

Linear Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”: Definition, Significance-Interpretation, and Application to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors

Yovani Marrero-Ponce*

Department of Pharmacy, Faculty of Chemical-Pharmacy, and Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba

Received February 3, 2004

This report describes a new set of molecular descriptors of relevance to QSAR/QSPR studies and drug design, atom linear indices $f_k(x_i)$. These atomic level chemical descriptors are based on the calculation of linear maps on \mathcal{R}^n [$f_k(x_i): \mathcal{R}^n \rightarrow \mathcal{R}^n$] in canonical basis. In this context, the k^{th} power of the molecular pseudograph’s atom adjacency matrix [$\mathbf{M}^k(\mathbf{G})$] denotes the matrix of $f_k(x_i)$ with respect to the canonical basis. In addition, a local-fragment (atom-type) formalism was developed. The k^{th} atom-type linear indices are calculated by summing the k^{th} atom linear indices of all atoms of the same atom type in the molecules. Moreover, total (whole-molecule) linear indices are also proposed. This descriptor is a linear functional (linear form) on \mathcal{R}^n . That is, the k^{th} total linear indices is a linear map from \mathcal{R}^n to the scalar \mathcal{R} [$f_k(x): \mathcal{R}^n \rightarrow \mathcal{R}$]. Thus, the k^{th} total linear indices are calculated by summing the atom linear indices of all atoms in the molecule. The features of the k^{th} total and local linear indices are illustrated by examples of various types of molecular structures, including chain-lengthening, branching, heteroatoms-content, and multiple bonds. Additionally, the linear independence of the local linear indices to other 0D, 1D, 2D, and 3D molecular descriptors is demonstrated by using principal component analysis for 42 very heterogeneous molecules. Much redundancy and overlapping was found among total linear indices and most of the other structural indices presently in use in the QSPR/QSAR practice. On the contrary, the information carried by atom-type linear indices was strikingly different from that codified in most of the 229 0D–3D molecular descriptors used in this study. It is concluded that the local linear indices are an independent indices containing important structural information to be used in QSPR/QSAR and drug design studies. In this sense, atom, atom-type, and total linear indices were used for the prediction of pIC_{50} values for the cleavage process of a set of flavone derivatives inhibitors of HIV-1 integrase. Quantitative models found are significant from a statistical point of view (R of 0.965, 0.902, and 0.927, respectively) and permit a clear interpretation of the studied properties in terms of the structural features of molecules. A LOO cross-validation procedure revealed that the regression models had a fairly good predictability (q^2 of 0.679, 0.543, and 0.721, respectively). The comparison with other approaches reveals good behavior of the method proposed. The approach described in this paper appears to be an excellent alternative or guides for discovery and optimization of new *lead* compounds.

1. INTRODUCTION

Agrochemical and pharmaceutical industries require, more than over, a new approach able to answer the challenge of discovering new lead drugs with minimum cost.¹ Economic pressures explain the interest of these industries in molecular design related methods in order to increase the effectiveness in lead generation and optimization.¹ This is manifested e.g., by gradually increasing interest shown by these firms in Quantitative Structure–Activity/Property Relationships (QSAR/QSPR) studies directed to rationalization of search for new biologically active molecules.

A crucial component of the QSAR/QSPR research is the identification of molecular descriptors relevant to the physical, chemical, or biological property of interest. Molecular descriptors permit the translation of the chemical structure into quantitative information.² That is, “the molecular

descriptor is the final result of logical and mathematical procedures which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment”.³

At present, there are a great number of molecular descriptors that can be used in QSAR/QSPR studies.⁴ Among them, the so-called topological indices (TIs) have found major applications in medicinal chemistry and molecular modeling.^{5–10} TIs are molecular descriptors derived from graph-theoretical invariants.^{3–6,10–15} These indices codify structural information contained in “molecular connectivities” and can be considered as structure-explicit descriptors.^{3–6,10–15}

The development of novel TIs continues in an accelerated way by using different approach. Recently, several molecular descriptors based on the two-dimensional topological structure of molecules have been defined and tested in QSAR models,^{15–20} showing that definition of novel molecular descriptors is a promising field in medicinal chemistry (see Todeschini and Consonni,³ Karelson,⁴ Devillers and Bala-

* Corresponding author phone: 53-42-281192, 281473; fax: 53-42-281130, 281455; e-mail: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es.

ban,¹¹ and Estrada and Uriarte¹⁰ for an exhaustive compilation). In this context, the present author has introduced the novel computer-aided molecular design scheme **TOMOCOMD** (acronym of **TO**pological **MO**lecular **COM**puter **D**esign). It calculates several new families of topologic molecular descriptors. One of these families has been defined as molecular quadratic indices by analogy with the quadratic mathematical forms.²¹ This point of view was successfully applied to the prediction of physical properties and Caco-2 permeability of organic compounds and drugs, respectively.^{21–23} The method is very flexible and makes possible the study of small molecules as well as macromolecules such as nucleic acids.²⁴

Interestingly, molecular quadratic indices can be generalized to allow the codification of 3D-structural features.²⁵

The main aim of this paper is to propose a total and local definition of linear indices of the “molecular pseudograph’s atom adjacency matrix” and to indicate the most important characteristic for these new indices by means of several structure changes in organic molecules, including chain-lengthening, branching, heteroatoms-content, and multiple bonds. Later, I will check if the information contained in the total and local linear indices is different from that of other 0D, 1D, 2D, and 3D molecular descriptors presently in use in QSPR/QSAR and drug design practice. Finally, to test the QSAR applicability of the present approach, I will develop quantitative models toward the prediction of negative log(IC₅₀) values for the cleavage process of a set of flavones inhibitors of HIV-1 integrase (HIV-1 IN). The leave-one-out (LOO) cross-validation procedure will be used to corroborate the predictive power of the models.

2. THEORETICAL APPROACH

Molecular vector (X) used to represent small-to-medium sized organic compounds have been explained in some detail elsewhere.^{21–25} However, this work gives an overview of this approach.

The molecular vector (X) is constructed in order to calculate the linear indices for a molecule where the components of this vector are numeric values, which represent a certain atomic property. These properties characterize each kind of atom within the molecule. Such properties can be the electronegativity, density, atomic radii, and so on. For instance, the Mulliken electronegativity (X_A)²⁶ of the atom A takes the values $X_H = 2.2$ for hydrogen, $X_C = 2.63$ for carbon, $X_O = 3.17$ for oxygen, $X_{Cl} = 3.0$ for chlorine, and so on.

Thus, a molecule having 5, 10, 15,..., n atoms can be represented by means of vectors, with 5, 10, 15,..., n components, belonging to the spaces \mathcal{R}^5 , \mathcal{R}^{10} , \mathcal{R}^{15} ,..., \mathcal{R}^n , respectively. Where n is the dimension of the real sets (\mathcal{R}^n).

This approach allows us encoding organic molecules such as acetic acid (suppressed H-atoms) throughout the molecular vector $X = [X_C, X_C, X_O, X_O] = [2.63, 2.63, 3.17, 3.17]$, in the X_A -electronegativity scale.²⁶ This vector belongs to the product space \mathcal{R}^4 . The use of other scales (atom properties) defines alternative molecular vectors.

2.1. Local (Atom) Linear Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”. If a molecule

consists of n atoms (vector of \mathcal{R}^n), then the k^{th} atom linear indices, $f_k(x_i)$, are calculated as linear maps on \mathcal{R}^n [$f_k(x_i): \mathcal{R}^n \rightarrow \mathcal{R}^n$; thus $f_k(x_i)$: endomorphism on \mathcal{R}^n] in canonical basis as shown in eq 1

$$f_k(x_i) = \sum_{j=1}^n {}^k a_{ij} X_j \quad (1)$$

where ${}^k a_{ij} = {}^k a_{ji}$ (symmetric square matrix), n is the number of atoms of the molecule, and X_j are the coordinates of the molecular vector (X) in a set of basis vectors of \mathcal{R}^n . One can choose the basis vectors; the coordinates of the same vector will be different.^{27–30} The values of the coordinates depend thus in an essential way on the choice of the basis. With the so-called canonical (“natural”) basis, e_j denotes the n -tuple having 1 in the j^{th} position and 0’s elsewhere. In the canonical basis, the coordinates of any vector X coincide with the components of this vector.^{27–30} For that reason, those coordinates can be considered as weights (atom labels) of the vertices of the molecular pseudograph.^{21–25}

The coefficients ${}^k a_{ij}$ are the elements of the k^{th} power of the matrix $\mathbf{M}(G)$ of the molecular pseudograph (G). Here, $\mathbf{M}(G) = \mathbf{M} = [a_{ij}]$ denotes the matrix of $f_k(x_i)$ with respect to the natural basis. In this matrix n is the number of vertices (atoms) of G, and the elements a_{ij} are defined as follows^{21–25}

$$\begin{aligned} a_{ij} &= P_{ij} \text{ if } i \neq j \text{ and } \exists e_k \in E(G) \\ &= L_{ii} \text{ if } i = j \\ &= 0 \text{ otherwise} \end{aligned} \quad (2)$$

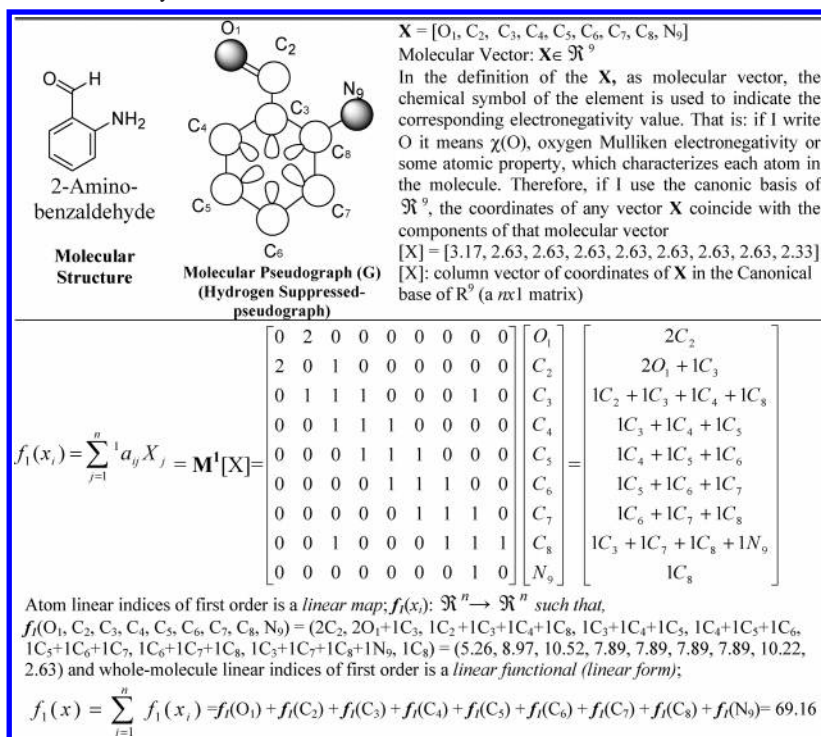
where $E(G)$ represents the set of edges of G. In this adjacency matrix $\mathbf{M}(G)$ the row i and column i correspond to vertex v_i from G. P_{ij} is the number of edges between vertices v_i y v_j , and L_{ii} is the number of loops in v_i .

Given that $a_{ij} = P_{ij}$, the elements a_{ij} of this matrix represent the number of bonds between an atom i and other j . The matrix \mathbf{M}^k provides the number of walks of length k that links the vertices v_i and v_j . For this reason, each edge in \mathbf{M}^1 represents 2 electrons belonging to the covalent bond between atoms v_i and v_j ; e.g. the inputs of \mathbf{M}^1 are equal to 1, 2, or 3 when simple, double, or triple bonds appear between vertices v_i and v_j , respectively. On the other hand, molecules containing aromatic rings with more than one canonical structure are represented by a pseudograph. It happens for substituted aromatic compounds such as pyridine, naphthalene, quinoline, and so on, where the presence $\text{PI}(\pi)$ electrons is accounted by means of loops in each atom of the aromatic ring. Conversely, aromatic rings having only one canonical structure, such as furan, thiophene, and pyrrol are represented by a multigraph.^{21–25}

Note, that atom’s linear indices are defined as a linear transformation $f_k(x_i)$ on an molecular vector space \mathcal{R}^n . This map is a correspondence that assigns to every vector X in \mathcal{R}^n a vector $f(x)$ in such a way that

$$f(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1 f(X_1) + \lambda_2 f(X_2) \quad (3)$$

for any scalar λ_1 , λ_2 and any vector X_1 , X_2 in \mathcal{R}^n .

Table 1. Definition and Calculation of Total (Whole-Molecule) and Local (Atom) Linear Indices of the Molecular Pseudograph's Atom Adjacency Matrix of the 2-Aminobenzaldehyde Molecule

atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$
Local (Atom) and Total (Whole-Molecule) Linear Indices of Order 0–5 ($k = 0-5$)						
O1	3.17	5.26	17.94	42.08	146.96	400.72
C2	2.63	8.97	21.04	73.48	200.36	676.25
C3	2.63	10.52	37.6	116.2	382.33	1193.57
C4	2.63	7.89	26.3	87.57	277.41	894.29
C5	2.63	7.89	23.67	73.64	234.55	739.87
C6	2.63	7.89	23.67	73.34	227.91	721.81
C7	2.63	7.89	26	80.93	259.35	820.73
C8	2.63	10.22	31.26	105.08	333.47	1080.23
N9	2.33	2.63	10.22	31.26	105.08	333.47
Total	23.91	69.16	217.7	683.58	2167.42	6860.94

The defining eq 1 for $f_k(x_i)$ may be written as the single matrix equation

$$f_k(x_i) = \begin{bmatrix} X'_1 \\ \mathbf{M} \\ X'_n \end{bmatrix}^k = \begin{bmatrix} a_{11} & \Lambda & a_{1n} \\ \mathbf{M} & & \mathbf{M} \\ a_{n1} & \Lambda & a_{nn} \end{bmatrix}^k \begin{bmatrix} X_1 \\ \mathbf{M} \\ X_n \end{bmatrix} \quad (4)$$

or in the more compact form

$$f_k(x_i) = [X']^k = \mathbf{M}^k[X] \quad (5)$$

where $[X]$ is a column vector (a $nx1$ matrix) of the coordinates of \mathbf{X} in the canonical basis of \mathcal{R}^n and \mathbf{M}^k the k^{th} power of the matrix \mathbf{M} of the molecular pseudograph (map's matrix).

Note, that this approach is rather similar to the **LCAO-MO** (Linear Combinations of Atomic Orbitals-Molecular Orbitals) method. Really, the approach (for $k = 1$) is a quite similar approximation to the Hückel MO method, due to the formalism each MO ψ_i is composed of n valence AOs of atoms in a molecule.

The main idea of the **LCAO-MO** method is that the electrons in a molecule are accommodated in definite MOs just as in an atom where they are accommodated in definite

AOs. Normally MOs made up as LCAO of atoms composing the system, i.e., are written in the form

$$\psi_i = \sum_{j=1}^n c_{ij} \varphi_j \quad (6)$$

where i is the number of the MO y [in our case, $f_i(x_i)$]; j are the numbers of atomic φ -orbitals (in our case, X_j); and c_{ij} (in our case, a_{ij}) are the numerical coefficients defining the contributions of individuals AOs into the given MO. Such a way of constructing a MO is based on the assumption that an atom represented by a definite set of orbitals remains distinctive in the molecule.

It is useful to perform a calculation on a molecule to illustrate the steps in the procedure. For this, I use the 2-aminobenzaldehyde molecule. Table 1 depicts the calculation of the linear indices of the molecular pseudograph's atom adjacency matrix for 2-aminobenzaldehyde. From Table 1, I extract the X -values (Mulliken electronegativity)²⁶ for each atom and the molecular vector \mathbf{X} , for encoding whole-organic molecule, is obtained. Additionally, all valence-bond electrons (c - and n -networks) in one step are revealed in \mathbf{M}^1 matrix. Then, the local (and total) linear indices of first-order values, $f_1(x_i)$, for each atom are calculated. Neverthe-

Table 2. Chemicals Data Set for Analysis of Principal Components³¹

no. chemical	no. chemical	no. chemical	no. chemical
01 methane	12 2-butene	23 nitrobenzene	34 purine
02 ethane	13 cyclopropane	24 fluorobenzene	35 dibenzofuran
03 <i>n</i> -propane	14 cyclobutane	25 chlorobenzene	36 ethanol
04 <i>n</i> -butane	15 cyclopentane	26 bromobenzene	37 trifluoroethanol
05 <i>n</i> -pentane	16 cyclohexane	27 iodobenzene	38 2-aminoethanol
06 <i>n</i> -hexane	17 cyclohexanone	28 benzamide	39 propanol
07 isobutane	18 benzene	29 naphthalene	40 2-propanone
08 neopentane	19 toluene	30 anthracene	41 2-propanol
09 2-methylpentane	20 phenol	31 pyrrole	42 2-propylamine
10 <i>cis</i> -2-butene	21 benzoic acid	32 furan	
11 <i>trans</i> -2-butene	22 aniline	33 thiophen	

less, the k^{th} ($k = 0-5$) local and total values are shown at the bottom of Table 1.

2.2. Total (Whole-Molecule) Linear Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”. Total linear indices are *linear functionals*²⁷⁻³⁰ (some mathematicians use the term *linear form*, which means the same as linear functional) on \mathcal{R}^n . That is, the k^{th} total linear index is a linear map from \mathcal{R}^n to the scalar \mathcal{R} [$f_k(x): \mathcal{R}^n \rightarrow \mathcal{R}$]. The mathematical definition of these molecular descriptors is the following

$$f_k(x) = \sum_{i=1}^n f_k(x_i) \quad (7)$$

where n is the number of atoms and $f_k(x_i)$ are the atom’s linear indices (linear maps) obtained by eq 1. Then, a linear form $f_k(x)$ can be written in matrix form

$$f_k(x) = [u]^t [X']^k \quad (8)$$

or

$$f_k(x) = [u]^t M^k [X] \quad (9)$$

for each molecular vector $X \in \mathcal{R}^n$. $[u]^t$ is a n -dimensional unitary row vector. As can be seen, the k^{th} total linear index is calculated by summing the local (atom) linear indices of all atoms in the molecule.

2.3. Local (Atom-type) Linear Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”. In addition to atom linear indices computed for each atom in the molecule, a local-fragment (atom-type) formalism can be developed. The k^{th} atom-type linear index of the molecular pseudograph’s atom adjacency matrix is calculated by summing the k^{th} atom linear indices of all atoms of the same atom type in the molecule. Consequently, if a molecule is partitioned in Z molecular fragments, the total linear indices can be partitioned in Z local linear indices $f_{kL}(x)$, $L = 1, \dots, Z$. That is to say, the total linear indices of order k can be expressed as the sum of the local linear indices of the Z fragments of the same order:

$$f_k(x) = \sum_{L=1}^Z f_{kL}(x) \quad (10)$$

In the atom-type linear indices formalism, each atom in the molecule is classified into an atom-type (fragment), such as heteroatoms (O, N, and S), H-bonding to heteroatoms, halogens atoms, aliphatic carbon chain, aromatic atoms

(aromatic rings), and so on. For all data sets, including those with a common molecular scaffold as well as those with very diverse structure, the k^{th} fragment (atom-type) linear indices provide much useful information.

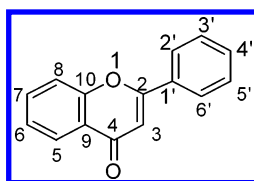
3. DATA AND METHODS

3.1. Data Sets. 3.1.1. Chemicals Data set for Analysis of Principal Components. The 42 chemicals used for this study were selected from DRAGON’s example data³¹ and are listed in Table 2. As evident, the sample is relatively small, but very heterogeneous, thus allowing for a general characterization of total and local linear indices information independently from individual chemical sets.

3.1.2. Biological Data Set for QSAR Study. Several key enzymes involved in the replication cycle of HIV can be targeted for chemotherapeutic intervention, most notably, reverse transcriptase and protease.^{32,33} HIV-1 IN is another such enzyme whose inhibition may be efficacious in the treatment of acquired immunodeficiency syndrome (AIDS), since this enzyme is required for viral replication, yet it is not indigenous to the human host.³⁴ IN catalyzes the integration of viral DNA into host DNA in two steps: 3'-processing (cleavage) and strand transfer (integration or end joining). First, IN cleaves the last two nucleotides from each 3'-end of the linear viral DNA. The subsequent DNA strand transfer reaction involves the nucleophilic attack of these 3'-ends on host chromosomal DNA.³⁵

Broadly, all inhibitors of HIV-1 IN can be classified as those containing catechol substructure and those lacking it. Quercetin, representing the flavone family, was initially identified with potent HIV-1 IN inhibitory activity.³⁶ Subsequently, a series of others flavones, mainly polyhydroxylated and glycosylated derivatives, were tested.³⁷ Electrotological state (E-state) chemical descriptors^{6,38} were used in QSAR studies of these flavones as inhibitors of HIV-1 IN in vitro.³⁹ E-state is an atomic level chemical descriptor. By this reason, to test the ability of the set of total and local (atom and atom-type) linear indices in QSAR studies and to compare our method with other previously reported approaches, this flavones data set has been investigated.³⁹ The chemical structures of the 15 flavones, together with negative Log(IC50) values [pIC₅₀] for cleavage step, are depicted in Table 3.³⁹

The pIC₅₀ values for both 3'-processing and 5'-strand transfer are strongly related [$R = 0.90$ and $s = 0.32$; if compound 2 is omitted (statistical outlier), a R of 0.981 and a s of 0.149 are obtained]. This is a logical result due to cleavage, and integration is a very similar reaction at the

Table 3. Structures and HIV-1 IN Inhibitory Activity of Flavones^a

flavones	-Log(IC ₅₀) cleavage	ring substituents								
		3	5	6	7	8	2'	3'	4'	5'
1 quercetagenin	6.10		OH	OH	OH	OH			OH	OH
2 baicalein	5.92		OH	OH	OH					
3 robinetin	5.23	OH			OH			OH	OH	OH
4 myricetin	5.12	OH	OH		OH			OH	OH	OH
5 quercetin	4.63	OH	OH		OH			OH	OH	
6 fisetin	4.55	OH			OH			OH	OH	
7 luteolin	4.48		OH		OH			OH	OH	
8 myricitrin	4.40	RH	OH		OH			OH	OH	OH
9 quercetrin	4.22	RH	OH		OH			OH	OH	
10 rhamnetin	4.21	OH	OH		OMe			OH	OH	
11 avicularin	4.18	AR	OH		OH			OH	OH	
12 gossypin	4.16	OH	OH		OH	GL		OH	OH	
13 morin	4.12	OH	OH		OH		OH		OH	
14 6-methoxyluteolin	4.03		OH	OMe	OH			OH	OH	
15 kaempferol	4.01	OH	OH		OH				OH	

^a GL= glucose; RH = rhamnose; and AR = arabinose.

chemical level. In this sense, stereochemical analysis of the reaction pathway has demonstrated that both 3'-processing and DNA strand transfer occur by a one-step transesterification mechanism.³⁵ For this reason, I only have taken into consideration the cleavage data set in order to validate the present approach and make a comparison with respect to the previously reported one.

3.2. Computational Methods. 3.2.1. TOMOCOMD-CARDD Approach. *TOMOCOMD* is an interactive program for molecular design and bioinformatics research.⁴⁰ It consists of four subprograms; they each allow drawing the structures (drawing mode) and calculating 2D and 3D molecular descriptors (calculation mode). The modules are named CARDD (Computed-Aided 'Rational' Drug Design), CAMPS (Computed-Aided Modeling in Protein Science), CANAR (Computed-Aided Nucleic Acid Research), and CABPD (Computed-Aided Bio-Polymers Docking). In this paper, I outline salient features concerned with only one of these subprograms: CARDD. This subprogram 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. a practicing pharmacist and organic chemist, teacher, university student, and so on).

The calculation of total and local linear indices for any organic molecule (or any drug-like compounds) was implemented in the *TOMOCOMD-CARDD* software.⁴⁰ The main steps for the application of this method in QSAR/QSPR can be briefly resumed as follows:

1. Draw the molecular pseudographs for each molecule of the data set, using the software drawing mode. This procedure is performed by a selection of the active atom symbol belonging to different groups of the periodic table.
2. Use appropriated weights in order to differentiate the molecular atoms. In this work, I used as the atomic property the Mulliken electronegativity²⁶ for each kind of atom.
3. Compute the total and local linear indices of the molecular pseudograph's atom adjacency matrix. They can

be carried out in the software calculation mode, where you can select the atomic properties and the family descriptor previously to calculate the molecular indices. This software generates a table in which the rows correspond to the compounds and columns correspond to the total and local linear indices or other family molecular descriptors implemented in this program.

4. Find a QSPR/QSAR equation by using mathematical techniques, such as multilinear regression analysis (MRA), Neural Networks (NN), Linear Discrimination Analysis (LDA), and so on. That is to say, I can find a quantitative relation between a property *P* and the linear indices having, for instance, the following appearance

$$P = a_0 f_0(x) + a_1 f_1(x) + a_2 f_2(x) + \dots + a_k f_k(x) + c \quad (11)$$

where *P* is the measurement of the property, *f_k(x)* is the *k*th total linear indices, and the *a_k*'s are the coefficients obtained by the linear regression analysis.

5. Test the robustness and predictive power of the QSPR/QSAR equation by using internal and external cross-validation techniques.

6. Develop a structural interpretation of the obtained QSAR/QSPR model using total and local (atom and atom-type) linear indices as molecular descriptors.

The *TOMOCOMD-CARDD* descriptors calculated in this study were the following: **chemicals data set**—(1) *k*th total (whole-molecule) linear indices [*f_k(x)*]; (2) *k*th local (atom-type) linear indices calculated for heteroatoms (E) in the molecules [*f_{kL}(x_E)*, where E = O, N, and S]; and (3) *k*th local (atom-type) linear indices calculated for H-bonding to heteroatoms (E) [*f_{kL}(x_{H-E})*, where E = O, N, and S] and **flavones data set**—(1) *k*th local (atom) linear indices calculated for some of the 17 skeletal atoms common to all molecules in the set as multivariate descriptors [*f_k(x_i)*] [Taken into consideration the relative importance of different atoms obtained by Buolamwini et al.³⁹ using E-state indices, the *k*th linear values at C₆, C_{5'}, C_{3'}, and O₄ were computed.]; (2)

Table 4. Description of Some of the 0D–3D DRAGON's Molecular Descriptors³¹ Used in the Preset Study

symbol	definition	class
MW	molecular weight	constitutional
Se	sum of atomic Sanderson electronegativities (scaled on carbon atom)	constitutional
Ss	sum of Kier-Hall electrotopological states	constitutional
Ui	unsaturation index	empirical
Hy	hydrophilic factor	empirical
ARR	aromatic ratio	empirical
MR	Ghose-Crippen molar refractivity	properties
PSA	fragment-based polar surface area	properties
MLOGP	Moriguchi octanol–water partition coeff. (logP)	properties
ZM1V	first Zagreb index by valence vertex degrees	topological
HNar	Narumi harmonic topological index	topological
TI2	second Mohar index TI2	topological
Rww	reciprocal hyper-detour index	topological
J	Balaban J index	topological
JhetZ	Balaban-type index from Z weighted distance matrix (Barysz matrix)	topological
X2v	valence connectivity index chi-2	topological
S1K	1-path Kier alpha-modified shape index	topological
PHI	Kier flexibility index	topological
PW2	path/walk 2 – Randic shape index	topological
PJI2	2D Petitjean shape index	topological
SIC2	structural information content (neighborhood symmetry of 2-order)	topological
SEigZ	Eigenvalue sum from Z weighted distance matrix (Barysz matrix)	topological
SRW05	self-returning walk count of order 05	mol. walk counts
BEHe6	highest eigenvalue n. 6 of Burden matrix/weighted by atomic Sand. elect.	BCUT
GGI1	topological charge index of order 1	Galvez charge ind.
ATS2e	Broto-Moreau autocorrelation of a topological structure – lag 2/Sand. elect.	2D autocorrelations
MATS1e	Moran autocorrelation – lag 1/weighted by atomic Sand. electronegativitie	2D autocorrelations
GATS3e	Geary autocorrelation – lag 3/weighted by atomic Sand. electronegativitie	2D autocorrelations
HOMA	Harmonic Oscillator Model of Aromaticity index	Aromat. indices
RCI	Jug RC index	Aromat. indices
AROM	aromaticity (trial)	Aromat. indices
DP10	molecular profile no. 10	Randic mol profiles
SHP2	average shape profile index of order 2	Randic mol profiles
J3D	3D-Balaban index	geometrical
TIE	E-state topological parameter	geometrical
FDI	folding degree index	geometrical
PJI3	3D Petitjean shape index	geometrical
RDF015u	Radial Distribution Function – 1.5/unweighted	RDF
RDF010e	Radial Distribution Function – 1.0/weighted by atomic Sand. Electroneg.	RDF
Mor02u	3D-MoRSE – signal 02/unweighted	3D-MoRSE
Mor20e	3D-MoRSE – signal 20/weighted by atomic Sanderson electronegativities	3D-MoRSE
E1e	first component accessibility directional WHIM index /atomic Sand. elect.	WHIM
Gu	G total symmetry index/unweighted	WHIM
HGM	geometric mean on the leverage magnitude	GETAWAY
H0e	H autocorrelation of lag 0/weighted by atomic Sand. electronegativities	GETAWAY
RCON	Randic-type R matrix connectivity	GETAWAY
R2e	R autocorrelation of lag 2/weighted by atomic Sand. electronegativities	GETAWAY

k^{th} local (atom-type) linear indices calculated for oxygen (O) atoms in the molecules [$f_{kL}(x_O)$]; and (3) k^{th} total linear indices [$f_k(x)$].

3.2.2. DRAGON Software. DRAGON is a very sophisticated program for the calculation of molecular descriptors³¹—among them, 0D (D = “dimensionality”) or constitutional, 1D (e.g., empirical descriptors and molecular properties), 2D (such as 2D-autocorrelations, topological indices, BCUT descriptors, Galvez topological charges indices, molecular walk counts), and 3D (aromaticity indices, Randic molecular profiles, charge-, geometrical-, RDF-, 3D-MoRSE-, GATEWAY-, and WHIM-descriptors) molecular descriptors. To compare the information carried by the total and local linear indices with 0D–3D QSPR/QSAR descriptors, all these structural indices were calculated by selecting all possible descriptors options from the “description selection” DRAGON-menu.³¹ The calculation was based on the set of 42 chemicals (Table 2) depicted in DRAGON's example data.³¹ However, I selected the “pair correlation checkbox” in order to exclude from the output file the descriptors containing redundant information (a correlation coefficient equal to or greater than 0.95). In addition, constant descriptors were also eliminated.

Subsequently, only the molecular indices “unweighted” or using Sanderson electronegativities such as atomic weights were saved. This procedure takes into account the fact that total and local linear indices were computed selecting only one atomic property (Mulliken electronegativity²⁶) in the TOMOCOMD-CARDD's properties menu⁴⁰ and in order to reduce the descriptor's number for the later chemometric analysis. Finally, to compare the total and local linear indices to other QSPR/QSAR descriptors, I have obtained a total of 229 (0D–3D) molecular descriptors. The symbols and definitions of some of these molecular indices are given in Table 4. The complete list of 0D–3D molecular descriptors used in this study as well as their description is given as Supporting Information. This list of molecular indices is representative of the three “dimension” of molecular descriptors reported in the literature.

3.3. Chemometric Methods. 3.3.1. Analysis of Principal Components for the Comparison of the Molecular Descriptors. One of the main objectives of this paper is the comparison of the information content of the k^{th} total and local linear indices with that of other descriptors used in QSPR/QSAR practice. The existence of linear independence

has been claimed by Randić⁴¹ as one of the desirable attributes for novel topological indices.

To conduct this analysis, I will carry out a factor analysis by using the principal components method. The theoretical aspects of this statistical technique have been extensively exposed in the literature including many chemical applications.^{42–50} The main applications of factor analytic techniques are (1) to *reduce* the number of variables and (2) to *detect structure* in the relationships between variables, that is to *classify* variables.^{45,51} In this approach, factor loadings (or “new” variables) are obtained from original (molecular descriptors) variables. Thus, these factors capture most of the “essence” of these molecular descriptors because they are a linear combination of the original items. Because each consecutive factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are independent of each other. Put another way, consecutive factors are uncorrelated or orthogonal to each other. In this sense, the first factor obtained is generally more highly correlated with the variables than the others factors. This is to be expected because, as previously described, these factors are extracted successively and will account for less and less variance overall. Finally, some of the most important conclusions that can be drawn from a factor analysis that will be of great usefulness in the present paper are the following:^{42–51} (1) variables with a high loading in the same factor are interrelated and will be the more so the higher the loadings, and (2) no correlation exists between variables having nonzero loadings only in different factors. These are the principal ideas that permit interpreting the *factor structure* obtained using the factor analysis as a classification method.

The factor analysis was performed with the STATISTICA software,⁵¹ and “varimax normalized” was used as a rotational strategy to obtain the factor loadings from the principal component analysis. The goal of this rotational procedure is to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others. The “varimax normalized” is the method that is most commonly used as “varimax” rotation.⁵¹ This rotation strategy is aimed at maximizing the variances of the squared *normalized factor loadings* (row factor loadings divided by squared roots of the respective communalities) across variable for each factors. This strategy makes the structure of factors pattern as simple as possible, permitting a clearer interpretation of the factors without loss of orthogonality between them.^{48–51}

3.3.2. Mathematical Tools for QSAR Modeling. Data sets of the k^{th} atom (C_6 , C_5 , C_3 , and O_4), atom-type (oxygen atoms), and total linear indices were used as molecular descriptors for derived QSARs. One of the difficulties with the large number of descriptors 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, I have used a genetic algorithm (GA) input selection.^{52–57} GAs are a class of algorithms inspired by the process of natural 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. Genetic

function approximation (GFA), a combination of GA and the linear polynomials, higher-order polynomials, splines (multivariate adaptive regression splines algorithm), or other nonlinear functions, provides multiple models with high predictive ability.^{52–60}

Build QSAR⁶¹ was employed to perform QSAR modeling. The mutation probability was specified as 35%. The length of the equations was set four terms and a constant. The population size was established as 100. The GA with an initial population size of 100 rapidly converges (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).⁶¹ The quality of models was also determined by examining the LOO press statistics (q^2 , s_{cv}).⁶² 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.

4. RESULTS AND DISCUSSION

4.1. Significance and Interpretations of the k^{th} Local and Total Linear Indices. **4.1.1. Influence of Structure Change on Total and Local Linear Indices.** The influence of structure on the k^{th} linear indices may be revealed by examining several sets of calculations in which features are systematically varied.⁶ In this sense, some effects of structure on k^{th} total and local linear indices are illustrated in several ways in the following examples: (a) effect of chain length, (b) effect due to branching, (c) effect across multiple bonds, and (d) effect due to heteroatom change. The influences of these structural features on our molecular descriptors are shown in Tables 5–8, respectively.

First, note that the mathematical linear maps' matrices, M^k , are graph-theoretic electronic-structure models, like an “extended Hückel” model. The M^1 matrix considers all valence-bond electrons (σ - and π -networks) in one step, and their power ($k = 0, 1, 2, 3, \dots$) can be considered as an interacting-electron chemical-network model in the k step. This model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas.⁶³

The present approach is based on a simple model for the intramolecular movement of all valence-bond electrons. Let us consider a hypothetical situation in which a set of atoms is free in space at an arbitrary initial time (t_0). In this time, the electrons are distributed around the atom nucleus. Alternatively, these electrons can be distributed around cores in discrete intervals of time t_k . In this sense, the electron in an arbitrary atom i can move to other atoms at different discrete time periods t_k ($k = 0, 1, 2, 3, \dots$) throughout the chemical-bonding network.

For this reason, each k^{th} total and local linear indices encodes particular information of the molecular structure. For instance, an atom-type linear index of zero-order has information on the molecular size of the fragment, and it depends on the number and type of atoms that are contained

Table 5. Changes in k^{th} Total and Local Linear Indices Due to Chain Lengthening in Alkanes

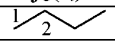
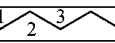
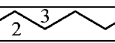
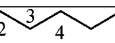
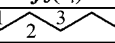
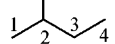
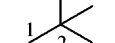
Atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$	$f_6(x_i)$	$f_7(x_i)$
								
C ₁	2.63	2.63	5.26	7.89	13.15	21.04	34.19	55.23
C ₂	2.63	5.26	7.89	13.15	21.04	34.19	55.23	89.42
Total	10.52	15.78	26.3	42.08	68.38	110.46	178.84	289.3
								
C ₁	2.63	2.63	5.26	7.89	15.78	23.67	47.34	71.01
C ₂	2.63	5.26	7.89	15.78	23.67	47.34	71.01	142.02
C ₃	2.63	5.26	10.52	15.78	31.56	47.34	94.68	142.02
Total	13.15	21.04	36.82	63.12	110.46	189.36	331.38	568.08
								
C ₁	2.63	2.63	5.26	7.89	15.78	26.3	49.97	86.79
C ₂	2.63	5.26	7.89	15.78	26.3	49.97	86.79	160.43
C ₃	2.63	5.26	10.52	18.41	34.19	60.49	110.46	197.25
Total	15.78	26.3	47.34	84.16	152.54	273.52	494.44	888.94
								
C ₁	2.63	2.63	5.26	7.89	15.78	26.3	52.6	89.42
C ₂	2.63	5.26	7.89	15.78	26.3	52.6	89.42	178.84
C ₃	2.63	5.26	10.52	18.41	36.82	63.12	126.24	215.66
C ₄	2.63	5.26	10.52	21.04	36.82	73.64	126.24	252.48
Total	18.41	31.56	57.86	105.2	194.62	357.68	662.76	1220.32

Table 6. Changes in k^{th} Total and Local Linear Indices Due to Branching in the Pentanes' Skeleton

Atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$	$f_6(x_i)$	$f_7(x_i)$
								
C ₁	2.63	2.63	5.26	7.89	15.78	23.67	47.34	71.01
C ₂	2.63	5.26	7.89	15.78	23.67	47.34	71.01	142.02
C ₃	2.63	5.26	10.52	15.78	31.56	47.34	94.68	142.02
Total	13.15	21.04	36.82	63.12	110.46	189.36	331.38	568.08
								
C ₁	2.63	2.63	7.89	10.52	26.3	36.82	89.42	126.24
C ₂	2.63	7.89	10.52	26.3	36.82	89.42	126.24	305.08
C ₃	2.63	5.26	10.52	15.78	36.82	52.6	126.24	178.84
C ₄	2.63	2.63	5.26	10.52	15.78	36.82	52.6	126.24
Total	13.15	21.04	42.08	73.64	142.02	252.48	483.92	862.64
								
C ₁	2.63	2.63	10.52	10.52	42.08	42.08	168.32	168.32
C ₂	2.63	10.52	10.52	42.08	42.08	168.32	168.32	673.28
Total	13.15	21.04	52.6	84.16	210.4	336.64	841.6	1346.56

in the fragment under study. In this connection, all the carbon-atom's linear indices [$f_0(x_i)$] have the same value (see Tables 5–8): 2.63 in Mulliken electronegativity-scale.²⁶

In a similar way, total linear indices of this same order [$f_0(x)$] encode information about size and heteroatom content, which can be easily observed in Tables 5 and 8, respectively. In this sense, with lengthening of the chain (from butane to heptane) the $f_0(x)$ value progressively increases. Therefore,

for a homologous series this descriptor increases 2.63 for the addition of each methylene group. In addition, the value of this index for pentane, for butan-1-ol, and for 1-fluorobutane is increased in this same order ($f_0(x)$ of 13.15, 13.69, and 14.43, respectively), in correspondence to the value of electronegativity of the following atoms: C, O, F.

Finally, total (and local) linear indices of zero order can be classified according to their “dimensionality” as one-

Table 7. Influences of Unsaturation on the k^{th} Total and Local Linear Indices in Hydrocarbons






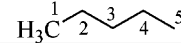
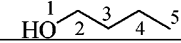
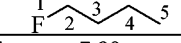
Atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$	$f_6(x_i)$	$f_7(x_i)$
								
C ₁	2.63	2.63	5.26	7.89	13.15	21.04	34.19	55.23
C ₂	2.63	5.26	7.89	13.15	21.04	34.19	55.23	89.42
Total	10.52	15.78	26.3	42.08	68.38	110.46	178.84	289.3
								
C ₁	2.63	5.26	15.78	31.56	84.16	168.32	441.84	883.68
C ₂	2.63	7.89	15.78	42.08	84.16	220.92	441.84	1157.2
C ₃	2.63	5.26	10.52	21.04	52.6	105.2	273.52	547.04
C ₄	2.63	2.63	5.26	10.52	21.04	52.6	105.2	273.52
Total	10.52	21.04	47.34	105.2	241.96	547.04	1262.4	2861.44
								
C ₁	2.63	7.89	31.56	86.79	323.49	883.68	3274.35	8939.37
C ₂	2.63	10.52	28.93	107.83	294.56	1091.45	2979.79	11035.48
C ₃	2.63	5.26	13.15	34.19	120.98	328.75	1212.43	3308.54
C ₄	2.63	2.63	5.26	13.15	34.19	120.98	328.75	1212.43
Total	10.52	26.3	78.9	241.96	773.22	2424.86	7795.32	24495.82
								
C ₁	2.63	2.63	7.89	18.41	44.71	107.83	260.37	628.57
C ₂	2.63	7.89	18.41	44.71	107.83	260.37	628.57	1517.51
Total	10.52	21.04	52.6	126.24	305.08	736.4	1777.88	4292.16
								
C ₁	2.63	2.63	10.52	34.19	113.09	373.46	1233.47	4073.87
C ₂	2.63	10.52	34.19	113.09	373.46	1233.47	4073.87	13455.08
Total	10.52	26.3	89.42	294.56	973.1	3213.86	10614.68	35057.9

Table 8. Influence of Heteroatoms on the k^{th} Total and Local Linear Indices Values

Atom (i)	$f_0(x_i)$	$f_1(x_i)$	$f_2(x_i)$	$f_3(x_i)$	$f_4(x_i)$	$f_5(x_i)$	$f_6(x_i)$	$f_7(x_i)$
								
C ₁ (CH ₃)	2.63	2.63	5.26	7.89	15.78	23.67	47.34	71.01
C ₂	2.63	5.26	7.89	15.78	23.67	47.34	71.01	142.02
C ₃	2.63	5.26	10.52	15.78	31.56	47.34	94.68	142.02
C ₄	2.63	5.26	7.89	15.78	23.67	47.34	71.01	142.02
C ₅	2.63	2.63	5.26	7.89	15.78	23.67	47.34	71.01
Total	13.15	21.04	36.82	63.12	110.46	189.36	331.38	568.08
								
O ₁ (OH)	3.17	2.63	5.8	7.89	16.86	23.67	50.04	71.01
C ₂	2.63	5.8	7.89	16.86	23.67	50.04	71.01	149.58
C ₃	2.63	5.26	11.06	15.78	33.18	47.34	99.54	142.02
C ₄	2.63	5.26	7.89	16.32	23.67	49.5	71.01	149.04
C ₅	2.63	2.63	5.26	7.89	16.32	23.67	49.5	71.01
Total	13.69	21.58	37.9	64.74	113.7	194.22	341.1	582.66
								
F ₁	3.91	2.63	6.54	7.89	18.34	23.67	53.74	71.01
C ₂	2.63	6.54	7.89	18.34	23.67	53.74	71.01	159.94
C ₃	2.63	5.26	11.8	15.78	35.4	47.34	106.2	142.02
C ₄	2.63	5.26	7.89	17.06	23.67	52.46	71.01	158.66
C ₅	2.63	2.63	5.26	7.89	17.06	23.67	52.46	71.01
Total	14.43	22.32	39.38	66.96	118.14	200.88	354.42	602.64

dimensional descriptors (1D). Subsequently, $f_0(x)$ includes “bulk” properties and physicochemical properties (such as

hydrophobicity,⁶⁴ molecular polar surface area,⁶⁵ molar refractivity,⁶⁶ molecular polarizability,⁶⁷ and atomic charge

summatory⁶⁸), if some atomic physicochemical parameters (such as atomic Log P,⁶⁴ surface contributions of polar atoms,⁶⁵ atomic molar refractivity,⁶⁶ atomic hybrid polarizabilities,⁶⁷ and Gasteiger–Marsilli atomic charge,⁶⁸ respectively) are considered as atom-property (atom-label) for building the n -dimensional molecular vector, X . Specifically, if the atomic mass is used to characterize each kind of atom within the molecule, $f_0(x)$ is the molecular weight.

However, this index does not take into consideration the effect of branching and unsaturated atoms. Some examples of these are shown in Tables 6 and 7, respectively. This is a logical result, because in this initial time (t_0) the set of atoms is free in space, and the electrons are distributed around atom nucleus.

On the other hand, total linear indices of first order $f_1(x)$ is capable of discriminating between saturated and unsaturated (double and triple bonds) isomers (see Table 7), even though it cannot be considered as a unique descriptor, and so some isomers have identical values. For instance, it does not discriminate the 1-butene (21.04) and 1-butyne (26.3) from their positional isomers 2-butene (21.04) and 3-butyne (26.3), respectively. Further, $f_1(x)$ is unable to differentiate between ramified isomers (see Table 6) and their value systematically varied due to chain lengthening in linear alkanes (see Table 5).

In another way, local (atom) linear indices of first order $f_1(x_i)$ are very influenced by the effect of branching in alkanes. Several examples illustrate this effect in Table 6. Methyl groups at a branch point have $f_1(x_i)$ values of 2.63. A significant increase in the $f_1(x_i)$ value is found on the carbon at the branch point (5.26, 7.89, and 10.52 for the carbon-atom with two, three, and four adjacent carbon atoms, respectively). The unsaturated atoms exhibit the same behavior (see Table 7). This calculated effect mirrors the inductive effect and also the reduction in topological freedom or a raise in steric crowding. Additionally, the introduction of a heteroatom into an alkane molecule produces an effect on the $f_1(x_i)$ of the adjacent carbon-atoms, which is proportionate with the Mulliken electronegativity value of the heteroatom. For example, $f_1(x_i)$ of the adjacent carbon-atom to the oxygen and fluorine atom are 5.8 and 6.54, respectively (see Table 8).

Conversely, the total and local linear indices of second order can be considered as branching and multiple bonds molecular descriptors. For example, with branching of the chain, the atom linear indices of second-order value of the methyl group connected with the n -, i -, and $tert$ -butyl group, steadily increase: $f_2(x_{C1})$ of 5.26, 7.89, and 10.52, respectively (see Table 6). This table depicts that total linear index of second-order $f_2(x)$ is able to discriminate among the pentane's branching isomers. In this case, $f_2(x)$ values are also increased due to branching in the skeleton: $f_2(x)$ of 36.82, 42.08, and 52.6 for pentane, for 2-methylbutane, and for 2,2-dimethylpropane, respectively (see Table 6). The unsaturated atoms exhibit the same behavior (see Table 7). For instance, these atoms (sp^2 and sp) have elevated $f_2(x_i)$ values, which depend of the nature and topology of the atoms involved. Furthermore, terminal unsaturation results in lower $f_2(x_i)$ values for these atoms relative to the unsaturated atoms in midchain. This behavior also mirrors the inductive effect and the reduction in topological freedom or a raise in steric crowding. That is, this local term represents the accessibility

Table 9. Results of the Factor Analysis by Using the Principal Component Method for 229 0D–3D Molecular Descriptors as Well as the 33 Total and Local (Atom-type) Linear Indices for 42 Very Heterogeneous Chemicals

factors	eigenvalue	% total variance	cumulative eigenvalue	% cumulative variance
F₁	87.23	33.29	87.23	33.29
F₂	38.25	14.60	125.47	47.89
F₃	27.29	10.42	152.77	58.31
F₄	15.53	5.93	168.30	64.24
F₅	14.50	5.53	182.80	69.77
F₆	11.26	4.30	194.06	74.07
F₇	9.64	3.68	203.69	77.75
F₈	8.31	3.17	212.01	80.92
F₉	7.69	2.93	219.70	83.85
F₁₀	5.34	2.04	225.03	85.89

from outside to atoms in a path of length 2 in the molecule. Thus, the second-order total and local linear indices are interpreted as a component of the “molecular accessibility” coming from contributions of paths of length 2 in the molecule.

In a similar way, we can interpret the effect of structure on “higher” order of total and local linear indices in a molecule, starting from contributions of “subgraphs” of different lengths 3, 4, 5, etc. The physical meaning of this approach is based in that the valence-shell electrons in an arbitrary atom i can move to other atoms at different discrete time periods t_k ($k = 0, 1, 2, 3, \dots$) throughout the chemical-bonding network.

In any case, whether a complete series of indices is considered, a specific characterization of the chemical structure is obtained (whole structure or fragment), which is not repeated in any other molecule. The generalization of the descriptors to “superior analogues” is necessary for the evaluation of situations where only one descriptor is unable to bring a good structural characterization.⁴¹ In this sense, the k^{th} atom, atom-type, and total linear indices can be used as variables in QSAR/QSPR and “rational” drug design studies.

4.1.2. Comparison of the Linear Indices with Other Molecular Descriptors. To check the existence or not of linear independence between the total and local linear indices and the other 0D–3D molecular descriptors calculated in this work, I carried out a factor analysis. The comparison was based on a set of 42 chemicals (see Table 2). Even though the number of chemicals is limited, the generality of the comparison was assured by the presence of diverse chemical functionalities and substructures.

The results of the factor analysis are summarized in Table 9, and the 10 principal factors explain approximately 86% of the variance. The first factor explains 33.29% of the variance in the molecular indices studied. The addition of the second factor increases to 47.89% of the variance explained, and the addition of the third factor allows 58.31% of the index variance to be accounted for. The other factors explain (% cumulative variance) the 5.93 (64.24), 5.53 (69.77), 4.30 (74.07), 3.68 (77.75), 3.17 (80.92), 2.93 (83.85), and 2.04% (85.89) of the variance in the molecular descriptors studied (see Table 9). Factor loadings from the principal component analysis, after a Varimax normalized rotation of the factors, are shown in Table 10. This table depicts a single portion of the obtained results, and the complete list of

Table 10. Factor Loadings (Varimax Normalized Rotation) for Some of the 229 0D–3D Molecular Descriptors as Well as Some of the 33 Total and Local (Atom-type) Linear Indices for 42 Very Heterogeneous Chemicals

index	F ₁	F ₂	F ₃	F ₄	F ₅	F ₆	F ₇	F ₈	F ₉	F ₁₀
$f_0(x)$	0.89	0.01	0.03	0.06	0.31	0.05	0.25	0.17	−0.02	0.07
$f_1(x)$	0.90	−0.15	0.03	0.00	0.32	0.05	0.08	0.19	−0.05	0.06
$f_7(x)$	0.95	−0.23	0.01	0.02	0.14	−0.01	−0.03	0.14	0.01	−0.04
$f_{10}(x)$	0.96	−0.19	0.00	0.03	0.08	−0.07	−0.04	0.10	0.03	−0.08
$f_{0L}(x_E)$	0.06	−0.28	0.32	−0.01	0.03	0.01	0.80	0.40	0.00	−0.02
$f_{1L}(x_E)$	0.13	−0.20	0.28	−0.03	0.06	−0.04	0.41	0.77	0.00	−0.02
$f_{5L}(x_E)$	0.25	−0.16	0.21	−0.01	0.09	−0.05	0.17	0.87	−0.02	0.03
$f_{10L}(x_E)$	0.30	−0.12	0.20	0.01	0.05	−0.08	0.10	0.87	−0.01	0.01
$f_{0L}(x_{H-E})$	−0.06	−0.08	0.95	0.03	−0.04	0.03	0.20	−0.07	0.04	−0.03
$f_{1L}(x_{H-E})$	−0.07	−0.09	0.93	0.04	−0.05	0.03	0.27	−0.10	0.04	−0.04
$f_{2L}(x_{H-E})$	−0.04	−0.06	0.96	0.01	−0.03	0.03	0.12	−0.02	0.03	−0.03
$f_{9L}(x_{H-E})$	0.10	−0.10	0.80	−0.02	0.10	0.09	−0.10	0.37	−0.04	0.08
MW	0.73	−0.08	−0.04	0.05	0.42	0.10	0.23	0.13	−0.02	0.19
Se	0.71	0.61	0.00	0.22	0.20	0.02	0.04	0.00	0.07	0.11
nSK	0.90	0.03	0.05	0.06	0.32	0.05	0.16	0.19	−0.02	0.08
nBM	0.89	−0.29	0.03	−0.03	0.26	0.15	−0.02	0.11	−0.02	0.05
Ui	0.71	−0.42	0.05	−0.01	0.41	0.23	−0.05	0.23	−0.03	0.13
Hy	−0.20	−0.16	0.81	−0.04	−0.20	−0.04	0.22	−0.16	0.04	−0.06
MR	0.88	0.09	−0.03	0.12	0.37	0.10	−0.05	0.06	0.02	0.12
PSA	0.05	−0.17	0.72	0.02	0.02	−0.06	0.37	0.34	0.05	0.03
MLOGP	0.40	0.41	−0.58	0.05	0.17	0.17	−0.20	−0.24	−0.10	0.16
ISIZ	0.71	0.63	−0.04	0.22	0.16	0.01	−0.05	−0.04	0.07	0.12
ZM2V	0.84	−0.22	0.10	−0.03	0.18	0.00	0.27	0.34	−0.03	0.01
HNar	0.48	0.05	0.02	−0.15	0.73	−0.21	−0.03	0.18	−0.34	−0.07
Ram	0.85	0.02	0.05	−0.10	0.01	0.03	0.30	0.19	0.32	−0.01
TI1	0.89	−0.15	−0.04	−0.21	−0.02	−0.20	−0.15	0.03	−0.08	−0.14
TI2	0.12	0.26	0.09	0.89	0.03	0.02	0.24	−0.08	0.12	−0.04
D/D	0.79	0.11	0.07	0.26	0.17	0.17	0.36	0.16	0.08	0.16
JhetZ	0.15	−0.22	0.04	0.05	0.75	0.05	0.39	0.15	0.24	0.05
Jhetv	0.24	0.05	−0.12	0.12	0.79	0.24	−0.11	0.03	0.23	0.04
X3v	0.81	0.13	−0.13	0.02	0.30	0.02	−0.04	0.10	−0.32	0.12
X5v	0.90	0.08	−0.09	−0.05	0.12	0.09	−0.01	0.01	−0.16	0.02
X2sol	0.85	0.02	0.02	−0.10	0.39	0.07	0.01	0.19	0.12	0.18
S2K	0.22	0.32	0.02	0.84	0.17	0.04	0.08	−0.12	−0.08	0.12
PHI	−0.08	0.38	0.01	0.85	0.01	−0.02	0.09	−0.20	0.01	0.10
PW2	0.36	0.22	0.11	−0.12	0.71	−0.04	0.32	0.14	0.21	0.01
PW4	0.70	−0.08	0.02	−0.06	0.42	0.03	−0.06	0.29	−0.25	0.33
PW5	0.82	−0.06	0.02	−0.03	0.27	0.27	0.01	0.20	−0.17	0.14
PJ12	0.11	0.31	0.12	0.33	−0.09	0.04	0.24	−0.16	0.70	0.00
AECC	0.79	0.13	0.05	0.36	0.39	0.14	0.17	0.10	−0.09	0.08
VAR	0.88	0.03	0.05	0.24	0.04	0.06	0.28	0.13	0.11	0.10
CIC0	0.37	0.71	−0.11	0.24	0.40	−0.04	−0.22	−0.14	0.10	−0.13
TIC4	0.72	0.18	0.25	0.38	0.22	−0.01	0.16	0.16	0.05	0.27
LP1	0.50	0.17	0.08	−0.07	0.75	−0.11	0.24	0.19	0.00	−0.06
STN	0.85	−0.13	0.00	−0.24	0.28	−0.09	−0.09	0.21	−0.22	0.02
Eig1p	0.81	0.05	0.08	0.13	0.09	−0.05	0.44	0.28	−0.02	0.05
SEigZ	0.03	−0.42	0.09	−0.04	0.24	0.05	0.71	0.23	−0.01	0.15
SEigp	−0.04	0.15	−0.24	−0.01	0.11	0.01	− 0.84	−0.25	0.00	0.13
VRZ1	0.96	−0.02	0.03	0.01	0.18	−0.02	0.10	0.16	−0.02	0.03
MPC07	0.91	−0.03	−0.05	0.00	−0.11	−0.29	−0.05	0.00	0.05	−0.15
piPC05	0.97	−0.11	−0.01	−0.03	0.04	−0.02	−0.03	−0.03	0.01	−0.10
D/Dr06	0.93	−0.09	0.01	−0.05	0.17	0.23	0.04	0.03	−0.09	0.02
SRW07	0.14	−0.13	0.00	−0.25	0.09	− 0.76	−0.15	0.41	−0.15	0.01
SRW09	0.29	−0.12	0.01	−0.16	0.04	− 0.72	−0.13	0.48	−0.08	0.03
BEHe6	0.78	0.34	−0.02	0.20	0.27	0.14	0.10	−0.06	−0.10	−0.02
BELe5	0.90	0.02	−0.12	0.01	−0.08	−0.25	−0.09	−0.19	0.03	−0.11
GGI3	0.83	−0.02	0.03	0.15	0.05	−0.11	0.10	0.30	0.00	0.19
GGI4	0.93	−0.05	0.00	0.01	0.03	0.10	0.04	−0.14	−0.01	−0.08
ATS2e	0.08	−0.27	0.19	−0.04	0.09	−0.11	0.81	0.21	−0.03	−0.09
ATS7e	0.77	0.10	0.08	0.24	−0.08	−0.12	0.04	−0.09	−0.03	0.12
MATS1e	0.28	0.15	−0.27	−0.03	0.76	−0.09	0.24	0.01	−0.04	−0.10
MATS3e	0.12	−0.10	0.14	0.09	−0.16	−0.14	0.04	0.09	− 0.80	0.02
GATS3e	0.00	0.15	0.00	−0.04	0.33	0.04	0.09	−0.09	0.76	−0.07
RCI	0.65	−0.30	0.11	−0.03	0.36	0.35	0.02	0.23	−0.07	0.18
AROM	0.58	−0.33	0.07	−0.07	0.41	0.41	−0.05	−0.01	−0.07	0.20
DP10	0.89	0.12	−0.08	0.28	−0.07	−0.19	0.04	−0.07	−0.01	0.09
AGDD	0.82	0.46	−0.04	0.27	0.14	−0.02	−0.05	−0.04	0.06	0.10
MAXDN	0.06	−0.14	0.24	0.01	−0.04	0.15	0.90	0.11	0.14	−0.04
TIE	0.79	0.18	0.13	0.10	0.18	0.10	0.14	0.18	0.03	0.14
G2	0.86	−0.09	0.04	0.00	0.35	0.03	0.19	0.24	−0.05	0.11
SPAM	0.01	− 0.80	0.14	0.39	0.16	−0.06	0.19	0.13	−0.06	0.11
ASP	0.16	−0.18	−0.04	0.80	0.23	0.06	−0.02	−0.10	−0.11	−0.02
FDI	0.42	−0.06	−0.02	0.10	0.81	0.00	0.13	0.21	−0.11	−0.03
RDF020u	0.21	−0.06	0.73	0.24	−0.02	−0.23	0.21	0.03	0.04	−0.10
RDF025u	0.48	0.72	−0.11	0.15	0.13	0.13	−0.03	−0.15	−0.10	0.23
RDF050u	0.86	0.09	−0.07	0.23	0.12	0.21	−0.02	−0.11	−0.09	0.03
RDF055u	0.72	0.13	−0.05	0.28	−0.09	−0.42	−0.07	0.19	0.03	0.12

Table 10 (Continued)

index	F ₁	F ₂	F ₃	F ₄	F ₅	F ₆	F ₇	F ₈	F ₉	F ₁₀
Mor02u	0.17	0.95	-0.02	0.07	0.05	-0.13	-0.04	-0.14	0.04	0.03
Mor07u	0.78	-0.20	0.07	-0.13	0.39	0.04	0.05	0.13	-0.31	0.05
Mor11u	0.36	-0.70	0.26	-0.10	0.28	0.04	0.14	0.16	0.06	0.04
Mor17u	0.11	-0.82	0.30	0.10	0.01	0.07	0.17	0.27	-0.08	-0.01
Mor20e	0.77	-0.39	0.12	-0.06	0.13	0.23	0.00	-0.08	0.07	0.00
L1u	0.78	0.09	0.00	0.58	0.09	-0.10	-0.05	0.02	0.06	0.00
P1u	0.28	-0.15	0.13	0.87	0.08	-0.16	0.04	0.04	0.09	-0.16
Ku	0.28	-0.23	0.09	0.82	0.34	-0.09	0.01	0.02	-0.04	-0.08
ITH	0.75	-0.06	0.24	0.18	0.18	0.01	0.27	0.36	0.04	0.21
HIC	0.54	0.72	0.01	0.16	0.36	0.03	0.01	-0.07	0.08	0.09
HGM	-0.39	-0.26	-0.06	-0.27	-0.81	0.04	-0.07	-0.01	-0.07	0.11
HTu	0.48	0.81	-0.04	0.15	0.17	0.05	-0.07	0.00	0.14	0.10
HATS5u	0.07	0.09	-0.08	0.75	0.08	0.28	-0.25	0.18	0.00	-0.22
RCON	0.17	0.93	0.07	0.09	0.19	-0.08	-0.02	0.08	0.01	-0.12
R3u	-0.23	0.87	0.06	-0.17	0.08	-0.15	0.17	-0.15	-0.13	-0.12
R5u	0.37	0.33	-0.04	0.73	0.05	0.19	-0.17	0.12	-0.09	0.02
R2e+	-0.39	-0.30	-0.07	-0.04	-0.71	-0.05	0.26	0.02	-0.03	-0.13

molecular descriptors' loadings is given as Supporting Information.

All k^{th} total linear indices [$f_k(x)$] are strongly loaded in factor 1 (F₁). Most of the constitutional (e.g., MW, Se, and nSK), empirical (Ui, and ARR), molecular properties (MR), topological (first-, second-, and some third-generation, such as ISIZ, ZM2V, Ram, TI1, D/D, X0-5v, X2sol, XMOD, PW4, PW5, AECC, VAR, TIC4, STN, Eig1p, VRZ1, MPC07, piPC05, piID, D/Dr06, D/Dr10, BEHe6, BELe4, and GGI4), 2D-autocorrelations, (ATS7e), Randic molecular profiles (DP01 and DP10), geometrical (i.e., AGDD, TIE, and G2), RDF (RDF045-55u), 3D-MoRSE (Mor07u and Mor20e), WHIM (L1u), and GETAWAY (ITH) are also robustly loaded (loadings > 0.70) in this factor. Thus, total linear indices and the others "F₁-indices" produce much redundancy and overlapping among them and their relations are very complex. The third factor (F₃) is almost exclusively an atom-type (H-atoms bonding to heteroatoms) linear indices [$f_{KL}(x_{H-E})$], a hydrophilic factor (Hy), a fragment-based polar surface area (PSA), and a Radial Distribution Function 2.0/unweighted (RDF020u) dimension. This result showed that $f_{KL}(x_{H-E})$, Hy, PSA, and RDF020u have a strongly parallel relation. In addition, these molecular descriptors and Moriguchi octanol-water partition coefficient (MLOGP) are also connected by the factor 3, as can be seen in Table 10. In this case, there is a weakly opposite relation between these molecular descriptors. This is a logical result, because these indices encode the hydrogen-bonding capabilities in opposite ways. The eighth factor (F₈) appears to be most significant for the atom-type linear indices of heteroatoms [$f_{KL}(x_E)$]. As previously stated, the indices with a high loading in the same factor are interrelated, while no correlation exists between indices having nonzero loadings only in different factors.⁴²⁻⁵⁰ Consequently, it is clear that the atom-type linear indices are orthogonal to most of the 0D-3D molecular descriptors. Thus, I can say that the atom-type (heteroatoms) linear indices contain structural information not contained in any other 0D-3D molecular descriptors.

4.2. QSAR Application. In the QSAR study, the structural features of each inhibitor were described numerically using descriptors in several categories: atom (C₆, C₅, C₃, and O₄), atom-type (oxygen atoms), and total linear indices (see TOMOCOMD-CARDD Approach). In this context, the best

models obtained together with its statistical parameters are given below:

$$\begin{aligned}
 -\text{Log(IC}_{50}) = & 28.36527(\pm 3.627859) - \\
 & 2.41666(\pm 0.346727)f_2(x_{C3'}) + 3.98 \times \\
 & 10^{-7}(\pm 6.05 \times 10^{-8})f_{15}(x_{C3'}) + \\
 & 0.000233(\pm 3.17 \times 10^{-5})f_{11}(x_{C5'}) - \\
 & 2.2(\pm 3.05 \times 10^{-6})f_{13}(x_{C5'}) \quad (12)
 \end{aligned}$$

$$N = 15, R = 0.965, q^2 = 0.679, s = 0.207, F(4,10) = 34.151$$

$$\begin{aligned}
 -\text{Log(IC}_{50}) = & 11.38486(\pm 2.04321) - 0.90169(\pm \\
 & 0.154808)f_{L2}(x_O) + 0.06646(\pm 0.01178)f_{L6}(x_O) - \\
 & 0.00695(\pm 0.001253)f_{L8}(x_O) + 3.92 \times 10^{-6}(\pm \\
 & 7.37 \times 10^{-7})f_{L13}(x_O) \quad (13)
 \end{aligned}$$

$$N = 15, R = 0.902, q^2 = 0.543, s = 0.342, F(4,10) = 10.890$$

$$\begin{aligned}
 -\text{Log(IC}_{50}) = & -39.90915(\pm 7.666448) - \\
 & 1.17926(\pm 0.20057)f_0(x) + 0.11653(\pm 0.021882)f_3(x) - \\
 & 0.00139(\pm 0.00055)f_7(x) + 0.00028(\pm 0.00013)f_8(x) \quad (14)
 \end{aligned}$$

$$N = 14, R = 0.927, q^2 = 0.721, s = 0.265, F(4,9) = 13.703$$

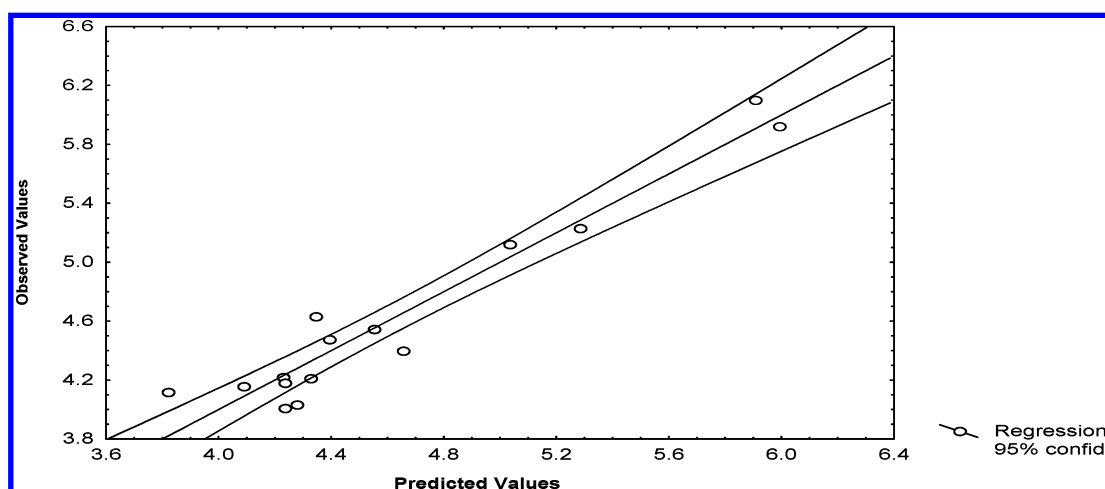
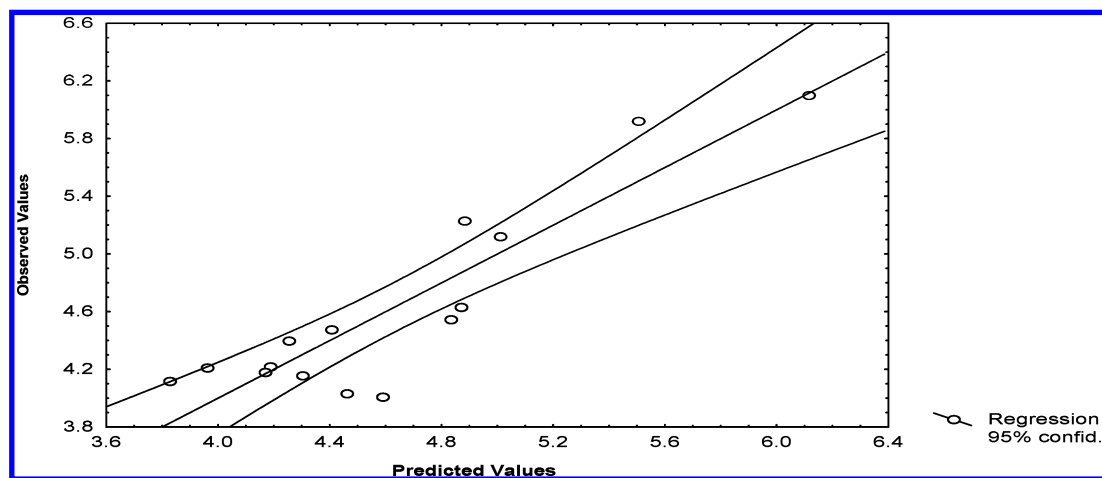
where R is the multiple regression coefficient, s is the standard deviation of the regression, q^2 is the multiple regression coefficient of the cross-validation procedure, and F is the Fisher ratio at the 95% confidence level. In Table 11 are depicted the values of experimental and calculated (by eqs 12-14) pIC₅₀ for 15 flavones and in Figures 1-3 are illustrated the linear relationships between them, respectively.

A rather similar equation for the cleavage process was reported by Buolamwini et al.³⁹ using E-states (E-state indices for the 17 skeletal atoms common to all molecules) and PLS as molecular descriptors and statistical techniques, respectively: $N = 14, R^2 = 0.975, q^2 = 0.513, s = 0.121$, and $F = 131$. This PLS model included 3 components for cleavage data.

Table 11. Observed versus Predicted pIC_{50} Values (Eqs 12, 13 and 14) for Flavones Data Set

no.	Obs ^a	Pred ^b	Res ^c	CV-Res ^d	Pred ^e	Res ^c	CV-Res ^d	Pred ^f	Res ^c	CV-Res ^d
1	6.10	5.91	0.19	0.57	6.11	-0.01	-0.04	5.80	0.30	1.44 ^g
2	5.92	5.99	-0.07	-0.82	5.50	0.42	0.82			
3	5.23	5.28	-0.05	-0.08	4.88	0.35	0.39	5.11	0.12	0.21
4	5.12	5.03	0.09	0.12	5.01	0.11	0.15	5.24	-0.12	-0.20
5	4.63	4.35	0.28	0.32	4.87	-0.24	-0.27	4.71	-0.08	-0.10
6	4.55	4.55	0.00	0.00	4.83	-0.28	-0.36	4.58	-0.03	-0.06
7	4.48	4.39	0.09	0.13	4.40	0.08	0.16	4.48	0.00	0.00
8	4.40	4.65	-0.25	-0.40	4.25	0.15	0.20	4.34	0.06	0.09
9	4.22	4.23	-0.01	-0.01	4.19	0.03	0.05	3.80	0.42	0.64
10	4.21	4.33	-0.12	-0.13	3.96	0.25	0.35	3.89	0.32	0.57
11	4.18	4.23	-0.05	-0.06	4.17	0.01	0.02	4.47	-0.29	-0.38
12	4.16	4.09	0.07	0.08	4.30	-0.14	-0.24	4.41	-0.25	-0.37
13	4.12	3.82	0.30	0.53	3.82	0.30	0.54	4.26	-0.14	-0.19
14	4.03	4.28	-0.25	-0.36	4.46	-0.43	-0.73	4.27	-0.24	-0.30
15	4.01	4.23	-0.22	-0.53	4.59	-0.58	-0.86	4.05	-0.04	-0.06

^a Experimental negative log of the molar concentration of compound that caused 50% inhibition of HIV-1 IN activity; taken from ref 39. ^b Calculated $-\text{Log}(\text{IC}_{50})$ values by eq 12. ^c Residuals: $[-\text{Log}(\text{IC}_{50})\text{Obs}] - [-\text{Log}(\text{IC}_{50})\text{Pred}]$. ^d Residuals of the LOO cross-validation procedure (deleted residual). ^e Calculated $-\text{Log}(\text{IC}_{50})$ values by eq 13. ^f Calculated $-\text{Log}(\text{IC}_{50})$ values by eq 14. ^g Statistical outlier of the LOO cross-validation procedure.

**Figure 1.** Correlation between experimental and calculated (by eq 12) pIC_{50} values for 3'-processing (cleavage) step of 15 flavones of the data set.**Figure 2.** Correlation between experimental and calculated (by eq 13) pIC_{50} values for 3'-processing (cleavage) step of 15 flavones of the data set.

In the development of the third quantitative model for the description of the 3'-processing step (eq 14), baicalein (compound 2) was detected as statistical outliers. Outliers detection was performed using the following standard statistical test: residual, standardized residuals, studentized residual, and Cooks' distance.^{51,69} Interestingly, the PLS results showed also that one compound (6-methoxyluteolin) was an outlier in the cleavage model based on the residuals.³⁹

In addition, the same molecule was also an outlier in parallel studies by Raghavan et al. using a very different QSAR method, comparative molecular field analysis (CoMFA).⁷⁰ In the flavones set, besides compound 2, there are only two other 6-position-substituted compounds (compound 1 and 14), the substituent being a hydroxyl and a methoxy group, respectively. In addition, baicalein is the compound with a minor number of substitutions in its skeleton: only 3

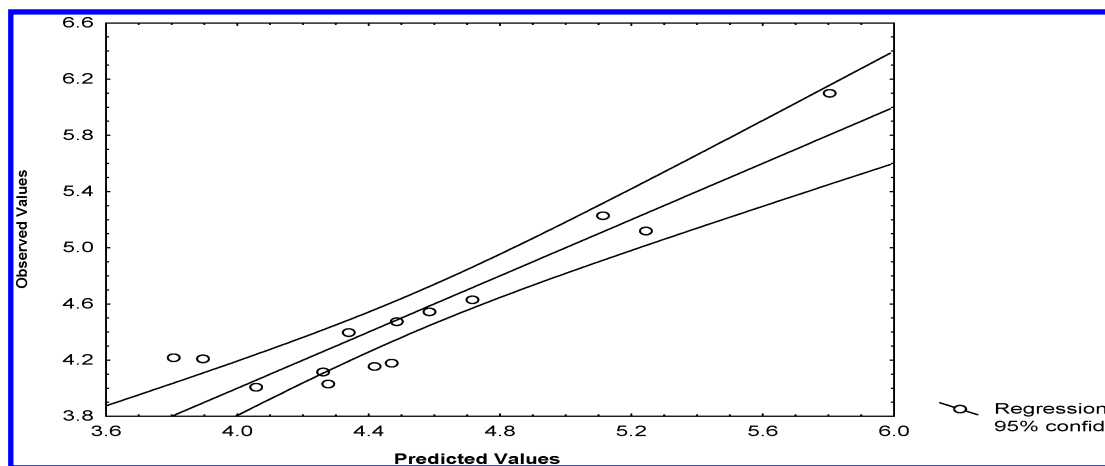


Figure 3. Correlation between experimental and calculated (by eq 14) pIC_{50} values for 3'-processing (cleavage) step of 14 flavones of the data set.

hydroxyl groups. The total linear indices appeared not to distinguish effectively between the H-atoms (H-substituent) in compound 2 and the activity-enhancing by the other hydroxyl group in compound 1.

On the other hand, validation is a crucial aspect of any QSAR/QSPR modeling.^{62,71} One of the most popular validation criteria is LOO cross-validated R^2 (LOO q^2 ; internal validation). For this reason, to assess the predictability of the models found, a LOO q^2 was carried out. This methodology systematically removed one data point at a time from the data set and then developed a number of parallel QSAR models from the reduced data with one of the compounds omitted. After developing a QSAR model, the omitted data point is used as a test set.^{62,71} That is, cross-validation simulates how well the model predicts new data. Using this approach, the model 12, 13, and 14 had a LOO q^2 of 0.679, 0.543, and 0.721 (compound 1 was detected as statistical outlier of the LOO cross-validation procedure for eq 14; see Table 11), respectively. These values of q^2 ($q^2 > 0.5$) can be considered as a proof of the high predictive ability of the models. In this sense, the equation obtained with the E-state indices³⁹ (eq 8 in ref 39) showed smaller predictive abilities ($q^2 = 0.513$ for $N = 14$) that the eq 12 ($q^2 = 0.679$ for $N = 15$) also achieved with atomic level chemical descriptors and atom linear indices.

Finally, several approaches can be used to extract a structural interpretation of an obtained model using linear indices. I used two different ways that permit an easy interpretation of the cleavage in terms of molecular structure. The first one is the "classical" way in which I do a direct analysis of the structural information presented by each molecular descriptor and how this contributes to the property under study. The second one is the way that the total contribution of different atoms in a specific molecule is expressed. In the second approach, a more compact additive scheme is obtained.^{22,47} The first approach permits estimating the relative contribution of different molecular factors to the biology activities. For example, in the eq 12, the atom linear indices at $C_{3'}$ and $C_{5'}$ were found to be more important for prediction of activity than those for any of other 2 (C_6 , and O_4) flavone skeletal atoms that are common to the molecules in the data set. Similar results were obtained by Buolamwini et al.³⁹ using E-state indices. These facts supported the important status of the atom linear indices at $C_{3'}$ and $C_{5'}$.

The second approach permits one to obtain the contribution of atoms in a specific molecule allowing the comparison among them in a more effective way. In this sense, the total contribution of the different atoms in a specific molecule can be obtain substituting the expression (eq 7) into a QSAR model (eq 11). The atoms' contribution is calculated from this procedure as shown in eq 15

$$P = b_0 + \sum_k a_k f_k(x) = b_0 + \sum_k \sum_i a_k f_k(x_i) \quad (15)$$

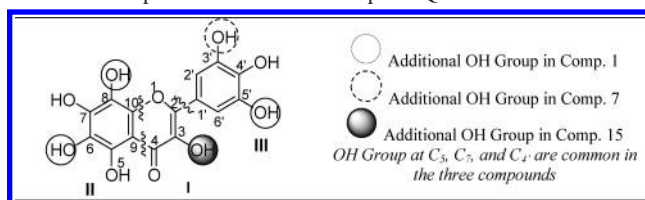
where i stands for the corresponding atom.

Considering the QSAR models obtained using total linear indices (eq 14) and the compound 7 (luteolin), a simple example is given here for calculation of these atoms contributions to cleavage. This flavones with its atom numbering and the total and local (atom) linear indices are depicted in Table 12.

Now, if the intercept values of QSAR models are divided with the number of atoms in the molecule ($n = 21$) and the atom linear indices are used as molecular descriptors in the model (eq 14), then the atom contribution for each specific atom is obtained (see Tables 12 and 13). Table 12 also includes this analysis for compounds 1 and 15. These compounds (1 and 15) were the highest (most potent in the assay) and lowest value of pIC_{50} (see Table 11). In this sense, compound 7 had a moderate activity.

Consequently, I can calculate the pIC_{50} of these compounds using two approaches. The first one is using the atom's linear indices, because it is clear that the sum of these atom contributions gives the value of the 3'-processing of the flavone molecule (see right columns in Table 12), and the second one is using the total linear indices (eq 14).

This approach permits building topological chemical representations of molecules (using a pseudograph) by combining molecular atom-fragments. In this sense, k^{th} total linear indices can be expressed as a "linear combination" of k^{th} fragment (atom) linear indices (subgraph). This way, the calculation of several molecules properties by combining distributions (atom contribution) of smaller fragments present in the molecule is carried out. This method is based on the assumption that contribution of a given molecular fragment to the complete molecular property should be quite similar in different molecules or in different locations of the same

Table 12. Basic Molecular Skeleton of Some Compounds Used To Develop the QSAR with Their Total and Local (Atom) Linear Indices

					-Log(IC ₅₀) [eq 14] ^a							-Log(IC ₅₀) [eq 14] ^a	
atom (i)	f ₀ (x _i)	f ₃ (x _i)	f ₇ (x _i)	f ₈ (x _i)	atom	fragment	atom (i)	f ₀ (x _i)	f ₃ (x _i)	f ₇ (x _i)	f ₈ (x _i)	atom	fragment
Compound 1													
O ₁	3.17	69.46	10275.39	37035.86	-1.26	I = 1.67	O ₈	3.17	35.81	4840.2	16451.6	-3.41	II = 1.99
C ₂	2.63	126.92	16896.37	53648.97	1.54		C ₉	2.63	146.34	21502.52	75494.18	3.53	
C ₃	2.63	92.05	13374.08	50176.71	1.39		C ₁₀	2.63	140.07	20139.49	68369	2.70	
C ₄	2.63	124.36	16383.97	52159.4	1.53		C _{1'}	2.63	124.69	16625.42	57261.11	2.67	III = 2.14
O ₄	3.17	57.86	8641.4	32767.94	-1.54		C _{2'}	2.63	90.5	11412.02	37789.75	0.46	
C ₅	2.63	126.38	17468.2	59397.02	2.29	II = 1.99	C _{3'}	2.63	85.78	9752.31	32196.36	0.65	
O ₅	3.17	35.27	5043.04	17468.2	-3.47		C _{4'}	2.63	103.18	11032.03	35946.91	1.95	
C ₆	2.63	117.95	15383.26	52568.41	2.29		O _{4'}	3.17	32.64	3390.69	11032.03	-3.28	
O ₆	3.17	35.81	4546.98	15383.26	-3.30		C _{5'}	2.63	105.81	11771.88	38740.4	2.01	
C ₇	2.63	118.49	15169.97	51462.72	2.34		O _{5'}	3.17	32.64	3609.19	11771.88	-3.38	
O ₇	3.17	35.81	4457.89	15169.97	-3.24		C _{6'}	2.63	99.47	12327.3	40724.6	1.06	
C ₈	2.63	120.58	16451.6	56601.26	2.25		total	64.81	2057.87	270495.2	919617.5	5.80	5.80
Compound 7													
O ₁	3.17	66.29	9531.77	34254.09	-1.54	I = 0.93	C ₉	2.63	140	19536.84	67925.84	3.24	II = 2.74
C ₂	2.63	126.92	16623.63	52187.4	1.34		C ₁₀	2.63	131.1	17630.46	58881.36	2.31	
C ₃	2.63	92.05	13062.62	49012.91	1.34		C _{1'}	2.63	124.69	16530.39	56862.72	2.52	
C ₄	2.63	124.36	15765.65	49248.82	1.41		C _{2'}	2.63	99.47	12311.99	40611.09	0.88	III = 0.81
O ₄	3.17	57.86	8324.68	31531.3	-1.61		C _{3'}	2.63	105.81	11768.71	38721.92	1.85	
C ₅	2.63	117.41	14992.89	50119.66	1.92	II = 2.74	O _{3'}	3.17	32.64	3609.19	11768.71	-3.55	
O ₅	3.17	32.1	4381.77	14992.89	-3.78		C _{4'}	2.63	103.18	11032.03	35940.57	1.78	
C ₆	2.63	91.58	11208.16	37732.58	0.69		O _{4'}	3.17	32.64	3390.69	11032.03	-3.45	
C ₇	2.63	100.55	11531.53	38369.26	1.46		C _{5'}	2.63	85.78	9749.14	32177.88	0.48	
O ₇	3.17	29.47	3447.28	11531.53	-3.76		C _{6'}	2.63	90.5	11396.71	37676.24	0.29	
C ₈	2.63	94.21	12182.29	41344.28	0.66		total	58.47	1878.61	238008.4	801923.1	4.48	4.48
Compound 15													
O ₁	3.17	72.63	10379.32	36276.94	-1.41	I = -0.46	C ₈	2.63	94.21	12389.21	42081.23	0.58	II = 2.54
C ₂	2.63	132.18	18171.35	59941.18	1.98		C ₉	2.63	143.17	20454.81	70694.04	3.11	
C ₃	2.63	118.96	16336.37	58230.66	2.51		C ₁₀	2.63	131.1	18105.59	61328.93	2.34	
O ₃	3.17	36.35	5001.47	16336.37	-3.77		C _{1'}	2.63	127.86	16889.12	57444.37	2.56	III = 1.96
C ₄	2.63	126.99	16886.49	55523.18	1.92		C _{2'}	2.63	90.5	11191.95	37134.28	0.42	
O ₄	3.17	64.2	9366	33772.98	-1.69		C _{3'}	2.63	82.61	9053.21	29652.48	0.37	
C ₅	2.63	117.41	15247.15	51469.81	1.94	II = 2.54	C _{4'}	2.63	94.21	9407.32	30434.77	1.45	
O ₅	3.17	32.1	4458.32	15247.15	-3.81		O _{4'}	3.17	29.47	2921.03	9407.32	-3.62	
C ₆	2.63	91.58	11309.53	38143.11	0.66		C _{5'}	2.63	82.61	9053.21	29652.48	0.37	
C ₇	2.63	100.55	11586.43	38745.13	1.49		C _{6'}	2.63	90.5	11191.95	37134.28	0.42	
O ₇	3.17	29.47	3459.96	11586.43	-3.76		total	58.47	1888.66	242859.8	820237.1	4.05	4.05

^a Results from eq 14 for the negative log of the molar concentration of compound that caused 50% inhibition of HIV-1 IN 3'-processing activity.

molecule, provided that the molecular environments are similar. That is to say, the atom or fragment contribution of several properties of molecular fragments is approximately "transferable".

Finally, the features of the formalism presented here are very much in accord with the ideas expressed by Milne,² who raised the challenge to consecrate more energy to use the QSAR model to address the *inverse problem*, also referred to as the following: graph reconstruction, inverse imaging, design of molecules from QSAR, back-projection, molecule building from QSAR equation, inverse structure generation, etc. It consists of the computational generation of candidate chemical structures and their selection according to a previously established QSAR model.¹ That is, when a good QSAR equation is developed, then this mathematical equation can be inverted to indicate those structures that would possess a specific range of property/activity values. The computational implementation in TOMOCOMD-CARDD

software and the applications of this approach to constructing a better drug-like molecule are now in progress and will be the subject of a future publication.

5. CONCLUDING REMARKS

The intensive interest of chemists in the creation of novel molecular indices to characterize the molecular structure of drugs is driven by the main paradigm in medicinal chemistry: "Biological activity of organic compounds depends on their molecular structure".⁷² Despite the great advances in the field of theoretical drug design, even today this paradigm is guiding the discovery of new lead compounds.¹⁰ Thus, the continuous definition of novel molecular descriptors that could explain different pharmacological properties by means of QSAR is necessary. Consequently, total and local (atom and atom-type) linear indices were proposed and used to the prediction of pIC₅₀ values for cleavage process of a set of flavone derivatives inhibitors of HIV-1 IN. Quantitative

Table 13. Equations for the Prediction the HIV-1 IN 3'-Processing Activity Using Atom Linear Indices in Compound 7

$-\text{Log(IC}_{50})_{O1} = (-39.9/21) - 1.179 f_0(x_{O1}) + 0.1165 f_3(x_{O1}) - 0.00139 f_7(x_{O1}) + 0.00028 f_8(x_{O1}) = -1.54M$
$-\text{Log(IC}_{50})_{C2} = (-39.9/21) - 1.179 f_0(x_{C2}) + 0.1165 f_3(x_{C2}) - 0.00139 f_7(x_{C2}) + 0.00028 f_8(x_{C2}) = 1.34M$
$-\text{Log(IC}_{50})_{C3} = (-39.9/21) - 1.179 f_0(x_{C3}) + 0.1165 f_3(x_{C3}) - 0.00139 f_7(x_{C3}) + 0.00028 f_8(x_{C3}) = 1.34M$
$-\text{Log(IC}_{50})_{C4} = (-39.9/21) - 1.179 f_0(x_{C4}) + 0.1165 f_3(x_{C4}) - 0.00139 f_7(x_{C4}) + 0.00028 f_8(x_{C4}) = 1.41M$
$-\text{Log(IC}_{50})_{O4} = (-39.9/21) - 1.179 f_0(x_{O4}) + 0.1165 f_3(x_{O4}) - 0.00139 f_7(x_{O4}) + 0.00028 f_8(x_{O4}) = -1.61M$
$-\text{Log(IC}_{50})_{C5} = (-39.9/21) - 1.179 f_0(x_{C5}) + 0.1165 f_3(x_{C5}) - 0.00139 f_7(x_{C5}) + 0.00028 f_8(x_{C5}) = 1.92M$
$-\text{Log(IC}_{50})_{O5} = (-39.9/21) - 1.179 f_0(x_{O5}) + 0.1165 f_3(x_{O5}) - 0.00139 f_7(x_{O5}) + 0.00028 f_8(x_{O5}) = -3.78M$
$-\text{Log(IC}_{50})_{C6} = (-39.9/21) - 1.179 f_0(x_{C6}) + 0.1165 f_3(x_{C6}) - 0.00139 f_7(x_{C6}) + 0.00028 f_8(x_{C6}) = 0.69M$
$-\text{Log(IC}_{50})_{C7} = (-39.9/21) - 1.179 f_0(x_{C7}) + 0.1165 f_3(x_{C7}) - 0.00139 f_7(x_{C7}) + 0.00028 f_8(x_{C7}) = 1.46M$
$-\text{Log(IC}_{50})_{O7} = (-39.9/21) - 1.179 f_0(x_{O7}) + 0.1165 f_3(x_{O7}) - 0.00139 f_7(x_{O7}) + 0.00028 f_8(x_{O7}) = -3.76M$
$-\text{Log(IC}_{50})_{C8} = (-39.9/21) - 1.179 f_0(x_{C8}) + 0.1165 f_3(x_{C8}) - 0.00139 f_7(x_{C8}) + 0.00028 f_8(x_{C8}) = 0.66M$
$-\text{Log(IC}_{50})_{C9} = (-39.9/21) - 1.179 f_0(x_{C9}) + 0.1165 f_3(x_{C9}) - 0.00139 f_7(x_{C9}) + 0.00028 f_8(x_{C9}) = 3.24M$
$-\text{Log(IC}_{50})_{C10} = (-39.9/21) - 1.179 f_0(x_{C10}) + 0.1165 f_3(x_{C10}) - 0.00139 f_7(x_{C10}) + 0.00028 f_8(x_{C10}) = 2.31M$
$-\text{Log(IC}_{50})_{C1'} = (-39.9/21) - 1.179 f_0(x_{C1'}) + 0.1165 f_3(x_{C1'}) - 0.00139 f_7(x_{C1'}) + 0.00028 f_8(x_{C1'}) = 2.52M$
$-\text{Log(IC}_{50})_{C2'} = (-39.9/21) - 1.179 f_0(x_{C2'}) + 0.1165 f_3(x_{C2'}) - 0.00139 f_7(x_{C2'}) + 0.00028 f_8(x_{C2'}) = 0.88M$
$-\text{Log(IC}_{50})_{C3'} = (-39.9/21) - 1.179 f_0(x_{C3'}) + 0.1165 f_3(x_{C3'}) - 0.00139 f_7(x_{C3'}) + 0.00028 f_8(x_{C3'}) = 1.85M$
$-\text{Log(IC}_{50})_{O3'} = (-39.9/21) - 1.179 f_0(x_{O3'}) + 0.1165 f_3(x_{O3'}) - 0.00139 f_7(x_{O3'}) + 0.00028 f_8(x_{O3'}) = -3.55M$
$-\text{Log(IC}_{50})_{C4'} = (-39.9/21) - 1.179 f_0(x_{C4'}) + 0.1165 f_3(x_{C4'}) - 0.00139 f_7(x_{C4'}) + 0.00028 f_8(x_{C4'}) = 1.78M$
$-\text{Log(IC}_{50})_{O4'} = (-39.9/21) - 1.179 f_0(x_{O4'}) + 0.1165 f_3(x_{O4'}) - 0.00139 f_7(x_{O4'}) + 0.00028 f_8(x_{O4'}) = -3.45M$
$-\text{Log(IC}_{50})_{C5'} = (-39.9/21) - 1.179 f_0(x_{C5'}) + 0.1165 f_3(x_{C5'}) - 0.00139 f_7(x_{C5'}) + 0.00028 f_8(x_{C5'}) = 0.48M$
$-\text{Log(IC}_{50})_{C6'} = (-39.9/21) - 1.179 f_0(x_{C6'}) + 0.1165 f_3(x_{C6'}) - 0.00139 f_7(x_{C6'}) + 0.00028 f_8(x_{C6'}) = 0.29M$

models found are significant from a statistical point of view and permits a clear interpretation of the studied properties in terms of the structural features of molecules. The observation that the total, atom-type and atom linear indices approach performs comparably to the E-state and the CoMFA methods for this data set is interesting in that the latter is more computationally intensive than the former.

In addition, evidence of significant information is presented in this paper in several ways. The variation of the k^{th} total and local linear indices values with alkyl-chain lengthening, branching, heteroatoms-content, and multiple bonds agrees with usual organic intuition. The principal component analysis presented in this work indicated that the information carried by local linear indices is markedly different from that codified in various 0D, 1D, 2D, and 3D molecular descriptors presently in QSPR/QSAR and drug design practice. On the contrary, much redundancy and overlapping was found among total linear indices and most of the other structural indices used in this study. The approach described here appears to be a very promising structural invariant, useful for QSPR/QSAR studies and shown to provide an excellent alternative or guides for discovery and optimization of new lead compounds, reducing the time and cost of traditional procedure.

ACKNOWLEDGMENT

I would like to offer my sincere thanks to the two anonymous referees for their critical opinions about the manuscript, which have significantly contributed to improving its presentation and quality. I am also indebted to the manuscript's Editor, Dr. A. J. Hopfinger (U.S.A.), for his kind attention.

Supporting Information Available: The complete list of 0D–3D DRAGON's molecular descriptors³¹ used in this study as well as their description and loadings (from factor analysis) This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) de Julián-Ortiz, J. V. Virtual Darwinian Drug Design: QSAR Inverse Problem, Virtual Combinatorial Chemistry, and Computational Screening. *Comb. Chem. High Throughput Screen.* **2001**, *4*, 295–310.
- (2) Milne, G. W. A. Mathematics as a Basis for Chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 639–644.
- (3) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Germany, 2000.
- (4) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.
- (5) Katritzky, A. R.; Gordееva, E. V. Traditional Topological Indexes vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835–857.
- (6) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: New York, 1999.
- (7) Balaban, A. Topological and Stereochemical Molecular Descriptors for Databases Useful in QSAR, Similarity/Dissimilarity and Drug Design. *SAR QSAR Environ. Res.* **1998**, *8*, 1–21.
- (8) Estrada, E. On the Topological Sub-Structural Molecular Design (TOSS-MODE) in QSPR/QSAR and Drug Design Research. *SAR QSAR Environ. Res.* **2000**, *11*, 55–73.
- (9) Julián-Ortiz, J. V.; Gálvez, J.; Muñoz-Collado, C.; García-Domenech, R.; Gimeno-Cardona, C. Virtual Combinatorial Synthesis and Computational Screening of New Potential Anti-Herpes Compounds. *J. Med. Chem.* **1999**, *42*, 3308–3314.
- (10) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1699–1714.
- (11) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999.
- (12) Randic, M. In *Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Ed.; John Wiley & Sons: New York, 1998; Vol. 5, pp 3018–3032.
- (13) Rouvray, D. H. In *Mathematical and Computational Concepts in Chemistry*; Trinajstić, N., Ed.; Ellis Horwood: Chichester, 1986; pp 295–306.
- (14) *From Chemical Graphs to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997.
- (15) *QSPR/QSAR Studies by Molecular Descriptors*; Diudea, M. V., Ed.; Nova Science, Huntington: New York, 2001.
- (16) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. Evaluation in Quantitative Structure–Property Relationship Models of Structural Descriptors Derived from Information-Theory Operators. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 631–643.
- (17) Balaban, T.; Mills, D.; Ivanciuc, O.; Basak, S. C. Reverse Wiener Indices. *Croat. Chem. Acta* **2000**, *73*, 923–941.
- (18) Ivanciuc, O.; Ivanciuc, T.; Klein, D. J.; Seitz, W. A.; Balaban, A. T. Wiener Index Extension by Counting Even/Odd Graph Distances. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 536–549.
- (19) Ivanciuc, O. Building-Block Computation of the Ivanciuc-Balaban Indices for the Virtual Screening of Combinatorial Libraries. *Internet Electron. J. Mol. Des.* **2002**, *1*, 1–9, <http://www.biochempress.com>.
- (20) Marino, D. J. G.; Peruzzo, P. J.; Castro, E. A.; Toropov, A. QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants. *Internet Electron. J. Mol. Des.* **2002**, *1*, 115–133, <http://www.biochempress.com>.
- (21) Marrero-Ponce, Y. Total and Local Quadratic Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds. *Molecules* **2003**, *8*, 687–726. <http://www.mdpi.org>

- (22) Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; Montero, L. A. Total and Local Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix". Application to Prediction of Caco-2 Permeability of Drugs. *Int. J. Mol. Sci.* **2003**, *4*, 512–536. www.mdpi.org/ijms/
- (23) Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. A New Topological Descriptors Based Model for Predicting Intestinal Epithelial Transport of Drugs in Caco-2 Cell Culture. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186–199.
- (24) Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. Nucleic Acid Quadratic Indices of the "Macromolecular Graph's Nucleotides Adjacency Matrix". Modeling of Footprints after the Interaction of Paromomycin with the HIV-1 Ψ -RNA Packaging Region. *CPS: physchem/0401004*.
- (25) Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. 3D-Chiral Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix" and their Application to Central Chirality Codification: Classification of ACE Inhibitors and Prediction of σ -Receptor Antagonist Activities. *Bioorg. Med. Chem.* DOI: 10.1016/j.bmc.2004.07.051
- (26) Cotton, F. A. *Advanced Inorganic Chemistry*; Ed. Revolucionaria: Havana, Cuba, p 103.
- (27) Browder, A. *Mathematical Analysis. An Introduction*. Springer-Verlag: New York, 1996; pp 176–296.
- (28) Axler, S. *Linear algebra Done Right*; Springer-Verlag: New York, 1996; pp 37–70.
- (29) Ross, K. A.; Wright, C. R. B. *Matemáticas discretas*; Prentice Hall Hispanoamericana: Mexico, 1990.
- (30) Maltsev, A. I. *Fundamentos del Álgebra Lineal*; Mir: Moscuw, 1976; pp 68–262.
- (31) Todeschini, R.; Consonni, V.; Pavan, M. Dragon. Software version 2.1. 2002. <http://www.disat.unimib.it/chm/Dragon.htm>.
- (32) De Clercq, E. New Developments in Anti-HIV Chemotherapy. *Curr. Med. Chem.* **2001**, *8*, 1543–1572.
- (33) De Clercq, E. Toward Improved Anti-HIV Chemotherapy: Therapeutic Strategies for Intervention with HIV Infections. *J. Med. Chem.* **1995**, *38*, 2491–2517.
- (34) Varmus, H.; Brown, P. Retroviruses. In *Mobile DNA*; Berg, D., Howe, M., Eds.; American Society of Microbiology: Washinton, DC, 1989; p 53.
- (35) Craigie, R. HIV Integrase, a Brief Overview from Chemistry to Therapeutics. *J. Biol. Chem.* **2001**, *276*, 23213–23216.
- (36) Fesen, M. R.; Kohn, K. W.; Leutertre, F.; Pommier, Y. Inhibitors of Human Immunodeficiency Virus Integrase. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 2399–2403.
- (37) Fesen, R.; Pommier, Y.; Leutertre, F.; Hiroguchi, S.; Yung, J.; Kohn, K. W. Inhibition of HIV-1 Integrase by Flavones, Caffeic Acid Phenethyl Ester (CAPE) and Related Compounds. *Biochem. Pharmacol.* **1994**, *48*, 595–608.
- (38) Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801–807.
- (39) Boulamwini, J. K.; Raghavan, K.; Fesen, M. R.; Pommier, Y.; Kohn, K. W.; Weinstein, J. N. Application of the Electrotopological State Index to QSAR Analysis of Flavone Derivates as HIV-1 Integrase Inhibitors. *Pharm. Res.* **1996**, *13*, 1892–1895.
- (40) Marrero-Ponce, Y.; Romero, V. *TOMOCOMD* software. Central University of Las Villas. 2002. *TOMOCOMD* (TOPOlogical MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version will be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es
- (41) Randic, M. Generalized Molecular Descriptors. *J. Math. Chem.* **1991**, *7*, 155–168.
- (42) Cramer, R. D., III. BC(DEF) Parameters. 2. An Empirical Structure Based Scheme for the Prediction of Some Physical Properties. *J. Am. Chem. Soc.* **1980**, *102*, 1849–1859.
- (43) Cramer, R. D., III. BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State. *J. Am. Chem. Soc.* **1980**, *102*, 1837–1849.
- (44) Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular Modeling of the Physical Properties of Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186–4194.
- (45) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley-Interscience: New York, 1980.
- (46) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984; pp 197–188.
- (47) Estrada, E.; González, H. What Are the Limits of Applicability for Graph Theoretic Descriptors in QSPR/QSAR? Modeling Dipole Moments of Aromatic Compounds with TOPS-MODE Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75–84.
- (48) Estrada, E.; Rodríguez, L. Edge-Connectivity Indices in QSPR/QSAR Studies. 1. Comparison to Other Topological Indices in QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1037–1041.
- (49) Estrada, E. Edge-Connectivity Indices in QSPR/QSAR Studies. 2. Accounting for Long-Range Bond Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1042–1048.
- (50) González-Díaz, H.; Marrero-Ponce, Y.; Hernández, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, U.; Castañedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Pérez, M. 3D-MEDNES: An Alternative "In Silico" Technique for Chemical Research in Toxicology. 1. Prediction of Chemically Induced Agranulocytosis. *Chem. Res. Toxicol.* **2003**, *16*, 1318–1327.
- (51) STATISTICA ver. 5.5, Statsoft, Inc. 1999.
- (52) Goldberg, D. E. *Genetic Algorithms*; Addison-Wesley, Reading, MA, 1989.
- (53) Willet, P. Genetic Algorithms in Molecular Recognition and Design. *Trends Biotechnol.* **1995**, *13*, 516–521.
- (54) So, S. S.; Karplus, M. Evolutionary Optimization in Quantitative Structure–Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521–1530.
- (55) So, S. S.; Karplus, M. Three-Dimensional Quantitative Structure–Activity Relationship from Molecular Similarity Matrices and Genetic Neural Networks. *J. Med. Chem.* **1997**, *40*, 4347–4359.
- (56) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure–Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- (57) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- (58) Senese, C. L.; Hopfinger, A. J. Receptor-Independent 4D-QSAR Analysis of a Set of Norstatine Derived Inhibitors of HIV-1 Protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1297–1307.
- (59) Liu, J.; Pan, D.; Tseng, Y.; Hopfinger, A. J. 4D-QSAR Analysis of a Series of Antifungal P450 Inhibitors and 3D-Pharmacophore Comparisons as a Function of Alignment. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2170–2179.
- (60) Senese, C. L.; Hopfinger, A. J. A Simple Clustering Technique to Improve QSAR Model Selection and Predictivity: Application to a Receptor Independent 4D-QSAR Analysis of Cyclic Urea Derived Inhibitors of HIV-1 Protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2180–2193.
- (61) De Oliveira, D. B.; Gaudio, A. C. BuildQSAR: A New Computer Program for QSAR Studies. *Quant. Struct.-Act. Relat.* **2000**, *19*, 599–601.
- (62) Wold, S.; Erikson, L. *Statistical Validation of QSAR Results. Validation Tools*; In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers: New York, 1995; pp 309–318.
- (63) Klein, D. J. Graph Theoretically Formulated Electronic-Structure Theory. *Internet Electron. J. Mol. Des.* **2003**, *2*, 814–834, <http://www.biochempress.com>.
- (64) Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-Additive Method. *Perspect. Drug Discov. Des.* **2000**, *19*, 47–66.
- (65) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
- (66) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure–Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.
- (67) Millar, K. J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.
- (68) Gasteiger, J.; Marsilli, M. A New Model for Calculating Atomic Charge in Molecules. *Tetrahedron Lett.* **1978**, *34*, 3181–3184.
- (69) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; Wiley: New York, 1980.
- (70) Raghavan, K.; Buolamwini, J. K.; Fesen, M. R.; Pommier, Y.; Kohn, K. W.; Weinstein, J. N. Three-Dimensional Structure–Activity Relationship (QSAR) of HIV Integrase Inhibition: A Comparative Molecular Field Analysis (CoMFA). *J. Med. Chem.* **1995**, *38*, 890–897.
- (71) Golbraikh, A.; Tropsha, A. Beware of q^2 ! *J. Mol. Graphic Modell.* **2002**, *20*, 269–276.
- (72) Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative Structure–Property Relationships in Pharmaceutical Research-Part 1. *Pharm. Sci. Technol. Today.* **2000**, *3*, 28–35.]