Relations Frequency Hypermatrices in Mutual, Conditional and Joint Entropy-Based Information Indices†

Stephen J. Barigye,[a] Yovani Marrero-Ponce,*[a,b,c] Yoan Martínez-López,[a,d] Francisco Torrens,[b] Luis Manuel Artiles-Martínez,[a] Ricardo W. Pino-Urias,[a] and Oscar Martínez-Santiago[a,e]

Graph-theoretic matrix representations constitute the most popular and significant source of topological molecular descriptors (MDs). Recently, we have introduced a novel matrix representation, named the duplex relations frequency matrix, **F**, derived from the generalization of an incidence matrix whose row entries are connected subgraphs of a given molecular graph **G**. Using this matrix, a series of information indices (IFIs) were proposed. In this report, an extension of **F** is presented, introducing for the first time the concept of a hypermatrix in graph-theoretic chemistry. The hypermatrix representation explores the *n*-tuple participation frequencies of vertices in a set of connected subgraphs of **G**. In this study we, however, focus on triple and quadruple participation frequencies, generating triple and quadruple relations frequency matrices, respectively. The introduction of hypermatrices allows us to redefine the recently proposed MDs, that is, the mutual, conditional, and joint entropy-based IFIs, in a generalized way. These IFIs are implemented in GT-STAF (acronym for Graph Theoretical Thermodynamic STAte Functions), a new module of the TOMOCOMD-CARDD program. Information theoretic-based variability analysis of the proposed IFIs suggests that the use of hypermatrices enhances the entropy and, hence, the variability of the previously proposed IFIs, especially the conditional and mutual entropy based IFIs. The predictive capacity of the proposed IFIs was evaluated by the analysis of the regression models, obtained for physico-chemical properties the partition coefficient (Log *P*) and the specific rate constant (Log *K*) of 34 derivatives of 2-furylethylene. The statistical parameters, for the best models obtained for these properties, were compared to those reported in the literature depicting better performance. This result suggests that the use of the hypermatrix-based approach, in the redefinition of the previously proposed IFIs, avails yet other valuable tools beneficial in QSPR studies and diversity analysis. © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23123

Introduction

The role of theoretical molecular descriptors (MDs) in QSPR/QSAR studies, similarity and diversity analysis, database mining and the virtual screening of combinatorial libraries cannot be denied. Among them, Topological Indices (TIs), regardless of their simplicity, are probably the most relevant ones because of their high correlational ability with a number of physical, chemical, physico-chemical, and biological properties.[1] The TIs are parameters mathematically derived from graph-theoretic invariants, which encode structural information contained in the bonding topology of the molecule, disregarding information on structural features like: bond lengths, bond angles, and torsion angles.[2–4] The TIs are divided into two categories: topo-structural and topo-chemical indices. Topo-structural indices are concerned with the adjacency and distance between

[a] *S. J. Barigye, Y. Marrero-Ponce, Y. Martínez-López, L. M. Artiles-Martínez, R. W. Pino-Urias, O. Martínez-Santiago*
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
E-mail: ymarrero77@yahoo.es (or) ymponce@gmail.com (or) yovanimp@uclv.edu.cu
Fax: +53 963543156

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

[c] *Y. Marrero-Ponce*
Departamento de Química Física, Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Facultad de Farmacia, Universitat de València, Spain

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

[e] *O. Martínez-Santiago*
Departament of chemical science, Faculty of Chemistry-Pharmacy. Universidad Central "Martha Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba

Contract/grant sponsor: Spanish Ministry of Science and Innovation (MICINN); Contract/grant numbers: SAF2009-10399; Contract/grant sponsor: VLIR (Vlaamse InterUniversitaire Raad, Flemish Interuniversity Council, Belgium; under the IUC Program VLIR-UCLV; Contract/grant sponsor: Spanish Ministerio de Ciencia e Innovacion; Contract/grant number: BFU2010-19118.

†Dedicated to the Research Group of Molecular Connectivity & Drug Design (University of Valencia) for its many seminal contributions to the definition of new topological indices and computational virtual screening, as well as their methodology for drug repositing.

© 2012 Wiley Periodicals, Inc.

vertices in the molecular graph (**G**); topo-chemical indices, in addition to measuring the information about topology, also quantify particular atomic properties such as chemical identity. This classification is, however, not entirely discriminative as there exists a subclass of TIs that could be both topo-structural and topo-chemical as well: the information indices (IFIs).[4] These indices will be the primary focus of the present report.

The IFIs form a subclass of TIs based on the application of concepts of information theory to a graph **G**. Since 1948, when Claude Elwood Shannon coined *The Mathematical Theory of Communication* (known nowadays as *Information Theory*),[5] the use of the information-theoretic approaches has been extended to a remarkably wide variety of fields as neuroscience, linguistics, sociology, taxonomy, psychology, and artificial intelligence among others.[6–10] The attempt to import information theory tools to these fields is not by sheer coincidence. This is given by the fact that information theory is accompanied by profound mathematical properties applicable to the core of various disciplines. Moreover, the birth of information theory paves way for a new perspective towards structures of any kind: as communication systems, which carry a certain quantity of information. Therefore, a graph **G** could also be viewed as a communication system and its information content subsequently measured or determined.

In an earlier publication, we proposed a series of information theory-based indices supported by the concepts of Shannon's, mutual, conditional, and joint entropies.[11] Additionally, a new matrix representation denominated frequency relations matrix was presented. This matrix representation arises from the exploration of duplex participation frequencies of connected subgraphs in the formation of a **G**. In addition, we introduced a set of invariants generalizing the definition of global or local invariants from atomic contributions (local vertex invariants, LOVIs).[11]

In the present report, we intend to introduce for the first time, the concept of a hypermatrix, conceived from the evaluation of the *n*-tuple ($n > 2$) participation frequencies of connected subgraphs 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 mutual, conditional and joint entropy-based IFIs in a "more generalized" way.

To evaluate the contribution of the hypermatrix approach in terms of the variability of the IFIs derived thereof, we compare the information content encoded by duplex, triple, and quadruple matrix-based IFIs using a Shannon's entropy-based approach proposed by Godden and Bajorath.[12,13] For this analysis, we use DRAGON's sample data consisting of 40 heterogeneous molecules. To assess the performance of the proposed MDs in modeling tasks, the 1-octanol/water partition coefficient (Log *P*) and 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 physico-chemical properties using the proposed IFIs are compared with those of other approaches [two-dimensional (2D)/3D, edge- and vertices-based connectivity indices, total and local spectral moments, topological and quantum chemical descriptors as well as some of TOMOCOMD-CARDD MDs, namely, bond-based nonstochastic (and stochastic) linear indices plus bond- and atom-based quadratic indices] reported in the literature.

Theoretical Framework

Frequency hypermatrix representation of a molecular graph

In an earlier publication,[11] we defined IFIs using the duplex participation frequency approach, which originates from the exploration of the participation of letters in a collection of words, called a dictionary or model. To present the definitions and notations in this report in a comprehensive way, we will go over those introduced in the previous publication, giving a brief summary of the important aspects presented.

To begin with, we define an event **E**, which we consider true when certain conditions of an examined procedure are fulfilled. The event **E** determines a bidimensional binary matrix $Q = [q_{ij}]_{m\times n}$, each column of which corresponds reciprocally to a condition included in at least one true event, and every row, to a collection of conditions in which the event occurs (the event **E** is true) and q_{ij} is equal to[14]:

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

In other words, every 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. We introduce the relation frequency matrix $\mathbf{F} = [f_{ij}]_{n\times n}$, which characterizes the model ψ with the incidence matrix $\boldsymbol{Q}(\psi) = [q_{ij}]_{m\times n}$[14]

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

From the definition of **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 an "incidence" matrix [$\boldsymbol{Q}$ (ψ)] for the model ψ.[14]

One could also arrive at this relation frequency matrix using a simple exploratory method. Let us consider a model where we include nine words of the English language in which no letter is repeated: TR**A**V**E**L OR**A**NG**E** **A**MB**E**R HOUSE MOUNTAIN OC**EA**N STR**EA**M M**E**RCH**A**NT P**EA**RL B**EA**CH DR**EA**M.

Our interest, in this case, is to find the number of times (frequency) that a subset of two letters participate in the formation of the same word (duplex participation frequency). If we look at letters {A, E} for example, these simultaneously

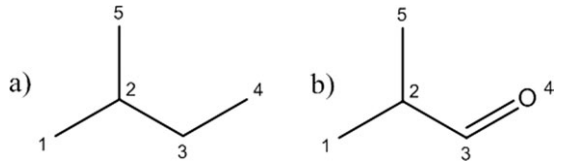


Figure 1. The chemical structure of the molecules of (the numbers correspond to the labels that are assigned to the non-hydrogen atoms (vertices) in the molecular structure): (a) isopentane, and b) 2-methylpropanal.

participate in the formation of the words: TRAVEL, ORANGE, AMBER, OCEAN, STREAM, MERCHANT, PEARL, BEACH, and DREAM, that is, they participate nine times in the formation of the same word, $f_{AE} = 9$. The participation frequencies of all possible two-component subsets of letters (f_{ij}) could be similarly explored, as well as the participation frequencies of each of the letters (f_i) that constitute these words. These frequencies are the components for the relation frequency matrix, **F**.[11]

We may also want to compute the number of times that a subset of three or four letters participates together in the formation of a word in the model, which constitutes the primary focus of the present article. Let us take as an example the participation of the letters {A, E, R} and {A, E, M, R} in the formation of the same word, respectively (see illustration below).

T**RAV**E**L** O**RA**NG**E** **A**MB**ER** HOUSE MOUNTAIN OCEAN ST**REA**M M**ER**CH**A**NT P**EAR**L BEACH D**REA**M
TRAVEL ORANGE **AM**B**ER** HOUSE MOUNTAIN ST**REAM MER**-CH**A**NT PEARL BEACH D**REAM**

It can be clearly seen that the letters {A, E, R} simultaneously participate in the formation of the words TRAVEL, ORANGE, AMBER, STREAM, MERCHANT, PEARL, and DREAM, that is, they participate seven times in the formation of the same word, $f_{AER} = 7$. If one inspects the words enlisted above, the letters A, E, M, and R participate concurrently in the formation of the words AMBER, STREAM, MERCHANT, and DREAM, that is, the frequency participation of these letters is 4, thus $f_{AEMR} = 4$. This analysis could be furthered to explore the number of times that subsets of 5, 6, 7, …, *n* letters participate in the same word. In this report, we will, however, stick to triple and quadruple participation frequencies.

These participation frequencies permit us generating 3D and 4D matrices, which we will denominate, triple, and quadruple hypermatrices, respectively. In the triple matrix, simultaneous three-letter participation frequencies are analyzed, while in the quadruple matrix simultaneous four-letter participation frequencies are examined. In all these matrices, the axis-labels are conditions denominated by letters. The algorithm used in this exploratory procedure taking is available as Supporting Information (SI1).

Given the difficulty in viewing all the elements in 3D and 4D matrices, these matrices will be separated into *n* sheets or slides (layers), where *n* is the total number of elements present in the universal set of letters that constitute the words.

Having introduced the ideas above, we could safely proceed to graph-theoretic chemistry and define logical analogies that could be of great use in the present definitions. Accordingly, the event (**E**) is the participation of connected subgraphs in the formation of a given molecular graph, **G**; the conditions (letters of the model) included in the event are the vertices (atomic nuclei) present in each collection of conditions (connected subgraphs, which are the words of the model). The graph-theoretic concepts of subgraphs orders and types, namely: path (p), cluster (c), and path-cluster (pc) are used as criteria to generate the connected subgraphs.

Consider as an illustrative example the molecular graph of isopentane (Fig. 1a), describing the carbon skeleton of this molecule.

First, we obtain connected subgraphs (words) of different orders from **G** based on the atomic relations.

Order 0: C_1, C_2, C_3, C_4, C_5
Order 1 (four paths): C_1–C_2, C_2–C_3, C_3–C_4, C_2–C_5
Order 2 (four 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 (two paths and one path-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 (one path-cluster): C_1–C_2–C_3–C_4–C_5

These connected subgraphs comprise the set of words, while the vertices for **G**, {C_1, C_2, C_3, C_4, C_5} are the letters. From the set of vertices (letters), all possible three- or four-element subsets of vertices, corresponding to triple and quadruple matrices, respectively, are formed. The participation frequencies of these subsets of vertices in the formation of the connected subgraphs are then computed. These frequencies are subsequently used to construct the hypermatrices. As an example, the set of vertices {C_1, C_2, C_4} concurrently participate in the formation of connected subgraphs: C_1–C_2–C_3–C_4 and C_1–C_2–C_3–C_4–C_5. Thus, the participation frequency of the set of vertices {C_1, C_2, C_4} is two [see entry (1, 2, 4)] in the triple matrix represented below. It should be noted that the matrix entries are commutative. Figure 2 shows an illustrative geometric representation of the triple matrix generated for the **G** of isopentane.

Below is the slide representation adapted for triple and quadruple matrices, respectively, for the **G** of isopentane.

Triple Matrix

$$1\begin{bmatrix}7&6&4&2&3\\6&6&4&2&3\\4&4&4&2&2\\2&2&2&2&1\\3&3&2&1&3\end{bmatrix}\quad 2\begin{bmatrix}6&6&4&2&3\\6&12&8&4&6\\4&8&8&4&4\\2&4&4&4&2\\3&6&4&2&6\end{bmatrix}\quad 3\begin{bmatrix}4&4&4&2&2\\4&8&8&4&4\\4&8&10&5&4\\2&4&5&5&2\\2&4&4&2&4\end{bmatrix}\quad 4\begin{bmatrix}2&2&2&2&1\\2&4&4&4&2\\2&4&5&5&2\\2&4&5&5&2\\1&2&2&2&2\end{bmatrix}\quad 5\begin{bmatrix}3&3&2&1&3\\3&6&4&2&6\\2&4&4&2&4\\1&2&2&2&1\\3&6&4&2&7\end{bmatrix}$$

Quadruple Matrix

1

1 1					1 2					1 3					1 4					1 5				
7	6	4	2	3	6	6	4	2	3	4	4	4	2	2	2	2	2	2	1	3	3	2	1	3
6	6	4	2	3	6	6	4	2	3	4	4	4	2	2	2	2	2	2	1	3	3	2	1	3
4	4	4	2	2	4	4	4	2	2	4	4	4	2	2	2	2	2	2	1	2	2	2	1	2
2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	1	1	1	1	1
3	3	2	1	3	3	3	2	1	3	2	2	2	1	2	1	1	1	1	1	3	3	2	1	3

2

2 1					2 2					2 3					2 4					2 5				
6	6	4	2	3	6	6	4	2	3	4	4	4	2	2	2	2	2	2	1	3	3	2	1	3
6	6	4	2	3	6	12	8	4	6	4	8	8	4	4	2	4	4	4	2	3	6	4	2	6
4	4	4	2	2	4	8	8	4	4	4	8	8	4	4	2	4	4	4	2	2	4	4	2	4
2	2	2	2	1	2	4	4	4	2	2	4	4	4	2	2	4	4	4	2	1	2	2	2	2
3	3	2	1	3	3	6	4	2	6	2	4	4	2	4	1	2	2	2	2	3	6	4	2	6

3

3 1					3 2					3 3					3 4					3 5				
4	4	4	2	2	4	4	4	2	2	4	4	4	2	2	2	2	2	2	1	2	2	2	1	2
4	4	4	2	2	4	8	8	4	4	4	8	8	4	4	2	4	4	4	2	2	4	4	2	4
4	4	4	2	2	4	8	8	4	4	4	8	10	5	4	2	4	5	5	2	2	4	4	2	4
2	2	2	2	1	2	4	4	4	2	2	4	5	5	2	2	4	5	5	2	1	2	2	2	2
2	2	2	1	2	2	4	4	2	4	2	4	4	2	4	1	2	2	2	2	2	4	4	2	4

4

4 1					4 2					4 3					4 4					4 5				
2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	1	1	1	1	1
2	2	2	2	1	2	4	4	4	2	2	4	4	4	2	2	4	4	4	2	1	2	2	2	2
2	2	2	2	1	2	4	4	2	2	2	4	5	5	2	2	4	5	5	2	1	2	2	2	2
2	2	2	2	1	2	4	4	4	2	2	4	5	5	2	2	4	5	6	2	1	2	2	2	2
1	1	1	1	1	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2

5

5 1					5 2					5 3					5 4					5 5				
3	3	2	1	3	3	3	2	1	3	2	2	2	1	2	1	1	1	1	1	3	3	2	1	3
3	3	2	1	3	3	6	4	2	6	2	4	4	2	4	1	2	2	2	2	3	6	4	2	6
2	2	2	1	2	2	4	4	2	4	2	4	4	2	4	1	2	2	2	2	2	4	4	2	4
1	1	1	1	1	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2	1	2	2	2	2
3	3	2	1	3	3	6	4	2	6	2	4	4	2	4	1	2	2	2	2	3	6	4	2	7

Information theory-based indices

The formalism used in the present report to define the information theory-based indices follows the subsequent steps:

Step 1. Construction of the frequency relations matrices for a given event (see above).

Step 2. Computation of the probability matrices whose elements p_{ijkl} (or p_{ijk}) for quadruple (or triple) matrices, respectively, are obtained as follows:

$$p_{ijkl}(\text{or } p_{ijk}) = \frac{f_{ijkl}(\text{or } f_{ijk})}{n(S)}$$

where, p_{ijkl} (or p_{ijk}) denotes the probability that vertices *i*, *j*, *k*, and *l* (or *i*, *j*, and *k*) simultaneously participate in the formation of the connected subgraphs, f_{ijkl} (or f_{ijkl}) the participation frequency of vertices *i*, *j*, *k*, and *j* (or *i*, *j*, and *k*) and *n* (*S*) the sample space. Note that the sample space is equal to the subgraph number (for proof, see Supporting Information, SI2).

Step 3. Calculation of the Mutual, Conditional and Joint entropies.

Mutual Entropy. The term mutual entropy (or information) refers to the mean information that two (or more) sources (or elements) have in common. Given four sources *i*, *j*, *k*, and *l* (quadruple matrix consideration), mutual entropy is defined as:

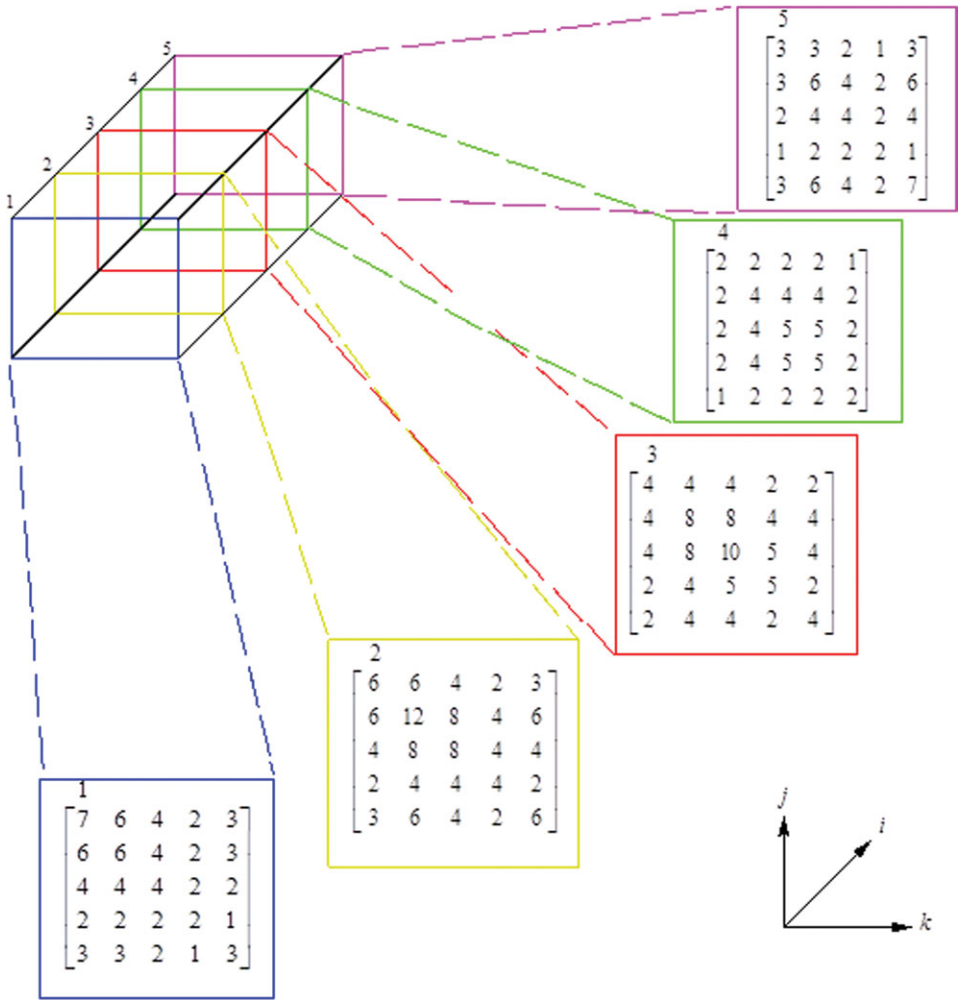


Figure 2. Illustrative geometric representation of the triple matrix generated for the **G** of isopentane. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

$$\mathrm{I}(i;j;k;l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} p(i,j,k,l)\log_2\frac{p(i,j,k,l)}{p(i)p(j)p(k)p(l)} \quad (1)$$

Mutual entropy could also be viewed as the measure of the distance between the joint probability $p(i, j, k, l)$ and the probability $p(i)p(j)p(k)p(l)$, which is the joint probability under the assumption of independence.

When at least two sources are identical (i.e., $k = l$), equation **1** reduces to a three-argument equation, used in a triple matrix:

$$\mathrm{I}(i;j;k) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} p(i,j,k)\log_2\frac{p(i,j,k)}{p(i)p(j)p(k)} \quad (2)$$

Conditional Entropy. Conditional entropy quantifies the remaining uncertainty of an element, when the other element(s) is (are) known. For instance, if i is a deterministic function of elements j and k then this conditional entropy is zero, since no more information is required to describe i when j and k are known.

On the contrary, if the elements above are independent, knowing j and k does not tell one anything about i and the conditional entropy is equal to the entropy itself. In hypermatrix-based conditional entropy calculations three possibilities for the derivation of IFIs do exist.

Triple matrix: Set theory provides a valuable tool for the comprehension of conditional entropy-based IFIs and their derivations could be properly explained using visual descriptions offered by Venn diagrams. Figure 3 illustrates Venn diagram representations for the possible conditional entropy-based IFIs offered by three-source events.

a. $H(i|j, k)$Figure 3a helps us to deduce the formula for conditional entropy $H(i|j, k)$ defined as:

$$H(i|j,k) = H(i,j,k) - H(j,k) \qquad i = 1,2,3,\ldots,\ n;\quad j = 1,2,3,\ldots,n;\ k = 1,2,3,\ldots,\ n \quad (3)$$

Note that though $H(i|j, k)$ as a property is mathematically different from $H(j|\ i, k)$ and $H(k|i, j)$, in our formalism these give the same outcome since i, j, and k are not essentially predetermined and span over the same values from *1* to n, where n is the vertex number of **G**.

b. $H(i, j \mid k)$Using the Venn diagram representation in Figure 3b the formula for $H(j \mid i, k)$ is established:

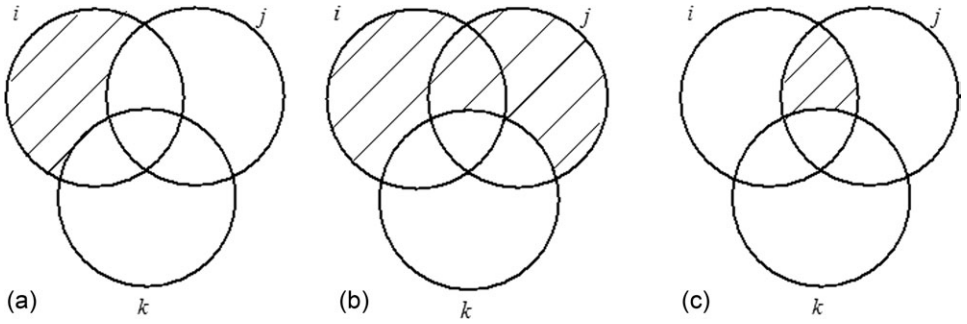


Figure 3. Venn diagram representations for the possible conditional entropy based IFIs offered by three source events: (a) $H(i|j,k)$, (b) $H(i,j \mid k)$, and (c) $H(i;j \mid k)$.

$$H(i,j|k) = H(i,j,k) - H(k) \qquad i = 1,2,3,\ldots,\ n;\ j = 1,2,3,\ldots,\ n;\ k = 1,2,3,\ldots,n \qquad (4)$$

As in the case of $H(i,j \mid k)$, the conditional entropies $H(i,j \mid k)$, $H(i,k|j)$, and $H(k,j \mid i)$ give the same result because of the already explained reason.

c. $H(i;j \mid k)$Finally from the visual reference of Venn diagram representation in Figure 3c, it is deduced that:

$$H(i;j|k) = H(i;j) - H(k) \qquad i = 1,2,3,\ldots,\ n;\ j = 1,2,3,\ldots,\ n;\ k = 1,2,3,\ldots,\ n \qquad (5)$$

Note that this conditional entropy type is denominated conditional mutual entropy.

The conditional entropies $H(i;j \mid k)$, $H(i;k|j)$, and $H(k;j \mid i)$ likewise give the same result as previously explained.

Quadruple Matrix: The quadruple-frequency matrices present one difficulty: the impossibility to construct Venn diagram representations for them on a bidimensional scale. Therefore, to define conditional entropy-based IFIs from such matrices, logical analogies using triple-matrix conditional entropy equations are crucial:

a. $$H(i|j,k,l) = H(i,j,k,l) - H(j,k,l) \quad i = 1,2,3,\ldots,n;\ j = 1,2,3,\ldots,n;\ k = 1,2,3,\ldots,n;\ l = 1,2,3,\ldots,n \qquad (6)$$

b. $$H(i,j,k|l) = H(i,j,k,l) - H(l) \qquad i = 1,2,3,\ldots,n;\ j = 1,2,3,\ldots,n;\ k = 1,2,3,\ldots,n;\ l = 1,2,3,\ldots,n \qquad (7)$$

c. $$H(i;j;k|l) = H(i;j;k) - H(l) \qquad i = 1,2,3,\ldots,n;\ j = 1,2,3,\ldots,n;\ k = 1,2,3,\ldots,n;\ l = 1,2,3,\ldots,n \qquad (8)$$

Joint Entropy. Joint entropy defines the total entropy of a system, whose elements (or variables) demonstrate interdependence (joint sets). Figure 4 illustrates the Venn diagram representation of the space covered by joint entropy in a three-source event.

Joint Entropy-based IFIs for triple and quadruple dimensional space are defined respectively as:

$$\mathrm{I}(i,j,k) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} p(i,j,k)\log_2 p(i,j,k) \qquad (9)$$

$$\mathrm{I}(i,j,k,l) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} p(i,j,k,l)\log_2 p(i,j,k,l) \qquad (10)$$

Note that when $i = j = k = l$, formulas (9) and (10) reduce to the classical formula for Shannon's entropy. Shannon's entropy-based IFIs will not be the basis of this report. They were defined in the previous publication using the principle diagonal elements of the frequency matrix and, since these elements are all equal for duplex, triple, and quadruple matrices, Shannon's entropy-based IFIs would give the same result.

Step 4: Computation of atomic entropies (LOVIs).

In duplex-matrix-based IFIs, this operation comprised the summation of the row or column entries of the respective entropy matrices, and the resulting values (atomic entropies) comprised a vector of LOVIs.[11] In triple-matrix considerations, to obtain the atomic entropies, the addition is carried out over j and k (with i constant) for each sheet (slide) that constitutes the matrix, while in the case of the quadruple matrix the summation is over j, k, and l (with i constant). Let us calculate the atomic joint entropy for the molecule of isopentane using the triple matrix approach as an illustration:

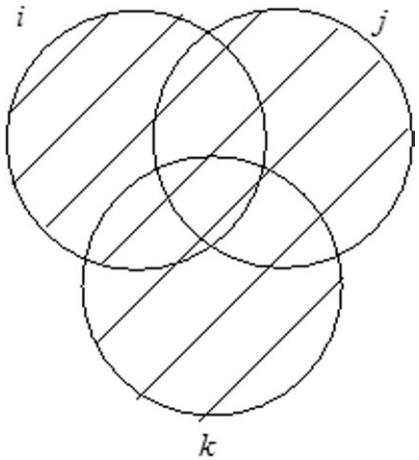


Figure 4. Venn diagram representation of the space covered by Joint Entropy in a three-source event.

1

$$\begin{bmatrix} 0.5271 & 0.5303 & 0.4912 & 0.3632 & 0.4416 \\ 0.5353 & 0.5303 & 0.4912 & 0.3632 & 0.4416 \\ 0.4912 & 0.4912 & 0.4912 & 0.3632 & 0.3632 \\ 0.3632 & 0.3632 & 0.3632 & 0.3632 & 0.2404 \\ 0.4416 & 0.4416 & 0.3632 & 0.2404 & 0.4416 \end{bmatrix}$$

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{IEJ}_1 = 10.5319$$

2

$$\begin{bmatrix} 0.5303 & 0.5303 & 0.4912 & 0.3632 & 0.4416 \\ 0.5353 & 0.3547 & 0.5117 & 0.4912 & 0.5303 \\ 0.4912 & 0.5117 & 0.5117 & 0.4912 & 0.4912 \\ 0.3632 & 0.4912 & 0.4912 & 0.4912 & 0.3632 \\ 0.4416 & 0.5303 & 0.4912 & 0.3632 & 0.5203 \end{bmatrix}$$

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{IEJ}_2 = 11.8284$$

3

$$\begin{bmatrix} 0.4912 & 0.4912 & 0.4912 & 0.3632 & 0.3632 \\ 0.4912 & 0.5117 & 0.5117 & 0.4912 & 0.4912 \\ 0.4912 & 0.5117 & 0.4503 & 0.5193 & 0.4912 \\ 0.3632 & 0.4912 & 0.5193 & 0.5193 & 0.3632 \\ 0.3632 & 0.4912 & 0.4912 & 0.3632 & 0.4912 \end{bmatrix}$$

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{IEJ}_3 = 11.6168$$

4

$$\begin{bmatrix} 0.3632 & 0.3632 & 0.3632 & 0.3632 & 0.2404 \\ 0.3632 & 0.4912 & 0.4912 & 0.4912 & 0.3632 \\ 0.3632 & 0.4912 & 0.5193 & 0.5193 & 0.3632 \\ 0.3632 & 0.4912 & 0.5193 & 0.5303 & 0.3632 \\ 0.2404 & 0.3632 & 0.3632 & 0.3632 & 0.3632 \end{bmatrix}$$

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{IEJ}_4 = 10.1101$$

5

$$\begin{bmatrix} 0.4416 & 0.4416 & 0.3632 & 0.2404 & 0.4416 \\ 0.4416 & 0.5303 & 0.4912 & 0.3632 & 0.5303 \\ 0.3632 & 0.4912 & 0.4912 & 0.3632 & 0.4912 \\ 0.2404 & 0.3632 & 0.3632 & 0.3632 & 0.3632 \\ 0.4416 & 0.5303 & 0.4912 & 0.3632 & 0.5271 \end{bmatrix}$$

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{IEJ}_5 = 10.5319$$

and the LOVIs vector with atomic joint entropy component can be obtained as:

$$V_{\mathrm{IEJ}} = (10.5319,\ 11.8284,\ 11.6168,\ 10.1101,\ 10.5319)$$

The application of the summation operator to this vector, analogous to the outer summation in i of eq. (9), gives IEJ = 54.6191 bits for the whole molecule.

Step 5: Application of invariants to the vector of LOVIs.

This procedure allows us to generalize the attainment of global (or local) invariants by summation of the LOVIs. These invariants were introduced in the previous publication and the contribution of this generalization was rather interesting.[11] These are classified in three major groups (see Table 1 for more information)[15,16]:

1. Norms (or Metrics): Minkowski's norms (N1, N2, N3) and Penrose's size (PN). It should be note that the summation operation is analogous to Minkowski's first norm (N1) in our case.

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

Codification of heteroatoms and insaturations In the previous report, we proposed a method that permitted us to discriminate adequately molecules with heteroatoms and insaturations.[11] Let us have a brief recapitulation of the concepts and the procedure utilized to achieve this important attribute for MDs.

Consider as an example an isomorph of isopentane, 2-methylpropanal molecule (Fig. 1b). As it is expected the mutual, conditional or joint entropy vector values for this **G** will be identical to that of isopentane. However, the molecular structure of 2-methylpropanal in contrast to that of isopentane contains a heteroatom and a double bond, and our objective is to achieve discrimination between structures of this nature.

Let us create a vector of weights V_p, in which weight (ϑ_i) corresponds reciprocally to element i for a given condition. The distinct weights for each atom (condition, in agreement with this event) can be determined according to the relationship $\vartheta_i = P/\delta$ (for this event based on atoms), where P represents a characteristic property of each atom (for example, atomic mass, electronegativity, etc.) and δ is the bond vertex degree.

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

$$\mathrm{p(O)} = \tfrac{3.44}{2} = 1.72 \qquad \mathrm{p(C3)} = \tfrac{2.55}{3} = 0.85$$

$$\mathrm{p(C1)} = \tfrac{2.55}{1} = 2.55 \qquad \mathrm{p(C5)} = \tfrac{2.55}{1} = 2.55$$

$$\mathrm{p(C2)} = \tfrac{2.55}{3} = 0.85$$

From the resulting values above we construct a vector of weights, V_p = (2.55, 0.85, 0.85, 1.72, 2.55). Then, we use the vector inner product (V_{IFI} [×] V_p) = ${}^{p}V_{\mathrm{IFI}}$ by multiplying each component of the vector of atomic entropies (V_{IFI}) by its corresponding element of vector V_p.

Consider the vector of triple-matrix-based atomic joint entropies (V_{IEJ}) calculated for the molecule of isopentane. The vector inner product (V_{IEJ} [×] V_p) would yield a vector of weighted atomic entropies (${}^{p}V_{\mathrm{IEJ}}$). Summation of the components of this vector gives the mean-weighted joint entropy for

Table 1. Invariant functions to derive molecular descriptors (total and local) from LOVIs.

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

The x_i is LOVI associated to the atoms v_i and *n* is the number of atoms.

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

2-methylpropanal (molecule b in Fig. 1). The vector for weighted atomic joint entropies in the case of isopentane would comprise the same values as for 2-methylpropanal, with an exception of the ones for the carbonyl group atoms (C_3 and O) in the latter case, which explains the importance of applying weights to achieve the much desired discrimination between isomeric molecules.

This procedure does not violate the ignore the substantial meaning prerequisite established for the measure of the degree of uncertainty, since parameters of properties intrinsic (or particular) to the vertices (atoms) are applied to values of atomic entropies already determined in the preceding steps, which would otherwise be violated if the weights were applied to frequency or probability values.

Finally, the introduction of the vector V_{IFI} in our method permits the definition of local IFIs for atom types or groups [(for example, in TOMOCOMD-CARDD program,[17] the following local indices can be calculated: Proton Acceptors (AH), Proton Donors (DH), Heteroatoms (HT), Halogens (HL), Carbons (Cb), Methyl Carbons (MC), and Unsaturated bonds (IS)]. It should be noted that the local definition capacity is one of the most important requisites for new MDs,[18,19] and all classic equivalence class-based IFIs are short of this quality.

A careful scrutiny of the equations for mutual, conditional and joint entropies reveals that no restrictions do exist on the values that *i*, *j*, *k*, and *l* could take. Their values span over all the range [1, *n*], where *n* is the number of vertices that contain **G**. We may however wish to introduce restrictions to these equations that permit us to define IFIs with definite particularities. Let us look at the possible restrictions that could be introduced.

Triple matrix

a. 3: only the matrix entries that satisfy the condition $i \neq j \neq k$ are selected.

b. 3, 2: only the matrix entries that satisfy the condition $i \neq j \neq k$ and $i = j \,||\, j = k \,||\, i = k$ are selected.

Quadruple matrix

a. 4: only the matrix entries that satisfy the condition $i \neq j \neq k \neq l$ are selected.

b. 4, 3: only the matrix entries that satisfy the condition $i \neq j \neq k \neq l$ and $i = j \,||\, j = k \,||\, k = l \,||\, i = j \,||\, i = k \,||\, i = l \,||\, j = l$ are selected.

c. 4, 2: only the matrix entries that satisfy the condition $i \neq j \neq k \neq l$ and $i = j = k \,||\, i = j = l \,||\, i = k = l \,||\, j = k = l$ are selected.

A full description of the nomenclature adapted for these particularities is available as Supporting Information (Table SI3).

It is crucial to point out that the invariants in Table 1 could be applied not only to the vector of original LOVIs but also to the vector of standardized LOVIs. That is to say, the global or local IFIs are calculated from a vector of standardized atomic LOVIs. In the standardization procedure, the original LOVI values are transformed to standardized ones as follows: Std. LOVIs = (Original LOVI − mean of LOVIs)/Std. deviation of original LOVIs. With this renormalization, the vector of standardized LOVIs has a mean of 0 and standard deviation of 1. In other words, standardized LOVIs have the same dimension, that is, are of comparable magnitude.

During the implementation of the proposed IFIs in the TOMOCOMD-CARDD program (GT-STAF module, see below), it turned out to be critical to incorporate the use of catalyst files (.CATA, .3MTX and .4MTX for duplex, triple and quadruple matrices, respectively) to optimize the computational process given that the search for connected subgraphs that make up a particular **G** is a No Polynomial problem by nature and would, otherwise, be time consuming and computationally costly. Table 2 shows a sample of the MD (IFI) values calculated for a dataset of 41 structurally diverse molecules.

It is evident that local IFIs (refer to $ILEM_{N1}$ [(HT) (T)] in Table 2 as an example) are quite degenerate as they are specifically sensitive to the presence of particular sections (fragments) of a molecular structure. Note that the use of a weighting scheme permits discriminating molecules with unsaturated bonds and heteroatoms, in contrast to the unweighted IFIs (compare $^{V}IEM_{N1}$[(3T)] and IEM_{N1}[(3T)] in Table 2). On the contrary, the proposed IFIs are not sensitive to stereoisomeric differences in molecular structures (see *cis*- and *trans*-2-butene in Table 2).

Variability Analysis of the Proposed IFIs

Highly variable MDs are doubtlessly ideal tools for any QSPR/QSAR application or diversity analysis. This is because such MDs are generally sensitive to changes in molecular structures and, thus, capable of effectively discriminating compounds with different chemical characteristics. Godden et al.[12] proposed an information theory-based approach, using the concept of Shannon's entropy, to evaluate and quantify the information content and, thus, the variability of MDs. To accomplish this goal, a discretization procedure (binning scheme) is presented, which uses histograms of descriptor distributions (equal interval width method) to enable the comparison of descriptors with different units and value ranges. It follows that p_i is the probability that a data point (case or variable-wise) adopts a value within specific data interval *i* (bin *i*). Therefore, a probability distribution function $P = (p1, p2, \ldots, pn)$ is formed. Shannon's fundamental equation is then applied to the resulting uniform data distribution. Moreover, a modification of this concept denominated differential Shannon's entropy was introduced in order to take into account both the variability and value range distributions of MDs.[13] To evaluate the quality of the IFIs proposed in the present report and demonstrate the potential of these IFIs as reliable tools in chemical structure related studies, we implemented this innovative application of information theory to variability analysis in an interactive software denominated IMMAN (acronym for Information Theory based CheMoMetric ANalysis), enriched with additional parameters, derived from modifications of Shannon's entropy as: standardized Shannon's entropy, Negentropy, Brillouin Redundancy Index, Gini index, and Information

Table 2. Sample of MD values obtained for a dataset of 40 structurally diverse molecules.

Compound	$IEM_{N1}[(3T)]$[a]	$^{V}IEM_{N1}[(3T)]$[b]	$ILEM_{N1}[(HT)(T)]$[c]	$IEJ_{N1}[(3T)]$[d]	$IEC_{N1}[(3T)(X\_YZ)]$[e]	$IEM_{N1}[(QD)]$[f]	$IEJ(N1)[(QD1)]$[g]
n-Propane	0.8050	13.6951	0.0000	12.8998	0.1950	4.1950	15.5098
n-Butane	3.6356	55.8603	0.0000	28.4923	1.9306	20.9723	65.3263
n-Pentane	9.4202	135.1053	0.0000	52.3649	5.4405	63.0092	173.0410
n-Hexane	19.1132	261.1795	0.0000	86.1506	11.2326	147.1328	367.2813
Isobutane	3.0562	52.3845	0.0000	29.4230	1.5870	17.6727	67.1639
Neopentane	6.1720	109.6548	0.0000	56.4300	4.2397	40.5890	185.2199
2-Methylpentane	17.7438	260.0639	0.0000	90.3515	9.8515	135.6509	387.3562
cis-2-butene	3.6356	47.2738	0.0000	28.4923	1.9306	20.9723	65.3263
trans-2-butene	3.6356	47.2738	0.0000	28.4923	1.9306	20.9723	65.3263
2-Butyne	3.6356	42.9805	0.0000	28.4923	1.9306	20.9723	65.3263
Cyclopropane	0.0630	0.7076	0.0000	13.0853	0.3618	1.2429	14.4378
Cyclobutane	1.5665	17.5839	0.0000	30.3098	1.9383	9.6353	66.6567
Cyclopentane	6.0745	68.1860	0.0000	58.7833	4.4966	39.1624	192.9808
Cyclohexane	15.1952	170.5658	0.0000	101.5246	8.1558	112.4514	440.8843
Cyclohexanone	24.1322	215.9486	1.7796	157.1318	13.9562	202.8332	830.9954
Benzene	15.1952	113.7105	0.0000	101.5246	8.1558	112.4514	440.8843
Toluene	24.1322	215.8517	0.0000	157.1318	13.9562	202.8332	830.9954
Phenol	24.1322	155.0222	1.7796	157.1318	13.9562	202.8332	830.9954
Benzoic acid	54.9013	316.5720	7.3320	329.6908	27.2989	584.1757	2147.4836
Aniline	24.1322	166.9502	1.7796	157.1318	13.9562	202.8332	830.9954
Nitrobenzene	54.9013	285.3981	12.7754	329.6908	27.2989	584.1757	2147.4836
Fluorobenzene	24.1322	177.3980	1.7796	157.1318	13.9562	202.8332	830.9954
Chlorobenzene	24.1322	215.8517	1.7796	157.1318	13.9562	202.8332	830.9954
Bromobenzene	24.1322	240.8393	1.7796	157.1318	13.9562	202.8332	830.9954
Iodobenzene	24.1322	263.2731	1.7796	157.1318	13.9562	202.8332	830.9954
Benzamide	54.9013	337.7389	7.3320	329.6908	27.2989	584.1757	2147.4836
Naphthalene	59.0144	417.1033	0.0000	475.9686	29.2139	688.3885	2147.4836
Anthracene	96.5206	676.3137	0.0000	1360.6463	27.9717	1577.8263	2147.4836
Pyrrole	6.0745	39.7676	0.4569	58.7833	4.4966	39.1624	192.9808
Furan	6.0745	37.2710	0.4569	58.7833	4.4966	39.1624	192.9808
Thiophen	6.0745	49.2012	0.4569	58.7833	4.4966	39.1624	192.9808
Purine	41.3921	196.4662	12.0869	343.8739	24.6906	433.6550	2147.4836
Dibenzofuran	79.8692	532.2397	3.1067	1074.6435	36.0755	1208.2343	2147.4836
Ethanol	0.8050	9.3455	0.0000	12.8998	0.1950	4.1950	15.5098
Trifluoroethanol	13.2919	103.9548	3.9586	94.0566	8.8941	99.8953	400.5121
2-Aminoethanol	3.6356	30.5127	0.4263	28.4923	1.9306	20.9723	65.3263
Propanol	3.6356	41.8088	0.2131	28.4923	1.9306	20.9723	65.3263
Propanone	3.0562	36.0888	0.1915	29.4230	1.5870	17.6727	67.1639
2-Propanol	3.0562	38.6068	0.1915	29.4230	1.5870	17.6727	67.1639
2-Propylamine	3.0562	41.3085	0.1915	29.4230	1.5870	17.6727	67.1639

[a] Triple matrix-based mutual information indices using invariant N1. [b] Weighted triple matrix-based mutual information indices using invariant N1and van der Waals Volume(V) as weight. [c] $i \neq j \neq k$ element triple matrix-based local mutual information Indices for Heteroatoms (HT) using invariant N1. [d] Triple matrix-based joint information indices using invariant N1. [e] Triple matrix-based X/YZ-type conditional information indices using invariant N1. [f] Quadruple matrix-based mutual information indices using invariant N1. [g] $i \neq j \neq k \neq l$ && $l = j \,||\, j = k \,||\, k = l \,||\, i = j \,||\, i = k \,||\, i = l \,||\, j = l$ element quadruple matrix-based joint information indices using invariant N1.

Energy Content, previously not used in the evaluation of the variability of MDs.[4]

The variability analysis performed in the present report comprised three main parts: first, a quantitative comparison in terms of entropy values of duplex, triple, and quadruple matrix-based IFIs; second, a comparison between the entire family of the proposed IFIs and DRAGON's IFIs; and third, a global comparison of the GT-STAF IFIs with MDs of other descriptor calculating programs. For this study, a small dataset of 40 structurally diverse compounds was used (Table 2), and the descriptor calculations were performed using GT-STAF (acronym for Graph Theoretical Thermodynamic STAte Functions), a new module of TOMOCOMD-CARDD program that offers rapid and low-computational-cost calculations of the proposed IFIs. With the resulting descriptor values, the respective SE values were determined using a binning scheme of 20 intervals (bins). The same number of variables was used for each case study to ensure an objective comparative analysis, with the class presenting the least number of variables determining the cut-off value. In the rest of the classes, the best variables up to the cut-off number were considered. The use of the same number of variables is preferred to the probability-based normalization procedure used by Hong et al.,[20] in which a probability scale (0.0–1.0) is used instead of a reduced the number of variables to obtain equal number of variables in each case, as this gives a biased graphic perspective when comparing cases with substantially unequal number of variables. However, when the same number of variables is used, the probability scale could be used as well. The local IFIs were not included in this study, essentially due to the fact that these account for particular features in a molecular structure and, thus, would depict low, if any, variability. All the variables employed in this

study are available as supporting information (WinRAR folder file SI4).

Comparative analysis of shannon's, mutual, conditional, and joint entropy-based ifis for duplex, triple and quadruple matrix approaches

The purpose of this study is to evaluate the contribution, if any, of the introduction of the hypermatrix-based approach in terms of an increase in the variability of the mutual-, conditional-, and joint entropy-based IFIs. Figure 5 shows a graphical comparison of the distribution of the best 126 variables for each family of IFIs, with Shannon's entropy-based IFIs determining the cut-off number. For a discretization scheme of 20 bins, the maximum entropy (Hartley's entropy) is given by $\log_2 20 = 4.3219$ bits. Figure 5 reveals that conditional entropy-based IFIs, derived from a quadruple matrix approach, interestingly present the best entropy distribution with 80% (102) of the variables presenting entropy over 3.50 bits, followed by triple-matrix-based IFIs belonging to the same family of indices (conditional entropy-based IFIs). Next in the order of entropy distribution are the triple and quadruple matrix-based mutual entropy IFIs, which show comparable behavior. These findings suggest that the use of hypermatrices seems to improve the entropy, and thus the variability of the previously proposed IFIs, particularly the conditional and mutual entropy-based IFIs.

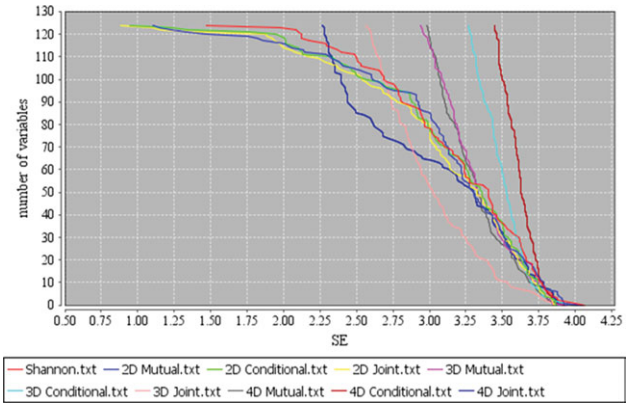


Figure 5. Shannon's entropy distribution for Shannon's, mutual, conditional, and joint entropy-based IFIs for duplex, triple, and quadruple matrix approaches.

Matrix-based outlook of the proposed IFIs

An analysis of the proposed IFIs from a matrix-based perspective was performed. In this case study, 472 variables were considered. As can be observed in Figure 6, duplex-matrix-based IFIs present the best entropy distribution, with 300 (approximately 64%) variables presenting entropy above 3.00 bits. Another remarkable observation is that, generally, whole matrix-based IFIs (3T and QD) seem to give better entropy values than variables calculated from matrices constructed from particular (reduced) elements, that is, T, QD1, QD2, and QD3 [refer to Supporting Information (Table SI3) for meanings].

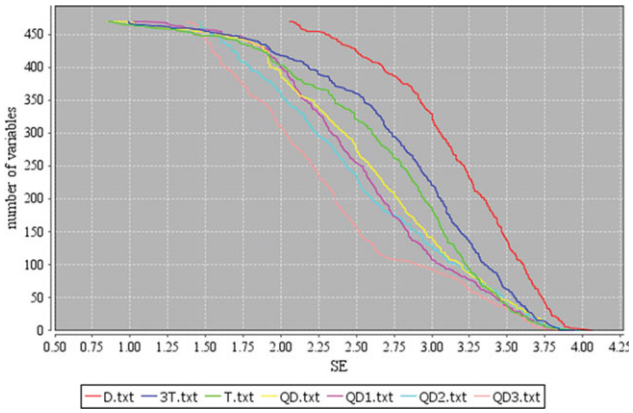


Figure 6. Shannon's entropy distribution for the GT-STAF IFIs from a matrix-based perspective.

Family wise comparison of DRAGON and GT-STAF indices

The DRAGON software, one of the most popular packages used in QSAR/QSPR studies, presents various families of MDs. Here, our concern is to compare the entropies of these descriptor families and the three families GT-STAF indices (classified on matrix-type base). Some DRAGON families were grouped together into bigger families, that is, OD-1D and others (functional group counts, atom-centered fragments, constitutional descriptors and molecular properties), 3D-Indices (charge descriptors, 3D-Morse descriptors, Randic molecular profiles and geometric descriptors), Topo-Indices (TIs, topological charge indices, connectivity indices), and Eigen-Indices (Burden eigenvalue descriptors, eigenvalue-based indices, 2D autocorrelations). The best 47 variables for each of these families were considered, with DRAGON's IFIs determining this cut-off number owing to its condition as the family with the fewest number of variables. The GT-STAF duplex, triple and quadruple-matrix-based IFIs presented comparable entropy distributions with 3D-Indices and GATEWAY descriptors. Additionally, better entropy distributions are observed when the GT-STAF duplex, triple and quadruple-matrix-based IFIs are compared with the rest of DRAGON's descriptor families (Fig. 7).

Comparison of GT-STAF software with other descriptor calculation packages

Finally, we performed a more generalized study, with the objective of comparing the GT-STAF software program with some of the relevant softwares used in for descriptor calculations in computational chemistry as: DRAGON,[21] MOLD2,[20] PADEL,[22] MODESLAB,[23–26] and CDK.[27] The cut-off number of 180 variables was provided by MODESLAB software. Figure 8 illustrates Shannon's entropy distribution for GT-STAF software and other descriptor computing packages. As can be observed, GT-STAF software presents similar to better entropy distribution than almost all the analyzed softwares. For example, the number of MDs with entropy values greater than 3.50 bits are 180 (100%) for the case of GT-STAF and DRAGON's MDs, 46 (26%) for MOLD2, 38 (21%) for CDK, an so forth.

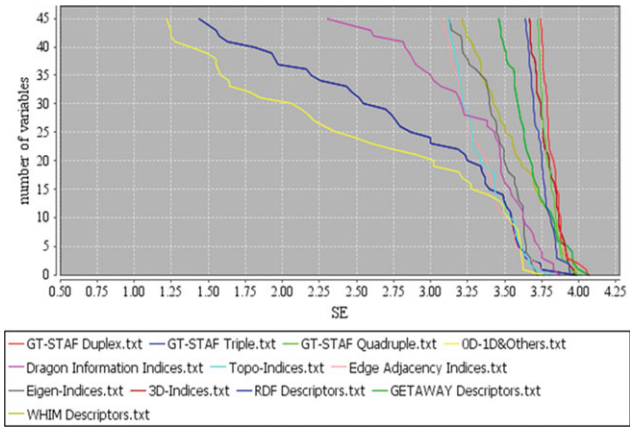


Figure 7. Shannon's entropy distribution for DRAGON's and GT-STAF descriptor families. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

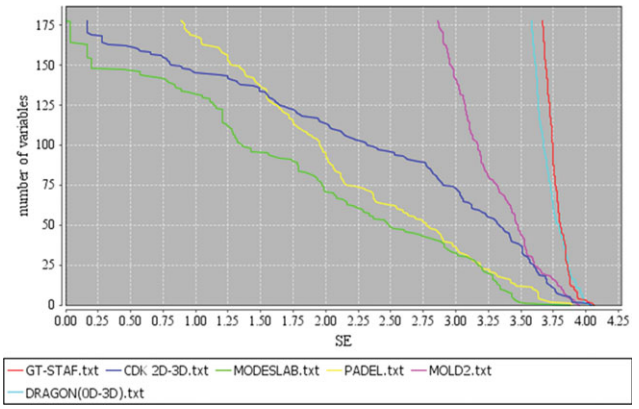


Figure 8. Shannon's entropy distribution for GT-STAF software and other MD calculating programs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

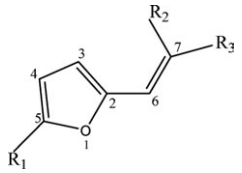
Moreover, in the case of DRAGON software where comparable behavior is observed, it is important to note that this software constitutes a series of substantially diverse MD families (0D–3D) derived from wide range of chemical and graph theoretic concepts. This result suggests that MDs calculated with the GT-STAF program may encode similar-to-better amount of structural information than the software packages compared in this study, and is possibly a relevant tool to take into account in QSPR/QSAR and similarity/dissimilarity analysis, at least according to Shannon's entropy-based variability analysis. Nevertheless, it is known that the structural variance is an important but not sufficient requisite, to postulate that an MD correlates with a particular physico-chemical, chemical or biological property. Therefore, the next section will be devoted to assessing the modeling power of the proposed IFIs.

QSPR Modeling of Physico-Chemical Properties of 2-Furylethylene Derivatives

Introduction

The descriptive or predictive capacity of molecular (or molecular-fragment) properties, offered by a particular MD, is imperative in claiming its relevance and possible application in QSPR/QSAR studies.[28] Consequently, to obtain a more profound insight of the contribution, if any, of hypermatrix-based IFIs in QSPR modeling, a search for the best regressions for the properties 1-octanol/water partition coefficient (Log *P*) and specific rate constant (Log *K*) for nucleophilic addition of a thiol group to the exo-cyclic double bond of the 34 2-furylethyelene derivates was performed. Subsequently, the performance of these indices was compared with the rest reported in the literature.[29,30] These 2-furylethylene

Table 3. Chemical structures and numbering of atoms in the furylethylenes compounds used in this study.



Number	R1	R2	R3	Number	R1	R2	R3
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	$CONHCH(CH_2)_3CH_3$	33	NO_2	H	H
17	NO_2	H	$CONHCH_2CH(CH_3)_2$	34	NO_2	CN	$COOCH_3$

derivatives have different substituents in position 5 of the furan ring as well as in position β of the exo-cyclic double bond (Table 3).

The values of Log P and Log *K* of these compounds have been experimentally determined and reported in the literature. The lipophilicity and the nucleophilic addition of the thiol groups of some enzymes to the exocyclic double bond of 2-furylethylene derivatives are critical for their antibacterial activity.[31] The Log *P* and Log *K* of nucleophilic addition of the mercaptoacetic acid to the exocyclic double bond are fundamental in the understanding of the biological behavior of these 2-furylethylene derivatives.[32,33] Thus, a study of these properties, using the proposed IFIs, permits us to get a general criterion about the applicability of these indices in QSPR studies.

The QSPRs were obtained with the software MOBYDIGS (version 1.0 – 2004).[34] This software allows searching for linear regression models (RLM), by developing optimal model populations using genetic algorithms (GAs). The theoretical basis of the GAs has been explained in detail elsewhere.[35–39] The population size was set at 100 and the reproduction/mutation trade-off ratio (*T*) at 0.70. The GAs with initial population sizes of 100 rapidly converged (200 generations) and achieved 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 both techniques "bootstrapping" (Q^2_{boot}) and "scrambling" [a (R^2), a (Q^2)]. The former evaluates the predictive power of the developed models and the latter checks the risk of fortuitous correlations (i.e., random correlations between the independent and response variables), a possibility when too many variables are screened relative to the number of available observations.[2,34]

The best models obtained, on the basis of the quality of the statistical parameters, using triple and quadruple matrix-based IFIs, respectively, are presented below:

$$\begin{aligned}&\textbf{\textit{Log P}} = 1.553(\pm0.092) + 0.765(\pm0.028)\ {}^{A}\textbf{\textit{SILEM}}_{P3}[(\textbf{\textit{IS}})((\textbf{\textit{T}})] + 0.027(\pm0.001)\ {}^{A}\textbf{\textit{ILEM}}_{G}[(\textbf{\textit{CB}})((\textbf{3\textit{T}})]\\&+ 1.889(\pm0.053)\ {}^{P}\textbf{\textit{ILEM}}_{CV}[(\textbf{\textit{HT}})((\textbf{3\textit{T}})] + 0.041(\pm0.004)\\&{}^{Z}\textbf{\textit{SILEC}}_{M}[(\textbf{\textit{IS}})((\textbf{3\textit{T}})(\textbf{\textit{XY\_Z}})] + 0.020(\pm0.004)\ {}^{A}\textbf{\textit{SILEC}}_{M}[(\textbf{\textit{HT}})((\textbf{3\textit{T}})(\textbf{\textit{XY\_Z}})]+\\&0.252(\pm0.049)\textbf{\textit{SIEC}}_{I50}[(\textbf{3\textit{T}})(\textbf{\textit{XnY\_Z}})] - 0.338(0.033)\textbf{\textit{SILEM}}_{P3}[(\textbf{\textit{IS}})(\textbf{3})(\textbf{3\textit{T}})]\\&N = 34\quad R^2 = 0.994\quad Q^2_{loo} = 0.990\quad Q^2_{boot} = 0.984\quad \mathrm{SECV} = 0.063\quad F = 608.29\end{aligned} \tag{11}$$

$$\begin{aligned}&\textbf{\textit{Log K}} = 3.065(\pm0.250) - 0.099(\pm0.009)\ {}^{V}\textbf{\textit{ILEM}}_{M}[(\textbf{\textit{IS}})((\textbf{\textit{T}})] - 0.045(\pm0.001)\\&{}^{A}\textbf{\textit{ILEC}}_{Q3}[(\textbf{\textit{IS}})((\textbf{\textit{T}})(\textbf{\textit{XnY\_Z}})] + 0.157(\pm0.005)\ {}^{P}\textbf{\textit{ILEC}}_{Q2}[(\textbf{\textit{CB}})((\textbf{\textit{T}})(\textbf{\textit{XnY\_Z}})] + 0.106(\pm0.011)\\&\textbf{\textit{ILEC}}_{MN}[(\textbf{\textit{HT}})((\textbf{3\textit{T}})(\textbf{\textit{X\_YZ}})] + 1.658(\pm0.093)\ {}^{A}\textbf{\textit{SILEC}}_{MX}[(\textbf{\textit{CB}})((\textbf{3\textit{T}})(\textbf{\textit{XY\_Z}})] - 1.124(\pm0.038)\\&{}^{V}\textbf{\textit{SILEC}}_{S}[(\textbf{\textit{IS}})(\textbf{6},\textbf{7},\textbf{8})(\textbf{\textit{T}})(\textbf{\textit{X\_YZ}})] + 0.279(0.061)\ {}^{V}\textbf{\textit{SIEC}}_{G}[(\textbf{6},\textbf{7},\textbf{8})(\textbf{\textit{T}})(\textbf{\textit{X\_YZ}})]\\&N = 34\quad R^2 = 0.998\quad Q^2_{loo} = 0.995\quad Q^2_{boot} = 0.991\quad \mathrm{SECV} = 0.081\quad F = 1453.20\end{aligned} \tag{12}$$

$$\begin{aligned}&\textbf{\textit{Log P}} = 4.180(\pm0.120) + 0.020(\pm0.001)\ {}^{P}\textbf{\textit{ILEM}}_{G}[(\textbf{\textit{HT}})(\textbf{\textit{QD}})] - 0.110(\pm0.016)\\&{}^{V}\textbf{\textit{ILEM}}_{S}[(\textbf{\textit{HT}})(\textbf{5},\textbf{6})(\textbf{\textit{QD1}})] + 0.001(\pm0.000)\ {}^{V}\textbf{\textit{ILEM}}_{Q3}[(\textbf{\textit{CB}})(\textbf{6},\textbf{7},\textbf{8})(\textbf{\textit{QD}})]-\\&3.139(\pm0.324)\ {}^{V}\textbf{\textit{SILEM}}_{I50}[(\textbf{\textit{HT}})(\textbf{6})(\textbf{\textit{QD1}})] + 0.810(\pm0.078)\ {}^{V}\textbf{\textit{SILEM}}_{Q2}[(\textbf{\textit{CB}})(\textbf{8})(\textbf{\textit{QD2}})]\\&- 0.002(\pm0.000)\ {}^{V}\textbf{\textit{ILEC}}_{Q2}[(\textbf{\textit{IS}})(\textbf{4})(\textbf{\textit{QD1}})(\textbf{\textit{X\_YZM}})] - 0.012(0.001)\\&{}^{V}\textbf{\textit{ILEC}}_{Q2}[(\textbf{\textit{HT}})(\textbf{4})(\textbf{\textit{QD2}})(\textbf{\textit{X\_YZM}})]\\&N = 34\quad R^2 = 0.991\quad Q^2_{loo} = 0.987\quad Q^2_{boot} = 0.983\quad \mathrm{SECV} = 0.078\quad F = 396.70\end{aligned} \tag{13}$$

$$\begin{aligned}&\textbf{\textit{Log K}} = 3.102(\pm0.245) - 0.361(\pm0.055)\ {}^{V}\textbf{\textit{SILEM}}_{N3}[(\textbf{\textit{HT}})(\textbf{6},\textbf{7},\textbf{8})(\textbf{\textit{QD2}})] + 0.249(\pm0.070)\\&{}^{V}\textbf{\textit{SILEM}}_{N2}[(\textbf{\textit{CB}})(\textbf{6},\textbf{7},\textbf{8})(\textbf{\textit{QD3}})] + 0.540(\pm0.040)\ {}^{V}\textbf{\textit{SILEM}}_{P3}[(\textbf{\textit{MC}})(\textbf{8})(\textbf{\textit{QD}})] - 0.034(\pm0.004)\\&{}^{V}\textbf{\textit{SILEM}}_{CV}[(\textbf{\textit{IS}})(\textbf{4})(\textbf{\textit{QD}})] - 0.004(\pm0.000)\ {}^{V}\textbf{\textit{ILEM}}_{MX}[(\textbf{\textit{MC}})(\textbf{4})(\textbf{\textit{QD3}})] + 3.170(\pm0.054)\\&{}^{V}\textbf{\textit{ILEJ}}_{S}[(\textbf{\textit{IS}})(\textbf{4})(\textbf{\textit{QD2}})] + 0.022(0.001)\ {}^{V}\textbf{\textit{ILEJ}}_{MN}[(\textbf{\textit{CB}})(\textbf{4})(\textbf{\textit{QD3}})]\\&N = 34\quad R^2 = 0.995\quad Q^2_{loo} = 0.991\quad Q^2_{boot} = 0.989\quad \mathrm{SECV} = 0.112\quad F = 759.43\end{aligned} \tag{14}$$

where, *N* is the number of compounds, R^2 is the determination coefficient, SECV 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, a (Q^2) is the intercept value, obtained from the validation technique scrambling, and *F* is the Fisher ratio.

The statistical parameters of the obtained models show satisfactory robustness and predictive capacity. A table showing the experimental and calculated values of Log *P* and Log *K* according to the models (11) and (12), respectively, is available as Supporting Information SI5.

Comparative study

Finally, a more comprehensive evaluation of the modeling power of the previously proposed IFIs was performed, using mixed models comprising of MDs derived from duplex, triple, and quadruple matrix-based approaches. The best models attained for 3, 4, 5, and 6 variables are given below:

$$\textbf{Log P} = 3.302(\pm0.261) + 0.029(\pm0.003)\ {}^{Z}\textbf{ILEM}_{A}[(\textbf{HT})((\textbf{T})] - 0.517(\pm0.063)\ \textbf{SILEM}(\textbf{N3})[(\textbf{IS})(\textbf{3})(\textbf{3T})] - 0.279(\pm0.026)\ {}^{A}\textbf{IEM}_{A}[\textbf{D}]$$
$$N = 34\ \ R^2 = 0.915\ \ Q^2_{loo} = 0.889\ \ Q^2_{boot} = 0.884\ \ \text{SECV} = 0.219\ \ F = 107.28 \quad (15)$$

$$\textbf{Log P} = 2.245(\pm0.167) + 0.028(\pm0.002)\ {}^{Z}\textbf{ILEM}_{PN}[(\textbf{HT})((\textbf{3T})] - 0.671(\pm0.068)\ \textbf{SILEM}_{P3}[(\textbf{IS})(\textbf{3})(\textbf{3T})] - 0.256(\pm0.039)\ {}^{E}\textbf{ILEM}_{K}[(\textbf{DH})(\textbf{D})] - 0.432(\pm0.027)\ {}^{A}\textbf{IEC}_{PN}[(\textbf{D})]$$
$$N = 34\ \ R^2 = 0.955\ \ Q^2_{loo} = 0.941\ \ Q^2_{boot} = 0.933\ \ \text{SECV} = 0.162\ \ F = 153.55 \quad (16)$$

$$\textbf{Log P} = -1.010(\pm0.137) + 0.337(\pm0.020)\ {}^{A}\textbf{SILEM}_{MX}[(\textbf{IS})((\textbf{QD3})] + 0.035(\pm0.002)\ {}^{Z}\textbf{ILEM}_{PN}[(\textbf{HT})((\textbf{T})] + 0.039(\pm0.005)\ {}^{E}\textbf{SILEC}_{M}[(\textbf{IS})((\textbf{3T})(\textbf{XY\_Z})] + 0.584(\pm0.067)\ \textbf{SIEC}_{I50}[(\textbf{3T})(\textbf{XnY\_Z})] - 0.547(\pm0.054)\ \textbf{SILEM}_{P3}[(\textbf{IS})(\textbf{3})(\textbf{3T})]$$
$$N = 34\ \ R^2 = 0.979\ \ Q^2_{loo} = 0.967\ \ Q^2_{boot} = 0.964\ \ \text{SECV} = 0.112\ \ F = 262.90 \quad (17)$$

$$\textbf{Log P} = -0.809(\pm0.095) + 0.758(\pm0.036)\ {}^{A}\textbf{SILEM}_{P3}[(\textbf{IS})((\textbf{T})] + 0.020(\pm0.001)\ {}^{V}\textbf{IEM}_{PN})(\textbf{T})] + 1.504(\pm0.065)\ {}^{P}\textbf{ILEM}_{CV}[(\textbf{HT})((\textbf{3T})] + 0.022(\pm0.005)\ {}^{A}\textbf{SILEC}_{M}[(\textbf{HT})((\textbf{3T})(\textbf{XY\_Z})] + 0.030(\pm0.004)\ {}^{E}\textbf{SILEC}_{M}(\textbf{IS})((\textbf{3T})(\textbf{XY}_{\textbf{Z}})] - 0.284(\pm0.042)\ \textbf{SILEM}_{P3}[(\textbf{IS})(\textbf{3})(\textbf{3T})]$$
$$N = 34\ \ R^2 = 0.989\ \ Q^2_{loo} = 0.984\ \ Q^2_{boot} = 0.965\ \ \text{SECV} = 0.082\ \ F = 415.37 \quad (18)$$

$$\textbf{Log K} = 5.926(\pm0.068) + 1.743(\pm0.118)\ {}^{V}\textbf{ILEJ}_{S}[(\textbf{IS})(\textbf{4})(\textbf{T})] + 0.576(\pm0.034)\ {}^{V}\textbf{ILEM}_{K}[(\textbf{CB})(\textbf{6},\textbf{7})(\textbf{T})] + 0.070(\pm0.006)\textbf{SILEJ}_{M}[(\textbf{IS})((\textbf{QD1})]$$
$$N = 34\ \ R^2 = 0.977\ \ Q^2_{loo} = 0.969\ \ Q^2_{boot} = 0.946\ \ \text{SECV} = 0.227\ \ F = 424.13 \quad (19)$$

$$\textbf{Log K} = 3.696(\pm0.428) + 1.837(\pm0.088)\ {}^{V}\textbf{SILEJ}_{S}[(\textbf{IS})(\textbf{4})(\textbf{T})] + 0.516(\pm0.027)\ {}^{V}\textbf{ILEM}_{K}[(\textbf{CB})(\textbf{6},\textbf{7})(\textbf{T})] + 2.140(\pm0.408)\textbf{SILEC}_{Q3}[(\textbf{IS})((\textbf{QD3})(\textbf{XYZ\_M})] + 0.062(\pm0.004)\ \textbf{SILEJ}_{M}[(\textbf{IS})((\textbf{QD1})]$$
$$N = 34\ \ R^2 = 0.988\ \ Q^2_{loo} = 0.984\ \ Q^2_{boot} = 0.969\ \ \text{SECV} = 0.166\ \ F = 605.57 \quad (20)$$

$$\textbf{Log K} = 3.873(\pm0.386) + 0.605(\pm0.605)\ {}^{V}\textbf{SIEC}_{Q2}[(\textbf{5})(\textbf{3T})(\textbf{XnY\_Z})] + 1.912(\pm0.082)\ {}^{V}\textbf{ILEJ}_{S}[(\textbf{IS})(\textbf{4})(\textbf{T})] + 0.471(\pm0.029)\ {}^{V}\textbf{ILEM}_{K}[(\textbf{CB})(\textbf{6},\textbf{7})(\textbf{T})] + 1.925(\pm0.371)\ \textbf{SILEC}_{Q3}[(\textbf{IS})((\textbf{QD3})(\textbf{XYZ\_M})] + 0.048(\pm0.006)\textbf{SILEJ}_{M}[(\textbf{IS})((\textbf{QD1})]$$
$$N = 34\ \ R^2 = 0.991\ \ Q^2_{loo} = 0.987\ \ Q^2_{boot} = 0.973\ \ \text{SECV} = 0.148\ \ F = 612.89 \quad (21)$$

$$\textbf{Log K} = 6.099(\pm0.079) + 0.016(\pm0.002)\ {}^{V}\textbf{ILEM}_{Q2}[(\textbf{CB})(\textbf{4})(\textbf{T})] + 0.439(\pm0.021)\ {}^{V}\textbf{ILEM}_{K}[(\textbf{CB})(\textbf{6},\textbf{7})(\textbf{T})] - 0.413(\pm0.058)\textbf{SILEC}_{N3}[(\textbf{CB})((\textbf{QD3})(\textbf{XnYnZ\_M})] + 0.045(\pm0.004)\textbf{SILEJ}_{M}[(\textbf{IS})((\textbf{QD1})] + 0.075(\pm0.009)\ {}^{V}\textbf{IEM}_{CV}[(\textbf{5},\textbf{6})(\textbf{QD2})] + 1.530(\pm0.065)\ {}^{V}\textbf{SILEJ}_{S}[(\textbf{IS})(\textbf{4})(\textbf{QD1})]$$
$$N = 34\ \ R^2 = 0.996\ \ Q^2_{loo} = 0.993\ \ Q^2_{boot} = 0.980\ \ \text{SECV} = 0.105\ \ F = 1005.61 \quad (22)$$

Regression parameters obtained by these models are compared with those of some of the most relevant indices or group of indices in QSPR studies 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.[33]

It is evident in Table 4 that triple and quadruple matrix-based IFIs show better performance, in modeling the considered properties than the rest of the MD families reported in the literature. Moreover, the statistical parameters for mixed models, obtained with fewer variables, are comparable-to-better than those of the other MD families. Unfortunately, authors of previous studies did not report the values of Q^2_{boot} and in some cases for Log *K*, nor Q^2_{loo} values.[29,30]

The correlations presented by these IFIs in respect of the physico-chemical properties Log *P* and Log *K* for 2-furylethylene derivatives, can be considered to be statistically significant. It can therefore be postulated that the GT-STAF indices, in general, seem to be a valuable tool to reckon on QSAR/QSPR studies and diversity analysis. The descriptor values for all the variables in models (11)–(22) are available as Supporting Information (Excel data spread sheet SI6).

Conclusions

The primary goal of the present report was to introduce the concept of a hypermatrix and its subsequent application to previously defined IFIs, permitting the "generalization" of the mutual, conditional and joint entropy-based IFIs. Although duplex-matrix-based IFIs presented the best entropy distribution in the

Table 4. Statistical parameters of QSPR models that describe physicochemical properties of 34 derivatives of 2-furyletilenes using different MDs.

Indices	*N* (size)	R^2	SECV	Q^2loo	Q^2boot	*F*
Partition Coefficient 1-octanol-water (*Log P*)						
Duplex matrix-based IFIs[11]	7	0.988	0.090	0.980	0.975	292.45
Triple matrix-based IFIs	7	0.994	0.063	0.990	0.984	608.29
Quadruple matrix-based IFIs	7	0.991	0.078	0.987	0.983	396.70
All Set of Novel IFIs	6	0.989	0.082	0.984	0.965	415.37
All Set of Novel IFIs	5	0.979	0.112	0.967	0.964	262.90
All Set of Novel IFIs	4	0.955	0.162	0.941	0.933	153.55
All Set of Novel IFIs	3	0.915	0.219	0.889	0.884	107.28
Bond-based NS LI[32]	7	0.975	0.127	0.951	*	146.80
Vertex and edge Conn. Indices[32,33]	7	0.939	0.199	*		56.9
Topological Descriptors[32,33]	7	0.964	0.155	*	*	84.6
Quantum Chemical Descriptors[32,33]	Used Rogers and Cammarata approach	0.875	0.319	*	*	45.5
Atom-based NS QI[32]	7	0.969	0.142	0.951	*	116.76
Atom-based NS LI[32]	7	0.968	0.143	0.938	*	113.38
Reactivity (*Log K*)						
Duplex matrix-based IFIs[11]	7	0.994	0.120	0.991	0.988	661.58
Triple matrix-based IFIs	7	0.998	0.081	0.995	0.991	1453.20
Quadruple matrix-based IFIs	7	0.995	0.112	0.991	0.989	759.43
All Set of Novel IFIs	6	0.996	0.105	0.993	0.980	1005.61
All Set of Novel IFIs	5	0.991	0.148	0.987	0.973	612.89
All Set of Novel IFIs	4	0.988	0.166	0.984	0.969	605.57
All Set of Novel IFIs	3	0.977	0.227	0.969	0.946	424.13
Connectivity Indices[32,33]	7	0.821	0.681	*	*	17.1
Global spectral moments[32,33]	7	0.843	0.655	*	*	18.8
Local spectral moments[32,33]	7	0.964	0.320	*	*	70.4
Quantum Chemical Descriptor[32,33]	7	0.968	0.288	*	*	112.2
Bond-based NS LI[32]	7	0.994	0.119	0.980		
Atom-based NS QI[32]	7	0.968	0.285	0.922	*	115.14
Bond-based NS QI[32]	7	0.967	0.292	0.940	*	108.79
Bond-based SS QI[32]	7	0.975	0.257	0.958	*	142.07

variability analysis, it was remarkably evident that the use of hypermatrices introduces numerous high entropy MDs to the pool of the previously proposed IFIs. This finding was further backed by the improved performance presented by the hypermatrix-based IFIs in the modeling of the physico-chemical properties of the derivatives of furyethylenes, with triple based IFIs presenting the best statistical parameters.

Future Prospects

Regardless of the promising behavior presented by these IFIs, additional studies with wider and more diverse databases are indispensable, to evaluate the genuine possibilities of the proposed IFIs in real QSAR/QSPR problems. This will be the subject of future works.

In the present report, we present yet another matrix representation. In forthcoming articles, we will intend to apply the most significant classical algorithms reported in the literature to the frequency matrices and all the matrices derived thereof (mutual, conditional and joint information matrices).[40]

Although the legend put it that Shannon chose the term entropy simply because his colleague von Neumann advised him to use it, since "no one" knew it and could thus win any argument easily, this term, however, originates from statistical thermodynamics and is one of the thermodynamic functions used to characterize molecules. In respect to this, in the future, we intend to "go back to the roots" and use other thermodynamic functions like: enthalpy, Gibb's energy and heat capacity to define a whole new family of indices. The resulting indices will be later extended to define geometric (3D) aspects of molecules and their applicability in inorganic molecules, chemical complexes, proteins as well as DNA and RNA molecules will be studied. This phase will conclude with the generalization of these indices in complex networks.

Acknowledgments

Marrero-Ponce, Y. thanks the program 'Estades Temporals per an Investigadors Convidats' *for a fellowship to work at Valencia University in 2012. Finally, but not least, the authors want to express their acknowledgements to Prof. Jorge Galvez (VU) and Prof. Ramón García-Domenech (VU) for his help and useful comments about these new MDs.*

Keywords: relations frequency matrix · hypermatrix · information index · variability analysis · physico-chemical property · 2-furylethylene derivative · QSPR study

How to cite this article: S. J. Barigye, Y. Marrero-Ponce, Y. M. López, F. Torrens, L. M. A. Martínez, R. W. Pino-Urias, O. M. Santiago, *J. Comput. Chem.* **2012,** *00,* 000–000. DOI: 10.1002/jcc.23123

Additional Supporting Information may be found in the online version of this article.

[1] A. R. Katritzky, E. V. Gordeevat, *J. Chem. Inf. Comput. Sci.* **1993**, *33,* 835.
[2] J. Devillers, A. T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR; Gordon and Breach: Amsterdam, The Netherlands, **1999**.
[3] O. Ivanciuc, T. Ivancuic, M. V. Diudea, *SAR QSAR Environ. Res.* **1997**, *7,* 63.
[4] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics; Wiley-VCH: Weinheim, **2009**.
[5] C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27,* 379.
[6] H. Quastler, Information Theory in Biology; University of Illinois Press: Urbana IL, **1953**.
[7] T. M. Cover, J. A. Thomas, Elements of Information Theory; Wiley: New York, **1991**.
[8] V. I. Dmítriev, Teoría de Información Aplicada; Mir: Moscow, **1989**.
[9] E. Desurvire, Classical and Quantum Information Theory; Cambridge University Press, **2009**.
[10] M. A. Hall, In Department of Computer Science; University of Waikato: Hamilton, New Zealand, New York, **1999**; p 178.
[11] S. J. Barigye, Y. Marrero-Ponce, O. Martínez Santiago, Y. M. López, F. Torrens, *Curr. Comput. Aided Drug Des.* **2011**, accepted for publication.
[12] J. W. Godden, F. L. Stahura, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 796.
[13] J. W. Godden, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2002**, *42,* 87.
[14] V. A. Gorbátov, Fundamentos de la Matematica discreta; Mir: Moscow, URSS, **1988**.
[15] M. M. Deza, E. Deza, Encyclopedia of Distances, Springer-Verlag: Berlin, Heidelberg, **2009**.
[16] C. D. Cantrell, Modern Mathematical Methods for Physicists and Engneers; Cambridge University Press: Cambridge, **2000**.
[17] Y. Marrero-Ponce, Y. Martínez-López, S. J. Barigye, O. Martínez-Santiago, Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), **2010**.
[18] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, WILEY-VCH Verlag GmbH: D-69469 Weinheim, Federal Republic of Germany, **2000**.
[19] L. B. Kier, L. Hall, In Topological Indices and Related Descriptors in QSAR and QSPR; J. Devillers, A. T. Balaban, Eds.; Gordon and Breach Sci. Pub.: Amsterdam, **1999**; pp. 491–562.
[20] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, *J. Chem. Inf. Comput. Sci.* **2008**, *48,* 1337.
[21] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, *MATCH Commun. Math. Comput. Chem.* **2006**, *56*, 237.
[22] C. W. Yap, *J. Comput. Chem.* **2011**, *32,* 1466.
[23] E. Estrada, *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 844.
[24] E. Estrada, *J. Chem. Inf. Comput. Sci.* **1999**, *39,* 1042.
[25] E. Estrada, *SAR QSAR Environ. Res.* **2000**, *11,* 55.
[26] E. Estrada, E. Molina, *J. Mol. Graphics Model.* **2001**, *20,* 54.
[27] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, The Chemistry Development Kit (CDK):? An Open-Source Java Library for Chemo-and Bioinformatics, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493.
[28] M. Randic, *J, Math, Chem,* **1991**, *7,* 155.
[29] E. Estrada, E. Molina *J. Mol. Graphics Model.* **2001**, *20,* 54.
[30] S. Wold, L. Erikson, In Chemometric Methods in Molecular Design; H. van de Waterbeemd, Ed.; VCH Publishers: Weinheim, Germany, **1995**.
[31] S. Balaz, E. Sturdik, M. Rosenberg, J. Augustin, B. Skara, *J. Comput. Aided Mol. Des.* **1988**, *131,* 115.
[32] Y. Marrero Ponce, E. R. Martinez-Albelo, G. M. Casanola-Martin, J. A. Castillo Garit, Y. Echeveria Diaz, *Mol Divers.* **2010**, *14*, 731.
[33] E. Estrada, E. Molina, *J. Chem. Inf. Comput. Sci.* **2001**, *41,* 791.
[34] R. Todeschini, V. Consonni, A. Mauri, M. Pavan. In *Genetic Algorithms and Artificial Neural Networks*; Leardi, R., Ed.; Elsevier: Amsterdam, The Netherlands, **2003**, p 141.
[35] D. E. Goldberg, Addison-Wesley: Reading, MA, **1989**.
[36] D. Rogers, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34,* 854.
[37] P. Willet, *Trends Biotechnol.* **1995**, *13,* 516.
[38] S. S. So, M. Karplus, *J Med Chem* **1996**, *39,* 1521.
[39] S. S. So, M. Karplus, *J Med Chem* **1997**, *40,* 4347.
[40] R. Todeschini, V. Consonni, *MATCH Commun. Math. Comput. Chem.* **2010**, *64,* 359.

Received: 9 May 2012
Revised: 5 July 2012
Accepted: 22 August 2012
Published online on