J Comput Aided Mol Des (2006) 20:685–701
DOI 10.1007/s10822-006-9089-4

ORIGINAL PAPER

Bond-based global and local (bond, group and bond-type) quadratic indices and their applications to computer-aided molecular design. 1. QSPR studies of diverse sets of organic chemicals

Yovani Marrero-Ponce · Francisco Torrens · Ysaias J. Alvarado · Richard Rotondo

Received: 18 August 2006 / Accepted: 18 October 2006 / Published online: 25 November 2006

Abstract The concept of atom-based quadratic indices is extended to a series of molecular descriptors (MDs) (both total and local) based on adjacency between edges. The *k*th edge-adjacency matrix ($\mathbf{E}^k$) denotes the matrix of bond-based quadratic indices (non-stochastic) with respect to the canonical basis set. The *k*th ''stochastic'' edge-adjacency matrix, $\mathbf{ES}^k$, is here proposed as a new molecular representation easily calculated from $\mathbf{E}^k$. Then, the *k*th stochastic bond-based quadratic indices are calculated using $\mathbf{ES}^k$ as operators of quadratic transformations. The study of six representative physicochemical properties of octane isomers was used to compare the ability of both series of MDs to produce significant quantitative structure–property relationship (QSPR) models. Moreover, the general performance of the new MDs in this QSPR study has been evaluated with respect to other 2D/3D well-known sets of indices and the obtained results shown a quite satisfactory behavior of the present method. The novel bond-level MDs were also used for the description and prediction of the boiling point of 28 alkyl-alcohols and to the modeling of the specific rate constant (log *k*) of 34 derivatives of 2-furylethylenes. These models were statistically significant and showed very good stability to data variation in leave-one-out (LOO) cross-validation experiment. The comparison with other approaches (edge- and vertices-based connectivity indices, total and local spectral moments, and quantum chemical descriptors as well as E-state/biomolecular encounter parameters) expose a good behavior of our method in this QSPR studies. The approach described in this report appears to be a very promising structural invariant, useful for QSPR/QSAR studies, similarity/diversity analysis, and computer-aided ''rational'' molecular (drug) design.

Keywords TOMOCOMD-CARDD software · Non-stochastic and stochastic bond-based quadratic indices · Edge-adjacency matrix · Stochastic matrix · QSPR study · QSPR model · Physicochemical property · Octane isomers · Alkyl-alcohol · 2-Furylethylene

''At present, most of researchers working in drug discovery with the use of TIs concentrate their efforts in the development of more powerful graph-theoretical descriptors.''

Estrada, E. and Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.

Y. Marrero-Ponce
Unit of Computer-Aided Molecular 'Biosilico' Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, Villa Clara 54830, Cuba

Y. Marrero-Ponce (✉) · F. Torrens
Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts
de Paterna, P.O. Box 22085,
E-46071 València, Spain
e-mail: ymarrero77@yahoo.es

Y. J. Alvarado
Laboratorio de Electrónica Molecular, Departamento de Química, Modulo II, grano de Oro, Facultad Experimental de Ciencias, La Universidad del Zulia (LUZ),
Maracaibo, Venezuela

R. Rotondo
Mediscovery, Inc. Suite 1050, 601 Carlson Parkway,
Minnetonka, MN 55305, USA

Background

During the past decade, a great explosion of molecular descriptors (MDs) has been observed. For instance, topological indices (TIs), surface areas, volume descriptors, charges, and quantum-chemical measures have been extensively enhanced and used as the MDs of the whole molecule [1–3]. However, local MDs has received very litter attention [4]. One exception in this sense is the electrotopological state (E-state) index [5]. Other ''global'' MDs such as spectral moments of the edge-adjacency matrix had been redefined in local form [4]. In this sense, in a manner similar to that for the atom- and atom-type level E-State, an E-State index for bonds and bond-type has been proposed. The bond-based E-State indices provided an improvement of 25% with regard to the atom-based E-State indices in the description of the boiling point (BP) of 372 alkanes, alcohols, and chloroalkanes [5].

The edge (bond)-adjacency relationships have also been used in the generation of new TIs [1, 3]. Their matrix form has been considered and explicitly defined in the chemical graph theory literature, but has received very little attention in both chemical and mathematical literature. Nevertheless, in the last decade Estrada rediscovered this matrix as an important source of graph theoretical invariants useful in the generation of new MDs [1]. For instance, first the ε index was defined by this author using the Randić-type graph-theoretical invariant [6]. That is to say, this new index is analogous to the Randić branching index but calculated by edge degrees instead of vertex degrees. In a second work, Estrada also extended the edge adjacency matrix **E** in molecular graph in a 3D-**E** matrix in order to generate the so-called topographic edge-connectivity index $\varepsilon(\rho)$ [7], also using the Randić-type graph-theoretical invariant. Later, this author used the same edge adjacency relationships in the generation of a new family of TIs, spectral moments of the E-matrix [8]. The analogous concept of spectral moments of vertex-adjacency matrix had also been discussed previously by different authors [9]. Afterward, Estrada et al. [10] introduced an extended set of edge connectivity indices, ${}^{m}\epsilon_t(G)$, using the same way in which the branching index of Randić was extended to the series of molecular connectivity indices. Finally, a novel graph theoretical polynomial, $P_{\epsilon}(G,x)$, counting the edge connectivity was introduced by the same researcher [11]. Such edge-adjacency relationships will be applied in the present report in order to generate a series of bond-based MDs to be used in drug design and chemoinformatic studies.

In recent years the latter task has been becoming a growing research field of great interest in the pharmaceutical industries, where MDs play an important role for the analysis of molecular diversity and lead to optimization through well-established quantitative structure–property/activity relationships (QSPR/QSAR) studies [1, 12]. Although a lot of different MDs have been proposed until now, both in QSAR/QSPR modeling and drug design studies there still exists the need of new MDs, because each class of descriptors encodes some specific structural features and thus it is useful to have an exhaustive description of the molecular structure [2].

In fact, recently one of the present authors, Y.M-P, has introduced a new set of atom-level MDs of relevance to QSAR/QSPR studies and 'rational' drug design, non-stochastic and stochastic quadratic indices [$\boldsymbol{q}_k(\bar{x})$ and $\boldsymbol{q}_k(\bar{x})$, respectively] [13, 14]. These local (atom, group and atom-type) and total chemical indices are based on the calculation of quadratic maps in $\Re^n$ in canonical basis set. The description of the significance-interpretation and the comparison to other MDs was also performed [14]. This approach describes changes along the time in the electronic distribution throughout the molecular backbone. Specifically, the features of the *k*th total and local quadratic indices were illustrated by examples of various types of molecular structures, including chain length and branching as well as content of heteroatoms, and multiple bonds [14]. Additionally, the linear independence of the atom-type quadratic fingerprints to 229 other 0D–3D MDs was demonstrated. In this sense, it was concluded that local (atom-based) quadratic fingerprints are independent indices, which contain important structural information to be used in QSPR/QSAR and drug design studies [14].

These MDs are easily and quickly calculated, thus being suitable for both QSAR/QSPR modeling and drug design studies of large chemical databases. This—in silico—method has been successfully applied to the prediction of several physical, physicochemical and chemical properties of organic compounds [13, 14]. These atom-level MDs, and their stochastic forms [15, 16], have also been useful for the selection of novel subsystems of compounds having a desired property/activity. In this sense, it was successfully applied to the virtual (computational) screening of novel anthelmintic compounds, which were then synthesized and in vivo evaluated on *Fasciola hepatica* [17]. Studies for the fast-track discovery of novel antibacterial [18], paramphistomicide [15], antimalarial [16, 19], trichomonicidal [20], and tripanocidals

[21] lead-like chemicals were also conducted with this theoretical approach. In addition, the atom-based quadratic indices have been extended to consider three-dimensional features of small/medium-sized molecules based on the trigonometric-3D-chirality-correction factor approach [22, 23]. This approach has also been successfully employed in QSAR and in silico ADME studies of Caco-2 Permeability of Drugs [24–26]. Finally, promising results have been found in the modeling of the interaction between drugs and HIV Ψ-RNA packaging-region in the field of bioinformatics using the nucleic acid's quadratic indices [27]. An alternative formulation of our approach for structural characterization of proteins was also carried out recently [28]. This extended method was used to encompass protein stability studies—specifically how alanine substitution mutation on Arc repressor wild-type protein affects protein stability—by means of a combination of protein quadratic indices (macromolecular fingerprints) and statistical (linear and non-linear model) methods.

The main aim of this paper is propose a new extended local (bond, group and bond-type) and total (whole) MDs based on the adjacency of edges and based on quadratic maps similar to that typically defined by mathematicians in linear algebra. We also propose in this manuscript a new matrix representation of the molecule on the ''stochastic'' adjacency of edges and quadratic indices derived from there. These MDs, called bond-based quadratic indices, encode topological information given by the molecular graph, weighted by chemical information encoded in selected bond weightings. Finally, the correlation ability of the new MDs is tested in QSPR studies of some physicochemical properties of octanes, alkyl-alcohols, and 2-furylethylenes.

Theoretical scaffold

''... but it is often very difficult and complicated question to decide in what part of the theory the improvement has to be made.''

Max Planck

The basis of the extension of quadratic indices that will be given here is the edge-adjacency matrix considered and explicitly defined in the chemical graph-theory literature [29, 30], and rediscovered by Estrada as an important source of new MDs [4, 6–8, 10, 11]. In this section, we first will define the nomenclature to be used in this work, then the atom-based molecular vector ($\bar{x}$) will be redefined for bond characterization using the same approach as previously reported, and finally some new definition of bond-based non-stochastic and stochastic quadratic indices with its peculiar mathematical properties will be given.

Background in edge-adjacency matrix and new edge-relations: stochastic edge-adjacency matrix

Let $G = (V, E)$ be a simple graph, with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots e_m\}$ being the vertex- and edge-sets of G, respectively. Then G represents a molecular graph having n vertices and m edge (bonds). The edge-adjacency matrix $\mathbf{E}$ of G (likewise called bond adjacency matrix, $\mathbf{B}$) is a square and symmetric matrix whose elements e_{ij} are 1 if and only if edge i is adjacent to edge j [1, 4, 8]. Two edges are adjacent if they are incidental to a common vertex. This matrix corresponds to the vertex-adjacency matrix of the associated line graph. Finally, the sum of the ith row (or column) of $\mathbf{E}$ is named the edge degree of bond i, $\delta(e_i)$ [1, 4, 6–8, 10, 11].

On the other hand, by using the edge (bond)-adjacency relationships we can find other new relation for a molecular graph that will be introduced here. The kth 'stochastic' edge-adjacency matrix, $\mathbf{ES}^k$ can be obtained directly from $\mathbf{E}^k$. Here, ${}^k = [{}^k es_{ij}]$ is a square table of order m (m = number of bonds) and the elements ${}^k es_{ij}$ are defined as follows:

$$ {}^k es_{ij} = \frac{{}^k e_{ij}}{{}^k \mathrm{SUM}(E^k)_i} = \frac{{}^k e_{ij}}{{}^k \delta(e)_i} \tag{1} $$

where, ${}^k e_{ij}$ are the elements of the kth power of $\mathbf{E}$ and the SUM of the ith row of $\mathbf{E}^k$ are named the k-order edge degree of bond i, ${}^k\delta(e)_i$. Note that the matrix $\mathbf{ES}^k$ in Eq. 1 has the property that *the sum of the elements in each row* is 1. An $m \times m$ matrix with nonnegative entries having this property is called a ''**stochastic matrix**'' [31]. Recently, some authors have introduced stochastic approach to atomic relationships for derived new MDs [15, 16, 32–36].

Chemical information and bond-based molecular vector

The atom-based molecular vector ($\bar{x}$) used to represent small-to-medium size organic chemicals has been explained in some detail elsewhere [13, 14]. In a manner parallel to the development of $\bar{x}$, we present the

expansion of the bond-based molecular vector ($\bar{w}$). The components ($\bar{w}$) of w are numeric values, which represent a certain standard bond property (bond-label). That is to say, these weights correspond to different bond properties for organic molecules. Thus, a molecule having $5, 10, 15, \ldots, m$ bonds can be represented by means of vectors, with $5, 10, 15, \ldots, m$ components, belonging to the spaces $\Re^5, \Re^{10}, \Re^{15}, \ldots, \Re^m$, respectively; where m is the dimension of the real sets (R^m). This approach allows us encoding organic molecules such as 2-hydroxybut-2-enenitrile through the molecular vector $\bar{w} = [w_{\text{Csp3–Csp2}}, w_{\text{Csp2=Csp2}}, w_{\text{Csp2–Osp3}}, w_{\text{H–Osp3}}, w_{\text{Csp2–Csp}}, w_{\text{Csp≡Nsp}}]$. This vector belongs to the product space $\Re^6$.

These properties characterize each kind of bond (and bond-types) within the molecule. Diverse kinds of bond weights (w) can be used in order to codify information related to each bond in the molecule. These bond labels are chemically meaningful numbers such as standard bond distance [37, 38], standard bond dipole [37, 38] or even mathematical expressions involving atomic weights such as atomic log P [39], surface contributions of polar atoms [40], atomic molar refractivity [41], atomic hybrid polarizabilities [42], Gasteiger-Marsilli atomic charge [43], atomic electronegativity in Pauling scale [44] and so on. Here, we characterized each bond with the following parameter:

$$w = x_i/\delta_i + x_j/\delta_j \tag{2}$$

which characterizes each bond. In this expression x_i can be any standard weight of the atom i bonded with atom j. δi is the vertex (atom) degree of atom i. The use of each scale (bond property) defines alternative molecular vectors, $\bar{w}$.

Theory of non-Stochastic and stochastic total (whole) and local (bond, group and bond-type) quadratic indices

If a molecule consists of m bonds (*vector of* $\Re^m$), then the kth total quadratic indices are calculated as quadratic maps (quadratic form) in $\Re^m$ in canonical basis set. Specifically, the kth total non-stochastic and stochastic quadratic indices, $q_k(\bar{w})$ and ${}^{s}q_k(\bar{w})$, are computed from these kth non-stochastic and stochastic edge adjacency matrices, $\mathbf{E}^k$ and $\mathbf{ES}^k$, as shown in Eqs. 3 and 4, correspondingly:

$$q_k(\bar{w}) = \sum_{i=1}^{m}\sum_{j=1}^{m} {}^{k}e_{ij}w^i w^j = [W]^t\mathbf{E}^k[W] \tag{3}$$

$$ {}^{s}q_k(\bar{w}) = \sum_{i=1}^{m}\sum_{j=1}^{m} {}^{k}es_{ij}w^i w^j = [W]^t\mathbf{ES}^k[W] \tag{4}$$

where, m is the number of bonds of the molecule, and $w^1, \ldots, w^m$ are the coordinates of the bond-based molecular vector ($\bar{w}$) in the so-called canonical ('natural') basis. In this basis system, the coordinates of any vector $\bar{w}$ coincide with the components of this vector [45, 46]. For that reason, those coordinates can be considered as weights (bond-labels) of the edge of the molecular graph. The coefficients ${}^{k}e_{ij}$ and ${}^{k}es_{ij}$ are the elements of the kth power of the matrix $\mathbf{E}$(G) and $\mathbf{ES}$(G), correspondingly, of the molecular graph. The defining Eqs. 3 and 4 for $\boldsymbol{q}_k(\bar{w})$ and ${}^{s}\boldsymbol{q}_k(\bar{w})$, respectively, may be also written as the single matrix equation (see Eqs. 3 and 4), where $[W]$ is a column vector (an $m \times 1$ matrix) of the coordinates of $\bar{w}$ in the canonical basis of $\Re^m$ and $[W]^t$ (an $1 \times m$ matrix) is the transpose of $[W]$. Here, $\mathbf{E}^k$ and $\mathbf{ES}^k$ denote the matrices of quadratic maps with respect to the natural basis set.

In addition to total bond-based quadratic indices, computed for the whole molecule, a local-fragment (bond, group and bond-type) formalism can be developed. These MDs are termed local non-stochastic and stochastic quadratic indices, ${}_{kL}(\bar{w})$ and ${}^{s}\boldsymbol{q}_{kL}(\bar{w})$, respectively. The definition of these descriptors is as follows:

$$q_{kL}(\bar{w}) = \sum_{i=1}^{m}\sum_{j=1}^{m} {}^{k}e_{ijL}w^i w^j = [W]^t\mathbf{E}_{\mathbf{L}}^{\boldsymbol{k}}[W] \tag{5}$$

$$ {}^{s}q_{kL}(\bar{w}) = \sum_{i=1}^{m}\sum_{j=1}^{m} {}^{k}es_{ijL}w^i w^j = [W]^t\mathbf{ES}_{\mathbf{L}}^{\boldsymbol{k}}[W] \tag{6}$$

where, m is the number of bonds and ${}^{k}e_{ijL}[{}^{k}es_{ijL}]$ is the kth element of the row "i" and column "j" of the local matrix $\mathbf{E}_{\mathbf{L}}^{k}[\mathbf{ES}_{\mathbf{L}}^{k}]$. This local matrix is extracted from the $\mathbf{E}^k[\mathbf{ES}^k]$ matrix and contains information referred to the edges (bonds) of the specific molecular fragments and also of the molecular environment in k steps. The matrix $\mathbf{E}_{\mathbf{L}}^{k}[\mathbf{ES}_{\mathbf{L}}^{k}]$ with elements ${}^{k}e_{ijL}[{}^{k}es_{ijL}]$ is defined as follows:

$$\begin{aligned} {}^{k}e_{ijL}[{}^{k}es_{ijL}] &= {}^{k}e_{ij}[{}^{k}es_{ijL}] \textit{ if } \text{both } e_i \text{ and } e_j \text{ are edges} \\ &\quad \text{(bonds) contained within the} \\ &\quad \text{molecular fragment} \\ &= 1/2\ {}^{k}e_{ij}[{}^{k}es_{ijL}] \textit{ if } e_i \text{ and } e_j \text{ are edges (bonds)} \\ &\quad \text{contained within the molecular fragment} \\ &\quad \text{but not both} \\ &= 0 \text{ otherwise} \end{aligned} \tag{7}$$

Notice that the above scheme follows the spirit of a Mulliken population analysis [47]. Note also that for every partitioning of a molecule into Z molecular fragments there will be Z local molecular fragment matrices. In this case, if a molecule is partitioned into Z molecular fragments, the matrices $\mathbf{E}^k[\mathbf{ES}^k]$ can be partitioned into Z local matrices $\mathbf{E}^k_{\mathrm{L}}[^k_{\mathrm{L}}], L = 1, \ldots Z$, and the kth power of matrix $\mathbf{E}[\mathbf{ES}]$ is exactly the sum of the kth power of the local Z matrices. In this way, the total non-stochastic and stochastic bond-based quadratic indices are the sum of the non-stochastic and stochastic bond-based quadratic indices, respectively, of the Z molecular fragments:

$$q_k(\bar{w}) = \sum_{L=1}^{Z} q_{kL}(\bar{w}) \tag{8}$$

$${}^{\mathrm{s}}q_k(\bar{w}) = \sum_{L=1}^{Z} {}^{\mathrm{s}}q_{kL}(\bar{w}) \tag{9}$$

Bond, group and bond-type quadratic fingerprints are specific cases of local bond-based quadratic indices. In this sense, the kth bond-type quadratic indices are calculated by adding the kth bond quadratic indices for all bonds of the same type in the molecule. That is to say, this extension of the bond quadratic index is similar to group additive schemes, in which an index appears for each bond type in the molecule together with its contribution based of the bond quadratic index.

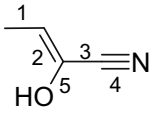
In the bond-type quadratic indices formalism, each bond in the molecule is classified into a bond-type (fragment). In this sense, bonds may be classified into bond types in terms of the characteristics of the two atoms which define the bond. For all data sets, including those with a common molecular scaffold as well as those with very diverse structure, the kth fragment (bond-type) quadratic indices provide much useful information. Thus, the development of the bond-type quadratic indices description provides the basis for application to a wider range of biological problems in which the local formalism is applicable without the need for superposition or a closely related set of structures. The bond-type descriptors combine three important aspects of structure information: (1) electron accessibility for the bonds of the same type, (2) presence/absence of the bond type, and (3) count of the bonds in the bond type.

Finally, these local MDs can be calculated by a chemical (or functional) group in the molecule, such as heteroatoms (O, N and S in all valence states and including the number of attached H-atoms), hydrogen bonding (H-bonding) to heteroatoms (O, N and S in all valence states), halogen atoms (F, Cl, Br and I), all aliphatic carbon chains (several bond-types), all aromatic bonds (aromatic rings), and so on. The group-level quadratic indices are the sum of the individual bond-level quadratic indices for a particular group of bonds. For all data set structures, the kth group-based quadratic indices provide also important information for QSAR/QSPR studies.

It is useful to perform a calculation on a molecule to illustrate the steps in the procedure. For this, in the next section we depict a pictorial representation of the calculus of the non-stochastic and stochastic quadratic indices of the bond matrix (both total and local) using a simple chemical example.

Sample calculation

The quadratic indices of the bond matrix are calculated in the following way. Considering the molecule of 2-hydroxybut-2-enenitrile as a simple example, we have the following labeled molecular graph and bond-based adjacency matrices (**E** and **ES**). The second ($k = 2$) and third ($k = 3$) power of these matrices and bond-based molecular vector, **W**, are also given:



$$E^0 = ES^0 = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} \quad E^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad E^2 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 3 & 1 & 1 & 1 \\ 1 & 1 & 3 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{bmatrix} \quad E^3 = \begin{bmatrix} 0 & 3 & 1 & 1 & 1 \\ 3 & 2 & 5 & 1 & 4 \\ 1 & 5 & 2 & 3 & 4 \\ 1 & 1 & 3 & 0 & 1 \\ 1 & 4 & 4 & 1 & 2 \end{bmatrix}$$

$$ES^1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \end{bmatrix} \quad ES^2 = \begin{bmatrix} 0.33 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.5 & 0 & 0.16 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0.16 & 0.16 & 0.16 & 0.16 & 0.33 \end{bmatrix} \quad ES^3 = \begin{bmatrix} 0 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.2 & 0.13 & 0.33 & 0.06 & 0.26 \\ 0.06 & 0.33 & 0.13 & 0.2 & 0.26 \\ 0.16 & 0.16 & 0.5 & 0 & 0.16 \\ 0.083 & 0.33 & 0.33 & 0.083 & 0.16 \end{bmatrix}$$

The molecule contains five localized bonds (corresponding to five edges in the H-suppressed molecular graph). To these we will associate the five "bond orbitals" (bond-labels) w_1, w_2, w_3, w_4, and w_5. Thus, $\bar{w} = [w_1, w_2, w_3, w_4, w_5] = [w_{(C-C)}, w_{(C=C)}, w_{(-C)}, w_{(C\equiv N)}, w_{(C-O)}]$ and each "bond orbitals" can be computed by Eq. 2 using, for instance, the atomic electronegativity in Pauling scale (x) [44] as atomic weight (atom-label):

$$
\begin{aligned}
w_1 &= x_C/1 + x_C/3 = 2.55/1 + 2.55/3 = 3.4\\
w_2 &= x_C/3 + x_C/4 = 2.55/3 + 2.55/4 = 1.4875\\
w_3 &= x_C/4 + x_C/4 = 2.55/4 + 2.55/4 = 1.275\\
w_4 &= x_C/4 + x_N/3 = 2.55/4 + 3.04/3 = 1.650833\\
w_5 &= x_C/4 + x_O/1 = 2.55/4 + 3.44/1 = 4.0775\\
&\text{and therefore, } \bar{w} = [3.4, 1.4875, 1.275, 1.650833, 4.0775].
\end{aligned}
$$

Each non-stochastic and stochastic total quadratic indices will have the form:

$$
\begin{aligned}
\boldsymbol{q}_k(\bar{w}) &= {}^k e_{11}(w^1)^2 + {}^k e_{21} w^1 w^2 + {}^k e_{31} w^1 w^3 + {}^k e_{41} w^1 w^4\\
&+ {}^k e_{51} w^1 w^5 + {}^k e_{12} w^1 w^2 + {}^k e_{22}(w^2)^2\\
&+ {}^k e_{32} w^2 w^3 + {}^k e_{42} w^2 w^4 + {}^k e_{52} w^2 w^5 + {}^k e_{13} w^1 w^3\\
&+ {}^k e_{23} w^2 w^3 + {}^k e_{33}(w^3)^2 + {}^k e_{43} w^3 w^4\\
&+ {}^k e_{53} w^3 w^5 + {}^k e_{14} w^1 w^4 + {}^k e_{24} w^2 w^4 + {}^k e_{34} w^3 w^4\\
&+ {}^k e_{44}(w^4)^2 + {}^k e_{54} w^4 w^5 + {}^k e_{15} w^1 w^5\\
&+ {}^k e_{25} w^2 w^5 + {}^k e_{35} w^3 w^5 + {}^k e_{45} w^4 w^5 + {}^k e_{55}(w^5)^2\\
&= \sum_{(i)} {}^k e_{ii}(w^i)^2 + 2\sum_{(i,j)} {}^k e_{ij} w^i w^j
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
{}^s\boldsymbol{q}_k(w) &= {}^k es_{11}(w^1)^2 + {}^k es_{21} w^1 w^2 + {}^k es_{31} w^1 w^3\\
&+ {}^k es_{41} w^1 w^4 + {}^k es_{51} w^1 w^5 + {}^k es_{12} w^1 w^2\\
&+ {}^k es_{22}(w^2)^2 + {}^k es_{32} w^2 w^3 + {}^k es_{42} w^2 w^4\\
&+ {}^k es_{52} w^2 w^5 + {}^k es_{13} w^1 w^3 + {}^k es_{23} w^2 w^3\\
&+ {}^k es_{33}(w^3)^2 + {}^k es_{43} w^3 w^4 + {}^k es_{53} w^3 w^5\\
&+ {}^k es_{14} w^1 w^4 + {}^k es_{24} w^2 w^4 + {}^k es_{34} w^3 w^4\\
&+ {}^k es_{44}(w^4)^2 + {}^k es_{54} w^4 w^5 + {}^k es_{15} w^1 w^5\\
&+ {}^k es_{25} w^2 w^5 + {}^k es_{35} w^3 w^5 + {}^k es_{45} w^4 w^5\\
&+ {}^k es_{55}(w^5)^2 = \sum_{(i)} {}^k es_{ii}(w^i)^2 + 2\sum_{(i,j)} {}^k es_{ij} w^i w^j
\end{aligned}
\tag{11}
$$

The ${}^k e_{ii}$'s and ${}^k es_{ii}$'s can be considered a measure of the attraction of an electron for a bond in the k step. The ${}^k e_{ij}$'s and ${}^k es_{ij}$'s are the terms of interaction between two bonds in the k step. The ${}^k e_{ij}$'s $= {}^k e_{ji}$'s are equal by symmetry (non-oriented molecular graph). However, ${}^k es_{ij} \neq {}^k es_{ji}$. This is a logical result because the kth es_{ij} elements are the transition probabilities with the 'electrons' moving from bond i to j at the discrete time periods t_k and it should be different in both senses. This result is in total agreement if the electronegativity of the two atom types in the bonds are taken into account. In this way, $\mathbf{E}^k$ and $\mathbf{ES}^k$ can be seen as graph-theoretic electronic-structure models [48]. In fact, quantum chemistry starts from the fact a molecule is made up of electrons and nuclei. The distinction here between bonded and non-bonded atoms is difficult to justify. Any two nuclei of a molecule interact directly and indirectly through the electrons present in the molecule. Only the intensity of this interaction varies in going from one pair of nuclei to another. In this sense, the electron in an arbitrary bond i can move (step-by-step) to other bonds at different discrete time periods t_k $(k = 0, 1, 2, 3, \ldots)$ through the chemical-bonding network. That is to say, the $\mathbf{E}^1$ and $\mathbf{ES}^1$ matrices consider the valence-bond electrons in one step and their power $(k = 0, 1, 2, 3 \ldots)$ can be considering as an interacting-electron chemical-network model in k step. This model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas [48].

On the other hand, the kth (k = 0–3) non-stochastic total quadratic indices can be expressed as the sum of the local (bond) quadratic indices for this molecule as follows:

$$
\begin{aligned}
\boldsymbol{q}_0(\bar{w}) &= q_{0L}(\bar{w}_1) + q_{0L}(\bar{w}_2) + q_{0L}(\bar{w}_3) + q_{0L}(\bar{w}_4)\\
&+ q_{0L}(\bar{w}_5) = 11.56 + 2.21265625\\
&+ 1.625625 + 2.72525069 + 16.6260063\\
&= 34.7495382\\
\boldsymbol{q}_1(\bar{w}) &= q_{1L}(\bar{w}_1) + q_{1L}(\bar{w}_2) + q_{1L}(\bar{w}_3) + q_{1L}(\bar{w}_4)\\
&+ q_{1L}(\bar{w}_5) = 5.0575 + 13.0193438\\
&+ 9.2001875 + 2.1048125 + 11.2640938\\
&= 40.6459375\\
\boldsymbol{q}_2(\bar{w}) &= q_{2L}(\bar{w}_1) + q_{2L}(\bar{w}_2) + q_{2L}(\bar{w}_3) + q_{2L}(\bar{w}_4)\\
&+ q_{2L}(\bar{w}_5) = 29.7585 + 17.0554271\\
&+ 16.30725 + 11.9121382 + 65.1108792\\
&= 140.144194\\
\boldsymbol{q}_3(\bar{w}) &= q_{3L}(\bar{w}_1) + q_{3L}(\bar{w}_2) + q_{3L}(\bar{w}_3) + q_{3L}(\bar{w}_4)\\
&+ q_{3L}(\bar{w}_5) = 38.9838333\\
&+ 55.7973646 + 44.17875 + 21.1141583\\
&+ 98.9031604 = 258.977267
\end{aligned}
$$

The terms in the summations for calculating the total quadratic indices are the so-called local (bond) quadratic indices. We have written these terms in the

consecutive order of the bond labels in the graph. For instance, the non-stochastic bond quadratic indices of order 0, 1, 2 and 3 for the bond labeled as 1 are 11.56, 5.0575, 29.7585, and 38.9838333, respectively.

The kth total stochastic quadratic indices values are also the sum of the kth local (bond) stochastic quadratic indices values for all bonds in the molecule:

$$
\begin{aligned}
{}^{s}\boldsymbol{q}_0(\bar{w}) &= {}^{s}q_{0L}(\bar{w}_1) + {}^{s}q_{0L}(\bar{w}_2) + {}^{s}q_{0L}(\bar{w}_3) \\
&\quad + {}^{s}q_{0L}(\bar{w}_4) + {}^{s}q_{0L}(\bar{w}_5) = 11.56 + 2.21265625 \\
&\quad + 1.625625 + 2.72525069 + 16.6260063 \\
&\quad = 34.7495382 \\
\boldsymbol{q}_1(\bar{w}) &= {}^{s}q_{1L}(\bar{w}_1) + {}^{s}q_{1L}(\bar{w}_2) + {}^{s}q_{1L}(\bar{w}_3) + {}^{s}q_{1L}(\bar{w}_4) \\
&\quad + {}^{s}q_{1L}(\bar{w}_5) = 3.37166667 \\
&\quad + 6.53105469 + 4.20156771 + 1.40320833 \\
&\quad + 4.6933724 = 20,2008698 \\
{}^{s}\boldsymbol{q}_2(\bar{w}) &= {}^{s}q_{2L}(\bar{w}_1) + {}^{s}q_{2L}(\bar{w}_2) + {}^{s}q_{2L}(\bar{w}_3) + {}^{s}q_{2L}(\bar{w}_4) \\
&\quad + {}^{s}q_{2L}(\bar{w}_5) = 8.40295833 \\
&\quad + 3.04720573 + 3.079125 + 3.20513877 \\
&\quad + 12.5680443 = 30.3024721 \\
{}^{s}\boldsymbol{q}_3(\bar{w}) &= {}^{s}q_{3L}(\bar{w}_1) + {}^{s}q_{3L}(\bar{w}_2) + {}^{s}q_{3L}(\bar{w}_3) \\
&\quad + {}^{s}q_{3L}(\bar{w}_4) + {}^{s}q_{3L}(\bar{w}_5) = 4.94428472 \\
&\quad + 4.80340608 + 3.65101563 + 2.80005408 \\
&\quad + 8.72457578 = 24.9233363
\end{aligned}
$$

Materials and methods

Dataset selection for QSPR studies

> *It is unusual for only one model to be compatible with experimental observations. Often data are not sufficiently extensive to discriminate among rival models and new experiments must be designed to answer the outstanding questions.*
>
> M. C. Kohn

The decisive criterion of quality for any MDs is its ability to describe structure-related properties of molecules. With this objective, first, we developed the QSPR models to describe the six physicochemical properties of octane isomers (see Table 1). The use of octane isomers as a very suitable data set for testing TIs has been advocated by Randić and Trinajstić [53, 54]. In fact, this dataset has been used by several researchers to evaluate the modeling power of their new MDs [2, 11, 55, 56]. This selection is recommended due to the fact that most of the physicochemical properties commonly studied in QSPR analyses with TIs are interrelated for data sets of compounds with different molecular weights, for instance for alkanes with two to nine carbon atoms. These correlations are not necessarily observed when the same descriptors are used in isomeric data sets of compounds, such as the octane database. Besides, these properties are hardly interrelated when octane isomers are used as a data set [57]. On the other hand, all TIs are designed to have (gradual) augmentation with the increments in the molecular mass. By this way, if we do the present study by using a series of chemicals having different molecular weights, we will find ''false'' interrelations between the descriptors by an overestimation of the size effects inherent to these indices [11, 55]. The same is also valid when the QSPR model is to be obtained.

It is not difficult to find ''good'' linear correlations between TIs and physicochemical properties of alkanes in data sets with great size variability [11, 55]. In fact, the simple use of the number of vertices in the molecular graph produced regression coefficients greater than 0.97 for most of the physicochemical properties of C_2–C_9 alkanes studied by Needham et al. [58]. However, when data sets of isomeric compounds are considered, typically correlations that have high correlation coefficients when molecules of different size were considered will no longer show such good linear correlation. In conclusion, if a new proposed MD is not able to model the variation of at least one property of octanes, then it probably does not contain any useful molecular information.

Moreover, octane isomers constituted a good set of compounds for comparative study, since many experimental data among their physicochemical properties are available. In this sense, we analyzed the quality of the QSPR models obtained to describe the BP, motor octane number (MON), heat of vaporization (HV), molar volume (MV), entropy (S), and heat of formation ($\Delta_f H$) of the octane isomers. In addition, regressions of octane properties based on the non-stochastic and stochastic bond-based quadratic indices will be compared to some regressions based on 2D (topologic) and 3D (geometric) descriptors taken from the literature [2, 11, 55]. Precisely, to evaluate the quality of the models based on our new bond level chemical descriptors we have taken as the reference: (a) the models published by Randić [52] based on diverse TIs such as the Wiener matrix invariants, (b) the equation published by Diudea [49] based on the SP indices, and (c) the best models obtained with a set constituted by the topological (69), WHIM (99), and GETAWAY descriptors (197) [2] (see Table 1).

Table 1 Statistical information for best QSPR models of selected physicochemical properties of octane isomers

Property	Method	Size	Q^2_{LOO}	R^2 (%)	s	F	Model descriptors	References
BP	Non-stochastic bond-based quadratic indices	3	97.50	98.41	0.828	304.73	$\mathbf{BP} = 194.62 - 0.55_1(w) + 0.024\boldsymbol{q}^{\mathbf{H}}_{4L}(w_{C-CH_3}) - 1.04\boldsymbol{q}_{0L}(w_{C-CH_3})$ **(12)**	
	Stochastic bond-based quadratic indices	3	96.8	98.41	0.865	278.69	$\mathbf{BP} = 91.30 + 43.62^{s}\boldsymbol{q}^{\mathbf{H}}_{2}(w) - 41.18^{s}\boldsymbol{q}^{\mathbf{H}}_{5}(w) - 1.91^{s}\boldsymbol{q}^{\mathbf{H}}_{0L}(w_{C-CH_3})$ **(13)**	
	Getaway + whim + top.	3	98.12	98.78	0.744		$^{2}\chi\,^{2}\bar{\chi}\,HATS_6(p)$	[2]
	Getaway	3	97.10	98.32	0.897		$HATS_2(v)\ R_4(u)\ R_6(v)$	[2]
	Getaway + whim + top.	2	96.62	97.58	1.013		$^{2}\chi HATS_6(p)$	[2]
	Topological	3		95.84	1.394		$S^3W\ S^4W\ SJ$	[49]
	Topological	2		94.78	1.508		$S^3W\ S^4W$	[49]
	Getaway	2	84.86	89.62	2.098		$HATS_2(m)\ R^+_4\ (u)$	[2]
	Topological	2		81.36	2.810		$WW\ x_1$	[50]
	Topological	1		78.85	2.90		Z	[51]
	Getaway + whim + top.	1	66.47	74.64	3.175		$HATS_2(m)$	[2]
	Topological	1		67.77	3.630		$^{2}\chi W$	[49]
MON	Non-stochastic bond-based quadratic indices	3	99.0	99.40	2.831	707.49	$\mathbf{MON} = 67.66 - 5.91\boldsymbol{q}^{\mathbf{H}}_{1}(w) + 2193.51\boldsymbol{q}^{\mathbf{H}}_{14}(w) - 2158.33\boldsymbol{q}^{\mathbf{H}}_{15}(w)$ **(14)**	
	Stochastic bond-based quadratic indices	3	97.9	98.60	4.289	305.89	$\mathbf{MON} = -793.95 + 12.27^{s}\boldsymbol{q}_0{}^{\mathbf{H}}(w) - 466.4^{s}\boldsymbol{q}_{4L}{}^{\mathbf{H}}(w_{C\text{-}CH3}) + 502.41^{s}\boldsymbol{q}_{6L}{}^{\mathbf{H}}(w_C\text{–}CH_3)$ **(15)**	
	Getaway + whim + top.	3	98.58	99.23	2.439		$^{v}I_D{}^{M}TsHATS_1(m)$	[2]
	Getaway	3	97.42	98.62	3.259		$HATS_4(u)\ \ HATS_7(v)\ R_7(p)$	[2]
	Topological	3		98.05	3.855		$S\ \chi^1W\chi^7W\chi^3W$	[49]
	Getaway + whim + top.	2	96.77	97.68	4.053		$Ts\ H_4(e)$	[2]
	Getaway	2	91.28	95.78	5.466		$HATS_7(m)\ \ R_4(u)$	[2]
	Topological	2		95.64	5.533		$S\chi^1W\ S\chi^3W$	[49]
	Topological	1		95.22	5.589		X^7W	[49]
	Getaway + whim + top.	1	90.83	92.40	7.069		Ts	[2]
	Topological	1		91.97	7.270		I_{wD}	[51]
	Getaway	1	85.64	88.98	8.515		REIG	[2]
HV	Non-stochastic bond-based quadratic indices	3	97.6	98.50	0.276	300.32	$\mathbf{HV} = 78.81\ 0.56\boldsymbol{q}_2{}^{\mathbf{H}}(w) – 0.19\boldsymbol{q}_3{}^{\mathbf{H}}(w)\ 1.17 \times 10^{-6}\boldsymbol{q}_{10}{}^{\mathbf{H}}(w)$ **(16)**	
	Stochastic bond-based quadratic indices	3	97.1	98.41	0.279	294.80	$\mathbf{HV} = 67.18 + 12.68^{s}\boldsymbol{q}_2{}^{\mathbf{H}}(w) – 14.75^{s}\boldsymbol{q}_3{}^{\mathbf{H}}(w)\ 0.098^{s}\boldsymbol{q}_1(w)$ **(17)**	
	Getaway + whim + top.	3	97.57	98.42	0.281		$^{0}\bar{\chi}^3\ \kappa\ R^+_6(u)$	[2]
	Getaway	3	95.46	97.18	0.375		$HATS_6(u)\ R_4(u)\ R^+_1(m)$	[2]
	Getaway + whim + top.	2	95.18	96.53	0.402		$^{2}\chi\ R^+_6(u)$	[2]
	Topological	3		95.65	0.459		$\chi^1W\ \chi^2W\ \chi^3W$	[49]
	Getaway	2	93.15	94.87	0.488		$HATS_4(u)\ R_6(e)$	[2]
	Topological	2		92.62	0.577		$^4W\ ^5W$	[49]
	Topological	1		91.78	0.429		Z	[51]
	Getaway + whim + top.	1	80.80	88.61	0.705		$^{2}\chi$	[2]
	Getaway	1	79.74	85.70	0.790		$R_2(m)$	[2]
	Topological	2		84.27	0.820		$WW\ x_1$	[50]
MV	Non-stochastic bond-based quadratic indices	3	98.6	99.20	0.245	575.01	$\mathbf{MV} = 214.18 + 1.29\boldsymbol{q}_1{}^{\mathbf{H}}(w) – 0.64\boldsymbol{q}_2{}^{\mathbf{H}}(w) + 0.095\boldsymbol{q}_{0L}(w_{C\text{-}CH3})$ **(18)**	
	Stochastic bond-based quadratic indices	3	98.2	99.00	0.287	417.95	$\mathbf{MV} = 236.71 + 1.91^{s}\boldsymbol{q}_1{}^{\mathbf{H}}(w) – 15.20^{s}\boldsymbol{q}_2{}^{\mathbf{H}}(w) + 0.43^{s}\boldsymbol{q}_{0L}{}^{\mathbf{H}}(w_C\text{–}CH_3)$ **(19)**	
	Getaway + whim + top.	3	75.96	92.01	1.825		$Ks\ R^+_6(u)\ RT^+(m)$	[2]
	Getaway	3	69.27	90.33	2.008		$HATS_6(p)\ RT^+(m)\ R_1(v)$	[2]
	Topological	3		88.29	2.210		$^5W^6W^7W$	[49]
	Getaway + whim + top.	2	54.49	84.96	2.419		$^{v}I_D{}^{M}R^+_6(u)$	[2]
	Getaway	2	45.49	81.79	2.662		$R^+_6(u)\ \ R_4(v)$	[2]
	Getaway + whim + top.	1	32.66	67.61	3.437		$R_6(v)$	[2]
	Topological	2		62.76	3.807		$^3W^4W$	[49]

Table 1 continued

Property	Method	Size	Q^2_{LOO}	R^2 (%)	s	F	Model descriptors	References
	Topological	1		60.85	3.780		^{7}W	[49]
Entropy (S)	Non-stochastic bond-based quadratic indices	3	89.7	93.90	1.236	71.05	$\mathbf{S} = 191.25 - 0.5 \times 10^{-3} {}^{\mathbf{H}}q_{10}(w) + 0.21 \times 10^{-3} {}^{\mathbf{H}}q_{11}(w) - 0.02 \times 10^{-3} {}^{\mathbf{H}}q_{12}(w)$ **(20)**	
	Stochastic bond-based quadratic indices	3	87.6	93.70	1.254	68.88	$\mathbf{S} = 120.99 - 6.76 {}^{s}{}^{\mathbf{H}}q_{1}(w) - 9.66 {}^{s}{}^{\mathbf{H}}q_{2}(w) + 13.41 {}^{s}{}^{\mathbf{H}}q_{15}(w)$ **(21)**	
	Getaway + whim + top.	3	97.17	97.96	0.711		$^{v}I_{D,deg}$ TWC $R^{+}{}_{2}(p)$	[2]
	Getaway + whim + top.	2	96.42	97.14	0.814		$^{v}I_{D,deg}$ TWC	[2]
	Getaway	3	93.45	95.84	1.016		I_{SH} $HATS_8(m)R_3(v)$	[2]
	Getaway	2	92.19	94.76	1.101		$I_{SH}R_3(v)$	[2]
	Getaway + whim + top.	1	89.86	92.51	1.274		$R_3(v)$	[2]
	Topological	1		91.10	1.400		$\chi^{[1/2]}$	[51]
	Topological	2		81.72	2.060		x_1x_2	[50]
Heat of	formation ($\Delta_f H$)		Non-				stochastic bond-based quadratic indices	3
	97.5		98.41	0.052	86.09		$\Delta_f\mathbf{H} = \boldsymbol{33.79} + \boldsymbol{0.03}{}^{\mathbf{H}}\boldsymbol{q}_2(\boldsymbol{w}) - 0.39\boldsymbol{q}_0(w) - 0.03\boldsymbol{q}_1(w)$ **(22)**	
	Stochastic bond-based quadratic indices	3	97.8	98.60	0.049	348.56	$\Delta_f\mathbf{H} = \boldsymbol{14.04} - \boldsymbol{0.08}{}^{s}{}^{\mathbf{H}}\boldsymbol{I}(\boldsymbol{w}) + 9.60 {}^{s}{}^{\mathbf{H}}\boldsymbol{q}_4(w) - 10.60{}^{\mathbf{H}}\boldsymbol{q}_5(w)$ **(23)**	
	Getaway + whim + top.	3	95.06	96.60	0.254		$HATS_5(m)$ $HATS_7(m)R_4(e)$	[2]
	Getaway + whim + top.	2	90.96	93.24	0.346		$^{2}\chi$ $HATS_2(e)$	[2]
	Getaway	2	90.18	92.87	0.356		$HATS_7(u)R_2(m)$	[2]
	Getaway + whim + top.	1	87.18	89.34	0.421		$HATS_2(m)$	[2]
	Topological	3		87.05	0.492		Ω_1 Ω_2 Ω_3	[52]
	Topological	2		86.86	0.478		Ω_1 Ω_2	[52]
	Topological	1		86.68	0.471		$1/^{2}\chi$	[51]
	Topological	2		78.70	0.570		WW x_1	[50]

Later, in order to illustrate the possibilities of our approach in the QSPR studies of heteroatomic molecules, we have selected the following two series to be investigated: (a) BP of 28 alkyl-alcohols (see Table 2) firstly studied by Kier and Hall using E-state/biomolecular encounter parameters [5] and recently by Estrada and Molina [4] using the local spectral moments of the edge adjacency matrix, and (b) a set of 34 2-furylethylene derivatives previously studied using total and local spectral moments, 2D/3D connectivity indices (vertex and edge ones) and to quantum chemical descriptors to model their specific rate constant (log k). These chemicals have different substituents at position 5 of the furan ring as well as at the β position of the exocyclic double bond [4, 59]. The structures of these 34 furylethylene derivatives are given in Table 3. The 2-furylethylene compounds have been well-known as antimicrobials, antitumoral, and cytotoxic during many years [60–62]. The values of the log k (for nucleophilic addition of the mercaptoacetic acid) of these compounds have been experimentally determined and reported in the literature [4, 59]. Tables 4 depict this value.

Computational strategies

''To calculate a molecule is not to understand a molecule.''

R. G. Parr

The total and local (bond-type) bond-based quadratic indices used to search for the best regression of the selected physicochemical properties of octanes were calculate by the ***TO***pological ***MO***lecular ***COM***puter ***D***esign-***C***omputer ***A***ided ''***R***ational'' ***D***rug ***D***esign ***TOMOCOMD-CARDD*** program [63]. This software is an interactive program for molecular design and bioinformatic research. The software was developed based on a user-friendly philosophy. That is to say, this computer graphics software shows a great efficiency of interaction with the user, without prior knowledge of programming skills (e.g. practicing pharmaceutic and organic chemist, teacher, university student, and so on). CARDD subprogram allows drawing the structures

Table 2 Experimental and predicted values of the boiling point of alcohols R-OH used in this study

Alcohol-R	Obsd.[a]	Pred.[b]	Res.$_{VC\text{-}LOO}$[c]	Pred.[d]	Res.$_{VC\text{-}LOO}$[e]
$(CH_3)_2CH–$	82.3	80.9	1.8	83.4	–1.6
$CH_3CH_2CH_2–$	97.2	98.8	–2.0	99.1	–2.4
$CH_3(CH_2)_3–$	117.7	118.0	–0.3	117.1	0.7
$CH_3CH(CH_3)CH_2–$	107.8	108.0	–0.2	107.6	0.3
$CH_3CH_2C(CH_3)_2–$	102.4	101.8	0.6	102.5	–0.1
$CH_3CH_2CH_2CH(CH_3)–$	119.3	119.5	–0.2	119.5	–0.2
$CH_3CH(CH_3)CH_2CH_2–$	131.1	130.2	1.0	128.8	2.5
$CH_3CH_2CH(CH_3)CH_2–$	128.0	130.4	–2.6	128.6	–0.7
$CH_3(CH_2)_4–$	137.9	137.2	0.8	136.8	1.3
$CH_3C(CH_3)_2CH(CH_3)–$	120.4	118.3	3.0	117.9	2.9
$CH_3(CH_2)_2C(CH_3)_2–$	121.1	121.3	–0.2	121.2	–0.2
$(CH_3CH_2)_2C(CH_3)–$	122.4	122.7	-0.3	122.6	–0.2
$CH_3CH_2C(CH_3)_2CH_2–$	136.5	138.1	-1.8	135.3	1.4
$CH_3CH(CH_3)CH_2CH(CH_3)–$	131.6	131.8	–0.3	131.3	0.3
$CH_3CH(CH_3)CH(CH_3CH_2)–$	126.5	128.1	–2.1	129.9	–3.7
$CH_3CH(CH_3)CH(CH_3)CH_2–$	144.5	141.8	2.9	141.2	3.6
$CH_3CH_2CH_2CH(CH_3)CH_2–$	149.0	149.6	–0.7	148.7	0.4
$CH_3(CH_2)_5–$	157.6	156.3	1.4	156.4	1.4
$(CH_3CH(CH_3))_2CH–$	138.7	137.1	2.4	140.6	–2.1
$CH_3CH(CH_3)CH_2CH(CH_3)CH_2–$	159.0	162.0	–3.3	161.7	–2.9
$(CH_3CH_2)_3C–$	142.0	143.7	–2.0	142.9	–1.2
$CH_3(CH_2)_6–$	176.4	175.5	1.0	176.0	0.4
$(CH_3CH_2CH_2)_2(CH_3)C–$	161.0	161.6	–0.7	160.7	0.4
$(CH_3(CH_2)_3)(CH_3CH_2)(CH_3)C–$	163.0	161.4	1.8	161.7	1.4
$CH_3CH(CH_3)CH_2(CH_2)_4–$	188.0	187.8	0.3	189.3	–1.5
$CH_3(CH_2)_7–$	195.1	194.7	0.5	195.7	–0.7
$CH_3(CH_2)_5C(CH_3)_2–$	178.0	177.1	1.0	178.7	–0.8
$(CH_3CH_2CH_2)_2(CH_3CH_2)C–$	182.0	182.7	–1.0	181.4	0.8

[a] Experimental values of BP (°C). [b,d] Predicted values using non-stochastic (Eq. 24) and stochastic (Eq. 25) bond-based linear indices. respectively. [c,e] Residual values of LOO CV process using non-stochastic and stochastic bond-based linear indices, correspondingly [$Res_{CV\text{-}LOO}$ = Bp(Obsd.) – Bp(Pred.$_{CV\text{-}LOO}$)]

(drawing mode) and calculating 2D (topologic), 3D-chiral (2.5D) and 3D (geometric and topographic) non-stocahstic and stochastic MDs (calculation mode).

In order to differentiate the atoms (bonds) in the molecule, the program use several appropriate weights (bond-labels). Here, we characterized each bond with the parameter described in Eq. 2, using the atomic electronegativity in Pauling scale [44] as atomic weights, x_i.

The bond-based *TOMOCOMD-CARDD* MDs computed in this study were the following:

1. kth (k = 15) total non-stochastic bond-based quadratic indices not considering and considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_k(\bar{w})$ and $\boldsymbol{q}_k(\bar{w})$, respectively].
2. kth (k = 15) total stochastic bond-based quadratic indices not considering and considering H-atoms in the molecular graph (G) [${}^{s}\boldsymbol{q}_k(\bar{w})$ and ${}^{s}\boldsymbol{q}_k^{H}(\bar{w})$, correspondingly].
3. kth (k = 15) bond-type (C-CH_3) non-stochastic and stochastic quadratic indices considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{kL}^{H}(\bar{w}_{C-CH_3})$ and ${}^{s}\boldsymbol{q}_{kL}^{H}$ $(\bar{w}_{C-CH_3})$, respectively]. These MDs were only calculated for octane isomers data sets.
4. kth (k = 15) group (heteroatoms: O, N, S and halogens) non-stochastic quadratic indices considering and non-considering H-atoms in the molecular graph (G) [$\boldsymbol{q}_{kL}^{H}(\bar{w}_E)$ and $\boldsymbol{q}_{kL}(\bar{w}_E)$, correspondingly]. These local MDs are putative molecular charge, dipole moment, and H-bonding acceptors.
5. kth (k = 15) group (heteroatoms: O, N, S and halogens) stochastic quadratic indices considering and non-considering H-atoms in the molecular graph (G) [${}^{s}\boldsymbol{q}_{kL}(\bar{w}_E)$ and ${}^{s}\boldsymbol{q}_{kL}(\bar{w})$, respectively]. These local MDs are putative molecular charge, dipole moment, and H-bonding acceptors.

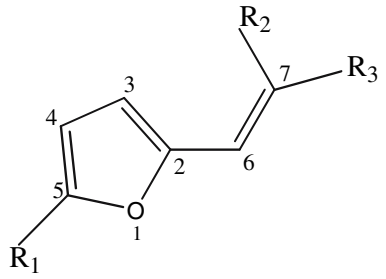
Chemometric analysis

The statistical, graph theoretical, and sensitivity analysis methods ... can identify the areas for further investigation that are likely to produce significant new results.

M. C. Kohn

These kth total and local bond-based quadratic indices were used as MDs for derived QSPRs. One of

Table 3 Chemical structures and numbering of atoms in the 2-furylethylene compounds used in this study



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

the difficulties with the large number of MDs is deciding which ones will provide the best regressions, considering both goodness of fit and the chemical meaning of the regression. In addition, as testing a large number of all possible combinations of variables would be a tedious task and time-consuming procedure, we have used a genetic algorithm (GA) input selection [64–71]. GAs are a class of algorithms inspired by the process of biological evolution in which species having a high fitness under some conditions can prevail and survive to the next generation; the best species can be adapted by crossover and/or mutation in the search for better individuals.

The software BuildQSAR [72] was employed to perform variable selection and QSAR modeling. The mutation probability was specified as 35%. The length of the equations was set three-four terms and a constant. The population size was established as 100. The GA with an initial population size of 100 rapidly converged (200 generations) and reached an optimal QSAR model in a reasonable number of GA generations.

The search for the best model can be processed in terms of the highest correlation coefficient (R) or F-test equations (Fisher-ratio's p-level [$p(F)$]), and the lowest standard deviation equations (s) [72]. The quality of models was also determined by examining the leave-one-out (LOO) cross-validation (CV) (q^2, s_{cv}) [73]. In recent years, the LOO press statistics (e.g., q^2) have been used as a means of indicating predictive ability. Many authors consider high q^2 values (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of a QSAR model.

Applications in QSPR studies

You know my methods. Apply them.

Conan Doyle

QSPR studies of octane isomers

The best linear models found using non-stochastic and stochastic total and bond-type quadratic indices are presented in Table 1. For each selected property of octane isomers, the statistical information for the best

Table 4 Experimental and calculated values of the specific rate constant for the reaction of nucleophilic addition of thiols (log k) to the exocyclic double bond of the studied 2-furylethylenes

No.	Obsd.[a]	Pred.[b]	Res.[c] $_{VC\text{-}LOO}$	Pred.[d]	Res.[e] $_{VC\text{-}LOO}$
1	6.591	6.591	0.000	6.493	0.128
2	6.518	6.694	–0.246	6.556	–0.060
3	6.914	6.968	–0.066	7.210	–0.371
4	6.982	6.996	–0.017	7.180	–0.243
5	7.176	6.511	0.835	7.198	–0.035
6	7.602	7.619	–0.027	7.627	–0.047
7	5.255	5.416	–0.212	5.052	0.275
8	6.763	6.499	0.492	6.604	0.317
9	5.623	5.497	0.248	5.470	0.354
10	3.813	3.784	0.036	3.577	0.398
11	3.840	3.875	–0.041	3.639	0.239
12	3.874	3.866	0.010	3.581	0.358
13	3.825	3.629	0.221	3.761	0.071
14	3.623	3.563	0.066	3.733	–0.129
15	3.751	3.630	0.136	3.774	–0.026
16	3.784	3.385	0.464	3.461	0.393
17	3.697	3.604	0.106	3.732	–0.040
18	3.705	3.583	0.143	3.683	0.025
19	3.697	3.951	–0.411	3.687	0.013
20	3.650	3.762	–0.155	3.718	–0.079
21	4.000	4.852	–0.962	4.278	–0.361
22	3.920	4.329	–0.438	4.111	–0.206
23	3.790	4.021	–0.252	4.005	–0.234
24	3.763	3.858	–0.106	3.792	–0.031
25	3.623	3.706	–0.098	3.715	–0.109
26	3.650	3.769	–0.136	3.912	–0.290
27	3.592	3.539	0.062	3.655	–0.072
28	3.584	3.516	0.105	3.571	0.016
29	3.590	3.415	0.255	3.403	0.263
30	2.987	2.821	0.624	2.920	0.229
31	3.273	3.524	–0.314	3.595	–0.409
32	2.140	2.288	–0.223	2.700	–0.731
33	3.553	3.120	0.772	3.436	0.204
34	5.557	5.523	0.038	4.876	0.768

[a] Experimental values of log k. [b,d] Predicted values using non-stochastic (Eq. 26) and stochastic (Eq. 27) bond-based linear indices, respectively. [c,e] Residual values of LOO CV process using non-stochastic and stochastic bond-based linear indices, respectively [$Res_{CV\text{-}LOO}$ = Bp(Obsd.) – Bp($Pred._{CV\text{-}LOO}$)]

regressions with 1, 2, and 3 MDs published so far are also depict in Table 1. Together with the LOO cross-validated explained variance (q^2 LOO), the determination coefficient (R^2), the standard estimate of the error (s), and Fischer ratio (F) are listed. The molecular descriptor symbols are reported in eighth column, and the last column in the table contains the references of the models taken from the literature.

As can be appreciated from the statistical parameters of regression equations in Table 1, all of the physicochemical properties were well described by bond-based quadratic indices. In this table we can observe that the statistical parameters for the models obtained with bond-based quadratic indices to describe MON (Eqs. 14 and 15; see Table 1), HV (Eqs. 16 and 17; see Table 1), heat of formation (ΔfH) (Eqs. 22 and 23; see Table 1) and MV (Eqs. 18 and 19; see Table 1) of octanes are better than those taken from the literature [2, 11, 55]. The last physicochemical property, that is, MV, is well-described exclusively by the bond-based quadratic indices. Note also that in the models based on the bond-level chemical quadratic indices, the two regressions for the BP (Eqs. 12 and 13; see Table 1) are better-to-similar than the models published so far [2, 11, 55]. Only the models found by us to describe entropy (S) (Eqs. 20 and 21, see Table 1) have significant differences with the precedent models obtained by applying the selection procedure to the set given by GETAWAY descriptors plus WHIM and TIs. According to the obtained QSPR results, it is possible to conclude that the new MDs encode some useful molecular information different from that of previous proposed indices. Moreover, they are quite diverse among themselves being able to describe well the variation of different properties of octanes.

Describing boiling points of 28 alkyl-alcohols

The first heteromolecules-based database of that will be studied here is composed by 28 alkyl-alcohols, 14 are primary, 6 secondary and 8 tertiary, for which the BP has been reported previously [4]. The best linear regression model obtained to describe the Bp of these chemicals using non-stochastic and stochastic bond-based quadratic indices is given below, respectively:

$$\begin{aligned}\mathbf{Bp}(^{\circ}\mathbf{C}) = &-98.33(\pm 2.82) - 1.08(\pm 0.03)\boldsymbol{q}_2^{\mathrm{H}}(\bar{w}) \\ &+ 6.19(\pm 0.11)\boldsymbol{q}_0(\bar{w}) \\ &+ 4.66\times 10^{-8}(\pm 7.55\times 10^{-9})\boldsymbol{q}_{13}(\bar{w})\end{aligned} \quad (24)$$

$$N = 28 \quad R^2 = 0.998 \; q^2 = 0.997 \quad s = 1.46$$
$$s_{CV} = 1.60 \quad F(3,24) = 3394.6 \quad P < 0.0001$$

$$\begin{aligned}\mathbf{Bp}(^{\circ}\mathbf{C}) = &-85.51(\pm 2.61) + 6.01(\pm 0.13)^{\mathrm{s}}\boldsymbol{q}_0(\bar{w}) \\ &- 30.83(\pm 0.77)^{\mathrm{s}}\boldsymbol{q}_3^{\mathrm{H}}(\bar{w}) \\ &+ 0.43(\pm 0.07)^{\mathrm{s}}\boldsymbol{q}_1(\bar{w})\end{aligned} \quad (25)$$

$$N = 28 \quad R^2 = 0.997 \quad q^2 = 0.996 \quad s = 1.59$$
$$s_{CV} = 1.66 \quad F(3,24) = 2857.1 \quad P < 0.0001$$

where N is the number of compounds, R^2 is the determination coefficient, s is the standard deviation of the regression, $q^2(s_{CV})$ is the square regression coefficient (standard deviation) obtained from the LOO cross validation procedure, and F is the Fisher ratio.

The values of experimental and calculated values of the Bp for the data set (both models) are given in

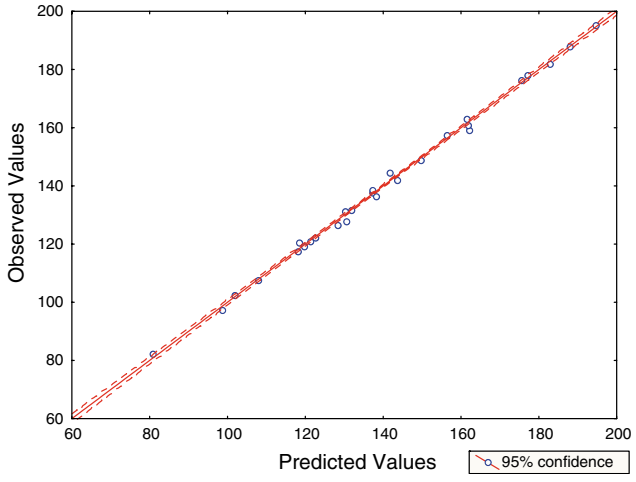


Fig. 1 Linear correlations of observed versus calculated boiling point according to the model obtained from non-stochastic bond-level quadratic indices (Eq. 24)

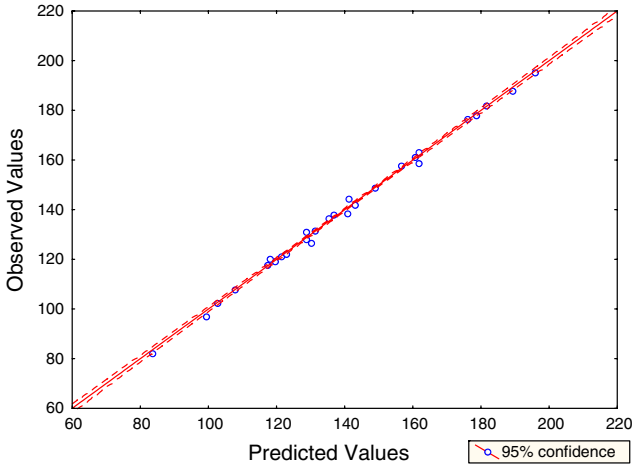


Fig. 2 Linear correlations of observed versus calculated boiling point according to the model obtained from non-stochastic bond-level quadratic indices (Eq. 25)

Table 2 and the linear relationships between them are illustrated in Figs. 1 and 2.

These models (Eqs. 24 and 25) explain more than 99% of the variance of the experimental Bp values. Similar-to-inferior equations were reported by Estrada and Molina [4] and Kier and Hall [5] using spectral moment and E-state as MDs, correspondingly. The statistical parameters of best equations obtained for those authors are given in Table 5. These models explain more than 98% and 92% of the variance of the experimental Bp values, respectively.

Predictability and stability of the obtained models using bond-based quadratic indices (Eqs. 24 and 25) to data variation is carried out here by means of LOO CV procedure. These models showed a q^2 (s_{CV}) of 0.997 (1.60) and 0.996 (1.66), correspondingly. Unfortunately, the authors (Estrada and Molina [4], and Kier and Hall [5]) do not report the result of the CV. It is remarkable that our models uses two variables less than the model obtained by Estrada and Molina [4] and the same variables than the model obtained by Kier and Hall [5]. However, Eqs. 24 and 25 explains a greater percent of the variance of the experimental Bp values than that the previously developed models do; showing a decrease in the standard error of 65.24% (62.14%) and 74.83% (72.59%), respectively, with regard to the results archived previously for Estrada and Molina [4, 5] (Kier and Hall [4, 5]). Table 5 summarizes the statistical parameters archived by all these approaches. As it can be observed in this Table, the obtained result using other atom-based ***TOMOCOMD-CARDD*** MDs were also included. The QSPR model derived with bond-level quadratic indices shown better results that obtained by some of the present author en previously studies [14, 74].

Modeling the specific rate constants (log k) of 34 2-furylethylenes derivatives

It has been clear from structure–activity relationship studies that the nucleophilic addition of thiol groups of some enzymes to the exocyclic double bond of 2-furylethylene derivatives are critical for the development of their antibacterial activity [59]. The specific rate constant (log k) of nucleophilic addition of the mercaptoacetic acid to the exocyclic double bond has an important role in the understanding of the biological behavior of these 2-furylethylene derivatives [59]. Consequently, I will study this parameter to compare the possibilities of molecular (bond-based) quadratic indices in QSPR and to compare these results to those obtained previously [4] using topological (total and local spectral moment and 2D connectivity indices), topographic and quantum chemical descriptors. This experiment will also possibility to compare the present result with the achieved using some atom-based ***TOMOCOMD-CARDD*** MDs [14, 74]. The molecular structures of such compounds are depicted in Table 3 and the overall result of the obtained QSPR models so far are summarized in Table 5 [4, 14, 74]

Many MDs are not useful to describe chemical reactions. In order to prove the applicability of this new approach in this kind of QSRR studies, we select a data set of 34 derivatives of 2-furylethylene and their specific rate constant of nucleophilic addition of the mercaptoacetic acid, k [4]. This reactive index was

Table 5 Statistical parameters for the models describing physicochemical properties of organic compounds using different MDs

Molecular descriptors	n	R^2	q^2	s	s_{CV}	F
Boiling point of alkyl-alcohols						
Local spectral moments [4]	5	0.982	*	4.2	*	23.8
E-State [5]	3	0.926	*	5.8	*	204
Atom-based linear indices [74]	2	0.984	0.981	3.78	3.91	748.57
Atom-based linear indices [74]	4	0.993	0.990	2.48	2.79	871.96
Atom-based quadratic indices [14]	2	0.990	0.989	2.91	2.969	1268.9
Non-stocastic bond-based quadratic indices (Eq. 24)	3	0.998	0.997	1.46	1.60	3394.6
Stocastic bond-based quadratic indices (Eq. 25)	3	0.997	0.996	1.59	1.66	2857.1
Reactivity index (log k) of 2-furylethylenes						
Conectivity indices [4]	7	0.821	*	0.681	*	17.1
Global spectral moments [4]	7	0.843	*	0.655	*	18.8
Local spectral moments [4]	7	0.964	*	0.320	*	70.4
Quantum chemical descriptors [4]	7	0.968	*	0.288	*	112.2
Atom-based linear indices [74]	6	0.973	0.948	0.26	0.33	161.22
Atom-based quadratic indices [14]	7	0.968	0.922	0.285	0.298	115.14
Non-stocastic bond-based quadratic indices (Eq. 26)	7	0.967	0.940	0.292	0.345	108.79
Stocastic bond-based quadratic indices (Eq. 27)	7	0.975	0.958	0.257	0.288	142.07

* Values are not reported in the literature

adequately described by Estrada and Molina using total and local spectral moments, connectivity indices and several quantum-chemical local MDs. All developed models had seven variables. The model obtained by these authors using the connectivity indices describes an 82% of the experimental values of log k, with a standard deviation of 0.681. In addition, these researchers obtained similar results using the global spectral moments as MDs in QSRR equation (R^2 = 84% and s = 0.655) [4]. The use of local MDs such as quantum-chemical or graph-theoretical (local spectral moments) produces a significant improvement in the statistical quality of the obtained models. In this sense, both models (quantum chemical and local spectral moments) explain more than 96% (96.8% and 96.4%) of the variance of the log k, with a standard deviation of 0.288 and 0.320, respectively. The MDs included in these equations clearly pointed to the identification of the reaction centers involved in the studied chemical interaction. That is to say, the molecular indices calculated for the atoms 2, 6 and 7 or for the bonds defined by these atoms (C_2–C_6 and C_6–C_7) were included in the obtained models [4]. These atoms are those involved in the exocyclic double bond of the 2-furylethylene and these are the ''target'' of the nucleophilic attack by thiol (mercapto) group.

Taking into account this logical result, we also calculated the kth bond-level quadratic indices for the bonds C_2–C_6 and C_6–C_7. The two best obtained models, using these local non-stochastic and stochastic quadratic indices as MDs, together with its statistical parameters, is given below, respectively:

$$\begin{aligned}\mathbf{log}\ \boldsymbol{k} = {} & 28.83(\pm 3.09) + 0.76(\pm 0.09)\boldsymbol{q}^{\mathrm{H}}_{3\mathrm{L}}(\bar{w}_{C_6-C_7}) \\ & - 0.05(\pm 0.01)\boldsymbol{q}_0(\bar{w}) \\ & - 1.72(\pm 0.18)\boldsymbol{q}_{2\mathrm{L}}(\bar{w}_{C_2-C_6}) - 0.06(\pm 0.008) \\ & \boldsymbol{q}^{\mathrm{H}}_{5\mathrm{L}}(\bar{w}_{C_6-C_7}) + 0.012(\pm 0.001)\boldsymbol{q}^{\mathrm{H}}_{4\mathrm{L}}(\bar{w}_{\mathrm{E}}) \\ & + 6.94\times 10^{-7}(\pm 9.69\times 10^{-8})\boldsymbol{q}^{\mathrm{H}}_{13\mathrm{L}}(\bar{w}_{C_6-C_7}) \\ & - 0.004(\pm 0.0005)\boldsymbol{q}^{\mathrm{H}}_{5\mathrm{L}}(\bar{w}_{\mathrm{E}})\end{aligned} \tag{26}$$

$$N = 34 \quad R^2 = 0.967 \quad q^2 = 0.940 \quad s = 0.292$$
$$s_{CV} = 0.345 \quad F(7,26) = 108.79 \quad P < 0.0001$$

$$\begin{aligned}\mathbf{log}\ \boldsymbol{k} = {} & 1.79(\pm 0.70)252.28(\pm 49.35)^{\mathrm{s}}\boldsymbol{q}^{\mathrm{H}}_{11\mathrm{L}}(\bar{w}_{rmC_6-C_7}) \\ & + 59.41(\pm 16.14)^{\mathrm{s}}\boldsymbol{q}^{\mathrm{H}}_{15\mathrm{L}}(\bar{w}_{C_6-C_7}) \\ & - 7.02(\pm 0.93)^{\mathrm{s}}\boldsymbol{q}^{\mathrm{H}}_{3\mathrm{L}}(\bar{w}_{C_6-C_7}) - 304.84(\pm 64.79)^{\mathrm{s}} \\ & \boldsymbol{q}^{\mathrm{H}}_{12\mathrm{L}}(\bar{w}_{C_6-C_7}) \\ & + 0.66(\pm 0.20)^{\mathrm{s}}\boldsymbol{q}^{\mathrm{H}}_{5\mathrm{L}}(\bar{w}_{\mathrm{E}}) - 0.39(\pm 0.15)^{\mathrm{s}}\boldsymbol{q}^{\mathrm{H}}_{3\mathrm{L}}(\bar{w}_{\mathrm{E}}) \\ & - 0.04(\pm 0.01)^{\mathrm{s}}\boldsymbol{q}_0(\bar{w})\end{aligned} \tag{27}$$

$$N = 34 \quad R^2 = 0.975 \quad q^2 = 0.958 \quad s = 0.257$$
$$s_{CV} = 0.288 \quad F(7,26) = 142.07 < 0.0001$$

Note, that our models, Eqs. 26 and 27, explained more than 96% and 97% of the observed variance, respectively. These statistics are better than those obtained previously (see Table 5) [4, 14, 74].

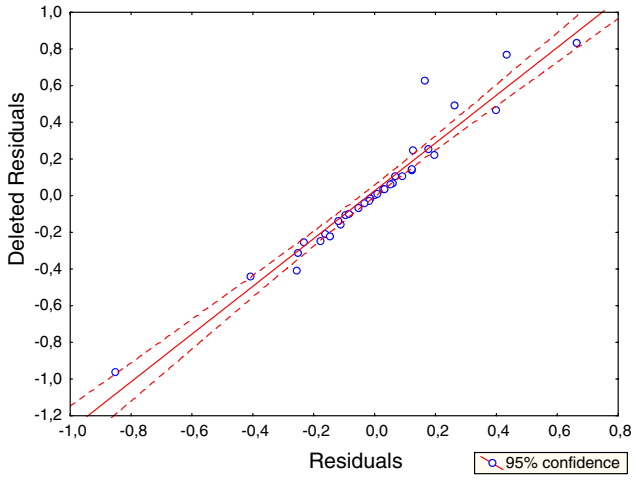


Fig. 3 Observed versus predicted (Eq. 26) log k of the specific rate constant for the reaction of nucleophilic addition of thiols to the exocyclic double bond of the 2-furylethylene derivatives

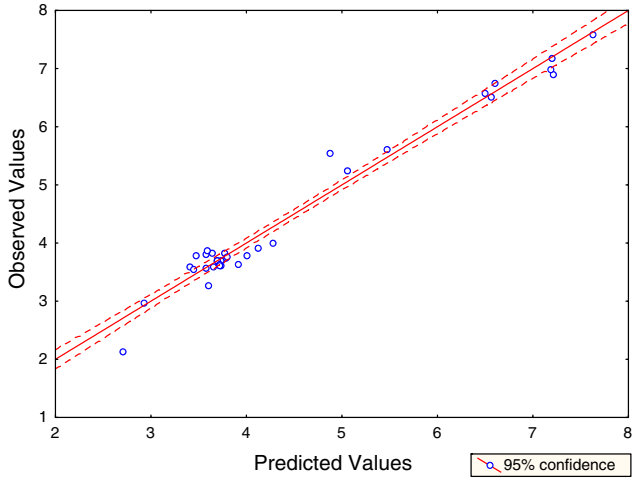


Fig. 4 Observed versus predicted (Eq. 27) log k of the specific rate constant for the reaction of nucleophilic addition of thiols to the exocyclic double bond of the 2-furylethylene derivatives

Table 4 depicts the experimental and calculated values of reactivity index (log k) from non-stochastic and stochastic bond-based quadratic indices. Plots of observed versus calculated log k from Eqs. 26 and 27 for data set of compounds are illustrated in Figs. 3 and 4, correspondingly.

Using LOO CV procedure, the models 26 and 27 had a LOO q^2 of 0.940 and 0.958, correspondingly. In this sense, the equations obtained with some vertex-based ***TOMOCOMD-CARDD*** MDs (Eqs. 11 and 16 in Ref. [14, 74], respectively) showed a smaller predictive abilities ($q^2(s_{CV})$ of 0.922 (0.298) and 0.948 (0.33), respectively) that Eq. 27 (q^2 = 0.958 and s_{CV} = 0.288), achieved with stochastic bond-level quadratic indices (see Table 5). Unfortunately, the previously study [4] do not report the result of the LOO CV experiment for log k (see Table 5).

Concluding remarks

> "*...It is lots of fun to blow bubbles but it is wiser to prick them yourself before someone else tries to.*"
>
> Oswald Avery

The non-stochastic and stochastic bond-based quadratic indices are novel sets of MDs. These indices can be used as a set of descriptors in QSPR (or QSAR) studies in the same way as they were used in the present work. However bond, bond-group, bond-type, and total quadratic indices can also be used in combination with pattern recognition techniques/artificial intelligent and in studies of similarity/dissimilarity features of molecules. The correlations found by these new sets of bond-level chemical descriptors for the description of six representative physicochemical properties of octane isomers can be considered as statistically significant. In addition, we need to take into account that the edge-based quadratic indices produce good correlations with all studied properties and for four of these properties this set of descriptors produced better models than the other 2D/3D TIs and geometrical set of indices previously tested by different researchers.

We also have shown here that total and local bond-based quadratic indices are useful MDs for modeling physicochemical properties of heteroatomic-organic chemicals. The obtained QSPR models for the description and prediction of the BP of alkyl-alcohols and the specific rate constant of nucleophilic addition of the mercaptoacetic acid to the exocyclic double bond of 34 2-furylethylene derivatives were statistically significant and better than other obtained previously using recognized methods, such as topological (total and local spectral moment and 2D (edge and vertex ones) connectivity indices), topographic and quantum chemical descriptors as well as some atom-level ***TOMOCOMD-CARDD MDs*** (see Table 5 for more details). This point is important because of the well-known broad applicability of these MDs in QSPR/QSAR studies. As a consequence, the bond-based quadratic indices represent a novel source for successful structure/activity-property models and drug design strategies. Applications of theses new bond(edge)-level MDs in molecular property/activity modeling, similarity/diversity analysis and biosilico drug discovery will be published in for coming papers.

Future outlooks

Science is to see what others have seen and to think what others have not thought.

Albert Szent-Gyorgi

At present, most of the researchers working in drug discovery with the use of TIs concentrate their efforts in the development of more powerful MDs. That is to say, although there have been many discoveries in the last years in the field of theoretical drug-design it is necessary to continue developing new MDs that can explain, by means of QSAR (or similar theoretical works) studies, different physicochemical properties and biological activities of chemical substances. Therefore, our research group is working in the definition of novel 2D/3D MDs based in algebra and group theory, geometric properties, and discrete mathematic, etc. We are also interested to apply our indices to codify central, planar and axial chirality as well as conformation alpha beta. An alternative formulation of our approach for structural characterization of macromolecules such as, proteins and nucleic acids will be also carried out proximately. In addition we have planned to concentrate our efforts in the use of more sophisticated statistical and artificial intelligent techniques to be used together with the *TOMOCOMD-CARDD* method in the derivation of QSPR models.

Acknowledgments

No person is an island; much is owed to many.

H. P. Schultz

Y. M-P thanks the program "*Estades Temporals per a Investigadors Convidats*" for a fellowship to work at Valencia University (2006–2007). The authors acknowledges financial support from the Spanish MEC DGI (Proyect No. CTQ2004-07768-C02-01/BQU) and Generalitat Valenciana (DGEUI INFO1-051 and INFRA03-047, and OCYT GRUPOS03-173). Comments of both reviewers contributed to a significant improvement of this material. Last, but not least, we are also indebted to the journal Editor Prof. Dr. Federico Gago for his kind attention.

"A bibliography is provided not only as a guide to further reading, but also in acknowledgment of works I have consulted and used"

R. A. Close

References

1. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Germany
2. Consonni V, Todeschini R, Pavan M (2002) J Chem Inf Comput Sci 42:682
3. Seybold PG, May M, Bagal UA (1987) J Chem Educ 64:575
4. Estrada E, Molina E (2001) J Mol Graph Model 20:54
5. Kier LB, Hall LH (1999) Molecular structure description. The Electrotopological state. Academic Press, New York
6. Estrada E (1995) J Chem Inf Comput Sci 35:31
7. Estrada E, Ramirez A (1996) J Chem Inf Comput Sci 36:837
8. Estrada E (1996) J Chem Inf Comput Sci 36:844
9. Marković S, Gutman I (1991) J Mol Struct (Theochem) 235:81
10. Estrada E, Guevara N, Gutman I (1998) J Chem Inf Comput Sci 38:428
11. Estrada E (1999) J Chem Inf Comput Sci 39:1042
12. Karelson M (2000) Molecular descriptors in QSAR/QSPR. John Wiley & Sons, New York
13. Marrero-Ponce Y (2003) Molecules 8:687
14. Marrero-Ponce Y (2004) Bioorg Med Chem 12:6351
15. Marrero-Ponce Y, Huesca-Guillen A, Ibarra-Velarde F (2005) J Mol Struct (Theochem) 717:67
16. Marrero-Ponce Y, Iyarreta-Veitia M, Montero-Torres A, Romero-Zaldivar C, Brandt CA, Avila PE, Kirchgatter K, Machado Y (2005) J Chem Inf Model 45:1082
17. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castañedo N, Ibarra-Velarde F, Huesca-Guillen A, Jorge E, del Valle A, Torrens F, Castro EA (2004) . J Comput Aided Mol Des 18:615
18. Marrero-Ponce Y, Medina-Marrero R, Torrens F, Martinez Y, Romero-Zaldivar V, Castro EA (2005) Bioorg Med Chem 13:2881
19. Montero-Torres A, García-Sánchez RN, Marrero-Ponce Y, Machado-Tugores Y, Nogal-Ruiz JJ, Martínez-Fernández AR, Arán VJ, Ochoa C, Meneses-Marcel A, Torrens F (2006) Eur J Med Chem 41:483–493
20. Meneses-Marcel A, Marrero-Ponce Y, Machado-Tugores Y, Montero-Torres A, Montero Pereira D, Escario JA, Nogal-Ruiz JJ, Ochoa C, Arán VJ, Martínez-Fernández AR, García Sánchez RN (2005) Bioorg Med Chem Lett 17:3838
21. Montero-Torres A, Celeste Vega M, Marrero-Ponce Y, Rolón M, Gómez-Barrio A, Escario JA, Arán VJ, Martínez-Fernández AR, Meneses-Marcel A (2005) Bioorg Med Chem 13:6264–6275
22. Marrero-Ponce Y, Díaz HG, Romero V, Torrens F, Castro EA (2004) Bioorg Med Chem 12:5331
23. Castillo-Garit JA, Marrero-Ponce Y, Torrens F (2006) Bioorg Med Chem 14:2398–2408
24. Marrero-Ponce Y, Cabrera MA, Romero V, Ofori E, Montero LA (2003) Int J Mol Sci 4:512
25. Marrero-Ponce Y, Cabrera MA, Romero V, González DH, Torrens F (2004) J Pharm Pharmaceut Sci 7:186
26. Marrero-Ponce Y, Cabrera MA, Romero-Zaldivar V, Bermejo M, Siverio D, Torrens F (2005) Internet Electron J Mol Des 4:124
27. Marrero-Ponce Y, Nodarse D, González HD, Ramos de Armas R, Romero-Zaldivar V, Torrens F, Castro E (2004) Int J Mol Sci 5:276
28. Marrero-Ponce Y, Medina R, Castro EA, de Armas R, González H, Romero V, Torrens F (2004) Molecules 9:1124
29. Rouvray DH (1976) Chemical applications of graph theory. Academic Press, London
30. Trinajstić N (1983) Chemical graph theory. CRC Press, Boca Raton FL
31. Edwards CH, Penney DE (1988) Elementary linear algebra. sPrentice-Hall, Englewood Cliffs, New Jersey, USA

32. Gonzalez-Diaz H, Tenorio E, Castanedo N, Santana L, Uriarte E (2005) Bioorg Med Chem 13:1523
33. Diaz HG, Bastida I, Castanedo N, Nasco O, Olazabal E, Morales A, Serrano HS, de Armas RR (2004) Bull Math Biol 66:1285
34. Gonzales-Diaz H, Gia O, Uriarte E, Hernadez I, Ramos R, Chaviano M, Seijo S, Castillo JA, Morales L, Santana L, Akpaloo D, Molina E, Cruz M, Torres LA, Cabrera MA (2003) J Mol Model 9:395
35. Marrero-Ponce Y, Torrens F (2006) Int J Mol Sci Accepted for publication
36. Marrero-Ponce Y, Medina-Marrero R, Martinez Y, Torrens F, Romero-Zaldivar V, Castro EA (2006) J Mol Mod 12:255
37. Vilar S, Estrada E, Uriarte E, Santana L, Gutierrez Y (2005) J Chem Inf Model 45:502
38. Potapov VM (1978) Stereochemistry. Mir, Moscow
39. Wang R, Gao Y, Lai L (2000) Perspect Drug Discov Des 19:47
40. Ertl P, Rohde B, Selzer P (2000) J Med Chem 43:3714
41. Ghose AK, Crippen GM (1987) J Chem Inf Comput Sci 27:21
42. Millar KJ (1990) J Am Chem Soc 112:8533
43. Gasteiger J, Marsilli MA (1978) Tetrahedron Lett 34:3181
44. Pauling L (1939) The nature of chemical bond. Cornell University Press, Ithaca (New York)
45. Browder A (1996) Mathematical analysis. An introduction. Springer-Verlag, New York Inc
46. Axler S (1996) Linear algebra done right. Springer-Verlag, New York
47. Walker PD, Mezey PG (1993) J Am Chem Soc 115:12423
48. Klein DJ (2003) Internet Electron J Mol Des 2:814
49. Diudea MV, Minailiuc OM, Katona G (1997) Rev Roum Chim 42:239
50. Randić M, Guo X, Oxley T, Krishnapriyan H, Naylor L (1994) . J Chem Inf Comput Sci 34:361
51. Randić M (1993) Croat Chim Acta 66:289
52. Randić M (1991) J Mol Struct (Theochem) 233:45
53. Randić M, Trinajstić N (1993) J Mol Struct (Theochem) 284:209
54. Randić M, Trinajstić N (1993) . J Mol Struct (Theochem) 300:551
55. Estrada E, Rodríguez L (1999) J Chem Inf Comput Sci 39:1037
56. Diudea MV (1996) J Chem Inf Comput Sci 36:535
57. Randić M (1991) J Math Chem 7:155
58. Needham DE, Wei I.-C, Seybold PG (1988) J Am Chem Soc 110:4186
59. Balaz S, Sturdik E, Rosenberg M, Augustin J, Skara B (1988) J Theor Biol 131:115
60. Dore JC, Viel C (1975) Farmaco 30:81
61. Sturdik E, Drobnica L, Balaz S (1985) Coll Czch Chem Comm 50:470
62. Blondeau JM,Castañedo N, Gonzalez O, Medina R, Silveira E (1999) Antimicrob Agents Chemother 11:1663
63. Marrero-Ponce Y, Romero V (2002) TOMOCOMD-CARDD software. TOMOCOMD (TOpological MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version can be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es Central University of Las Villas, Santa Clara, Villa Clara
64. Senese CL, Hopfinger AJ (2003) J Chem Inf Comput Sci 43:1297
65. Goldberg DE (1989) Genetic algorithms. Addison Wesley, Reading, MA
66. Willet P (1995) Trends Biotechnol 13:516–521
67. So SS, Karplus M (1996) J Med Chem 39:1521
68. So SS, Karplus M (1997) J Med Chem 40:4347
69. Rogers D, Hopfinger AJ (1994) J Chem Inf Comput Sci 34:854
70. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C (1997) J Am Chem Soc 119:10509
71. Senese CL, Hopfinger AJ (2003) J Chem Inf Comput Sci 43:2180
72. De Oliveira DB, Gaudio AC (2000) Quant Struct-Act Relat 19:599
73. Wold S, Erikson L (1995) In: van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH Publishers, New York, p 309
74. Marrero-Ponce Y, Castillo-Garit JA, Torrens F, Romero-Zaldivar V, Castro E (2004) Molecules 9:1100