MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **73** (2015) 397-420

ISSN 0340 - 6253

Extending Graph (Discrete) Derivative Descriptors to *N*-Tuple Atom-Relations

Oscar Martínez Santiago[a,b], Yovani Marrero-Ponce[a,c,d*], Reisel Millán Cabrera[a,b,e], Stephen J. Barigye[a], Yoan Martínez-López[a,f], Luis M. Artiles Martínez[a], José O. Guerra de León[b], Facundo Perez-Giménez[c], Francisco Torrens[g]

[a]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Universidad Central "Martha Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba.* **ymarrero77@yahoo.es or ymponce@gmail.com URL: http://www.uv.es/yoma/**
[b]*Departament of Chemical Science, Faculty of Chemistry-Pharmacy. Universidad Central "Martha Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*
[c]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.*
[d]*Facultad de Química Farmacéutica, Universidad de Cartagena, Cartagena de Indias, Bolívar, Colombia.*
[e]*Department of Chemical Science, Camaguey University, Camaguey City, 74650, Camaguey Cuba.*
[f]*Department of Computer Sciences, Camaguey University, Camaguey City, 74650, Camaguey Cuba.*
[g]*ICMOL, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071, València, Spain.*

(Received June 8, 2014)

Abstract

In the present manuscript, an extension of the previously defined Graph Derivative Indices (GDIs) is discussed. To achieve this objective, the concept of a hypermatrix, conceived from the calculation of the frequencies of triple and quadruple atom relations in a set of connected sub-graphs, is introduced. This set of subgraphs is generated following a predefined criterion, known as the event (S), being in this particular case the connectivity among atoms. The triple and quadruple relations frequency matrices serve as a basis for the computation of triple and quadruple discrete derivative indices, respectively. The GDIs are implemented in a computational program denominated DIVATI (acronym for DIscrete DeriVAtive Type Indices), a module of TOMOCOMD-CARDD program. Shannon's entropy-based variability analysis demonstrates that the GDIs show major variability than others indices used in QSAR/QSPR researches. In addition, it can be appreciated when the indices are extended over n-elements from the graph, its quality increases, principally when they are used in a combined way. QSPR modeling of the physicochemical properties Log P and Log K of the 2-furylethylenes derivatives reveals that the GDIs obtained using the triple

and quadruple matrix approaches yield superior performance to the duplex matrix approach. Moreover, the statistical parameters for models obtained with the GDI method are superior to those reported in the literature by using other methods. It can therefore be suggested that the GDI method, seem to be a promissory tool to reckon on in QSAR/QSPR studies, virtual screening of compound datasets and similarity/dissimilarity evaluations.

1 Introduction

A representation of an object that only provides information on the number of elements composing it and their connectivity is known as a *topological representation*. For a molecular structure, a topological representation is achieved through the so-called molecular graphs [1-3]. However, such representations are non-numeric in nature, which makes computational treatment of chemical information provided by these rather difficult. In this sense, *topological indices* (TIs) constitute a practical aperture as they provide a numeric interpretation of information codified by molecular graphs (G) [1]. The TIs (also known as *graph-theoretical invariants*) account for structural information contained in bi-dimensional representations of molecules and are among the most important molecular descriptors (MDs) used nowadays in the theoretical description of chemical, physicochemical and biological properties of molecular structures [4-11].

The TIs are divided into 2 categories: topo-structural and topo-chemical indices. The topo-structural indices are concerned with the adjacency and distances between vertices in a **G** while topo-chemical indices in addition to offering the information on the topology of a **G**, also codify information on the nature of the vertices such as their chemical identity or hybridization state.

Recently, a new family of TIs has been published that could be defined as an overlap of both classifications, previously noted [12, 13]. This set of MDs, collectively denominated as ***Graph Derivative Indices*** (GDIs), is based on the concept of derivatives in discrete mathematics, metaphorical to the derivative concept in classical mathematical analysis [12]. The Discrete Derivative $\partial G/\partial S$ is defined on a weighted undirected graph <V,(U, P)>, whose bearer coincides with that of a model determined by a chosen *event* (S) [14]. These new TIs have been applied over the data set proposed by the Mathematical-Chemistry Academy for the validation of new MDs, and they have been used in several applications with good results [12, 13]. The GDIs can be obtained in a local way, detecting a relation between the values of the atomic indices [Local Vertex invariants (LOVIs)] and the chemical nature of each atom in the molecular structure. It has been probed in previous experiments that the LOVI value for each atom has a direct relation with the electronic environment of those atoms. The GDs can be obtained in a global way applying many different mathematical strategies (aggregation operators), as it is going to be showed in other epigraph. The obtaining of GDIs in a global way also allows establishing a differentiation among isomers of chain, position and geometrical isomers [13]. In addition, a new matrix representation denominated the relations

frequency matrix, **F**, has been presented. This matrix representation arises from the exploration of *duplex* participation frequencies of ***connected sub-graphs*** (*event* initially used) in the formation of a **G** [12].

The present report is aimed at introducing the concept of a *hypermatrix*, conceived from the evaluation of the *n*-tuple (*n*>2) participation frequencies of *connected sub-graphs* in the formation of a **G**. Although the participation frequencies are unbounded, our attention will be focused on triple (*n*=3) and quadruple (*n*=4) participation frequencies. These hypermatrix representations will permit us to "redefine" the GDIs in a "more generalized" way (for *n*-tuples).

With the aim of evaluating the quality of the GDIs in terms of their sensitivity to variations in the molecular structure, Shannon's entropy-based variability analysis [15] is performed and comparisons between duplex and higher dimensional GDIs are carried out. On the other hand, in order to assess the performance of the proposed MDs in modeling tasks, the 1-octanol/water partition coefficient (Log P) and the specific rate constant for nucleophilic addition of a thiol group to the exo-cyclic double bond (Log K) of the 34 derivatives of 2-furylethylenes are studied; and the statistical parameters of the best models obtained for these physicochemical properties using the proposed GDIs are compared with those of other approaches reported in the literature.

2 Theoretical Scaffold

2.1 Frequency hypermatrix representation of a molecular graph

First, a brief recapitulation of the aspects presented in a previous publication will be performed to ease in the definitions and notations discussed in this report.

To begin with, an *event* **S** is defined as the criterion followed in the generation of a collection of conditions representative of a predefined model. In other words, the event provides the context for the model. Consider the following paragraph as a model.

*"In any r**ea**ction, **e**nth**a**lpy and entropy ch**a**ng**e** when the r**ea**ctants are used up to obtain products."*

The event in this case would be a thermodynamic description of chemical reactions. Each description is comprised of a collection of conditions (words) as "building blocks" which form a model in the defined event space. The key interest in this case is to analyze the contributions of the different characters (or combinations of these) in the set of conditions that collectively constitute the model, as a means of acquiring knowledge about the diversity of the model. The contributions (frequencies) of the different characters constitute a *frequency relations matrix*.

Let us consider **two letter** participation frequencies in the formation of **words** (duplex participation frequencies) using the model above, as an example. The **letters** {a, e} simultaneously contribute in the formation of the **words**: *r***ea***ction*, *ch***a***ng***e***s*, ***e****nth***a***lpy*, *r***ea***ct*, *i.e.* participate four times to the creation of words that comprise the model, f_{ae} = 4. The participation frequencies of all possible two component subsets of **letters** (f_{ij}) are similarly explored, as well as the participation frequencies of each **letter** (f_i) that constitute these **words**. These contributions (frequencies) are the components for the *frequency relations matrix*, **F**. In a preceding manuscript, only duplex participation frequencies were considered, although a purely mathematical interpretation based on generalized incidence matrices was used [12]. The preference for the exploratory method discussed in the present manuscript is because it offers a simpler interpretation of the event-based method, favoring the generalization of the graph derivative method to higher dimensions.

A natural extension of the duplex-based approach involves the evaluation of the *concurrent participation frequencies of n (n ≥ 3) letters in the formation of words in the model.* The present manuscript enunciates the procedure adopted for the analysis of the event space on a higher dimensional scale. Consider the participation of the **letters** {a, e, n} and {e, n, t, y} in the formation of **words** in the model given above: the **letters** {a, e, n} simultaneously contribute to the **words**: *r***ea***ctio***n,** *ch***an***g***e***s*, ***en****th***a***lpy*, *i.e.* participate three times in the formation of **word** of the model, f_{aen} = 3. On the other hand, the **letters** {e, n, t, y} participate concurrently in the formation of the **words**: ***ent****hal***py**, ***ent****rop***y**, *i.e.* possess a participation frequency of two (f_{enty}). An exploration of the participation of subsets of *5, 6, 7…n* letters in the formation of words of the model could be performed as well. Nevertheless, this analysis will be limited to triple and quadruple participation frequencies for simplicity.

In view of the ease provided by the matrix-based operations in computational chemistry, these participation frequencies are condensed in three- and four-dimensional matrices, which we will designate triple and quadruple hypermatrices (3- and 4-order tensors), respectively.

This set of rules is applicable to any system of discrete macroarrangements that are in turn comprised of microunits such as genetic codes or chemical structures. Our interest is in the latter. Given a molecular structure **G**, this is partitioned into a set of substructures, according to a predefined criterion. This criterion is the event (S) and provides the context in which the substructures are formed. In the preceding manuscript, *connectivity* was used as a rule and thus the ensuing substructures were denominated *connected subgraphs*. The concepts of sub-graphs orders and types (according to Kier-Hall's nomenclature, namely: path (*p*), cluster (*c*) and path-cluster (*pc*) were taken into account). In this sense, the ***conditions*** (**letters**

of the model) are the ***vertices (atomic nuclei)*** that comprise the substructures, while the ***collection of conditions*** is the ***connected sub-graphs*** analogical to the **words** of the model. An example of the application of this set of rules in chemical structure characterization, on a higher dimensional perspective, will now be given. Take as an example the molecular graph of N,N-dimethylpropionamide (*see* Figure 1) describing the skeleton of this molecule.

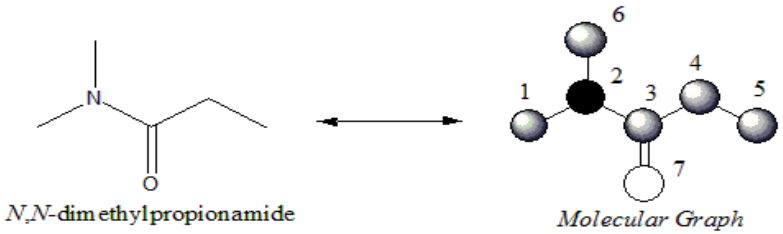

Fig. 1. The chemical structure [H (implicit)-depleted structure] and molecular graph of N,N-dimethylpropionamide [the numbers correspond to the labels that are assigned to the atoms (vertices) in the molecular structure].

Following the connectivity criterion, a set of sub-graphs of different orders and types are obtained (*see* Table 1). Accordingly, these ***connected sub-graphs*** are the set of **words** (sub-structures) that constitute the model space, while the ***vertices*** (atoms) for **G,** [C_1, N_2, C_3, C_4, C_5, C_6 and O_7] are the **letters**. From this set of connected sub-graphs, an exploration of the

Table 1: Connected sub-graphs of different orders and types from N,N-dimethylpropionamide.

Order	Type	Sub-graph	Order	Type	Sub-graph
Order 0	paths	C1	Order 3	paths	C1-N2-C3-C4
	paths	N2		paths	C1-N2-C3-O7
	paths	C3		cluster	C1-N2-C3-C6
	paths	C4		paths	N2-C3-C4-C5
	paths	C5		paths	N2-C3-C6-O7
	paths	C6		paths	N2-C3-C4-C6
Order 1	paths	O7		cluster	N2-C3-C4-O7
	paths	C1-N2		paths	C3-C4-C5-O7
	paths	N2-C3		paths	C1-N2-C3-C4-C5
	paths	N2-C6		paths-cluster	C1-N2-C3-C4-O7
	paths	C3-C4		paths-cluster	C1-N2-C3-C4-C6
	paths	C3-O7	Order 4	paths	N2-C3-C4-C5-C6
	paths	C4-C5		paths-cluster	N2-C3-C4-C5-O7
Order 2	paths	C1-N2-C3		paths-cluster	N2-C3-C4-C6-O7
	paths	C1-N2-C6		paths-cluster	C1-N2-C3-C6-O7
	paths	N2-C3-C6		paths-cluster	C1-N2-C3-C4-C5-C6
	paths	N2-C3-C4	Order 5	paths-cluster	C1-N2-C3-C4-C5-O7
	paths	N2-C3-O7		paths-cluster	N2-C3-C4-C5-C6-O7
	paths	C3-C4-C5		paths-cluster	C1-N2-C3-C4-C6-O7
	paths	C3-C4-O7	Order 6	paths-cluster	C1-N2-C3-C4-C5-C6-O7

frequency of concurrent participation of three- or four vertex subsets, corresponding to triple and quadruple matrices, respectively, is performed. These frequencies are components of the respective hypermatrices.

Triple Matrix

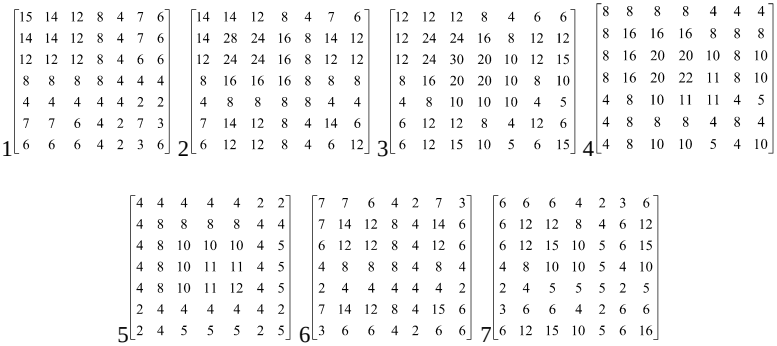

Fig. 2. The triple matrix generated for the Graph of N,N-dimethylpropionamide.

As an example, the set of vertices [C_1, N_2, O_7] are included in connected sub-graphs: C_1-N_2-C_3-O_7, C_1-N_2-C_3-C_4-O_7, C_1-N_2-C_3-C_6-O_7, C_1-N_2-C_3-C_4-C_5-O_7, C_1-N_2-C_3-C_4-C_6-O_7 and C_1-N_2-C_3-C_4-C_5-C_6-O_7. Thus the participation frequency of the set of vertices [C_1, N_2, O_7] is six [see entry (1, 2, 7)] in the triple matrix represented below (see Figure 2). For clarity, the slides (bi-dimensional matrices) that comprise this matrix are "extracted". For bi-dimensional representation of quadruple matrix see supporting information SI1.

2.2 Graph's discrete derivative-based indices

The triple and quadruple relations frequency matrices serve as a basis for the computation of triple and quadruple discrete derivative indices, respectively and these are expressed as follows:

Triple derivative index

$$\frac{\partial G}{\partial S}(m_1, m_2, m_3) = \frac{1}{f_{m1,m2,m3}} \left(\sum_{i=m1,m2,m3} f_i - 2 \cdot \sum_{i \neq j, i, j=m1,m2,m3} f_{ij} + 3 \cdot \sum_{\substack{i \neq j, i \neq k, j \neq k \\ i,j,k=m1,m2,m3}} f_{ijk} \right) \qquad \textbf{(2.1)}$$

where, f_i is the participation frequency of vertex i, f_{ij} is the simultaneous participation frequency of vertices i and j, and f_{ijk} is the participation frequency of vertices i, j and k.

Quadruple derivative index

$$\frac{\partial G}{\partial S}(m_a, m_b, m_c, m_d) = \frac{1}{f_{m_a m_b m_c m_d}}\left(\sum_i f_i - 2\cdot\sum_{\substack{i,j \\ i\neq j}} f_{ij} + 3\cdot\sum_{\substack{i,j,k \\ i\neq j, i\neq k, j\neq k}} f_{ijk} - 4\cdot\sum_{\substack{i,j,k,l \\ i\neq j, i\neq k, i\neq l, j\neq k, j\neq l, k\neq l}} f_{ijkl}\right) \quad \textbf{(2.2)}$$

where, f_i is the participation frequency of vertex i, f_{ij} is the simultaneous participation frequency of vertices i and j, f_{ijk} is the participation frequency of vertices i, j and k and f_{ijkl} is the participation frequency of vertices i, j, k and l.

Generalizing the formula for n-tuples derivatives we obtain:

$$\frac{\partial G}{\partial S}(m_1, m_2, \ldots m_n) = \frac{1}{f_{m_1 m_2 \ldots m_n}}\left(\sum_{i_1} f_i - 2\cdot\sum_{\substack{i_1,i_2 \\ i_1\neq i_2}} f_{i_1 i_2} + \ldots + (-1)^{\alpha+1}\cdot\alpha\cdot\sum_{\substack{i_1,i_2\ldots i_\alpha \\ i_1\neq i_2,\ldots i_{\alpha-1}\neq i_\alpha}} f_{i_1 i_2 \ldots i_\alpha} + \ldots + (-1)^{n+1}\cdot n\cdot\sum_{\substack{i_1,i_2,\ldots i_n \\ i_1\neq i_2,\ldots i_{n-1}\neq i_n}} f_{i_1 i_2 \ldots i_n}\right) \quad \textbf{(2.3)}$$

An illustration for the computation of the derivative index will now be given, using the triple relations frequency matrix generated for the molecular structure of N,N-dimethylpropionamide represented in the figure 1. Using the equation **2.1** the derivatives for all three vertex combinations are computed. Let us take (1, 2, 3) as an example:

$$\frac{\partial G}{\partial S}(C_1, N_2, C_3) = \frac{1}{12}[(15 + 28 + 30) - 2(14 + 12 + 24) + 3(12)] = 0.75.$$

The rest of the values for the derivatives over three vertex combinations for the event space generated for the molecular structure of N,N-dimethylpropionamide are:

$$\frac{\partial G}{\partial S}(C_1, N_2, C_4) = 1.625;\ \frac{\partial G}{\partial S}(C_1, N_2, C_5) = 3.750;\ \frac{\partial G}{\partial S}(C_1, N_2, C_6) = 1.286;\ \frac{\partial G}{\partial S}(C_1, N_2, O_7) = 2.167$$

$$\frac{\partial G}{\partial S}(C_1, C_3, C_4) = 1.375;\ \frac{\partial G}{\partial S}(C_1, C_3, C_5) = 4.250;\ \frac{\partial G}{\partial S}(C_1, C_3, C_6) = 2.667;\ \frac{\partial G}{\partial S}(C_1, C_3, O_7) = 2.167$$

$$\frac{\partial G}{\partial S}(C_1, C_4, C_5) = 3.750;\ \frac{\partial G}{\partial S}(C_1, C_4, C_6) = 4.50;\ \frac{\partial G}{\partial S}(C_1, C_4, O_7) = 4.250;\ \frac{\partial G}{\partial S}(C_1, C_5, C_6) = 9.00$$

$$\frac{\partial G}{\partial S}(C_1, C_5, O_7) = 9.50;\ \frac{\partial G}{\partial S}(C_1, C_6, O_7) = 5.667$$

$$\frac{\partial G}{\partial S}(N_2, C_3, C_4) = 0.50;\ \frac{\partial G}{\partial S}(N_2, C_3, C_5) = 1.250;\ \frac{\partial G}{\partial S}(N_2, C_3, C_6) = 0.750;\ \frac{\partial G}{\partial S}(N_2, C_3, O_7) = 0.667$$

$$\frac{\partial G}{\partial S}(N_2, C_4, C_5) = 2.00;\ \frac{\partial G}{\partial S}(N_2, C_4, C_6) = 1.625;\ \frac{\partial G}{\partial S}(N_2, C_4, O_7) = 1.750;\ \frac{\partial G}{\partial S}(N_2, C_5, C_6) = 3.750$$

$$\frac{\partial G}{\partial S}(N_2, C_5, O_7) = 4.50;\ \frac{\partial G}{\partial S}(N_2, C_6, O_7) = 2.167$$

$$\frac{\partial G}{\partial S}(C_3, C_4, C_5) = 1.20; \frac{\partial G}{\partial S}(C_3, C_4, C_6) = 1.375; \frac{\partial G}{\partial S}(C_3, C_4, O_7) = 0.80$$

$$\frac{\partial G}{\partial S}(C_3, C_5, C_6) = 4.250; \frac{\partial G}{\partial S}(C_3, C_5, O_7) = 2.60; \frac{\partial G}{\partial S}(C_3, C_6, O_7) = 2.167$$

$$\frac{\partial G}{\partial S}(C_4, C_5, C_6) = 3.750; \frac{\partial G}{\partial S}(C_4, C_5, O_7) = 2.60$$

$$\frac{\partial G}{\partial S}(C_4, C_6, O_7) = 4.250; \frac{\partial G}{\partial S}(C_5, C_6, O_7) = 9.50$$

From the triple (or quadruple) derivatives Local Vertex Invariants **(LOVIs)** also known as atomic derivatives are computed using the following formulas, for triple and quadruple derivatives, respectively:

$$\Delta_i = \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial G}{\partial S}(i, j, k) \qquad \textbf{(2.4)}$$

$$\Delta_i = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\partial G}{\partial S}(i, j, k, l) \qquad \textbf{(2.5)}$$

Therefore, for the molecule of N,N-dimethylpropionamide the atomic derivatives obtained following triple vertex relations are:

$$\Delta_{C1} = 56.704;\ \Delta_{N2} = 28.537; \Delta_{C3} = 26.768; \Delta_{C4} = 35.350$$

$$\Delta_{C5} = 65.650; \Delta_{C6} = 56.704; \Delta_{O7} = 54.752$$

With these atomic derivatives Δ_i, a vector of LOVIs is constructed:

V_L = (56.704, 28.537, 26.768, 35.350, 65.650, 56.704, 54.752)

It is interesting to note that atomic derivatives (atomic indices) for the vertexes 1 and 6 are the same. These results coincide with chemical reality, because both atoms are chemically equivalent. Note that peripheral atoms possess high atomic derivative values, while the lowest value belongs to atom 2 (nitrogen), which is the most buried within the molecular structure. In previous paper is evident that the increase of the electronic density in a region of the molecule implies major values of LOVIs for involved atoms, respect to a same structure without double or triple bonds.[13] This is a logical result, taking into account the probed relation among GDIs and the electronic properties of atoms and molecules [13]. In fact, for the N,N-dimethylpropionamide molecule, the value of the atomic indices (LOVIs) from the atoms of oxygen (O_7) and carbon 3 (C_3) is higher than the value hoped if it wouldn't exists a double bond among them, however there isn't an alteration in the regularity of the LOVIs values from each atom of the molecule respect to another, as it was exposed in the beginning of the paragraph.

A linear combination of the components of V_L yields the global discrete derivative index as defined in equation 2.1. Therefore, for the molecule of N,N-dimethylpropionamide yields a global discrete derivative value is 324.465.

2.3 Application of aggregation operators as a generalization of the linear combination of vector components to yield global derivative indices

Motivated by the understanding that global characterizations of chemical systems do not necessary follow the additive rule (nonextensive systems), aggregation operators are applied to the vector of LOVIs as a generalization of the use of the summation as exclusive operator applicable to the vector V_L to obtain global (or local) derivative indices. Indeed it has been demonstrated in previous reports that global indices obtained with the summation do not necessary provide the best correlations when modeling molecular properties [16]. These aggregation operators are classified into three major groups (see Table 2 for more information):

- **Norms (or Metrics)**: Minkowski's norms (N1, N2, N3) and Penrose's size (PN). *Note that the summation operation is analogous to Minkowski's first norm (N1) in our case.*

- **Mean Invariants (first statistical moment)**: Geometric Mean (G), Arithmetic Mean (M), Quadratic Mean (P2), Potential Mean (P3) and Harmonic Mean (A).

- **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), Maximum X (MX) and Minimum X (MN).

2.2.1 Codification of heteroatoms and multiple bonds: The degeneracy of several topostructural indices generally arises from the inability to discriminate isomorphic structures with distinct functional groups or bond types. This is mainly because the majority of the topostructural indices are globally defined and thus do not permit the assignation of weights on the vertices (or bonds). Therefore, with the aim of bolstering the discriminating power of the GDIs, a scheme for weighting vertices and/or bonds is introduced.

Consider as example the N,N-dimethylpropionamide molecule (see Figure 1). The notion here is to assign labels to the atoms based on their physicochemical, chemical or topological properties in order to achieve the discrimination of these structures from their topological isomorphs.

Given a vector of weights V_w, the component (ϑ_i) included in V_w reciprocally corresponds to element i for a given property. The distinct weights for each atom are determined according to the relationship $\vartheta_i = P_i/\delta_i$, where P represents a characteristic property of each atom (for example: atomic mass, electronegativity, etc.) and δ is the vertex degree (number of bonds).

Atomic electronegativity (according to Pauling's scale) will be used as an example of weights for each atom. The weights for the different atoms in the molecular structure of N,N-dimethylpropionamide are calculated as follows:

$$\vartheta(C_1) = \frac{2.5}{1} = 2.5, \vartheta(N_2) = \frac{3.0}{3} = 1.0, \vartheta(C_3) = \frac{2.5}{4} = 0.625, \vartheta(C_4) = \frac{2.5}{2} = 1.25,$$

$$\vartheta(C_5) = \frac{2.5}{1} = 2.5, \vartheta(C_6) = \frac{2.5}{1} = 2.5, \vartheta(O_7) = \frac{3.5}{2} = 1.75$$

Therefore the resulting vector of weights, V_w = (2.5, 1.0, 0.625, 1.25, 2.5, 2.5, 1.75). The inner product of the vectors V_L and V_w (V_L [x] V_W) yields the weighted vector of atomic derivatives ($^{W}V_L$). The atomic weight ϑ_i may also be directly introduced into the frequency hypermatrix ($F = [f_{ijk\ldots n}]_n{}^n$), by multiplying each frequency with the weights of the involved atoms. In this way, a weighted relations frequency hypermatrix ($F^w = [f^w{}_{ijk\ldots n}]_n{}^n$) is constructed and the equations for computing the *n*-tuple discrete derivative applied. Using the equations **2.4** or **2.5**, a weighted vector of atomic derivatives ($^{w}V_{L\text{-}f}$) is obtained.

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

No.	Group	Name	ID	Formula[a]
1		Minkowski norms (p = 1) Manhattan norm	N1	$N1 = \sum_{i=1}^{n} \lvert L_i \rvert$
2		Minkowski norms (p = 2) Euclidean norm	N2	$N2 = \sqrt{\sum_{i=1}^{n} \lvert L_i \rvert^2}$
3	**Norms** (Metrics)	Minkowski norms (p = 3)	N3	$N3 = \sqrt[3]{\sum_{i=1}^{n} \lvert L_i \rvert^3}$
4		Chebyshev distance	NI	$NI = \lim_{n\to\infty} \left[\sum_{i=1}^{n} L_i^n \right]^{1/n}$
5		Penrose size	PN	$PN = \sqrt{\frac{1}{n^2} \left[\sum_{i=1}^{n} (L_i) \right]^2}$

6		Geometric Mean	G	$G = \sqrt[n]{\prod_{i=1}^{n} L_i}$
7	**Mean** (first statistical moment)	Arithmetic Mean (Power mean of degree α = 1)	M (or M_1)	$M_\alpha = \left(\frac{L_1^\alpha + L_2^\alpha + \ldots + L_n^\alpha}{n} \right)^{\frac{1}{\alpha}}$
8		Quadratic Mean (Power mean of degree α = 2)	P2(or M_2)	
9		Power mean of degree α = 3	P3(or M_3)	
10		Harmonic Mean (Power mean of degree α = -1)	A(or M_{-1})	
11		Variance	V	$V = \frac{\sum_{i=1}^{n} (L_i - M)^2}{n-1}$
12		Skewness	S	$S = n(X_3)/[(n-1)(n-2)(DE)^3]$ n, number of vertices. $X_3 = \sum_{i=1}^{n} (L_i - M)^3$ M, arithmetic mean DE, standard deviation
13		Kurtosis	K	$K = [n(n+1)(X_4) - 3(X_2)(X_2)(n-1)] / [(n-1)(n-2)(n-3)(DE)^4]$ n, number of vertices $X_j = \sum_{i=1}^{n} (L_i - M)^j$ M, arithmetic mean DE, standard deviation
14	**Statistical** (highest statistical moments):	Standard Deviation	DE	$DE = \sqrt{\frac{(\sum L_i - M)^2}{n-1}}$
15		Variation Coefficient	CV	$CV = DE / M$
16		Range	R	$R = L_{max} - L_{min}$
17		Percentile 25	Q1	$Q1 = \left[\frac{N}{4} + \frac{1}{2} \right]$ N, L_i number
18		Percentile 50	Q2	$Q2 = \left[\frac{N}{2} + \frac{1}{2} \right]$ N, L_i number
19		Percentile 75	Q3	$Q3 = \left[\frac{3N}{4} + \frac{1}{2} \right]$ N, L_i number
20		Inter-quartile Range	I50	$I50 = Q3 - Q1$
21		Maximum value	MX	$MX = L_i$ max
22		Minimum value	MN	$MN = L_i$ min

Note: The formulae used in these invariants, are simplified forms of general equations given that the vector $\bar{y}$ is constituted of the coordinates of the origin. For example, in the case of the Euclidean norm (N2), the general formula is: $\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2 + (x_j - y_j)^2 + (x_z - y_z)^2}$ However given that $\bar{y}$ = (0, 0, 0), this formula reduces to $\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$.

Note that changing the step where the weighting scheme may be applied along the GDI algorithm yields non-identical vectors of LOVIs (i.e. applying the weights to the vector of LOVIs does not yield the same result as when applied directly to the relations frequency hypermatrix).

The use of the vectors of LOVIs (V_L, ${}^{w}V_L$, ${}^{w}V_{L\text{-}f}$) in the GDI method allows the computation of *local GDIs* for types or groups of atom [(for example, in ***TOMOCOMD-CARDD*** program [17] the following local indices may be calculated: Proton Acceptors (AH), Proton Donors (DH), Heteroatoms (HT), Halogens (HL), Carbons (Cb), Methyl Carbons (MC) and Unsaturations (IS)].

Finally, the equations for triplex and quadruple derivatives are analyzed, it is evident that restrictions on the span for the values of *i*, *j*, *k* and *l* are not considered (i.e. their values extend over the entire range [1, *n*], where *n* is the number of vertices that contain **G**). Restrictions to these equations may be included allowing to define GDIs that codify particular type of information on atom relations in the event space (See Table 3).

Note that the aggregation operators introduced in section **2.2** (see Table 2) may be applied not only to the vector of original LOVIs but also to the vector of standardized LOVIs. In the standardization procedure, the original LOVI values are converted to standardized ones using the following formula: Std. LOVIs = (Original LOVI – mean of LOVIs)/Std. With this normalization procedure, the vector of standardized LOVIs has a mean of 0 and standard deviation of 1.

Table 3: Calculation´s conditions and nomenclature of each shape of hipermatrix based calculations.

Calculation´s Conditions	Nomenclature
Triple Matrix	
Only the matrix entries that satisfy the condition $i \neq j \neq k$ are selected.	T
Only the matrix entries that satisfy the condition $i \neq j \neq k$ and $i=j \parallel j=k \parallel i=k$ are selected.	T2
Quadruple Matrix	
Only the matrix entries that satisfy the condition $i \neq j \neq k \neq l$ are selected.	C
Only the matrix entries that satisfy the condition $i \neq j \neq k \neq l$ and $i=j \parallel j=k \parallel k=l \parallel i=j \parallel i=k \parallel i=l \parallel j=l$ are selected.	C3
Only the matrix entries that satisfy the condition $i \neq j \neq k \neq l$ and $i=j=k \parallel i=j=l \parallel i=k=l \parallel j=k=l$ are selected	C2

While the theoretical algorithms adopted in the definition of MDs may seem to be borne out of deep creativity, the true beauty of novel MDs lies in the quality of the information codified and their ability to correlate with properties inherent to molecular structures. The following sections are dedicated to this analysis.

3 Shannon Entropy Based Variability Analysis

3.1 General information and purpose

One of the desirable requirements of novel MDs, as proposed by Randic [18], is that they should gradually change with gradual variations in structures. Shannon's entropy-based variability analysis offers a practical procedure for evaluating this attribute [15, 19]. This is an unsupervised feature selection method, based on the examination of case-wise statistical distribution of MD values in equal discrete intervals (bins). The use of histograms of descriptor distributions permits the comparison of descriptors with different units and value ranges. To this end, the variable range is determined (x_{max} - x_{min}) and divided by the number of desired intervals (maximum number of intervals is equal to number of cases) to determine the interval size.

Shannon's entropy formula (see equation 2.6) is applied to the resulting probability distribution function for each variable:

$$SE = \sum_{i=1}^{n} p_i \log_2 p_i \qquad (\mathbf{2.6})$$

where, p_i is the probability that a case adopts a value within data interval i (bin i).

Note that this entropic measure is not related to the chemical graph entropy [20], but is rather a measure of the variables' information content (i.e. relevance)[15, 21] in the sense that a variable that changes progressively with changes in chemical structures, as a desirable attribute of molecular descriptors [8], possesses high SE values, while low SE values correspond to redundant variables (with similar values for chemically different structures).

To compute the SE of the GDIs proposed in this manuscript, an interactive *in house* software denominated ***IMMAN*** (acronym for **I**nformation Theory based Che**M**o**M**etric **AN**alysis) was used [22]. The IMMAN software has been used by some of the present authors in an earlier study to examine the variability of MDs, and is freely available upon request to the authors [23].

The aim of the present study is to compare, in variability terms, the performance of families of known 2D MDs and the proposed GDIs. For this experiment, a small data set of 41 structurally diverse compounds was used (see Supporting Information SI2), and the descriptor calculations carried out using ***DIVATI***, a new module of ***TOMOCOMD-CARDD*** program. The following configurations were used for the GDI computation: 1) weighting scheme: Pauling's Electronegativity and the non-weighted option, 2) Dimensions: duplex,

triple and quadruple matrix dimensions, and 3) Aggregation Operators: Norms, Means and Statistical Invariants.

With the MDs obtained, the corresponding SE values were computed using a binning scheme of 41 intervals (bins), in which case the maximum entropy value, $SE_{max} = log_2 41 =$ 5.358 bits, corresponding to equiprobable discrete intervals ($p_i = 1/N$).

3.2 Comparison for duplex, triple and quadruple dimensional GDIs and other 2D-MDs

For this experiment, a comparison of SE values for the best overall 100 variables (in SE terms) of matrix representation (duplex, triple, quadruple) was performed. Better SE distributions are obtained for triple and quadruple approaches, although the distribution for the duplex matrix approach equally shows satisfactory results. Uniting all approaches based on *n*-dimentional matrix representations as a unique family, the quality's distribution increases, (see figure 3) showing the major variability from all the compared families.

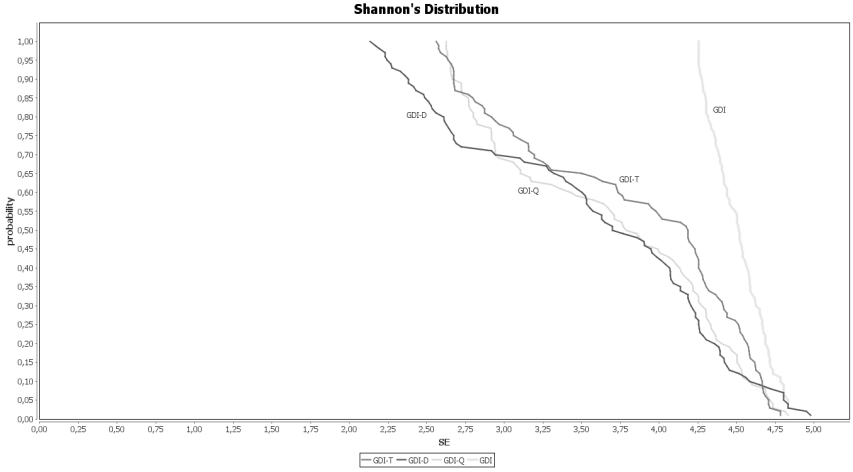

Fig. 3. Shannon's entropy distribution for GDI families.

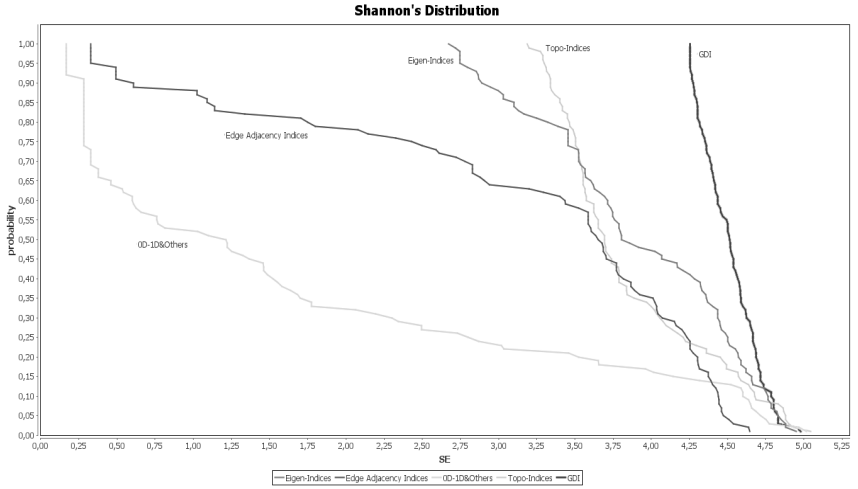

Fig. 4. Shannon's entropy distribution for DRAGON's and GDI descriptor families.

As it can be observed, it is perfectly justifiable to extend the derivative indices for triplex (GDI-T) and quadruple (GDI-Q) matrix representations, obtaining as result a qualitative and quantitative improvement, when all the matrix representation are combined as a unique indices family (GDIs family).

Besides, it was developed the comparison of GDIs as a unique family [12 % are Duplex relations frequency matrix-based (D), 33 % are Triple frequency matrix-based (T) and 55 % Quadruple relations frequency matrix-based (88% for *n*-tuples GDIs)] respect to the principal 2D indices used in relation structure-activity researches.

It can be observed in figure 4 that the GDIs express a major variability respect to the rest of the analyzed indices.

4 QSPR modeling of physico-chemical properties of 2-furylethylene derivatives

4.1 QSPR study

One of the key applications of theoretical MDs is in QSPR/QSAR studies. It is thus natural that the practical contribution of these parameters be evaluated in standard modeling paradigms to assess their true capacity in codifying useful chemical information. In view of the fact that the main aim of the present manuscript is to introduce extensions of the GDI method to higher dimensions, it is of interest to assess the contribution, if any, of

hypermatrix-based GDIs in QSPR modeling. In this experiment, a dataset of 34 2-furylethyelenes was used (see Supporting Information SI3), and a search for the best regressions for the properties Log P and Log K was performed. Subsequently, the performance of these indices was compared with the rest reported in the literature [24, 25].

The 2-furylethylene derivatives have different substituents in position 5 of the furan ring as well as in position β of the exo-cyclic double bond (see Supporting Information SI3). The Log P and Log K values of these chemicals have been experimentally determined and reported in the literature[26-28]. The lipophilicity and the nucleophilic addition of the mercaptoacetic acid to the exocyclic double bond of 2-furylethylene derivatives play a significant role in comprehending of the antibacterial activity of these chemicals. Therefore modeling such properties using the proposed GDIs provides a preliminary overview about the applicability of these indices in QSPR studies.

The search for optimal QSPR models was performed using Multiple Linear Regression method coupled with genetic algorithms (GAs) as the search strategy. This approach is implemented in the MOBYDIGS software (version 1.0 – 2004) [29]. The theoretical basis of the GAs has been explained in detail elsewhere [30-34]. The population size was set at 100 and the reproduction/mutation trade-off (T) at 0.70. The models were optimized using as objective function the statistical parameter Q^2_{loo} ("leave one out" cross-validation) and validated using the strategies "*bootstrapping*" (Q^2_{boot}) and "*scrambling*" (y-SC) [a (R^2), a (Q^2)]. The former checks the predictive power of the obtained models and the latter evaluates the possibility of random correlations, a usual phenomenon when dealing with high dimensional data [1, 29].

The best QSPR models attained using triple and quadruple matrix-based GDIs, and all GDIs-families respectively, are presented below:

Triple matrix-based GDIs models

$$\log K = 10.03(\pm 0.098) - 0.003(\pm 1.7 \cdot 10^{-4})\left[{}^{N/Pd}_{2}\boldsymbol{CV}(\boldsymbol{IS})\right]_S^T + 0.19(\pm 0.00)\left[{}^{A/Pl}_{2}\boldsymbol{N_1}(\boldsymbol{HT})\right]_S^T - 0.001(\pm 2 \cdot 10^{-5})\left[{}^{T/Pd}_{3}\boldsymbol{MX}(\boldsymbol{IS})\right]_S^{T/f} + 0.41(\pm 0.08)\left[{}^{C/In}_{3}\boldsymbol{Q}^{\boldsymbol{E}}_{\boldsymbol{3}}(\boldsymbol{Cb})\right]_S^T + 4 \cdot 10^{-5}(\pm 0.00)\left[{}^{T/Pl}_{7}\boldsymbol{Q_3}(\boldsymbol{HT})\right]_S^T - 5.65(\pm 0.09)\left[{}^{T/Pl}_{2}\boldsymbol{Q}^{\boldsymbol{E}}_{\boldsymbol{2}}(\boldsymbol{AH})\right]_S^{T2/f} + 0.001(\pm 1 \cdot 10^{-4})\left[{}^{Y/Pl}_{2}\boldsymbol{CV}^{\boldsymbol{E}}(\boldsymbol{Cb})\right]_S^{T2} \quad (\mathbf{4.1})$$

N = 34 R^2(%) = 99.72 s = 0.086 Q^2_{Loo}(%) = 99.53 s_{CV} = 0.097 Q^2_{Boot}(%) = 98.64 Y-SC = 0.143 F = 1302

$$\log P = -0.22(\pm 0.03) + 0.29(\pm 0.00)\left[{}^{T/In}_{2}\boldsymbol{K}(\boldsymbol{IS})\right]_{S}^{\frac{T}{f}} + 1 \cdot 10^{-5}(\pm 0.00)\left[{}^{Y/Pd}_{5}\boldsymbol{MN}(\boldsymbol{HT})\right]_{S}^{T} + 3 \cdot 10^{-5}(\pm 0.00)\left[{}^{V/Pd}_{1}\boldsymbol{N_2}\right]_{S}^{\frac{T2}{f}} - 0.99(\pm 0.04)\left[{}^{Sp}_{2}\boldsymbol{Q}^{E}_{2}(\boldsymbol{HL})\right]_{S}^{\frac{T2}{f}} + 0.5(\pm 0.02)\left[{}^{A/Pl}_{6}\boldsymbol{P}^{E}_{3}(\boldsymbol{IS})\right]_{S}^{\frac{T2}{f}} - 1.1(\pm 0.00)\left[{}^{A/Pl}_{6}\boldsymbol{MN}^{E}(\boldsymbol{Cb})\right]_{S}^{T2} + 0.11(\pm 0.00)\left[{}^{A/Pl}_{8}\boldsymbol{K}(\boldsymbol{IS})\right]_{S}^{T2} \quad \textbf{(4.2)}$$

N = 34 R^2(%) = 99.46 s = 0.06 Q^2_{Loo}(%) = 99.1 s_{CV} = 0.067 Q^2_{Boot}(%) = 98.95 Y-SC = 0.153 F = 678.64

Quadruple matrix-based GDIs models

$$\log K = 9.22(\pm 0.22) + 0.27(\pm 0.00)\left[{}^{P/In}\boldsymbol{K}(\boldsymbol{IS})\right]_{S}^{\frac{C}{f}} - 1.50(\pm 0.03)\left[{}^{T/In}_{4}\boldsymbol{Q}^{E}_{2}(\boldsymbol{HT})\right]_{S}^{\frac{C}{f}} - 0.58(\pm 0.05)\left[{}^{L/Pl}_{7}\boldsymbol{Q}^{E}_{2}(\boldsymbol{HT})\right]_{S}^{\frac{CT}{f}} - 33.15(\pm 0.35)\left[{}^{T/Pl}_{3}\boldsymbol{Q}^{E}_{2}\right]_{S}^{\frac{C2}{f}} - 0.33(\pm 0.04)\left[{}^{C/Pl}_{4}\boldsymbol{K}(\boldsymbol{Cb})\right]_{S}^{C2} + 35.25(\pm 0.00)\left[{}^{T/Pl}_{3}\boldsymbol{Q}^{E}_{2}(\boldsymbol{IS})\right]_{S}^{C3} - 0.33(\pm 0.02)\left[{}^{P/In}_{5}\boldsymbol{I}^{E}_{50}(\boldsymbol{AH})\right]_{S}^{C3/f} \quad \textbf{(4.3)}$$

N = 34 R^2(%) = 99.77 s = 0.078 Q^2_{Loo}(%) = 99.60 s_{CV} = 0.090 Q^2_{Boot}(%) = 99.31 Y-SC = 0.151 F = 1621.43

$$\log P = -6.5(\pm 0.16) - 5113.63(\pm 159.42)\left[{}^{T/In}_{2}\boldsymbol{CV}^{E}(\boldsymbol{Cb})\right]_{S}^{C/f} - 3.3 \cdot 10^{-4}(\pm 10^{-5})\left[{}^{A/In}_{2}\boldsymbol{M}(\boldsymbol{HT})\right]_{S}^{\frac{C}{f}} + 0.01(\pm 0.003)\left[{}^{A/In}_{2}\boldsymbol{Q_3}(\boldsymbol{HT})\right]_{S}^{\frac{C2}{f}} - 2.14(\pm 0.20)\left[{}^{P/In}_{2}\boldsymbol{Q_3}(\boldsymbol{HT})\right]_{S}^{C2} + 2 \cdot 10^{-5}(\pm 0.00)\left[{}^{L/Pl}_{3}\boldsymbol{K}(\boldsymbol{HT})\right]_{S}^{C2} + 0.65(\pm 0.14)\left[{}^{T/Pl}_{3}\boldsymbol{V}(\boldsymbol{AH})\right]_{S}^{C2} + 0.09(\pm 0.00)\left[{}^{T/In}_{4}\boldsymbol{P}^{E}_{3}(\boldsymbol{AH})\right]_{S}^{C2/f} \quad \textbf{(4.4)}$$

N = 34 R^2(%) = 99.47 s = 0.059 Q^2_{Loo}(%) = 99.23 s_{CV} = 0.062 Q^2_{Boot}(%) = 98.84 Y-SC = 0.138 F = 690.65

Triplex and Quadruple Matrix-based GDI Combined Models

$$\log K = 17.98(\pm 0.33) + 0.26(\pm 0.00)\left[{}^{P/In}\boldsymbol{K}(\boldsymbol{IS})\right]_{S}^{\frac{C}{f}} - 3.05(\pm 0.05)\left[{}^{L/Pd}{}_{5}\boldsymbol{Q}_{3}^{E}(\boldsymbol{Cb})\right]_{S}^{\frac{C}{f}} - 8.16(\pm 0.24)\left[{}^{T/Pl}{}_{8}\boldsymbol{P}_{3}^{E}\right]_{S}^{\frac{C}{f}} - 1.88(\pm 0.24)\left[{}^{C/Pd}{}_{5}\boldsymbol{Q}_{1}^{E}(\boldsymbol{IS})\right]_{S}^{\frac{C2}{f}} + 1.74(\pm 0.10)\left[{}^{N/Pd}{}_{1}\boldsymbol{M}^{E}(\boldsymbol{IS})\right]_{S}^{\frac{CT}{f}} + 0.0015(\pm 0.00)\left[{}^{P/In}{}_{3}\boldsymbol{MN}(\boldsymbol{IS})\right]_{S}^{\frac{T2}{f}} - 1.2 \cdot 10^{-4}(\pm 2 \cdot 10^{-5})\left[{}^{L/Pl}{}_{6}\boldsymbol{Q}_{1}\right]_{S}^{\frac{T2}{f}} \quad \textbf{(4.5)}$$

N = 34 R^2(%) = 99.79 s = 0.074 Q^2_{Loo}(%) = 99.70 s_{CV} = 0.079 Q^2_{Boot}(%) = 99.54 Y-SC = 0.150 F = 1735.42

$$\log P = -4.89(\pm 0.31) - 1.26(\pm 0.07)\left[{}^{L/Pl}{}_{2}\boldsymbol{K}(\boldsymbol{HT})\right]_{S}^{\frac{C3}{f}} - 5197.8(\pm 120.51)\left[{}^{T/In}{}_{2}\boldsymbol{CV}^{E}(\boldsymbol{Cb})\right]_{S}^{\frac{C}{f}} - 2.7 \cdot 10^{-4}(\pm 10^{-5})\left[{}^{A/In}{}_{2}\boldsymbol{Q}_{3}(\boldsymbol{HT})\right]_{S}^{\frac{C2}{f}} - 0.18(\pm 0.05)\left[{}^{P/In}{}_{6}\boldsymbol{Q}_{3}{}^{E}(\boldsymbol{IS})\right]_{S}^{\frac{T}{f}} + 0.55(\pm 0.08)\left[{}^{E/In}{}_{2}\boldsymbol{Q}_{1}^{E}(\boldsymbol{IS})\right]_{S}^{\frac{T}{f}} + 10^{-5}(\pm 0.00)\left[{}^{A/In}{}_{7}\boldsymbol{M}\right]_{S}^{T2/f} - 0.87(\pm 0.15)\left[{}^{R/Pl}{}_{8}\boldsymbol{P}_{3}^{E}\right]_{S}^{T2} \quad \textbf{(4.6)}$$

N = 34 R^2(%) = 99.53 s = 0.056 Q^2_{Loo}(%) = 99.34 s_{CV} = 0.057 Q^2_{Boot}(%) = 99.19 Y-SC = 0.153 F = 781.09

Duplex, Triplex and Quadruple Matrix-based GDI Combined Models

$$\log k = 2.61(\pm 0.14) + 0.11(\pm 0.02)\left[{}^{L/In}{}_{1}\boldsymbol{Q}_{1}^{E}(\boldsymbol{AH})\right]_{S}^{\frac{D}{f}} - 0.13(\pm 0.05)\left[{}^{T/In}{}_{4}\boldsymbol{3TSKQ}_{1}^{E}(\boldsymbol{AH})\right]_{S}^{\frac{D}{f}} + 0.09(\pm 0.03)\left[{}^{L/In}{}_{4}\boldsymbol{TTSKQ}_{3}^{E}\right]_{S}^{\frac{D}{f}} - 1.14(\pm 0.02)\left[{}^{T/In}{}_{4}\boldsymbol{5ACM}^{E}(\boldsymbol{IS})\right]_{S}^{\frac{D}{f}} + 0.03(\pm 0.002)\left[{}^{T/In}{}_{4}\boldsymbol{DE}(\boldsymbol{IS})\right]_{S}^{\frac{D}{f}} + 0.22(\pm 0.02)\left[{}^{M/Pl}{}_{7}\boldsymbol{S}^{E}\right]_{S}^{\frac{CT}{f}} + 0.89(\pm 0.07)\left[{}^{N/In}{}_{6}\boldsymbol{N}_{2}^{E}(\boldsymbol{HT})\right]_{S}^{C2}$$

(4.7)

N = 34 R^2(%) = 99.89 s = 0.055 Q^2_{Loo}(%) = 99.85 s_{CV} = 0.055 Q^2_{Boot}(%) = 99.73 Y-SC = 0.15 F = 3171.65

$$log\ P = -6.44(\pm 0.14) - 0.36(\pm 0.01)\left[{}^{T/In}\boldsymbol{8GI}(\boldsymbol{Q}_{3})^{E}\right]_{S}^{\frac{D}{f}} - 0.02(\pm 0.004)\left[{}^{V/In}{}_{1}\boldsymbol{TGI}(\boldsymbol{M})^{E}\right]_{S}^{\frac{D}{f}} - 1.81(\pm 0.03)\left[{}^{L/Pl}{}_{1}\boldsymbol{S}(\boldsymbol{HT})^{E}\right]_{S}^{CT} + 2.68(\pm 0.10)\left[{}^{C/Pd}{}_{8}\boldsymbol{N}_{3}(\boldsymbol{AH})^{E}\right]_{S}^{C} - 9 \cdot 10^{-5}(\pm 10^{-5})\left[{}^{R/In}{}_{6}\boldsymbol{I}_{50}(\boldsymbol{AH})\right]_{S}^{\frac{T}{f}} + 8.6 \cdot 10^{-4}(\pm 1.8 \cdot 10^{-4})\left[{}^{P/In}{}_{3}\boldsymbol{MN}(\boldsymbol{IS})\right]_{S}^{\frac{T2}{f}} + 0.85(\pm 0.09)\left[{}^{A/Pl}{}_{6}\boldsymbol{P}_{3}(\boldsymbol{HT})^{E}\right]_{S}^{T2} \quad \textbf{(4.8)}$$

N = 34 R^2(%) = 99.67 s = 0.047 Q^2_{Loo}(%) = 99.44 s_{CV} = 0.053 Q^2_{Boot}(%) = 99.26 Y-SC = 0.152 F = 1108.65

where, N is the number of compounds, R^2 is the determination coefficient, S_{CV} is the standard deviation of the regression, Q^2_{loo} and Q^2_{boot} are the regression coefficients obtained from the cross-validation procedures LOO and bootstrapping, respectively; y-SC is the intercept value, obtained from the validation technique scrambling, and F is the Fisher ratio.

The main statistic parameters of equations **4.7** and **4.8** show a better performance once all the GDIs families are mixed. All the models since 3 until 7 variables with respectively statistic parameters are showed in Supporting Information SI4.

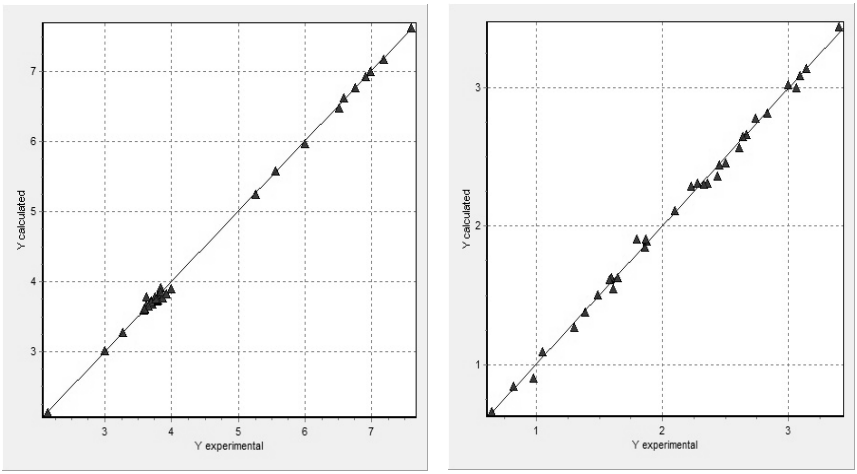

Fig. 5. The linear correlations between the calculated and experimental values of Log K and Log P for 2-furylethylenes. (Equations **4.7** and **4.8,** respectively).

The experimental and calculated values of Log K and Log P according to the models **4.7** and **4.8,** as well the corresponding residual values, are showed on supporting information SI5. The linear correlations existing between the calculated and experimental values of Log K and Log P for 2-furylethylenes derivatives are illustrated in Figure 5, corresponding to models **4.7** (Q^2_{loo} = 99.85) and **4.8** (Q^2_{loo} = 99.44), respectively. The statistical parameters of these models show satisfactory robustness and predictive capacity.

4.2 Comparative study

Finally, regressions of the physicochemical properties (Log P and Log K) for the 2-furylethylenes obtained using the novel GDIs are compared with those of some of the most relevant indices or group of indices in QSPR studies such as: connectivity indices (both 2D and 3D as well as edge- and vertex-based), total (global) spectral moment (sum of the trace of the bond matrix), local (fragment) spectral moment (partial sum of the trace of the bond matrix), linear indices (bond-based stochastic and non-stochastic), atom- and bond-based quadratic indices, 2/3D ITs and quantum chemical descriptors (See Table 4) [28].

As can be seen in Table 4, triple and quadruple matrix-based GDIs show superior performance, in modeling the considered properties than the rest of the MD families reported in the literature. When all the Graph Derivative Indices (GDIs) families are combined, models with better performance are obtained. This result reaffirms the theoretical and practical contribution of extending the GDI method to higher dimensions.

Despite the models reported for each GDIs family and group of families have until 7 independent variables, it can be noted that in almost all cases with only 4 descriptors there are founded regression equations more statistically robust that equations reported in the specialized literature with 7 parameters.

Unfortunately, authors of previous studies did not report the values of Q^2_{boot}, y-SC and in some cases for Log K and log P, or Q^2_{loo} values [24, 25].

It can therefore be suggested that the **DIVATI** indices, in general, seem to be a promissory tool to reckon on in QSAR/QSPR studies, virtual screening of compound datasets as well as in similarity/dissimilarity evaluations.

5 Concluding Remarks

A generalization of the previously proposed GDIs to higher dimensions, considering triple and quadruple atom relations is discussed. The frequencies of these atom relations are condensed in triple and quadruple relations frequency hypermatrices, respectively. A Shannon's entropy-based variability analysis reveals that the extension of the GDI method to

Table 4: Statistical Parameters of QSPR Models that Describe Physicochemical Properties of 34 Derivatives of 2-furyletilenes by using Different MDs.

Indices	N	R^2	*s*	Q^2_{loo}	s_{CV}	Q^2_{boot}	y-sc	F
1-octanol/water partition coefficient (Log P)								
GDI-D^{13}	7	99.19	0.072	98.72	0.080	98.34	*	453.84
GDI-D^{13}	6	98.81	0.086	98.16	0.096	97.68	*	374.08
GDI-D^{13}	5	98.44	0.097	97.81	0.105	97.41	*	352.67
GDI-D^{13}	4	97.10	0.13	96.19	0.138	95.79	*	242.48
GDI-T (Eq. 4.2)	7	99.46	0.06	99.1	0.067	98.95	0.153	678.64
GDI-T	6	99.11	0.075	98.6	0.083	98.42	0.116	503.94
GDI-T	5	98.28	0.102	97.46	0.113	97.25	0.077	319.33

GDI-T	4	97.07	0.131	95.81	0.145	95.66	0.065	240.61
GDI-T	3	95.42	0.161	94.12	0.151	93.89	0.028	208.4
GDI-Q (Eq. 4.4)	7	99.47	0.059	99.23	0.062	98.84	0.138	690.65
GDI-Q	6	99.31	0.066	99.09	0.067	98.3	0.04	651.05
GDI-Q	5	99.0	0.078	98.49	0.087	97.67	0.087	554.94
GDI-Q	4	98.26	0.101	97.59	0.11	97.35	0.055	408.67
GDI-Q	3	97.18	0.126	96.27	0.136	96.18	0.022	344.91
GDI-*[T & Q]*(Eq. 4.6)	7	99.53	0.056	99.34	0.057	99.19	0.153	781.09
***GDI-*[T & Q]**	6	99.36	0.063	99.09	0.067	98.88	0.125	697.42
***GDI-*[T & Q]**	5	99.05	0.076	98.62	0.083	98.44	0.089	585.47
***GDI-*[T & Q]**	4	98.45	0.095	97.84	0.104	97.7	0.043	460.95
***GDI-*[T & Q]**	3	97.19	0.126	96.36	0.135	96.22	0.044	345.29
GDI-*[D, T & Q]*(Eq. 4.8)	7	99.67	0.047	99.44	0.053	99.26	0.152	1108.65
***GDI-*[D, T & Q]**	6	99.47	0.058	99.19	0.064	99.08	0.113	840.73
***GDI-*[D, T & Q]**	5	99.11	0.073	98.72	0.08	98.55	0.108	624.97
***GDI-*[D, T & Q]**	4	98.48	0.094	97.69	0.107	97.66	0.065	469.64
***GDI-*[D, T & Q]**	3	96.39	0.143	95.56	0.149	95.38	0.04	266.69
Bond-based NS LI	7	97.5	0.127	0.951	*	*	*	146.80
Vertex and edge Conn. Indices	7	93.9	0.199	*	*	*	*	56.9
Topological Descriptors	7	96.4	0.155	*	*	*	*	84.6
Quantum Chemical Descriptors	**	87.5	0.319	*	*	*	*	45.5
Atom-based NS QI	7	96.9	0.142	95.1	*	*	*	116.76
Atom-based NS LI	7	96.8	0.143	93.8	*	*	*	113.38
Specific rate constant (Log K)								
GDI-D^{I3}	7	99.81	0.069	99.7	0.111	98.47	*	2003.08
GDI-D^{I3}	6	98.86	0.169	98.26	0.186	97.73	*	389.39
GDI-D^{I3}	5	98.47	0.192	97.79	0.209	97.21	*	359.88
GDI-D^{I3}	4	97.92	0.22	97.29	0.231	96.74	*	341.08
GDI-T (Eq. 4.1)	7	99.72	0.086	99.53	0.097	98.64	0.143	1302
GDI-T	6	99.52	0.11	99.30	0.118	98.17	0.126	932.46
GDI-T	5	98.91	0.163	98.43	0.178	96.51	0.092	508.38
GDI-T	4	97.64	0.236	96.78	0.254	95.71	0.076	299.67
GDI-T	3	95.20	0.331	93.51	0.361	91.73	0.034	198.17
GDI-Q (Eq. 4.3)	7	99.77	0.078	99.60	0.090	99.31	0.150	1621.43
GDI-Q	6	99.55	0.107	99.17	0.130	98.72	0.127	990.43
GDI-Q	5	99.34	0.127	99.06	0.137	98.37	0.097	842.10
GDI-Q	4	99.09	0.146	98.71	0.161	98.02	0.072	789.91
GDI-Q	3	98.16	0.205	97.19	0.238	96.86	0.032	533.30
GDI-*[T & Q]*(Eq. 4.5)	7	99.79	0.074	99.70	0.079	99.54	0.150	1735.42
***GDI-*[T & Q]**	6	99.75	0.080	99.66	0.084	99.49	0.146	1732.47
***GDI-*[T & Q]**	5	99.59	0.101	99.42	0.109	99.22	0.162	1321.14
***GDI-*[T & Q]**	4	99.15	0.143	98.70	0.163	98.28	0.064	813.22
***GDI-*[T & Q]**	3	98.39	0.194	97.67	0.218	96.92	0.029	589.95
GDI-*[D, T & Q]*(Eq.4.7)	7	99.89	0.055	99.85	0.055	99.73	0.150	3171.65
***GDI-*[D, T & Q]**	6	99.85	0.062	99.81	0.062	99.74	0.118	2942.74
***GDI-*[D, T & Q]**	5	99.82	0.066	99.76	0.070	99.70	0.085	3052.76
***GDI-*[D, T & Q]**	4	99.71	0.084	99.60	0.090	99.05	0.070	2368.70
***GDI-*[D, T & Q]**	3	98.98	0.152	98.53	0.173	98.43	0.047	938.02
Connectivity Indices	7	82.1	0.681	*	*	*	*	17.1
Global spectral moments	7	84.3	0.655	*	*	*	*	18.8
Local spectral moments	7	96.4	0.320	*	*	*	*	70.4
Quantum Chemical Descriptor	7	96.8	0.288	*	*	*	*	112.2
Atom-based NS QI	7	96.8	0.922	28.5	*	*	*	115.14
Bond-based NS QI	7	96.7	0.940	29.2	*	*	*	108.79
Bond-based SS QI	7	97.5	0.958	25.7	*	*	*	142.07

GDI-X [X: D(Duplex Matrix-based), T(Triplex Matrix-based), Q(Quadruple Matrix-based), T&Q (Triple and Quadruple Matrix-based), D,T&Q (Duplex, Triplex and Quadruple Matrix-based)]

*Not reported

**Used Rogers and Cammarata approach.

higher dimensions increases the entropy of the proposed MDs, superior to that of the t commonly used MDs in chemoinformatics tasks. Moreover, the GDIs are employed in the modeling of the physicochemical properties Log P and Log K of the 2-furylethylenes derivatives, obtaining robust models whose respective statistical parameters are comparable to superior to those reported in the literature.

The generalization scheme discussed in the present report could be applied in the extension of any family of MDs defined on the basis of atom-pair relations.

The DIVATI program is free multiplatform software, built following a Master/Worker pattern to utilize multiple CPU cores. The software is available upon request to the authors (ymarrero77@yahoo.es).

Supplementary Data Available: The Quadruple Hypermatrix for the N,N-dimethylpropionamide molecule (SI1), the data set of 41 diverse molecular structures for the variability study (SI2), chemical structures and numbering of atoms in the furylethylenes compounds used QSPR study (SI3), models since 3 until 7 variables (SI4) and experimental and calculated values of Log K and Log P (SI5) are available free of charge via the Internet.

Acknowledgement: Marrero-Ponce, Y. thanks to the program '*International Professor'* for a fellowship to work at Cartagena University in 2013-2014.

References

[1] A. T. Balaban, O. Ivanciuc, J. Devillers, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon & Breach, Amsterdam, 1999, pp. 21–57.

[2] I. Gutman, O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer, Berlin, 1987.

[3] N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, 1992.

[4] E. Estrada, E. Uriarte, Recent advances on the role of topological indices in drug discovery research, *Curr. Med. Chem.* **8** (2001) 1573–1588.

[5] A. R. Katritzky, E. V. Gordeeva, Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research, *J. Chem. Inf. Comput. Sci.* **33** (1993) 835–857.

[6] L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.

[7] L. B. Kier, L. H. Hall, *Molecular Structure Description. The Electrotopological State*, Academic Press, San Diego, 1999.

[8] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.

[9] J. L. Mokrosz, B. Duszynska, L. Strekowski, Topological indices in correlation analysis, Part 3: The modelling of hydrophobic properties using molecular connectivity and shape indices, *Pharmazie* **47** (1992) 538–541.

[10] A. T. Balaban, Topological and stereochemical molecular descriptors for databases useful in QSAR, similarity/dissimilarity and drug design, *SAR QSAR Environ. Res.* **8** (1998) 1–21.

[11] J. Devillers, New trends in (Q)SAR modeling with topological indices, *Curr. Op. Drug Discov. Dev.* **3** (2000) 275–279.

[12] Y. Marrero–Ponce, O. Martínez Santiago, Y. Martínez López, S. J. Barigye, F. Torrens, Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors I. Theory and QSPR application, *J Comput. Aided Mol. Des.* **26** (2012) 1229–1246.

[13] O. Martínez–Santiago, R. Millán–Cabrera, Y. Marrero–Ponce, S. J. Barigye, Y. Martínez–López, F. Torrens, F. Pérez–Giménez, Discrete derivatives for atom-pairs as a novel graph–theoretical invariant for generating new molecular descriptors: Orthogonality, interpretation and QSARs/QSPRs on benchmark databases, *Mol. Inform*. **33** (2014) 343–368.

[14] V. A. Gorbátov, *Fundamentos de la Matematica Discreta*, Mir, Moscú, 1988.

[15] J. W. Godden, F. L. Stahura, J. Bajorath, Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations, *J. Chem. Inf. Comput. Sci.* **40** (2000) 796–800.

[16] S. J. Barigye, Y. Marrero–Ponce, O. Martínez Santiago, Y. Martínez López, F. Torrens, Shannon's, mutual, conditional and joint entropy-based information indices. Generalization of global indices defined from local vertex invariants, *Curr. Comput. Aided Drug Des.* **9** (2013) 164–183.

[17] Y. Marrero–Ponce, **TOMOCOMD** (TOpological MOlecular COMputational Design) for Windows Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), 2010.

[18] M. Randić, Molecular bonding profiles, *J. Math. Chem.* **19** (1996) 375–392.

[19] J. W. Godden, J. Bajorath, Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis, *J. Chem. Inf. Comput. Sci.* **42** (2002) 87–93.

[20] S. J. Barigye, Y. Marrero–Ponce, F. Pérez–Giménez, D. Bonchev, Trends in information theory based chemical structure codification, *Mol. Divers.* **18** (2014) 673-686.

[21] H. Hong, Q. Xie, Q. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, T. W., Mold2, Molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics, *J. Chem. Inf. Model.* **48** (2008) 1337–1344.

[22] S. J. Barigye, R. W. Pino Urias, Y. Marrero–Ponce, IMMAN (Information Theory based Chemometric Analysis) 2011.

[23] S. J. Barigye, Y. Marrero–Ponce, Y. Martínez–Lopez, F. Torrens, L. M. Artiles–Martínez, R. W. Pino–Urias, O. Martínez–Santiago, Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices, *J. Comput. Chem.* **34** (2012) 259–274.

[24] E. Estrada, E. Molina, Novel local (fragment-based) topological molecular descriptors for QSPR/QSAR and molecular design, *J. Mol. Graphics Model*. **20** (2001) 54–64.

[25] S. E. Wold, Statistical validation of QSAR results. Validation tools, in: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995.

[26] S. S. Balaz, M. Rosenberg, J. Augustin, B. Skara, Kinetics of drug activities as influenced by their physicochemical properties: antibacterial effects of alkylating 2-furylethylenes, *J. Comput. Aided Mol. Des.* **131** (1988) 115–134.

[27] Y. Marrero Ponce, E. R. Martinez–Albelo, G. M. Casanola–Martin, J. A. Castillo Garit, Y. Echeveria Diaz, V. Romero Zaldivar, J. Tygat, J. E. Rodriguez Borges, R. García–Domenech, F. Torrens, F. Pérez–Giménez, Bond–based linear indices of the non-stochasitic and stochastic edge-adjacency matrix. 1. Theory and modeling of ChemPhys properties of organic molecules, *Mol. Divers.* **14** (2010) 731–753.

[28] E. Estrada, E. Molina, 3D conectivity indices in QSPR/QSAR studies, *J. Chem. Inf. Comput. Sci.* **41** (2001) 791–797.

[29] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, MOBYDIGS version 1.0, Milano, 2005.

[30] D. E. Goldberg, *Genetic Algorithms*. Addison-Wesley, Reading, 1989.

[31] D. H. Rogers, Application of genetic function approximation to quantitative structure–property relationships, *J. Chem. Inf. Comput. Sci.* **34** (1994) 854–866.

[32] P. Willet, Genetic algorithms in molecular recognition and design, *Trends Biotechnol.* **13** (1995) 516–521.

[33] S. S. So, M. Karplus, Evolutionary optimization in quantitative structure–activity relationship: An application of genetic neural networks, *J. Med. Chem.* **39** (1996) 1521–1530.

[34] S. S. So, M. Karplus, Three–dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations, *J. Med. Chem.* **40** (1997) 4347–4359.