SAR and QSAR in Environmental Research
Vol. 24, No. 1, January 2013, 3–34

Taylor & Francis
Taylor & Francis Group

Event-based criteria in GT-STAF information indices: theory, exploratory diversity analysis and QSPR applications

S.J. Barigye[a], Y. Marrero-Ponce[abc]*, Y. Martínez López[ad], O. Martínez Santiago[ae], F. Torrens[b], R. García Domenech[c] and J. Galvez[c]

[a]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Universidad Central "Marta Abreu" de Las Villas, Villa Clara, Cuba;* [b]*Institut Universitari de Ciència Molecular, Universitat de València, València, Spain;* [c]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, València, Spain;* [d]*Department of Computer Sciences, Faculty of Informatics, Camaguey University, Camaguey, Cuba;* [e]*Departament of Chemical Science, Faculty of Chemistry-Pharmacy, Universidad Central "Marta Abreu" de Las Villas, Villa Clara, Cuba*

(*Received 23 May 2012; in final form 26 August 2012*)

Versatile event-based approaches for the definition of novel information theory-based indices (IFIs) are presented. An event in this context is the criterion followed in the "discovery" of molecular substructures, which in turn serve as basis for the construction of the generalized incidence and relations frequency matrices, **Q** and **F**, respectively. From the resultant **F**, Shannon's, mutual, conditional and joint entropy-based IFIs are computed. In previous reports, an event named *connected subgraphs* was presented. The present study is an extension of this notion, in which we introduce other events, namely: terminal paths, vertex path incidence, *quantum* subgraphs, walks of length k, Sach's subgraphs, MACCs, E-state and substructure fingerprints and, finally, Ghose and Crippen atom-types for hydrophobicity and refractivity. Moreover, we define magnitude-based IFIs, introducing the use of the magnitude criterion in the definition of mutual, conditional and joint entropy-based IFIs. We also discuss the use of information-theoretic parameters as a measure of the dissimilarity of codified structural information of molecules. Finally, a comparison of the statistics for QSPR models obtained with the proposed IFIs and DRAGON's molecular descriptors for two physicochemical properties $\log P$ and $\log K$ of 34 derivatives of 2-furylethylenes demonstrates similar to better predictive ability than the latter.

Keywords: event; Shannon's, mutual; conditional and joint entropy; magnitude criterion; GT-STAF; TOMOCOMD-CARDD; cluster analysis; QSPR

*Corresponding author. Email: ymarrero77@yahoo.es

ISSN 1062–936X print/ISSN 1029–046X online

http://dx.doi.org/10.1080/1062936X.2012.728994
http://www.tandfonline.com

''There are no restrictions on the design of structural invariants; the limiting factor is one's own imagination.''

M. Randić (1996)

1. Introduction

It has long been known that there exists an intimate connection between the structure of a molecule and its physical, chemical and biological properties [1,2]. However, profitable analysis of this relationship requires ''*condensation*'' of molecular structure information into some kind of numeric representation to facilitate structure–property correlation evaluations. This constitutes one of the primary objectives of mathematical chemistry; that is, the development of simple numeric representations for molecular skeletons, also known as *theoretical molecular descriptors* [3–5]. Although in recent years the number of molecular descriptors (MDs) has increased considerably, to more than 3000 MDs [3], the search for new MDs continues, with the hope of increasing diversity and potentially codifying structural information not adequately captured by the existing MD pool. Indeed, it is unrealistic to infer that a particular MD (or group of MDs) could satisfactorily describe all chemical, physical, physicochemical and biological activities of molecules. The lack of a sole numeric model that *completely* represents a molecular structure is reflected in a statement by Weiner: ''The best model of a cat is another cat or, better, the cat itself.'' [6]. This statement poses an intricate dilemma to mathematical chemists. How best can we approximate to the *complete* description of a molecular structure without need of another identical molecular structure? The widely accepted approach by chemometricians for this challenge is a *case-wise* analysis; that is, to form models sensitive to the structural features of interest for a specific case of study. Such an approach calls for the creation of a widely diverse collection of MDs, to rely on as a source of mathematical tools for different chemometric studies. It is, hence, natural that the need for the definition of novel MDs will continue deserving the attention of researchers. Basically, this goal constitutes the extrapolation of different concepts or theories from different fields, such as quantum chemistry, information theory, organic chemistry, graph theory, algebra, physics, and so on; and applying them to problems related to the codification of molecular structures; or redefining – or probably more precisely, generalizing – already existing concepts to improve their performance.

In initial reports [7,8], we introduced a series of novel information indices (IFIs), derived from a *connected subgraphs* event-based approach, applying the concepts of Shannon's, mutual, conditional and joint entropy over duplex, triple and quadruple frequency relations, condensed in respective matrix representations. The present work is aimed at extending the notion of *events*, in order to broaden the perspective towards a molecular graph (**G**). This in turn allows us to generate relations frequency matrices from different points of view, with the hope of codifying different structural information, if any, of a **G**.

2. Theoretical approach

2.1 *Overview*

An *event*, in simple terms, refers to the criterion from a graph-theoretic, chemical and physicochemical point of view, followed in the generation of subgraphs (or substructure, atom-type) from a given **G**. The set of the generated subgraphs subsequently serves as basis for the construction of the relations frequency matrices, through the computation of

the participation frequencies of vertices that constitute **G** in the formation of the corresponding set of subgraphs.

Before we proceed, let us have a brief mathematical recapitulation of the concepts introduced in the previous articles, critical for the comprehension of the definitions and notations presented in this report.

To begin with, we define an *event* **E**, which is true when certain conditions of an examined procedure are fulfilled. The *event* **E** determines a bi-dimensional matrix $\mathbf{Q}=[q_{ij}]_{m\times n}$, each column of which corresponds reciprocally to a *condition*, true in the event, and every row, to a collection of conditions, in which the *event* occurs (in which the *event* **E** is true) and in a Boolean matrix representation, q_{ij} is equal to [9]:

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

Accordingly, every *event* defines a model for the incidence matrix **Q**; the conditions included in the event are the letters corresponding to the model, and the collection of conditions true in the *event* would be the words for the model. In other words, each *event* serves to establish the set of words that a model will have. For example, an event could be "*the words of the English language related with nature*".

Subsequently, we introduce the *relation frequency matrix* $\mathbf{F}=[f_{ij}]_{n\times n}$ to characterize the model ψ, with the incidence matrix $\mathbf{Q}(\psi)=[q_{ij}]_{m\times n}$ [9]. 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 $i\neq j$. On the other hand, 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 the **F**, one notices that it is symmetric with respect to the principal diagonal, that is $f_{ij}=f_{ji}$, and the individual frequency of each letter is greater than (or equal to) the reciprocal frequency of this letter with any other letter, $f_i\geq f_{ij}$. It can also be demonstrated that: $\mathbf{F}=\mathbf{Q}^{\mathrm{T}}\times\mathbf{Q}$, $\mathbf{Q}^{\mathrm{T}}$ being the transpose matrix of an "incidence" matrix [**Q** (ψ)] for the model ψ [9].

One could also arrive at this *relation frequency matrix* using a simple exploratory method. Let us consider a model where we have nine words of the English language related with nature in which no letter is repeated: **EARTH ORANGE MINERAL RIVER MOUNTAIN OCEAN STREAM VALLEY PEARL BEACH WATER**.

Our interest in this case is to find the number of times (frequency) that a subset of two letters participates in the formation of the same word (duplex participation frequency). If we look at letters {A, E} for example, these simultaneously participate in the formation of the words: EARTH, ORANGE, MINERAL, OCEAN, STREAM, VALLEY, PEARL, BEACH and WATER, i.e. 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** [7].

This analysis could be furthered to explore the number of times that subsets of 3, 4, 5, 6, 7 . . . *n* letters participate in the same word. In this paper, we will solely utilize duplex participation frequencies.

2.2 *Event-based criteria in graphs theory and cheminformatics*

Having introduced these concepts and procedure in the preceding section, ground is provided for their application to chemical graph theory in the codification of molecular structures. We will discuss 10 new events, some entirely graph-theoretical while others incorporate organic chemistry concepts and physicochemical considerations. However, let us first go over the procedure followed in the definition of event previously introduced, which we will consider as the first event.

2.2.1 *Connected subgraphs (CS)*

The criterion followed in this event to generate subgraphs for a given **G** is *connectivity*, and thus the event is denominated connected subgraphs. Accordingly, the conditions (letters of the model) included in the event are the vertices (atom-nuclei) present in each collection of conditions (connected subgraphs, which are the words of the model).

Let us take the molecule of 4-methylcyclobuta-1,3-dienamine as a simple example (see Figure 1a), where the numbers correspond to the labels that are assigned to the carbon atoms (vertices) in the molecular structure.

First we obtain all possible connected subgraphs of different orders from **G** based on the atomic relations.

Order 0 C_1, C_2, C_3, C_4, N_5, C_6

Order 1 C_1-C_2, C_1-C_4, C_1-C_6, C_2-C_3, C_3-C_4, C_4-N_5

Order 2 C_1-C_2-C_3, C_4-C_1-C_2, C_1-C_2-C_6, C_3-C_4-C_1, C_1-C_4-C_3, C_1-C_4-N_5, C_4-C_1-C_6, C_2-C_3-C_4, C_3-C_4-N_5

Order 3 C_1-C_2-C_3-C_4, C_2-C_3-C_4-N_5, C_3-C_4-C_5-C_1, N_5-C_4-C_1-C_6, N_5-C_4-C_1-C_2, C_4-C_1-C_2-C_6, C_6-C_1-C_2-C_4, C_6-C_1-C_2-C_3, C_6-C_1-C_4-C_3

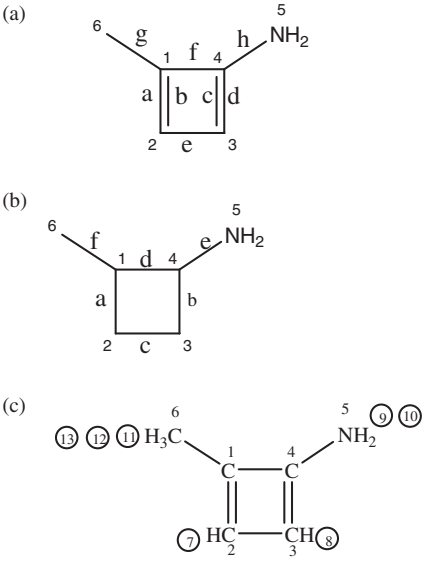

Figure 1. (a) The chemical structure for the molecule of 4-methylcyclobuta-1,3-dienamine; (b) The molecular graph for 1,2-dimethylcyclobutane; and (c) H-filled molecular graph for the molecule of 4-methylcyclobuta-1,3-dienamine.

Order 4 C_1-C_2-C_3-C_4-N_5, C_6-C_1-C_2-C_3-C_4, C_6-C_1-C_4-N_5-C_2, C_6-C_1-C_4-N_5-C_3
Order 5 C_6-C_1-C_2-C_3-C_4-N_5

These connected subgraphs are subsequently represented in an *incidence matrix* ($\mathbf{Q}_{CS}$), from which we obtain the corresponding *relation frequency matrix* ($\mathbf{F}_{CS}$) as shown below.

$$Q_{CS} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}
\qquad
F_{CS} = \begin{bmatrix}
22 & 12 & 10 & 16 & 8 & 11 \\
12 & 16 & 9 & 10 & 5 & 6 \\
10 & 9 & 16 & 12 & 6 & 5 \\
16 & 10 & 12 & 22 & 11 & 8 \\
8 & 5 & 6 & 11 & 12 & 4 \\
11 & 6 & 5 & 8 & 4 & 12
\end{bmatrix}$$

The number of inclusions of each vertex in the carbon skeletons permits us to establish the required frequencies. For example, vertex (letter) 1 participates 22 times in the formation of the subgraphs (words).

The graph-theoretical concepts of subgraph orders and types, namely: path (*p*), cluster (*c*) and path-cluster (*pc*), according to Kier–Hall nomenclature [3,10–12], could be employed as specific criteria to generate the connected subgraphs. A particular case where only subgraphs of *order 1* (pairs of vertices or edges in **G**) are considered, the formed matrix **Q** coincides with the common incidence matrix used in graph theory.

2.2.2 *Quantum (Q)*

The *quantum* event is based on the removal of the edges joining vertices *vi* and *vj* from **G**, with replacement. The use of the term "*quantum*" is not remotely related to quantum theory. It is simply chosen in reference to the removal of discrete units (i.e. edges) from **G**, analogous to the classical postulation by Ludwig Boltzmann of the discrete nature of energy states of a physical system. The resulting subgraphs are used to construct the incidence matrix. This is a successive procedure which could be performed for 1 up to $n-1$ edges. In the conception of this event, we however opted to only work with the removal of to six edges as maximum, taking into account the computational cost that would arise in the implementation of an event of this nature. This event is defined for simple graphs (graphs that do not take into account multiples, neither bonds nor heteroatoms). In effect, we will use for illustrative purposes the molecule of 1,2-dimethylcyclobutane, an isomorph of 4-methylcyclobuta-1,3-dienamine (see Figure 1b). Let us denominate the edges (C_1-C_2), (C_2-C_3), (C_3-C_4), (C_4-C_1), (C_4-N_5), (C_1-C_6) as ***a***, ***b***, ***c***, ***d***, ***e*** and ***f*** respectively. Table 1 shows the subgraphs formed upon the removal of up to three edges from **G**.

Accordingly, these subgraphs represented in Table 1 are used to construct the quantum incidence matrix ($\mathbf{Q}_Q$), which in turn gives the corresponding quantum frequency matrix ($\mathbf{Q}_F$).

As in the case of the connected subgraphs events, we may wish to compute quantum incidence matrices with reduced dimensions, in the sense that not all the generated subgraphs are considered. As in the preceding event, the graph-theoretical concepts of subgraph orders and types, namely: path (*p*), cluster (*cl*), path-cluster (*pc*) and cycles (*c*), according to Kier–Hall nomenclature, are employed as criteria to generate these particular subgraphs.

2.2.3 *Terminal paths (T)*

This is a peculiar event which traces *i-j* path subgraph types in a given **G** that fulfil the condition that the terminal vertices (v_i and v_j) have a valence vertex degree (δ) of one [13]. In the case of the molecule of 4-methylcyclobuta-1,3-dienamine as a simple example (see Figure 1a), the terminal paths present in this **G** are: N_5-C_4-C_1-C_6 and C_6-C_1-C_2-C_3-C_4-N_5. These terminal paths subgraphs consequently constitute the respective incidence ($\mathbf{Q}_T$) and the relations frequency ($\mathbf{F}_T$) matrices.

2.2.4 *Vertex path incidence (**VP**)*

This event is derived from the definition proposed by Janezic et al. of a vertex-path incidence matrix, **VP** [14]. It follows that given that **V** is the set of vertices $\{v_i\}$ and **P** the set

Table 1. Subgraphs formed upon the removal of edges from the **G** of 1,2-dimethylcyclobutane.

Removed edge	*Subgraphs*	*Removed edge*	*Subgraphs*
a	C_2-C_3-C_4-C_1-C_6-N_5	abc	C_2, C_3, N_5-C_4-C_1-C_6
b	N_5-C_4-C_1-C_2-C_3-C_6	abd	C_2-C_3, C_1-C_6,C_4-N_5
c	C_6-C_1-C_2-C_4-C_3-N_5	abe	C_2-C_3, C_6, C_1-C_4-N_5
d	C_6-C_1-C_2-C_3-C_4-N_5	abf	C_2-C_3, N_5, C_6-C_1-C_4
e	C_6, N_5-C_4-C_1-C_2-C_3-C_4	acd	C_2, C_1-C_6, C_3-C_4-N_5
f	N_5, C_6-C_1-C_2-C_3-C_4-C_1	ace	C_2, C_6, C_1-C_4-C_3-N_5
ab	C_2-C_3, N_5-C_4-C_1-C_6	acf	C_2, N_5, C_3-C_4-C_1-C_6
ac	C_2,C_6-C_1-C_4-C_3-N_5	ade	C_6, C_1, C_2-C_3-C_4-N_5
ad	C_1-C_6, C_2-C_3-C_4-N_5	aef	C_6, N_5, C_1-C_4-C_3-C_2
ae	C_6, C_2-C_3-C_4-C_1-N_5	adf	C_6-C_1, N_5, C_2-C_3-C_4
af	N_5, C_2-C_3-C_4-C_1-C_6	bcd	C_3, C_4-N_5, C_6-C_1-C_2
bc	C_3, C_6-C_1-C_2-C_4-N_5	bce	C_3, C_6, N_5-C_4-C_1-C_2
bd	C_4-N_5,C_6-C_1-C_2-C_3	bcf	C_3, N_5, C_6-C_1-C_2-C_4
be	C_6, N_5-C_4-C_1-C_2-C_3	bde	C_6, C_4-N_5,C_1-C_2-C_3
bf	N_5, C_3-C_2-C_1-C_4-C_6	bdf	C_4, N_5, C_6-C_1-C_2-C_3
cd	C_6-C_1-C_2, C_3-C_4-N_5	bef	C_6, N_5, C_4-C_1-C_2-C_3
ce	C_6, C_2-C_1-C_4-C_3-N_5	cde	C_6, C_1-C_2, C_3-C_4-N_5
cf	N_5, C_3-C_4-C_1-C_2-C_6	cdf	C_6-C_1-C_2, N_5, C_3-C_4
de	C_6, C_1-C_2-C_3-C_4-N_5	cef	C_6, N_5, C_2-C_1-C_4-C_3
df	N_5, C_6-C_1-C_2-C_3-C_4	def	C_6, N_5, C_1-C_2-C_3-C_4
ef	C_6, N_5, C_1-C_2-C_3-C_4-C_1		

of paths $\{p_j\}$, then the VP incidence matrix is defined as

$$[\mathbf{VP}]_{ij} = \begin{cases} n(i,j) \text{ if } \{v_i\} \text{ and } \{p_j\} & \text{have non} - \text{zero intersections} \\ 0 & \text{otherwise} \end{cases}$$

where $n(i,j)$ is the number of incidences between two sets $\{v_i\}$ and $\{p_j\}$.

Note that in order to maintain the same scheme followed in the definitions presented so far, we work with the transpose of **VP**, i.e. $\mathbf{VP}^{\mathrm{T}}$, represented as **VP***, in the sense that the row entries represent the subgraphs participation intensities and the column entries the vertex participation intensities, contrary to the initial matrix definition presented by Janezic et al.

This approach is quite different from the particularity presented in the connected subgraphs incidence matrix approach in which vertex paths are exclusively considered in that, in this case, the row entries follow the computation of the number of times that vertex (v_i) is included vertex paths of a particular order, while in the former the row entries are simple Boolean representations of the participation of the vertices in a given subgraph. As in the previous event, let us use the molecule of 4-methylcyclobuta-1,3-dienamine as an example. The corresponding incidence ($\mathbf{Q}_{\mathrm{VP}*}$) and the relations frequency ($\mathbf{F}_{\mathrm{VP}*}$) matrices are shown below:

$$Q_{VP*} = \begin{array}{c} \\ \mathrm{p0} \\ \mathrm{p1} \\ \mathrm{p2} \\ \mathrm{p3} \\ \mathrm{p4} \\ \mathrm{p5} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 2 & 2 & 3 & 1 & 1 \\ 3 & 3 & 3 & 3 & 2 & 2 \\ 1 & 2 & 2 & 2 & 3 & 3 \\ 1 & 1 & 1 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \qquad F_{VP*} = \begin{bmatrix} 21 & 19 & 19 & 22 & 15 & 15 \\ 19 & 19 & 19 & 21 & 17 & 17 \\ 19 & 19 & 19 & 21 & 17 & 17 \\ 22 & 21 & 21 & 24 & 18 & 18 \\ 15 & 17 & 17 & 18 & 20 & 20 \\ 15 & 17 & 17 & 18 & 20 & 20 \end{bmatrix}$$

2.2.5 *Walks of length k (K)*

This event arises from the exploration of walks of length k in a given **G**. However, in a **G** with many vertices, walks of this kind are exceptionally numerous and could give rise to redundancy. We thus deemed it necessary to set the maximum walks order, $m_{\max}$, of the generated walks at 10. Let us continue with the same example of the molecule of 4-methylcyclobuta-1,3-dienamine. We will use walks of orders 0, 1 and 2, only, for convenience. In this example we take into account self-avoiding as well as self-returning walks. For example in case of order 2, there exist four self-avoiding walks between *v1* and *v3* (i.e. *ae*, *be*, *fd*, *fc*) and six self-returning walks to *v1* (see Table 2).

As can be seen in Table 2, this event presents a peculiar characteristic in that the walk counts are thermodynamic in nature rather than kinetic, in the sense that only the initial and final states (vertices) are taken into account. The incidence matrix for the walks of length k event ($\mathbf{Q}_W$) and subsequent the frequency matrix ($\mathbf{F}_W$) for walks of lengths k 0, 1 and 2 are as shown below.

$$Q_W = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 1-2 \\ 2-3 \\ 3-4 \\ 1-4 \\ 1-6 \\ 4-5 \\ 1-3 \\ 1-5 \\ 1-1 \\ 2-4 \\ 2-6 \\ 2-2 \\ 3-3 \\ 3-5 \\ 4-4 \\ 4-6 \\ 5-5 \\ 6-6 \end{array} \begin{bmatrix} v1 & v2 & v3 & v4 & v5 & v6 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 4 & 0 & 4 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 4 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 2 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$F_W = \begin{bmatrix} 60 & 4 & 16 & 1 & 1 & 1 \\ 4 & 51 & 1 & 16 & 0 & 4 \\ 16 & 1 & 51 & 4 & 4 & 0 \\ 1 & 16 & 4 & 60 & 1 & 1 \\ 1 & 0 & 4 & 1 & 8 & 0 \\ 1 & 4 & 0 & 1 & 0 & 8 \end{bmatrix}$$

Table 2. Walks of length k for the **G** of 4-methylcyclobuta-1,3-dienamine.

Orders	*Walk*	*Type*
0	C_1	
	C_2	
	C_3	
	C_4	
	N_5	
	C_6	
1	a, b	
	e	
	c, d	
	f	
	g	
	h	
2	ae, be, fd, fc	self-avoiding
	fh	self-avoiding
	gg, ff, aa, ab, bb, ba	self-returning
	af, bf, ed, ec	self-avoiding
	ag, bg	self-avoiding
	aa, bb, ab, ba, ee	self-returning
	ee, cc, dd, cd, dc	self-returning
	ch, dh	self-avoiding
	ff, hh, cc, dd, cd, dc	self-returning
	fg	self-avoiding
	hh	self-returning
	gg	self-returning

2.2.6 *Sach's subgraphs (S)*

Since the introduction of the renowned Sach's theorem by the German mathematician Horst Sach in 1964, it has had a wide range of applications in theoretical chemistry [15–19]. This is yet another application, quite different, as it is neither related to the reconstruction of structural graphs nor the analysis of the relationships between the spectral and structural properties of graphs. Our interest lies in the use of the subgraphs generated using Sach's theorem as criterion to construct generalized incidence matrices and their corresponding frequency matrices. In this criterion **G** is ''partitioned'' into subgraphs, using the following considerations [16,17]:

(a) isolated subgraphs constituted of two vertices connected by an edge;
(b) isolated cycles or ring-type components (with multiplicity $m \geq 3$)

Using these considerations, sets of subgraphs (S_k) are constructed, where k indicates the number of vertices that constitute a given subgraph. These subgraphs, constructed from considerations of no other than isolated edges and/or isolated cycles are called ''Sachs' graphs''. Let us use the same illustrative example of the **G** of 4-methylcyclobuta-1,3-dienamine. Figure 2 shows the Sachs' subgraphs constructed from **G**.

The Sachs' graphs are drawn within parentheses and the set of all Sachs graphs S_k are denoted in brackets. The obtained subgraphs are used to construct the

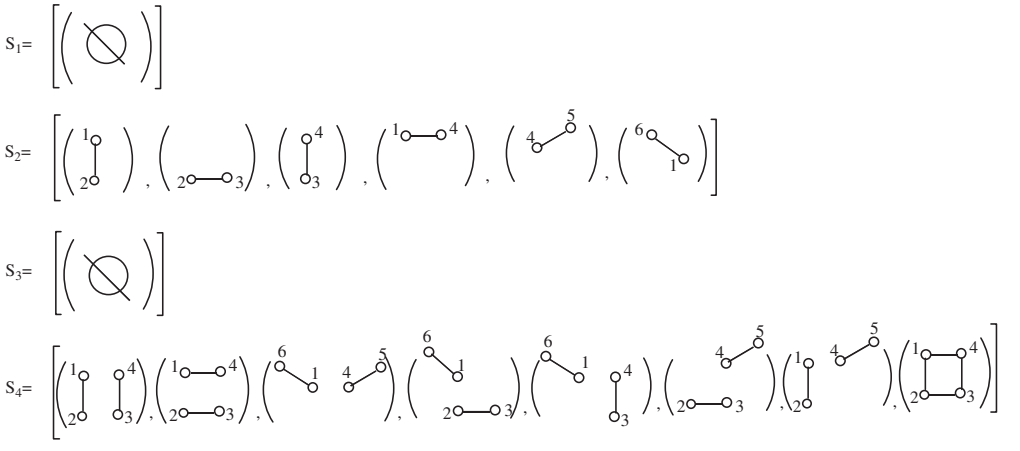

Figure 2. Sachs' subgraphs of the **G** for 4-methylcyclobuta-1,3-dienamine.

generalized incidence matrix, $\mathbf{Q}_s$, from which the corresponding frequency matrix, $\mathbf{F}_s$, is later obtained.

$$Q_S = \begin{bmatrix} v1 & v2 & v3 & v4 & v5 & v6 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \quad F_S = \begin{bmatrix} 10 & 6 & 5 & 7 & 2 & 4 \\ 6 & 8 & 6 & 5 & 2 & 1 \\ 5 & 6 & 8 & 6 & 1 & 2 \\ 7 & 5 & 6 & 10 & 4 & 2 \\ 2 & 2 & 1 & 4 & 4 & 1 \\ 4 & 1 & 2 & 2 & 1 & 4 \end{bmatrix}$$

Sach's theorem can be extended to analyse heteroatomic and multiple bond systems. In order to achieve this, the heteroatoms are depicted by the use of a self-loop representation, according to graphical method proposed by Mallion et al. [20]. In this way, the self-loop can be considered as a hypothetical edge connecting a *real* and an *imaginary* vertex. The loop represents a lone pair of electrons, where if X has more than one lone pair, we have more than one loop. The self-loop representation has been previously used in extensions of Sach's theorem to heteroconjugated systems but with different considerations that will not be used in this approximation [19]. In the case of multiple bond

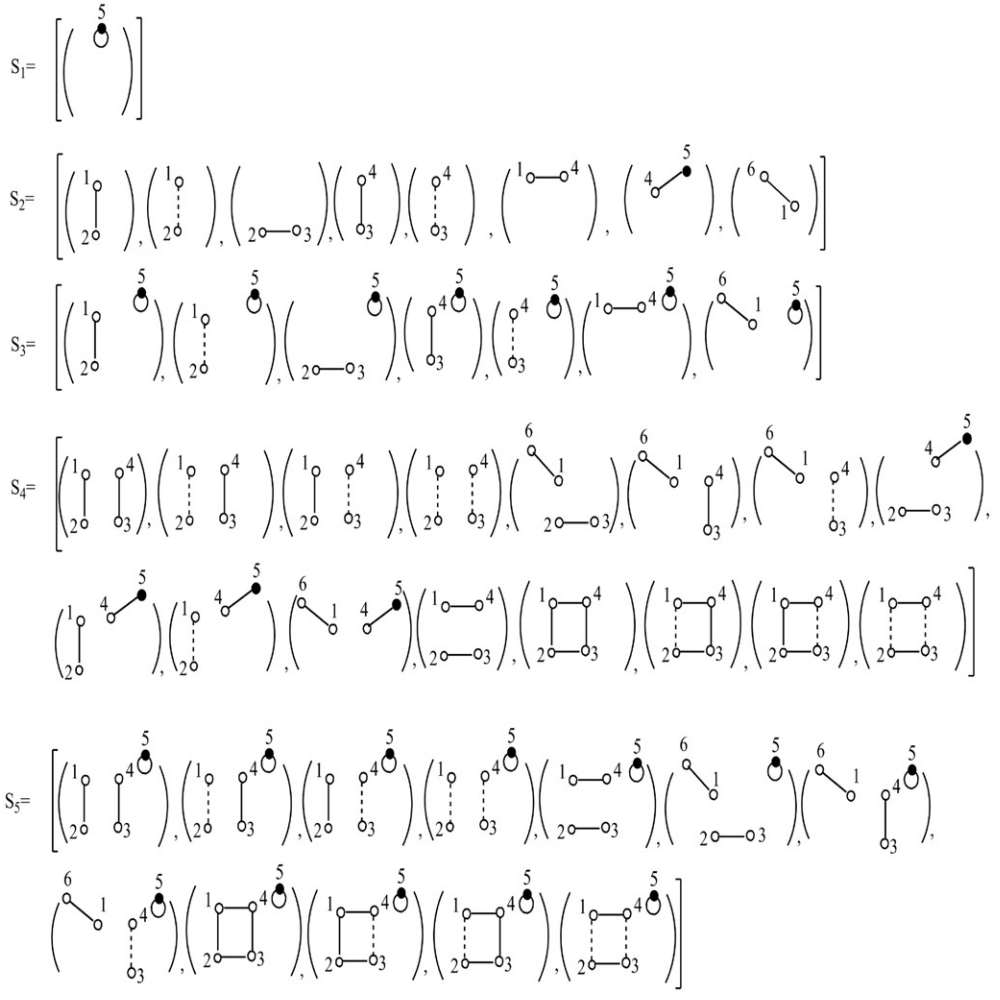

Figure 3. Sachs' subgraphs of the **G** for 4-methylcyclobuta-1,3-dienamine with heteroatomic and multiple bond considerations.

systems, a π bond is considered to constitute an imaginary edge between vertices v_i and v_j, in the sense that a multiple bond is composed of a real and an imaginary edge(s). This approximation is derived from considerations used in determining chiral centre configurations in stereochemistry, according to Cahn et al. [21], and could be viewed as a supplement to Sach's theorem. Figure 3 shows what the resulting subgraph sets in this case would be.

The computational cost incurred in the implementation of such an event dictates that we work with sets S_k with maximum k value of eight.

2.2.7 *Fingerprints*

In simple terms, a chemical fingerprint is a list of binary values (known as a bit string), which respond to the query about the presence (or no) of given features (i.e. atom types or

fragments) in a molecular structure. There exist various fingerprints, most of which are implemented in Structural Chemometric and Bioinformatic libraries such as CDK [22,23], Joelib [24], RDKit [25], etc. Our intention in this study is not to use these fingerprints for the classical similarity comparisons and database filtering, but rather as criteria for introducing new events, and from these we intend to select three representative fingerprints, namely: MACCS, E-state and Substructure fingerprints. This approach differs from all the previously presented events in the sense that, contrary to all the previous events which commence with a search of the possible sub-structures according to the criterion in question, here we rely on a fixed number of substructures and the exploration is aimed at verifying their existence or not in a given **G**.

2.2.7.1. *MACCS fingerprints (MA)*. Various keyset lengths for MACCS fingerprints have been reported in the literature [26,27]. Among these the most popular are the 960 bit and 166 bit keyset lengths based on 2D descriptors. In the present report we use the 166 bit keyset lengths. To benefit from the MACCS fingerprints query, it is crucial to adapt the algorithm used in encoding the substructure occurrences to suit our purpose. Accordingly, in place of a bit string, we employ a vector constituted of numbers that indicate the positions of the activated prints in the search performed. These properties or substructures are subsequently identified and used to construct the incidence matrix. As in the previous events we will use an illustrative example of the **G** of 4-methylcyclobuta-1,3-dienamine.

The positions of MACCS fingerprints activated in the 166 bit keyset for this molecule as well as their respective SMART patterns are shown in the Table 3.

2.2.7.2. *E-state fingerprints (ES)*. These fingerprints are derived from a count of the electrotopological state fragments [10–12,28,29]. A query of the E-state fingerprints yields a 79 bit keyset length. As in the case of MACCS fingerprints, our interest is to identify the positions of the activated fragments, which in turn serve to construct the vector of fragments, used for the incidence matrix. In the case of the molecule of 4-methylcyclobuta-1,3-dienamine, a one-component vector is formed, i.e. $\text{E-state}_{\textit{4-methylcyclobuta-1,3-dienamine}} = [15]$, represented as [CD3H0](=*)(−*)−*.

2.2.7.3. *Substructure fingerprints (SS)*. The substructure fingerprints, comprising a set of a 307 bit keyset, is rather peculiar in the sense that it entails substructures representative of practically all functional groups, organic and inorganic, known in molecular medicinal chemistry, contrary to the MACCS and E-state fingerprints. It is not surprising therefore that these fingerprints are the most understandable and interpretable in organic chemistry terms. Table 4 shows the substructure fingerprints activated for the molecule of 4-methylcyclobuta-1,3-dienamine.

Truly, it is evident from Table 4 that a more comprehensible structural corroboration of the activated fingerprints can be traced in the **G**. For example, a quick analysis of the molecule of 4-methylcyclobuta-1,3-dienamine ascertains the presence of conjugated double bonds, a C-N bond, in addition to axial chirality. The sensitivity to chirality confers to this event an exceptionally beneficial attribute, in the sense that it would permit discrimination between stereoisomers. The substructure incidence matrix for the molecule of

Table 3. MACCS fingerprints activated for the molecule of 4-methylcyclobuta-1,3-dienamine.

Position	*Fingerprint (SMART)*
10	[Be,Mg,Ca,Sr,Ba,Ra] 0 |Group IIa (Alkaline earth)
21	[#6]=[#6](~[!#6;!#1])~[!#6;!#1] 0 |C=C(Q)Q
44	? 0 |OTHER
49	[!+0] 0 |CHARGE
75	*!@[#7]@* 0 |A!N$A
98	[!#6;!#1]1~*~*~*~*~*~*~1 0 |QAAAAA@1
132	[#8]~*~[CH2]~*',0), # OACH2A
149	[C;H3,H4] 1 |CH3 > 1
155	*!@[CH2]!@* 0 |A!CH2!A
157	[#6]–[#8] 0 |C–O
160	[C;H3,H4] 0 |CH3
164	[#8] 0 |O

Table 4. Substructure fingerprints activated for the molecule of 4-methylcyclobuta-1,3-dienamine.

Position	*Name*	*Fingerprint (SMART)*
286	Conjugated double bond	*=*[*]=,#,:[*]
294	C ONS bond	[#6]~[#7,#8,#16]
306	Chiral centre specified	[$([*@](~*)(~*)(*)*),$([*@H](*)(*)*),$([*@](~*)(*)*), $([*@H](~*)~*)]

4-methylcyclobuta-1,3-dienamine and the respective frequency matrix are presented below (see also Table 4).

$$Q_{SS} = \begin{bmatrix} v1 & v2 & v3 & v4 & v5 & v6 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad F_{SS} = \begin{bmatrix} 2 & 1 & 1 & 2 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Note that in structural terms, the substructure fingerprint *chiral center specified* **(306)** is represented in the incidence matrix ($\mathbf{Q}_{SS}$) by the vertices that confer axial chirality to the molecular structure, in this case being the amine (NH_2) and methyl ($–CH_3$) groups in positions 5 and 6, respectively (see row 3 entries of matrix $\mathbf{Q}_{SS}$).

2.2.8 *Atomic refractivity (AMR) and hydrophobicity (ALOGP)*

This event follows a unique approach, different from all the preceding ones, as it is related to determined atomic physicochemical properties. The conceptual structure of this event proposes an evaluation of the contribution of the chemical entities (or groups) in the

vicinity of a vertex on its atomic molar refractivity (MR) and hydrophobicity ($\log P$), the latter expressed in terms of partition coefficient values. To achieve this objective, Ghose and Crippen introduce a set of 110 atom types representing commonly occurring atomic states of carbon, hydrogen, oxygen, nitrogen, halogens, and sulphur in organic molecules [30]. Accordingly, the construction of the incidence matrix follows an exploration of the atom type most representative of the vicinity of a given vertex neighbourhood. The resulting incidence matrix is defined as:

$$[\mathbf{Q}]_{ij} = \begin{cases} \log P \text{ or MR}, & \text{if there exists an atom} - \text{type analogous to the examined vertex} \\ 0 & \text{otherwise} \end{cases}$$

Let us present an illustration of this event, using the molecule of 4-methylcyclobuta-1,3-dienamine. Note that contrary to all the previously presented events, this event uses H-filled molecular graphs (see Figure 1c).

The atom types activated for each vertex in this molecular graph and their respective positions in the Ghose–Crippen container of atomic contributions to $\log P$ and molar refractivity (MR) are: C_1 (17); C_2, C_3 (16); C_4 (19); N_5 (66); C_6 (1); H_7, H_8 (47); H_9, H_{10} (50); H_{11}, H_{12}, H_{13} (46). These are then used to construct the incidence matrix, $\mathbf{Q}^M_{MR(\text{or } AP)}$, from which the corresponding frequency matrix, $\mathbf{F}^M_{MR(\text{or } AP)}$, is obtained as shown below, taking MR as example:

$$Q^M_{MR} = \begin{array}{c} \\ 1 \\ 16 \\ 17 \\ 19 \\ 46 \\ 47 \\ 50 \\ 66 \end{array} \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & N_5 & C_6 & H_7 & H_8 & H_9 & H_{10} & H_{11} & H_{12} & H_{13} \\ 0 & 0 & 0 & 0 & 0 & 2.968 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4.265 & 4.265 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3.939 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.487 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.845 & 0.845 & 0.845 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.894 & 0.894 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.800 & 0.800 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.622 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$F^M_{MR} = \begin{bmatrix} 15.516 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 18.190 & 18.190 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 18.190 & 18.190 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20.133 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6.875 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8.809 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.799 & 0.799 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.799 & 0.799 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.640 & 0.640 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.640 & 0.640 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.714 & 0.714 & 0.714 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.714 & 0.714 & 0.714 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.714 & 0.714 & 0.714 \end{bmatrix}$$

3. Frequency-based Shannon's, mutual, conditional and joint IFIs

The relations frequency matrices discussed in Sections 2.2.1–2.2.8 serve as the basis for the computation of Shannon's, mutual, conditional and joint entropy-based information indices. The theory of these IFIs has been extensively explained elsewhere [7,8]. However, the concepts and formalism followed in the computation of these indices is provided as supporting information (SI1) available via the Supplementary Content tab on the article's online page at http://dx.doi.org/

4. Magnitude-based Shannon's, mutual, conditional and joint IFIs

The refractivity and hydrophobicity event-based approach (discussed in Section 2.2.8) paves the way to the prospect of definition of magnitude-based Shannon's, mutual, conditional and joint IFIs, using the non-Boolean matrix approach. The equivalence classes in this case are obtained according to the magnitude criterion [3,31,32], which postulates that each element is considered as an equivalence class whose cardinality, that is, number of elements, is equal to the magnitude of the element.

4.1 *Magnitude-based total information content, Shannon's entropy and standardized information content*

These IFIs are defined using the diagonal elements of the refractivity and hydrophobicity event-based frequency matrices. The general formulae for the calculation of magnitude-based total information content (negentropy), mean information content (Shannon's entropy) and standardized Shannon's entropy are the following:

$$I_m = \text{Total magnitude} \cdot \log_2 \text{Total magnitude} - \sum_{i=1}^{n} f \cdot \text{magnitude}_i \cdot \log_2 \text{magnitude}_i \quad (1)$$

$$\bar{I}_m = -\sum_{i=1}^{n} f \cdot \frac{\text{magnitude}_i}{\text{Total magnitude}} \cdot \log_2 \frac{\text{magnitude}_i}{\text{Total magnitude}} \quad (2)$$

$$I_m^* = \frac{\bar{I}_m}{\log_2 \text{Total magnitude}} \quad (3)$$

This magnitude could be graph theoretical, such as vertex distance, vertex degree, edge degree, edge cyclic degree, among others, or a given atomic physical or physicochemical property such as the atomic mass, refractivity and hydrophobicity. The frequency matrix values computed for the refractivity and hydrophobicity event are a product of the vertex frequencies and their respective magnitudes or property values, i.e. $f \times property$ (for hypothetical illustration see supplementary information SI2 available via the Supplementary Content tab on the article's online page at http://dx.doi.org/).

Accordingly, Equations (1)–(3) are expressed as:

$$I_m = w_T \cdot \log\, 2w_T - \sum_{i=1}^{n} f \cdot w_i^2 \cdot \log_2\, w_i^2 \text{ where } w_T = \sum_{i=1}^{n} w_i^2 \quad (4)$$

$$\bar{I}_m = -\sum_{1=1}^{n} f \cdot \frac{w_i^2}{w_T} \cdot \log_2 \frac{w_i^2}{w_T} \tag{5}$$

$$I_m^* = \frac{\bar{I}_m}{\log_2 w_T} \tag{6}$$

The magnitude-based IFIs could be applied to any event in which the incidence matrix values are replaced with atomic property values.

4.2 *Magnitude-based mutual, conditional and joint entropy*

For the definition of these IFIs, it is necessary to introduce the magnitude-based probability matrix. Given weights w_{ij} of the magnitude frequency matrix, the probabilities p_{ij} are defined as:

$$p_{ij} = \frac{w_i \cdot w_j}{\sum_{i=1}^{n} \sum_{j\geq 1}^{n} c \cdot w_i \cdot w_j} \text{ where } c = \begin{cases} 1 & \text{if } f_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Remember that the off-diagonal elements of the F_M matrix $w_{ij} = f_{ij} \times (w_i \cdot w_j)$, where w_i and w_j are the magnitude entries of vertices i and j in the incidence matrix.

When dealing with atomic properties, it is usual to have negative values. It is, however, mathematically incorrect to have negative probability values. This makes the use of a rescaling scheme necessary, when there exists at least one negative value in a given set of property values, for example the atomic hydrophobic values. We employ a rescaling scheme that gives the property values in the interval [0, 1].

The mutual, conditional and joint entropy-based IFIs on vertex magnitudes are defined in Equations (8), (9) and (10), respectively, as follows:

$$I_m(u; v) = H_m(u; v) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_{ij} \cdot p_{ij} \log_2 \frac{p_{ij}}{p_i p_j} \tag{8}$$

$$\begin{aligned} H_m(u; v) &= H_m(u) - H_m(u/v \text{ or } H_m(u) - H_m(v/u) \\ H_m(u/v) &= H_m(u) - H_m(u; v) \\ H_m(v/u) &= H_m(v) - H_m(u; v) \end{aligned} \tag{9}$$

$$I_m(u, v) = H_m(u, v) = -\sum_{i=1}^{N} \sum_{j=1}^{N} f_{ij} \cdot p_{ij} \log_2 p_{ij} \tag{10}$$

Let us illustrate the computation of the magnitude-based IFIs with the same example used in previous sections, the molecular structure of 4-methylcyclobuta-1,3-dienamine, using the refractivity event criterion.

To compute the magnitude-based negentropy, Shannon's entropy and standardized Shannon's entropy indices, first we determine w_T, defined as the sum of the principal

diagonal elements of the $\mathbf{F}_{MR}$ matrix (i.e. the .principal diagonal elements $w_{ii} = w_i^2$, where w_i is the magnitude entry of vertex i in the incidence matrix).

$$w_T = 15.516 + 18.190 + 18.190 + 20.133 + 6.875 + 8.809 + 0.799 + 0.799$$
$$0.640 + 0.640 + 0.714 + 0.714 + 0.714 = 92.733.$$

Applying the Equations (4), (5) and (6) to the elements w_{ii} of the F_{MR} matrix we obtain, respectively: $I_m = 260.789$ bits, $\bar{I}_m = 2.812$ bits per element and $I_m^* = 0.430$.

The first step in the calculation of the magnitude-based mutual, conditional and joint entropies is the computation of the magnitude-based probability matrix, P_M, using Equation (7).

$$P_M = \begin{bmatrix}
0.1355 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.1589 & 0.1589 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.1589 & 0.1589 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.1758 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.0600 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.0769 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.0070 & 0.0070 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.0070 & 0.0070 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0056 & 0.0056 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0056 & 0.0056 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0062 & 0.0062 & 0.0062 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0062 & 0.0062 & 0.0062 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0062 & 0.0062 & 0.0062
\end{bmatrix}$$

Subsequently Equations (8), (9) and (10) are applied to matrix P_M to obtain the values of $I_m(u;v)$, $I_m(u/v)$ and $I_m(u,v)$, respectively.

$I_m(u; v) = 3.825$ bits, $I_m(u/v) = 28.985$ bits, $I_m(u, v) = 3.825$ bits

Note that the mutual and joint entropy values are equal. This is not coincidence. It is attributed to the fact that the off-diagonal elements of the P_M are equal to their respective i and j principal diagonal elements, in which case the formula for mutual entropy reduces to the one for joint entropy (see Equations 8 and 10).

As in the case of the frequency-based IFIs, the defined norms, means, statistical invariants and classical algorithms (see supplementary information SI1) could be applied to the magnitude-based IFIs defined at atomic (vertex) level, i.e. Shannon's, mutual, conditional and joint entropies.

5. Materials and methods

5.1 *Molecular descriptor computation*

5.1.1 *Dataset for cluster analysis*

For each event, 84 IFIs (variables) are calculated, for a total of 924 IFIs over DRAGON's sample data consisting of 41 heterogeneous molecules (see supporting information SI3 available via the Supplementary Content tab on the article's online page), using the ***GT-STAF*** (acronym for **G**raph **T**heoretical Thermodynamic **STA**te **F**unctions) software, a new module of ***TOMOCOMD-CARDD*** program. The computed variables were filtered to

exclude the ones with zero variance, and a total of 909 variables remained. Local IFIs for atom-types or groups were not included in this study.

5.1.2 *Dataset for QSPR/QSAR modelling*

A dataset comprising 34 2-furylethylene derivatives was used for this study. These 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, SI4 available via the Supplementary Content tab on the article's online page). The GT-STAF software permits the computation of a remarkably large number of MDs. It would obviously be tedious to explore the entire MD space. Consequently, a reduced number of GT-STAF IFIs was calculated for each event and further filtered to eliminate the low-variance MDs.

A total of 3224 DRAGON's MDs were calculated. After eliminating variables with correlation coefficient (x/x) of 1.0, 1392 MDs remained.

5.2 *Chemometric methods*

5.2.1 *Cluster analysis*

These are simple data mining techniques that seek to explore the "natural" relationship that exists among objects (or variables) and allocate to the same classes the similar ones, on the basis of predefined similarity (or dissimilarity) measures [33–36]. Critics of clustering algorithms argue that one of the most important weaknesses of these algorithms is that they will always attempt to find clusters in any data, even where clusters do not naturally exist [37,38].

Bearing this flaw in mind, we first performed hierarchical agglomerative clustering using Ward's method (hierarchical analogue of k-means) and the squared Euclidean distance as amalgamation rule and proximity function, respectively, to have preliminary insight of the possible "optimum" number of subsets contained in the examined data, to serve as guide in determining the k-value to use k-Means Cluster Analysis. For cluster analysis STATISTICA software was employed.

5.2.2 *Multiple linear regression and genetic algorithm*

An indispensable tool in the building of QSAR models is the set of independent numerical quantifiers (variables) of the molecular structure. From this set of variables, subsets that best describe (fit) the experimentally measured molecular parameters are selected, following defined optimization functions. In this study, our aim is to evaluate the predictive capacity of the GT-STAF indices of the 1-octanol/water partition coefficient ($\log P$) and rate constant ($\log K$; for nucleophilic addition of a thiol group to the exo-cyclic double bond) of the 34 2-furylethylene derivatives.

The values of these compounds have been experimentally determined and reported in the literature [39,40]. In this report, we use Multiple Linear Regression analysis coupled with the Genetic Algorithm (MLR-GA), using MobyDigs software (version 1.0 – 2004) [41]. Applications of GA in computational chemistry have demonstrated its efficiency in global optimization. As may have been noted, computations with the GT-STAF software yield high MD dimensional space, justifying the need for data reduction. Accordingly, the MobyDigs GA operator, tabu, was used as the strategy to exclude variables with a high correlation coefficients (x/x). The MDs with zero variance were also eliminated.

The population size was set at 100 and the reproduction/mutation trade-off (T) at 0.70. For each event, the best seven, six and five variable models for the physicochemical properties $\log P$ and $\log K$ were constructed, using as objective function (optimization function) the statistical parameter Q^2_{loo} (''leave one out'' cross-validation). Later, the best variables, for each event and property, were grouped together into a single set and seven, six, five, four, three and two variable models developed (variables used to obtain final models for $\log P$ and $\log K$ available as supporting information (Excel data spread sheet SI5, available via the Supplementary Content tab on the article's online page). A similar procedure was also carried out for DRAGON's MDs. The peculiarity of the GA is that for every exploration, a population of models is obtained, instead of producing a single model, as most statistical methods. From the population of generated models, the ''best'' 10 in each case were retained for validation using the techniques ''bootstrapping'' (Q^2_{boot}) and ''scrambling'' (a (R^2), a (Q^2)). The former evaluates the predictive power of the developed models, while the latter checks the risk of fortuitous correlations (i.e. random correlations between the independent and response variables), a factual possibility when too many variables are screened relative to the number of available observations [41,42]. In addition, Fisher-ratio's p-level (p (F)) and the standard deviation (SECV) were taken into account. Thus, using a multi-criteria perspective, the best model for each case was selected.

6. Results and discussion

6.1 *Cluster analysis*

In this report, a huge volume of MDs is proposed. There is evidently a need for sorting such high-dimensional data into logical and easy-to-work with structures, without loss of vital information. However, a logical question arises: to what extent are variables in a data dissimilar (or similar) to each other in terms of the information that they codify? To respond to this interrogative, cluster analysis techniques are performed.

Figure 4 is a dendrogram for agglomerative hierarchical cluster analysis using Ward's method. As can be seen, the dendrogram shows a clear and consistent tree structure. From a cut-off agglomerative distance of 2250 (27% of maximum distance), eight clusters are obtained and k-means cluster analysis is performed with the same number of clusters.

The loadings for each cluster (C) are available as supporting information (Excel data spread sheet SI6, available via the Supplementary Content tab on the article's online page). The C-3 seems to be particularly important for the terminal paths-based IFIs as they possess remarkable representativity. Fingerprint-based IFIs (i.e. E-state, substructure and MACC) IFIs are strongly grouped in C-4, constituting 88% of the members in the 10-class cluster, which suggests the existence of a close relationship in terms of the information captured by these events. On the other hand, C-5 seems to be important for graph-theoretic events as CS, S, VP and Q-based IFIs are the key contributors to the cluster members, representing 86% of the members in the 8-class cluster. This is a logical result, since these events use the atom–atom connectivity perspective as the starting point, despite the differences in the posterior treatment of molecular graphs. AMR-based IFIs are highly loaded in C-7, suggesting that this event captures structural information not adequately represented by other event-based IFIs. It is worth noting that, although there is no dominance of a particular event in C-1, the majority of the IFIs grouped in this cluster (80% of the IFIs) use the invariant Kurtosis as the global characterizing parameter,

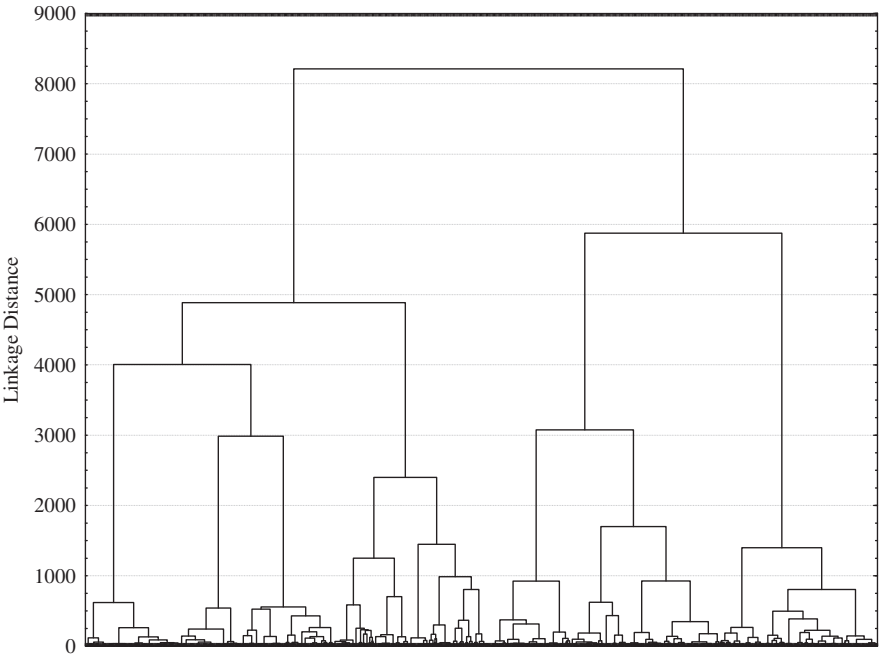

Figure 4. Dendogram for agglomerative hierarchical cluster analysis using Ward's method.

corroborating the results obtained in a previous report that suggested that this invariant captures structural information not captured by the rest of the invariants.

6.1.1 *Information–theoretic cluster evaluation and interpretation*

The use of Shannon's entropy in the evaluation of the goodness of the obtained clusters has been previously proposed by several authors [37,43], following this procedure. Given cluster c with objects (or variables) n, distributed in classes $s_1 \ldots s_m$, according to a determined criterion of similarity, where m is the number of classes, it follows that p_s is the probability that a randomly chosen object i belongs to class s. Thus a probability distribution function (pdf) is constructed. From this pdf, the entropy of each cluster is calculated, using Equation (11):

$$\bar{I} = -\sum_{1=1}^{m} p_s \cdot \log_2 p_s \tag{11}$$

Zero-entropy values represent perfect clustering, in the sense that each cluster comprises elements that naturally belong together, while maximum entropy is obtained when these elements are evenly distributed over all the clusters. Bearing in mind the difference in cluster sizes, the cluster entropies were normalized with respect to the maximum cluster entropy to guarantee plausible comparability. Table 5 shows the cluster entropy values and their corresponding normalized entropy values (wSE).

Table 5. Shannon's entropy for clusters C-1 to C-8.

Clusters	*C-3*	*C-4*	*C-5*	*C-8*	*C-6*	*C-2*	*C-7*	*C-1*
SE	1.219	2.171	2.518	1.887	2.861	3.262	3.035	2.939
wSE	0.197	0.281	0.338	0.371	0.401	0.448	0.463	0.641

As can be observed, the best approximation to desirable clustering is exhibited by C-4 (0.197), which is important for terminal paths-based IFIs; followed by C-4 (0.281) and C-5 (0.338), with high representativity for fingerprint-based IFIs (ES, SS and MACC) and graph theoretic-based IFIs (CS, S, VP and Q), respectively.

Up to this point we have discussed intra-cluster diversity analysis. This approach, however, serves only for cluster evaluation, i.e. it simply gives criterion about the degree of uniformity of data in a cluster, and therefore does not respond to interrogatives about the degree of representativity of a particular class of objects (or variables) in a set of clusters (k).

Accordingly, a formalism to evaluate the fidelity (or the promiscuity) of a particular class objects (or variables) in a set of clusters is introduced. Given a set of n elements grouped in classes $c_1 \ldots c_m$, formed on the basis of a pre-defined criteria of similarity, for example: dimensional (in the case of MDs), pharmacological/toxicological action (in the case of molecular structures) etc, and distributed among a set of clusters $S = \{s_1, s_2, \ldots, s_{n-1}, s_l\}$ using hierarchical or non-hierarchical amalgamation methods, it follows that ${}^{s}p_{(c_i)}$, is the probability that a randomly selected element i belongs to class c of cluster s. Thus for each class c_i, a pdf with respect to the set of cluster $\{{}^{s1}p(c_i), {}^{s2}p(c_i), {}^{s3}p(c_i) \ldots {}^{sl}p_{(c_i)}\}$ is created. Likewise, from this pdf, the entropy of each class is calculated, using Equation (11). In an ideal case, cases (or variables) belonging to the same pre-determined class will have zero-entropy values as they will be loaded in the same cluster, while the promiscuous ones will have maximum entropy as they will be equally distributed in all the clusters.

In this report, diverse events-based criteria are used to derive IFIs. Each of these events thus led to a "family" of IFIs. In this sense, we would like to discover the extent of similarity, or dissimilarity (as a non-linear parameter of independence) of each of these approaches, and also to have a wider understanding of the possible similarities that may exist among the proposed events. Using the formalism proposed in the previous paragraph, entropy for each of the events is calculated. Table 6 shows the resulting entropy values for the eleven events used to derive the novel IFIs. In this case, since for each the events 84 MDs were calculated, the normalization procedure was not necessary.

It is clear that IFIs derived from the E-state event present the highest tendency to assemble in a particular (or a reduced number of) cluster(s). The apparent displacement of the terminal paths-based IFIs from the first position (first study) to a competitive third position suggests that despite possessing the best approximation to "ideal" clustering (C-3), these IFIs are more dispersed in other clusters than ES and SS events-based IFIs (predominantly grouped in C-4). In general terms, fingerprint-based IFIs tend to clump together in a reduced number of clusters compared with graph-theoretic and magnitude-based IFIs, respectively. This finding seem to indicate that fingerprint-based IFIs, as broader family of IFIs, depict a certain degree of similarity and could possibly codify structural information neither codified by graph-theoretic nor magnitude-based IFIs.

Table 6. Shannon's entropy for each event.

Event	*SE*	*Max. SE*[a]	*% of Max. SE*
E-state fingerprints	0.833	2.998	27.785
Substructure fingerprints	1.282	2.998	42.762
Terminal paths	1.553	2.998	51.801
MACC fingerprints	1.615	2.998	53.869
Walks of length *K*	1.943	2.998	64.810
Quantum	2.074	2.998	69.179
Sach's subgraphs	2.082	2.998	69.446
Connected subgraphs	2.132	2.998	71.114
Vertex path incidence	2.155	2.998	71.881
AMR	2.231	2.998	74.416
ALogP	2.299	2.998	76.684

[a]Max SE is the maximum entropy for 8 clusters with 84 elements

The high level of dispersion of the AMR and AlogP-based IFIs among the set of clusters could carry different interpretations depending on the underlying interest: (1) these events codify information contained in most of the other events; and (2) data obtained from these event does not possess a well-defined structure.

6.2 *QSPR/QSAR modelling*

6.2.1 *In-house comparative analysis of the IFIs*

Here, the primary aim is to compare the performance of the proposed IFIs, from an event-based point of view, in the modelling of the physicochemical properties $\log P$ and $\log K$.

6.2.1.1 *log P (1 – octanol/water partition coefficient)*. Figure 5 is a plot of the statistical parameter, Q^2_{loo} for 7, 6 and 5-variable models obtained for the physicochemical property $\log P$ (a table showing all statistical parameters for the best seven, six and five variable models for each event is available as supporting information SI7, available via the Supplementary Content tab on the article's online page). As can be seen, the best models are obtained with the substructure-based IFIs, depicting with even fewer variable models comparable to superior behaviour with respect to the rest of the events. On the other hand, graph-theoretic-based events, as a super-family, yield better regression models than fingerprint and magnitude-based IFIs. This comportment is consistent with results of comprehensive studies, reported by various authors in the literature, which infer better predictive power to the substructure (fragmental) approach in the prediction of $\log P$ of molecular structures in comparison with atom-centred (ALOGP) and topological approaches. In a number of studies [44–47], comparisons with different datasets of drugs and simple organics reveal that fragmental methods correlate more favourably with experimental $\log P$ than atom-based and whole molecule approaches. Similarly, Ghose et al. [48] report marginal superiority of CLOGP (substructure approach) over ALOG P (atom-centred approach). Later on, in a comparative study between the ACD/$\log P$ (fragmental) and topological (2D) descriptor approaches, Petrauskas and Kolovanonv [49] observe better performance with the former. In the definition of SS-based IFIs,

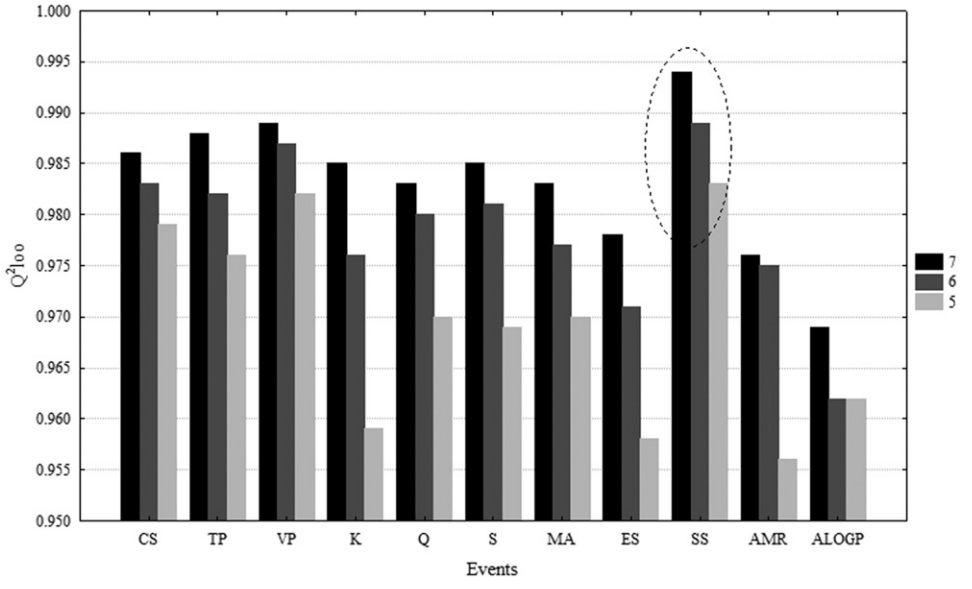

Figure 5. Q^2_{loo} values for $\log P$ models obtained for each family of event-based IFIs.

predetermined values (fragment constants) are not assigned to the group contributions towards the global molecular $\log P$ present in a molecular graph. Instead, Boolean codifications of vertices that contribute to the activated fragment(s) are used. Nonetheless similar behaviour with results reported in the literature substantiates the GT-STAF mathematical approach, as a promising tool in the prediction of physicochemical properties, and it could be inferred that the SS-based IFIs ''latently'' codify structural electronic and hydrophobic interactions, in addition to atomic connectivity features. In addition, all classical procedures used to theoretically determine the molecular hydrophobicity from structural sub-unit contributions follow the use of linear combinations of the fragments (or atoms). It is, however, evident that molecular structures are not necessarily collections of substructures, and in an attempt to counter the error inherent to this assumption, correction factors are used. It is intuitive that these correction rules are not the perfect fix to the assumption of linearity, since they are not applicable to the entire molecular space. In the definition of the GT-STAF IFIs, a series of the so-called *invariants* which are alternative mathematical and graph-theoretic operators that characterize global molecular features from substructure (fragments, atoms) contributions are used, yielding a wider ''span'' of the global structural space (detailed explanation of the invariants given in supporting information SI1, available via the Supplementary Content tab on the article's online page). Furthermore, in a previous report [7] it was demonstrated, using the GT-STAF approach, that the operational rule ''sum of parts, makes the total'' is not necessarily the function that best characterizes the $\log P$ of molecules.

6.2.1.2 *log K (rate constant)*. As can be observed in Figure (6), there is no marked difference in the seven-variable models, that is, they generally demonstrate comparable behaviour in all the events. Six and five-variable models, on the other hand, show

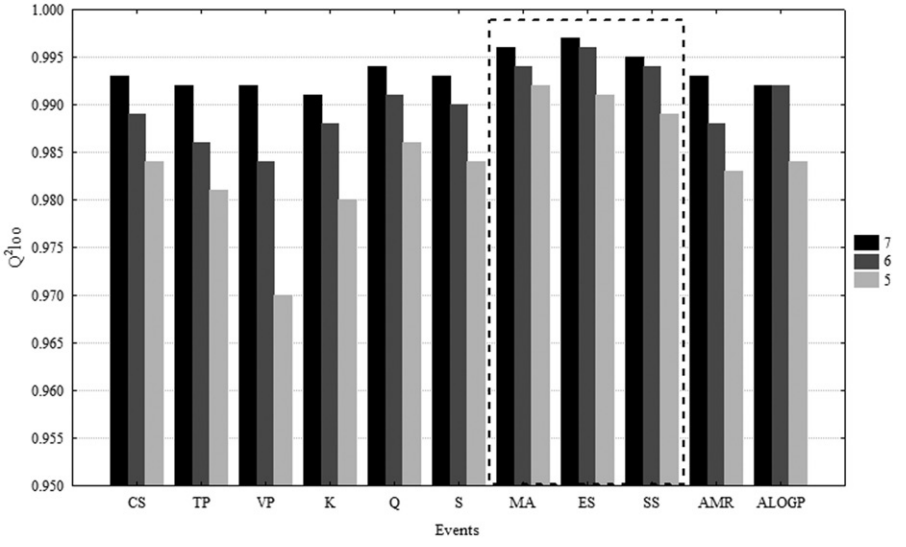

Figure 6. Q^2_{loo} values for $\log K$ models obtained for each family of event-based IFIs.

pronounced differences, with the latter in greater magnitude. A table of all statistical parameters for the best seven, six and five-variable models for each event is available as supplementary information (SI6), available via the Supplementary Content tab on the article's online page).

Fingerprint-based IFIs yield superior correlation coefficients than the rest of the approaches. On the other hand, the correlation coefficients obtained for magnitude (atom-centred) and graph-theoretic IFIs are comparable, with exception of the vertex-path-based IFIs which produce noticeably lower Q^2_{loo} values. It is interesting to note that the best correlation coefficients for $\log K$ are obtained with the ES-based IFIs. This is a logical result, taking into account that nucleophilic addition of a thiol group to the exo-cyclic double bond is primarily dependent on ability of the latter to accommodate the high electron density due to the π electrons and its accessibility, aspects adequately characterized by ES subgraphs.

It is worth noting that although *k*-means clustering is carried out with IFIs calculated over a different database (see Section 4.1.1), the results obtained with QSPR modelling of the physicochemical properties of furylethylenes uphold the outcomes observed using the information-theoretic-based evaluation and interpretation of cluster patterns, in that events with the lowest entropic values (i.e. ES and SS fingerprints) as the desired attribute give the best correlations with the physicochemical properties $\log K$ and $\log P$, respectively, followed by graph-theoretic and magnitude-based (atom-centred) events.

6.2.2 *Comparison with DRAGON molecular descriptors*

The rise in the number of theoretical MDs schemes has been paralleled with the implementation of corresponding software to enable their computation. Of all these software options, in our opinion the most comprehensive collection of MDs, in terms of

number and diversity, is availed by the DRAGON software. We considered it informative to compare the GT-STAF formalism with DRAGON's MDs, in order to have a preliminary notion of the "position" occupied by these novel MDs. For this section, the final models were obtained from a set constituting the best variables in each of the events. Tables 7 and 8 show the equations and the corresponding statistical parameters of the best models for seven, six, five, four, three and two variables for GT-STAF IFIs and DRAGONs MDs as a whole. Complete description of the nomenclature adopted for the GT-STAF IFIs is available as supporting information, SI8, available via the Supplementary Content tab on the article's online page.

As can be seen, the GT-STAF approach yields statistically significant models for the two physicochemical properties $\log P$ and $\log K$, with (almost entirely) comparable to better statistics than those obtained with the entire set of DRAGON MDs. Marginally better performance with DRAGON's MDs is only observed in the two-variable model for $\log P$ (see Equation 23) comprising the MDs MLOP (Moriguchi octanol-water partition coefficient) and MATS4v (Moran autocorrelation - lag 4/weighted by atomic van der Waals volumes). From an empirical perspective, MLOP is in fact an algorithm rather than a simple variable, since it is a regression model comprised of 13 structural parameters (MDs) [50]. Note that participation of MLOG P in the best subsets is limited to only up to three variable models because it does not combine suitably with other DRAGON MDs. It is interesting to highlight that in both the GT-STAF and DRAGON models for $\log K$, there is prevalence of variables that codify (or include in the case of the former) information about the presence of unsaturated bonds and heteroatoms in the furylethylene molecules, with the GT-STAF approach yielding superior correlation coefficients in all the cases. This is a logical result, since these functional groups are critical in determining the rate constant of the furylethylene molecules.

Generally, the use of variables from different events gives models with better statistical parameters for both properties in various cases with fewer variables than in event-wise modeling, though there a few models, especially for six and seven variables, in which the statistics obtained with variables using specific events are not improved (see Equations 12, 13 and 25).

The results obtained in this QSPR study suggest that the GT-STAF approach could be a useful tool in the modelling and prediction of physicochemical, chemical and biological properties of molecular structures. As a concluding remark for this sub-section, we point out that the GT-STAF approach is persuasive as a valuable tool for the calculation of both properties. Moreover, the versatility of the GT-STAF approach, through the use of different events is demonstrated.

7. Conclusions

The aim of this research work was to explore diverse graph-theoretic, physicochemical and chemical approaches that could be applied in the construction of the generalized incidence and relations frequency matrices, **Q** and **F**, respectively. Previously defined Shannon's, mutual, conditional and joint entropy-based IFIs were then calculated on these "extended" **F** matrices, incorporating new insights in the definition of the GT-STAF IFIs. Information-theoretic cluster analysis suggests that the GT-STAF methodology permits to codify dissimilar chemical information of molecular structures. Moreover, the GT-STAF approach demonstrates satisfactory modelling capability, providing ground to

Table 7. Statistical parameters for best seven, six, five, four, three and two variable models for each event obtained for physicochemical property log P of 2-Furylethylenes.

MDs	*N*	r^2	*SECV*	Q^2loo	Q^2boot	*F*	*Model*	*Eq*
All GT-STAF IFIs	7	0.996	0.051	0.994	0.991	936.06	$\log P = -3.0417\ (\pm 0.2128) - 0.7172\ (\pm 0.1176)$ ^{V}***ILEC(1AC(N1))[(IS)((D)]SS*** + 6.1154 (±0.4123) ^{V}***ILE(Q3)[(HT)((D)]SS*** − 1.3509 (±0.2926) ^{C}***ILEC(PN)[(IS)((D)]SS*** + 5.1341 (±0.1588) ^{R}***ILEC(V)[(HT)((D)]SS*** + 7.0094 (±0.2299) ^{C}***IEJ(2AC(MN))[(D)]SS*** − 4.3269 (±0.1552) ^{C}***ILEJ(N3)[(IS)((D)]SS*** + 1.3609 (±0.2426) ^{C}***ILEJ(1AC(N1))[(IS)((D)]SS***	12
	6	0.993	0.064	0.989	0.985	680.00	$\log P = -0.7221\ (\pm 0.2351) + 3.2317\ (\pm 0.4860)$ ^{V}***ILE(Q3)[(HT)((D)]SS*** + 3.3929 (±0.0695) ^{R}***ILEC(V)[(HT)((D)]SS*** + 5.5286 (±0.2931) ^{C}***IEJ(2AC(MN))[(D)]SS*** − 1.7878 (±0.2103) ^{C}***IEJ(4TSK(PN))[(D)]SS*** − 2.1211 (±0.3392) ^{C}***ILEJ(N3)[(IS)((D)]SS*** + 1.7919 (±0.1907) ^{C}***ILEJ(1AC(N1))[(IS)((D)]SS***	13
	5	0.990	0.076	0.986	0.984	578.55	$\log P = 0.5906\ (\pm 0.1605) - 3.2778\ (\pm 0.1783)$ ^{C}***IEJ(4TSK(PN))[(D)]SS*** − 0.0001 (±0.0000) ^{V}***IEC(8AC(A))[(D)]S*** + 1233.4036 (±116.5450) ^{R}***IEM(3GI(V))[(D)]CS*** + 23.2437 (±1.5029) ^{C}***ILEM(P3)[(AH)((D)]Q*** + 8.3494 (±0.7172) ^{C}***IEC(1AC(CV))[(D)]Q***	14
	4	0.981	0.106	0.972	0.970	366.56	$\log P = 0.0668\ (\pm 0.0765) - 0.0021\ (\pm 0.0001)$ ^{V}***IEM(TGI(N1))[(2,3,4)(D)]S*** + 104.6023 (±5.3657) ^{C}***ILE(M)[(IS)((D)]CS*** + 5.6247 (±0.3527) ^{C}***IEM(7TSK(MN))[(D)]CS*** − 5045.9310 (±0.0144) ^{C}***IE(8AC(P2))[(D)]CS***	15

	3	0.960	0.151	0.948	0.946	237.52	$\log P = 1.3945\ (\pm 0.0751) - 5.8574\ (\pm 0.2890)$ ***^{C}IEJ(7TSK(MN))[(D)]Q*** $- 10.6509\ (\pm 0.4189)$ ***^{C}IEJ(8TSK(P2))[(D)]Q*** $- 2.2203\ (\pm 0.2412)$ ***^{R}ILEC(P2)[(DH)((SR,NSR)(D)]K***	16
	2	0.921	0.208	0.902	0.902	179.66	$\log P = -0.3917\ (\pm 0.1387) + 0.0034\ (\pm 0.0002)$ ***^{V}IEM(TGI(N1))[(2,3,4)(D)]S*** $+ 117.5212\ (\pm 10.4663)$ ***^{C}ILE(M)[(IS)((D)]CS***	17
DRAGON's MDs	7	0.994	0.060	0.991	0.988	659.74	$\log P = 0.4031\ (\pm 0.0785) - 1.2424\ (\pm 0.0562)$ GATS4m $+ 0.3151\ (\pm 0.0716)$ EEig09x $+ 0.0683\ (\pm 0.0152)$ RDF085m $- 0.0273\ (\pm 0.0028)$ Vs $- 1.0119\ (\pm 0.0387)$ BLTA96 $+ 0.2528\ (\pm 0.0182)$ F01 [C-C] $- 0.2358\ (\pm 0.0238)$ F01 [C-N]	18
	6	0.992	0.071	0.989	0.986	557.64	$\log P = -0.9394\ (\pm 0.1764) + 0.1526\ (\pm 0.0334)$ piPC07 $- 0.8064\ (\pm 0.0772)$ CIC1 $+ 0.4477\ (\pm 0.0261)$ RDF010u $- 0.2268\ (\pm 0.0405)$ C-016 $- 1.2562\ (\pm 0.0358)$ BLTA96 $- 0.7119\ (\pm 0.0291)$ F01 [C-C]	19
	5	0.990	0.079	0.985	0.983	538.72	$\log P = -3.9163\ (\pm 0.3646) + 3.7416\ (\pm 0.3836)$ SIC1 $+ 0.4220$ (± 0.0238) RDF010e $- 0.2169\ (\pm 0.0450)$ C-016 $- 1.2966$ (± 0.0395) BLTD48 $- 0.7426\ (\pm 0.0311)$ F02 [N-O]	20
	4	0.977	0.115	0.968	0.965	310.18	$\log P = -5.4993\ (\pm 1.0263) - 3.4447\ (\pm 0.2008)$ GATS1p $+ 1.9404\ (\pm 0.2708)$ EEig01d $+ 0.1260\ (\pm 0.0043)$ ADDD $- 0.1385\ (\pm 0.0052)$ G(N..N)	21
	3	0.958	0.153	0.949	0.943	230.49	$\log P = 3.1486\ (\pm 0.1393) - 1.6203\ (\pm 0.1224)$ GATS1p $- 0.7294\ (\pm 0.1573)$ nRCONR2 $+ 0.5618$ (± 0.0232) MLOGP2	22
	2	0.932	0.192	0.919	0.919	213.32	$\log P = 1.6334\ (\pm 0.0597) + 1.6029\ (\pm 0.1458)$ MATS4v $+ 0.5833\ (\pm 0.0290)$ MLOGP2	23

Statistical parameters for best seven and six variable models obtained with SS-based IFIs are not improved using a set of best variables from all events and therefore the statistical parameters of the former are reported (see Equations 12 and 13).

Table 8. Statistical parameters for best seven, six, five, four, three and two variable models for each event obtained for physicochemical property log K of 2-Furylethylenes.

MDs (log K)	*N*	r^2	*SECV*	Q^2loo	Q^2boot	*F*	*Model*	*Eq*
All GT-STAF IFIs	7	0.999	0.061	0.998	0.997	2588.26	$\log K = -4.0406\ (\pm 0.3148) + 25.9292\ (\pm 2.7535)$ $^{R}IE(5TSK(DE))[(D)]SS + 1.1668\ (\pm 0.0622)$ $^{V}IE(5TSK(N3))[(D)]AP + 28.4112\ (\pm 0.5376)$ $^{V}ILE(N1)[(IS)((D)]MA + 0.0106\ (\pm 0.0024)$ $IEJ(TCN(K))[(D)]MA - 0.5800\ (\pm 0.0442)$ $^{V}IE(Q3)[(D)]V - 0.9657\ (\pm 0.0674)$ $^{V}IEM(1TSK(I50))[(2,3,4)(D)]S - 0.0375\ (\pm 0.0015)$ $^{V}IEC(7TSK(V))[(D)]CS$	24
	6	0.998	0.076	0.996	0.995	1960.29	$\log K = -\ 5.2578\ (\pm 0.1934) + 0.8316\ (\pm 0.0584)$ $^{C}ILEJ(6GI(P2))[(HT)((D)]ES - 65.0624\ (\pm 2.4755)$ $^{C}ILEM(EE(Q3))[(HT)((D)]ES + 3.1393\ (\pm 0.0968)$ $ILEM(G)[(HT)((D)]ES + 25.5880\ (\pm 0.3815)$ $IE(2TSK(Q2))[(D)]ES + 0.1236\ (\pm 0.0101)$ $^{R}IE(5TSK(K))[(D)]ES + 0.0213\ (\pm 0.1915)$ $^{C}IEC(7CN(K))[(D)]ES$	25
	5	0.995	0.115	0.992	0.986	1024.17	$\log K = -2.2697\ (\pm 0.5940) + 0.4905\ (\pm 0.0428)\ ^{V}ILEC$ $(IM(K))[(HT)((D)]MA + 36.1376\ (\pm 0.8812)$ $^{V}ILE(N1)[(IS)((D)]MA - 0.7586\ (\pm 0.0640)$ $^{C}ILEC(N3)[(IS)((D)]MA - 106.4318\ (\pm 9.5060)$ $^{C}ILEM(Q2)[(IS)((D)]SS - 3.1966\ (\pm 0.2326)$ $^{V}IEM(1CN(MN))[(D)]SS$	26
	4	0.993	0.131	0.989	0.987	992.14	$\log K = -6.2313\ (\pm 0.7102) + 6.3813\ (\pm 0.5747)$ $ILEJ(P3)[(HT)((D)]AP + 35.0818\ (\pm 0.9914)$ $^{V}ILE(N1)[(IS)((D)]MA - 1.2177\ (\pm 0.1300)$ $^{V}IEM(1TSK(I50))[(2,3,4)(D)]S - 0.0007\ (\pm 0.0001)$ $^{V}IEJ(7AC(V))[(D)]CS$	27

	3	0. 979	0.220	0. 969	0. 958	462.26	$\log K = -11.2874\ (\pm 0.7131) + 40.3715\ (\pm 1.3089)\ {}^{V}$***ILE(N1)[(IS)((D)]MA*** $- 134.2182\ (\pm 19.6324)$ ^{C}***ILEM(Q2)[(IS)((D)]SS*** $+ 11.5671\ (\pm 0.9641)$ ^{V}***IEC(0CN(Q3))[(D)]SS***	28
	2	0. 948	0.339	0. 929	0. 921	281.18	$\log K = -12.5333\ (\pm 0.7679) + 44.4985\ (\pm 1.9446)\ {}^{V}$***ILE(N1)[(IS)((D)]MA*** $+ 104.1157\ (\pm 13.8546)$ ^{C}***ILE(I50)[(HT)((D)]SS***	29
DRAGON's MDs	7	0.998	0.08	0.996	0.993	1482.69	$\log K = -1.2423\ (\pm 0.2711) + 0.8656\ (\pm 0.0367)$ nDB $+ 0.5229$ (± 0.1337) Jhetv $- 1.4491\ (\pm 0.1060)$ MATS6e $+ 0.0459$ (± 0.0103) RDF070e $+ 16.0161\ (\pm 2.0124)$ R8p+ $- 0.6947$ (± 0.0723) nArNO2 $- 1.5268\ (\pm 0.0713)$ B04 [N-O]	30
	6	0.996	0.097	0.995	0.994	1185.6	$\log K = -14.5980\ (\pm 0.9718) + 4.5728\ (\pm 0.2505)$ piPC02 $+ 1.2627\ (\pm 0.1063)$ GATS6e $- 0.0514\ (\pm 0.0135)$ RDF065e $- 1.6815\ (\pm 0.2922)$ Mor13p $+ 0.8088\ (\pm 0.0907)$ nR=Ct $+ 2.3538\ (\pm 0.0605)$ nArNO2	31
	5	0. 995	0.108	0.993	0.989	1156.46	$\log K = -42.5509\ (\pm 3.2662) - 1.7755\ (\pm 0.1215)$ MATS6e $+ 23.7063\ (\pm 1.7531)$ BELp1 $+ 2.6451\ (\pm 0.2770)$ R8m $+ 5.4982\ (\pm 0.5259)$ R8p $+ 1.5899\ (\pm 0.0966)$ B04 [N-O]	32
	4	0. 988	0.165	0.981	0.980	608.9	$\log K = 2.1705\ (\pm 0.1161) + 0.5221\ (\pm 0.0962)$ Mor12e $- 1.0199$ (± 0.0947) B05[C-C] $+ 0.9312\ (\pm 0.0471)$ F05[C-N] $+ 0.9005$ (± 0.0245) F05[O-O]	33
	3	0. 984	0.191	0.974	0.972	608.55	$\log K = -16.8907\ (\pm 1.2173) + 4.8261\ (\pm 0.2869)$ piPC03 $- 1.9646\ (\pm 0.2049)$ MATS6e $+ 2.3391\ (\pm 0.0991)$ nRNO2	34
	2	0. 947	0.338	0.930	0.927	278.71	$\log K = 0.7798\ (\pm 0.4081) + 0.7473\ (\pm 0.1037)$ nDB $+ 2.6183$ (± 0.1397) B04 [N-O]	35

Statistical parameters for best six variable model obtained with ES-based IFIs are not improved using a set of best variables in each of the events (see Equation 25).

recommend this approach as a tool to consider in QSAR/QSPR modelling, diversity analysis and other chemoinformatic tasks.

Acknowledgements

Y. Marrero-Ponce thanks the 'Estades Temporals per an Investigadors Convidats' program for a fellowship to work at Valencia University in 2012. The authors acknowledge the partial financial support from Spanish Ministry of Science and Innovation (MICINN, project reference: SAF2009-10399). Finally, this work was also partially supported by VLIR (Vlaamse InterUniversitaire Raad, Flemish Interuniversity Council, Belgium) under the IUC Program VLIR-UCLV.

References

[1] A. Crum-Brown, *On an application of mathematics to chemistry*, Proc. Roy. Soc. VI (1867), pp. 89–90.
[2] A. Crum-Brown and T.R. Fraser, *On the connection between chemical constitution and physiological action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia*, Trans. Roy. Soc. 25 (1868), pp. 151–203.
[3] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 1st ed., Vol. 1, Wiley-VCH, Weinheim, 2009, p. 667.
[4] S.D. Brown, T.B. Blank, S.T. Sum, and L.G. Weyert, *Chemometrics*, Anal. Chem. 66 (1994), pp. 315R–359R.
[5] B.K. Lavine, *Chemometrics*, Anal. Chem. 70 (1998), pp. 209R–228R.
[6] E.V. Krishna Murty, *Computer Modeling of Complex Biological Systems*, CRC Press, Boca Raton, FL, 1985, p. 77.
[7] S.J. Barigye, Y. Marrero-Ponce, O. Martínez-Santiago, Y. Martínez-López, and F. Torrens, *Shannon's, mutual, conditional and joint entropy-based information indices. Generalization of molecular descriptors defined from LOVIs*, Curr. Comput.-Aided Drug Des. (2012).
[8] S.J. Barigye, Y. Marrero-Ponce, Y. Martínez López, L.M. Artiles Martínez, R.W. Pino-Urias, O. Martínez Santiago, and F. Torrens, *Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices*, J. Comput. Chem. (2012), DOI:10.1002/jcc.23123.
[9] V.A. Gorbátov, *Fundamentos de la Matematica discreta*, URSS, Mir, Moscow, 1988.
[10] L.H. Hall and L.B. Kier, *Molecular Connectivity and Substructure Analysis*, J. Pharm. Sci. 67 (1978), pp. 1743–1747.
[11] L.H. Hall and L.B. Kier, *A comparative analysis of molecular connectivity, Hansch, Free-Wilson and Darc-Pelco methods in the SAR of halogenated phenols*, Eur. J. Med. Chem. 13 (1978), pp. 89–92.
[12] L.B. Kier and L.H. Hall, *Molecular Structure Description. The Electrotopological State*, Academic Press, San Diego, 1999.
[13] L. Jäntschi, *Graph Theory. 1. Fragmentation of Structural Graphs*, Leonardo Electron. J. Pract. Technol. 1 (2002), pp. 19–36.
[14] D. Janezic, A. Milicevic, S. Nikolic, and N. Trinajstic, *Graph Theoretical Matrices in Chemistry*, University of Kragujevac, Faculty of Science, Kragujevac, Serbia, 2007.
[15] M.I. Skvortsova and I.V. Stankevich, *Eigenvectors of weighted graphs: A supplement to Sachs' theorem*, J. Mol. Struct. 719 (2005), pp. 213–223.

[16] H. Sachs, *Beziehungen zwischen den in einem graphen enthaltenen Kreisen und seinem characteristischen Polynom*, Publ. Math. 11 (1964), pp. 119–134.
[17] D. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs-Theory and Application*, Academic Press, 1979.
[18] I. Gutman, *Impact of the Sachs Theorem on theoretical chemistry: A participant's testimony*, MATCH Commun. Math. Comput. Chem. 48 (2003), pp. 17–34.
[19] A. Jun-ichi, *General rules for constructing Huckel molecular orbital characteristic polynomials*, J. Am. Chem. Soc. 98 (1975), pp. 6840–6844.
[20] R.B. Mallion, A.J. Schwenk, and N. Trinajstic, *A graphical study of heteroconjugated molecules*, Croat. Chem. Acta 46 (1974), p. 171.
[21] R.T. Morrison and R.N. Boyd, *Organic Chemistry*, 6th ed., Prentice-Hall, New Delhi, 1992.
[22] R. Guha, The CDK Descriptor Calculator, 0.94 ed., Indiana, 1991, available at http://cheminfo.informatics.indiana.edu/~rguha/code/java/cdkdesc.html
[23] C. Steinbeck, Y.Q. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E.L. Willighagen, *The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 493–500.
[24] R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J.K. Wegner, and E.L. Willighagen, *The Blue Obelisk-Interoperability in Chemical Informatics*, J. Chem. Inf. Model. 46 (2006), pp. 991–998.
[25] G. Landrum, *RDKit: Open-source cheminformatics*, available at http://www.rdkit.org
[26] J.L. Durant, B.A. Leland, D.R. Henry, and J.G. Nourse, *Reoptimization of MDL keys for use in drug discovery*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 1273–1280.
[27] *MACCS Drug Data Report*, 2000.2, MDL Information Systems, Inc. 14600 Catalina Street, San Leandro, CA 94577, 2000.
[28] L.B. Kier and L.H. Hall, *An electrotopological-state index for atoms in molecules*, Pharm. Res. 7 (1990), pp. 801–807.
[29] L.H. Hall and L.B. Kier, *Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 1039–1045.
[30] A.K. Ghose and G.M. Crippen, *Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions*, J. Chem. Inf. Comput. Sci. 27 (1987), pp. 21–35.
[31] D. Bonchev, *My life-long journey in mathematical chemistry*, Internet Electron. J. Mol. Des. 4 (2005), pp. 434–490.
[32] D. Bonchev, *Information-Theoretic Indices for Characterization of Chemical Structures*, Research Studies Press, Chichester, UK, 1983.
[33] J.M. Barnard and G.M. Downs, *Clustering of chemical structures on the basis of two-dimensional similarity measures*, J. Chem. Inf. Comput. Sci. 32 (1992), pp. 644–649.
[34] S.C. Basak, B.D. Gute, and A.T. Balaban, *Interrelationship of major topological indices evidenced by clustering*, Croat. Chem. Acta 77 (2004), pp. 331–344.
[35] R.D. Brown and Y.C. Martin, *Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection*, J. Chem. Inf. Comput. Sci. 36 (1996), pp. 572–584.
[36] J. MacCuish, C. Nicolaou, and N.E. MacCuish, *Ties in proximity and clustering compounds*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 134–146.
[37] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, New Jersey, March 1988.
[38] G.W. Milligan, *Clustering validation: Results and implications for applied analyses*, in *Clustering and Classification*, P. Arabie, L. Hubert, and G.D. Soete, eds., World Scientific, Singapore, 1996, pp. 345–375.
[39] E. Estrada and E. Molina, *Novel local (fragment-based) topological molecular descriptors for QSPR/QSAR and Molecular Design*, J. Mol. Graphics Model 20 (2001), pp. 54–64.

[40] S. Wold and L. Erikson, *Statistical validation of QSAR results. Validation tools*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH Publishers, Weinheim, Germany, 1995, pp. 309–318.
[41] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, *MOBYDIGS version 1.0*, Milano, 2005; available at http://www.talete.mi.it/mobydigs.htm
[42] J.G. Topliss and R.P. Edwards, *Chance factors in studies of quantitative structure-activity relationships*, J. Med. Chem. 22 (1979), pp. 1238–1244.
[43] Y. Zhao and G. Karypis, *Empirical and theoretical comparisons of selected criterion functions for document clustering*, Machine Learning 55 (2004), pp. 311–331.
[44] R. Mannhold, K. Dross, and R.F. Rekker, *Drug lipophilicity in QSAR practice.1. A comparison of experimental with calculative approaches*, Quant. Struct. Act. Relat. 9 (1990), pp. 21–28.
[45] R.F. Rekker, R. Mannhold, and A.M. ter Laak, *On reliability of calculated P-values: Rekker,Hansch/Leo and Suzuki approach*, Quant. Struct. Act. Relat. 12 (1993), pp. 152–157.
[46] R. Mannhold, R.F. Rekker, C. Sonntag, A.M. ter Laak, K. Dross, and E.E. Polymeropoulos, *Comparative evaluation of the predictive power of calculation procedures for molecular lipophilicity*, J. Pharm. Sci. 84 (1995), pp. 1410–1419.
[47] R. Mannhold, G. Cruciani, K. Dross, and R.F. Rekker, *Multivariate analysis of experimental and computational descriptors of molecular lipophilicity*, J. Comput. Aided Mol. Des. 12 (1998), pp. 573–581.
[48] A.K. Ghose, V.N. Viswanadhan, and J.J. Wendoloski, *Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods*, J. Phys. Chem. A 102 (1998), pp. 3762–3772.
[49] A.A. Petrauskas and E.A. Kolovanov, *ACD approaches for PHYS-Chem data prediction*, 13th Eur. Symp on QSAR, Düsseldorf, 2000, Quant. Struct. Act. Relat., p. 84.
[50] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita, *Simple method of calculating octanol/water partition coefficient*, Chem. Pharm. Bull. 40 (1992), pp. 127–130.