Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201300173

Discrete Derivatives for Atom-Pairs as a Novel Graph-Theoretical Invariant for Generating New Molecular Descriptors: Orthogonality, Interpretation and QSARs/QSPRs on Benchmark Databases

Oscar Martínez-Santiago,[a] Reisel Millán-Cabrera,[a] Yovani Marrero-Ponce,*[a, d] Stephen J. Barigye,[a] Yoan Martínez-López,[a, e] Francisco Torrens,[b] and Facundo Pérez-Giménez[c]

Abstract: This report presents a new mathematical method based on the concept of the *derivative of a molecular graph* (G) with respect to a given *event (S)* to codify chemical structure information. The *derivate over each pair of atoms* in the molecule is defined as $\partial G/\partial S(v_i, v_j) = (f_i - 2f_{ij} + f_j)/f_{ij}$, where f_i (or f_j) and f_{ij} are the *individual frequency* of atom *i* (or *j*) and the *reciprocal frequency* of the atoms *i* and *j*, respectively. These frequencies characterize the participation intensity of atom pairs in ***S***. Here, the event space is composed of molecular sub-graphs which participate in the formation of the **G** skeleton that could be complete (representing all possible connected sub-graphs) or comprised of sub-graphs of certain orders or types or combinations of these. The *atom level graph derivative index*, Δ_i, is expressed as a linear combination of all atom pair derivatives that include the atomic nuclei *i*. Global [total or local (group or atom-type)] indices are obtained by applying the so called *invariants* over a vector of Δ_i values. The novel MDs are validated using a data set of 28 alkyl-alcohols and other *benchmark* data sets proposed by the International Academy of Mathematical Chemistry. Also, the boiling point for the alcohols, the adrenergic blocking activity of *N*,*N*-dimethyl-2-halo-phenethylamines and physicochemical properties of polychlorinated biphenyls and octanes are modeled. These models exhibit satisfactory predictive power compared with other 0–3D indices implemented successfully by other researchers. In addition, tendencies of the proposed indices are investigated using examples of various types of molecular structures, including chain-lengthening, branching, heteroatoms-content, and multiple bonds. On the other hand, the relation of atom-based derivative indices with ^{17}O NMR of a series of ethers and carbonyls reflects that the new MDs encode electronic, topological and steric information. Linear independence between the graph derivative indices and other 0-3D MDs is demonstrated by using principal component analysis on a dataset of 41 heterogeneous molecules. It is concluded that the graph derivative indices are independent indices containing important structural information to be used in QSPR/QSAR and drug design studies, and permit obtaining easier, more interpretable and robust mathematical models than the majority of those reported in the literature.

Keywords: Generalized incidence matrix · Frequency matrix · Event · Sub-graph · Invariant · Molecular descriptors · DIVATI · TOMOCOMD-CARDD · Genetic algorithm · QSPR

1 Introduction

In any process of molecular modeling (e.g., QSPR/QSAR studies, ligand-based virtual screening, and so on), the need for molecular structure representation is critical and its role is significant in finding appropriate predictive

[a] *O. Martínez-Santiago, R. Millán-Cabrera, Y. Marrero-Ponce, S. J. Barigye, Y. Martínez-López*
Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Facultad de Química y Farmacia,. Universidad Central "Marta Abreu" de Las Villas
Carretera a Camajuani Km 5 1/2, Santa Clara, 54830, Villa Clara, Cuba.
fax: 963543156; phone: 963543156
**e-mail: ymarrero77@yahoo.es*
ymponce@gmail.com

[b] *F. Torrens*
Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna
Polígon la Coma s/n, E-46071 Valencia, Spain

[c] *F. Pérez-Giménez*
Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València
Spain

[d] *Y. Marrero-Ponce*
Doctorado en Toxicología Ambiental, Facultad de Química Farmacéutica, Universidad de Cartagena
Cartagena de Indias, Bolívar, Colombia

[e] *Y. Martínez-López*
Department of Computer Sciences, Faculty of Informatics, Camaguey University
Camaguey City, 74650 Camaguey Cuba

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201300173.

models. An information-rich representation rapidly computed and readily manipulated is essential.[1] This is the case with the *so*-called topological (as well as topo-chemical) indices (TIs), which are among the most useful *molecular descriptors* (MDs) known nowadays.[2] The TIs are "*numerical values associated with chemical constitution for correlation of chemical structures with various physical properties, as well as chemical or biological activities*" and these are derived from graph-theoretical invariants.[3] That is, TIs are numbers calculated from a molecular graph representing a molecule, which does not depend on the numbering of the graph vertices or edges.

Several TIs have been introduced to date. A compilation by Todeschini and Consonni systematizes more than 1600 MDs for small-molecule drug discovery.[1] There are two main sources of TIs, the distance (**D**) and adjacency (**A**) matrices, but the number and diversity of graph invariants is so wide that this makes it difficult to find general relations for the indices so derived. However, some of these MDs are redundant or have certain communalities. For instance, many researchers define TIs from graphs by using vector-matrix-vector (VMV) procedures, a fact that indicates significant similarities between these systems.[4] Indeed, the first TI ever defined in a chemical context, the Wiener index (W) can be calculated by using the same mathematical formalism (VMV) and it is related to the invariants based on the sum of vertex degree products, e.g., Randić index;[5] although there is no apparent relation between these invariants. One of the present authors also proposed new MDs families by using a more elaborate approach in terms of the algebraic space; which from a matrix point of view can also be expressed like VMV.[4]

Nevertheless it is well-known that MDs defined to the moment do not behave satisfactorily in the solution of all the problems in which they are applied. It is for this reason that it continues to be necessary to find novel and more complete methods of coding chemical structural information.

In the present report, new TIs based on the concept of the molecular graph derivative are obtained from the generalization of the traditional incidence matrix. This matrix has not had much use in the definition of MDs for not being symmetrical. However, this matrix can also offer valuable information on the molecular structure. Also, we introduce the use of norms (distances), means and statistical invariants as an interesting way of obtaining local and total indices from atomic indices (local vertex invariants, LOVIs). Besides, in order to evaluate the performance of the proposed MDs in QSPR/QSAR studies, we model the physicochemical and biological properties of four benchmark datasets. Three of them have been proposed by the *International Academy of Mathematical Chemistry* to check the behavior of new MDs. Finally, other principal objectives of this paper are: 1) to investigate the most important characteristics of these novel indices by means of several structural changes in organic molecules, including chain-lengthening, branching, heteroatoms-content, and multiple bonds, and 2) to check if, the information contained in the total and local (atom and atom-type) derivative indices is different from that of other 0-3D MDs presently in use in QSPR/QSAR and drug design practice.

2 Preliminary Concepts on Discrete Derivative

As it is well-known in mathematical analysis, the derivative concept characterizes the degree of variation in a function on carrying out a small variation in its argument; this derivative concept is based on one for the limit. In discrete mathematics, the limit concept does not exist and, therefore, it is impossible to transfer the derivative concept like it is known, from continuous to discrete mathematics. Before proposing the definition of the derivative concept in discrete mathematics, let us first define certain important concepts.

To start with, we define an *event* (*S*), which is true when certain conditions of the examined process are fulfilled.[6] Every *S* determines a bi-dimensional binary matrix $\boldsymbol{Q}=[q_{ij}]_{m\times n}$, each column of which corresponds reciprocally to a *condition*, included in at least a true event, and every row, a collection of conditions, in which the *event* occurs (in which the event is true) and q_{ij} is equal to:[6]

1, if the *j*-th condition is included in the *i*th collection of conditions, in which the event is true.
0, otherwise

In other words, every ***S*** determines a model (ψ) for the *incidence matrix* ***Q***; the conditions included in the *event* are *letters* corresponding to the model and the *collection of conditions* in which the *event* is true would be the *words* for the model ψ. Therefore, it is important to introduce the *relations frequency matrix* $\boldsymbol{F}=[f_{ij}]_{n\times n}$ that characterizes the model ψ, with the incidence matrix $\boldsymbol{Q}(\psi)=[q_{ij}]_{m\times n}$.[6]

We denominate *relations frequency matrix* $\boldsymbol{F}=[f_{ij}]_{n\times n}$, one in which each row and column correspond reciprocally to a *condition*, and element f_{ij} is equal to the number of *words* that contain the *letters* *i* and *j*, respectively, if $i \neq j$. If $i=j$ then f_i corresponds to the number of *words* that contain *letter i*. The term f_i is known as the *individual frequency* of letter *i* and f_{ij} the *reciprocal frequency* of the letters *i* and *j*.

From the definition of the ***F***, one notices that it is symmetric with respect to the principal diagonal, that is $f_{ij}=f_{ji}$, and the individual frequency of each letter is greater than the reciprocal frequency of this letter with any other letter, $f_i \geq f_{ij}$. It can also be demonstrated that: $\boldsymbol{F}=\boldsymbol{Q}^{T}\times\boldsymbol{Q}$, $\boldsymbol{Q}^{T}$ being the transpose matrix of the incidence matrix [$\boldsymbol{Q}$ (ψ)] for the model ψ.[6]

We are, therefore, in condition of determining the heterogeneity grade of the graph's components with respect to a given event and we will characterize this heterogeneity

by means of the graph's derivative $\partial G/\partial S$ with respect to the event S.

The derivative $\partial G/\partial S$ (for duplexes or a pair of edges in this case) of a graph (G) with respect to an event (***S***) is a non-oriented weighted graph $<V,(U, P)>$, whose labels coincide with those of a model determined by this event and a pair of edges (e_i, e_j) is weighted by the frequency ratio $(f_i - f_{ij}) + (f_j - f_{ij})$ of its incompatible participation frequency to the participation frequency f_{ij} compatible to the event ***S***:

$$\frac{\partial G}{\partial S}(e_i, e_j) = \frac{(f_i - 2f_{ij} + f_j)}{f_{ij}} \quad (1)$$

with the particularity that

1) if $(e_i, e_j) \notin U$, then $(\partial G/\partial S)(e_i, e_j) = \infty$
2) if $(e_i, e_j) \in U$, then $(\partial G/\partial S)(e_i, e_j) =$ a finite magnitude different from zero
3) if $(e_i = e_j)$ then, $(\partial G/\partial S)(e_i, e_j) = 0$

Let us therefore illustrate the derivative concept of a graph with an example. Given **G** in Figure 1A, we would like to determine the participation frequency of the different edges in the formation of the graph skeletons. The **G** has 8 skeletons [sub-graphs of order 3, without differentiating the type (Figure 1, B)]. The required frequency can be determined, for example, by determining the number of inclusions of each edge in the skeletons. For example, the edge "a" participates 5 times in the formation of the skeletons, the edge "c" 4 times, etc. The required frequency can be better characterized, if in addition to the pair of previously indicated numbers, we determine numbers that characterize the non-uniform participation grade of graph edge

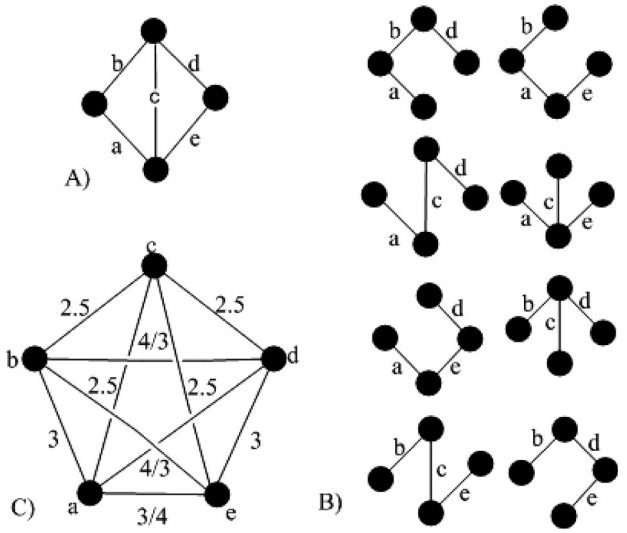


Figure 1. A) Molecular graph, B) sub-graphs (words) according to event S (connected sub-graphs of order 3 based on edge (letters) relations). C) Derivative graph.

pairs (graph derivative for a pair of elements), in the formation of the graph skeletons, from which we should obtain the corresponding incidence and frequency matrices for the model determined by our event (formation of the graph-skeleton by the different edges), and in this way calculate the derivative values $\partial G/\partial S$ for graph edge pairs. The incidence and frequency matrices, respectively, for this model, are:

$$Q = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix} \qquad F = Q^T x Q = \begin{Vmatrix} 5 & 2 & 2 & 3 & 3 \\ 2 & 5 & 2 & 3 & 3 \\ 2 & 2 & 4 & 2 & 2 \\ 3 & 3 & 2 & 5 & 2 \\ 3 & 3 & 2 & 2 & 5 \end{Vmatrix}$$

The elements for the matrix (***F***) determine the $\partial G/\partial S$, which is a weighted graph with labels.[7] It follows that two edges of this graph are adjacent, if the derivative value over the arc formed by these vertices is different from zero or infinity. The derivative values for the edge pairs of the graph are:

$$(\partial G/\partial S)(a, b) = 3,\ (\partial G/\partial S)(a, c) = 2.5,\ \ldots \rightarrow$$
$$\ldots,\ (\partial G/\partial S)(d, e) = 3$$

and with these values we can form the $(\partial G/\partial S)$ (Figure 1C).

As can be observed, to determine the graph's derivative, according to the event (S), it is necessary to:[6]

1) construct a model determined by a previously chosen event, S, which determines an incidence matrix, **Q**.
2) find the relations frequency matrix, **F**, corresponding to the model.
3) calculate the derivative values $(\partial G/\partial S)$ over a pair of elements (vertices or edges) of the graph.

Below, we will define two categories of derivatives that extend and generalize the derivative concept analogous to the development of the derivative concept when mathematical analysis is applied. Note that in this report we only apply the derivative concept for duplexes in the generation of novel MDs.

3 Theory of New Molecular Descriptors

3.1 New Atom-Relations: Extended Incidence Matrix

Let us take the molecule of 2,3-dimethylpentane as a simple example (see Figure 2), where the numbers correspond to the labels that are assigned to the carbon atoms (vertices) in the molecular structure and graph.

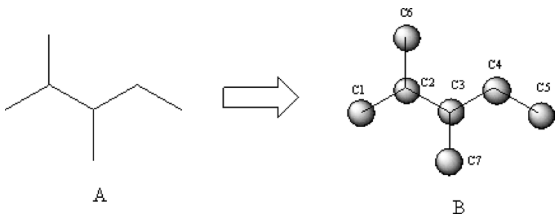


Figure 2. The chemical structure and molecular graph of [the numbers correspond to the labels that are assigned to the atoms (vertices) in the molecular structure]: A) 2,3-dimethylpentane (H-depleted structure), B) molecular graph of 2,3-dimethylpentane.

This graph is in correspondence with the chemical structure. In the same, the carbon atoms labeled C_1, C_2, C_3, C_4, C_5, C_6 and C_7 are represented as the **G** vertices.

Let us therefore define a new event *the formation of the molecular structure from the connected sub-structures (sub-graphs) of distinct orders and types*, based on atomic relations. Applying this event to the molecular structure of 2,3-dimethylpentane (see Figure 2B), the following sub-structures are obtained, organized according to their orders (see Table 1).

These sub-graphs are represented in an *incidence matrix* (***Q***), from which we obtain the corresponding *relations frequency matrix* (***F***). The number of inclusions of each vertex in the carbon skeletons permits us to establish the required frequencies.

For example, vertex 1 participates 15 times in the formation of the sub-graphs (see ***Q*** and ***F*** matrix for 2,3-dimethylpentane).

Table 1. Incidence matrix for 2,3-dimethylpentane.

Order	Type	Sub-graph	Incidence matrix (**Q**)						
			C_1	C_2	C_3	C_4	C_5	C_6	C_7
Order 0	paths	C1	1	0	0	0	0	0	0
	paths	C2	0	1	0	0	0	0	0
	paths	C3	0	0	1	0	0	0	0
	paths	C4	0	0	0	1	0	0	0
	paths	C5	0	0	0	0	1	0	0
	paths	C6	0	0	0	0	0	1	0
	paths	C7	0	0	0	0	0	0	1
Order 1	paths	C1–C2	1	1	0	0	0	0	0
	paths	C2–C3	0	1	1	0	0	0	0
	paths	C2–C6	0	1	0	0	0	1	0
	paths	C3–C4	0	0	1	1	0	0	0
	paths	C3–C7	0	0	1	0	0	0	1
	paths	C4–C5	0	0	0	1	1	0	0
Order 2	paths	C1–C2–C3	1	1	1	0	0	0	0
	paths	C1–C2–C6	1	1	0	0	0	1	0
	paths	C2–C3–C6	0	1	1	0	0	1	0
	paths	C2–C3–C4	0	1	1	1	0	0	0
	paths	C2–C3–C7	0	1	1	0	0	0	1
	paths	C3–C4–C5	0	0	1	1	1	0	0
	paths	C3–C4–C7	0	0	1	1	0	0	1
Order 3	paths	C1–C2–C3–C4	1	1	1	1	0	0	0
	paths	C1–C2–C3–C7	1	1	1	0	0	0	1
	cluster	C1–C2–C3–C6	1	1	1	0	0	1	0
	paths	C2–C3–C4–C5	0	1	1	1	1	0	0
	paths	C2–C3–C6–C7	0	1	1	0	0	1	1
	paths	C2–C3–C4–C6	0	1	1	1	0	1	0
	cluster	C2–C3–C4–C7	0	1	1	1	0	0	1
	paths	C3–C4–C5–C7	0	0	1	1	1	0	1
Order 4	paths	C1–C2–C3–C4–C5	1	1	1	1	1	0	0
	paths-cluster	C1–C2–C3–C4–C7	1	1	1	1	0	0	1
	paths-cluster	C1–C2–C3–C4–C6	1	1	1	1	0	1	0
	paths	C2–C3–C4–C5–C6	0	1	1	1	1	1	0
	paths-cluster	C2–C3–C4–C5–C7	0	1	1	1	1	0	1
	paths-cluster	C2–C3–C4–C6–C7	0	1	1	1	0	1	1
	paths-cluster	C1–C2–C3–C6–C7	1	1	1	0	0	1	1
Order 5	paths-cluster	C1–C2–C3–C4–C5–C6	1	1	1	1	1	1	0
	paths-cluster	C1–C2–C3–C4–C5–C7	1	1	1	1	1	0	1
	paths-cluster	C2–C3–C4–C5–C6–C7	0	1	1	1	1	1	1
	paths-cluster	C1–C2–C3–C4–C6–C7	1	1	1	1	0	1	1
Order 6	paths-cluster	C1–C2–C3–C4–C5–C6–C7	1	1	1	1	1	1	1

This new matrix representation is a generalization of the incidence matrix, and this matrix could be complete (representing all possible related sub-graphs) or constitute sub-graphs of determined orders or types (according to Kier and Hall nomenclature) as well as a combination of these (see Table 1). A particular case where only sub-graphs of *Order 1* (pairs of vertices or

edges in **G**) are considered, ***Q*** coincides with the common incidence matrix used in graph theory.

3.2 Derivative of Molecular Graph. Local and Total Definition

In this section, we will define novel indices which apply the Equation 1 for each pair of vertices in the **G**. Let us continue with the example of the 2,3-dimethylpentane molecule for which we have already obtained its corresponding frequency matrix according to the event proposed in the present report.

We characterize the participation intensities of different pairs of elements [atoms (vertices) in the molecule (graph)] from the calculation of the derivative for a pair of elements (see Equation 1):

$$\frac{\partial G}{\partial S}(C1,C2) = \frac{14-2(13)+27}{13} = 1.15$$
$$\frac{\partial G}{\partial S}(C1,C3) = \frac{14-2(11)+29}{11} = 1.90$$

In the same way, the rest of the values for pairs of elements of the graph are successively determined, as shown below:

$$\frac{\partial G}{\partial S}(C1,C4) = 2.50;\ \frac{\partial G}{\partial S}(C1,C5) = 4.50;\ \frac{\partial G}{\partial S}(C1,C6) = 2.67;$$
$$\frac{\partial G}{\partial S}(C1,C7) = 3.80;\ \frac{\partial G}{\partial S}(C2,C3) = 0.43$$

$$\frac{\partial G}{\partial S}(C2,C4) = 1.06;\ \frac{\partial G}{\partial S}(C2,C5) = 2.87;\ \frac{\partial G}{\partial S}(C2,C6) = 1.15;$$
$$\frac{\partial G}{\partial S}(C2,C7) = 1.82;\ \frac{\partial G}{\partial S}(C3,C4) = 0.55$$

$$\frac{\partial G}{\partial S}(C3,C5) = 2.10;\ \frac{\partial G}{\partial S}(C3,C6) = 1.90;\ \frac{\partial G}{\partial S}(C3,C7) = 1.14;$$
$$\frac{\partial G}{\partial S}(C4,C5) = 1.09;\ \frac{\partial G}{\partial S}(C4,C6) = 2.50$$

$$\frac{\partial G}{\partial S}(C4,C7) = 1.70;\ \frac{\partial G}{\partial S}(C5,C6) = 4.50;$$
$$\frac{\partial G}{\partial S}(C5,C7) = 3.40;\ \frac{\partial G}{\partial S}(C6,C7) = 3.80$$

All these pair derivatives will be organized in matrix form (**£** matrix), whose entries *ij* are the derivative values for the *i* and *j* vertices.

We now introduce a new concept with the purpose of obtaining new LOVIs from the derivatives for duplexes, which we will denominate the differential for atom *i*(Δ_i). The Δ_i for each of the elements of the **G** (i.e. each atomic nucleus) is defined as the summation over all the derivative values that include the element *i* (linear combination):

$$\Delta_i \sum_{j=1}^{n} \frac{\partial G}{\partial S}(i,j) = [\Delta_i] = [\pounds] \times [I] \quad (2)$$

where, *n* is the number of atoms in the molecule, and $\partial G/\partial S(i,j)$ is the derivative for vertices *i* and *j*. Equation 2 for Δ_i, may also be written in the matrix form, where [***I***] is a column unitary vector (an $n\times1$ matrix) and [**£**] is the derivative matrix (entries *ij* are the derivatives for *i* and *j* vertices). We obtain the atomic derivative value (LOVI) for each element, which would be: $\Delta_1 = 15.65$, $\Delta_2 = 8.35$, $\Delta_3 = 7.78$, $\Delta_4 = 9.87$, $\Delta_5 = 19.39$, $\Delta_6 = 15.65$ and $\Delta_7 = 14.47$.

If we thoroughly observe the values for each Δ_i, it can be noted that each value for the first, second, third, fourth, fifth and seventh atoms (from 1–5 and 7) are different, while the first and sixth are equal. This is logical behavior if we consider the chemical nature of each of these atoms, given that it is precisely the carbon atoms numbered 1 and 6 that exclusively possess identical chemical surroundings (terminal methyl groups). More so, the values for each Δ_i can be organized in the same order of their steric-electronic chemical surroundings. Like for example, the greatest value of Δ_i is possessed by the least enclosed atoms while the smallest value is presented by atom 3 that suffers the greatest steric hindrance. This also coincides with the nature of the concept of the graph derivative since the atom that most suffers hindrance is the one that most contributes to the formation of the molecule.

3.3 Codification of Heteroatoms and Unsaturated Bonds

We propose an approach, in this report, that permits us characterize adequately molecules with heteroatoms and unsaturated bonds. As an example we choose an isomer of, *N*-methylisobutyramide molecule (see Figure 3). According to the procedure previously explained, we can assert that

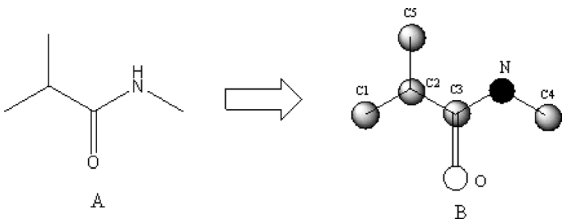


Figure 3. The chemical structure and molecular graph of [the numbers correspond to the labels that are assigned to the atoms (vertices) in the molecular structure]: A) *N*-methylisobutyramide [H (implicit)-depleted structure], B) molecular graph of *N*-methylisobutyramide.

the ***Q*** and ***F*** matrices for the **G** represented in Figure 2 are identical to those of 2,3-dimethylpentane.

Nonetheless, it can be easily perceived by simple inspection that the molecular structure of this new molecule contains a heteroatom and a double bond. Let us create a vector of weights V_p, in which the weight (p_i) corresponds reciprocally to element p_i for a given condition. The distinct weights for each atom (condition, according to this event) can be determined according to the relationship $p_i = P/\delta$, where P represents a characteristic property of each atom (for example: atomic mass, electronegativity, etc.) and δ is the vertex degree.

As an example we use the electronegativity (according to Pauling's scale) as weight for each atom (condition). The weights or labels for the different atoms are:

$$p(C1) = \frac{2.5}{1} = 2.5 \quad p(O) = \frac{3.5}{2} = 1.75 \quad p(C5) = \frac{2.5}{1} = 2.5$$

$$p(C2) = \frac{2.5}{3} = 0.833 \quad p(N) = \frac{3.0}{2} = 1.5$$

$$p(C3) = \frac{2.5}{4} = 0.625 \quad p(C4) = \frac{2.5}{1} = 2.5$$

From these resulting values we construct a vector of weights, $V_p = (2.5, 0.833, 0.625, 1.75, 1.5, 2.5, 2.5)$. In the same way, we can obtain this vector by means of a weighted matrix.

$$P = \begin{pmatrix} 2.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.833 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.625 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.75 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2.5 \end{pmatrix}$$

Multiplying the incidence matrix with the weighted matrix, we obtain the *weighted incidence matrix* $\boldsymbol{Q}_P = [\mu_{ij}]$, which is similar to ***Q*** in its form only that this new matrix captures more particular information of each of the atoms in the molecule on top of the atom-atom connectivity with others in the mentioned molecule, from which it follows that:

$\mu_{ij} = p_i$, if the j^{th} condition is included in the ith collection of conditions, in which the event is true

$\mu_{ij} = 0$, otherwise.

We can now continue with the method previously proposed for determining the derivative values over the pairs of graph elements. That is, we obtain the matrix $\boldsymbol{Q}_P$ and its transpose $\boldsymbol{Q}_P^T$, followed by the corresponding multiplication operation as seen in the previous example ($\boldsymbol{Q}_P^T \times \boldsymbol{Q}_P = \boldsymbol{F}_P$). The *weighted frequency matrix* $\boldsymbol{F}_P$ captures information on the number of times each element participates in the formation of the **G** (according to the predetermined event), on top of its participation characteristic, that may be interpreted as its identity or relative capacity (with respect to other atoms of the molecule) to form the molecular structure.

The derivative values for the pairs of elements of the molecular graph are the following:

$$\frac{\partial G}{\partial S}(C1, C2) = 1.92; \quad \frac{\partial G}{\partial S}(C1, C3) = 3.75; \quad \frac{\partial G}{\partial S}(C1, O) = 4.10;$$
$$\frac{\partial G}{\partial S}(C1, N) = 2.57; \quad \frac{\partial G}{\partial S}(C1, C4) = 4.50$$

$$\frac{\partial G}{\partial S}(C1, C5) = 2.67; \quad \frac{\partial G}{\partial S}(C2, C3) = 0.51; \quad \frac{\partial G}{\partial S}(C2, O) = 2.03;$$
$$\frac{\partial G}{\partial S}(C2, N) = 1.41; \quad \frac{\partial G}{\partial S}(C2, C4) = 3.63$$

$$\frac{\partial G}{\partial S}(C2, C5) = 1.92; \quad \frac{\partial G}{\partial S}(C3, O) = 1.74; \quad \frac{\partial G}{\partial S}(C3, N) = 1.24;$$
$$\frac{\partial G}{\partial S}(C3, C4) = 3.52; \quad \frac{\partial G}{\partial S}(C3, C5) = 3.75$$

$$\frac{\partial G}{\partial S}(O, N) = 1.63; \quad \frac{\partial G}{\partial S}(O, C4) = 3.53; \quad \frac{\partial G}{\partial S}(O, C5) = 4.10;$$
$$\frac{\partial G}{\partial S}(N, C4) = 1.02; \quad \frac{\partial G}{\partial S}(N, C5) = 2.57$$

$$\frac{\partial G}{\partial S}(C4, C5) = 4.50$$

With these values we can also obtain the derivatives of each atom in the molecule: ${}^P\Delta_{c1} = 19.51$, ${}^P\Delta_{c2} = 11.42$, ${}^P\Delta_{c3} = 14.51$, ${}^P\Delta_O = 17.13$, ${}^P\Delta_N = 10.44$, ${}^P\Delta_{c4} = 20.7$ and ${}^P\Delta_{c5} = 19.51$. These weighted Δ_i, $p\Delta_i$, may also be written in the matrix form by using the Equation 2, where [I] is a column unitary vector (an $n \times 1$ matrix) but [p£] is used instead of an unweighted derivative matrix, £. The weighted derivative matrix, p£, is obtained from the weighted frequency matrix, $\boldsymbol{F}_P$, employed in Equation 1.

A second weighting scheme will be used in order to obtain other weighted LOVIs, designed as ${}_p\Delta_i$. Here, the vector of weights, $\boldsymbol{V}_p$, is multiplied by the *unweighted derivative matrix*, £. Therefore $[{}_p\Delta_i] = [\pounds] \times [\boldsymbol{V}_p]$. This formula (${}_p\Delta_i$) is similar to the one defined in Equation 2 for Δ_i, only that [$\boldsymbol{V}_p$] is used instead of [$\boldsymbol{I}$], where [$\boldsymbol{V}_p$] is a column weighting vector (an $n \times 1$ matrix) whose elements are weights (atom-labels) of the vertices of the **G**. It is important to remark that in this second weighting approach the derivative matrix £ is obtained from an unweighted ***F***. Alternatively, atomic weighted descriptors could be obtained by multiplying the vector of unweighted LOVIs, $\mathbf{V}^{sp}$ by the weighting matrix **P**, $\mathbf{V}^{sp} \times \mathbf{P} = \mathbf{V}^{p}$, and thus we can obtain a new vector whose elements would be weighted LOVIs. In the example introduced in this epigraph, $\boldsymbol{V}_p = (2.5, 0.833, 0.625, 1.75, 1.5, 2.5, 2.5)$.

3.4 Applying *Invariants* (Operators) to Atomic Derivative: Generalization of the Procedure for Obtaining Global and Local (Group and Atom-Type) Indices from LOVIs

Over the years, it has been generally accepted that the definition of global (or local) indices from LOVIs implies the summation of the contributions of the elements that constitute a given *G*.[7–8] In fact in quantum chemistry, the notion that "the summation of the parts makes the total" is applied. For instance, LCAO (Linear Combination of Atomic Orbitals) is a means of forming molecular orbitals (MOs) by taking linear combinations of functions associated with the

Table 2. Invariants functions to derive molecular descriptors (total and local) from local vertex invariants (LOVIs). The L_i is LOVI associated to the atoms v_i and n is the number of atoms.

No.	Group	Name	ID	Formula
1	Norms (metrics)	Minkowski's norms ($p=1$) Manhattan norm	N1	$\|\bar{x}\|_1=\sum_{i=1}^{n}\|L_i\|$
2		Minkowski's norm ($p=2$) Euclidean norm	N2	$\|\bar{x}\|_2=\sqrt{\sum_{i=1}^{n}\|L_i\|^2}$
3		Minkowski's norm ($p=3$)	N3	$\|\bar{x}\|_3=\sqrt[3]{\sum_{i=1}^{n}\|L_i\|^3}$
4		Penrose's size	PN	$d_i=\sqrt{\frac{1}{n^2}\left[\sum_{i=1}(L_i)\right]^2}$
5	Mean (first statistical moment)	Geometric mean	G	$\bar{\xi}=\sqrt[n]{\prod_{i=1}^{n}L_i}$
6		Arithmetic mean (potential with $\alpha=1$)	M	$m_a=\left(\frac{L_1^a+L_2^a+\cdots L_n^a}{n}\right)^{\frac{1}{a}}$
7		Quadratic mean (potential with $\alpha=2$)	P2	
8		Potential mean (potential with $\alpha=3$)	P3	
9		Harmonic mean (potential con $\alpha=-1$)	A	
10	Statistical (highest statistical moments)	Variance	V	$v=\frac{\sum_{i=1}^{n}(L_i-\bar{L})^2}{n-1}$
11		Skewness	S	$S=n\,M_3/[(n-1)\,(n-2)\,s^3]$ $M_3=\sum_{i=1}^{n}(L_i-\bar{L})^3$ s^3 is the standard deviation raised to the 3rd power n is the number of atoms.
12	Statistical (highest statistical moments)	Kurtosis	K	$K=[n\,(n+1)\,M_4-3\,M_2\,M_2\,(n-1)]/[(n-1)\,(n-2)\,(n-3)\,s^4]$ $M_j=\sum_{i=1}^{n}(L_i-\bar{L})^j$ n is the number of atoms; s^4 is the standard deviation raised to the fourth power
13		Standard deviation	DE	$\sigma=\sqrt{\frac{(\sum L_i-\bar{L})^j}{i=1}}$
14		Variation coefficient	CV	$c_v=s/\bar{L}$
15		Range	R	$R=L_{max}-L_{min}$
16		Percentile 25	Q1	$P25=\left[\frac{N}{4}+\frac{1}{2}\right]$ N is the number of values
17		Percentile 50	Q2	$P50=\left[\frac{N}{2}+\frac{1}{2}\right]$ N is the number of values
18		Percentile 75	Q3	$P75=\left[\frac{3N}{4}+\frac{1}{2}\right]$ N is the number of values
19		Inter-quartile Range	I50	I50=P75−P25
20		X max	MX	L_i maximum
21		X min	MN	L_j minimum

[a] The formulae used in these invariants, are simplified forms of general equations given that the vector $\bar{y}$is constituted by the coordinates of the origin. For example, in the case of the Euclidean norm (N2), the general formula is:

$$\|\bar{x}\|_2=\sqrt{\sum_{i=1}^{n}(x_i-y_1)^2+(x_j-y_j)^2+(x_z-y_z)^2}$$

but given that $\bar{y}=(0, 0, 0)$, this formula reduces to

$$\|\bar{x}\|_2=\sqrt{\sum_{i=1}^{n}|x_i|^2}$$

different atoms in the molecule.[9] Therefore, MOs are made up as LCAO of atoms composing the system, i.e. they are written in the form, $\varphi_i = \sum_{j=1}^{n} c_{ij} Y_j$, where, *i* is the number of the MO, φ; *j* are the numbers of atom *Y*-orbital; c_{ij} are the numerical coefficients defining the contributions of individual AOs to the given MO. Such a way of constructing a MO is based on the assumption that an atom represented by a definite set of orbitals remains distinctive in the molecule. However, summation (in our case, Minskowski's first norm, N1, see below) is just one of the many invariants capable of globally characterizing given LOVIs.

In this work, we introduce a series of *invariants* that generalize the traditional method of obtaining global (or local) indices by summation of the LOVIs. These are classified into three major groups (see Table 2),[1,10]

1) Norms (or Metrics): Minkowski's norms (N1, N2, N3) and Penrose's size (PN).
2) Mean Invariants (first statistical moment): Geometric Mean (G), Arithmetic Mean (M), Quadratic Mean (P2), Potential Mean (P3) and Harmonic Mean (A).
3) Statistical Invariants (highest statistical moments): Variance (V), Skewness (S), Kurtosis (K), Standard Deviation (DE), Variation Coefficient (CV), Range (R), Percentile 25 (Q1), Percentile 50 (Q^2), Percentile 75 (Q3), Inter-quartile Range (I50), Maximum X (MX) and Minimum X (MN).

It should be noted that these invariants are only applied with the purpose of generalizing the summation i.e., there exists other forms of obtaining indices from LOVIs but these are not related to the summation but rather with other procedures, e.g., the use of the atomic derivative (Δ_i ${}^{p}\Delta_i$ and ${}_{p}\Delta_i$) as a LOVI in Randić's equation, M_1 and M_2 Zagreb formula, among others.

The application of "classical" algorithms (invariants) on these LOVIs (Δ_i, ${}^{p}\Delta_i$ and ${}_{p}\Delta_i$), and on the **F** matrix, will be presented in forthcoming papers.

The application of these invariants to the vector of weighted or unweighted atomic derivative values (Δ_i, ${}^{p}\Delta_i$ and ${}_{p}\Delta_i$) enables us obtain a series of global indices using atomic derivative values as LOVIs. In the same way, these invariants could be applied to a vector comprised of a particular class of local (group and atom-type) derivative values obtaining *local derivative-based indices* for atom-types or groups (for example, in TOMOCOMD-CARDD software,[11] the following local indices can be calculated: Proton Acceptors (AH), Proton Donors (DH), Heteroatoms (HT), Halogens (HL) and Carbon (Cb) and Unsaturated bonds (IS). It should be noted that *local definition* capacity is one of the most important requisite for new MDs.[12]

It is rather important to emphasize that all invariants in Table 2 will be applied not only to the original vector of LOVIs (in the tree possible forms: Δ_i ${}^{p}\Delta_i$ and ${}_{p}\Delta_i$) but also to standardized LOVIs. That is, global or local indices will be calculated from a vector of standardized atomic LOVIs (${}^{s}\Delta_i$, ${}^{sp}\Delta_i$ and ${}^{s}{}_{p}\Delta_i$). In the standardization procedure, all values of *original* LOVIs (Δ_i, ${}^{p}\Delta_i$ and ${}_{p}\Delta_i$) are replaced by standardized LOVI values (${}^{s}\Delta_i$ ${}^{sp}\Delta_i$ and ${}^{s}{}_{p}\Delta_i$), which are computed as follows: Std. LOVIs = (Original LOVI – mean of LOVIs)/Std. deviation of original LOVIs. With this re-scaling, the components of a vector of LOVI have a mean of 0 and a standard deviation of 1 (this means that standardized LOVIs have the same dimensions, i.e., are of comparable magnitude).

4 QSAR/QSPR Applications

4.1 Database Selection

Four *benchmark* datasets have been used to evaluate the QSPR/QSAR behavior of the new TOMOCOMD-CARDD MDs. With this objective in mind, we developed QSPR models to describe several physicochemical properties of octane isomers (FIRST ROUND),[13] to analyze the boiling point of 28 alkyl-alcohols (SECOND ROUND),[8b,14] four physicochemical properties for a total of 209 polychlorobiphenyls (THIRD ROUND),[15] and a biological activity of 22 Phenetylamines (FOURTH ROUND).[15] The data set used in first, third and fourth round have been proposed by the International Academy of Mathematical Chemistry[13a] as ideal for the evaluation of new MDs.

The use of octanes (FIRST ROUND), as a suitable data set for testing TIs, has been advocated for Randić and Trinajstić and at present is considered by International Academy of Mathematical Chemistry as a benchmark database for comparison among old (well-known) and new MDs.[3a,16] In fact, this dataset has been used by several researchers to evaluate the modeling power of their new MDs.[13b,17] The physicochemical properties studied in this paper are boiling point (*Bp*), motor octane number (MON), heat of vaporization (*HV*), molar volume (*MV*), entropy (S) and heat of formation (ΔH_f). This selection is recommended, because the physicochemical properties commonly studied in QSPR analyses with TIs are interrelated for data sets of compounds with different molecular weights, for instance for alkanes having from two to nine carbon atoms. These correlations are not necessarily observed when the same indices are used in isomeric data sets of compounds, such as the octane data set. In addition, these properties are hardly interrelated when octanes are used as a data set.[12b] On the other hand, all TIs are designed to have (gradual) increments with the increments in the molecular weight. In this way, if we do the present study by using a series of compounds having different molecular weights, we will find "false" interrelations among the indices by an overestimation of the size effects inherent to these descriptors.[17a,b] The same is also valid when a QSPR model is to be obtained. It is not difficult to find "good" linear correlations between TIs and physicochemical properties of alkanes in data sets with great size variability.[17a,b] In fact, the simple use of the number of vertices in the **G** produced regression coefficients greater than 0.97 for most of the physicochem-

ical properties of C2–C9 alkanes studied by Needham et al.[18] However, when data sets of isomeric compounds are considered, MDs that typically have high correlation coefficients when molecules of different sizes are considered may no longer show such good linear correlations. In conclusion, if a newly proposed MD is not able to model the variation of at least one property of octanes, then it probably does not contain any useful molecular information.

In order to illustrate the possibilities of our approach in the QSPR studies of heteroatomic molecules, we have selected the boiling point of 28 alkyl-alcohols (SECOND ROUND) to be investigated.[8b,14] This data set was firstly studied by Kier and Hall[8b] using E-state/biomolecular encounter parameters and later by Estrada and Molina[14a] employing the local spectral moments of the edge adjacency matrix. This heteromolecule-based database is composed of 28 alkyl-alcohols, 14 are primary, 6 secondary and 8 tertiary, for which the boiling point (Bp) has been reported previously. Alcohols constituted a good set of chemicals for comparative study, because it is an isomeric data set, comprised of heteroatomic compounds and the boiling point not only depends on *gradual* variation of molecular weight, but also of H-bonding capacity and the R-group type. Additionally, QSPR studies are available for comparison purposes.[8b,14a]

The third heteromolecule-based database that will be studied here consists of a total of 209 polychlorobiphenyls (THIRD ROUND), early studied by using dragon descriptors.[15] It is well-known that polychlorinated biphenyls (PCB) are widespread and persistent organic contaminants. Moreover, it has been demonstrated that they are toxic and lipophilic and tend to be bioaccumulated. Four of the physicochemical properties of environmental relevance for PCB congeners have been chosen: melting point (*mp*), octanol water partition coefficient (log K_{OW}), Henry's law constant (*H*), and aqueous water coefficient, expressed as the negative logarithm (log Y_W).[15]

We finish this study with a *N*,*N*-dimethyl-2-halo-phenethylamines family (FOURTH CASE). This data set has been frequently used in QSAR studies and it is also proposed by the International Academy of Mathematical Chemistry for the evaluation of new ITs. The response is the antagonism of these compounds to epinephrine (adrenergic blocking activity) in the rat (log1/ED_{50}).[15]

4.2 Computational Approach

TOMOCOMD is an interactive program for molecular design and bioinformatic research.[11] It consists of four subprograms; each one of them allows drawing the structures (drawing mode) and calculating molecular 2D/3D (calculation mode) descriptors. The modules are named CARDD (Computed-Aided 'Rational' Drug Design), CAMPS (Computed-Aided Modeling in Protein Science), CANAR (Computed-Aided Nucleic Acid Research) and CABPD (Computed-Aided Bio-Polymers Docking). The DIVATI (Discrete deriVAtive Type Indices) module used in the present report forms part of the CARDD section.

4.2.1 Computational In-House Software

The total and local (in this case, group-type) molecular graph derivative indices used to search the best regression of several physicochemical and biological properties of four data sets of organic compounds broadly used were calculated by an *in-house* software developed on the JAVA platform. The novel indices are implemented in DIVATI,[19] a new module of TOMOCOMD-CARDD program to facilitate their automatic computation.

In this case, in order to distinguish saturated hydrocarbons from chemical structures with heteroatoms and unsaturated bonds, we used a weighting scheme conformed by five atomic properties: Atomic number (Z), Atomic Mass (*A*), Van der Waals Volume (*V*), Polarizability (*P*) and Pauling Electronegativity (*E*). These atomic-labels are shown in Table 3. The MDs computed in this study adopted the following format:

$$ {}^{PIS}_{Qorder}Inv^{E}(localtype) $$

where, P means ponderation (see Table 3) and S ponderation position, where '*In*' means that the atomic weighting is made in **Q** matrix, 'Pd' means that the atomic weighting is made in derivative Matrix, 'Pl' means that the atomic weighting is made in LOVIs vector and 'Sp' means that it is not pondered. Q_{order} will appear only if a specific order *k* (or any combination of these) is used to compute the total or local MDs. *Inv* means the invariant used to compute the MDs from atomic LOVIs. The superscript E stands for standardized LOVIs. Finally, the parenthesis will appear if the MDs are computed for a particular group of atoms (local indices). In this sense, in the parenthesis will appear the group-type indices, namely, Cb for carbon atoms, HT for Heteroatoms, AH for proton acceptor, DH for proton donor, HL for halogens, MC for methyl carbons and IS for unsaturated bonds. Finally, note that the particularities of subgraph types were not taken into account and only orders were considered.

Work Methodology. The main steps for the application of the present method in QSAR/QSPR and drug design can be briefly summarized in the following algorithm: 1) Draw the molecular-graphs for each molecule in the data set. The software DIVATI[19] accepts the mol or sdf file formats. 2) Choose the order and type of sub-graphs with which the incidence matrix will be built. 3) Use appropriate atomic properties in order to weight and differentiate the molecular atoms. The properties used are those previously proposed for the calculation of DRAGON descriptors,[1,13b,20] i.e., atomic mass (*M*), atomic polarizability (*P*), atomic electronegativity (*K*), Van der Waals atomic volume (*V*).The values of these atomic labels are shown in Table 3. In this step we

Table 3. Values of the atomic weights used for MDs calculation. (VdW: van der Waals)

ID	Atomic mass	VdW volume (Å^3)	Polarizability (Å^3)	Pauling electronegativity
H	1.01	6.709	0.667	2.2
B	10.81	17.875	3.030	2.04
C	12.01	22.449	1.760	2.55
N	14.01	15.599	1.100	3.04
O	16.00	11.494	0.802	3.44
F	19.00	9.203	0.557	3.98
Al	26.98	36.511	6.800	1.61
Si	28.09	31.976	5.380	1.9
P	30.97	26.522	3.630	2.19
S	32.07	24.429	2.900	2.58
Cl	35.45	23.228	2.180	3.16
Fe	55.85	41.052	8.400	1.83
Co	58.93	35.041	7.500	1.88
Ni	58.69	17.157	6.800	1.91
Cu	63.55	11.494	6.100	1.9
Zn	65.39	38.351	7.100	1.65
Br	79.90	31.059	3.050	2.96
Sn	118.71	45.830	7.700	1.96
I	126.90	38.792	5.350	2.66

should also indicate position (step) where the weighting scheme is to be applied. 4) Select standardized or non-standardized LOVIs and the invariants to obtain total or local descriptors. 5) Compute the total and local indices. The local indices implemented are (Cb, AH, DH, HT, HL, MC, IS). 6) Find a QSPR/QSAR equation by using several multivariate analytical techniques, such as multilinear regression analysis (MRA), neural networks (NN), linear discrimination analysis (LDA), and so on. 7) Test the robustness and predictive power of the QSPR/QSAR equation by using internal (cross-validation) and external (using a test set and an external predicting set) validation techniques.

4.3 Chemometric Analysis

The whole set of new MDs were used as independent variables for deriving QSPRs by means of multiple linear regression (MLR) technique. The MOBYDIGS (version 1.0-2004).[21] was employed to perform variable selection and QSPR modeling.

This software allows searching for linear regression models by developing optimal model populations using genetic algorithms (GAs). The GAs[21–22] are a class of algorithms inspired by the process of natural evolution in which species having a high fitness under some conditions can prevail and survive to the next generation; the best species can be adapted by crossover and/or mutation in the search for better individuals. The GAs uses a population of individuals as a model search for the globally optimum solution to a problem. The GAs constitute an optimization procedure that permits the search of the best values of a set of parameters able to optimize an objective function. The GA evolution consists in the replication of the GA operators such as reproduction, mutation, epidemy, predatory and tabu, in such a way that the global quality of the population individuals (the models) increases and the best subset of models could be found. The population size was set at 100 and the reproduction/ mutation trade-off (T) at 0.50. The GAs with initial population sizes of 100 rapidly converge (200 generations) and achieve optimum QSAR models in a reasonable number of GA generations.

The models were optimized using as objective function (optimization function) the statistical parameter Q_{Loo}^2 ("leave one out" crossed validation) and they were validated using techniques "*bootstrapping*" (Q_{boot}^2) and "*scrambling*" (*Ysc*). The search for the best model can be processed in terms of the highest square correlation coefficient (R^2) or *F*-test equations (Fisher-ratio's *p*-level [*p* (*F*)]) and the lowest standard deviation equations (s). We analyzed statistical parameters Q_{Loo}^2 ("leave one out" crossed validation) and Q_{boot}^2 to evaluate the quality of the models. In the recent years, the LOO press statistics (e.g., Q_{Loo}^2) have been used as a means of indicating predictive ability. Many authors consider high Q_{Loo}^2 values (for instance, $Q_{Loo}^2 > 0.5$) as an indicator or even as the ultimate proof of the high-predictive power of a QSAR model. However, it is known that this affirmation is only true for small data (< 100 cases), and that data with wide dimensionality is just a necessary but not sufficient condition to affirm that a model possesses adequate predictive power.

We calculated all the possible indices, using all the weights, graph-orders and graph-types. The best models in each case were selected, taking in consideration the quality of their corresponding statistical parameters.

4.4 QSARs/QSPRs

4.4.1 Case 1. Physicochemical Properties of Octane Isomeres

In a previous study by Consonni et al. several physicochemical properties for octane isomers were analyzed.[13b] However, to evaluate the quality of the models based on our new GDIs (*acronym of Graph Derivative Indices*) we have taken as reference only six physicochemical properties selected in the previous study. Therefore, we analyzed the quality of the QSPR models obtained to describe the boiling point (*Bp*), motor octane number (*MON*), heat of vaporization (*HV*), molar volume (*MV*), entropy (*S*), and heat of formation (ΔH_f) of the octane isomers. The regressions of octane properties, based on the GDIs, will be compared to some regressions based on 2D (topological/topo-chemical) and 3D (geometrical) MDs, taken from the literature.[13b]

The best linear models, of three parameters for each property, found using GDIs are next presented:

Boiling point (*Bp*)

$$Bp = 22.83\ (\pm 14.87) + 19.01\ (\pm 1.50)(A1/In)De(Cb) - 8.32\ (\pm 0.495)\ (E1/Pd)P_1 + 57.31\ (\pm 5.32)\ (E1/Pd)\ M(Cb) \quad (3)$$

$N = 18 \quad R^2 = 97.35 \quad s = 1.097 \quad {Q_{Loo}}^2 = 95.59 \quad s_{CV} = 1.249$
${Q_{boot}}^2 = 95.01 \quad Ysc = 0.088 \quad F = 171.42$

Heat of formation (ΔH_f)

$$\Delta H_f = 4.95\ (\pm 0.23) - 0.58\ (\pm 0.052)\ (E1/In)\ Ra(Cb) - 1.75(\pm 0.15)\ (E2/In)\ S^E + 0.065\ (\pm 0.005)\ (A2/Pd)\ DE(Cb) \quad (4)$$

$N = 18 \quad R^2 = 93.12 \quad s = 0.245 \quad {Q_{Loo}}^2 = 87.06 \quad s_{CV} = 0.296$
${Q_{boot}}^2 = 87.20 \quad Ysc = 0.112 \quad F = 63.17$

Heat of vaporization (*HV*)

$$HV = 75.578\ (\pm 1.35) - 0.53\ (\pm 0.03)\ (V1/Pd)PN + 6.74\ (\pm 0.89)\ (P1/Pd)\ M(Cb) + 0.60\ (\pm 0.06)\ (E2/In)PN \quad (5)$$

$N = 18 \quad R^2 = 98.47 \quad s = 0.277 \quad {Q_{Loo}}^2 = 97.73 \quad s_{CV} = 0.297$
${Q_{boot}}^2 = 97.00 \quad Ysc = 0.073 \quad F = 299.47$

Motor octane number (**MON**)

$$MON = -798.27\ (\pm 33.81) - 2.90\ (\pm 0.18)\ (A/Pd)\ N_3 + 42.45\ (\pm 5.08)\ (P/In)N_3\ (Cb) + 19.05\ (\pm 2.88)\ (E5/Pd)\ M \quad (6)$$

$N = 18 \quad R^2 = 99.16 \quad s = 2.536 \quad {Q_{Loo}}^2 = 98.53 \quad s_{CV} = 2.912$
${Q_{boot}}^2 = 98.20 \quad Ysc = 0.098 \quad F = 473.22$

Entropy (*S*)

$$S = 111.92\ (\pm 8.96) + 713.21\ (\pm 40.1)\ (A3/Pd)\ MX^E\ (Cb) - 490.72\ (\pm 24.38)\ (P3/Pd)\ P_3{}^E\ (Cb) + 7.18.14\ (\pm 40.19)\ (P3/Pd)\ MN^E\ (Cb) \quad (7)$$

$N = 18 \quad R^2 = 97.41 \quad s = 0.802 \quad {Q_{Loo}}^2 = 96.23 \quad s_{CV} = 0.852$
${Q_{boot}}^2 = 95.48 \quad Ysc = 0.049 \quad F = 175.24$

Molar volume (*MV*)

$$MV = 141.33\ (\pm 2.33) - 33.26\ (\pm 3.9)\ (P2/Pd)\ Q_1^E\ (Cb) - 1.24\ (\pm 0.24)\ (P1/Pd)K^E - 0.94\ (\pm 0.23)\ (V2/Pd)\ M^E \quad (8)$$

$N = 18 \quad R^2 = 84.49 \quad s = 2.3 \quad {Q_{Loo}}^2 = 52.55 \quad s_{CV} = 4.04$
${Q_{boot}}^2 = 55.04 \quad Ysc = 0.123 \quad F = 25.41$

For each selected property of octane isomers, the statistical information for the best regressions with one, two and three MDs published so far[13b] are also depicted in Table 4, together with the LOO cross-validation-explained variance (${Q_{Loo}}^2$), the square of correlation coefficient (R^2, given in percentages), the standard error of estimate (*s*), the standard deviation the error in LOO cross-validation (s_{CV}), the bootstrap average predictive power (${Q_{boot}}^2$), the *Y*-scrambling parameter (*Ysc*) and Fischer ratio (*F*). As it can see from the statistical parameters of regression equations in Table 4, all of the physicochemical properties were well described by the GDIs. In all the cases the models obtained with the GDIs demonstrated statistical robustness, with statistical parameters comparable with the best models proposed in the literature.[13b] For instance, we can observe that the statistical parameters for the model obtained with GDIs to describe heat of vaporization (*HV*) (Equation 5) of octanes are better than those taken from the literature using 2D and 3D MDs.[13b] It should be pointed out that in the models based on the GDIs, both regressions for the motor octane number (*MON*) (see Equation 6) are better-to-similar than the models published so far.[13b] The models obtained, using the GDIs, to describe boiling point (*Bp*) (Equation 3), entropy (*S*) (Equation 7), heat of formation (ΔH_f) (Equation 4) and molar volume (MV) (Equation 8) have significant differences with those obtained with WHIM, GETAWAY and TIs altogether. However, these properties were better described with our approach than several TIs.

According to the obtained QSPR results, it is possible to conclude that the novel MDs encode some useful molecular information different from that of previously proposed descriptors. Moreover, they depict considerable diversity being able to adequately describe the variation of different properties of octanes.

Table 4. Statistical information for best multiple regression models of selected physicochemical properties of octane isomers.

Property	Index	Descriptor	*n*	R^2	*s*	Q_{Loo}^2	s_{CV}	Q_{boot}^2	*Ysc*	*F*
Boiling point (*Bp*)	GETAWAY + WHIM + topological	$^{2}\chi$ $^{2\overline{\chi}}$ *HATS*$_6$(p)	3	98.78	0.744	98.12	–	–	–	–
	GETAWAY	*HATS*$_2$(v) R$_4$(u) R$_6$(v)	3	98.32	0.897	97.10	–	–	–	–
	GETAWAY + WHIM + topological	$^{2}\chi$ *HATS*$_6$(p)	2	97.58	1.013	96.62	–	–	–	–
	GDIs	[$^{A/In}_{1}$DE(Cb)] [$^{E/Pd}_{1}$P$_2$] [$^{E/Pd}$1 M(Cb)]	3	97.35	1.097	95.59	1.249	95.01	0.088	171.4
	Topological	S^{3}W S^{4}W SJ	3	95.84	1.394	–	–	–	–	–
	Topological	S^{3}W S^{4}W	2	94.78	1.508	–	–	–	–	–
	GDIs	[$^{E/Pd}$G] [$^{V/In}$P$_1$]	2	91.4	1.91	86.08	2.218	86.22	0.01	79.7
	GETAWAY	*HATS*$_2$(m) R$^+_4$(u)	2	89.62	2.098	84.86	–	–	–	–
	Topological	WW x$_1$	2	81.36	2.810	–	–	–	–	–
	Topological	Z	1	78.85	2.90	–	–	–	–	–
	GETAWAY + WHIM + topological	*HATS*$_2$(m)	1	74.64	3.175	66.47	–	–	–	–
	GDI	[$^{P/In}$6 M(Cb)]	1	67.26	3.608	56.75	3.91	59.32	−0.012	32.9
	Topological	$^{2}\chi$ W	1	67.77	3.630	–	–	–	–	–
Motor octane number (*MON*)	GETAWAY + WHIM + topological	$^{v}I_D^M$ Ts *HATS*$_1$(m)	3	99.23	2.439	98.58	–	–	–	–
	GDIs	[$^{A/Pd}$N$_3$] [$^{P/In}$N$_3$(Cb)] [$^{E/Pd}$M$_5$]	3	99.16	2.536	98.53	2.912	98.20	0.098	473.2
	GETAWAY	*HATS*$_4$(u) *HATS*$_7$(v) R$_7$(p)	3	98.62	3.259	97.42	–	–	–	–
	GDIs	[$^{E/In}_{6}$PN(Cb)] [$^{P/Pd}_{6}$RA(Cb)]	2	98.03	3.73	97.19	4.024	97.11	0.021	323.7
	Topological	Sχ^1W χ^7W χ^3W	3	98.05	3.855	–	–	–	–	–
	GETAWAY + WHIM + topological	Ts *H*$_4$(e)	2	97.68	4.053	96.77	–	–	–	–
	GETAWAY	*HATS*$_7$(m) R$_4$(u)	2	95.78	5.466	91.28	–	–	–	–
	Topological	Sχ^1W Sχ^3W	2	95.64	5.533	–	–	–	–	–
	Topological	$^{\overline{X}^7}$W	1	95.22	5.589	–	–	–	–	–
	GETAWAY + WHIM + top.	Ts	1	92.40	7.069	90.83	–	–	–	–
	GDI	[$^{P/In}$G(Cb)]	1	92.33	7.10	89.76	7.68	90.00	−0.042	168.5
	Topological	I$_{wD}$	1	91.97	7.270	–	–	–	–	–
	GETAWAY	RElG	1	88.98	8.515	85.64	–	–	–	–
Heat of vaporization (*HV*)	GDIs	[$^{V/Pd}_{1}$PN] [$^{P/Pd}$1 M(Cb)][$^{E/In}_{2}$PN]	3	98.47	0.277	97.73	0.279	97.00	0.073	299.5
	GETAWAY + WHIM + top.	$^{0}\overline{\chi}^{3}\kappa$ R$^+_6$(u)	3	98.42	0.281	97.57	–	–	–	–
	GETAWAY	*HATS*$_6$(u) R$_4$(u) R$^+_1$(m)	3	97.18	0.375	95.46	–	–	–	–
	GETAWAY + WHIM + topological	$^{2}\chi$ R$^+_6$(u)	2	96.53	0.402	95.18	–	–	–	–
	GDIs	[$^{E/Pd}_{1}$CV(Cb)] [$^{E/In}_{2}$S(Cb)]	2	96.27	0.416	93.74	0.493	94.09	−0.004	193.8
	Topological	χ^1W χ^2W χ^3W	3	95.65	0.459	–	–	–	–	–
	GETAWAY	*HATS*$_4$(u) R$_6$(e)	2	94.87	0.488	93.15	–	–	–	–
	Topological	^{4}W ^{5}W	2	92.62	0.577	–	–	–	–	–
	Topological	Z	1	91.78	0.429	–	–	–	–	–
	GDI	[$^{V/In}_{5}$G(Cb)]	1	85.75	0.788	81.19	0.854	82.36	−0.029	96.3
	GETAWAY + WHIM + topological	$^{2}\chi$	1	88.61	0.705	80.80	–	–	–	–
	GETAWAY	R^2(m)	1	85.70	0.790	79.74	–	–	–	–
	Topological	WW x$_1$	2	84.27	0.820	–	–	–	–	–
Heat of formation (ΔH_f)	GETAWAY + WHIM + topological	*HATS*$_5$(m) *HATS*$_7$(m) *R*$_4$(e)	3	96.60	0.254	95.06	–	–	–	–
	GETAWAY + WHIM + topological	$^{2}\chi$ *HATS*$_2$(e)	2	93.24	0.346	90.96	–	–	–	–
	GETAWAY	*HATS*$_7$(u) R^2(m)	2	92.87	0.356	90.18	–	–	–	–
	GETAWAY + WHIM + topological	*HATS*$_2$(m)	1	89.34	0.421	87.18	–	–	–	–
	GDIs	[$^{E/In}_{1}$RA(Cb)] [$^{E/In}_{2}$S^E] [$^{A/Pd}_{2}$DE(Cb)]	3	93.12	0.245	87.06	0.296	87.20	0.112	63.2
	Topological	Ω_1 Ω_2 Ω_3	3	87.05	0.492	–	–	–	–	–
	Topological	Ω_1 Ω_2	2	86.86	0.478	–	–	–	–	–
	Topological	$1/^{2}\chi$	1	86.68	0.471	–	–	–	–	–
	GDIs	[$^{V/In}_{1}$DE(Cb)] [$^{P/Pd}_{1}$CV(Cb)]	2	79.15	0.412	71.40	0.44	71.58	0.049	28.5
	Topological	WW x$_1$	2	78.70	0.570	–	–	–	–	–
Entropy (*S*)	GETAWAY + WHIM + topological	$^{v}I_{D,deg}$ TWC R$^+_2$(p)	3	97.96	0.711	97.17	–	–	–	–
	GETAWAY + WHIM + topological	$^{v}I_{D,deg}$ TWC	2	97.14	0.814	96.42	–	–	–	–
	GDIs	[$^{E/In}_{1}$N$_2$(Cb)] [$^{P/In}$3 MX][$^{P/Pd}_{3}$K]	3	97.41	0.802	96.23	0.852	95.48	0.049	175.24
	GDIs	[$^{E/In}_{1}$RA] [$^{E/In}_{1}$P$_3$(Cb)]	2	95.6	1.008	94.10	1.067	93.44	0.009	163.08
	GETAWAY	I_{SH} *HATS*$_8$(m) *R*$_3$(v)	3	95.84	1.016	93.45	–	–	–	–
	GETAWAY	I_{SH} *R*$_3$(v)	2	94.76	1.101	92.19	–	–	–	–
	GDI	[$^{A/Pd}_{2}$G(Cb)]	1	92.64	1.263	91.31	1.294	91.45	−0.038	201.36

Table 4. (Continued)

Property	Index	Descriptor	*n*	R^2	*s*	Q_{Loo}^2	s_{CV}	Q_{boot}^2	*Ysc*	*F*
	GETAWAY + WHIM + topological	$R_3(v)$	1	92.51	1.274	89.86	–	–	–	–
	Topological	$\chi^{[1/2]}$	1	91.10	1.400	–	–	–	–	–
	Topological	$x_1\ x_2$	2	81.72	2.060	–	–	–	–	–
	GETAWAY + WHIM + topological	$^{v}I_{D,deg}$ TWC $R^{+}{}_2(p)$	3	97.96	0.711	–	–	–	–	–
Molar volume (MV)	GETAWAY + WHIM + topological	Ks $R^{+}{}_6(u)$ $RT^{+}(m)$	3	92.01	1.825	75.96	–	–	–	–
	GETAWAY	$HATS_6(p)$ $RT^{+}(m)$ $R_1(v)$	3	90.33	2.008	69.27	–	–	–	–
	Topological	$^{5}W\ ^{6}W\ ^{7}W$	3	88.29	2.210	–	–	–	–	–
	GETAWAY + WHIM + topological	$^{v}I_D^M\ R^{+}{}_6(u)$	2	84.96	2.419	54.49	–	–	–	–
	GDIs	$[^{E/Pd}{}_2Q_1^E(Cb)]$ $[^{V/Pd}{}_1K^E]$ $[^{P/Pd}2\ M^E]$	3	84.49	2.3	52.55	4.04	55.04	0.123	25.41
	GETAWAY	$R^{+}{}_6(u)$ $R_4(v)$	2	81.79	2.662	45.49	–	–	–	–
	GETAWAY + WHIM + topological	$R_6(v)$	1	67.61	3.437	–	–	–	–	–
	GDIs	$[^{P/Pd}{}_2Q_1^E(Cb)]$ $[^{P/Pd}{}_1K^E]$	2	65.62	3.43	19.08	5.12	25.42	0.069	13.48
	Topological	$^{3}W\ ^{4}W$	2	62.76	3.807	–	–	–	–	–
	Topological	^{7}W	1	60.85	3.780	–	–	–	–	–

4.4.2 Case 2. Boiling Point of 28 Alkyl-Alcohols

The boiling point (*Bp*) of a set of 28 alkyl-alcohols (see Table 5) compiled by Kier and Hall[8b] was examined using the new GDIs. The statistical information for the best regressions with two, three, four and five parameters is depicted in Table 6.

It is interesting to observe that the models of two, four and five parameters include local indices for hydrogen atoms bonded to the oxygen atom and in the case of the

Table 5. Experimental and predicted values of the boiling point of alcohols R–OH used in this study.

Alcohol-R	Found (°C)	Calculated (°C)					
		A	B	C	D	E	F
$(CH_3)_2CH–$	82.3	91.12	86.15	83.88	82.66	82.9	91.1
$CH_3CH_2CH_2–$	97.2	102.76	101.13	99.47	100.37	96.0	97.4
$CH_3(CH_2)_3–$	117.7	114.88	116.37	117.83	119.08	115.2	113.6
$CH_3CH(CH_3)CH_2–$	107.8	112.75	109.72	109.46	109.73	108.0	109.0
$CH_3CH_2C(CH_3)_2–$	102.4	103.89	101.89	102.15	102.81	105.4	112.4
$CH_3CH_2CH_2CH(CH_3)–$	119.3	115.62	119.05	117.03	114.76	114.4	120.3
$CH_3CH(CH_3)CH_2CH_2–$	131.1	126.78	128.20	131.7	131.0	134.5	127.4
$CH_3CH_2CH(CH_3)CH_2–$	128.0	124.97	123.93	126.86	126.32	127.3	125.2
$CH_3(CH_2)_4–$	137.9	130.64	134.62	136.51	135.33	134.3	131.8
$CH_3C(CH_3)_2CH(CH_3)–$	120.4	129.27	119.95	119.08	121.26	129.3	123.0
$CH_3(CH_2)_2C(CH_3)_2–$	121.1	120.59	124.24	119.71	119.92	124.9	128.9
$(CH_3CH_2)_2C(CH_3)–$	122.4	118.02	121.85	121.69	122.10	121.9	126.3
$CH_3CH_2C(CH_3)_2CH_2–$	136.5	139.98	134.14	134.15	137.12	142.5	138.4
$CH_3CH(CH_3)CH_2CH(CH_3)–$	131.6	129.23	131.27	132.63	130.33	133.9	133.4
$CH_3CH(CH_3)CH(CH_3CH_2)–$	126.5	127.33	129.58	129.35	128.31	121.9	128.7
$CH_3CH(CH_3)CH(CH_3)CH_2–$	144.5	138.75	135.73	139.61	139.81	146.7	138.3
$CH_3CH_2CH_2CH(CH_3)CH_2–$	149.0	141.29	151.02	144.64	148.25	146.4	143.4
$CH_3(CH_2)_5–$	157.6	156.23	155.95	159.62	158.59	153.4	169.8
$(CH_3CH(CH_3))_2CH–$	138.7	145.49	144.90	147.23	142.04	136.4	139.0
$CH_3CH(CH_3)CH_2CH(CH_3)CH_2–$	159.0	156.47	156.70	158.59	158.49	165.5	157.7
$(CH_3CH_2)_3C–$	142.0	143.37	141.09	142.76	143.20	138.6	138.5
$CH_3(CH_2)_6–$	176.4	174.33	175.11	175.74	173.95	172.5	172.2
$(CH_3CH_2CH_2)_2(CH_3)C–$	161.0	158.08	162.53	159.36	160.77	160.9	161.3
$(CH_3(CH_2)_3)(CH_3CH_2)(CH_3)C–$	163.0	160.11	164.08	163.14	163.32	160.5	162.7
$CH_3CH(CH_3)CH_2(CH_2)_4–$	188.0	191.38	191.92	189.72	190.92	191.6	188.3
$CH_3(CH_2)_7–$	195.1	202.78	197.26	196.61	195.08	191.6	193.0
$CH_3(CH_2)_5C(CH_3)_2–$	178.0	181.15	177.63	176.33	179.95	182.2	188.4
$(CH_3CH_2CH_2)_2(CH_3CH_2)C–$	182.0	179.22	180.48	181.49	180.54	177.6	177.0

Calculated values using: A) GDI, Equation 9; B) GDI, Equation 10; C) GDI, Equation 11; D) Equation 12; E) Spectral Moments; F) E-State.

Table 6. Result obtained by modeling boiling point of 28 alkyl-alcohols.

Index	*n*	R^2	*s*	Q_{Loo}^2	Q_{boot}^2	*Ysc*	s_{CV}	*F*
GDIs (Equation 9)	2	97.18	4.91	96.42	96.39	0.003	5.30	436.61
GDIs (Equation 10)	3	99.07	2.91	98.77	98.71	0.03	3.09	848.57
GDIs (Equation 11)	4	99.53	2.14	99.38	99.29	0.091	2.2	1207.9
GDIs (Equation 12)	5	99.61	2.15	99.4	99.3	0.106	2.15	1136.6
Local spectral moments	5	0.982	4.2	*	*	*	*	23.8
E-State/encounter parameters	3	0.926	5.8	*	*	*	*	204

* Value not reported

bivariate models one of variables is the local index for the oxygen atom. This is perfectly explainable according to chemistry, because the boiling point depends directly on the strength of the hydrogen bridges and this is certainly quantified with the local indices for the hydrogen atom once directly bonded to the oxygen atom.

In conclusion, the best linear regression model obtained to describe the BP of these chemicals, by using indices of derivative molecular graph is given below, respectively:

$$Bp = 59.26\ (\pm 2.88) + 2.29\ (\pm 0.09)\ (E/In)\Delta_{(H-O)} + 1.01\ (\pm 0.09)\ (E/In)\Delta_O \quad (9)$$

$N = 28 \quad R^2 = 97.18 \quad Q_{Loo}^2 = 96.42 \quad s = 4.91\,^{\circ}C$

$s_{CV} = 5.3\,^{\circ}C \quad Q_{boot}^2 = 96.39 \quad Ysc = 0.003 \quad F = 436.61$

$$Bp = 68.37\ (\pm 3.64) + 1.00\ (\pm 0.09)\ (P/In)N_3(Cb) + 2.26\ (\pm 0.21)\ (P2/IN)N_1(Cb) - 3.74\ (\pm 0.40)\ (P3/In)G \quad (10)$$

$N = 28 \quad R^2 = 99.07 \quad Q_{Loo}^2 = 98.77 \quad s = 2.91\,^{\circ}C$

$s_{CV} = 3.09\,^{\circ}C \quad Q_{boot}^2 = 98.71 \quad Ysc = 0.03 \quad F = 848.57$

$$Bp = 59.37\ (\pm 3.59) + 18.82\ (\pm 1.37)\ (P/In)N_3(Cb) - 16.05\ (\pm 1.27)\ (P/In)N_3 + 1.87\ (\pm 0.15)\ (A/In)\Delta_{(H-O)} \quad (11)$$

$N = 28 \quad R^2 = 99.53 \quad Q_{Loo}^2 = 99.38 \quad s = 2.14\,^{\circ}C$

$s_{CV} = 2.2\,^{\circ}C \quad Q_{boot}^2 = 99.29 \quad Ysc = 0.091 \quad F = 1207.89$

$$Bp = 75.79\ (\pm 5.13) + 14.26\ (\pm 1.37)\ (P/In)N_3\ (Cb) + 1.22\ (\pm 0.28)\ (P2/In)N_1 - 16.62\ (\pm 1.24)\ (E/In)\ G(Cb) - 9.68\ (\pm 1.68)\ (P/In)N_3 + 0.94\ (\pm 0.24)\ (A/In)\Delta_{(H-O)} \quad (12)$$

$N = 28 \quad R^2 = 99.61 \quad Q_{Loo}^2 = 99.4 \quad s = 1.95^{\circ} \quad s_{CV} = 2.15^{\circ}$

$Q_{boot}^2 = 99.3 \quad Ysc = 0.106 \quad F = 1136.6$

The values of experimental and calculated values of the Bp for the data set (fourth models) are given in Table 5, and the linear relationships between them (equation of five parameters) are illustrated in Figures 4.

These models (Equations 10 and 12) explain from 98.8% to 99.5% of the variance of the experimental *Bp* values. Similar results were reported by Estrada and Molina,[14a] and Kier and Hall[8b] by using spectral moment and E-state MDs, respectively. The statistical parameters of the best equa-

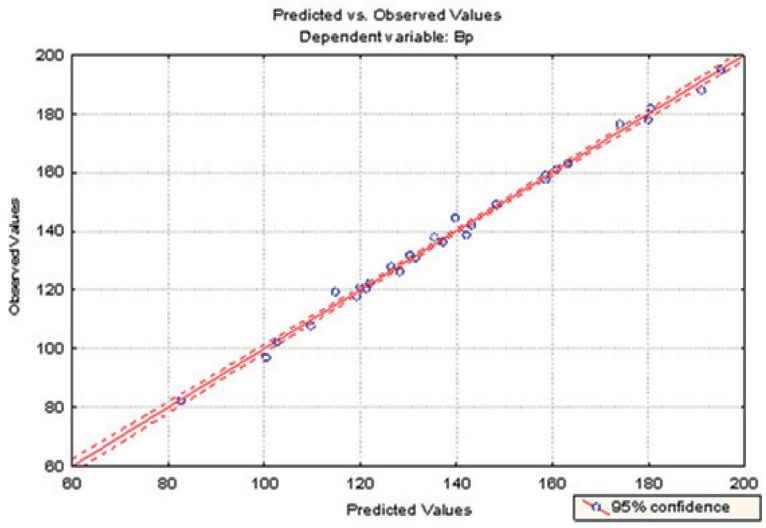


Figure 4. Scatter plot between experimental and calculated (by Equation 12) boiling points of data set containing 28 alcohols.

tions obtained by these authors are given in Table 6. Unfortunately, these authors (Estrada and Molina,[14a] as well as Kier and Hall[8b]) did not report the results of the LOO.

However, it is remarkable that our models explain a higher percentage of the variance of the experimental Bp values than the previously developed models, showing a decrease in the standard error of 46.92% and 44.14% (models of 3 and 5 parameters), with regard to the results previously achieved by Estrada and Molina[14a] and Kier and Hall,[8b] respectively (see Table 6).

4.4.3 Case 3. Phisicochemical Property of Polychlorinated bi-Phenyls (pcbs)

It is well-known that polychlorinated biphenyls are organic contaminants. It has been demonstrated that they are toxic, lipophilic and tend to be bioaccumulated. Four of the physicochemical properties of environmental relevance for PCB congeners have been chosen: melting point (*mp*), octanol water partition coefficient (log *P*), Henry's law constant (*H*), and aqueous water coefficient, expressed as the negative logarithm (log Y_W).

The statistical parameters of the best four, three, two, and one variable models for the four physicochemical properties are reported in the Table 7, 8, 9 and 10, ordered with respect to the decreasing values of the predictive ability (Q_{Loo}^2).

Below are the best linear regression models (Equations 13, 14, 15 and 16) for each of the studied properties:

Partition coefficient *n*-octanol-water (log *P*)

$$\begin{aligned}\log P = &-20.05\ (\pm 1.01) + 9.10\ (\pm 0.34)\ (A3/Pd)N_2^E \\ &+ 0.007\ (\pm 5.6 \cdot 10^{-4})\ (P3/Pd)N_1 \\ &+ 0.12\ (\pm 0.05)\ (A3/In)N_3^E\ (HL) \\ &- 0.23\ (\pm 0.01)\ (V3/In)\ A(Cb)\end{aligned} \quad (13)$$

$N = 139 \quad R^2 = 96.1 \quad s = 0.154 \quad Q_{Loo}^2 = 95.8 \quad s_{CV} = 0.157$

$Q_{boot}^2 = 95.75 \quad Ysc = -0.01 \quad F = 816.22$

Table 7. Data set PCB: Molecular descriptors and statistical information for the best regressions of the log *P* with one, two, three, and four variables.

Index	Descriptor	*n*	R^2	*s*	Q_{Loo}^2	s_{CV}	Q_{boot}^2	*Ysc*	*F*
All	λ_1^{LP} (P/W)^4 L1m Ts	4	96.4	–	96.2	–	–	–	–
GETAWAY + WHIM	Ts $HATS_6$(m) R_5(u) R_4(m)	4	96.2	–	96.0	–	–	–	–
All	ATS4m L1m Ts	3	96.1	–	95.9	–	–	–	–
GETAWAY	H_5(m) H_2(e) R_6(e) $R^+{}_4$(p)	4	96.2	–	95.9	–	–	–	–
GDIs	$[^{A/Pd}{}_3N_2^E][^{P/Pd}{}_3N_1]$ $[^{A/In}{}_3N_3^E$(HL)] $[^{V/In}{}_3$A(Cb)]	4	96.1	0.154	95.8	0.157	95.8	−0.01	816.2
GETAWAY + WHIM	Ts As R_4(m)	3	96.0	–	95.8	–	–	–	–
Topological	$^2X^v$ BIC λ_1^{LP} PCR	4	96.0	–	95.7	–	–	–	–
GETAWAY	H_2(p) $R^+{}_4$(m) R_6(e)	3	95.9	–	95.7	–	–	–	–
GDIs	$[^{A/Pd}{}_3N_2^E]$ $][^{P/Pd}{}_3N_1]$ $[^{V/In}{}_3$PN(Cb)]	3	95.9	0.157	95.6	0.16	95.6	−0.016	1048.7
Topological	$^2X^v$ SIC PCR	3	95.9	–	95.6	–	–	–	–
BCUT	BELm8 BEHp1 BELp2 BELp8	4	95.9	–	95.6	–	–	–	–
WHIM	E1u L2m Ts Av	4	95.7	–	95.4	–	–	–	–
WHIM	E1m Ts Au	3	95.7	–	95.4	–	–	–	–
All (topological)	$^2X^v$ PCR	2	95.6	–	95.4	–	–	–	–
BCUT	BEHp1 BELp2 BELp8	3	95.5	–	95.2	–	–	–	–
GETAWAY + WHIM	L1u H_2(p)	2	95.2	–	95.0	–	–	–	–
Broto-Moreau	ATS6m ATS6v ATS8v ATS8e	4	95.4	–	95.0	–	–	–	–
GETAWAY	HATS(u) H_2(e)	2	95.2	–	95.0	–	–	–	–
WHIM	L1u As	2	95.2	–	95.0	–	–	–	–
Broto-Moreau	ATS6m ATS6v ATS8e	3	95.0	–	94.7	–	–	–	–
Broto-Moreau	ATS4m ATS7e	2	94.5	–	94.2	–	–	–	–
All (WHIM)	Tu	1	94.1	–	93.9	–	–	–	–
BCUT	BELe2 BELe4	2	94.0	–	93.7	–	–	–	–
Broto-Moreau	ATS7e	1	93.8	–	93.6	–	–	–	–
BCUT	BELv2	1	93.3	–	93.1	–	–	–	–
topological	DECC	1	92.8	–	92.6	–	–	–	–
GETAWAY	R^2(v)	1	92.6	–	92.5	–	–	–	–
GDI	$[^{E/In}{}_2N_2^E]$ $][^{E/In}{}_3$G]	2	91.4	0.227	90.9	0.229	90.9	−0.025	720.62
GDI	$[^{E/Pd}{}_5N_2^E]$	1	85.5	0.293	85.1	0.294	85.35	−0.026	806.23
Constitutional	AMW	1	84.8	–	84.5	–	–	–	–

Aqueous water coefficient (log Y_W)

$$\log Yw = -14.23\ (\pm 1.17) + 0.51\ (\pm 0.10)\ (A9/PdI)MN^E - 0.03\ (\pm 0.005)\ (E3/Pd)M + 5.87\ (\pm 0.36)\ (P4/Pd)N_2^E + 0.64\ (\pm 0.15)\ (E4/Pd)\ N_3^E(HT) \quad (14)$$

$N = 87$ $R^2 = 89.1$ $s = 0.304$ $Q_{Loo}{}^2 = 88.1$ $s_{CV} = 0.308$ $Q_{boot}{}^2 = 85.73$ $Ysc = 0.002$ $F = 167.52$

Henry's law constant (H)

$$H = 44.51\ (\pm 2.18)\ -0.25(0.06)\ (A7/Pd)K(HT) - 14.64\ (\pm 0.66)\ (A9/(In)N_2^E\ (HL) - 4.79\ (\pm 0.93)\ (V9/In)CV + 0.13\ (\pm 0.008)\ (V3/In)Q_1(Cb) \quad (15)$$

$N = 21$ $R^2 = 98.0$ $s = 0.386$ $Q_{Loo}{}^2 = 96.2$ $s_{CV} = 0.466$ $Q_{boot}{}^2 = 95.87$ $Ysc = 0.12$ $F = 195.63$

Melting point (Mp)

$$Mp = 456.62\ (\pm 85.20) + 2.32\ (\pm 0.21)\ (E3/In)Q_3 - 13.76\ (\pm 2.60)\ (V/In)\ RA - 34.24\ (\pm 6.38)\ (A3/In)MX^E\ (HT) - 0.033\ (0.008)\ (P3/Pd)V\ (Cb) \quad (16)$$

$N = 79$ $R^2 = 84.3$ $s = 1.94$ $Q_{Loo}{}^2 = 81.6$ $s_{CV} = 2.86$ $Q_{boot}{}^2 = 81.24$ $Ysc = 0.01$ $F = 100.9$

It is interesting to highlight that in all cases the models achieved with the new GDIs show statistical parameters comparable to the best models reported in the literature (see Table 7, 8, 9 and 10) and in some cases (see Table 8 and 9) four variable models obtained using the GDI method possess statistical parameters superior to those reported by other authors so far.[15]

4.4.4 Case 4. Biological Activities of 22 N,N-Dimethyl-2-halophenethylamines

This final data set is comprised of 22 *N,N*-dimethyl-2-halophenethylamines and has been broadly used by other spe-

Table 8. Data set PCB: Molecular descriptors and statistical information for the best regressions of the aqueous activity coefficient (logYw) with one, two, three, and four variables.

Index	Descriptor	n	R^2	s	$Q_{Loo}{}^2$	s_{CV}	$Q_{boot}{}^2$	Ysc	F
GDIs	$[^{A/Pd}9MN^E][^{E/Pd}3\ M][^{P/Pd}{}_4N_2{}^E][^{E/Pd}{}_4N_3{}^E(HT)]$	4	89.1	0.304	88.1	0.308	85.73	0.002	167.52
All (GETAWAY + WHIM)	Tv I_{TH} $R_4(v)$ $R^+{}_1(p)$	4	89.0	–	87.6	–	–	–	–
GETAWAY	I_{SH} $HATS_4(u)$ $H_2(e)$ $R^+{}_1(p)$	4	88.6	–	87.3	–	–	–	–
All	^2X I_{SH} $RT^+(e)$	3	87.8	–	86.6	–	–	–	–
GETAWAY + WHIM	Tp $HATS_5(e)$ $RT^+(e)$	3	87.7	–	86.3	–	–	–	–
GDIs	$[^{E/Pd}3\ M]$ $[^{A/Pd}3\ MN(Cb)]$ $[^{P/Pd}{}_4N_2{}^E]$	3	87.1	0.328	85.8	0.336	85.33	−0.01	187.14
GETAWAY	$HATS_5(u)$ $H_2(m)$ $RT^+(e)$	3	86.8	–	85.3	–	–	–	–
WHIM	L1u P1p Gm Kv	4	86.5	–	84.8	–	–	–	–
All	^2X $R^+{}_5(e)$	2	85.8	–	84.7	–	–	–	–
GETAWAY + WHIM	I_{SH} $R^2(m)$	2	85.1	–	84.1	–	–	–	–
Topological	^2X $^{V}\overline{I}^{E}{}_D$,deg $\lambda_1{}^{LP}$ PCD	4	85.8	–	84.0	–	–	–	–
WHIM	L1u Gu Ke	3	85.4	–	83.9	–	–	–	–
Topological	$^1\overline{X}$CIC $\lambda_1{}^{LP}$	3	85.3	–	83.8	–	–	–	–
Topological	$^1\overline{X}\lambda_1{}^{LP}$	2	84.5	–	83.5	–	–	–	–
GDIs	$[^{E/Pd}3M]$ $[^{P/Pd}{}_4N_2{}^E]$	2	84.5	0.358	83.3	0.365	82.43	−0.021	229.46
BCUT	BEHm7 BEHv3 BEHe4 BEHp5	4	84.3	–	82.1	–	–	–	–
All (topological)	^2X	1	82.6	–	81.8	–	–	–	–
WHIM	E2p Tu	2	82.6	–	81.5	–	–	–	–
BCUT	BEHm4 BELm2 BEHp5	3	82.8	–	81.4	–	–	–	–
BCUT	BELv2 BEHp5	2	82.4	–	81.2	–	–	–	–
Broto-Moreau	ATS4v ATS7e	2	82.5	–	81.0	–	–	–	–
Broto-Moreau	ATS8m ATS7v ATS5e	3	82.7	–	80.7	–	–	–	–
Broto-Moreau	ATS1e	1	81.0	–	80.2	–	–	–	–
Broto-Moreau	ATS4m ATS5v ATS8v ATS8e	4	82.6	–	80.1	–	–	–	–
Constitutional	Sv	1	80.9	–	80.1	–	–	–	–
GETAWAY + WHIM	$R^2(m)$	1	80.9	–	80.0	–	–	–	–
BCUT	BELe2	1	80.8	–	79.9	–	–	–	–
WHIM	Ae	1	78.3	–	77.4	–	–	–	–
GDI	$[^{P/Pd}{}_4N_2{}^E]$	1	79.2	0.412	77.2	0.427	77.18	−0.03	324.03

Table 9. Data set PCB: Molecular descriptors and statistical information for the best regressions of Henry's law constant (*H*) with one, two, three, and four variables.

Index	Descriptor	*n*	R^2	*s*	Q_{Loo}^2	s_{CV}	Q_{boot}^2	*Ysc*	*F*
GDIs	$[^{A/Pd}{}_7K(HT)]$ $[^{A/In}{}_9N_2{}^E(HL)]$ $[^{V/In}{}_9CV]$ $[^{V/In}{}_3Q_1(Cb)]$	4	98.0	0.386	96.2	0.466	95.87	0.12	195.6
All (GETAWAY + WHIM)	Du $HATS_4(m)$ $R+_7(v)$ $R_7(p)$	4	97.0	–	95.1	–	–	–	–
GETAWAY	$H_4(p)$ $R_4(m)$ $R_7(v)$ $R^+{}_7(e)$	4	97.2	–	94.8	–	–	–	–
GDIs	$[^{A/In}{}_9N_2{}^E(HL)]$ $[^{V/In}{}_3Q_1(Cb)]$ $[^{V/In}{}_10\ M^E(HT)]$	3	95.9	0.538	94.0	0.586	93.44	0.065	131.5
All (GETAWAY)	$HATS_7(v)$ $R_4(m)$ $R^+{}_7(e)$	3	95.5	–	93.4	–	–	–	–
GDIs	$[^{A/In}{}_9N_2{}^E(AH)]$ $[^{P/In}3\ V(Cb)]$	2	92.7	0.693	90.4	0.739	89.91	−0.015	114.8
Topological	S_G $(P/W)^4$ $(P/W)^5$ $\overline{W}$	4	93.1	–	88.0	–	–	–	–
All (GETAWAY)	$HATS_4(m)$ $R_3(e)$	2	91.7	–	86.4	–	–	–	–
WHIM	E1u L2v E1e P1s	4	92.2	–	85.9	–	–	–	–
Topological	UNIP $(P/W)^4$ $\overline{W}$	3	91.6	–	85.3	–	–	–	–
Broto-Moreau	ATS4m ATS6m ATS8m	3	89.9	–	81.1	–	–	–	–
Broto-Moreau	ATS4m ATS3m ATS8m ATS8e	4	91.0	–	80.1	–	–	–	–
WHIM	P1m P2e E1e	3	87.7	–	78.3	–	–	–	–
Topological	$^2\overline{X}$ $(P/W)^4$	2	84.4	–	76.4	–	–	–	–
BCUT	BELv6 BEHe4 BEHp7 BEHp8	4	85.3	–	71.7	–	–	–	–
All (GETAWAY)	$R_8(u)$	1	74.3	–	66.3	–	–	–	–
BCUT	BEHe4 BEHp7 BELp6	3	78.4	–	63.8	–	–	–	–
Broto-Moreau	ATS4m ATS8m	2	76.6	–	63.2	–	–	–	–
Topological	$(P/W)^4$	1	70.4	–	61.4	–	–	–	–
WHIM	E1e E3e	2	71.2	–	58.7	–	–	–	–
GDIs	$[^{E/In}5\ V(Cb)]$	1	60.8	1.57	54.4	1.61	54.66	−0.015	29.51
WHIM	E1e	1	64.1	–	52.9	–	–	–	–
BCUT	BELm6 BEHp7	2	50.2	–	24.3	–	–	–	–
Broto-Moreau	ATS4m	1	36.6	–	19.8	–	–	–	–
BCUT	BEHp7	1	23.1	–	5.4	–	–	–	–

Table 10. Data Set PCB: Molecular descriptors and statistical information for the best regressions of the melting point (*mp*) with one, two, three, and four variables.

Index	Descriptor	*n*	R^2	s	Q_{Loo}^2	s_{CV}	Q_{boot}^2	*Ysc*	*F*
All	TIC ATS7e G1s Tu	4	84.6	–	82.0	–	–	–	–
GDIs	$[^{E/In}{}_3Q_3][^{V/In}RA][^{A/In}3\ MX^E(HT)][^{P/Pd}3\ V(Cb)]$	4	84.3	2.07	81.6	2.86	81.24	0.01	100.9
GETAWAY + WHIM	Gm $H_3(u)$ $R_1(e)$ $R^+{}_2(e)$	4	83.9	–	81.3	–	–	–	–
GETAWAY	I_{TH} $H_3(u)$ $HATS_1(u)$ $R^+{}_2(e)$	4	83.7	–	80.8	–	–	–	–
All	$V\overline{I}^E{}_{D,deg}$ Tu $R^+{}_2(m)$	3	83.2	–	80.8	–	–	–	–
WHIM	G1mL2vTuAm	4	83.7	–	80.4	–	–	–	–
GETAWAY + WHIM	Tu Gm $R^+{}_2$ (v)	3	82.4	–	80.0	–	–	–	–
GETAWAY	I_{TH} HT(e) $R_1(p)$	3	79.0	–	81.6	–	–	–	–
WHIM	Tu Am Gm	3	81.7	–	78.9	–	–	–	–
Topological	$V\overline{I}^E{}_{D,deg}$ CIC $(P/W)^2$ $(P/W)^5$	4	81.2	–	77.9	–	–	–	–
All	$V\overline{I}^E{}_{D,deg}$ Tu	2	79.9	–	77.5	–	–	–	–
GDIs	$[^{E/In}{}_3Q_3][^{V/In}RA][^{A/In}3\ MX^E(HT)]$	3	80.5	6.52	77.4	6.87	72.06	0.001	74.9
GETAWAY + WHIM	Tu Gs	2	79.8	–	77.4	–	–	–	–
Topological	^2X $V\overline{I}^E{}_{D,deg}$ $\overline{\sigma}$	3	79.3	–	76.3	–	–	–	–
GETAWAY	I_{sH} $R_3(e)$	2	74.4	–	77.3	–	–	–	–
GDIs	$[^{E/In}{}_3Q_3][^{V/In}RA]$	2	77.5	7.32	74.8	7.19	70.08	−0.01	101.7
Topological	TIC UNIP	2	75.7	–	73.0	–	–	–	–
BCUT	BELm5 BELv4 BELp3 BELp8	4	63.7	–	69.3	–	–	–	–
All (WHIM)	Tu	1	71.2	–	69.0	–	–	–	–
BCUT	BELm5 BELv4 BELp3	3	71.9	–	68.5	–	–	–	–
BCUT	BELm5 BELp3	2	69.8	–	67.0	–	–	–	–
Broto-Moreau	ATS5m ATS7v	2	69.1	–	65.8	–	–	–	–
Broto-Moreau	ATS7e	1	68.1	–	65.7	–	–	–	–
Topological	UNIP	1	67.4	–	65.0	–	–	–	–
BCUT	BELv2	1	66.5	–	64.0	–	–	–	–
GETAWAY	$R_7(v)$	1	66.1	–	63.5	–	–	–	–
Constitutional	AMW	1	61.5	–	58.7	–	–	–	–
GDI	$[^{V/In}{}_3Q_1(Cb)]$	1	61.4	11.86	58.6	11.59	54.49	−0.02	97.83

cialists in QSAR studies.[15] The response is the antagonism of these compounds to epinephrine in rat models (log 1/ED_{50}). Table 11 shows the statistical information for the best regressions with one, two, three and four MDs, ordered with respect to decreasing values of predictive ability (Q_{Loo}^2).

The best linear regression models obtained to describe the Bp of these chemicals, using GDI are given below, respectively:

$$log\left(\frac{1}{ED_{50}}\right) = 7.46\ (\pm 0.38) - 0.22\ (\pm 0.02)\ (V/In)Q_2 - 9 \times 10^{-5}\ (\pm 2 \times 10^{-5})\ (V/Pd)S(AH) + (0.024\ (\pm 7.8 \times 10^{-4})\ (V/Pd)P_3(Cb) + 0.57\ (\pm 0.10)\ (A7/Pd)DE^E(Cb) \quad (17)$$

$N = 22 \quad R^2 = 98.44 \quad s = 0.0789 \quad Q_{Loo}^2 = 97.10$

$s_{CV} = 0.0945 \quad Q_{boot}^2 = 94.83 \quad Ysc = 0.118 \quad F = 267.62$

$$log\left(\frac{1}{ED_{50}}\right) = 8.67\ (\pm 0.47) - 0.19\ (\pm 0.03)\ (V/In)Q_2 + 0.023\ (\pm 0.001)\ (V/Pd)P_3(Cb) + 1.79\ (\pm 0.63)\ (E/In)Q_1^E \quad (18)$$

$N = 22 \quad R^2 = 96.65 \quad s = 0.112 \quad Q_{Loo}^2 = 95.06$

$s_{CV} = 0.123 < + < Q_{boot}^2 = 94.30 \quad Ysc = 0.053 \quad F = 173.2$

$$Log\left(\frac{1}{ED_{50}}\right) = 8.42\ (\pm 0.54) - 0.23\ (\pm 0.02)\ (V/In)Q_2 + 0.024\ (\pm 0.001)\ (V/Pd)P_2(Cb) \quad (19)$$

$N = 22 \quad R^2 = 95.21 \quad s = 0.13 \quad Q_{Loo}^2 = 93.56 \quad s_{CV} = 0.141$

$Q_{boot}^2 = 93.53 \quad Ysc = 0.021 \quad F = 188.95$

$$log\left(\frac{1}{ED_{50}}\right) = +4.21\ (\pm 0.61) + 0.014\ (\pm 0.002)\ (V/Pd)P_3(Cb) \quad (20)$$

$N = 22 \quad R^2 = 73.34 \quad s = 0.30 \quad Q_{Loo}^2 = 67.65 \quad s_{CV} = 0.3157$

$Q_{boot}^2 = 69.54 \quad Ysc = -0.02 \quad F = 55.03$

Table 11. Data set phenetylamines: Molecular descriptors and statistical information for the best regressions of the adrenergic blocking activity log(1/ED_{50}) with one, two, three, and four variables.

Index	Descriptor	*n*	R^2	*s*	Q_{Loo}^2	s_{CV}	Q_{boot}^2	*Ysc*	*F*
All	MSD As H_2(v) R_4(u)	4	98.5	–	97.7	–	–	–	–
GETAWAY + WHIM	E2v P2s Tv $HATS_1$(m)	4	98.4	–	97.4	–	–	–	–
WHIM	E1v G2p P1s Tv	4	98.2	–	97.4	–	–	–	–
GDIs	$[^{V/In}Q^2][^{V/In}S(AH)][^{V/Pd}P_3(Cb)][^{A/Pd}{}_7DE(Cb)]$	4	98.4	0.079	97.1	0.095	94.83	0.118	267.6
All	$(P/W)^4$ Tv R_4(e)	3	97.6	–	96.4	–	–	–	–
GETAWAY + WHIM	P1s Tv R_4(e)	3	97.2	–	95.9	–	–	–	–
WHIM	E1v P1s Tv	3	97.2	–	95.8	–	–	–	–
Topological	ICPX IAC 3AECC	4	97.2	–	95.8	–	–	–	–
GDIs	$[^{V/In}Q^2]$ $[^{V/Pd}P_3(Cb)][^{E/In}Q^E{}_1)]$	3	96.7	0.112	95.1	0.123	94.3	0.053	173.2
BCUT	BELm2 BELm5 BEHv5 BEHv6	4	96.1	–	93.8	–	–	–	–
GDIs	$[^{V/In}Q^2]$ $[^{V/Pd}P_2(Cb)]$	2	95.2	0.13	93.6	0.141	93.53	0.021	188.9
GETAWAY	HATS3(u) HATS(u) H4(m) H1(v)	4	95.8	–	93.1	–	–	–	–
Topological	IAC IDVE 3	3	94.4	–	92.5	–	–	–	–
All	Ms MSD	2	94.3	–	92.2	–	–	–	–
GETAWAY	$HATS_6$(v) H_4(e) $HATS_7$(p)	3	94.1	–	91.2	–	–	–	–
BCUT	BEHm5 BEHv6 BEHp5	3	93.8	–	91.2	–	–	–	–
GETAWAY + WHIM	E3u L1v	2	93.9	–	90.8	–	–	–	–
BCUT	BEHv6 BEHp5	2	91.5	–	89.2	–	–	–	–
Broto-Moreau	ATS2e ATS7e ATS6p	3	92.5	–	86.8	–	–	–	–
GETAWAY	H_4(v) H_3(p)	2	85.6	–	81.0	–	–	–	–
Topological	MSD VM1	2	85.1	–	80.1	–	–	–	–
All (WHIM)	Tv	1	83.2	–	79.4	–	–	–	–
Broto-Moreau	*ATS6v ATS8v ATS5e ATS8e*	*4*	*91.1*	–	*76.3*	–	–	–	–
GDI	$[^{V/Pd}P_3(Cb)]$	1	73.3	0.3	67.7	0.316	69.54	−0.02	55.03
Constitutional	Sv Mv	2	75.0	–	66.7	–	–	–	–
BCUT	BEHv6	1	70.1	–	66.3	–	–	–	–
Constitutional	Sp	1	71.5	–	64.9	–	–	–	–
Constitutional	Sv Se nCl	3	75.3	–	64.5	–	–	–	–
Broto-Moreau	ATS3v ATS5e	2	70.4	–	61.2	–	–	–	–
GETAWAY	H4(m)	1	65.2	–	57.0	–	–	–	–
Topological	VIDE	1	56.6	–	48.2	–	–	–	–
Broto-Moreau	ATS8v	1	44.0	–	31.3	–	–	–	–

Table 12. Value of log(1/ED_{50}) calculated with Equations 17, 18, 19, 20 and their residual ones.

No	Observed	Calculated				Residual			
		Eq. 17	Eq. 18	Eq. 19	Eq. 20	Eq. 17	Eq. 18	Eq. 19	Eq. 20
1	7.46	7.48	7.47	7.49	7.82	0.02	0.01	0.03	0.36
2	8.16	8.15	8.16	8.11	7.8	−0.01	0	−0.05	−0.36
3	8.68	8.84	8.71	8.67	8.31	0.16	0.03	−0.01	−0.37
4	8.89	8.89	8.99	8.91	8.64	0	0.1	0.02	−0.25
5	9.25	9.16	9.24	9.11	8.94	−0.09	−0.01	−0.14	−0.31
6	9.3	9.22	9.16	9.09	8.6	−0.08	−0.14	−0.21	−0.70
7	7.52	7.5	7.58	7.61	7.84	−0.02	0.06	0.09	0.32
8	8.16	8.1	8.06	8.07	8.32	−0.06	−0.1	−0.09	0.16
9	8.3	8.22	8.29	8.26	8.64	−0.08	−0.01	−0.04	0.34
10	8.4	8.49	8.5	8.41	8.93	0.09	0.1	0.01	0.53
11	8.46	8.45	8.42	8.43	8.55	−0.01	−0.04	−0.03	0.09
12	8.19	8.28	8.3	8.41	8.35	0.09	0.11	0.22	0.16
13	8.57	8.62	8.5	8.52	8.65	0.05	−0.07	−0.05	0.08
14	8.82	8.72	8.66	8.72	8.55	−0.1	−0.16	−0.1	−0.27
15	8.89	8.78	8.79	8.87	8.82	−0.11	−0.1	−0.02	−0.07
16	8.92	8.93	8.97	8.99	9.12	0.01	0.05	0.07	0.2
17	8.96	9.03	9.13	9.19	9.02	0.07	0.17	0.23	0.06
18	9	9.01	9.03	9.05	9.13	0.01	0.03	0.05	0.13
19	9.35	9.33	9.21	9.17	9.42	−0.02	−0.14	−0.18	0.07
20	9.22	9.29	9.36	9.38	9.32	0.07	0.14	0.16	0.1
21	9.3	9.32	9.43	9.5	9.2	0.02	0.13	0.2	−0.1
22	9.52	9.49	9.36	9.37	9.36	−0.03	−0.16	−0.15	−0.16

Table 12 shows the experimental (or observed) and calculated values attained with the equations 17, 18, 19 and 20, as well as their respective residual values.

4.5 External Validation for QSPR Models

While internal cross–validation methods give criteria on a model's robustness, the true predictive power should be evaluated on a set of compounds not employed during its development, i.e. following an external validation work flow. However, with difficulty in obtaining new experimentally tested compounds for external validation purposes, an alternative involves the splitting the initial dataset into training and test sets. A proper splitting method is essential to ensure representatives, diversity and independence of the training and test sets. In this section, QSPR models were obtained for the Polychlorinated bi-phenyls (PCBs) and *N*,*N*-dimethyl-2-halo-phenethylamines (Phenet) datasets, respectively, following an external validation scheme. In both cases, the datasets were split using the statistical technique *k*-Means Cluster Analysis (*k*-MCA). Clustering quality was taken into account using the intra and inter cluster sum of the squared errors (*SSE*), also known as the scatter. Then the construction of the training and test sets was randomly performed, selecting compounds from each cluster. In order to check for possible variability in the results due to biased selection of training and test compounds, a 10-fold external validation was carried out. The statistical parameters obtained with the repetitions were practically identical and thus parameters from the first external validation were considered.

Note that although the PCB dataset is comprised of 209 compounds, many of these do not have the corresponding experimental values for the 4 modeled physicochemical properties reported. As a result, the sizes of the training and test sets for the PCB data vary according to the modeled property (see Table 13).

Tables 14 and 15 show the statistic parameters of the best models with two, three and four MDs founded in each case. At can be observed in all the cases the obtained models show high predictive power.

Table 13. Partition of the *PCB* and *Phenet* data set for each property, for the external validation.

Set	PCB				Phenet
	H	Log *Yw*	Log *P*	*MP*	log 1/ED_{50}
Total compounds	209	209	209	209	22
Reported values	21	87	139	79	22
Missing values	188	122	70	130	0
Training objects	16	66	105	60	17
Test objects	5	21	34	19	5

Table 14. Data set PCB: molecular descriptors and statistical information for the best regressions of the melting point (*mp*) with two, three, and four variables *with external validation*.

Index	Descriptor	*N*	R^2	*s*	Q^2	s_{CV}	Q_{boot}^2	Q^2_{ext}	*Ysc*	F
Partition coefficient *N*-octanol-water (log *P*)										
GDI	$[^{P/Pd}{}_3N_1][^{A/In}{}_3N_3{}^E(HL)][^{V/In}{}_3A(Cb)][^{A/Pd}{}_3N_2{}^E]$	4	94.85	0.146	94.38	0.148	94.34	**97.87**	−0.005	460.22
GDI	$[^{P/Pd}{}_3N_1]$ $[^{A/Pd}{}_3N_2{}^E]$ $[^{V/In}{}_3PN(Cb)]$	3	94.68	0.147	94.28	0.150	94.26	**97.70**	−0.013	599.32
GDI	$[^{V/In}{}_3A(Cb)]$ $[^{A/Pd}{}_3N_2{}^E]$	2	88.58	0.215	87.70	0.220	87.63	**94.39**	−0.020	395.56
Aqueous water coefficient (log *YW*)										
GDI	$[^{A/Pd}9\ MN^E][^{E/Pd}3\ M][^{P/Pd}{}_4N_2{}^E][^{E/Pd}{}_4N_3{}^E(HT)]$	4	82.98	0.339	80.52	0.349	78.95	**93.84**	0.022	75.58
GDI	$[^{E/Pd}3\ M]$ $[^{P/Pd}{}_4N_2{}^E]$ $[^{A/Pd}3\ MN(Cb)]$	3	81.55	0.350	79.54	0.358	79.45	**93.39**	0.007	92.83
GDI	$[^{A/Pd}9\ MN^E]$ $[^{P/Pd}{}_4N_2{}^E]$	2	77.80	0.381	75.74	0.389	75.59	**91.50**	−0.009	112.17
Henry's law constant (*H*)										
GDI	$[^{A/Pd}{}_7K(HT)]$ $[^{A/In}{}_9N_2{}^E(HL)]$ $[^{V/In}{}_9CV]$ $[^{V/In}{}_3Q_1(Cb)]$	4	96.48	0.435	91.64	0.548	88.42	**98.75**	0.208	68.52
GDI	$[^{A/In}{}_9N_2{}^E(HL)]$ $[^{V/In}{}_9CV]$ $[^{P/In}3\ V(Cb)]$	3	93.91	0.546	88.55	0.641	86.75	**92.37**	0.103	56.54
GDI	$[^{A/In}{}_9N_2{}^E(HL)]$ $[^{P/In}3\ V(Cb)]$	2	91.18	0.629	85.80	0.714	85.51	**88.86**	0.035	62.00
Melting point (*Mp*)										
GDI	$[^{V/Pd}S(Cb)]$ $[^{V/In}RA]$ $[^{E/In}{}_3Q_3]$ $[^{A/In}3\ MX^E(HT)]$	4	81.13	17.01	76.83	18.03	75.68	**77.06**	0.032	58.04
GDI	$[^{A/In}{}_4Q_3{}^E]$ $[^{V/In}RA]$ $[^{E/In}{}_3Q_3]$	3	73.56	19.95	69.33	20.74	68.56	**77.11**	0.016	51.01
GDI	$[^{V/In}RA]$ $[^{E/In}{}_3Q_3]$	2	69.84	21.11	65.49	22.00	65.25	**78.33**	−0.049	64.86

Table 15. Data set phenetylamines: Molecular descriptors and statistical information for the best regressions of the adrenergic blocking activity [log(1/ED_{50})] with one, two, three, and four variables *with external validation.*

Index	Descriptor	*n*	R^2	*s*	Q^2	s_{CV}	Q_{boot}^2	Q^2_{ext}	*Ysc*	F
GDI	$[^{V/Pd}P_3(Cb)][^{V/Pd}P_2(Cb)][^{V/In}Q^2][^{E/In}Q_1{}^E]$	4	97.88	0.086	96.03	0.099	94.24	**95.79**	0.143	138.4
GDI	$[^{V/Pd}P_3(Cb)][^{V/In}Q^2][^{A/Pd}{}_7DE^E(Cb)]$	3	96.69	0.104	94.56	0.116	94.14	**95.78**	0.095	126.47
GDI	$[^{V/Pd}P_3(Cb)][^{V/In}Q^2]$	2	94.79	0.125	92.42	0.137	92.16	**94.74**	0.027	127.30
GDI	$[^{V/Pd}P_3(Cb)]$	1	77.79	0.250	68.58	0.279	71.75	**65.39**	−0.016	52.54

The values for the MDs used to construct the models in this section are available as Supporting Information (SI2).

5 Linear Independence and Structure/ Physicochemical Interpretation of New MDS

Although theoretical MDs have attracted increasing attention in recent years as valuable tools for different chemometric tasks, little progress has been made in the effort to address the interrogative on their interpretation in structural, or even better, physicochemical terms. This is probably due to the elusive nature of the task of rationalizing trends observed as a result of the application of strictly mathematical algorithms to chemical structure representations. Nonetheless a couple of papers could be mentioned in the literature dedicated to structural and physicochemical interpretation of well-known MDs, such as the molecular connectivity indices. Certainly, it is "healthy" for any new family MDs to have structural and physicochemical interpretations, but as it has been explained in several reports, this is a desirable property rather than an imperative one and the lack of interpretation for MDs does not demerit their usability.[12] On the other hand, more than finding interpretations, of far greater importance is the orthogonality of new MDs with respect to the existing ones as this delineates between novelty and "preexistence" in the sense that MDs collinear with existing ones do not codify distinct structural information, while the opposite is true. In this manuscript, we attempt to carry out both studies, first to evaluate the possible linear independence of the GDI with respect to all families of DRAGON's MDs (0D-3D) and preliminary efforts to give structural and physicochemical interpretations to the novel MDs.

5.1 Analysis of Molecular Information Captured by GDIs and Their Linear Independence

The primary objective of this section is to compare the information contained in the GDIs and those implemented in the DRAGON software,[1,21] as a measure of the practical utility of the discrete derivative approach in characterizing molecular structures. The choice of DRAGON's software is not arbitrary, as it is comprised of probably the most diverse collection of MDs defined so far in the literature, and correlation or even better orthogonality between a novel set of MDs and DRAGON's MDs would award reasonable credibility to the former. For this analysis, we use 41 heterogeneous molecules of DRAGON's sample data (methane not considered). The descriptor calculations were performed using DIVATI software, a new module of *TOMOCOMD-CARDD* program that offers rapid and low-computational-cost calculations of the proposed MDs.[19]

For this study, *factor analysis* using the principal components method is performed. This is a powerful tool used to condense the information contained in several variables into a reduced number of weighted composites. The theoretical aspects of this statistical technique have been extensively explained elsewhere.[23] The general objectives of factor analytical techniques are (1) data *reduction* and (2) *interpretation* of the underlying relationship between variables, i.e., to *classify* variables. In this context, factor loadings (or artificial variables) are obtained from original (MDs) variables. These factors capture most of the "essence" of the MDs because they are a linear combination of the original items. Because each factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are orthogonal to each other. Therefore, the first factor is generally more highly correlated with the variables than the other factors. Two important inferences could be made from this study (1) variables with a high loading in the same factor are correlated and this correlation will be greater the higher the loadings, (2) no correlation exists between variables having nonzero loadings in only different factors. The existence of linear independence has been claimed by Randic as one of the desirable attributes for novel TIs.

Factor analysis is performed with the STATISTICA software[24] and the "varimax normalized" is used as the rotational strategy to obtain a clear pattern of loadings, i.e., factors that are clearly marked by high loadings for some variables and low loadings for others. This rotation strategy maximizes the variances of the square *normalized factor loadings* (row factor loadings divided by square roots of the respective communalities) across variables for each factor, permitting a clearer interpretation of the factors without loss of orthogonality among them. In this analysis, only factor loadings greater than 0.60 are considered as "meaningful".

Table that reflects the factor loadings of all the MDs used in this study is available as Supporting Information (Table SI3). Table 16 shows the eigenvalues and the percentages of the explained variance by 10 principal factors of this analysis, which explain approximately 80.64% of the cumulative variance.

Table 16. Eigenvalues and percentages of the explained variance by ten principle factors.

Factors	Eigenvalue	% Total variance	% Cumulative variance
F1	1069.67	30.22	30.22
F2	483.65	13.66	43.88
F3	350.54	9.90	53.78
F4	190.79	5.39	59.17
F5	171.38	4.84	64.01
F6	148.75	4.20	68.21
F7	128.61	3.63	71.85
F8	117.81	3.33	75.18
F9	104.44	2.95	78.13
F10	89.03	2.51	80.64

An analysis of the factor loadings in Table SI3 reveals robust representativity in Factor 1 (30.21%) for GDIs and DRAGON MDs, particularly in the case of the former: molecular path counts, Broto-Moreau autocorrelation indices, eigenvalue-based indices, Burden eigenvalue descriptors, Randic molecular profiles, RDF Descriptors, *A* and *V* total size indices (WHIM descriptors), and to a lesser extent, molecular properties (Hypnotic-80, Infective-80 and GVWAI-80) and 2D frequency fingerprints. A similar trend is observed in Factor 2 (13.66%) with strong loadings for GDIs and DRAGON MDs (constitutional descriptors, topological descriptors, walk and paths counts, connectivity indices, information indices, 2D autocorrelations, spectral moments, Burden eigenvalue descriptors, eigenvalue-based descriptors, Randic Molecular profiles, geometric descriptors, 3D-MoRSE descriptors, WHIM descriptors and GETAWAY descriptors). The existence of correlation between GDIs and DRAGON MDs suggests that the former are able to capture structural information codified by the latter, although as can be appreciated the DRAGON MDs encompass a wide range of theories, formalisms and dimensions.

On the other hand, GDIs are solely loaded in Factor 3 (9.90%), Factor 4 (5.38%) and Factor 10 (2.51%); and almost exclusively in Factor 7 (3.63%) and F9 (2.95%) with very minimal loading from DRAGON's MDs (constitutional descriptors, eigenvalue-based indices, WHIM descriptors). The existence of orthogonality between GDIs and DRAGON's MDS, suggests that the former codify structural information not captured by the latter, which rationalizes the contribution of the new mathematical approach for the codification of the geometric space of a molecular structure.

An important inference from this study is that GDIs codify structural information not adequately described by DRAGON's MDs, in addition to capturing all the information codified by the latter indices, which suggests practical relevance of the discrete derivative approach in characterizing molecular structures.

5.2 Structural and Physicochemical Interpretation

5.2.1 Interpretation and Influence of Structural Changes on Total and Local (Atom) Derivative Indices

The structural influence on the total (whole molecule) and local (atom-based) GDIs may be revealed by examining several sets of calculations in which features are systematically varied. In this sense, the following structural effects on total and local derivative indices are investigated: a) effect of chain length, b) effect due to branching, c) effect across multiple bonds, and d) effect due to heteroatom change. The GDI tendencies due to these structural features are shown in Tables 17–20, respectively. These tables show the values of the atom-level GDIs (first using all possible connected-subgraphs and followed by all first-, second-, and third connected-subgraphs orders, separately). Besides, some total invariants were computed (N1, N2 and G). All

Table 17. The changes in GDIs due to chain lengthening in alkanes.

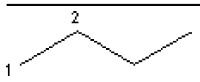
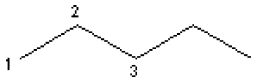
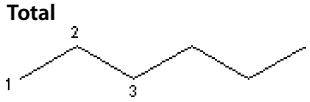
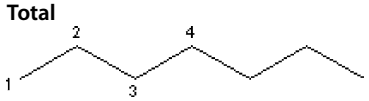
Atom (*i*)	${}^{E}\Delta$	${}^{E}\Delta_1$	${}^{E}\Delta_2$	${}^{E}\Delta_3$	$\|\|{}^{E}\overline{X}\|\|_1$	$\|\|{}^{E}\overline{X}\|\|_2$	$\|\|{}^{E}\overline{\xi}\|\|$
							
C_1	11.17	1.00	2.00	1.00	–	–	–
C_2	6.17	3.00	2.00	1.00	–	–	–
Total	–	–	–	–	**34.67**	**18.04**	**8.3**
							
C_1	17.33	1.00	2.50	3.00	–	–	–
C_2	9.33	3.00	3.50	2.00	–	–	–
C_3	7.33	4.00	4.00	2.00	–	–	–
Total	–	–	–	–	**60.67**	**28.79**	**11.39**
							
C_1	24.40	1.00	2.50	4.00	–	–	–
C_2	13.32	3.00	4.50	4.00	–	–	–
C_3	9.58	4.00	6.00	4.00	–	–	–
Total	–	–	–	–	**94.60**	**41.58**	**14.60**
							
C_1	32.30	1.00	2.50	4.50	–	–	–
C_2	18.03	3.00	4.50	5.50	–	–	–
C_3	12.58	4.00	7.00	6.33	–	–	–
C_4	11.17	4.00	8.00	6.67	–	–	–
Total	–	–	–	–	**137.00**	**56.38**	**17.95**

Table 18. Changes in GDIs due to branching in the pentanes' skeleton.

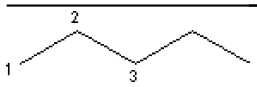
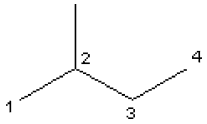
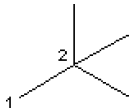
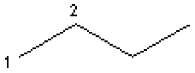
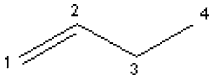
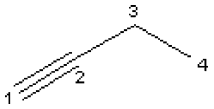
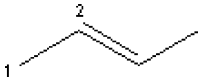
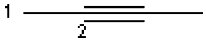
Atom (*i*)	${}^{E}\Delta$	${}^{E}\Delta_1$	${}^{E}\Delta_2$	${}^{E}\Delta_3$	$\|\|{}^{E}\overline{X}\|\|_1$	$\|\|{}^{E}\overline{X}\|\|_2$	$\|\|{}^{E}\overline{\xi}\|\|$
							
C_1	17.33	1.00	2.50	3.00	–	–	–
C_2	9.33	3.00	3.50	2.00	–	–	–
C_3	7.33	4.00	4.00	2.00	–	–	–
Total	–	–	–	–	**60.67**	**28.79**	**11.39**
							
C_1	12.08	2.00	7.17	6.25	–	–	–
C_2	8.71	7.00	6.06	4.67	–	–	–
C_3	7.78	4.00	8.89	2.42	–	–	–
C_4	13.90	1.00	3.83	6.25	–	–	–
Total	–	–	–	–	**54.55**	**24.93**	**10.66**
							
C_1	10.50	3.00	14.50	5.33	–	–	–
C_2	12.00	12.00	10.00	9.33	–	–	–
Total	–	–	–	–	**54.00**	**24.19**	**10.78**

Table 19. The influences of unsaturation on the GDIs in hydrocarbons.

Atom (*i*)	$^{E}\Delta$	$^{E}\Delta_1$	$^{E}\Delta_2$	$^{E}\Delta_3$	$\|\|^{E}\bar{X}\|\|_1$	$\|\|^{E}\bar{X}\|\|_2$	$\|\|^{E}\bar{\xi}\|\|$
							
C_1	11.17	1.00	2.00	1.00	–	–	–
C_2	6.17	3.00	2.00	1.00	–	–	–
Total	–	–	–	–	**34.67**	**18.04**	**8.30**
							
C_1	12,33	0.83	1.83	0.67	–	–	–
C_2	7.58	3.17	2.67	1.67	–	–	–
C_3	5.92	3.33	2.17	0.67	–	–	–
C_4	14.67	1.00	2.67	2.33	–	–	–
Total	–	–	–	–	**40.50**	**21.44**	**9.49**
							
C_1	16.44	0.83	2.50	1.58	–	–	–
C_2	9.78	3.83	3.83	2.83	–	–	–
C_3	7.25	4.00	3.17	1.17	–	–	–
C_4	19.75	1.00	3.50	4.08	–	–	–
Total	–	–	–	–	**53.22**	**28.44**	**12.32**
							
C_1	13.67	1.67	3.33	2.67	–	–	–
C_2	8.67	3.67	3.33	2.67	–	–	–
Total	–	–	–	–	**44.67**	**22.89**	**10.88**
							
C_1	16.58	2.50	5.00	4.50	–	–	–
C_2	11.58	4.50	5.00	4.50	–	–	–
Total	–	–	–	–	**56.33**	**28.61**	**13.86**

MDs were calculated using Pauling-electronegativity as atom-

As can be observed in Tables 17–20, the GDIs encode information about size, branching, multiple bonds and heteroatom content. Firstly, chain lengthening (from butane to heptane) is accompanied with a progressive increase in the whole-molecule GDI values. Therefore, for a homologous series the GDI increases adequately with the addition of each methylene group (see the last column in Table 17).

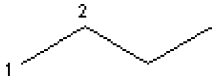
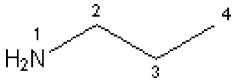
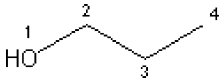
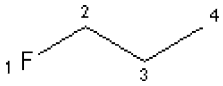
The influence of branching on the local (atom) and total derivative indices in alkanes is illustrated in Table 18 using the isomers of pentane as an example. As can be observed, terminal methyl groups (for instance C_1) show a decrease in their values ($^{E}\Delta$ equal of 17.33, 12.08 and 10.50 for pentane, 2-methyl-butane and *t*-pentane, respectively). Likewise, the total indices steadily decrease with branching: $\|^{E}\bar{X}\|_1$ of 60.67, 54.55, and 54, respectively (see Table 18). This table shows that total and local (atom-based) derivative indices are able to discriminate among the pentane's branching isomers.

On the other hand, the introduction of multiple bonds yields higher $^{E}\Delta$ values in this order: $^{E}\Delta(Csp) \square$ $^{E}\Delta(Csp^2) \square$ $^{E}\Delta(Csp^3)$, demonstrating the dependence of the $^{E}\Delta$ values on the nature and topology of the atoms involved(see Table 19). Furthermore, atoms belonging to terminal multiple bonds possess lower $^{E}\Delta$ values relative to the inner ones. This behavior also mirrors the inductive effect and the reduction in topological freedom or an increase in steric hindrance, that is, this local parameter represents the accessibility from outside of atoms in subgraphs of order 2 in the molecule.

Thus, the GDIs are interpreted as a component of the "molecular accessibility" coming from contributions of subgraphs of length 2 in the molecule. On the other hand, the global invariants are capable of discriminating between saturated and unsaturated (double and triple bonds) isomers (see Table 19). For instance, the molecules of 1-butene (40.50) and 1-butyne (53.22) are adequately discriminated from their isomers 2-butene (44.67) and 3-butyne (56.33), respectively.

Finally, the value of these indices for Butane, for Propylamine, Propan-1-ol and for 1-Fluoro-propane increase in this same order, in correspondence to the electronegativity value of the following atoms: C, N, O, F (see Table 20). Ad-

Table 20. The influence of heteroatoms on the GDI values.

Atom (*i*)	$^{E}\Delta$	$^{E}\Delta_1$	$^{E}\Delta_2$	$^{E}\Delta_3$	$\|\|^{E}\overline{X}\|\|_1$	$\|\|^{E}\overline{X}\|\|_2$	$\|\|^{E}\overline{\xi}\|\|$
							
C_1	11.17	1.00	2.00	1.00	–	–	–
C_2	6.17	3.00	2.00	1.00	–	–	–
Total	–	–	–	–	**34.67**	**18.04**	**8.30**
							
N	12.17	1.22	2.45	1.64	–	–	–
C_2	6.52	3.22	2.22	1.30	–	–	–
C_3	6.69	3.00	2.22	1.30	–	–	–
C_4	11.29	1.00	2.00	1.03	–	–	–
Total	–	–	–	–	**36.67**	**19.05**	**8.80**
							
O	13.21	1.44	2.88	2.23	–	–	–
C_2	6.84	3.44	2.44	1.57	–	–	–
C_3	7.17	3.00	2.44	1.57	–	–	–
C_4	11.53	1.00	2.00	1.09	–	–	–
Total	–	–	–	–	**38.75**	**20.14**	**9.30**
							
F	14.81	1.76	3.52	3.09	–	–	–
C_2	7.30	3.76	2.76	1.94	–	–	–
C_3	7.87	3.00	2.76	1.94	–	–	–
C_4	11.97	1.00	2.00	1.20	–	–	–
Total	–	–	–	–	**41.96**	**21.86**	**10.05**

ditionally, the introduction of a heteroatom into an alkane molecule produces an effect on the $^{E}\Delta$ of the adjacent carbon-atom (C_2) proportional to Pauling's electronegativity value for the heteroatom. For example, $^{E}\Delta$ of the carbon-atom adjacent to the nitrogen, oxygen and fluorine atom are 6.52, 6.84, and 7.30, respectively (see Table 20).

These simple examples demonstrate that variation of the GDIs values with alkyl-chain lengthening, branching, heteroatoms-content, and multiple bonds is consistent with logical trends due to structural changes in molecules.

Similarly, we may interpret the effect of "higher" order total and local derivative indices, starting from contributions of "subgraphs" of different orders 3, 4, 5, etc. In any case, whether a complete series of indices is considered, or a specific characterization of the chemical structure, the generalization of the descriptors to "superior analogs" is necessary for the evaluation of situations where only one descriptor is unable to produce good structural characterization.[12]

5.2.2 Preliminary Trends in Electronic and Steric Influence

A good test of the validity of the information encoded by new indices is to evaluate the correlation between GDI and nuclear magnetic resonance (RMN) chemical shifts. This experimental property reflects the environment of an atom in a molecule due to electronic, topologic and steric influences. In fact, Kier and Hall previously used this approach in order to discover and/or indicate that E-state values encode both attributes using E-state of oxygen atoms in a series of ethers and carbonyl chemicals.[8b] Here, will use this data in order to validate the codification of these structural attributes by the GDIs. These indices were compared with the ^{17}O chemical shifts as shown in Tables 21 and 22 (also see Equation 21 and Equation 22, respectively). The best correlations are depicted below

$$^{17}\text{O RMN (carbonyls)} = 588.88\ (\pm 3.17) - 4.12\ (\pm 0.41)\ ^{E}\Delta_3(C_1) \quad (21)$$

$$R^2 = 0.94 \quad s = 5.0218 \quad Q^2 = 0.94 \quad s_{CV} = 2.34 \quad F = 102.55$$

$$^{17}\text{O RMN (ethers)} = -247.01\ (\pm 18.39) + 41.87\ (\pm 2.86)\ ^{V}\Delta_1(O) \quad (22)$$

$$R^2 = 0.96 \quad s = 8.01 \quad Q^2 = 0.94 \quad s_{CV} = 4.49 \quad F = 214.60$$

The correlation between the atom-based GDI and chemical shifts are rather close. Therefore, the new MDs encode

Table 21. Graph derivative index for C_α-atom of carbonyls and ^{17}O RMN chemical shifts.

No	Compound	$^{E}\Delta_3(C_1)$ [a]	^{17}O RMN [b]	Cal. (Equation 21)
1	CH_3CHO	0.0	592.0	588.9
2	C_2H_5CHO	2.0	579.5	580.5
3	*i*–C_3H_7CHO	3.8	574.5	573.3
4	$(CH_3)_2CO$	5.6	569.0	565.8
5	$CH_3COC_2H_5$	6.5	557.0	561.9
6	CH_3CO–*i*–C_3H_7	9.1	557.0	551.6
7	$(C_2H_5)_2CO$	7.9	547.0	556.5
8	C_2H_5CO–*i*–C_3H_7	10.8	543.5	544.4
9	(*i*–C_3H_7)CO	13.9	535.0	531.4

[a] Graph derivative index for Cα-atom of carbonyls using 3-order path-type subgraph. [b] Measures ^{17}O RMN chemical shifts.[8b]

Table 22. Graph derivative index for O-atom of ethers and ^{17}O RMN chemical shifts.

No	Compound	$^{V}\Delta_1(O)$ [a]	^{17}O RMN [b]	Calc. (Equation 22)
1	Dimetil éter	4.84	−52.2	−53.12
2	Etil metal-	5.35	−22.5	−22.64
3	Isopropil metal-	5.86	−2	−1.56
4	t-Butil metil	6.37	8.5	9.36
5	Dietil	5.86	6.5	7.72
6	Isopropil etil	6.37	28	28.75
7	t-Butil etil	6.88	40.5	39.50
8	Diisopropil	6.88	52.5	50.84
9	t-Butil isopropyl-	7.40	62.5	62.59
10	Di-t-Butil-	7.91	76	76.37

[a] Graph derivative index for O-atom of ethers using 1-order path-type subgraph, [b] measures ^{17}O RMN chemical shifts.[8b]

relevant structural information for this property, mainly reflecting the electronic, topologic and steric environment of an atom in a molecule.

6 Conclusions

The approach described in this report appears to be a prominent method to find quantitative models for the description of physical, thermodynamic, or biological properties. The novel MDs proposed here have shown to have some interesting features, such as: 1) their functional definitions are based on novel algorithms and mathematical formulae. These novel atom-based MDs are based on the graph derivative for vertex pairs similar to the one defined in discrete mathematics. The atom- and group-level approach as well as atom-type formalism will permit to expedite the investigation of molecular mechanisms and rational design of molecules at the local level. 2) These local indices together with global ones are now added as a new set of MDs to the significant arsenal of whole-molecule indices. Moreover, we also define strategies that generalize the definition of global or local invariants from atomic contributions (LOVIs). In respect to this, metric (norms), means and statistical *invariants* are introduced. These invariants are applied to a vector whose components express the atomic indices. 3) This approach stems from a new matrix representation of a G derived from the generalization of an incidence matrix whose row entries correspond to connected sub-graphs of a given G. 4) These MDs can be easily and quickly calculated. That is, the calculation is simple and straightforward, requiring only 2D information. The novel indices are implemented in DIVATI, a new module of *TOMOCOMD-CARDD* program to facilitate their computation. 6) The novel indices show good predictive power in the modeling of physicochemical properties. Furthermore, it was clearly demonstrated that this set of descriptors produced similar to better models than the other 2/3D TIs and geometric indices previously used by different researchers. 7) In addition, principal component analysis indicates that the information carried by the GDIs is markedly different from that codified in various 0–3D MDs presently in QSPR/QSAR and drug design practice. The variation of the GDIs values with alkyl-chain lengthening, branching, heteroatoms-content, and multiple bonds agrees with usual organic intuition. The relation of atom-based derivative indices with ^{17}O NMR of a series of ethers and carbonyl chemicals reflects that the new MDs encode electronic, topological and steric information.

7 Future Perspective

Despite these positive features of atom-based derivative indices, additional studies have to be performed to further investigate their meaning and behavior with respect to the

structural features of the molecules. The applications of the present method to QSPR/QSAR and drug-design studies as well as in similarity/diversity analysis of several classes of organic compounds are now in progress and will be subject of future publications.

In forthcoming articles, we will define derivative indices using the relations frequency hyper-matrix (derivative for *n*-tuples relations). We will also introduce new *events* derived from other graph-theoretic and geometric concepts that permit us to define other relations frequency matrices. In addition, we intend to apply all the invariants that have been extensively used in definition of indices reported in the literature up to date to the frequency matrix and all the matrices derived thereof. Other extensions of original concepts will be aimed at the definition of indices based on mixed and higher order-derivatives on a G.

Supplementary Data Available

The molecular descriptor values for the four data sets (octane isomers, alcohols, phenetylamines and polychlorobiphenyls) in excel file (SI1), the molecular descriptor values for external validation for two data sets (phenetylamines and polychlorobiphenyls)in excel file (SI2), and factor loadings from PCA (excel file, SI3) are available free of charge via the Internet.

Acknowledgements

Y. Marrero-Ponce thanks the program 'International Professor' for a fellowship to work at *Cartagena University* in 2013–2014. Finally, but not least, the authors want to express their acknowledgements to Prof. *Jorge Galvez* (VU) and Prof. *Ramón García-Domenech* (VU) for their help and useful comments about these new MDs.

References

[1] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Germany, Weinheim, **2000**.
[2] A. R. Katritzky, E. V. Gordeeva, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835 – 857.
[3] a) M. T. Randić, N. Trinajstić, *J. Mol. Struct. (THEOCHEM)* **1993**, *300*, 551 – 572; b) J. Devillers, A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Scientific, Amsterdam, **1999**, 21 – 57.
[4] E. Estrada, *Chem. Phys. Lett.* **2001**, *336*, 248 – 252.
[5] M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609 – 6615.
[6] V. A. Gorbátov, *Fundamentos de la Matematica Discreta*, Mir, Moscú, URSS, **1988**.
[7] R. Todeschini, V. Consonni, *MATCH Commun. Math. Comput. Chem.* **2010**, *64*, 359 – 372.
[8] a) L. B. Kier, L. H. Hall, *Pharm. Res.* **1990**, *7*, 801 – 807; b) L. B. Kier, L. H. Hall, *Molecular Structure Description. The Electrotopological State*, Academic Press, San Diego, **1999**; c) A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398 – 402; d) Y. Marrero-Ponce, *Bioorg. Med. Chem.* **2004**, *12*, 6351 – 6369; e) Y. Marrero-Ponce, F. Torres; Alvarado, Y. R. Rotondo, *J. Comput.-Aided Mol. Des.* **2006**, *20*, 685 – 701.
[9] R. Daudel, C. Moser, *Quantum Chemistry: Methods and Applications*, Wiley, New York, USA, **1984**.
[10] M. M. Deza, E. Deza, *Encyclopedia of Distances*, Springer, Heidelberg, **2009**.
[11] Y. Marrero-Ponce, V. T. Romero, *TOMOCOMD-CARDD (TOpological MOlecular COMputational Design) Software*, Version 1.0; an academic version can be obtained upon request to Y. Marrero-Ponce: ymarrero77@yahoo.es version 1.0. *Central University of Las Villas: Santa Clara, Villa Clara,* **2002**.
[12] a) M. Randic, *J. Math. Chem.* **1996**, *19*, 375 – 392; b) M. Randić, *J. Math. Chem.* **1991**, *7*, 155 – 168.
[13] a) P. K. Agarwal, *Proteins* **2004**, *56*, 449 – 463; b) V. Consonni, R. Todeschini, M. Pavan, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682 – 692.
[14] a) E. Estrada, E. Molina, *J. Mol. Graphics Model.* **2001**, *20*, 54 – 64; b) L. H. Hall, B. Mohney, L. B. Kier, *Quant. Structure-Activity Relat.* **1991**, *10*, 43 – 51.
[15] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693 – 705.
[16] M. T. Randić, N. Trinajstić, *J. Mol. Struct. (THEOCHEM)* **1993**, *284*, 209 – 221.
[17] a) E. Estrada, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1042 – 1048; b) E. Estrada, L. Rodriguez, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1037 – 1041; c) M. Randić, *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45 – 59; d) M. Randić, *Croat. Chim. Acta.* **1993**, *66*, 289 – 312; e) M. Randić, X. Guo, T. Oxley, H. Krishnapriyan, L. Naylor, *J. Chem. Inf. Comput .Sci.* **1994**, *34*, 361 – 367; f) M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 535 – 540; g) M. V. Diudea, O. M. Minailiuc, G. Katona, *Rev. Roum. Chim.* **1997**, *42*, 239 – 249.
[18] D. E. Needham, I.-C. Wei, P. G. Seybold, *J. Am. Chem. Soc.* **1988**, *110*, 4186 – 4194.
[19] O. Martínez-Santiago, Y. Martinez, S. Barigye, Y. Marrero-Ponce, *DIVATI (Discrete deriVAtive Type Indices)*, Unit of Computer-Aided Molecular "*Bio-Silico*" Discovery and Bioinformatic Research (CAMD-BIR Unit), Santa Clara, Villa Clara, Cuba, **2011**.
[20] R. Todeschini, P. Gramatica, *Perspect. Drug Dis. Des.* **1998**, *9–11*, 355 – 380.
[21] R. Todeschini, V. Consoni, A. Mauri, M. Pavan, Talete, Milano, Italy **2005**.
[22] a) E. Bayram, P. Santago, R. Harris, Y. D. Xiao, A. J. Clauset, J. D. Schmitt, *J. Comput.-Aided Mol. Des.* **2004**, *18*, 483 – 493; b) D. E. Goldberg, *Genetic Algorithms*, Addison-Wesley, Reading, MA **1989**.
[23] a) A. Basilevsky, *Statistical Factor Analysis and Related Methods*, Wiley, New York, **1994**; b) E. F. Ildiko, J. H. Friedman, *Technometrics,* **1993**, *35*, 109 – 135.
[24] *STATISTICA (data analysis software system)*, Statsoft, Tulsa, **2008**.

Received: December 9, 2013
Accepted: January 29, 2014
Published online: May 12, 2014

