modeling. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum: New York, 1997; pp 1−24.
(5) Balaban, A. T. Using Real Numbers as Vertex Invariants for Third Generation Topological Indexes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 23−28.
(6) Balaban, A. T. Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398−402.
(7) Balaban, A. T. Local (Atomic) and Global (Molecular) Graph-Theoretical Descriptors. *SAR QSAR Environ. Res.* **1995**, *3*, 81−95.
(8) Ivanciuc, O.; Balaban, A. T. Graph Theory in Chemistry. 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, U.K., 1998; pp 1169−1190.
(9) Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp. 21−57.
(10) Ivanciuc, O.; Balaban, A. T. The Graph Description of Chemical Structures. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp. 59−167.
(11) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Vertex- and Edge-Weighted Molecular Graphs and Derived Structural Descriptors. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 169−220.
(12) Ivanciuc, O.; Ivanciuc, T. Matrixes and Structural Descriptors Computed from Molecular Graph Distances. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: The Netherlands, 1999; pp 221−277.
(13) Ivanciuc, O.; Ivanciuc, T.; Diudea, M. V. Molecular Graph Matrices and Derived Structural Descriptors. *SAR QSAR Environ. Res.* **1997**, *7*, 63−87.
(14) Diudea, M. V.; Gutman, I. Wiener-Type Topological Indices. *Croat. Chem. Acta* **1998**, *71*, 21−51.
(15) Mallion, R. B.; Schwenk, A. J.; Trinajstić, N. A Graphical Study of Heteroconjugated Molecules. *Croat. Chem. Acta* **1974**, *46*, 171−182.
(16) Mallion, R. B.; Trinajstić, N.; Schwenk, A. J. Graph Theory in Chemistry. Generalization of Sachs' Formula. *Z. Naturforsch.* **1974**, *29a*, 1481−1484.
(17) Graovac, A.; Polansky, O. E.; Trinajstić, N.; Tyutyulkov N. Graph Theory in Chemistry. II. Graph-Theoretical Description of Heteroconjugated Molecules. *Z. Naturforsch.* **1975**, *30a*, 1696−1699.
(18) Trinajstić, N. Computing the Characteristic Polynomial of a Conjugated System Using the Sachs Theorem. *Croat. Chem. Acta* **1977**, *49*, 593−633.
(19) Ivanciuc, O. Chemical Graph Polynomials. Part 1. The Polynomial Description of Generalized Chemical Graphs. *Rev. Roum. Chim.* **1988**, *33*, 709−717.
(20) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(21) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43−51.
(22) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(23) Basak, S. C.; Gieschen, D. P.; Harriss, D. K.; Magnuson, V. R. Physicochemical and Topological Correlates of the Enzymatic Acetyltransfer Reaction. *J. Pharm. Sci.* **1983**, *72*, 934−937.
(24) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. A Quantitative Structure−Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneim.-Forsch. (Drug Res.)* **1983**, *33*, 501−503.
(25) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Comparative Study of Lipophilicity versus Topological Molecular Descriptors in Biological Correlations. *J. Pharm. Sci.* **1984**, *73*, 429−437.
(26) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.
(27) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(28) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking Into Account Periodicities of Element Properties. *MATCH (Commun. Math. Chem.)* **1986**, *21*, 115−122.
(29) Balaban, A. T.; Ivanciuc, O. FORTRAN 77 Computer Program for Calculating the Topological Index J for Molecules Containing Heteroatoms. In *MATH/CHEM/COMP 1988*, *Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Sciences*, Dubrovnik, Yugoslavia, June 20−25 1988; Graovac, A., Ed.; Studies in Physical and Theoretical Chemistry 63; Elsevier: Amsterdam,1989; pp 193−212.
(30) Balaban, A. T.; Catana, C.; Dawson, M.; Niculescu-Duvaz, I. Applications of Weighted Topological Index J for QSAR of Carcinogenesis Inhibitors (Retinoic Acid Derivatives). *Rev. Roum. Chim.* **1990**, *35*, 997−1003.

(31) Bonchev, D.; Mountain, C. F.; Seitz, W. A.; Balaban, A. T. Modeling the Anticancer Action of Some Retinoid Compounds by Making Use of the OASIS Methodol. *J. Med. Chem.* **1993**, *36*, 1562−1569.
(32) Balaban, A. T.; Kier, L. B.; Joshi, N. Correlations between Chemical Structure and Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals, and Their Sulfur Analogues. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237−244.
(33) Lall, R. S.; Srivastava, V. K. Edge-Weighted Graphs in Biological Systems. *MATCH (Commun. Math. Chem.)* **1982**, *13*, 325−332.
(34) Tvaruzek, P.; Komenda, J. Calculation of Graph Distance Matrix for Heteroatomic Systems. *MATCH (Commun. Math. Chem.)* **1989**, *24*, 317−322.
(35) Filip, P. A.; Balaban, T.-S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61−83.
(36) Ivanciuc, O.; Balaban, T.-S.; Filip, P.; Balaban, A. T. Design of Topological Indices. Part 7. Analytical Formulae for Local Vertex Invariants of Linear and Monocyclic Molecular Graphs.*MATCH (Commun. Math. Chem.)* **1992**, *28*, 151−164.
(37) Diudea, M. V.; Silaghi-Dumitrescu, I. Molecular Topology. I. Valence Group Electronegativity as a Vertex Discriminator. *Rev. Roum. Chim.* **1989**, *34*, 1175−1182.
(38) Klopman, G.; Raychaudhury, C.; Henderson, R. V. A New Approach to Structure−Activity Using Distance Information Content of Graph Vertices: A Study with Phenylalkylamines. *Math. Comput. Model.* **1988**, *11*, 635−640.
(39) Klopman, G.; Raychaudhury, C. Vertex Indices of Molecular Graphs in Structure−Activity Relationships: A Study of the Convulsant−Anticonvulsant Activity of Barbiturates and the Carcinogenicity of Unsubstituted Polycyclic Aromatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 12−19.
(40) Wang, S.; Milne, G. W. A.; Klopman, G. Graph Theory and Group Contributions in the Estimation of Boiling Points. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1242−1250.
(41) Yu, F.-B.; Xun, Y.-M.; Cheng, W.-T. Structure/Property Correlation of Substituted Compounds Based on Graph Theory. *Anal. Chim. Acta* **1990**, *234*, 487−491.
(42) Randić, M. On Computation of Optimal Parameters for Multivariate Analysis of Structure−Property Relationshsip. *J. Comput. Chem.* **1991**, *12*, 970−980.
(43) Randić, M. Novel Graph Theoretical Approach to Heteroatoms in Quantitative Structure−Activity Relationships. *Chemom. Intell. Lab. Syst.* **1991**, *10*, 213−227.
(44) Balaban, A. T.; Bonchev, D.; Seitz, W. A. Topological/Chemical Distances and Graph Centers in Molecular Graphs with Multiple Bonds. *J. Mol. Struct. (THEOCHEM)* **1993**, *280*, 253−260.
(45) Yang, Y.-Q.; Xu, L.; Hu, C.-Y. Extended Adjacency Matrix Indices and Their Applications. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1140−1145.
(46) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstić, N. On the Distance Matrix of Molecules Containing Heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, **1983**; pp 222−227.
(47) Medić-Šarić, M.; Nikolić, S.; Matijević-Sosa, J. A QSPR Study of 3-(Phthalimidoalkyl)-pyrazolin-5-ones. *Acta Pharm.* **1992**, *42*, 153−167.
(48) Jurić, A.; Gagro, M.; Nikolić, S.; Trinajstić, N. Molecular Topological Index: An Application in the QSAR Study of Toxicity of Alcohols. *J. Math. Chem.* **1992**, *11*, 179−186.
(49) Nikolić, S.; Medić-Šarić, M.; Matijević-Sosa, J. A QSAR Study of 3-(Phthalimidoalkyl)-pyrazolin-5-ones. *Croat. Chem. Acta* **1993**, *66*, 151−160.
(50) Nikolić, S.; Trinajstić, N.; Mihalić, Z. Molecular Topological Index: An Extension to Heterosystems. *J. Math. Chem.* **1993**, *12*, 251−264.
(51) Medić-Sarić, M.; Rendić, S.; Vestemar, V.; Sarić, S. A Comparative Study of Some Topological Indices and log P in Structure−Property-Activity Analysis of Phenylalkylamines. *Acta Pharm.* **1993**, *43*, 15−26.
(52) Amić, D.; Davidović-Amić, D.; Trinajstić, N. Calculation of Retention Times of Anthocyanins with Orthogonalized Topological Indices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 136−139.
(53) Ivanciuc, O.; Balaban, A. T. Design of Topological Indices. Part 5. Precision and Error in Computing Graph Theoretic Invariants for Molecules Containing Heteroatoms and Multiple Bonds. *MATCH (Commun. Math. Chem.)* **1994**, *30*, 117−139.
(54) Ivanciuc, O.; Ivanciuc, T.; Balaban A. T. Design of Topological Indices. Part 10. Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395−401.
(55) Ivanciuc, O. Design of Topological Indices. Part 12. Parameters for Vertex- and Edge-Weighted Molecular Graphs. *Rev. Roum. Chim.*, in press.
(56) Ivanciuc, O. Design of Topological Indices. Part 19. Computation of Vertex and Molecular Graph Structural Descriptors with Operators. *Rev. Roum. Chim.*, in press.
(57) Balaban, T. S.; Filip, P. A.; Ivanciuc, O. Computer Generation of Acyclic Graphs Based on Local Vertex Invariants and Topological Indices. Derived Canonical Labeling and Coding of Trees and Alkanes. *J. Math. Chem.* **1992**, *11*, 79−105.
(58) Plavšić, D.; Nikolić, S.; Trinajstić, N.; Mihalić, Z. On the Harary Index for the Characterization of Chemical Graphs. *J. Math. Chem.* **1993**, *12*, 235−250.
(59) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318.
(60) Diudea, M. V.; Ivanciuc, O.; Nikolić, S.; Trinajstić, N. Matrices of Reciprocal Distance, Polynomials and Derived Numbers. *MATCH (Commun. Math. Comput. Chem.)* **1997**, *35*, 41−64.
(61) Ivanciuc, O.; Diudea, M. V.; Khadikar, P. V. New Topological Matrices and Their Polynomials. *Ind. J. Chem.* **1998**, *37A*, 574−585.
(62) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 339−353.
(63) Pearlman, R. S.; Smith, K. M. Software for Chemical Diversity in the Context of Accelerated Drug Discovery. *Drugs Future* **1998**, *23*, 885−895.
(64) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 11−20.
(65) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(66) Balaban, A. T.; Balaban, T.-S. New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8*, 383−397.
(67) Balaban, A. T.; Balaban, T.-S. Correlations Using Topological Indices Based on Real Graph Invariants. *J. Chim. Phys.* **1992**, *89*, 1735−1745.
(68) Balaban, A. T.; Feroiu, V. Correlations Between Structure and Critical Data or Vapor Pressures of Alkanes by Means of Topological Indices. *Rep. Mol. Theor.* **1990**, *1*, 133−139.
(69) Ivanciuc, O.; Balaban, A. T. Design of Topological Indices. Part 20. Molecular Structure Descriptors Computed with Information on Distances Operators. *Rev. Roum. Chim.* **1999**, *44*, 479−489.
(70) Ivanciuc, O.; Balaban, A. T. Design of Topological Indices. Part 21. Molecular Graph Operators for the Computation of Geometric Structural Descriptors. *Rev. Roum. Chim.* **1999**, *44*, 539−547.
(71) Woloszyn, T. F.; Jurs, P. C. Prediction of Gas Chromatographic Retention Data for Hydrocarbons from Naphthas. *Anal. Chem.* **1993**, *65*, 582−587.
(72) Georgakopoulos, C. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Stimulants and Narcotics. *Anal. Chem.* **1991**, *63*, 2021−2024.
(73) Georgakopoulos, C. G.; Tsika, O. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Anabolic Steroids. *Anal. Chem.* **1991**, *63*, 2025−2028.
(74) Woloszyn, T. F.; Jurs, P. C. Quantitative Structure-Retention Relationship Studies of Sulfur Vesicants. *Anal. Chem.* **1992**, *64*, 3059−3063.
(75) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.
(76) Buryan, P.; Nabivach, V. M.; Dmitrikov, V. P. Structure-Retention Correlations of Isomeric Alkylphenols in Gas−Liquid Chromatography. *J. Chromatogr.* **1990**, *509*, 3−14.
(77) Ivanciuc, O. Design of Topological Indices. Part 15. The Szeged Index of Vertex- and Edge-Weighted Molecular Graphs. *Rev. Roum. Chim.*, in press.
(78) Nagle, J. K. Atomic Polarizability and Electronegativity. *J. Am. Chem. Soc.* **1990**, *112*, 4741−4747.
(79) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure−Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066−1078.
(80) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(81) Wold, S.; Eriksson, L. Validation Tools. In *QSAR: Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; Methods and Principles in Medicinal Chemistry, Vol. 2; Verlag Chemie: Weinheim, Germany, 1995; pp 309−318.
(82) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517−525.

(83) Randić, M.; Seybold, P. G. Molecular Shape as a Critical Factor in Structure−Property-Activity Studies. *SAR QSAR Environ. Res.* **1993**, *1*, 77−85.
(84) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. A Structure−Property Study of the Solubility of Aliphatic Alcohols in Water. *Croat. Chem. Acta* **1995**, *68*, 417−434.
(85) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. A Novel QSAR Approach to Physicochemical Properties of the α-Amino Acids. *Croat. Chem. Acta* **1995**, *68*, 435−450.
(86) Amić, D.; Davidović-Amić, D.; Trinajstić, N. Calculation of Retention Times of Anthocyanins with Orthogonalized Topological Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 136−139.
(87) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D. The Structure−Property Models Can Be Improved Using the Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538.
(88) Šoškić, M.; Plavšić, D.; Trinajstić, N. 2-Difluoromethylthio-4,6-bis-(monoalkylamino)-1,3,5-triazines as Inhibitors of Hill Reaction: A QSAR Study with Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 146−150.
(89) Šoškić, M.; Plavšić, D.; Trinajstić, N. Link between Orthogonal and Standard Multiple Linear Regression Models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829−832.
(90) Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. Hierarchical Orthogonalization of Descriptors. *Int. J. Quantum Chem.* **1997**, *63*, 215−222.
(91) Amić, D.; Davidović-Amić, D.; Bešlo, D.; Lučić, B.; Trinajstić, N. Calculation of Retention Times of Anthocyanins with Orthogonalized Topological Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 136−139.
(92) Lučić, B.; Trinajstić, N. New Developments in QSPR/QSAR Modeling based on Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45−62.
(93) Nikolić, S.; Trinajstić, N. Modeling the Aqueous Solubility of Aliphatic Alcohols. *SAR QSAR Environ. Res.* **1998**, *9*, 117−126.
(94) Lučić, B.; Trinajstić, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.
(95) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.
(96) Randić, B.; Trinajstić, N., M. Similarity Based on Extended Basis Descriptors. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 686−692.
(97) Burden, F. R. Molecular Identification Number for Substructure Searches.*J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(98) Estrada, E. Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31−33.
(99) Estrada, E. Edge Adjacency Relationships in Molecular Graphs Containing Heteroatoms: A New Topological Index Related to Molar Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 701−707.
(100) Estrada, E.; Ramirez, A. Edge Adjacency Relationships and Molecular Topographic Descriptors. Definition and QSAR Applications. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 837−843.
(101) Mohar, B. Laplacian Matrixes of Graphs. *Stud. Phys. Theor. Chem.* **1989**, *63*, 1−8.
(102) Ivanciuc, O. Chemical Graph Polynomials. Part 3. The Laplacian Polynomial of Molecular Graphs. *Rev. Roum. Chim.* **1993**, *38*, 1499−1508.
(103) Trinajstić, N.; Babić, D.; Nikolić, S.; Plavšić, D.; Amić, D.; Mihalić, Z. The Laplacian Matrix in Chemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 368−376
(104) Diudea, M. V. Walk Numbers $^{e}W_{M}$: Wiener-Type Numbers of Higher Rank. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 535−540.
(105) Diudea, M. V. Wiener and Hyper-Wiener Numbers in a Single Matrix. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 833−836.
(106) Klein, D. J.; Randić, M. Resistance Distance. *J. Math. Chem.* **1993**, *12*, 81−95.
(107) Bonchev, D.; Balaban, A. T.; Liu, X.; Klein, D. J. Molecular Cyclicity and Centricity of Polycyclic Graphs. I. Cyclicity Based on Resistance Distances or Reciprocal Distances. *Int. J. Quantum Chem.* **1994**, *50*, 1−20.
(108) Ivanciuc, O.; Balaban, A. T. Design of Topological Indices. Part 8. Path Matrixes and Derived Molecular Graph Invariants. *MATCH (Commun. Math. Chem.)* **1994**, *30*, 141−152.
(109) Randić, M. On Characterization of Cyclic Structures.*J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1063−1071.
(110) Randić, M. Novel Molecular Descriptor for Structure−Property Studies. *Chem. Phys. Lett.* **1993**, *211*, 478−483.
(111) Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H. Wiener Matrix: Source of Novel Graph Invariants. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 709−716.
(112) Diudea, M. V.; Minailiuc, O.; Katona, G.; Gutman, I. Szeged Matrixes and Related Numbers. *MATCH (Commun. Math. Comput. Chem.)* **1997**, *35*, 129−143.
(113) Diudea, M. V. Cluj Matrix Invariants. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 300−305.
(114) Diudea, M. V.; Pârv, B.; Topan, M. I. Derived Szeged and Cluj Indices. *J. Serb. Chem. Soc.* **1997**, *62*, 267−276.
(115) Diudea, M. V.; Pârv, B.; Gutman, I. Detour-Cluj Matrix and Derived Invariants. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1101−1108.
(116) Diudea, M. V. Cluj Matrix, CJ_u: Source of Various Graph Descriptors. *MATCH (Commun. Math. Comput. Chem.)* **1997**, *35*, 169−183.
(117) Kiss, A. A.; Katona, G.; Diudea, M. V. Szeged and Cluj Matrixes within the Matrix Operator $W_{(M1,M2,M3)}$. *Collect. Sci. Pap. Fac. Sci. Kragujevac* **1997**, *19*, 95−107.

CI990129W