J Comput Aided Mol Des
DOI 10.1007/s10822-012-9591-9

Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application

Yovani Marrero-Ponce · Oscar Martínez Santiago · Yoan Martínez López · Stephen J. Barigye · Francisco Torrens

Received: 9 May 2012 / Accepted: 19 July 2012

Abstract In this report, we present a new mathematical approach for describing chemical structures of organic molecules at atomic-molecular level, proposing for the *first time* the use of the concept of *the derivative* (∂) *of a molecular graph* (MG) *with respect to a given event* (*E*), to obtain a new family of molecular descriptors (MDs). With this purpose, a new matrix representation of the MG, which generalizes graph's theory's traditional incidence matrix, is introduced. This matrix, denominated the generalized incidence matrix, ***Q***, arises from the Boolean representation of *molecular sub-graphs* that participate in the formation of the graph molecular skeleton MG and could be complete (representing all possible connected sub-graphs) or constitute sub-graphs of determined orders or types as well as a combination of these. The *Q* matrix is a non-quadratic and unsymmetrical in nature, its columns (*n*) and rows (*m*) are *conditions* (letters) and *collection of conditions* (words) with which the event occurs. This non-quadratic and unsymmetrical matrix is transformed, by algebraic manipulation, to a quadratic and symmetric matrix known as *relations frequency matrix*, ***F***, which characterizes *the participation intensity of the conditions (letters) in the events (words)*. With ***F***, we calculate the derivative over a pair of atomic nuclei. The local index for the atomic nuclei *i*, Δ_i, can therefore be obtained as a linear combination of all the pair derivatives of the atomic nuclei *i* with all the rest of the *j*'*s* atomic nuclei. Here, we also define new strategies that generalize the present form of obtaining global or local (group or atom-type) invariants from atomic contributions (local vertex invariants, LOVIs). In respect to this, metric (norms), means and statistical *invariants* are introduced. These invariants are applied to a vector whose components are the values Δ_i for the atomic nuclei of the molecule or its fragments. Moreover, with the purpose of differentiating among different atoms, an atomic weighting scheme (atom-type labels) is used in the formation of the matrix ***Q*** or in ***LOVIs*** state. The obtained indices were utilized to describe the partition coefficient (Log P) and the reactivity index (Log K) of the 34 derivatives of 2-furyl-ethylenes. In all the cases, our MDs showed better statistical results than those previously obtained using some of the most used families of MDs in chemometric practice. Therefore, it has been demonstrated to that the proposed MDs are useful in molecular design and permit obtaining easier and robust mathematical models than the majority of those reported in the literature. All this range of mentioned possibilities open "the doors" to the creation of a new family of MDs, using the graph derivative, and avail a new

Y. Marrero-Ponce (✉) · O. M. Santiago · Y. M. López · S. J. Barigye
Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, 54830 Santa Clara, Villa Clara, Cuba
e-mail: ymarrero77@yahoo.es; ymponce@gmail.com; yovanimp@uclv.edu.cu
URL: http://www.uv.es/yoma/; http://ymponce.googlepages.com/home

Y. Marrero-Ponce · F. Torrens
Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, 46071 València, Spain

O. M. Santiago
Department of Chemical Science, Faculty of Chemistry-Pharmacy, Central University of Las Villas, 54830 Santa Clara, Villa Clara, Cuba

Y. M. López
Department of Computer Sciences, Faculty of Informatics, Camaguey University, 74650, 70100 Camagüey City, Camagüey, Cuba

Published online: 04 November 2012

tool for QSAR/QSPR and molecular diversity/similarity studies.

Keywords Discrete mathematics · Chemical graph theory · Molecular graph · Sub-graph · Incidence matrix · Frequency matrix · Derivative of molecular graph · Invariant · QSPR study · Physicochemical property · Derivatives of 2-furylethylene

"When you have a deep truth, then the opposite of a deep truth may again be a deep truth."

N. Bohr.

Introduction

A **molecular descriptor** (MD) is "*the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*."[1] One important characteristic of MDs is that, until now, no single MD could be universally used to yield optimal correlations; therefore, more than 1,000 MDs (many of which are implemented in programs DRAGON [2], CODESSA, [3] [4] MOLGEN [5], etc.), are in existence. Among them, the so-called topological indices (TIs) [6] have found major applications in quantitative structure–activity (or structure–property) relationships (QSAR/QSPR) and drug design studies as well as in similarity/dissimilarity studies. The TIs are numbers associated with chemical structures and they are obtained from the molecular graph (MG) of the molecule on the basis of local graph invariants (LOVIs) [7]. Such *invariants* are numbers associated with vertices symbolizing atoms in such a way that they are independent of arbitrary vertex numbering.

A large number of TIs are generally determined through the use of several *invariants* applied to different algebraic representations of MGs. The most used of such representations are the adjacency (**A**) and distance (**D**) matrices (or a combination of these two), which turn out to be very similar, i.e.; the indices defined so far do not depict great diversity due to the reduced number of matrices used in molecular representation and the close relationship that exists among the invariants used to extract structural information [6, 8, 9]. Despite the fact that these two matrices (**A** and **D**) are unrelated, the great variability and the *ad hoc* nature (defined by mathematical constructions) of the structural *invariants* used to define TIs gives the impression of a profound lack of *unity* among all these MDs. However, it is well know that most TIs are related by their analytic definition. Even indices so diverse in their analytic definition such as Schultz molecular topological index (MTI) and Kier-Hall connectivity indices can be expressed in the same form using the procedure vector -matrix–vector transpose [10, 11]. There is, therefore, need to define new matrix representations and invariants in order to develop better and different global and local indices.

Another important aspect in TIs, is that the greater part are global (molecule) indices and they can be calculated for whole molecules [12] . This fact is a crucial problem because many properties or even biological activity depend more on structural features of determined molecular zones than on the molecule as a whole and in this case the whole-molecule indices may tend to *obscure* this fact and prevent clear identification of such atomic features (individual atoms or localized molecular regions). In this sense, this pre-requisite depicts an important weakness of TIs defined so far, given that most of them are global definition and are not used in the definition of fragments or determined zones of molecules [13]. There is, therefore, a need to explore the possibility of obtaining more directly applicable atom-level indices (as well as their uses as LOVIs).

In addition, it is important to highlight that all locally defined indices proposed so far use the procedure that *the sum of parts makes the total* to obtain corresponding global (or local group or atom-type) indices. For example, the electro-topological state indices of a molecule or fragment are calculated as the sum of the electro-topological states of each atom, *i.e.;* as a linear combination of the atomic indices [14].

The main purpose of the present paper is to present new sets of MDs, namely *derivatives of MG* and establish their abilities (both total and local) in the description of the molecular structure by correlating them with physicochemical properties of furylethylenes. We also propose new matrix representations of the MG based on the relations frequency of vertices and from these we derive atom-level derivative indices. These MDs encode topological information given by the MG, weighted by chemical information encoded in selected atom-weightings.

In this report, we also define new strategies that generalize the present form of obtaining global or local (group) *invariants* from atomic contributions (LOVIs). In respect to this, metric (norms), means and statistical *invariants* are introduced. These invariants are applied to a vector constituted by atomic-level indices. Finally, in order to evaluate the performance of the proposed MDs in QSPR studies, we model the partition coefficient (Log P) and the specific rate constant (Log K) of the derivatives of 2-furylethylenes. This experiment permits us to compare our best models directly with those of others (edge- and vertices-based connectivity indices, total and local spectral moments, as well as topological and quantum chemical descriptors) reported in the literature.

The structure of this report will be as follows: A background-review of preliminary concepts about chemical graph-theory and the derivative will be described in the following section ("Background-review of preliminary concepts: chemical graph-theory and discrete derivative"). After, an outline and definition of our procedures will be illustrated in "Theoretical scaffold: theory of new molecular descriptors" section, as well as the generalization of the method for obtaining global and local (group and atom-type) indices from LOVIs. Next, the correlation equations for selected properties and molecules, as well as the statistical considerations on the obtained results will be developed in "QSPR Study. A Comparative Analysis" section. Finally, the conclusions and future outlooks will be presented in the last sections.

Background-review of preliminary concepts: chemical graph-theory and discrete derivative

Graph-theoretical matrix representation, invariant and TIs

A complete review of TIs is almost impossible due to the great quantity of such MDs that are published in the literature. Therefore, in this section we do not intend to perform an exhaustive compilation of graph theory and/or TIs. We only pretend here to introduce some important concepts about this topic that will be useful in explaining our procedure.

A topological representation of a molecule can be carried out through the so-called molecular graphs (MG). The MGs are non-directed chemical graphs which represent, in different conventions, molecules. Usually, in MGs ***vertices*** correspond to atoms and lines are named ***edges*** and represent covalent bonds between atoms.

A graph G, $G = (V,E)$, is defined as an ordered pair consisting of two sets $V = V(\mathrm{G})$ (V is the set of vertices) and $E = E(\mathrm{G})$ (E is the set of edges.), where the elements of the set E define the binary relationship between the elements of the set V. However, in order to give a more realistic representation of the topology of a molecule we need to identify the different atoms in the molecule by labeling them with their chemical symbols. In doing so we represent the MG as a weighted graph of the form $G = (V, E, \mathrm{f}, \mathrm{V})$, where V is a set containing the entire chemical symbols of the elements and f is a subjective mapping of the elements of V onto the set V [10, 15].

The number of vertices in the graph is n and the number of edges m. In a connected graph G every pair of vertices is joined by a path. A multi-graph contains pairs of vertices connected by more than one edge. A multi-edge of multiplicity m is a set of m edges incident with the same pair of distinct vertices. The vertex degree (or valency of atom i), $\delta(v_i)$ is equal to the number of vertices adjacent to vertex v_i. A path is a sequence of vertices $v_{i0}, v_{i1}, v_{i2},\ldots,v_{il}$ of a graph, such that v_{ij-1} and v_{ij} are adjacent. The length of this path is equal to l. The graph distance d_{ij} between a pair of vertices v_i and v_j from a connected graph G is defined as the length (number of edges) of the shortest path connecting the two vertices. A sub-graph of a graph G is a graph $G^* = (V^*, E^*)$, where V^* is a subset of V and E^* is a subset of E. An important classification of sub-graphs was proposed by Kier and Hall for calculating molecular connectivity indices.[16] Their scheme classifies the sub-graphs by order m (number of edges in sub-graph) or type t (path, clusters, path-clusters, and chains, which are designed as p, C, pC, and Ch, respectively, according to their original definitions).

MGs are widely used to represent the chemical structure of covalent compounds in a graphical form. The MG is, however, a non-numerical representation of the chemical structure, and the computation of TIs requires a numerical description of graphs. Graphs can be represented in algebraic form as matrices. The main matrices in graph theory are adjacency, distance and incidence matrices. ***Adjacency matrix A*** is a square and symmetric matrix of order n whose elements a_{ij} are ones or zeros if the corresponding vertices i and j are adjacent or not [1, 6, 17]. ***Distance matrix D*** is a square and symmetric matrix of order n whose elements d_{ij} correspond to the topological distances between atoms i and j. ***Incidence matrix C*** expresses the linkage between the vertices and the edges or bonds in the molecule, in which c_{ij} are equal to 1 or 0. If the jth edge and the ith vertex are adjacent, $c_{ij} = 1$, otherwise, $c_j = 0$. For the molecules with n atoms and m bonds, $\boldsymbol{C}$ is an $n \times m$ matrix [1, 6, 17].

As we see in the previous definitions, TIs are MDs derived from graph-theoretical *invariants* [1, 6, 13, 17]. For instance, adjacency or distance matrices are not graph invariants as they depend on the numbering of graph vertices. However, a simple *invariant* can be the adjacency matrix order, that is, the number of vertices in the graph, which does not depend on the graph elements numbering. TIs have been classified according to their nature in first, second and third generation [10, 18].

At present, the development of novel TIs continues. However, the trend in this field is to modify the well-known TIs other than to define new ones. For instance, one of these approaches has been named the "variable molecular descriptors" and mainly developed by M. Randic. In particular, a generalization of "variable molecular descriptors" approach which permits the calculation of several known TIs by using the same graph invariant have also been recently introduced by Estrada [11, 19, 20]. According to this generalized algorithm, Wiener index, Zagreb group

indices, Balaban J index, Randic c index and some others are particular cases of an infinite set of TIs. Another topic in development is the modification of TIs to permit the codification of molecular chirality, therefore, "chirality TIs" [21–23]. The main purpose of developing these MDs is to be able to account for chiral molecules, which are very well known to have an important role in medicinal chemistry. Thus, the current aim is to unify and understand the well-known TIs obtained by traditional procedures. There is, therefore, a need to define new matrix representation and invariants as well as to explore the possibility of obtaining more directly applicable atom-level indices (as well as their uses as LOVIs) in order to develop better and different global and local indices that open new possibilities in the development of TIs to be applied in the drug discovery process.

Derivatives in discrete mathematics (graph theory)

As it is well-known a mathematical analysis of the derivative concept characterizes the degree of variation in a function on carrying out a small variation in its argument, based on the limit and continuity theorems. In discrete mathematics, however, the limit as a 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's first define certain important concepts.

In order to start, we define an *event* (E), which is true when certain conditions of the examined process are fulfilled [24]. Every E determines a bi-dimensional binary matrix $Q = [q_{ij}]_{m\times n}$, 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: [24]

- 1, *if the jth condition is included in the ith collection of conditions, in which the event is true.*
- 0, otherwise.

In other words, every E *event* determines a model (ψ) for the *incidence matrix* $\boldsymbol{Q}$; 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* $\mathbf{F} = [f_{ij}]_{n\times n}$ that characterizes the model ψ, with the incidence matrix $\boldsymbol{Q}(\psi) = [q_{ij}]_{m\times n}$ [24].

We denominate *relations frequency matrix* $\mathbf{F} = [f_{ij}]_{n\times n}$, 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 *individual frequency* of letter i and f_{ij} the *reciprocal frequency* of the letters i and j.

From the definition of the $\mathbf{F}$, 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: $\mathbf{F} = \boldsymbol{Q}^{\mathrm{T}} \times \boldsymbol{Q}$, $\boldsymbol{Q}^{\mathrm{T}}$ being the transpose matrix of the incidence matrix [$\boldsymbol{Q}$ (ψ)] for the model ψ [24].

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 E.

Let us call 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 (E) a non-oriented weighted graph <V,(U, P) >, whose label coincides with that of a model determined by this event and a pair of edges (e_i, e_j) is weighted by the ratio of its incompatible participation frequency $(f_i - f_{ij}) + (f_j - f_{ij})$ to the participation frequency f_{ij} compatible to the event E:

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

with the particularity that,

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

Let us therefore illustrate the derivative concept of a graph with an example. Given graph G in Fig. 1A, we would like to determine the participation frequency of the different edges in the formation of the graph skeletons. The graph G has 8 skeletons [sub-graphs of order 3, without differentiating the type (Fig. 1B)]. 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 to the pair of previously indicated numbers, we determine numbers that characterize the non uniform participation grade of pairs of graph edges (graph derivative for a pair of elements), in the formation of the graph skeletons, for 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 the pairs of graph

edges. The incidence and frequency matrices, for this model, are:

$$Q = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

$$F = Q^T \times Q = \begin{Vmatrix} 5 & 2 & 2 & 3 & 3 \\ 2 & 5 & 2 & 3 & 3 \\ 2 & 2 & 4 & 2 & 2 \\ 3 & 3 & 2 & 5 & 2 \\ 3 & 3 & 2 & 2 & 5 \end{Vmatrix}$$

The elements for the matrix (F) determine the $\partial G/\partial S$, which is a weighted graph, with labels [7] and 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:

$$\frac{\partial G}{\partial S}(a,b) = 3, \frac{\partial G}{\partial S}(a,c) = 2,5 \ldots\ldots\ldots\ldots\ldots\ldots \frac{\partial G}{\partial S}(d,e) = 3$$

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

As can be observed, to determine the graph's derivative, according to the event (E), it is necessary to [24]:

1. Construct a model determined by an event, E, previously chosen, 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. However, in this report, we only apply the derivative concept for duplexes in the generation of novel MDs.

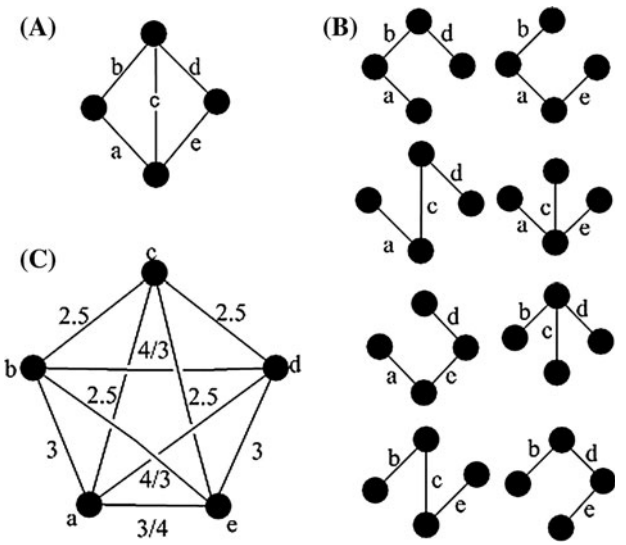


Fig. 1 (**A**) Molecular graph, (**B**) sub-graphs (*words*) according to event E (connected subgraphs of order 3 based on edge (*letters*) relations), (**C**) derivative graph

Theoretical scaffold: theory of new molecular descriptors

New atom-relations: extended incidence matrix

Let's take the molecule of methylbutane as a simple example (see Fig. 2), where the numbers correspond to the labels that are assigned to the carbon atoms (vertices) in the molecular structure and graph.

This graph is in correspondence with the previous chemical structure. In the same, the carbon atoms labeled C_1, C_2, C_3, C_4 *and* C_5 are represented as the MG vertices and *a, b, c* and *d* constitute the edges that represent the established chemical bonds among the mentioned atoms.

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 previously introduced MG, the following sub-structures are obtained, organized according to their orders:

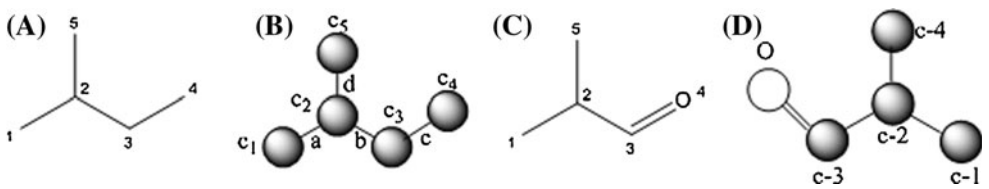


Fig. 2 The chemical structure and molecular graph of [the *numbers* correspond to the labels that are assigned to the atoms (*vertices*) in the molecular structure]: (**A**) 2-methylbutane (H-depleted structure), (**B**) molecular graph of 2-methylbutane, (**C**) 2-methylpropanal (H-depleted structure) and (**D**) molecular graph of 2-methylpropanal

Order 0:	C_1, C_2, C_3, C_4, C_5
Order 1 (4 paths):	C_1–C_2, C_2–C_3, C_3–C_4, C_2–C_5
Order 2 (4 paths):	C_1–C_2–C_3, C_1–C_2–C_5, C_2–C_3–C_4, C_2–C_3–C_5
Order 3 (2 paths and 1 cluster):	C_1–C_2–C_3–C_4, C_2–C_3–C_4–C_5, C_1–C_2–C_3–C_5
Order 4 (1 path-cluster):	C_1–C_2–C_3–C_4–C_5

These sub-graphs are represented in an *incidence matrix* (***Q***), from which we obtain the corresponding *relation 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 7 times in the formation of the sub-graphs (see Q and F matrix for isopentane).

The preselected event determines the corresponding incidence and frequency matrices, which are shown below:

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad F = \begin{Vmatrix} 7 & 6 & 4 & 2 & 3 \\ 6 & 12 & 8 & 4 & 6 \\ 4 & 8 & 10 & 5 & 4 \\ 2 & 4 & 5 & 6 & 2 \\ 3 & 6 & 4 & 2 & 7 \end{Vmatrix}$$

The 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 (Kier and Hall nomenclature) as well as a combination of these. A particular case where only sub-graphs of *order 1* (pairs of vertices or edges in G) are considered, the formed matrix Q coincides with the common incidence matrix used in graph theory. The incidence matrix, in view of the fact that it is non-quadratic and unsymmetric, has had few applications in the definition of MDs, since the present invariants are not designed for matrices of this nature.

Derivative of molecular graph. Local and total definition

In this section, we will define novel indices which apply the Eq. 1 for each pair of vertices in the MG. Let us therefore continue with the example of the methylbutane molecule for which we have already obtained its corresponding frequency matrix according to the event proposed in the present report.

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

$$\frac{\partial G}{\partial S}(C_1, C_2) = \frac{7 - 2(6) + 12}{6} = \frac{7}{6}$$
$$\frac{\partial G}{\partial S}(C_1, C_3) = \frac{7 - 2(4) + 10}{4} = 2.25$$

In the same way, values of pairs of elements of the graph can be determined successively, as shown below:

$$\frac{\partial G}{\partial S}(C_1, C_4) = 4.5 \quad \frac{\partial G}{\partial S}(C_3, C_4) = 1.2$$
$$\frac{\partial G}{\partial S}(C_1, C_5) = \frac{8}{3} \quad \frac{\partial G}{\partial S}(C_3, C_5) = 2.25$$
$$\frac{\partial G}{\partial S}(C_2, C_3) = 0.75 \quad \frac{\partial G}{\partial S}(C_4, C_5) = 4.5$$
$$\frac{\partial G}{\partial S}(C_2, C_4) = 2.5 \quad \frac{\partial G}{\partial S}(C_2, C_5) = \frac{7}{6}$$

All these pair derivatives will be organized in matrix form (£ matrix), whose entries *ij* are the derivative 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*. Let us name the derivative (differential) of atom *i* as Δ_i for each one of the elements of the graph (that is for each atomic nucleus), as the summation over all the derivative values that include the element *i* (linear combination):

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

where, *n* is the number of atoms in the molecule, and $\frac{\partial G}{\partial S}(i,j)$ is the derivative for vertices *i* and *j*. The defining Eq. (2) for Δ_i, may also be written in the matrix form, where [***I***] is a column unitary vector (an $n \times 1$ matrix) and [£] is the derivative matrix (entries *ij* are the derivatives for *i* and *j* vertices). We obtain the atomic derivative values (LOVI) for each element, which would be: $\Delta_1 = 10.58$, $\Delta_2 = 5.58$, $\Delta_3 = 6.45$, $\Delta_4 = 12.7$ and $\Delta_5 = 10.58$.

If we thoroughly observe the values for each Δ_i, it can be noted that each value for the first four atoms (from 1 to 4) are different, while the first and the fifth 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 5 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 2 that suffers the greatest steric hindrance. This also coincides with the nature of the concept of the derivative since the atom that most suffers hindrance is the one that most contributes to the formation of the molecule.

Derivative for heteroatomic molecules. Codification of heteroatoms and insaturations

We propose an approach in this report that permits us to adequately characterize molecules with heteroatoms and unsaturated bonds. Let's take as an example an isomer of isopentane, 2-methylpropanal molecule (see Fig. 2). According to the procedure previously explained, we can assert that the Q and F matrices for the MG represented in Fig. 2 are identical to that of isopentane.

It can be easily perceived by simple inspection that the molecular structure of this new molecule contains a heteroatom and a double bond. Let's create a vector of weights V_p, in which weight (ϑ_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 $\vartheta_i = {}^{P}/_{\delta}$ (for this event based in atoms), where P represents a characteristic property of each atom (for example: atomic mass, electronegativity, etc.) and δ is the vertex degree.

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

$$p(o) = \frac{3.5}{2} = 1.75 \quad p(c3) = \frac{2.5}{3} = 0.833$$
$$p(c1) = \frac{2.5}{1} = 2.5 \quad p(c4) = \frac{2.5}{1} = 2.5$$
$$p(c2) = \frac{2.5}{3} = 0.833$$

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

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

Multiplying the incidence matrix with the weighted matrix, we obtain the *weighted incidence matrix* $Q_P = [\mu_{ij}]_{m\times n}$, which is similar to Q in its form only that this new matrix captures specific information of each of the atoms in the molecule on top of the connectivity with others in the mentioned molecule, from which it follows that:

- $\mu_{ij} = p_i$, *if the jth condition is included in the ith collection of conditions, in which the event is true.*
- $\mu_{ij} = 0$, otherwise.

Let us find the weighted incidence matrix Q_P in our case,

$$Q_P = \begin{vmatrix} 2.5 & 0 & 0 & 0 & 0 \\ 0 & 0.833 & 0 & 0 & 0 \\ 0 & 0 & 0.833 & 0 & 0 \\ 0 & 0 & 0 & 1.75 & 0 \\ 0 & 0 & 0 & 0 & 2.5 \\ 2.5 & 0.833 & 0 & 0 & 0 \\ 0 & 0.833 & 0.833 & 0 & 0 \\ 0 & 0.833 & 0 & 0 & 2.5 \\ 0 & 0 & 0.833 & 1.75 & 0 \\ 2.5 & 0.833 & 0.833 & 0 & 0 \\ 2.5 & 0.833 & 0 & 0 & 2.5 \\ 0 & 0.833 & 0.833 & 0 & 2.5 \\ 0 & 0.833 & 0.833 & 1.75 & 0 \\ 2.5 & 0.833 & 0.833 & 1.75 & 0 \\ 2.5 & 0.833 & 0.833 & 0 & 2.5 \\ 0 & 0.833 & 0.833 & 1.75 & 0 \\ 2.5 & 0.833 & 0.833 & 1.75 & 2.5 \end{vmatrix}$$

We can now proceed with the method previously proposed for determining the derivative values over the pairs of graph elements. That is, we obtain the matrix Q_P and its transpose Q_P^T, followed by the corresponding multiplication operation as seen in the previous example ($Q_P^T \times Q_P = F_P$). The obtained *weighted frequency matrix*, F_P, captures information about the number of times that each element participates in the formation of the molecular graph (according to the predetermined event), on top of its participation characteristic, that we can interpret as its identity or relative capacity (in 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:

$$\frac{\partial G}{\partial S}(C_1, C_2) = 2.17 \quad \frac{\partial G}{\partial S}(C_2, O) = 2.57$$
$$\frac{\partial G}{\partial S}(C_1, C_3) = 4.08 \quad \frac{\partial G}{\partial S}(C_2, C_4) = 2.17$$
$$\frac{\partial G}{\partial S}(C_1, O) = 5.12 \quad \frac{\partial G}{\partial S}(C_3, O) = 1.46$$
$$\frac{\partial G}{\partial S}(C_1, C_4) = 2.67 \quad \frac{\partial G}{\partial S}(C_3, C_4) = 4.08$$
$$\frac{\partial G}{\partial S}(C_2, C_3) = 0.75 \quad \frac{\partial G}{\partial S}(O, C_4) = 5.12$$

With these calculated values we can also obtain the derivatives of each atom in the molecule: ${}^{P}\Delta_{c1} = 14.07$,

${}^{P}\Delta_{c2} = 7.63$, ${}^{P}\Delta_{c3} = 10.36$, ${}^{P}\Delta_{O} = 14.30$ and ${}^{P}\Delta_{c4} = 14.07$. These weighted Δ_i, ${}^{P}\Delta_i$, may also be written in the matrix form by using defining Eq. (2), where $[\boldsymbol{I}]$ 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 weighted frequency matrix, F_P, employed Eq. 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, V_p, is multiply by the *unweighted derivative matrix*, £. Therefore, $\left[{}_{p}\Delta_i\right] = [£] \times \left[V_p\right]$. This formula (${}_{p}\Delta_i$) is similar to the one defined in Eq. (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 MG. For the example introduced in this epigraph, $V_p = (2.5,\ 0.833,\ 0.833,\ 1.75,\ 2.5)$. It is important to remark that in this second weighting approach the derivative matrix £ is obtained from an unweighted **F**.

Applying *invariants* to atomic derivative: generalization of 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 MG [1, 7]. In fact, also in quantum chemistry "the summation of the parts makes the total" is applied. For instance, **LCAO** (**L**inear **C**ombination of **A**tomic **O**rbitals) is way of forming molecular orbitals by taking linear combinations of functions associated with the different atoms in the molecule [25]. Therefore, MOs are made up as LCAO of atoms composing the system, *i.e.* they are written in the form, $\phi_i = \sum_{j=1}^{n} c_{ij} Y_j$, 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 a set of 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 1),

1. **Norms (or Metrics)**: Minkowski's norms (N1, N2, N3), Chebyshev's distance (NI, Minkowski's norms for potency equal to infinitum, 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 (Q2), Percentile 75 (Q3), Inter-quartile Range (I50), X max (MX) and X min (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 Randic's equation, M_1 and M_2 Zagreb formulae, 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 published 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-type or group (for example, in ***TOMOCOMD-CARDD*** software [26], the following local indices can be calculated: Proton Acceptors (AH), Proton Donors (DH), Heteroatoms (HT), Halogens (HL) and Carbon (Cb). 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 1 will be applied not only to the original vector of LOVIs (in the three 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, each new LOVI has a mean of 0 and a standard deviation of 1 (this means that standard LOVIs have the same dimensions, i.e., are of comparable magnitude).

Table 1 Invariants

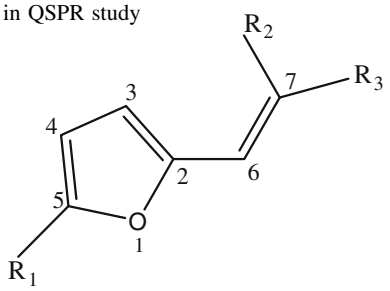
No.	Group	Name	ID	Formula[a]
1	**Norms** (Metrics)	Minkowski's norms ($p = 1$) Manhattan norm	N1	$\|\bar{x}\|_1 = \sum_{i=1}^{n} \|x_i\|$
2		Minkowski's norms ($p = 2$) Euclidean norm	N2	$\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$
3		Minkowski's norms ($p = 3$)	N3	$\|\bar{x}\|_3 = \sqrt[3]{\sum_{i=1}^{n} \|x_i\|^3}$
4		Cheybshev's distance	NI	$\|\bar{x}\|_\infty = \lim_{k\to\infty} \left[\sum_{i=1}^{k} x_i^k\right]^{1/k}$
5		Penrose's size	PN	$d_i = \sqrt{\frac{1}{n^2}\left[\sum_{i=1}(x_i)\right]^2}$
6	**Mean** (first statistical moment)	Geometric mean	G	$\bar{\xi} = \sqrt[n]{\prod_{i=1}^{n} x_i}$
7		Arithmetic mean	M	(potential with $\alpha = 1$)
8		Quadratic mean	P2	(potential with $\alpha = 2$)
9		Potential mean	P3	$c_\alpha = \left(\frac{x_1^\alpha + x_2^\alpha + \ldots + x_n^\alpha}{n}\right)^{\frac{1}{\alpha}}$
10		Harmonic mean	A	(potential with α = -1)
11	**Statistical** (highest statistical moments):	Variance	V	$V = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$
12		Skewness	S	$S = n*M_3/[(n-1)*(n-2)*s^3]$ M_3 is equal to $\sum_{i=1}^{n}(x_i - \bar{x})^3$ s^3 is the standard deviation raised to the third power n is the number of atoms
13		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 is equal to $\sum_{i=1}^{n}(x_i - \bar{x})^j$ n is the number of atoms. s^4 is the standard deviation raised to the fourth power
14		Standard deviation	DE	$\sigma = \sqrt{\frac{\left(\sum x - \bar{x}\right)^2}{n}}$
15		Variation coefficient	CV	$c_v = s/\bar{x}$
16		Range	R	Range $= x_{max} - x_{min}$
17		Percentile 25	Q1	$P25 = \left[\frac{N}{4} + \frac{1}{2}\right]$ N is the number of values
18		Percentile 50	Q2	$P50 = \left[\frac{N}{2} + \frac{1}{2}\right]$ N is the number of values
19		Percentile 75	Q3	$P75 = \left[\frac{3N}{4} + \frac{1}{2}\right]$ N is the number of values
20		Inter-quartile range	I50	I50 = P 75 − P 25
21		X max	MX	X maximum
22		X min	MN	X 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

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

But given that $\bar{y} = (0, 0, 0)$, this formula reduces to $\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n}|x_i|^2}$

Table 2 Chemical structures of compounds used in QSPR study



No.	R_1	R_2	R_3	No.	R_1	R_2	R_3
1	H	NO_2	$COOCH_3$	18	NO_2	H	$CONHCH(CH_3)C_2H_5$
2	CH_3	NO_2	$COOCH_3$	19	NO_2	H	$CONHC(CH_3)_3$
3	Br	NO_2	$COOCH_3$	20	NO_2	H	$CONHCH_2C(CH_3)_3$
4	I	NO_2	$COOCH_3$	21	NO_2	H	$COOCH_3$
5	$COOCH_3$	NO_2	$COOCH_3$	22	NO_2	H	$COOC_2H_5$
6	NO_2	NO_2	$COOCH_3$	23	NO_2	H	$COO(CH_2)_2CH_3$
7	NO_2	$COOC_2H_5$	$COOC_2H_5$	24	NO_2	H	$COOCH(CH_3)_2$
8	NO_2	H	NO_2	25	NO_2	H	$COO(CH_2)_3CH_3$
9	H	H	NO_2	26	NO_2	H	$COOCH_2CH(CH_3)_2$
10	NO_2	H	$CONH_2$	27	NO_2	H	$COOCH(CH_3)C_2H_5$
11	NO_2	H	$CONHCH_3$	28	NO_2	H	$COOC(CH_3)_3$
12	NO_2	H	$CON(CH_3)_2$	29	NO_2	H	$COO(CH_2)_4CH_3$
13	NO_2	H	$CONHC_2H_5$	30	NO_2	H	Br
14	NO_2	H	$CONH(CH_2)_2CH_3$	31	NO_2	H	CN
15	NO_2	H	$CONHCH(CH_3)_2$	32	NO_2	H	OCH_3
16	NO_2	H	$CONH(CH_2)_3CH_3$	33	NO_2	H	H
17	NO_2	H	$CONHCH_2CH(CH_3)_2$	34	NO_2	CN	$COOCH_3$

QSPR study. A comparative analysis

Data set

The decisive criterion of quality for any MD is its ability to describe structure-related properties of molecules. With this objective, we developed the QSPR models to describe *n*-octanol/water partition coefficient (Log P) and rate constant (Log K; for nucleophilic addition of a thiol group to the exo-cyclic double bond) of 34 2-furylethylenes derivatives. These 2-furylethylenes derivatives have different substituents in position 5 of the furan ring as well as in position β of the exo-cyclic double bond (see Table 2).

This chemometric study will permit us to analyze the behavior of each class of MDs defined in this paper and regressions of 2-furylethylenes properties based on the new MDs will be compared to those of well-known indices taken from the literature [27, 28]. The Log P and Log K have an important role in the understanding of the biological behavior of these 2-furylethylene derivatives [29]. Consequently, in order to evaluate the real possibilities of the proposed derivative indices in QSPRs studies, we shall examine these parameters and compare these results with those reported in the literature [27, 28].

This data set was first studied by Estrada and Molina [27] using local (fragment-based) spectral moments of the edge adjacency matrix for Log K description. The model of best fit achieved accounts for more than 96 % of the variance in Log K. Two other models were also obtained by using molecular connectivity indices and total spectral moments of the bond matrix and account for <84 % of the variance in this reactivity index. A model based on quantum chemical descriptors accounts for the same variance as that obtained with bond moments. Later, the same authors using several topographic (3D) molecular connectivity indices based on molecular graphs weighted with quantum chemical parameters are used in modeling Log P of these 2-furylethylene derivatives.[28] These descriptors were compared to 2D connectivity indices (vertex and edge ones) and to quantum chemical descriptors in modeling Log P.

Computational *in house* software

The total and local (in this case, group-type) molecular graph derivative indices used to search for the best

Table 3 Values of the atomic weights used for MDs calculation

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

VdW van der Waals

regression of Log P and Log K of 2-furylethylene derivatives were calculated by an *in-House* software developed in a JAVA script. The novel indices are implemented in **DIVATI** (**DI**screte deri**VA**tive-**T**ype **I**ndices), [30] a new module of ***TOMOCOMD-CARDD*** program [26] to facilitate their automatic computation.

In this case, in order to distinguish every kind of atom in molecule, we used a weight scheme conformed by four atomic properties: 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:

$$ {}^{P/S}_{Qorder}Inv^{E}(localtype) $$

where, P represents ponderations (see Table 3) and S ponderation position, in that *In* means that the atomic weighting is made in Q matrix (^{P}Q, and from this ^{P}£ is obtained) while *Pd* means that the atomic weighting is made on derivative matrix ($[£] \times [V_p]$), and the absence of a label in this position stands for unweighted MDs ($^{P}\Delta_i$, $_{P}\Delta_i$ and Δ_i, respectively). Q_{order} will appear only if a specific order *k* (or any combination of these) is used to compute the total or local MDs. The *Inv* mean the *invariant* used to compute the new MDs from atomic LOVIs (see Table 1). The superscript E stands for standardized MDs. Therefore, global or local indices calculated from a vector of standardized atomic LOVIs ($^{s}\Delta_i$ $^{sp}\Delta_i$ and $^{s}_{p}\Delta_i$ instead of Δ_i $^{P}\Delta_i$ and $_{p}\Delta_i$, respectively). Finally, the parentheses 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 acceptors, DH for proton donors (some plural some singular, unify) and HL for halogens. Finally, it is rather important to note that particularity of sub-graph types was not taken into account and only orders were considered.

Chemometric analysis

The whole set of new MDs were used as independent variables for deriving QSPRs by using multiple linear regression (MLR) technique. The MOBYDIGS (version 1.0—2004) [31] was employed to perform variable selection and QSPR modeling.

This software allows searching for RLM by developing optimal model populations using genetic algorithms (GAs). The GAs [32] 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 use a population of individuals as a model search for the globally optimum solution to a problem. The GAs is optimization procedures searching for 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 [33, 34]. The population size was set at 100 and the reproduction/mutation trade-off (T) at 0.50. 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^2_{Loo} ("leave-one-out" cross-validation) and they were validated using techniques "*bootstrapping*" (Q^2_{boot}) y "*scrambling*" [a (R^2), a (Q^2)]. 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^2_{Loo} and Q^2_{boot} to evaluate the quality of the models. In the recent years, the Loo press statistics (e.g., Q^2) have been used as a means of indicating predictive ability. Many authors consider high Q^2 values (for instance, $Q^2 > 0.5$) 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 in high dimensional data this is just a necessary but insufficient

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. We obtained models of up to 4–7 variables and a constant with the software MobyDigs, taking Log P and Log K as the dependent variables. The best models in each case were selected, taking in consideration the quality of the corresponding statistical parameters.

Modeling physicochemical properties of 2-furylethylenes derivatives and comparison with other 2/3D molecular descriptors

The best linear regression models obtained to describe the Log P of 2-furylethylenes derivatives by using four, five, six and seven parameters are given below, respectively:

$$\begin{aligned}\log P = &-5.05(\pm0.39) + 1.49(\pm0.13)\left[{}^{V/In}_{1}N_3^E\right]\\ &- 0.33(\pm0.04)\left[{}^{V/Pd}_{1}MX^E(DH)\right]\\ &+ 0.19(\pm0.01)\left[{}^{V/In}_{1}PN(DH)\right]\\ &+ 2.39(\pm0.21)\left[{}^{V/In}_{4}PN^E(DH)\right] \qquad (3)\end{aligned}$$

$$N = 34\ R^2 = 97.1\ s = 0.13\ Q^2 = 96.19\ F = 242.48$$

$$\begin{aligned}\log P = &-7.54(\pm0.30) + (\pm0.09)\left[{}^{V/In}_{1}N_3^E\right]\\ &- 1.02(\pm0.08)\left[{}^{A/Pd}_{1}I_{50}^E(DH)\right]\\ &- 0.51(\pm0.03)\left[{}^{V/Pd}_{1}MX^E(DH)\right]\\ &- 1.10(\pm0.15)\left[{}^{E/Pd}_{1}DE^E(Cb)\right]\\ &+ 0.34(\pm0.01)\left[{}^{V/In}_{1}PN(AH)\right] \qquad (4)\end{aligned}$$

$$N = 34\ R^2 = 98.44\ s = 0.097\ Q^2 = 97.81\ F = 352.67$$

$$\begin{aligned}\log P = &-7.26(\pm0.30) + 0.33(\pm0.01)\left[{}^{V/In}_{1}A(AH)\right]\\ &+ 2.48(\pm0.10)\left[{}^{V/In}_{1}N_3^E\right] - 0.94(\pm0.07)\left[{}^{A/Pd}_{1}I_{50}^E\right]\\ &- 0.54(\pm0.03)\left[{}^{V/Pd}_{1}MX^E(Dh)\right]\\ &- 1.06(\pm0.14)\left[{}^{E/Pd}_{1}P_2^E(Cb)\right]\\ &+ 0.18(\pm0.06)\left[{}^{E/Pd}_{3}P_3^E(DH)\right] \qquad (5)\end{aligned}$$

$$N = 34\ R^2 = 98.81\ s = 0.086\ Q^2 = 98.16\ F = 374.08$$

$$\begin{aligned}\log P = &-5.06(\pm0.38) + 1.59(\pm0.11)\left[{}^{V/In}_{1}N_3^E\right]\\ &- 0.47(\pm0.09)\left[{}^{A/Pd}_{1}I_{50}^E(DH)\right]\\ &- 0.35(\pm0.03)\left[{}^{V/Pd}_{1}MX^E(AH)\right]\\ &+ 0.18(\pm0.02)\left[{}^{V/In}_{1}PN(DH)\right]\\ &+ 1.79(\pm0.23)\left[{}^{V/In}_{4}PN^E(DH)\right]\\ &+ 5.6\cdot10^{-4}(\pm1.2\cdot10^{-4})\left[{}^{V/In}_{5}Q_3\right]\\ &+ 0.08(\pm0.02)\left[{}^{A/Pd}_{6}N_1^E(Cb)\right] \qquad (6)\end{aligned}$$

$$N = 34\ R^2 = 99.19\ s = 0.072\ Q^2 = 98.72\ F = 453.84$$

where, N is the number of compounds, R^2 is the determination coefficient, s is the standard deviation of the regression, Q^2 is the square regression coefficient obtained from the Loo cross-validation procedure, and F is the Fisher ratio.

The values of experimental and calculated values of the Log P for the data set (four models) are given in Table 4, and the linear relationships between them (Eq. 6) are illustrated in Fig. 3.

The statistical information for the best regressions with 4, 5, 6 and 7 parameters (predictor variables) are depicted in Table 5. These models were fitted using results from calculations of all possible sets of the new MDs employing the entire weighting scheme).

Models for the reactivity index (Log K) for 2-furylethylenes derivatives were also generated. The best models of 4–7 variables, on the basis of the quality of the statistical parameters, were retained. Below, we report the best models obtained for Log K for 4–7 variables, respectively, using the novel derivative-based indices, as well as the corresponding statistical parameters:

$$\begin{aligned}\log K = &\ 6.98(\pm0.23) - 0.013(\pm0.003)\left[{}^{A/Pd}_{1}I_{50}(HT)\right]\\ &- 4.91(\pm0.22)\left[{}^{V/In}_{3}A^E(AH)\right]\\ &- 1.85(\pm0.10)\left[{}^{P/In}_{3}I_{50}^E(AH)\right]\\ &+ 0.006(\pm0.0005)\left[{}^{A/Pd}_{3}I_{50}(AH)\right] \qquad (7)\end{aligned}$$

$$N = 34 \quad R^2 = 97.92 \quad s = 0.22 \quad Q^2 = 97.29$$
$$F = 341.08$$

Table 4 Experimental and predicted (by using 4–7 variables) values of log P of 34 2-furylethylenes

No	Obsd.[a]	Eq. 3[b]	Eq. 4[c]	Eq. 5[d]	Eq. 6[e]	Res[f]	Res[g]	Res[h]	Res[i]
1	1.88	2.19	1.95	1.91	1.89	−0.31	−0.07	−0.03	−0.01
2	2.44	2.6	2.38	2.38	2.44	−0.16	0.06	0.06	0
3	2.74	2.79	2.85	2.8	2.89	−0.05	−0.11	−0.06	−0.15
4	3	2.78	3.06	3	2.93	0.22	−0.06	0	0.07
5	1.87	1.76	1.74	1.81	1.88	0.11	0.13	0.06	−0.01
6	1.6	1.59	1.6	1.59	1.68	0.01	0	0.01	−0.08
7	2.5	2.35	2.52	2.49	2.43	0.15	−0.02	0.01	0.07
8	1.3	1.03	1.11	1.11	1.24	0.27	0.19	0.19	0.06
9	1.58	1.6	1.58	1.53	1.54	−0.02	0	0.05	0.04
10	0.65	0.65	0.64	0.65	0.61	0	0.01	0	0.04
11	0.98	1.05	1.12	1.07	1.05	−0.07	−0.14	−0.09	−0.07
12	0.82	0.78	0.71	0.78	0.84	0.04	0.11	0.04	−0.02
13	1.39	1.39	1.49	1.49	1.43	0	−0.1	−0.1	−0.04
14	1.86	1.79	1.94	1.94	1.85	0.07	−0.08	−0.08	0.01
15	1.8	1.86	2.02	1.98	1.94	−0.06	−0.22	−0.18	−0.14
16	2.36	2.19	2.17	2.19	2.19	0.17	0.19	0.17	0.17
17	2.23	2.32	2.27	2.28	2.26	−0.09	−0.04	−0.05	−0.03
18	2.28	2.32	2.41	2.31	2.31	−0.04	−0.13	−0.03	−0.03
19	2.33	2.23	2.3	2.3	2.28	0.1	0.03	0.03	0.05
20	2.61	2.69	2.41	2.56	2.58	−0.08	0.2	0.05	0.03
21	1.65	1.68	1.71	1.7	1.66	−0.03	−0.06	−0.05	−0.01
22	2.1	2.13	2.07	2.11	2.11	−0.03	0.03	−0.01	−0.01
23	2.67	2.53	2.5	2.55	2.58	0.14	0.17	0.12	0.09
24	2.64	2.56	2.52	2.54	2.57	0.08	0.12	0.1	0.07
25	2.83	2.97	2.93	2.93	2.9	−0.14	−0.1	−0.1	−0.07
26	3.14	3.11	3.05	3.09	3.15	0.03	0.09	0.05	−0.01
27	3.09	3.07	3.09	3.03	3.13	0.02	0	0.06	−0.04
28	3.06	2.96	3.16	3.15	3.06	0.1	−0.1	−0.09	0
29	3.4	3.43	3.38	3.4	3.4	−0.03	0.02	0	0
30	2.45	2.46	2.42	2.48	2.46	−0.01	0.03	−0.03	−0.01
31	1.05	1.25	1.11	1.07	1.11	−0.2	−0.06	−0.02	−0.06
32	1.59	1.73	1.66	1.66	1.58	−0.14	−0.07	−0.07	0.01
33	1.61	1.55	1.48	1.6	1.57	0.06	0.13	0.01	0.04
34	1.49	1.59	1.51	1.5	1.43	−0.1	−0.02	−0.01	0.06

[a] Experimental values of Log P

[b,c,d,e] Predicted (calculated) values by using Eqs. 3, 4, 5 and 6, respectively

[f,g,h,i] Residual values of by using Eqs. 3, 4, 5 and 6, correspondingly [Res = Log P(Obsd.) − Log P(Pred.)]

$$\log K = 7.56(\pm 0.27) - 0.014(\pm 0.003)\left[{}^{A/Pd}_{1}I_{50}(HT)\right] - 5.00(\pm 0.19)\left[{}^{V/In}_{3}A^{E}(AH)\right] - 1.92(\pm 0.09)\left[{}^{P/In}_{3}I^{E}_{50}(AH)\right] - 0.34(\pm 0.11)\left[{}^{P/Pd}_{3}V^{E}(Cb)\right] + 0.005(\pm 0.0004)\left[{}^{A/Pd}_{3}I_{50}(DH)\right] \quad (8)$$

$N = 34 \quad R^2 = 98.47 \quad s = 0.192 \quad Q^2 = 97.79$

$F = 359.88$

$$\log K = 5.76(\pm 0.15) - 4.57(\pm 0.18)\left[{}^{V/In}_{3}A^{E}(AH)\right] - 1.44(\pm 0.09)\left[{}^{P/In}_{3}I^{E}_{50}(AH)\right] + 0.006(\pm 0.0005)\left[{}^{A/Pd}_{3}I_{50}(AH)\right] + 1.31(\pm 0.28)\left[{}^{E/Pd}_{3}PN^{E}(HT)\right] - 0.47(\pm 0.13)\left[{}^{V/Pd}_{3}G^{E}(DH)\right] + 0.49(\pm 0.099)\left[{}^{V/In}_{9}Q^{E}_{2}(Cb)\right] \quad (9)$$

$N = 34 \quad R^2 = 98.86 \quad s = 0.169 \quad Q^2 = 98.26$

$F = 389.39$

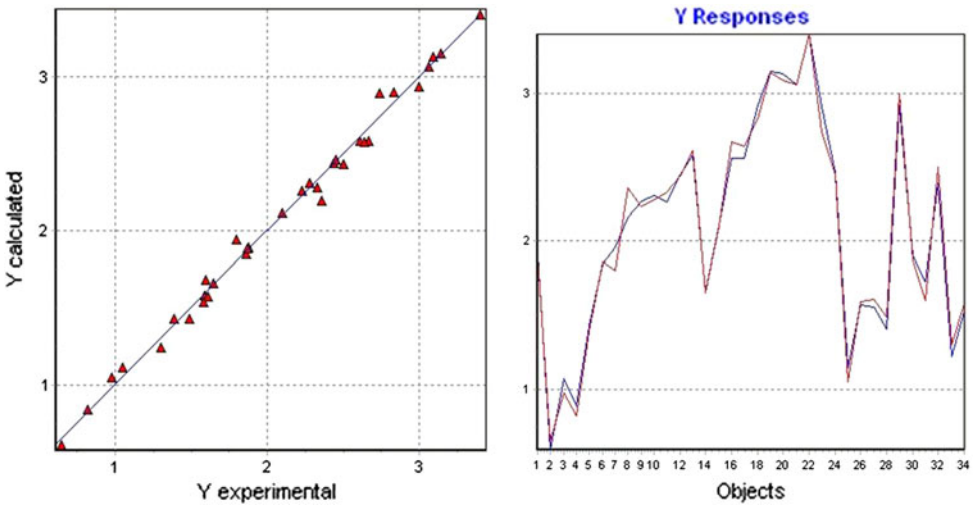


Fig. 3 Correlation between experimental and calculated (by Eq. 3) log P of 34 derivatives of 2-furylethylenes of the data set

Table 5 Statistical information for best multiple regression (4–7 parameters) QSPR models that describe Log P of 34 derivatives of 2-furylethylenes by using different MDs

Index	n	R^2	s	q^2	q^2_{boot}	s_{CV}	F
Derivative Indices (Eq. 3)	4	97.10	0.13	96.19	95.79	0.138	242.48
Derivative Indices (Eq. 4)	5	98.44	0.097	97.81	97.41	0.105	352.67
Derivative Indices (Eq. 5)	6	98.81	0.086	98.16	97.68	0.096	374.08
Derivative Indices (Eq. 6)	7	99.19	0.072	98.72	98.34	0.080	453.84
Vertex and edge 2/3D conn. Indices [28]	7	93.9	0.199	*	*	0.247	56.9
Topographic descriptors [28]	7	96.4	0.155	*	*	0.176	84.6
Quantum chemical descriptors[28]	R & C[a]	87.5	0.319	*	*	0.37	45.5

$$\begin{aligned}\log K = {} & 4.37(\pm 0.17) - 0.017(\pm 0.004)\left[{}^{P/In}_{2}M^{E}(HT)\right] \\ & - 0.54(\pm 0.12)\left[{}^{P/In}_{3}DE^{E}(AH)\right] \\ & - 0.78(\pm 0.04)\left[{}^{V/In}_{5}N^{E}_{1}(HT)\right] \\ & - 0.123(\pm 0.003)\left[{}^{V/In}_{5}CV^{E}(HT)\right] \\ & + 1.98(\pm 0.05)\left[{}^{E/Pd}_{5}Q^{E}_{3}(AH)\right] \\ & + 0.03(\pm 9.1\cdot 10^{-4})\left[{}^{E/In}_{5}Q_{1}(DH)\right] \\ & + 0.002(\pm 0.0001)\left[{}^{V/In}_{5}I_{50}(Cb)\right] \qquad (10)\end{aligned}$$

$$N = 34 \quad R^2 = 99.81 \quad s = 0.069 \quad Q^2 = 99.70 \quad F = 2003.08$$

Table 6 reflects the experimental and calculated values Log K according to the models 7–10 as well the corresponding residual values. The linear correlations that exist between the calculated (Eq. 10) and experimental values of Log K for 2-furylethylenes derivatives are illustrated in Fig. 4. The statistical parameters of these models are depicted in Table 7.

As can be seen, all regression equations showed rather good behavior in the description of Log P and Log K of 2-furylethylenes derivatives. It should be remarked that although the four-parameter regression models (Eqs. 3, 7) had rather good predictive power (Q^2 of 96.19 and 97.92, respectively), the inclusion of new MDs depicted a *dramatic* improvement of statistical R^2, s, and Q^2 values from the four-parameter equations. Similarly, the best seven-parameter combination that yields correlations with Log P and Log K was searched to give the highest *statistical parameters values* (Q^2 of 98.72 and 99.70, respectively).

Similar-to-inferior equations were reported by Estrada and Molina by using seven variables in the models (2/3D edge- and vertices-based connectivity indices, total and local spectral moments, as well as topological and quantum chemical descriptors). The statistical parameters of best equations obtained by these authors in the description of Log P and Log K are given in Tables 5 and 7, respectively. These models explain more than 93.9 % (2/3D edge- and

Table 6 Experimental and predicted (by using 4–7 variables) values of log K of 34 2-furylethylenes

No.	Obsd.[a]	Eq. 7[b]	Eq. 8[c]	Eq. 9[d]	Eq. 10[e]	Res[f]	Res[g]	Res[h]	Res[i]
1	6.591	6.54	6.35	6.53	6.59	0.051	0.241	0.061	0.001
2	6.518	6.77	6.61	6.45	6.53	−0.252	−0.092	0.068	−0.012
3	6.914	6.73	6.82	6.76	6.85	0.184	0.094	0.154	0.064
4	6.982	6.74	6.93	6.77	7	0.242	0.052	0.212	−0.018
5	7.176	6.98	6.99	7.2	7.22	0.196	0.186	−0.024	−0.044
6	7.602	7.68	7.66	7.83	7.55	−0.078	−0.058	−0.228	0.052
7	5.255	5.55	5.61	5.55	5.26	−0.295	−0.355	−0.295	−0.005
8	6.763	6.57	6.69	6.64	6.79	0.193	0.073	0.123	−0.027
9	5.623	5.68	5.62	5.63	5.59	−0.057	0.003	−0.007	0.033
10	3.813	3.83	3.87	3.76	3.79	−0.017	−0.057	0.053	0.023
11	3.84	3.82	3.79	3.64	3.76	0.02	0.05	0.2	0.08
12	3.874	4	3.81	3.88	4.01	−0.126	0.064	−0.006	−0.136
13	3.825	3.82	3.85	3.82	3.78	0.005	−0.025	0.005	0.045
14	3.623	3.61	3.65	3.7	3.83	0.013	−0.027	−0.077	−0.207
15	3.751	3.68	3.58	3.61	3.76	0.071	0.171	0.141	−0.009
16	3.784	3.37	3.4	3.48	3.74	0.414	0.384	0.304	0.044
17	3.697	3.63	3.76	3.72	3.73	0.067	−0.063	−0.023	−0.033
18	3.705	3.53	3.54	3.56	3.7	0.175	0.165	0.145	0.005
19	3.697	4.47	4.25	4.11	3.79	−0.773	−0.553	−0.413	−0.093
20	3.65	3.62	3.75	3.78	3.67	0.03	−0.1	−0.13	−0.02
21	4	4.19	4.13	4.11	4.01	−0.19	−0.13	−0.11	−0.01
22	3.92	3.86	3.87	3.89	3.88	0.06	0.05	0.03	0.04
23	3.79	3.8	3.84	3.94	3.75	−0.01	−0.05	−0.15	0.04
24	3.763	3.67	3.56	3.68	3.74	0.093	0.203	0.083	0.023
25	3.623	3.65	3.67	3.72	3.61	−0.027	−0.047	−0.097	0.013
26	3.65	3.65	3.77	3.72	3.58	0	−0.12	−0.07	0.07
27	3.592	3.34	3.34	3.39	3.54	0.252	0.252	0.202	0.052
28	3.584	3.65	3.45	3.42	3.67	−0.066	0.134	0.164	−0.086
29	3.59	3.45	3.5	3.57	3.53	0.14	0.09	0.02	0.06
30	2.987	3	3.06	2.98	2.96	−0.013	−0.073	0.007	0.027
31	3.273	3.17	3.28	3.35	3.23	0.103	−0.007	−0.077	0.043
32	2.14	2.43	2.34	2.33	2.09	−0.29	−0.2	−0.19	0.05
33	3.553	3.58	3.63	3.58	3.63	−0.027	−0.077	−0.027	−0.077
34	5.557	5.65	5.74	5.59	5.53	−0.093	−0.183	−0.033	0.027

[a] Experimental values of Log K

[b,c,d,e] Predicted (calculated) values by using Eqs. 7, 8, 9 and 10, respectively

[f,g,h,i] Residual values of by using Eqs. 7, 8, 9 and 10, correspondingly [Res = Log K(Obsd.) − Log K(Pred.)]

vertices-based connectivity indices), 96.4 % (topological and topographic indices) and 87.5 % (quantum chemical descriptors) of the variance of the experimental Log P values (see Table 5). Our worst model (Eq. 3, four variables) explains more than 97 % of the variance of the experimental Log P values. For Log K, similar behavior is depicted, i.e., our four-parameter model (Eq. 7, Q^2 of 97.92) shows better performance than the best equation (Q^2 of 96.8) reported using seven quantum parameters. Unfortunately, Estrada and Molina do not report the results for the cross-validation. It is remarkable that our models, Eqs. 6 and 10, with the same number of variables as those for the best models previously obtained by these authors explain a greater percentage of the variance of the experimental Log P and Log K values; showing a decrease in the standard error of 53.54 % and 76.04 %, respectively. Tables 5 and 7 summarize the statistical parameters achieved by all these approaches.

Concluding remarks

The approach described in this report appears to be a promising method for finding quantitative models that describe physical, thermodynamic, or biological properties. The novel MDs proposed here have been shown to have some interesting features, such as:

1. Their functional definitions are based on novel algorithms and formulae in mathematics. These novel

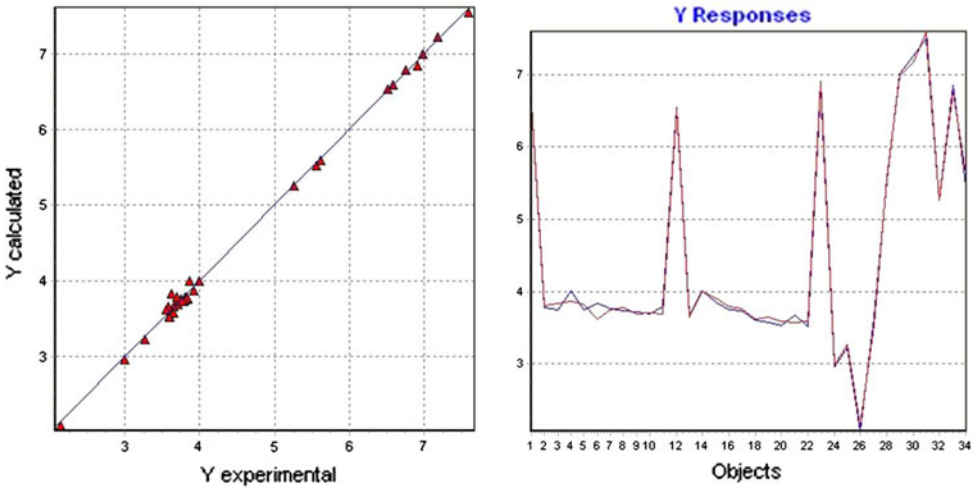


Fig. 4 Correlation between experimental and calculated (by Eq. 7) log K of 34 derivatives of 2-furylethylenes of the data set

Table 7 Statistical information for best multiple regression (4–7 parameters) QSPR models that describe Log K of 34 derivatives of 2-furiletilenes by using different MDs

Index	n	R^2	s	q^2	q^2_{boot}	s_{CV}	F
Derivative indices (Eq. 7)	4	97.92	0.22	97.29	96.74	0.231	341.08
Derivative indices (Eq. 8)	5	98.47	0.192	97.79	97.21	0.209	359.88
Derivative indices (Eq. 9)	6	98.86	0.169	98.26	97.73	0.186	389.39
Derivative indices (Eq. 10)	7	99.81	0.069	99.7	98.47	0.111	2,003.08
Conn. indices [27]	7	82.1	0.681	*	*	*	17.1
Global spectral moments [27]	7	84.3	0.655	*	*	*	18.8
Local spectral moments [27]	7	96.4	0.32	*	*	*	70.4
Quantum chemical descriptors [27]	7	96.8	0.288	*	*	*	112.2

atom-based MDs are based on a graph derivative for pairs of vertex rather similar to those defined in discrete mathematics. The atom- and group-level as well as atom-type formalisms will permit to expedite the investigation of molecular mechanisms and rational design of molecules at local level.

2. These local and global indices are incorporated as a new set of MDs to the significant arsenal of whole-molecule indices. We also define, ***for the first time,*** 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 MG derived from the generalization of an incidence matrix whose row entries correspond to connected sub-graphs of a given MG (Q matrix). Also, from this matrix, other new representations can be obtained: a) the relations frequency **F** matrix and b) the derivate £ matrix.
4. The calculation of the proposed indices is simple and straightforward, requiring only the 2D information. The novel indices are implemented in **DIVATI**, [30] a new module of ***TOMOCOMD-CARDD*** program [26] to facilitate their quick and easy computation.
5. These indices show good prediction power in physicochemical-properties modeling. The QSPR models presented here demonstrate a good statistical account of the 2-furylethylenes derivatives data. Further, it was clearly demonstrated that this set of descriptors produced better models than the other 2/3D TIs and geometric set of indices previously tested by different researchers.

Future perspective

Despite these positive features of atom-based derivative indices, additional study has to be done 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 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 relation frequency hyper-matrix (derivative for *n*-tuple). We will also introduce new *events* derived from other graph-theoretical and geometric concepts that permit us to define other relation 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.

Finally, structural and physico-chemistry interpretations (for both the new LOVIs and invariants introduced here) as well as the significance (similarity/dissimilarity analysis) of these new MDs will be subject of posterior studies. Specifically, the physical meaning of these indices is a rather important aspect because this attribute (interpretability) has been the source of criticism for using TIs in chemometrics tasks. In this study, our MDs showed better statistical results than those obtained previously using some of the most successful families of MDs in chemometric and drug design practice. It is obvious that this good behavior is related to the chemical information codified by these indices, and for this reason, separation of this success in describing the molecular structure from the structural interpretation of these indices would be inadequate. Therefore, if the new indices are successful in predicting properties that depends on molecular structure, then it is obvious that these indices have an important and interesting physical meaning [35–39]. As recently concluded by Estrada [40] "Because molecules are "physical" objects, i.e., objects in a real world, topological indices are mathematical representations of a physical reality. Consequently, they necessarily have to have a physical basic meaning." If in "continuous" mathematics, the ***derivative*** concepts is used as a mathematical representation of the reality (as the first derivative of distance with respect to time is to velocity), we will hope that discrete derivative on chemical graphs introduced here (in a mathematical language) to codifying chemical information is a good mathematical representation of this reality but should have physical interpretation as well.

The next papers will present an effort to solve this "mystery" and with these results, we hope to increase the interest in the application of these new indices to describe physical and biological properties as well as to provide a physical place for *atom-based derivative indices* among the pool of MDs. In conclusion, we are sure that the necessity of considering physical interpretation of these new MDs ***is a deep truth***, but also ***is a deep truth*** that we can codify chemical information of molecules by using only our new 2D MDs, that is the *atom-based global and local derivative indices*, and these should have physical interpretation. This fact reminds us of the phrase stated by Bohr that, "When you have a deep truth, then the opposite of a deep truth may again be a deep truth".

Acknowledgments Marrero-Ponce, Y. thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University in 2011. The authors acknowledge also the partial financial support from Spanish *"Comisión Interministerial de Ciencia y Tecnología"* (CICYT) (Project Reference: SAF2009-10399). Finally, but not least, this work was supported in part by VLIR (Vlaamse InterUniversitaire Raad, Flemish Interuniversity Council, Belgium) under the IUC Program VLIR-UCLV.

References

1. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Germany
2. Todeschini R, Consonni V, Pavan M (2002) 2.1 ed., Milano Chemometric and QSAR Research Group, Milano, Italy
3. Katritzky AR, Perumal S, Petrukhin R, Kleinpeter E (2001) J Chem Inf Comput Sci 41:569
4. 2.13 ed. 7204 Mullen, Shawnee, KS 66216, USA
5. Gugisch R, Kerber A, Laue R, Meringer M, Weidinger J University of Bayreuth, D-95440 Bayreuth, Germany
6. Topological Indices and Related Descriptors in QSAR and QSPR; Devillers, J. B., A. T, Ed.; Gordon and Breach Amsterdam, The Netherlands, 1999
7. Todeschini R, Consonni V (2010) MATCH Commun. Math. Comput. Chem 64:359
8. Devillers J (2000) Curr Opin Drug Discovery Dev. 3
9. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley, New York
10. Estrada E, Uriarte E (2001) Curr Med Chem 8:1573
11. Estrada E, Rodríguez L (1997) Comm Math Chem (MATCH) 35:157
12. Randić M (1991) J Math Chem 7:155
13. Randić M, Trinajstić NJ (1993) Mol Struct (Theochem) 300:551
14. Kier LB, Hall LH (1999) Molecular structure description. The electrotopological state. Academic Press, New York
15. Rouvray DH (1976) Chemical applications of graph theory. Academic Press, London
16. Kier LB, Hall LH (1986) Molecular connectivity in structure—activity analysis. Research Studies Press, Letchworth
17. Ivanciuc O, Gasteiger J (2003) Ed. Wiley-VCH, Weinheim, p 103
18. Balaban AT (1997) From chemical graphs to three-dimensional geometry. Plenum Press, New York
19. Estrada E (2001) Chem Phys Lett 336:248
20. Estrada E, Rodríguez L, Gutierrez A (1997) Commun Math Chem (MATCH) 35:145
21. Aires-de-Sousa J, Gasteiger J (2002) J Mol Graph Model 20:373
22. Golbraikh A, Bonchev D, Tropsha A (2001) J Chem Inf Comput Sci 41:147
23. Marrero-Ponce Y, Castillo-Garit JA, Castro EA, Torrens F, Rotondo RJ (2008) Math. Chem. doi:10.1007/s10910-008-9386-3
24. Gorbátov VA (1988) Fundamentos de la Matematica discreta, Mir, Moscow, URSS
25. Daudel R, Lefebre R, Moser C (1984) Quantum chemistry: methods and applications. Wiley, New York
26. Marrero-Ponce Y (2002) version 1.0 ed., Unit of computer-aided molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit): Santa Clara, Villa Clara

27. Estrada E, Molina E (2001) J Mol Graph Model 20:54
28. Estrada E, Molina E (2001) J Chem Inf Comput Sci 41:791
29. Dore JC, Viel C (1975) Farmaco 30:81
30. Martínez Santiago O, Martínez-López Y, Marrero-Ponce Y (2010) version 1.0 ed., Unit of computer-aided molecular "Biosilico" discovery and bioinformatic research (CAMD-BIR Unit), Santa Clara, Villa Clara, Cuba
31. Todeschini R, Consonni V, Mauri A, Pavan M (2005) 1.0 ed., Talete, Milano
32. Goldberg DE (1989) Genetic algorithms. Addison Wesley, Reading
33. Rogers D, Hopfinger AJJ (1994) Chem Inf Comput Sci 34:854
34. So SS, Karplus M (1996) J Med Chem 39:1521
35. Stankevich V, Skvortsova MI, Zefirov NSJ (1995) Mol Struct (THEOCHEM) 342:173
36. Galvez JJ (1998) Mol Struct (THEOCHEM) 429:255
37. Kier LB, Hall LHJ (2000) Chem Inf Comput Sci 40:792
38. Kier LB, Hall LH (2001) J Mol Graph Model 20:76
39. Randic M, Hansen PJ, Jurs PCJ (1988) Chem Inf Comput Sci 28:60
40. Estrada EJ (2002) Phys Chem A 106:9085