J Math Chem (2015) 53:2028–2064
DOI 10.1007/s10910-015-0533-3

ORIGINAL PAPER

Novel global and local 3D atom-based linear descriptors of the Minkowski distance matrix: theory, diversity–variability analysis and QSPR applications

Néstor Cubillán[1,2] · Yovani Marrero-Ponce[3,4] · Harold Ariza-Rico[1] · Stephen J. Barigye[5] · César R. García-Jacas[6] · José R. Valdes-Martini[7] · Ysaías J. Alvarado[8]

Received: 30 May 2015 / Accepted: 3 July 2015 / Published online: 18 July 2015

Abstract A new family of alignment-free 3D descriptors based on TOMOCOMD-CARDD framework has been designed, namely 3D-linear indices. In this report, we have proposed the use of a generalized form of the geometric pairwise atom-atom

Electronic supplementary material The online version of this article (doi:10.1007/s10910-015-0533-3) contains supplementary material, which is available to authorized users.

✉ Néstor Cubillán
ncubillan@fec.luz.edu.ve

Yovani Marrero-Ponce
ymarrero77@yahoo.es; ymponce@gmail.com; yovanimp@uclv.edu.cu

1 Laboratorio de Electrónica Molecular, Departamento de Química, Facultad Experimental de Ciencias, Universidad del Zulia, Maracaibo, Bolivarian Republic of Venezuela

2 Departamento de Física, Facultad Experimental de Ciencias, Centro de Modelado Científico, Universidad del Zulia, Maracaibo, Bolivarian Republic of Venezuela

3 Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research International Network (CAMD-BIR IN), Los Laureles L76MD, Nuevo Bosque, 130015 Cartagena de Indias, Bolívar, Colombia

4 Grupo de Investigación en Estudios Químicos y Biológicos, Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar (UTB), Parque Industrial y Tecnológico Carlos Vélez Pombo, Km 1 vía Turbaco, 130010 Cartagena de Indias, Bolívar, Colombia

5 Departamento de Química, Universidade Federal de Lavras, Lavras, MG, Brazil

6 Departamento de Bioinformática, Universidad de las Ciencias Informáticas, Havana, Cuba

7 Laboratorio de Inteligencia Artificial, Facultad de Matemática, Física y Computación, Centro de Estudios de Informática (CEI), Universidad Central "Marta Abreu" de Las Villas, Santa Clara 54830, Villa Clara, Cuba

8 Laboratorio de Caracterización Molecular y Biomolecular, Departamento de Investigación en Tecnología de los Materiales y el Ambiente (DITeMA), Instituto Venezolano de Investigaciones Científicas (IVIC), Avenida 74 con calle 14A, Maracaibo, Bolivarian Republic of Venezuela

distance matrix as structural information matrix. This matrix, denominated as non-stochastic, uses as matrix form of linear maps as well as their algebraic transformations: stochastic, double stochastic and mutual probabilities matrices. The methodology for 3D-QSAR studies is based on the combined use of global and local approaches. Principal component analysis reveals that the novel indices are capable of capturing structural information not codified by the indices implemented in the DRAGON's software. Moreover, Shannon's entropy based variability analysis comparing the 3D-linear indices with some relevant descriptors suggests that the former encode similar-to-better amount of structural information than these descriptors. Finally, a search for the best regressions for congeneric databases in QSPR modeling was performed. The overall results demonstrates satisfactory behavior.

Keywords TOMOCOMD-CARDD · 3D-linear index · Variability analysis · QSPR study

1 Introduction

The pharmaceutical industry needs to address the increasing cost and time for drug development [1,2], and *in silico* lead identification and optimization (two main steps in the new lead discovery) are increasingly becoming important means of tackling these challenges [3,4]. This drug discovery pipeline often involves quantitative structure-activity relationships (QSARs) [5–7], which focus on deriving correlations between the properties/molecular descriptors and their pharmacological activities and ADME-TOX endpoints. These methods can be broadly categorized as two-dimensional (2D) or three-dimensional (3D) QSARs. The 2D-QSAR methods commonly use chemical information derived from constitutional and topological information of molecules, fingerprints, and traditional physicochemical properties (0D–2D molecular descriptors) [8]. The 3D-QSAR methods consider the physicochemical properties of the ligands in their hypothesized bioactive conformations [5–7].

At present, many outstanding QSAR methods based on 2D properties of the molecule have a comparable to better quality than the 3D methods [9–13]. Although 0D–2D molecular descriptors (MDs) are routinely applied in endpoint predictions, many properties have been shown to require more detailed 3D information to properly capture the relevant structural features responsible for the description or modeling of endpoints of interest [14–22]. It has become evident that physical, chemical, or biological properties of a compound depend on the three-dimensional (3D) arrangements of the atoms in the molecule. Hence, 3D MDs seem to be indispensable for a reliable structural characterization and adequate model generation. For that reason, in recent years, the QSAR methods based on the 3D structures of the molecules, such as CoMFA [23], GRID [24], COMPASS [25], and GERM [26], have been widely used in several scientific fields.

However, the CoMFA-like (field-based approaches) methods suffer from a number of limitations such as: (a) the requirement of molecular superposition, (b) dependence of the statistical quantities on the grid-point distance, and (c) the use of partial charges in representing electrostatic interactions. In addition, the implementing these methods

based on 3D structures are in general difficult and time-consuming because of the difficulty in generating optimal 3D conformations of molecules under study [16]. Typically, molecules are aligned by performing an overlap of common structural units. The possibility of molecular alignment and their types is an inherent, and often critical, element of many 3D QSAR methods [27–29]. This "alignment schedule" is adequate for a data set, that is closely related structurally, but is far more difficult to apply to either a diverse data set or on the basis of some structural property other than shape, even for sterically similar molecules. In addition, the alignment rules can introduce user bias and the resultant model being dependent upon and sensitive to the alignment used. In conclusion, molecular superposition remains an *intrinsic* and problematic requirement of many 3D QSAR methods, which can only be eliminated if the binding mode of a ligand with respect to the receptor, i.e., the alignment of the ligand with respect to the receptor and the conformation of the receptor-bound ligand, is experimentally known. This involves knowledge of the structure of the ligand-receptor complex, a condition that is unfortunately seldom fulfilled.

One of the approaches for overcoming this problem is by using a 3D model from Cartesian coordinates, which is invariant to rotation and translation of the molecule and alignment-independent (free) [24,30]. The MS-WHIM (weighted holistic invariant molecular) (and also WHIM descriptors [31]) approach overcomes the problem of molecular alignment by calculating statistical parameters (eigenvalue proportion, skewness and kurtosis) from the score matrix obtained from weighted principal component analysis (PCA) [32]. Methods based on autocorrelation of certain molecular properties represent another type of approaches that are alignment insensitive [32,33]. Another example of a novel 3D QSAR approach *invariant* to molecular orientation, and therefore, that does *not* require *alignment rules*, is comparative spectra analysis (CoSA) [34]; in which molecular spectra are used as three-dimensional MDs for the prediction of biological activities. The grid-independent descriptors (GRIND) are also another significant example of this kind of MDs [24]. The vibration-based descriptors (EVA) represent another important method (these MDs appear to be even less sensitive to conformation) which unlike 3D-QSAR methods such as CoMFA, provides a conformationally sensitive but, superposition-free descriptors that have been shown to perform well with a wide range of datasets and biological endpoints [35,36]. The distance atomic physicochemical parameter energy relationships (DAPPER) method is sensitive to the 3D structure, but has an advantage over field-based 3D QSAR methods in as much as it is invariant to both translation and rotation of the structures concerned and thus structural superposition is not required [37]. Another alignment-insensitive novel code, the 3D MoRSE (molecule representation of structures based on electron diffraction) code was proposed for representing the 3D structure [38]. The CoMMA method provides 3D descriptors independent of the orientation of the molecules in space as well [39]. Finally, GETAWAY (geometry, topology, and atom-weights assembly) and RDF (radial distribution functions) MDs are two appropriate families of parameters rather useful to coding the 3D molecular structure [40].

On the other hand, during past decade there has been a tendency to extend traditional topological indices (TIs) to account for 3D representation of the molecule by including geometrical information. Among such indices, are the 3D-Wiener, 3D-Harary, 3D-Balaban, 3D-Gravitational indices, the 3D-Petitjean shape indices, the Randić

molecular profiles, BCUT descriptors, etc., [40–48]. These MDs are alignment-free, easily and quickly calculated (and also faster and easier to implement in an automated fashion), and are typically characterized by the same or better statistics; thus being suitable for QSARs. In addition, chiral indices derived from molecular graphs (as modified conventional (nonchiral) topological descriptors) have been proposed and applied in several QSAR studies of several benchmark datasets [9,10,49–54]. In all of these studies chirality MDs were characterized by similar or better statistics and predictive power compared with CoMFA and/or other 3D-QSAR models reported in the literature for the same datasets.

In previous papers, one of the authors, introduced new sets of atom- and bond-level MDs of relevance to QSAR/QSPR studies and "rational" drug design, *atom- and bond-based 2D-linear indices* [12,55,56]. These local (atomic, group and atom-type) indices are based on the calculation of linear maps (and linear functional) in R^n, using canonical basis sets. The description of the significance-interpretation and the comparison with other MDs were also performed. This approach describes changes in the electronic distribution with time, throughout the molecular backbone. Specifically, features of the k^{th} total and local linear indices upon variations in the molecular structure, including chain lengthening and branching as well as content of heteroatoms, and multiple bonds were illustrated by various examples. This *in silico*-method has been successfully applied to the prediction of several physical and chemical properties of organic compounds [12,56,57]. These MDs, and their stochastic forms [12,58], have also been useful in the selection of novel subsystems of compounds with desired properties/activities in virtual screening protocols [59–64]. In addition, the molecular linear indices (2D) have been extended to consider three-dimensional features of small/medium-sized molecules based on the *trigonometric–3D–chirality–correction factor* approach (2.5 GBT-like indices) [53,65].

In this report we present alignment-free *global and local 3D-linear descriptors*, derived in a similar way, (accounting for 3D molecular information, although based on topological approaches previously defined in 2D and 2.5-atom based linear indices). Moreover, we propose novel total and local (atom, atom-type and group) MDs based on the *extended and generalized 3D (geometric) distance matrices*. Algebraic transformations on these matrix representations yield "stochastic", "double-stochastic" and "mutual probabilistic" distance matrices for atom-pairs, from which 3D (geometric)-linear indices are obtained. In order to evaluate the contribution of the extended and generalized 3D distance matrices on the variability of the 3D-linear indices derived thereof, we compare the information content encoded by these MDs using a methodology proposed by Godden and Bajorath based on the concept of Shannon's entropy [66,67]. For this analysis, we use DRAGON's sample data consisting of 41 heterogeneous molecules. This data was also used to compare the information content codified by the new 3D-linear indices with the MDs implemented in the DRAGON program using principal component analysis (PCA). In addition, the correlation ability of the new MDs is tested in QSPR studies of selected physicochemical properties of octanes (first experiment), in the description of the boiling point of 28 alkyl-alcohols (second experiment), as well as in the modeling of the specific rate constant (log k) and partition coefficient (log p) of 34 derivatives of 2-furylethylenes (third experiment). Comparisons with other approaches (vertex- and edge-based connectivity indices, total

and local spectral moments, quantum-chemical descriptors, plus E-state/biomolecular encounter parameters) are carried out in order to analyze the behavior of the novel method in these QSPR studies with regard to most of the MDs reported in the literature to date.

2 Theoretical scaffold

In the classical TOMOCOMD-CARDD approach, the linear descriptor of a molecule composed of m atoms is a linear function (form) of the atom linear maps from R^m to the scalar R [$f(\vec{\mathbf{x}})$: $R^m \rightarrow R$], and it is expressed in matrix form as [56]:

$$f(\vec{\mathbf{x}}) = \sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}x_j = \vec{\mathbf{u}}^t \cdot \mathbf{G} \cdot \vec{\mathbf{x}} \tag{1}$$

where, m is the number of atoms in the molecule, $\vec{\mathbf{u}}^t$ is a m-dimensional row vector whose components are unity, $\vec{\mathbf{x}}$ is the molecular vector and $\mathbf{G}$ is the structural information matrix that characterizes the molecular graph [56,58,68]. The coefficients g_{ij} are the elements of matrix $\mathbf{G}$ and x_j are the coordinates of the atom-based molecular vector ($\vec{\mathbf{x}}$) in the so-called canonical ('natural') basis set.

Note that Eq. (1) is defined as a linear form (global index). However, these linear indices can be defined as linear transformations (linear applications) $f(\vec{\mathbf{x}}_i)$ in the molecular vector space $\Re^m$. This map is a correspondence that assigns a vector $f(\vec{\mathbf{x}}_i)$to every vector $\vec{x}$ in R^m. That is, if a molecule consists of m atoms, then the atom linear indices for atom i are calculated as linear maps in R^m (endomorphism in R^m), in canonical basis set. Specifically, the atom-level linear indices, $f(\vec{\mathbf{x}}_i)$, are computed as shown in Eq. (2):

$$f(\vec{\mathbf{x}}_i) = \sum_{j=1}^{m} g_{ij}x_j = \mathbf{G} \cdot \vec{\mathbf{x}} = \mathbf{Y} \tag{2}$$

In this way, the total linear index (whole-molecule), $f(\vec{\mathbf{x}})$, is calculated from local (atom) linear indices as shown in Eq. (3):

$$f(\vec{\mathbf{x}}) = \sum_{i=1}^{m} f(\vec{\mathbf{x}}_i) = \vec{\mathbf{u}}^t \cdot \mathbf{Y} \tag{3}$$

2.1 Molecular vector

The molecular vector contains information about of the properties of the atoms or entities of a molecule or macromolecule. Given a molecule with m atoms, the components of vector $\vec{\mathbf{x}}$ are numerical values corresponding to particular atomic property, that is [56,58,64]:

$$\vec{\mathbf{x}} = \{x_1, x_2, x_3, \ldots, x_m\} \tag{4}$$

The atomic properties used in this work were Mulliken electronegativity (χ), polarizability (α), atomic mass (m) and van der Waals volume (r). For values of these properties see Table S1 of the Supplementary Information.

2.2 Total and local structural information matrices

The TOMOCOMD-CARDD 2D approach uses graph-theoretical matrices in order to codify structural information of molecules, mainly adjacency matrices of molecular pseudographs, macromolecular graphs [68] and bond-based graphs [12] which contain information about long- and short-range covalent interactions within molecule, but these exclude the non-covalent interactions. With the aim of including all possible interactions, we propose to associate the molecular structural information to a function $\times$ expressed as Taylor or McLaurin series of the interatomic distance g:

$$X(g) = \sum_{i=0}^{\infty} a_i \cdot g^i \tag{5}$$

The general definition of distance depends on the space and metric. If a molecule is in an Euclidean space, it is possible to generalize the distance between the atoms i and j through Minkowski distance [69]:

$$g_{ij}^{n,p} = \left(|x_i - x_j|^n + |y_i - y_j|^n + |z_i - z_j|^n\right)^{p/n} \tag{6}$$

where x, y and z represent the coordinates in Cartesian axis; n is the Minkowski distance norm, order or metric ($e.g.$, $n = 1$ is the Manhattan, or city-block distance, and $n = 2$ the well-known Euclidean distance). The p value, [i in the Eq. (5)], is an analogue to the topological *step-count* and it measures the range of the interaction. In this report, similar to the graph-theoretical framework, we propose a matrix $\mathbf{G}^{n,p}$ to codify the structural information of a molecule. Henceforth, it will be known as *non-stochastic Minkowski distance matrix*.

It should be noted that this matrix, $\mathbf{G}^{n,p}$, is the more general case (extended or expanded) of the well-known geometry matrix, $\mathbf{G}$ (if $n = 2$ and $p = 1$, then $\mathbf{G}^{n,p} = \mathbf{G}$). The geometry matrix (or geometric distance matrix) of a molecule is a square symmetric matrix $m \times m$ whose entry r_{ij} is the *geometric distance* calculated as the *Euclidean distance* between the atoms i and j; diagonal entries are always zero. Geometric distances are intramolecular (interatomic) distances [70]. Like the molecular matrix forms, the geometry matrix contains information about molecular configurations and conformations.

A normalization of our "extended" geometric distance matrix $\mathbf{G}^{n,p}$ can be obtained by using probability-based matrices derived from $\mathbf{G}^{n,p}$. These representations are based on a probability matrix to describe the interatomic interactions [9,61,71–76]. In TOMOCOMD-CARDD 2D, the *pseudograph's stochastic adjacency matrix* describes changes in the electron distribution over time throughout the molecular backbone. In this scheme, a hypothetical situation in which a set of atoms are initially free in space is considered (discrete object in the space). Later, outer shell electrons of atoms are distributed around atomic cores in discrete intervals of time. In this sense, the electrons in an arbitrary atom can move to other atoms at different discrete time periods throughout the chemical-bonding network. In our geometrical approach, this matrix can be interpreted as the changes in the probability of atoms in a molecule to interact

with each other—equivalent to the electron transfer process in a bond. Consequently, we can consider this probability as a measure of the spreading of the atoms (taken as discrete objects) in the space. On this basis, we have defined the *stochastic Minkowski distance matrix* (${}_{ss}\mathbf{G}^{n,p}$), which can be obtained from $\mathbf{G}^{n,p}$ as follows:

$$ {}_{ss}g_{\mathrm{ij}}^{\mathrm{n,p}} = \frac{g_{\mathrm{ij}}^{\mathrm{n,p}}}{\sum_{\mathrm{j}} g_{\mathrm{ij}}^{\mathrm{n,p}}} \tag{7} $$

These matrices are not necessarily symmetric, therefore, in an interaction process between atoms i and j, the probability of an electron to move from atom i to j could be different to the probability to move from j to i, e.g. the transfer of electrons in a carbonyl bond is more probable from carbon to oxygen than the reverse process. Additionally, with the aim of equalizing the probabilities in both senses [77], we have proposed the use of a *doubly-stochastic matrix*, defined as a matrix with real nonnegatives entries whose column and row sums are 1 [78,79]. Henceforth, these matrices are referred to *doubly stochastic Minkowski distance matrices* (${}_{ds}\mathbf{G}^{n,p}$).

The process to find the doubly stochastic matrix associated to a non-stochastic matrix is not trivial. According to Sinkhorn, a strictly positive matrix $\mathbf{A}$ can be scaled to a doubly stochastic matrix $\mathbf{B}$ by,

$$ \mathbf{B} = \mathbf{D} \times \mathbf{A} \times \mathbf{D} \tag{8} $$

where, $\mathbf{D}$ is a diagonal matrix [80]. In 1967, Sinkhorn and Knopp extended this theorem to nonnegative matrices and, also proposed the well-known iteration algorithm for matrix balancing that bears their names [81]. Finally, Johnson et al. [82], have considered the problem of the scaling euclidean distance matrices to doubly stochastic matrices, demonstrating that it is possible to scale them using the Eq. (8). On this basis, it is possible to find a *doubly stochastic Minkowski distance matrix* (${}_{ds}\mathbf{G}^{n,p}$) from a *non-stochastic Minkowski distance matrix* ($G^{n,p}$) through the Eq. (8) and the Sinkhorn–Knopp algorithm.

Finally, the *Minkowski distance-based mutual probability* matrix is introduced. The elements ${}_{mp}g_{ij}^{n,p}$ are obtained as follows:

$$ {}_{mp}g_{ij}^{n,p} = \frac{g_{ij}^{n,p}}{m(S)} = \frac{g_{ij}^{n,p}}{\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}^{n,p}} \tag{9} $$

where, ${}_{mp}g_{ij}^{n,p}$ denotes the mutual probability between vertices i and j, and m(S) the *sample space*. The sample space is computed by summing all elements of $\mathbf{G}^{n,p}$.

As an extension, these matrix approaches should be suitable for representing a molecular fragment, L, of the whole molecule. In this sense, it is possible obtain the structural information matrix of the molecular fragment L, $\mathbf{G}_L^{n,p}$ from the total matrix, $\mathbf{G}^{n,p}$:

$$\begin{aligned} g_{ij}(L) &= g_{ij} \quad \text{if } i \wedge j \in \text{L} \\ &= \frac{1}{2} g_{ij} \quad \text{if } i \vee j \in \text{L} \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{10}$$

Consequently, if a molecule is partitioned into Z molecular fragments, the total matrix can be partitioned in Z local matrices, i.e., the total matrix can be expressed as the sum of the local matrices of the Z fragments.

2.3 Definition of new descriptors: 3D-linear descriptors of the Minkowski distance matrices

If a molecule consists of m atoms, then the p^{th} linear indices of order n for atom i in a molecule are calculated as linear maps in R^m (endomorphism in R^m), in canonical basis set as shown in Eq. (2). Specifically, the p^{th} non-stochastic, stochastic, doubly-stochastic as well as mutual probabilistic Minkowski distance 3D-linear descriptors of order n are computed from their p^{th} non-stochastic, stochastic and doubly-stochastic and mutual probabilistic Minkowski distance matrices of order n as shown in Eqs. (11)–(14), respectively:

$$f_i^{n,p}(\mathbf{x}) = \sum_{\mathrm{j}=1}^{m} g_{ij}^{n,p} x_{\mathrm{j}} = \mathbf{G}^{n,p}[\vec{\mathbf{x}}] \tag{11}$$

$${}_{ss}f_{\mathrm{i}}^{\mathrm{n,p}}(\mathbf{x}) = \sum_{\mathrm{j}=1}^{\mathrm{m}} {}_{ss}g_{\mathrm{ij}}^{\mathrm{n,p}} x_{\mathrm{j}} = {}_{ss}\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{12}$$

$${}_{ds}f_{\mathrm{i}}^{\mathrm{n,p}}(\mathbf{x}) = \sum_{\mathrm{j}=1}^{\mathrm{m}} {}_{ds}g_{\mathrm{ij}}^{\mathrm{n,p}} x_{\mathrm{j}} = {}_{ds}\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{13}$$

$${}_{mp}f_{\mathrm{i}}^{\mathrm{n,p}}(\mathbf{x}) = \sum_{\mathrm{j}=1}^{\mathrm{m}} {}_{mp}g_{\mathrm{ij}}^{\mathrm{n,p}} x_{\mathrm{j}} = {}_{mp}\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{14}$$

where, m is the number of atoms in the molecule, and x_j are the coordinates of the molecular vector ($\vec{\mathbf{x}}$) in the so-called canonical ("natural") basis set. In this basis set, the coordinates of any vector coincide with the components of this vector [83–85]. Therefore, these coordinates can be considered as weights (labels) of the atom-atom distance.

Note that atom-level linear indices are defined as a linear transformation $f_i^{n,p}(\vec{\mathbf{x}})$ in the molecular vector space R^m. This map is a correspondence that assigns a vector $f^{\mathrm{n,p}}(\vec{\mathbf{x}})$ to every vector $\vec{x}$ in R^m, in such a way that:

$$f(\lambda_1 \vec{\mathbf{x}}_1 + \lambda_2 \vec{\mathbf{x}}_2) = \lambda_1 f(\vec{\mathbf{x}}_1) + \lambda_2 f(\vec{\mathbf{x}}_2) \tag{15}$$

for any λ_1, λ_2 and vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$ in R^m.

The total (whole molecule) atom-based non-stochastic, stochastic, doubly-stochastic as well as mutual probabilistic linear indices, $f^{n,p}(\vec{\mathbf{x}})$, ${}_{ss}f^{\mathrm{n,p}}(\vec{\mathbf{x}})$ and ${}_{ds}f^{\mathrm{n,p}}(\vec{\mathbf{x}})$, ${}_{mp}f^{\mathrm{n,p}}(\vec{\mathbf{x}})$, are calculated from atomic linear indices as shown in Eqs. (16)–(19), respectively:

$$f^{n,p}(\vec{\mathbf{x}}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} f_{\mathrm{i}}^{\mathrm{n,p}}(\vec{\mathbf{x}}) = [\vec{\mathbf{u}}]\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{16}$$

$${}_{ss}f^{\mathrm{n,p}}(\vec{\mathbf{x}}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} {}_{ss}f_{\mathrm{i}}^{\mathrm{n,p}}(\vec{\mathbf{x}}) = [\vec{\mathbf{u}}]_{ss}\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{17}$$

$${}_{ds}f^{\mathrm{n,p}}(\vec{\mathbf{x}}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} {}_{ds}f_{\mathrm{i}}^{\mathrm{n,p}}(\vec{\mathbf{x}}) = [\vec{\mathbf{u}}]_{ds}\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{18}$$

$${}_{mp}f^{\mathrm{n,p}}(\vec{\mathbf{x}}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} {}_{mp}f_{\mathrm{i}}^{\mathrm{n,p}}(\vec{\mathbf{x}}) = [\vec{\mathbf{u}}]_{mp}\mathbf{G}^{\mathrm{n,p}}[\vec{\mathbf{x}}] \tag{19}$$

Finally, in addition to the total and atomic 3D-linear indices computed for each atom in the molecule, local-fragment (atom-type or group) formalism can be developed. The k^{th} atom-type 3D-linear index is calculated by adding the k^{th} atomic 3D-linear indices for all atoms of the same type in the molecule. To be precise, this extension of the atom-level linear index is similar to the group additive scheme, in which an index appears for each atom type in the molecule, together with its contribution based on the atom linear index. Consequently, if a molecule is partitioned into Z molecular fragments, the total linear indices can be partitioned into Z local linear indices, L = 1,…, Z. Furthermore, the total 3D-linear indices can be expressed as the sum of the local 3D-linear indices of the Z fragments:

$$f(\vec{\mathbf{x}}) = \sum_{\mathrm{i}=1\cdots Z} f_{\mathrm{i}}(\vec{\mathbf{x}}) \tag{20}$$

In the atom-type (or group) 3D-linear index formalism, each atom in the molecule is classified into an atom-type (fragment). To this effect, atoms may be classified into atom types, in terms of the characteristics of the two atoms that define the bond. For all data sets, including those with a common molecular scaffold as well as those with very diverse structures, the fragment linear-indices provide a lot of useful information. Thus, the development of the atom-type and group 3D-linear indices provides the basis for application to a wider range of biological problems, in which the local formalism is adequate, without the need for superposition of a closely related set of structures.

2.4 Sample calculation

Up to the preceding section, the theoretical framework of the novel 3D-linear indices has been described. It is natural to perform a calculation on a molecule

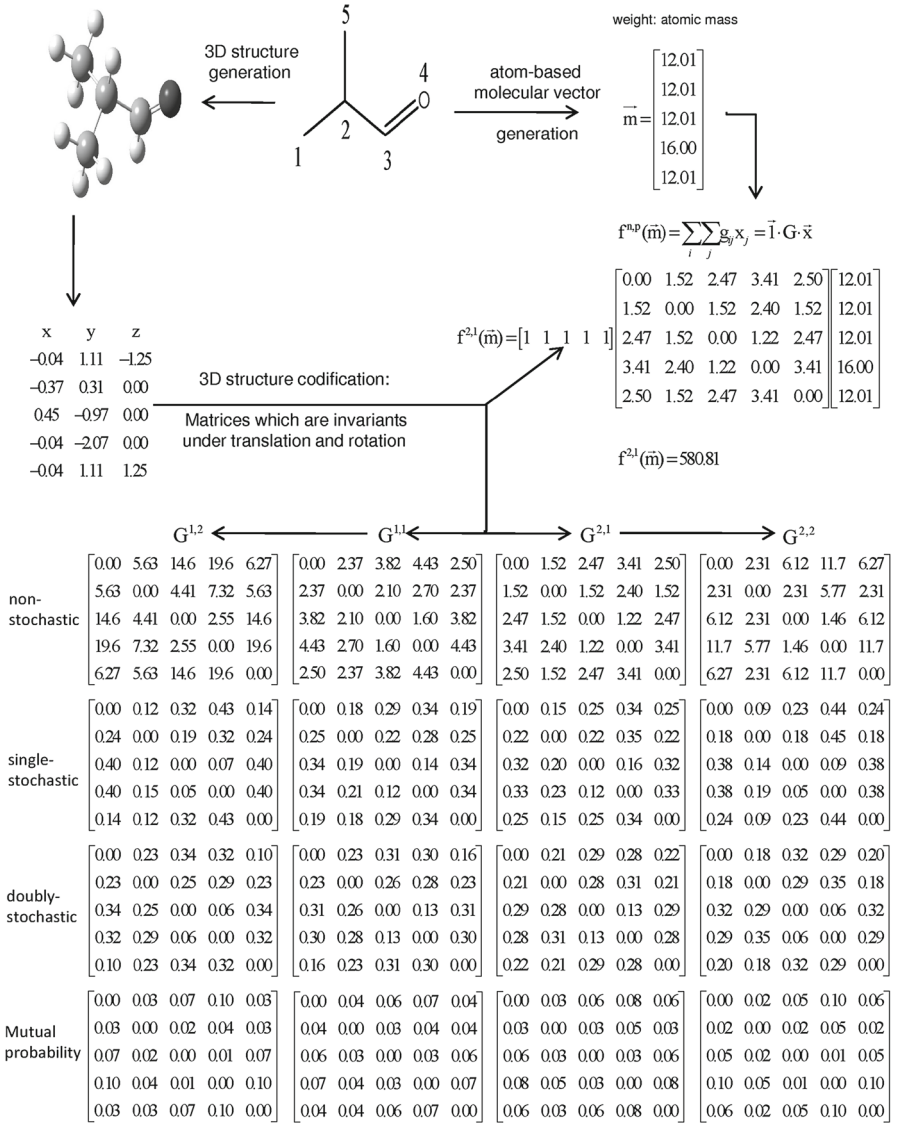

Fig. 1 Workflow for 3D-linear indices (total MDs) calculation

to illustrate the steps in the procedure. To this end, we depict a pictorial representation of the computation of the non-stochastic, stochastic, doubly-stochastic and mutual probabilistic 3D-linear indices (both total and local), using a simple chemical example. Considering the molecule of 2-methyl-propanaldehyde as a simple example, we illustrate the workflow for the proposed methodology in Fig. 1.

3 Analysis of molecular information captured by the proposed 3D-linear indices and their linear dependence

In this section, we compare the information contained in the 3D-linear and DRAGON's software MDs [86]. For this analysis, we use 41molecules of DRAGON's sample data (methane not considered, see Table 1). As can be observed, though this chemical data is relatively small, it is rather heterogeneous, thus allowing the comparison of the information codified by MDs. The descriptor calculations were performed using QuBiLs-MiDAS (quadratic, bilinear and linear maps based on Minkowski distance matrices and atomic weightings), a new module of TOMOCOMD-CARDD program that offers fast and low-computational-cost calculations of the proposed MDs. Note that mutual probabilistic 3D-linear indices were neither considered neither for this study nor in the ensuing sections.

In order to conduct this study, we carry out *factor analysis* using the principal components method. This is a versatile data analysis method for summarizing the information contained in several variables into a small number of weighted composites. The theoretical aspects of this statistical technique have been extensively explained elsewhere [87–91]. The general objectives of factor analytical techniques are (1) data *reduction* and (2) *interpretation* of the underlying relationship between variables, i.e., to *classify* variables. In this context, factor loadings (or "new" variables) are obtained from original (MDs) variables [87–91]. These factors capture most of the "essence" of the MDs because they are a linear combination of the original items. Because each factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are orthogonal to each other. Therefore, the first factor is generally more highly correlated with the variables than the other factors. Some of the most valuable conclusions that can be drawn from factor analysis are: (1) variables with high loadings in the same factor are correlated and this correlation will be greater the higher the loadings, (2) no correlation exists between variables having nonzero loadings in only different factors. The existence of linear independence has been claimed by Randić as one of the desirable attributes for novel TI's [40].

Factor analysis is performed with the STATISTICA software [92] and "varimax normalized" is used as the rotational strategy to obtain the factor loadings from the principal component analysis [92]. The goal of this rotational procedure is to obtain a clear pattern of loadings, i.e., factors that are clearly marked by high loadings for some variables and low loadings for others. This rotation strategy maximizes the variances of the square *normalized factor loadings* (row factor loadings divided by square roots of the respective communalities) across variables for each factor. This strategy makes the factors pattern structure as simple as possible, permitting a clearer interpretation of the factors without loss of orthogonality among them. In this analysis, factor loadings greater than 0.60 are considered.

The table reflecting the factor loadings of all the MDs used in this study is available as supporting information (Table S2).

Table 2 shows the eigenvalues and the percentages of the explained variance by 8 principal factors of this analysis, which explain approximately 91.05 % of the cumulative variance.

Table 1 Sample of 3D-linear indices values (eight best variables, in entropy terms) in QuBiLs-MIDAS software obtained for a dataset of 41 structurally diverse molecules

Compound	${}^{*}_{SS}f^{1,10}(r)$	${}^{*}_{SS}f^{1,8}(\alpha)$	${}^{*}_{SS}f^{1,11}(r)$	${}^{*}_{SS}f^{3,8}(r)$	${}^{*}_{SS}f^{3,4}(\alpha)$	${}^{*}_{SS}f^{1,5}(\alpha)$	${}^{*}_{SS}f^{1,9}(\alpha)$	${}^{*}_{SS}f^{3,9}(r)$
n-ethane	54.838	5.511	54.457	59.162	6.267	5.848	5.456	57.969
n-propane	63.160	5.840	62.352	66.448	6.588	6.182	5.782	65.115
n-butane	63.482	6.282	62.758	67.345	6.918	6.634	6.215	65.709
n-pentane	69.327	6.381	67.880	70.948	7.335	6.972	6.274	68.313
n-hexane	71.762	6.785	71.234	73.992	8.022	7.624	6.617	72.587
Isobutane	73.378	7.203	71.826	77.547	8.368	7.797	7.085	73.834
Neopentane	73.606	7.363	72.375	78.211	8.425	8.022	7.221	75.490
2-Methylpentane	73.830	7.399	72.604	79.652	8.813	8.146	7.229	77.276
Cis-2-butene	78.374	7.744	77.278	83.178	8.936	8.306	7.667	81.421
Trans-2-butene	80.423	7.881	79.556	83.539	8.976	8.498	7.754	81.491
2-Butyne	81.944	8.537	80.106	88.851	9.550	9.077	8.435	86.961
Cyclopropane	85.900	8.607	84.475	92.023	9.660	9.222	8.480	89.753
Cyclobutane	87.315	8.686	85.927	92.602	9.697	9.569	8.511	90.402
Cyclopentane	87.540	8.817	86.116	93.099	10.055	9.663	8.625	90.572
Cyclohexane	88.837	8.855	87.905	94.490	10.236	9.698	8.635	91.587
Cyclohexanone	90.610	9.008	88.212	97.926	10.273	9.834	8.767	94.177
Benzene	90.859	9.027	89.282	98.124	10.509	10.175	8.853	95.374
Toluene	93.124	9.197	90.577	100.720	10.861	10.304	8.921	98.284
Phenol	95.031	9.469	93.422	103.544	11.157	10.518	9.309	99.991

Table 1 continued

Compound	${}^{*}_{SS}f^{1,10}(r)$	${}^{*}_{SS}f^{1,8}(\alpha)$	${}^{*}_{SS}f^{1,11}(r)$	${}^{*}_{SS}f^{3,8}(r)$	${}^{*}_{SS}f^{3,4}(\alpha)$	${}^{*}_{SS}f^{1,5}(\alpha)$	${}^{*}_{SS}f^{1,9}(\alpha)$	${}^{*}_{SS}f^{3,9}(r)$
Benzoic acid	100.400	9.903	97.895	108.105	11.536	11.026	9.641	105.404
Aniline	101.847	10.213	100.170	109.521	11.791	11.107	10.033	105.753
Nitrobenzene	102.520	10.275	100.756	115.043	11.901	11.155	10.087	112.060
Fluorobenzene	106.857	10.670	104.161	117.956	12.737	11.981	10.391	113.850
Chlorobenzene	110.924	10.968	108.735	117.989	12.741	12.206	10.632	114.509
Bromobenzene	114.355	11.146	111.318	123.874	12.965	12.324	10.910	118.029
Iodobenzene	115.313	11.302	111.602	130.815	13.166	12.462	11.063	126.188
Benzamide	119.971	11.493	116.914	133.675	13.991	12.876	11.196	129.476
Naphthalene	125.117	11.876	123.029	134.629	14.122	13.224	11.588	130.301
Anthracene	129.215	11.992	126.400	139.925	14.181	13.248	11.680	136.566
Pyrrole	132.667	12.508	130.145	143.584	14.544	13.510	12.287	139.071
Furan	132.825	12.912	130.818	144.901	14.845	14.085	12.594	139.295
Thiophen	136.632	13.105	133.302	145.986	15.071	14.315	12.857	142.293
Purine	136.747	13.224	133.499	149.845	15.089	14.549	12.900	146.037
Dibenzofuran	138.815	13.694	135.673	151.307	15.530	15.035	13.373	150.173
Ethanol	144.577	14.492	141.990	168.815	17.169	15.446	14.238	161.338
Trifluoroethanol	154.808	15.356	149.183	172.240	17.528	16.776	14.827	167.927
2-Aminoethanol	158.102	15.584	154.685	178.496	17.879	17.105	15.273	173.953
Propanol	165.182	16.046	161.762	187.024	18.354	17.446	15.756	187.688
Propanone	174.624	18.926	173.017	217.424	22.644	21.381	18.272	207.958
2-Propanol	194.113	21.252	186.730	226.425	24.236	21.384	21.159	228.895
2-Propylamine	227.413	22.032	218.836	254.831	26.232	24.769	21.287	244.186

Table 2 Results of the factor analysis by using the principal component method for 0D–3D DRAGON MDs as well as the total and local (atom type) 3D-linear indices for 41 heterogeneous chemicals

Factor	Eigenvalue	% Total variance	Cumulative eigenvalue	Cumulative (%)
F1	2504.15	62.60	2504.15	62.60
F2	466.60	11.67	2970.76	74.27
F3	256.27	6.41	3227.03	80.68
F4	148.04	3.70	3375.07	84.38
F5	89.40	2.24	3464.47	86.61
F6	69.37	1.73	3533.84	88.35
F7	58.34	1.46	3592.18	89.80
F8	50.01	1.25	3642.19	91.05

With a simple examination of the principal components, it is intuitive that the TOMOCOMD-CARDD (namely, QuBiLs-MiDAS) MDs are strongly loaded in factor one (62.60 %) and factor three (6.41 %); while factor two (11.67 %) is particularly important for DRAGON's MDs with robust and exclusive factor loadings. This result suggests that there exists orthogonality between the novel the 3D-linear indices and DRAGON's MDS as a whole. In other words, the former codify structural information not codified by the latter, which rationalizes the contribution of the new mathematical approach in the codification of the geometric space of a molecular structure. In order to have deeper comprehension of the relationship among the 3D-linear indices, a more painstaking analysis of factors one and three (collectively explain 69.01 % of the total variance) is performed, revealing curious behavior. First of all, total and fragment-based 3D-linear indices for aromatic rings (ARM), sp^2 (SP2) and sp^3 (SP3) hybridized carbons, pnictides or nitrogen group elements (NIT), methine (CH), methylene (CH2) and methyl groups (CH3) are strongly loaded in Factor one which suggests the existence of collinearity among the indices defined for these atom-types. Logical explanations could be given to this outcome: (1) the presence of sp2 hybridized ring carbons is a necessary, but insufficient condition, for the existence of aromaticity and it is thus natural that collinearity exists between these two groups, (2) methyl carbons are sp3 hybridized, while methine and methylene carbons can be sp^2 or sp^3 hybrids. Factor three seems to be relevant to fragment-based 3D-linear indices for heteroatoms (HET) given their robust loadings in this factor. The rest of the atom-based based 3D-linear indices present comparable representativity in both factors with no clear-cut exclusiveness. Another important aspect that could be highlighted is that for any group of indices calculated over H-suppressed molecular graphs loaded in a particular factor, analogous indices over H-explicit molecular graphs (MG) are loaded in the same factor with representativity of equal magnitude. This outcome suggests that there is no orthogonality in terms of captured structural information when H-suppressed (leftsuperscript $= *$) or H-explicit MG are used. A yet keener scrutiny of Table 2 reveals another interesting and consistent pattern. The majority of the 3D-Linear index groups (with respect to the atom- and matrix-types) derived from step-counts between 1 and 6 (associated to short-range non-covalent interatomic

interactions) are loaded in Factor three while those from step-counts greater than 6 (long-range interactions) in Factor one. This suggests that there exists a 'critical point' in the force-interatomic distance curve beyond which there is a change in the electronic potential which could affect in the spatial configuration of a molecular structure, leading to the gain of "new" structural information.

From this study two fundamental conclusions are drawn. (1) The different approaches used in the definition of the 3D-linear indices contribute to generation of orthogonal MDs. (2) The proposed 3D-linear indices codify structural information not described by DRAGON's MDs, a whole.

3.1 Variability analysis of the proposed atom-based 3D-linear indices

The identification of the most suitable variables from a high-dimensional MD space to be incorporated in a computational model is a non-trivial challenge because an exhaustive exploration of the entire descriptor space is time-consuming and unpractical. Several dimension reduction approaches have been reported in the literature [40,93]. Principally, these methods seek to filter out the MDs most representative for a molecular data set. However, most of these methods are based on the assumption of linearity, which is not necessarily fulfilled for a given MD space.

Godden et al. [68] 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. Since MD value ranges may substantially differ for a given data set, to guarantee comparability, the first step in this approach is to apply a scaling procedure (binning scheme) forming histograms of descriptor distributions (*equal interval width method*). It is to the resulting uniform data distribution that Shannon's fundamental Eq. [94] is subsequently applied.

With the aim of evaluating the quality of the MDs proposed in the present report and demonstrate the potential of these 3D indices as a reliable tool in chemoinformatic studies, some of the authors implemented this innovative application of information theory to variability analysis in an interactive software denominated *IMMAN* (acronym for Information Theory based chemometric analysis) [95], enriched with additional parameters, derived from modifications of Shannon's entropy as: standardized Shannon's entropy (sSE), Negentropy (nSE), Brillouin Redundancy Index (rSE), Gini index (gSE) and Information Energy Content (iSE), previously not used in the evaluation of the variability of MDs [40].

The study carried out in this section was sub-divided in four main parts: (1) matrix-oriented analysis of the 3D-linear indices (2) comparison between 3D- and 2D-linear indices (3) Family-wise evaluation of DRAGON and QuBiLs-MiDAS 3D-linear indices (4) QuBiLs-MiDAS (3D-linear indices) software *versus* other MD calculating packages.

A binning scheme of 41 intervals (bins) was used for the SE computation of the MDs. For this discretization scheme, the maximum entropy (Hartley's entropy) is given by $\log_2 41 = 5.358$ bits. 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. For the rest of the classes, the best

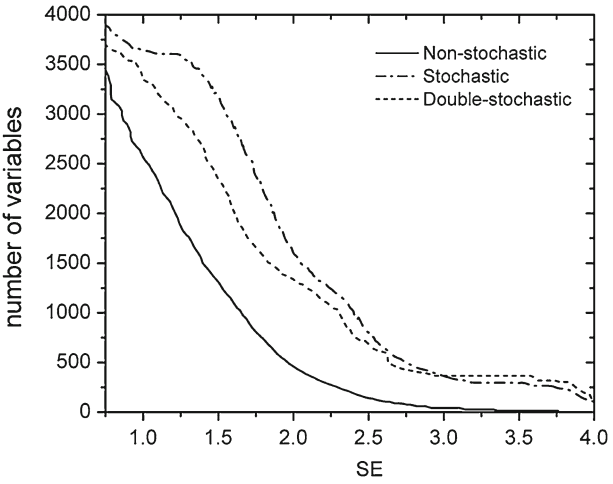

Fig. 2 Shannon's entropy distribution for non-, simple-, and doubly-stochastic 3D-linear indices

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. [96] as this gives a subjective graphic perspective when comparing cases with markedly unequal number of variables. However, in the case that the same number of variables be used, the probability scale could be used as well.

3.2 Comparative analysis of 3D-linear indices for non-, simple-, and doubly-stochastic matrix approaches

The purpose of the present study is to evaluate the contribution, in terms of the variability, of the different matrix-based approaches i.e. non-, simple-, and doubly-stochastic matrices in the definition of the 3D-linear indices. Figure 2 illustrates a graphical comparison of Shannon's distribution for the best 4000 variables of non-, simple-, and doubly-stochastic 3D-linear indices. Comparable behavior is observed for higher entropy (>3.0 bits) simple-stochastic and no-stochastic linear indices although below this value a marginally better distribution pattern for simple-stochastic linear indices is observed. On the other hand, the two groups of indices present better distribution patterns than double-stochastic 3D-linear indices. This result suggests that a greater percentage of highly variable MDs are obtained with simple-stochastic and non-stochastic matrix-based approaches than the double-stochastic formalism.

3.3 3D-linear indices versus 2D counterpart

In previous reports [97,98], Marrero-Ponce et al. defined 2D-linear indices calculated as linear maps on R^m. Bearing in mind that this report focuses on a dimensional extension of these indices, an analysis of the possible contribution, in variability terms,

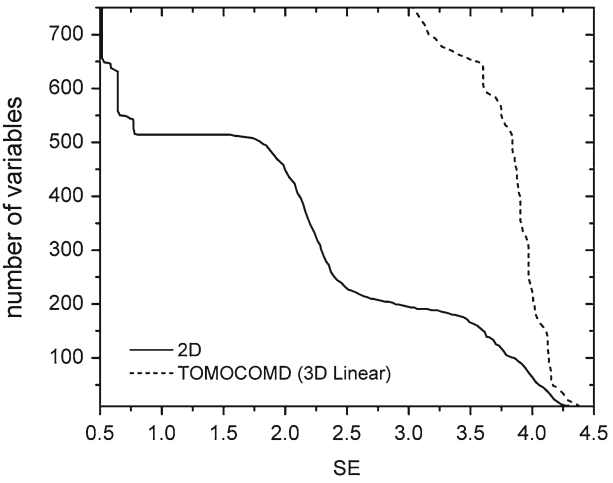

Fig. 3 Shannon's entropy distribution for 3D- versus 2D-linear indices

of this geometric approach is necessary. Figure 3 shows the Shannon's distribution graph of the best 768 variables for two families of indices (cut-off number provided by 2D-linear indices). As can be seen a strikingly better entropy distribution is observed with 3D-linear indices in comparison to the 2D-linear indices, with 85 % of the former presenting entropy values over 3.5 bits (66 % of maximum entropy) in comparison with 21 % of the latter at the same level. This result suggests that the incorporation of information on the spatial configuration to the linear index formalism improves the global variability of the linear indices, and thus better discriminating power for molecular datasets may be obtained.

3.4 Family-wise comparison of DRAGON and the 3D-linear indices

The DRAGON software, one of the most popular packages used in QSAR/QSPR studies, is comprised of various MDs families. Here, our goal is to compare the entropies of these descriptor families and the 3D-linear indices. Some DRAGON families were grouped together into bigger *families*, i.e., OD-1D and others (functional group counts, atom-centered fragments, constitutional descriptors and molecular properties), 3D-Indices (charge descriptors, 3D-Morse descriptors, Randić molecular profiles and geometric descriptors), Topo-Indices (topological indices, 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 information indices determining this cut-off number as the family with the fewest number of variables. The 3D linear indices presented comparable entropy distributions with 3D-indices and GATEWAY descriptors. On the other hand, better entropy distributions are observed when the 3D linear indices are compared with the rest of DRAGON's descriptor families (see Fig. 4).

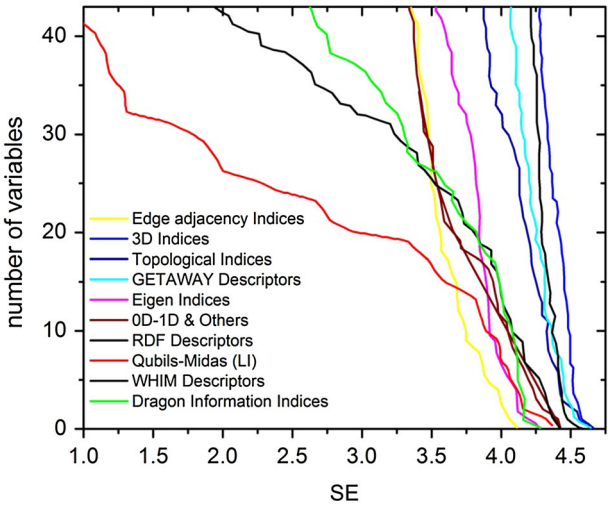

Fig. 4 Shannon's entropy distribution for DRAGON's and 3D-linear descriptor families

3.5 Comparison of QuBiLs-MiDAS (3D-linear indices) software with other descriptor calculation packages

The fourth study constitutes a broader analysis, with the objective of comparing variability of the QuBiLs-MiDAS software program and some of the relevant programs used in for descriptor calculations in chemoinformatics such as: DRAGON [86], MOLd^2 [96], PADEL [99], MODESLAB [88,100–102], BLUECAL [103], MOLCONNZ [104], POWER MV [105], and CDK [106]. The cut-off number of 170 variables was provided by BLUECAL software. The top 25 descriptors obtained for each of these softwares listed in Table 3 show relatively comparable entropy distribution.

Figure 5 is a graphic illustration of Shannon's entropy distribution for these softwares, where QuBiLs-MiDAS software (represented by only 3D-linear indices) demonstrates similar to better entropy distribution than most of the analyzed softwares. For example, the number of MDs with entropy values greater than 4.00 bits are 170 (100 %) for the case of 3D-linear indices in QuBiLs-MiDAS and DRAGON's MDs, 44 (26 %) for MOLD2, 39 (23 %) for CDK, etc.

It is worth noting that, in the case of DRAGON software where comparable behavior is observed, this software encompasses a series of substantially diverse MD families (0D–3D) derived from wide range of chemical and graph-theoretic concepts. This outcome suggests that MDs calculated with the QuBiLs-MiDAS program may capture similar-to-better amount of structural information than the software packages compared in this study, and may possibly be an important tool in QSPR/QSAR and similarity/dissimilarity analysis. Although high variability is a desirable quality for MDs, it is not the ultimate requirement for good correlations with a particular physicochemical, chemical or biological property to be obtained. Therefore, the next section will be devoted to assessing the modeling power of the proposed 3D-linear indices.

Table 3 Shannon entropy for top 25 molecular descriptors calculated with different softwares

BLUECAL	CDK	DRAGON	MODESLAB	MOLD2	MOLZ	PADEL	POWER MV	QuBiLs-MIDAS (3D-linear MDs)
4.653	4.571	4.675	4.257	4.428	5.227	4.623	4.412	4.571
4.601	4.522	4.642	4.220	4.382	4.322	4.391	4.233	4.458
4.556	4.495	4.623	4.181	4.375	4.275	4.322	4.226	4.440
4.428	4.495	4.574	4.145	4.342	4.251	4.306	4.217	4.422
4.425	4.486	4.556	4.092	4.320	4.224	4.257	4.209	4.422
4.410	4.458	4.553	3.983	4.287	4.196	4.251	4.192	4.410
4.391	4.436	4.537	3.940	4.263	4.053	4.199	4.154	4.401
4.379	4.391	4.534	3.929	4.249	4.029	4.185	4.129	4.391
4.358	4.342	4.526	3.928	4.240	4.001	4.163	4.108	4.391
4.356	4.342	4.507	3.922	4.240	3.979	4.063	4.108	4.387
4.342	4.340	4.507	3.915	4.224	3.964	4.028	4.093	4.361
4.309	4.339	4.505	3.894	4.214	3.940	3.991	4.092	4.357
4.275	4.330	4.504	3.815	4.212	3.910	3.986	4.013	4.357
4.263	4.312	4.489	3.813	4.206	3.839	3.983	3.987	4.352
4.251	4.279	4.489	3.794	4.196	3.805	3.969	3.979	4.352
4.249	4.279	4.489	3.791	4.187	3.737	3.964	3.978	4.342
4.226	4.267	4.477	3.780	4.181	3.666	3.948	3.959	4.330
4.212	4.263	4.474	3.754	4.163	3.592	3.931	3.958	4.318
4.194	4.242	4.470	3.744	4.159	3.555	3.882	3.922	4.312
4.181	4.233	4.468	3.736	4.157	3.509	3.861	3.882	4.294
4.181	4.233	4.458	3.725	4.147	3.498	3.748	3.843	4.294
4.175	4.230	4.446	3.721	4.136	3.301	3.676	3.818	4.294

Table 3 continued

BLUECAL	CDK	DRAGON	MODESLAB	MOLD2	MOLZ	PADEL	POWER MV	QuBiLs-MIDAS (3D-linear MDs)
4.166	4.230	4.444	3.705	4.129	3.297	3.669	3.791	4.279
4.166	4.212	4.440	3.696	4.112	3.281	3.463	3.783	4.278
4.163	4.163	4.440	3.672	4.110	3.275	3.462	3.550	4.278
4.316	**4.340**	**4.513**	**3.886**	**4.226**	**3.869**	**4.013**	**4.025**	**4.364**

Numbers in bold are the average Shannon's entropy values

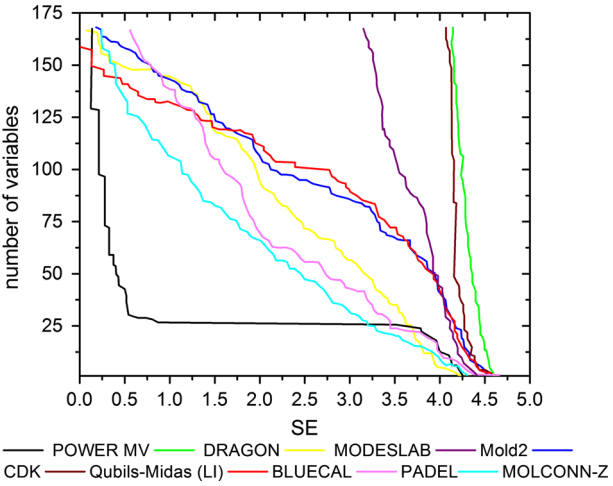

Fig. 5 Shannon's entropy distribution for QuBiLs-MiDAS software and other MD calculating programs

3.6 QSPR modeling of benchmark datasets

QSAR modeling is based on the premise that the properties of molecules are a function of their structural features. Thus proper codification of the information contained in molecular structures should enable the description and prediction of molecular (or molecular-fragment) properties. Consequently, in order to obtain deeper insight of the contribution, if any, of our 3D-linear indices in the codification of molecular structural information, several QSPR modeling tasks of well-known physicochemical properties are performed.

3.7 Datasets

To perform the QSPR studies, we have selected the following datasets to be investigated: (1) 18 octane isomers, (2) 28 alkyl alcohols and, (3) 34 furylethylenes derivatives. The first has been advocated by Randić and Trinajstić [107,108] and used by several researchers to evaluate the modeling power of molecular descriptors [88,109–115]. Presently, it is considered by International Academy of Mathematical Chemistry as one of the benchmark databases for comparing old (well-known) MDs with new ones [107,108]. These datasets are recommended due to the fact that most of the physicochemical properties commonly studied with MDs in QSPR analyses are interrelated for data sets of compounds with different molecular weights. When isomeric data sets are used these correlations with the same descriptors are not necessarily observed.

On the other hand, all MDs are designed to have (gradual) augmentation with increments in the molecular mass. In this way, if we perform the present study by using a series of chemicals having different molecular weights, we may find "false" interrelations between the descriptors by an overestimation of the size effects inherent to these indices [114,116]. The same is also valid when a QSPR model is to be obtained.

In conclusion, if a newly proposed MD is not able to model the variation of at least one property of octanes (FIRST EXPERIMENT), then it probably does not contain any useful molecular information. Precisely, to evaluate the quality of the models based on our MDs, we analyzed the statistical parameters for the best QSPR models obtained in the description of the boiling point (BP), motor octane number (MON), heat of vaporization (HV), molar volume (MV), entropy (S), and heat of formation (ΔH_f) of the octane isomers. The regressions for these models were compared with: (1) parameters for models published by Randić [113–116] based on diverse topological indices such as the Wiener matrix invariants, (2) results for models reported by Diudea [111] based on the SP indices, and (3) parameters for the best models obtained with a set consisting of the topological, WHIM, and GETAWAY descriptors [108].

The SECOND EXPERIMENT focused on the examination of the possibilities of our MDs in the QSPR studies with heteroatomic molecules, selecting the boiling point of 28 alkyl-alcohols, as the property to be investigated [102,117]. This data set was first studied by Kier and Hall [117] using E-state/biomolecular encounter parameters and later by Estrada and Molina [102] employing the local spectral moments of the edge adjacency matrix. This heteromolecule-based database is composed of 28 alkyl-alcohols, in which 14 are primary, 6 secondary and 8 tertiary. The boiling points (Bp) of these compounds have been experimentally determined and reported in the literature. This isomeric dataset of heteroatomic compounds is suitable for comparative studies of MDs, since the boiling point not only depends of gradual variation of molecular weight, but also on the H-bonding capacity and R-group type. Additionally, results of QSPR studies are available for comparison purpose [102,117].

The third dataset, and second heteromolecule-based database, consists of a set of 34 2-furylethylene derivatives (THIRD EXPERIMENT), studied earlier with total and local spectral moments, 2D/3D vertex- and edge- connectivity indices and two quantum-chemical descriptors [102,118]. These chemicals, whose chemical structures are shown in the Table S3, have different substituents at position 5 of the furan ring, as well as at the β position of the exocyclic double bond [119]. The values of the n-octanol/water partition coefficient (log p) and rate constant (log k) (for nucleophilic addition of the mercaptoacetic acid) of these compounds have been experimentally determined and reported in the literature [119], (see Table S3 of the electronic supplementary material). 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 [119]. 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 [12,118]. Thus, a study of these properties, using the proposed MDs, permits us to obtain a general criterion about the applicability of these indices in QSPR studies.

The local fragments considered in the calculation of the local descriptors were designed according to the structural characteristics and the properties of the molecules in the datasets. Table 4 shows a summary of the local fragments used in this work. Additionally, in 2-furylethylenes the substituents R1, R2 and R3 (described in electronic supplementary material Table S3), were also considered for the calculation of local indices.

Table 4 Molecular and atomic fragments used to calculate local descriptors

Fragment	Symbol	Experiment		
		1	2	3
Methyl, methylene, methyne and quaternary carbon	CH3, CH2, CH1, QC	X	X	
Hybridisation	SP3, SP2, SP1	X	X	X
Heteroatoms	HET		X	X
pnictides, Chalcogens and Halogens	NIT, SUL and HAL			X
Individual element	Atomic symbol		X	X
Aromatic fragment	AR			X
H bond-acceptor and donor	HBA and HBD			X

3.8 Descriptor calculations

The molecules contained in each dataset were modeled and their geometries optimized with Mopac2009 software at semi-empirical level using PM6 Hamiltonian [120]. The total and local non-stochastic, stochastic and doubly-stochastic 3D molecular linear descriptors on Minkowski atom-atom distance matrices [Eqs. (16)–(18)] were codified in an experimental version of TOMOCOMD-CARDD software (QuBiLs-MiDAS module) and calculated for the datasets mentioned above. In this work, the norm (n) and step-count (p) were ranged between 1–3 and 1–15, respectively.

3.9 Statistical analysis

With the large number of MDs generated by the TOMOCOMD-CARDD approach, it is difficult to predict which descriptor subsets are most suitable for providing the best regressions, considering both goodness of fit and the chemical meaning of the regression. Therefore, machine learning techniques like genetic algorithms (GAs) and neural networks (NNs) are often used to facilitate descriptor selection, which can be done by systematically exploring various descriptor combinations and further refining those that give best intermediate results. In the present work, we have used GA variable selection [28,121–125]. inspired by the process of Darwinian evolution, in which individuals of high fitness in an initial population prevail and/or survive to the next generations; the best individuals can be adapted by crossover and/or mutation in the search for better individuals.

The software MOBYDIGS (version 1.0—2004) [126] was used to perform variable selection and QSPR modeling. The mutation and reproduction probabilities were fixed at 10 %. The size of the models was set between three and six descriptors plus the independent variable depending on dataset. The population size was established as 100 and a maximum of 10,000 generations were allowed to find an optimal QSPR model.

The models were optimized using as objective function (optimization function) the statistical parameter Q2LOO (“leave one out” cross-validation) and they were

validated using both techniques "bootstrapping" (Q2BOOT) and "y-scrambling" [a (R2), a (Q2)]. The former evaluates the predictive power of the developed models and the latter checks the risk of chance correlations (common occurrence when too many variables are screened relative to the number of available observations) [126,127]. The selection of the best model was processed in terms of the highest determination coefficient (R2), F test (Fisher ratio's p-level [p(F)]) or leave-one-out (LOO) cross-validation (Q^2_{LOO}), and the lowest standard deviation (s).

3.10 QSPRs and comparison with other MDs

3.10.1 Experiment 1

Here, several physicochemical properties of the octane isomers were analyzed. However, to evaluate the quality of the models based on our new atom-level chemical descriptors we have taken as reference only six physicochemical properties selected in the previous study [128]. The regressions of octane properties [boiling point (BP), motor octane number (MON), heat of vaporization (HV), molar volume (MV), entropy (S), and heat of formation (ΔfH)], based on the non-, simple- and doubly-stochastic atom-based 3D linear indices, will be compared to some regressions based on 2D (topological/topo-chemical) and 3D (geometrical) MDs, taken from the literature [128].

The best models, found using our 3D-linear indices are presented in Table 5. For each selected property of octane isomers, the statistical information for the best regressions with 1, 2, and 3 MDs published so far [128] are also depicted in Table 5, together with the LOO cross-validation-explained variance (Q2LOO), the correlation coefficient (R2), the standard deviation the error (s), and Fischer ratio (F) are listed.

As can be appreciated from the statistical parameters of regression equations in Table 5, all of the physicochemical properties were adequately described by atom-based 3D-linear indices. From this table it is evident that the statistical parameters for the models, obtained with our MDs to describe motor octane number (MON), molar volume (MV) and heat of vaporization (HV) of octanes, are better than those reported in the literature. Only the models that describe entropy (S) have subtle differences with the precedent models.

According to the obtained QSPR results, it is possible to conclude that the new MDs encode useful molecular information and exhibit considerable diversity, being able to adequately describe the variation in different properties of octanes.

3.10.2 Experiment 2

In this study, we select QSPR models for non-, simple and doubly-stochastic atom-level MDs that best describe the boiling point of 28 alkyl alcohols, represented by

Table 5 Statistical information of best multiple regression models of selected physicochemical properties of 18 octane isomers

Approach	Descriptors	N	R^2	F	s	Q^2_{LOO}
Boiling point (BP)						
Non-stochastic 3D linear indices	$f^{1,2}_{QC}(m)$ $^{*}f^{1,2}_{CH}(m)$ $^{*}f^{1,1}_{CH}(r)$	3	0.944	79.07	1.591	0.916
Stochastic 3D linear indices	$_{ss}f^{1,3}_{CH}(m)$ $^{*}_{ss}f^{3,10}_{QC}(m)$ $^{*}_{ss}f^{1,4}_{CH}(\chi)$	3	0.976	190.1	1.043	0.967
Doubly-stochastic 3D linear indices	$^{*}_{ds}f^{2,6}_{CH}(\alpha)$ $^{*}_{ds}f^{2,6}_{CH2}(m)$ $^{*}_{ds}f^{2,3}_{CH3}(r)$	3	0.939	72.25	1.660	0.889
GETAWAY + WHIM + Topological [128]	$^{2}\chi$ $^{2}\bar{\chi}$ $HATS_6(p)$	3	0.988		0.744	0.981
GETAWAY [128]	$HATS_2(v)$ $R_4(u)$ $R_6(v)$	3	0.983		0.897	0.971
GETAWAY + WHIM + Topological [128]	$^{2}\chi$ $HATS_6(p)$	2	0.978		1.013	0.966
Topological [111]	S_3W S_4W SJ	3	0.958		1.394	
Topological [111]	S_3W S_4W	2	0.948		1.508	
GETAWAY [128]	$HATS_2(m)$ $R^{+}_4(u)$	2	0.896		2.098	0.849
Topological [115]	WW x_1	2	0.814		2.810	
Topological [114]	Z	1	0.789		2.900	
GETAWAY + WHIM + Topological [128]	$HATS_2(m)$	1	0.746		3.175	0.665
Topological [111]	χ_1W	1	0.678		3.630	
Motor octane number (MON)						
Non-stochastic 3D linear indices	$f^{1,15}_{CH3}(\alpha)$ $f^{1,5}(m)$ $^{*}f^{1,4}_{CH3}(m)$	3	0.997	1238	1.572	0.995
Stochastic 3D linear indices	$_{ss}f^{2,2}_{CH3}(\alpha)$ $^{*}_{ss}f^{2,2}_{CH3}(r)$ $_{ss}f^{3,3}(m)$	3	0.991	440.1	2.629	0.985
Doubly-stochastic 3D linear indices	$^{*}_{ds}f^{1,7}_{CH3}(m)$ $^{*}_{ds}f^{2,8}_{CH3}(r)$ $^{*}_{ds}f^{3,9}_{CH3}(\alpha)$	3	0.973	141.2	4.598	0.955
GETAWAY + WHIM + Topological [128]	$^{v}\bar{I}^{M}_{D}$ Ts $HATS_1(m)$	3	0.992		2.439	0.986

Table 5 continued

Approach	Descriptors	N	R^2	F	s	Q^2_{LOO}
GETAWAY [128]	$HATS_4(u)$ $HATS_7(v)$ $R_7(p)$	3	0.986		3.259	0.974
Topological [111]	$S\,\chi^1\,W\,\chi^7\,W\,\chi^3\,W$	3	0.981		3.855	
GETAWAY + WHIM + Topological [128]	$Ts\,H_4(e)$	2	0.977		4.053	0.968
GETAWAY [111]	$HATS_7(m)\,R_4(u)$	2	0.958		5.466	0.913
Topological [111]	$S\,\chi^1\,W\,S\,\chi^3\,W$	2	0.956		5.533	
Topological [111]	$\chi^7 W$	1	0.952		5.589	
GETAWAY + WHIM + Topological [128]	Ts	1	0.924		7.069	0.908
Topological [114]	I_{WD}	1	0.920		7.270	
GETAWAY [128]	REIG	1	0.890		8.515	0.856
Heat of vaporization (HV)						
Non-stochastic 3D linear indices	$f^{2,9}_{QC}(m)$ $^{*}f^{1,3}_{QC}(m)$ $^{*}f^{1,2}_{CH}(m)$	3	0.973	168.0	0.367	0.961
Stochastic 3D linear indices	${}_{ss}f^{1,3}_{CH}(m)$ ${}^{*}_{ss}f^{3,10}_{QC}(m)$ ${}^{*}_{ss}f^{1,4}_{CH}(\chi)$	3	0.984	292.5	0.280	0.974
Doubly-stochastic 3D linear indices	${}_{ds}f^{1,3}_{CH3}(r)$ ${}^{*}_{ds}f^{2,2}_{QC}(\chi)$ ${}^{*}_{ds}f^{1,1}_{CH2}(\chi)$	3	0.973	168.9	0.366	0.958
GETAWAY + WHIM + Topological [128]	${}^{0}\bar{\chi}$ ${}^{3}\kappa$ $R^{+}_{6}(u)$	3	0.984		0.281	0.976
GETAWAY [128]	$HATS_6(u)\,R_4(u)\,R^{+}_{1}(m)$	3	0.972		0.375	0.955
GETAWAY + WHIM + Topological [128]	${}^{2}\chi$ $R^{+}_{6}(u)$	2	0.965		0.402	0.952
Topological [111]	$\chi^1\,W\,\chi^2\,W\,\chi^3\,W$	3	0.957		0.459	
GETAWAY [128]	$HATS_4(u)\,R_6(e)$	2	0.949		0.488	0.932
Topological [111]	${}^{4}W^{5}W$	2	0.926		0.577	
Topological [114]	Z	1	0.918		0.429	

Table 5 continued

Approach	Descriptors	N	R^2	F	s	Q^2_{LOO}
GETAWAY + WHIM + Topological [128]	χ^2	1	0.886		0.705	0.808
GETAWAY [128]	$R_2(m)$	1	0.857		0.790	0.797
Topological [115]	WW x_1	1	0.843		0.820	
Molar volume (MV)						
Non-stochastic 3D linear indices	$f^{2,4}_{QC}(\chi)$ $f^{2,2}_{QC}(\alpha)$ $f^{2,4}_{QC}(\alpha)$	3	0.909	46.81	1.944	0.865
Stochastic 3D linear indices	${}^{*}_{ss}f^{3,1}_{QC}(m)$ ${}^{*}_{ss}f^{1,15}_{QC}(\chi)$	2	0.921	87.95	1.176	0.931
Doubly-stochastic 3D linear indices	${}_{ds}f^{2,15}_{CH2}(m)$ ${}_{ds}f^{1,5}_{CH3}(\alpha)$ ${}^{*}_{ds}f^{2,1}_{QC}(\chi)$	3	0.967	136.0	1.749	0.944
GETAWAY + WHIM + Topological [128]	Ks R^{+}_{6}(u) $RT^{+}(m)$	3	0.920		1.825	0.760
GETAWAY [128]	$HATS_6(p)$ $RT^{+}(m)$ $R_1(v)$	3	0.903		2.008	0.693
Topological [111]	${}^{5}W{}^{6}W{}^{7}W$	3	0.883		2.210	
GETAWAY + WHIM + Topological [128]	${}^{v}\bar{I}^{M}_{D}$ R^{+}_{6}(u)	2	0.850		2.419	0.545
GETAWAY [128]	R^{+}_{6}(u) $R_4(v)$	2	0.818		2.662	0.455
GETAWAY + WHIM + Topological [128]	$R_6(v)$	1	0.676		3.437	0.327
Topological [111]	${}^{3}W{}^{4}W$	2	0.628		3.807	
Topological [111]	${}^{7}W$	1	0.609		3.780	
Entropy (S)						
Non-stochastic 3D linear indices	${}^{*}f^{2,8}_{CH3}(\alpha)$ ${}^{*}f^{1,2}(m)$ ${}^{*}f^{1,1}(\alpha)$	3	0.961	115.6	0.981	0.939
Stochastic 3D linear indices	${}_{ss}f^{1,12}_{QC}(m)$ ${}_{ss}f^{1,12}_{QC}(\chi)$ ${}^{*}_{ss}f^{1,9}_{CH}(\chi)$	3	0.972	164.9	0.825	0.954
Doubly-stochastic 3D linear indices	${}_{ds}f^{3,13}_{CH3}(m)$ ${}_{ds}f^{3,13}_{CH3}(\alpha)$	2	0.927	95.40	1.298	0.905

Table 5 continued

Approach	Descriptors	N	R^2	F	s	Q^2_{LOO}
GETAWAY + WHIM + Topological [128]	${}^{v}\bar{I}^{E}_{D,deg}$ TWC R^{+}_{2}(p)	3	0.980		0.711	0.972
GETAWAY + WHIM + Topological [128]	${}^{v}\bar{I}^{E}_{D,deg}$ TWC	2	0.971		0.814	0.964
GETAWAY [128]	ISH $HATS_8$(m) R_3(v)	3	0.958		1.016	0.935
GETAWAY [128]	ISH R_3(v)	2	0.948		1.101	0.922
GETAWAY + WHIM + Topological [128]	R_3(v)	1	0.925		1.274	0.899
Topological [114]	$\chi^{[1/2]}$	1	0.911		1.400	
Topological [111]	$x_1 x_2$	2	0.817		2.060	
Heat of formation (ΔH_f)						
Non-stochastic 3D linear indices	$f^{1,3}_{CH3}$(m) $*f^{2,4}_{CH}(\chi)$ $*f^{1,8}_{CH3}$(r)	3	0.814	20.39	0.403	0.745
Stochastic 3D linear indices	${}_{ss}f^{1,4}_{CH2}$(m) ${}^{*}_{ss}f^{2,3}_{CH2}(\chi)$ ${}^{*}_{ss}f^{3,8}_{CH3}$(m)	3	0.907	45.29	0.285	0.857
Doubly-stochastic 3D linear indices	${}^{*}_{ds}f^{1,7}_{CH}$(m) ${}^{*}_{ds}f^{1,11}_{CH}$(r) ${}^{*}_{ds}f^{1,1}(\chi)$	3	0.892	38.58	0.307	0.799
GETAWAY + WHIM + Topological [128]	$HATS_5$(m) $HATS_7$(m) R_4(e)	3	0.966		0.254	0.951
GETAWAY + WHIM + Topological [128]	${}^{2}\chi$ $HATS_2$(e)	2	0.932		0.346	0.910
GETAWAY [128]	$HATS_7$(u) R_2(m)	2	0.929		0.356	0.902
GETAWAY + WHIM + Topological [128]	$HATS_2$(m)	1	0.893		0.421	0.872
Topological [113]	Ω_1 Ω_2 Ω_3	3	0.871		0.492	
Topological [113]	Ω_1 Ω_2	2	0.869		0.478	
Topological [114]	$1/{}^{2}\chi$	1	0.867		0.471	
Topological [111]	WW x_1	2	0.787		0.570	

Eqs. (21–23), respectively:

$$\begin{aligned}\mathbf{BP}(^{\circ}\mathbf{C}) = {} & 54.3(\pm 2.8) + 3.36(\pm 0.18) \times 10^{-1} \cdot {}^{1,1}\mathrm{f}_{\mathrm{O}}(\vec{\mathrm{r}}) \\ & + 1.73(\pm 0.26) \times 10^{-2} \cdot {}^{2,1}\mathrm{f}(\vec{\mathrm{m}}) - 2.69(\pm 0.25) \times 10^{-3} \cdot {}^{*1,1}\mathrm{f}_{\mathrm{CH3}}(\vec{\mathrm{r}}) \\ & - 3.56(\pm 0.22) \times 10^{-2} \cdot {}^{*1,2}\mathrm{f}_{\mathrm{O}}(\vec{\chi})\end{aligned} \tag{21}$$

$$\mathrm{N} = 28 \quad \mathrm{R}^2 = 0.993 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.990 \quad \mathrm{s} = 2.6\,^{\circ}\mathrm{C} \quad \mathrm{sCV} = 9.9\,^{\circ}\mathrm{C}$$
$$\mathrm{F}(4,23) = 800.0 \quad p < 0.0001$$

$$\begin{aligned}\mathbf{BP}(^{\circ}\mathbf{C}) = {} & -3.56(\pm 2.60) - 42.0(\pm 2.7) \cdot {}_{\mathrm{ss}}{}^{2,15}\mathrm{f}_{\mathrm{QC}}(\vec{\alpha}) - 1.00(\pm 0.07) \cdot {}_{\mathrm{ss}}{}^{1,4}\mathrm{f}_{\mathrm{CH}}(\vec{\mathrm{r}}) \\ & + 0.62(\pm 0.11) \cdot {}_{\mathrm{ss}}{}^{1,9}\mathrm{f}_{\mathrm{CH3}}(\vec{\mathrm{m}}) + 8.88(\pm 1.20) \cdot {}_{\mathrm{ss}}{}^{1,15}\mathrm{f}_{\mathrm{O}}(\vec{\alpha}) \\ & + 2.17(\pm 0.06) \cdot {}^{*}_{\mathrm{ss}}\mathrm{f}^{2,11}(\vec{\chi})\end{aligned} \tag{22}$$

$$\mathrm{N} = 28 \quad \mathrm{R}^2 = 0.999 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.997 \quad \mathrm{s} = 1.1\,^{\circ}\mathrm{C} \quad \mathrm{sCV} = 2.1\,^{\circ}\mathrm{C}$$
$$\mathrm{F}(5,22) = 3458 \quad p < 0.0001$$

$$\begin{aligned}\mathbf{BP}(^{\circ}\mathbf{C}) = {} & -8.06(\pm 0.80) \times 10^{3} - 6.14(\pm 0.74) \cdot {}_{\mathrm{ds}}{}^{1,7}\mathrm{f}_{\mathrm{QC}}(\vec{\mathrm{r}}) \\ & - 32.9(\pm 4.79) \cdot {}_{\mathrm{ds}}{}^{1,6}\mathrm{f}_{\mathrm{CH}}(\vec{\alpha}) + 1.45(\pm 0.02) \cdot {}^{*}_{\mathrm{ds}}\mathrm{f}^{1,4}_{\mathrm{CH2}}(\vec{\mathrm{m}}) \\ & + 4.91(\pm 0.48) \cdot {}^{*}_{\mathrm{ds}}\mathrm{f}^{1,2}_{\mathrm{CH3}}(\vec{\mathrm{m}}) + 2.55(\pm 0.26) \times 10^{3} \cdot {}^{*}_{\mathrm{ds}}\mathrm{f}^{1,2}_{\mathrm{O}}(\vec{\chi})\end{aligned} \tag{23}$$

$$\mathrm{N} = 28 \quad \mathrm{R}^2 = 0.997 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.996 \quad \mathrm{s} = 1.7\,^{\circ}\mathrm{C} \quad \mathrm{sCV} = 4.6\,^{\circ}\mathrm{C}$$
$$\mathrm{F}(5,22) = 1504 \quad p < 0.0001$$

where, N is the number of compounds, R^2 is the correlation coefficient, s is the standard deviation of the regression, $\mathrm{Q}^2_{\mathrm{LOO}}$ (sCV) is the square correlation coefficient (standard deviation) obtained from the LOO cross-validation procedure, and F is the Fisher ratio. As can be seen, all three models show good performance in the description of Bp of alkyl-alcohols.

These models, Eqs. (21)–(23), explain more than 99 % of the variance of the experimental Bp values. Similar results were reported by Estrada and Molina [102], and Kier and Hall [117] by using spectral moments and E-states as MDs, where more than 98 and 92 % of the variance of the experimental Bp values was explained, respectively. The statistical parameters for the best equations obtained for these sets of MDs are given in Table 6. Unfortunately, the cross-validation results for the precedent models were not reported by the respective authors.

However, it is remarkable that our models explain a higher percentage of the variance of the experimental Bp values than the previously developed models, showing a decrease in the standard error between 38 and 81 %, with regard to the results previously achieved by Estrada and Molina [102], and Kier and Hall [117].

The QSPR model derived with 3D-level linear indices showed similar-to-better results than those obtained by some of the present authors in previous studies [98]. To conclude, it is important to highlight that the models (21)–(23) reveal the importance of methyl groups and quaternary carbons which represent the presence and absence of

Table 6 Statistical parameters of the models describing the boiling point of 28 alkyl-alcohols, log p and log k of 34 2-fyrylethylenes using different MDs

Molecular descriptors	N	R^2	Q^2_{LOO}	s	s_{cv}	F
Boiling point of 28 alkyl alcohols						
Non-stochastic 3D-linear indices	4	0.993	0.990	2.60	9.85	800.0
Stochastic 3D-linear indices	5	0.999	0.997	1.12	2.08	3458
Doubly-stochastic 3D-linear indices	5	0.997	0.996	1.70	4.58	1504
Local spectral moments [102]	5	0.982	–	4.2	–	23.8
Non-stochastic 3D-linear indices	3	0.984	0.978	3.84	17.9	482.3
Stochastic 3D-linear indices	3	0.996	0.994	1.87	4.38	2070
Doubly-stochastic 3D-linear indices	3	0.989	0.984	3.14	12.6	724.8
E-State [117]	3	0.926	–	5.8	–	204
*Partition coefficient n-octanol/water (*log P*) of 34 2-furylethylenes*						
Non-stochastic 3D-linear indices	6	0.948	0.925	0.180	0.049	82.60
Stochastic 3D-linear indices	6	0.951	0.933	0.175	0.044	87.99
Doubly-stochastic 3D-linear indices	7	0.946	0.933	0.187	0.052	65.45
Vertex and edge Conn. Indices [118]	7	0.939	–	0.199	0.247	56.9
Topographic descriptors [118]	7	0.964	–	0.155	0.176	84.6
Quantum chemical descriptors [118]	7	0.875	–	0.319	0.370	45.5
*Reactivity (*log *k) of 34 2-furylethylenes*						
Non-stochastic 3D-linear indices	7	0.950	0.918	0.359	0.222	71.09
Stochastic 3D-linear indices	6	0.980	0.972	0.221	0.072	225.33
Doubly-stochastic 3D-linear indices	7	0.990	0.983	0.164	0.046	354.37
Connectivity indices [102]	7	0.821	–	0.681	–	17.1
Global spectral moments [102]	7	0.843	–	0.655	–	18.8
Local spectral moments [102]	7	0.964	–	0.320	–	70.4
Quantum chemical descriptors [102]	7	0.968	–	0.288	–	112.2

branching, and the oxygen atom in the prediction of the boiling point of the aliphatic alcohols.

3.10.3 Experiment 3

In this section, we evaluate the possibilities of the 3D-linear indices in QSPRs for the partition coefficient (log p) and the specific rate constant (log k) of 34 derivatives of 2-furylethylenes and compare these results to those obtained by Estrada and Molina [102,118] by using topological (total and local spectral moments and 2D connectivity indices), plus topographic and quantum chemical descriptors. The MDs, included in these equations, clearly pointed to the identification of the reaction centers involved in the studied chemical interaction [102,118]. That is to say, the atoms 2, 6 and 7 or the bonds defined by these atoms (C2–C6 and C6–C7) were selected as the most significant ones; because these are the ones involved in the exocyclic double bond of

the 2-furylethylene and these are the "target" of the nucleophilic attack by the thiol (mercapto) group. Taking into account this logical result, we also calculated the local linear indices for these atoms (bonds C2–C6 and C6–C7). The best models obtained, by using these 3D-linear indices, together with their statistical parameters, are given below:

$$\begin{aligned}\log p = {} & 1.02(\pm 0.19) + 6.65(\pm 0.57) \times 10^{-4} \cdot {}^{1,1}_{\mathrm{AR}}\mathrm{f}(\vec{\mathrm{m}}) \\ & -1.00(\pm 0.16) \times 10^{-7} \cdot {}^{1,5}_{\mathrm{NIT}}\mathrm{f}(\vec{\chi}) + 1.15(\pm 0.10) \times 10^{-4} \cdot {}^{2,2}_{\mathrm{sp3}}\mathrm{f}(\vec{\mathrm{m}}) \\ & +3.54(\pm 1.34) \times 10^{-16} \cdot {}^{*}{}^{1,12}_{\mathrm{CH}}\mathrm{f}(\vec{\alpha}) - 3.51(\pm 0.28) \times 10^{-5} \cdot {}^{*}{}^{3,2}_{\mathrm{sp2}}\mathrm{f}(\vec{\chi}) \\ & -2.73(\pm 1.35) \times 10^{-14} \cdot {}^{*}{}^{3,12}_{\mathrm{sp}}\mathrm{f}(\vec{\chi}) \end{aligned} \tag{24}$$

$$\mathrm{N} = 34 \quad \mathrm{R}^2 = 0.948 \quad \mathrm{s} = 0.180 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.925 \quad \mathrm{s}_{cv} = 0.049$$
$$\mathrm{F}(6,27) = 82.60 \quad p < 1.00 \times 10^{-14}$$

$$\begin{aligned}\log p = {} & 3.48(\pm 0.39) - 5.17(\pm 0.53) \times 10^{-2} \cdot {}_{\mathrm{ss}}\mathrm{f}^{2,1}_{\mathrm{sp2}}(\vec{\mathrm{r}}) \\ & + 1.31(\pm 0.16) \times 10^{-1} \cdot {}_{\mathrm{ss}}\mathrm{f}^{3,5}_{\mathrm{sp2}}(\vec{\chi}) + 2.11(\pm 0.14) \times 10^{-2} \cdot {}_{\mathrm{ss}}\mathrm{f}^{1,2}(\vec{\mathrm{r}}) \\ & - 0.39(\pm 0.13) \cdot {}^{*}_{\mathrm{ss}}\mathrm{f}^{1,2}_{\mathrm{sp3}}(\vec{\mathrm{r}}) + 0.71(\pm 0.23) \cdot {}^{*}_{\mathrm{ss}}\mathrm{f}^{1,12}_{\mathrm{sp3}}(\vec{\mathrm{r}}) \\ & - 2.10(\pm 0.38) \times 10^{-2} \cdot {}^{*}_{\mathrm{ss}}\mathrm{f}^{3,3}_{\mathrm{NIT}}(\vec{\mathrm{r}}) \end{aligned} \tag{25}$$

$$\mathrm{N} = 34 \quad \mathrm{R}^2 = 0.951 \quad \mathrm{s} = 0.175 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.933 \quad \mathrm{s}_{\mathrm{cv}} = 0.044$$
$$\mathrm{F}(6,27) = 87.99 \quad p < 1.00 \times 10^{-14}$$

$$\begin{aligned}\log p = {} & 6.18(\pm 0.66) + 1.57(\pm 0.38) \times 10^{-1} \cdot {}_{\mathrm{ds}}\mathrm{f}^{2,2}_{\mathrm{NIT}}(\vec{\mathrm{m}}) \\ & - 1.67(\pm 0.28) \times 10^{-1} \cdot {}_{\mathrm{ds}}\mathrm{f}^{2,12}_{\mathrm{NIT}}(\vec{\mathrm{r}}) + 1.18(\pm 0.19) \times 10^{-1} \cdot {}_{\mathrm{ds}}\mathrm{f}^{3,6}_{\mathrm{sp3}}(\vec{\mathrm{m}}) \\ & - 1.58(\pm 0.46) \times 10^{-1} \cdot {}_{\mathrm{ds}}\mathrm{f}^{1,1}(\vec{\chi}) + 5.10(\pm 1.60) \times 10^{-2} \cdot {}^{*}_{\mathrm{ds}}\mathrm{f}^{2,12}_{\mathrm{HBD}}(\vec{\mathrm{m}}) \\ & + 6.56(\pm 3.18) \times 10^{-3} \cdot {}^{*}_{\mathrm{ds}}\mathrm{f}^{3,12}_{\mathrm{NIT}}(\vec{\mathrm{r}}) - 1.95(\pm 0.45) \times 10^{-1} \cdot {}^{*}_{\mathrm{ds}}\mathrm{f}^{2,5}_{\mathrm{NIT}}(\vec{\chi}) \end{aligned} \tag{26}$$

$$\mathrm{N} = 34 \quad \mathrm{R}^2 = 0.946 \quad \mathrm{s} = 0.187 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.933 \quad \mathrm{s}_{\mathrm{cv}} = 0.052$$
$$\mathrm{F}(7,26) = 65.45 \quad p < 1.00 \times 10^{-14}$$

$$\begin{aligned}\log k = {} & 4.15(\pm 0.38) + 4.75(\pm 0.60) \times 10^{-4} \cdot {}^{3,3}_{\mathrm{AR}}\mathrm{f}(\vec{\chi}) \\ & + 2.78(\pm 0.52) \times 10^{-4} \cdot {}^{1,2}_{\mathrm{HET}}\mathrm{f}(\vec{\chi}) - 3.82(\pm 0.69) \times 10^{-3} \cdot {}^{2,1}_{\mathrm{R1}}\mathrm{f}(\vec{\chi}) \\ & - 3.50(\pm 0.36) \times 10^{-9} \cdot {}^{3,7}_{\mathrm{R3}}\mathrm{f}(\vec{\mathrm{m}}) - 2.35(\pm 0.28) \times 10^{-3} \cdot {}^{*}{}^{3,2}_{\mathrm{AR}}\mathrm{f}(\vec{\alpha}) \\ & - 1.95(\pm 0.93) \times 10^{-5} \cdot {}^{*}{}^{3,2}_{\mathrm{sp3}}\mathrm{f}(\vec{\mathrm{r}}) + 1.22(\pm 0.18) \times 10^{-3} \cdot {}^{*}{}^{2,3}_{\mathrm{C2C6}}\mathrm{f}(\vec{\alpha}) \end{aligned} \tag{27}$$

$$\mathrm{N} = 34 \quad \mathrm{R}^2 = 0.950 \quad \mathrm{s} = 0.359 \quad \mathrm{Q}^2_{\mathrm{LOO}} = 0.918 \quad \mathrm{s}_{\mathrm{cv}} = 0.222$$

$$F(7,26) = 71.09 \quad p < 1.00 \times 10^{-14}$$

$$\begin{aligned}\log k = & -2.09(\pm 0.97) + 6.47(\pm 0.60) \times 10^{-2} \cdot {}_{ss}f_{R3}^{3,1}(\vec{m}) \\ & -3.22(\pm 0.42) \times 10^{-1} \cdot {}_{ss}f_{R3}^{2,9}(\vec{\alpha}) + 2.70(\pm 0.48) \times 10^{-1} \cdot {}_{ss}f_{sp2}^{1,2}(\vec{\alpha}) \\ & + 9.45(\pm 0.80) \times 10^{-1} \cdot {}_{ss}^{*}f_{AR}^{3,12}(\vec{\chi}) - 1.54(\pm 0.12) \cdot {}_{ss}^{*}f_{AR}^{3,4}(\vec{\alpha}) \\ & + 1.92(\pm 0.35) \times 10^{-2} \cdot {}_{ss}^{*}f_{R2}^{1,10}(\vec{r}) \end{aligned} \quad (28)$$

$$N = 34 \quad R^2 = 0.980 \quad s = 0.221 \quad Q^2_{LOO} = 0.972 \quad s_{cv} = 0.072$$
$$F(6,27) = 225.33 \quad p < 1.00 \times 10^{-14}$$

$$\begin{aligned}\log k = & -49.4(\pm 5.54) + 1.68(\pm 0.15) \times 10^{-1} \cdot {}_{ds}f_{HET}^{1,12}(\vec{\chi}) \\ & + 2.12(\pm 0.26) \cdot {}_{ds}f_{NIT}^{1,8}(\vec{\alpha}) - 4.51(\pm 0.95) \cdot {}_{ss}f_{C6C7}^{1,3}(\vec{\chi}) \\ & + 6.11(\pm 0.23) \cdot {}_{ss}^{*}f_{AR}^{3,2}(\vec{\chi}) - 6.16(\pm 0.65) \cdot {}_{ss}^{*}f_{NIT}^{1,9}(\vec{m}) \\ & + 5.97(\pm 0.63) \cdot {}_{ss}^{*}f_{NIT}^{1,10}(\vec{m}) - 4.54(\pm 0.58) \cdot {}_{ss}^{*}f_{C2C6}^{2,8}(\vec{\chi}) \end{aligned} \quad (29)$$

$$N = 34 \quad R^2 = 0.990 \quad s = 0.164 \quad Q^2_{LOO} = 0.983 \quad s_{cv} = 0.046$$
$$F(7,26) = 354.37 \quad p < 1.00 \times 10^{-14}$$

These equations, obtained by using non-stochastic, stochastic and doubly-stochastic 3D-linear indices, explained (94.8, 95.1, 94.6 %) and (95.0, 98.0, 99.0 %) of the variance of log k and log p, respectively. These statistics are rather better than those previously obtained (see Table 6 for more details) [102,118].

The LOO cross-validation procedure was used in order to assess the predictive ability of the developed models. Using this approach, models (24)–(29) yielded a Q^2_{LOO} of 0.948, 0.951, 0.946, 0.950, 0.980, and 0.990, respectively. These values of Q^2_{LOO} can be considered as proof of the high predictive ability of the models [129]. On the other hand, the equations obtained with vertex- and edge-connectivity indices, topographic descriptors, and quantum chemical indices showed lower predictive abilities (scv of 0.247, 0.176, and 0.370, respectively) than Eqs. (24) (scv = 0.049), (25) (scv = 0.044) and (26) (scv = 0.052), achieved with the total and local 3D-based linear indices, respectively, for description of the log *p* values (see Table 6 for more details). Unfortunately, the authors [102] of the previous work did not report the results for the LOO cross-validation experiment for log k. However, in Table 6 can be easily observed that our obtained models, Eqs. (27)–(29), explain greater percentages of the variance of the experimental log k values than the previously developed models, showing decreases in the standard error between 23 and 76 % with regard to the results previously obtained by Estrada and Molina [102], using connectivity indices (both 2D and 3D as well as edge- and vertex-based), total (global) spectral moments (sum of the trace of the bond matrix), local (fragment) spectral moments (partial sum of the trace of the bond matrix) and quantum chemical descriptors, respectively.

The equations predicting the octanol/water partition coefficient have a great representation from local fragments related to hybridization and presence of aromatic groups and heteroatoms. The doubly-stochastic MD models, as a particular case, have an additional contribution of H-bond acceptor fragments. The atom properties mainly

reflected in the equations are of electronic nature, associated with indices related to interactions of short-to-medium range (p =1–6). Therefore, the models (24)–(26) are capable of interpreting the distribution of furylethylenes in water and/or octanol as a consequence of purely structural characteristics of the atoms in the molecule. On the contrary, the models (27)–(29) that predict the specific rate constant of double-bond addition take the local MDs representing the C2–C6 and C6–C7 bonds and fragments. This is a logical result, if we take into account that these atoms are involved in the exocyclic double bond of the 2-furylethylenes, and that these are the "target" of the nucleophilic attack by the thiol (mercapto) group. Nevertheless, the total MDs included in the attained models also indicate that the best description of the properties will be obtained by using a combination of local features of every molecular structure included in the analysis. From this point of view, it is of great importance to have atom level as well as total molecular indices in the molecular space, to obtain a better description than using the local and global sets of MDs separately. The indices that appear more frequently in the final equation are the short (p = 1–3) and long range ($p > 8$). This shows that the interaction of electrons in atoms belonging to the molecular environment of the exocyclic double bond, determines the chemical reactivity of furylethylenes. Finally, the whole weighting-schedule [Mulliken electronegativity (χ), polarizability (α), atomic mass (m) and van der Waals volume (r)], included in every model, showed the importance of the use of adequate combinations of chemical-labels, in order to predict properties and activities of different nature.

3.11 Final conclusions

The application of the concepts of discrete mathematics and linear algebra to chemistry permitted us to define a new family of 3D-indices based in the concepts of linear maps and functions on geometric-based matrices. Here, we defined new total and local (atom, atom-type and group) MDs based on the extended and generalized 3D (geometric) distance matrices. We also propose algebraic transformations on these matrix representations to yield "stochastic", "double-stochastic" and "mutual probabilistic" distances of atom-pairs, from which 3D (geometric)-linear indices are defined. It was demonstrated that the novel 3D-linear indices codify structural information not captured by other descriptor families in DRAGON software, possess similar-to-better variability, according to Shannon's entropy based variability analysis, than known MD calculating software packages (DRAGON, MOLd2, PADEL, MODESLAB, BLUECAL, MOLCONNZ, POWER MV, and CDK), and are useful in the modeling of physicochemical properties of molecules. Therefore, they can be considered as a relevant tool to take into account in QSPR/QSAR and similarity/dissimilarity analysis.

3.12 Future outlook

Although the first results with these 3D-linear indices show promissory behavior, additional studies with wider and more diverse databases are indispensable, in order to evaluate the genuine possibilities of the proposed MDs in real QSAR/QSPR problems. This will be the subject of future works. The steroid benchmark data set and eight

datasets (aligned ligands plus data) compiled by Sutherland et al. [7] have gained considerable acceptance as ideal for such extensive QSPR/QSAR studies. It would also be interesting to perform multiple non-parametric analyses for statistical significance of the results obtained with 2D, 2.5 and 3D approaches.

We also intend to extend other previously defined 2D-TOMOCOMD indices like quadratic and bilinear indices, whose effectiveness has been demonstrated in the literature, to codify 3D-molecular features and in the same manner evaluate the contribution of this "generalization" to the improvement of the correlation with molecular properties.

The MD computations in this study were performed with a preliminary version of the TOMOCOMD-CARDD software. In the future better weighting schemes, atom-types and fragments, plus an improved visual platform will be implemented. We also intend to implement cut-offs on geometric distances, yielding yet other local (or fragment-based) 3D-linear indices.

As mentioned in this report, the total (global) 3D-linear indices are expressed as the sum of the local 3D-linear indices of the Z fragments, similar to the extended Hückel MO method. In posterior works we will generalize this procedure through the introduction of a series of metric, mean and statistical invariants.

The resulting indices will be later extended to define other geometric (3D) aspects of molecules (for instance, using others metrics and several modified similarity coefficients) 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 Cubillán, N. thanks Consejo de Desarrollo Científico y Humanístico (CONDES-LUZ, Grant CC-0593-10), Fondo Nacional de Ciencia, Tecnología e Innovación (FONACIT, Grant G-2005000403), Misión Ciencia (Grant 2007000881) and Instituto Zuliano de Investigaciones Tecnológicas (INZIT, Project LOCTI) for partial financial support of this work. Marrero-Ponce, Yovani thanks the program 'Estades Temporals per an Investigadors Convidats' for a fellowship to work at Valencia University. The authors acknowledge also the partial financial support from Spanish Ministry of Science and Innovation (MICINN, Project Reference: SAF2009-10399). Last, but not least, the authors want to express their acknowledgements to Prof. Jorge Galvez (VU) and Prof. Ramón García-Domenech (VU) for their help and useful comments about these new MDs.

References

1. D.E. Clark, S.D. Pickett, Drug Discov. Today **5**, 49 (2000)
2. B.L. Claus, D.J. Underwood, Drug Discov. Today **7**, 957 (2002)
3. So Jonsdottir, F.S. Jorgensen, S. Brunak, Bioinformatics **21**, 2145 (2005)
4. R. Perkins, H. Fang, W. Tong, W.J. Welsh, Environ. Toxicol. Chem. **22**, 1666 (2003)
5. C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, Nature **194**, 178 (1962)
6. H. Kubinyi, Drug Discov. Today **2**, 457 (1997)
7. J.J. Sutherland, L.A. O'Brien, D.F. Weaver, J. Med. Chem. **47**, 5541 (2004)
8. A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Chem. Rev. (Washington, DC, United States) **110**, 5714 (2010)
9. J.A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. Rotondo, J. Mol. Graph. Model. **26**, 32 (2007)
10. A. Golbraikh, A. Tropsha, J. Chem. Inf. Comput. Sci. **43**, 144 (2003)
11. A.R. Katritzky, E.V. Gordeeva, J. Chem. Inf. Comput. Sci. **33**, 835 (1993)

12. Y. Marrero-Ponce, E. Martínez-Albelo, G. Casañola-Martín, J. Castillo-Garit, Y. Echeverría-Díaz, V. Zaldivar, J. Tygat, J. Rodríguez Borges, R. García-Domenech, F. Torrens, F. Pérez-Giménez, Mol Divers **14**, 731 (2010)
13. W. Tong, D.R. Lowis, R. Perkins, Y. Chen, W.J. Welsh, D.W. Goddette, T.W. Heritage, D.M. Sheehan, J. Chem. Inf. Comput. Sci. **38**, 669 (1998)
14. M. Atabati, K. Zarei, A. Borhani, Fluid Phase Equilib. **293**, 219 (2010)
15. M. Devereux, P.L.A. Popelier, I.M. McLay, J. Chem. Inf. Model. **49**, 1497 (2009)
16. M. Hechinger, K. Leonhard, W. Marquardt, J. Chem. Inf. Model. **52**, 1984–1993 (2012)
17. A. Klamt, F. Eckert, M. Hornig, M.E. Beck, T. Bürger, J. Comput. Chem. **23**, 275 (2002)
18. K. Masuch, A. Fatemi, H. Murrenhoff, K. Leonhard, Lubr. Sci. **23**, 249 (2011)
19. G. Occhipinti, H.-R. Bjørsvik, V.R. Jensen, J. Am. Chem. Soc. **128**, 6952 (2006)
20. J.J. Panek, A. Jezierska, M. Vracko, J. Chem. Inf. Model. **45**, 264 (2005)
21. O.M. Rivera-Borroto, Y. Marrero-Ponce, J.M. García-de la Vega, RdC Grau-Ábalo, J. Chem. Inf. Model. **51**, 3036 (2011)
22. S.-S. Yang, W.-C. Lu, T.-H. Gu, L.-M. Yan, G.-Z. Li, QSAR Comb. Sci. **28**, 175 (2009)
23. K. Kim, G. Greco, E. Novellino, Perspect. Drug Discovery Des. **12–14**, 257 (1998)
24. M. Pastor, G. Cruciani, I. McLay, P. Pickett, S. Clementi, J. Med. Chem. **43**, 3233 (2000)
25. A.N. Jain, K. Koile, D. Chapman, J. Med. Chem. **37**, 2315 (1994)
26. D.E. Walters, *3D QSAR in Drug Design* (Springer, Netherlands, 1998)
27. M.F. Parretti, R.T. Kroemer, J.H. Rothman, G.W. R, J. Comput. Chem. **18**, 1344 (1997)
28. S.-S. So, M. Karplus, J. Med. Chem. **40**, 4347 (1997)
29. Y. Tominaga, I. Fujiwara, J. Chem. Inf. Comput. Sci. **37**, 1158 (1997)
30. C.T. Klein, D. Kaiser, G. Ecker, J. Chem. Inf. Comput. Sci. **44**, 200 (2004)
31. R. Todeschini, M. Lasagni, E. Marengo, J. Chemometrics **8**, 263 (1994)
32. G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A.J. Zaliani, J. Comput. Aided Mol. Des. **11**, 79 (1997)
33. M. Wagener, J. Sadowski, J. Gasteiger, J. Am. Chem. Soc. **117**, 7769 (1995)
34. R. Bursi, D. Dao, T. van Wijk, M. de Gooyer, M. Kellenbach, P. Verwer, J. Chem. Inf. Comput. Sci. **39**, 861 (1999)
35. D.B. Turner, P. Willett, Eur. J. Med. Chem. **35**, 367–375 (2000)
36. D.B. Turner, P. Willett, A.M. Ferguson, T.W. Heritage, J. Comput. Aided Mol. Des. **13**, 271 (1999)
37. S.A. Wildman, G.M. Crippen, J. Mol. Graphics Modell. **21**, 161 (2002)
38. G. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, J. Chem. Inf. Comput. Sci. **36**, 1030 (1996)
39. B.D. Silverman, D. Platt, M. Pitman, I. Rigoutsos, Perspect. Drug Discov. Des. **12–14**, 183 (1998)
40. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, vol. 1. *Alphabetical Listing*, vol 2. Appendices, References (Wiley-VCH, Weinheim, 2009), p. 2125
41. A.T. Balaban, *From Chemical Topology to Three-Dimensional Geometry* (Plenum Press, New York, 1997), p. 420
42. B. Bogdanov, S. Nikolic, N. Trinajstic, J. Math. Chem. **3**, 299 (1989)
43. B. Bogdanov, S. Nikolic, N. Trinajstic, J. Math. Chem. **5**, 305 (1990)
44. O. Mekenyan, D. Peitchev, D. Bonchev, N. Trinajstic, I.P. Bangov, Arzneim. Forsch. **36**, 176 (1986)
45. M. Randić, New J. Chem. **19**, 781 (1995)
46. M. Randić, J. Chem. Inf. Comput. Sci. **35**, 373 (1995)
47. M. Randić, New J. Chem. **20**, 1001 (1996)
48. M. Randić, M. Razinger, J. Chem. Inf. Comput. Sci. **35**, 594 (1995)
49. J. Aires-de-Sousa, J. Gasteiger, I. Gutman, D. Vidovic, J. Chem. Inf. Comput. Sci. **44**, 831 (2004)
50. R. Benigni, M. Cotta-Ramusino, G. Gallo, F. Giorgi, A. Guliani, M.R. Vari, J. Med. Chem. **43**, 3699 (2000)
51. A. Golbraikh, D. Bonchev, A. Tropsha, J. Chem. Inf. Comput. Sci. **41**, 147 (2001)
52. J.V. Julian-Ortiz, C.G. Alapont, I. Rıos-Santamarina, R. Garcıa-Domenech, J. Galvez, J. Mol. Graphics Model. **16**, 14 (1998)
53. Y. Marrero-Ponce, J. Castillo-Garit, E. Castro, F. Torrens, R. Rotondo, J. Math. Chem. **44**, 755 (2008)
54. H.B. Schulz, E.B. Schulz, T.P. Schulz, J. Chem. Inf. Comput. Sci. **35**, 864 (1995)
55. Y. Marrero Ponce, J. Chem. Inf. Comput. Sci. **44**, 2010 (2004)
56. Y. Marrero-Ponce, J. Castillo-Garit, F. Torrens, V. Romero Zaldivar, E. Castro, Molecules **9**, 1100 (2004)

57. Y. Marrero-Ponce, J. Chem. Inf. Comput. Sci. **44**, 2010 (2004)
58. Y. Marrero-Ponce, J.A. Castillo-Garit, E. Olazabal, H.S. Serrano, A. Morales, N. Castañedo, F. Ibarra-Velarde, A. Huesca-Guillen, A.M. Sánchez, F. Torrens, E.A. Castro, Bioorg. Med. Chem. **13**, 1005 (2005)
59. G.M. Casañola-Martín, M.T.H. Khan, Y. Marrero-Ponce, A. Ather, M.N. Sultankhodzhaev, F. Torrens, Bioorg. Med. Chem. Lett. **16**, 324 (2006)
60. G.M. Casañola-Martín, Y. Marrero-Ponce, M.T.H. Khan, A. Ather, S. Sultan, F. Torrens, R. Rotondo, Bioorg. Med. Chem. **15**, 1483 (2007)
61. J.A. Castillo-Garit, M.C. Vega, M. Rolon, Y. Marrero-Ponce, V.V. Kouznetsov, D.F.A. Torres, A. Gómez-Barrio, A.A. Bello, A. Montero, F. Torrens, F. Pérez-Giménez, Eur. J. Pharm. Sci. **39**, 30 (2010)
62. Y. Marrero-Ponce, Y. Machado-Tugores, D.M. Pereira, J.A. Escario, A.G. Barrio, J.J. Nogal-Ruiz, C. Ochoa, V.J. Aran, A.R. Martinez-Fernandez, R.N. Garcia Sanchez, Curr. Drug Discov. Technol. **2**, 245 (2005)
63. Y. Marrero-Ponce, A. Meneses-Marcel, O.M. Rivera-Borroto, R. Garcia-Domenech, J.V. De Julian-Ortiz, A. Montero, J.A. Escario, A.G. Barrio, D.M. Pereira, J.J. Nogal, R. Grau, F. Torrens, C. Vogel, V.J. Aran, J. Comput. Aided Mol. Des. **22**, 523 (2008)
64. Y. Marrero-Ponce, A. Montero-Torres, C. RomeroZaldivar, M. IyarretaVeitıa, M. MayonPerez, R. GarcıaSanchez, Bioorg. Med. Chem. **13**, 1293 (2005)
65. J.A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. García-Domenech, V. Romero-Zaldivar, J. Comput. Chem. **29**, 2500 (2008)
66. J.W. Godden, J. Bajorath, J. Chem. Inf. Comput. Sci. **42**, 87 (2001)
67. J.W. Godden, F.L. Stahura, J. Bajorath, J. Chem. Inf. Comput. Sci. **40**, 796 (2000)
68. Y. Marrero Ponce, J.A. Castillo Garit, D. Nodarse, Bioorg. Med. Chem. **13**, 3397 (2005)
69. B. Vargas-Quesada, F.M. Anegón, *Visualizing the Structure of Science* (Springer, New York, 2007)
70. S. Nikolic, N. Trinajstic, Z. Mihalic, S. Carter, Chem. Phys. Lett. **179**, 21 (1991)
71. J.A. Castillo-Garit, O. Martinez-Santiago, Y. Marrero-Ponce, G.M. Casañola-Martín, F. Torrens, Chem. Phys. Lett. **464**, 107 (2008)
72. Y. Marrero Ponce, A. Montero-Torres, C. Romero Zaldivar, M. Iyarreta Veitía, M. Mayón Peréz, R.N. García Sánchez, Bioorg. Med. Chem. **13**, 1293 (2005)
73. Y. Marrero-Ponce, A. Huesca-Guillén, F. Ibarra-Velarde, J. Mol. Struct. (Thoechem) **717**, 67 (2005)
74. Y. Marrero-Ponce, R. Medina-Marrero, F. Torrens, Y. Martinez, V. Romero-Zaldivar, E.A. Castro, Bioorg. Med. Chem. **13**, 2881 (2005)
75. A. Montero-Torres, R.N. García-Sánchez, Y. Marrero-Ponce, Y. Machado-Tugores, J.J. Nogal-Ruiz, A.R. Martínez-Fernández, V.J. Arán, C. Ochoa, A. Meneses-Marcel, F. Torrens, Eur. J. Med. Chem. **41**, 483 (2006)
76. A. Montero-Torres, M.C. Vega, Y. Marrero-Ponce, M. Rolón, A. Gómez-Barrio, J.A. Escario, V.J. Arán, A.R. Martínez-Fernández, A. Meneses-Marcel, Bioorg. Med. Chem. **13**, 6264 (2005)
77. S.I. Sandler, *An Introduction to Applied Statistical Thermodynamics* (Wiley, New Jersey, USA, 2010)
78. G.A.F. Seber, *A Matrix Handbook for Statisticians* (Wiley-Interscience, New Jersey, USA, 2008)
79. D. Serre, *Matrices: Theory and Applications* (Springer, New York, 2002)
80. R. Sinkhorn, Ann. Math. Stat. **35**, 876 (1964)
81. R. Sinkhorn, P. Knopp, Pac. J. Math. **21**, 343 (1967)
82. C.R. Johnson, R.D. Masson, M.W. Trosset, Linear Algebra Appl. **397**, 253 (2005)
83. S.J. Axler, *Linear Algebra Done Right* (Springer, New York, 1997)
84. A. Browder, *Mathematical Analysis: An Introduction* (Springer, New York, 1996)
85. C.H. Edwards, D.E. Penney, *Elementary Linear Algebra* (Prentice Hall, Englewoods Cliffs, 1988)
86. R. Todeschini, V. Consonni, M. Pavan, *Dragon Software* (Talete, Italy, 2002)
87. A. Basilevsky, *Statistical Factor Analysis Rel Method: Theory and Applications* (Wiley, New York, USA, 1994)
88. E. Estrada, J. Chem. Inf. Comput. Sci. **39**, 1042 (1999)
89. I.E. Frank, J.H. Friedman, Technometrics **35**, 109 (1993)
90. R. Franke, *Theoretical Drug Design Methods* (Elsevier Science Amsterdam, The Netherlands, 1984)
91. E.R. Malinowski, *Factor Analysis in Chemistry* (Wiley Interscience, Hoboken, USA, 1991)
92. Statsoft, STATISTICA, (Statsoft, Tulsa, 2001)
93. B.A. Bunin, *Chemoinformatics: Theory, Practice & Products* (Springer, Dordrecht, 2007), p. 295
94. C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948)

95. R.P. Urias, S. Barigye, Y. Marrero-Ponce, C. García-Jacas, J. Valdes-Martiní, F. Perez-Gimenez, Mol. Divers **19**, 305 (2015)
96. H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, J. Chem. Inf. Model. **48**, 1337 (2008)
97. Y. Marrero-Ponce, Molecules **8**, 687 (2003)
98. Y. Marrero-Ponce, R. Marrero, E. Castro, R. Ramos de Armas, H.G. Díaz, V. Zaldivar, F. Torrens, Molecules **9**, 1124 (2004)
99. C.W. Yap, J. Comput. Chem. **32**, 1466 (2011)
100. E. Estrada, J. Chem. Inf. Comput. Sci. **36**, 844 (1996)
101. E. Estrada, SAR QSAR Environ. Res. **11**, 55 (2000)
102. E. Estrada, E. Molina, J. Mol. Graph. Model. **20**, 54 (2001)
103. H. Georg, *BlueDesc-Molecular Descriptor Calculator* (University of Tübingen, Tübingen, Germany, 2008)
104. L. H. Hall, L. B. Kier, *Molconn-Z* 4.00 (Hall Associates Consulting, Quincy, 2002)
105. J. Liu, J. Feng, A. Brooks, S. Young, *PowerMV: A Software Environment for Statistical Analysis, Molecular Viewing, Descriptor Generation, and Similarity Search* (National Institute of Statistical Sciences, North Carolina, USA, 2005)
106. R. Guha, *The CDK Descriptor Calculator* (NIH Chemical Genomics Center, Indiana, USA, 1991)
107. M. Randic, N. Trinajstic, J. Mol. Struct. (Thoechem) **284**, 209 (1993)
108. M. Randic, N. Trinajstic, J. Mol. Struct. **300**, 551 (1993)
109. V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, J. Chem. Inf. Comput. Sci. **42**, 693 (2002)
110. M.V. Diudea, J. Chem. Inf. Comput. Sci. **36**, 535 (1996)
111. M.V. Diudea, O.M. Minailiuc, G. Katona, Revue roumaine de chimie **42**, 239 (1997)
112. E. Estrada, L. Rodríguez, J. Chem. Inf. Comput. Sci. **39**, 1037 (1999)
113. M. Randic, J. Mol. Struct. (Thoechem) **233**, 45 (1991)
114. M. Randic, Croat. Chem. Acta **66**, 289 (1993)
115. M. Randic, X. Guo, T. Oxley, H. Krishnapriyan, L. Naylor, J. Chem. Inf. Comput. Sci. **34**, 361 (1994)
116. S. Markovic, I. Gutman, J. Mol. Struct. (Thoechem) **235**, 81 (1991)
117. L.B. Kier, L.H. Hall, *Molecular Structure Description: The Electrotopological State* (Academic Press, San Diego, 1999)
118. E. Estrada, E. Molina, J. Chem. Inf. Comput. Sci. **41**, 791 (2001)
119. Š. Baláž, E. Šturdík, M. Rosenberg, J. Augustín, B. Škára, J. Theor. Biol. **131**, 115 (1988)
120. J. Stewart, J. Mol. Model. **13**, 1173 (2007)
121. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley Pub. Co., Reading, 1989)
122. A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami, J. Am. Chem. Soc. **119**, 10509 (1997)
123. D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. **34**, 854 (1994)
124. S.-S. So, M. Karplus, J. Med. Chem. **39**, 1521 (1996)
125. P. Willett, Trends Biotechnol. **13**, 516 (1995)
126. R. Todeschini, V. Consonni, A. Mauri, M. Pavan, R. Leardi, *Data Handling in Science and Technology* (Elsevier, Amsterdam, 2003)
127. J. Devillers, A.T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR* (Gordon and Breach, Amsterdam, The Netherland, 1999)
128. V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. **42**, 682 (2002)
129. A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. **20**, 269 (2002)