Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors

Viviana Consonni, Roberto Todeschini,* and Manuela Pavan

Department of Environmental Sciences, Milano - Bicocca University, P.za della Scienza 1,
20126 Milano, Italy

Received December 21, 2001

Novel molecular descriptors based on a leverage matrix similar to that defined in statistics and usually used for regression diagnostics are presented. This leverage matrix, called *Molecular Influence Matrix* (MIM), is here proposed as a new molecular representation easily calculated from the spatial coordinates of the molecule atoms in a chosen conformation. The proposed molecular descriptors are called *GETAWAY* (GEometry, Topology, and Atom-Weights AssemblY) as they try to match 3D-molecular geometry provided by the molecular influence matrix and atom relatedness by molecular topology, with chemical information by using different atomic weightings (atomic mass, polarizability, van der Waals volume, and electronegativity, together with unit weights). A first set of molecular descriptors, called *H-GETAWAY*, is derived by using only the information provided by the molecular influence matrix, while a second set, called *R-GETAWAY*, combines this information with geometric interatomic distances in the molecule. The prediction ability in structure−property correlations of the new descriptors was tested by analyzing regressions of these descriptors for selected properties of octanes.

INTRODUCTION

During the past decade, a great explosion of molecular descriptors has been observed. Surface areas, volume descriptors, charges, and quantum-chemical measures have been extensively enhanced and used as the descriptors of the whole molecule. Moreover, a tendency to extend traditional topological indices accounting for three-dimensional representation of the molecule by including geometrical information and/or physicochemical atomic properties has been apparent.[1] Among such indices, usually called topographic descriptors, are the 3D-Wiener index,[2−5] the Randic molecular profiles,[6−9] and BCUT descriptors,[10,11] all examples of molecular descriptors accounting for 3D molecular information, although based on topological approaches.

These descriptors are easily and quickly calculated, thus being suitable for both QSAR modeling and similarity/diversity analysis of large chemical databases. In recent years the latter task has been becoming a growing research field of great interest in the pharmaceutical and agrochemical industries, where combinatorial chemistry and High-Throughput Screening (HTS) are effective approaches to lead discovery. The main objective of such approaches is to select a subset which best represents the full range of chemical diversity present in a large population of compounds. It is obvious that the results of this selection strictly depend on how the chemical diversity is described. Therefore, molecular descriptors catching as much chemical and structural information as possible are desirable. However, they also need to be calculated easily in order to avoid long computational time for screening large populations of compounds.

Although a lot of different molecular descriptors have been proposed until now, both in QSAR modeling and similarity/diversity analysis there still exists the need of new descriptors, because each class of descriptors encodes some specific structural features and thus it is useful to have an exhaustive description of the molecular structure.

In this paper novel 3D molecular descriptors are presented, based on an influence or leverage matrix similar to that typically defined by statisticians in regression diagnostics.[12] These descriptors, called *GEometry, Topology, and Atom-Weights AssemblY* (GETAWAY), encode both the geometrical information given by the influence molecular matrix and the topological information given by the molecular graph, weighted by chemical information encoded in selected atomic weightings.

Two sets of molecular descriptors have been devised: H-GETAWAY descriptors have been calculated from the *molecular influence matrix* **H**, while R-GETAWAY descriptors are from the *influence/distance matrix* **R** where the elements of the molecular influence matrix are combined with those of the geometry matrix. With the aim of catching relevant chemical information and in some cases also molecular complexity, these new descriptors have been defined by applying some traditional matrix operators, concepts of the information theory, and spatial autocorrelation formulas.

This paper is mainly dedicated to the theory of the GETAWAY descriptors. In the first section, the novel molecular influence matrix is defined with its peculiar mathematical properties. In the second and third sections, definitions and formulas of H-GETAWAY and R-GETAWAY descriptors, respectively, are given. In the fourth section, the correlation ability of the new descriptors is tested

* Corresponding author e-mail: roberto.todeschini@unimib.it.

10.1021/ci015504a CCC: $22.00 © 2002 American Chemical Society
Published on Web 04/20/2002

Table 1. Molecular Influence Matrix of Chlorobenzene[a]

ID	C1	C2	C3	C4	C5	C6	H7	H8	H9	H10	H11	Cl
C1	**0.065**	0.031	−0.036	−0.069	−0.036	0.031	0.057	−0.062	−0.123	−0.062	0.057	0.148
C2	0.031	**0.075**	0.042	−0.034	−0.077	−0.044	0.134	0.076	−0.059	−0.136	−0.079	0.071
C3	−0.036	0.042	**0.079**	0.039	−0.039	−0.077	0.075	0.141	0.068	−0.071	−0.138	−0.082
C4	−0.069	−0.034	0.039	**0.075**	0.039	−0.034	−0.061	0.067	0.132	0.067	−0.061	−0.159
C5	−0.036	−0.077	−0.039	0.039	**0.079**	0.042	−0.138	−0.071	0.068	0.141	0.075	−0.082
C6	0.031	−0.044	−0.077	−0.034	0.042	**0.075**	−0.079	−0.136	−0.059	0.076	0.134	0.071
H7	0.057	0.134	0.075	−0.061	−0.138	−0.079	**0.242**	0.135	−0.108	−0.246	−0.141	0.130
H8	−0.062	0.076	0.141	0.067	−0.071	−0.136	0.135	**0.250**	0.118	−0.129	−0.246	−0.143
H9	−0.123	−0.059	0.068	0.132	0.068	−0.059	−0.108	0.118	**0.232**	0.118	−0.108	−0.280
H10	−0.062	−0.136	−0.071	0.067	0.141	0.076	−0.246	−0.129	0.118	**0.250**	0.135	−0.143
H11	0.057	−0.079	−0.138	−0.061	0.075	0.134	−0.141	−0.246	−0.108	0.135	**0.242**	0.130
Cl	0.148	0.071	−0.082	−0.159	−0.082	0.071	0.130	−0.143	−0.280	−0.143	0.130	**0.337**

[a] The atom numbering refers to Figure 1.

in a QSPR study of some physicochemical properties of octanes.

MOLECULAR INFLUENCE MATRIX

Let **M** be the molecular matrix constituted by *A* rows corresponding to the atoms in a molecule (hydrogen atoms included) and three columns corresponding to the Cartesian coordinates x, y, z of each atom in some optimized molecular structure. Atomic coordinates are assumed to be calculated with respect to the geometrical center of the molecule in order to obtain translational invariance.

The *Molecular Influence Matrix* (MIM), denoted by **H** and resembling the leverage (or influence) matrix defined in regression diagnostics,[12] is calculated from the molecular matrix **M** as

$$\mathbf{H} = \mathbf{M}\cdot(\mathbf{M}^{\mathbf{T}}\cdot\mathbf{M})^{-1}\cdot\mathbf{M}^{\mathbf{T}} \qquad (1)$$

where the superscript T refers to the transposed matrix.

This is a symmetric $A \times A$ matrix, where A represents the number of atoms, with the following mathematical properties

$$0 \leq h_{ii} \leq 1 \quad \sum_{i=1}^{A} h_{ii} = D \quad \bar{h} = D/A$$

$$D = 1, 2, \text{ or } 3 \quad \sum_{j=1}^{A} h_{ij} = 0$$

where h denotes elements of the molecular influence matrix, D is the rank of the molecular matrix **M** equal to 1, 2, and 3 for linear, planar, and 3D-molecules, respectively, and $\bar{h}$ is the average value of the diagonal terms. Note that in all the cases where the matrix rank D is lower than three, the generalized Penrose inverse[13] is required to calculate the matrix **H**. Another very important property of the molecular influence matrix is the rotational invariance with respect to the molecule coordinates.

The diagonal elements h_{ii} of the molecular influence matrix, called *leverages*, encode atomic information and represent the "influence" of each molecule atom in determining the whole shape of the molecule; in fact mantle atoms always have higher h_{ii} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in the molecule depends on the size and shape of the molecule itself. Lower leverages can be found for atoms in molecules of spherical shape, while higher leverages for atoms in more linear molecules. In a series of molecules with almost the same shape, the maximum leverage decreases as the molecular size (number of atoms) increases.

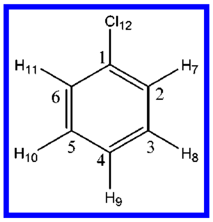

Figure 1. Atom numbering of chlorobenzene.

Each off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interactions with the *i*th atom, or, in other words, the attitude of the two considered atoms to interact between themselves. Negative sign of the off-diagonal elements means that the two atoms occupy opposite molecular regions with respect to the center and hence their mutual degree of accessibility should be low.

As derived from the geometry of the molecule, leverage values are effectively sensitive to significant conformational changes and to the bond lengths that account for atom types and bond multiplicity. Table 1 collects all the values of the molecular influence matrix of chlorobenzene, whose three-dimensional structure has been optimized by minimizing the conformational energy. Atom numbering of chlorobenzene is shown in Figure 1. Moreover, leverage values of the atoms of chlorobenzene, bromobenzene, and iodobenzene are shown in Figure 2. It can be noted that the outer atoms have larger leverage values than the carbon atoms of the aromatic ring. Then, among the outer atoms, the halogens have the largest value, and this value increases from chlorobenzene to bromobenzene and to iodobenzene since it is sensitive to the increase in bond length. Note also that equivalent atoms have equal leverage values.

H-GETAWAY DESCRIPTORS

As pointed out above, the molecular influence matrix **H** contains some useful information on the molecular geometry and especially the diagonal elements (leverages) of the matrix allow to discriminate among the atoms according to their position in the 3D molecular space with respect to the molecule center. The first set of new molecular descriptors has been derived from this molecular matrix and hence called H-GETAWAY descriptors.

Most of these descriptors are simply calculated only by the leverages used as the atomic weightings in the molecular graph, thus obtaining a vertex-weighted molecular graph as

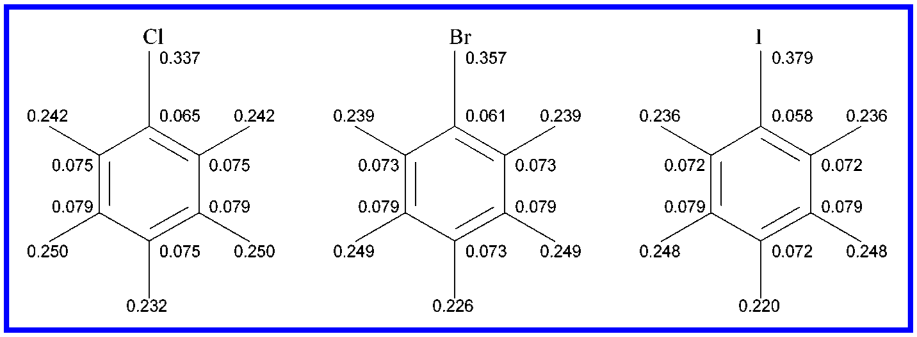

Figure 2. Leverage values of the atoms of chlorobenzene, bromobenzene, and iodobenzene.

the representation of the molecule from which traditional descriptors have been derived.

The most simple H-GETAWAY descriptor has been defined as the *geometric mean on the leverage magnitude*

$$H_{GM} = 100\cdot(\prod_{i=1}^{A} h_{ii})^{1/A} \tag{2}$$

where A represents the number of atoms and the factor 100 scales the index values between 0 and 100. High values of this index are obtained when all the leverages have similar values and this is the case of molecules of almost spherical shape. However, it has to be pointed out that in all the cases where one atom is located exactly in or next to the molecular space center, the H_{GM} value will be zero or very low although the molecule is very compact and branched, since there will be one leverage so small to drop the geometric mean. To avoid this drawback causing a loss of information, an inner protection sphere is used in such a way to exclude from the geometric mean the atoms with leverage value equal to or smaller than 10^{-6}.

In the isomeric series of hydrocarbons, the H_{GM} index increases from linear to more branched molecules; it is also inversely related to molecular size, decreasing as the number of atoms in the molecule increases.

The other H-GETAWAY descriptors have been conceptually divided in three main groups reported below: information indices, autocorrelation descriptors, and local-vertex invariants.

Information Indices. The concept of *molecular complexity* was introduced into chemistry about 20 years ago and is based on the information content of molecules. Several different measures of complexity can be obtained according to the diversity of the considered structural elements such as atom types, bonds, connections, cycles, etc. The first attempts to quantify molecular complexity were based on the elemental composition of molecules; later other molecular characteristics were considered such as the symmetry of molecular graphs, molecular branching, molecular cyclicity, and molecular centricity. The concept of molecular complexity became a hierarchically defined concept to which further discrimination is provided by geometric interatomic distances and spatial molecular symmetry.[14-18]

As the diagonal values of the leverage matrix are sensitive to the whole molecule structure, they automatically contain information on molecular complexity, which is a function of the size, symmetry, elemental molecular composition, molecular branching, and centricity. 3D complexity descriptors have been defined as the *total* and *standardized information content on the leverage equality* as

$$I_{\mathrm{TH}} = A_0\cdot\log_2 A_0 - \sum_{g=1}^{G} \mathrm{N_g}\cdot\log_2 \mathrm{N_g}$$

$$\text{and} \qquad I_{\mathrm{SH}} = \frac{I_{\mathrm{TH}}}{A_0\cdot\log_2 A_0} = 1 - \frac{\sum_{g=1}^{G} \mathrm{N_g}\cdot\log_2 \mathrm{N_g}}{A_0\cdot\log_2 A_0} \tag{3}$$

where $\mathrm{N_g}$ is the number of atoms with the same leverage value and G is the number of equivalence classes into which the atoms are partitioned according to the leverage equality (a 4-digit approximation is assumed). For these two descriptors, the equivalence classes are derived from only the non-hydrogen atoms in order to avoid biased estimates of the symmetry due to the hydrogen degrees of freedom. Hence, A_0 represents the number of non-hydrogen atoms in the molecule.

If all the atoms have different leverage values, i.e., the molecule does not show any element of symmetry, $I_{\mathrm{TH}} = A_0\cdot \log A_0$ and $I_{\mathrm{SH}} = 1$; otherwise, if all the atoms have equal leverage values (a perfectly symmetric theoretical case), $I_{\mathrm{TH}} = 0$ and $I_{\mathrm{SH}} = 0$.

These indices encode information on the molecule entropy (thermodynamic entropy) and hence should be useful in modeling physicochemical properties related to entropy and symmetry.

Another interesting information index has been defined as the *mean information content on the leverage magnitude* (HIC)

$$HIC = \bar{I}_{\mathrm{H}} = -\sum_{i=1}^{A} \frac{h_{ii}}{D}\cdot\log{}_2\frac{h_{ii}}{D} \tag{4}$$

where D is the matrix rank (i.e. the sum of all leverages) defined above and A is the total number of atoms, hydrogens included.

To better understand the information contained in GETAWAY descriptors, a first analysis of the values of some descriptors was performed by simple comparison within a small data set of structurally diverse compounds. The values of the analyzed molecular descriptors for 40 simple diverse compounds are collected in Table 2. First of all, it can be noted that the *standardized information content on the leverage equality* I_{SH} is quite degenerate, and it can be considered useful to identify molecules having all atoms equivalent and therefore largely symmetric. As the *total information content on the leverage equality* I_{TH} is more

Table 2. Values of Some GETAWAY Descriptors for 40 Structurally Diverse Molecules

compound	H_{GM}	I_{TH}	I_{SH}	*HIC*	$HATS_1$(u)	*RCON*	*RARS*	*REIG*	R_1(u)	*RT*(u)
ethane	28.836	0	0	2.770	0.185	6.657	1.160	1.204	0.979	9.276
propane	22.673	2.755	0.579	3.268	0.229	9.718	1.119	1.152	1.278	12.311
n-butane	17.64	4	0.5	3.617	0.200	11.389	1.028	1.066	1.321	14.396
n-pentane	13.685	7.610	0.655	3.874	0.191	12.477	0.938	0.991	1.357	15.942
n-hexane	11.855	13.51	0.871	4.113	0.168	13.927	0.879	0.928	1.384	17.586
isobutane	18.155	3.245	0.406	3.649	0.253	12.325	1.075	1.108	1.490	15.045
neopentane	17.974	3.610	0.311	3.947	0.271	12.538	1.011	1.076	1.624	17.181
cis-2-butene	20.01	4	0.5	3.324	0.253	9.865	1.005	1.042	1.372	12.062
trans-2-butene	18.037	4	0.5	3.278	0.225	8.813	0.946	1.014	1.243	11.356
2-butyne	15.042	4	0.5	2.864	0.266	6.302	0.828	0.970	1.179	8.281
cyclopropane	27.255	0	0	2.954	0.300	9.201	1.158	1.182	1.370	10.422
cyclobutane	21.186	0	0	3.400	0.259	11.805	1.108	1.129	1.464	13.295
cyclopentane	17.337	11.610	1	3.741	0.230	14.247	1.063	1.084	1.512	15.942
cyclohexane	14.815	0	0	4.029	0.214	16.775	1.031	1.049	1.600	18.559
cyclohexanone	15.412	15.651	0.796	3.924	0.221	15.507	1.011	1.032	1.559	17.194
methanol	34.466	2	1	2.320	0.235	4.558	1.191	1.270	0.885	7.144
ethanol	25.712	4.755	1	2.938	0.245	7.441	1.124	1.174	1.168	10.115
trifluoroethanol	25.065	13.510	0.871	2.910	0.221	6.677	1.021	1.069	1.012	9.186
2-aminoethanol	21.472	8	1	3.220	0.239	9.106	1.079	1.130	1.314	11.868
propanol	20.201	8	1	3.370	0.218	9.918	1.062	1.106	1.296	12.744
benzene	14.193	0	0	3.376	0.159	7.500	0.669	0.677	1.114	8.022
toluene	13.13	19.651	1	3.386	0.247	10.668	0.823	0.956	1.414	12.345
phenol	12.77	19.651	1	3.475	0.169	7.715	0.643	0.661	1.175	8.360
benzoic acid	10.305	28.529	1	3.660	0.157	8.073	0.594	0.625	1.167	8.911
aniline	11.72	15.651	0.796	3.587	0.159	8.218	0.637	0.659	1.187	8.920
nitrobenzene	11.508	22.529	0.790	3.574	0.158	7.831	0.610	0.633	1.118	8.537
F-benzene	14.074	15.651	0.796	3.366	0.155	7.331	0.655	0.664	1.078	7.858
Cl-benzene	13.812	15.651	0.796	3.341	0.148	7.037	0.631	0.641	1.024	7.569
Br-benzene	13.694	12.897	0.656	3.329	0.145	6.923	0.621	0.632	1.005	7.454
I−benzene	13.564	15.651	0.796	3.315	0.142	6.805	0.611	0.623	0.986	7.336
2-propanone	23.743	6	0.750	3.076	0.237	8.300	1.062	1.104	1.210	10.622
2-propanol	20.349	8	1	3.395	0.281	10.309	1.090	1.134	1.435	13.077
2-propylamine	19.687	8	1	3.543	0.283	11.815	1.109	1.141	1.539	14.419
2-fluoropropane	22.37	6	0.75	3.252	0.222	9.357	1.082	1.114	1.222	11.897
2-iodopropane	21.837	8	1	3.215	0.212	8.852	1.022	1.054	1.158	11.245
2-propanethiol	20.201	8	1	3.389	0.304	9.765	1.030	1.071	1.334	12.358
methylamine	32.619	2	1	2.557	0.254	6.097	1.214	1.263	1.075	8.496
dimethylamine	25.061	2.755	0.579	3.126	0.257	9.151	1.152	1.182	1.311	11.522
naphthalene	8.77	15.219	0.458	3.936	0.142	9.902	0.567	0.596	1.240	10.214
anthracene	6.657	27.303	0.512	4.343	0.123	12.083	0.504	0.532	1.312	12.028

discriminating than I_{SH} for its dependence on molecular size, it is a more suitable measure of molecular complexity. This means that molecules with all equivalent atoms are distinguished by I_{TH} according to their different sizes.

Moreover, both I_{TH} and I_{SH} are invariant to the presence of multiple bonds in the molecule (see, for example, *n*-butane, 2-butene, 2-butyne) and to the conformational changes (see *cis*- and *trans*-2-butene). This is due to the fact that they are based on equivalencies among leverage values and not on their absolute values. From this point of view, the *mean information content on the leverage magnitude HIC* seems to catch more information related to molecular complexity. Note that, differently from I_{TH} and I_{SH}, *HIC* can recognize the different substituents in the series of monosubstituted benzenes as well as the two *cis*/*trans*-2-butene isomers. It is also sensitive to the presence of multiple bonds.

Note also that the *geometric mean on the leverage magnitude* H_{GM} is sensitive to the molecular shape and decreases as the size of the molecule increases.

Autocorrelation Descriptors. In the H-GETAWAY descriptors defined above, chemical properties of the molecule atoms are accounted for only implicitly; in fact, the leverage values depend on the molecular geometry, this also being dependent on the atom chemical properties.

If chemical information wants to be explicitly considered, geometrical information provided by leverage values has to be combined with atomic weightings accounting for specific physicochemical properties of molecule atoms, thus resulting in new molecular descriptors.

The simplest quadratic molecular descriptor *P* can be obtained by summing the squared atomic property values. Mathematically

$$P = \mathbf{w}^{\mathbf{T}} \cdot \mathbf{I} \cdot \mathbf{w} = \mathbf{w}^{\mathbf{T}} \cdot \mathbf{w} = \sum_{i=1}^{A} w_i^2 \qquad (5)$$

where **I** is the identity matrix of size *A*×*A*, **w** is the *A*-dimensional property vector of the atoms of a molecule, and *P* is the global property descriptor. It mainly depends on the kind of molecule atoms and not on the molecular structure.

An extension of this global property descriptor that combines chemical information contained in the atomic weightings and structural information is given by the autocorrelation descriptors.

Autocorrelation descriptors constitute a well-known set of molecular descriptors derived from a conceptual dissection of the molecular topology (or the 3D molecular geometry) and taking into account chemical information by specified weights of the molecule atoms.

The most known autocorrelation descriptors are those defined by Moreau-Broto,[19−21] called Autocorrelation of a

Topological Structure (*ATS*), and recently revisited by Wagener *et al.*[22] The spatial autocorrelation is evaluated by considering separately all the contributions of each different path length (*lag*) in the molecular graph, as collected in the topological distance matrix. In other words, the total spatial autocorrelation at *lag k* is obtained by summing all the products $w_i \cdot w_j$ of all the pairs of atoms *i* and *j*, for which the topological distance equals the *lag*.

Autocorrelation descriptors calculated for 3D-spatial molecular geometry are based on interatomic distances collected in the geometry matrix **G**. In this case, the variable *x*, which is an interatomic distance *r*, is divided into elementary distance intervals of 0.5 Å, and all the interatomic distances falling in the same interval are considered identical.[21] The autocorrelation functions at *lag k* are obtained by summing all the products $w_i \cdot w_j$ of all the pairs of atoms *i* and *j*, for which the interatomic distance r_{ij} falls within the considered interval $[x, x + 0.5]_k$.

Mathematically, the global spatial autocorrelation of the considered molecule is calculated substituting the identity matrix in eq 5 by the unity matrix **U** (a matrix with all elements equal to one) as

$$ATS = \mathbf{w}^{\mathbf{T}} \cdot \mathbf{U} \cdot \mathbf{w} = (\sum_{i=1}^{A} w_i)^2 = \sum_{i=1}^{A} w_i^{\,2} + 2 \cdot \sum_{i=1}^{A-1} \sum\nolimits_{j>i} w_i \cdot w_j = ATS_0 + 2 \cdot \sum_{k=1}^{d} ATS_k \quad (6)$$

where ATS_0 is the zero-order Moreau-Broto autocorrelation descriptor, ATS_k is the higher-order Moreau-Broto autocorrelation descriptor, representing the interactions between atoms at topological distance *k*, and *d* is the topological diameter, i.e., the maximum topological distance in the molecule. Note that the common notation to represent the Moreau-Broto autocorrelation descriptors is

$$ATS_0 = \sum_{i=1}^{A} w_i^{\,2}$$

$$ATS_k = \sum_{i=1}^{A-1} \sum\nolimits_{j>i} w_i \cdot w_j \cdot \delta(k;d_{ij}) \qquad k = 1,2,\ldots,d \quad (7)$$

where d_{ij} is the topological distance between atoms *i* and *j*, and $\delta(k; d_{ij})$ is a Dirac-delta function defined as

$$\delta(k;d_{ij}) = \begin{Bmatrix} 1 & if\, d_{ij} = k \\ 0 & if\, d_{ij} \neq k \end{Bmatrix}$$

In analogy with the Moreau-Broto autocorrelation descriptors, the new *HATS indices* have been defined weighting each atom of the molecule by

$$w'_i = w_i \cdot h_{ii}$$

thus obtaining the weight vector **w′**. They are calculated as

$$HATS_0(w) = \sum_{i=1}^{A} (w_i \cdot h_{ii})^2$$

$$HATS_k(w) = \sum_{i=1}^{A-1} \sum\nolimits_{j>i} (w_i \cdot h_{ii}) \cdot (w_j \cdot h_{jj}) \cdot \delta(k;d_{ij}) \qquad k = 1,2,\ldots,d \quad (8)$$

Formally similar to the Moreau-Broto autocorrelations, these descriptors also take into account the 3D molecular geometry by using the leverage values as the atomic weightings. Note that the $HATS_k$(u) descriptors calculated by unit weights w_i correspond to the ATS_k(h) descriptors calculated by using the leverage values as the atomic weightings. Moreover, the $HATS_1$(u) descriptor is a bond-additive index given by the sum over all bonds of the products of the leverages of bonded atoms. As can be observed in Table 2, it seems to be sensitive to the molecular branching and cyclicity.

The *HATS total index* is obtained by summing the $HATS_k$ indices for increasing values of *k*

$$HATS(w) = (\mathbf{w}')^{\mathbf{T}} \cdot \mathbf{U} \cdot \mathbf{w}' = \sum_{i=1}^{A} (w_i \cdot h_{ii})^2 + 2 \cdot \sum_{i=1}^{A-1} \sum\nolimits_{j>i} w_i \cdot h_{ii} \cdot w_j \cdot h_{jj} = HATS_0(w) + 2 \cdot \sum_{k=1}^{d} HATS_k(w) \quad (9)$$

where **U** is an unit matrix of size *A*×*A*. For each weighting scheme **w**, an ordered sequence of $HATS_k(w)$ descriptors plus a total autocorrelation index *HATS*(*w*) can be calculated for each molecule.

By fixing an upper value *L* of the topological distance, uniform length descriptors suitable for similarity/diversity analysis and property modeling of a set of compounds can be calculated.

The weights used in this work are those previously proposed for the calculation of the WHIM descriptors,[23,24] i.e., atomic mass (m), atomic polarizability (p), atomic electronegativity (e), van der Waals atomic volume (v), plus the unit weight (u). The electrotopological weight is not used; moreover, all the weights are scaled on the corresponding carbon atom value (Table 3).

HATS indices are based only on the diagonal elements of the molecular influence matrix (MIM), which account for the relative position of each atom in the 3D molecular space. To consider also the MIM off-diagonal elements which provide information on the degree of interaction between atom pairs, the *H indices* have been defined, modifying the Moreau-Broto autocorrelations (7) in the following way

$$H_0(w) = \sum_{i=1}^{A} h_{ii} \cdot w_i^{\,2}$$

$$H_k(w) = \sum_{i=1}^{A-1} \sum\nolimits_{j>i} h_{ij} \cdot w_i \cdot w_j \cdot \delta(k;d_{ij};h_{ij}) \qquad k = 1,2,\ldots,d \quad (10)$$

where d_{ij} is the topological distance between atoms *i* and *j* and *d* is the topological diameter. The function $\delta(k; d_{ij}; h_{ij})$ is a Dirac-delta function defined as

$$\delta(k;d_{ij};h_{ij}) = \begin{Bmatrix} 1 & if\, d_{ij} = k & and & h_{ij} > 0 \\ 0 & if\, d_{ij} \neq k & or & h_{ij} \leq 0 \end{Bmatrix}$$

Table 3. Unscaled and Scaled Values of the Atom Weights Used for GETAWAY Descriptor Calculation

	atomic mass		VdW volume		electronegativity		polarizability	
ID	m	m/m(C)	v	v/v(C)	e	e/e(C)	p	p/p(C)
H	1.01	0.084	6.709	0.299	2.592	0.944	0.667	0.379
B	10.81	0.900	17.875	0.796	2.275	0.828	3.030	1.722
C	12.01	1.000	22.449	1.000	2.746	1.000	1.760	1.000
N	14.01	1.166	15.599	0.695	3.194	1.163	1.100	0.625
O	16.00	1.332	11.494	0.512	3.654	1.331	0.802	0.456
F	19.00	1.582	9.203	0.410	4.000	1.457	0.557	0.316
Al	26.98	2.246	36.511	1.626	1.714	0.624	6.800	3.864
Si	28.09	2.339	31.976	1.424	2.138	0.779	5.380	3.057
P	30.97	2.579	26.522	1.181	2.515	0.916	3.630	2.063
S	32.07	2.670	24.429	1.088	2.957	1.077	2.900	1.648
Cl	35.45	2.952	23.228	1.035	3.475	1.265	2.180	1.239
Fe	55.85	4.650	41.052	1.829	2.000	0.728	8.400	4.773
Co	58.93	4.907	35.041	1.561	2.000	0.728	7.500	4.261
Ni	58.69	4.887	17.157	0.764	2.000	0.728	6.800	3.864
Cu	63.55	5.291	11.494	0.512	2.033	0.740	6.100	3.466
Zn	65.39	5.445	38.351	1.708	2.223	0.810	7.100	4.034
Br	79.90	6.653	31.059	1.384	3.219	1.172	3.050	1.733
Sn	118.71	9.884	45.830	2.042	2.298	0.837	7.700	4.375
I	126.90	10.566	38.792	1.728	2.778	1.012	5.350	3.040

The *H* indices have been defined following the basic principles of the spatial autocorrelation as above; however, for a given *lag* (i.e. topological distance) the product of the atom weightings is multiplied by the corresponding MIM value h_{ij} and only those contributions with a positive MIM value are considered. This means that, for a given atom *i*, only those atoms *j* at topological distance d_{ij} with a positive h_{ij} value have the chance to interact with the *i*th atom.

The terms, H_1, H_2, ..., H_d, represent autocorrelation quantities of *lag* 1, 2, ..., *d*, weighted by the molecular influence matrix. The maximum *lag* coincides with the molecule topological diameter *d*, i.e., with the maximum topological distance in the molecule.

For each weighting scheme, the *H total index* is obtained as

$$HT(w) = \mathbf{w^T}\cdot(\mathbf{H}\otimes\mathbf{B}^+)\cdot\mathbf{w} = \sum_{i=1}^{A} h_{ii}\cdot w_i^{\,2} + 2\cdot\sum_{i=1}^{A-1}\sum\nolimits_{j>i} h_{ij}\cdot w_i\cdot w_j = H_0(w) + 2\cdot\sum_{k=1}^{d} H_k(w) \tag{11}$$

where **w** is an atomic property vector, **H** is the molecular influence matrix, and $\mathbf{B}^+$ is a sparse binary matrix of size $A\times A$ whose elements $i-j$ are equal to one if corresponding to positive h_{ij} values, and zero otherwise; the symbol • indicates the Hadamard matrix product.

It can be observed that the indices H_0(u) and *HATS*(u), where u is the unit weight, are trivially related to the rank *D* of the molecular matrix **M** and should be excluded from every correlation analysis.

In conclusion, for each **w** vector of chemical weights, the following autocorrelation H-GETAWAY descriptors have been defined

$$\langle HATS, HATS_0, HATS_1, HATS_2, ..., HATS_L\rangle_w$$

$$\langle HT, H_0, H_1, H_2, ..., H_L\rangle_w$$

where *L* is a user-defined value for the maximum *lag*. It is chosen so as to obtain uniform length descriptors for a set of molecules and usually depends on the average size of the molecules in the data set. Obviously, for each *lag* greater than the topological diameter of the considered molecule ($k > d$), the corresponding autocorrelation terms are set to zero (i.e. $HATS_k = 0$ and $H_k = 0$).

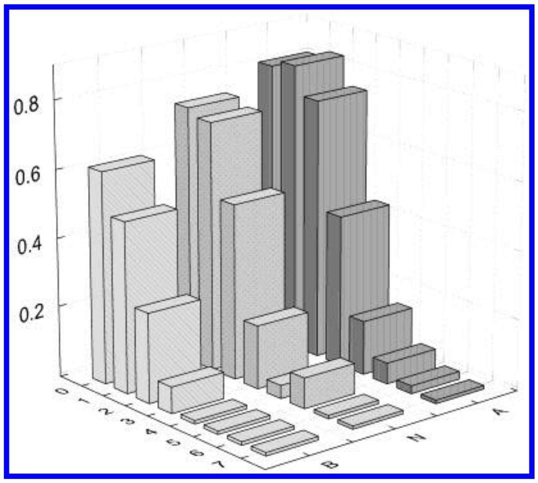

Figure 3. *H* profiles weighted by atomic van der Waals volumes (v), *lag* from 0 to 7, for benzene (B), naphthalene (N), and anthracene (A).

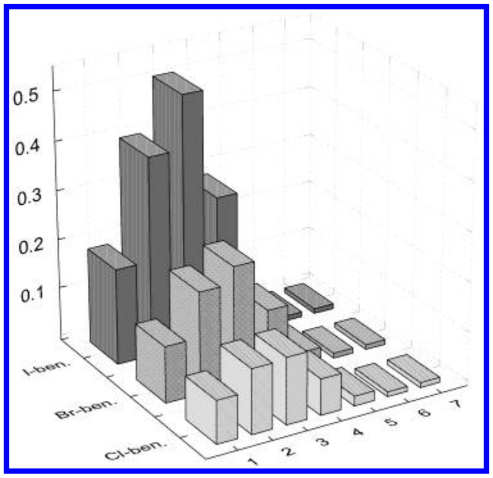

Figure 4. Profiles of mass-weighted autocorrelation indices *HATS*(m) for chlorobenzene, bromobenzene, and iodobenzene. *Lag* values 1−7.

HATS and *H* indices are molecular descriptors for structure−property correlations, but they can also be used as molecular profiles suitable for similarity/diversity analysis studies. These molecular profiles can be used all together or separately, and there is no need to take all the weights into account. In Figure 3, a simple comparison among benzene, naphthalene, and anthracene is shown. The *H* indices, from *lag* 0 to *lag* 7, weighted by the van der Waals volume (v) are used to perform the comparison. Figure 4 shows a comparison between chloro-, bromo-, and iodobenzene using the *HATS* profiles, from *lag* 1 to *lag* 7, weighted by the atomic masses (m). Finally, Figure 5 shows the comparison between Moreau-Broto *ATS* and *HATS* descriptors calculated for chlorobenzene, from *lag* 1 to *lag* 7, weighted by the atomic masses (m). As can be noted, the leverage values included in *HATS* indices influence both the trend along the *lag* axis and the absolute scale of the descriptors, performing a nonuniform smoothing dependent on the 3D relative atom location in the molecule.

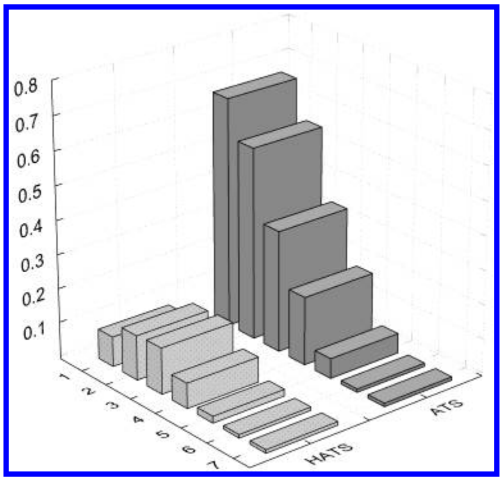

Figure 5. *HATS* and *ATS* profiles weighted by atomic masses (m), *lag* from 1 to 7, for chlorobenzene.

Local Vertex Invariants (LOVIs). The simple product **H·w** produces an *A*-dimensional vector of local vertex invariants containing a great amount of information related to the chemical identity and physicochemical properties of the atoms, their location in the molecular space, and the degree of accessibility to intra- and intermolecular interactions. They are real numbers able to uniquely characterize the atoms in a molecule and therefore suitable for canonical numbering of molecule atoms as well as for calculating all the already proposed molecular descriptors derived from local vertex invariants.[25−28] Moreover, leverage values have already been proposed as a tool for identifying the center of a molecular graph.[29]

R-GETAWAY DESCRIPTORS

Different molecular descriptors can be defined following the same approach as the H-GETAWAY descriptors by substituting the molecular influence matrix **H** by other kinds of molecular matrices. In particular, a new matrix here defined is based on both the matrix **H** and the geometry matrix **G**, whose elements r_{ij} are the 3D geometric distances between each pair of atoms *i* and *j*, **G** being a symmetric $A \times A$ matrix. From the geometry matrix and the reciprocal geometry matrix, several molecular descriptors have already been calculated. For example, the *gravitational indices* G_1 and G_2[30] were defined as the following

$$G_1 = \sum_{i=1}^{A-1} \sum\nolimits_{j>i} \frac{\mathrm{m}_i \cdot \mathrm{m}_j}{{r_{ij}}^2} \qquad G_2 = \sum_{b=1}^{B} \left(\frac{\mathrm{m}_i \cdot \mathrm{m}_j}{{r_{ij}}^2} \right)_b$$

where *A* and *B* are the number of atoms and bonds in the molecule, respectively, m is the atomic mass, and *r* is the geometric interatomic distance. The product of the masses of two atoms is divided by the square of their interatomic distance in order to make less significant contributions from pairs of atoms far apart, according to the basic idea that interactions between atoms in the molecule decreases as their distance increases.

The *influence/distance matrix* **R** is the new symmetric $A \times A$ molecular matrix here proposed whose elements resemble the single terms in the sums of the gravitational indices, defined as the following

$$[\mathbf{R}]_{ij} \equiv \left[\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} \right]_{ij} \quad i \neq j \qquad (12)$$

where h_{ii} and h_{jj} are the leverages of the two considered atoms and r_{ij} is their geometric distance. The diagonal elements of the matrix **R** are zero, while each off-diagonal element $i-j$ is calculated by the ratio of the geometric mean of the corresponding *i*th and *j*th diagonal elements of the matrix **H** to the interatomic distance r_{ij} provided by the geometry matrix **G**. Obviously, the largest values of the matrix elements derive from the most external atoms (i.e. with high leverages) and simultaneously next to each other in the molecular space (i.e. small interatomic distance).

The row sums of the influence/distance matrix encode some useful information that could be related to the presence of significant substituens or fragments in the molecule. In fact, it has been observed that larger row sums correspond to terminal atoms that are located very next to other terminal atoms such as those in substituents on a parent structure. Consequently, the average row sum of the influence/distance matrix (*RARS*) has been proposed as the first R-GETAWAY descriptor, defined as the following

$$RARS = \frac{1}{A} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} = \frac{1}{A} \cdot \sum_{i=1}^{A} RS_i \qquad (13)$$

where *A* is the number of atoms in the molecule and RS_i is the *i*th row sum.

Moreover, another descriptor proposed here is an index referred to *R-connectivity index* (*RCON*) because it is defined in analogy with the Randic connectivity index[31] based on the vertex degrees derived as the row sums of the adjacency matrix. It is a bond-additive index calculated by summing the squared root products of the **R** matrix row sums for all pairs of adjacent vertices

$$RCON = \sum_{b=1}^{B} (RS_i \cdot RS_j)_b^{1/2} \qquad (14)$$

where the sum runs over all bonds in the molecule and RS_i and RS_j indicate the row sums of two adjacent vertices. Note that the simple squared root is used instead of the inverse squared root of the Randic connectivity index since in the case of the influence/distance matrix the terminal vertices have larger row sums and thus make a larger contribution to *RCON* index than centrally located ones. In other words, the function used to weight bonds has been chosen so as to give more weight to more external bonds and less weight to internal shielded bonds.

The third R-GETAWAY descriptor has been defined in analogy with the Lovasz-Pelikan index[32] that is an index of molecular branching calculated as the first eigenvalue of the adjacency matrix. Therefore, the first eigenvalue of the influence/distance matrix (*REIG*) has been calculated and is here proposed as a molecular descriptor since it seems to have some structural interpretation; it has larger values for more branched molecules.

Calculated values of *RARS*, *RCON*, and *REIG* indices for a small set of diverse compounds are collected in Table 2. It can be noted that the *RCON* index is very sensitive to the

Table 4. GETAWAY Descriptors[a]

symbol	descriptor	geometry	topology	atoms	eq
H_{GM}	geometric mean on the leverage magnitude	H	*n*	*n*	(2)
I_{TH}	total information content on the leverage equality	H	*n*	*n*	(3)
I_{SH}	standardized information content on the leverage equality	H	*n*	*n*	(3)
HIC	mean information content on the leverage magnitude	H	*n*	*n*	(4)
$HATS_0$, $HATS_1$, ...	*HATS* indices	H	*y*	*y*	(8)
HATS	*HATS* total index	H	*y*	*y*	(9)
H_0, H_1, ...	*H* indices	H	*y*	*y*	(10)
HT	*H* total index	H	*y*	*y*	(11)
RARS	*R* matrix average row sum	R	*n*	*n*	(13)
RCON	*R*-connectivity index	R	*y*	*n*	(14)
REIG	first eigenvalue of the *R* matrix	R	*n*	*n*	
R_1, R_2, ...	*R* indices	R	*y*	*y*	(15)
RT	*R* total index	R	*y*	*y*	(16)
R_1^+, R_2^+, ...	maximal *R* indices	R	*y*	*y*	(17)
RT^+	maximal *R* total index	R	*y*	*y*	(18)

[a] H and R represent the *molecular influence matrix* and the *influence/distance matrix*, respectively; *n* and *y* represent *no* and *yes* dependence on topology via *lag* and on atom via atom weightings.

molecular size as well as to conformational changes and cyclicity. *RARS* and *REIG* indices are closely related; their values decrease as the molecular size increases and seem to be a little more sensitive to molecular branching than to cyclicity and conformational changes.

In analogy with the H-GETAWAY descriptors derived from the molecular influence matrix **H**, the autocorrelation R-GETAWAY descriptors have been defined based on the influence/distance matrix **R**. Namely, following the same definition scheme as the *H* indices (10), the *R indices* have been defined as

$$R_k(w) = \sum_{i=1}^{A-1} \sum_{j>i} \frac{\sqrt{h_{ii}\cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(k;d_{ij}) \quad k = 1,2,\ldots,d \qquad (15)$$

where $R_k(w)$ is the *w*-weighted *k*th order autocorrelation index, d_{ij} is the topological distance between atoms *i* and *j*, *d* is the topological diameter, and δ is the delta Dirac function defined as above. As the **R** diagonal terms are equal to zero, the term $R_0(w)$ has not to be considered. Note that the R_1(u) index is a bond-additive index very similar to the gravitational index G_2; its values for a small set of compounds are shown in Table 2

The corresponding *R total index* is

$$RT(w) = \mathbf{w}^{\mathbf{T}}\cdot\mathbf{R}\cdot\mathbf{w} = 2\cdot\sum_{i=1}^{A-1} \sum_{j>i} \frac{\sqrt{h_{ii}\cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j = 2\cdot\sum_{k=1}^{d} R_k(w) \qquad (16)$$

where **w** is the atomic property vector and *d* is the topological diameter. In the case of unit weights, i.e., *w* = u, the *R total index* is twice the Wiener-type index derived from the influence-distance matrix as the half-sum of all the matrix elements. Moreover, it is strictly related to the gravitational index G_1. In Table 2, values of the *RT*(u) index are collected for a small set of compounds; for linear molecular structures, the values of the *RT*(u) index are not so large as the contributions to the autocorrelation involving high leverage atoms far apart are smoothed by their interatomic distances.

To take into account local aspects of the molecule, from eq 15 the maximal contribution to the autocorrelation at each *lag* (i.e. topological distance) has also been proposed as a molecular descriptor

$$R_k^+(w) = \max_{ij}\left(\frac{\sqrt{h_{ii}\cdot h_{jj}}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(k;d_{ij})\right) \qquad i \neq j \text{ and } k = 1,2,\ldots,d \quad (17)$$

where R_k^+ (*w*) is the *w*-weighted *k*th order maximal *R* index.

Moreover, the maximum value among all the *k*th order maximal indices R_k^+ (*w*) is called *maximal R total index* defined as

$$RT^+(w) = \max{}_k(R_k^+(w)) \qquad (18)$$

Therefore, for each weighting scheme **w**, the following ordered sequences of molecular descriptors have been defined

$$\langle RT, R_1, R_2,\ldots, R_L\rangle_w \langle RT^+, R_1^+, R_2^+,\ldots, R_L^+\rangle_w$$

where *L* is a user-defined upper value of the topological distance.

The first terms R_1 and R_2 are expected to have a lower dependence on conformational changes as encoding information on pairs of atoms very near each other.

The complete list of the proposed GETAWAY descriptors is reported in Table 4.

If N_W is the number of weights and *L* is the user-defined *lag* upper value, the total number N of GETAWAY descriptors is given by

$$\mathrm{N} = 7 + 4\cdot\mathrm{N_W} + 2\cdot\mathrm{N_W} + 4\mathrm{L}\cdot\mathrm{N_W} = 7 + 2\cdot\mathrm{N_W}\cdot(3 + 2\mathrm{L})$$

where 7 is the number of single descriptors; for example, if $N_W = 5$ and $L = 8$, N = 197. The time required to calculate on a Windows/PC 800 MHz 197 GETAWAY descriptors for a small acyclic molecule of 17 atoms (i.e. neopentane) is 0.16 s and for a cyclic molecule of 24 atoms (i.e. anthracene) is 0.38 s.

STRUCTURE−PROPERTY CORRELATIONS

To have a deeper insight into the modeling power of these new molecular descriptors some regressions of selected

Table 5. Statistical Parameters for the Best 1, 2, and 3 Variables Regressions of Some Physicochemical Properties of Octane Isomers[a]

property	approach	size	Q^2_{LOO}	R^2	s	model descriptors	ref
boiling point (BP)	getaway + whim + top.	3	98.12	98.78	0.744	${}^2\chi$ ${}^2\bar{\chi}$ **$HATS_6$(p)**	
	getaway	3	97.10	98.32	0.897	**$HATS_2$(v) R_4(u) R_6(v)**	
	getaway + whim + top.	2	96.62	97.58	1.013	${}^2\chi$ **$HATS_6$(p)**	
	topological	3		95.84	1.394	S^3W S^4W SJ	37
	topological	2		94.78	1.508	S^3W S^4W	37
	getaway	2	84.86	89.62	2.098	**$HATS_2$(m) $R^+{}_4$(u)**	
	topological	2		81.36	2.810	WW x_1	35
	topological	1		78.85	2.900	Z	34
	getaway + whim + top.	1	66.47	74.64	3.175	**$HATS_2$(m)**	
	topological	1		67.77	3.630	χ^1W	37
motor octane number (MON)	getaway + whim + top.	3	98.58	99.23	2.439	${}^V\bar{I}_D{}^M$ Ts **$HATS_1$(m)**	
	getaway	3	97.42	98.62	3.259	**$HATS_4$(u) $HATS_7$(v) R_7(p)**	
	topological	3		98.05	3.855	$S\chi^1W$ χ^7W χ^3W	37
	getaway + whim + top.	2	96.77	97.68	4.053	Ts **H_4(e)**	
	getaway	2	91.28	95.78	5.466	**$HATS_7$(m) R_4(u)**	
	topological	2		95.64	5.533	$S\chi^1W$ $S\chi^3W$	37
	topological	1		95.22	5.589	χ^7W	37
	getaway + whim + top.	1	90.83	92.40	7.069	Ts	
	topological	1		91.97	7.270	I_{WD}	34
	getaway	1	85.64	88.98	8.515	**REIG**	
heat of vaporization (HV)	getaway + whim + top.	3	97.57	98.42	0.281	${}^0\bar{\chi}$ ${}^3\kappa$ **$R^+{}_6$(u)**	
	getaway	3	95.46	97.18	0.375	**$HATS_6$(u) R_4(u) $R^+{}_1$(m)**	
	getaway + whim + top.	2	95.18	96.53	0.402	${}^2\chi$ **$R^+{}_6$(u)**	
	topological	3		95.65	0.459	χ^1W χ^2W χ^3W	37
	getaway	2	93.15	94.87	0.488	**$HATS_4$(u) R_6(e)**	
	topological	2		92.62	0.577	${}^4W$ ${}^5W$	37
	topological	1		91.78	0.429	Z	34
	getaway + whim + top.	1	80.80	88.61	0.705	${}^2\chi$	
	getaway	1	79.74	85.70	0.790	**R_2(m)**	
	topological	2		84.27	0.820	WW x_1	35
molar volume (MV)	getaway + whim + top.	3	75.96	92.01	1.825	Ks **$R^+{}_6$(u) RT^+(m)**	
	getaway	3	69.27	90.33	2.008	**$HATS_6$(p) RT^+(m) R_1(v)**	
	topological	3		88.29	2.210	${}^5W$ ${}^6W$ ${}^7W$	37
	getaway + whim + top.	2	54.49	84.96	2.419	${}^V\bar{I}_D^M$ **$R^+{}_6$(u)**	
	getaway	2	45.49	81.79	2.662	**$R^+{}_6$(u) R4(v)**	
	getaway + whim + top.	1	32.66	67.61	3.437	**R_6(v)**	
	topological	2		62.76	3.807	${}^3W$ ${}^4W$	37
	topological	1		60.85	3.780	${}^7W$	37
entropy (S)	getaway + whim + top.	3	97.17	97.96	0.711	${}^V\bar{I}^E_{D,\mathrm{deg}}$ TWC **$R^+{}_2$(p)**	
	getaway + whim + top.	2	96.42	97.14	0.814	${}^V\bar{I}^E_{D,\mathrm{deg}}$ TWC	
	getaway	3	93.45	95.84	1.016	**I_{SH} $HATS_8$(m) R_3(v)**	
	getaway	2	92.19	94.76	1.101	**I_{SH} R_3(v)**	
	getaway + whim + top.	1	89.86	92.51	1.274	**R_3(v)**	
	topological	1		91.10	1.400	$\chi^{[1/2]}$	34
	topological	2		81.72	2.060	x_1 x_2	35
heat of formation (Δ_fH)	getaway + whim + top.	3	95.06	96.60	0.254	**$HATS_5$(m) $HATS_7$(m) R_4(e)**	
	getaway + whim + top.	2	90.96	93.24	0.346	${}^2\chi$ **$HATS_2$(e)**	
	getaway	2	90.18	92.87	0.356	**$HATS_7$(u) R_2(m)**	
	getaway + whim + top.	1	87.18	89.34	0.421	**$HATS_2$(m)**	
	topological	3		87.05	0.492	Ω_1 Ω_2 Ω_3	33
	topological	2		86.86	0.478	Ω_1 Ω_2	33
	topological	1		86.68	0.471	$1/{}^2\chi$	34
	topological	2		78.70	0.570	WW x_1	35

[a] Q^2_{LOO}: *leave-one-out* cross-validated explained variance; R^2: determination coefficient; *s*: standard estimate of the error. Molecular descriptors not cited in the text are ${}^0\bar{\chi}$, ${}^2\chi$, and ${}^2\bar{\chi}$ (Kier-Hall connectivity indices[54,55]); S^3W, S^4W, SJ, $S\chi^1W$, and $S\chi^3W$ (SP indices[37]); WW (hyper-Wiener index[56]); x_1 and x_2 (first and second eigenvalues of the Wiener matrix[35]); Z (Hosoya Z index[57]); χ^1W, χ^2W, χ^3W, and χ^7W (walk connectivity indices[58]); ${}^V\bar{I}_D{}^M = I_{WD}$ (mean information content on the distance magnitude[59]); Ts and Ks (WHIM descriptors[24]); ${}^3\kappa$ (3-path alpha-modified Kier shape index[60]); ${}^3W$, ${}^4W$, ${}^5W$, ${}^6W$, and ${}^7W$ (counts of walks of different length in the molecular graph[37]); ${}^V\bar{I}^E_{D,\mathrm{deg}}$ (mean information content on the distance degree equality[59]); TWC (total walk count[29]); $\chi^{[1/2]}$ (Altenburg's $p = 1/2$ index[61]); Ω_1, Ω_2, and Ω_3 (orthogonal connectivity indices[33]).

physicochemical properties of octanes have been searched for. Comprehensive studies of numerous physicochemical properties of the 18 octane isomers have been already published,[33−37] based on the idea that the properties of a series of isomeric compounds, being independent of molecular size, are suitable for the first evaluation of the modeling capabilities of different mathematical descriptors. If a new proposed molecular descriptor is not able to model the variation of at least one property of octanes, then it probably does not contain any useful molecular information. Moreover, octanes constitute a good set of compounds for comparative study, since many experimental data among their physicochemical properties are available.

The properties of the octane isomers studied in this work are boiling point (BP), motor octane number (MON), heat of vaporization (HV), molar volume (MV), entropy (S), and heat of formation (Δ_fH). Their experimental values have been taken from refs 35 and 37.

Molecular Descriptors. The molecular descriptors used to search for the best regressions of the selected physicochemical properties of octanes were calculated by the *Dragon* program.[38] *Dragon* is a new free available software for molecular descriptors calculation, able to provide more than 800 descriptors. These include, together with the constitutional descriptors and the traditional topological and information indices, various geometrical descriptors such as 3D-Wiener index,[2,3] folding degree index,[39,40] radius of gyration,[41,42] span,[42] spherosity index,[43] and asphericity.[44] Moreover, 3D-Morse,[45,46] Randic molecular profiles,[7,9] Moreau-Broto autocorrelations,[19−21] WHIM,[23,24] Galvez topological charge indices,[47,48] and BCUT descriptors[10,11] are also calculated by *Dragon*. For each block of descriptors, principal components can be calculated, automatically selecting as the most significant components those with eigenvalues greater than one. *Dragon* can accept as the input files for molecular geometries both HyperChem files (.hin), Sybyl files (.mol-,.mol2), and SD files (.sdf). In this work, the molecular geometries of octanes have been optimized by *HyperChem* package[49] (PM3 semiempirical method).

Variable Selection. Commonly used over the years, the Genetic Algorithm − Variable Subset Selection (GA-VSS) method[50,51] has been adopted to search for the best linear regressions, optimizing the prediction power Q^2_{LOO} (*leave-one-out* procedure). This algorithm provides the subsets of the most predictive molecular descriptors for the selected property, automatically chosen among all the available descriptors. It can be applied also to a large set of molecular descriptors. The software *MobyDigs/Evolution*,[52] developed by our research group, has been used to perform the variable selection. To avoid chance correlation, the QUIK rule[53] has been adopted, thus only the models with a K multivariate correlation calculated on the X+Y block of the 5% greater than the K correlation of the X-block have been considered statistically significant. The K correlation index measures the correlation of a set of variables taking into account all the variables together, instead of single pairs of variables.

Comparative Study. Regressions of octane properties based on the GETAWAY descriptors have been compared to some regressions based on topological descriptors taken from the literature. Precisely, to evaluate the quality of the models based on our new descriptors we have taken as the reference the models published by Randic[33−35] based on diverse topological indices such as the Wiener matrix invariants and those published by Diudea[37] based on the SP indices. Moreover, to the set constituted by the topological (69), WHIM (99), and GETAWAY descriptors (197) calculated by our program *Dragon* the GA-VSS method has been applied in order to search for the best models.

For each selected property of octanes, the statistical information for the best regressions with 1, 2, and 3 molecular descriptors has been reported in Table 5. Together with the leave-one-out cross-validated explained variance (Q^2_{LOO}), the determination coefficient (R^2) and the standard estimate of the error (s) are listed. The regressions have been sorted in such a way to have decreasing values of R^2, since the cross-validated explained variance Q^2_{LOO} is not available for all the models. The GETAWAY descriptor symbols are reported in boldface, and the last column contains the bibliographic references of the models taken from the literature.

For all the studied properties, GETAWAY descriptors seem to give satisfactory results. It can be easily seen in Table 5 that the models based on the GETAWAY descriptors are in almost all the cases better than those taken from the literature. Moreover, the models calculated by applying the selection procedure to the set given by GETAWAY descriptors plus WHIM and topological indices contain at least one of the new proposed molecular descriptors.

According to the obtained QSPR results, it is possible to conclude that the new descriptors encode some useful molecular information different from that of previous proposed descriptors. Moreover, they are quite diverse among themselves being able to describe well the variation of different properties of octanes. Note also that in two regressions for the entropy the descriptor I_{SH}, encoding information on the molecular symmetry, has been selected by the GA-VSS method as it was expected.

CONCLUDING REMARKS

GETAWAY descriptors are mathematical quantities calculated from two new molecular representations depending on the molecular geometry: the molecular influence matrix and the influence/distance matrix. The new descriptors proposed here have been shown to have some interesting characteristics: (a) they contain 3D information, (b) their functional definitions are based on well-known and accepted algorithms and formulas (e.g. autocorrelations, connectivity indices, information theory concepts), (c) they can be easily and quickly calculated, (d) they show good prediction power in physicochemical property modeling, and (e) they give uniform length descriptors suitable for similarity/diversity analysis.

Despite these positive characteristics of GETAWAY descriptors, additional work has to be done to further investigate their meaning and behavior with respect to the structural features of the molecules. Applications of these new descriptors in molecular property modeling and similarity/diversity analysis will be published in subsequent papers.

ACKNOWLEDGMENT

The work was sponsored by the Commission of the European Union (R&D project “Beam”, EVK1-CT1999-00012).

REFERENCES AND NOTES

(1) *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997.

(2) Mekenyan, O.; Peitchev, D.; Bonchev, D.; Trinajstic, N.; Bangov, I. P. Modelling the Interaction of Small Organic Molecules with Biomacromolecules. I. Interaction of Substituted Pyridines with anti-3-azopyridine Antibody. *Arzneim. Forsch.* **1986**, *36*, 176−183.

(3) Bogdanov, B.; Nikolic, S.; Trinajstic, N. On the Three-Dimensional Wiener Number. *J. Math. Chem.* **1989**, *3*, 299−309.

(4) Mekenyan, O.; Peitchev, D.; Bonchev, D.; Trinajstic, N.; Dimitrova, J. Modelling the Interaction of Small Organic Molecules with Biomacromolecules. III. Interaction of Benzoates with anti-p-(p′-axophenylazo)-benzoate Antibody. *Arzneim. Forsch.* **1986**, *36*, 629−635.

(5) Bogdanov, B.; Nikolic, S.; Trinajstic, N. On the Three-Dimensional Wiener Number. A Comment. *J. Math. Chem.* **1990**, *5*, 305−306.

(6) Randic, M. Molecular Profiles. Novel Geometry-Dependent Molecular Descriptors. *New J. Chem.* **1995**, *19*, 781−791.

(7) Randic, M. Molecular Shape Profiles. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373−382.

(8) Randic, M.; Razinger, M. On Characterization of Molecular Shapes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 594−606.

(9) Randic, M. Quantitative Structure−Property Relationship − Boiling Points of Planar Benzenoids. *New J. Chem.* **1996**, *20*, 1001−1009.

(10) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. In *3D QSAR in Drug Design − Vol. 2*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, The Netherlands, 1998.

(11) Pearlman, R. S. Novel Software Tools for Addressing Chemical Diversity. *Internet Communication*; 1999; http://www.netsci.org/Science/Combichem/feature08.html.

(12) Atkinson, A. C. *Plots, Transformations, and Regression*; Clarendon Press: Oxford, U.K., 1985.

(13) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, U.K., 1988.

(14) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.

(15) Bonchev, D.; Polansky, O. E. On the Topological Complexity of Chemical Systems. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier: Amsterdam, The Netherlands, 1987.

(16) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: a QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609.

(17) Bonchev, D.; Seitz, W. A. The Concept of Complexity in Chemistry. In *Concepts in Chemistry: Contemporary Challenge*; Rouvray, D. H., Ed.; Research Studies Press: Taunton, U.K., 1996.

(18) Bonchev, D. Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res.* **1997**, *7*, 23−43.

(19) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.

(20) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures, Application to SAR Studies. *Nouv. J. Chim.* **1980**, *4*, 757−764.

(21) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66−70.

(22) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling *Corticosteroid Binding Globulin* and Cytosolic *Ah* Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(23) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D- and 3D-Structures. Theory. *J. Chemom.* **1994**, *8*, 263−273.

(24) Todeschini, R.; Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113−119.

(25) 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 Correlation Ability. *J. Math. Chem.* **1987**, *1*, 61−83.

(26) Balaban, A. T. Numerical Modelling of Chemical Structures: Local Graph Invariants and Topological Indices. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier: Amsterdam, The Netherlands, 1987.

(27) 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.

(28) Balaban, A. T. Local vs Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398−402.

(29) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(30) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(31) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(32) Lovasz, L.; Pelikan, J. On the Eigenvalue of Trees. *Period. Math. Hung*. **1973**, *3*, 175−182.

(33) Randic, M. Correlation of Enthalpy of Octanes with Orthogonal Connectivity Indices. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45−59.

(34) Randic, M. Comparative Regression Analysis. Regressions Based on a Single Descriptor. *Croat. Chem. Acta* **1993**, *66*, 289−312.

(35) Randic, M.; Guo, X.; Oxley, T.; Krishnapriyan, H.; Naylor, L. Wiener Matrix Invariants. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 361−367.

(36) Diudea, M. V. Walk Numbers $^{e}W_{M}$: Wiener-Type Numbers of Higher Rank. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 535−540.

(37) Diudea, M. V.; Minailiuc, O. M.; Katona, G. Molecular Topology. 26. SP Indices: Novel Connectivity Descriptors. *Rev. Roum. Chim.* **1997**, *42*, 239−249.

(38) Todeschini, R.; Consonni, V. *Dragon, rel. 1.12 for Windows*; Milano, Italy, 2001. Program for the calculation of molecular descriptors from HyperChem, Sybyl, and SD file formats. Free download at http://www.disat.unimib.it/chm/.

(39) Randic, M.; Kleiner, A. F.; DeAlba, L. M. Distance/Distance Matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−286.

(40) Randic, M.; Krilov, G. On a Characterization of the Folding of Proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.

(41) Tanford, C. *Physical Chemistry of Macromolecules*; Wiley: New York, 1961.

(42) Volkenstein, M. V. *Configurational Statistics of Polymeric Chains*; Wiley-Interscience: New York, 1963.

(43) Robinson, D. D.; Barlow, T. W.; Richards, W. G. Reduced Dimensional Representations of Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939−942.

(44) Arteca, G. A. Molecular Shape Descriptors. In *Reviews in Computational Chemistry − Vol. 9*; Lipkowitz, K. B., Boyd, D., Eds.; VCH Publishers: New York, 1991.

(45) Schuur, J.; Gasteiger, J. 3D-MoRSE Code − A New Method for Coding the 3D Structure of Molecules. In *Software Development in Chemistry − Vol. 10*; Gasteiger, J., Ed.; Fachgruppe Chemie-Information-Computer (CIC): Frankfurt am Main, Germany, 1996.

(46) Schuur, J.; Gasteiger, J. Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation. *Anal. Chem.* **1997**, *69*, 2398−2405.

(47) Gálvez, J.; Garcìa, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520−525.

(48) Gálvez, J.; Garcìa-Domenech, R.; De Julián-Ortiz, V.; Soler, R. Topological Approach to Drug Design. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272−284.

(49) *HyperChem, rel. 4 for Windows*; Autodesk Inc.: Sausalito, CA, 1995.

(50) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Massachusetts, MA, 1989.

(51) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267−281.

(52) Todeschini, R. *Moby Digs/Evolution, rel. 1.0 for Windows*; Talete srl: Milano, Italy, 2001. Software for multilinear regression analysis and variable subset selection by Genetic Algorithm.

(53) Todeschini, R.; Consonni, V.; Maiocchi, A. The K Correlation Index: Theory Development and its Applications in Chemometrics. *Chemom. Intell. Lab. Syst.* **1998**, *46*, 13−29.

(54) Kier, L. B.; Hall, L. H. The Nature of Structure-Acitivity Relationships and their Relation to Molecular Connectivity. *Eur. J. Med. Chem.* **1977**, *12*, 307−312.

(55) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press − Wiley: Chichester, U.K., 1986.

(56) Randic, M. Novel Molecular Descriptor for Structure−Property Studies. *Chem. Phys. Lett.* **1993**, *211*, 478−483.

(57) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

(58) Razinger, M. Discrimination and Ordering of Chemical Structures by the Number of Walks. *Theor. Chim. Acta* **1986**, *70*, 365−378.

(59) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K., 1983.

(60) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.

(61) Altenburg, K. Eine Bemerkung zu dem Randicschen “Molekularen Bindungs-Index”. *Z. Phys. Chem.* **1980**, *261*, 389−393.

CI015504A